*YOUR TA’S NAME*:

*Lecture Worksheet*

*Tuesday 12/7/2020*

**MAIN POINTS OF LECTURE**

1. Multiple Regression Analysis is a statistical technique for estimating the relationship between a continuous dependent variable and two or more continuous or discrete independent, or predictor, variables
2. The multivariate (2 independent variable) regression prediction equation is $\hat{Y}\_{i}=a+b\_{1}X\_{1i}+b\_{2}X\_{2i}$, where a is the estimated intercept, X1 and X2 are independent variables, and b1 and b2 are estimated slopes (or coefficients)
3. For the multivariate (2 independent variable) regression prediction equation, the intercept and slopes are estimated as:

$b\_{1}=\left(\frac{s\_{Y}}{s\_{X\_{1}}}\right)\frac{r\_{YX\_{1}}-r\_{YX\_{2}}r\_{X\_{1}X\_{2}}}{1-r\_{X\_{1}X\_{2}}^{2}}$ ; $b\_{2}=\left(\frac{s\_{Y}}{s\_{X\_{2}}}\right)\frac{r\_{YX\_{2}}-r\_{YX\_{1}}r\_{X\_{1}X\_{2}}}{1-r\_{X\_{1}X\_{2}}^{2}}$ ; $a=\bar{Y}-\left(b\_{1}\bar{X}\_{1}+b\_{2}\bar{X}\_{2}\right)$

1. Interpretations:
* The intercept, a, is interpreted as the predicted value of Y when both X1 and X2 equal 0
* Regression coefficient (or slope) b1 is interpreted as the expected change in Y associated with a one unit increase in X1, controlling for X2
* Regression coefficient (or slope) b2 is interpreted as the expected change in Y associated with a one unit increase in X2, controlling for X1
1. R2 expresses the proportion of variation in Y that is accounted for by the predictor variables, and is calculated as $R\_{Y•X\_{1}X\_{2}}^{2}=\frac{r\_{YX\_{1}}^{2}+r\_{YX\_{2}}^{2}-2r\_{YX\_{1}}r\_{YX\_{2}}r\_{X\_{1}X\_{2}}}{1-r\_{X\_{1}X\_{2}}^{2}}$. R2 is an estimate of population parameter 2Y•X1X2
* The hypothesis test for 2Y•X1X2 is an F test with dfNUM=2 (the number of predictors in the model) and dfDENOM=N—3 (N-1 minus the number of predictors in the model)
* The test statistic for hypothesis tests about 2Y•X1X2 is $F\_{2,N-3}=\frac{SS\_{REGRESSION}/2}{SS\_{ERROR}/N-3}=\frac{MS\_{REGRESSION}}{MS\_{ERROR}}$, where $SS\_{TOTAL}=(s\_{Y}^{2})(N-1)$, $SS\_{REGRESSION}=(R\_{Y•X\_{1}X\_{2}}^{2})(SS\_{TOTAL})$, and $SS\_{ERROR}=SS\_{TOTAL}-SS\_{REGRESSION}$
1. b1 and b2 are estimates of population parameters 1 and 2
* Hypothesis tests for 1 and 2 are t tests with N-3 degrees of freedom (because MSERROR has N-3 degrees of freedom when there are two predictor variables)
* The tests statistics for hypothesis tests about 1 and 2 are:
* For 1: $t\_{N-3}=\frac{b\_{1}-0}{s\_{b\_{1}}}=\frac{b\_{1}-0}{\sqrt{\frac{MS\_{ERROR}}{\left(s\_{X\_{1}}^{2}\right)\left(N-1\right)\left(1-R\_{X\_{1}•X\_{2}}^{2}\right)}}}$
* For 2: $t\_{N-3}=\frac{b\_{2}-0}{s\_{b\_{2}}}=\frac{b\_{2}-0}{\sqrt{\frac{MS\_{ERROR}}{\left(s\_{X\_{2}}^{2}\right)\left(N-1\right)\left(1-R\_{X\_{2}•X\_{1}}^{2}\right)}}}$

**QUESTIONS**

**From the recorded lecture**

Example: *How is income affected by education and IQ?*

Y = The adult income of 1,000 people (in $1,000s)

X1 = The number of years of school they completed

X2 = Their IQ

 Y X1 X2 Mean SD

Y 1.00 35.0 12.0

X1 0.50 1.00 12.0 3.0

X2 0.30 0.60 1.00 100.0 15.0

1. **Compute and interpret the intercept and slopes of the multiple regression prediction equation**

$$\hat{Y}\_{i}=11+2X\_{1i}+0X\_{2i}$$

$$b\_{2}=\left(\frac{s\_{Y}}{s\_{X\_{2}}}\right)\frac{r\_{YX\_{2}}-r\_{YX\_{1}}r\_{X\_{1}X\_{2}}}{1-r\_{X\_{1}X\_{2}}^{2}}=\left(\frac{12.0}{15.0}\right)\frac{0.30-(0.50)(0.60)}{1-0.60^{2}}=0.0$$

$$b\_{1}=\left(\frac{s\_{Y}}{s\_{X\_{1}}}\right)\frac{r\_{YX\_{1}}-r\_{YX\_{2}}r\_{X\_{1}X\_{2}}}{1-r\_{X\_{1}X\_{2}}^{2}}=\left(\frac{12}{3}\right)\frac{0.50-(0.30)(0.60)}{1-0.60^{2}}=2.0$$

$$a=\bar{Y}-\left(b\_{1}\bar{X}\_{1}+b\_{2}\bar{X}\_{2}\right)=35-\left[\left(2\right)\left(12\right)+\left(0\right)\left(100\right)\right]   =11$$

When both years of schooling (X1) and IQ (X2) are 0.0, the expected value of income (Y) is the intercept … 11.

Holding constant IQ (X2) — that is, among people with the same value of IQ (X2) — a one unit increase in years of schooling (X1) is associated with a 2.0 increase in income (Y).

Holding constant years of schooling (X1) — that is, among people with the same value of years of schooling (X1) — a one unit increase in IQ (X2) is associated with a 0.0 increase in income (Y).

1. **Test the hypothesis that 2Y•X1X2 = 0 … or, that X1 and X2 explain none of the variability in Y** (Note: R2Y•X1X2 = 0.25); **use  = 0.05**

Thus, we reject H0 that 2Y•X1X2 = 0. The two X variables do explain some of the variation in Y in the population.

$$R\_{Y•X\_{1}X\_{2}}^{2}=\frac{r\_{YX\_{1}}^{2}+r\_{YX\_{2}}^{2}-2r\_{YX\_{1}}r\_{YX\_{2}}r\_{X\_{1}X\_{2}}}{1-r\_{X\_{1}X\_{2}}^{2}}=\frac{0.5^{2}+0.3^{2}-2(0.5)(0.3)(0.6)}{1-0.6^{2}}=0.25$$

$$SS\_{ERROR}=SS\_{TOTAL}-SS\_{REGRESSION}=143,856-35,964=107,892$$

$$SS\_{REGRESSION}=(R\_{Y•X\_{1}X\_{2}}^{2})(SS\_{TOTAL})=(0.25)(143,856)=35,964$$

$$Critical value = F\*=3.04$$

$$F\_{2,N-3}=\frac{SS\_{REGRESSION}/2}{SS\_{ERROR}/N-3}=\frac{35,964/2}{107,892/997}=166.167$$

$$SS\_{TOTAL}=(s\_{Y}^{2})(N-1)=(144)(1000-1)=143,856$$

1. **Test the hypotheses that 1 and 2 equal zero** (Note: Use b1 and b2 from above; MSerror=108.2 and R2Y•X1X2 = 0.25); **use  = 0.05**

$$Critical value = t\*=1.984$$

$$t\_{N-3}=\frac{b\_{2}-0}{s\_{b\_{2}}}=\frac{b\_{2}-0}{\sqrt{\frac{MS\_{ERROR}}{\left(s\_{X\_{2}}^{2}\right)\left(N-1\right)\left(1-R\_{X\_{2}•X\_{1}}^{2}\right)}}}=\frac{0}{0.027}=0$$

$$t\_{N-3}=\frac{b\_{1}-0}{s\_{b\_{1}}}=\frac{b\_{1}-0}{\sqrt{\frac{MS\_{ERROR}}{\left(s\_{X\_{1}}^{2}\right)\left(N-1\right)\left(1-R\_{X\_{1}•X\_{2}}^{2}\right)}}}=\frac{2}{0.137}=14.6$$

Given that the value of t for b1 exceeds |1.984|, we reject the null hypothesis that b1=0. However, the value of t for b2 does not exceed |1.984| so we fail to reject the null hypothesis that b2=0.

**From the synchronous session**

1. In the example we worked through in the synchronous class session: What was the zero-order correlation between X1 and Y? What was the correlation between X1 and Y after controlling for X2? How do we interpret the latter conditional correlation?

The zero-order correlation—the correlation before we control for anything—was 0.15. The correlation after controlling for X2 was 0.08. This means that about half of the association between X1 and Y is accounted for by controlling for X2.