Unit 10 HW Solutions

## Question 1 (60 points total)

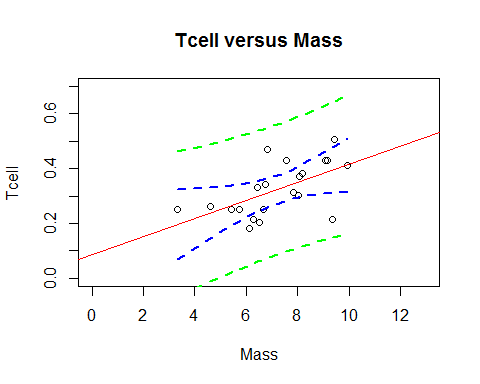
Black-eared wheatears are small birds of Spain and Morocco. Males of the species demonstrate an exaggerated sexual display by carrying many heavy stones to nesting cavities. This 35-gram bird transports, on average, 3.1 kg of stones per nesting season! Different males carry somewhat different sized stones, prompting a study of whether larger stones may be a signal of higher health status. M. Soler et al. calculated the average stone mass (g) carried by each of 21 male black-eared wheatears, along with T-cell response measurements (in mm) reflecting their immune system strengths. Analyze the data and write a statistical report (by answering the questions below); treat the T-cell as the response and the stone mass as the explanatory variable. You may assume all criteria for regression and related t-tests are met. You can find the data for this problem on 2DS. (Male Display Data Set)

Analyze the data, providing the following:

### Part A (6 points)

Provide a scatterplot with 99% confidence intervals of the regression line and 99% prediction intervals of the regression line. Please do this in R.

display <- read.csv("C:/Users/Charles/Documents/SMU/Online Teaching/MSDS 6371 - Statistical Foundations for Data Science/UNIT 10/Male Display Data Set.csv")  
  
##Initial scatterplot  
plot(display[,1], display[,2], xlim = c(0, 13), ylim =c(0, 0.7), xlab = "Mass",   
ylab = "Tcell", main = "Tcell versus Mass")  
displayDF <- data.frame(display)  
  
##Regression Model  
displaylm <- lm(Tcell ~ Mass, data = display)  
  
##Add the regression line to the existing scatterplot  
abline(displaylm, col = "red")  
  
##Create "new" data to make confidence and prediction intervals  
newx <- displayDF$Mass  
newx <- sort(newx)  
  
##Confidence Interal  
prd\_c <- predict(displaylm, newdata = data.frame(Mass = newx), interval = c("confidence"),   
type = c("response"), level = .99)   
  
##Prediction Interval  
prd\_p <- predict(displaylm,newdata = data.frame(Mass = newx), interval = c("predict"),   
type = c("response"), level = .99)  
  
##Add prediction and confidence intervals to the scatterplot  
lines(newx, prd\_c[,2], col = "blue", lty = 2, lwd = 2)  
lines(newx,prd\_c[,3], col = "blue", lty = 2, lwd = 2)  
lines(newx,prd\_p[,2], col = "green", lty = 2, lwd = 2)  
lines(newx,prd\_p[,3], col = "green", lty = 2, lwd = 2)



### Part B (2 points)

Provide a table showing the t-statistics and p-values for the significance of the regression parameters and (as different from 0). Please do this in R.

summary(displaylm)

##   
## Call:  
## lm(formula = Tcell ~ Mass, data = display)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.18138 -0.04673 0.01796 0.04219 0.15999   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.08750 0.07868 1.112 0.27996   
## Mass 0.03282 0.01064 3.084 0.00611 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.08102 on 19 degrees of freedom  
## Multiple R-squared: 0.3336, Adjusted R-squared: 0.2986   
## F-statistic: 9.513 on 1 and 19 DF, p-value: 0.006105

### Part C (12 points)

Using the output in (b), show all 6 steps of each hypothesis test. (That’s one test for and one test for ). Find critical values in R. Your conclusion should include a confidence interval. Use alpha = 0.01.

confint(displaylm, level=0.99)

## 0.5 % 99.5 %  
## (Intercept) -0.137591407 0.31258537  
## Mass 0.002376893 0.06326609

qt(0.995, 21-2)

## [1] 2.860935

**First, the slope (though order doesn’t matter).**

**(1 point) Step 1 - Hypotheses:**

**(1 point) Step 2 - Identification of Critical Value:**

**(1 point) Step 3 - Value of Test Statistic: :**

**(1 point) Step 4 - Give p-value:**

**(1 point) Step 5 - Decision: Reject**

**(1 point) Step 6 - Conclusion: There is sufficient evidence at the level of significance ( from the t-test) to conclude that there is a linear relationship between the average stone mass and Tcell response. A 99% confidence interval for the slope of the regression line is .**

**Next, the intercept.**

**(1 point) Step 1 - Hypotheses:**

**(1 point) Step 2 - Identification of Critical Value:**

**(1 point) Step 3 - Value of Test Statistic: :**

**(1 point) Step 4 - Give p-value:**

**(1 point) Step 5 - Decision: Fail to Reject**

**(1 point) Step 6 - Conclusion: There is not sufficient evidence at the level of significance ( from the t-test) to conclude that the y-intercept is not equal to 0. A 99% confidence interval for this value is .**

### Part D (2 points)

State the regression equation. Be careful to use the mean Tcell or predicted Tcell, rather than just Tcell.

*Note: Either option below is acceptable.*

### Part E (2 points)

Interpret the slope in the model (regression equation).

**For every increase of 1 gram in the stone mass, the estimated T-cell response increases by 0.03282 mm.**

### Part F (2 points)

Interpret the y-intercept in the model (regression equation).

**When the stone mass is equal to 0, the estimated T-cell response is 0.0875 mm.**

### Part G (4 points)

Find and interpret the 99% **confidence** interval for the mean t-cell response conditional on a stone mass of 4.5 grams. Please do this directly in R.

*Note: 2 points for the interval, 2 points for the interpretation.*

newpoint <- data.frame(Tcell=NA, Mass=4.5)  
predict(displaylm, newpoint, interval="confidence", level=0.99)

## fit lwr upr  
## 1 0.2351937 0.1385665 0.3318209

**With 99% confidence, the mean t-cell response for a stone mass of 4.5 grams is between 0.139 and 0.332.**

### Part H (4 points)

Find and interpret the 99% **prediction** interval for the predicted t-cell response conditional on a stone mass of 4.5 grams. Please do this directly in R.

*Note: 2 points for the interval, 2 points for the interpretation.*

newpoint <- data.frame(Tcell=NA, Mass=4.5)  
predict(displaylm, newpoint, interval="prediction", level=0.99)

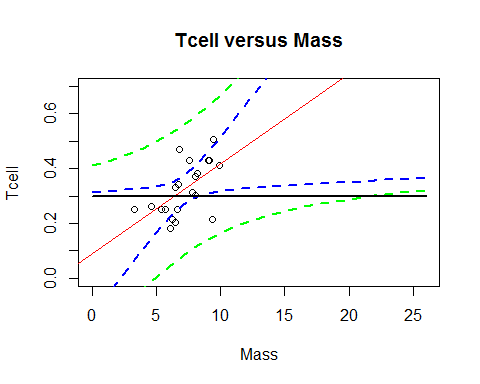
## fit lwr upr  
## 1 0.2351937 -0.01593192 0.4863193

**A 99% prediction interval for a t-cell response given a stone mass of 4.5 grams is -0.016 to 0.486. Because a response less than zero is nonsensical, a logical interval is (0 mm, 0.486 mm).**

### Part I (18 points total)

1. Using the graphical method (using your best judgment using the graphs from part (a)), find the following using R, as part (a) was done in R.
2. (3 points) 99% calibration interval for the **mean** t-cell response of 0.3.

##Extend the CIs and PIs so that calibration is more clear, though the results are doubtful outside the range of the data.  
  
##Scatterplot  
plot(display[,1], display[,2], xlim = c(0,26), ylim = c(0, 0.7), xlab = "Mass",   
ylab = "Tcell", main = "Tcell versus Mass")  
abline(displaylm, col = "red")  
  
##New data  
newrange <- 0:26  
newy <- rep(0.3, 27)  
  
##CI and PI  
prd\_c <- predict(displaylm, newdata = data.frame(Mass = newrange),   
interval = c("confidence"), type = c("response"), level = .99)   
prd\_p <- predict(displaylm, newdata = data.frame(Mass = newrange),   
interval = c("predict"), type = c("response"), level = .99)  
  
##Add CI and PI to scatterplot  
lines(newrange, prd\_c[,2], col = "blue",lty = 2, lwd = 2)  
lines(newrange, prd\_c[,3], col = "blue", lty = 2, lwd = 2)  
lines(newrange, prd\_p[,2], col = "green", lty = 2, lwd = 2)  
lines(newrange, prd\_p[,3], col = "green", lty = 2, lwd = 2)  
lines(newrange, newy, col = "black", lwd = 2)



**Considering that a mass less than zero is nonsensical, it looks like a 99% calibration interval for the mean t-cell response of 0.3 is (0g, 8g).**

1. (3 points) 99% calibration interval for a **single** t-cell response of 0.3.

**Considering that a mass less than zero is nonsensical, it looks like a 99% calibration interval for an individual t-cell response of 0.3 is (0g, 22g). Because the upper limit is so far out of the range of data, we should not put much “confidence” in its accuracy.**

1. Using **software directly**, find the following using R, as SAS does not provide calibration intervals directly. (R: package investr)
   1. (3 points) 99% calibration interval for the **mean** t-cell response of 0.3.

library(investr)  
calibrate(displaylm, y0=0.3, interval='inversion', mean.response=T, level=0.99)

## estimate lower upper   
## 6.474508 -4.389857 8.342649

**Considering that a mass less than 0 is nonsensical, the 99% calibration interval for a mean response of 0.3 is (0 g, 8.3 g).**

* 1. (3 points) 99% calibration interval for a **single** t-cell response of 0.3.

library(investr)  
calibrate(displaylm, y0=0.3, interval='inversion', level=0.99)

## estimate lower upper   
## 6.474508 -17.968869 21.921661

**Considering that a mass less than 0 is nonsensical, the 99% calibration interval for a single response of 0.3 is (0 g, 21.9 g).**

**Calculation in R is slightly different than that of the method described below and more exact than the visual method of course.**

\*\*Additional reference: [https://journal.r-project.org/archive/2014-1/greenwell-kabban.pdf\*\*](https://journal.r-project.org/archive/2014-1/greenwell-kabban.pdf**)

1. Interpret the following using the results from (1) and (2) above.
2. (3 points) 99% calibration interval for the **mean** t-cell response of 0.3.

**We are 99% confident that the estimated mass that would be needed to have a mean t-cell response of 0.3 mm is between 0g and 8.3g.**

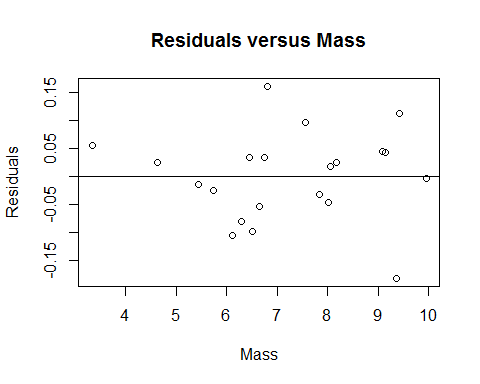
1. (3 points) 99% calibration interval for a **single** t-cell response of 0.3.

**We are 99% confident that the estimated mass that would be needed to have an individual t-cell response of 0.3 mm is between 0g and 21.9g.**

### Part J (3 points)

Provide a scatterplot of residuals. Please do this in R.

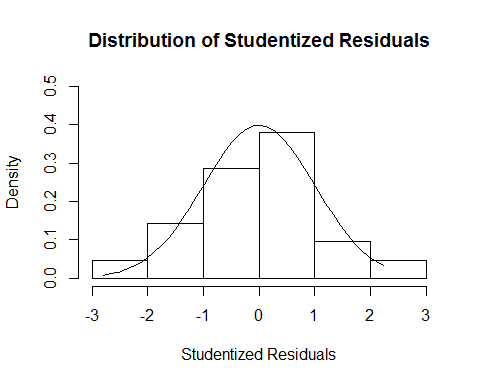
displayres <- resid(displaylm)  
plot(display$Mass, displayres, ylab="Residuals", xlab="Mass", main="Residuals versus Mass")  
abline (0,0)



### Part K (3 points)

Provide a histogram of residuals with a normal distribution superimposed. It might be helpful to use studentized residuals, rather than regular residuals, with a normal curve overlay. Use R. (You may need to research this, such as googling “histogram with normal curve in R.”)

##Store studentized residuals  
studresdisplay <- rstudent(displaylm)  
  
##Histogram  
hist(studresdisplay, freq=FALSE, main="Distribution of Studentized Residuals",   
xlab="Studentized Residuals", ylab="Density", ylim=c(0,0.5))  
  
##Create range of x-values for normal curve  
xfit <- seq(min(studresdisplay), max(studresdisplay), length=40)  
  
##Generate values from the normal distribution at the specified values  
yfit <- (dnorm(xfit))  
  
##Add the normal curve  
lines(xfit, yfit, ylim=c(0,0.5))



\*\*Additional Resource: [https://www.r-bloggers.com/visualising-residuals/\*\*](https://www.r-bloggers.com/visualising-residuals/**)

### Part L (2 points)

Provide a measure of the **proportion** of variation in the response that is accounted for by the explanatory variable. **Interpret** this measure. Use R.

cor(display$Mass, display$Tcell)^2

## [1] 0.3336345

**33.363% of the variation in Tcell response is accounted for by the mass.**

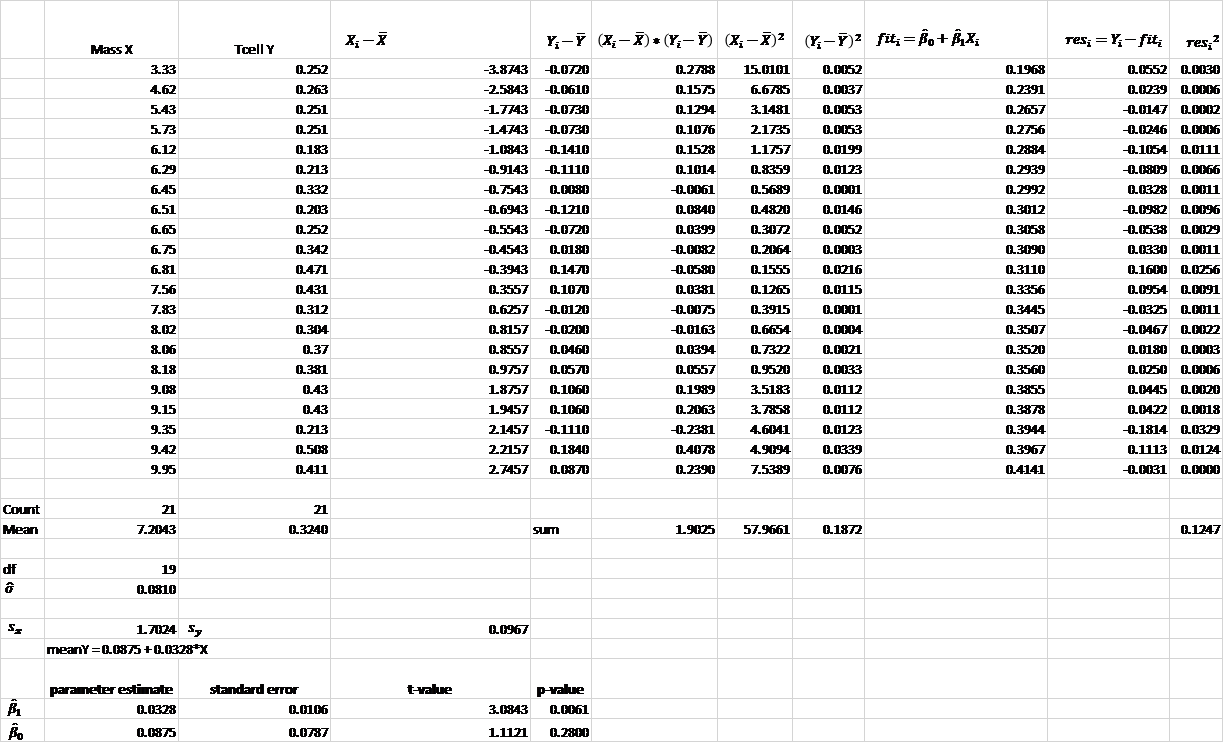
## Question 2 (40 points total)

Using the data for Black-eared Wheatears, calculate by “hand” (using Excel) the following elements. (An example of much of this was in the PowerPoints and in the videos below.)

### Part A (8 points)

and

*Note: 4 points are for and 4 points are for . You can receive partial credit if you make a mistake but have some of the process correct.*



### Part B (4 points)

The t-statistics and p-values for the hypothesis tests and .

**See table above. Give 1 point for each test statistic (t-value) and p-value.**

### Part C (7 points)

99% confidence intervals for the mean of Y when X = {3,4,5,6,7,8,9} grams. You do NOT need to make a Bonferroni (or any other type of) multiple interval correction, as the primary purpose of these intervals is to be able to plot confidence interval bands.

*Note from part 1c that the t-multiplier is 2.8609. Assign 1 point for each confidence interval.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | ) | 99% Lower Confidence Limits | 99% Upper Confidence Limits |
| 3 | 0.1860 | 0.0481 | 0.0483 | 0.3236 |
| 4 | 0.2188 | 0.0384 | 0.1089 | 0.3287 |
| 5 | 0.2516 | 0.0294 | 0.1676 | 0.3356 |
| 6 | 0.2844 | 0.0218 | 0.2220 | 0.3469 |
| 7 | 0.3172 | 0.0178 | 0.2663 | 0.3682 |
| 8 | 0.3501 | 0.0196 | 0.2940 | 0.4062 |
| 9 | 0.3829 | 0.0260 | 0.3084 | 0.4574 |

### Part D (7 points)

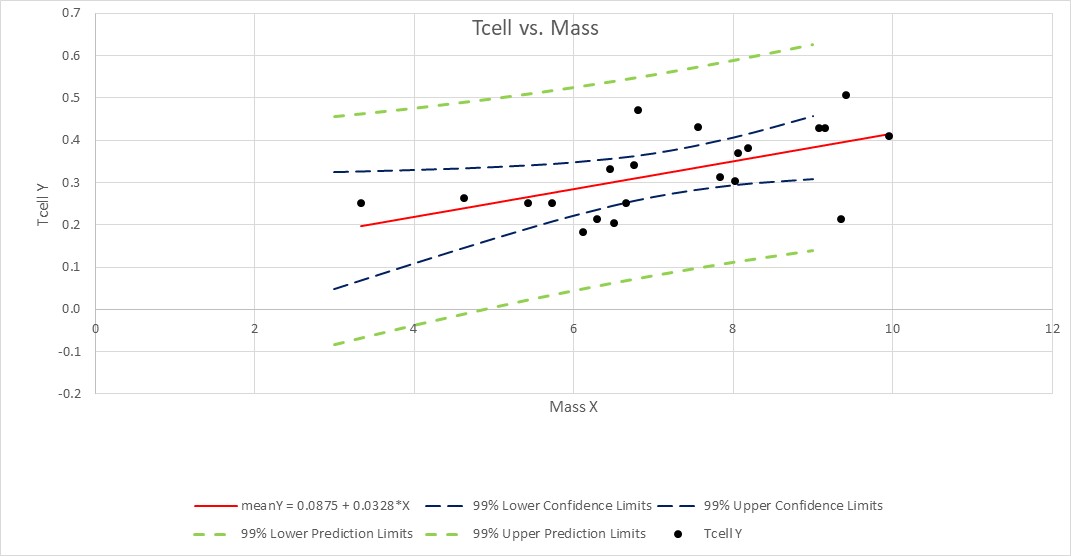
99% prediction intervals for the predicted Y when X = {3,4,5,6,7,8,9} grams. You do NOT need to make a Bonferroni (or any other type of) multiple interval correction, as the primary purpose of these intervals is to be able to plot prediction interval bands.

*Again, assign 1 point for each prediction interval.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | ) | 99% Lower Prediction Limits | 99% Upper Prediction Limits |
| 3 | 0.1860 | 0.0942 | -0.0836 | 0.4555 |
| 4 | 0.2188 | 0.0897 | -0.0377 | 0.4753 |
| 5 | 0.2516 | 0.0862 | 0.0050 | 0.4982 |
| 6 | 0.2844 | 0.0839 | 0.0444 | 0.5245 |
| 7 | 0.3172 | 0.0830 | 0.0799 | 0.5546 |
| 8 | 0.3501 | 0.0834 | 0.1116 | 0.5885 |
| 9 | 0.3829 | 0.0851 | 0.1394 | 0.6264 |

### Part E (2 points)

Provide a plot for the confidence intervals and prediction intervals using Excel. Fully label your graph. (Use the regression equation and parts (c) and (d) above to create the plot.)



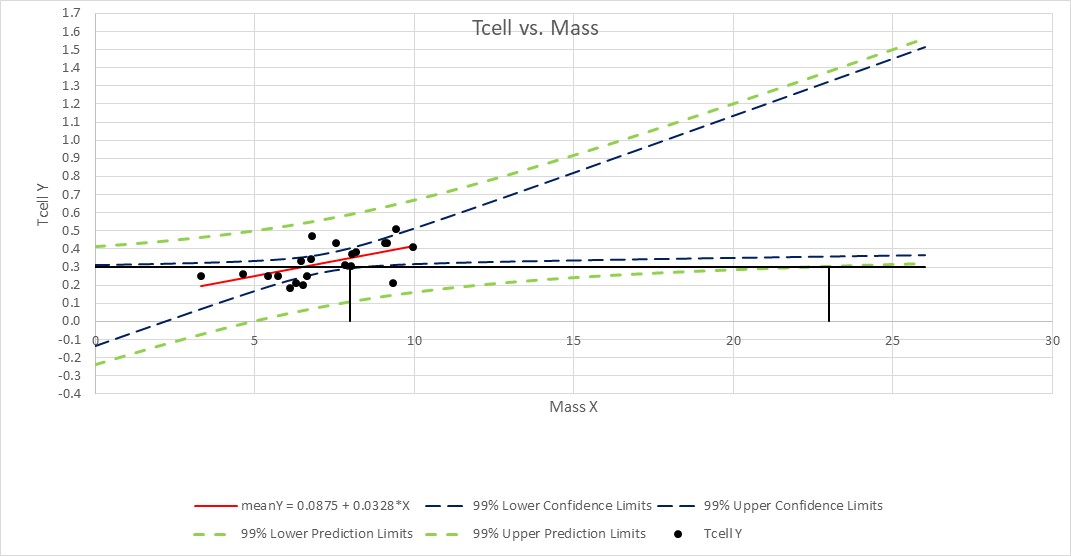
### Part F (12 points total)

1. Using the SE equations given in class and in the book (Version 3 page 194), find the following **analytically**. (Use Excel for calculations.)
2. (3 points) 99% calibration interval for the mean t-cell response of 0.3.



1. (3 points) 99% calibration interval for a single t-cell response of 0.3.



1. Using the Excel graphs, find the following. You may want to add data points to parts (c) and (d) so that the confidence and prediction limits extend well beyond the data range (although their interpretation is questionable outside the range).
2. (3 points) 99% calibration interval for the mean t-cell response of 0.3. 

**See black lines in plot. It looks like a logical interval is (0 g, 8 g). Responses will vary.**

1. (3 points) 99% calibration interval for a single t-cell response of 0.3.

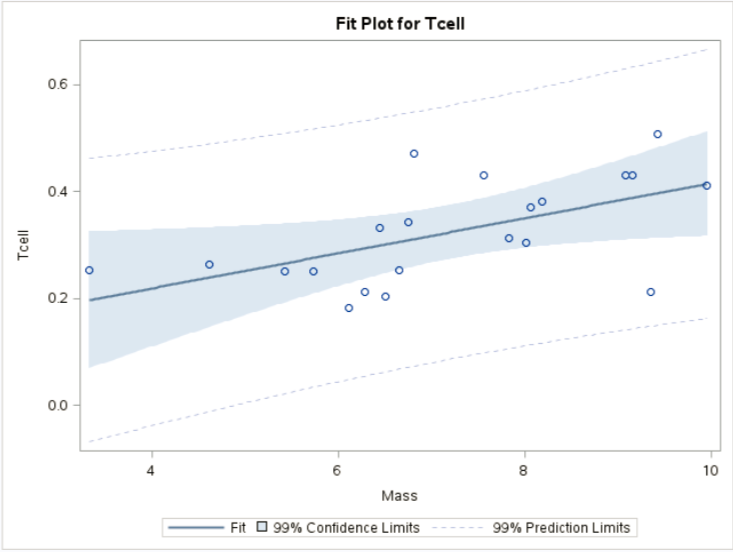
**See black lines in plot. It looks like a logical interval is (0 g, 23 g). Responses will vary.**

## Bonus (+8 points total)

### Part A (+1 point)

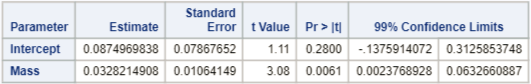
Repeat 1 (a) using SAS.

\*Q3 Parts a,b,c;  
proc glm data = displaydata plots = all alpha = 0.01;  
model Tcell=mass / solution clparm;  
run;



### Part B (+1 point)

Repeat 1 (b) using SAS.



### Part C (+1 point)

Repeat 1 (c) using SAS.

\*To get critical values for hypothesis tests;  
data mycritval;  
cv = quantile(“t”, 0.995, 21-2);  
run;

proc print data = mycritval;  
run;



*Note: the results for the test should be exactly the same, just verify the output in SAS.*

### Part D (+1 point)

Repeat 1 (g) using SAS.

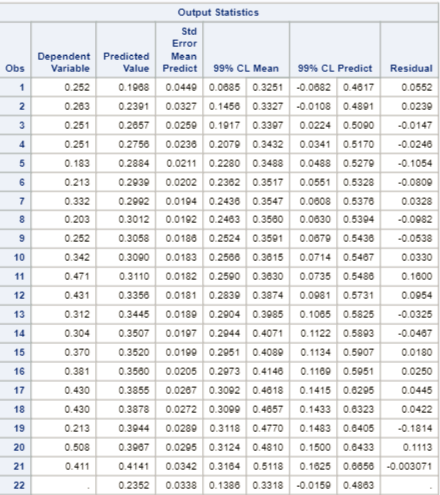
\*To create a record with mass = 4.5 but missing Tcell;  
data display2;  
input Mass Tcell;  
datalines;  
4.5 .  
;  
run;

proc print data = display2;  
run;

\*To add the dummy record to the original display set;  
data combined;  
set displaydata display2;  
run;

proc print data = combined;  
run;

\*To get confidence intervals and prediction intervals at every value of mass (especially 4.5, which is what we are looking for);  
proc reg data = combined alpha = 0.01;  
model Tcell=mass / clm cli;  
run;



**A 99% confidence interval for the mean Tcell when mass = 4.5 g is (0.1386 mm, 0.3318 mm).**

### Part E (+1 point)

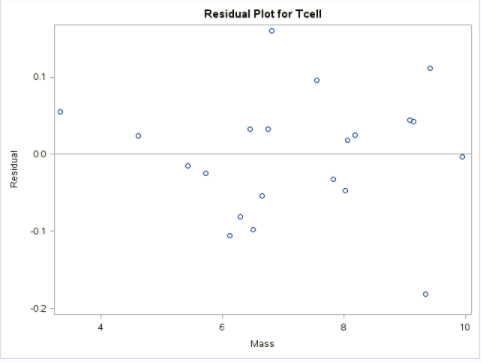
Repeat 1 (g) using SAS.

*Note: output is above.*

**A 99% prediction interval for an individual Tcell when mass = 4.5 g is (-0.0159 mm, 0.4863 mm). Because a response less than zero is nonsensical, a logical interval is (0 mm, 0,4863 mm).**

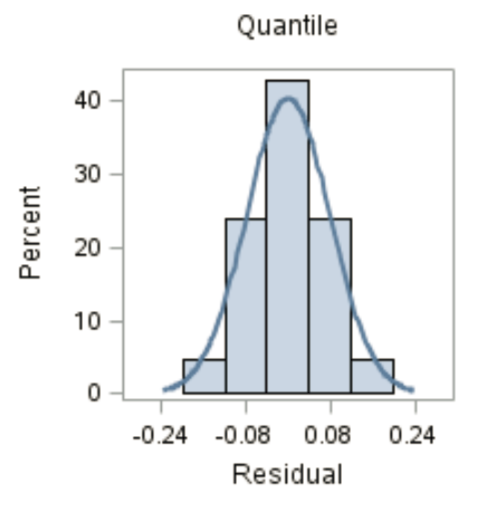
### Part F (+1 point)

Repeat 1 (j) using SAS.



### Part G (+1 point)

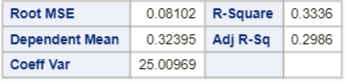
Repeat 1 (k) using SAS.



**There is no evidence against residuals being normally distributed.**

### Part H (+1 point)

Repeat 1 (l) using SAS.



**The percent of variation in Tcell that can be explained by the mass is 33.36%.**

Videos for using Excel:

and \_1: <http://screencast.com/t/ztSxTImiOk6s>

SE of and \_1 and RMSE: <http://screencast.com/t/V9gnhSwb>

Confidence Intervals: <https://www.screencast.com/t/ELiUGTe7Kc>

Prediction Intervals: <https://www.screencast.com/t/ap8WETxsGUqN>

CI and PI Plotting: <https://www.screencast.com/t/efrpHrqgYZnG>

Calibration Mean Gross: <https://www.screencast.com/t/Yu7eqiiH0X>

Calibration Single Movie: <https://www.screencast.com/t/2vS1lGqtJ>