Regression Coursework

Chloe Hilton, 9848535

21 November 2018

Introduction

In this report, the aim is to analyse different regression models constructed for the data set on Viscosity of Elastomer blends, with explanatory variables Naphthenic Oil and Filler. We believe that Viscosity follows a Normal Distribution meaning we can conduct a number of statistical test such as confidence intervals to estimate the variance and mean of the data, as well as create fitted models of the data by estimating the parameters. For the purpose of this report, many of the calculations and functions were completed in R to acquire accurate and concise results and have been provided in the Appendix as evidence. Parts will be referenced to throughout where appropriate.

Analysis

Part A

The first regression model produced was fitted with a constant term, 2 regressor terms and an interaction term. The model was set up in the following way:

$$Viscosity = \beta_0 + \beta_1 * Oil + \beta_2 * Filler + \beta_3 * Oil * Filler + \epsilon$$

Firstly, it was necessary to create the full design matrix of the data, where the columns were fitted with the data for Oil and Filler provided in the Viscosity table. The format of the design matrix can be found in the Appendix. The first 3 rows of the design matrix were

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 12 & 0 \\ 1 & 0 & 24 & 0 \\ . & . & . & . \end{pmatrix}$$

This matrix was labelled $\mathbf{X}$ and will be referred to as this from now on. For clarity, $\mathbf{y}$ is a column vector consisting of the corresponding values of Viscosity. In order to estimate the individual parameters of the model, $\mathbf{X^T}$ along with $\mathbf{X^T X}$ and $\mathbf{X^T X}^{-1}$ were calculated by simple matrix multiplication. Then, the matrix $\mathbf{X^T y}$ was computed which allowed us to find

$$\hat{\beta} = (X^T X)^{-1} X^T y = \begin{pmatrix} 2.7995 \\ -0.0958 \\ 0.5248 \\ -0.0102 \end{pmatrix}$$

where the values of this vector correspond to the estimates of the model's parameters.
Hence, using the values given in the matrix $\hat{\beta}$ above, the fitted model is

$$Viscosity = 2.79954350 - 0.09582438 * Oil + 0.52482098 * Filler - 0.01015172 * Oil * Filler + \epsilon$$

We now wished to estimate the variance of the response, meaning we wanted to calculate a value for $\hat{\sigma}^2$. From prior knowledge and some derivation, it can be seen that $\hat{\sigma}^2 = \frac{SSE}{(n-p)}$, where SSE is the residual sum of

squares and n-p is the degrees of freedom of SSE. In the Appendix for Part A, the matrix calculations for SSE, SST and SSR can be found. Using these values, it was found that

$$\hat{\sigma}^2 = \frac{135.2173}{23-4} = 7.116698$$

This is accurate assuming that the fitted model is correct. Otherwise $\hat{\sigma}$ overestimates the true value of σ.

A measure of how well the model fitted the data was to analyse the Coefficient of Determination, denoted R^2. Since there was an intercept in the fitted model, we needed to adjust the values for SSR and SST before computing this coefficient $\implies$ SSR_c = SSR - $n\bar{y}^2$, SST_c= SST - $n\bar{y}^2$. Hence,

$$R^2 = \frac{SSR_c}{SST_c} = \frac{1843.734}{1978.951} = 0.9316723$$

The Coefficient of Determination represents the proportion of variability, meaning that if the value is close to 1, the model is better at predicting response values accurately. Therefore, we interpret 0.9316 as 93.16% of the data being explained by this model.

Now, it was necessary to test if the true values of the model parameters were likely to be 0. This was done by creating confidence intervals for each of the β_i estimates. It is known that $\hat{\beta} \sim N_p(\beta, \sigma^2(X^TX)^{-1})$ and so we use this fact to derive the 100(1-α)% confidence interval for $\hat{\beta}$. The 100(1-α)% confidence interval is

$$\hat{\beta}_i \pm t_{n-p,\frac{1-\alpha}{2}} \hat{\sigma} \sqrt{g^{ii}}$$

where g^{ii} denotes the ith diagonal element of the matrix $\mathbf{X^TX}^{-1}$ and t_{n-p} is the critical observation value with n-p degrees of freedom. Below, the results of the four 90% confidence intervals calculated are tabulated:

Parameter	Lower bound	Upper bound
β_0	-0.05890774	5.657995
β_1	-0.2707184	0.07906962
β_2	0.4467116	0.6029304
β_3	-0.01475311	-0.005550324

It was observed that 0 is contained within the confidence intervals for β_0 and β_1, so there is insufficient evidence to say that these parameters are not equal to 0 and thus conclude it is plausible they could be 0 at the 10% significance level. However, since β_2 and β_3 do not contain 0 in their respective confidence intervals, we cannot draw the same conclusions for them.

Part B

For comparison, a new model was fitted, this one including quadratic terms, Oil^2 and $Filler^2$. The second regression model was set up in the following way:

$$Viscosity = \beta_0 + \beta_1 * Oil + \beta_2 * Filler + \beta_3 * Oil * Filler + \beta_4 * Oil^2 + \beta_5 * Filler^2 + \epsilon$$

The design matrix was adapted to include new columns for the extra regressors. Following a similar method

to the first model, new $\mathbf{X^T}$, $\mathbf{X^TX}$, $\mathbf{X^TX}^{-1}$ and $\mathbf{X^Ty}$ were computed which meant that

$$\hat{\beta} = (X^T X)^{-1} X^T y = \begin{pmatrix} 6.01664 \\ -0.20657 \\ 0.14346 \\ -0.01232 \\ 0.006879 \\ 0.006597 \end{pmatrix}$$

The new quadratic fitted model now looked like this, using the values computed in $\hat{\beta}$ above:

$$Viscosity = 6.01664 - 0.20657 * Oil + 0.14346 * Filler - 0.01232 * Oil * Filler + 0.006879 * Oil^2 + 0.006597 * Filler^2 + \epsilon$$

Again, to estimate the variance of the response, a value for $\hat{\sigma}$ was calculated.

$$\hat{\sigma}^2 = \frac{SSE}{(n-p)} = \frac{12.53844}{23-6} = 0.7375553$$

Matrix calculations for SSE, SST and SSR of the quadratic model can be found in Part B of the Appendix. Here, we see that the variance of the response is much smaller than the model fitted in A, which indicates that this model is better than the first since a smaller variance means that the data values for Oil and Filler tend to be closely distributed about the true mean.

Using the same method as before with new values for SSR and SST appropriately adjusted since there was an intercept in the quadratic model, R^2 was found to be

$$R^2 = \frac{SSR_c}{SST_c} = \frac{1966.413}{1978.951} = 0.9936641$$

The R^2 Coefficient of Determination is somewhat closer to 1 than the previous model stated in Part A and 99.37% of the data is explained by the quadratic model. Hence, it could be argued that this model is better at predicting Viscosity values at certain Naphthenic Oil and Filler levels than the first model used.

Since we believed that this model is better, it seemed appropriate to again test whether any of the model parameters were likely to be 0 so more confidence intervals were found. Below, the results of the six 90% confidence intervals calculated are tabulated:

Parameter	Lower bound	Upper bound
β_0	4.978363	7.054927
β_1	-0.3102674	-0.1028747
β_2	0.084974196	0.201961304
β_3	-0.013849190	-0.010801681
β_4	0.003709628	0.010049722
β_5	0.005689401	0.0075054227

None of the above confidence intervals included 0, meaning they all have significant values in the fitted model and so all the regressor terms contribute in the fitted quadratic model.

Overall, the quadratic model was significantly better that the original model fitted to the data as it had a much smaller variance for the response and a higher proportion of the variability meaning it would be more reliable in accurately predicting Viscosity values of Elastomer blends.

Part C

A chemist believes that a elastomer blend viscosity of 21M can be achieved by using 10 phr Oil and 50 phr Filler. We want to test if this is in faact true and therefore the best way to do this is to produce a confidence interval at the 5% significance level under the stated settings. The 100(1-α)% confidence interval for the mean viscosity is

$$f_0^T \hat{\beta} \pm t_{n-p,\frac{\alpha}{2}} \hat{\sigma} \sqrt{f_0^T (X^T X)^{-1} f_0}$$

where f_0 is a column vector containing values of the explanatory variables under the conditions we are testing, t_{n-p} is the critical observation value with n-p degrees of freedom and $\hat{\beta}$, $\hat{\sigma}$ and $(X^T X)^{-1}$ have all been found previously for the quadratic model. For our calculations, $f_0 = \begin{pmatrix} 1 \\ 10 \\ 50 \\ 500 \\ 100 \\ 2500 \end{pmatrix}$ since this uses the settings the chemist has suggested for each explanatory variable. Full calculations are provided in Part C of the Appendix, however, the 95% confidence interval is

$$(21.45403, 22.83217)$$

This does not include the required 21M viscosity, hence the settings the chemist believes are incorrect.

If a quality inspector then decides to take a single observation using the values of 10 phr Oil and 50 phr Filler like the chemist suggested, we then require to compute a prediction interval that will contain 21M. We aim to see if 21M is now in this interval because like with all predictions, there is a level of uncertainty and this is captured in an error term, ϵ_0, which results in a wider interval. A prediction interval is calculated using the following formula:

$$f_0^T \hat{\beta} \pm t_{n-p,\frac{\alpha}{2}} \hat{\sigma} \sqrt{1 + f_0^T (X^T X)^{-1} f_0}$$

The same value for f_0 is used here. The found 95% prediction interval under the stated condition is

$$(20.20457, 24.08163)$$

This interval does include the 21M we require, so by taking a single observation, 95% of the time, a viscosity of 21M will be achieved.

Conclusion

In conclusion, out of the two regression models created in this report, it was clear after interpretation that the quadratic model in Part B was the better model. The second fitted model in theory was much more successful at accurately predicting the Viscosity of Elastomer blends at different phr levels of Naphthenic Oil and Filler. We observed that the simpler model fitted in Part A didn't provide enough contributing factors and as a result, it was likely that some of the parameters were 0, whereas the model fitted in Part B has no parameters likely to be 0. The Coefficient of Determination was also much closer to 1, which is ideal when trying to create a useful model to represent the data.

Appendix

```r
# Part A R code
visc.data <- read.table("Viscos.txt", sep="", header=TRUE)
colnames(visc.data)
```

```
## [1] "Visc"   "Oil"    "Filler"
```

```r
Visc <- visc.data$Visc
Oil <- visc.data$Oil
Filler <- visc.data$Filler

Xrows <- matrix( c(1, 1, 1, 0, 0, 0, 0, 12, 24, 0, 0, 0), nrow=3, ncol=4, byrow = FALSE)
Xrows # first 3 rows of the design matrix
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    0    0    0
## [2,]    1    0   12    0
## [3,]    1    0   24    0
```

```r
model.fit <- lm(formula = Visc ~ Oil + Filler + I(Oil*Filler))
```

```r
summary(model.fit) # This has been used to cross-check step by step calculations were correct.
anova(model.fit)
```

```r
# step by step
y <- Visc
Int <- rep(1,length(Oil))
X <- cbind(Int, Oil, Filler, Oil*Filler)
XT <- t(X)
XTX <- XT %*% X
XTXI <- solve(XTX)
XTy <- XT %*% y
beta.hat <- XTXI %*% XTy
beta.hat # same as summary
```

```
##                [,1]
## Int     2.79954350
## Oil    -0.09582438
## Filler  0.52482098
##        -0.01015172
```

```r
yhat <- X %*% beta.hat
e <- yhat - y
SSR <- t(beta.hat) %*% XTy
SSRc <- SSR - (length(y)*mean(y)^2)
SSRc
```

```
##          [,1]
## [1,] 1843.734
```

```r
SSE <- sum(e^2) # same as anova
SSE
```

```
## [1] 135.2173
```

```r
SST <- t(y) %*% y
SSTc <- SST - (length(y)*mean(y)^2)
```

```
SSTc
```

```
##          [,1]
## [1,] 1978.951
```

```
sigma2.1 <- SSE/(length(y)-length(beta.hat)) # estimate of sigma squared
sigma2.1
```

```
## [1] 7.116698
```

```
sig1 <- summary(model.fit)[[6]]
sig2.1 <- sig1^2
sig2.1
```

```
## [1] 7.116698
```

```
R2 <- SSRc/SSTc # coefficient of determination
R2
```

```
##           [,1]
## [1,] 0.9316723
```

```
summary(model.fit)$r.squared # Used to compare and check
```

```
## [1] 0.9316723
```

```
alpha <- 0.1
tobs1 <- qt(1-alpha/2, length(y)-length(model.fit[[1]]))

lb1 <- beta.hat[1] - tobs1*sig1*sqrt(XTXI[1,1])
ub1 <- beta.hat[1] + tobs1*sig1*sqrt(XTXI[1,1])
b1cf <- cbind(lb1, ub1)
b1cf
```

```
##              lb1      ub1
## [1,] -0.05890774 5.657995
```

```
lb2 <- beta.hat[2] - tobs1*sig1*sqrt(XTXI[2,2])
ub2 <- beta.hat[2] + tobs1*sig1*sqrt(XTXI[2,2])
b2cf <- cbind(lb2, ub2)
b2cf
```

```
##             lb2        ub2
## [1,] -0.2707184 0.07906962
```

```
lb3 <- beta.hat[3] - tobs1*sig1*sqrt(XTXI[3,3])
ub3 <- beta.hat[3] + tobs1*sig1*sqrt(XTXI[3,3])
b3cf <- cbind(lb3, ub3)
b3cf
```

```
##            lb3       ub3
## [1,] 0.4467116 0.6029304
```

```
lb4 <- beta.hat[4] - tobs1*sig1*sqrt(XTXI[4,4])
ub4 <- beta.hat[4] + tobs1*sig1*sqrt(XTXI[4,4])
b4cf <- cbind(lb4, ub4)
b4cf
```

```
##              lb4          ub4
## [1,] -0.01475311 -0.005550324
```

```r
confint(model.fit, level=0.9, interval="confidence") # to confirm working out is correct.
```

```
##                          5 %         95 %
## (Intercept)     -0.05890774  5.657994742
## Oil             -0.27071838  0.079069618
## Filler           0.44671160  0.602930364
## I(Oil * Filler) -0.01475311 -0.005550324
```

```r
# Part B R code
model.fit2 <- lm(formula = Visc ~ Oil + Filler + Oil*Filler +I(Oil^2)+ I(Filler^2))
```

```r
summary(model.fit2)
anova(model.fit2)
```

```r
y2 <- Visc
Int2 <- rep(1,length(Oil))
X2 <- cbind(Int2, Oil, Filler, Oil*Filler, Oil^2, Filler^2)
X2T <- t(X2)
X2TX2 <- X2T %*% X2
X2TX2I <- solve(X2TX2)
X2Ty2 <- X2T %*% y2
beta.hat2 <- X2TX2I %*% X2Ty2
beta.hat2
```

```
##                [,1]
## Int2    6.016644748
## Oil    -0.206571079
## Filler  0.143467750
##        -0.012325436
##         0.006879675
##         0.006597411
```

```r
yhat2 <- X2 %*% beta.hat2
e2 <- yhat2 - y2
SSR2 <- t(beta.hat2) %*% X2Ty2
SSRc2 <- SSR2 - (length(y2)*mean(y2)^2)
SSRc2
```

```
##          [,1]
## [1,] 1966.413
```

```r
SSE2 <- sum(e2^2)
SSE2
```

```
## [1] 12.53844
```

```r
SST2 <- t(y2) %*% y2
SSTc2 <- SST2 - (length(y2)*mean(y2)^2)
SSTc2
```

```
##          [,1]
## [1,] 1978.951
```

```r
sigma2.2 <- SSE2/(length(y2)-length(beta.hat2)) # estimate of sigma squared
sigma2.2
```

```
## [1] 0.7375553
```

```r
sig2 <- summary(model.fit2)[[6]]
sig2.2 <- sig2^2
sig2.2
```

```
## [1] 0.7375553
```

```r
R2.2 <- SSRc2/SSTc2 # coefficient of determination
R2.2
```

```
##           [,1]
## [1,] 0.9936641
```

```r
summary(model.fit2)$r.squared
```

```
## [1] 0.9936641
```

```r
alpha <- 0.1
tobs2 <- qt(1-alpha/2, length(y2)-length(model.fit2[[1]]))

lb5 <- beta.hat2[1] - tobs2*sig2*sqrt(X2TX2I[1,1])
ub5 <- beta.hat2[1] + tobs2*sig2*sqrt(X2TX2I[1,1])
b1cf2 <- cbind(lb5, ub5)
b1cf2
```

```
##           lb5      ub5
## [1,] 4.978363 7.054927
```

```r
lb6 <- beta.hat2[2] - tobs2*sig2*sqrt(X2TX2I[2,2])
ub6 <- beta.hat2[2] + tobs2*sig2*sqrt(X2TX2I[2,2])
b2cf2 <- cbind(lb6, ub6)
b2cf2
```

```
##             lb6        ub6
## [1,] -0.3102674 -0.1028747
```

```r
lb7 <- beta.hat2[3] - tobs2*sig2*sqrt(X2TX2I[3,3])
ub7 <- beta.hat2[3] + tobs2*sig2*sqrt(X2TX2I[3,3])
b3cf2 <- cbind(lb7, ub7)
b3cf2
```

```
##            lb7       ub7
## [1,] 0.0849742 0.2019613
```

```r
lb8 <- beta.hat2[4] - tobs2*sig2*sqrt(X2TX2I[4,4])
ub8 <- beta.hat2[4] + tobs2*sig2*sqrt(X2TX2I[4,4])
b4cf2 <- cbind(lb8, ub8)
b4cf2
```

```
##              lb8         ub8
## [1,] -0.01384919 -0.01080168
```

```r
lb9 <- beta.hat2[5] - tobs2*sig2*sqrt(X2TX2I[5,5])
ub9 <- beta.hat2[5] + tobs2*sig2*sqrt(X2TX2I[5,5])
b5cf <- cbind(lb9, ub9)
b5cf
```

```
##              lb9        ub9
## [1,] 0.003709628 0.01004972
```

```r
lb10 <- beta.hat2[6] - tobs2*sig2*sqrt(X2TX2I[6,6])
ub10 <- beta.hat2[6] + tobs2*sig2*sqrt(X2TX2I[6,6])
b6cf <- cbind(lb10, ub10)
b6cf
```

```
##             lb10        ub10
## [1,] 0.005689401 0.007505422
```

```r
confint(model.fit2, level=0.9, interval="confidence") # to confirm working out is correct.
```

```
##                       5 %         95 %
## (Intercept)  4.978362696  7.054926800
## Oil         -0.310267413 -0.102874746
## Filler       0.084974196  0.201961304
## I(Oil^2)     0.003709628  0.010049722
## I(Filler^2)  0.005689401  0.007505422
## Oil:Filler  -0.013849190 -0.010801681
```

```r
# Part C R code
alpha2 <- 0.05
n <- length(y2)
p <- length(model.fit2[[1]])
tobs <- qt(1-alpha2/2, n-p)
tobs
```

```
## [1] 2.109816
```

```r
f0 <- c(1, 10, 50, 500, 100, 2500)
f0T <- t(f0)

lowerbound <- (f0T %*% beta.hat2) - tobs*sig2*sqrt(f0T %*% X2TX2I %*% f0)
upperbound <- (f0T %*% beta.hat2) + tobs*sig2*sqrt(f0T %*% X2TX2I %*% f0)
conf.int <- cbind(lowerbound, upperbound)
conf.int
```

```
##          [,1]     [,2]
## [1,] 21.45403 22.83217
```

```r
lowerbound2 <- (f0T %*% beta.hat2) - tobs*sig2*sqrt(1 + f0T %*% X2TX2I %*% f0)
upperbound2 <- (f0T %*% beta.hat2) + tobs*sig2*sqrt(1 + f0T %*% X2TX2I %*% f0)
pred.int <- cbind(lowerbound2, upperbound2)
pred.int
```

```
##          [,1]     [,2]
## [1,] 20.20457 24.08163
```