STAT 420: Homework 05

Fall 2020, D. Unger

Due: Tuesday, October 13 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 `hw05_yourNetID.zip`, which contains:
  - Your RMarkdown file which should be saved as `hw05_yourNetID.Rmd`. For example `hw05_dunger.Rmd`.
  - The result of knitting your RMarkdown file as `hw05_yourNetID.html`. For example `hw05_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 (Using `lm`)

For this exercise we will use the data stored in `nutrition.csv`. It contains the nutritional values per serving size for a large variety of foods as calculated by the USDA. It is a cleaned version totaling 5,138 observations and is current as of September 2015.

The variables in the dataset are:

- `ID`
- `Desc` - Short description of food
- `Water` - in grams
- `Calories`
- `Protein` - in grams
- `Fat` - in grams
- `Carbs` - Carbohydrates, in grams
- `Fiber` - in grams
- `Sugar` - in grams
- `Calcium` - in milligrams
- `Potassium` - in milligrams
- `Sodium` - in milligrams
- `VitaminC` - Vitamin C, in milligrams
- `Chol` - Cholesterol, in milligrams
- `Portion` - Description of standard serving size used in analysis

(a) Fit the following multiple linear regression model in `R`. Use `Calories` as the response and `Carbs`, `Fat`, and `Protein` as predictors.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i.$$

Here,

- Y_i is `Calories`.
- x_{i1} is `Carbs`.
- x_{i2} is `Fat`.
- x_{i3} is `Protein`.

Use an F-test to test the significance of the regression. Report the following:

- The null and alternative hypotheses
- The value of the test statistic
- The p-value of the test
- A statistical decision at $\alpha = 0.01$
- A conclusion in the context of the problem

When reporting these, you should explicitly state them in your document, not assume that a reader will find and interpret them from a large block of `R` output.

(b) Output only the estimated regression coefficients. Interpret all $\hat{\beta}_j$ coefficients in the context of the problem.

(c) Use your model to predict the amount of `Calories` in a Big Mac. According to McDonald's publicized nutrition facts, the Big Mac contains 47g of carbohydrates, 28g of fat, and 25g of protein.

(d) Calculate the standard deviation, s_y, for the observed values in the Calories variable. Report the value of s_e from your multiple regression model. Interpret both estimates in the context of this problem.

(e) Report the value of R^2 for the model. Interpret its meaning in the context of the problem.

(f) Calculate a 90% confidence interval for β_2. Give an interpretation of the interval in the context of the problem.

(g) Calculate a 95% confidence interval for β_0. Give an interpretation of the interval in the context of the problem.

(h) Use a 99% confidence interval to estimate the mean Calorie content of a small order of McDonald's french fries that has 30g of carbohydrates, 11g of fat, and 2g of protein. Interpret the interval in context.

(i) Use a 90% prediction interval to predict the Calorie content of new healthy menu item that has 11g of carbohydrates, 1.5g of fat, and 1g of protein. Interpret the interval in context.

Exercise 2 (More `lm`)

For this exercise we will again use the nutrition data.

(a) Fit a model with Calories as the response and `Carbs`, `Sodium`, `Fat`, and `Protein` as predictors. Use an F-test to test the significance of the regression. Report the following:

- The null and alternative hypotheses
- The value of the test statistic
- The p-value of the test
- A statistical decision at $\alpha = 0.01$
- A conclusion in the context of the problem

(b) For each of the predictors in part **(a)**, perform a t-test for the significance of its regression coefficient. Report the following for each:

- The null and alternative hypotheses
- The value of the test statistic
- The p-value of the test
- A statistical decision at $\alpha = 0.01$

(c) Based on your results in part **(b)**, do you still prefer the model in part **(a)**, or is there instead a model with three predictors that you prefer? Briefly explain.

Exercise 3 (Comparing Models)

For this exercise we will use the data stored in `goalies_cleaned.csv`. It contains career data for 462 players in the National Hockey League who played goaltender at some point up to and including the 2014 - 2015 season. The variables in the dataset are:

- `W` - Wins
- `GA` - Goals Against
- `SA` - Shots Against
- `SV` - Saves
- `SV_PCT` - Save Percentage
- `GAA` - Goals Against Average
- `SO` - Shutouts
- `MIN` - Minutes
- `PIM` - Penalties in Minutes

(a) Fit a multiple linear regression model with Wins as the response and all other variables as the predictors.

Use an F-test to test the significance of the regression. Report the following:

- The null and alternative hypotheses
- The value of the test statistic
- The p-value of the test
- A statistical decision at $\alpha = 0.10$
- A conclusion in the context of the problem

When reporting these, you should explicitly state them in your document, not assume that a reader will find and interpret them from a large block of `R` output.

(b) Calculate the RMSE of this full model. Report the residual standard error of this full model. What is the relationship of these two values?

Recall, we have defined RMSE as,

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

(c) Fit a model with Wins as the response and with Goals Against, Goals Against Average, Saves, and Save Percentage as the predictors. Calculate the RMSE of this model.

(d) Fit a model with Wins as the response and with Goals Against Average and Save Percentage as the predictors. Calculate the RMSE of this model.

(e) Based on the previous three models, which model is most helpful for predicting wins? Briefly explain.

(f) Conduct an ANOVA F-test comparing the models in parts **(c)** and **(d)**. Report the following:

- The null and alternative hypotheses
- The value of the test statistic
- The p-value of the test
- A statistical decision at $\alpha = 0.10$
- A conclusion in the context of the problem

When reporting these, you should explicitly state them in your document, not assume that a reader will find and interpret them from a large block of `R` output.

Exercise 4 (Regression without `lm`)

For this exercise use the `prostate` dataset from the `faraway` package. Use `?prosate` to learn about the dataset. The goal of this exercise is to fit a model with `lpsa` as the response and the remaining variables as predictors.

(a) Obtain the estimated regression coefficients **without** the use of `lm()` or any other built-in functions for regression. That is, you should use only matrix operations. Store the results in a vector `beta_hat_no_lm`. To ensure this is a vector, you may need to use `as.vector()`. Return this vector as well as the results of `sum(beta_hat_no_lm)`.

(b) Obtain the estimated regression coefficients **with** the use of `lm()`. Store the results in a vector `beta_hat_lm`. To ensure this is a vector, you may need to use `as.vector()`. Return this vector as well as the results of `sum(beta_hat_lm)`.

(c) Use the `all.equal()` function to verify that the results are the same. You may need to remove the names of one of the vectors. The `as.vector()` function will do this as a side effect, or you can directly use `unname()`.

(d) Calculate s_e without the use of `lm()`. That is, continue with your results from **(a)** and perform additional matrix operations to obtain the result. Output this result. Also, verify that this result is the same as the result obtained from `lm()`.

(e) Calculate R^2 without the use of `lm()`. That is, continue with your results from **(a)** and **(d)** and perform additional operations to obtain the result. Output this result. Also, verify that this result is the same as the result obtained from `lm()`.