STAT 420: Homework 11

Fall 2020, D. Unger

Due: Tuesday, December 1 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 `hw11_yourNetID.Rmd`. For example `hw11_dunger.Rmd`.
  - The result of knitting your RMarkdown file as `hw11_yourNetID.html`. For example `hw11_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 (`longley` Macroeconomic Data)

The data set `longley` from the `faraway` package contains macroeconomic data for predicting employment.

```
library(faraway)
```

```
View(longley)
?longley
```

(a) Find the correlation between each of the variables in the dataset.

(b) Fit a model with `Employed` as the response and the remaining variables as predictors. Calculate the variance inflation factor (VIF) for each of the predictors. What is the largest VIF? Do any of the VIFs suggest multicollinearity?

(c) What proportion of the observed variation in `Population` is explained by a linear relationship with the other predictors?

(d) Calculate the partial correlation coefficient for `Population` and `Employed` **with the effects of the other predictors removed**.

(e) Fit a new model with `Employed` as the response and the predictors from the model in **(b)** that were significant. (Use $\alpha = 0.05$.) Calculate the variance inflation factor for each of the predictors. What is the largest VIF? Do any of the VIFs suggest multicollinearity?

(f) Use an F-test to compare the models in parts **(b)** and **(e)**. Report the following:

- The null hypothesis
- The test statistic
- The distribution of the test statistic under the null hypothesis
- The p-value
- A decision
- Which model you prefer, **(b)** or **(e)**

(g) Check the assumptions of the model chosen in part **(f)**. Do any assumptions appear to be violated?

Exercise 2 (`odor` Chemical Data)

Use the `odor` data from the `faraway` package for this question.

(a) Fit a complete second order model with `odor` as the response and the three other variables as predictors. That is, use each first order term, their two-way interactions, and the quadratic term for each of the predictors. Perform the significance of the regression test. Use a level of $\alpha = 0.10$. Report the following:

- The test statistic

- The distribution of the test statistic under the null hypothesis
- The p-value
- A decision

(b) Fit a model with the same response, but now excluding any interaction terms. So, include all linear and quadratic terms. Compare this model to the model in **(a)** using an appropriate test. Use a level of $\alpha = 0.10$. Report the following:

- The test statistic
- The distribution of the test statistic under the null hypothesis
- The p-value
- A decision

(c) Report the proportion of the observed variation of `odor` explained by the two previous models.

(d) Use adjusted R^2 to pick from the two models. Report both values. Does this decision match the decision made in part **(b)**?

Exercise 3 (`teengamb` Gambling Data)

The `teengamb` dataset from the `faraway` package contains data related to teenage gambling in Britain.

(a) Fit an additive model with `gamble` as the response and the other variables as predictors. Use backward AIC variable selection to determine a good model. When writing your final report, you may wish to use `trace = 0` inside of `step()` to minimize unneeded output. (This advice is also useful for future questions that use `step()`.)

(b) Use backward BIC variable selection to determine a good model.

(c) Use a statistical test to compare these two models. Use a level of $\alpha = 0.10$. Report the following:

- The test statistic
- The distribution of the test statistic under the null hypothesis
- The p-value
- A decision

(d) Fit a model with `gamble` as the response and the other variables as predictors with *all* possible interactions, up to and including a four-way interaction. Use backward AIC variable selection to determine a good model.

(e) Compare the values of adjusted R^2 for the each of the five previous models. Which model is the "best" model out of the five? Justify your answer.

Exercise 4 (`prostate` Data)

Using the `prostate` dataset from the `faraway` package, fit a model with `lpsa` as the response and the other variables as predictors. For this exercise only consider first order predictors.

(a) Find the model with the **best** AIC. Report the predictors that are used in the resulting model.

(b) Find the model with the **best** BIC. Report the predictors that are used in the resulting model.

(c) Find the model with the **best** adjusted R^2. Report the predictors that are used in the resulting model.

(d) Of the four models you just considered, some of which *may* be the same, which is the best for making predictions? Use leave-one-out-cross-validated MSE or RMSE to decide.

Exercise 5 (Goalies, Revisited)

(a) Use the data found in `goalies_cleaned.csv` to find a "good" model for wins, `W`. Use any methods seen in class. The model should reach a `Multiple R-squared` above `0.99` using fewer than 37 parameters. Hint: You may want to look into the ability to add many interactions quickly in `R`.