

# STAT 420: Homework 06

Fall 2020, D. Unger

Due: Tuesday, October 20 by 11:30 PM CT

## Contents

<b>Directions</b>	<b>1</b>
<b>Assignment</b>	<b>2</b>
Exercise 1 (Regression for Explanation) . . . . .	2
Exercise 2 (Regression for Prediction) . . . . .	2
Exercise 3 (Simulating Multiple Regression) . . . . .	3

## 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 **hw06\_yourNetID.zip**, which contains:
  - Your RMarkdown file which should be saved as **hw06\_yourNetID.Rmd**. For example **hw06\_dunger.Rmd**.
  - The result of knitting your RMarkdown file as **hw06\_yourNetID.html**. For example **hw06\_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 (Regression for Explanation)

For this exercise use the `prostate` dataset from the `faraway` package. Use `?prostate` to learn about the dataset. The goal of this exercise is to find a model that is useful for **explaining** the response `lpsa`.

Fit a total of five models.

- One must use all possible predictors.
- One must use only `lcavol` as a predictor.
- The remaining three you must choose. The models you choose must be picked in a way such that for any two of the five models, one is nested inside the other.

Argue that one of the five models is the best among them for explaining the response. Use appropriate methods and justify your answer.

## Exercise 2 (Regression for Prediction)

For this exercise use the `Boston` dataset from the `MASS` package. Use `?Boston` to learn about the dataset. The goal of this exercise is to find a model that is useful for **predicting** the response `medv`.

When evaluating a model for prediction, we often look at RMSE. However, if we both fit the model with all the data, as well as evaluate RMSE using all the data, we're essentially cheating. We'd like to use RMSE as a measure of how well the model will predict on *unseen* data. If you haven't already noticed, the way we had been using RMSE resulted in RMSE decreasing as models became larger.

To correct for this, we will only use a portion of the data to fit the model, then we will use leftover data to evaluate the model. We will call these datasets **train** (for fitting) and **test** (for evaluating). The definition of RMSE will stay the same

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

where

- $y_i$  are the actual values of the response for the given data
- $\hat{y}_i$  are the predicted values using the fitted model and the predictors from the data

However we will now evaluate it on both the **train** set and the **test** set separately. So each model you fit will have a **train** RMSE and a **test** RMSE. When calculating **test** RMSE, the predicted values will be found by predicting the response using the **test** data with the model fit using the **train** data. *Test data should never be used to fit a model.*

- Train RMSE: Model fit with train data. Evaluate on **train** data.
- Test RMSE: Model fit with train data. Evaluate on **test** data.

Set a seed of 42 and then split the `Boston` data into two datasets, one called `train_data` and one called `test_data`. The `train_data` dataframe should contain 400 randomly chosen observations. `test_data` will contain the remaining observations. Hint: consider the following code:

```
library(MASS)
set.seed(114)
train_index = sample(1:nrow(Boston), 400)
```

Fit a total of five models using the training data.

- One must use all possible predictors.
- One must use only `crim` as a predictor.
- The remaining three you can pick to be anything you like. One of these should be the best of the five for predicting the response.

For each model report the **train** and **test** RMSE. Argue that one of your models is the best for predicting the response.

### Exercise 3 (Simulating Multiple Regression)

For this exercise we will simulate data from the following model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i$$

Where  $\epsilon_i \sim N(0, \sigma^2)$ . Also, the parameters are known to be:

- $\beta_0 = 2$
- $\beta_1 = 3$
- $\beta_2 = 4$
- $\beta_3 = 0$
- $\beta_4 = 1$
- $\sigma^2 = 16$

We will use samples of size  $n = 25$ .

We will verify the distribution of  $\hat{\beta}_1$  as well as investigate some hypothesis tests.

(a) We will first generate the  $X$  matrix and data frame that will be used throughout the exercise. Create the following 9 variables:

- `x0`: a vector of length `n` that contains all 1
- `x1`: a vector of length `n` that is randomly drawn from a uniform distribution between 0 and 10
- `x2`: a vector of length `n` that is randomly drawn from a uniform distribution between 0 and 10
- `x3`: a vector of length `n` that is randomly drawn from a uniform distribution between 0 and 10
- `x4`: a vector of length `n` that is randomly drawn from a uniform distribution between 0 and 10
- `X`: a matrix that contains `x0`, `x1`, `x2`, `x3`, `x4` as its columns
- `C`: the  $C$  matrix that is defined as  $(X^T X)^{-1}$
- `y`: a vector of length `n` that contains all 0
- `ex_4_data`: a data frame that stores `y` and the **four** predictor variables. `y` is currently a placeholder which we will update during the simulation

Report the diagonal of `C` as well as the 10th row of `ex_4_data`. For this exercise we will use the seed 42.

```
set.seed(114)
n = 25
```

(b) Create three vectors of length 1500 that will store results from the simulation in part (c). Call them `beta_hat_1`, `beta_2_pval`, and `beta_3_pval`.

(c) Simulate 1500 samples of size `n = 25` from the model above. Each time update the `y` value of `ex_4_data`. Then use `lm()` to fit a multiple regression model. Each time store:

- The value of  $\hat{\beta}_1$  in `beta_hat_1`
- The p-value for the two-sided test of  $\beta_2 = 0$  in `beta_2_pval`
- The p-value for the two-sided test of  $\beta_3 = 0$  in `beta_3_pval`

(d) Based on the known values of  $X$ , what is the true distribution of  $\hat{\beta}_1$ ?

(e) Calculate the mean and variance of `beta_hat_1`. Are they close to what we would expect? Plot a histogram of `beta_hat_1`. Add a curve for the true distribution of  $\hat{\beta}_1$ . Does the curve seem to match the histogram?

(f) What proportion of the p-values stored in `beta_3_pval` are less than 0.05? Is this what you would expect?

(g) What proportion of the p-values stored in `beta_2_pval` are less than 0.05? Is this what you would expect?