Name: Brad Hall
ISDS 7024
Homework 4

1) Variables created by counting people, objects, events, etc., are usually characterized by:

- **(a)** a right skew
- **(b)** a left skew
- **(c)** a symmetrical shape
- **(d)** none of these

Solution: **(d)** None of these

2) Transformations are commonly used to change all of the attributes of data distributions below except:

- **(a)** the center
- **(b)** scaling
- **(c)** shapes, skewness, and kurtosis
- **(d)** rank order of elements

Solution: **(d)** rank order of elements

3) When $\lambda = 0.5$, Y^{λ} is equivalent to $\sqrt{Y}$.

- **(a)** True
- **(b)** False

Solution: **(a)** True

4) The log() function is at the high end of the ladder of powers between $\lambda = 1.5$ and $\lambda = 2.5$.

- **(a)** True
- **(b)** False

Solution: **(b)** False

Use `dataset29.xlsx` to answer questions 5-9.

5) Run the model $Y = X_1 + X_2$.

- **(a)** What is the value of R^2? (Round to 4 decimal places.)

  Solution: $R^2 = 0.7064$

(b) Is the assumption of normal residuals violated? Why or why not?

Solution: **Yes.** The assumption of normal residuals is violated, because the p-value from the Shapiro-Wilk test yields a p-value less than .0001. This corresponds to rejecting the null hypothesis which states that the data is from a normal distribution.

6) Using the model $Y = X_1 + X_2$, linearity and non-constant variance seem to be issues. Which variable is responsible? (Plot X_1 and X_2 against Y separately. Plot Y, X_1, and X_2 against the fitted values. You may need to standardize X_1 and X_2 to make them easier to compare.)

(a) X_1

(b) X_2

(c) X_1 and X_2

Solution: **(c)** X_1 and X_2

7) Use the Box-Cox approach to transform Y, X_1, and X_2 in an attempt to reduce the curvilinearity and non-constant variance. Do not drop any variables or delete any cases. What is the best lambda for each of the following variables. (Round to 1 decimal place.)

(a) X_1 - *Solution:* $\lambda = -0.5$

(b) X_2 - *Solution:* $\lambda = -0.3$

(c) Y - *Solution:* $\lambda = -0.1$

(d) Run the regression analysis (transformed Y = transformed X_1 + transformed X_2) using the transformations obtained from the Box-Cox approach. What is the R^2 value? (Be sure to use the coefficients from the Box-Cox method rounded to one decimal here.)

Solution: $R^2 = 0.8$

(e) Have the following violations satisfactorily been improved?

- Linearity Assumption? Yes or No

  Solution: **Yes.** Found by comparing the scatterplot of externally studentized residuals vs predicted values for both models.

- Equal Variance Assumption? Yes or No

  Solution: **Yes.** For the original model, the Breusch-Pagan Test yields a p-value of $\text{CHISQ.DIST.RT}(37.5237, 2) = 0$. For the transformed model, the Breusch-Pagan Test yields a p-value of $\text{CHISQ.DIST.RT}(8.5812, 2) = 0.01$.

- Normality of Residuals? Yes or No

  Solution: **Yes.** This is because the Shapiro-Wilk test statistic increases from .9733 to .9940, which increases the p-value to 0.1216 and allows us to conclude normality of the residuals.

For problems 8 and 9, do not drop any variables or delete any cases. Using the bulging rule, you could use $\log(Y)$ or $\sqrt{Y}$, and $\log(X)$ or $\sqrt{X}$ for each of the X's, or any other combinations.

8) The model $\log(Y) = \log(X_1) + \log(X_2)$ is ______________.

(a) better than the model using the original untransformed variables.

(b) worse than the model using the original untransformed variables.

(c) about the same as the model using the original untransformed variables.

Solution: **(a)** better than the model using the original untransformed variables. (Larger R^2 in transformed model.)

9) The model $\sqrt{Y} = \sqrt{X_1} + \sqrt{X_2}$ is ______________.

(a) better than the model using the original untransformed variables.

(b) worse than the model using the Box-Cox transformed variables.

(c) worse than the model using the LOG-transformed variables.

(d) All of the above statements are true.

(e) None of the above statements are true.

Solution: **(d)** All of the above statements are true. (Produces $R^2 = .76112$.)

Have the following violations satisfactorily been improved?

- Linearity Assumption? Yes or No

  Solution: **Yes.** Found by comparing the scatterplot of externally studentized residuals vs predicted values for both models.

- Equal Variance Assumption? Yes or No

  Solution: **Yes.** For the original model, the Breusch-Pagan Test yields a Chi-squared test statistic value of 37.5237. For the transformed model, the Breusch-Pagan Test yields a Chi-squared test statistic value of 28.7275. The smaller test statistic produces a slightly larger p-value, which improves (but does not quite show) equal variance.

- Normality of Residuals? Yes or No

  Solution: **Yes.** This is because the Shapiro-Wilk test statistic increases from .9733 to .9889, which increases the p-value to 0.0045.

Use the dataset `set6f.xlsx` to answer questions 10 and 11.

10) The scatterplot of Y versus X shows two distinct groups in `set6f.xlsx`. Create a dummy variable (d) and assign 0 or 1 to each case to separate the groups. Run the model $Y = d$. Select the range that contains the regression coefficient for d.

(a) Coeff(d) < 5.0

(b) $5.0 \leq$ Coeff(d) < 6.0

(c) $6.0 \leq$ Coeff(d) < 7.0

(d) $7.0 \leq$ Coeff(d) < 8.0

(e) Coeff(d) ≥ 8.0

Solution: **(e)** Coeff(d) ≥ 8.0 [$\widehat{\beta} = 8.4861$]

11) Compare the coefficients and standard errors for the models $Y = X$, $Y = d$, and $Y = X + d$. Run the diagnostics. Which assumptions are violated by $Y = X + d$?

(a) Constant variance

Solution: **Not violated.** The Breusch-Pagan Test yields a Chi-squared test statistic value of 2.536215, which results in a p-value of 0.281364. We fail to reject the null hypothesis, which implies constant variance of the model.

Solution:

(b) Normality of residuals

Solution: **Not violated.** The Shapiro-Wilk Test yields a p-value of 0.4123, which means we fail to reject the null hypothesis and thus conclude with normality of the residuals.

(c) No outliers or influential cases

Solution: **Violated.** Row 38 gives us an influential outlier with a Cook's D of 0.13, above the threshold of $4/n = 4/43 \approx 0.093$.