HOMEWORK 7 MATH 484-564, REGRESSION

DUE OCTOBER 18TH 2024, FRIDAY, 11:59PM. SEE THE SUBMISSION INSTRUCTIONS ON CANVAS.

(1) (6 points) You are given a dataset that contains information on employees in a company, including their salary, years of experience, and whether they have a management position (coded as 1 for management and 0 for non-management). You are tasked with analyzing the relationship between **salary** (in thousands of dollars) and **years of experience**, but you also want to determine if being in a **management position** modifies this relationship.

You are asked to fit the following linear regression model with an interaction term between **Years of Experience** and **Management**:

$$\text{Salary} = \beta_0 + \beta_1 \cdot \text{Years of Experience} + \beta_2 \cdot \text{Management} + \beta_3 \cdot (\text{Years of Experience} \times \text{Management}) + \epsilon$$

where

- Salary: Continuous variable (in thousands of dollars)
- Years of Experience: Continuous variable (in years)
- Management: Categorical variable (1 = Management, 0 = Non-management)

(a) Explain what the coefficient β_3 represents in the context of this problem. What does it mean for the relationship between years of experience and salary?

(b) You run the regression and obtain the following estimated coefficients:

$$\hat{\beta}_0 = 35, \quad \hat{\beta}_1 = 1.5, \quad \hat{\beta}_2 = 10, \quad \hat{\beta}_3 = 2.5$$

(i) For an employee with 5 years of experience in a management position, what is their predicted salary?

(ii) For an employee with 5 years of experience in a non-management position, what is their predicted salary?

(c) Let us study the effect of experience on salary for different groups:

(i) Based on the model, what is the effect of an additional year of experience on salary for non-management employees?

(ii) How does the effect of an additional year of experience on salary differ for management employees compared to non-management employees?

(2) (6 points) You are tasked with analyzing the effect of **employee training hours**, X_1 and **company experience (years)**, X_2 on **employee productivity**, Y. After collecting the data, you run a regression model that includes both the main effects and their interaction term. Your goal is to evaluate how the combination of training hours and years of experience affects productivity.

The regression model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2) + \epsilon$$

After fitting the model, the p-values for the estimated coefficients are as follows:

Term	Coefficient($\hat{\beta}$)	p-value
Intercept(β_0)	50	0.001
Training Hours(β_1)	0.2	0.35
Experience (Years)(β_2)	1.5	0.40
Interaction(β_3)	5	0.02

(a) Explain the hierarchy principle in the context of this problem. Based on the p-values, should you include the interaction term without the main effects? Justify your decision.

(b) Why are the p-values for X_1 and X_2 large, while the p-value for the interaction term is small? What does this imply about the relationship between training hours, experience, and productivity?

3. (6 points) Using the simulated data from the file HW7-data-1.csv, follow the code below to compute the residuals and fitted values. The code includes two manual plots: one showing the residuals on the regression line and another displaying the residuals versus the fitted values.

```r
# Load necessary library
library(ggplot2)

# Read the data from CSV file
data <- read.csv("HW7-data-1.csv")

# Fit a linear model
model <- lm(y ~ x, data = data)

# Calculate fitted values and residuals
data$fitted <- fitted(model)
data$residuals <- resid(model)

# Print the first few rows of the data frame with fitted values and residuals
head(data)

# Plot the regression line, observed values, and residuals
ggplot(data, aes(x = x, y = y)) +
  geom_point(color = "blue", alpha = 0.6) +  # observed points
  geom_smooth(method = "lm", color = "black", se = FALSE) +  # regression line
  geom_segment(aes(x = x, xend = x, y = fitted, yend = y),
               color = "red", linetype = "dashed", alpha = 0.5) +  # residuals as
vertical lines
  labs(title = "Regression Line with Residuals",
       x = "X",
       y = "Y") +
  theme_minimal()


# Plot residuals vs fitted values
ggplot(data, aes(x = fitted, y = residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Residuals vs Fitted Values",
       x = "Fitted Values",
       y = "Residuals") +
  theme_minimal()

```

What conclusions can you draw from the residual vs. fitted plot? Specifically, can you determine from the plot whether:

(i) x and y have a linear relationship?

(ii) The variance of the residuals (errors) remains constant across all levels of the independent variable?

4. (6 points) Using the same dataset as in Problem 3, fit the regression line and use the `plot(model)` function to generate the four diagnostic plots. What can you infer about the distribution of the residuals from the QQ plot?

   Additionally, we can also assess the distribution by creating a histogram of the residuals. Follow the code below to do this and see what you can conclude from the histogram plot.

```r
# Load necessary library
library(ggplot2)

# Read the data from CSV file
data <- read.csv("HW7-data-1.csv")

# Fit a linear model
model <- lm(y ~ x, data = data)

plot(model)

# Calculate fitted values and residuals
data$fitted <- fitted(model)
data$residuals <- resid(model)


# Plot histogram of residuals
ggplot(data, aes(x = residuals)) +
  geom_histogram(binwidth = 1, fill = "blue", color = "black", alpha = 0.7) +
  labs(title = "Histogram of Residuals",
       x = "Residuals",
       y = "Frequency") +
  theme_minimal()
```

5. (6 points) Please apply the same analysis you performed in Problem 4 to the new dataset, HW7-data-2.csv. Afterward, discuss the distribution of the residuals. Do the residuals appear to follow a normal distribution?