Linear Models

Chiranjit Mukhopadhyay
Indian Institute of Science

1 Introduction

All the models that we will fit are linear models. They are called linear because they are linear in parameters. That is for instance even if we fit a model of the type $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$ with $\epsilon \sim N(0, \sigma^2)$, it will be called a linear model. In this model though we are expressing a quadratic, and thus non-linear, relationship between X and Y, since the model is linear in the model parameters β_0, β_1 and β_2, it will be called a linear model.

The types of linear models we will deal with here will essentially fall in three categories: Analysis of Variance (ANOVA) Models, Regression Models and Analysis of Covariance (ANCOVA) Models. Strictly speaking all these models are basically Linear Regression models. But we shall make the following distinctions. When all the factors are qualitative (or they may be quantitative, but the quantitative nature of their values are not exploited in the model) and only the nature of their effects (main effect/interactions) are considered, then we shall refer to such models as ANOVA models. When the quantitative natures of the factors are exploited by explicitly accounting for them as some quantitative variable in the model equation, we shall refer to such models as regression models. And when we have both intrinsically qualitative (and thus the numerical labels we associate with their levels do not have any quantitative meaning) and quantitative factors then we use a special type of regression model with so-called dummy variables, which will be referred to as ANCOVA models.

2 ANOVA Models

Here we have a quantitative response Y and one or more qualitative factors. The different values assumed by a quantitative factor do not explicitly appear into the model, rather they are treated the same way as the levels of a pure qualitative factor is treated, namely just as distinct levels of the quantitative factor. For a basic understanding of analysis of ANOVA models we first begin with the so-called one-way analysis of variance in which we just compare different treatment means. A treatment in general is of course a combination of different levels of the factors under consideration, but for understanding one-way ANOVA, it is better to think of it as an analysis of response when we have only one factor assuming multiple levels. The factorial structure of the treatment and thus enhancement of the one-way ANOVA analysis will be taken up next, which as a pre-requisite requires an understanding of the one-way ANOVA.

2.1 One-Way ANOVA

As mentioned above, for the sake of simplicity, assume that we have only one factor X, which takes on values at multiple levels, or to be concrete let X have k levels $1, 2, \ldots, k$ [1] with $k \geq 2$. Also assume that we have n_i observations on response Y for the i-th level of X or when X takes the value i, $i = 1, 2, \ldots, k$. Let these n_i response value be denoted by $Y_{i1}, Y_{i2}, \ldots, Y_{ik}$. We shall assume that $Y_i \sim N(\mu_i, \sigma^2)$, where Y_i denotes the (population) response variable when $X = i$. Sometimes the μ_i's are written as $\mu_i = \mu + \alpha_i$ with $\sum_{i=1}^{k} \alpha_i = 0$, to be consistent with the general factorial ANOVA models of which the one-way ANOVA model may be thought of as the one of first order. This is the probability model we put forth for analyzing the responses $Y_{11}, Y_{12}, \ldots, Y_{1n_1}; Y_{21}, Y_{22}, \ldots, Y_{2n_2}; \ldots, Y_{k1}, Y_{k2}, \ldots, Y_{kn_k}$. Note that according to this formulation, it is a straight forward generalization of §3.3 of Session 2 notes to k Normal populations from two Normal populations. Also note that, barring the Normality and Homoscedasticity assumption this is the most general model one can put forth for a response for k levels of a factor X. Normality may be informally (tested) using NPP plots of the observations for the k Y_i's. However as in Case III of §3.3 of Session 2 notes, here also we first require to validate the homoscedasticity assumption in this k-population problem, before we can proceed any further.

Test for Homoscedasticity (Bartlett's Test):

The test for homoscedasticity in this general k-sample case is slightly more complex than the corresponding F-test in the two-sample problem. However as in Case III of §3.3 of Session 2 notes, here also we are interested in testing the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$, where $V[Y_i] = \sigma_i^2$ for $i = 1, 2, \ldots, k$. Let s_i^2 denote the sample variance in the i-th sample, $i = 1, 2, \ldots, k$. That is for $i = 1, 2, \ldots, k$ let $s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)^2$, where $\overline{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$ denotes the i-th sample mean. If the null hypothesis of homoscedasticity is true, then we shall expect $s_1^2, s_2^2, \ldots, s_k^2$ to be all close to one another. One way to measure this closeness would be to compare the Arithmetic Mean (AM) to the Geometric Mean (GM) of $s_1^2, s_2^2, \ldots, s_k^2$. It is well-known that in general AM of k numbers is always $\geq$ their GM, with equality following if and only if all the k numbers are identical to one another. Thus Bartlett's test statistics for checking for homoscedasticity of k Normal populations is given by $B = \frac{n-k}{C} \log_e \left(\frac{AM_s}{GM_s} \right)$, where AM_s and GM_s respectively denote the AM and GM of $s_1^2, s_2^2, \ldots, s_k^2$, $n = \sum_{i=1}^{k} n_i$ the total sample size, and $C = 1 + \frac{1}{3(k-1)} \left(\sum_{i=1}^{k} \frac{1}{n_i - 1} - \frac{1}{n-k} \right)$. Note that since s_i^2 is computed based on $(n_i - 1)$ and in general $n_1 \neq n_2 \neq \ldots \neq n_k$ the appropriate way of computing AM_s and GM_s would be using weighted means with weights $n_1 - 1, n_2 - 1, \ldots, n_k - 1$. Thus $AM_s = \frac{1}{n-k} \sum_{i=1}^{k} (n_i - 1) s_i^2$ and $GM_s = \left[\prod_{i=1}^{k} (s_i^2)^{n_i - 1} \right]^{\frac{1}{n-k}}$.

If H_0 is true, AM_s and GM_s would be close to each other yielding a small positive value for B (in the extreme case of $s_1^2 = s_2^2 = \cdots = s_k^2$, $AM_s = GM_s$ resulting in $B = 0$, which is the smallest possible value of B). Thus we should reject H_0, the null hypothesis of homoscedasticity, for "large" values of B. The answer to this now familiar question of how

[1] Note that though we are using positive integers for denoting the levels of X, the values of these integers themselves as numbers will never appear in the subsequent analysis. They are simply treated as notationally convenient labels for distinguishing the distinct levels of X.

"large" is "large" is given by the sampling distribution of B under H_0, which can be shown to have *approximately* a χ^2_{k-1} distribution for large samples. Thus the decision rule would be to reject H_0 if $B > \chi^2_{k-1,1-\alpha}$ or p-value$=P(\chi^2_{k-1} > B)$.

Suppose H_0 above is not rejected. Then we are in an easy to interpret case (analogous to Case III of §3.3 of Session 2 notes) and next we are to decide whether the means of Y for the k levels of X are different from one another. That is we now need to test the main ANOVA null hypothesis of interest $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$. The logic of this test is as explained below.

Under the ANOVA null hypothesis, $\overline{Y}_1, \overline{Y}_2, \ldots, \overline{Y}_k$ should be close to each other. For location parameters like mean, from now on, the closeness will be measured using the standard measure called variability. Simple variability of k numbers $x_1, x_2, \ldots, x_k$ is given by $\sum_{i=1}^k (x_i - \overline{x})^2$ where $\overline{x} = \frac{1}{k}\sum_{i=1}^k x_i$. If $x_1, x_2, \ldots, x_k$ are close to each other their variability would be small (=0 in case they are all same). However since our $\overline{Y}_1, \overline{Y}_2, \ldots, \overline{Y}_k$ are measured respectively using $n_1, n_2, \ldots, n_k$ observations, their variability is measured using the weights $n_1, n_2, \ldots, n_k$. That is the variability *between* the group means is given by $\sum_{i=1}^k n_i(\overline{Y}_i - \overline{Y})^2$ where $\overline{Y} = \frac{1}{n}\sum_{i=1}^k n_i \overline{Y}_i$ is the overall grand mean. The quantity $\sum_{i=1}^k n_i(\overline{Y}_i - \overline{Y})^2$ is typically denoted by SSB (Sum of Squares Between groups) or $SSTr$ (Sum of Squares due to Treatments).

As mentioned above, under the ANOVA null hypothesis thus we should expect to see a "small" value of SSB. For determining this "small" value we refer to the result that, under $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$, $\frac{SSB}{\sigma^2} \sim \chi^2_{k-1}$ where σ^2 is the common (across the k groups and possibly unknown for most practical cases) value of the variance $\sigma^2 = \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$. Thus if σ^2 were known we could employ a χ^2-test for testing the ANOVA null hypothesis $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ by rejecting H_0 for $SSB/\sigma^2 > \chi^2_{k-1,1-\alpha}$ or by computing the p-value$=P(\chi^2_{k-1} > SSB/\sigma^2)$.

However since σ^2 would be unknown for most practical applications, we shall need to replace it by its estimate. We estimate this common variance σ^2 using the same approach adopted in Case III of §3.3 of Session 2 notes. That is here also the UMVUE of σ^2 is given by $\frac{\sum_{i=1}^k (n_i-1)s_i^2}{\sum_{i=1}^k (n_i-1)} = \frac{1}{n-k}\sum_{i=1}^k (n_i-1)s_i^2$, which is same as the pooled variance s_p^2 of Case III of §3.3 of Session 2 notes when $k = 2$. The quantity $\sum_{i=1}^k (n_i-1)s_i^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)^2$ is denoted by SSW (Sum of Squares Within groups) or SSE (Sum of Squares due to Error). This is because each Y_{ij} may be thought of being modeled as $Y_{ij} = \mu_i + \epsilon_{ij}$ with $\epsilon_{ij} \sim N(0, \sigma^2)$ and $\overline{Y}_i$ being an estimate of μ_i, $Y_{ij} - \overline{Y}_i$ may be thought of as the deviation from the estimated model mean and thus a proxy for the error ϵ_{ij}.

Now as before, $\frac{SSE}{\sigma^2}$ may be shown to have a χ^2 distribution with $n-k$ d.f. Thus replacing σ^2 by its estimate in the test statistic $\frac{SSTr}{\sigma^2}$ for the ANOVA hypothesis, results in the refined test statistic $\frac{SSTr/(k-1)}{SSE/(n-k)}$ which after dividing both the numerator and denominator by σ^2 results in an F statistic with $k-1$ and $n-k$ as the respective numerator and denominator d.f. under H_0. Thus the test of the ANOVA null hypothesis $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ may be described as follows. Reject H_0 if $\frac{SSTr/(k-1)}{SSE/(n-k)} > F_{k-1,n-k,1-\alpha}$ or compute p-value$=P(F_{k-1,n-k} > \frac{SSTr/(k-1)}{SSE/(n-k)})$.

Apart from the above distributional logic, the F-statistic $\frac{SSTr/(k-1)}{SSE/(n-k)}$ can also be under-

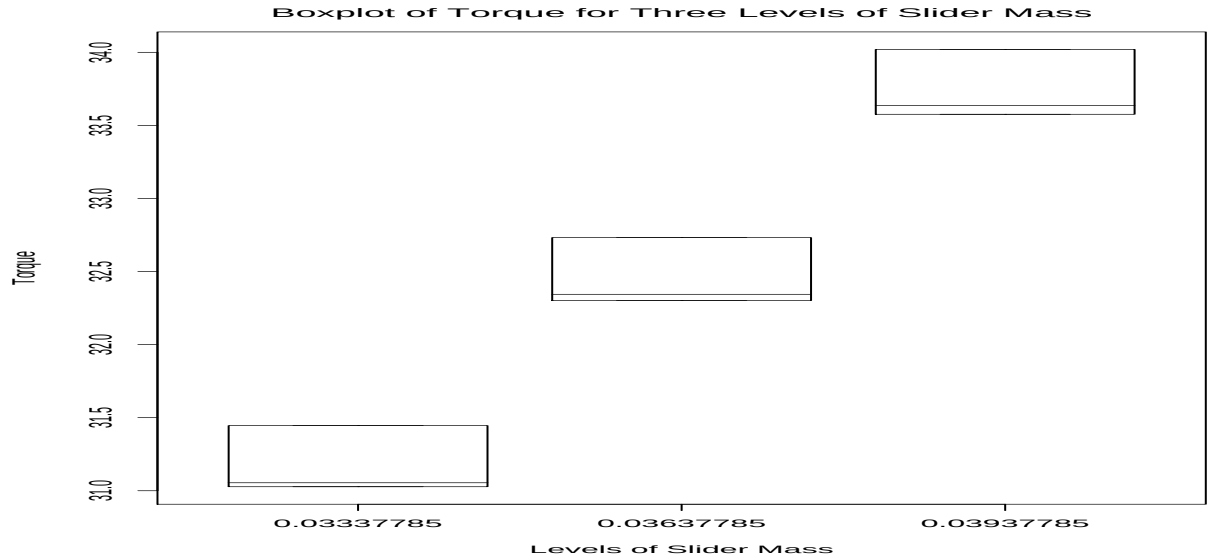
stood from an intuitive point of view as follows. $SSTr/(k-1)$ or $SSB/(k-1)$ gives the average amount of variability that exists *between* the group means $\overline{Y}_1, \overline{Y}_2, \ldots, \overline{Y}_k$, while $SSE/(n-k)$ or $SSW/(n-k)$ gives the average amount of variability that exists *within* each group (variability within the i-th group consisting of the observations $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$ being measured by $\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_i)^2$). Now if the variability between the groups is "large" compared to the variability within the groups, then the k sample means must be *significantly* different from each other, and the question of how "large" is "large" is settled by the $F_{k-1,n-k}$ distribution, which is the sampling distribution of the test statistic $\frac{SSTr/(k-1)}{SSE/(n-k)}$ under H_0. leading to the test procedure or decision rule explained above.

All these computations are typically presented using an Analysis of Variance Table as follows:

ANOVA Table

Source of Variation	$D.F.$	SS	$MSS = SS/D.F.$	F	p-value
Treatment	$k-1$	$SSTr = \sum_{i=1}^{k} n_i(\overline{Y}_i - \overline{Y})^2$	$MSTr = \frac{SSTr}{k-1} = \frac{1}{k-1}\sum_{i=1}^{k} n_i(\overline{Y}_i - \overline{Y})^2$	$F_{obs} = \frac{MSTr}{MSE}$	$P(F_{k-1,n-k} > F_{obs})$
Error	$n-k$	$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2$	$MSE = \frac{SSE}{n-k} = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2$		
Total	$n-1$	$SSTotal = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y})^2$			

Example 3: In an experiment it was of interest to determine whether varying the value of the quantitative variable Slider Mass, say `sm`, over its three possible levels 0.03337785, 0.03637785 and 0.03937785, changes the resulting response of Torque on Crank Slider Mechanism, denoted by `torque`. As in Example 2, we begin with the descriptive analysis of box-plots.



As can be seen from the above plot that the `torque` values indeed change for changing values of `sm`. This is formally justified by the following ANOVA analysis. (NPP were not made

because though there are 9 observations on `torque` for each level of `sm`, there are only three distinct values of `torque` making the NPP redundant.)

```
> anova(aov(torque~sm))
Analysis of Variance Table

Response: torque
          Df  Sum Sq Mean Sq F value    Pr(>F)
sm         2 29.6704 14.8352  350.51 < 2.2e-16 ***
Residuals 24  1.0158  0.0423
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.2 Two-Factor ANOVA

Now suppose we have two factors X_1 and X_2 having respectively say s_1 and s_2 levels. Thus we have $l = s_1 \times s_2$ treatments. Also suppose each of these l treatments is replicated r times. [2] Thus let Y_{ijk} denote the k-th response wits X_1 at level i and X_2 at level j, $i = 1, 2, \ldots, s_1$. $j = 1, 2, \ldots, s_2$ and $k = 1, 2, \ldots, r$. We shall assume that $Y_{ijk} \sim N(\mu_{ij}, \sigma^2)$ with μ_{ij} having the structure $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$, with $\sum_{i=1}^{s_1} \alpha_i = 0$, $\sum_{j=1}^{s_2} \beta_j = 0$, $\sum_{i=1}^{s_1} \gamma_{ij} = 0 \ \forall j = 1, 2, \ldots, s_2$ and $\sum_{j=1}^{s_2} \gamma_{ij} = 0 \ \forall i = 1, 2, \ldots, s_1$. This is the full-factorial model with two factors. Note that with the above constraints there are exactly $s_1 s_2$ many free parameters in total which are equivalent to the $s_1 s_2$ many μ_{ij}'s. Thus the full-factorial model is a general model except it writes each mean in terms of the main effects and interaction between the two factors.

The first thing one checks is whether the two factors indeed interact with one another. That is whether the effect of X_1 depends on the level of X_2. This is because, in case there are interactions, optimality of a factor in terms of one of its levels may be meaningless. One level of X_1 might be the best when the level of X_2 is say j_0, and another level of X_1 might be the best when the level of X_2 is say j_0'. In presence of interaction, optimality is typically expressed in terms of the best treatment or combination of levels of the two factors. On the other hand, if there is no interaction, life is very simple and in this case one can meaningfully talk about the best levels for each individual factor.

In terms of the model parameters the null hypothesis of no interaction effect is expressed as $H_{0\gamma}: \begin{array}{ccccc} \gamma_{11} = \gamma_{12} = \cdots = \gamma_{1s_2} = 0 \\ \gamma_{21} = \gamma_{22} = \cdots = \gamma_{2s_2} = 0 \\ \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \\ \gamma_{s_1 1} = \gamma_{s_1 2} = \cdots = \gamma_{s_1 s_2} = 0 \end{array}$. All hypotheses of such types and also for the main-effects in the ANOVA model are tested using F-tests. The logic of these F-tests is exactly same as the one-way ANOVA F-test discussed above in §2.1. We find an appropriate variability or SS (Sum of Squares), say SSH_0, that one expects to be small (ideally 0) under the null hypothesis. These SSH_0/σ^2 typically have χ^2 distributions under the null hypothesis

[2] If the treatments are replicated unequal number of times, then the resulting design is called unbalanced. We shall only deal with balanced designs here thus it assumed that each treatment has been replicated an equal r number of times.

of interest, where σ^2 is the common inherent variability parameter of the model. That is even if two experimental units receive the same treatment, one cannot expect to observe the same response for the two units. This inherent variability of the experimental units (which is assumed to be the same regardless of the kind of treatment an experimental unit receives - the homoscedasticity assumption) is parameterized using σ^2. Thus for known σ^2 one could do a χ^2 test for all such hypotheses. However as σ^2 is never known, it is replaced by its estimate. The UMVUE of σ^2 also has a standard common structure across the board for all linear models, and thus in particular for the ANOVA models as well. Whatever way the data may be classified (like one, two or multiple way) if $Y_1, Y_2, \ldots, Y_n$ denote the observed responses and $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n$ denote the corresponding predicted responses using the fitted model, σ^2 is always naturally estimated as $SSE/e.d.f.$, where the Sum of Squares due to Errors or SSE is given by $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ and $e.d.f.$ denotes the error degrees of freedom which invariably equals (n - the number of estimated model parameters used for predicting the $\hat{Y}_i$'s). The reader should check that this general approach of estimating σ^2 gives the same estimate of σ^2's that we have obtained earlier in the cases of single Normal population (§3.2 of Session 2 notes), two Normal populations (case III of §3.2 of Session 2 notes) and k Normal populations (§2.1 above). Using this UMVUE of σ^2, the hypothesis of interest is then tested using an F-statistics, which has the formula $\frac{SSH_0/\text{its } d.f.}{SSE/e.d.f.}$ and an $F_{H_0 d.f., e.d.f.}$ distribution under H_0. As in §2.1, intuitively, the numerator of this F-statistic is the average standard "small" amount of variability that one expects to see if H_0 were true, while what should be considered to be "small" intrinsically depends on the phenomenon and the response variable under consideration, a reasonable value of which is given by an estimate of the average inherent amount variability, which is precisely the denominator of the F-statistic, and thus if this ratio is "large" that leads to the suspicion about the truth of H_0. The question of what is "large" is as usual answered using the sampling distribution of the test statistics which is $F_{H_0 d.f., e.d.f.}$ in this case.

Coming back to testing the null hypothesis $H_{0\gamma}$ of no two-factor interaction, all we need to find is the corresponding $SSH_{0\gamma}$, because in this case in general, the predicted value of Y_{ijk} would be given by $\hat{Y}_{ijk} = \hat{\mu}_{ij} = \overline{Y}_{ij} = \frac{1}{r}\sum_{k=1}^{r} Y_{ijk}$, and thus SSE would be given by $\sum_{i=1}^{s_1}\sum_{j=1}^{s_2}\sum_{k=1}^{r}(Y_{ijk} - \overline{Y}_{ij})^2$ and since we are thus estimating $l = s_1 s_2$ parameters for obtaining $\hat{Y}_{ijk}$, $e.d.f.$ would equal $n - l$, where $n = s_1 s_2 r$ denotes the total number of observations. Now the interaction Sum of Squares, denoted by $SS(X_1 * X_2)$, which is same as the sought $SSH_{0\gamma}$ is found as follows.

First find $SSTr$ using the l treatments as explained in §2.1 above. Thus $SSTr = r\sum_{i=1}^{s_1}\sum_{j=1}^{s_2}(\overline{Y}_{ij} - \overline{Y})^2$, where $\overline{Y} = \frac{1}{n}\sum_{i=1}^{s_1}\sum_{j=1}^{s_2}\sum_{k=1}^{r} Y_{ijk}$ is the grand mean of all the n observations. Note that $SSTr$ measures the variability that exists in the responses due to receiving different treatments, which are nothing but all possible combinations of levels of the two factors X_1 and X_2. There are three major sources of this variability namely the two main effects of X_1 and X_2 and the interaction effect of $X_1 * X_2$. The variability or SS due to the main effects are found again using the same argument as in §2.1. For $i = 1, 2, \ldots, s_1$ let $\overline{Y}_{i\cdot} = \frac{1}{rs_2}\sum_{j=1}^{s_2}\sum_{k=1}^{r} Y_{ijk}$ denote the mean response for receiving the i-th level of factor X_1, and similarly for $j = 1, 2, \ldots, s_2$ let $\overline{Y}_{\cdot j} = \frac{1}{rs_1}\sum_{i=1}^{s_1}\sum_{k=1}^{r} Y_{ijk}$ denote the mean response for receiving the j-th level of factor X_2. Thus following the same argument as in §2.1, SS due to the main effects of factor X_1 and X_2, denoted respectively by SSX_1 and SSX_2, are found

as $SSX_1 = rs_2 \sum_{i=1}^{s_1} (\overline{Y}_{i.} - \overline{Y})^2$ and $SSX_2 = rs_1 \sum_{j=1}^{s_2} (\overline{Y}_{.j} - \overline{Y})^2$. Now since we already have the variability or SS due to changing treatments in $SSTr$, which consists of the two main effects and the two-factor interaction, and have measured the variability or the SS due to the two main effects in SSX_1 and SSX_2, the variability or SS due to interaction is found as $SS(X_1 * X_2) = SSTr - SSX_1 - SSX_2$. Discussion about a number of points is in order at this point of time.

First let us revisit the two main effect SS, SSX_1 and SSX_2. Note that having no main effect due to factor X_1 is same as saying $\alpha_1 = \alpha_2 = \cdots = \alpha_{s_1} = 0$. If the null hypothesis $H_{0\alpha} : \alpha_1 = \alpha_2 = \cdots = \alpha_{s_1} = 0$ were true, we will see little difference in the values of $\overline{Y}_{1.}, \overline{Y}_{2.}, \ldots, \overline{Y}_{s_1.}$. The variability in these $\overline{Y}_{1.}, \overline{Y}_{2.}, \ldots, \overline{Y}_{s_1.}$ values is precisely what is being measured in SSX_1 and thus alternatively SSX_1 can also be viewed as $SSH_{0\alpha}$. In particular note that α_i can be estimated as $\hat{\alpha}_i = \overline{Y}_{i.} - \overline{Y}$ for $i = 1, 2, \ldots, s-1$, and thus $SSH_{0\alpha}$ can be interpreted as $rs_2 \sum_{i=1}^{s_1} \hat{\alpha}_i^2$. Likewise for the null hypothesis of no main effect of X_2, $H_{0\beta} : \beta_1 = \beta_2 = \cdots = \beta_{s_2} = 0$ its appropriate SS is given by $SSH_{0\beta} = SSX_2 = rs_1 \sum_{j=1}^{s_2} \hat{\beta}_j^2$, where $\hat{\beta}_j = \overline{Y}_{.j} - \overline{Y}$.

Viewed as above, our $SSH_{0\gamma}$ or the interaction SS should equal $r \sum_{i=1}^{s_1} \sum_{j=1}^{s_2} \hat{\gamma}_{ij}^2$, and $\hat{\gamma}_{ij}$ should equal $Y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j$. With $\overline{Y}$ as $\hat{\mu}$, and $\hat{\alpha}_i$ and $\hat{\beta}_j$ as above, $\hat{\gamma}_{ij}$ may be found as $\overline{Y}_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} + \overline{Y}$. Interestingly the interaction $SS(X_1 * X_2)$ obtained above as $SSTr - SSX_1 - SSX_2$ coincides with $r \sum_{i=1}^{s_1} \sum_{j=1}^{s_2} (\overline{Y}_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} + \overline{Y})^2$.

Thus now we have all the SS that is of interest for the complete analysis of the full factorial design with two factors. To summarize, we have three hypotheses which are of interest namely $H_{0\alpha}$, $H_{0\beta}$ and $H_{0\gamma}$ and we have also derived the three SS that would be of use to test these hypotheses together with the SSE. These computations are summarized in the form of an Analysis of Variance (ANOVA) table as follows:

ANOVA Table

Source of Variation	$D.F.$	SS	$MSS = SS/D.F.$	F	p-value
X_1	$s_1 - 1$	SSX_1	MSX_1	$F_\alpha = \frac{MSX_1}{MSE}$	$P(F_{s_1-1,n-l} > F_\alpha)$
X_2	$s_2 - 1$	SSX_2	MSX_2	$F_\beta = \frac{MSX_2}{MSE}$	$P(F_{s_2-1,n-l} > F_\beta)$
$X_1 * X_2$	$(s_1-1) \times (s_2-1)$	$SS(X_1 * X_2)$	$MS(X_1 * X_2)$	$F_\gamma = \frac{MS(X_1*X_2)}{MSE}$	$P(F_{(s_1-1)(s_2-1),n-l} > F_\gamma)$
Treatment	$s_1 s_2 - 1$	$SSTr$	$MSTr$	$F_{Tr} = \frac{MSTr}{MSE}$	$P(F_{s_1 s_2-1,n-l} > F_{Tr})$
Error	$n - l$	SSE	MSE		
Total	$n - 1$	$SSTotal$			

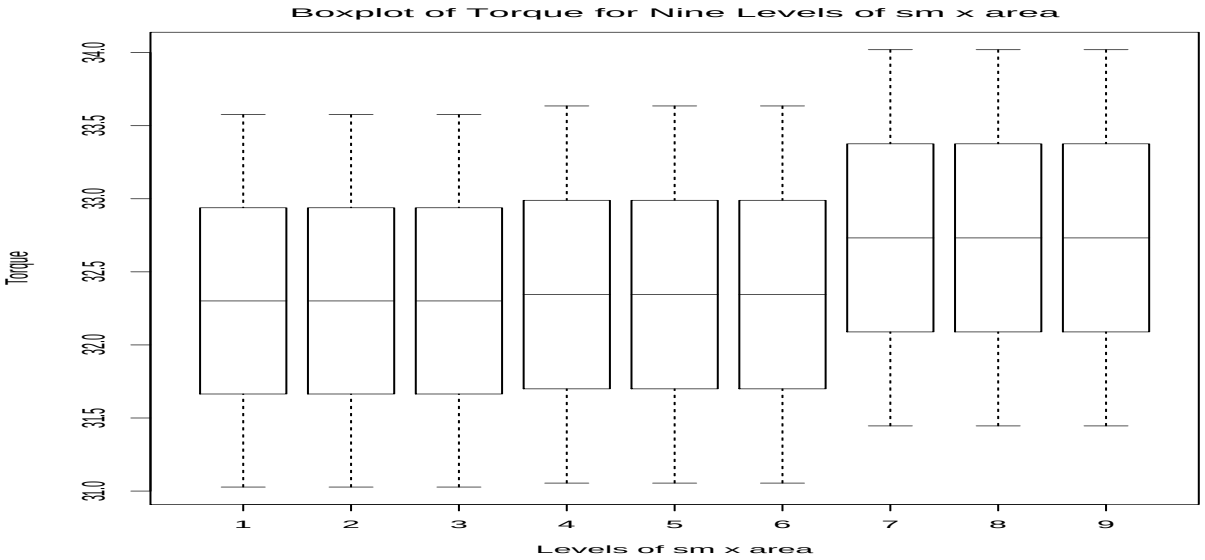
where $SSX_1 = rs_2 \sum_{i=1}^{s_1} (\overline{Y}_{i.} - \overline{Y})^2$, $SSX_2 = rs_1 \sum_{j=1}^{s_2} (\overline{Y}_{.j} - \overline{Y})^2$, $SS(X_1 * X_2) = r \sum_{i=1}^{s_1} \sum_{j=1}^{s_2} (\overline{Y}_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} + \overline{Y})^2$, $SSTr = r \sum_{i=1}^{s_1} \sum_{j=1}^{s_2} (\overline{Y}_{ij} - \overline{Y})^2$, $SSE = \sum_{i=1}^{s_1} \sum_{j=1}^{s_2} \sum_{k=1}^{r} (Y_{ijk} - \overline{Y}_{ij})^2$ and $SSTotal = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y})^2$.

As mentioned in the beginning of this sub-section, one first tests for $H_{0\gamma}$. If it is not rejected then one tests for the main effects with the F-tests as given in the ANOVA table

above. Otherwise one absorbs the $SS(X_1 * X_2)$ and the corresponding d.f. into the SSE and $e.d.f.$ respectively and then tests for the main effects.

Note that if the replication number for each treatment (r) equals 1, SSE as given above reduces to 0. In such situations, there is no way to test hypothesis about all the three hypotheses$H_{0\alpha}$, $H_{0\beta}$ and $H_{0\gamma}$. Under such circumstances, typically it is assumed that there is no interaction and thus $SS(X_1 * X_2)$ and its d.f. is treated as the SSE and $e.d.f.$ respectively for testing for the main effects. This technique of ignoring interactions and thus computing SSE assuming only the main effects, will be used through out while analyzing data from fractional factorial designs.

Example 3 (Continued): Along with the variable `sm` now we also wish to include a second factor called `area` in modeling and analysis of the response `torque`. `area` is again a quantitative variable which is experimented with three possible values 28.27433×10^{-6}, 29.27433×10^{-6} and $30.27433\times10^{E-6}$. These constitute the three levels of the factor `area`. That is in this experiment, $s_1 = s_2 = 3$. Thus with the two factors `sm` and `area`, each at three levels, there are nine treatments in total. Also here the experiment is replicated three times for each of the nine treatment combinations, *i.e.* the replication number $r = 3$ in this experiment. As a preliminary analysis as usual we first create a box-plot as follows:

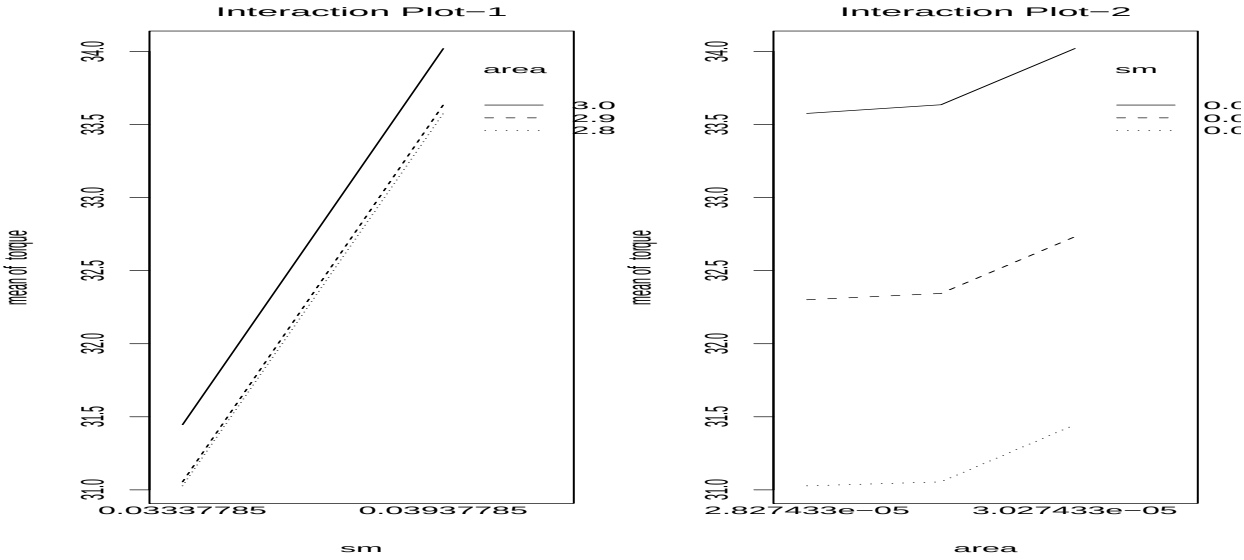


From this plot it appears that the nine treatment means differ from one another. To obtain further break-up of how the main effects and the interaction of `sm` and `area` are affecting the response `torque` we perform a two-way ANOVA as follows.

```
> anova(aov(torque~sm*area))
Analysis of Variance Table

Response: torque
          Df    Sum Sq   Mean Sq    F value    Pr(>F)
sm         2   29.6704   14.8352 5.0974e+31 < 2.2e-16 ***
area       2    1.0149    0.5075 1.7437e+30 < 2.2e-16 ***
sm:area    4    0.0009    0.0002 7.4384e+26 < 2.2e-16 ***
```

```
Residuals 18 5.239e-30 2.910e-31
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From this analysis it follows that all the three effects are highly significant. However more precisely it may be stated that of the total amount of variability in `torque`, given by $SSTotal = 30.68622$; `sm` alone, or its main-effect, explains about 96.6897% of this variability with SS(`sm`)=29.6704; `area` alone, or its main-effect, explains about 3.3073% of this variability with SS(`area`)=1.0149; and the interaction between these two factors account for about 0.0029% of this variability with SS(`sm`* `area`)=0.0009; with a minuscule amount of 0.0001% left unexplained or attributed to the error. These percentages are also called PCR or Percentage Contribution Ratio, which are fairly useful statistics which may accompany an ANOVA table. The lesson learned from this ANOVA model is that, for the subsequent quantitative regression model of `torque` in terms of `sm` and `area`, we must start with both the appropriate main effect and interaction terms. We conclude this two-factor analysis of the response `torque` by studying the following interaction plots, which plots the mean level of a response against the changing levels of one factor, while holding the level of the other factor(s) constant.



These depict the interaction as well as the quantitative nature of the effects of the two factors `sm` and `area` on `torque`. Though the three sets of curves look nearly parallel in both the plots, it is for avoiding falling into such type of subjective trap, that we need to perform the formal Analysis of Variance as in the previous page. The ANOVA says that though the interaction SS of 0.0009 is very small, it is nonetheless significant and thus should get at least preliminary consideration while building the subsequent quantitative regression model for `torque` in terms of the quantitative levels of `sm` and `area`, which were not explicitly exploited in the ANOVA model building exercise above. Furthermore notice that the plots also provide hints to the nature of this quantitative relationship. In particular it appears that the effect of `sm` on `torque` is linear, while that of `area` is non-linear. With three distinct levels of `area`, the maximal polynomial model we would be able to employ to capture this non-linearity would be quadratic. A second option would be to try to model `torque` using

other non-linear transformation of `area` like its logarithm or some power, which is possible to do using even two distinct levels, though would be hard to distinguish from a linear one in case there are only two distinct levels compared to the present case of three levels, which just adds an additional degree of freedom.

2.3 Multi-Factor ANOVA

Here we shall not attempt to unnecessarily give the general mathematical formulation, which though notationally complex (and hence the reason for the avoidance) is a straight-forward generalization of the two-factor ANOVA model of §2.2 above. Except now the full-factorial model with replication allows one to estimate the higher order interactions, the concept of which is more general than the two-factor interaction of §2.2 and thus merits some discussion.

To understand the concept of three-factor or higher order interaction, we shall start with a three-factor model assuming that we have a full-factorial design with (equal number of) replications of the experiment at each treatment combination. Let X_1, X_2 and X_3 denote the three factors having s_1, s_2 and s_3 levels respectively. The three factor interaction $X_1 * X_2 * X_3$ explains how the two-factor interactions change with the changing level of the third factor. To understand this concept geometrically, think of the interaction plot of X_1 and X_2 for a fixed level of X_3 which is a 2D plot as in the previous example. Now stack these 2D interaction plots of X_1 and X_2 on each other on the third dimension for changing levels of X_3. The curves in the interaction plot of X_1 and X_2 (for a fixed level of X_3) may be non-parallel, indicating a presence of two-factor interaction between X_1 and X_2. However the nature of this non-parallelism itself might change for changing levels of X_3. This is what the three-factor interaction tries to capture.

In general, a k-factor interaction might be understood recursively as the nature by which the $k-1$-factor interaction changes for the changing levels of a k-th factor. As it can be seen that the higher order interactions quickly loose their practical interpretablity and thus they are better avoided in the model building process. This also has the practical advantage in the design issues. For example, if we are willing to sacrifice the highest order interaction, we can get away with just a full-factorial design with just 1 replication. In this case, strictly speaking $SSE = 0$ if we attempt to isolate the SS corresponding to the highest order interaction. But since we pretend that there is no highest-order interaction, we treat the SS corresponding to the highest order interaction as the SSE, which then enables us to check for significance of all the lower order effects. This point has also been mentioned in the last paragraph of §2.2 immediately preceding the example in the context of two-factor full factorial design.

Here we take it even a step further. For example consider the three factor experiment with temperature, pressure and catalyst as the three factors for modeling the yield of a chemical process discussed in Session 1. Table 1 of Session 1 notes give the lay-out of the full-factorial design. This design will enable us to estimate all the two factor interactions but the three-factor interaction will remain confounded with the SSE. This is same as the point mentioned in the last paragraph for $r = 1$ with full-factorial design, namely sacrificing the highest order interaction, which is the three-factor interaction in this example. Now consider the design given in Table 2 of Session 1 notes. In this design we not only sacrifice the three-factor interaction, but also all the two-factor interactions as well! But this design (the design given in Table 2 of Session 1 notes) allows us to estimate the three main effects.

This is the philosophy behind constructing the fractional factorial designs. That is first decide which effects you want to estimate. Then use a minimal design which captures only those effects, sacrificing the higher-order effects by clubbing them into SSE.

Now turning back to the problem of figuring out the SS for the higher order interactions, we shall explain it in the context of a three-factor full-factorial design with replication with X_1, X_2 and X_3 as the three factors having s_1, s_2 and s_3 levels. First find $SSTr$ (as in §2.1) due to the $l = s_1 \times s_2 \times s_3$ treatments, denoted by $SS(X_1 \times X_2 \times X_3)$. Now find $SS(X_1)$, $SS(X_2)$, and $SS(X_3)$ as the SS due to the main-effects of the three factors by comparing their group means. [3] Likewise compute $SS(X_1 \times X_2)$, $SS(X_2 \times X_3)$ and $SS(X_3 \times X_1)$. Then the two factor interaction SS are found as $SS(X_1 * X_2) = SS(X_1 \times X_2) - SS(X_1) - SS(X_2)$, $SS(X_2 * X_3) = SS(X_2 \times X_3) - SS(X_2) - SS(X_3)$ and $SS(X_3 * X_1) = SS(X_3 \times X_1) - SS(X_3) - SS(X_1)$; and the three-factor interaction SS, $SS(X_1 * X_2 * X_3)$ is found as $SS(X_1 * X_2 * X_3) = SS(X_1 \times X_2 \times X_3) - SS(X_1) - SS(X_2) - SS(X_3) - SS(X_1 * X_2) - SS(X_2 * X_3) - SS(X_3 * X_1)$. The d.f. for $SS(X_1)$, $SS(X_2)$, $SS(X_3)$, $SS(X_1 * X_2)$, $SS(X_2 * X_3)$, $SS(X_3 * X_1)$ and $SS(X_1 * X_2 * X_3)$ are given by $s_1 - 1$, $s_2 - 1$, $s_3 - 1$, $(s_1 - 1)(s_2 - 1)$, $(s_2 - 1)(s_3 - 1)$, $(s_3 - 1)(s_2 - 1)$, and $(s_1 - 1)(s_2 - 1)(s_3 - 1)$ respectively, totaling to $s_1 s_2 s_3 - 1$ the d.f. of $SSTr$ or $SS(X_1 \times X_2 \times X_3)$.

Example 3 (Continued): Now bring in a third factor called Crank-Mass (`cm`) having three levels 0.03337785, 0.03347785, 0.03357785 like `sm`. Now we have a full-factorial design with 27 (=3×3×3) treatments each replicated exactly only once. This allows us to build a model with all the three two-factor interactions but sacrificing the three-factor interaction. This is accomplished as follows.

```
> anova(aov(torque~sm*area*cm-sm:area:cm))
Analysis of Variance Table

Response: torque
          Df    Sum Sq   Mean Sq    F value Pr(>F)
sm         2   29.6704   14.8352 5.3128e+31 <2e-16 ***
area       2    1.0149    0.5075 1.8174e+30 <2e-16 ***
cm         2 6.237e-31 3.119e-31 1.1168e+00 0.3734
sm:area    4    0.0009    0.0002 7.7528e+26 <2e-16 ***
sm:cm      4 1.109e-30 2.772e-31 9.9270e-01 0.4641
area:cm    4 1.264e-30 3.161e-31 1.1319e+00 0.4067
Residuals  8 2.234e-30 2.792e-31
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that the third factor `cm` now does not have any effect on `torque` and thus should be dropped, and the model must be refitted with only `sm` and `area` as the two factors. But this has already been done in §2.2, and that model is the one we should go by for modeling `torque` in terms of `sm`, `area` and `cm`.

[3] At this stage we shall define SS(Effect) due to any effect as follows. Whatever may be the effect, like X_1 or $X_1 \times X_2 \times X_3$, suppose the effect has k levels $1, 2, \ldots, k$. Now letting $Y_1, Y_2, \ldots, Y_n$ denote the n responses, let $G = \sum_{i=1}^{n} Y_i$, and for $j = 1, 2, \ldots, k$ let $n_j = \#\{i : \text{the Effect} = j\}$ and $T_j = \sum_{i:\text{the Effect}=j} Y_i$. Then $SS(\text{Effect}) = \sum_{j=1}^{k} \frac{T_j^2}{n_j} - C.F.$, where the correction factor $C.F. = \frac{G^2}{n}$.

3 Regression Models

In this section we shall assume that we have k independent variables or k factors $X_1, X_2, \ldots, X_k$, all of which are quantitative, and we wish to build a quantitative model for the response Y, utilizing the (quantitative) values or the levels of the k factors. As mentioned in the Introduction, we shall only build linear regression models, though they are capable of capturing non-linear relationships between Y and the X's as well.

One very important point that is to be noted for the regression models is that, the model just describes the conditional distribution of $Y|X_1, X_2, \ldots, X_k$. That is in the regression models (or for that matter all the linear models we are concerned with here) the response Y is assumed to be random, having (typically a Normal) p.d.f. in the population; while since the factors are under our control or we vary their values or levels at our will, they are assumed to be non-random; and since we are interested in the way the population of Y values change with changing levels of $X_1, X_2, \ldots, X_k$, we are concerned with modeling this conditional distribution of $Y|X_1, X_2, \ldots, X_k$. The models are called regression models because, the **definition** of *regression* of Y on $X_1, X_2, \ldots, X_k$ is nothing but the conditional mean of Y given $X_1, X_2, \ldots, X_k$, or notationally $E[Y|X_1, X_2, \ldots, X_k]$, which is a non-stochastic function of $X_1, X_2, \ldots, X_k$ called the regression function.

There are numerous theoretical motivations for studying or modeling the regression function of Y on $X_1, X_2, \ldots, X_k$. Simply put, the regression function is the "best" possible predictor of Y given the factors $X_1, X_2, \ldots, X_k$. Two such theoretical results motivating studying the regression analysis for predicting Y in terms of $X_1, X_2, \ldots, X_k$ are as follows. First, if we wish to construct a predictor for Y based on $X_1, X_2, \ldots, X_k$ which minimizes the mean square prediction error, then regression function comes out as the answer for such optimal predictor. Second, if we seek a predictor for Y based on $X_1, X_2, \ldots, X_k$ which has maximum correlation with the observed values of Y, then also the answer to such optimal predictor is the regression function.

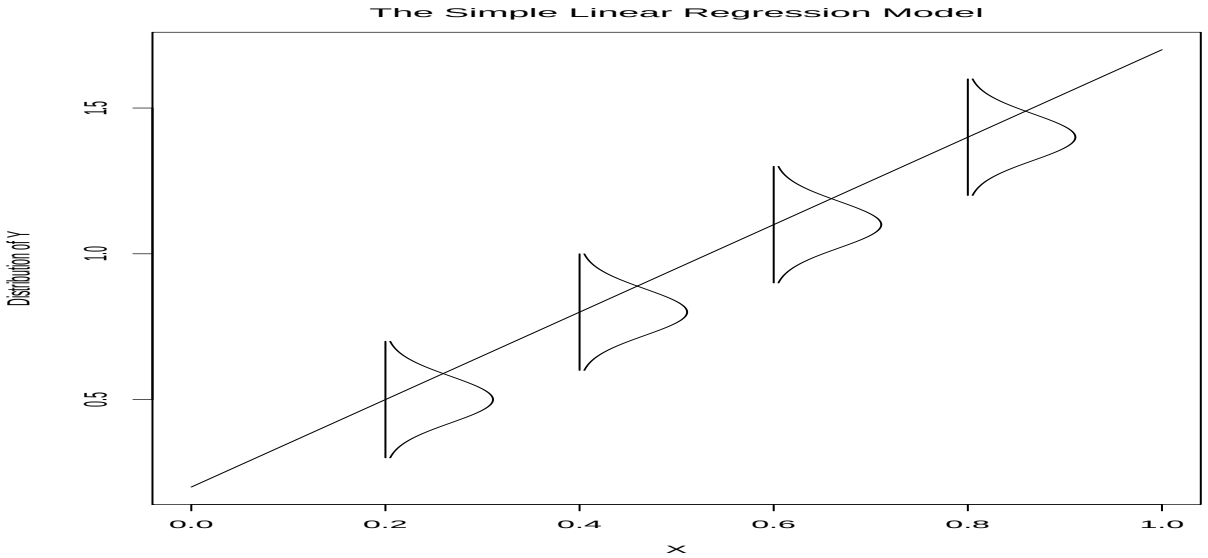
3.1 Simple Linear Regression

We begin with the simplest possible model and then move on to discussing more complex models. Suppose we have only one X, which takes at the minimum two distinct values or has two levels or more. In simple linear regression models we build a straight-line relationship between such an X and the response Y. More precisely the model we wish to build is given by $Y = \beta_0 + \beta_1 X + \epsilon$ with $\epsilon \sim N(0, \sigma^2)$. What this model says is that for every single X, there is a population of Y values characterized by the conditional distribution of Y given X denoted by $Y|X$. As X changes this conditional distribution of $Y|X$ changes. But the only thing that changes in this distribution is its mean, its family (Normal distribution) and the variance remains unchanged as X changes. Or in other words, for a simple linear regression model, the conditional distribution of $Y|X$ is $N(\beta_0 + \beta_1 X, \sigma^2)$.

Unlike the ANOVA model the regression model also says how the mean of Y changes with X. The ANOVA model only says that the means of Y are different for different levels of X. It does not precisely state how this mean changes in quantitative terms, and thus is more general in nature. In contrast, the regression model goes a step further and attempts to model the exact quantitative nature of this change in the mean level of Y by making it

depend on the exact quantitative value of X. Both models however leave the family of the distribution and variance unchanged (the homoscedasticity assumption).

For instance, the simple linear regression model says that as X changes the only thing that changes is the mean of Y, and it changes linearly following the expression $\beta_0 + \beta_1 X$. It however also states that the variance of $Y|X$ does not depend on X, or it remains constant as X changes, which precisely is the assumption of homoscedasticity. It finally states that the distribution of Y for a given value of X is also Normal with its mean and variance as just described above. This simple linear regression model may be graphically depicted as in the following figure:



The straight line $y = \beta_0 + \beta_1 X$ in this figure depicts how the conditional mean of $Y|X$ changes with X. It then goes on to show the distributions of the populations of Y values for changing X values. These distributions (there are infinitely many of them, depending on the possible values X can theoretically take) are all Normal, indicated by the Normal p.d.f.'s centered around their respective mean values. And all these Normal distributions have the same shape or the same variance σ^2 owing to the homoscedasticity assumption.

Now suppose we have n observations $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ coming from the above depicted probability model. The standard statistical problem of interest is that of inference, which consists of estimation and hypothesis testing, about these population parameters β_0, β_1 and of course σ^2. Of these three parameters, the key parameter of interest is β_1, which ties up X with the conditional distribution of $Y|X$. In particular β_1 gives the *expected* amount of change in Y for a unit increase in X.

The UMVUE of β_1 is given by $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$, where $S_{xy} = \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$ and $S_{xx} = \sum_{i=1}^{(} X_i - \overline{X})^2$. Similarly the UMVUE of β_0 is given by $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$. [4] The

[4] These estimators can basically be derived as follows. Consider the likelihood function $L(\beta_0, \beta_1, \sigma^2)$ of β_0, β_1 and σ^2. For any probability model the likelihood function of the population parameters is given by the probability of observing the data at hand, for a given value of the parameter values, viewed as a function of these parameter values. Thus for this model, $L(\beta_0, \beta_1, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2\right\}$. Maximum Likelihood Estimate (or MLE) of β_0, β_1 and σ^2 may be found as those values $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}^2$ which maximizes the function $L(\beta_0, \beta_1, \sigma^2)$ of β_0, β_1 and σ^2. Now note that $L(\beta_0, \beta_1, \sigma^2)$ is maximized

sampling distribution of $\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \sim N_2\left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \frac{\sigma^2}{S_{xx}} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n X_i^2 & -\overline{X} \\ -\overline{X} & 1 \end{bmatrix}\right)$, where $N_2, (\cdot, \cdot)$ denotes a bivariate Normal distribution. Thus using this sampling distribution one could test hypothesis or construct confidence intervals for the β_j's for known σ^2. However as usual since σ^2 in general is unknown, this calls for first estimating this variance.

Note that σ^2 is the conditional variance of Y given X and the general procedure of finding the UMVUE of σ^2 as discussed in the third paragraph of §2.2 still holds. Thus the numerator of this estimate is found by $SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$, where $\hat{Y}_i$ is the predicted value of Y_i using the estimated model given by $\hat{\beta}_0 + \hat{\beta}_1 X_i$. Note that $\hat{\beta}_0 + \hat{\beta}_1 X_i$ is also nothing but the estimated conditional mean of Y_i given X_i, and thus it is no surprise that it appears as the subtracter of the observation Y_i in the estimate of the conditional variance. Now the *e.d.f.* is also found from the general principle discussed in the third paragraph of §2.2, namely (n - the number of estimated model parameters used for predicting the $\hat{Y}_i$'s). Since we are estimating 2 parameters β_0 and β_1 for predicting the $\hat{Y}_i$'s, the required *e.d.f.* is given by $n-2$ and the UMVUE of of σ^2 is found as $SSE/(n-2)$. A computationally efficient formula for computing SSE in this case is given by $SSE = S_{yy} - \hat{\beta}_1 S_{xy}$, where $S_{yy} = \sum_{i=1}^n (Y_i - \overline{Y})^2$. Finally, it may be shown that $\frac{SSE}{\sigma^2} \sim \chi^2_{n-2}$ and is independent of $\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$.

Using these estimates and their sampling distributions we now provide the following inference procedure for the main parameter of interest β_1 as follows. The basic t-statistics for β_1 is given by $\frac{\hat{\beta}_1 - \beta_1}{\widehat{SE}(\hat{\beta}_1)} \sim t_{n-2}$ where $\widehat{SE}(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2/S_{xx}}$ with $\hat{\sigma}^2 = SSE/(n-2)$, and the test of hypothesis and the confidence interval for β_1 are as follows:

Hypotheses	Fixed Significance Level Testing	p-value
$H_0: \beta_1 \geq \beta_{10}$ $H_{a1}: \beta_1 < \beta_{10}$	Reject H_0 if $t_{obs} = \frac{\hat{\beta}_1 - \beta_{10}}{\widehat{SE}(\hat{\beta}_1)} < t_{n-2,\alpha}$	$P(t_{n-2} < t_{obs})$
$H_0: \beta_1 \leq \beta_{10}$ $H_{a2}: \beta_1 > \beta_{10}$	Reject H_0 if $t_{obs} = \frac{\hat{\beta}_1 - \beta_{10}}{\widehat{SE}(\hat{\beta}_1)} > t_{n-2,1-\alpha}$	$P(t_{n-2} > t_{obs})$
$H_0: \beta_1 = \beta_{10}$ $H_{a2}: \beta_1 \neq \beta_{10}$	Reject H_0 if $t_{obs} = \left\lvert\frac{\hat{\beta}_1 - \beta_{10}}{\widehat{SE}(\hat{\beta}_1)}\right\rvert > t_{n-2,1-\alpha/2}$	$2P(t_{n-2} > t_{obs})$

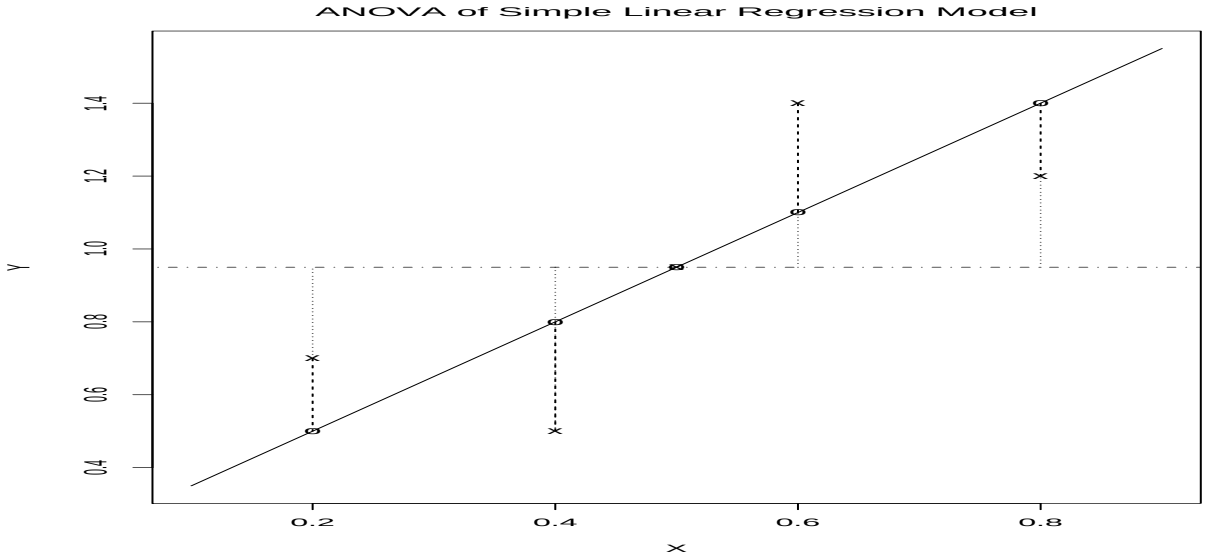
A $100(1-\alpha)\%$ confidence interval for β_1 is given by $\hat{\beta}_1 \pm t_{n-2,1-\alpha/2}\widehat{SE}(\hat{\beta}_1)$.

The test of H_{a3} above for $\beta_{10} = 0$ can also be viewed from an ANOVA perspective with an F-test. One way of interpreting $H_0 : \beta_1 = 0$ is the straight-line model is not doing a good job or the straight-line model is basically useless. On the other hand $H_{a3} : \beta_1 \neq 0$ would mean that the straight-line model has some use, though it does not mean that it is adequate. Since we already have a t-test for testing this hypothesis, it might seem redundant to develop an F-test for the same. However it is instructive to consider this test in this simplest possible

for those values of β_0 and β_1 which minimizes $\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$. This minimization problem can be viewed as searching for that straight-line through the scatter of points $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, which minimizes the sum of vertical distances of these points from the sought straight-line. This approach of fitting a straight-line through a scatter of points is also called the method of Ordinary Least Squares, or OLS. Thus the OLS estimates of β_0 and β_1 are same as their respective MLE, both of which are found by minimizing $\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$. The solution to this problem can now be obtained by using straightforward calculus *i.e.* differentiating $\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$ w.r.t. β_0 and β_1 and equating them to 0 yields $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ and $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$.

case, where it happens to coincide with a t-test. This is because in the case of general regression models, the test for the question of whether a model is useful or not (which in the simple linear regression model is same as $\beta_1 \neq 0$ and $\beta_1 = 0$ respectively) is different from testing hypothesis about a single β, leading to an F-test. We discuss this F-test here in the case of simple linear regression, because the geometrical logic of this F-test is easiest to understand here, which can then be used to understand the logic of ANOVA F-test in the general regression models.

Thus consider the hypothesis which says that a simple linear regression model has some use in predicting Y from an X. This under the given model is same as saying $\beta_1 \neq 0$. This is a point which we must establish if we are to proceed any further with the simple linear regression model. Since this is the point we wish to prove, we must take this as the alternative hypothesis and then the *de facto* null becomes $\beta_1 = 0$ which states that the model is basically useless. Now when should we say that the model has some use in predicting Y's? If the X's can explain a "significant" amount of variability in the Y's through the straight-line model, then we should say that the model has some use. This leads to analyzing the variance of Y and decomposing it into the constituent components. For having a geometrical feeling of this decomposition consider the following plot:



In the above plot the crosses indicate the observed data points $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, the circles indicate the predicted points $(X_1, \hat{Y}_1), (X_2, \hat{Y}_2), \ldots, (X_n, \hat{Y}_n)$ using the straight-line model, the solid line gives the estimated regression line $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ and the dot-dashed line represents $Y = \overline{Y}$.

To begin with, the Total amount of variability in Y can be measured by $SSTotal = S_{yy} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$. This is represented by the sum of squares of the dotted distances of the crosses from the dot-dashed line in the above figure. Now these distances $(Y_i - \overline{Y})$, as is evident from the figure, can be decomposed into two components - one from the observed point to the predicted point, given by $(Y_i - \hat{Y}_i)$, represented by the dashed lines; and the other from the predicted point to the mean line, given by $(\hat{Y}_i - \overline{Y})$, represented by the dotted lines. Algebraically of course $(Y_i - \overline{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \overline{Y})$. The deviations $(Y_i - \hat{Y}_i)$ are the errors of the model, which should ideally be small for the model to be useful. The deviations $(\hat{Y}_i - \overline{Y})$ on the other hand represents the deviations of the predicted values $\hat{Y}_i$'s. This is

because it may be shown that $\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i = \overline{Y}$. That is $\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$ gives the variability of the predicted values $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n$, and $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ is the already familiar SSE. Now it can be shown that $SSTotal = SSR + SSE$, where $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$ is called the Sum of Squares due to Regression. This is the basic ANOVA result in case of regression.

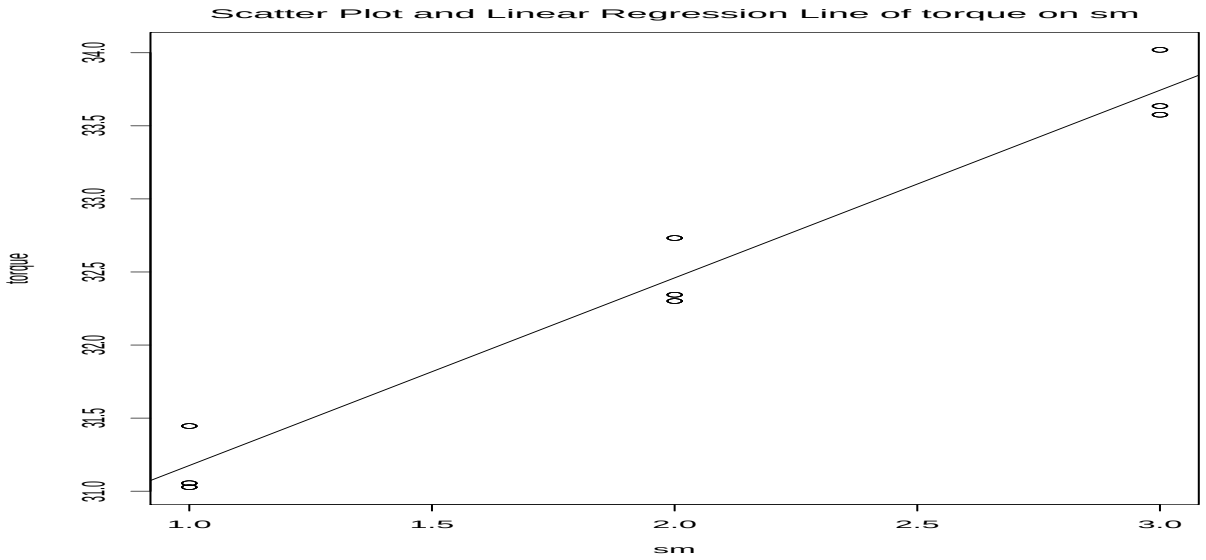
This ANOVA result may be understood as follows. As such the responses $Y_1, Y_2, \ldots, Y_n$ (the crosses in the figure) are different and the total amount of their variability may be measured using $SSTotal$ (sum of squares of the distances of the crosses from the dot-dashed mean line). But why are these Y_i's different? Part of the reason is because possibly the corresponding X_i's are different. So how much of this variability in the Y_i's can be explained by the simple linear regression model or the linear effect of the X's on the corresponding Y's? To answer this question see the kind of Y's one would *expect* to see for changing X values using the straight-line model and then measure the variability of these *expected* Y values. These *expected* Y values are nothing but $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n$, represented by the circles in the figure, which naturally all line up on the regression line. These $\hat{Y}_i$ values are not all same and thus, as explained in the last paragraph, their variability is given by SSR. Thus SSR basically gives that amount of variability in the Y_i's that is due to the linear effect of X on Y. That is the regression model predicts that the Y values would be different, and also gives the amount of variability that may be attributed due to the regression effect (which for the simple linear regression model is same as the linear effect of the X's on the corresponding Y's) as SSR. But that does not explain all the variability in Y. According to the ANOVA result, what is left in $SSTotal$ is nothing but SSE, which gives the amount of variability in Y that remains unexplained in the model, which is again nothing but the variability in the errors of the model given by $Y_1 - \hat{Y}_1, Y_2 - \hat{Y}_2, \ldots, Y_n - \hat{Y}_n$, which have mean 0.

Thus if SSR is large compared to SSE we should say that the model has been successfully able to explain the major portion of the variability in the responses, and thus we should reject $H_0 : \beta_1 = 0$, the model is useless, in favor of $H_a : \beta_1 \neq 0$, the model is useful. However SSR and SSE are not directly comparable because different d.f. are associated with the two SS. For figuring out the d.f. of SSR note that SSR is the amount of variability that exists in $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n$. But all these $\hat{Y}_i$'s are required to fall in a straight-line and since any two points determine a straight-line, though they are n numbers, fundamentally one just has two d.f. with them. (Because as soon as you fix two of them the others automatically get fixed.) But for measuring their variability, we have to first estimate their mean and in the process loosing one more d.f. Thus the d.f. of SSR equal 1. That the d.f. of SSE or $e.d.f.$ equals $n - 2$, has already been discussed above. However it is illustrative to see how this $n - 2$ is arising in this case. SSE is the variability of $e_1, e_2, \ldots, e_n$, where $e_i = Y_i - \hat{Y}_i$ for $i = 1, 2, \ldots, n$. The first degree of freedom (from a totality of n e_i's) is lost due to the fact the $\sum_{i=1}^{n} e_i = 0$, and the second degree of freedom is lost because the e_i's also need to satisfy a second constraint of $\sum_{i=1}^{n} e_i X_i = 0$. That is finally after adjusting for the respective d.f. the decision rule would be, reject $H_0 : \beta_1 = 0$ if $\frac{SSR/1}{SSE/(n-2)}$ is "large".

To decide how "large" is "large", we need to derive the sampling distribution of $\frac{SSR/1}{SSE/(n-2)}$. It can be shown that $SSR = \hat{\beta}_1 S_{xy} = \hat{\beta}_1^2 S_{xx}$. Thus under $H_0 : \beta_1 = 0$, $\frac{SSR}{\sigma^2} \sim \chi_1^2$ and in general $\frac{SSE}{\sigma^2} \sim \chi_{n-2}^2$ which is independent of $\hat{\beta}_1$. Thus under H_0, $\frac{SSR/1}{SSE/(n-2)} \sim F_{1,n-2}$ and we have an ANOVA F-test for testing $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$ by rejecting H_0 if $\frac{SSR/1}{SSE/(n-2)} . F_{1,n-2,1-\alpha}$ or computing the p-value as $P(F_{1,n-2} > \frac{SSR/1}{SSE/(n-2)})$. These computations

are usually represented in a ANOVA table, which is omitted here for the sake of brevity. It can be shown that the F-value obtained above would be identical to the square of the t-value that one would obtain for the t-test described for testing H_{a3} in page 14, and this should not come as a surprise because we already know that $t_\nu^2 \equiv F_{1,\nu}$, and thus the two tests (t-test and F-test) for testing $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$ are equivalent.

Example 3 (Continued): In §2.1 we built an ANOVA model for `torque` for changing levels of the quantitative factor `sm` and found it to be significant. However there we did not explicitly model the value of `torque` in terms of the quantitative values of `sm`. This is what we take up in the regression modeling. As usual we start with some graphical analysis. However here we plot the values of `torque` against the quantitative values of `sm` as opposed to studying the general problem of how the distribution of `torque` changes with changing levels of `sm` in §2. This plot of a response against a quantitative factor is called *scatter plot* which is as follows for the response `torque` and quantitative factor `sm`:



From this plot the relationship appears fairly linear and thus we proceed to formally fit a simple linear Regression model for `torque` in terms of `sm` as `torque`$=\beta_0 + \beta_1$`sm`$+\epsilon$, with $\epsilon \sim N(0, \sigma^2)$.

```
> lms<-lm(torque~as.numeric(sm))
> summary(lms)

Call:
lm(formula = torque ~ as.numeric(sm))

Residuals:
    Min      1Q  Median      3Q     Max
-0.1675 -0.1482 -0.1154  0.2701  0.2765

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)    29.89184    0.10264  291.24   <2e-16 ***
as.numeric(sm)  1.28388    0.04751   27.02   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2016 on 25 degrees of freedom
Multiple R-Squared: 0.9669,     Adjusted R-squared: 0.9656
F-statistic: 730.2 on 1 and 25 DF,  p-value: < 2.2e-16
```

Thus the estimated values of β_0 and β_1 are obtained as 29.89184 and 1.28388 respectively, both of which are highly significant. The estimate of σ is given by 0.2016. The R-output also reports two quantities called R^2 and Adjusted-R^2. While we shall discuss the concept of Adjusted-R^2 in the next sub-section, let us elaborate a little bit on the concept of R^2 in the simple linear regression context.

In general $R^2 = \frac{SSR}{SSTotal}$. Thus R^2 gives the proportion of variability explained by the model and typically it is the first key quantity one looks for in a regression output. However in the context of simple linear regression, R^2 has an additional meaning. It is the square of the correlation coefficient of the response Y and factor X, and in general a correlation coefficient measures the degree of linear association between the two variables. [5] Its numerical value is still interpreted in terms of its square, as explained above, and in this case it says what proportion of variability in Y can be attributed to the linear effect of X. For the example at hand, thus it says that 96.69% of the variability in `torque` can be explained away due to the linear effect of `sm`, which is quite large.

Though an R^2 value of 0.9669 is quite large, the immediate question is, is it significant? Or in other words is the straight line model doing a useful job? This is answered in terms of the ANOVA F-test, which is the last line of the bare minimum output produced by R. It gives an F-value of 730.2 with 1 and 25 d.f. with a p-value of nearly 0. Thus the model is indeed useful. It may be noted that F-statistic given by $\frac{SSR}{SSE/(n-2)}$ may be alternatively written in terms of R^2 as $= \frac{R^2}{(1-R^2)/(n-2)}$. Thus the ANOVA F-test is really same as checking for the significance of R^2, which in a nut-shell describes the usefulness of the straight-line model. For the sake of completeness the ANOVA table is presented below.

```
> anova(lms)
Analysis of Variance Table

Response: torque
               Df  Sum Sq Mean Sq F value    Pr(>F)
as.numeric(sm)  1 29.6704 29.6704  730.22 < 2.2e-16 ***
Residuals      25  1.0158  0.0406
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that through-out we are very cautiously using the term "usefulness" of the model instead of "adequacy" or "goodness of fit" of the model, while describing the ANOVA F-

[5] The correlation coefficient between two variables X and Y, typically denoted by r_{XY} is given by $\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$.

test. Though many times, albeit erroneously or being a little more harsh we may add very callously, the ANOVA F-test is called by such names. All the F-test does is check whether the model as an approximation may serve a useful purpose. It never says anything about the adequacy of the model, nor does it say anything about how good the fit is. True it says whether the fit has been useful or not, but not how good the fit is. But this is one of the major challenges of the regression model building, namely to assess whether the model has been adequate or how good the fit is. This is because, if not, then we have to revise and rebuild the model. These issues are taken up next.

Model Diagnostics:

The way the adequacy or goodness of fit of a model is judged is essentially by checking whether the model assumptions have been met. For example the simple linear regression model $Y = \beta_0 + \beta_1 X + \epsilon$ with $\epsilon \sim N(0, \sigma^2)$ states that once the model has been fitted the error or residual part computed using $e_1, e_2, \ldots, e_n$ where $e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$ should behave as though they are coming from a $N(0, \sigma^2)$ population. This statement has to be observed a little more carefully. This sates

1. $e_1, e_2, \ldots, e_n$ do not have anything to do with the X's any more.
2. $e_1, e_2, \ldots, e_n$ are homoscedastic, and
3. $e_1, e_2, \ldots, e_n$ have a Normal distribution.

There is also the assumption of $e_1, e_2, \ldots, e_n$ being independent. But in our applications here, in the context of AutoDOE, they will be independent by design and thus we will not bother about this assumption. In many applications the data are observed in a temporal or spatial sequence making consecutive observations auto-correlated. But here the runs in the CAE tool are independent and thus checking for any dependence on the run sequence is basically redundant.

However careful attention must be paid to the above three assumptions. The first assumption loosely translates to the issue of when the residuals are plotted against the X's they should not exhibit any pattern. If there is a distinct way by which the residuals are changing with the X's, that means there is still some residual pattern left, which the model has not been able to capture adequately, and thus there is scope for improvement of the model in terms of the way we let the X enter into the model. Non-linear effects of X on Y are thus typically diagnosed from such a plot for a straight-line model.

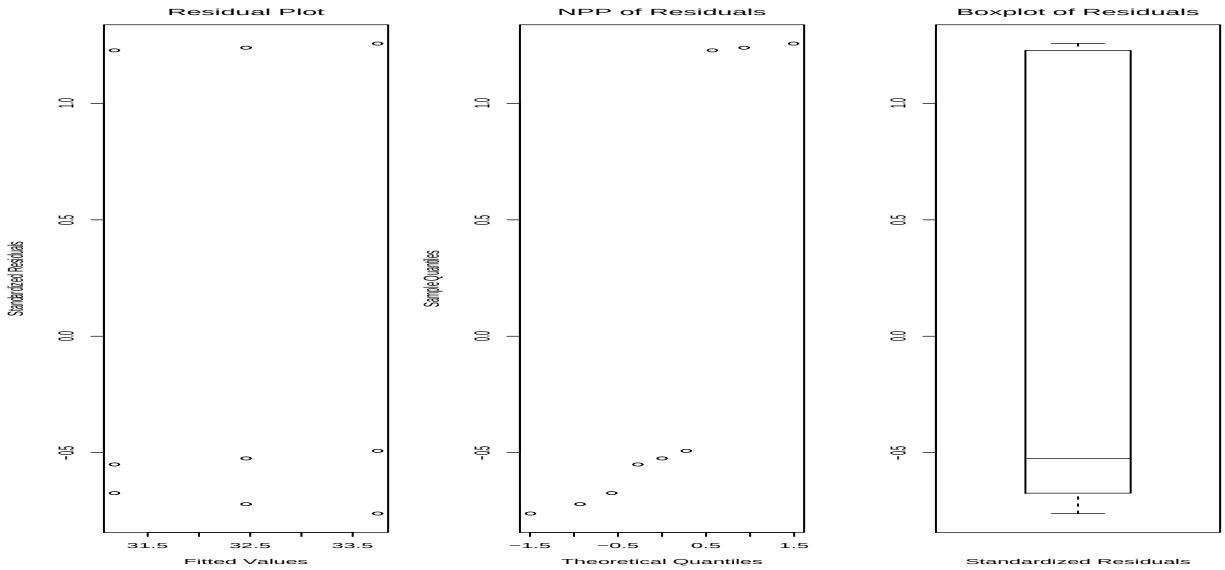
The plot of residuals against the X's also allows one to keep an eye on the second assumption of homoscedasticity. If the residuals exhibit unequal spread of values for changing levels of the X's then the homoscedasticity assumption is a suspect. Typically such situations arise together with the problem of Normality of the residuals, which can be checked using the NPP of residuals. If that is the case *i.e.* both assumptions two and three are problematic, then the issue is typically attempted to settle using a non-linear transformation of Y. The standard transformation one uses is some power or logarithm of Y. Note that this is different from allowing X to enter the model in a non-linear fashion. Otherwise in the very rare case of satisfactory NPP but heteroscedastic residuals, there is no other option but to model the σ^2 using the X's.

There is a third plot apart from the residual plot, defined below, and the NPP of the residuals, which is also useful. This is the box and whisker's plot of the residuals. This plot not only gives a rough check for Normality, and a feel for the distribution of the residuals,

it is also an indispensable tool for detecting outliers. An outlier in our AutoDOE context might help us capture tabulation errors or mis-specification of factor levels or any other unintentional error which might have crept into the runs.

Though we discussed plotting the residuals against the X's, in standard terms, this is not called the *residual plot*. In residual plots one plots the residuals $e_1, e_2, \ldots, e_n$ against the fitted values $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n$ instead of $X_1, X_2, \ldots, X_n$. Note however that in the simple linear regression model, there is a perfect linear relationship between $X_1, X_2, \ldots, X_n$ and $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n$, and thus both the plots would look identical. The plot of residuals against the fitted values is the preferred way of plotting because it is easily generalizable in the context of multiple X's. A second point to note is it is usually desirable to deal with the standardized residuals say $se_1, se_2, \ldots, se_n$ instead of the raw residuals $e_1, e_2, \ldots, e_n$, where $se_i = \frac{e_i}{\hat{\sigma}}$. This is because we know that if the regression assumptions are satisfied we shall expect to see values within ± 3 for the standardized residuals, while no such bound can as such be given for the raw residuals.

These plots are constructed for the example 3 and are as follows:



There does not appear to be any problem with the residual plot or the box-plot, though the NPP is clearly bad. However since our methods are least sensitive to the Normality assumption compared to the others, we accept the simple linear Regression model for `torque` in terms of `sm` and proceed to the next topic.

3.2 Multiple Regression

Here we wish to model Y using multiple X's like $X_1, X_2, \ldots, X_k$. This is a straight-forward generalization of the simple linear Regression model of §3.1 where we model Y with the multiple X's using a linear model. More precisely, Y is modeled as $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$ with $\epsilon \sim N(0, \sigma^2)$. Like in the case of simple linear regression, here also we are modeling the conditional distribution of Y, but this time given multiple X's, $X_1, X_2, \ldots, X_k$ instead of a single X. The conditional distribution of $Y|X_1, X_2, \ldots, X_k$ in the multiple regression model is assumed to be $N(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k, \sigma^2)$. Other than the mean structure, all other assumptions and concepts like a population of Y values

for given values of $X_1, X_2, \ldots, X_k$, their homoscedasticity and Normality etc. are exactly carried through as before in the multiple regression model as well. In the simple linear regression model, the conditional mean of Y depended on a single X through a straight-line relationship. Here also the conditional mean of Y depend on the multiple X's, $X_1, X_2, \ldots, X_k$ through a linear relationship.

Though this linear structure of the mean may look like a very restrictive model, it will be so only if all these X_j's are thought to be different factors. Multiple regression models allow one to build non-linear relationships between the X's and the Y as well. For example, suppose we have a single X but its relationship with Y appears to be cubic. Such a cubic model may be built using multiple regression technique with $X_1 = X$, $X_2 = X^2$ and $X_3 = X^3$. Likewise if there are two X's X_1 and X_2, and they seem to affect Y in a quadratic manner, such a quadratic model may be built using multiple regression technique with $X_1 = X_1$, $X_2 = X_2$, $X_3 = X_1X_2$, $X_4 = X_1^2$ and $X_5 = X_2^2$.

The complexity of such polynomial relationships that one can model using multiple regression, however is limited to the underlying design of experiment used for collecting observations on the X's and Y, which among other things determine the number of levels the different X_j's have been experimented with, the number of runs, the kinds of effects that have been allowed to estimate and so on. For example if in an experiment there are two factors X_1 and X_2 and each one of them has been allowed to assume only two possible values, then one cannot fit a full quadratic model in two variables, as discussed above. If there are replications at each treatment combination in this experiment, then the maximal model one can fit is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$, allowing additional room only for modeling the interaction effect. For being able to fit a cubic model, the factor must be allowed to assume at least four different levels. For the design given in Table 2 of Session 1 notes, the maximal model one would be able to fit is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$, where X_1, X_2, X_3 (a 0-1 valued dummy variable) and Y denote the temperature, pressure, catalyst and the chemical yield respectively, with no room for even accommodating the error term ϵ!

Thus assume that we have multiple X's $X_1, X_2, \ldots, X_k$, where some $X_{j'}$'s may be a function of some other X_j's. Now suppose we have n observations on the response Y, with the i-th observation denoted by Y_i, $i = 1, 2, \ldots, n$, for the independent variables $X_1, X_2, \ldots, X_k$ taking values $X_{i1}, X_{i2}, \ldots, X_{ik}$ respectively. Then according to the multiple regression model, $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \epsilon_i$ for $i = 1, 2, \ldots, n$. This model for all the n observations may be written in matrix notation as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where for $p = k + 1$,

$$\boldsymbol{Y}_{n\times 1} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix} \boldsymbol{X}_{n\times p} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1j} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2j} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{i1} & X_{i2} & \cdots & X_{ij} & \cdots & X_{ik} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nj} & \cdots & X_{nk} \end{bmatrix} \boldsymbol{\beta}_{p\times 1} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_k \end{pmatrix} \boldsymbol{\epsilon}_{n\times 1} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_i \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

The model errors $\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$, where $N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a n-variate Normal distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, $\boldsymbol{0}$ is an $n \times 1$ vector of all 0's and $\boldsymbol{I}_n$ denotes the $n \times n$ identity matrix.

Now the statistical problem of multiple regression analysis is exactly same as discussed in §3.1, except now for a more general model. Thus all we need to do is just jot down the formula and discuss only those concepts that are different in the multiple regression setting.

The UMVUE of $\boldsymbol{\beta}$, which may be derived from the MLE or equivalently OLS consideration as before, denoted by $\hat{\boldsymbol{\beta}}$ is given by $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. Furthermore it may be shown that the sampling distribution of $\hat{\boldsymbol{\beta}}$ is given by $\hat{\boldsymbol{\beta}} \sim N_p\left(\boldsymbol{\beta}, \sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right)$. The fitted or predicted values of the observations $\boldsymbol{Y}$ is given by $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$, so that the SSE in this case is obtained as $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = (\boldsymbol{Y} - \hat{\boldsymbol{Y}})^T(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = \boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}$, where the so-called $n \times n$ hat-matrix $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$. Here also it may be shown that $\frac{SSE}{\sigma^2} \sim \chi^2_{n-p}$ and is independent of $\hat{\boldsymbol{\beta}}$. Note that the $e.d.f.$ equals $n-p$ because one need to estimate $p = k+1$ β_j's for obtaining the predicted values $\hat{\boldsymbol{Y}}$. Thus the UMVUE of σ^2 is given by $\hat{\sigma}^2 = s^2 = MSE = SSE/(n-p)$.

With the above point estimates and sampling distributions of the model parameters $\boldsymbol{\beta}$ and σ^2, we are now ready to discuss the inferential issues. First consider inference about an individual β_j for $j = 0, 1, 2, \ldots, k$. The relevant quantity here is given by $\frac{\hat{\beta}_j - \beta_j}{\widehat{SE}(\hat{\beta}_j)} \sim t_{n-p}$ where $\widehat{SE}(\hat{\beta}_j)$ gives the estimated standard error of $\hat{\beta}_j$ as $s^2 \times \{(j+1, j+1)$-th element of the $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$ matrix$\}$, which can be used to test hypothesis or construct confidence intervals for the individual β_j's. The test statistics for testing $H_0 : \beta_j = \beta_{j0}$, where typically $\beta_{j0} = 0$, is given by $t_{obs} = \frac{\hat{\beta}_j - \beta_{j0}}{\widehat{SE}(\hat{\beta}_j)}$ with the decision rule of rejecting H_0 if $t_{obs} < t_{n-p,\alpha}$, $t_{obs} > t_{n-p,1-\alpha}$ or $|t_{obs}| > t_{n-p,1-\alpha/2}$ against the three three kinds of alternatives $H_{a1} : \beta_j < \beta_{j0}$, $H_{a2} : \beta_j > \beta_{j0}$ or $H_{a1} : \beta_j \neq \beta_{j0}$ respectively, or by computing the respective p-values as $P(t_{n-p} < t_{obs})$, $P(t_{n-p} > t_{obs})$ or $2P(t_{n-p} > |t_{obs}|)$. Similarly a 100(1-α)% confidence interval for β_j would be given by $\hat{\beta}_j \pm t_{n-p,1-\alpha/2}\widehat{SE}(\hat{\beta}_j)$.

Inference about individual β_j's are important and interesting for various reasons. β_j essentially gives the *partial* effect of X_j on Y in presence of the other $X_{j's}$'s. Thus for example in a quadratic model $Y = \beta_1 + \beta_1 X + \beta_2 X^2 + \epsilon$ a test for $H_0 : \beta_2 = 0$ against $H_a : \beta_2 \neq 0$ yields the result about whether a quadratic term is useful in modeling Y, over and above the inclusion of a linear term. In the model for a chemical yield $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, where X_1 and X_2 denote the temperature (say in ^{0} Kelvin) and pressure respectively, β_1 gives the *expected increment* in yield for increasing the temperature by 1^0 Kelvin for any given pressure level. If there were an interaction term $\beta_3 X_1 X_2$ in the model, then of course this *expected increment* would have also depended on the pressure level X_2 and would have been given by $\beta_1 + \beta_3 X_2$, [6] but that is not what is being discussed here for explaining the *partial* effect. β_1 in the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ says how X_1 is affecting Y when X_2 is also there in the model. This effect in general is different from how as such X_1 affects Y, which can be estimated using a model for Y involving X_1 alone and with no other variable, and is called the *marginal* effect. However fortunately, for orthogonal designs, as in the case of AutoDOE, these two are the same for different factors, and thus we need not worry too

[6] Such linear functions of $\boldsymbol{\beta}$, say $\boldsymbol{l}^T\boldsymbol{\beta}$, thus also arise naturally, inference for which may be drawn using the quantity $\frac{\boldsymbol{l}^T\boldsymbol{\beta} - \boldsymbol{l}^T\hat{\boldsymbol{\beta}}}{\sqrt{s^2\boldsymbol{l}^T\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{l}}} \sim t_{n-p}$.

much about the *marginal* and *partial* effect issues. However for one factor, when we decide between a higher and lower order model, as in the case of linear versus quadratic example above, it should be borne in mind that what we are really interested in is this *partial* effect of higher order terms over and above the presence of the lower order terms, and not the effect of the higher order terms as such, which we are hardly interested in, and the t-test would precisely test for this *partial* effect.

Next let us look at the issue of usefulness of the model. This issue was investigated in terms of the ANOVA F-test in §3.1. Here also we shall do the same. But here for the model $Y = \beta_1 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$ the model being useless is expressed in terms of the null hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$. Note that for $k = 1$ as in §3.1, this is same as $H_0 : \beta_1 = 0$, which is what we tested there. And since there was only one X in §3.1, numerically the ANOVA F-test was equivalent to the two-sided t-test for H_{a3}. But in general the question of usefulness of the model is assessed in terms of the totality of all the terms present in the model. This in this case translates to the hypothesis $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ which states that the model is not serving any useful purpose.

The logic of the ANOVA F-test for testing $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ is exactly same as in §3.1. The total amount of variability in the observed Y values are given by $SSTotal = S_{yy} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \boldsymbol{Y}^T(\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{J}_n)\boldsymbol{Y}$, where $\boldsymbol{J}_n$ is an $n \times n$ matrix of all 1's, which has in total $n-1$ d.f. The amount of variability that is explained by the model is given by the amount of variability in the fitted or predicted values $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n$ or $\hat{\boldsymbol{Y}}$. As in §3.1 here also it can be shown that the mean of these $\hat{\boldsymbol{Y}}$ values $\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i = \overline{Y}$. Thus the amount of variability in the observed Y values $\boldsymbol{Y}$ that has been explained by the model or Sum of Squares due to Regression is given by $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = \boldsymbol{Y}^T(\boldsymbol{H} - \frac{1}{n}\boldsymbol{J}_n)\boldsymbol{Y}$. Since $\hat{\boldsymbol{Y}}$ values must lie in a k-dimensional plane totally there are $k+1$ d.f. in the $\hat{\boldsymbol{Y}}$ values, but one d.f. is lost for having to estimate their mean in order to compute the SS, resulting in k d.f. for the SSR. The amount of variability that remains unexplained is measured by the variability in the residual values $\boldsymbol{e}$, where $\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$. Now again it can be shown that $\frac{1}{n}\sum_{i=1}^{n} e_i = 0$ resulting in the same SSE formula obtained above as $SSE = \sum_{i=1}^{n} e_i^2 = (\boldsymbol{Y} - \hat{\boldsymbol{Y}})^T(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = \boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}$. That the $e.d.f.$ equals $n-p$ has already been discussed above. As in §3.1 an alternative way of interpreting this $n-p$ d.f. associated with the errors $\boldsymbol{e}$ can be understood by noting the number of constraints that must be satisfied by $\boldsymbol{e}$, which are p many, given by $\sum_{i=1}^{n} e_i = 0$ and $\sum_{i=1}^{n} e_i X_{ij} = 0 \; \forall j = 1, 2, \ldots, k$, or in other words the errors and all the n values of all the X_k's are uncorrelated. These computations and the F-test are typically presented in the following tabular form:

ANOVA Table

Source of Variation	$D.F.$	SS	$MSS = SS/D.F.$	F	p-value
Regression	$k = p-1$	SSR	$MSR = SSR/k$	$F_{obs} = \frac{MSR}{MSE}$	$P(F_{k,n-p} > F_{obs})$
Error	$n-p$	SSE	$MSE = SSE/(n-p)$		
Total	$n-1$	$SSTotal$			

As before $R^2 = \frac{SSR}{SSTotal} = 1 - \frac{SSE}{SSTotal}$, which provides the proportion of variability in $\boldsymbol{Y}$ that has been explained by the model and provides a single-statistic descriptive measure of

how good a job the model is being able to do. The F-statistic can be expressed in terms R^2 as $F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$ and thus the F-test can be equivalently treated as a test of significance for the population value of R^2. If R^2 is small the overall performance of the model is poor, and a large value of R^2 indicates that a major proportion of variability has been explained by the model and thus the model is doing a decent job or is useful. R^2 is also called the Multiple Correlation Coefficient of Y and $X_1, X_2, \ldots, X_k$. It is called a correlation coefficient because the square of the correlation coefficient between $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$ computed using the formula provided in the footnote of page 18 coincides with R^2.

R^2 however has the undesirable property of showing a larger value for a larger model. That is for example, the R^2 of the model $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$ will always be larger than the model $Y = \beta_0 + \beta_1 X + \epsilon$ for the same set of data, no matter what the true relationship between X and Y may be - be it linear, quadratic or something else. This is because adding additional terms always reduces the SSE. This problem is circumvented by considering a slightly different measure called Adjusted-R^2 which is given by $1 - \frac{MSE}{MSTotal} = 1 - \frac{SSE/(n-k-1)}{SSTotal/(n-1)}$. This quantity also takes the number of additional terms that has been thrown into the model into account and thus is a slightly better measure for assessing the usefulness of a model.

Example 3 (Continued): At this stage I detected that it was a "bad" example in the sense that the response `torque` was exactly identical for three different levels of the factor `cm` for all the 9 combination of levels of `sm` and `area`. Thus I pruned it down to 9 observations for the 9 combination of levels of `sm` and `area`. Thus we have a full factorial experiment with two factors each at three levels with no replication. This will not allow us to illustrate modeling interaction, but since there are two factors and three levels each, this gives us ample scope to demonstrate multiple regression with non-linear terms. To begin with we shall re-do the basic ANOVA model and then proceed. In the sequel `t` denotes the 9 (distinct) values of `torque`, `s` denotes (numerical) `sm` and `a` denotes (numerical) `area`.

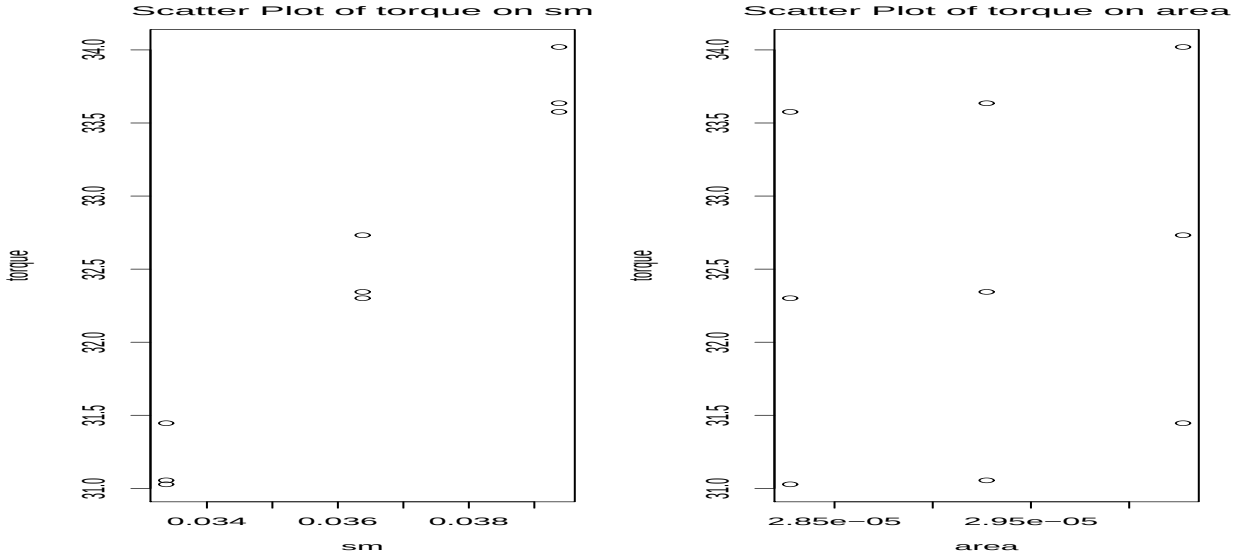
```
> anova(aov(t~as.factor(s)+as.factor(a)))
Analysis of Variance Table

Response: t
             Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(s)  2 9.8901  4.9451 68528.2 8.517e-10 ***
as.factor(a)  2 0.3383  0.1692  2344.1 7.267e-07 ***
Residuals     4 0.0003  0.0001
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Thus we see both the main effects are significant. Thus we begin by fitting the maximal model possible and then drop the insignificant terms as we go along. Since `s` and `a` takes values at three levels, it is possible to fit quadratic models for both the variables. Thus letting `s2` and `a2` respectively denote $\mathtt{s}^2$ and $\mathtt{a}^2$, the maximal model is given by $\mathtt{t} = \beta_0 + \beta_1 \mathtt{s} + \beta_2 \mathtt{s2} + \beta_3 \mathtt{a} + \beta_4 \mathtt{a2} + \epsilon$.

```
> lms2a2<-lm(t~s+s2+a+a2)
> summary(lms2a2)
```

```
Call:
lm(formula = t ~ s + s2 + a + a2)

Residuals:
         1          2          3          4          5          6          7
 9.644e-03 -2.222e-05 -9.622e-03 -6.422e-03  1.111e-05  6.411e-03 -3.222e-03
         8          9
 1.111e-05  3.211e-03

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.583e+02  5.219e+00  30.335 7.03e-06 ***
s            4.222e+02  4.857e+01   8.692 0.000965 ***
s2           7.963e+01  6.674e+02   0.119 0.910781
a           -9.879e+06  3.517e+05 -28.090 9.56e-06 ***
a2           1.724e+11  6.007e+09  28.704 8.77e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.008495 on 4 degrees of freedom
Multiple R-Squared:     1,      Adjusted R-squared: 0.9999
F-statistic: 3.544e+04 on 4 and 4 DF,  p-value: 2.389e-09
```

In this model we see that the quadratic term involving `s2` is insignificant while that involving `a2` is. This is not surprising because of the following scatter-plots of the values of `t` against `s` and `a`.

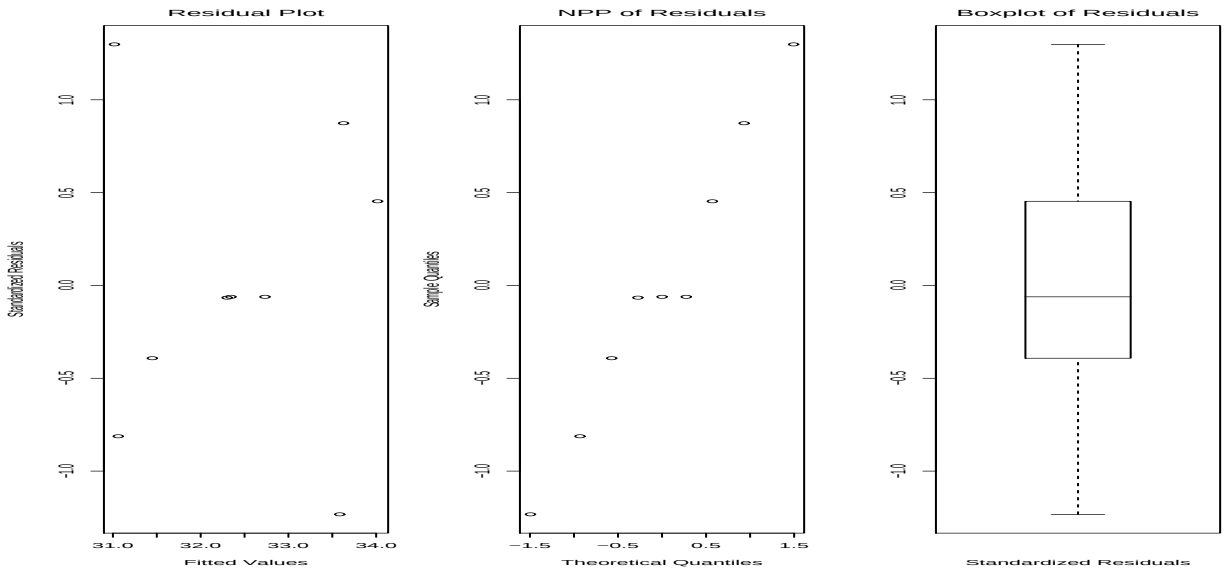


Now we refit the model by dropping the insignificant terms as follows.

```
> lmsa2<-lm(t~s+a+a2)
```

```
> summary(lmsa2)

Call:
lm(formula = t ~ s + a + a2)

Residuals:
         1          2          3          4          5          6          7
 0.0098833 -0.0005000 -0.0093833 -0.0061833 -0.0004667  0.0066500 -0.0029833
         8          9
-0.0004667  0.0034500

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.582e+02  4.610e+00   34.32 3.95e-07 ***
s            4.280e+02  1.036e+00  413.17 1.58e-12 ***
a           -9.879e+06  3.151e+05  -31.35 6.20e-07 ***
a2           1.724e+11  5.382e+09   32.03 5.57e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.007611 on 5 degrees of freedom
Multiple R-Squared:     1,      Adjusted R-squared:     1
F-statistic: 5.885e+04 on 3 and 5 DF,  p-value: 8.694e-12

> anova(lmsa2)
Analysis of Variance Table

Response: t
          Df Sum Sq Mean Sq  F value    Pr(>F)
s          1 9.8901  9.8901 170712.9 1.576e-12 ***
a          1 0.2789  0.2789   4813.3 1.178e-08 ***
a2         1 0.0595  0.0595   1026.2 5.567e-07 ***
Residuals  5 0.0003  0.0001
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This model now looks satisfactory modulo the check on model diagnostics. Towards this end we conduct a residual analysis by constructing the three diagnostic plots involving the standardized residuals of the above linear model `lmsa2` as in §3.2. These plots are given in the next page. As the plots look quite satisfactory we finally take the above model as the final multiple regression model for `torque` in terms of `sm` and `area`.



4 ANCOVA Models

In contrast to the approach of §3, here all the factors or independent variables or the X's are not assumed to be quantitative. There might be some pure qualitative factors assuming some discrete levels, which may be labeled using some numbers but it would be meaningless to let these labeled numbers themselves appear in the model equation. The way such qualitative variables are handled are using the so-called dummy variable approach. In this approach, the different levels of a qualitative factor are handled using 0-1 valued variables. For example if a factor X assumes two possible levels say A and B, then to distinguish these two levels we introduce a dummy variable D as $D = \begin{cases} 1 & \text{if } X \text{ assumes the value } A \\ 0 & \text{otherwise} \end{cases}$. If X assumes three levels A, B and C then to distinguish between these three levels we need two dummy variables D_1 and D_2 as follows. Let $D_1 = \begin{cases} 1 & \text{if } X \text{ assumes the value } A \\ 0 & \text{otherwise} \end{cases}$ and $D_2 = \begin{cases} 1 & \text{if } X \text{ assumes the value } B \\ 0 & \text{otherwise} \end{cases}$. Then by looking at the combination of values of (D_1, D_2) we will know which level of X are we talking about. For instance, a (1,0) value for (D_1, D_2) would indicate that X has level A, a (0,1) value for (D_1, D_2) would indicate that X has level B, while a (0,0) value for (D_1, D_2) would indicate that X has level C.

Next by keeping the quantitative variables as they are, one proceeds to build a multiple regression model with the quantitative variables and the dummy variables introduced for distinguishing the levels of the qualitative factors. Though the basic procedure of model building and the inference techniques are exactly same as in multiple regression of §3.2, there are some subtle nuances in terms of interpretation of the parameters in ANCOVA models. Instead of attempting to address the general nature of these models, we shall learn this model using a case study example, which will bring out all the concepts that are required for building ANCOVA models. The extra theoretical developments will be introduced as and when required while discussing the case study.

Example 2 (Continued): In this problem we are interested in modeling Load on column joint of steering, say Y in terms of three quantitative factors Tilt Mass, say X_1, IC Steering, say X_2, and Mass Torso, say X_3 and two qualitative factors Column Stroke Stiffness, say X_4, and Lower Column Revo joint Stiffness, say X_5. X_1 and X_2 have been experimented with three levels while X_3 have been experimented with five levels. The qualitative factor X_4 assumes three possible levels whileX_4 assumes two possible levels. Thus in total there are $3 \times 3 \times 5 \times 3 \times 2 = 270$ treatments. A full factorial design with no replication was employed for the experimentation.

Thus we begin with a maximal ANOVA model, which in this case will allow us to estimate all interaction effects up to and including the fourth order. The only interaction we shall not be able to estimate is the five-factor interaction. The result of fitting such an ANOVA model is given below:

```
x1                          ***
x2                          *
x3                          ***
x4                          ***
x5                          *
x1:x2
x1:x3
x2:x3                       ***
x1:x4                       ***
x2:x4                       ***
x3:x4                       ***
x1:x5                       ***
x2:x5
x3:x5
x4:x5                       ***
x1:x2:x3
x1:x2:x4
x1:x3:x4
x2:x3:x4                    ***
x1:x2:x5
x1:x3:x5                    *
x2:x3:x5
x1:x4:x5                    ***
x2:x4:x5
x3:x4:x5
x1:x2:x3:x4
x1:x2:x3:x5                 ***
x1:x2:x4:x5
x1:x3:x4:x5
x2:x3:x4:x5
Residuals
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From this maximal model we see that all the main effects are significant, and six out of ten two-factor interactions are also significant. Though three out of the ten three-factor interactions and one four-factor interaction is significant, to keep things tractable, interpretable and for the purpose of illustration, we shall demonstrate the example using terms only up to two-factor interactions. This is no loss of generality because the higher order terms can be accommodated exactly in the same fashion following the method described below.

Now while considering models with just two-factor interactions, it would not be wise to consider all the terms which came out to be significant in the maximal model above. This is because if we ignore third and higher order interactions, the SS corresponding to these effects get absorbed in the SSE which might result in a larger MSE. Since the MSS of the main effects and two-factor interactions remain unchanged in a smaller model, this might result in some insignificant effects in a model with only two-factor interactions. Thus we refit a model with just two factor interactions as follows.

```
Analysis of Variance Table

Response: y
               Df       Sum Sq     Mean Sq     F value    Pr(>F)
x1              2        2.490       1.245     53.2527    < 2.2e-16   ***
x2              2        0.002       0.001      0.0402    0.9605760
x3              4        0.024       0.006      0.2587    0.9041197
x4              2       46.641      23.320    997.6705    < 2.2e-16   ***
x5              1        0.001       0.001      0.0365    0.8486565
x1:x2           4    2.015e-04   5.037e-05      0.0022    0.9999907
x1:x3           8    3.437e-04   4.296e-05      0.0018    1.0000000
x1:x4           4        0.159       0.040      1.7052    0.1499971
x1:x5           2        6.522       3.261    139.5150    < 2.2e-16   ***
x2:x3           8        0.138       0.017      0.7372    0.6585230
x2:x4           4        0.111       0.028      1.1851    0.3183484
x2:x5           2    2.222e-06   1.111e-06   4.753e-05    0.9999525
x3:x4           8        0.654       0.082      3.4955    0.0008176   ***
x3:x5           4    8.370e-05   2.093e-05      0.0009    0.9999984
x4:x5           2        0.036       0.018      0.7805    0.4595046
Residuals     212        4.955       0.023
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From this analysis we see that the interaction between only X_1 and X_5 and X_3 and X_4 are significant, while the main-effects are significant only for X_1 and X_4. This somewhat reduces the search space of all possible models, among which we are to choose the best one. First, since X_2 does not figure in any of the main-effect or interaction terms we may safely drop it for further analysis. Note that even for the (proper) maximal model with terms up to four-factor interaction, the level of significance of the main-effect of X_2 was small compared to others. Thus indeed there would be no loss in dropping X_2 from further consideration for the ANCOVA model building for Y. Second, we shall restrict ourselves in considering the interaction terms involving only X_1 & X_5 and X_3 & X_4. Third, though the main-effects of

X_3 and X_5 are not significant, their interactions with other factors are. Thus we should not ignore them in building the ANCOVA model for Y. If their main-effect terms turn out to be insignificant in the process then be it, but they must receive our initial consideration.

As mentioned in the beginning of this section, the qualitative variables X_4 and X_5 will now be handled using 0-1 valued dummy variables. Since X_4 has three levels, -1, 0 and 1 we need two dummies to handle it. Thus define $D_{-1} = \begin{cases} 1 & \text{if } X_4 \text{ assumes the value } -1 \\ 0 & \text{otherwise} \end{cases}$ and $D_1 = \begin{cases} 1 & \text{if } X_4 \text{ assumes the value 1} \\ 0 & \text{otherwise} \end{cases}$. Likewise for two possible levels of X_5, -1 and 0 define $(DX)_5 = \begin{cases} 1 & \text{if } X_5 \text{ assumes the value } -1 \\ 0 & \text{otherwise} \end{cases}$. The quantitative variable X_1 has value at three possible levels. Thus the maximal model may have a quadratic term in X_1. Similarly since the quantitative variable X_3 has value at five possible levels, the maximal model may have up to fourth degree terms in X_3. Accounting for the two-factor interactions between X_1 & X_5 and X_3 & X_4 the maximal model may thus be written as $Y = \beta_0+\beta_1X_1+\beta_2X_1^2+\beta_3X_3+\beta_4X_3^2+\beta_5X_3^3+\beta_6X_3^4+\beta_7D_{-1}+\beta_8D_1+\beta_9(DX_5)+\beta_{10}(DX_5)X_1+\beta_{11}(DX_5)X_1^2+\beta_{12}D_{-1}X_3+\beta_{13}D_{-1}X_3^2+\beta_{14}D_{-1}X_3^3+\beta_{15}D_{-1}X_3^4+\beta_{16}D_1X_3+\beta_{17}D_1X_3^2+\beta_{18}D_1X_3^3+\beta_{19}D_1X_3^4+\epsilon$. What this ANCOVA model does is it depicts different kinds of regression relationships between Y and the two quantitative variables X_1 and X_3 for different combinations of levels of the two qualitative variables X_4 and X_5, which has been expressed in terms of their corresponding dummies D_{-1}, D_1 and $(DX)_5$. Thus for the six possible combinations of levels of X_4 and X_5 the regression relationships of Y on X_1 and X_3 for this maximal model are as follows:

X_4	X_5	Regression of Y on X_1 and X_3
0	0	$Y = \beta_0 + \beta_1X_1 + \beta_2X_1^2 + \beta_3X_3 + \beta_4X_3^2 + \beta_5X_3^3 + \beta_6X_3^4 + \epsilon$
0	-1	$Y = (\beta_0 + \beta_9) + (\beta_1 + \beta_{10})X_1 + (\beta_2 + \beta_{11})X_1^2 + \beta_3X_3 + \beta_4X_3^2 + \beta_5X_3^3 + \beta_6X_3^4 + \epsilon$
-1	0	$Y = (\beta_0 + \beta_7) + \beta_1X_1 + \beta_2X_1^2 + (\beta_3 + \beta_{12})X_3 + (\beta_4 + \beta_{13})X_3^2 + (\beta_5 + \beta_{14})X_3^3 + (\beta_6 + \beta_{15})X_3^4 + \epsilon$
-1	-1	$Y = (\beta_0 + \beta_7 + \beta_9) + (\beta_1 + \beta_{10})X_1 + (\beta_2 + \beta_{11})X_1^2 + (\beta_3 + \beta_{12})X_3 + (\beta_4 + \beta_{13})X_3^2 + (\beta_5 + \beta_{14})X_3^3 + (\beta_6 + \beta_{15})X_3^4 + \epsilon$
1	0	$Y = (\beta_0 + \beta_8) + \beta_1X_1 + \beta_2X_1^2 + (\beta_3 + \beta_{16})X_3 + (\beta_4 + \beta_{17})X_3^2 + (\beta_5 + \beta_{18})X_3^3 + (\beta_6 + \beta_{19})X_3^4 + \epsilon$
1	-1	$Y = (\beta_0 + \beta_8 + \beta_9) + (\beta_1 + \beta_{10})X_1 + (\beta_2 + \beta_{11})X_1^2 + (\beta_3 + \beta_{16})X_3 + (\beta_4 + \beta_{17})X_3^2 + (\beta_5 + \beta_{18})X_3^3 + (\beta_6 + \beta_{19})X_3^4 + \epsilon$

Now we are to build an appropriate model containing only the significant terms from the above model involving 19 terms. There are several ways by which one may approach this problem. They may be broadly classified as automated methods and intuitive methods. I shall demonstrate only the automated approach. However among the automated methods also there are several approaches. Out of these I shall demonstrate only one called the stepwise regression, which has been implemented in the AutoDOE. Again among stepwise regression procedures also there are two primary choices called backward elimination and forward selection. In backward elimination we start with the maximal model and then keep dropping terms one at a time, starting with the one which is least significant, till we arrive at

a model with all terms being significant. This is the most popular choice of model selection and has been implemented in the AutoDOE. In contrast, in the forward selection we start with the smallest possible model with no X's and then keep adding terms one at a time, starting with the one which is most significant, till we arrive at a model with all terms being significant with no more addition of significant terms being possible. This usually involves going back a step at a time checking for the significance of the old terms in presence of the newly added term and thus adds complexity in the search procedure. Thus we shall not discuss the forward selection procedure. Moreover since we are starting with a maximal model, by first fitting an ANOVA model, the natural choice of stepwise regression in our case would be the backward elimination. There is a third more full-proof approach of considering all possible models, which for the example at hand would involve fitting $2^{19} = 524,288$ or more than half a million models, which by no means is a practical solution and thus is not even attempted.

By fitting the full model involving all the 19 terms we get

```
lm(formula = y ~ x1 + x12 + x3 + x32 + x33 + x34 + d_1 + d1 +
    dx5 + dx5x1 + dx5x12 + d_1x3 + d_1x32 + d_1x33 + d_1x34 +
    d1x3 + d1x32 + d1x33 + d1x34)
Residuals:
      Min        1Q    Median        3Q       Max
-0.314815 -0.108843 -0.003370  0.102185  0.384185
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.4880568 20.3843980   0.956    0.340
x1          -1.1804444  0.1084696 -10.883   <2e-16 ***
x12          0.3237778  0.0268393  12.064   <2e-16 ***
x3           3.7130822  7.4625786   0.498    0.619
x32         -0.5293156  1.0062004  -0.526    0.599
x33          0.0334429  0.0592677   0.564    0.573
x34         -0.0007849  0.0012879  -0.609    0.543
d_1          4.7007370 28.8275784   0.163    0.871
d1          20.2162256 28.8275784   0.701    0.484
dx5         -2.2128889  0.1350880 -16.381   <2e-16 ***
dx5x1        2.6431111  0.1533991  17.230   <2e-16 ***
dx5x12      -0.6593333  0.0379564 -17.371   <2e-16 ***
d_1x3       -1.6616910 10.5536798  -0.157    0.875
d_1x32       0.2359875  1.4229822   0.166    0.868
d_1x33      -0.0148235  0.0838171  -0.177    0.860
d_1x34       0.0003455  0.0018214   0.190    0.850
d1x3        -7.3528866 10.5536798  -0.697    0.487
d1x32        1.0580557  1.4229822   0.744    0.458
d1x33       -0.0674336  0.0838171  -0.805    0.422
d1x34        0.0015944  0.0018214   0.875    0.382
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.147 on 250 degrees of freedom
Multiple R-Squared: 0.9125,     Adjusted R-squared: 0.9058
F-statistic: 137.2 on 19 and 250 DF,  p-value: < 2.2e-16
```

with $D_{-1}X_3$ being the least significant. Thus we next fit a model by just dropping this term as follows:

```
lm(formula = y ~ x1 + x12 + x3 + x32 + x33 + x34 + d_1 + d1 +
    dx5 + dx5x1 + dx5x12 + d_1x32 + d_1x33 + d_1x34 + d1x3 +
    d1x32 + d1x33 + d1x34)

Residuals:
      Min        1Q    Median        3Q       Max
-0.314815 -0.108476 -0.003132  0.101762  0.384185

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.176e+01  1.439e+01   1.512    0.132
x1          -1.180e+00  1.083e-01 -10.904   <2e-16 ***
x12          3.238e-01  2.679e-02  12.087   <2e-16 ***
x3           2.882e+00  5.267e+00   0.547    0.585
x32         -4.173e-01  7.104e-01  -0.587    0.557
x33          2.686e-02  4.190e-02   0.641    0.522
x34         -6.420e-04  9.125e-04  -0.704    0.482
d_1          1.639e-01  8.826e-01   0.186    0.853
d1           1.795e+01  2.492e+01   0.720    0.472
dx5         -2.213e+00  1.348e-01 -16.413   <2e-16 ***
dx5x1        2.643e+00  1.531e-01  17.264   <2e-16 ***
dx5x12      -6.593e-01  3.788e-02 -17.405   <2e-16 ***
d_1x32       1.204e-02  4.337e-02   0.278    0.782
d_1x33      -1.650e-03  5.041e-03  -0.327    0.744
d_1x34       5.989e-05  1.612e-04   0.371    0.711
d1x3        -6.522e+00  9.122e+00  -0.715    0.475
d1x32        9.461e-01  1.230e+00   0.769    0.443
d1x33       -6.085e-02  7.249e-02  -0.839    0.402
d1x34        1.452e-03  1.576e-03   0.921    0.358
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1467 on 251 degrees of freedom
Multiple R-Squared: 0.9125,     Adjusted R-squared: 0.9062
F-statistic: 145.4 on 18 and 251 DF,  p-value: < 2.2e-16
```

Now the least significant term is D_{-1} and hence it gets dropped at this step yielding the next model

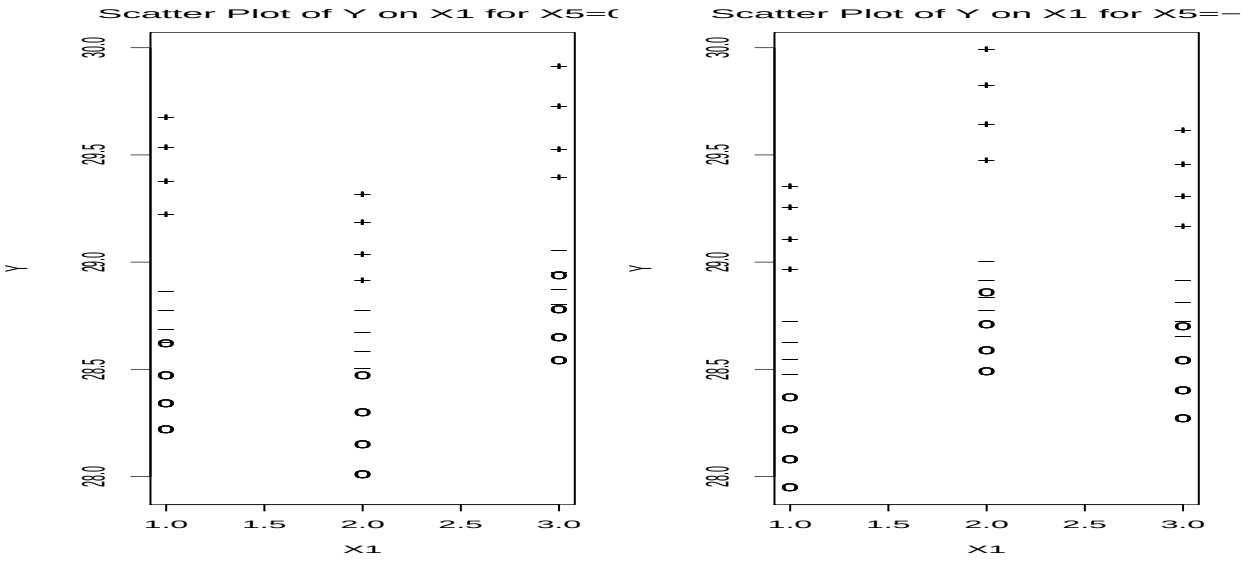
```
lm(formula = y ~ x1 + x12 + x3 + x32 + x33 + x34 + d1 + dx5 +
    dx5x1 + dx5x12 + d_1x32 + d_1x33 + d_1x34 + d1x3 + d1x32 +
    d1x33 + d1x34)

Residuals:
      Min        1Q    Median        3Q       Max
-0.314815 -0.106367 -0.002037  0.101818  0.384185

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.184e+01  1.436e+01   1.521 0.129528
x1          -1.180e+00  1.081e-01 -10.925  < 2e-16 ***
x12          3.238e-01  2.674e-02  12.110  < 2e-16 ***
x3           2.882e+00  5.256e+00   0.548 0.583957
x32         -4.213e-01  7.088e-01  -0.594 0.552723
x33          2.732e-02  4.175e-02   0.654 0.513503
x34         -6.566e-04  9.073e-04  -0.724 0.469924
d1           1.787e+01  2.487e+01   0.718 0.473179
dx5         -2.213e+00  1.346e-01 -16.445  < 2e-16 ***
dx5x1        2.643e+00  1.528e-01  17.297  < 2e-16 ***
dx5x12      -6.593e-01  3.781e-02 -17.438  < 2e-16 ***
d_1x32       2.004e-02  5.310e-03   3.773 0.000201 ***
d_1x33      -2.572e-03  8.792e-04  -2.926 0.003752 **
d_1x34       8.909e-05  3.562e-05   2.502 0.012999 *
d1x3        -6.522e+00  9.104e+00  -0.716 0.474436
d1x32        9.501e-01  1.228e+00   0.774 0.439691
d1x33       -6.131e-02  7.231e-02  -0.848 0.397320
d1x34        1.466e-03  1.571e-03   0.933 0.351680
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1464 on 252 degrees of freedom
Multiple R-Squared: 0.9125,     Adjusted R-squared: 0.9066
F-statistic: 154.5 on 17 and 252 DF,  p-value: < 2.2e-16
```

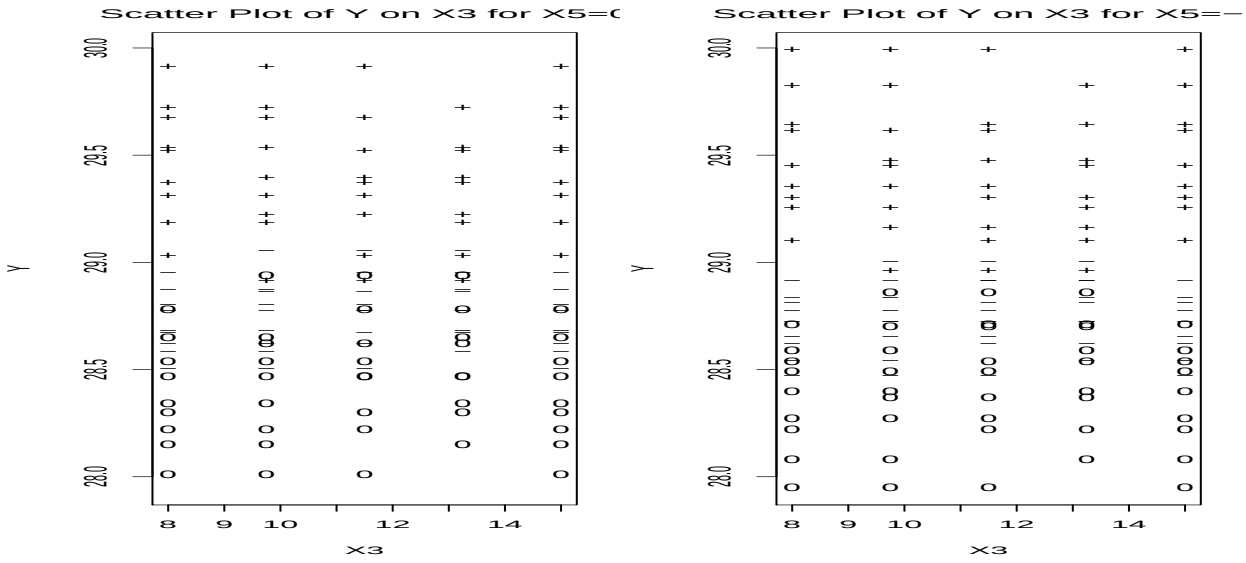
Proceeding in this manner we arrive at a model with all significant terms in just six steps, with $\alpha = 0.05$ which is as follows:

```
lm(formula = y ~ x1 + x12 + x33 + x34 + dx5 + dx5x1 + dx5x12 +
    d_1x32 + d_1x33 + d_1x34 + d1x32 + d1x33 + d1x34)

Residuals:
      Min        1Q    Median        3Q       Max
-0.347168 -0.108510 -0.003795  0.106093  0.387807

Coefficients:
```

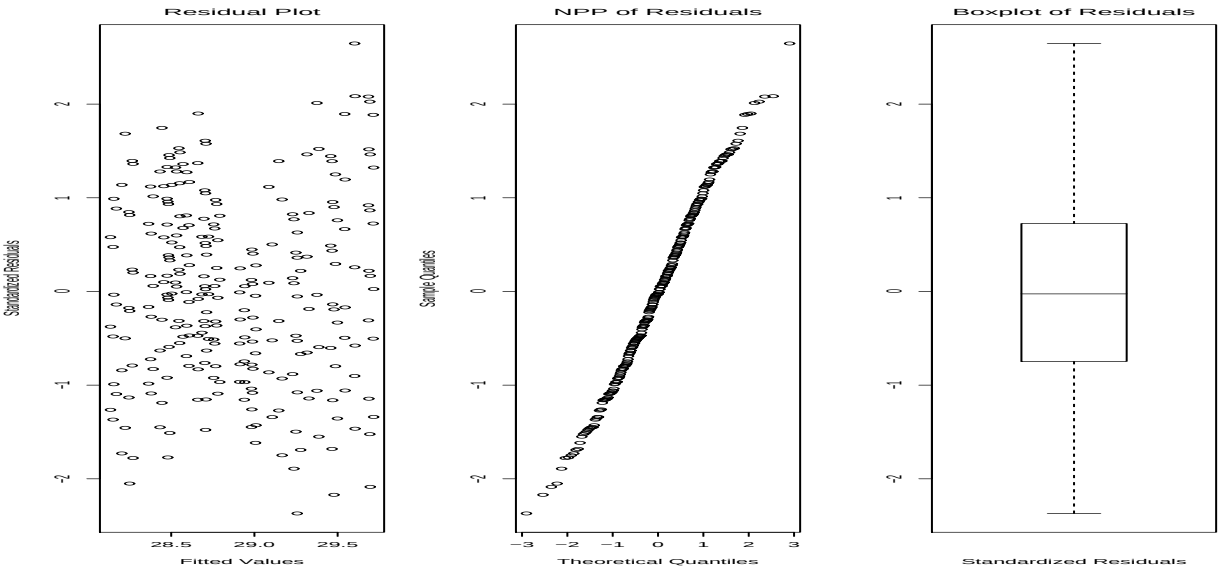
```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.906e+01  1.210e-01 240.125  < 2e-16 ***
x1          -1.180e+00  1.081e-01 -10.923  < 2e-16 ***
x12          3.238e-01  2.674e-02  12.108  < 2e-16 ***
x33          6.007e-04  1.927e-04   3.117 0.002039 **
x34         -3.663e-05  1.177e-05  -3.112 0.002067 **
dx5         -2.213e+00  1.346e-01 -16.441  < 2e-16 ***
dx5x1        2.643e+00  1.528e-01  17.293  < 2e-16 ***
dx5x12      -6.593e-01  3.782e-02 -17.434  < 2e-16 ***
d_1x32       1.855e-02  5.253e-03   3.532 0.000489 ***
d_1x33      -2.328e-03  8.698e-04  -2.676 0.007926 **
d_1x34       7.932e-05  3.524e-05   2.251 0.025266 *
d1x32        7.251e-02  5.253e-03  13.804  < 2e-16 ***
d1x33       -9.693e-03  8.698e-04 -11.143  < 2e-16 ***
d1x34        3.458e-04  3.524e-05   9.811  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1465 on 256 degrees of freedom
Multiple R-Squared: 0.911,      Adjusted R-squared: 0.9065
F-statistic: 201.7 on 13 and 256 DF,  p-value: < 2.2e-16
```

To have a final appreciation of this model we begin with the scatter plots of the quantitative variables for different levels of X_4 and X_5. In these plots the three levels of X_4, -1, 0 and +1 have been indicated by the plotting symbols '-', '0' and '+' respectively.





These plots show the non-linearity in Y for changing values of X_1 and X_3 as well as the interaction between X_1 & X_5 and X_3 & $X-4$. Next we perform a residual analysis to check whether the model assumptions have been satisfied through the following plots.



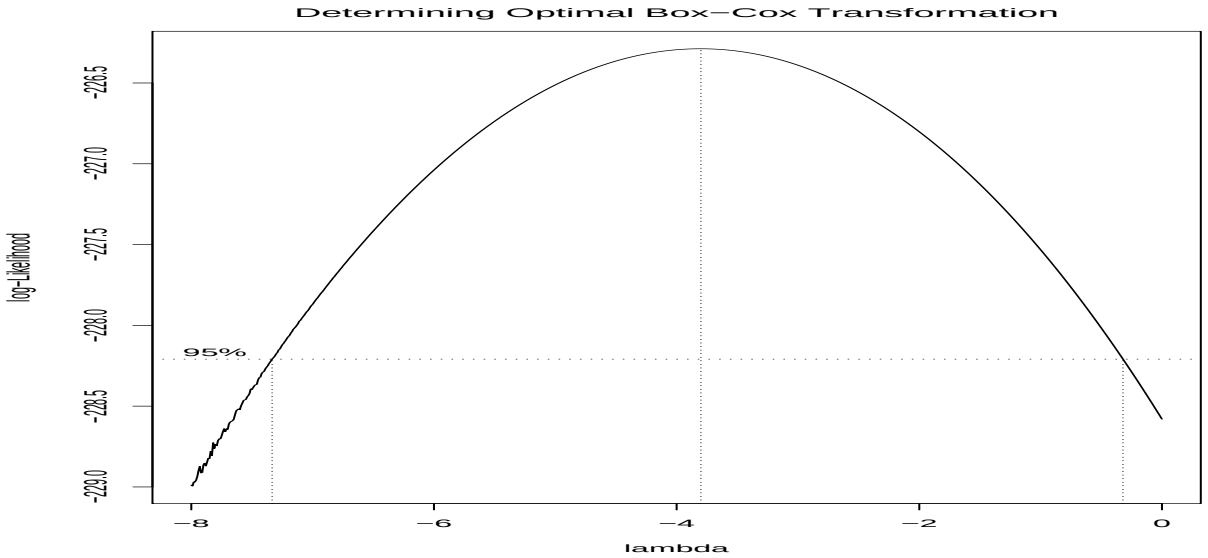
Not only the plots look very satisfactory, they also indeed survive the objective statistical tests for homoscedasticity (Bartlett's test) and Normality as well. Thus they satisfy all the regression assumptions and thus we may feel satisfied with the above model and stop our model building effort here to proceed with the next stage of optimization using the above model.

Transformations:

However for the purpose of demonstration we shall take up the issue of what to do in case the model assumptions are not satisfied. The nature of non-linearity that is to be considered in the independent variable in our case is chiefly determined by the levels of the quantitative factors and thus there is no need to discuss that issue any further. Furthermore as mentioned in point 1 of page 19 for §3.1, this issue very much depends on the nature of the residual

plot, which indeed looks so good in this example that we cannot take it up any further in that direction.

However sometimes one may need to transform the response Y to satisfy the homoscedasticity and Normality assumptions. A general class of these transformations is called the Box-Cox transformation where one tries to model $\frac{Y^{\lambda}-1}{\lambda}$ instead of Y, where the optimal value of λ is determined by its likelihood function. A non-zero value of λ gives a power (and hence sometimes Box-Cox transformations are also called power transformation) of Y while for $\lambda = 0$ it is same as considering $\log(Y)$. The likelihood of λ of the Box-Cox transformation for the above data set with the final model is as follows:

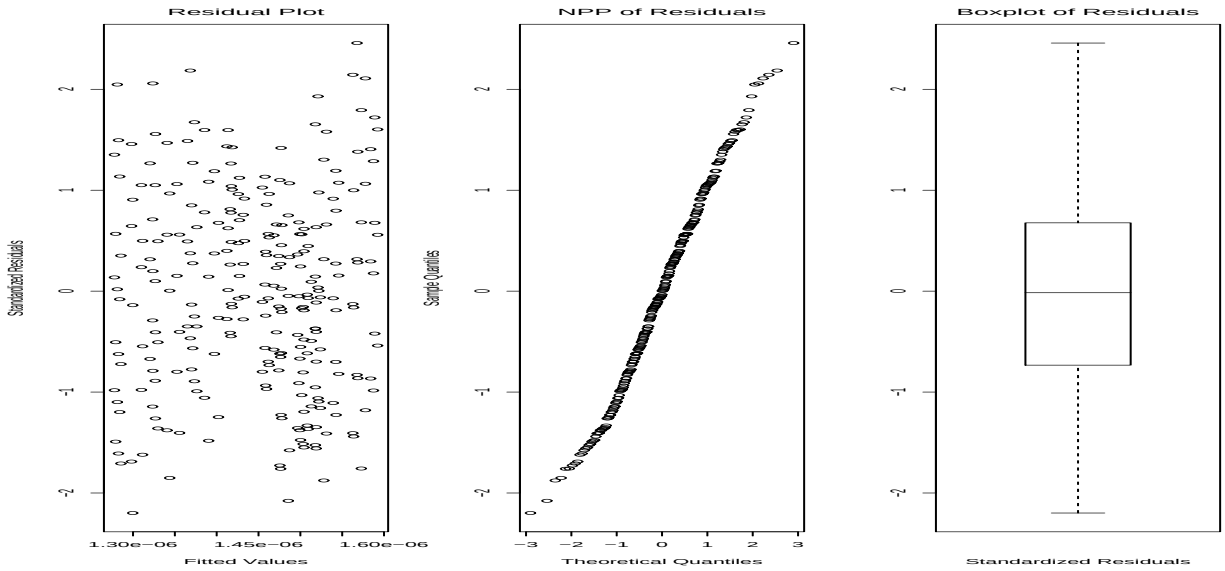


which says that an "optimal" value of this transformation would be -4, or it advises us to model $1/Y^4$ instead of Y itself. The summary of this model is as follows:

```
lm(formula = y^(-4) ~ x1 + x12 + x33 + x34 + dx5 + dx5x1 + dx5x12 +
    d_1x32 + d_1x33 + d_1x34 + d1x32 + d1x33 + d1x34)

Residuals:
       Min         1Q     Median         3Q        Max
-6.345e-08 -2.109e-08 -4.538e-10  1.953e-08  7.095e-08

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.409e-06  2.385e-08  59.085  < 2e-16 ***
x1           2.350e-07  2.129e-08  11.038  < 2e-16 ***
x12         -6.446e-08  5.268e-09 -12.234  < 2e-16 ***
x33         -1.284e-10  3.797e-11  -3.381 0.000835 ***
x34          7.827e-12  2.319e-12   3.375 0.000853 ***
dx5          4.397e-07  2.652e-08  16.580  < 2e-16 ***
dx5x1       -5.234e-07  3.011e-08 -17.383  < 2e-16 ***
dx5x12       1.303e-07  7.451e-09  17.495  < 2e-16 ***
d_1x32      -4.024e-09  1.035e-09  -3.888 0.000129 ***
```

```
d_1x33       5.087e-10  1.714e-10   2.969 0.003275 **
d_1x34      -1.743e-11  6.944e-12  -2.511 0.012662 *
d1x32       -1.430e-08  1.035e-09 -13.819  < 2e-16 ***
d1x33        1.910e-09  1.714e-10  11.148  < 2e-16 ***
d1x34       -6.813e-11  6.944e-12  -9.812  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.886e-08 on 256 degrees of freedom
Multiple R-Squared: 0.9111,     Adjusted R-squared: 0.9066
F-statistic: 201.9 on 13 and 256 DF,  p-value: < 2.2e-16
```

The residual analysis of this transformed model is as follows:



This model is indeed an improvement over the previous model, though the amount of improvement is nominal, because the original model in terms of Y itself had already satisfied all the required assumptions.