K300 (4392) Statistical Techniques (Fall 2007)

Assignment 8: Linear Regression Models (155 points, Due December 5)

Instructor: Hun Myoung Park
kucc625@indiana.edu, (317) 274-0573

Please first read the following instructions and questions carefully.

- Do not use a word processor or other computer software packages.
- Explicitly indicate question numbers (e.g., Q1.2, Q2.4, etc.).
- Hand in this assignment **by Wednesday, December 5**. Due to the final exam, **late assignment WILL NOT BE ACCEPTED** after the due date. Answer key will be released after 5:00 P.M. on December 5.
- You **MAY NOT discuss with other classmates** in any circumstance when answering questions. If you have any problem with any of the questions, just talk to me.

The dependent variable Y (`owncar`) in your model is the propensity (simply probability) that a college student will own his car. This is a continuous variable. There are three independent variables (regressors) X_1 through X_3. X_1 (`offcamp`) is a binary variable or dummy variable, which is set as 1 if a student lives off campus, 0 otherwise. X_2 (`income`) is the amount of money that a student can spend per month; I guess this is the sum of his/her salary and money that he/she receives from his/her parents or relatives. This disposable income is measured in \$1,000 (1 means \$1,000.00). Obviously, X_2 is a continuous, more specifically ratio, variable. Finally, X_3 (`male`) is set 1 for male students and 0 for female students. Of course, X3 is another dummy variable. Research question here is "*What are important factors that determine whether a college student has his/her own car*."

Question 1. (95 points) You are regressing Y on X_1 and X_2 first. Let us call it **Model 1**. See the first SPSS output attached below.

Q1.1 (5 points) Write down your linear regression model. You need to use Greek letters βs. Do not forget to add ε (not *e*) to the model. See Question 4.1 of assignment 7 to get some ideas.

Q1.2 (10 points) Report R^2. Can you show me how R^2 is computed? How would you like to say about the goodness-of-fit of this model using this information? **You need to interpret R^2 substantively** by examining the proportion of SSM of the total variance. Does your model fit the data well? See slides if you have no idea.

Q1.3 (15 points) Report the F statistic and its p-value. Test the null hypothesis and draw a conclusion. You must follow all five steps as you did in assignment 7 (see Question 4.11). Pay attention to the alternative hypothesis. How would you like to say about the goodness-of-fit of this model on the basis of this hypothesis test?

Q1.4 (5 points) Compare your conclusions of Q1.2 and Q1.3. Are they consistent or not? Which conclusion do you think is more plausible or appealing? And why?

Q1.5 (5 points) Report b_2 (estimator of parameter β_2) and test the null hypothesis of $\beta_2=0$ (not $b_2=0$). See Q4.5 of assignment 7; do not omit any one of five steps. The p-value approach will do. Do you think X_2 (`income`) is an important determinant of student's car ownership?

Q1.6 (10 points) Interpret b_2 substantively. (In fact, if β_2 turns out zero in Q1.5, you do not need to interpret b_2. But **please do so** regardless of the result of Q1.5; this is an exercise.). You need to add the scale ($1,000) to make it clear and add the p-value in parentheses at the end of the sentence; *for $1,000.00 increase in … (p<.xxx)*. See the slides for hints.

Q1.7 (5 points)) Report b_1 (estimator of parameter β_1) and test the hypothesis of $\beta_1=0$ (not $b_1=0$). See Q4.5 of assignment 7; do not omit any one of five steps. Do you conclude that X_1 (`offcamp`) is an important determinant of student's car ownership?

Q1.8 (10 points) Write down two regression equations: one for off-campus students and the other for on-campus students. Equations should contain X_2 without X_1. Show how you got these equations. Coefficient b_1 needs to be incorporated into the intercept. See slides for an example.

Q1.9 (5 points) Draw the two regression lines on a plot. Write down regression equations near the proper regression lines. See slides for an example.

Q1.10 (10 points) Interpret b_1 substantively. For example, the coefficient of `male` can be interpreted as "*A male student is bbb percent more likely to own a car than female students, holding other variables X_1 and X_2 constant*" or "*The probability that a male student owns a car is about bbb percent higher than that of female students, holding other variables X_1 and X_2 constant (p<.ppp)*." Note that you may or may not add the p-value at the end of the sentence.

Q1.11 (5 points) Go back to Q1.7 and Q1.10. Would you like to conclude that students who live off campus have a significantly different intercept (or significant vertical distance between two regression lines) than those who live on campus? Note that this is another way to interpret the coefficient of a dummy variable.

Q1.12 (10 points) Consider Q1.2, Q1.3, Q1.4, Q1.5, and Q1.7. How would you say about your model? I want you to evaluate your model as a whole? Is your Model 1 a *lemon* or *peach*?

Question 2. (60 points) Now, you come across that gender may make a big difference in predicting college student's car ownership. That is, you want to add a regressor `male` (X_3) to Model 1. Let us call it **Model 2**, which regresses Y on X_1, X_2, and X_3. See the second SPSS output below.

Q2.1 (5 points) Write down this linear regression model, Model 2, as you did in Q1.1 above. You should use Greek letters.

Q2.2 (5 points) Report and compare R^2 of two models: Model 1 and 2. Which one is larger? Is there any big difference? Based on this result, which model do you prefer, and why? (Hint: adding any regressor will increase R^2 somehow).

Q2.3 (5 points) Report and compare adjusted R^2 of two models. Which one is larger? Is there any big difference? Which model do you prefer, and why?

Q2.4 (10 points) Report and compare F statistics and their p-values of two models. Which F statistic is larger? Is there any big difference in the p-value of these models? (Keep in mind that F statistics, in fact, are not comparable in a strict sense because they have different degrees of freedom. However, their p-values are comparable. This is a reason why I emphasized the p-value approach over test statistic and confidence interval approaches). Which model do you think looks better? And why? Note that there is no single answer for this question.
Q2.5 (5 points) Report and compare b_1 and b_2 and their p-values of two models. Ignore the intercept and b_3. Is there any big difference?
Q2.6 (5 points) Report and compare SSE of two models. Compute SSE_2-SSE_1 and report the result. This is the change in error variance component when adding X_3 to Model 1. (The positive sign indicates increase in error variance component, while the negative sign means reduction in error variance component.) Note that SSE_2 is the sum of squares due to error of Model 2.
Q2.7 (5 points) Report and compare degrees of freedom of SSE of two models. Which one is larger? (Hint: adding regressors ends up with loss of degrees of freedom).
Q2.8 (5 points) Report and compare SSM of two models. Compute SSM_2-SSM_1 and report the result. This is the increase or decrease of variance of Y that the model can explain when adding X_3 to Model 1. Compare this difference with one you got in Q2.6. Can you get what happened in SSM and SSE when adding a regressor to a model?
Q2.9 (5 points) Report and compare SST of two models. Is there any difference? (Hint: Q2.6 and Q2.8 should report the same difference; one is plus and the other is minus.) Again, you should understand that how adding a regressor changes the partition of variance components.
Q2.10 (10 points) Consider Q2.1 through Q2.9 and then decide if adding a variable to Model 1 is valuable in terms of improving goodness-of-fit (see Q2.6). You just need to eyeball two models (Yes, there are formal ways to test this difference, but the test is not required in K300). If there is, according to your subjective criterion, a large reduction of error variance or a large increase of variance of Y that the model can explain, addition, Model 2, deserves improvement of goodness-of-fit at the expense of one degree of freedom. Otherwise, you need to take a parsimonious model, Model 1, assuming that two models do not have a big difference in terms of goodness-of-fit. Note that there is no single answer. I want to check if you are able to evaluate linear models correctly and justify your reasoning.

Regression

```
[DataSet1]
```

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	income, offcamp[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: owncar

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.229[a]	.052	.048	.466

a. Predictors: (Constant), income, offcamp

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5.219	2	2.610	12.021	.000[a]
	Residual	94.213	434	.217		
	Total	99.432	436			

a. Predictors: (Constant), income, offcamp

b. Dependent Variable: owncar

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.014	.153		.090	.929
	offcamp	.669	.136	.229	4.903	.000
	income	-.024	.125	-.009	-.190	.850

a. Dependent Variable: owncar

Regression

```
[DataSet1]
```

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	male, income, offcamp[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: owncar

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.246[a]	.061	.054	.464

a. Predictors: (Constant), male, income, offcamp

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6.027	3	2.009	9.313	.000[a]
	Residual	93.406	433	.216		
	Total	99.432	436			

a. Predictors: (Constant), male, income, offcamp

b. Dependent Variable: owncar

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	-.022	.153		-.147	.884
	offcamp	.655	.136	.225	4.810	.000
	income	-.024	.124	-.009	-.192	.847
	male	.087	.045	.090	1.935	.054

a. Dependent Variable: owncar