Midterm Exam: Linear Regression Models

Clearly write down your solution on a paper and upload your solution to Courseworks.

1. (50 points) Write TRUE or FALSE with brief justification.

   1.1. The usual ANOVA F-test of $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ in simple linear regression is actually a special case of a more general framework for testing

   $$H_0 : \text{Reduced model} \quad \text{vs.} \quad H_a : \text{Full model}$$

   based on the test statistic

   $$F = \frac{SSE(\text{Reduced}) - SSE(\text{Full})}{df_{\text{Reduced}} - df_{\text{Full}}} \Bigg/ \frac{SSE(\text{Full})}{df_{\text{Full}}}$$

   1.2. We have a dataset $\{(x_i, y_i)\}_{i=1}^n$ with the ground truth $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, and we run linear regression on this dataset. Let b_1 and b_0 be the least square estimations for β_1 and β_0 respectively, then we have

   $$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \leq \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2.$$

   1.3. We have a dataset $\{(x_i, y_i)\}_{i=1}^n$ we run linear regression on this dataset. Adding in a new point, $(\bar{x}, \bar{y})$ to this dataset will not change our regression fit.

   1.4. Adding a new predictor increases R^2 both if it is positively correlated and negatively correlated with the response.

   1.5. In simple linear regression, we can use both the ANOVA F-test and the t-test for the hypothesis testing $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ with level α, but the F-test might reject the null hypothesis while the t-test might accept the null hypothesis.

   1.6. The least squares estimators can be unbiased for β_0 and β_1 even if we do not assume Gaussian errors, as long as the expectation of the errors are 0.

   1.7. The Q-Q plot is a useful graphical tool for assessing the assumption that the error term are normally distributed.

   1.8. The least squares estimates of the slope and intercept in simple linear regression are the values of β_1 and β_0 that minimize the sum of the squared horizontal distances between the points (x_i, y_i) and the line $y = \beta_0 + \beta_1 x$ in a scatterplot of y versus x.

   1.9. For simple linear regression, in general, the sum of residuals will be zero only if we include an intercept term in the regression.

   1.10. In a multiple linear regression with $p > 1$ predictors, the sign of β_j must be same as the sign of the correlation between predictor $x^{(j)}$ and response y, even if the value is different.

2. (30 points) We are interested in learning the relationship of x = the unemployment rate with respect to the S& P 500, the data available is the monthly data from Jan 2016 to Dec 2017. Notice that the range of the price of S& P 500 is $[719, 1464]$, and the range of the unemployment rate is $[5.3\%, 6.2\%]$. We are interested in fitting the following model:

$$Y_n = \beta_0 + \beta_1 x_n + \epsilon_n \text{ with } \epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \text{ for } n = 1, \ldots, 24.$$

The corresponding R output is:

```
Coefficients:

            Estimate    Std. Error t value   Pr(>|t|)
(Intercept) 4471.3393         A     14.696      0.00
x           -588.9621     52.602        B       0.00

Response: y

           Df   Sum Sq   Mean Sq   F value   Pr(>F)
x           1 868925.20 868925.20   125.36   1.49e-10
Residuals 22 152490.63   6931.39
```

2.1. (6 points) Find A and B in the output above. What can you tell from the slope of your regression relationship of the price of S & P and the unemployment rate?

2.2. (10 points) Clearly state the null hypothesis to test whether there exists a linear relationship between the price of S & P 500 and the unemployment rate. With level $\alpha = 10\%$, are you going to reject the null hypothesis using the ANOVA F-test? Calculate the R^2 for this model.

2.3. (4 points) Given that t(.975, 22) = 2.074, calculate a double-sided 95% confidence interval for the slope.

2.4. (10 points) Now we are fitting a new model:

$$Y_n = \beta_1^{new} x_n + \epsilon_n^{new} \text{ with } \epsilon_i^{new} \overset{iid}{\sim} N(0, \sigma_{new}^2) \text{ for } n = 1, \ldots, 24.$$

The corresponding R output is:

```
Coefficients:

            Estimate    Std. Error t value   Pr(>|t|)
x           182.8802     9.451       19.350       0.00

Response: y

           Df     Sum Sq   Mean Sq   F value   Pr(>F)
x           1 26853804.84 26853804.84   374.4   9.92e-16
Residuals 23  1649491.16    71717.01
```

What can you tell from the slope of your new regression relationship of the price of S & P and the unemployment rate? Calculate the R^2_{new} for the new model and compare to R^2 in 2.2, which model you are going to choose to predict the price of S & P 500 using only the unemployment rate and what's your explanation?