Project2

Joshua Wiseman

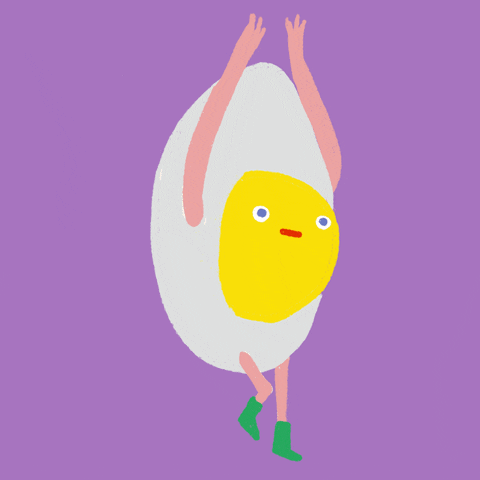
19 April, 2023

This project is all about applications of SLR to real data using R. The data used here is over the EGGS dataset and the main goal of this project is to find a predictor variable of the thickness of an egg so that a higher ratio of eggs can make it to market. What was found is that there is not enough data to confidently conclude that the trait of strength is a strong predictor of the thickness variable (due to lack of R^2 value, shape of the plots, and most importantly the p-value not rejecting the null hypothesis). However, with a different model and more data the opposite could be the case.

Table of Contents

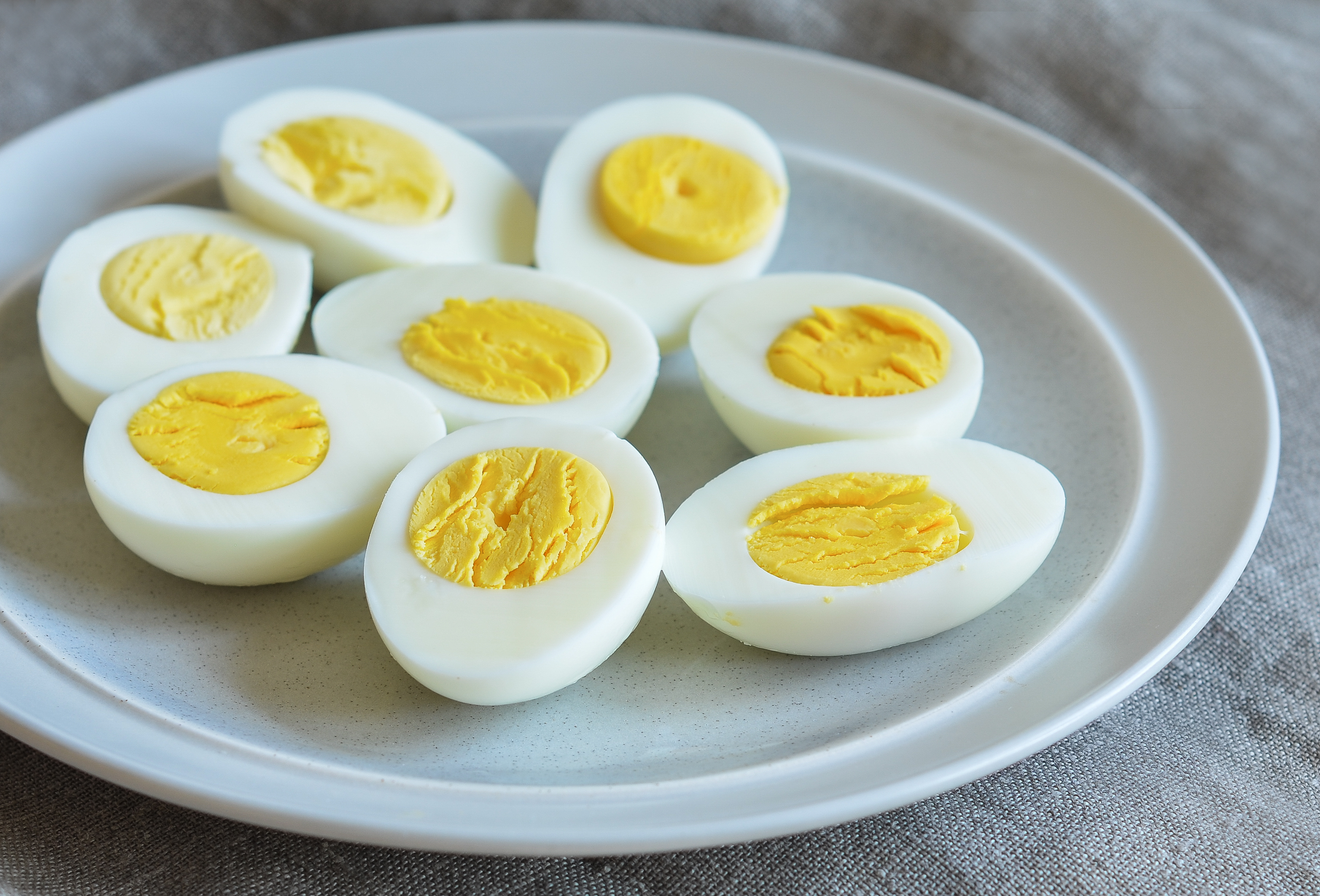
# Introduction of the Chicken Egg Problem :)

EGGCITING STUFF MAN!!!



🥚

Hello, My name is Joshua Wiseman and I’m working with the egg data set (EGGS.XLS). The main reason is eggs are a good source of protein, which helps build muscle and they also taste good. It’s a two for one deal! What’s not to like? But with this data I will try and predict what factors will allow for the most eggs to survive as possible so that we can all consume more greatness! But what factor will give us this information? Well, for this we will use the Strength Metric as the factor which we are trying to predict. Why? Well it’s because the more strength the egg has, the less likely the egg is going to break compared to normal allowing for more eggs to be produced and consumed by us people who love eggs so much!



Lovely Eggs, Amazin!

## What are the variables we need?

eggs <- read.csv("EGGS.csv")  
eggs

## HOUSING THICKNESS OVERRUN STRENGTH WTCLASS  
## 1 CAGE 0.47 495 36.9 M  
## 2 CAGE 0.43 462 39.2 M  
## 3 CAGE 0.38 488 40.2 M  
## 4 CAGE 0.47 471 33.0 M  
## 5 CAGE 0.44 471 39.0 M  
## 6 CAGE 0.40 502 36.6 L  
## 7 CAGE 0.41 472 37.5 L  
## 8 CAGE 0.45 474 38.1 L  
## 9 CAGE 0.41 492 37.8 L  
## 10 CAGE 0.37 479 34.9 L  
## 11 FREE 0.55 520 31.5 L  
## 12 FREE 0.50 531 39.7 L  
## 13 FREE 0.47 513 37.8 M  
## 14 FREE 0.48 510 33.5 M  
## 15 FREE 0.51 521 39.9 L  
## 16 FREE 0.50 510 40.6 M  
## 17 BARN 0.49 515 40.0 M  
## 18 BARN 0.50 516 37.6 M  
## 19 BARN 0.51 514 39.6 M  
## 20 BARN 0.48 526 40.3 L  
## 21 BARN 0.52 501 38.3 L  
## 22 BARN 0.50 508 40.2 L  
## 23 ORGANIC 0.43 530 34.5 L  
## 24 ORGANIC 0.49 544 36.8 L  
## 25 ORGANIC 0.52 531 32.6 L  
## 26 ORGANIC 0.49 532 38.5 M  
## 27 ORGANIC 0.52 511 40.2 M  
## 28 ORGANIC 0.44 527 33.2 M

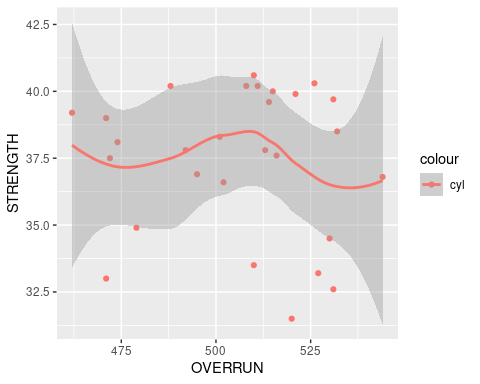
# Traits to each egg  
# eggs$STRENGTH  
# eggs$OVERRUN  
# eggs$THICKNESS  
# eggs$HOUSING  
# eggs$WTCLASS

As you can see there are multiple variables to choose to be our “x” variable or the variable to predict our “y”, aka the Strength.

### Different data plots

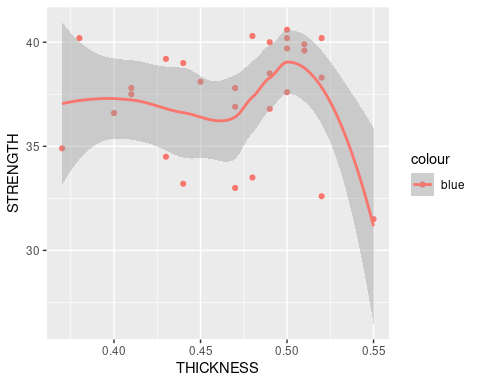
This is OVERRUN to STRENGTH correlation. As you can see there is not really a correlation here.

library(ggplot2)  
  
# OVERRUN predicting STRENGTH  
g = ggplot(eggs, aes(x = OVERRUN, y = STRENGTH, color = "cyl")) + geom\_point()  
g = g + geom\_smooth(method = "loess", formula = y ~ x)  
g



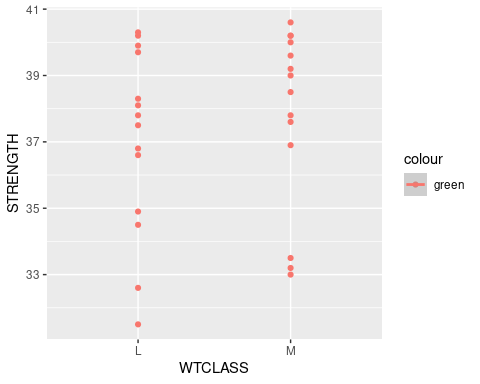
This is THICKNESS to STRENGTH correlation. As you can see there is somewhat of a correlation, but it’s not strong.

# THICKNESS predicting STRENGTH  
g = ggplot(eggs, aes(x = THICKNESS, y = STRENGTH, color = "blue")) + geom\_point()  
g = g + geom\_smooth(method = "loess", formula = y ~ x)  
g



This is WTCLASS to STRENGTH correlation. As you can see there is not really a correlation here.

# WTCLASS predicting STRENGTH  
g = ggplot(eggs, aes(x = WTCLASS, y = STRENGTH, color = "green")) + geom\_point()  
g = g + geom\_smooth(method = "loess", formula = y ~ x)  
g



This is HOUSING to STRENGTH correlation. As you can see the correlation is not that strong, however, there is still somewhat of a correlation here.

# HOUSING predicting STRENGTH  
g = ggplot(eggs, aes(x = HOUSING, y = STRENGTH, color = "red")) + geom\_point()  
g = g + geom\_smooth(method = "loess", formula = y ~ x)  
g



## How was the data collected?

Four housing systems were investigated: (1) cage, (2) barn, (3) free range, and (4) organic. Twenty-eight commercial grade A eggs were randomly selected from supermarkets — 10 of which were produced in cages, 6 in barns, 6 with free range, and 6 organic.

## What is the story behind the data?

Food Chemistry (Vol. 106, 2008) published a study of the quality of commercial eggs produced from different housing systems for chickens.

## Why was it gathered?

To find a number of quantitative characteristics were measured for each egg, including penetration strength (Newtons).

## What is my interest in the data?

My interest in the data is because eggs are cool and taste good.



## What problem do I wish to solve?

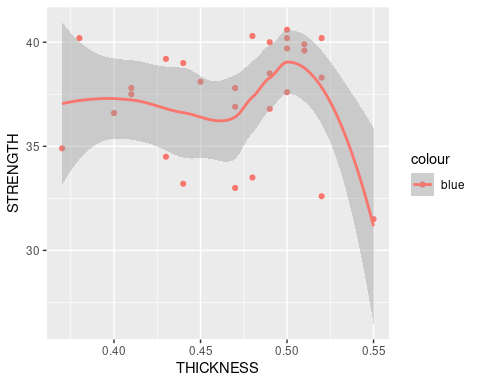
I wish to solve the problem to produce more strength in eggs so that more eggs make it to market.

# Theory needed to carry out SLR, aka Prediction

## Main result 1 (What is the x value?)

As stated previously we need our “x” value which will predict our “y” value. Our “y” value being what we want to find, Strength of the egg and the “x” value being what will be the best predictor of the “y” value. As shown in the four plots above in the plot data section. Our two best options are Housing and Thickness for the “x” value. However, since Housing is Nominal data and Thickness is Ratio data we will go ahead and choose the Thickness variable to be our “x” value.

# THICKNESS predicting STRENGTH  
g = ggplot(eggs, aes(x = THICKNESS, y = STRENGTH, color = "blue")) + geom\_point()  
g = g + geom\_smooth(method = "loess", formula = y ~ x)  
g



## Main result 2 (Models are not looking good)

Huge Problem though, the data size of 28 is not a lot of data :(. Which means that this prediction is not going to be the most reliable because trends could be found by randomness (probability 0.5, but 8 out of 10 are true when it should be 5 for example). Another official term for this is the null hypothesis and will be shown later in this document. So because of this, using a deterministic model is out of the question (calculate future event exactly, without involvement of randomness). However a questionable probabilistic model is still something which we can make from this data (uses the effect of random occurrences or actions to forecast the possibility of future results).

A Simple Linear Regression Model (SLR) is one type of probabilistic model that assumes that the data goes in a straight line. If we had data that looked like it goes in a straight line this would make life easy, but sadly this is not true.

This is mathematical notation for SLR

The main thing which you need to know is that is the random error & is the mean value of y given x.

Then there is a Quadratic Linear Regression Model which is a type of probabilistic model that assumes that the data goes in a quadratic type curve. Again, this is not the case for our egg data :(.

This is mathematical notation for Quadratic Linear Regression Model

The main thing which you need to know is that is the random error & is the mean value of y given x. Same as SLR, but just, you know, quadratic.

## Checks on validity

library(s20x)  
eggs.lm=lm(STRENGTH~THICKNESS,data=eggs)  
summary(eggs.lm)

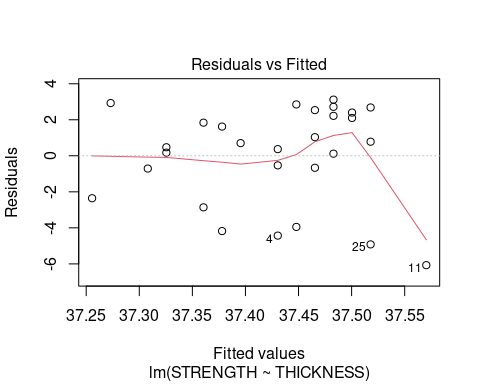
##   
## Call:  
## lm(formula = STRENGTH ~ THICKNESS, data = eggs)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.0704 -1.1199 0.5895 2.2627 3.1171   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 36.608 5.438 6.732 3.83e-07 \*\*\*  
## THICKNESS 1.749 11.543 0.152 0.881   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.772 on 26 degrees of freedom  
## Multiple R-squared: 0.0008822, Adjusted R-squared: -0.03755   
## F-statistic: 0.02296 on 1 and 26 DF, p-value: 0.8807

eggs.qdlm=lm(STRENGTH~THICKNESS + I(THICKNESS ^ 2),data=eggs)  
summary(eggs.qdlm)

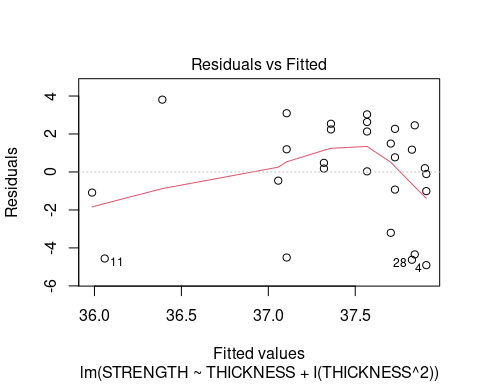
##   
## Call:  
## lm(formula = STRENGTH ~ THICKNESS + I(THICKNESS^2), data = eggs)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.9088 -1.0282 0.3397 2.2476 3.8097   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -12.04 49.85 -0.241 0.811  
## THICKNESS 216.84 219.42 0.988 0.333  
## I(THICKNESS^2) -235.25 239.65 -0.982 0.336  
##   
## Residual standard error: 2.774 on 25 degrees of freedom  
## Multiple R-squared: 0.03796, Adjusted R-squared: -0.039   
## F-statistic: 0.4933 on 2 and 25 DF, p-value: 0.6164

P-value for the qdlm is smaller, therefore better, but still a long shot away to invalidating the null hypothesis (random chance). Essentially, both still suck :( and are so high in fact that the null hypothesis cannot be rejected (chance of randomness).

plot(eggs.lm, which = 1)



plot(eggs.qdlm, which = 1)



Residuals (red line) are not close to the dotted line (closeness to mean) which means the data is worse.

ciReg(eggs.lm)

## 95 % C.I.lower 95 % C.I.upper  
## (Intercept) 25.43065 47.78625  
## THICKNESS -21.97711 25.47497

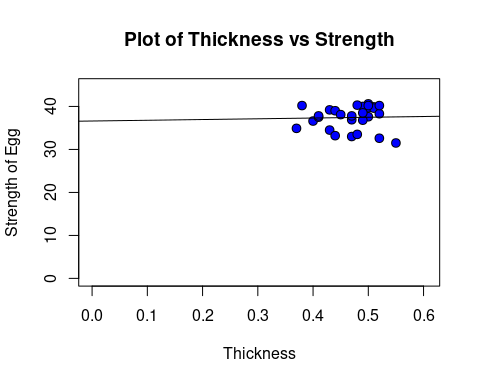
ciReg(eggs.qdlm)

## 95 % C.I.lower 95 % C.I.upper  
## (Intercept) -114.7114 90.6377  
## THICKNESS -235.0575 668.7300  
## I(THICKNESS^2) -728.8292 258.3232

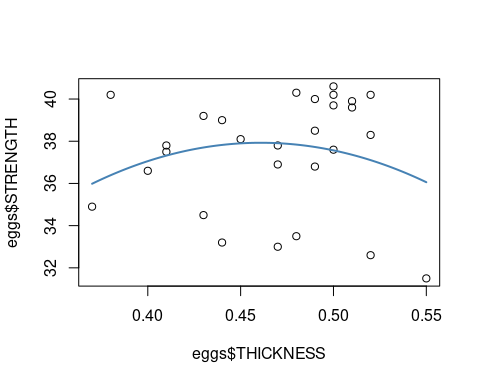
This range is where 95% of the data shows up.

### Straight trend line for eggs data

layout(1)  
plot(eggs$STRENGTH~eggs$THICKNESS, bg="Blue",pch=21, xlim = c(0, 1.1\*max(eggs$THICKNESS)), xlab = "Thickness", ylim = c(0, 1.1\*max(eggs$STRENGTH)), ylab = "Strength of Egg", cex = 1.2, main = "Plot of Thickness vs Strength", data=eggs)  
abline(eggs.lm)

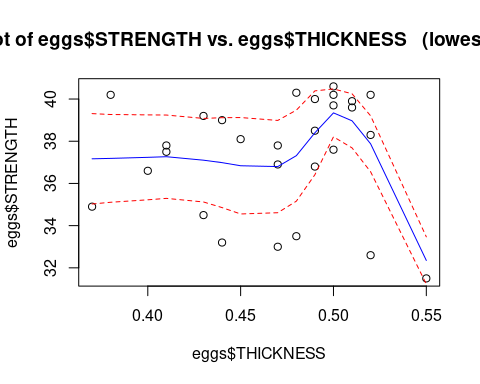


plot(eggs$STRENGTH~eggs$THICKNESS)  
myplot = function(x){  
 eggs.qdlm$coef[1] + eggs.qdlm$coef[2] \* x + eggs.qdlm$coef[3] \* x ^ 2  
}  
  
curve(myplot, lwd = 2, col = "steelblue", add = TRUE)



#### Trendscatter of eggs data

trendscatter(eggs$STRENGTH~eggs$THICKNESS, data = eggs, f = 0.5)

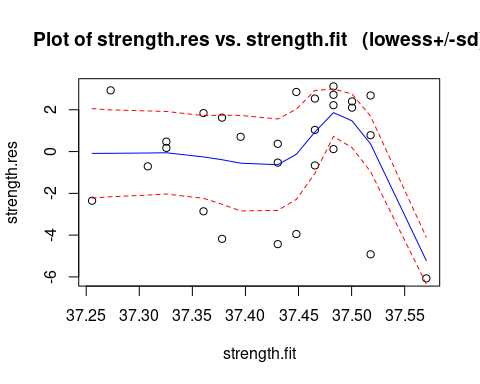


#### Residual vs fitted values for eggs data

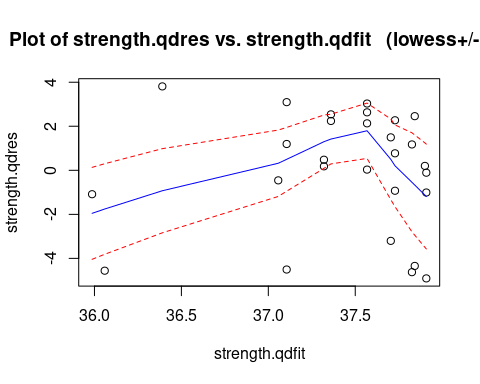
strength.res = residuals(eggs.lm)  
strength.fit = fitted(eggs.lm)  
  
strength.qdres = residuals(eggs.qdlm)  
strength.qdfit = fitted(eggs.qdlm)

#### Trendscatter on Residual Vs Fitted for eggs data

trendscatter(strength.res~strength.fit)



trendscatter(strength.qdres~strength.qdfit)



The trend must go back to the middle dotted line and stay their to be reliable, this is not the case as you can see here.

### Independence of data

t.test(eggs$STRENGTH,eggs$THICKNESS)

##   
## Welch Two Sample t-test  
##   
## data: eggs$STRENGTH and eggs$THICKNESS  
## t = 71.859, df = 27.016, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 35.90434 38.01495  
## sample estimates:  
## mean of x mean of y   
## 37.4285714 0.4689286

They come from the same egg so no, there is no independence.

# Which model is better?

As seen previously, these models are really not great, but based on the p variable and closeness of the data points to the mean, the quadratic model will be best. (Faraway 2016)

## Adjusted Check

# Linear Model  
RSS1 = with(eggs, sum((THICKNESS-strength.fit)^2))  
MSS1 = with(eggs, sum((strength.fit-mean(THICKNESS))^2))  
TSS1 = with(eggs, sum((THICKNESS-mean(THICKNESS))^2))  
R1 = MSS1/TSS1  
R1

## [1] 663256.9

# Quadratic Model  
RSS2 = with(eggs, sum((THICKNESS-strength.qdfit)^2))  
MSS2 = with(eggs, sum((strength.qdfit-mean(THICKNESS))^2))  
TSS2 = with(eggs, sum((THICKNESS-mean(THICKNESS))^2))  
R2 = MSS2/TSS2  
R2

## [1] 663385.4

Here is further proof. The Higher the R^2 value, the more reliable the model is and the quadratic model is a little bit higher than the linear model, honestly the difference is negligible because they are both not really reliable, but it is still there. (Faraway 2016)

## Summary lmobject

summary(eggs.lm)

##   
## Call:  
## lm(formula = STRENGTH ~ THICKNESS, data = eggs)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.0704 -1.1199 0.5895 2.2627 3.1171   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 36.608 5.438 6.732 3.83e-07 \*\*\*  
## THICKNESS 1.749 11.543 0.152 0.881   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.772 on 26 degrees of freedom  
## Multiple R-squared: 0.0008822, Adjusted R-squared: -0.03755   
## F-statistic: 0.02296 on 1 and 26 DF, p-value: 0.8807

summary(eggs.qdlm)

##   
## Call:  
## lm(formula = STRENGTH ~ THICKNESS + I(THICKNESS^2), data = eggs)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.9088 -1.0282 0.3397 2.2476 3.8097   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -12.04 49.85 -0.241 0.811  
## THICKNESS 216.84 219.42 0.988 0.333  
## I(THICKNESS^2) -235.25 239.65 -0.982 0.336  
##   
## Residual standard error: 2.774 on 25 degrees of freedom  
## Multiple R-squared: 0.03796, Adjusted R-squared: -0.039   
## F-statistic: 0.4933 on 2 and 25 DF, p-value: 0.6164

# Conclusion



Burger Chicken?!

## Answer can we get BETTER EGGS?!

With the data given from this data set, sadly I must conclude that this data is not reliable enough to make such a conclusion. There is not enough data given and the data given looks pretty random for the most part. Could you condlude that there is some correlation? Sure, but that corralation could litteraly also be chance. However I do have hope that we can get find some corralation with more data to result in more eggs with higher thinkness so that more eggs make it to market.

## Ways to improve the model or experiment

I think a concave up quadratic with removing outliers would make the model more reliable. However, as previously stated, there should be more data in this data set because otherwise the trend is not as obvious and it’s 100% not obvious here.

# References

Faraway, Julian James. 2016. *Extending the Linear Model with r : Generalized Linear, Mixed Effects and Nonparametric Regression Models / Julian j. Faraway, University of Bath, UK.* Second edition.. Texts in Statistical Science.