Viscosity of Elastomer Blends

MATH38141

Yann McLatchie
Student ID: 10161640

Introduction

In this project, we

A chemist reports that adding naphthenic oil (phr) and filler (phr) can be used to control the viscosity (M) of elastomer blends. It is believed that the viscosity follows a normal distribution with homogenous variance for any oil and filler level within the design region. In this project, we will analyse the data and fit various regression models to them to best model the viscosity of elastomer blends.

```
# retrieve data
df = read.table(
  "./data/Viscos.txt",
  header=TRUE
  )
# build predictor and regressors
visc = df$Visc
ones = seq(1,1,length.out=length(visc))
oil = df$Oil
filler = df$Filler
n = length(visc)
```

Part A

Question 1.

When performing the regression analysis, we assume that the errors are independant and identically distributed with zero mean and homogeneous variance.

First, we define an interaction term and the data matrix needed for this analysis.

```
interact = oil*filler
l.X = cbind(ones, oil, filler, interact)
```

We calcluate an estimate for our regressors $\hat{\vec{\beta}}$ by first calculating, with the help of R, $(\mathbf{X}^T\mathbf{X})^{-1}$ as

```
##                   ones           oil        filler      interact
## ones      0.3839957035 -1.831364e-02 -0.0087271751  4.162191e-04
## oil      -0.0183136412  1.437522e-03  0.0004162191 -3.267096e-05
## filler   -0.0087271751  4.162191e-04  0.0002867287 -1.324740e-05
## interact  0.0004162191 -3.267096e-05 -0.0000132474  9.950471e-07
```

and $\mathbf{X}^T\vec{y}$ as

```
##            [,1]
## ones        301
## oil        3315
## filler    12783
## interact 142920
```

from which we can calculate $\overrightarrow{\hat{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\overrightarrow{y}$ as

```
##                 [,1]
## ones       2.79954350
## oil       -0.09582438
## filler     0.52482098
## interact -0.01015172
```

We are able to verify these calculations with the standard R function for linear models

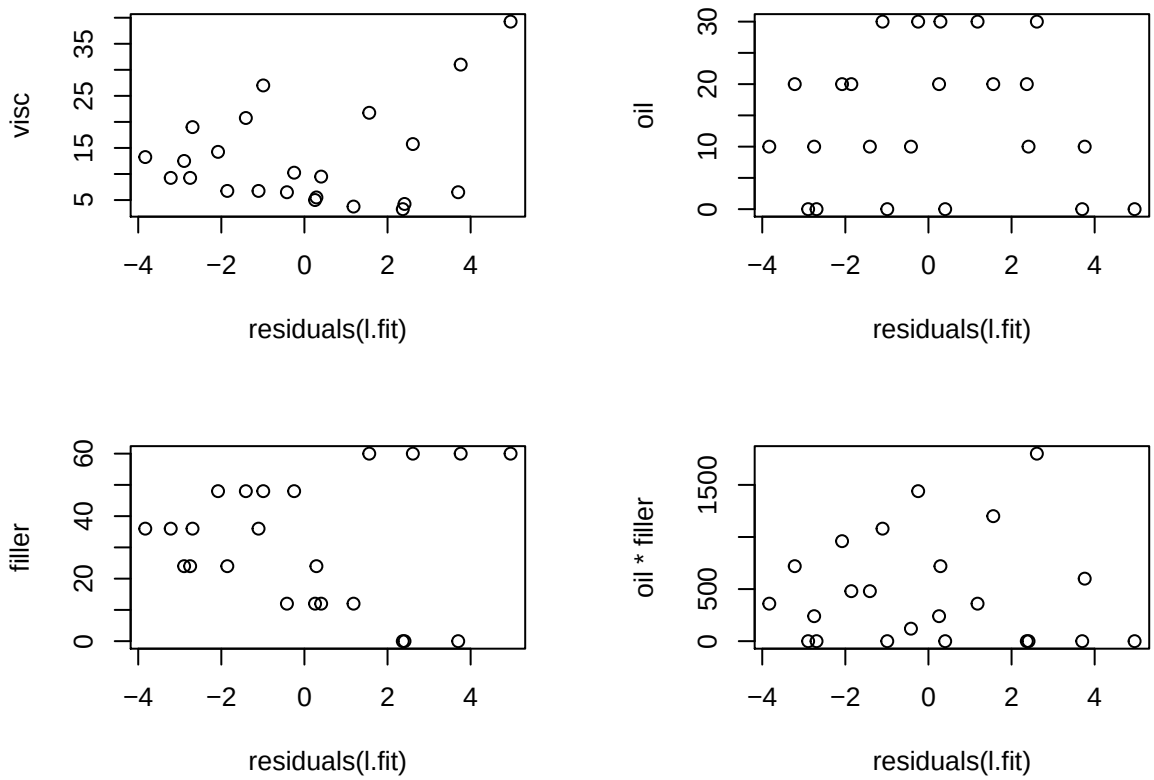
```
l.fit = lm(visc ~ oil*filler)
summary(l.fit)
```

```
##
## Call:
## lm(formula = visc ~ oil * 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
## oil         -0.095824   0.101145  -0.947  0.35533
## filler       0.524821   0.045173  11.618 4.46e-10 ***
## oil: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
```

and thus define our final fitted model as

$$\hat{y} = 2.799544 - 0.095824 \cdot \text{oil} + 0.524821 \cdot \text{filler} - 0.010152 \cdot \text{oil} \cdot \text{filler}.$$

We may verify our assumptions on the residuals by plotting them against the response and the regressors.



We find that there is perhaps evidence for heterogeneity when we plot them against the response variable, although not for any of the regressors. We can assume thus that our assumptions are largely satisfied.

Question 2.

The first three rows of our data matrix are as follows.

ones	oil	filler	interact
1	0	0	0
1	0	12	0
1	0	24	0

Question 3.

The variance of our response is given by

$$\mathrm{Var}(\overrightarrow{y}) = \mathrm{Var}(\mathbf{X}\overrightarrow{\beta} + \overrightarrow{\varepsilon})$$

and since the deterministic component of the model, $\mathbf{X}\overrightarrow{\beta}$, has zero variance,

$$\mathrm{Var}(\overrightarrow{y}) = \mathrm{Var}(\overrightarrow{\varepsilon})$$

The residuals are given as

$$\begin{aligned}
\overrightarrow{e} &= \overrightarrow{y} - \overrightarrow{\hat{y}} \\
&= \overrightarrow{y} - \mathbf{X}\overrightarrow{\beta} \\
&= (\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\overrightarrow{y} \\
&= (\mathbf{I} - \mathbf{H})\overrightarrow{y} \\
&= \mathbf{M}\overrightarrow{y}.
\end{aligned}$$

Note that these residuals are not the errors of our models, since $\hat{y}(x_i)$ differs from y_i due to the ε_i. If the model is chosen correctly, then $\varepsilon_i = e_i$. LS estimates have the lowest sum of squared residuals, which we calculate as

$$\begin{aligned} SSE &= \sum_{i=1}^{n} e_i^2 = \overrightarrow{e}^T \overrightarrow{e} \\ &= \overrightarrow{y}^T \mathbf{M}^T \mathbf{M} \overrightarrow{y} \\ &= \overrightarrow{y}^T \mathbf{M} \overrightarrow{y} \\ &= \overrightarrow{y}^T (\mathbf{I} - \mathbf{H}) \overrightarrow{y} \\ &= \overrightarrow{y}^T \overrightarrow{y} - \overrightarrow{y}^T \mathbf{H} \overrightarrow{y} \\ &= \overrightarrow{y}^T \overrightarrow{y} - \hat{\overrightarrow{\beta}}^T \mathbf{X}^T \overrightarrow{y}. \end{aligned}$$

Thus we calculate the SSE in our model to be

```
l.sse = t(visc)%*%visc - t(l.beta)%*%t(l.X)%*%visc
l.sse
```

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

and use this to estimate the variance of the response, $\hat{\sigma}^2$ via

$$\hat{\sigma}^2 = \frac{SSE}{n-p}$$

which we find to be

```
p = length(l.fit[[1]])
l.S2.hat = l.sse/(n-p)
l.S2.hat
```

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

which is also provided in the R output of the linear model summary.

```
(summary(l.fit)$sigma)**2
```

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

Note that this estimate for $\hat{\sigma}^2$ assumes that the fitted model is correct, otherwise we overestimate σ^2. We will find later that this is in fact the case.

Question 4.

We introduce the adjusted sums of squares

$$SST_c = SST - n\bar{y}^2 \quad \text{and} \quad SSR_c = SSR - n\bar{y}^2$$

in order to account for the presence of an intercept in our model. The coefficient of determination is thus calculated via

$$R^2 = \frac{SST_c - SSE}{SST_c} = \frac{SSR_c}{SST_c}.$$

We see from the linear model summary R output that our coefficient of determination is

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

This is quite close to one, although could be closer, and means that around 93% of the variation in the data is explained by the model, or in other words, the data fits the model well but not perfectly. This model would thus be good for many applications.

It is worth noting however that we are not necessarily looking for a model that perfectly explains all the variation in the data, as we would likely be overfitting our training data, or in a situation where $n = p$ which is uninformative. Thus a model with a higher R^2 than this model may not be favourable if this were the case.

Question 5.

When computing confidence intervals using the t statistic, we must further assume that the errors in our model are normally distributed.

We know that each regressor estimate $\hat{\beta}_i$ is normally distributed with distribution

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2 g^{ii})$$

where g^{ii} is the i^{th} diagonal element of the matrix $\mathbf{G}^{-1} = (\mathbf{X}^T\mathbf{X})^{-1}$. Using the fact that $\hat{\beta}_i$ is independant from the SSE and that $(n-p)\hat{\sigma}^2 \sim \sigma^2\chi^2_{n-p}$, we may standardise $\hat{\beta}_i$ with the transformation

$$\frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}\sqrt{g^{ii}}} \sim t_{n-p}$$

where t_{n-p} is the Student's t-distribution in $n-p$ degrees of freedom. Knowing this, we may contruct a $100\gamma\%$ confidence interval for $\hat{\beta}_i$ as

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

Let us illustrate this by calculating a 95% confidence interval for the intercept by hand. We achieve the following lower and upper bounds.

```
l.intercept_lower = (
  l.beta[1] - qt(1-(1-0.95)/2, df=n-p)*sqrt(l.S2.hat)*sqrt(diag(l.XtX_inv)[1])
)
l.intercept_upper = (
  l.beta[1] + qt(1-(1-0.95)/2, df=n-p)*sqrt(l.S2.hat)*sqrt(diag(l.XtX_inv)[1])
)
c(l.intercept_lower, l.intercept_upper)
```

```
## [1] -0.6604609  6.2595479
```

We may cross reference this result for the output of the R function `confint` which returns the 95% confidence interval for all regressors, given as follows.

	2.5 %	97.5 %
(Intercept)	-0.6604609	6.2595479
oil	-0.3075243	0.1158756
filler	0.4302737	0.6193683
oil:filler	-0.0157215	-0.0045820

As we can see, the bounds for the intercept we calculated manually match those of the `confint` output.

We note also that the confidence inervals for the intercept and oil both include zero, suggesting that these regressors may not be informative.

Part B

Question 1.

Using the same methodology and the same assumptions as for Part A, we build the new $\mathbf{X}$ matrix as

```
oil2 = oil^2
filler2 = filler^2
q.X = cbind(ones, oil, filler, interact, oil2, filler2)
```

and estimate $\hat{\vec{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{y}$ as

```
q.XtX_inv = solve(t(q.X)%*%q.X)
q.Xty = t(q.X)%*%visc
q.beta = q.XtX_inv%*%q.Xty
q.beta
```

```
##                  [,1]
## ones      6.016644748
## oil      -0.206571079
## filler    0.143467750
## interact -0.012325436
## oil2      0.006879675
## filler2   0.006597411
```

Which we can check against the R output

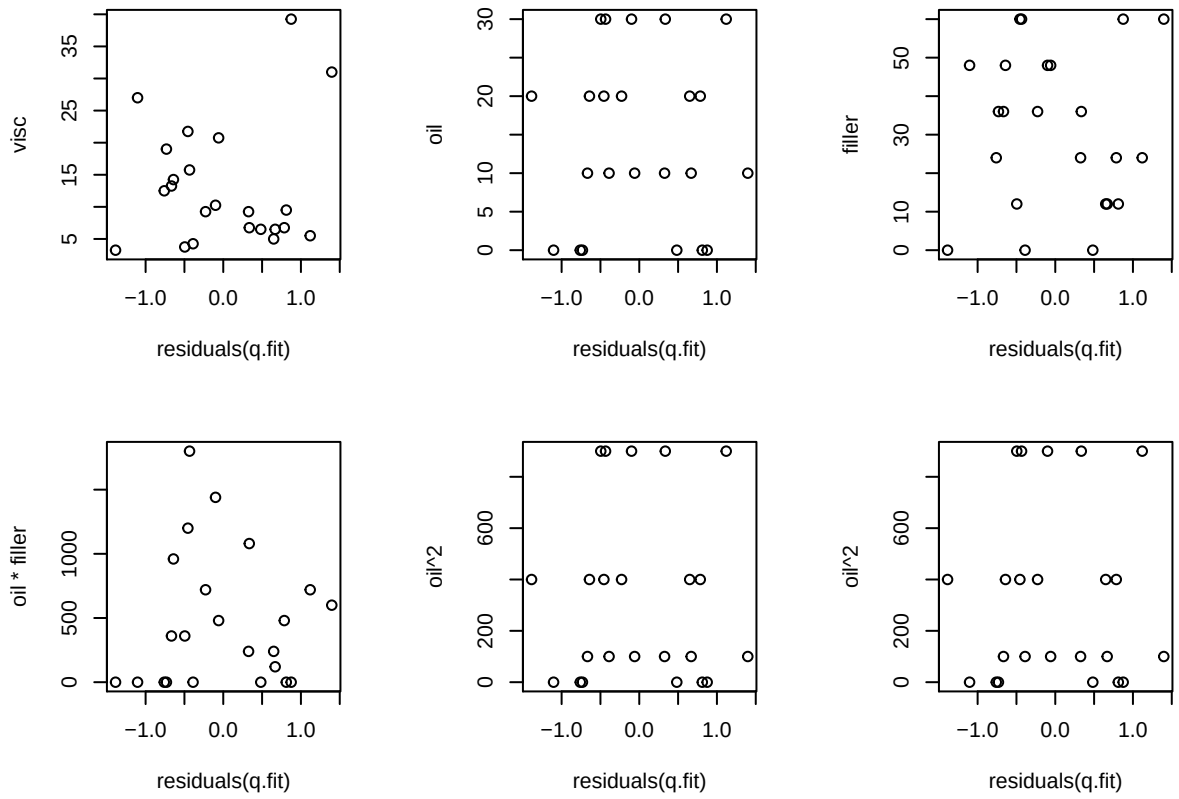
```
q.fit = lm(visc ~ oil*filler + I(oil2) + I(filler2))
summary(q.fit)
```

```
##
## Call:
## lm(formula = visc ~ oil * filler + I(oil2) + I(filler2))
##
## 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 ***
## I(oil2)      0.0068797  0.0018223   3.775 0.001510 **
## I(filler2)   0.0065974  0.0005220  12.640 4.53e-10 ***
## oil:filler  -0.0123254  0.0008759 -14.071 8.50e-11 ***
## ---
## 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
```

And so our fitted model becomes

$$\hat{y} = 6.0166447 - 0.2065711 \cdot \text{oil} + 0.1434677 \cdot \text{filler} - 0.0123254 \cdot \text{oil} \cdot \text{filler} + 0.0068797 \cdot \text{oil}^2 + 0.0065974 \cdot \text{filler}^2.$$

We may once more test our assumptions on the errors of our model.



Here, we see no evidence of heterogeneity, and we can be confident in our assumptions on $\overrightarrow{\varepsilon}$.

Question 2.

We use the same methodology as in Part A to estimate the variance of the response as

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

The new quadratic variables explain more of the variation in the data, resulting in the variance of the response decreasing.

Question 3.

The coefficient of determination can also be calculated the same way as in Part A, or equivalently taken directly from the R output, as

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

This is clearly much closer to one than our model in Part A, because our new model is more informative, although less parsimonious, as it has more variables. It is possible however that our model is overfitting our data when we have such a high R^2 value.

It is worth noting that when we add more parameters to our model, we expect the R^2 to increase since new variables will explain some more of the variation in the data. R^2 may thus not be an ideal measure of model performance as a result of this, since optimising for it will often result in building a highly complex model that may not generalise, instead of a more parsimonious model.

Question 4.

The methodology for calculating the confidence intervals for each of the model parameters is identical to that performed in Part A and with the same assumptions. Similarly to Part A, we are able to retrieve the output directly from R with the `confint` function.

	2.5 %	97.5 %
(Intercept)	4.7574040	7.2758855

	2.5 %	97.5 %
oil	-0.3323352	-0.0808069
filler	0.0725261	0.2144094
I(oil2)	0.0030350	0.0107243
I(filler2)	0.0054962	0.0076987
oil:filler	-0.0141735	-0.0104774

This time we note that none of the regressors has a confidence interval including zero, suggesting that they are all important to prediction and our model has a good structure.

Question 5.

Our quadratic model has a lower variance that our linear model as well as a higher R^2 since it has more variables. We may thus suggest that since it explains more of the variation in our data, the quadriatic model is better.

We can further this claim by noting that in our linear model, the confidence intervals for two of the regressors include zero. This means that two regressors are not statistically significant in the linear model, and the model is wrong. The same is not true in the quadratic case, where no regressors have a 95% confidence interval including zero. The quadratic model then has a better structure with higher R^2 and lower variance in the response.

We thus say that our quadratic model is better than our linear model.

Part C

We assume that the errors in our models are normally distributed for the statistical tests.

We have two models from parts A and B, which we will call the reduced model and the full model respectively,

$$\hat{y}_{\text{Reduced}} = \overrightarrow{\beta}_0 - \overrightarrow{\beta}_1 \cdot \text{oil} + \overrightarrow{\beta}_2 \cdot \text{filler} - \overrightarrow{\beta}_3 \cdot \text{oil} \cdot \text{filler}$$
$$\hat{y}_{\text{Full}} = \overrightarrow{\beta}_0 - \overrightarrow{\beta}_1 \cdot \text{oil} + \overrightarrow{\beta}_2 \cdot \text{filler} - \overrightarrow{\beta}_3 \cdot \text{oil} \cdot \text{filler} + \overrightarrow{\beta}_4 \cdot \text{oil}^2 + \overrightarrow{\beta}_5 \cdot \text{filler}^2$$

and we wish to test the null hypothesis that our quadratic terms might reasonably be zero, and thus that the more parsimonious reduced model is sufficient for effective prediction.

Let $\overrightarrow{\beta}_{\text{quad}} = (\overrightarrow{\beta}_4 \overrightarrow{\beta}_5)^T$, our null hypothesis is thus $H_0 : \overrightarrow{\beta}_{\text{quad}} = \overrightarrow{0}$, with the alternative $H_A : \overrightarrow{\beta}_{\text{quad}} \neq \overrightarrow{0}$. The test statistic for this hypothesis is

$$F = \frac{(\frac{SSE_R - SSE_F}{q})}{(\frac{SSE_F}{n-p})}$$

which follows the $F_{q,n-p}$ distribution, where $q = d_R - d_F$ and $d_F = n - p$. We may use our past estimates to calculate this F statistic as

```
n = length(visc)
p = length(q.fit[[1]])
q = length(q.fit[[1]]) - length(l.fit[[1]])
F.stat = ((l.sse - q.sse) / q) / (q.sse/(n-p))
F.stat
```

```
##          [,1]
## [1,] 83.16585
```

We will test this hypothesis at the 95% confidence level, meaning that our critical value is

```
## [1] 3.591531
```

As we can see, our F statistic is far greater than our critical value and thus we reject the null hypothesis in favour of the alternative and conclude that at least one of the quadratic terms in the model is statistically significant.

We can double check this using the native ANOVA table function in R

```
anova(l.fit, q.fit)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
19	135.21725	NA	NA	NA	NA
17	12.53844	2	122.6788	83.16585	0

where we find again that the supplementary quadratic terms contribute significantly to the variation in the model, and thus are statistically significant. The p-value for this test is very close to zero, and thus we can conclude the above with confidence.

Part D

Question 1.

Given the result achieved in Part C, we will use the regression model with quadratic terms as our fitted model for the rest of this coursework. Using 10 phr of oil and 50 phr of filler, we can calculate the mean response via $\hat{\overrightarrow{y}} = \overrightarrow{f}_0^T \hat{\overrightarrow{\beta}}$, where $\overrightarrow{f}_0^T = (1 \quad 10 \quad 50 \quad 500 \quad 100 \quad 2500)$. In R, this is done as follows.

```
x.oil = 10
x.filler = 50
f0 = matrix(
  c(1, x.oil, x.filler, x.oil*x.filler, x.oil^2, x.filler^2),
  nrow=6,
  ncol=1,
  byrow = FALSE
)
y.hat = t(f0) %*% q.beta
y.hat
```

```
##          [,1]
## [1,] 22.1431
```

We calculate the standard deviation of the response for the model from the estimate of variance we calculated earlier.

```
q.S.hat = sqrt(q.S2.hat)
q.S.hat
```

```
## [1] 0.8588104
```

Using these information, we can calculate the confidence interval for our response for the vector $\overrightarrow{f}_0^T$, via

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

Implemented in R, we calculate this interval to be

```
df = df.residual(q.fit)
CI_L = y.hat - qt(0.975,df) * q.S.hat*sqrt(t(f0) %*% q.XtX_inv %*% f0)
CI_U = y.hat + qt(0.975,df) * q.S.hat*sqrt(t(f0) %*% q.XtX_inv %*% f0)
c(CI_L, CI_U)
```

```
## [1] 21.45403 22.83217
```

We may also retrieve these values automatically from the native `predict` function.

```
predict(
  q.fit,
  data.frame(
    oil=x.oil,
    filler=x.filler,
    interact=x.oil*x.filler,
    oil2=x.oil^2,
    filler2=x.filler^2
  ),
  interval="confidence"
)
```

```
##       fit      lwr      upr
## 1 22.1431 21.45403 22.83217
```

and find that they match our manual calculations.

Question 2.

Similarly, we can calculate the prediction interval for our response for the vector $\overrightarrow{f}_0^T = (1 \quad 10 \quad 50 \quad 500 \quad 100 \quad 2500)$, via

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

Implemented in R, this is

```
PI_L = y.hat - qt(0.975,df) * q.S.hat*sqrt(1+ t(f0) %*% q.XtX_inv %*% f0)
PI_U = y.hat + qt(0.975,df) * q.S.hat*sqrt(1+ t(f0) %*% q.XtX_inv %*% f0)
c(PI_L, PI_U)
```

```
## [1] 20.20457 24.08163
```

We also perform these calculations automatically with the native `predict` function to cross-reference our answers.

```
predict(
  q.fit,
  data.frame(
    oil=x.oil,
    filler=x.filler,
    interact=x.oil*x.filler,
    oil2=x.oil^2,
    filler2=x.filler^2),
  interval="predict"
)
```

```
##       fit      lwr      upr
## 1 22.1431 20.20457 24.08163
```

Question 3.

Since our confidence interval does not include 21 at the 95% level while our prediction interval does, the quality inspector will not be able to find out whether the settings suggested by the chemist are satisfactory from a single obervation since the single obervation may not yield the desired result of 21 even though the prediction interval at the same level does indeed contain 21. The quality inspector thus may not find the truth because they can observe a value around 21, but it is also possible that they will because the prediction interval includes it. The quality inspector would need to take multiple obervations and calculate their average so that their errors cancel out in order to accurately conclude.