University of California, Los Angeles
Department of Statistics

Statistics 13

Instructor: Nicolas Christou

Exam 1
25 January 2016

Name: SOLUTIONS UCLA ID: _____ Section: _____

Problem 1 (25 points)
Answer the following questions:

a. In simple regression we use n pairs of $(y_i, x_i), i = 1, \ldots, n$. Suppose one pair is deleted from the data set. Give an expression of the new sample mean of x in terms of the old sample mean $\bar{x}$ and the deleted x_i. Your answer should be in the form $\bar{x}_{\text{new}} = a\bar{x} + bx_i$.

$$(n-1)\bar{x}' + x_i = n\bar{x}$$

$$\bar{x}' = \frac{n}{n-1}\bar{x} - \frac{1}{n-1}x_i$$

b. Refer to question (a). Let's denote $S_{XY} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ and $S_{XX} = \sum_{i=1}^{n}(x_i - \bar{x})^2$. It can be shown that after we delete pair (y_i, x_i) from the original data set that the new quantities of S_{XY} and S_{XX} are given by: $S_{XY\text{new}} = S_{XY} - \frac{n}{n-1}(x_i - \bar{x})(y_i - \bar{y})$ and $S_{XX\text{new}} = S_{XX} - \frac{n}{n-1}(x_i - \bar{x})^2$. Note: S_{XY} and S_{XX} are the quantities from the original full data set. Find an expression of the new $\hat{\beta}_1$ in terms of $\hat{\beta}_1$ of the full data set and the leverage value h_{ii}. Note: h_{ii} is the leverage value associated with the deleted point i.

$$\hat{\beta}_1' = \frac{S_{XY\,\text{new}}}{S_{XX\,\text{new}}} = \frac{S_{XY} - \frac{n}{n-1}(x_i - \bar{x})(y_i - \bar{y})}{S_{XX} - \frac{n}{n-1}(x_i - \bar{x})^2}$$

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{XX}} \qquad (x_i - \bar{x})^2 = S_{XX}\left[h_{ii} - \frac{1}{n}\right]$$

$$= \frac{S_{XY} - \frac{n}{n-1}(x_i - \bar{x})(y_i - \bar{y})}{S_{XX}\,\frac{n(1-h_{ii})}{n-1}} = \frac{(n-1)S_{XY} - n(x_i - \bar{x})(y_i - \bar{y})}{S_{XX}\,n(1-h_{ii})}$$

$$\frac{n-1}{n(1-h_{ii})}\hat{\beta}_1 - \frac{1}{1-h_{ii}}\frac{(x_i - \bar{x})(y_i - \bar{y})}{S_{XX}}.$$

c. Consider the simple regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. Assume $E(\epsilon_i) = 0$ and $var(\epsilon_i) = \sigma^2$. In addition, assume that x_i is not random, and of course the parameters β_0 and β_1 are not random as well. Find $E(y_i)$ and $var(y_i)$.

$$E Y_i = E(\beta_0 + \beta_1 x_i + \epsilon_i) = \beta_0 + \beta_1 x_i$$

$$VAR(Y_i) = VAR(\beta_0 + \beta_1 x_i + \epsilon_i) = VAR(\epsilon_i) = \sigma^2$$

d. Refer to question (c). Define $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$. Find $E(\bar{y})$ and $var(\bar{y})$.

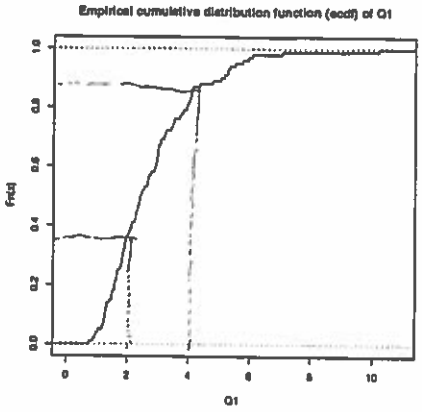
$$\bar{y} = \frac{\sum(\beta_0 + \beta_1 x_i + \epsilon_i)}{n} = \frac{n\beta_0 + \beta_1 \sum x_i + \sum \epsilon_i}{n} = \beta_0 + \beta_1 \bar{x} + \frac{\sum \epsilon_i}{n}$$

$$E\bar{Y} = \beta_0 + \beta_1 \bar{x}$$

$$VAR(\bar{Y}) = \frac{\sigma^2}{n}.$$

Problem 2 (25 points)
Answer the following questions:

a. Consider the North Carolina SIDS (Sudden Infant Death Syndrome) data. The variable of interest here is the number of SIDS per 1000 births, $Q_1 = 1000\frac{\#\text{SIDS}+1}{\#\text{BIRTHS}}$ for each of the 100 counties in North Carolina. The graph below shows the ecdf of Q_1. Approximately how many counties have between 2-4 SIDS per 1000 births?



PROPORTION 2-4 : $0.83 - 0.38 \approx 0.45$

$0.45 \times 100 = 45$

APPROXIMATELY 45 COUNTIES

b. When the variable of interest is time until death or failure (e.g. guinea pigs' survival after time t when they were infected with a certain virus) we often work with the survival function instead of the ecdf. The survival function is defined as $S_n(t) = \frac{\#\text{survived}>t}{n}$ (number of guinea pigs survived after time t divided by the total number of guinea pigs when the experiment started). What is the connection between the ecdf and the survival function? Give a typical plot of the survival function.

$S_n(t) = 1 - F_n(x)$

c. Another measure of central tendency is the "trimmed" sample mean. This is computed by discarding the lowest and highest portions of the data after they are sorted from smallest to largest. In order to compute the trimmed sample mean, we discard the lowest and highest 10% of the data. Compute the trimmed sample mean of the following 20 values (they represent lead concentration of soil in parts per million (ppm)). DISCARD: 65, 72 AND 252, 462

226 462 75 72 237 73 96 81 219 244 81 148 158 141 132 252 65 148 76 110

$\Sigma x_i = 2245$ $\quad \bar{x}_{\text{TRIMMED}} = \frac{2245}{18} = 140.3125$

d. Consider the permutation test in simple regression for testing the hypothesis $H_0 : \beta_1 = 2$ against the alternative $H_a : \beta_1 \neq 2$. Suppose the table below is your original data on y and x:

y	x
11.16	11.64
8.90	9.78
10.18	11.87
14.63	13.17
8.71	7.39

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

$Y_i - 2X_i = \beta_0 + \beta_1 X_i - 2X_i + \varepsilon_i$

$Y_i - 2X_i = \beta_0 + (\beta_1 - 2) X_i + \varepsilon_i$

$Y-2X$
-12.12
-10.66
-13.56
~~-11.71~~
~~-15.07~~

Give a possible data set under the null hypothesis when the permutation test is performed.

→ SHUFFLE THESE VALUES

A POSSIBLE NEW SAMPLE IS:

-10.66
-13.56
-11.71
-12.12
-6.07

Problem 3 (25 points)
Answer the following questions:

a. Consider the simple regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. Suppose we want to regress the residuals (e_i) against the fitted values ($\hat{y}_i$). Find the estimates of the intercept and slope of this regression.

$e_i = \alpha_0 + \alpha_1 \hat{y}_i + \varepsilon_i$

$\hat{\alpha}_1 = \sum (e_i - \bar{e}) \hat{y}_i$

$= \sum e_i \hat{y}_i - \bar{e} \sum \hat{y}_i$

$= \sum e_i [\bar{y} + \hat{\beta}_1 (x_i - \bar{x})] - 0$

$= \bar{y} \sum e_i + \hat{\beta}_1 \sum e_i (x_i - \bar{x})$

$= 0$

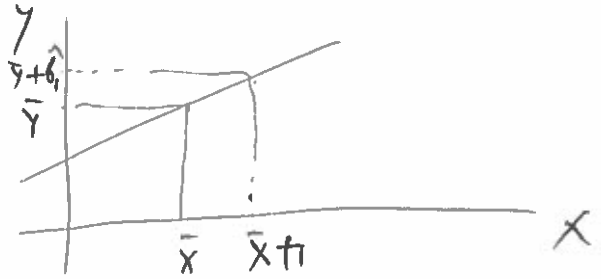
$\hat{\alpha}_0 = \bar{e} - \hat{\alpha}_1 \bar{\hat{y}} = 0$

b. Consider the simple regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. One of the goals of regression is to make predictions on y given a new x value. Is it true that when $x = \bar{x} + 1$ the predicted value of y will be $\bar{y} + \hat{\beta}_1$? Explain your answer mathematically.

$\hat{y}_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x})$

LET $x_i = \bar{x} + 1$

$\hat{y}_i = \bar{y} + \hat{\beta}_1$ YES



c. Suppose the price of two products A and B are random variables X and Y and they are independent. Suppose a person wants to buy three of product A and four of product B. Find the variance of this purchase.

$VAR(3X + 4Y) = 3^2 VAR(X) + 4^2 VAR(Y)$

$= 9\sigma_X^2 + 16\sigma_Y^2$

d. It can happen that the list of value $x_1, x_2, \ldots, x_n$ will involve duplications. Suppose that there are k different values and that we name them as $v_1, v_2, \ldots, v_k$. Let's say that v_1 occurs n_1 times, v_2 occurs n_2 times, and so on. Give a general formula that computes $\bar{x}$ in terms of the v_i and n_i values.

$\bar{X} = \frac{n_1 V_1 + n_2 V_2 + \cdots + n_k V_k}{n_1 + n_2 + \cdots + n_k} = \frac{\sum n_i V_i}{\sum n_i}$

Problem 4 (25 points)
Answer the following questions:

a. Consider the simple regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. In this example, y represents the concentration of lead in ppm and x represents the concentration of zinc in ppm of soil at a particular area of interest. The sample size is $n = 15$. These data gave the following results:
$\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y}) = 70767.08$, $\sum_{i=1}^{n}(y_i - \bar{y})^2 = 73327.6$.
$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 1560112$, $\sum_{i=1}^{n} x_i^2 = 5072016$, and $\bar{y} = 161.4$.

Find $\hat{\beta}_1$ and $\hat{\beta}_0$. $\sum[y_i - \bar{y}][\bar{y} + \hat{\beta}_1(x_i - \bar{x}) - \bar{y}] = 70767.08$

$\hat{\beta}_1 \sum(y_i - \bar{y})(x_i - \bar{x}) = 70767.08 \Rightarrow \hat{\beta}_1^2 \sum(x_i - \bar{x})^2 = 70767.08 \Rightarrow \hat{\beta}_1 = 0.2130$

$\sum(x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2 = 1560112 \Rightarrow 5072016 - 15\bar{x}^2 = 1560112 \Rightarrow \bar{x} = 483.67$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 161.4 - 0.2130(483.67) \Rightarrow \hat{\beta}_0 = 58.38$

b. Refer to question (a). Find R^2.

$R^2 = \frac{SSR}{SST} = \frac{\hat{\beta}_1^2 \sum(x_i - \bar{x})^2}{\sum(y_i - \bar{y})^2} = \frac{0.2130^2 (1560112)}{73327.6} = 0.9689$

c. For five individuals the blood platelet aggregation was measured before (x_1) and after (y_1) they smoked a cigarette. These five pairs of values gave the following results: $\bar{x}_1 = 30.2, \bar{y}_1 = 39.0, \text{cov}(x_1, y_1) = 88.5$. For another six individuals the blood platelet aggregation was also measured before (x_2) and after (y_2) they smoked a cigarette. These six pairs of values gave the following results: $\bar{x}_2 = 52.17, \bar{y}_2 = 63.67, \text{cov}(x_2, y_2) = 148.67$. Find the covariance of the combined 11 pairs.

$Cov(X,Y) = \frac{1}{n-1}\left[\sum XY - \frac{1}{n}(\sum X)(\sum Y)\right] = \frac{1}{n-1}\left[\sum XY - n\bar{X}\bar{Y}\right]$

$\sum XY = \sum X_1 Y_1 + \sum X_2 Y_2$ $\quad Cov(X_1,Y_1) = \frac{1}{n_1}\left[\sum X_1 Y_1 - n_1 \bar{X}_1 \bar{Y}_1\right] \Rightarrow \sum X_1 Y_1 = (n_1 - 1) Cov(X_1, Y_1) + n_1 \bar{X}_1 \bar{Y}_1$

SIMILARLY, $\sum X_2 Y_2 = (n_2 - 1) Cov(X_2, Y_2) + n_2 \bar{X}_2 \bar{Y}_2$

$\bar{X} = (n_1 \bar{X}_1 + n_2 \bar{X}_2)/(n_1 + n_2)$

$\bar{Y} = (n_1 \bar{Y}_1 + n_2 \bar{Y}_2)/(n_1 + n_2)$

$n_1 = 5, \quad n_2 = 6, \quad \bar{X}_1 = 30.2, \quad \bar{Y}_1 = 39.0$

$Cov(X_1, Y_1) = 88.5 \quad \bar{X}_2 = 52.17, \quad \bar{Y}_2 = 63.67$

$Cov(X_2, Y_2) = 148.67$

SOLVE TO FIND $\boxed{Cov(X,Y) = 257.5091}$

d. Suppose in the simple regression of y on x the units of y are hours and the units of x are inches. What are the units of the following quantities?

$\hat{\beta}_1$	R^2	e_i	h_{ii}	SSE	s_e
HOURS/INCH	NONE	HOURS	NONE	HOURS^2	HOURS