**Name - vvgupta2 ( Vishesh Gupta )**

# Homework 3

## Homework Instructions

For questions that require code, please create a code chunk directly below the question and type your code there. Your knitted pdf will show both your code and your output. You are encouraged to knit your file as you work to check that your coding and formatting is done so appropriately.

For written responses or multiple choice questions, please bold your (selected) answer.

## Grading Details

All questions will be graded full credit (1 point), half credit (0.5 point) or no credit (0 points).

Full credit responses should have the correct response and appropriate code (if applicable). Half credit responses will have a reasonable attempt (typically no more than one small error or oversight), and no credit responses will be either non-attempts or attempts with significant errors.

### Exercise 1

Consider the simple linear regression model

$$Y = -3 + 2.5x + \epsilon$$

where

$$\epsilon \sim N(0, \sigma^2 = 4).$$

What is the expected value of $Y$ given that $x = 5$? That is, what is $\text{E}[Y \mid X = 5]$?

Use the solution code chunk to calculate your answer.

```r
#solution
expected_value = function (n, x){
  epsilon = rnorm (n = n, mean = 0, sd = 2)
  y = - 3 + (2.5 * x) + epsilon
  conditional =  - 3 + (2.5 * x)
  print (conditional)
}
```

```r
expected_value (10,5)
```

```
## [1] 9.5
```

### Exercise 2

Which **2** of the following statements are **not** assumptions we need to make before fitting a simple linear model? Please **bold** your answers.

- The variance in Y is constant across each value of x
- **The variance of Y is equal to the variance of X**
- **The Y variable is normally distributed**
- The residuals are independent of one another
- The residuals at any given x are normally distributed

### Exercise 3

For this Exercise, use the built-in `trees` dataset in `R`. Fit a simple linear regression model with `Girth` as the response and `Height` as the predictor.

*Tip: The response variable should always be listed first inside the `lm` command.*

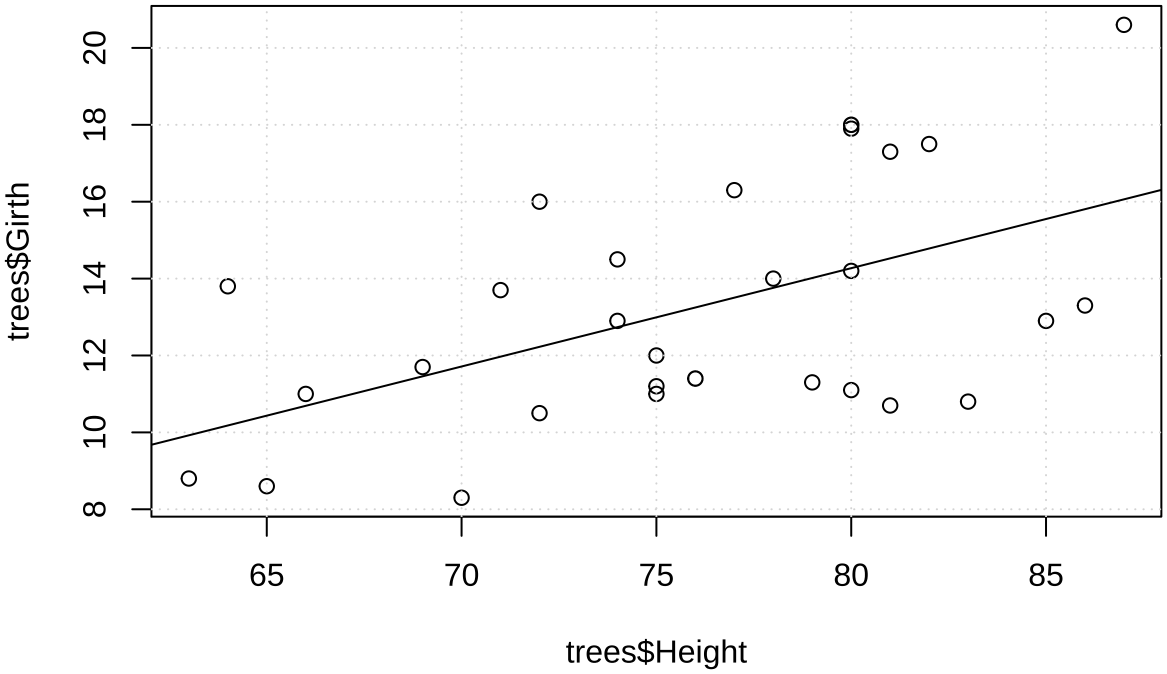
What is the slope of the fitted regression line?

*Hint: To extract the slope specifically, you'll need to use double brackets to select the exact coefficient you want.*

```
#solution
#head(trees)
slr = lm (trees$Girth ~ trees$Height, data = trees)
slr$coefficients[2]
```

```
## trees$Height
##    0.2557471
```

```
plot (trees$Girth ~ trees$Height, data = trees)
grid()
abline(slr)
```



**Exercise 4**

Continue using the SLR model you fit in the previous exercise. What is the value of $R^2$ for this fitted SLR model?

```
#solution
names(summary(slr))
```

```
##  [1] "call"          "terms"         "residuals"     "coefficients"
##  [5] "aliased"       "sigma"         "df"            "r.squared"
##  [9] "adj.r.squared" "fstatistic"    "cov.unscaled"
```

```
summary(slr)$r.squared
```

```
## [1] 0.2696518
```

**Exercise 5**

Consider the simple linear regression model

$$Y = 10 + 5x + \epsilon$$

where

$$\epsilon \sim N(0, \sigma^2 = 16).$$

Calculate the probability that $Y$ is less than 6 given that $x = 0$.

```
#solution
x=0
sigma=4
mu=10+5*x
pnorm(6,mu,sigma)
```

```
## [1] 0.1586553
```

```
# how do you get the probability?
```

### Exercise 6

Using the SLR model in the previous exercise, what is the probability that $Y$ is greater than 3 given that $x = -1$?

```
#solution
x=-1
sigma=4
mu=10+5*x
pnorm(3,mu,sigma, lower.tail = FALSE)
```

```
## [1] 0.6914625
```

### Exercise 7

Consider the fitted regression model:

$$\hat{y} = -1.5 + 2.3x$$

Indicate all of the following that **must** be true:

- The difference between the $y$ values of observations at $x = 10$ and $x = 9$ is 2.3.

**- From this information, our best estimate for the mean of $Y$ when $x = 0$ is -1.5.**

- There are observations in the dataset used to fit this regression with negative $y$ values.

**- As the x variable increases, the y variable tends to also increase.**

### Exercise 8

For the remaining exercises, use the `faithful` dataset, which is built into `R`.

*Tip: You should print it and/or look at the documentation for it in an r chunk for context! But try to delete it before knitting your final pdf for submission.*

Suppose we would like to predict the duration of an eruption of the Old Faithful geyser in Yellowstone National Park based on the waiting time before an eruption.

Fit a simple linear model in `R` that uses the Waiting time between eruptions to predict the Duration of the eruptions.

Use the fitted model to predict the duration of an eruption based on a waiting time of **80** minutes.

```
#solution
?faithful
head(faithful)
```

```
##   eruptions waiting
## 1     3.600      79
## 2     1.800      54
## 3     3.333      74
## 4     2.283      62
## 5     4.533      85
## 6     2.883      55
```

```
model = lm (eruptions ~ waiting, data = faithful)
predict(model, data.frame(waiting = 80))
```

```
##       1
## 4.17622
```

### Exercise 9

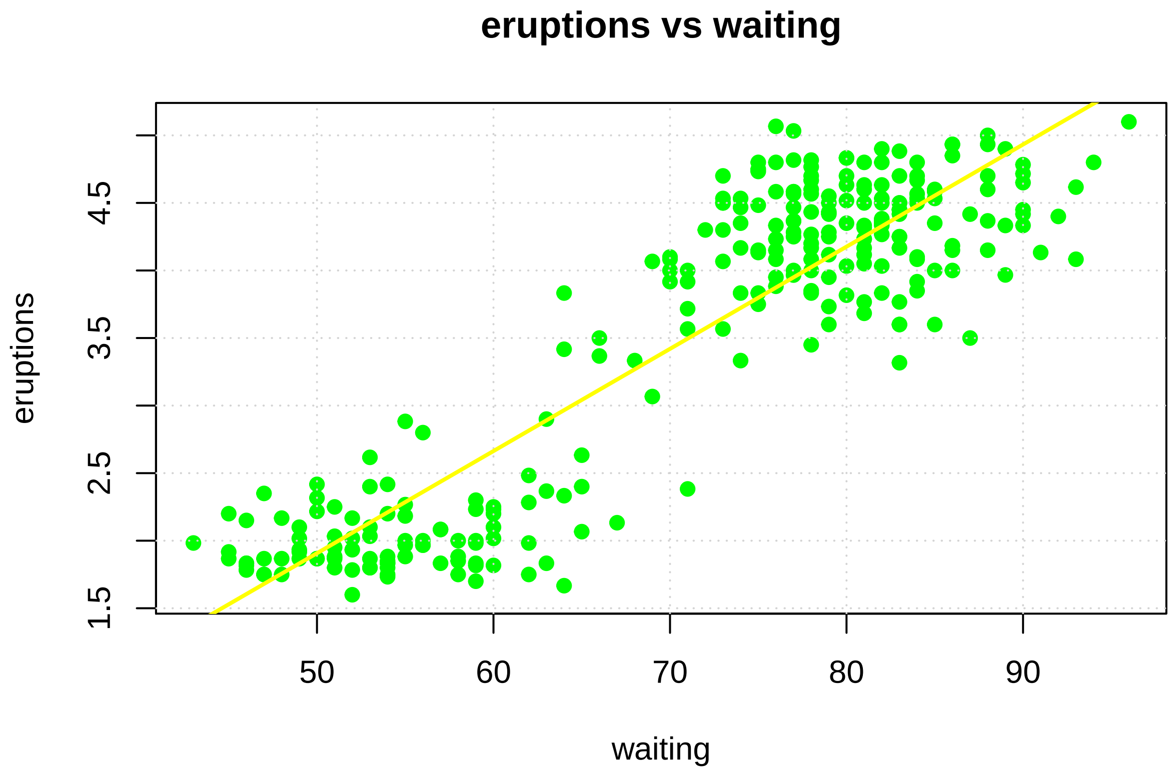
Before going too far, let's plot the data in a scatterplot and visually check our assumptions.

Let's also make this scatterplot look presentable. It should have a) an appropriate title, b) appropriate axis labels, c) use some color, and d) use a different point symbol by changing [pch] (http://www.sthda.com/english/wiki/r-plot-pch-symbols-the-different-point-shapes-available-in-r)

You should also add e) a grid and f) a best fit line from your previous model. Give the line a width of 3 and color.

*And yes... all assumptions appear reasonable based on the plot.*

```
#solution
plot (faithful$eruptions ~ faithful$waiting, data = faithful, lwd = 3, col = "green", xlab = "waiting",
grid()
abline(model , lwd = 2, col= "yellow")
```



**Exercise 10**

Would it be appropriate to use this data to predict the duration of an eruption after a waiting period of **120** minutes?

- Yes, since all we need is an x value to insert into our equation
- Not necessarily, only if we have a data point where waiting period is exactly 120. **- Not necessarily, if our x range doesn't extend that far, it is extrapolation**

**Exercise 11**

What proportion of the variation in eruption duration is explained by the linear relationship with waiting time?

```
#solution
summary(model)$r.squared
```

```
## [1] 0.8114608
```

**Exercise 12**

Calculate the standard deviation of the residuals of the fitted model with the `faithful` data.

_Hint: You'll need to complete an additional operation on an extracted value from the model.

```r
#solution
sd(model$residuals)
```

```
## [1] 0.495596
```