HOMEWORK 5 MATH 484-564, REGRESSION

DUE OCTOBER 4TH 2024, FRIDAY, 11:59PM. SEE THE SUBMISSION INSTRUCTIONS ON CANVAS.

(1) (6 points) Using the advertising data set, we'll perform a regression analysis to examine how sales are influenced by the advertising budgets for TV, radio, and newspapers. The code below provides the 95% confidence intervals for the regression parameters, βs. Please use this information to answer the following questions.

```
model1<-lm(df$sales~df$TV+df$radio+df$newspaper)
confint(model1)
```

```
                  2.5 %     97.5 %
(Intercept)   2.32376228 3.55401646
df$TV         0.04301371 0.04851558
df$radio      0.17154745 0.20551259
df$newspaper -0.01261595 0.01054097
```

(a) By writing

$$\text{sales} = \beta_0 + \beta_1 \text{ TV} + \beta_2 \text{ radio} + \beta_3 \text{ newspaper} + \epsilon$$

how should we interpret the confidence interval for β_1?

(b) Based on the 95% confidence interval for the newspaper coefficient, explain why we would fail to reject the null hypothesis $H_0 : \beta_3 = 0$ in favor of the alternate hypothesis $H_1 : \beta_3 \neq 0$ at the significant level of $\alpha = 0.05$.

(2) (6 points)

(a) Given the regression model

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

explain why we interpret $\widehat{\beta}_1$ as the **average** or **expected** change in y for a one-unit increase in x, rather than simply the change in y for one unit increase in x.

(b) In multiple linear regression, represented by the equation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon,$$

how is the value of $\widehat{\beta}_1$ interpreted?

3. (6 points) In the previous homework, you learned that one key feature of R^2 is that adding an extra predictor will never decrease its value, regardless of the predictor's significance. In fact, R^2 usually increases when additional predictors are included. To overcome this limitation, we use adjusted R^2 (refer to Notes, Sept 22nd version, page 72).

   (a) Recall the muscle mass example from the previous homework, where we obtained an R^2 of 0.7501 for the regression model:

   $$\text{muscle mass} = \beta_0 + \beta_1 \text{Age} + \epsilon.$$

   Now, let's consider a second regression model that includes data on estrogen levels:

   $$\text{muscle mass} = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Estrogen} + \epsilon.$$

   In this case, the adjusted R^2 is 0.8223. What can we conclude about the predictor estrogen?

   (b) Suppose we also have data on the number of hours a person sleeps (Sleep), protein intake (Protein), and the number of hours a person exercises per day (Exercise). We can set up a third model:

   $$\text{muscle mass} = \beta_0 + \beta_1 \text{Sleep} + \beta_2 \text{Protein} + \beta_3 \text{Exercise} + \epsilon.$$

   In this model, the adjusted R^2 is 0.8511, which is higher than the adjusted R^2 for the second model. What can we infer from this adjusted R^2 value of 0.8511 when compare to the value of 0.8223?

4. (6 points) Upload the Advertising dataset into R using your R Markdown file, then set up the following regression model:

$$\text{sales} = \beta_0 + \beta_1 \text{ TV} + \beta_2 \text{ radio} + \epsilon$$

Using this model, answer the following questions:

(a) What is the average sales when \$100,000 is spent on TV advertising and \$20,000 on radio advertising?

(b) What is the predicted sales for a specific market when \$100,000 is spent on TV advertising and \$20,000 on radio advertising?

(c) Construct a 95% confidence interval for estimating the average sales under the same advertising expenditures as in (a) and (b).

(d) Construct a 95% prediction interval for predicting the sales of a specific market under the same advertising expenditures as in (a) and (b).

Hint:

```r
df<-read.csv("Advertising.CSV", header=TRUE, sep=",")

model1 <- lm(sales ~ TV + radio, data=df)

new_data <- data.frame(TV = 100, radio = 20)

prediction <- predict(model1, newdata=new_data, interval="confidence", level=0.95)
prediction_interval <- predict(model1, newdata=new_data, interval="prediction", level=0.95)

# Display the results
print(prediction)          # For confidence interval
print(prediction_interval) # For prediction interval
```

The *R*-code above will be useful. Note that because the Advertising data has sales in thousand of units and advertising budget in thousand of dollars, we set "TV =100" and "radio=20" in the R-code.

5. (6 points) If you compare the lengths of the 95% confidence interval and the prediction interval, you will notice that the prediction interval is longer. This is a well-known result, and we have mathematically demonstrated why this happens in the notes from September 22nd, pages 74-76.

   However, if your client or manager—who has an MBA and a not data science degree—asks you to explain why this occurs, how would you describe this phenomenon without using advanced statistical knowledge?