

# MATH38141 Regression Analysis Coursework

Name: Josh Mottley

May 11, 2020

## Introduction

A chemist has reported that adding naphthenic oil and filler can be used to control the viscosity of elastomer blends. I have been given data of various viscosities with the amount of oil and filler added in. It is believed that the viscosity follows a normal distribution with homogenous variance for any oil and filler level within the design region. I will be analysing the data given and creating 2 regression models that fits the data given. After this I will try to make a statistical guess on whether the a new idea by a chemist of a viscosity with a given oil/filler is correct. I will then reanalyse my 2 models, with there associated results with the statistical programming language R, and then compare the results between manual vs programming. Finally I will draw a conclusion from the original statement of the chemist.

For this report I will be using bold letters to represents vectors (e.g:  **$\mathbf{A}$** ), bold and underlined letters to show matrices (e.g:  **$\underline{\mathbf{B}}$** ), and a hat to show estimates (e.g:  $\hat{\beta}$ ). A full example may look like:

$$\hat{\mathbf{y}} = \hat{\beta} \underline{\mathbf{X}}$$

I will also be using R for the matrix calculations.

Futhermore when using reference, if the reference has the layout Listing-(#) where # is a number, then the reference is referring to the appendix section of the report. Otherwise it has been stated previously.

# 1 Analyse of data without R

Please note that the matrix calculation are done in R. If you wish to see the programming please refer to Listing-(1).

## 1.1 Creating a regression model

### 1.1.1 Linear regression

When creating the model, I choose the viscosity to be the response variable, and the naphthenic oil and filler as the input variables. I will take into account a constant term, main effects for naphthenic oil and for filler, and an interaction term. This gives our model as:

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + \beta_3 \mathbf{X}_1 \mathbf{X}_2 + \boldsymbol{\varepsilon} \quad (1)$$

where  $\mathbf{X}_1$  is the vector of samples for the naphthenic oil,  $\mathbf{X}_2$  is the vector of samples for the filler,  $\boldsymbol{\varepsilon}$  is the error vector,  $\mathbf{y}$  is the vector of samples for the corresponding viscosity, and lastly  $\beta_i$  are our model parameters. Note that the errors  $\varepsilon_i$  are assumed to be independently distributed, with 0 mean  $\mu$  and homogeneous variance  $\sigma^2$ . We can also write this in its matrix form:

$$\mathbf{y} = \underline{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where:  $\underline{\mathbf{X}} = (\mathbf{1} \quad \mathbf{X}_1 \quad \mathbf{X}_2 \quad \mathbf{X}_1 \mathbf{X}_2)$ ,  $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$ .

So now using the data given in the Viscos.txt the first 3 rows of the matrix  $\underline{\mathbf{X}}$ , with its correspond response variable  $\mathbf{y}$  would be:

$$\underline{\mathbf{X}} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 12 & 0 \\ 1 & 0 & 24 & 0 \end{pmatrix}, \mathbf{y} = \begin{pmatrix} 6.5 \\ 9.5 \\ 12.50 \end{pmatrix}$$

Now we want to estimate the values of  $\boldsymbol{\beta}$  which we will call  $\hat{\boldsymbol{\beta}}$ , so we will use the formula:

$$\hat{\boldsymbol{\beta}} = (\underline{\mathbf{X}}^T \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^T \mathbf{y}$$

First we will work out  $(\underline{\mathbf{X}}^T \underline{\mathbf{X}})^{-1}$ :

$$\underline{\mathbf{X}}^T \underline{\mathbf{X}} = \begin{pmatrix} 23 & 330 & 720 & 10800 \\ 330 & 7500 & 10800 & 252000 \\ 720 & 10800 & 31680 & 475200 \\ 10800 & 252000 & 475200 & 11088000 \end{pmatrix} \quad (2)$$

$$\Rightarrow (\underline{\mathbf{X}}^T \underline{\mathbf{X}})^{-1} = \begin{pmatrix} 0.3839957035 & -1.831364 \times 10^{-2} & -0.0087271751 & 4.162191 \times 10^{-4} \\ -0.0183136412 & 1.437522 \times 10^{-3} & 0.0004162191 & -3.267096 \times 10^{-5} \\ -0.0087271751 & 4.162191 \times 10^{-4} & 0.0002867287 & -1.324740 \times 10^{-5} \\ 0.0004162191 & -3.267096 \times 10^{-5} & -0.0000132474 & 9.950471 \times 10^{-7} \end{pmatrix} \quad (3)$$

Next we will find the value of  $\underline{\mathbf{X}}^T \mathbf{y}$ :

$$\underline{\mathbf{X}}^T \mathbf{y} = \begin{pmatrix} 301 \\ 3315 \\ 12783 \\ 142920 \end{pmatrix}$$

Which means our final solution to our estimate of the coefficient vector  $\hat{\boldsymbol{\beta}}$  is:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\underline{\mathbf{X}}^T \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^T \mathbf{y} \\ &= \begin{pmatrix} 0.3839957035 & -1.831364 \times 10^{-2} & -0.0087271751 & 4.162191 \times 10^{-4} \\ -0.0183136412 & 1.437522 \times 10^{-3} & 0.0004162191 & -3.267096 \times 10^{-5} \\ -0.0087271751 & 4.162191 \times 10^{-4} & 0.0002867287 & -1.324740 \times 10^{-5} \\ 0.0004162191 & -3.267096 \times 10^{-5} & -0.0000132474 & 9.950471 \times 10^{-7} \end{pmatrix} \begin{pmatrix} 301 \\ 3315 \\ 12783 \\ 142920 \end{pmatrix} \\ \hat{\boldsymbol{\beta}} &= \begin{pmatrix} 2.79954350 \\ -0.09582438 \\ 0.52482098 \\ -0.01015172 \end{pmatrix} \end{aligned}$$

Which will result in our estimated fitted model (when provided with the given data) now looking like:

$$\hat{\mathbf{y}} = \begin{pmatrix} 1 & \mathbf{X}_1 & \mathbf{X}_2 & \mathbf{X}_1 \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} 2.79954350 \\ -0.09582438 \\ 0.52482098 \\ -0.01015172 \end{pmatrix}$$

### 1.1.2 Estimating the variance of the response

To find an estimate of the variance of the response of my model, I will find the LS estimate for  $\sigma^2$ . To do this, we must first calculate the SSE of my model, given by:

$$\begin{aligned} \text{SSE} &= \mathbf{y}^T \mathbf{y} - \hat{\boldsymbol{\beta}}^T \underline{\mathbf{X}}^T \mathbf{y} \\ &= 5918.125 - 5782.908 \\ &= 135.2173 \end{aligned} \tag{4}$$

Hence my estimate our  $\sigma^2$  is:

$$\begin{aligned} \sigma^2 &= \frac{\text{SSE}}{n - p} \\ &= \frac{\text{SSE}}{23 - 4} = \frac{135.2173}{19} \\ &= 7.116698 \end{aligned} \tag{5}$$

### 1.1.3 Finding the coefficient of determination $R^2$

To find the coefficient of determination I first need to find the value of  $SST_C$ . This can be given by the equation:  $SST_C = SST - ny^2$ . To value of the SST is given by:

$$\begin{aligned} SST &= \sum_{i=1}^n y_i^2 \\ &= \mathbf{y}^T \mathbf{y} \\ &= 5918.125 \end{aligned} \tag{6}$$

Next we need to work out the mean  $\bar{y}$  with the following:

$$\begin{aligned} \bar{y} &= \frac{\sum_{i=1}^n y_i}{n} \\ &= \frac{\sum_{i=1}^{23} y_i}{23} \\ &= \frac{301}{23} \\ &= 13.08696 \end{aligned} \tag{7}$$

So finally the value of  $SST_C$  with (6) and (7) subbed in is:

$$\begin{aligned} SST_C &= SST - n\bar{y}^2 \\ &= 5918.125 - 23(13.08696)^2 \\ &= 1978.951 \end{aligned} \tag{8}$$

Now to find the value of the coefficient of determination, sub (8) and (4) into the following:

$$\begin{aligned} R^2 &= \frac{SST_C - SSE}{SST_C} \\ &= \frac{1978.951 - 135.2173}{1978.951} \\ &= 0.9316723 \end{aligned} \tag{9}$$

The value of  $R^2$  is very close to 1, which suggests that that the model is very successful in predicting the observed values of  $Y$ .

### 1.1.4 Testing for if the true values of the model parameters equal to zero

Test for significance for each model parameters, i.e:  $H_0 : \beta_i = \mathbf{0}, H_1 : \beta_i \neq \mathbf{0}$ . I choose  $\alpha = 1 - \gamma = 2.5\%$  significance test. The critical value  $t_{n-p,\gamma} = t_{19,0.975} = 2.093$ . Use the test statistic equation to find the t-value for each parameter:

$$\frac{\hat{\beta}_i - c_i}{\hat{\sigma}\sqrt{g^{ii}}} = \frac{\hat{\beta}_i - 0}{\sqrt{7.116698}\sqrt{g^{ii}}} = \frac{\hat{\beta}_i}{\sqrt{7.116698}\sqrt{g^{ii}}} \sim t_{19}$$

Note that  $g^{ii}$  is the  $i$ th diagonal element of the matrix  $\mathbf{G}^{-1} = (\mathbf{X}^T \mathbf{X})^{-1}$  which was figured out in (3). Next find the test statistic of each parameter and compare against the critical value:

$$\frac{\hat{\beta}_0}{\sqrt{7.116698}\sqrt{g^{11}}} = \frac{2.79954350}{\sqrt{7.116698}\sqrt{0.3839957035}} = 1.6934984 \tag{10}$$

$$\frac{\hat{\beta}_1}{\sqrt{7.116698}\sqrt{g^{22}}} = \frac{-0.09582438}{\sqrt{7.116698}\sqrt{1.437522 \times 10^{-3}}} = -0.9473915 \quad (11)$$

$$\frac{\hat{\beta}_2}{\sqrt{7.116698}\sqrt{g^{33}}} = \frac{0.52482098}{\sqrt{7.116698}\sqrt{0.0002867287}} = 11.6181327 \quad (12)$$

$$\frac{\hat{\beta}_3}{\sqrt{7.116698}\sqrt{g^{44}}} = \frac{-0.01015172}{\sqrt{7.116698}\sqrt{9.950471 \times 10^{-7}}} = -3.8148590 \quad (13)$$

While the critical value  $t_{19,0.975} = 2.093 > |1.6934984|, |-0.9473915|$  which mean that (10) and (11) pass, (12) and (13) do not pass as  $t_{19,0.975} = 2.093 < |11.6181327|, |-3.8148590|$ . Therefore we reject  $H_0$  and conclude that the model parameters are significant/non-zero.

## 1.2 Creating a quadratic model

Please note that the matrix calculation are done in R. If you wish to see the programming please refer to Listing-(2).

### 1.2.1 Quadratic regression

When creating a quadratic regression model, I will use the same random input variables as when creating the linear regression in (1). Instead I will use the following model for my regression:

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + \beta_3 \mathbf{X}_1 \mathbf{X}_2 + \beta_4 \mathbf{X}_1^2 + \beta_5 \mathbf{X}_2^2 + \boldsymbol{\varepsilon} \quad (14)$$

In matrix form this would look like the following:

$$\mathbf{y} = \underline{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where:  $\underline{\mathbf{X}} = (\mathbf{1} \quad \mathbf{X}_1 \quad \mathbf{X}_2 \quad \mathbf{X}_1 \mathbf{X}_2 \quad \mathbf{X}_1^2 \quad \mathbf{X}_2^2)$ ,  $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix}$ .

Now to find estimates of  $\hat{\boldsymbol{\beta}}$  we will use the same method as with the linear model. So we want to first work out the value of  $(\underline{\mathbf{X}}^T \underline{\mathbf{X}})^{-1}$ :

$$\begin{aligned} \underline{\mathbf{X}}^T \underline{\mathbf{X}} &= \begin{pmatrix} 23 & 330 & 720 & 10800 & 7500 & 31680 \\ 330 & 7500 & 10800 & 252000 & 189000 & 475200 \\ 720 & 10800 & 31680 & 475200 & 252000 & 1555200 \\ 10800 & 252000 & 475200 & 11088000 & 6480000 & 23328000 \\ 7500 & 189000 & 252000 & 6480000 & 5070000 & 11088000 \\ 31680 & 475200 & 1555200 & 23328000 & 11088000 & 81202176 \end{pmatrix} \\ \Rightarrow (\underline{\mathbf{X}}^T \underline{\mathbf{X}})^{-1} &= \begin{pmatrix} 0.4829853691 & -2.814401 \times 10^{-2} & -1.778522 \times 10^{-2} & 3.493342 \times 10^{-4} & 4.257768 \times 10^{-4} & 1.583991 \times 10^{-4} \\ -0.0281440079 & 4.817593 \times 10^{-3} & 3.141439 \times 10^{-4} & -2.602882 \times 10^{-5} & -1.224108 \times 10^{-4} & 9.632377 \times 10^{-7} \\ -0.0177852211 & 3.141439 \times 10^{-4} & 1.532919 \times 10^{-3} & -7.127103 \times 10^{-6} & -5.573936 \times 10^{-6} & -2.144988 \times 10^{-5} \\ 0.0003493342 & -2.602882 \times 10^{-5} & -7.127103 \times 10^{-6} & 1.040240 \times 10^{-6} & -2.876870 \times 10^{-7} & -1.070264 \times 10^{-7} \\ 0.0004257768 & -1.224108 \times 10^{-4} & -5.573936 \times 10^{-6} & -2.876870 \times 10^{-7} & 4.502301 \times 10^{-6} & 1.248641 \times 10^{-7} \\ 0.0001583991 & 9.632377 \times 10^{-7} & -2.144988 \times 10^{-5} & -1.070264 \times 10^{-7} & 1.248641 \times 10^{-7} & 3.693898 \times 10^{-7} \end{pmatrix} \quad (15) \end{aligned}$$

Now find the vector  $\underline{\mathbf{X}}^T \mathbf{y}$ :

$$\underline{\mathbf{X}}^T \mathbf{y} = \begin{pmatrix} 301 \\ 3315 \\ 12783 \\ 142920 \\ 70400 \\ 640044 \end{pmatrix} \quad (16)$$

Now we can find the estimates of  $\hat{\beta}$ , sub (15) and (16) into the following:

$$\begin{aligned} \hat{\beta} &= (\underline{\mathbf{X}}^T \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^T \mathbf{y} \\ &= \begin{pmatrix} 0.4829853691 & -2.814401 \times 10^{-2} & -1.778522 \times 10^{-2} & 3.493342 \times 10^{-4} & 4.257768 \times 10^{-4} & 1.583991 \times 10^{-4} \\ -0.0281440079 & 4.817593 \times 10^{-3} & 3.141439 \times 10^{-4} & -2.602882 \times 10^{-5} & -1.224108 \times 10^{-4} & 9.632377 \times 10^{-7} \\ -0.0177852211 & 3.141439 \times 10^{-4} & 1.532919 \times 10^{-3} & -7.127103 \times 10^{-6} & -5.573936 \times 10^{-6} & -2.144988 \times 10^{-5} \\ 0.0003493342 & -2.602882 \times 10^{-5} & -7.127103 \times 10^{-6} & 1.040240 \times 10^{-6} & -2.876870 \times 10^{-7} & -1.070264 \times 10^{-7} \\ 0.0004257768 & -1.224108 \times 10^{-4} & -5.573936 \times 10^{-6} & -2.876870 \times 10^{-7} & 4.502301 \times 10^{-6} & 1.248641 \times 10^{-7} \\ 0.0001583991 & 9.632377 \times 10^{-7} & -2.144988 \times 10^{-5} & -1.070264 \times 10^{-7} & 1.248641 \times 10^{-7} & 3.693898 \times 10^{-7} \end{pmatrix} \begin{pmatrix} 301 \\ 3315 \\ 12783 \\ 142920 \\ 70400 \\ 640044 \end{pmatrix} \\ \hat{\beta} &= \begin{pmatrix} 6.016644748 \\ -0.206571079 \\ 0.143467750 \\ -0.012325436 \\ 0.006879675 \\ 0.006597411 \end{pmatrix} \quad (17) \end{aligned}$$

Which will result in our fitted quadratic model (with provided data) now looking like:

$$\hat{\mathbf{y}} = \begin{pmatrix} 1 & \mathbf{X}_1 & \mathbf{X}_2 & \mathbf{X}_1 \mathbf{X}_2 & \mathbf{X}_1^2 & \mathbf{X}_2^2 \end{pmatrix} \begin{pmatrix} 6.016644748 \\ -0.206571079 \\ 0.143467750 \\ -0.012325436 \\ 0.006879675 \\ 0.006597411 \end{pmatrix}$$

### 1.2.2 Estimating the variance of the response

Similary to the linear model, I will use a similar method to find the estimate of the variance of the response. So first we need to calculate the SSE of the model:

$$\begin{aligned} \text{SSE} &= \mathbf{y}^T \mathbf{y} - \hat{\beta}^T \underline{\mathbf{X}}^T \mathbf{y} \\ &= 5918.125 - 5905.587 \\ &= 12.53844 \quad (18) \end{aligned}$$

Hence, substitute (18) into the equation for  $\sigma^2$  to find our LS estimate:

$$\begin{aligned} \sigma^2 &= \frac{\text{SSE}}{n - p} \\ &= \frac{12.53844}{23 - 5} \\ &= 0.7375553 \quad (19) \end{aligned}$$

The estimate for  $\sigma^2$  for our quadratic model (19) is much smaller than the estimate from our linear model (5). The reason for this is that the fit of the quadratic model is much closer to the data points given in the sample than for the linear model, resulting in a much smaller SSE, leading to a much smaller estimate of  $\sigma^2$ .

### 1.2.3 Finding the coefficient of determination $R^2$

To find the coefficient of determination, we can use the value of the  $SST_C$  that we worked out in (8) previously. Therefore the coefficient of determination is now:

$$\begin{aligned} R^2 &= \frac{SST_C - SSE}{SST_C} \\ &= \frac{1978.951 - 12.53844}{1978.951} \\ &= 0.9936641 \end{aligned} \quad (20)$$

The coefficient of determination  $R^2$  is very close to 1, which suggests that the model is very successful in predicting the observed values of  $y$ . The coefficient of determination for the quadratic model (20) is greater than the linear model (9), which suggests that the quadratic model is more successful in predicting the observed values of  $y$ . We can say that there is a  $\frac{(20)-(9)}{1} * 100 = \frac{0.9936641-0.9316723}{1} * 100 = 6.19918\%$  increase in prediction of the observed values for the coefficient of determination of the quadratic model.

### 1.2.4 Testing for if the true values of the model parameters equal to zero

Test for significance for each model parameters, i.e:  $H_0 : \beta_i = \mathbf{0}, H_1 : \beta_i \neq \mathbf{0}$ . I choose  $\alpha = 1 - \gamma = 2.5\%$  significance test. The critical value  $t_{n-p, \gamma} = t_{17, 0.975} = 2.110$ . Use the test statistic equation to find the t-value for each parameter:

$$\frac{\hat{\beta}_i - c_i}{\hat{\sigma} \sqrt{g^{ii}}} = \frac{\hat{\beta}_i - 0}{\sqrt{0.7375553} \sqrt{g^{ii}}} = \frac{\hat{\beta}_i}{\sqrt{0.7375553} \sqrt{g^{ii}}} \sim t_{19}$$

Note that  $g^{ii}$  is the  $i$ th diagonal element of the matrix  $\mathbf{G}^{-1} = (\mathbf{X}^T \mathbf{X})^{-1}$  which was figured out in (15). Next find the test statistic of each parameter and compare against the critical value:

$$\frac{\hat{\beta}_0}{\sqrt{0.7375553} \sqrt{g^{11}}} = \frac{6.016644748}{\sqrt{0.7375553} \sqrt{0.4829853691}} = 10.080686 \quad (21)$$

$$\frac{\hat{\beta}_1}{\sqrt{0.7375553} \sqrt{g^{22}}} = \frac{-0.206571079}{\sqrt{0.7375553} \sqrt{4.817593 \times 10^{-3}}} = -3.465431 \quad (22)$$

$$\frac{\hat{\beta}_2}{\sqrt{0.7375553} \sqrt{g^{33}}} = \frac{0.143467750}{\sqrt{0.7375553} \sqrt{1.532919 \times 10^{-3}}} = 4.266752 \quad (23)$$

$$\frac{\hat{\beta}_3}{\sqrt{0.7375553} \sqrt{g^{44}}} = \frac{-0.012325436}{\sqrt{0.7375553} \sqrt{1.040240 \times 10^{-6}}} = -14.071432 \quad (24)$$

$$\frac{\hat{\beta}_4}{\sqrt{0.7375553} \sqrt{g^{55}}} = \frac{0.006879675}{\sqrt{0.7375553} \sqrt{4.502301 \times 10^{-6}}} = 3.775316 \quad (25)$$

$$\frac{\hat{\beta}_5}{\sqrt{0.7375553}\sqrt{g^{66}}} = \frac{0.006597411}{\sqrt{0.7375553}\sqrt{0.006597411}} = 12.639615 \quad (26)$$

(21), (22), (23), (24), (25), (26) do not pass as:

$$t_{19,0.975} = 2.093 < |10.080686|, |-3.465431|, |4.266752|, |-14.071432|, |3.775316|, |12.639615|.$$

Therefore we reject  $H_0$  and conclude that the model parameters are significant/non-zero.

### 1.2.5 Comparison between the 2 models

When comparing between the linear model and the quadratic model, we have found that the quadratic model has a much smaller estimated L.S variance  $\sigma^2$  than the linear models estimate, which means that the data points are much closer to the quadratic regression model on average than the linear model. Furthermore, the quadratic model coefficient of determination  $R^2$  is closer to 1 than the linear models, suggesting that the model is more successful at predicting the observed values. Therefore it appears that the quadratic model is a superior fit for the given data than the linear model, and hence is more likely to have a higher accuracy when predicting future unknown values of viscosity for Oil/Filler input.

## 1.3 Prediction using confidence intervals

I will choose to use the quadratic regression model (14) when working out the confidence intervals. “An elastomer blend with viscosity equal to 21 M is required. A chemist believes that this can be achieved by using 10 phr oil and 50 phr filler.”

### 1.3.1 95% confidence interval for the mean viscosity of elastomer blends manufactured as suggested by the chemist

To create this confidence interval, I will use a  $(1 - \alpha)100\%$  confidence interval for mean  $y_0$ , i.e:

$$\mathbf{f}_0^T \hat{\boldsymbol{\beta}} \pm t_{n-p, 1-\frac{\alpha}{2}} \hat{\sigma} \sqrt{\mathbf{f}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{f}_0} \quad (27)$$

The value of  $\mathbf{f}_0 = \begin{pmatrix} 1 \\ 10 \\ 50 \\ 500 \\ 100 \\ 2500 \end{pmatrix}$ , which uses the input data given by the chemist, that is believed to be true.

We already know the values for the quadratic regression model of  $\hat{\boldsymbol{\beta}}$  (17),  $\hat{\sigma}$  (19) and  $(\mathbf{X}^T \mathbf{X})^{-1}$  (15).



So that gives us the confidence interval as:

$$\begin{aligned}
& \mathbf{f}_0^T \hat{\boldsymbol{\beta}} \pm t_{n-p, 1-\frac{\alpha}{2}} \hat{\sigma} \sqrt{\mathbf{f}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{f}_0} \\
& \Rightarrow \begin{pmatrix} 1 \\ 10 \\ 50 \\ 500 \\ 100 \\ 2500 \end{pmatrix}^T \begin{pmatrix} 6.016644748 \\ -0.206571079 \\ 0.143467750 \\ -0.012325436 \\ 0.006879675 \\ 0.006597411 \end{pmatrix} \pm t_{23-6, 1-\frac{0.05}{2}} \\
& \cdot 0.7375553 \sqrt{\begin{pmatrix} 1 \\ 10 \\ 50 \\ 500 \\ 100 \\ 2500 \end{pmatrix}^T \begin{pmatrix} 0.4829853691 & -2.814401 \times 10^{-2} & -1.778522 \times 10^{-2} & 3.493342 \times 10^{-4} & 4.257768 \times 10^{-4} & 1.583991 \times 10^{-4} \\ -0.0281440079 & 4.817593 \times 10^{-3} & 3.141439 \times 10^{-4} & -2.602882 \times 10^{-5} & -1.224108 \times 10^{-4} & 9.632377 \times 10^{-7} \\ -0.0177852211 & 3.141439 \times 10^{-4} & 1.532919 \times 10^{-3} & -7.127103 \times 10^{-6} & -5.573936 \times 10^{-6} & -2.144988 \times 10^{-5} \\ 0.0003493342 & -2.602882 \times 10^{-5} & -7.127103 \times 10^{-6} & 1.040240 \times 10^{-6} & -2.876870 \times 10^{-7} & -1.070264 \times 10^{-7} \\ 0.0004257768 & -1.224108 \times 10^{-4} & -5.573936 \times 10^{-6} & -2.876870 \times 10^{-7} & 4.502301 \times 10^{-6} & 1.248641 \times 10^{-7} \\ 0.0001583991 & 9.632377 \times 10^{-7} & -2.144988 \times 10^{-5} & -1.070264 \times 10^{-7} & 1.248641 \times 10^{-7} & 3.693898 \times 10^{-7} \end{pmatrix} \begin{pmatrix} 1 \\ 10 \\ 50 \\ 500 \\ 100 \\ 2500 \end{pmatrix}} \\
& \Rightarrow 22.1431 \pm 2.110 \cdot 0.7375553 \sqrt{0.1446243} \\
& \Rightarrow 22.1431 \pm 0.5918309 \\
& \Rightarrow (21.55127, 22.73493) \tag{28}
\end{aligned}$$

As the required value of viscosity is 21M, which exists outside of the 95% confidence interval (28), then we reject the chemist guess that this can be achieved by using 10 phr oil and 50 phr filler.

### 1.3.2 Calculating an interval that will contain the measured value 95% of the time

To create an interval that will contain the measured value 95% of the time, I will use a  $(1 - \alpha)100\%$  prediction interval given by the following:

$$\mathbf{f}_0^T \hat{\boldsymbol{\beta}} \pm t_{n-p, 1-\frac{\alpha}{2}} \hat{\sigma} \sqrt{1 + \mathbf{f}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{f}_0} \tag{29}$$

As we have used most of the values when working out the confidence interval, then our prediction interval is:

$$\begin{aligned}
& \mathbf{f}_0^T \hat{\boldsymbol{\beta}} \pm t_{n-p, 1-\frac{\alpha}{2}} \hat{\sigma} \sqrt{1 + \mathbf{f}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{f}_0} \\
& \Rightarrow \mathbf{f}_0^T \hat{\boldsymbol{\beta}} \pm t_{23-6, 1-\frac{0.05}{2}} \hat{\sigma} \sqrt{1 + \mathbf{f}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{f}_0} \\
& \Rightarrow 22.1431 \pm 2.110 \cdot 0.7375553 \sqrt{1 + 0.1446243} \\
& \Rightarrow 22.1431 \pm 1.664978 \\
& \Rightarrow (20.47812, 23.80808) \tag{30}
\end{aligned}$$

Therefore the the prediction interval (30) will contain the measured viscosity using the settings the chemist provided 95% of the time.

## 2 Analyse of data with R

For this section I will be using R functions to analyse the data using the same 2 models in the previous section.

### 2.1 Recreating the linear model, with R analysis

So for this I will use the regression model in (1). The code used for this is given under Listing-(3). The result from R is then given in Listing-(4). From the summary, under the “Estimate” column, we can conclude that the estimate of  $\beta$  is:

$$\hat{\beta} = \begin{pmatrix} 2.799544 \\ -0.095824 \\ 0.524821 \\ -0.010152 \end{pmatrix} \quad (31)$$

This would result in the estimated linear model now looking like:

$$\hat{y} = (1 \quad X_1 \quad X_2 \quad X_1 X_2) \begin{pmatrix} 2.799544 \\ -0.095824 \\ 0.524821 \\ -0.010152 \end{pmatrix}$$

#### 2.1.1 Find the estimate of the variance of response

The various estimates of variance for the different explanatory/response variables can be shown in the anova table under the Listing-(4). As we want to find the estimate of the variance of response, we look at the value of the “Residuals” row, under the “Mean Sq” column. This gives us that  $\hat{\sigma}^2 = 7.12$ . Below the anov table we can find a more precise value of  $\hat{\sigma}^2 = 7.116698$ .

#### 2.1.2 Finding the coefficient of determination

The coefficient of determination  $R^2$  can be found in the summary table under Listing-(4). The value can be given in the line “Multiple R-squared: 0.9317”, i.e giving  $R^2 = 0.9317$ . As  $R^2$  is very close to 1, then we can say that the model in R is very succesful in predicting the observed values of the response.

#### 2.1.3 Testing for if the true values of the model parameters equal to zero

Test for significance for each model parameters, i.e:  $H_0 : \beta_i = 0, H_1 : \beta_i \neq 0$ . I choose  $\alpha = 1 - \gamma = 2.5\%$  significance test. The critical value  $t_{n-p,\gamma} = t_{19,0.975} = 2.093$ . The t values for each input variable is given in the “t value” column of the summary table in Listing-(4). While the critical value  $t_{19,0.975} = 2.093 > |1.693|, |-0.947|$ , so the (Intercept) and “tab\$Oil” pass, “tab\$Filler” and “I(tab\$Oil \* tab\$Filler)” do not pass as  $t_{19,0.975} = 2.093 < |11.618|, |-3.815|$ . Therefore we reject  $H_0$  and conclude that the model parameters are significant.

## 2.2 Recreating the quadratic model, with R analysis

So for this I will use the regression model in (14). The code used for this is given under Listing-(5). The result from R is then given in Listing-(6). From the summary, under the “Estimate” column, we can conclude that the estimate of  $\beta$  is:

$$\hat{\beta} = \begin{pmatrix} 6.0166447 \\ -0.2065711 \\ 0.1434677 \\ -0.0123254 \\ 0.0068797 \\ 0.0065974 \end{pmatrix} \quad (32)$$

This would result in the estimated quadratic model now looking like:

$$\mathbf{y} = \begin{pmatrix} 1 & \mathbf{X}_1 & \mathbf{X}_2 & \mathbf{X}_1\mathbf{X}_2 & \mathbf{X}_1^2 & \mathbf{X}_2^2 \end{pmatrix} \begin{pmatrix} 6.0166447 \\ -0.2065711 \\ 0.1434677 \\ -0.0123254 \\ 0.0068797 \\ 0.0065974 \end{pmatrix}$$

### 2.2.1 Find the estimate of the variance of response

The various estimates of variance for the different explanatory/response can be shown in the anova table under the Listing-(6). As we want to find the estimate of the variance of response, we look at the value of the “Residuals” row, under the “Mean Sq” column. This gives us that  $\hat{\sigma}^2 = 0.74$ . Below the anov table we can find a more precise value of  $\hat{\sigma}^2 = 0.7375553$ . The variance estimate of response for the quadratic model is smaller than the linear model, because the quadratic model has a much closer fit to the data points given in the sample than the linear model.

### 2.2.2 Finding the coefficient of determination

The coefficient of determination  $R^2$  can be found in the summary table under Listing-(6). The value can be given in the line “Multiple R-squared: 0.9937”, i.e giving  $R^2 = 0.9937$ . As  $R^2$  is very close to 1, then we can say that the model in R is very succesful in predicting the observed values of the response.

### 2.2.3 Testing for if the true values of the model parameters equal to zero

Test for significance for each model parameters, i.e:  $H_0 : \beta_i = \mathbf{0}, H_1 : \beta_i \neq \mathbf{0}$ . I choose  $\alpha = 1 - \gamma = 2.5\%$  significance test. The critical value  $t_{n-p,\gamma} = t_{17,0.975} = 2.110$ . The t values for each input variable is given in the “t value” column of the summary table in Listing-(6). None of the input variables / (Intercept) pass as:  $t_{19,0.975} = 2.093 < |10.081|, |-3.465|, |4.267|, |-14.071|, |3.775|, |12.640|$ .

Therefore we reject  $H_0$  and conclude that the model parameters are significant.

## 2.3 Prediction using confidence intervals

I will choose to use the quadratic regression model (14) when working out the confidence intervals. “An elastomer blend with viscosity equal to 21 M is required. A chemist believes that this can be achieved by using 10 phr oil and 50 phr filler.”

### 2.3.1 95% confidence interval for the mean viscosity of elastomer blends manufactured as suggested by the chemist

The 95% confidence interval can be found in the Listing-(6), as the variable “ci”. This gives us a confidence interval of: (21.45403, 22.83217). As the chemist guess of 21M for the viscosity exists outside of the confidence interval then we reject his proposal that it can be achieved by using 10 phr oil and 50 phr filler.

### 2.3.2 Calculating an interval that will contain the measured value 95% of the time

The 95% confidence interval can be found in the Listing-(6), as the variable “pi”. This gives us a prediction interval of: (20.20457, 24.08163). Therefore the measured viscosity, when using the settings provided by the chemist, will be in the interval (20.20457, 24.08163) 95% of the time.

### 3 Conclusion

In conclusion, I believe that the chemist is correct in the fact that by adding naphthenic oil (phr) and filler (phr), you can control the viscosity (M) of elastomer blends. My reasoning for this is that I have found that both the linear and quadratic estimated models exist, and succesful can predict the observed values of viscosity, as shown with both coefficient of determination being greater than 0.9, with the quadratic model have a even greater coefficient of determination than the linear models. Futhermore, I am 95% confident that the true model parameters are not equal to zero for the linear/quadratic models as shown by previous tests, leading to fact that a model does indeed exist for estimating the viscosity by the amount of naphthenic oil and filler added, for the linear and quadratic models. I am confident in saying that quadratic model is a better fit for the given data than the linear model, due to its much smaller estimated response variance and its higher coefficient of determination.

## 4 Appendix

Listing 1: Linear regression equation method

```
1 # Set working directory
2 dir<-getwd()
3 if(!is.null(dir)) setwd(dir) else stop("Working directory is incorrect")
4
5 # Load data as a table
6 tab<-read.table("Viscos.txt", header = TRUE)
7
8 # Create X matrix
9 X<-cbind(
10   rep(1, time = nrow(tab)),
11   tab$Oil,
12   tab$Filler,
13   tab$Oil*tab$Filler
14 )
15
16 # Get/store transpose of X
17 XT<-t(X)
18
19 # Get value of n
20 n<-nrow(X)
21
22 # Times transpose by X
23 prod<-XT %*% X
24
25 # Find the inverse of prod
26 invProd<-solve(prod)
27
28 # Create y response vector
29 y<-tab$Visc
30
31 # Calculate beta vector
32 beta<-invProd%*%XT%*%y
33
34 # Calculate SSE
35 SSE<-t(y)%*%y-t(beta)%*%XT%*%y
36
37 # Calculate estimate of variance
38 var<-SSE/(nrow(X) - ncol(X))
39
40 # Calculate the SST
41 SST<-t(y)%*%y
42
43 # Calculate mean of y
44 my = mean(y)
45
46 # Calculate value of SST_C
47 SST_C<-SST-n*(my^2)
48
49 # Calculate coefficient of regression
50 R<-(SST_C-SSE)/SST_C
51
52 # Get diagonal of (X^TX)^-1
53 dia<-diag(invProd)
54
55 # Get values of test statistics
56 ts<-beta/(sqrt(var)[1]*sqrt(dia))
```

Listing 2: Quadratic regression equation method

```
1 # Set working directory
2 dir<-getwd()
3 if(!is.null(dir)) setwd(dir) else stop("Working directory is incorrect")
4
5 # Load data as a table
6 tab<-read.table("Viscos.txt", header = TRUE)
7
8 # Create X matrix
9 X<-cbind(
10   rep(1, time = nrow(tab)),
11   tab$Oil,
12   tab$Filler,
```

```

13 |   tab$Oil*tab$Filler,
14 |   tab$Oil^2,
15 |   tab$Filler^2
16 | )
17 |
18 | # Get/store transpose of X
19 | XT<-t(X)
20 |
21 | # Get value of n
22 | n<-nrow(X)
23 |
24 | # Times transpose by X
25 | prod<-XT %*% X
26 |
27 | # Find the inverse of prod
28 | invProd<-solve(prod)
29 |
30 | # Create y response vector
31 | y<-tab$Visc
32 |
33 | # Get value of X^Ty
34 | XTy<-XT%*%y
35 |
36 | # Calculate beta vector
37 | beta<-invProd%*%XTy
38 |
39 | # Calculate SSE
40 | SSE<-t(y)%*%y-t(beta)%*%XT%*%y
41 |
42 | # Calculate estimate of variance
43 | var<-SSE/(nrow(X) - ncol(X))
44 |
45 | # Calculate the SST
46 | SST<-t(y)%*%y
47 |
48 | # Calculate mean of y
49 | my = mean(y)
50 |
51 | # Calculate value of SST_C
52 | SST_C<-SST-n*(my^2)
53 |
54 | # Calculate coefficient of regression
55 | R<-(SST_C-SSE)/SST_C
56 |
57 | # Get diagonal of (X^TX)^-1
58 | dia<-diag(invProd)
59 |
60 | # Get values of test statistics
61 | ts<-beta/(sqrt(var)[1]*sqrt(dia))
62 |
63 | # Store guessed values from chemist
64 | f_0=c(1, 10, 50, 500, 100, 2500)
65 |
66 | # Calculate f_0%*%beta
67 | ciMean<-f_0%*%beta
68 |
69 | # Calculate f_0(X^TX)^-1f_0
70 | ciP<-f_0%*%invProd%*%f_0
71 |
72 | # 97.5% t distro
73 | ciT<-2.110
74 |
75 | # Calculate CI +-
76 | ciPM<-ciT*var*sqrt(ciP)
77 |
78 | # Calculate confidence interval as 2 vec
79 | ci<-c(ciMean-ciPM, ciMean+ciPM)
80 |
81 | # Calculate PI +-
82 | piPM<-ciT*var*sqrt(1+ciP)
83 |
84 | # Calculate prediction interval as 2 vec
85 | pi<-c(ciMean-piPM, ciMean+piPM)

```

Listing 3: Linear regression summary method

```

1 # Set working directory
2 dir<-getwd()
3 if(!is.null(dir)) setwd(dir) else stop("Working directory is incorrect")
4
5 # Load data as a table
6 tab<-read.table("Viscos.txt", header = TRUE)
7
8 # Fit the model as a linear model
9 model.fit<-lm(tab$Visc ~ tab$Oil + tab$Filler + I(tab$Oil*tab$Filler))
10
11 # Get summary of model
12 summa<-summary(model.fit)
13 summa
14
15 # Get anova of model
16 anov<-anova(model.fit)
17 anov
18
19 # Find variance of response
20 var<-anov[1]$'Mean Sq'[4]
21 var

```

Listing 4: Linear regression summary results

```

1 Call:
2 lm(formula = tab$Visc ~ tab$Oil + tab$Filler + I(tab$Oil * tab$Filler))
3
4 Residuals:
5      Min       1Q   Median       3Q      Max
6 -3.8302 -1.9674 -0.2477  1.9633  4.9612
7
8 Coefficients:
9              Estimate Std. Error t value Pr(>|t|)
10 (Intercept)    2.799544    1.653113   1.693  0.10669
11 tab$Oil        -0.095824    0.101145  -0.947  0.35533
12 tab$Filler      0.524821    0.045173  11.618 4.46e-10 ***
13 I(tab$Oil * tab$Filler) -0.010152    0.002661  -3.815  0.00117 **
14 ---
15 Signif. codes:  0   ***    0.001   **    0.01   *    0.05   .    0.1    1
16
17 Residual standard error: 2.668 on 19 degrees of freedom
18 Multiple R-squared:  0.9317, Adjusted R-squared:  0.9209
19 F-statistic: 86.36 on 3 and 19 DF, p-value: 2.97e-11
20
21 >
22 > # Get anova of model
23 > anov<-anova(model.fit)
24 > anov
25 Analysis of Variance Table
26
27 Response: tab$Visc
28      Df Sum Sq Mean Sq F value    Pr(>F)
29 tab$Oil      1  364.31   364.31   51.191 8.442e-07 ***
30 tab$Filler    1 1375.85  1375.85  193.327 2.074e-11 ***
31 I(tab$Oil * tab$Filler) 1  103.57   103.57   14.553  0.001169 **
32 Residuals   19  135.22     7.12
33 ---
34 Signif. codes:  0   ***    0.001   **    0.01   *    0.05   .    0.1    1
35 >
36 > # Find variance of response
37 > var<-anov[1]$'Mean Sq'[4]
38 > var
39 [1] 7.116698

```

Listing 5: Quadratic regression summary method

```

1 # Delete previous variables
2 rm(list=ls())
3
4 # Set working directory
5 dir<-getwd()
6 if(!is.null(dir)) setwd(dir) else stop("Working directory is incorrect")
7
8 # Load data as a table
9 tab<-read.table("Viscos.txt", header = TRUE)
10

```



```

11 # Store variables from table
12 Visc<-tab$Visc
13 Oil<-tab$Oil
14 Filler<-tab$Filler
15 Interaction<-tab$Oil*tab$Filler
16 OilSquared<-tab$Oil^2
17 FillerSquared<-tab$Filler^2
18
19 # Fit the model as a quadratic model
20 model.fit<-lm(Visc ~ Oil + Filler + Interaction + OilSquared + FillerSquared)
21
22 # Get summary of model
23 summa<-summary(model.fit)
24 summa
25
26 # Get anova of model
27 anov<-anova(model.fit)
28 anov
29
30 # Find variance of response
31 var<-anov[,"Mean Sq"][6]
32 var
33
34 newdata<-data.frame(
35   Oil=10,
36   Filler=50,
37   Interaction = 500,
38   OilSquared = 100,
39   FillerSquared = 2500
40 )
41
42 # Find the confidence interval
43 ci<-predict(model.fit, newdata, interval="confidence", level=0.95)
44 ci
45
46 # Find the prediction interval
47 pi<-predict(model.fit, newdata, interval="prediction", level=0.95)
48 pi

```

Listing 6: Quadratic regression summary results

```

1 Call:
2 lm(formula = Visc ~ Oil + Filler + Interaction + OilSquared +
3   FillerSquared)
4
5 Residuals:
6     Min       1Q   Median       3Q      Max
7 -1.38709 -0.56863 -0.09948  0.65894  1.39761
8
9 Coefficients:
10             Estimate Std. Error t value Pr(>|t|)
11 (Intercept)   6.0166447   0.5968487   10.081 1.38e-08 ***
12 Oil          -0.2065711   0.0596091    -3.465 0.002958 **
13 Filler        0.1434677   0.0336246    4.267 0.000521 ***
14 Interaction  -0.0123254   0.0008759  -14.071 8.50e-11 ***
15 OilSquared    0.0068797   0.0018223    3.775 0.001510 **
16 FillerSquared 0.0065974   0.0005220   12.640 4.53e-10 ***
17 ---
18 Signif. codes:  0   ***    0.001   **    0.01   *    0.05   .    0.1    1
19
20 Residual standard error: 0.8588 on 17 degrees of freedom
21 Multiple R-squared:  0.9937, Adjusted R-squared:  0.9918
22 F-statistic: 533.2 on 5 and 17 DF, p-value: < 2.2e-16
23
24 >
25 > # Get anova of model
26 > anov<-anova(model.fit)
27 > anov
28 Analysis of Variance Table
29
30 Response: Visc
31      Df Sum Sq Mean Sq F value    Pr(>F)
32 Oil    1  364.31   364.31  493.9469 5.303e-14 ***
33 Filler  1 1375.85  1375.85 1865.4200 < 2.2e-16 ***
34 Interaction  1  103.57   103.57  140.4239 1.219e-09 ***
35 OilSquared  1    4.85    4.85    6.5718  0.02014 *

```

```

36| FillerSquared    1   117.83   117.83   159.7599 4.528e-10 ***
37| Residuals       17    12.54     0.74
38| ---
39| Signif. codes:  0   ***    0.001    **    0.01    *    0.05    .    0.1    1
40| >
41| > # Find variance of response
42| > var<-anov[1]$'Mean Sq'[6]
43| > var
44| [1] 0.7375553
45| >
46| > newdata<-data.frame(
47| +   Oil=10,
48| +   Filler=50,
49| +   Interaction = 500,
50| +   OilSquared = 100,
51| +   FillerSquared = 2500
52| + )
53| >
54| > # Find the confidence interval
55| > ci<-predict(model.fit, newdata, interval="confidence", level=0.95)
56| > ci
57|      fit      lwr      upr
58| 1 22.1431 21.45403 22.83217
59| >
60| > # Find the prediction interval
61| > pi<-predict(model.fit, newdata, interval="prediction", level=0.95)
62| > pi
63|      fit      lwr      upr
64| 1 22.1431 20.20457 24.08163

```