Statistics 500: Midterm 1

Name: ________________

We randomly collected n samples and used the linear model as follows.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon,$$

1. [5pt] The matrix form of linear model is simpler, it can be written as followings.

$$Y = X\beta + \epsilon.$$

State the vector β and the first row of X.

$$\beta = (\beta_0, \beta_1, \beta_1)^T,$$

$$X = (1, X_1, X_2).$$

2. [5pt] Recall that the Least Square Estimator (LSE) is $\hat{\beta} = (X^T X)^{-1} X^T Y$. The following italicized statement is either true or false.

If sample size n is smaller than the number of predictors p, the least square estimator does not exist.

Circle whether the statement is True or False.

True; $(X^T X)^{-1}$ does ont exist.

3. [5pt] The following italicized statement is either true or false.

The LSE is the best estimator for β (i.e., it has the minimum mean square of errors).

Circle whether the statement is True or False.

False; it is best among the unbiased estimator.

4. [5pt] The following italicized statement is either true or false.

   $Y = \beta_0 + \beta_1 \log(X) + \epsilon$ ***is a linear model.***

   Circle whether the statement is True or False.

   True; Linear in parameter.

5. (10pt) Suppose for a linear regression the predictors was x_1, x_2, x_3. For $j = 1, 2, 3$, after regressing x_j on the remaining predictors, we get R^2 values of $R_1^2 = 0.75$, $R_2^2 = 0.80$, $R_3^2 = 0.90$. Based on this, what can you infer about collinearity of the predictor variables x_1, x_2, x_3? **Justify your answer**.

   The biggest $VIF_3 = \frac{1}{1-0.90} = 10$. Predictors may have a (weak) collinearity issue.

We performed an experiment concerned with assessing the toxic effect of dioxin. Every separate fish tanks were maintained with different dioxin concentrations (x). A single fish was placed in each tank, and the length of time until the fish died was recorded in days (y). Of interest is a single linear regression model of the form $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. We assume $\epsilon_i \sim_{i.i.d.} N(0, \sigma^2)$ and all assumptions required are satisfied. Suppose that the significance level $\alpha = 0.05$ and the following summary table is faithful.

```
Coefficients:
Estimate  Std. Error t value Pr(>|t|)
(Intercept) 506.4864    50.1074  10.108    <2e-16
x            -0.9281     0.5140  -1.806    0.07406
```

6. [10pt] The following italicized statement is either true or false.

   If hypotheses are $H_0 : \beta_1 = -0.9281$ *v.s.* $H_A : \beta_1 \neq -0.9281$***, We fail to reject the*** H_0

   Circle whether the statement is True or False and explain your choice.

   False; because for given the hypotheses, $t_{stat} = \frac{\hat{\beta}_1 - \beta_1}{\hat{se}(\beta_1)} = \frac{-0.9281-(-0.9281)}{0.514} = 0$. Since $H_A : \beta_1 \neq -0.9281$, the corresponding p-value is $P(|t_{df}| > 0) = 1$. Furthermore because the significance level is 0.05, we fail to reject H_0. We do not have strong evidence that $\beta_1 \neq -0.9281$.

7. [10pt] Suppose that we have a new observation $x_{new} = 1000$. Then, the predicted value is $506.4864 - 0.9281 \times 1000 = -421.6136$. State the interpretation of the predicted value.

   For the new observation $x_{new} = 1000$, the expected value of the response variable is -421.6136. Using the context of the response variable, the length of days until the fish died that must be the non-negative, if the dioxin concentration 1000, the expected length of days until the fish died is 0 day. Or extrapolation

8. [10pt] We believe that samples have errors in dioxin concentration (x). Can we expect the true coefficient for dioxin concentration is smaller than -0.9281 (i.e., $\beta_1 < -0.9281$)? Justify your answer.

   Yes; because errors in predictor causes the estimated coefficient shrinks toward zero.

An analyst studying a chemical process expects the yield to be affected by the levels of two factors, X_1 and X_2. Observations recorded for various levels of the two factors are shown in the following table. The analyst wants to fit a first order regression model to the data. Interaction between X_1 and X_2 is not expected based on knowledge of similar processes. Of interest is a simple linear regression model of the form $y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i$ and a multiple linear regression model of the form $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$. We assume $\epsilon_i \sim_{i.i.d.} N(0, \sigma^2)$ and all assumptions required are satisfied.

Simple Linear Regression

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.0089     0.7347  12.262 7.23e-10 ***
x1            2.1936     0.8195   2.437    0.001 ***
```

Multiple Linear Regression

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.0089     0.7347  12.262 7.23e-10 ***
x1            1.1936     0.6195   0.837    0.444
x2           -0.3488     0.2195  -1.589    0.131
```

9. [10pt] The following italicized statement is either true or false.

Predictors are independent

Circle whether the statement is True or False and explain your choice.

No; coefficients are changed

10. [10pt] Suppose that correlation between X_1 and X_2 is 0.5 (i.e., $cor(X_1, X_2) = 0.5$). The following italicized statement is either true or false.

Since predictors are dependent, a model with the interaction between X_1 and X_2 is better than the considered model. In other words,

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

is better.

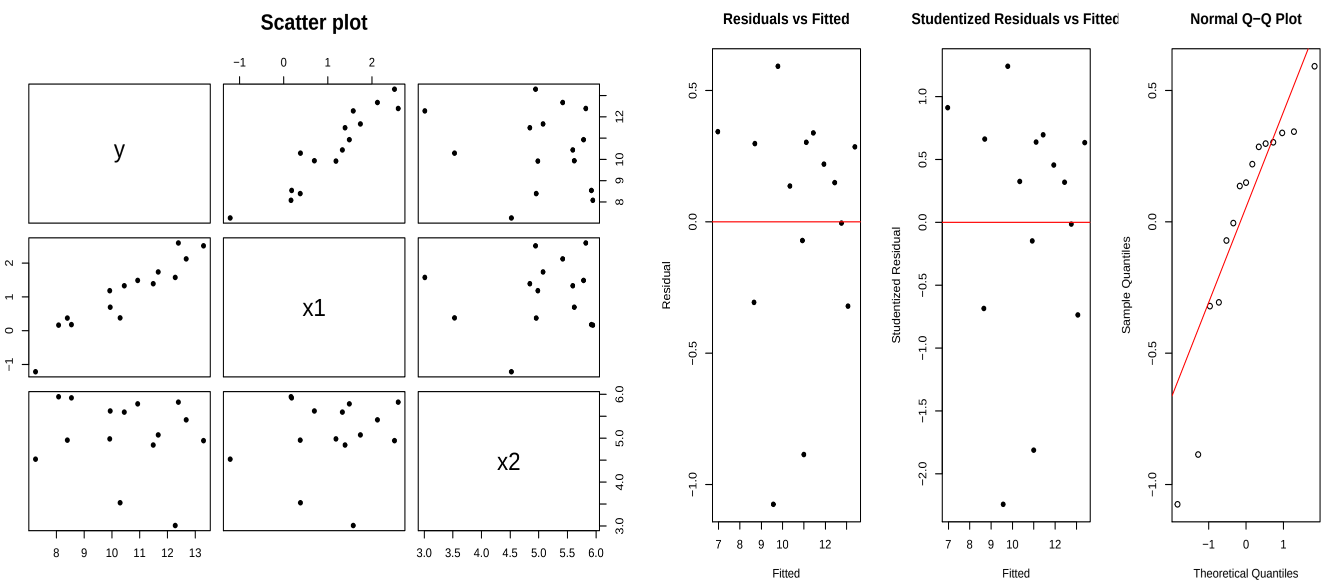
Circle whether the statement is True or False and explain your choice.

No; prior information said interaction can be ignored.

Suppose that we have two predictors x_1 and x_2. Of interest is a multiple linear regression model of the form $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$ for $i = 1, 2, ..., 15$. Suppose that the significance level $\alpha = 0.05$ and the following summary table and plots are faithful.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.5665	0.8137	14.22	0.0000
x1	1.7844	0.1337	13.34	0.0000
x2	-0.5382	0.1590	-3.39	0.0054

Table 1: Summary Table



11. [10pt] Do a test $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 < 0$, and justify your answer based on the above diagnostic plots and summary table.

12. [10pt] Do a test $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 > 0$, and justify your answer based on the above diagnostic plots and summary table.

    Since all assumptions especially linearity and normality are satisfied, we can conclude that X_1 is significant in the model.

    Becuase nomarlity assumption is not satisfied, we cannot trust p-value in the summary table. If you do not discuss diagnostic plots, there is no partial credits.