ESE305-Homework 2

Due: Friday, September 29. 11:59PM

1 Conceptual

1.1 Question 1

Looking at the table below, describe the null hypothesis to which the p-values correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of `sales`, `TV`, `radio`, and `newspaper`, rather than in terms of the coefficients of the linear model.

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	<0.0001
TV	0.046	0.0014	32.81	<0.0001
radio	0.189	0.0086	21.89	<0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

1.2 Question 2

Suppose we have a data set with five predictors, X_1 = GPA, X_2 = IQ, X_3 = Gender (1 for Female and 0 for Male), X_4 = Interaction between GPA and IQ, and X_5 = Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50$,$\hat{\beta}_1 = 20$,$\hat{\beta}_2 = 0.07$,$\hat{\beta}_3 = 35$,$\hat{\beta}_4 = 0.01$,$\hat{\beta}_5 = -10$.

1. Which answer is correct, and why?

    1. For a fixed value of IQ and GPA,males earn more on average than females.
    2. For a fixed value of IQ and GPA, females earn more on average than males.
    3. For a fixed value of IQ and GPA, males earn more on average than females provided that GPA is high enough.
    4. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

2. Predict the salary of a female with IQ of 110 and a GPA of 4.0.

3. True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

1.3 Question 3

I collect a set of data ($n = 100$ observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$

1. Suppose that the true relationship between X and Y is linear,$Y = \beta_0 + \beta_1 X$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

2. Answer (a) using test rather than training RSS.

3. Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

4. (d) Answer (c) using test rather than training RSS.

1.4 Question 4

Using the equations for $\hat{\beta}_0$ and $\hat{\beta}_1$ given below, argue that in the case of simple linear regression, the least squares line always passes through the point $(\bar{x}, \bar{y})$

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

2 Applied

2.1 Question 5

In this exercise you will create some simulated data and will fit simple linear regression models to it. Make sure to use `numpy.random.seed(seed = 42)` every time you run a code block with a random number generator. This is to ensure consistent and reproducible results.

1. Using the `numpy.random.normal()` function create a vector x, containing 100 observations drawn from a $N(0, 1)$ distribution. This represents a feature, X.

2. Now, using the same function create a vector, `eps`, containing 100 observations drawn from a $N(0, 0.25)$ distribution i.e. a normal distribution with mean zero and variance 0.25.

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

$$Y = -1 + 0.5X + \epsilon$$

What is the length of the vector `y`? What are the values of β_0 and β_1 in this linear model?

4. Create a scatter plot displaying the relationship between `x` and `y`. Comment on what you observe.

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

6. Display the least squares line on the scatter plot obtained in 5, Draw the population regression line on the plot, in a different color. Create an appropriate legend.

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.

8. Repeat 1-6 after modifying the data generation process in such a way that there is **less** noise in the data. The model should remain the same. You do this by decreasing the variance of the normal distribution used to generate the error term ϵ in 2. Describe your results.

9. Repeat 1-6 after modifying the data generation process in such a way that there is **more** noise in the data. The model should remain the same. You do this by increasing the variance of the normal distribution used to generate the error term ϵ in 2. Describe your results.

10. What are the confidence intervals for β_0 and β_1 based on the original data set, the noisier set, and the less noisy data set? comment on your results.

2.2 Question 6

This question involves the use of multiple linear regression on the `Auto` data set.

1. Produce a scatter plot matrix which includes all of the variables in the data set.

2. Compute the matrix of correlations between the variables using the function `df.corr()`.

3. Perform a multiple regression with `mpg` as the response. Print the results of your regression analysis. Comment on the output, for instance:

    1. Is there a relationship between the predictors and the response
    2. Which predictors appear to have a statistically significant relationship to the response?
    3. What does the coefficient for the `year` variable suggest?

4. Produce diagnostic plots (Q-Q plot and residual plot). Comment on any problems you see with the fit. Do the residual plot suggest any unusually large outliers?

5. Fit a linear regression model with interaction effects. Do any interactions appear to be statistically significant?

6. Try a few transformations of the variables, such as $\log(X)$, $\sqrt{X}$, X^2, comment of your findings with help of your fit statistics, and diagnostic plots.