ISyE4031 Regression and Forecasting
Practice Problems 1
Spring 2016

1. Consider the output below with several missing values. Answer the following multiple choice questions (circle the correct/closest answer).

The regression equation is Y = 254 + 2.77 X1 – 3.58 X2

Predictor	Coef	SE Coef	T	P
Constant	253.81	4.781	53.087	
X1	2.7738	0.1846	**a**	
X2	-3.5753	0.1526		
Source	DF	SS	MS	F
Regression	2	22784	11392	**b**
Residual Error	12	307	25.5833	
Total	14	23091		

a. What is **a**?

b. What is **b**?

c. Calculate the estimated standard deviation, *s*.

d. Calculate the coefficient of determination, R^2.

2. Assume that x and y are related according to a simple linear regression model which was estimated as $\hat{y} = 2.1 + 3.4x$. If $SS_{xy} = 16.22$ and $SSE = 4.062$,

a. What is SS_{xx}?

b. What is SS_{yy}?

3. A sociologist investigating the recent increase in the incidence of homicide throughout the US studied the extent to which the homicide rate per 100,000 population (Y) is associated with the city's population size (X_1), the percentage of families with low income (X_2), and the rate of unemployment (X_3). A first-order multiple linear regression model, $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$, was studied and the following output is provided for a sample of 19 cities.

```
The regression equation is
Y = - 40.7 + 0.00362 X1 + 1.23 X2 + 4.76 X3

Predictor      Coef   SE Coef      T      P
Constant    -40.667     6.073  -6.70  0.000
X1         0.003622  0.001170   3.10  0.007
X2           1.2281    0.4733   2.60  0.020
X3            4.764     1.289   3.70  0.002

S = 3.86588   R-Sq = 87.7%   R-Sq(adj) = 85.2%

Analysis of Variance
Source          DF       SS      MS      F      P
Regression       3  1592.24  530.75  35.51  0.000
Residual Error  15   224.18   14.95
Total           18  1816.42
```

a. Is the model useful as a whole? Apply an appropriate F test (state the hypothesis, critical F value, conclusion, etc.) and confirm your conclusion by stating the $p-$value. Use $\alpha = 0.05$.

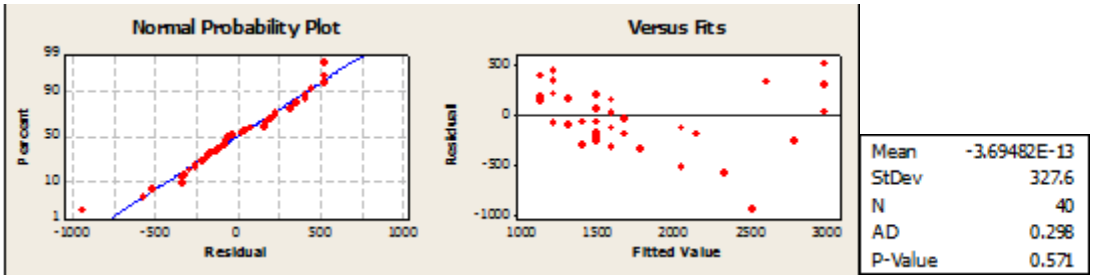
b. If the observation, $Y = 29$ when $X_1 = 1531$, $X_2 = 21.3$, and $X_3 = 7.6$, what is the prediction error?

c. What percent of the total variability in homicide rate is explained by the regression?

d. Is the percentage of families with low income (X_2) a significant variable to predict Y? Apply a t-test (state the hypothesis, critical t value, conclusion, etc.). Use $\alpha = 0.05$.

e. Suppose that $\alpha = 0.01$. By looking at the p-values, are the predictors, X_1, X_2, X_3, significant?

4. Error assumptions. Consider the following normality plot and Residuals vs Fits plot of a regression study. Answer the following True/False questions (circle either True or False).



a. True or False? The assumption $E(\varepsilon_i) = 0$ holds, since the "Mean" of residuals is basically zero.

b. True or False? According to the Anderson-Darling test result, the normality assumption is violated, because the p-value = 0.571 > AD = 0.298.

c. True or False? Since the Residual vs Fitted value plot does not depict any obvious pattern, we do not reject that the error terms are independent.

d. True or False? In general, the error terms in a simple linear regression model should be distributed $N(0,1)$.

e. True or False? According to the above plots, we can say that the error terms of this model do not have the constant variance.

5. Consider the property: The sum of the residuals in any regression model should be zero, i.e., $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$.

a. Prove or disprove that the estimators, $\hat{y}_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x})$ where $\bar{y} = \sum_{i=1}^{n} y_i / n$ and $\bar{x} = \sum_{i=1}^{n} x_i / n$, satisfy this property for a simple linear regression model.

b. Prove or disprove that the estimators, $\hat{y}_i = \bar{y} + \hat{\beta}_1 x_i$, satisfy this property for a simple linear regression model.

6. Miscellaneous short-answer questions. Answer the following questions (circle the correct answer whenever required).

a. In a multiple regression what is the common effect of adding a totally irrelevant (non-significant) variable to the model?

i. *SSE* goes up
ii. Total variation of the response variable increases
iii. Parameter estimates become biased
iv. R^2 goes up

b. Consider a simple linear regression model to predict the sales of a product (y) by using advertising expenditures (x). If we want to predict the sales of a new product at a given amount x_p of the advertising expenditure, then the following holds:

i. The closer x_p is to $\bar{x}$ the less accurate the prediction will be.
ii. The length of the estimated prediction interval would be decreased as the value of x_p gets closer to $\bar{x}$.
iii. The length of the estimated prediction interval is independent of x_p.
iv. A prediction interval of a single observation in a simple linear regression is always shortest at the origin (x=0, y=0).

c. True or False? The point ($\bar{x}$, $\bar{y}$) always lies on the regression line.

d. True or False? Some regression situations imply that the true line passes through the origin (x=0, y=0). For some of those situations a no-intercept model may provide a superior fit, if $\bar{x}$ is close to zero.

e. True or False? Regression analysis cannot be used to establish cause-and-effect relationships even if the associations between the response variable and the predictors are strongly significant.

f. True or False? Consider a simple linear regression model that was developed to study the relationship between the starting annual salaries of certain college graduates and their cumulative grade point averages (CGPA). Suppose that one of your friends has recently graduated from this college with a CGPA of 2.83, and the regression model produces a point estimate of $53,552 annual salary with a 95% prediction interval of (41842, 65262) when CGPA = 2.83. If she got a job offer that pays $50,000 per year and asked your opinion (should she accept the offer?), you should recommend that she *reject* the offer.

g. True or False? Suppose a fitted linear regression model is $\hat{y} = 10 + x_1 - x_2$. If $x_1 = 1$ and $x_2 = 2$, and the corresponding observed value of $y = 8$, the residual at this observation is 0.