MATH 456/565 – Sample EXAM 1

SHOW ALL YOUR WORK.

Please show all the formulas and facts you are using. Please write down all the steps. It makes it easier to grade and give partial credit.

NAME:____________________

1. [15 pts] Describe the Simple Linear Regression model, with all the assumptions. Describe the purpose of the model. Give an example.
   The Simple Linear Regression model expresses the response variable Y as a linear relationship with the explanatory variable X plus an error term, not accounted for by the linear relationship between X and Y.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad 1 \le i \le n$$

   The assumptions are:
   a. There is a linear relationship between X and Y
   b. The error terms $\{\varepsilon_i\}$ are independent, normally distributed, mean 0 and constant variance σ^2, that is, they are i.i.d $N(0, \sigma^2)$

   Example: relationship between

   - Height and weight
   - Advertising and sales

2. [15 pts]
   a. Describe how the least square estimators for the intercept and slope in the simple regression model are obtained. That is, explain what do we do to find them.
   Estimators for β_0 and β_1 are obtained by minimizing the distance between the Y observation and the corresponding value in the regression line, that is

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 + \beta_1 x_i)^2$$

   b. Give expressions for the estimate of the i) slope, ii) intercept and iii) variance of the error terms.

$$b_0 = \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$b_1 = \hat{\beta}_1 = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

$$\hat{\sigma}^2 = s^2 = \text{MSE} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{\sum {e_i}^2}{n-2}$$

3. [20 pts]
   a. Give expressions for SST, SSE, SSR.

$$\text{SST} = \sum (y_i - \bar{y})^2$$
$$\text{SSE} = \sum (y_i - \hat{y}_i)^2$$
$$\text{SSR} = \sum (\hat{y}_i - \bar{y})^2$$

   b. Show that SST=SSE+SSR

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$
$$= \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2 + 2\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

But

$$\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$
$$= \sum (y_i - b_0 - b_1 x_i)(b_0 + b_1 x_i - \bar{y})$$
$$= \sum (y_i - (\bar{y} - b_1\bar{x}) - b_1 x_i)(\bar{y} - b_1\bar{x} + b_1 x_i - \bar{y})$$
$$= \sum ((y_i - \bar{y}) - b_1(x_i - \bar{x}))\ b_1(x_i - \bar{x}) =$$
$$= b_1\left[\sum (y_i - \bar{y})(x_i - \bar{x}) - b_1 \sum (x_i - \bar{x})^2\right] = 0$$

because

$$b_1 = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

Thus

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

Or

SST=SSE+SSR

   c. Define R^2 in terms of the expressions in part a). Use b) to find another expression of R^2 in terms of the expressions in part a) . (Hint: re-write the numerator).

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

   d. State the interpretation of R^2.
   R^2 is the proportion of the variability in the response variable explained by the regression model.

4. [30 pts] A criminologist studying the relationship between level of education and crime rate in medium sized US counties collected data from a random sample of 84 counties. X = the percentage of individuals in the county having at least a high school diploma, and Y = the crime rate per year.

```
Call:
lm(formula = Y ~ X)

Coefficients:
```

	Estimate	Std. Error	T value	Pr(>|t|)
(Intercept)	20517.6	3277.64	$\frac{20517.6}{3277.64}$=6.26	1.67E-08
X	-170.575	41.57	$\frac{-170.575}{41.57}$= -4.1	9.57E-05

Residual standard error: __2356.29_____ on __82_ degrees of freedom

R-squared: __0.17___,

Analysis of Variance Table
Response: Y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SSR	1	93462942.27	93462942.27	$\frac{93462942.27}{5552111.77} = 16.83$	9.59E-05
SSE	82	455273165.33	$\frac{455273165.33}{82} = 5552111.77$		
SST	83	548736107.6			

a. Obtain the estimated regression function.

$$Y = 20517.6 - 170.575X$$

b. Fill in the above tables.

c. Obtain a point estimate for σ^2. =5552111.77

d. Run a test of hypothesis to see whether there is a linear relationship between the variables. Describe the statistic you are using and explain your conclusions.

$H_0: Y = \beta_0 + \varepsilon,$

$H1: Y = \beta_0 + \beta_1 X + \varepsilon,$

Test statistic:

$F = \frac{93462942.27}{5552111.77} = 16.83$, p-value = 9.59E-05 thus, there is a linear relationship between the variables.

e. Find the 95% confidence interval for the slope. $t_{alpha/2} \sim 1.989$

$$-170.575 \pm 1.989 * 41.57$$

f. Find a confidence interval for the mean response at when the percentage of individuals in the country having at least a high school diploma is at the level = 100.

Mean prediction:

$$Y = 20517.6 - 170.575 * 100 = 3460.1$$

$$se_{pred} = s\sqrt{\frac{1}{84} + \frac{(100-\bar{x})^2}{S_{XX}}} = 2356.29\sqrt{\frac{1}{84} + \frac{(100-\bar{x})^2}{S_{XX}}}$$

We are missing $\bar{x}$

and with some effort we can compute S_{XX}

Confidence interval: $3460.1 \pm 1.989\ se\_pred$

g. What is the significance of R^2? What you would conclude in this case?

R^2=0.17 which indicates that the regression line is a poor fit, or it explains a small percentage of the variability in Y.

5. (5 pts) Which of the following can never be 0 (unless the population standard deviation σ= 0)?

A. The estimated intercept, β_0

B. A residual

C. The variance of the prediction error, $\sigma^2\{pred\}$

D. The estimate of $E\{Y_h\}$, $\hat{Y}_h$

6. (5 pts) In the context of simple linear regression, the point $(\bar{X}, \bar{Y})$ ____________

Circle ALL answers that apply to the blank above:

a) will always be one of the points in the data set.

b) will always fall on the fitted line.

c) is not informative

7. (10 pts) When we run the following model $E\{Y_i\} = \beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\beta_3X_{i3}$, we obtained

$$(X^tX)^{-1} = \begin{bmatrix} 0.506174 & 0.001158 & -0.050611 & -0.077902 \\ 0.001158 & 0.000022 & -0.000030 & -0.000702 \\ -0.050611 & -0.000030 & 0.86947 & -0.036324 \\ -0.077902 & -0.000702 & -0.036324 & 0.048116 \end{bmatrix}$$

The estimated variance was: 6.310144

a. Write the formula for the covariance matrix of the coefficients, $\hat{\beta}^t = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$, and its estimated value:

$$\sigma^2(\hat{\beta}) = \sigma^2\ (X^tX)^{-1}$$

$$se^2(\hat{\beta}) = 6.310144^*\begin{bmatrix} 0.506174 & 0.001158 & -0.050611 & -0.077902 \\ 0.001158 & 0.000022 & -0.000030 & -0.000702 \\ -0.050611 & -0.000030 & 0.86947 & -0.036324 \\ -0.077902 & -0.000702 & -0.036324 & 0.048116 \end{bmatrix}$$

b. Find the standard error of $\hat{\beta}_2$. $= \sqrt{6.310144 * 0.86947}$

c. Find the correlation between $\hat{\beta}_1$ and $\hat{\beta}_2$.$= 6.310144 * (-0.000030)$

8. In a small scale regression study, the following data were obtained:

Y	X_1	X_2	X_3
360	30	4	0
340	20	2	0
250	17	3	0
205.5	16	2	1
275.5	22	3	0

8a. (10 pts) Set up the design matrix X, the vectors Y, β and ϵ, for the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$

$$X = \begin{bmatrix} 1 & 30 & 4 & 0 \\ 1 & 20 & 2 & 0 \\ 1 & 17 & 3 & 0 \\ 1 & 16 & 2 & 1 \\ 1 & 22 & 3 & 0 \end{bmatrix}, Y = \begin{bmatrix} 360 \\ 340 \\ 250 \\ 205.5 \\ 275.5 \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix}$$

8b. (10 pts) Find $\hat{\beta}$ and the hat matrix H. Round them to 2 decimal places.

$$\hat{\beta} = (X^t X)^{-1} X^t Y$$

$$H = X(X^t X)^{-1} X^t Y$$

We need software to be able to do these computations. (this exam was for a year when we had a classroom with computers)