**STA 6126**
**Practice questions for exam 3**

The computer printout at the end refers to regression models for recent county-wide data in the state of Florida on $Y$ = CRIME (crime rate, measured as the number of crimes in past year per 1000 population), $X_1$ = HS (education, measured as the percentage of adult residents of that county having at least a high school education), and $X_2$ = URBAN (urbanization, measured as the percentage of residents of that county living in an urban environment). Problems 1-7 refer to the printout. You should be able to tell which model a question refers to by the wording of that question.

1. (a) For the prediction equation for the bivariate regression equation $E(Y) = \alpha + \beta X_1$, interpret carefully the slope estimate for HS.

   (b) For the prediction equation for the multiple regression equation $E(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2$, interpret carefully the partial slope estimate for $X_1$ = HS.

   (c) Explain carefully how the estimated effects of HS on CRIME could be so different in the bivariate and multiple regression models.

2. Give all steps for testing $H_0 : Y$ is independent of $X_1$, against the alternative of a <u>positive</u> bivariate association. Report the value of the test statistic, degrees of freedom, and $P$-value, and interpret.

3. Using the printout, report the value of

   (a) Estimated standard deviation of crime rate, ignoring other variables. ______________

   (b) Predicted change in crime rate for a 10% increase in urbanization. ______________

   (c) The Pearson correlation between education and urbanization. ______________

   (d) Predicted number of standard deviation change in $Y = CRIME$ for a one standard deviation change in $X_1 = HS$. ______________

   (e) Estimated standard deviation of crime rate, at fixed value for $HS$. ______________

For questions 4 and 5, worth 2 points for each part, indicate whether each statement is true (T) or false (F). Question 4 refers to the printout.

4. (a) _______ If $X_3$ = INCOME were added to the model, it is possible that the prediction equation $\hat{Y} = a + b_1 X_1 + b_2 X_2 + b_3 X_3$ could have $b_1 = b_2 = 0$.

(b) ______ If $X_3$ = INCOME were added to the model, $R^2$ could decrease compared to its value with only $X_1$ and $X_2$ in the model.

5. The following are general true – false questions about association and about regression not pertaining to the printout.

   (a) ______ If $r^2_{YX_1} = r^2_{YX_2} = .50$, it is possible that $R^2_{Y(X_1,X_2)} = .50$.

   (b) ______ If the $F$-test of $H_0 : \beta_1 = \beta_2 = 0$ gives $P < .05$, then necessarily both of $t = b_1/\hat{\sigma}_{b_1}$ and $t = b_2/\hat{\sigma}_{b_2}$ gives $P < .05$ for testing $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$.

   (c) ______ The ordinal measure of association called Gamma and the correlation are similar in that they can only take values between -1 and +1, with statistical independence of $X$ and $Y$ implying a value of 0.

   (d) ______ For a given set of data on two quantitative variables $X$ and $Y$, the slope of the least squares prediction equation and the correlation must have the same sign.

   (e) ______ *Simpson's paradox*, named after a statistician named O. J. Simpson, states that it is possible to find a linear prediction equation that goes exactly through every single point in a scatter diagram.

   (f) ______ There is said to be *interaction* between $X_1$ and $X_2$ in their effects on $Y$ if the following holds: $Y$ depends on $X_1$, which itself depends on $X_2$, so that there is a bivariate association between $Y$ and $X_2$ which completely disappears when we control for $X_1$.

```
data florida;
input county $  income unemp hs urban crime ;
income = income/1000; crime = crime*1000;
cards;
     ALACHUA  22084  47 82.7 73.21527 0.104035358
       BAKER  25816  93 64.1 21.45407 0.019504723
........
      WASHING 18266  80 60.9 22.85005 0.020642593
;
proc corr; var income unemp hs urban crime ;
proc reg; model crime = hs ;
proc reg; model crime = urban ;
proc reg; model crime = hs urban ;
run;
```

Simple Statistics

| Variable | N | Mean | Std Dev | Sum | Minimum | Maximum |
|---|---|---|---|---|---|---|
| INCOME | 67 | 24.5081 | 4.6850 | 1642.0 | 15.3800 | 35.6370 |
| UNEMP | 67 | 84.0448 | 24.0979 | 5631.0 | 40.0000 | 162.0 |
| HS | 67 | 69.4896 | 8.8588 | 4655.8 | 54.5000 | 84.9000 |
| URBAN | 67 | 49.5561 | 33.9725 | 3320.3 | 0 | 99.5974 |
| CRIME | 67 | 52.4205 | 28.2694 | 3512.2 | 0 | 128.2 |

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 67

| | INCOME | UNEMP | HS | URBAN | CRIME |
|---|---|---|---|---|---|
| INCOME | 1.00000 | -0.11906 | 0.79275 | 0.73029 | 0.43242 |
| | 0.0 | 0.3372 | 0.0001 | 0.0001 | 0.0003 |
| UNEMP | -0.11906 | 1.00000 | -0.25020 | -0.05310 | -0.00062 |
| | 0.3372 | 0.0 | 0.0411 | 0.6695 | 0.9960 |
| HS | 0.79275 | -0.25020 | 1.00000 | 0.79074 | 0.46771 |
| | 0.0001 | 0.0411 | 0.0 | 0.0001 | 0.0001 |
| URBAN | 0.73029 | -0.05310 | 0.79074 | 1.00000 | 0.67781 |
| | 0.0001 | 0.6695 | 0.0001 | 0.0 | 0.0001 |
| CRIME | 0.43242 | -0.00062 | 0.46771 | 0.67781 | 1.00000 |
| | 0.0003 | 0.9960 | 0.0001 | 0.0001 | 0.0 |

```
Dependent Variable: CRIME

                           Sum of          Mean
Source          DF        Squares        Square       F Value        Prob>F
Model            1    11537.75473   11537.75473        18.200        0.0001
Error           65    41206.56790     633.94720
C Total         66    52744.32263

    Root MSE        25.17831     R-square        0.2187
                                 Adj R-sq        0.2067

                    Parameter      Standard    T for H0:
 Variable   DF       Estimate         Error   Parameter=0    Prob > |T|
 INTERCEP    1     -51.292769   24.50469105        -2.093        0.0402
 HS          1       1.492502    0.34984916         4.266        0.0001
```

```
Dependent Variable: CRIME

                           Sum of          Mean
Source          DF        Squares        Square       F Value        Prob>F
Model            1    24232.04510   24232.04510        55.242        0.0001
Error           65    28512.27752     438.65042
C Total         66    52744.32263

    Root MSE        20.94398     R-square        0.4594
                                 Adj R-sq        0.4511

                    Parameter      Standard    T for H0:
 Variable   DF       Estimate         Error   Parameter=0    Prob > |T|
 INTERCEP    1      24.469871    4.54852504         5.380        0.0001
 URBAN       1       0.564021    0.07588559         7.433        0.0001
```

Dependent Variable: CRIME

| Source | DF | Sum of Squares | Mean Square | F Value | Prob>F |
|---|---|---|---|---|---|
| Model | 2 | 24888.03642 | 12444.01821 | 28.590 | 0.0001 |
| Error | 64 | 27856.28621 | 435.25447 | | |
| C Total | 66 | 52744.32263 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 20.86275 | R-square | 0.4719 |
| | | Adj R-sq | 0.4554 |

| Variable | DF | Parameter Estimate | Standard Error | T for H0: Parameter=0 | Prob > \|T\| |
|---|---|---|---|---|---|
| INTERCEP | 1 | 58.928049 | 28.43156603 | 2.073 | 0.0422 |
| HS | 1 | -0.581364 | 0.47355547 | -1.228 | 0.2241 |
| URBAN | 1 | 0.683896 | 0.12348568 | 5.538 | 0.0001 |

# FORMULAS

## Bivariate regression models

$E(Y) = \alpha + \beta X \qquad \hat{Y} = a + bX \qquad r = b(s_X/s_Y) \qquad r^2 = (TSS - SSE)/(TSS)$

$$b \pm t\hat{\sigma}_b \qquad t = \frac{b}{\hat{\sigma}_b} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \quad (df = n-2), \quad \hat{\sigma}_b = \hat{\sigma}/\sqrt{\Sigma(x-\overline{x})^2} = \hat{\sigma}/s_x\sqrt{n-1}$$

$$\hat{\sigma} = \sqrt{SSE/(n-2)} = \text{Root MSE}$$

## Multiple regression models

$E(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \qquad \hat{Y} = a + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k$

$R^2 = (TSS - SSE)/(TSS) \qquad TSS = \sum(Y - \bar{Y})^2 \qquad SSE = \sum(Y - \hat{Y})^2$

$$F = \frac{R^2/k}{(1-R^2)/[n-(k+1)]} = \text{MS(model)/MSE} \qquad df_1 = k,\ df_2 = n-(k+1)$$

$$t = b_i/\hat{\sigma}_{b_i} \quad df = n-(k+1) \qquad b_i \pm t\hat{\sigma}_{b_i}$$

**Answers:**

1. a. $b$ = 1.49. We estimate that, on the average, for a 1% increase in the county's percentage of residents with at least a high school education, crime increases by 1.49 crimes per 1000 residents.

b. $b_1$ = -0.58. Controlling for urbanization, we estimate that on the average, for a 1% increase in the county's percentage of residents with at least a high school education, crime decreases by .58 crimes per 1000 residents.

c. Simpson's paradox. The strong correlation of .79 between HS and URBAN and .68 between CRIME and URBAN explains this. More highly urbanized counties tend to have both more crime and higher percents of high school graduates.

2. $H_0 : \beta = 0$, $H_a : \beta > 0$. Test statistic $t = b/(stderror) =$ 1.49/0.35 = 4.27, $df = n - 2 = 65$, P-value = 0.0001/2 for one-sided alternative. Very strong evidence of a positive association between CRIME and HS.

3. a. $s_y$ = 28.269, b. 10(0.564) = 5.64, c. 0.79, d. 0.47 (This is the correlation), e. $\hat{\sigma}$ = root MSE = 25.18.

4. a. T, b. F (R-squared cannot decrease when variables are added)

5. a. T (If $X_1$ and $X_2$ are perfectly correlated)
b. F (not if there is multicollinearity)
c. T
d. T
e. F
f. F (This is a chain relationship)