MichaelY - HW7 - Introduction to linear regression

Michael Y.

April 14th, 2019

```
###setwd("c:/Users/Michael/DROPBOX/priv/CUNY/MSDS/201902-Spring/DATA606-Jason/Homework")
```

Homework - Chapter 7 - Introduction to Linear Regression (pp.331-355)

Practice : 7.23, 7.25, 7.29, 7.39 (pp.356-371)

Datasets:

7.23 - tourism

7.25 - coast_starlight

7.29 - murders

7.39 - urban_owner

Exercises: 7.24, 7.26, 7.30, 7.40 (pp.356-371)

Datasets:

7.24 - starbucks

7.26 - bdims

7.30 - cats

7.40 - prof_evals –NB: the actual name of this data set is "prof.evaltns.beauty.public"

#.#

Exercise 7.24 Nutrition at Starbucks, Part I.

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.

```
# look at the exact data set referenced
data("starbucks")
summary(starbucks)
```

```
##      item              calories          fat              carb           fiber            protein
##  Length:77          Min.   : 80.00   Min.   : 0.000   Min.   :16.00   Min.   :0.0000   Min.   : 0.000
##  Class :character   1st Qu.:300.00   1st Qu.: 9.000   1st Qu.:31.00   1st Qu.:0.0000   1st Qu.: 5.000
##  Mode  :character   Median :350.00   Median :13.000   Median :45.00   Median :2.0000   Median : 7.000
##                     Mean   :338.83   Mean   :13.766   Mean   :44.87   Mean   :2.2208   Mean   : 9.480
##                     3rd Qu.:420.00   3rd Qu.:18.000   3rd Qu.:59.00   3rd Qu.:4.0000   3rd Qu.:15.000
##                     Max.   :500.00   Max.   :28.000   Max.   :80.00   Max.   :7.0000   Max.   :34.000
##
```

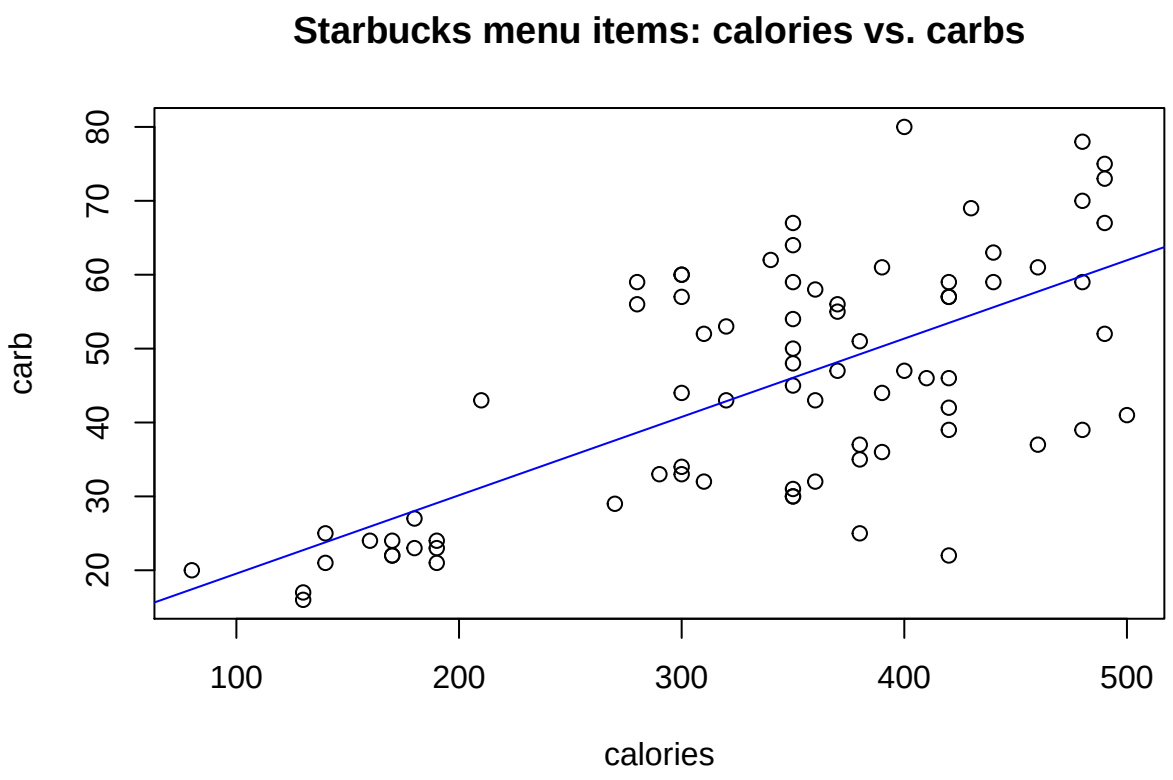
```
sbux_model <- lm(carb ~ calories, data=starbucks)
summary(sbux_model)
```

```
##
## Call:
## lm(formula = carb ~ calories, data = starbucks)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -31.4765  -7.4765  -1.0291  10.1266  28.6441
##
## Coefficients:
##             Estimate Std. Error t value         Pr(>|t|)
## (Intercept) 8.943560   4.746003  1.8844          0.06338 .
## calories    0.106031   0.013383  7.9229 0.00000000001673 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.293 on 75 degrees of freedom
## Multiple R-squared:  0.45562,    Adjusted R-squared:  0.44837
## F-statistic: 62.772 on 1 and 75 DF,  p-value: 0.000000000016725
```

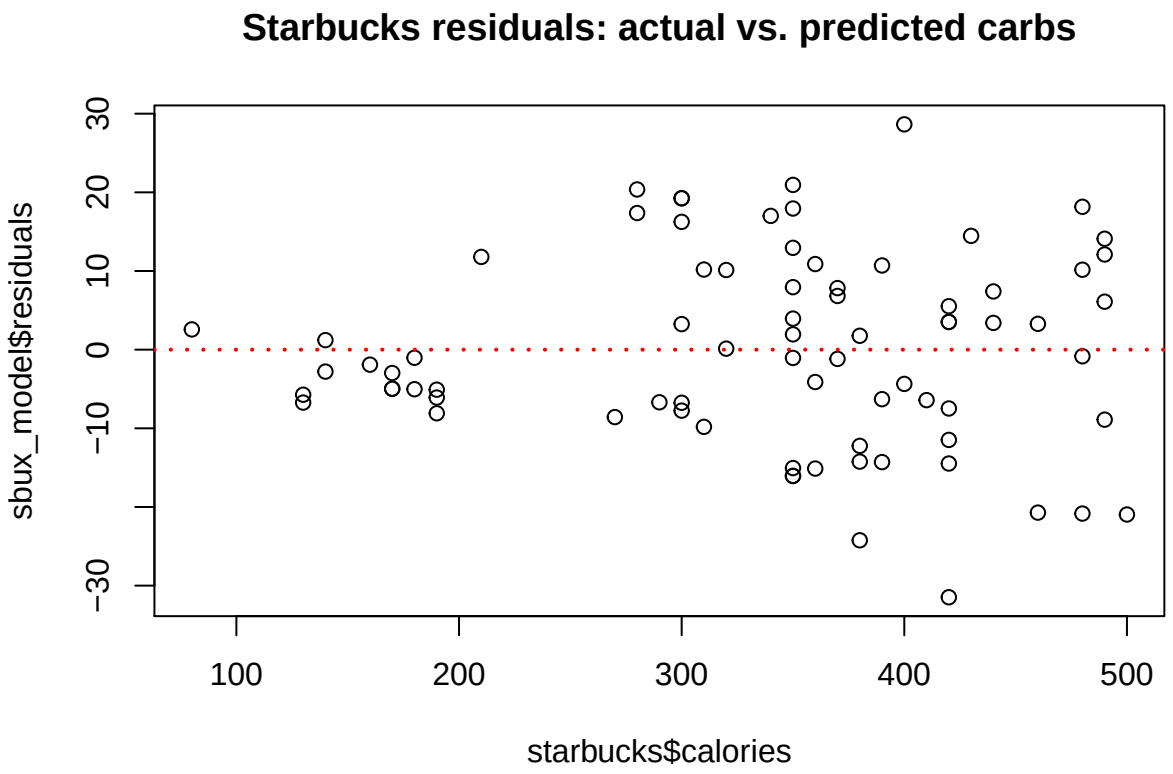
```
anova(sbux_model)
```

```
## Analysis of Variance Table
##
## Response: carb
##           Df  Sum Sq Mean Sq F value            Pr(>F)
## calories   1  9486.4 9486.40 62.7723 0.000000000016725 ***
## Residuals 75 11334.3  151.12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

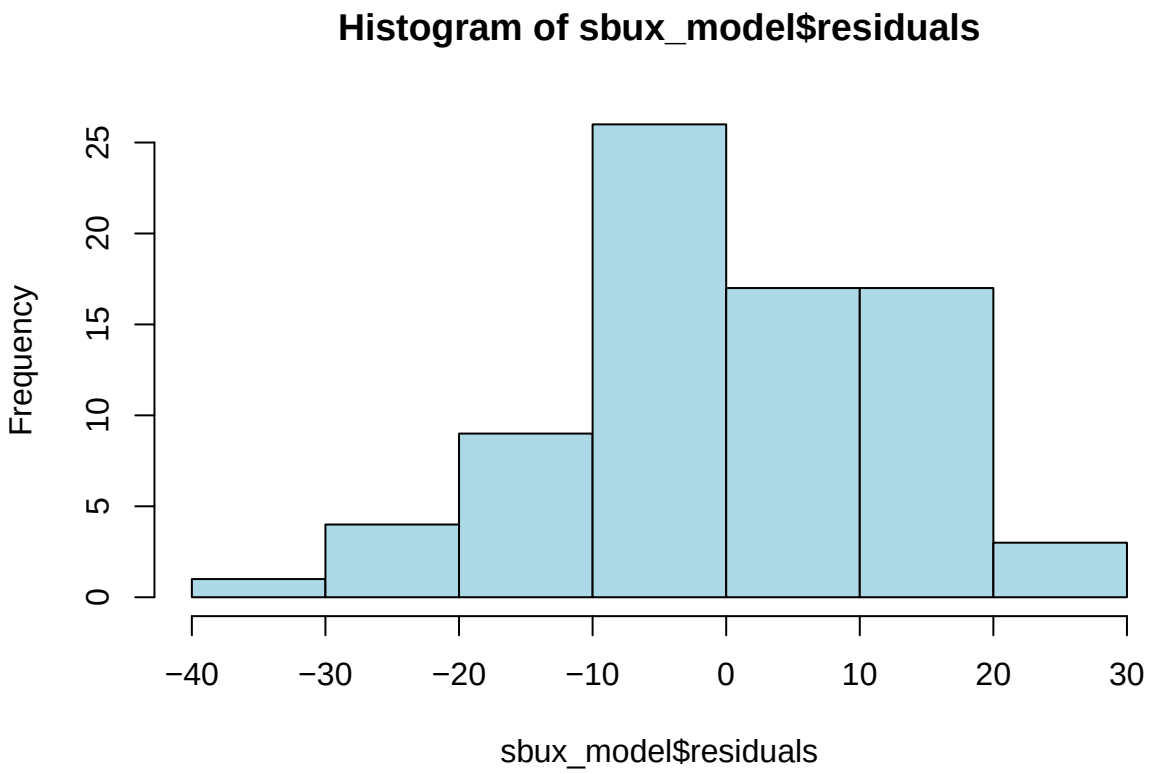
```r
plot(carb ~ calories, data=starbucks)
abline(sbux_model, col="blue")
title(main="Starbucks menu items: calories vs. carbs")
```




```r
sbux_model <- lm(carb ~ calories, data=starbucks)
plot(sbux_model$residuals ~ starbucks$calories)
abline(h = 0, lty = 3, col="red", lwd=2)  # adds a horizontal dashed line at y = 0
title(main="Starbucks residuals: actual vs. predicted carbs")
```




```
hist(sbux_model$residuals, col="lightblue")
```



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

There is an increasing relationship between the number of calories and the amount of carbohydrates that Starbucks menu items contain.

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

The explanatory variable is the amount of calories, while the response variable is the amount of carbs.

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

We may want to evaluate the slope of such line (it's 0.106031) to understand the general relationship between calories and carbs. Also, we may want to evaluate the goodness-of-fit between the actual values and the predicted values by examining the residuals.

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

The data appear to indicate a linear trend; the residuals do not show a pattern.

They appear to be nearly normal.

Regarding heteroscedasticity, the carbs corresponding to low calorie counts show smaller residuals than carbs corresponding to high calorie counts. This suggests non-constant variance, which is noted in the low p-value in the Breusch-Pagan test below. One way to sidestep this problem may be to model the logarithm of the data series rather than the series itself. This is the sole test which does not "pass."

```
require(lmSupport)    ## Note:  the "S" is capitalized in the package name
```

```
## Loading required package: lmSupport
```

```
## Warning: package 'lmSupport' was built under R version 3.5.3
```

```
modelAssumptions(sbux_model,"LINEAR")
```

```
##
## Call:
## lm(formula = carb ~ calories, data = starbucks)
##
## Coefficients:
## (Intercept)     calories
##     8.94356      0.10603
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = Model)
##
##                       Value p-value                Decision
## Global Stat        1.807940 0.77103 Assumptions acceptable.
## Skewness           0.017401 0.89505 Assumptions acceptable.
## Kurtosis           0.511569 0.47446 Assumptions acceptable.
## Link Function      1.189958 0.27534 Assumptions acceptable.
## Heteroscedasticity 0.089013 0.76544 Assumptions acceptable.
```

```
shapiro.test(sbux_model$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sbux_model$residuals
## W = 0.990507, p-value = 0.84255
```

```
ks.test(sbux_model$residuals,"pnorm",0,sd(sbux_model$residuals))
```

```
## Warning in ks.test(sbux_model$residuals, "pnorm", 0, sd(sbux_model$residuals)): ties should not be p
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  sbux_model$residuals
## D = 0.0605478, p-value = 0.94034
## alternative hypothesis: two-sided
```

```
require(nortest)
```

```
## Loading required package: nortest
```

```
ad.test(sbux_model$residuals)
```

```
##
##  Anderson-Darling normality test
##
## data:  sbux_model$residuals
## A = 0.265636, p-value = 0.68343
```

```
require(tseries)
```

```
## Loading required package: tseries
```

```
## Warning: package 'tseries' was built under R version 3.5.3
```

```
jarque.bera.test(sbux_model$residuals)
```

```
##
##  Jarque Bera Test
##
## data:  sbux_model$residuals
## X-squared = 0.52897, df = 2, p-value = 0.7676
```

```
require(olsrr)
```

```
## Loading required package: olsrr
```

```
## Warning: package 'olsrr' was built under R version 3.5.3
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
ols_test_breusch_pagan(sbux_model)
```

```
##
##  Breusch Pagan Test for Heteroskedasticity
##  -----------------------------------------
##  Ho: the variance is constant
##  Ha: the variance is not constant
##
##               Data
##  ---------------------------------
##  Response : carb
##  Variables: fitted values of carb
##
##         Test Summary
##  ----------------------------
##  DF            =    1
##  Chi2          =    5.0494237
##  Prob > Chi2   =    0.02463413
```

#.#

Exercise 7.26 Body measurements, Part III.

Exercise 7.15 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.

```
# look at the exact data set referenced
data("bdims")
#summary(bdims)
p726_shogi_mean <- mean(bdims$sho.gi)
p726_shogi_mean
```

```
## [1] 108.19507
```

```
p726_shogi_sd <- sd(bdims$sho.gi)
p726_shogi_sd
```

```
## [1] 10.374834
```

```
p726_hgt_mean <- mean(bdims$hgt)
p726_hgt_mean
```

```
## [1] 171.14379
```

```
p726_hgt_sd   <- sd(bdims$hgt)
p726_hgt_sd
```

```
## [1] 9.4072052
```

```
p726_correl   <- cor(bdims$sho.gi, bdims$hgt)

p726_slope    <- p726_hgt_sd / p726_shogi_sd * p726_correl
p726_slope
```

```
## [1] 0.60364419
```

```
b1 <- p726_slope

p726_intercept <- p726_hgt_mean - b1 * p726_shogi_mean
p726_intercept
```

```
## [1] 105.83246
```

$$y - \bar{y} = b_1(x - \bar{x})$$

$$y - HeightMean = b_1(x - shoulderGirthMean)$$

$$y - HeightMean = b_1(x - shoulderGirthMean) = b_1 * x - b_1 * shoulderGirthMean$$

$$y = b_1 x + (HeightMean - b_1 shoulderGirthMean)$$

```
###

mod <- lm(bdims$hgt ~ bdims$sho.gi)
summary(mod)
```

```
##
## Call:
## lm(formula = bdims$hgt ~ bdims$sho.gi)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -19.22968  -4.79756  -0.11418   4.78854  21.09789
##
## Coefficients:
##                Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  105.832462   3.272451  32.340 < 0.00000000000000022 ***
## bdims$sho.gi   0.603644   0.030108  20.049 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.0265 on 505 degrees of freedom
## Multiple R-squared:  0.4432, Adjusted R-squared:  0.4421
## F-statistic: 401.97 on 1 and 505 DF,  p-value: < 0.00000000000000222
```

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

Given the actual data in the dataset, the equation would be:

$$height = 105.832462 + .603644 * ShoulderGirth$$

Given the summary data in the textbook (which contains a typo !!!!), we obtain:

```
b1 = 9.41 / 10.37 * 0.67
b1
```

```
## [1] 0.60797493
```

```
b0 = 171.14 -b1 * 107.20
b0
```

```
## [1] 105.96509
```

$$Height = 105.965 + .60797 * ShoulderGirth$$

It is worth noting that the textbook incorrectly indicates that the mean shoulder girth is 107.20 when the dataset indicates that it is actually 108.20 . This typo will yield differing results.

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

The slope of about 0.60 indicates that each 1 centimeter increase in shoulder girth is associated with an increase in height of 0.60 cm.

The intercept of about 105.8 or 105.9 (depending on whether one is using the actual dataset, or relying on the typo in the textbook) indicates that ShoulderGirth of zero predicts a height of about 105.8 cm. (Of course, it is not possible to have a ShoulderGirth of zero. The smallest ShoulderGirth in the data set is 85.90cm.)

(c) Calculate R2 of the regression line for predicting height from shoulder girth, and interpret it in the context of the application.

Given that the correlation is .67, the R^2 would be $.67^2 = .44$

From the data:

```
summary(mod)$r.squared
```

```
## [1] 0.44320349
```

This means that the Shoulder Girth explains 44 percent of the variability of the height.

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

```
predict_exact <- 105.832462 + .603644*100
predict_exact
```

```
## [1] 166.19686
```

```
predict_texterror <- 105.965 + .60797*100
predict_texterror
```

```
## [1] 166.762
```

The model indicates that the student has predicted height of 166.2 cm (using the exact data from the dataset) or the student has predicted height of 166.76cm (using the incorrect summary data in the textbook.)

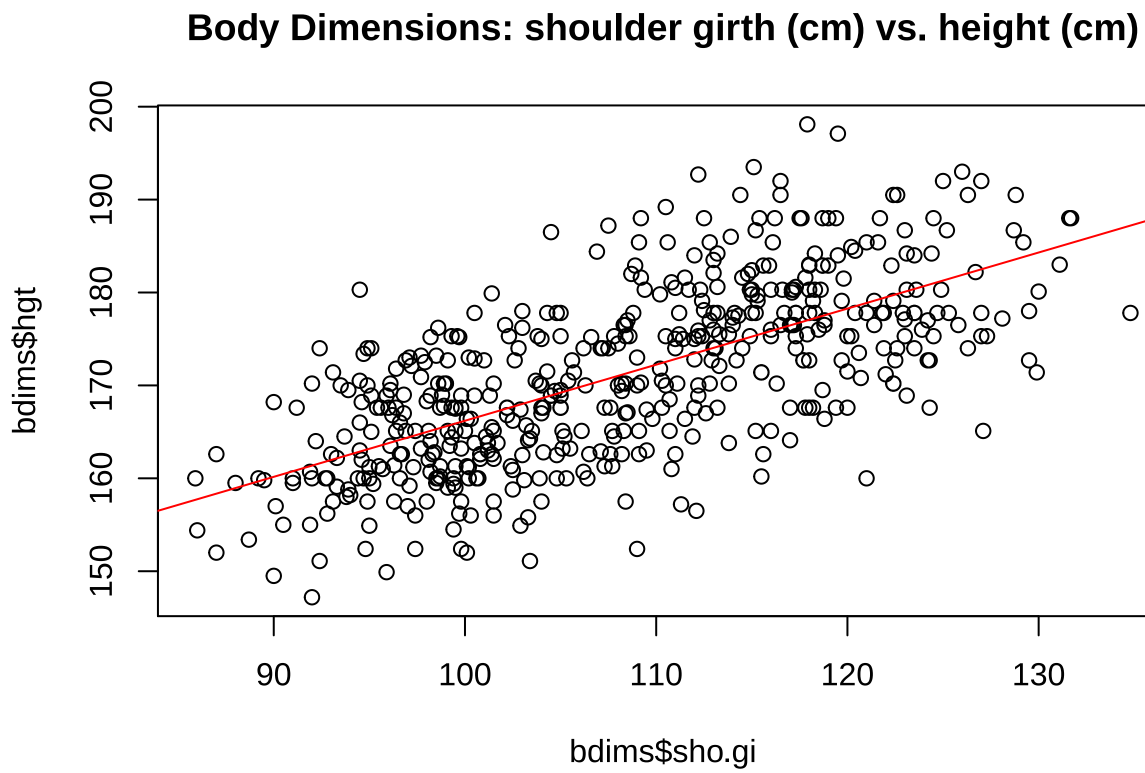
(e) The student from part (d) is 160 cm tall. Calculate the residual, and explain what this residual means.

The residual is the difference between the student's actual height (160 cm) and the predicted height as fitted by the model (either 166.2 or 166.76 , depending on which source you use.) This means that the residual is negative 6.2 cm (or, negative 6.76 cm) as the model has over-predicted this student's height.

(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?

No, it would not be appropriate to use this model because, as indicated above, the smallest shoulder girth in this dataset is 85.9 cm. Doing so would require extrapolation well outside of the range of known data, which is ill-advised.

```
plot(bdims$hgt ~ bdims$sho.gi)
abline(mod, col="red")
title(main="Body Dimensions: shoulder girth (cm) vs. height (cm)")
```



#.#

Exercise 7.30 Cats, Part I.

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.

```
# look at the exact data set referenced
data("cats")
summary(cats)
```

```
##  Sex         Bwt             Hwt
##  F:47   Min.   :2.0000   Min.   : 6.300
##  M:97   1st Qu.:2.3000   1st Qu.: 8.950
##         Median :2.7000   Median :10.100
##         Mean   :2.7236   Mean   :10.631
##         3rd Qu.:3.0250   3rd Qu.:12.125
##         Max.   :3.9000   Max.   :20.500
```

```
catmodel <- lm(cats$Hwt ~cats$Bwt)
summary(catmodel)
```

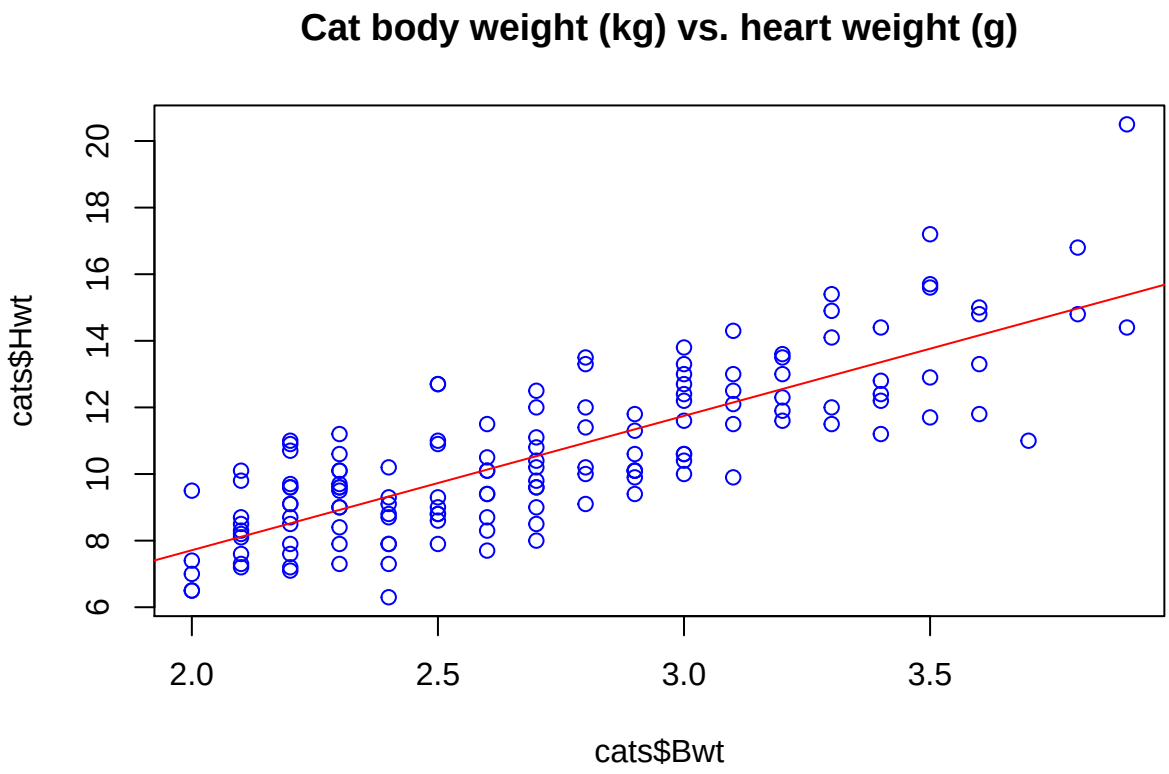
```
##
## Call:
## lm(formula = cats$Hwt ~ cats$Bwt)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.56937 -0.96341 -0.09212  1.04255  5.12382
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) -0.35666    0.69228 -0.5152              0.6072
## cats$Bwt     4.03406    0.25026 16.1194 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.4524 on 142 degrees of freedom
## Multiple R-squared:  0.64662,    Adjusted R-squared:  0.64413
## F-statistic: 259.83 on 1 and 142 DF,  p-value: < 0.000000000000000222
```

```
anova(catmodel)
```

```
## Analysis of Variance Table
##
## Response: cats$Hwt
##            Df  Sum Sq Mean Sq F value                Pr(>F)
## cats$Bwt    1 548.092 548.092 259.835 < 0.000000000000000222 ***
## Residuals 142 299.533   2.109
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# plot the data
plot(cats$Hwt ~cats$Bwt, col="blue")
abline(catmodel, col="red")
title(main = "Cat body weight (kg) vs. heart weight (g)")
```



(a) Write out the linear model.

$$HeartWeight(g) = 4.034 * BodyWeight(kg) - 0.357$$

(b) Interpret the intercept.

The intercept means that if we had a cat with zero body weight, then we would predict such cat to have heart weight of negative 0.357 grams. This is, of course, absurd, because we would never extrapolate so far as to consider a weightless cat.

(c) Interpret the slope.

The slope of 4.034 indicates that if we were to consider two cats whose body weights differ by 1kg, then we would expect that the weights of their hearts would differ by 4.034 grams (with, obviously, the heavier cat possessing the heavier heart.)

Some may believe that this equation indicates that if an individual cat's body weight were to increase by 1kg, then we would expect the weight of that cat's heart to increase by 4.034 grams. However, this was not an experiment which evaluated the weights of cats and their hearts over time – it was an observational study which evaluated the weights across 144 cats, presumably at the same time. Therefore it would not be appropriate to make this statement.

(d) Interpret R2.

As the R^2 measures more than 64 percent, this indicates a strong association between the overall weight of a cat vs. the weight of its heart, with the body weight explaining 64 percent of the variability of the heart weight.

(e) Calculate the correlation coefficient.

```r
p730_R2 <- .6466
p730_correlation = sqrt(p730_R2)
p730_correlation
```

```
## [1] 0.80411442
```

```r
cor(cats$Bwt,cats$Hwt)
```

```
## [1] 0.80412742
```

The correlation coefficient is 0.8041 .

#.#

Exercise 7.40 Rate my professor.

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.

Daniel S Hamermesh and Amy Parker. "Beauty in the classroom: Instructors pulchritude and putative pedagogical productivity". In: Economics of Education Review 24.4 (2005), pp. 369–376.

The scatterplot below shows the relationship between these variables, and also provided is a regression output for predicting teaching evaluation score from beauty score.

```r
# look at the exact data set referenced
data("prof.evaltns.beauty.public")
summary(prof.evaltns.beauty.public)
```

```
##     tenured          profnumber        minority             age          beautyf2upper       beautyflowe
##  Min.    :0.00000   Min.   : 1.000   Min.    :0.00000   Min.   :29.000   Min.    : 1.0000   Min.   :1.00
##  1st Qu.:0.00000   1st Qu.:20.000   1st Qu.:0.00000   1st Qu.:42.000   1st Qu.: 4.0000   1st Qu.:2.00
##  Median :1.00000   Median :44.000   Median :0.00000   Median :48.000   Median : 5.0000   Median :4.00
##  Mean   :0.54644   Mean   :45.434   Mean   :0.13823   Mean   :48.365   Mean   : 5.2138   Mean   :3.96
```

```
##  3rd Qu.:1.00000   3rd Qu.:70.500   3rd Qu.:0.00000   3rd Qu.:57.000   3rd Qu.: 6.0000   3rd Qu.:5.0
##  Max.   :1.00000   Max.   :94.000   Max.   :1.00000   Max.   :73.000   Max.   :10.0000   Max.   :8.0
##  beautymlowerdiv  beautymupperdiv    btystdave          btystdf2u           btystdfl            bty
##  Min.   :1.0000   Min.   :1.0000   Min.   :-1.538843   Min.   :-2.096532   Min.   :-1.668032   Min.
##  1st Qu.:2.0000   1st Qu.:3.0000   1st Qu.:-0.744618   1st Qu.:-0.665002   1st Qu.:-1.136523   1st Qu
##  Median :3.0000   Median :4.0000   Median :-0.156363   Median :-0.187825   Median :-0.073507   Median
##  Mean   :3.4125   Mean   :4.1469   Mean   :-0.088349   Mean   :-0.085794   Mean   :-0.093022   Mean
##  3rd Qu.:5.0000   3rd Qu.:5.0000   3rd Qu.: 0.457253   3rd Qu.: 0.289352   3rd Qu.: 0.458002   3rd Qu
##  Max.   :7.0000   Max.   :9.0000   Max.   : 1.881674   Max.   : 2.198059   Max.   : 2.052527   Max.
##     btystdml          btystdmu           class1             class2             class3
##  Min.   :-1.487607   Min.   :-1.57312   Min.   :0.000000   Min.   :0.0000000   Min.   :0.000000   Min
##  1st Qu.:-0.900065   1st Qu.:-0.65465   1st Qu.:0.000000   1st Qu.:0.0000000   1st Qu.:0.000000   1st
##  Median :-0.312523   Median :-0.19542   Median :0.000000   Median :0.0000000   Median :0.000000   Med
##  Mean   :-0.070146   Mean   :-0.12797   Mean   :0.010799   Mean   :0.0043197   Mean   :0.017279   Mea
##  3rd Qu.: 0.862562   3rd Qu.: 0.26381   3rd Qu.:0.000000   3rd Qu.:0.0000000   3rd Qu.:0.000000   3rd
##  Max.   : 2.037647   Max.   : 2.10074   Max.   :1.000000   Max.   :1.0000000   Max.   :1.000000   Max
##      class6             class7              class8              class9             class10
##  Min.   :0.000000   Min.   :0.0000000   Min.   :0.0000000   Min.   :0.000000   Min.   :0.000000   Min
##  1st Qu.:0.000000   1st Qu.:0.0000000   1st Qu.:0.0000000   1st Qu.:0.000000   1st Qu.:0.000000   1st
##  Median :0.000000   Median :0.0000000   Median :0.0000000   Median :0.000000   Median :0.000000   Med
##  Mean   :0.012959   Mean   :0.0086393   Mean   :0.0043197   Mean   :0.017279   Mean   :0.010799   Mea
##  3rd Qu.:0.000000   3rd Qu.:0.0000000   3rd Qu.:0.0000000   3rd Qu.:0.000000   3rd Qu.:0.000000   3rd
##  Max.   :1.000000   Max.   :1.0000000   Max.   :1.0000000   Max.   :1.000000   Max.   :1.000000   Max
##     class13            class14             class15             class16            class17
##  Min.   :0.0000000   Min.   :0.0000000   Min.   :0.0000000   Min.   :0.0000000   Min.   :0.000000   M
##  1st Qu.:0.0000000   1st Qu.:0.0000000   1st Qu.:0.0000000   1st Qu.:0.0000000   1st Qu.:0.000000   1
##  Median :0.0000000   Median :0.0000000   Median :0.0000000   Median :0.0000000   Median :0.000000   M
##  Mean   :0.0064795   Mean   :0.0064795   Mean   :0.0043197   Mean   :0.0086393   Mean   :0.015119   M
##  3rd Qu.:0.0000000   3rd Qu.:0.0000000   3rd Qu.:0.0000000   3rd Qu.:0.0000000   3rd Qu.:0.000000   3
##  Max.   :1.0000000   Max.   :1.0000000   Max.   :1.0000000   Max.   :1.0000000   Max.   :1.000000   M
##     class20            class21             class22             class23            class24
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.000000   Min.   :0.000000   Min.   :0.0000000   Min
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.0000000   1st
##  Median :0.000000   Median :0.000000   Median :0.000000   Median :0.000000   Median :0.0000000   Medi
##  Mean   :0.010799   Mean   :0.030238   Mean   :0.023758   Mean   :0.010799   Mean   :0.0064795   Mean
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.0000000   3rd
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.000000   Max.   :1.000000   Max.   :1.0000000   Max
##     class27            class28             class29             class30         courseevaluation dido
##  Min.   :0.0000000   Min.   :0.0000000   Min.   :0.0000000   Min.   :0.000000   Min.   :2.1000   Min
##  1st Qu.:0.0000000   1st Qu.:0.0000000   1st Qu.:0.0000000   1st Qu.:0.000000   1st Qu.:3.6000   1st
##  Median :0.0000000   Median :0.0000000   Median :0.0000000   Median :0.000000   Median :4.0000   Medi
##  Mean   :0.0043197   Mean   :0.0086393   Mean   :0.0043197   Mean   :0.017279   Mean   :3.9983   Mean
##  3rd Qu.:0.0000000   3rd Qu.:0.0000000   3rd Qu.:0.0000000   3rd Qu.:0.000000   3rd Qu.:4.4000   3rd
##  Max.   :1.0000000   Max.   :1.0000000   Max.   :1.0000000   Max.   :1.000000   Max.   :5.0000   Max
##      formal           fulldept           lower          multipleclass       nonenglish          onecred
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.000000   Min.   :0
##  1st Qu.:0.00000   1st Qu.:1.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0
##  Median :0.00000   Median :1.00000   Median :0.00000   Median :0.00000   Median :0.000000   Median :0
##  Mean   :0.16631   Mean   :0.89417   Mean   :0.33909   Mean   :0.33909   Mean   :0.060475   Mean   :0
##  3rd Qu.:0.00000   3rd Qu.:1.00000   3rd Qu.:1.00000   3rd Qu.:1.00000   3rd Qu.:0.000000   3rd Qu.:0
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.000000   Max.   :1
##  profevaluation      students        tenuretrack      blkandwhite      btystdvariance      btystdavep
##  Min.   :2.3000   Min.   :  8.000   Min.   :0.0000   Min.   :0.00000   Min.   :0.085029   Min.   :0.0
##  1st Qu.:3.8000   1st Qu.: 19.000   1st Qu.:1.0000   1st Qu.:0.00000   1st Qu.:0.828371   1st Qu.:0.0
```

```
##  Median :4.3000   Median : 29.000   Median :1.0000   Median :0.00000   Median :1.565791   Median :0.
##  Mean   :4.1747   Mean   : 55.177   Mean   :0.7797   Mean   :0.16847   Mean   :1.842626   Mean   :0.
##  3rd Qu.:4.6000   3rd Qu.: 60.000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:2.682287   3rd Qu.:0.
##  Max.   :5.0000   Max.   :581.000   Max.   :1.0000   Max.   :1.00000   Max.   :5.791667   Max.   :1.
```

(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.

The slope can be computed from the summary table by multiplying the t-value by the standard error, i.e., 4.13 * .0322 = 0.133

Alternatively, the slope can also be computed by recognizing that it is

$$b_1 = \frac{y - \bar{y}}{x - \bar{x}} = \frac{4.010 - 3.9983}{0 - (-0.0883)} = \frac{0.0117}{0.0883} = 0.133$$

```
print("Beauty:")
```

```
## [1] "Beauty:"
```

```
beauty = prof.evaltns.beauty.public$btystdave
t(t(summary(beauty)))
```

```
##         [,1]
## Min.    -1.538843
## 1st Qu. -0.744618
## Median  -0.156363
## Mean    -0.088349
## 3rd Qu.  0.457253
## Max.     1.881674
```

```
print("Evaluations:")
```

```
## [1] "Evaluations:"
```

```
eval = prof.evaltns.beauty.public$courseevaluation
t(t(summary(eval)))
```

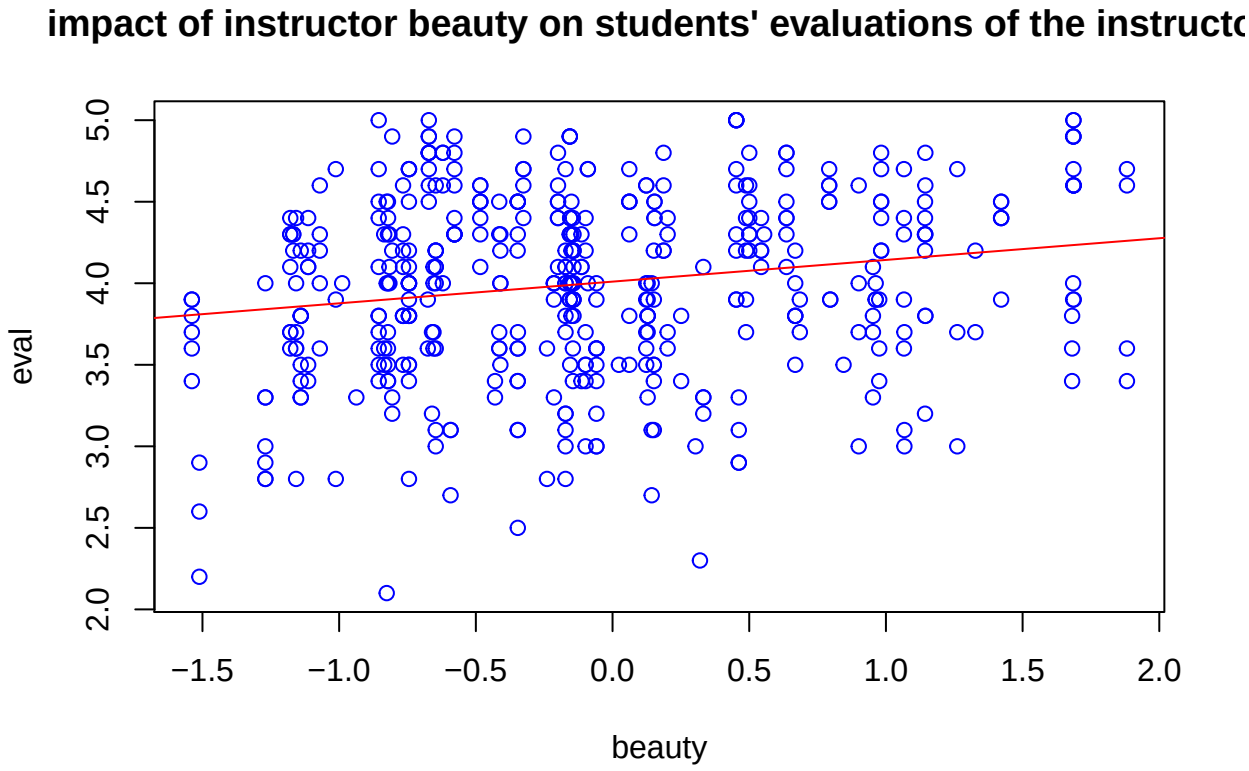
```
##         [,1]
## Min.    2.1000
## 1st Qu. 3.6000
## Median  4.0000
## Mean    3.9983
## 3rd Qu. 4.4000
## Max.    5.0000
```

```
ratingmodel <- lm(eval~beauty)
summary(ratingmodel)
```

```
##
## 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) 4.010023   0.025508 157.2052 < 0.00000000000000022 ***
## beauty      0.133001   0.032178   4.1334            0.00004247 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.54545 on 461 degrees of freedom
## Multiple R-squared:  0.035736,   Adjusted R-squared:  0.033644
## F-statistic: 17.085 on 1 and 461 DF,  p-value: 0.000042471
```

```
plot(eval~beauty, col="blue")
abline(ratingmodel, col="red")
title(main="impact of instructor beauty on students' evaluations of the instructor")
```



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

Yes, because the slope of 0.133 has a standard error of 0.0322 which means that it is 4.13 standard deviations away from zero (which is also indicated). This means that a confidence interval about the point estimate of the slope would be approximately (.0686 , .1974), again, not covering zero. Furthermore, the p-value shown in the regression summary is 0.0000 which means that the null hypothesis – that the slope is not different from zero – is rejected in favor of the alternative.

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

The conditions required for linear regression include:

(1) Linearity - the data should show a linear trend. The data does appear to show a slight trend. However, certain tests do not pass:

```
require(lmSupport)    ## Note:  the "S" is capitalized in the package name
modelAssumptions(ratingmodel,"LINEAR")
```

```
##
## Call:
## lm(formula = eval ~ beauty)
##
## Coefficients:
## (Intercept)       beauty
##       4.010        0.133
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = Model)
##
##                      Value     p-value                   Decision
## Global Stat        17.54617 0.001513318 Assumptions NOT satisfied!
## Skewness           15.85720 0.000068306 Assumptions NOT satisfied!
## Kurtosis            0.54822 0.459045812    Assumptions acceptable.
## Link Function       0.92552 0.336029551    Assumptions acceptable.
## Heteroscedasticity  0.21522 0.642705825    Assumptions acceptable.
```

(2) Nearly Normal Residuals - the low p-values indicate that all of the below tests fail.

```
shapiro.test(ratingmodel$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ratingmodel$residuals
## W = 0.981504, p-value = 0.000012576
```

```
ks.test(ratingmodel$residuals,"pnorm",0,sd(sbux_model$residuals))
```

```
## Warning in ks.test(ratingmodel$residuals, "pnorm", 0, sd(sbux_model$residuals)): ties should not be p
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  ratingmodel$residuals
## D = 0.463993, p-value < 0.00000000000000222
## alternative hypothesis: two-sided
```

```
require(nortest)
ad.test(ratingmodel$residuals)
```

```
##
##  Anderson-Darling normality test
##
## data:  ratingmodel$residuals
## A = 1.99014, p-value = 0.00004483
```

```
require(tseries)
jarque.bera.test(ratingmodel$residuals)
```

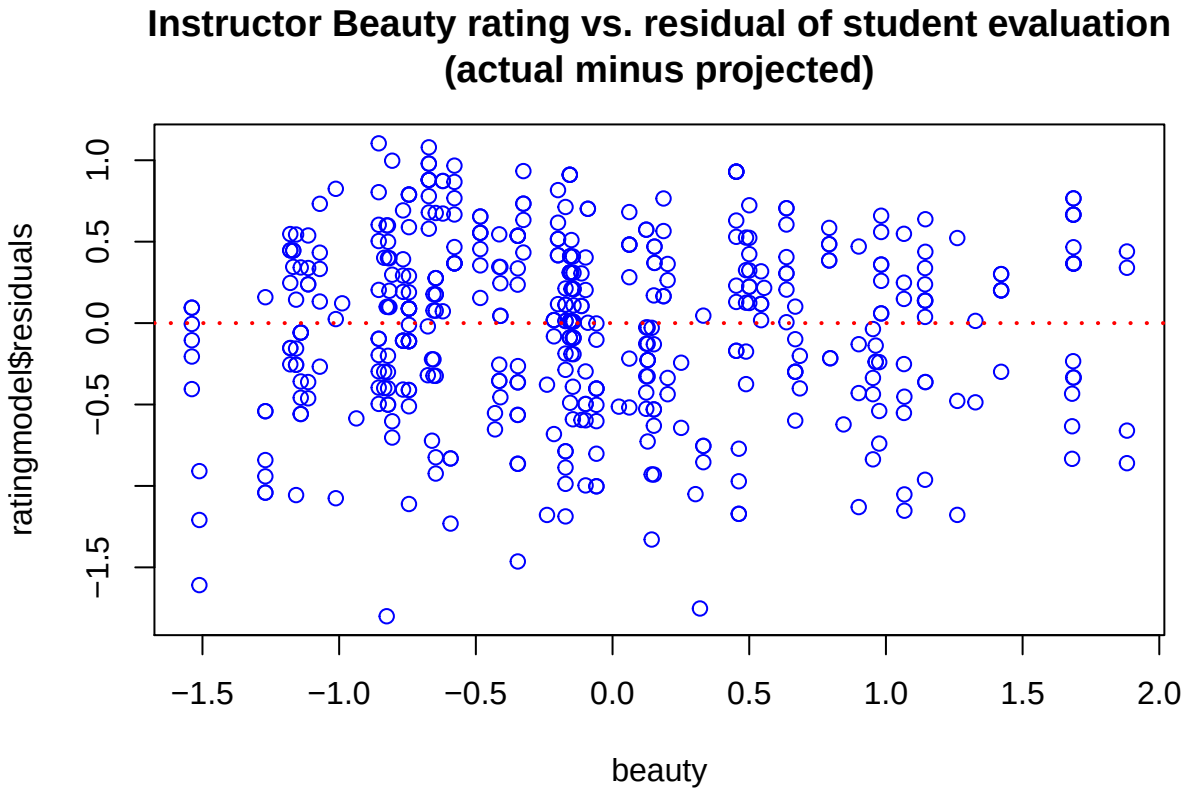
```
##
##  Jarque Bera Test
##
## data:  ratingmodel$residuals
## X-squared = 16.4054, df = 2, p-value = 0.00027391
```

(3) Constant variability: The below test assumes Normality, which failed above, so we shouldn't use it. But, just for fun, let's have a look:

```
require(olsrr)
ols_test_breusch_pagan(ratingmodel)
```

```
##
##  Breusch Pagan Test for Heteroskedasticity
##  -----------------------------------------
##  Ho: the variance is constant
##  Ha: the variance is not constant
##
##               Data
##  ---------------------------------
##  Response : eval
##  Variables: fitted values of eval
##
##         Test Summary
##  -----------------------------
##  DF            =    1
##  Chi2          =    0.93015634
##  Prob > Chi2   =    0.3348223
```

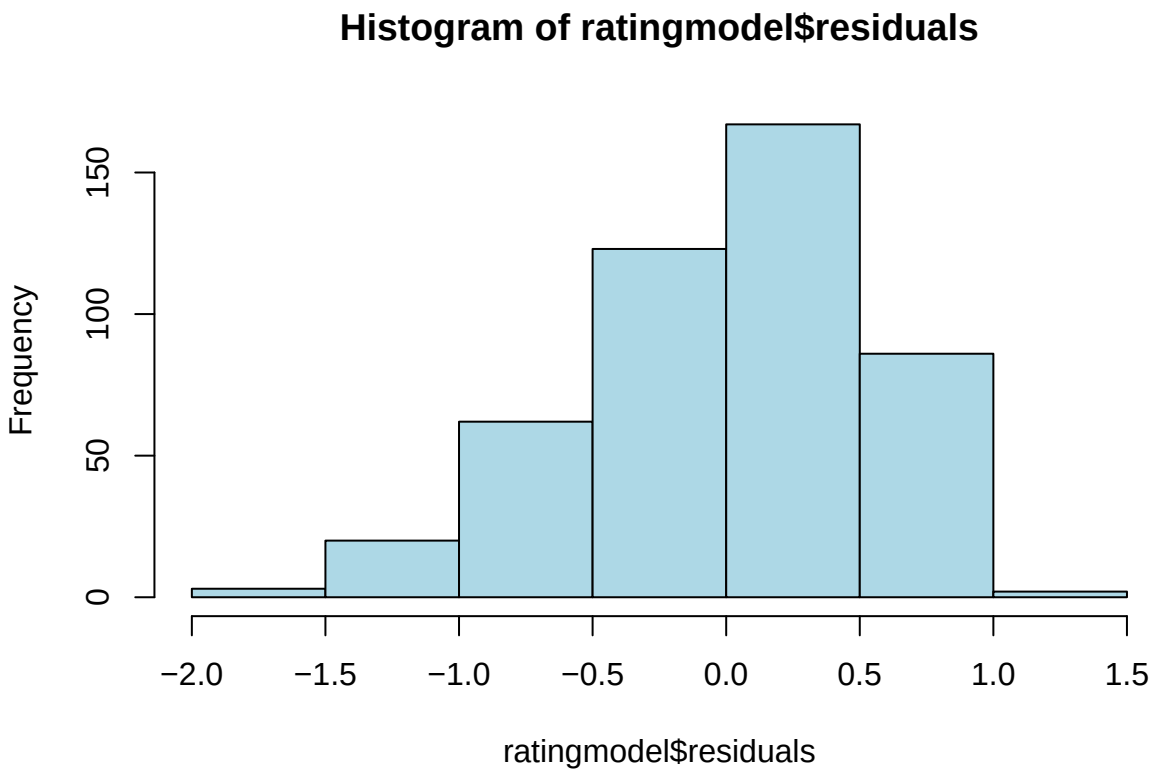
```
plot(ratingmodel$residuals ~ beauty, col="blue")
abline(h = 0, lty = 3, col="red", lwd=2)  # adds a horizontal dashed line at y = 0
title(main="Instructor Beauty rating vs. residual of student evaluation\n(actual minus projected)")
```



Above, the residuals do not show an obvious pattern, so that is OK.

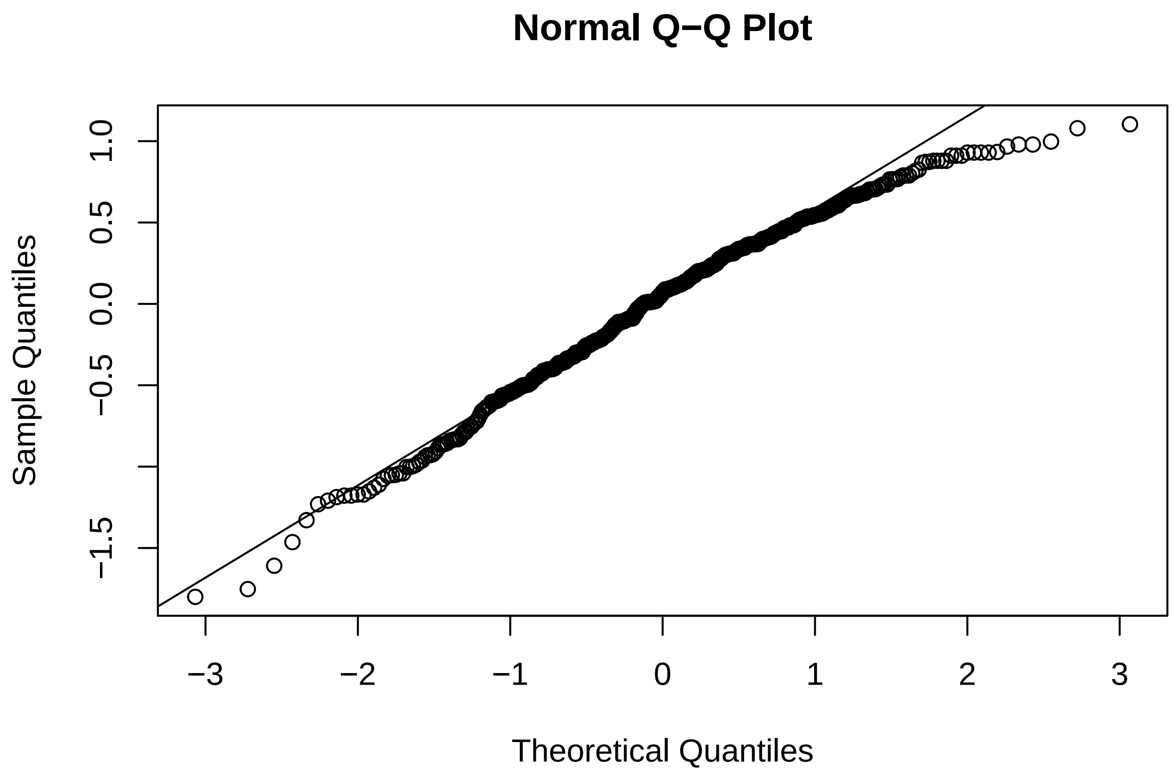
However, there are significantly more residual points above the line (255/463 = 55%) than there are below the line (208/463 = 45%). Because the sum of all residuals must add up to zero, it is necessary that the average value of those residual points below the line (-0.494) is significantly larger (in absolute value) that the average value of those residuals above the line (here, +0.403).

```
hist(ratingmodel$residuals, col="lightblue")
```



Above, the histogram of the residuals indicates a strong skew, which is not consistent with normality.

```
qqnorm(ratingmodel$residuals)
qqline(ratingmodel$residuals)  # adds diagonal line to the normal prob plot
```



The QQ-plot reveals tails which differ significantly from normality.

(4) Independent Observations

Finally, the "Order of Data Collection" plot (which I am uncertain how to reproduce) does not appear to show any pattern or bias in regard to the impact of such sequence on the corresponding residuals.

#.#