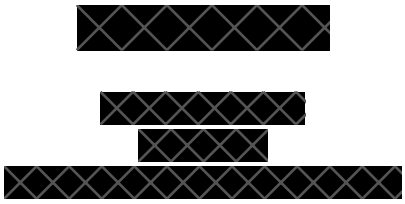
Predictive Analytics



October 24, 2020

Contents

1 Exploratory Data Analysis

1.1 Question 1

The Distribution of the Somatic Cell Count. From these we can see the that the SCC is very unevenly distributed. we can see the max value is 3199 and that 75% of the data is below 99. The plots (Figure: 1 2) are difficult to read and infer actual data from due to this huge range. The summary statistics do a much better job at describing the data in my opinion. We can read that the Minimum, mean, meadian and 1st Quartile values are 6, 126, 29, 16 respectively.

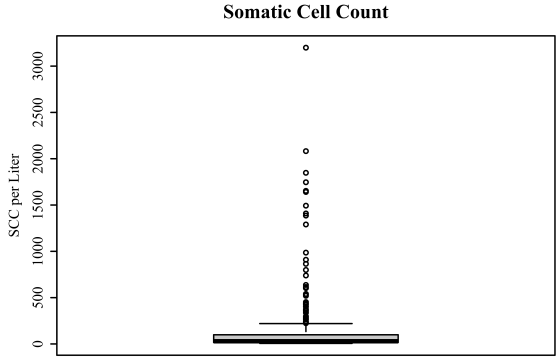


Figure 1: SCC boxplot

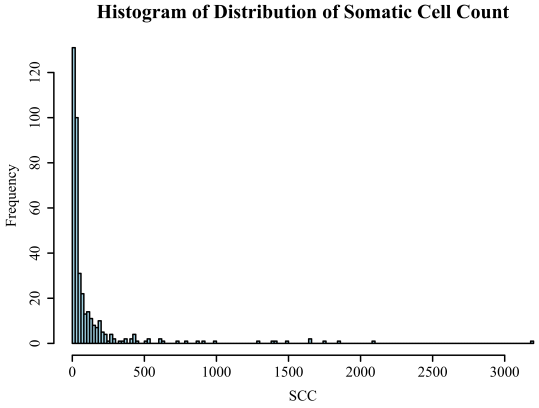


Figure 2: SCC histogram

Min	1st Quartile	Median	Mean	3rd Quartile	Max
6.0	16.0	29.0	126.4	99.0	3199.0

1.2 Question 2

The Distribution of the Log of The Somatic Cell Count is Described in 3 ways below.
The Boxplot (Figure: 3) shows that there are alot of outliers and that more data lies in the 3rd Quartile than the 2nd Quartile. The Histogram (Figure: 4) Shows us that the Data is Right Skewed. The Descriptive Statistics show us at a glance what the mean, median, min and max are. They are 3.773, 3.367, 1.792, and 8.071 respectively.

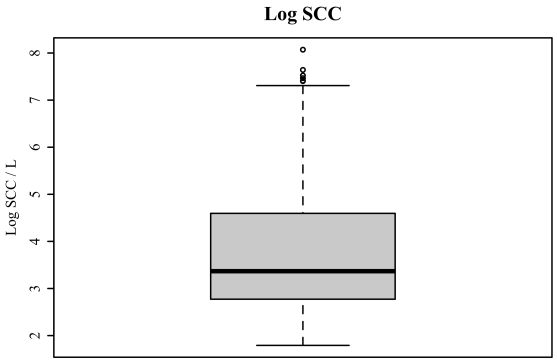


Figure 3: Log SCC boxplot

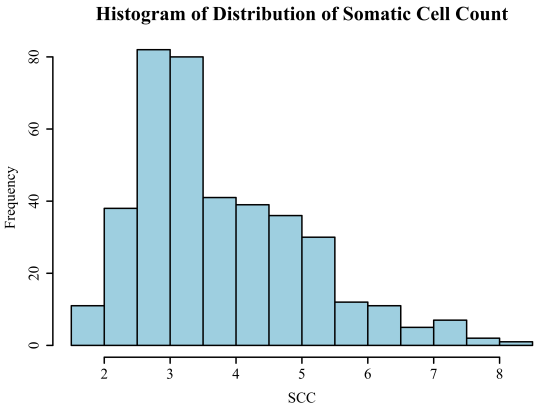


Figure 4: Log SCC histogram

Min	1st Quartile	Median	Mean	3rd Quartile	Max
1.792	2.773	3.367	3.773	4.595	8.071

1.3 Question 3

The Distribution of Protein is Described in 3 ways below.
The Boxplot (Figure: 5) shows us that there are only 3 outliers and that the data is roughly evenly distributed. The Histogram (Figure: 6) shows us that the Data is normally distributed and we can see the outlier at the very left. The Descriptive Statistics show us at a glance any information that we couldn't infer easily from the Graphs. It shows that the Minimum value is 2.110, The 1st Quartile lies below 3.425, the median is 3.610, the mean is 3.606, 75% of the data lies below 3.810, and the max value is 4.390

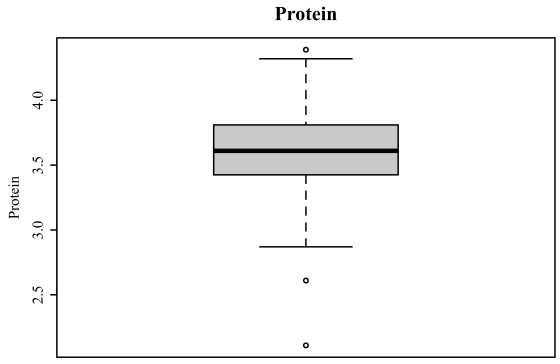


Figure 5: Protein boxplot

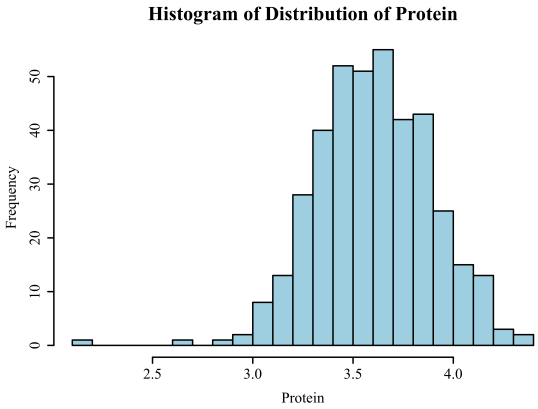


Figure 6: Protein histogram

Min	1st Quartile	Median	Mean	3rd Quartile	Max
2.110	3.425	3.610	3.606	3.810	4.390

1.4 Question 4

The Distribution of Casein is Described in 3 ways below.
The Boxplot (Figure: 7) shows us that there are only 3 outliers and that the data is roughly evenly distributed. The Histogram (Figure: 8) shows us that the Data is normally distributed and we can see the outlier at the very left. The Descriptive Statistics show us at a glance any information that we couldn't infer easily from the Graphs. It shows that the Minimum value is 1.2, The 1st Quartile lies below 2.640, the median is 2.8, the mean is 2.793, 75% of the data lies below 2.97, and the max value is 3.440.

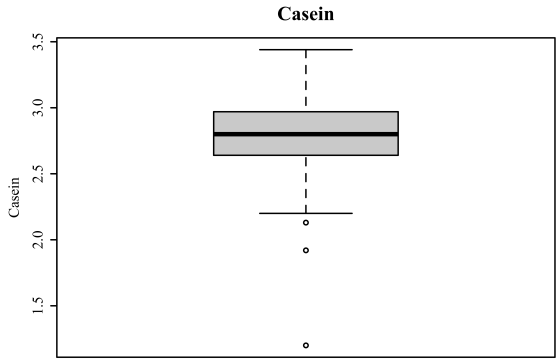


Figure 7: Casein boxplot

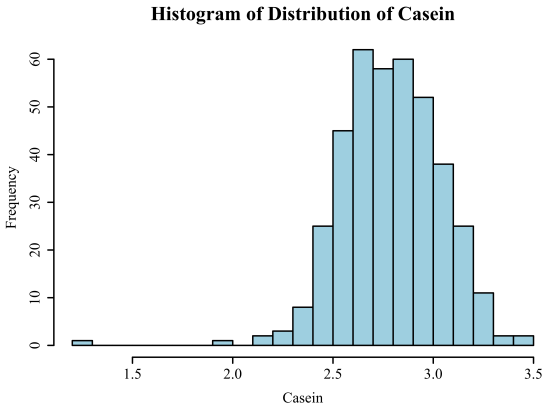


Figure 8: Casein histogram

Min	1st Quartile	Median	Mean	3rd Quartile	Max
1.200	2.640	2.800	2.793	2.970	3.440

1.5 Question 5

Below I have Illustrated the Frequency of the Concentrated Feed Figure 9 and the Proportion of the Concentrated Feed Figure 10. We can see both these graphs pretty much are showing us the same thing. That there are over twice the amount of cows on 2.5% Concentrate Feed than cows on 0%.

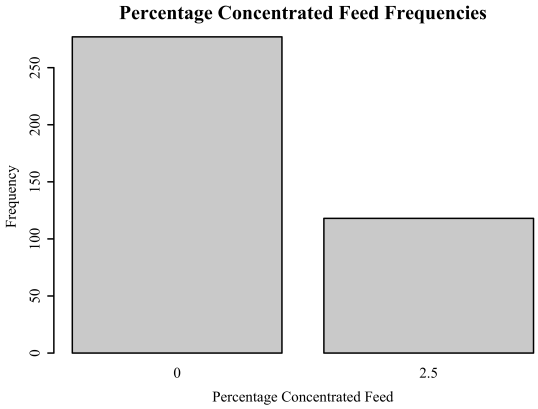


Figure 9: Concentrated Feed Frequency

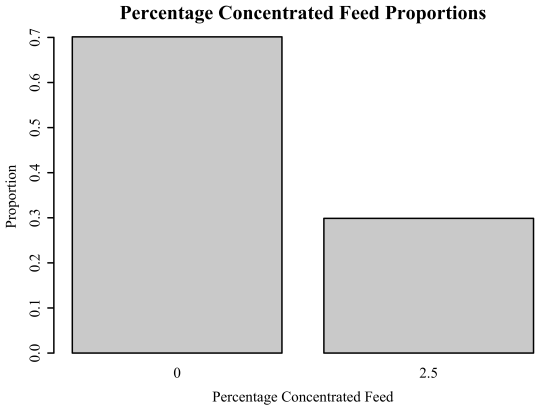


Figure 10: Concentrated Feed Proportions

1.6 Question 6

in Figure 11 we see how the log of somatic cell counts scc varies with respect to the variable concentrate feed. from the boxplots we can see that for the Concentrated Feed percent 0 that there is only one outlier and that there are more points within the 3rd Quartile than in the 2nd Quartile. for the Concentrated Feed percent 1 that there are quite a few outliers and that the data is roughly evenly Distributed if we ignore the outliers.

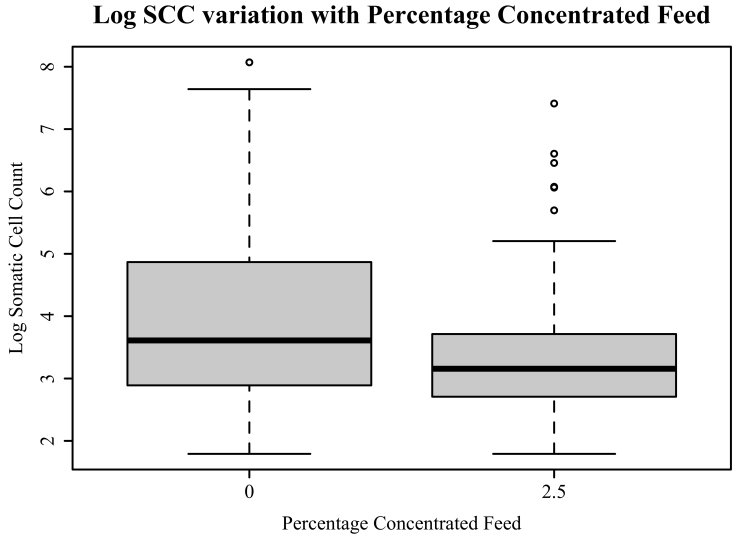


Figure 11: Concentrated Feed with Respect to Log SCC

The Descriptive Statistics show us so many different measuers about hwo the Log SCC varies with respect to the concentrated Feed. We can for instance see that the means, standard deviations, and median are quite close. We also can see at a glance that there is over 150 more datpoints in the gorup 0 than in group 2.5. We can also see the skew at a glace. Finally something of interest is that the Standard Error (se) is very small for both.

Group	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
0	277	3.94	1.32	3.61	3.81	1.34	1.79	8.07	6.28	0.78	0.01	0.08
2.5	118	3.38	1.03	3.16	3.26	0.67	1.79	7.41	5.62	1.4	2.27	0.09

1.7 Question 7

1.7.1 Question 7 - 1

Below are 2 plots showing Casein vs Log SCC (Figure 12 13) To be honest the Variables are not really in a linear relationship. The correlation between Protein and the Log SCC is essentially 9% which is not surprising as when we look at the plots especially the Scatter matrix No relationship jumps out of us. we can see some clustering but that is about it.

Correlation Between Protein and Log SCC: 0.09013953

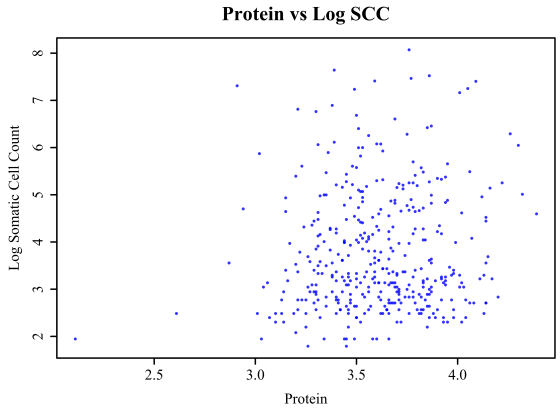


Figure 12: Protein vs Log SCC Plot

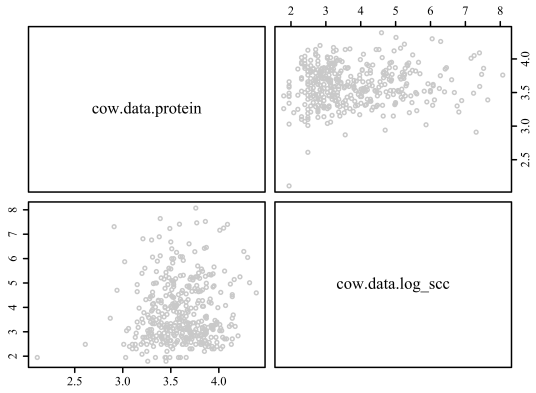


Figure 13: Scatter Matrix for Protein vs Log SCC

1.7.2 Question 7 - 2

Below are 2 plots showing Casein vs Log SCC (Figure 14 15) To be honest the Variables are not really in a linear relationship. The correlation between Casein and the Log SCC is very little wand that is not very surprising as we see in the figures that there isn't much of a relationship at all

Correlation Between Casein and Log SCC: 0.05871938

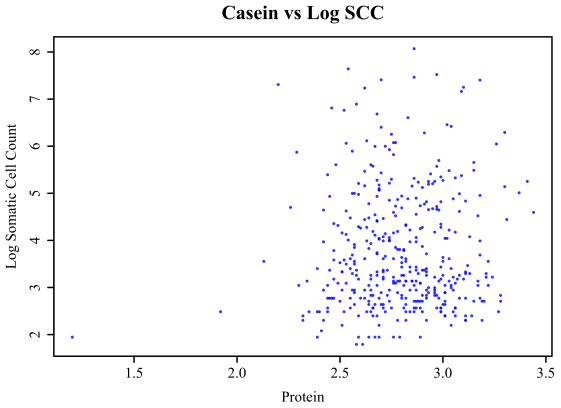


Figure 14: Casein vs Log SCC Plot

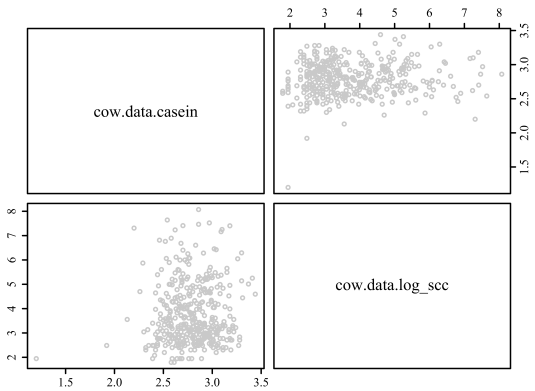


Figure 15: Scatter Matrix for Casein vs Log SCC

1.8 Question 8

From both the Casein and Protein versus the Log SCC I chose Protein as we can see from the Correlation that Protein is more correlated with Log SCC than the Casein is with the Log SCC.

2 Regression Model

2.1 Question 1

We fit the Regression Model as Outlined Below in the R code.

$$y = \beta_0 + \beta_1 * + \epsilon \tag{1}$$

This is the General form where β_0 is the Y-Intercept. i.e. the y value at the position $x = 0$. For our model $\beta_0 = 2.3462121$. β_1 is the slope of the model, Essentially for every 1 Unit increase in x we get a β_1 unit increase in y. For our model $\beta_1 = 0.3955028$. (We ignore ϵ as we assume its 0 for Simple Linear Regression) So we end up with the equation:

$$y_i = 2.3462121 + 0.3955028 \cdot x_i \tag{2}$$

$$Y = 2.3462121 + 0.3955028 \cdot X \tag{3}$$

We can see this Line Plotted below in (Figure 13)

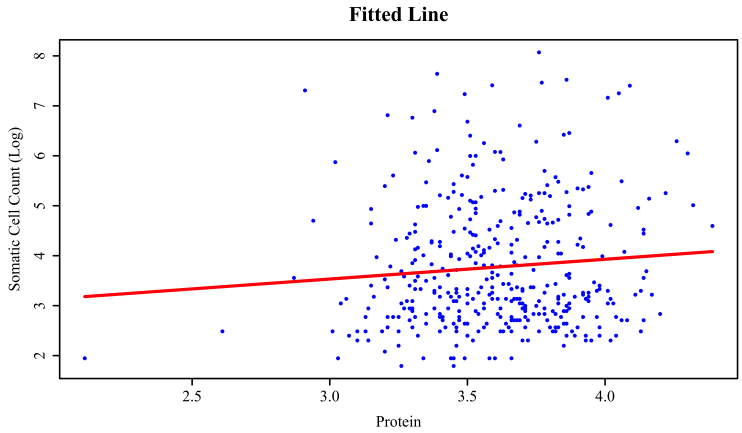


Figure 16: Fitted Regression Line

2.1.1 Code

```
linearMod <- lm(log_scc ~protein, data=cow.data)
linearMod$coefficients
summary(linearMod)
```

2.2 Question 2

The Estimate of the intercept term is 2.3462121 which we can interpret as the Log of the Somatic Cell count when the protein is 0.

2.3 Question 3

The Estimate of the Slope term is 0.3955028 which we can interpret as the increase in Log SCC for every one unit increase Protein

2.4 Question 4

Calculations for the Variances:

For β_0 we calculate the variance with the following Equations:

$SXX = \sum_{i=1}^{n}(x_i - \bar{X})^2$
$MSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
$Var(\beta_0) = MSE \cdot (\frac{1}{n} + \frac{\bar{X}^2}{SXX})$

Then we simply plug in our Values
$\bar{X} = 3.606278$
$\bar{X}^2 = 13.00524$
$SXX = 32.62063$
$MSE = 1.584986$
$Var(\beta_0) = 1.584986 \cdot (\frac{1}{395} + \frac{13.00524}{32.62063}) = 0.6359172$

For β_1 we calculate the variance as follows:

$MSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
$SXX = \sum_{i=1}^{n}(x_i - \bar{X})^2$
$Var(\beta_1) = \frac{MSE}{SXX}$

Then we simply plug in our Values again
$\bar{X} = 3.606278$
$SXX = 32.62063$
$MSE = 1.584986$
$Var(\beta_1) = \frac{1.584986}{32.62063} = 0.04858844$

Interpretations of the Variances:

The Variance for β_0 is $Var(\beta_0) = 0.6359172$ We can interpret this as saying that there is very Little deviation around β_0 in which the true Value of beta 0 might lie. We Like our Estimators to have little variance.

The Variance for β_1 is $Var(\beta_1) = 0.04858844$ We can interpret this the same as β_0 as saying that there is very Little deviation around β_1 in which the true Value of beta 0 might lie. Again this is good because We Like our Estimators to have little variance.

2.5 Question 5

Then Confidence Interval for β_0 is:

$$\beta_0 \pm t_{1-\frac{\alpha}{2},n-2} \cdot \sqrt{Var(\beta_0)} \tag{4}$$

We can Plug in our values to get:
$t_{1-\frac{\alpha}{2},n-2} = 1.966019$
$2.346212 \pm 1.966019 \cdot \sqrt{2.346212}$

So the Values of Confidence Interval for β_0 is 0.7784221 to 3.9140022 this means that we are 97.5% that the true value for β_0 is inside this range. which in the context of this problem meas we are 97.5% Confident that the Y-Intercept, The Log SCC when the protein is measured at 0, is withing the range 0.7784221 to 3.9140022

2.5.1 Code

```
N    = length(cow.data$protein)
MSE = sum(linearMod$residuals^2/(N-2))
SXX = sum((cow.data$protein-mean(cow.data$protein))^2)
VARB0 = MSE*(1/ N + (mean(cow.data$protein)^{2}/SXX))

alpha=0.05
beta0 = linearMod$coefficients[1]
c(beta0 - qt(1-alpha/2,N-2)*sqrt(VARB0),
  beta0 + qt(1-alpha/2,N-2)*sqrt(VARB0))

confint(linearMod)
```

2.6 Question 6

Then Confidence Interval for β_1 is:

$$\beta_1 \pm t_{1-\frac{\alpha}{2},n-2} \cdot \sqrt{Var(\beta_1)} \tag{5}$$

We can Plug in our values to get:
$t_{1-\frac{\alpha}{2},n-2} = 1.966019$
$0.6359172 \pm 1.966019 \cdot \sqrt{0.6359172}$

So the Values of Confidence Interval for β_1 is -0.03786245 to 0.82886809 this means that we are 97.5% that the true value for β_1 is inside this range. which in the context of this problem meas we are 97.5% Confident that Slope, The increase in Log SCC with every one unit increase in protein, is within the range -0.03786245 to 0.82886809

2.6.1 Code

```
N    = length(cow.data$protein)
SSE  = sum((cow.data$log_scc - linearMod$fitted.values)^2)
MSE  = SSE/(N-2)
SXX  = sum((cow.data$protein - mean(cow.data$protein))^2)

VARB1 = MSE/SXX

beta1= linearMod$coefficients[2]
alpha=0.05
c(beta1 - qt(1-alpha/2,N-2)*sqrt(VARB1),
  beta1 + qt(1-alpha/2,N-2)*sqrt(VARB1))
confint(linearMod)
```

2.7 Question 7

***Hypothesis Test*:** $H_0 : \beta_0 = 0, H_a : \beta_0 \neq 0$

The Hypothesis H_0 is essentially stating that β_0 is not significantly Different from 0. In that regard H_a is stating that β_0 is significantly different from 0.

We are going to use a T-Test to test the Hypothesis

We can Calculate the Test Statistic with:

$T = \frac{\beta_0 - 0}{\sqrt{Var(\beta_0)}}$

The *Test Statistic* is $\frac{2.346212 - 0}{\sqrt{Var(2.346212)}} = 2.942165$

And we Calculate the Distribution Value:
$t_{1-\frac{0.05}{2}, n-2} = 1.966019$

The Hypothesis Test H_0 will be Rejected if the Absolute Value of our Test Statistic is Greater than the Real Distribution Value.

Since $|T| < t_{1-\frac{0.05}{2}, n-2}$ from above we can Reject H_0

The *P-Value* of this test is 0.003452356

This *P-Value* is the probability that the Null Hypothesis H_0 is actually true. By this we can say it is very unlikely that the True value for β_0 is actually 0 and can safely assume the best estimate is in fact 2.346212

In the context of this Problem we can Now safely assume that the Y-Intercept, The Log SCC, when the protein is 0, is 2.346212 and that probability that the Y-Intercept is actually 0 is 0.3%

2.8 Question 8

***Hypothesis Test*:** $H_1 : \beta_0 = 0, H_a : \beta_1 \neq 0$

The Hypothesis H_0 is essentially stating that β_1 is not significantly Different from 0. In that regard H_a is stating that β_1 is significantly different from 0.

We are going to use a T-Test to test the Hypothesis

We can Calculate the Test Statistic with:

$T = \frac{\beta_1 - 0}{\sqrt{Var(\beta_1)}}$

The *Test Statistic* is $\frac{0.3955028 - 0}{\sqrt{Var(0.3955028)}} = 1.794251$

And we Calculate the Distribution Value:
$t_{1-\frac{0.05}{2}, n-2} = 1.966019$

The Hypothesis Test H_0 will be Rejected if the Absolute Value of our Test Statistic is Greater than the Real Distribution Value.

Since $|T| < t_{1-\frac{0.05}{2}, n-2}$ as $|1.794251| < 1.966019$ from above, so we Fail to Reject H_0

The *P-Value* of this test is 0.07354166

This *P-Value* is the probability that the Null Hypothesis H_0 is actually true. While the P-Value is still very small (7%) we Still can use this to say that β_0 is not statistically significant from 0

In the context of this Problem we can Now safely assume that the Slope is 0. Which means for every 1 unit increase in protein we will have a 0 unit increase in the log SCC. Also as per the P-Value we are 7% we are right and while it may not seem much generally if $P < 0.05$ we fail to reject H_0

2.9 Question 9

The F-Test is a Test that Tests the following Hypothesis:

$$H_0 : \hat{Y}_i = \hat{\beta}_0 + \epsilon_i$$

$$H_a : \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \hat{X}_i + \epsilon_i$$

The Null Hypothesis States that the Model for the line that is $\hat{Y}_i = \hat{\beta}_0 + \epsilon_i$ is the best fit. The Alternative Hypothesis States That The Line $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \hat{X}_i + \epsilon_i$ is The better fit for the Model.
To test the Hypothesis we employ the F-test to check if we reject H_0 or we fail to reject H_0.

To get the Test Statistic we do the following Calculations:

$MSR = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{1}$
$MSE = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}$
$F = \frac{MSR}{MSE}$

We Plug in our values to get out Test Statistic: $F = \frac{5.1026}{1.584986} = 3.219335$

The F distribution value is: $F_{0.95,1,n-2} = 3.865229$

The *P-Value* is 0.9264583

What do These values mean in the context in the of the problem? Well as $F < F_{0.95,1,n-2} \rightarrow 3.219335 < 3.865229$ so now we Fail to Reject H_0 so this means that we assume $H_0 : \hat{Y}_i = \hat{\beta}_0 + \epsilon_i$ is correct and that The line $\hat{Y}_i = \hat{\beta}_0 + \epsilon_i$ is The better Model for the Problem. If we look at the P-Value we see that we are very certain of this. The Probability of The model is better Described by $\hat{Y}_i = \hat{\beta}_0 + \epsilon_i$ is 92.6%. In other words. There is only an 7.4% chance we are wrong to fail to reject H_0

2.10 Question 10

To Calculate The R^2 Value for the Model we use the following Formula:

$SSY = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$
$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
$R^2 = SSY - SSE/SSY$

When we plug in our Data we get the $R^2 = 0.008125134$. This is a measure that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, R^2 shows how well the data fit the regression model (the goodness of fit). Our R^2 reveals that only 0.8% of the data fit the regression model

2.11 Question 11

The Residual Standard Error (or the Root Mean Squared Error [RMSE]) is 1.258962. Which is Calculated by $RMSE = \sqrt{\frac{SSE}{n-2}}$. The Root Mean Squared Error is the average amount that the real values of Y differ from the predictions provided by the regression line. So if it is close to 0 it means the real values do not Differ from the predictions by much whereas when it is a large number it indicates the Real values are not closely aligned with the plot.

2.12 Question 12

What is the confidence intervals for the estimated values of Y. The confidence Intervals for the Estimated value of Y is the Interval in which we believe there is a 95% probability that the true value of Y lies in.

Firstly we must Calculate the Confidence Interval for $\hat{Y}$

$$\hat{y} \pm t_{1-\frac{\alpha}{n-2}} \cdot \sqrt{Var(y)} \tag{6}$$

The Variation in Y is Calculated with $var(Y) = MSE \cdot (\frac{1}{n} + \frac{(X-\bar{X})^2}{SXX})$

Below in Figure 14 I have plotted the Regression line with the Confidence Interval for the Line (which is the confidence Interval for the estimations of Y)

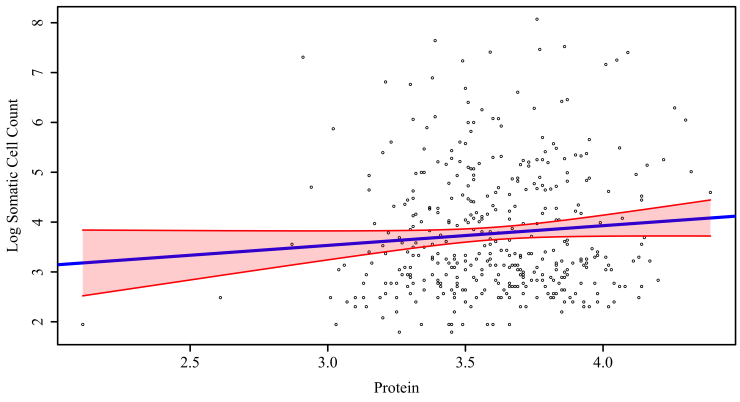


Figure 17: Error Bars - Confidence Interval for Fitted Line

Based on this plot I would say the confidence Intervals are surprisingly small they also have a squeezed nature in the middle where the Confidence interval gets closer to the Regression Line and then Fans out again. So we can see that the Confidence Interval Range is the area That our model is 97.5% (because of $1-\frac{\alpha}{2}, where \alpha = 0.05$) confident that the true Value of Y is in there.

2.12.1 Code

```
alpha = 0.05

N = length(cow.data$protein)
N

SXX = sum((cow.data$protein - mean(cow.data$protein))^2)
SXX

MSE = sum(linearMod$residuals^2/(N-2))
MSE

VAR_Y = MSE*(1/N+(cow.data$protein-mean(cow.data$protein))^2/SXX)
VAR_Y

Yhat = fitted(linearMod)
Yhat

YCI_HI = Yhat + qt(1-alpha/2,N-2)*sqrt(VAR_Y)
YCI_HI

YCI_LO = Yhat - qt(1-alpha/2,N-2)*sqrt(VAR_Y)
YCI_LO

png(file = pathf("errorbars.png"),
    width     = 5,
    height    = 3,
    units     = "in",
    res       = 1200,
    pointsize = 5)
par(mfrow=c(1,1))
par(family = 'Times New Roman', cex.axis=1.5, cex.lab=1.5, cex.main=2)
plot(cow.data$protein,cow.data$log_scc,
     xlab="Protein", ylab="Log Somatic Cell Count",
     col="black", type = "p", cex=0.5, pch=1, lwd=0.5)
abline(linearMod, col="blue", lwd=2)

df1 = data.frame(cbind(cow.data$protein,YCI_HI))
orderidx = order(df1[, "V1"])
df1 = df1[orderidx, ,drop=FALSE]
lines(df1$V1,df1$YCI_HI,col="red", type = "l", lwd=1 )

df2 = data.frame(cbind(cow.data$protein,YCI_LO))
  orderidx = order(df2[, "V1"])
df2 = df2[orderidx, ,drop=FALSE]
lines(df2$V1,df2$YCI_LO,col="red", type = "l", lwd=1 )

polygon(c(df1$V1, rev(df2$V1)), c(df1$YCI_HI, rev(df2$YCI_LO)), col=rgb(1, 0, 0,
    0.2), border = NA)
dev.off()
```

2.13 Question 13

The Assumptions of the linear regression modelrequired for small sample inference

Assumption 1 - Variation in X

Assumption 2 - Random Sampling

Assumption 3 - Linearity in Parameters

Assumption 4 - Zero Conditional Mean

Assumption 5 - Homoskedacity

Assumption 6 - Normality of Errors

2.14 Question 14

I do not think the residuals satisfy the Assumptions of the Linear Model
Supporting Figures Below
To Explain my Rationale I will go through each Asumption and check it against the Linear Model Residuals.

Assumption 1 - Variation in X?
Yes we do have Variation in X we can see from Figure 18 Below. X is not completely Normally Distributed but it is not far off.

Assumption 2 - Random Sampling?
Yes I would also Agree That we do have Random Sampling.

Assumption 3 - Linearity in Parameters?
Now, It could be argued That there is some sort of upward trend I think but I would say a look at Figure 20 shows there's not much Linear about the Scatter Plot from what we have seen throughout this Report That the R^2 value is very small and that $\hat{Y}_i = \hat{\beta}_0 + \epsilon_i$ is a good "fit" for this model that there not linearity in the parameters.

Assumption 4 - Zero Conditional Mean?
If we look at the Summary Statistics table for the Residuals:

Min	1st Quartile	Median	Mean	3rd Quartile	Max
-1.9189	-0.9434	-0.3983	0.0000	0.8048	4.2373

We can see that there is a mean of 0. also if we look at Figure 19 we see alot of Variance nothing to show that it is skewed so it has generally constant Variance.

Assumption 5 - Homoskedacity?
To check for Homoskedacity we should take a look at Figure 20 and ask what is the Variation about it? if the Variation about the fitted line is constant then we have Homoskedacity. From a look at the Figure we can say that there is not a Constant Variance About the Fitted line.

Assumption 6 - Normality of Errors?
To Check for Normality of Errors we must look at Figure 21. If there was Normality of Errors we should see the Error Density plot fit a normal Distribution. However it is obvious from the plot that the erros do not fit a normal distribution so we do not have Normality of Errors

Having Stepped Through all the Assumptions we See that the residuals do not satisfy them all.

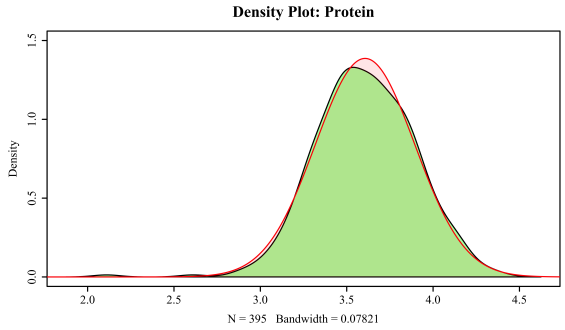


Figure 18: Protein Density Plot

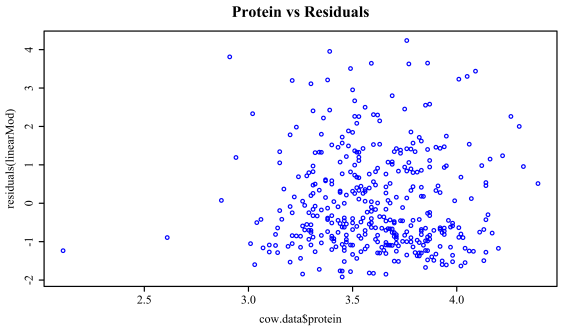


Figure 19: Protein vs Residuals

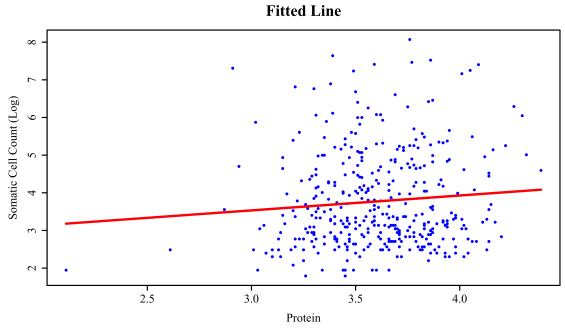


Figure 20: Fitted Line

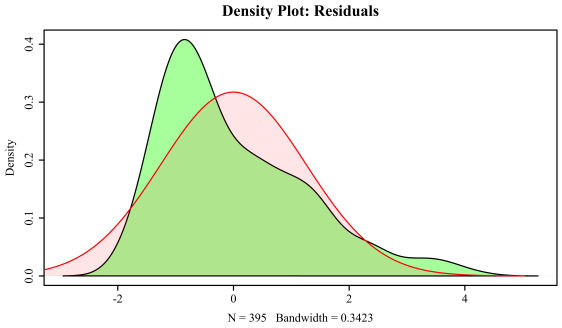


Figure 21: Noramility of Errors

I Eoghan Hogan confirm that this assignment is my own work. I have not copied in part or whole or otherwise plagiarised the work of other students and/or persons. I confirm that I have read and understood the UCD School of Mathematics and Statistics regulations on plagiarism in the Week 5 folder on bright space.