1

A

$518.4 - 5.61 \times 22 = 394.98$

The regressions prediction for that classrooms average test score is 394.98.

B

The slope -5.61 indicates that the value of TS(average test scores) decreases by 5.61 units for every increase in CS(class size) of 1 unit.

C

For $n = 22$, $df = n - 2 = 20$.

$Se = 1.78, R^2 = 0.18, n = 22$

$SST = 77.27805 = S_{yy}, SSR = 13.91005 = \frac{S_{xy}^2}{S_{xx}}$

For $\hat{b}_1 = \frac{S_{xy}}{S_{xx}} = -5.61$, $S_{xy} = -2.47951, S_{xx} = 0.4419804$

$\alpha = 0.1$, $t_{\alpha/2} = 1.724718$.

$t_{\alpha/2} \cdot \frac{S_e}{\sqrt{S_{xx}}} = 4.617816$

$-5.61 + 4.617816 = -0.992184, -5.61 - 4.617816 = -10.22782$

$(-10.22782, -0.992184)$ is a 90% confidence interval for the slope of the population regression line.

D

I would like to add the variable "the average time that students spend on homework" into the model to raise R^2.

E

H_0 : the coefficient on class size is -4;

H_1 : the coefficient on class size is not -4.

$t = \frac{-5.61-(-4)}{\frac{S_e}{\sqrt{S_{xx}}}} = \frac{-1.61}{\frac{1.78}{\sqrt{0.4419804}}} = -0.6013224$

for df=20, P-value=0.5543804

Since P-value is large, there is no significant reason to reject H_0.

2

Suppose $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

Since $\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}$, and $\bar{X} = \bar{Y} = 0$. $\hat{\beta}_0 = 0$

$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2} = \frac{\sum(X_iY_i)}{\sum X_i^2} = \frac{5}{3}$

The OLS estimates of the regression is $\hat{Y} = \frac{5}{3}\hat{X}$

3

The answer is shown in 4.

4

Suppose $y_i = \alpha + \beta x_i + \varepsilon_i, \hat{y} = a + bx_i$.

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

$$\begin{aligned} E(b) &= E(\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}) \\ &= E(\frac{\sum (x_i y_i) - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}}) \end{aligned}$$

$$\begin{aligned} x_i y_i &= \alpha x_i + \beta x_i^2 + \varepsilon_i x_i \\ \sum (x_i y_i) - n\bar{x}\bar{y} &= n\alpha\bar{x} + \beta \sum x_i^2 + \sum \varepsilon_i x_i - n\bar{x}(\alpha + \beta\bar{x} + \bar{\varepsilon}) \\ &= \beta(\sum x_i^2 - n\bar{x}) + \sum \varepsilon_i x_i - n\bar{\varepsilon} \end{aligned}$$

$$\begin{aligned} E(b) &= \frac{\beta(\sum x_i^2 - n\bar{x}) + \sum \varepsilon_i x_i - n\bar{\varepsilon}}{\sum x_i^2 - n\bar{x}} \\ &= \beta + E(\frac{\sum \varepsilon_i x_i - n\bar{\varepsilon}}{\sum x_i^2 - n\bar{x}}) \\ &= \beta \end{aligned}$$

$a = \bar{y} - b\bar{x}$

$$\begin{aligned} E(a) &= E(\bar{y}) - E(b\bar{x}) \\ &= E(\alpha + \beta\bar{x} + \bar{\varepsilon}) - E((\beta + \frac{\sum \varepsilon_i x_i - n\bar{\varepsilon}}{\sum x_i^2 - n\bar{x}})\bar{x}) \\ &= \alpha + \beta\bar{x} - \beta\bar{x} \\ &= \alpha \end{aligned}$$

5

R^2 from the regression of Y on X:

$$R_a^2 = \frac{SSR_a}{SST_a} = \frac{\frac{[\sum x_i y_i - (\sum x_i)(\sum y_i)/n]^2}{\sum x_i^2 - (\sum x_i)^2/n}}{\sum y_i^2 - (\sum y_i)^2/n}$$

R^2 from the regression of X on Y:

$$R_b^2 = \frac{SSR_b}{SST_b} = \frac{\frac{[\sum x_i y_i - (\sum x_i)(\sum y_i)/n]^2}{\sum y_i^2 - (\sum y_i)^2/n}}{\sum x_i^2 - (\sum x_i)^2/n} = R_a^2$$

6

A

If the data is from one of these four regions, then if all four region variables are included in the model, the model will be wrong because of the collinearity.

B

One could pick the one that he/she thins makes the most interesting contrast as the reference category.

C

The new estimated regression will be in the form like $\hat{Y} = a + b_1 X + b_2 NE + b_3 MW + b_4 South$.

7

A

"a" indicates the income of a female whose years of education is 0;

"b_1" indicates that income increase by b_1 units for any increase in years of education of 1 unit.

"b_2" indicates that if a person is male he will earn b_2 more compare to if the person is female.

B

$H_0 : b_2 = 0$;
$H_1 : b_2 \neq 0$.

C

$$\hat{Inc} = a + b_{1m}Educ + b_{1f}Educ + b_2Male$$

D

The model assumes that the nature of the relationship between education and income is linear.

E

If more able individuals have higher incomes and are more likely to have more education, the actual relationship between education and income is nonlinear. Then the linear assumption of the model above will make the model less accurate.

F

Assign b1 with discrete values such as $0, 1, 2, \cdots$ and each value represents an education category(the number of values should be 1 less than the number of education levels, the last education category would be assigned to no value).

8