MATH38141 Regression Analysis Coursework

Name: Josh Mottley   Student ID: 10136392

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: $\boldsymbol{A}$), bold and underlined letters to show matrices (e.g: $\underline{\boldsymbol{B}}$), and a hat to show estimates (e.g: $\hat{\boldsymbol{\beta}}$). A full example may look like:

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

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 refering 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:

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{X_1} + \beta_2 \boldsymbol{X_2} + \beta_3 \boldsymbol{X_1} \boldsymbol{X_2} + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{X_1}$ is the vector of samples for the naphthenic oil, $\boldsymbol{X_2}$ is the vector of samples for the filler, $\boldsymbol{\varepsilon}$ is the error vector, $\boldsymbol{y}$ is the vector of samples for the corresponding viscosity, and lastly β_i are our model parameters. Note that the errors ε_i are assumed to be independently distributed, with 0 mean μ and homogeneous variance σ^2. We can also write this in its matrix form:

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

where: $\underline{\boldsymbol{X}} = \begin{pmatrix} \mathbf{1} & \boldsymbol{X_1} & \boldsymbol{X_2} & \boldsymbol{X_1 X_2} \end{pmatrix}$, $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$.

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

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

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

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

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

$$\underline{\boldsymbol{X}}^T \underline{\boldsymbol{X}} = \begin{pmatrix} 23 & 330 & 720 & 10800 \\ 330 & 7500 & 10800 & 252000 \\ 720 & 10800 & 31680 & 475200 \\ 10800 & 252000 & 475200 & 11088000 \end{pmatrix} \tag{2}$$

$$\implies (\underline{\boldsymbol{X}}^T \underline{\boldsymbol{X}})^{-1} = \begin{pmatrix} 0.3839957035 & -1.831\,364 \times 10^{-2} & -0.0087271751 & 4.162\,191 \times 10^{-4} \\ -0.0183136412 & 1.437\,522 \times 10^{-3} & 0.0004162191 & -3.267\,096 \times 10^{-5} \\ -0.0087271751 & 4.162\,191 \times 10^{-4} & 0.0002867287 & -1.324\,740 \times 10^{-5} \\ 0.0004162191 & -3.267\,096 \times 10^{-5} & -0.0000132474 & 9.950\,471 \times 10^{-7} \end{pmatrix} \tag{3}$$

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

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

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

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\underline{\boldsymbol{X}}^T\underline{\boldsymbol{X}})^{-1}\underline{\boldsymbol{X}}^T\boldsymbol{y} \\ &= \begin{pmatrix} 0.3839957035 & -1.831\,364\times10^{-2} & -0.0087271751 & 4.162\,191\times10^{-4} \\ -0.0183136412 & 1.437\,522\times10^{-3} & 0.0004162191 & -3.267\,096\times10^{-5} \\ -0.0087271751 & 4.162\,191\times10^{-4} & 0.0002867287 & -1.324\,740\times10^{-5} \\ 0.0004162191 & -3.267\,096\times10^{-5} & -0.0000132474 & 9.950\,471\times10^{-7} \end{pmatrix} \begin{pmatrix} 301 \\ 3315 \\ 12783 \\ 142920 \end{pmatrix} \\ \hat{\boldsymbol{\beta}} &= \begin{pmatrix} 2.79954350 \\ -0.09582438 \\ 0.52482098 \\ -0.01015172 \end{pmatrix}\end{aligned}$$

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

$$\hat{\boldsymbol{y}} = \begin{pmatrix} \boldsymbol{1} & \boldsymbol{X_1} & \boldsymbol{X_2} & \boldsymbol{X_1}\boldsymbol{X_2} \end{pmatrix} \begin{pmatrix} 2.79954350 \\ -0.09582438 \\ 0.52482098 \\ -0.01015172 \end{pmatrix}$$

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 σ^2. To do this, we must first calculate the SSE of my model, given by:

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

Hence my estimate our σ^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: $\mathrm{SST}_C = \mathrm{SST} - ny^2$. To value of the SST is given by:

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

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:

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

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

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

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 ith diagonal element of the matrix $\boldsymbol{G}^{-1} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$ which was figured out in (3). Next find the test statistic of each parameter and compare against the critical value:

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

$$\frac{\hat{\beta}_1}{\sqrt{7.116698}\sqrt{g^{22}}} = \frac{-0.09582438}{\sqrt{7.116698}\sqrt{1.437\,522 \times 10^{-3}}} = -0.9473915 \tag{11}$$

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

$$\frac{\hat{\beta}_3}{\sqrt{7.116698}\sqrt{g^{44}}} = \frac{-0.01015172}{\sqrt{7.116698}\sqrt{9.950\,471 \times 10^{-7}}} = -3.8148590 \tag{13}$$

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:

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{X_1} + \beta_2 \boldsymbol{X_2} + \beta_3 \boldsymbol{X_1 X_2} + \beta_4 \boldsymbol{X_1}^2 + \beta_5 \boldsymbol{X_2}^2 + \boldsymbol{\varepsilon} \tag{14}$$

In matrix form this would look like the following:

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

where: $\underline{\boldsymbol{X}} = \begin{pmatrix} \boldsymbol{1} & \boldsymbol{X_1} & \boldsymbol{X_2} & \boldsymbol{X_1 X_2} & \boldsymbol{X_1}^2 & \boldsymbol{X_2}^2 \end{pmatrix}$, $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix}$.

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 $(\underline{\boldsymbol{X}}^T\underline{\boldsymbol{X}})^{-1}$:

$$\underline{\boldsymbol{X}}^T\underline{\boldsymbol{X}} = \begin{pmatrix} 23 & 330 & 720 & 10800 & 7500 & 31680 \\ 330 & 7500 & 10800 & 252000 & 189000 & 475200 \\ 720 & 10800 & 31680 & 475200 & 252000 & 1555200 \\ 10800 & 252000 & 475200 & 11088000 & 6480000 & 23328000 \\ 7500 & 189000 & 252000 & 6480000 & 5070000 & 11088000 \\ 31680 & 475200 & 1555200 & 23328000 & 11088000 & 81202176 \end{pmatrix}$$

$$\implies (\underline{\boldsymbol{X}}^T\underline{\boldsymbol{X}})^{-1} = \begin{pmatrix} 0.4829853691 & -2.814\,401 \times 10^{-2} & -1.778\,522 \times 10^{-2} & 3.493\,342 \times 10^{-4} & 4.257\,768 \times 10^{-4} & 1.583\,991 \times 10^{-4} \\ -0.0281440079 & 4.817\,593 \times 10^{-3} & 3.141\,439 \times 10^{-4} & -2.602\,882 \times 10^{-5} & -1.224\,108 \times 10^{-4} & 9.632\,377 \times 10^{-7} \\ -0.0177852211 & 3.141\,439 \times 10^{-4} & 1.532\,919 \times 10^{-3} & -7.127\,103 \times 10^{-6} & -5.573\,936 \times 10^{-6} & -2.144\,988 \times 10^{-5} \\ 0.0003493342 & -2.602\,882 \times 10^{-5} & -7.127\,103 \times 10^{-6} & 1.040\,240 \times 10^{-6} & -2.876\,870 \times 10^{-7} & -1.070\,264 \times 10^{-7} \\ 0.0004257768 & -1.224\,108 \times 10^{-4} & -5.573\,936 \times 10^{-6} & -2.876\,870 \times 10^{-7} & 4.502\,301 \times 10^{-6} & 1.248\,641 \times 10^{-7} \\ 0.0001583991 & 9.632\,377 \times 10^{-7} & -2.144\,988 \times 10^{-5} & -1.070\,264 \times 10^{-7} & 1.248\,641 \times 10^{-7} & 3.693\,898 \times 10^{-7} \end{pmatrix} \tag{15}$$

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

$$\underline{\boldsymbol{X}}^T \boldsymbol{y} = \begin{pmatrix} 301 \\ 3315 \\ 12783 \\ 142920 \\ 70400 \\ 640044 \end{pmatrix} \tag{16}$$

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

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

$$= \begin{pmatrix} 0.4829853691 & -2.814\,401 \times 10^{-2} & -1.778\,522 \times 10^{-2} & 3.493\,342 \times 10^{-4} & 4.257\,768 \times 10^{-4} & 1.583\,991 \times 10^{-4} \\ -0.0281440079 & 4.817\,593 \times 10^{-3} & 3.141\,439 \times 10^{-4} & -2.602\,882 \times 10^{-5} & -1.224\,108 \times 10^{-4} & 9.632\,377 \times 10^{-7} \\ -0.0177852211 & 3.141\,439 \times 10^{-4} & 1.532\,919 \times 10^{-3} & -7.127\,103 \times 10^{-6} & -5.573\,936 \times 10^{-6} & -2.144\,988 \times 10^{-5} \\ 0.0003493342 & -2.602\,882 \times 10^{-5} & -7.127\,103 \times 10^{-6} & 1.040\,240 \times 10^{-6} & -2.876\,870 \times 10^{-7} & -1.070\,264 \times 10^{-7} \\ 0.0004257768 & -1.224\,108 \times 10^{-4} & -5.573\,936 \times 10^{-6} & -2.876\,870 \times 10^{-7} & 4.502\,301 \times 10^{-6} & 1.248\,641 \times 10^{-7} \\ 0.0001583991 & 9.632\,377 \times 10^{-7} & -2.144\,988 \times 10^{-5} & -1.070\,264 \times 10^{-7} & 1.248\,641 \times 10^{-7} & 3.693\,898 \times 10^{-7} \end{pmatrix} \begin{pmatrix} 301 \\ 3315 \\ 12783 \\ 142920 \\ 70400 \\ 640044 \end{pmatrix}$$

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} 6.016644748 \\ -0.206571079 \\ 0.143467750 \\ -0.012325436 \\ 0.006879675 \\ 0.006597411 \end{pmatrix} \tag{17}$$

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

$$\hat{\boldsymbol{y}} = \begin{pmatrix} \mathbf{1} & \boldsymbol{X_1} & \boldsymbol{X_2} & \boldsymbol{X_1}\boldsymbol{X_2} & \boldsymbol{X_1}^2 & \boldsymbol{X_2}^2 \end{pmatrix} \begin{pmatrix} 6.016644748 \\ -0.206571079 \\ 0.143467750 \\ -0.012325436 \\ 0.006879675 \\ 0.006597411 \end{pmatrix}$$

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:

$$\begin{aligned} \text{SSE} &= \boldsymbol{y}^T \boldsymbol{y} - \hat{\boldsymbol{\beta}}^T \boldsymbol{X}^T \boldsymbol{y} \\ &= 5918.125 - 5905.587 \\ &= 12.53844 \end{aligned} \tag{18}$$

Hence, substiture (18) into the equation for σ^2 to find our LS estimate:

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

The estimate for σ^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 σ^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:

$$\begin{aligned} R^2 &= \frac{\text{SST}_C - SSE}{\text{SST}_C} \\ &= \frac{1978.951 - 12.53844}{1978.951} \\ &= 0.9936641 \end{aligned} \tag{20}$$

The coeffecient 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 coeffecient 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 coeffecient 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 ith diagonal element of the matrix $\boldsymbol{G}^{-1} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$ which was figured out in (15). Next find the test statistic of each parameter and compare against the critical value:

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

$$\frac{\hat{\beta}_1}{\sqrt{0.7375553}\sqrt{g^{22}}} = \frac{-0.206571079}{\sqrt{0.7375553}\sqrt{4.817\,593 \times 10^{-3}}} = -3.465431 \tag{22}$$

$$\frac{\hat{\beta}_2}{\sqrt{0.7375553}\sqrt{g^{33}}} = \frac{0.143467750}{\sqrt{0.7375553}\sqrt{1.532\,919 \times 10^{-3}}} = 4.266752 \tag{23}$$

$$\frac{\hat{\beta}_3}{\sqrt{0.7375553}\sqrt{g^{44}}} = \frac{-0.012325436}{\sqrt{0.7375553}\sqrt{1.040\,240 \times 10^{-6}}} = -14.071432 \tag{24}$$

$$\frac{\hat{\beta}_4}{\sqrt{0.7375553}\sqrt{g^{55}}} = \frac{0.006879675}{\sqrt{0.7375553}\sqrt{4.502\,301 \times 10^{-6}}} = 3.775316 \tag{25}$$

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

(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 σ^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. Futhermore, the quadratic model coefficient of determination R^2 is closer to 1 than the linear models, suggesting that the model is more successful at prediciting 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:

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

The value of $\boldsymbol{f}_0 = \begin{pmatrix} 1 \\ 10 \\ 50 \\ 500 \\ 100 \\ 2500 \end{pmatrix}$, 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 $(\underline{\boldsymbol{X}}^T\underline{\boldsymbol{X}})^{-1}$ (15).

So that gives us the confidence interval as:

$$
\begin{aligned}
&\boldsymbol{f}_0^T\hat{\boldsymbol{\beta}} \pm t_{n-p,1-\frac{\alpha}{2}}\hat{\sigma}\sqrt{\boldsymbol{f}_0^T(\underline{\boldsymbol{X}}^T\underline{\boldsymbol{X}})^{-1}\boldsymbol{f}_0}\\
\implies &\begin{pmatrix}1\\10\\50\\500\\100\\2500\end{pmatrix}^T\begin{pmatrix}6.016644748\\-0.206571079\\0.143467750\\-0.012325436\\0.006879675\\0.006597411\end{pmatrix} \pm t_{23-6,1-\frac{0.05}{2}}\\
&\cdot 0.7375553\sqrt{\begin{pmatrix}1\\10\\50\\500\\100\\2500\end{pmatrix}^T\begin{pmatrix}0.4829853691 & -2.814\,401\times10^{-2} & -1.778\,522\times10^{-2} & 3.493\,342\times10^{-4} & 4.257\,768\times10^{-4} & 1.583\,991\times10^{-4}\\ -0.0281440079 & 4.817\,593\times10^{-3} & 3.141\,439\times10^{-4} & -2.602\,882\times10^{-5} & -1.224\,108\times10^{-4} & 9.632\,377\times10^{-7}\\ -0.0177852211 & 3.141\,439\times10^{-4} & 1.532\,919\times10^{-3} & -7.127\,103\times10^{-6} & -5.573\,936\times10^{-6} & -2.144\,988\times10^{-5}\\ 0.0003493342 & -2.602\,882\times10^{-5} & -7.127\,103\times10^{-6} & 1.040\,240\times10^{-6} & -2.876\,870\times10^{-7} & -1.070\,264\times10^{-7}\\ 0.0004257768 & -1.224\,108\times10^{-4} & -5.573\,936\times10^{-6} & -2.876\,870\times10^{-7} & 4.502\,301\times10^{-6} & 1.248\,641\times10^{-7}\\ 0.0001583991 & 9.632\,377\times10^{-7} & -2.144\,988\times10^{-5} & -1.070\,264\times10^{-7} & 1.248\,641\times10^{-7} & 3.693\,898\times10^{-7}\end{pmatrix}\begin{pmatrix}1\\10\\50\\500\\100\\2500\end{pmatrix}}\\
\implies &22.1431 \pm 2.110\cdot 0.7375553\sqrt{0.1446243}\\
\implies &22.1431 \pm 0.5918309\\
\implies &(21.55127, 22.73493) \qquad (28)
\end{aligned}
$$

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:

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

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

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

Therefore the the predicition 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 anaylsis

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 $\boldsymbol{\beta}$ is:

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} 2.799544 \\ -0.095824 \\ 0.524821 \\ -0.010152 \end{pmatrix} \tag{31}$$

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

$$\hat{\boldsymbol{y}} = \begin{pmatrix} \mathbf{1} & \boldsymbol{X_1} & \boldsymbol{X_2} & \boldsymbol{X_1 X_2} \end{pmatrix} \begin{pmatrix} 2.799544 \\ -0.095824 \\ 0.524821 \\ -0.010152 \end{pmatrix}$$

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: $H_0 : \beta_i = \mathbf{0}, H_1 : \beta_i \neq \mathbf{0}$. I choose $\alpha = 1 - \gamma =$ 2.5% significance test. The critical value $t_{n-p,\gamma} = t_{19,0.975} = 2.093$. 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 anaylsis

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 $\boldsymbol{\beta}$ is:

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} 6.0166447 \\ -0.2065711 \\ 0.1434677 \\ -0.0123254 \\ 0.0068797 \\ 0.0065974 \end{pmatrix} \tag{32}$$

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

$$\boldsymbol{y} = \begin{pmatrix} \mathbf{1} & \boldsymbol{X_1} & \boldsymbol{X_2} & \boldsymbol{X_1 X_2} & \boldsymbol{X_1}^2 & \boldsymbol{X_2}^2 \end{pmatrix} \begin{pmatrix} 6.0166447 \\ -0.2065711 \\ 0.1434677 \\ -0.0123254 \\ 0.0068797 \\ 0.0065974 \end{pmatrix}$$

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

```
# Set working directory
dir<-getwd()
if(!is.null(dir)) setwd(dir) else stop("Working directory is incorrect")

# Load data as a table
tab<-read.table("Viscos.txt", header = TRUE)

# Create X matrix
X<-cbind(
  rep(1, time = nrow(tab)),
  tab$Oil,
  tab$Filler,
  tab$Oil*tab$Filler
)

# Get/store transpose of X
XT<-t(X)

# Get value of n
n<-nrow(X)

# Times transpose by X
prod<-XT %*% X

# Find the inverse of prod
invProd<-solve(prod)

# Create y response vector
y<-tab$Visc

# Calculate beta vector
beta<-invProd%*%XT%*%y

# Calculate SSE
SSE<-t(y)%*%y-t(beta)%*%XT%*%y

# Calculate estimate of variance
var<-SSE/(nrow(X) - ncol(X))

# Calculate the SST
SST<-t(y)%*%y

# Calculate mean of y
my = mean(y)

# Calculate value of SST_C
SST_C<-SST-n*(my^2)

# Calculate coefficient of regression
R<-(SST_C-SSE)/SST_C

# Get diagonal of (X^TX)^-1
dia<-diag(invProd)

# Get values of test statistics
ts<-beta/(sqrt(var)[1]*sqrt(dia))
```

Listing 2: Quadratic regression equation method

```
# Set working directory
dir<-getwd()
if(!is.null(dir)) setwd(dir) else stop("Working directory is incorrect")

# Load data as a table
tab<-read.table("Viscos.txt", header = TRUE)

# Create X matrix
X<-cbind(
  rep(1, time = nrow(tab)),
  tab$Oil,
  tab$Filler,
  tab$Oil*tab$Filler,
  tab$Oil^2,
  tab$Filler^2
)

# Get/store transpose of X
XT<-t(X)

# Get value of n
n<-nrow(X)

# Times transpose by X
prod<-XT %*% X

# Find the inverse of prod
invProd<-solve(prod)

# Create y response vector
y<-tab$Visc

# Get value of X^Ty
XTy<-XT%*%y

# Calculate beta vector
beta<-invProd%*%XTy

# Calculate SSE
SSE<-t(y)%*%y-t(beta)%*%XT%*%y

# Calculate estimate of variance
var<-SSE/(nrow(X) - ncol(X))

# Calculate the SST
SST<-t(y)%*%y

# Calculate mean of y
my = mean(y)

# Calculate value of SST_C
SST_C<-SST-n*(my^2)

# Calculate coefficient of regression
R<-(SST_C-SSE)/SST_C

# Get diagonal of (X^TX)^-1
dia<-diag(invProd)

# Get values of test statistics
ts<-beta/(sqrt(var)[1]*sqrt(dia))

# Store guessed values from chemist
f_0=c(1, 10, 50, 500, 100, 2500)

# Calculate f_0%*%beta
ciMean<-f_0%*%beta

# Calculate f_0(X^TX)^-1f_0
ciP<-f_0%*%invProd%*%f_0

# 97.5% t distro
ciT<-2.110

# Calculate CI +-
ciPM<-ciT*var*sqrt(ciP)

# Calculate confidence interval as 2 vec
ci<-c(ciMean-ciPM, ciMean+ciPM)

# Calculate PI +-
piPM<-ciT*var*sqrt(1+ciP)

# Calculate predicition interval as 2 vec
pi<-c(ciMean-piPM, ciMean+piPM)
```

Listing 3: Linear regression summary method

```
# Set working directory
dir<-getwd()
if(!is.null(dir)) setwd(dir) else stop("Working directory is incorrect")

# Load data as a table
tab<-read.table("Viscos.txt", header = TRUE)

# Fit the model as a linear model
model.fit<-lm(tab$Visc ~ tab$Oil + tab$Filler + I(tab$Oil*tab$Filler))

# Get summary of model
summa<-summary(model.fit)
summa

# Get anova of model
anov<-anova(model.fit)
anov

# Find variance of response
var<-anov[]$‘Mean Sq‘[4]
var
```

Listing 4: Linear regression summary results

```
Call:
lm(formula = tab$Visc ~ tab$Oil + tab$Filler + I(tab$Oil * tab$Filler))

Residuals:
    Min      1Q  Median      3Q     Max
-3.8302 -1.9674 -0.2477  1.9633  4.9612

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              2.799544   1.653113   1.693  0.10669
tab$Oil                 -0.095824   0.101145  -0.947  0.35533
tab$Filler               0.524821   0.045173  11.618 4.46e-10 ***
I(tab$Oil * tab$Filler) -0.010152   0.002661  -3.815  0.00117 **
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1       1

Residual standard error: 2.668 on 19 degrees of freedom
Multiple R-squared:  0.9317,  Adjusted R-squared:  0.9209
F-statistic: 86.36 on 3 and 19 DF,  p-value: 2.97e-11

>
> # Get anova of model
> anov<-anova(model.fit)
> anov
Analysis of Variance Table

Response: tab$Visc
                        Df  Sum Sq Mean Sq F value    Pr(>F)
tab$Oil                  1  364.31  364.31  51.191 8.442e-07 ***
tab$Filler               1 1375.85 1375.85 193.327 2.074e-11 ***
I(tab$Oil * tab$Filler)  1  103.57  103.57  14.553  0.001169 **
Residuals               19  135.22    7.12
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1       1
>
> # Find variance of response
> var<-anov[]$‘Mean Sq‘[4]
> var
[1] 7.116698
```

Listing 5: Quadratic regression summary method

```
# Delete previous variables
rm(list=ls())

# Set working directory
dir<-getwd()
if(!is.null(dir)) setwd(dir) else stop("Working directory is incorrect")

# Load data as a table
tab<-read.table("Viscos.txt", header = TRUE)

```

```
# Store variables from table
Visc<-tab$Visc
Oil<-tab$Oil
Filler<-tab$Filler
Interaction<-tab$Oil*tab$Filler
OilSquared<-tab$Oil^2
FillerSquared<-tab$Filler^2

# Fit the model as a quadratic model
model.fit<-lm(Visc ~ Oil + Filler + Interaction + OilSquared + FillerSquared)

# Get summary of model
summa<-summary(model.fit)
summa

# Get anova of model
anov<-anova(model.fit)
anov

# Find variance of response
var<-anov[]$‘Mean Sq‘[6]
var

newdata<-data.frame(
  Oil=10,
  Filler=50,
  Interaction = 500,
  OilSquared = 100,
  FillerSquared = 2500
)

# Find the confidence interval
ci<-predict(model.fit, newdata, interval="confidence", level=0.95)
ci

# Find the predicition interval
pi<-predict(model.fit, newdata, interval="prediction", level=0.95)
pi
```

Listing 6: Quadratic regression summary results

```
Call:
lm(formula = Visc ~ Oil + Filler + Interaction + OilSquared +
    FillerSquared)

Residuals:
     Min       1Q   Median       3Q      Max
-1.38709 -0.56863 -0.09948  0.65894  1.39761

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.0166447  0.5968487  10.081 1.38e-08 ***
Oil           -0.2065711  0.0596091  -3.465 0.002958 **
Filler         0.1434677  0.0336246   4.267 0.000521 ***
Interaction   -0.0123254  0.0008759 -14.071 8.50e-11 ***
OilSquared     0.0068797  0.0018223   3.775 0.001510 **
FillerSquared  0.0065974  0.0005220  12.640 4.53e-10 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Residual standard error: 0.8588 on 17 degrees of freedom
Multiple R-squared:  0.9937,  Adjusted R-squared:  0.9918
F-statistic: 533.2 on 5 and 17 DF,  p-value: < 2.2e-16

>
> # Get anova of model
> anov<-anova(model.fit)
> anov
Analysis of Variance Table

Response: Visc
              Df  Sum Sq Mean Sq   F value    Pr(>F)
Oil            1  364.31  364.31  493.9469 5.303e-14 ***
Filler         1 1375.85 1375.85 1865.4200 < 2.2e-16 ***
Interaction    1  103.57  103.57  140.4239 1.219e-09 ***
OilSquared     1    4.85    4.85    6.5718   0.02014 *
```

```
FillerSquared  1  117.83  117.83  159.7599 4.528e-10 ***
Residuals     17   12.54    0.74
---
Signif. codes:  0     ***     0.001     **     0.01     *     0.05     .     0.1          1
>
> # Find variance of response
> var<-anov[]$‘Mean Sq‘[6]
> var
[1] 0.7375553
>
> newdata<-data.frame(
+   Oil=10,
+   Filler=50,
+   Interaction = 500,
+   OilSquared = 100,
+   FillerSquared = 2500
+ )
>
> # Find the confidence interval
> ci<-predict(model.fit, newdata, interval="confidence", level=0.95)
> ci
      fit      lwr      upr
1 22.1431 21.45403 22.83217
>
> # Find the predicition interval
> pi<-predict(model.fit, newdata, interval="prediction", level=0.95)
> pi
      fit      lwr      upr
1 22.1431 20.20457 24.08163
```