|  |
| --- |
| title: "Statistical Foundations for Data Science - Assignment8 " |
| author: "Travis Deason" |
| date: "Oct 22nd 2017" |
| output: word\_document |

# MSDS 6371 UNIT 10 HW

* These are the same data from last week’s HW. Now, we are going to use them for simple linear regression.

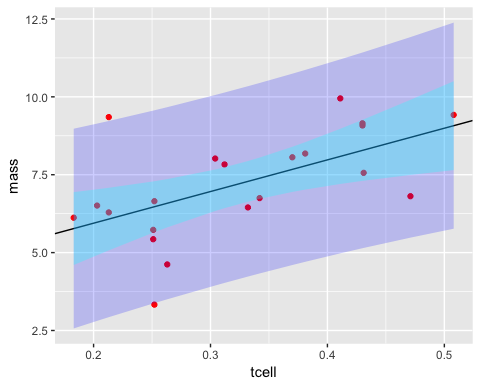
# UNIT 10: STAT 6371: Chapter 7

1. 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

#Perform Regression  
model = lm(mass~tcell, data=wheaters)  
#Extract Regression line coefficents  
wheaters$resid <- model$residuals  
slope <- model$coefficients['tcell']  
intercept <- model$coefficients[1]  
#Prediction intervals  
pred.int = predict(model, interval="prediction")

## Warning in predict.lm(model, interval = "prediction"): predictions on current data refer to \_future\_ responses

#Confidence intervals  
conf.int = predict(model, interval="confidence")  
# Add predictions and CI to original data  
wheaters$pred.lower <- pred.int[,2]  
wheaters$pred.upper <- pred.int[,3]  
wheaters$ci.upper <- conf.int[,2]  
wheaters$ci.lower <- conf.int[,3]  
## plot  
ggplot(data=wheaters, aes(x=tcell, y=mass, main='Scatterplot of data with regression line and Confidence intervals')) +   
 geom\_point(color= 'red') +  
 geom\_abline(intercept=intercept, slope=slope, color='black') +  
 geom\_ribbon(data=wheaters, aes(ymin= pred.lower, ymax= pred.upper), fill = "blue", alpha = 0.2) +  
 geom\_ribbon(data=wheaters, aes(ymin= ci.lower, ymax= ci.upper), fill = "cyan", alpha = 0.3)



summary(model)

##   
## Call:  
## lm(formula = mass ~ tcell, data = wheaters)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.1429 -0.7327 0.3448 0.7472 3.2736   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.911 1.112 3.517 0.00230 \*\*  
## tcell 10.165 3.296 3.084 0.00611 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.426 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

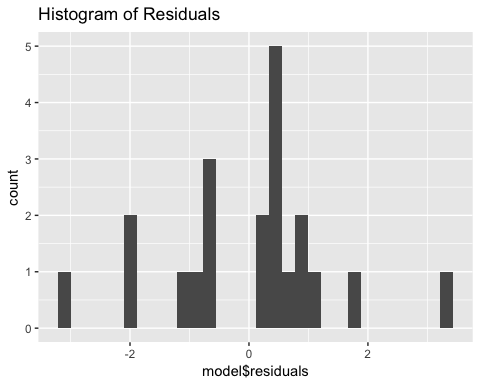
print(confint(model, level=.975))

## 1.25 % 98.75 %  
## (Intercept) 1.205079 6.617461  
## tcell 2.145068 18.185176

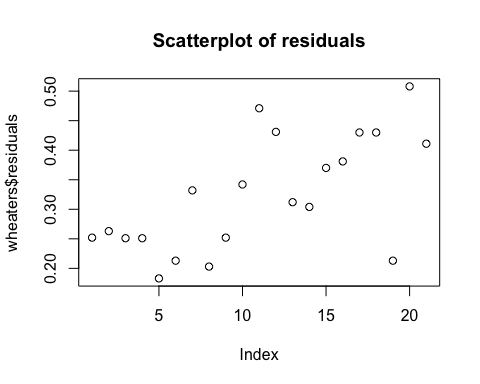
# Find critical t-value for slope  
qt(.975, 19)

## [1] 2.093024

qplot(model$residuals, geom="histogram", main='Histogram of Residuals')



plot(wheaters$tcell, wheaters$residuals, main='Scatterplot of residuals')



* In the study of tcell count verses weight transported for Black-eared wheatears. The null hypothesis is that there is no correlation between t cell counts and weights whihc can be carried by a male black-eared wheater.
* The alternative hypothesis is that there is a correlation between t-cell counts and weight carried in black-eared wheaters.
* The assumptions required to apply a t-test on our linear regression are that the residuals are normally distributed, the relationship between the variables is linear and the datapoints are independent. All of these assumptions are satisfied in this model.
* For a significence level of .95, the critical t-value is 2.09
* Using an t-test with 19 degrees of freedom on a linear regression model, the LMS slope is equal to 10.165, meaning for every 1% increase in t-cell ratio, the weight a wheater can carry in a season increased by 100 grams, and the intercept is equal to 3.911. This means a wheater who has no t-cells could theoretically carry 3.911 Kg.
* The 95% confidence interval for the slope of tcells verses carried weight is is 2.14 to 18.18. The p-value for the t-test is .00611.
* Based on the t-test, we will reject the null hypothesis that there is no coorelation between tcell count and weight carried by Black-Eared Wheaters. There is evidence to show a correlation between t-cell count and weight. Since this study was observational, we cannot assume this relationship to be causal.

## 2. Using the data in problem 1: Calculate by “hand” (using Excel) the following: (An example of this was in the PowerPoints and in the videos below.)=

### a. β ̂\_(0 ) and β ̂\_1

### b, The t-statistics and p-values for each.

### c. 99% Confidence Interval Calculations for X = {3,4,5,6,7,8,9} grams.

### d. 99% Prediction Interval Calculations for X = {3,4,5,6,7,8,9} grams.

### e. Calibration Interval Calculations for Y ̅=0.3 and for Y = 0.3 mm.

#### i. Graphical Method

#### ii, Analytical Method

* This has been worked out on the attached .xlsx document.

### f. Provide Plots for the Confidence Intervals and Prediction Intervals (From Excel … fully labeled!) … an example is shown for the Movie data in the videos below:

* Beta1 and Beta0 <http://screencast.com/t/ztSxTImiOk6s>
* SE of Beta1 and Beta 0 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>

## 3. For question 1, provide a measure of the amount of variation in the response that is accounted for by the explanatory variable. Interpret this measure clearly indicating the units of the response and the explanatory variables. Use any method (software) you choose.

* Performing a linear regression on the Black-eared wheatear dataset shows there is a linear relationship between the tcell count of the birds, and the amount of nesting rock weight in Kilograms the birds can carry in a year. The regression shows that for every 1% increase in t-cell response, the bird can carry an additional 100 grams of nesting rocks. If the bird had no t-cell response, it would be able to carry 3.9 Kg in nesting eggs. The percent of variance in nesting stone weight that is explained by the t-cell response is 33%. This means that after applying a t-cell based linear regression model, 2/3 of the original varaince in the nesting rock weight is still present in the residuals.