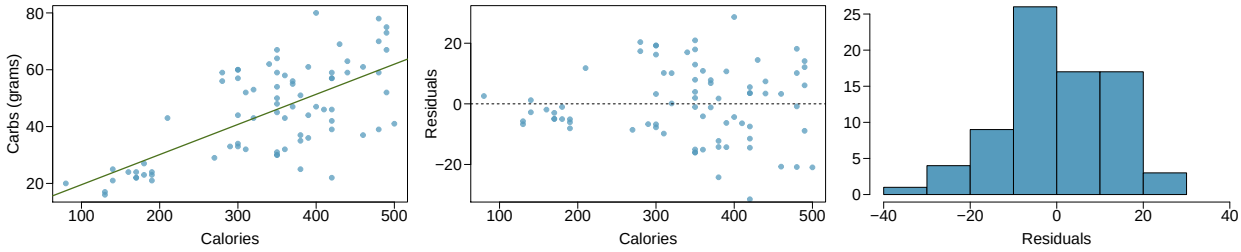
Chapter 8 - Introduction to Linear Regression

Nutrition at Starbucks, Part I. (8.22, p. 326) The scatterplot below shows the relationship between the number of calories and amount of carbohydrates (in grams) Starbucks food menu items contain. Since Starbucks only lists the number of calories on the display items, we are interested in predicting the amount of carbs a menu item has based on its calorie content.

```
##
## Call:
## lm(formula = carb ~ calories, data = starbucks)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.477  -7.476  -1.029  10.127  28.644
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.94356    4.74600   1.884   0.0634 .
## calories     0.10603    0.01338   7.923 1.67e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.29 on 75 degrees of freedom
## Multiple R-squared:  0.4556, Adjusted R-squared:  0.4484
## F-statistic: 62.77 on 1 and 75 DF,  p-value: 1.673e-11
```



(a) Describe the relationship between number of calories and amount of carbohydrates (in grams) that Starbucks food menu items contain.

A gram of carbohydrates usually contains about 4 calories but not all calories contain carbs. So the relationship indicates the percentage of calories that come from carbs.

(b) In this scenario, what are the explanatory and response variables?

To me, the true predictor is the ingredient. But in our model, we have calories as the x value or explanatory value.

(c) Why might we want to fit a regression line to these data?

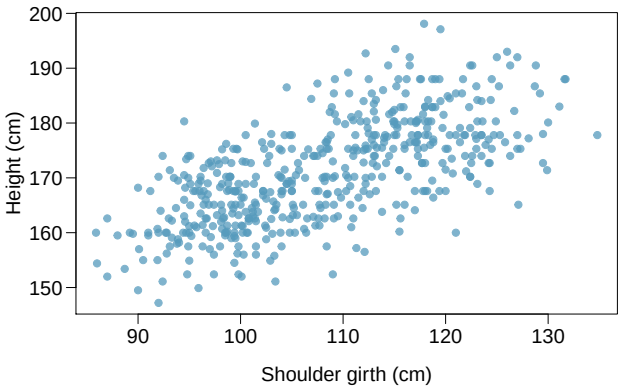
If starbucks were to offer a new item, and we knew there were x calories, then we can predict y carbs.

(d) Do these data meet the conditions required for fitting a least squares line?

The conditions would be 1) Linearity, 2) Residuals distributed normally, 3) Fairly Constant Variability, 4) Independence.

The conditions are almost met. The variability seems to increase quite a bit over 300 calories.

Body measurements, Part I. (8.13, p. 316) Researchers studying anthropometry collected body girth measurements and skeletal diameter measurements, as well as age, weight, height and gender for 507 physically active individuals. The scatterplot below shows the relationship between height and shoulder girth (over deltoid muscles), both measured in centimeters.



(a) Describe the relationship between shoulder girth and height.

```
R<-cor(bdims$hgt , bdims$sho_gi)

sprintf("The correlation is %.2f. We say that %.2f is the variablity of height explained by girth",
        R, R^2)
```

```
## [1] "The correlation is 0.67. We say that 0.44 is the variablity of height explained by girth"
```

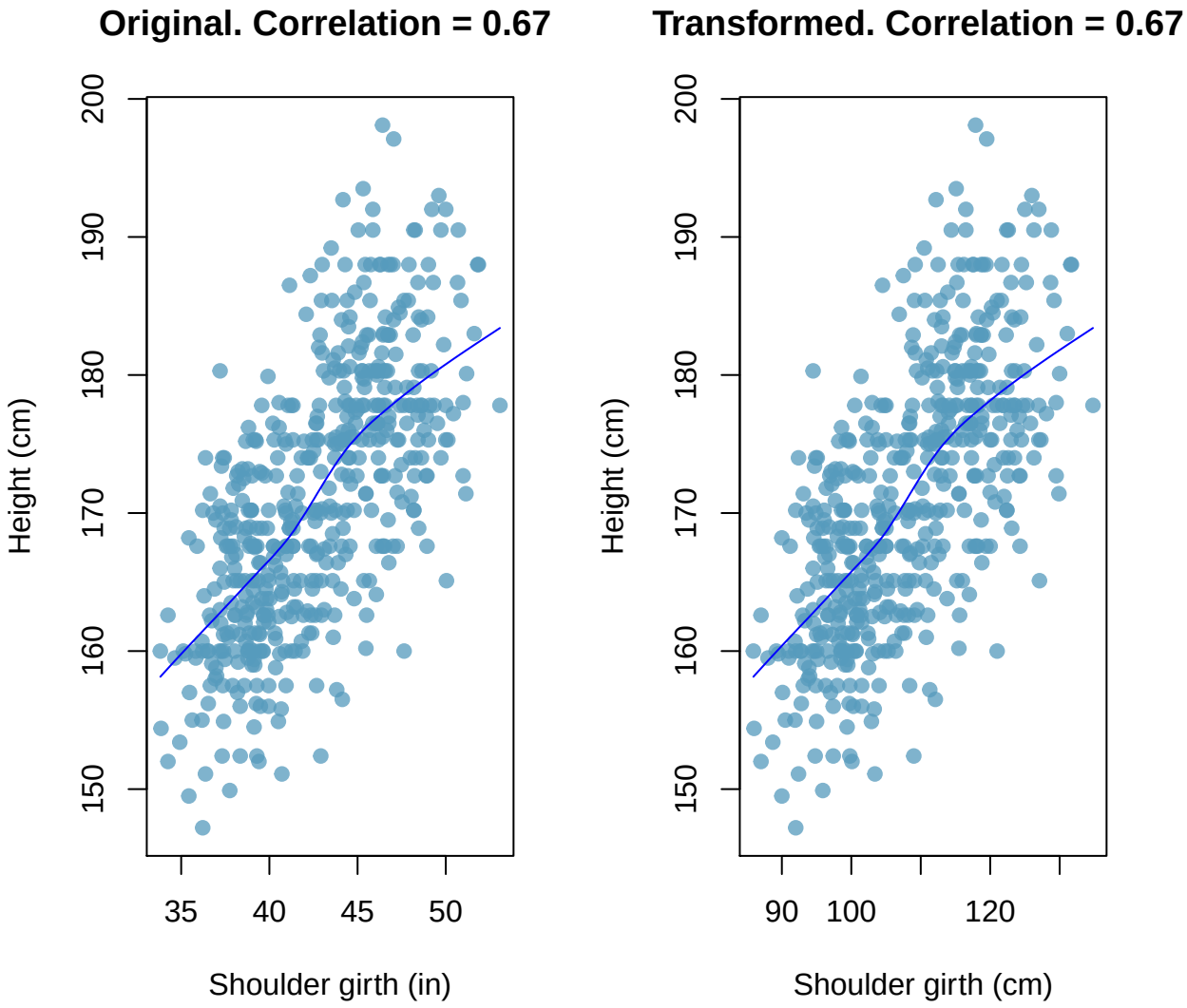
As girth increases, height increases althought the relationship is not completely linear.

(b) How would the relationship change if shoulder girth was measured in inches while the units of height remained in centimeters?

Ill plot them side by side.

```
bdims <- bdims %>%
  mutate(sho_gi_in = sho_gi * 0.3937)

par(mfrow=c(1,2))

R_in<-cor(bdims$hgt,bdims$sho_gi_in)

title<-sprintf("Original. Correlation = %.2f ", R)
title_in<-sprintf("Transformed. Correlation = %.2f ", R_in)

plot(bdims$hgt ~ bdims$sho_gi_in,
     xlab = "Shoulder girth (in)", ylab = "Height (cm)",
     pch = 19, col = COL[1,2], main=title)
abline(bdims$hgt ~ bdims$sho_gi_in, col = "red", lwd = 2, las=1)
lines(lowess(bdims$sho_gi_in ,bdims$hgt), col="blue")
```

```
plot(bdims$hgt ~ bdims$sho_gi,
     xlab = "Shoulder girth (cm)", ylab = "Height (cm)",
     pch = 19, col = COL[1,2], main=title_in)
abline(bdims$hgt ~ bdims$sho_gi, col = "red", lwd = 2, las=1)
lines(lowess(bdims$sho_gi ,bdims$hgt), col="blue")
```



This is an example of a linear transformation. The distribution is not effected.

Body measurements, Part III. (8.24, p. 326) Exercise above introduces data on shoulder girth and height of a group of individuals. The mean shoulder girth is 107.20 cm with a standard deviation of 10.37 cm. The mean height is 171.14 cm with a standard deviation of 9.41 cm. The correlation between height and shoulder girth is 0.67.

(a) Write the equation of the regression line for predicting height.

```
bdims_lm<-lm(hgt ~ sho_gi, data=bdims)
bdims_lm_sum<-summary(bdims_lm)


mean_sho_gi<-mean(bdims$sho_gi)
mean_height<-mean(bdims$hgt)
sd_sho_gi<-sd(bdims$sho_gi)
sd_height<-sd(bdims$hgt)

m <- R * (sd_height / sd_sho_gi)           # m is correlation times Xsd/Ysd

b <- mean_height - (m * mean_sho_gi)

sprintf("The slope is %.2f and the intercept is %.2f ", m,b)
```

```
## [1] "The slope is 0.60 and the intercept is 105.83 "
```

```
s1_b<-bdims_lm_sum$coefficients[1,1]
s1_m<-bdims_lm_sum$coefficients[2,1]


sprintf("The slope is %.2f and the intercept is %.2f", s1_m,s1_b)
```

```
## [1] "The slope is 0.60 and the intercept is 105.83"
```

```
# View(cbind(bdims$hgt,bdims$sho_gi))
```

$$height = 0.60 girth + 105.832$$

(b) Interpret the slope and the intercept in this context.

Its useful to calculate the slope as a product of the R and the standard deviations of x and y.

The intercept simply adjusts the line up and down the y axis. It shouldnt effect the correlation of x and y.

(c) Calculate R^2 of the regression line for predicting height from shoulder girth, and interpret it in the context of the application.

```
R<-cor(bdims$hgt , bdims$sho_gi)
R2<-R^2

sprintf("Using the cor function, R Squared = %.4f", R2)
```

```
## [1] "Using the cor function, R Squared = 0.4432"
```

```r
R2<-bdims_lm_sum$r.squared

sprintf("Extracting it from the linear model, R Squared = %.4f", R2)
```

```
## [1] "Extracting it from the linear model, R Squared = 0.4432"
```

```r
SST <- var( bdims$hgt ) * (nrow(bdims)-1)
SSE<- sum( bdims_lm$residuals ^2 )
R2<-(1-SSE/SST)


sprintf("Calculating it manually from the Sum of Squares, R Squared = %.4f", R2)
```

```
## [1] "Calculating it manually from the Sum of Squares, R Squared = 0.4432"
```

R Squared is a metric of variability of y from a known x. Its agnostic of direction, i.e. it doesnt matter if y is greater or lesser than x.

(d) A randomly selected student from your class has a shoulder girth of 100 cm. Predict the height of this student using the model.

```r
predicted_hgt<-m*100 + b

sprintf("Plug 100 into our slope and intercept returns %.2f", predicted_hgt)
```

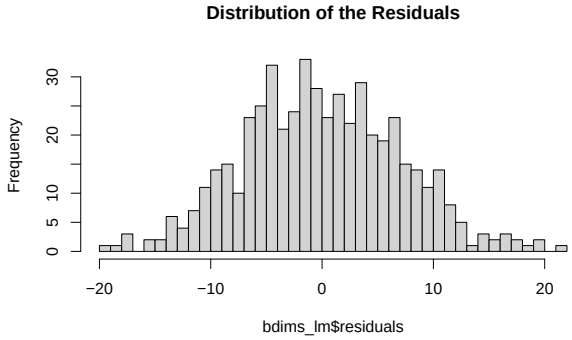
```
## [1] "Plug 100 into our slope and intercept returns 166.20"
```

```r
new_sho_gi<-data.frame(sho_gi=c(100))

predicted_hgt<-predict(bdims_lm, newdata = new_sho_gi , interval = "prediction")[1]

sprintf("The predict function returns %.2f", predicted_hgt)
```

```
## [1] "The predict function returns 166.20"
```

(e) The student from part (d) is 160 cm tall. Calculate the residual, and explain what this residual means.

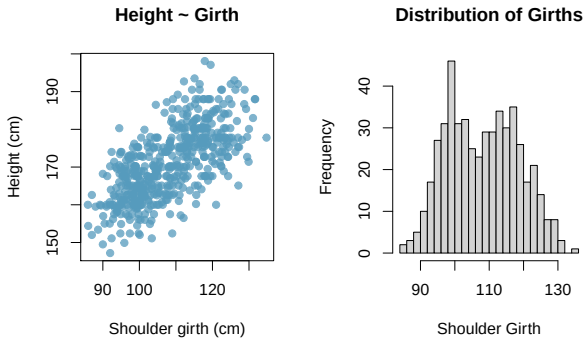
```r
# something not right with this
# sprintf("The average residual is %.24f",mean(bdims_lm_sum$residuals))

hist( bdims_lm$residuals, breaks=30, main = "Distribution of the Residuals")

# lm_xy<-data.frame(y = bdims_lm$residuals + bdims_lm$fitted.values,
#   residuals = bdims_lm$residuals, fitted = bdims_lm$fitted.values )
# bdims_xy<-data.frame(y = bdims$hgt, x = bdims$sho_gi )
# merged_xy<-merge(lm_xy, bdims_xy, by="y")
```



The residual is -6.2. Reviewing the other residuals I would such a variation is not unusual.

(f) A one year old has a shoulder girth of 56 cm. Would it be appropriate to use this linear model to predict the height of this child?

```
par(mfrow=c(1,2))


plot(bdims$hgt ~ bdims$sho_gi,
     xlab = "Shoulder girth (cm)", ylab = "Height (cm)",
     pch = 19, col = COL[1,2], main="Height ~ Girth")
abline(bdims$hgt ~ bdims$sho_gi, col = "red", lwd = 2, las=1)


hist( bdims$sho_gi, breaks=30, main = "Distribution of Girths", xlab = "Shoulder Girth")
```

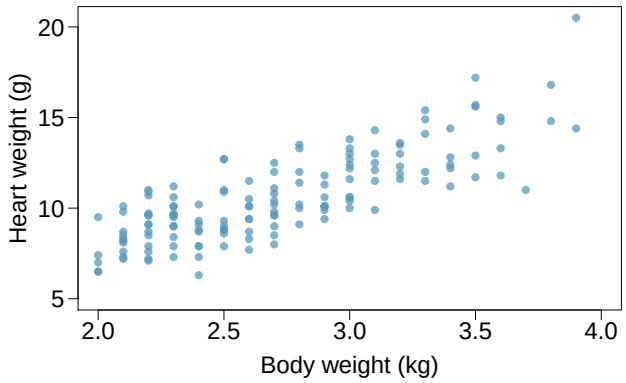


Looking at the distributions, I would say constant variability is good, normality is good, and linearity is good.

Cats, Part I. (8.26, p. 327) The following regression output is for predicting the heart weight (in g) of cats from their body weight (in kg). The coefficients are estimated using a dataset of 144 domestic cats.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.357	0.692	-0.515	0.607
body wt	4.034	0.250	16.119	0.000

$s = 1.452$ $R^2 = 64.66\%$ $R^2_{adj} = 64.41\%$



(a) Write out the linear model.

```
sum_m_cats_hwt_bwt<-summary(m_cats_hwt_bwt)

cats_b<-sum_m_cats_hwt_bwt$coefficients[1,1]
cats_m<-sum_m_cats_hwt_bwt$coefficients[2,1]


sprintf("The slope is %.2f and the intercept is %.2f", cats_m,cats_b)
```

```
## [1] "The slope is 4.03 and the intercept is -0.36"
```

(b) Interpret the intercept.

The intercept reduces the heart weight by .36 after the slope product is applied.

Its not really an intercept in this case since a cat cant weigh 0 kg's

(c) Interpret the slope.

For each 1 kilogram increase in the weight of a cat, the heart weight increases by 4 grams

(d) Interpret R^2.

```
sprintf("RSquared is %.4f", sum_m_cats_hwt_bwt$r.squared)
```

```
## [1] "RSquared is 0.6466"
```

The relationship between a cats overall weight and heart weight is pretty reliable.

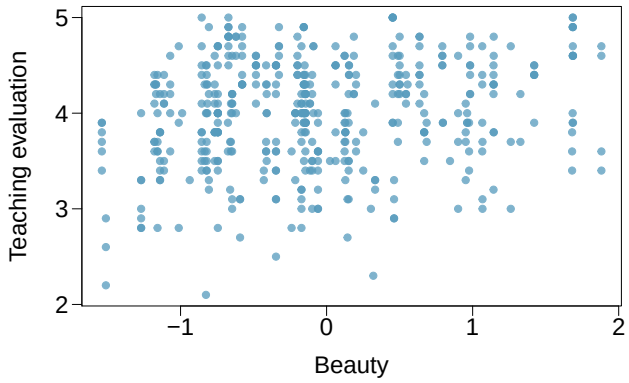
(e) Calculate the correlation coefficient.

```r
R<-cor(cats$Bwt, cats$Hwt)

sprintf("The correlation coefficient is %.2f", R)
```

```
## [1] "The correlation coefficient is 0.80"
```

Rate my professor. (8.44, p. 340) Many college courses conclude by giving students the opportunity to evaluate the course and the instructor anonymously. However, the use of these student evaluations as an indicator of course quality and teaching effectiveness is often criticized because these measures may reflect the influence of non-teaching related characteristics, such as the physical appearance of the instructor. Researchers at University of Texas, Austin collected data on teaching evaluation score (higher score means better) and standardized beauty score (a score of 0 means average, negative score means below average, and a positive score means above average) for a sample of 463 professors. The scatterplot below shows the relationship between these variables, and also provided is a regression output for predicting teaching evaluation score from beauty score.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.010	0.0255	157.21	0.0000
beauty	-45.2808	0.0322	4.13	0.0000



(a) Given that the average standardized beauty score is -0.0883 and average teaching evaluation score is 3.9983, calculate the slope. Alternatively, the slope may be computed using just the information provided in the model summary table.

```
b<-4.01002
mean_score<-3.9983
mean_beauty<--0.0883


# mean_score = mean_beauty * m + b
m<-(mean_score-b)/mean_beauty

sprintf("As a product of the 2 means, slope is %.2f", m)
```

```
## [1] "As a product of the 2 means, slope is 0.13"
```

```
R<-cor(eval, beauty)
m <- R * (sd(eval)/sd(beauty))           # m is correlation times Xsd/Ysd

sprintf("As a product of the correlation and standard deviations, its %.2f", m)
```

```
## [1] "As a product of the correlation and standard deviations, its 0.13"
```

```
summary(m_eval_beauty)
```

```
##
## Call:
## lm(formula = eval ~ beauty)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.80015 -0.36304  0.07254  0.40207  1.10373
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.01002    0.02551 157.205  < 2e-16 ***
## beauty       0.13300    0.03218   4.133 4.25e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5455 on 461 degrees of freedom
## Multiple R-squared:  0.03574,    Adjusted R-squared:  0.03364
## F-statistic: 17.08 on 1 and 461 DF,  p-value: 4.247e-05
```

(b) Do these data provide convincing evidence that the slope of the relationship between teaching evaluation and beauty is positive? Explain your reasoning.

The correlation is closer to 0 than it is to 1. I dont think there is much relationship.

(c) List the conditions required for linear regression and check if each one is satisfied for this model based on the following diagnostic plots.

(d) Linearity

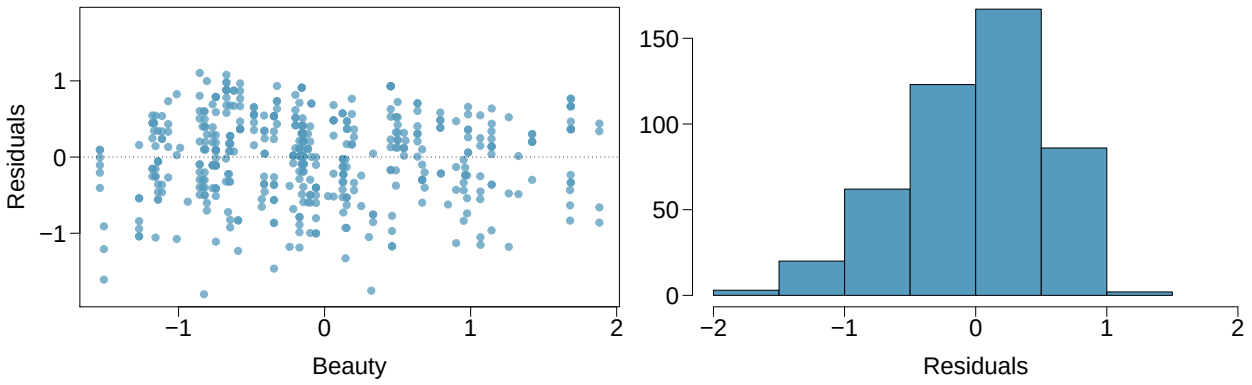
(e) Constant variablity. The *range* of y values for any given x should be consistent.

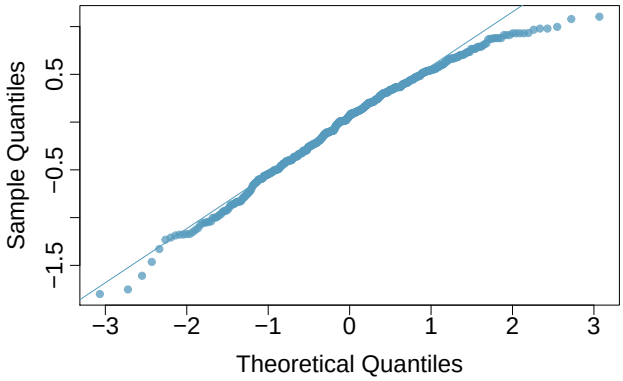
(f) The residuals should have a distribution that is fairly normal.

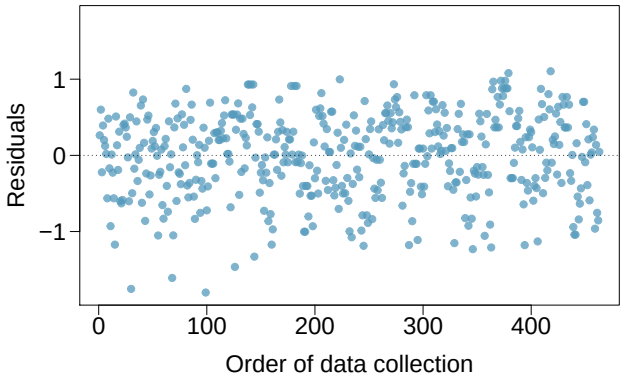
(g) Independence.

I have trouble saying there is linearity. There is some positive relationship but it looks insignificant.

The other 3 requirements look good.




Sample Quantiles
0.5
−0.5
−1.5
−3
−2
−1
0
1
2
3
Theoretical Quantiles


Residuals
1
0
−1
0
100
200
300
400
Order of data collection