STAT 252

Week 6

May 13, 2019 – May 17, 2019

**Transformation**

**Goals od Transformation**

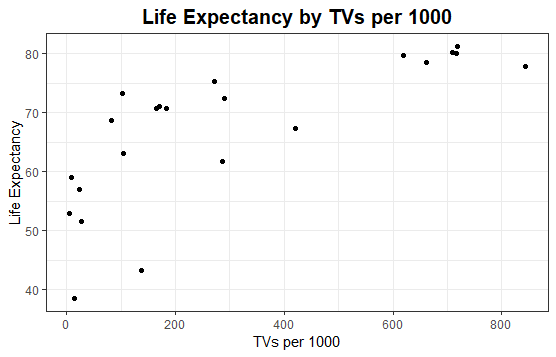
**1.**

**2.**

**3.**

**4.**

**Example 1:** There was a study to see whether the number of TVs for every 1000 people in a country has an impact on the country’s life expectancy. Below is a scatter plot that shows the relationship between these variables.

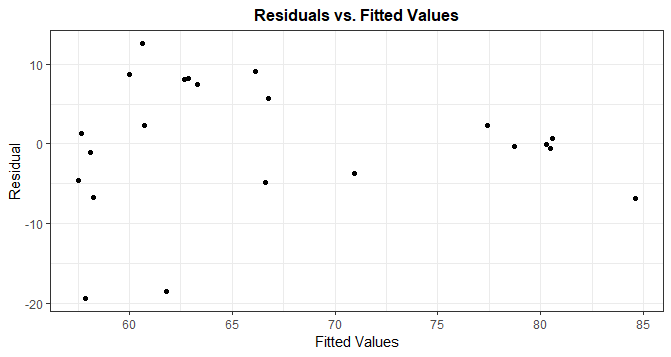


Question 1: What is the observation unit?

Question 2: What is the explanatory variable?

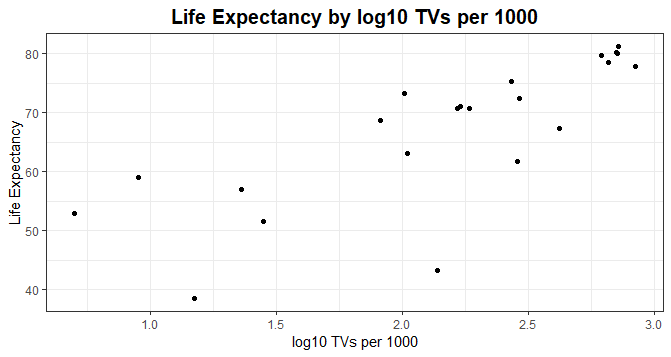
Question 3: What is the response variable?

Question 4: Would a linear regression line appear to summarize the relationship well? Explain.



Question 5: Above is the a plot of the residuals vs. fitted. Does this residual plot reveal random scatter or a pattern?

**Transform explanatory variable by using on TVs per 1000**



Question 6: What do you notice in this plot?

Below is the R output when using on TVs per 1000 as the explanatory variable.

Call:

lm(formula = life\_expectancy ~ log\_tv, data = tv)

Residuals:

Min 1Q Median 3Q Max

-23.407 -1.558 2.558 4.051 9.391

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 35.848 5.885 6.092 5.92e-06 \*\*\*

log\_tv 14.421 2.611 5.524 2.09e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.903 on 20 degrees of freedom

Multiple R-squared: 0.6041, Adjusted R-squared: 0.5843

F-statistic: 30.51 on 1 and 20 DF, p-value: 2.086e-05

Analysis of Variance Table

Response: life\_expectancy

Df Sum Sq Mean Sq F value Pr(>F)

log\_tv 1 1905.7 1905.68 30.515 2.086e-05 \*\*\*

Residuals 20 1249.0 62.45

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Question 7: Report the equation of this regression line and interpret.

Question 8: Report and interpret the model metrics.

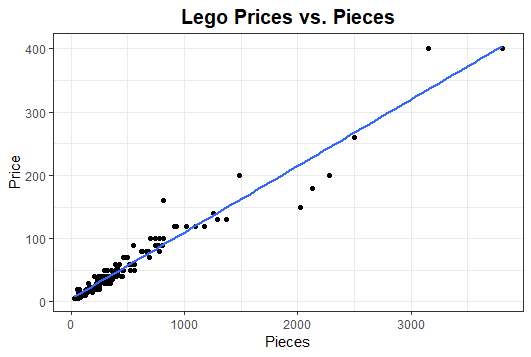
Question 9: Use this regression equation to predict the life expectancy for a country with 800 televisions per thousand people. Then repeat for a country with 80 televisions per thousand people. [*Hint*: Remember to first convert 800 and 80 to transformed values.]

Question 11: Calculate the difference in these predictions. Does the result look familiar? Explain why this makes sense.

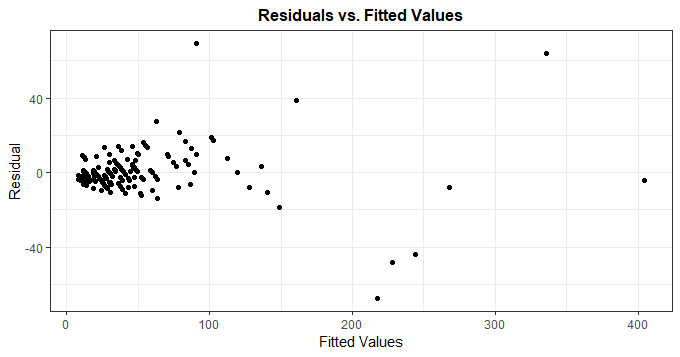
With a *log*-transformed explanatory variable, the slope coefficient reveals the predicted change in the response variable for a *10-times* increase in the explanatory variable.

**Example 2:** A data set exists that has the prices and number of pieces of a Lego sets advertised for sale the Lego website in 2012.

The correlation between the two variables is 0.9739.

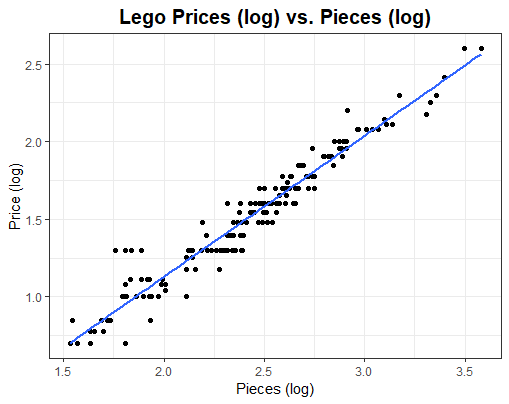


Question 12: Report the value of *­.* Does the line appear to provide a reasonable model for the relationship?



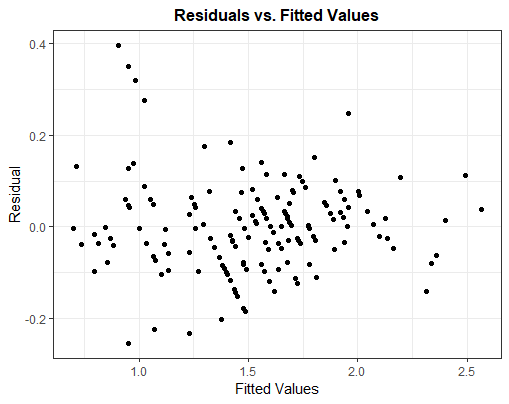
Question 13: Examine residual plots above, does the assumption that the residuals have constant variance across the fitted values appear to be satisfied?

**To address the concern about constant variance of residuals, we will apply a log transformation (base 10) to *both* variables.**



Question 14: The above plot depicts the log transformation applied to both variables. Does this model fit the data better than the first model?

Question 15: Below is the residual plot of the model that used log transformation to both variables. What are some conclusions that you can make from this graphic?



Below is the regression output for the log transformation from R :

Call:

lm(formula = log\_price ~ log\_pieces, data = lego)

Residuals:

Min 1Q Median 3Q Max

-0.25363 -0.05701 -0.00262 0.04846 0.39483

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.69092 0.04542 -15.21 <2e-16 \*\*\*

log\_pieces 0.90946 0.01848 49.21 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1001 on 155 degrees of freedom

Multiple R-squared: 0.9398, Adjusted R-squared: 0.9394

F-statistic: 2421 on 1 and 155 DF, p-value: < 2.2e-16

Analysis of Variance Table

Response: log\_price

Df Sum Sq Mean Sq F value Pr(>F)

log\_pieces 1 24.2416 24.242 2421.3 < 2.2e-16 \*\*\*

Residuals 155 1.5518 0.010

Question 16: Report the regression equation and the model metrics.

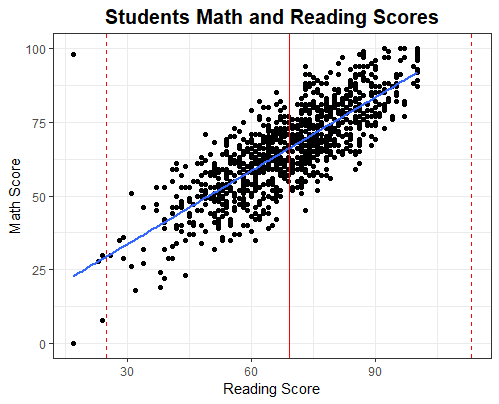
Question 17: Use the regression model (with the transformed variables) to predict (by hand) the price for a Lego set with 400 pieces. [*Hint*: First take the log of the number of pieces, then plug into the equation to predict the log of the price, then take 10 to that power in order to “back-transform” to predict the price rather than the log price.] Confirm that your prediction seems reasonable, based on the original scatterplot.

**Outliers, Leverage and Influence**

An **outlier** is when a point’s y-value is far from the regression line.

A point has **high leverage** point when a point’s x-value is far from the mean of x-values

A point is **influential** if it is removed and the model is very different from the original model.



**Original regression model with point**

**New regression model without point**

Question 18: Circle the points that are outliers.

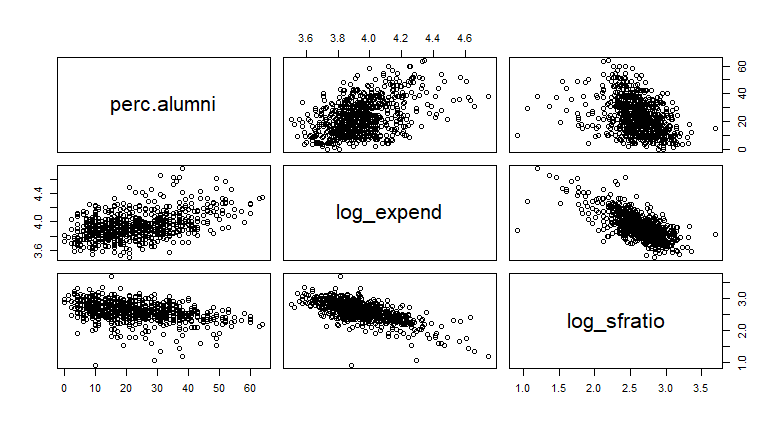
Question 19: Put an X on the points that have high leverage.

Question 20: Discuss whether the point that has an X on is influential.

**Beginning of Multiple Regression**

A dataset exists that investigates 777 colleges in terms of 15 characteristics. We will analyze three of the 15 variables here:

1. perc.alumni – % alumni who donate
2. log\_expend – log based 10 of instructional expenditure per student
3. log\_sfratio – log based 10 of student/faculty ratio



Question 21: Explain the relationship between log\_expend and log\_sfratio.

Question 22: Explain the relationship between log\_expend and perc.alumni.

Question 23: Explain the relationship between log\_sfratio. and perc.alumni.

Question 24: Of the three variables, what makes the most sense to be the response variable?

**Below is the R output for predicting perc.alumni from log\_sfratio**

Call:

lm(formula = perc.alumni ~ log\_sfratio, data = college)

Residuals:

Min 1Q Median 3Q Max

-39.352 -8.632 -1.397 7.428 34.848

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 63.801 3.579 17.82 <2e-16 \*\*\*

log\_sfratio -15.770 1.366 -11.55 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.45 on 775 degrees of freedom

Multiple R-squared: 0.1468, Adjusted R-squared: 0.1457

F-statistic: 133.3 on 1 and 775 DF, p-value: < 2.2e-16

**Below is the R output for predicting perc.alumni from log\_expend**

Call:

lm(formula = perc.alumni ~ log\_expend, data = college)

Residuals:

Min 1Q Median 3Q Max

-24.672 -8.554 -1.131 7.638 33.572

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -104.314 8.786 -11.87 <2e-16 \*\*\*

log\_expend 32.218 2.226 14.48 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11 on 775 degrees of freedom

Multiple R-squared: 0.2128, Adjusted R-squared: 0.2118

F-statistic: 209.6 on 1 and 775 DF, p-value: < 2.2e-16

Question 25: Write the regression model and for predicting perc.alumni from log\_expend.

Question 26: Write the regression model and for predicting perc.alumni from log\_sfratio.

Question 27: Which model should be used to predict percent of alumni that donate?

**Multiple regression allows for more than one explanatory variable to be used for predicting the response variable.**

**Below is the R output for predicting perc.alumni from log\_expend and log\_sfratio**

Call:

lm(formula = perc.alumni ~ log\_expend + log\_sfratio, data = college)

Residuals:

Min 1Q Median 3Q Max

-26.710 -8.347 -1.006 7.609 33.489

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -67.791 15.737 -4.308 1.86e-05 \*\*\*

log\_expend 26.291 3.069 8.567 < 2e-16 \*\*\*

log\_sfratio -5.050 1.809 -2.792 0.00537 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.95 on 774 degrees of freedom

Multiple R-squared: 0.2207, Adjusted R-squared: 0.2187

F-statistic: 109.6 on 2 and 774 DF, p-value: < 2.2e-16

Question 28: Write the regression model for predicting perc.alumni from log\_expend and log\_sfratio.

Question 29: Interpret the Regression coefficients.

Question 30: Report the for all 3 models ran. How does the for the multiple regression model compare to the other 2?

Question 31: Predict the percent alumni when expenditure is 25000 and student per faculty ratio is 35. Determine if the values are plausible.

indicates the proportion of variability in the response variable explained by the multiple regression model.

As more explanatory variables are added to the model, can only increase (or stay the same).

, the proportion of variability in the response variable that is explained by the regression model with the explanatory variables

Larger is better, but adding variables to the model cannot reduce .

**Adjusted**  tries to compensate for adding variables that might not be very useful to the model, calculated as .