STAT 420: Homework 10

Fall 2020, D. Unger

Due: Tuesday, November 17 by 11:30 PM CT

Contents

Directions

Students are encouraged to work together on homework. However, sharing, copying or providing any part of a homework solution or code is an infraction of the University's rules on Academic Integrity. Any violation will be punished as severely as possible.

Final submissions must be uploaded to our Compass 2g site on the Homework page. No email, hardcopy, or late submissions will be accepted.

- Your assignment must be submitted through the submission link on **Compass 2g.** You are required to attach one `.zip` file, named `hw09_yourNetID.zip`, which contains:
  - Your RMarkdown file which should be saved as `hw10_yourNetID.Rmd`. For example `hw10_dunger.Rmd`.
  - The result of knitting your RMarkdown file as `hw10_yourNetID.html`. For example `hw10_dunger.html`.
  - Any raw data supplied by me for the assignment. For example `nutrition.csv`.
- Your resulting `.html` file will be considered a "report" which is the material that will determine the majority of your grade. Be sure to visibly include all `R` code and output that is relevant to answering the exercises. (You do not need to include irrelevant code you tried that resulted in error or did not answer the question correctly.)
- You are granted an unlimited number of submissions, but only the last submission *before* the deadline will be viewed and graded.
- If you use this `.Rmd` file as a template, be sure to remove the directions section. Consider removing `eval = FALSE` from any code chunks provided in the template, if you would like to run that code as part of your assignment.
- Your `.Rmd` file should be written such that, if it is placed in a folder with any data your are asked to import, it will knit properly without modification.

- Unless otherwise stated, you may use `R` for each of the exercises.
- Be sure to read each exercise carefully!
- Include your Name and NetID in the final document, not only in your filenames.

Assignment

Exercise 1 (TV Is Healthy?)

For this exercise we will use the `tvdoctor` data, which can be found in the `faraway` package. After loading the `faraway` package, use `?tvdoctor` to learn about this dataset.

```
library(faraway)
```

(a) Fit a simple linear regression with `life` as the response and `tv` as the predictor. Plot a scatterplot and add the fitting line. Check the assumptions of this model.

(b) Fit higher order polynomial models of degree 3, 5, and 7. For each, plot a fitted versus residuals plot and comment on the constant variance assumption. Based on those plots, which of these three models do you think are acceptable? Use a statistical test(s) to compare the models you just chose. Based on the test, which is preferred? Check the normality assumption of this model. Identify any influential observations of this model.

Exercise 2 (Brains)

The data set `mammals` from the `MASS` package contains the average body weight in kilograms (x) and the average brain weight in grams (y) for 62 species of land mammals. Use `?mammals` to learn more.

```
library(MASS)
```

(a) What are the smallest and largest body weights in the dataset?

(b) What are the smallest and largest brain weights in the dataset?

(c) Plot average brain weight (y) versus average body weight (x).

(d) Fit a linear model with `brain` as the response and `body` as the predictor. Test for significance of regression. Do you think this is an appropriate model?

Recall, *the log rule*: if the values of a variable range over more than one order of magnitude and the variable is strictly positive, then replacing the variable by its logarithm is likely to be helpful.

(e) Since the body weights do range over more than one order of magnitude and are strictly positive, we will use log(body weight) as our *predictor*, with no further justification. Use the Box-Cox method to verify that log(brain weight) is then a "recommended" transformation of the *response* variable. That is, verify that $\lambda = 0$ is among the "recommended" values of λ when considering,

$$g_\lambda(y) = \beta_0 + \beta_1 \log(\text{body weight}) + \epsilon$$

Please include the relevant plot in your results, using an appropriate zoom onto the relevant values.

(f) Fit the model justified in part **(e)**. That is, fit a model with log(brain weight) as the response and log(body weight) as a predictor. Plot log(brain weight) versus log(body weight) and add the regression line to the plot. Does a linear relationship seem to be appropriate here?

(g) Use a Q-Q plot to check the normality of the errors for the model fit in part **(f)**.

(h) Use the model from part **(f)** to predict the brain weight of a male Pikachu which, has a body weight of 13.4 pounds. (Pikachu would be mammals, right?) Construct a 99% prediction interval.

Exercise 3 (EPA Emissions Data, Redux)

For this exercise we will again use the data stored in `epa2015.csv`. It contains detailed descriptions of 4,411 vehicles manufactured in 2015 that were used for fuel economy testing as performed by the Environment Protection Agency.

(a) Recall the model we had finished with last time:

```
epa2015 = read.csv("epa2015.csv")
epa2015$type = as.factor(epa2015$type)
co2_int = lm(CO2 ~ horse * type, data = epa2015)
```

Which looked like this:

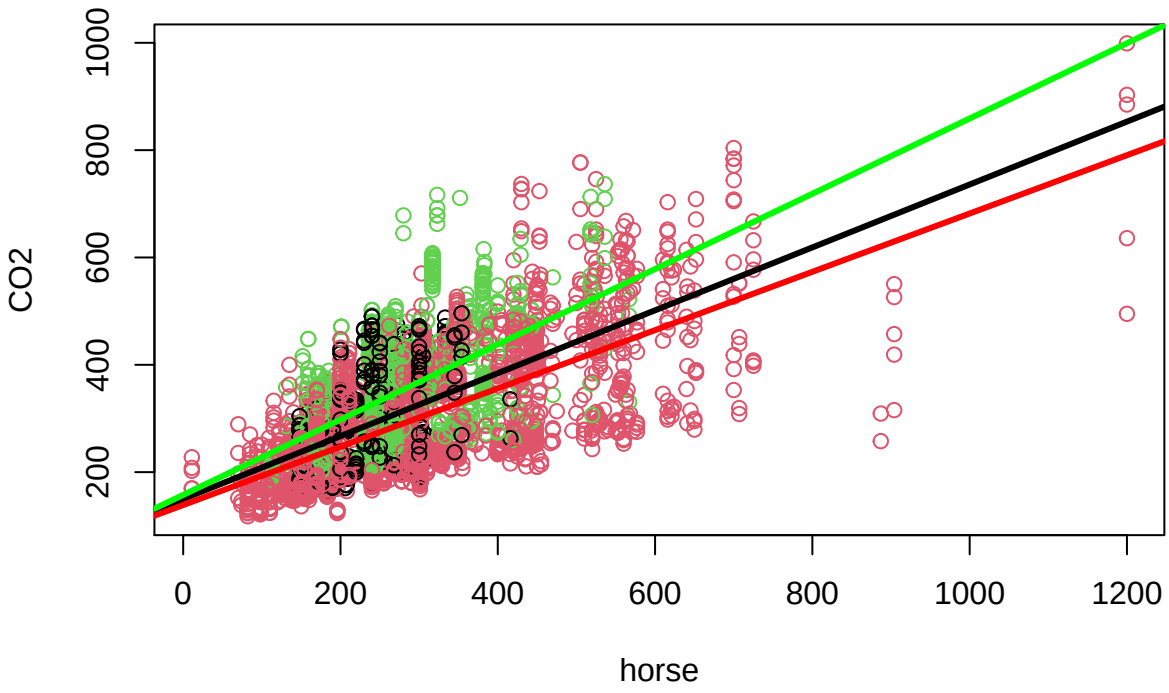
```
plot(CO2 ~ horse, data = epa2015, col = type)

int_coef = summary(co2_int)$coef[,1]

int_both    = int_coef[1]
int_car     = int_coef[1] + int_coef[3]
int_truck   = int_coef[1] + int_coef[4]

slope_both  = int_coef[2]
slope_car   = int_coef[2] + int_coef[5]
slope_truck = int_coef[2] + int_coef[6]

abline(int_both, slope_both, lwd = 3, col = "black")
abline(int_car, slope_car, lwd = 3, col = "red")
abline(int_truck, slope_truck, lwd = 3, col = "green")
```



Create a fitted vs residuals plot for this model. Do you believe the constant variance assumption has been violated?

(b) Fit the same model as **(a)** but with a logged response. Create a fitted vs residuals plot for this model. Compare to the previous. Do you believe the constant variance assumption has been violated? Any other assumptions?

(c) Fit a model that has all of the terms from the model in **(b)** as well as a quadratic term for `horse`. Use `log(CO2)` as the response. Create a fitted vs residuals plot for this model. Compare to the previous. Comment on model assumptions.

(d) Perform further analysis of the model fit in part **(c)**. Can you find any violations of assumptions?

Exercise 4 (Bigger Is Better?)

Consider the true model,

$$Y = 3 - 4x + \epsilon,$$

where $\epsilon \sim N(\mu = 0, \sigma = 9)$.

We can simulate observations from this model. We choose a sample size of 40.

```
n = 40
set.seed(114)
x = runif(n, 0 , 10)
y = 3 - 4 * x + rnorm(n, 0 , 3)
```

Consider two models, one small, one big. The small fits a SLR model. The big fits a polynomial model of degree 10.

```
fit_slr = lm(y ~ x)
fit_big = lm(y ~ poly(x, 10))
```

The big model has a smaller RMSE.

```
mean(resid(fit_slr) ^ 2)
```

```
## [1] 7.984
```

```
mean(resid(fit_big) ^ 2)
```

```
## [1] 7.067
```

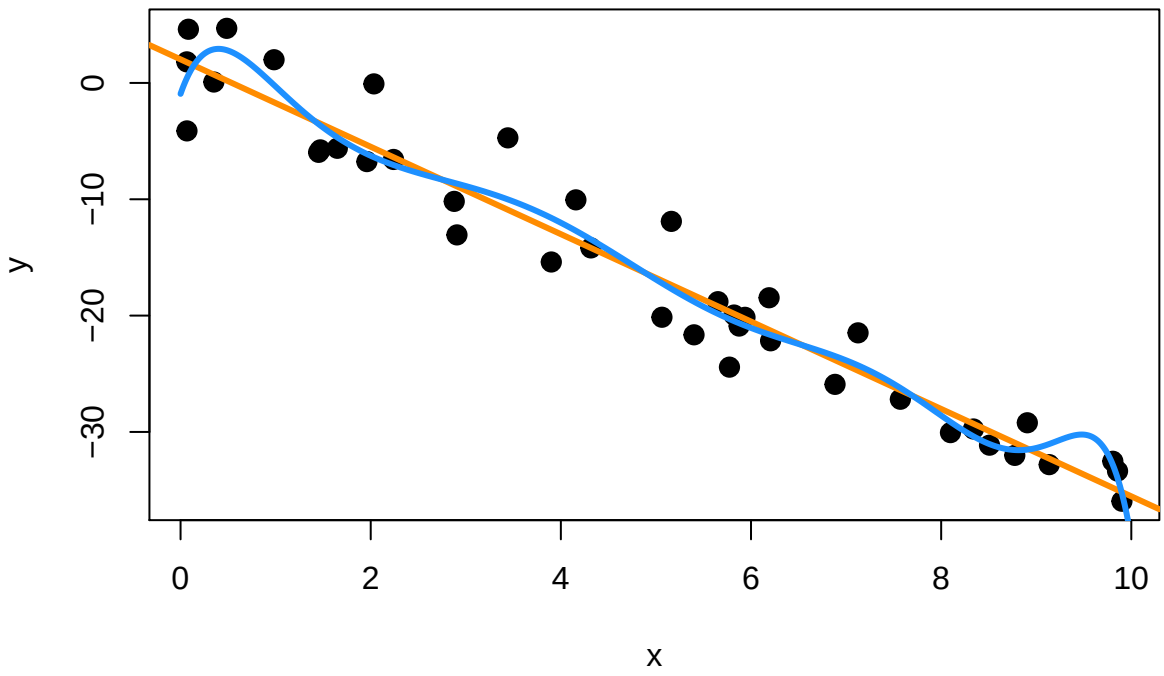
However, it is not significant when compared to the small.

```
anova(fit_slr, fit_big)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x
## Model 2: y ~ poly(x, 10)
##   Res.Df RSS Df Sum of Sq    F Pr(>F)
## 1     38 319
## 2     29 283  9      36.7 0.42   0.91
```

By plotting the data and adding the two models, we see the the degree 10 polynomial is *very* wiggly.

```
plot(x, y, pch = 20, cex = 2)
abline(fit_slr, col = "darkorange", lwd = 3)
lines(seq(0, 10, 0.01),
      predict(fit_big, newdata = data.frame(x = seq(0, 10, 0.01))),
      col = 'dodgerblue', lwd = 3)
```



(a) Use the following code after changing `birthday` to your birthday.

```
num_sims = 1000
rmse_slr = rep(0, num_sims)
rmse_big = rep(0, num_sims)
pval     = rep(0, num_sims)
birthday = 18760613
set.seed(birthday)
```

Repeat the above process, keeping `x` the same, then re-generating `y` and fitting the SLR and big models `1000` times. Each time, store the RMSE of each model, and the p-value for comparing the two. (In the appropriate variables defined above.)

(b) What proportion of the RMSEs of the SLR model are smaller than the big model?

(c) What proportion of the p-values are less than 0.05?

(d) Do you think bigger is better?