HOMEWORK 3 — MATH 484-564, REGRESSION

DUE SEPTEMBER 20TH 2024, FRIDAY, 11:59PM. SEE THE SUBMISSION INSTRUCTIONS ON CANVAS.

(1) (6 points) Recalled that S_{xy} is defined as

$$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}).$$

Show that S_{xy} can also be written as

$$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x}) y_i$$

therefore

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n} (x_i - \overline{x}) y_i}{S_{xx}} = \sum_{i=1}^{n} c_i y_i$$

where

$$c_i = \frac{x_i - \overline{x}}{S_{xx}}, \quad i = 1, \cdots, n.$$

(2) (6 points) In our class discussion and on pages 37 and 38 of the August 19th version of the notes, we established that $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are unbiased estimators of β_0 and β_1, respectively. This means that:

$$E(\widehat{\beta}_0) = \beta_0 \quad \text{and} \quad E(\widehat{\beta}_1) = \beta_1.$$

In the proof, we assumed that $E(\epsilon_i) = E(\epsilon \mid x_i) = 0$. However, what if $E(\epsilon_i) = E(\epsilon \mid x_i) = \alpha$, where α is a non-zero constant? How would this assumption affect the expectations of $\widehat{\beta}_0$ and $\widehat{\beta}_1$?

To receive full credit, you need to support your claim with a clear and rigorous mathematical proof.

3. (6 points) This exercise involves using R. Start by opening R Studio and navigating to File − > New File − > R Script to create a new R script. In the R script, enter the code below to generate a simulated data set with 1000 data points. Save this data set as a CSV file named simulated_data.csv.

```r
1  # Set seed for reproducibility
2  set.seed(123)
3
4  # Number of data points
5  n <- 1000
6
7  # Generate x values (you can specify the range or generate random values)
8  x <- runif(n, min = 0, max = 10)  # Random values between 0 and 10
9
10 # Generate epsilon values (normally distributed with mean 0 and variance 3)
11 epsilon <- rnorm(n, mean = 0, sd = sqrt(3))
12
13 # Compute y values based on the equation y = 2 + 7x + epsilon
14 y <- 2 + 7 * x + epsilon
15
16 # Combine x and y into a data frame
17 data <- data.frame(x = x, y = y)
18
19 # Write data to a CSV file
20 write.csv(data, "simulated_data.csv", row.names = FALSE)
21
22 # Confirmation
23 cat("Data saved to 'simulated_data.csv'")
24
```

Figure 1: Note: In line 8, we use a uniform distribution to generate values for x ranging from 0 to 10. Once x is generated and recorded in the CSV file, it is treated as fixed and non-random. This fixed x value is a key assumption in our regression model.

Based on the code from Homework 2, Problem 5, please complete the following tasks:

(a) Create an R Markdown document and load the CSV file simulated_data.csv.

(b) Set up a least squares regression model using the lm() function.

(c) Use the summary function to display the results of the regression model.

(d) Plot the data using the plot() function.

(e) Add the least squares regression line to the plot using the abline() function.

4. (6 points) Repeat the steps in Question 3 and enter the code below to generate another simulated data set with 1000 data points. Save this data set as a CSV file named simulated_data_nonconstant_variance.CSV.

```
1  # Set seed for reproducibility
2  set.seed(123)
3
4  # Number of data points
5  n <- 1000
6
7  # Generate x values (you can specify the range or generate random values)
8  x <- runif(n, min = 0, max = 10)  # Random values between 0 and 10
9
10 # Generate epsilon values (normally distributed with mean 0 and variance x^2)
11 epsilon <- rnorm(n, mean = 0, sd = x^2)
12
13 # Compute y values based on the equation y = 2 + 7x + epsilon
14 y <- 2 + 7 * x + epsilon
15
16 # Combine x and y into a data frame
17 data <- data.frame(x = x, y = y)
18
19 # Write data to a CSV file
20 write.csv(data, "simulated_data_nonconstant_variance.csv", row.names = FALSE)
21
22 # Confirmation
23 cat("Data saved to 'simulated_data_nonconstant_variance.csv'")
24
```

Figure 2: Note: In line 11, we observe that the variance of the error term is no longer constant. Instead, it increases as x increases, following the relationship $\text{Var}(\epsilon) = x^2$.

Based on the code from Homework 2, Problem 5, please complete the following tasks:

(a) Create an R Markdown document and load simulated_data_nonconstant_variance.CSV.

(b) Set up a least squares regression model using the lm() function.

(c) Use the summary function to display the results of the regression model.

(d) Plot the data using the plot() function.

(e) Add the least squares regression line to the plot using the abline() function.

(f) Describe how the plot in 3(e) differ from 4(e). Write this comment on your R Markdown document.

5. (6 points) Consider the true relationship between x and y given by

$$y = \beta_0 + \beta_1 x + \epsilon.$$

When we collect a data set (x_i, y_i), $i = 1, \cdots n$, this relationship can be expressed as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

What assumptions about the error terms ϵ_i are necessary to obtain the following results?

(a) $E(\widehat{\beta}_0) = \beta_0$

(b) $\mathrm{Var}(\widehat{\beta}_1) = \sigma^2 / S_{xx}$

(c) $\widehat{\beta}_1 \sim \mathrm{N}(\beta_1, \sigma^2 / S_{xx})$