*YOUR TA’S NAME*:

*Lecture Worksheet*

*Thursday 12/3/2020*

**MAIN POINTS OF LECTURE**

1. MAIN TOPIC: Models with *k* continuous predictor variables (X1 through Xk) and a continuous response variable
	1. Sample Prediction Equation:
	2. The ordinary least squares (OLS) method is used to estimate a and b1 through bj … again, this method minimizes the sum of the squared prediction errors
	3. The intercept, *a*, equals the predicted value of Y when each of the k predictor variables (X1 through Xj) equal 0
	4. Multiple regression coefficient *b*k represents the expected change in Y associated with a one unit increase in *X*j, controlling for all other predictors in the model
	5. R2 expresses the proportion of variation in Y that is accounted for by the predictor variables taken as a whole
	6. We use R2Y•X1…Xk to estimate 2Y•X1…Xk; Hypothesis tests about 2Y•X1…Xk are F tests with dfnum=k and dfdenom= n-k-1. The test statistic is the same as for two variable regression (except for the different numbers of degrees of freedom)
	7. We use bk to estimate k; Hypothesis tests about k are t tests with n-k-1 degrees of freedom. The test statistic is 
2. BONUS TOPIC #1: Call the model that contains the full set of X variables the complete model; it has k2 impendent variables. Call a model that contains a subset of those X variables the reduced model; it contains k1 independent variables. Question: Does the addition of the k2-k1 new predictor variables in the complete model improve our ability to predict Y (relative to the reduced model)? Answer: Test the null hypothesis that the additional variables explain no additional variation in Y. This test is an F test, with dfnum=k2-k2, dfdenom=n-k2-1, and 
3. BONUS TOPIC #2: In a model with a discrete independent variable X that has j categories, X should be represented by a series of j-1 “dummy variables” that indicate whether individuals belong to categories of X. This is directly analogous to ANOVA.
4. BONUS TOPIC #3: Interaction terms … a strategy for allowing the effect of X1 on Y to vary across levels of X2 and simultaneously allowing the effect of X2 on Y to vary across levels of X1 … can be modeling by adding a new variable that equals X1×X2

**QUESTIONS**

**From the recorded lecture**





1. Interpret the intercept, the slopes, and R2

Intercept: When the value of all the X variables is 0, the predicted value of Y is 53.93007

Slope for ProblemSets: Among people who are the same with respect to the other X variables, a 1 unit increase in Problem Set scores is associated with a 0.167 increase in Exam scores.

Slope for InClass: Among people who are the same with respect to the other X variables, a 1 unit increase in In Class scores is associated with a 0.088 decline in Exam scores.

Slope for InLab: Among people who are the same with respect to the other X variables, a 1 unit increase in In Lab scores is associated with a 0.089 increase in Exam scores.

R2: The three X variables in the model explain 25.72% of the variation in Exam Scores.

1. Test the hypothesis that 2=0; use =0.05

R2 = 0.2572

The critical value of F with =0.05, dfNUM = k = 3 and dfDENOM = n-k-1 = 155-3-1 = 151 is 2.70

$$SS\_{REGRESSION}=(R\_{}^{2})(SS\_{TOTAL})=(0.2572)(21701.779)=5,581.7$$

$$SS\_{TOTAL}=(s\_{Y}^{2})(N-1)=(11.871^{2})(155-1)=21,701.779$$

$$SS\_{ERROR}=SS\_{TOTAL}-SS\_{REGRESSION}=21,701.779-5,581.7 = 16,120=107,892$$

$$F\_{K,N-K-1}=\frac{SS\_{REGRESSION}/K}{SS\_{ERROR}/N-K-1}=\frac{5,581.7/3}{16,120/151}=17.4$$

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

1. Test the hypothesis that ProblemSets=0; use a=0.05

The critical value of t with a=0.05 and N-K-1 = 151 degrees of freedom is 1.984

The value of t is just the coefficient (0.167) divided by the standard error (0.032) from the regression table above. That means t = 0.167/0.032 = 5.219

We thus reject the null hypothesis that the slope for this variable equals 0 in the population