Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

HOMEWORK 2: Wonky Residuals and Transformations

WSCI 6390 – 002: Population Parameter Estimation

Due 11:59 PM Tuesday, February 6

Let’s practice (1) accounting for residuals that do not meet linear model assumptions and (2) using transformations within linear models that have non-constant variance

INSTRUCTIONS:

We will be using two datasets for this assignment:

1. A dataset “income.csv” that describes daily income (in dollars) as a function of years in the work force (years).
2. A dataset “gala.csv” from 30 Galapagos Islands. The number of plant species on each island is reported, and we are interested in modeling the number of plant species based on several geographic variables. You can use the command ?gala after loading the faraway package to learn more.

Be sure to install the packages faraway and MASS before completing this assignment. Also, be sure to read in these two datasets using read.csv(). You can call the first one “income” and the second one “gala” (or whatever you want!).

ASSIGNMENT:

1. Can you please re-draw that model from last week that was equivalent to ? Can you draw “fake” data points around that line and draw the residuals? Can you remind me what the residuals represent, and why we are trying to minimize the sum of their squares in linear regression?
2. What are the three assumptions that residuals must follow in a linear model, and what are the approaches you can use when residuals do not follow those assumptions?

|  |  |
| --- | --- |
| Residual assumption | Statistical approach when this assumption is not met |
| 1. | 1. |
| 2. | 2. |
| 3. | 3. |

1. Let’s load the (very fake dataset that I made) called “income.” Let’s run the most basic model ever: . Use the lm() argument in R to build this model. Hint: use summary(MODEL\_NAME) to see model output.
2. What is the value of the intercept?
3. What is the value of the intercept standard error?
4. What does this intercept mean in regular-speak?
5. A. What is the value of the beta coefficient for years?
6. What is the standard error?
7. What does this slope mean in regular-speak?
8. Similar to last week’s homework assignment, produce a plot of residuals versus fitted values. Add a line at for perspective:

plot(MODEL\_NAME$fitted.values, MODEL\_NAME$residuals)

abline(h = 1)

Paste the plot here. Do we meet the assumption of constant variance?

1. Let’s test the assumption of constant variance one other way. Fit the regression model Code is:

model.test <- lm(sqrt(abs(MODEL\_NAME $residuals)) ~ MODEL\_NAME $fitted)

summary(model.test)

Do we reach a different conclusion here?

1. Let’s run a weighted least squares (WLS) regression to account for non-normal variance in residuals. First of all, our weights will be proportional to the variance, such that data points with large variance will have smaller weights. This line of code is:

weights <- 1 / lm(abs(MODEL\_NAME$residuals) ~ MODEL\_NAME$fitted.values)$fitted.values^2

Once you have your weights, please run the weighted least squares regression using this line:

model.2 <- lm(income ~ years, weights=weights, data=income)

Can you tell me what the intercept and slope terms are now?

Intercept:

Slope:

1. In order to see if we now have constant variance, we will have to use a kind of residual called “standardized residuals.” Standardized residuals plot residuals in terms of standard deviations rather than raw values, and be helpful in pinpointing outlier values. We can plot our fitted values against standardized residuals using

plot(fitted(model.weighted), rstandard(model.weighted))

Please paste the plot below. Does it look like we’ve corrected for unequal variance (heteroscedasticity) with our weights?

1. Now let’s move onto the “gala” dataset. Let’s build this model:

model <- lm(Species ~ Area + Nearest + Scruz + Adjacent, data=gala)

What are the intercept term and beta coefficients for Area, Nearest, Scruz, and Adjacent?

Intercept:

Area:

Nearest:

Scruz:

Adjacent:

1. Given these intercept and slope terms, can you summarize what factors are impacting the number of plant species on these islands, and in what direction? (e.g., “Number of species on these islands is positively associated with X, Y but negatively associated with Z…”)
2. Similar to last week, let’s produce a Q-Q plot to test for normality of residuals. Paste it here. Does this model meet the assumption of normally-distributed errors? As a reminder, use this code:

qqnorm(residuals(MODEL\_NAME), main= "")

qqline(residuals(MODEL\_NAME))

1. Let’s do a Box-Cox transformation to transform our y-variable (Species) such that we can try to meet this assumption. Remember, we transform our y-variable using the equation (Y ^ lambda – 1) / lambda, but we need to know lambda first. Ensure that you have the MASS package installed.

boxcox <- boxcox(MODEL\_NAME,

lambda = seq(-1, 1, by=.1),

plotit = TRUE,

xlab = expression(lambda),

ylab = "log-Likelihood")

#the below line tells you what value of x is associated with the maximum value of y on that curve

lambda <- boxcox$x[which.max(boxcox$y)]

What is lambda?

1. Let’s transform our y-variable using this new lambda using

gala$Y\_TRANSFORM <- (gala$Species ^ lambda - 1) / lambda

Now re-run our model using this new y-variable and tell me what those coefficients are:

model\_new <- lm(Y\_TRANSFORM ~ Area + Nearest + Scruz + Adjacent, data=gala)

Intercept:

Area:

Nearest:

Scruz:

Adjacent:

1. FINALLY, let’s make a new Q-Q plot and see if we have normally distributed errors. Paste it here.
2. Do we now conform to this assumption? Should we move on in a peaceful state of bliss, or should we have checked and made sure that it conforms to the other two assumptions of independent errors and non-constant variance before even doing this transformation? [Note that I didn’t include all these steps for the sake of demonstrating correcting for this assumption in particular]