Assignment 1

CS 750/850 Machine Learning

Landon Buell

29 January 2020

- **Due**: Monday 2/3 at 11:59PM
- **Submisssion**: Turn in as a **PDF** and the **source code** (R,Rmd,py,ipynb) on MyCourses
- **Questions**: Piazza and Office hours: *Marek*: Wed 1:30-3:00pm, Soheil: Mon 2-4pm, Xihong: Thu 1:30-3:30pm
- **Extra credit**: Especially good questions or helpful answers on Piazza regarding the assignment earn up to 5 points extra credit towards the assignment grade.

The instructions are geared towards R users. The comments in [P: xxx] are meant as hints to Python users.

Feel free to achieve the results using commands other than the ones mentioned. R, especially, shines when it comes to processing and visualization of structured data, but you would not be able to tell from the ISL book. It uses the oldest and simplest (and ugliest) subset of R. I recommend checking out `dplyr` for data processing [3,4] and GGPlot [1,4] for plotting.

Problem 1 [25%]

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)` [P: `np.random.seed(1)`] prior to starting part (1) to ensure consistent results.

1. Using the `rnorm()` [P: `np.random.normal`] function, create a vector, `x`, containing 100 observations drawn from a $\mathcal{N}(0, 3)$ distribution (Normal distribution with the mean 0 and the **standard deviation** $\sqrt{3}$). This represents a feature, X.

```
set.seed(1)
x <- rnorm(n=100,mean=0,sd=sqrt(3))
```

2. Using the `rnorm()` function, create a vector, `eps`, containing 100 observations drawn from a $\mathcal{N}(0, 0.5)$ distribution i.e. a normal distribution with mean zero and standard deviation $\sqrt{0.5}$.

```
eps <- rnorm(n=100,mean=0,sd=sqrt(0.5))
```

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

$$Y = -2 + 0.6X + \epsilon$$

What is the length (number of elements) of `y`? What are the values of β_0, β_1 in the equation above (intercept and slope)?

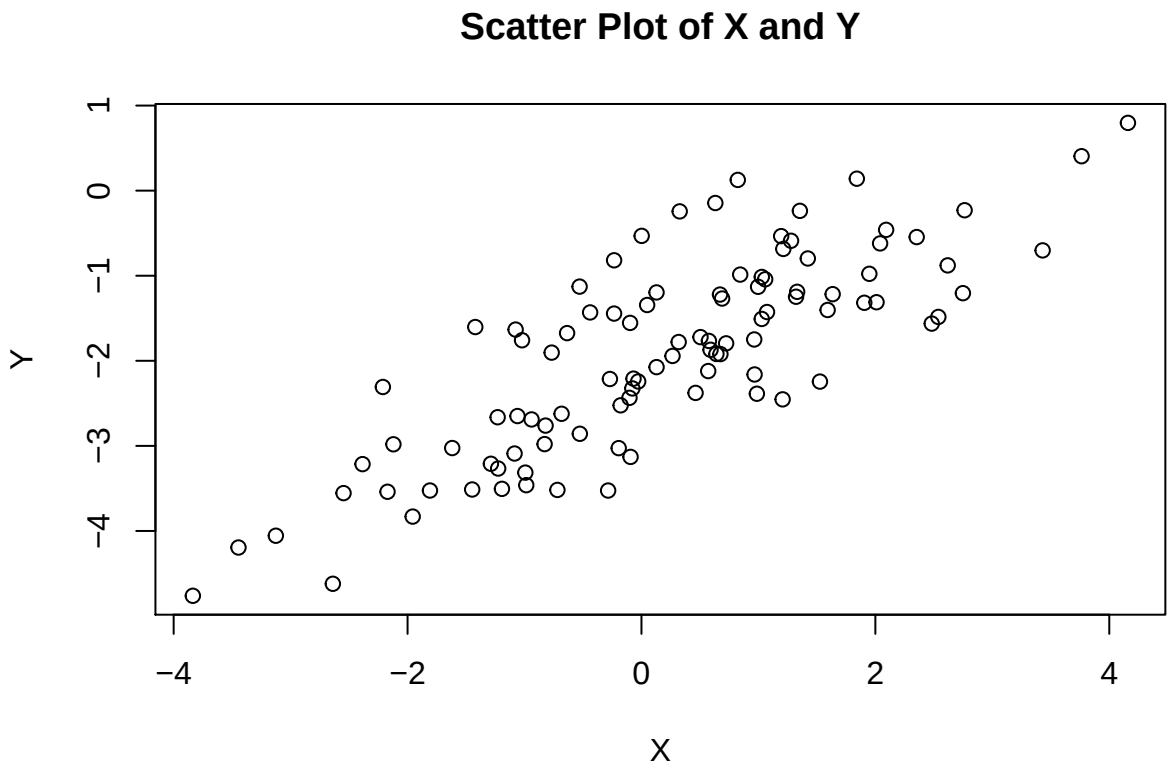
```r
Y <- -2 + 0.6*x + eps
message("Elements in Y:",length(Y))
```

```
## Elements in Y:100
```

In this model above, the intercept is given by $\beta_0 = -2$ and the slope is given by $\beta_1 = +0.6$.

4. Create a scatterplot displaying the relationship between `x` and `y`. Comment on what you observe. [P: see [2]]

```r
plot(x,Y,main='Scatter Plot of X and Y',
     xlab='X',ylab='Y')
```



The data seemes to be losely correlated, following a roughly linear relationship. (As I would expect it to because of our linear model)

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

```r
lin_fit <- lm(Y~x) # lm = linear model
lin_fit
```

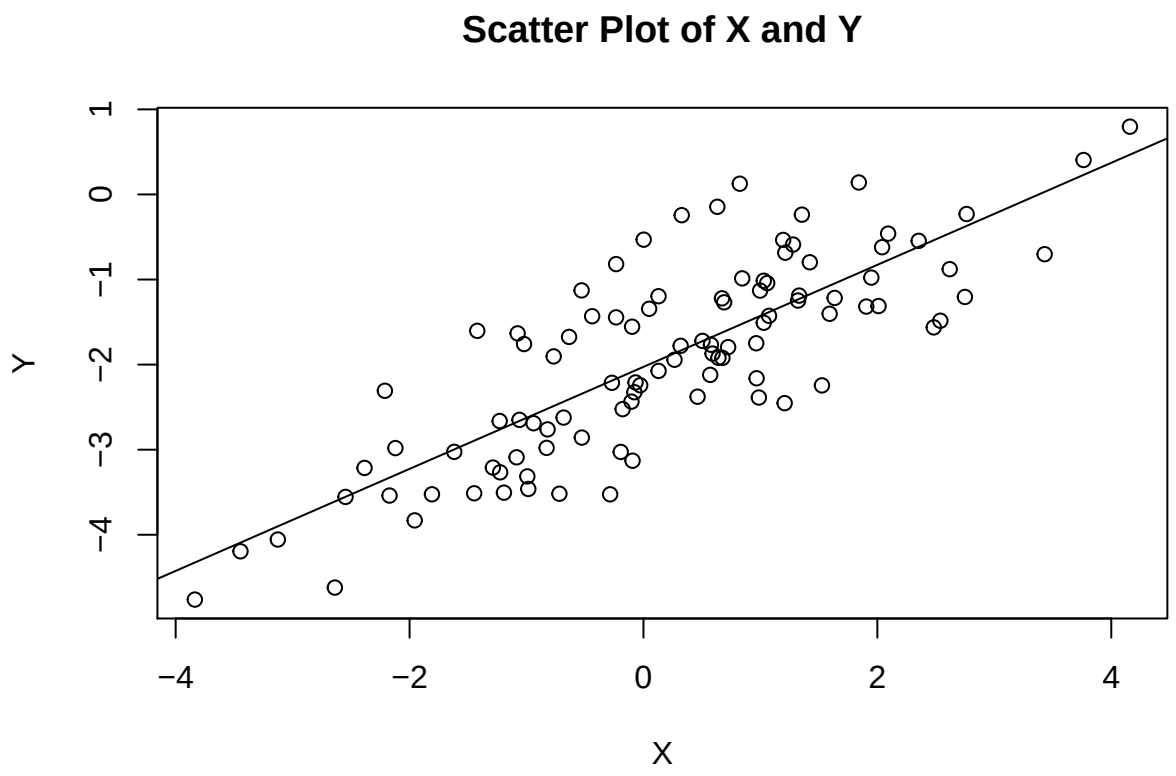
```
##
## Call:
```

```
## lm(formula = Y ~ x)
##
## Coefficients:
## (Intercept)            x
##     -2.0267       0.5996
```

The regression line is a linear relationsip (1st degree polynomial) as anticipated. It produces a fit such that the sum of the vertical distances between each data point and the line has been minimized.
In our equation, we set the intercept and slope to be: $\beta_0 = -2$ and $\beta_1 = +0.6$. repectivly. In the fit model, the linear model found the approximations of the slpoe and intercept to be roughly: $\hat{\beta_0} = -2.0267$ and $\hat{\beta_1} = +0.5996$ as shown by the coeffefcients in the fit variable above.

6. Display the least squares line on the scatterplot obtained in 4.

```
plot(x,Y,main='Scatter Plot of X and Y',
     xlab='X',ylab='Y')
abline(lin_fit)
```



7. 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.

```
quad_fit = lm(Y~poly(x,2))
quad_fit
```

```
##
```

```
## Call:
## lm(formula = Y ~ poly(x, 2))
##
## Coefficients:
## (Intercept)  poly(x, 2)1  poly(x, 2)2
##     -1.9136       9.2809      -0.9504
```

The quadratic model seems to be much further removed from the data set. Ideally, I would have expected the intercept and slope coefficients to remain fairly close to the repesctive linear model and then the quadratic coefficient to be close to zero. TOo me this would indicate that the quadratic term was not important, but still allow for the retention of the properties of the linear model. Instead, the addition of another polynomial term completely changes the best fit model to a point where is differs greatly from the expected value. Now we have the predictions: $\hat{\beta}_0 = -19.136$, $\hat{\beta}_1 = 9.2809$ and $\hat{\beta}_2 = -0.9504$.

Optional Problem O1 [30%]

This problem can be substituted for Problem 1 above, for up to 5 points extra credit. At most one of the problems 1 and O1 will be considered.

Read Chapter 1 and solve Exercises *1.6* and *1.10* in [Bishop, C. M. (2006). Pattern Recognition and Machine Learning].

Problem 2 [25%]

Read through Section 2.3 in ISL. Load the `Auto` data set and *make sure to remove missing values from the data*. Then answer the following questions:

```
# Below is from James, pg. 49
autodata <- read.csv(file='Auto.csv',header=T,na.strings="?")
#fix(autodata) # I'm not sure what "fix()" does
dim(autodata) # dimensions of array (nrows,ncols)
```

```
## [1] 397   9
```

```
autodata = na.omit(autodata) # eliminate "na's"
dim(autodata) # dimensions of array (nrows,ncols)
```

```
## [1] 392   9
```

```
# check variable names:
names(autodata)
```

```
## [1] "mpg"          "cylinders"    "displacement" "horsepower"   "weight"
## [6] "acceleration" "year"         "origin"       "name"
```

1. Which predictors are *quantitative* and which ones are *qualitative*?

```
summary(autodata)
```

```
##       mpg          cylinders      displacement     horsepower        weight
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613
##  1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225
##  Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Median :2804
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140
##
##   acceleration        year           origin                      name
##  Min.   : 8.00   Min.   :70.00   Min.   :1.000   amc matador       :  5
##  1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000   ford pinto        :  5
##  Median :15.50   Median :76.00   Median :1.000   toyota corolla    :  5
##  Mean   :15.54   Mean   :75.98   Mean   :1.577   amc gremlin       :  4
##  3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000   amc hornet        :  4
##  Max.   :24.80   Max.   :82.00   Max.   :3.000   chevrolet chevette:  4
##                                                  (Other)           :365
```

The *name*, *origin*, and perhaps even the *year* are all quantitative. The *MPG*, *Cylinders*, *Displacement*, *horsepower*, *weight* and *acceleration* categories are all made up of quantitative data. 2. What is the range, mean, and standard deviation of each predictor? Use `range()` [`pandas.DataFrame.min` and `max`] function.

```
print("Mins & Maxes:")
```

```
## [1] "Mins & Maxes:"
```

3. Investigate the predictors graphically using plots. Create plots highlighting relationships between predictors. See [1] for a ggplot cheatsheet.
4. Compute the matrix of correlations between variables using the function `cor()` [P: `pandas.DataFrame.corr`]. Exclude the `name` variable.
5. Use the `lm()` function to perform a multiple linear regression with `mpg` as the response. [P: using `rpy` package is acceptable] Exclude `name` as a predictor, since it is qualitative. Briefly comment on the output: What is the relationship between the predictors? What does the coefficient for `year` variable suggest?
6. Use the symbols `*` and `:` to fit linear regression models with interaction effects. What do you observe?
7. Try a few different transformations of variables, such as $\log(X)$, $\sqrt{X}$, X^2. What do you observe?

Problem 3 [25%]

Using equation (3.4) in ISL, argue that in the case of simple linear regression, the least squares line always passes through the point $(\bar{x}, \bar{y})$. Equation (3.4) in ISL:

$$\bar{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Problem 4 [25%]

It is claimed in the ISL book that in the case of simple linear regression of Y onto X, the R^2 statistic (3.17) is equal to the square of the correlation between X and Y (3.18). Prove that this is the case. For simplicity, you may assume that $\bar{x} = \bar{y} = 0$.

References

Each reference is a link. Please open the PDF in a viewer if it is not working on the website.

1. R GGPlot cheat sheet
2. Python Pandas data visualization
3. R For Data Science
4. Cheatsheets fff