Problem Set 3: Least Squares Regression

Stat 154, Fall 2017, Prof. Sanchez

Due date: Th Oct-5 (before midnight)

Problem 1 (10 pts)

Consider a simple linear regression model with OLS fitted values given by:

$$\hat{y}_i = b_0 + b_1 x_i$$

Show that the sum of residuals equals zero, that is: $\sum_{i=1}^{n} e_i = 0$

Problem 2 (30 pts)

We examine a response variable Y in terms of two predictors X and Z. There are n observations. Let $\mathbf{X}$ be a matrix formed by a constant term of $\mathbf{1}$, and the vectors $\mathbf{x}$ and $\mathbf{z}$. Consider the cross-product matrix $\mathbf{X}^\mathsf{T}\mathbf{X}$ given below:

$$\mathbf{X}^\mathsf{T}\mathbf{X} = \begin{bmatrix} 30 & 0 & 0 \\ ? & 10 & 7 \\ ? & ? & 15 \end{bmatrix}$$

a. Complete the missing values denoted by "?", and determine the value of n? (10 pts)

b. Calculate the linear correlation coefficient between X and Z. (10 pts)

c. If the OLS regression equation is: $\hat{y}_i = -2 + x_i + 2z_i$, What is the value of $\bar{y}$?

d. If the residual sum of squares (RSS) is 12, What is the value of R^2?

Problem 3 (30 pts)

In this exercise you will create some simulated data and will fit simple linear regression models to it. Make sure to use `set.seed(1)` prior to starting part (a) to ensure consistent results.

- Parts (a) - (g) are worth 10 pts
- Part (h) is worth 10 pts
- Part (i) is worth 10 pts

a. Using the `rnorm()` function, create a vector, `x`, containing 100 observations drawn from a $N(0, 1)$ distribution. This represents a feature, X.

b. Using the `rnorm()` function, create a vector, `eps`, containing 100 observations drawn from a $N(0, 0.25)$ distribution i.e. a normal distribution with mean zero and variance 0.25.

c. Using `x` and `eps`, generate a vector `y` according to the model:

$$Y = -1 + 0.5X + \epsilon$$

d. Create a scatterplot displaying the relationship between `x` and `y`. Comment on what you observe.

e. Fit a least squares linear model to predict `y` using `x`. Comment on the model obtained. How do $\hat{\beta}_0$ and $\hat{\beta}_1$ compare to β_0 and β_1?

f. Display the least squares line on the scatterplot obtained in (d). Draw the theoretical regression line on the plot, in a different color. Use the `legend()` command to create an appropriate legend.

g. Now fit a polynomial regression model that predicts `y` using `x` and `x^2`. Is there evidence that the quadratic term improves the model fit? Explain your answer.

h. Repeat (a)-(f) after modifying the data generation process in such a way that there is *less* noise in the data. The theoretical model $Y = -1 + 0.5X + \epsilon$ should remain the same. Describe your results.

i. Repeat (a)-(f) after modifying the data generation process in such a way that there is *more* noise in the data. The theoretical model $Y = -1 + 0.5X + \epsilon$ should remain the same. Describe your results.

Problem 4 (10 pts)

Write a function `ols_fit` that computes the (ordinary) least squares solution, via QR decomposition, given an input model matrix $\mathbf{X}$ and a response vector $\mathbf{y}$. The function `ols_fit` should take two arguments:

- `X`: a model (or design) matrix of predictors
- `y`: a vector for the response variable

The output should be a `list` with elements:

- `coefficients`: estimated regression coefficients
- `y_values`: vector of observed values
- `fitted_values`: vector of fitted (hat) values
- `residuals`: vector of residuals
- `n`: number of observations n, and

- `q`: number of columns in the model matrix `X`

Unless all the variables (i.e. the predictors and response) are mean-centered, `X` will be a matrix of dimension $n \times (p+1)$, where n is the number of observations, and p is the number of predictors. In other words, `X` should include the column of 1's to account for the intercept term, except for when all variables are mean-centered.

Consider the data frame `mtcars` that comes in R. Assume a vector `y` for the response variable `mpg`, and the following matrix `X` formed by a column of 1's (intercept), `disp`, and `hp`:

	intercept	disp	hp
Mazda RX4	1	160	110
Mazda RX4 Wag	1	160	110
Datsun 710	1	108	93
⋮	⋮	⋮	⋮
Maserati Bora	1	301	335
Volvo 142E	1	121	109

You should be able to call `ols_fit` as follows:

```
fit <- ols_fit(X, y)
names(fit)
```

```
## [1] "coefficients"  "y_values"      "fitted_values" "residuals"
## [5] "n"             "q"
```

```
fit$coefficients
```

```
##              [,1]
## [1,] 30.73590425
## [2,] -0.03034628
## [3,] -0.02484008
```

```
summary(fit$fitted_values)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.32   15.16   21.64   20.09   24.70   27.15
```

```
summary(fit$residuals)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -4.7950 -2.3040 -0.8246  0.0000  1.8580  6.9360
```

Test your function with the example above (using the `mtcars` data set), and confirm you get the same `coefficients`, and the same `summary()` statistics for `fitted_values` and `residuals`.

Problem 5 (10 pts)

Write auxiliary functions `R2()` and `RSE()` to compute the coefficient of determination R^2, and the Residual Standard Error (RSE), respectively.

a. The function `R2` computes the coefficient of determination R^2:

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

The function `R2()` should take the output of `ols_fit()` as the only input. This means that you should be able to call `R2()` like:

```
fit <- ols_fit(X, y)
R2(fit)
```

```
## [1] 0.7482402
```

b. The function `RSE` computes the Residual Standard Error given by the formula:

$$\text{RSE} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n - p - 1}}$$

where p is the number of predictors.

Instead of using p, you can use:

$$\text{RSE} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n - q}}$$

where q is the number of columns in the model matrix `X`, returned by `ols_fit()`.

The function `RSE()` should take the output of `ols_fit()` as the only input. This means that you should be able to call `RSE()` like:

```
fit <- ols_fit(X, y)
RSE(fit)
```

```
## [1] 3.126601
```

Problem 6 (20 pts)

Consider the dataset `prostate` available in the `data/` folder of the course github repository. This dataset comes from a study on 97 men with prostate cancer who were due to receive a radical prostatectomy. Use your `ols_fit()` and `R2()` functions to:

- Fit a model with `lpsa` as the response and `lcavol`as the predictor. Record the residual standard error and the R^2.
- Add `lweight`, `svi`, `lbph`, `age`, `lcp`, `pgg45` and `gleason` to the model one at a time. For each model record the residual standard error and the R^2.
- Make plots of the trends for each of these two statistics.

Problem 7 (20 pts)

This question involves the use of multiple linear regression on the `Auto` data set. The associated file is in the website of the textbook "An Introduction to Statistical Learning":

http://www-bcf.usc.edu/~gareth/ISL/Auto.data

a. Produce a scatterplot matrix which inlcudes all of the variables in the data set.

b. Compute the matrix of correlations between the variables using the function `cor()`. You will need to exclude the `name` variable, which is qualitative.

c. Use the `lm()` function to perform a multiple linear regression with `mpg` as the response and all other variables except `name` as the predictors. Use the `summary()` function to print the results. Comment on the output. For instance:

   - Is there a relationship between the predictors and the response?
   - Which predictors appear to have a statistically significant relationship to the response?
   - What does the coefficient for the `year` variable suggest?

d. Use the `plot()` function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusual high leverage?

e. Use the `*` and `:` symbols to fit a linear regression model with an interaction effect. Does the interaction that you chose appear to be statistically significant?

f. Fit another linear regression model by trying a few different transformations of the variables, such as $log(X)$, $\sqrt{X}$, X^2. Comment on your findings.