# 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.

### Instructor's comment

Some of you still appear to have difficulty interpreting coefficients of regressors and conducting hypothesis test for individual coefficients. PLEASE read through slides for OLS; it has ONLY 34 pages with BIG fonts!!! Let me hit the highlights.

- $b_0$ (or a) is an estimator of its parameter $\beta_0$, the Y intercept (a point where the regression line intersects Y-axis), and the value of Y when X=0.
- $b_1$, $b_2$, … are estimators of their parameters $\beta_1$, $\beta_2$, …, coefficients of regressors $X_1$, $X_2$, …, and slopes of the regression line with respect to the regressors of interest.
- Interpretation: *For 1 unit increase in X, Y will increase by b, holding all other variables constant*. All you need is to replace X with a specific regressor of interest and b with its coefficient. If the coefficient of regressor GNP (measured in $1 billion) is .234, you may interpret as "*For $1 billion increase in GNP, Y will increase by .234, holding all other variables constant.*" Note that $ 1 billion is 1 unit in GNP.

The difficulty you have when conducting hypothesis test comes from confusion about the test statistic approach and p-value approach. Some of you still fail to fully understand the concept of the p-value. PLEASE read the lecture note 1, which describes all necessary information regarding hypothesis testing. I am copying some key parts from the note.

- The *p-value* is the probability that you get even more unlike sample than test statistic. Technically speaking, the p-value is the area underneath the probability distribution curve from the test statistic to the (positive or negative) infinity.
- The *test size* or *significance level* ($\alpha$) is the probability of rejecting the null hypothesis that is true.

- This approach asks, "*What is the probability that we get a more unlikely (extreme or odd) test statistic if the null hypothesis is true?*" What does "a p-value is smaller than the significance level" mean? If a p-value is smaller than the significance level or falls into the rejection region, you have a smaller risk of being wrong than your subjective criterion (significance level) (if you reject the null hypothesis). Thus, you become more confident to reject the null hypothesis in favor of the alternative (research) hypothesis because it is less risky than you thought to reject the null hypothesis. Keep in mind that the .05 significance level, for example, means you are willing to take a risk of making a wrong conclusion by that amount. When you reject the null hypothesis, there is still 5 percent chance that your conclusion is wrong (rejecting the true null hypothesis).
- Put it differently, a small p-value, say .003, indicates that your sample mean is vary far away from the hypothesized mean. You just have only .3 percent chance to observe sample means less than the one you get if the hypothesized mean is true. Obviously, this is an extremely unlikely event. Therefore, it is less risky than you thought to reject the null hypothesis or conclude that the hypothesized mean is wrong.
- Which one is better than the others? It depends on researchers' purposes. The (classical) test statistic approach is easy to understand. The modern approach (confidence interval approach) does not require computing the test statistic and p-value. The p-value approach reports correct information about making a wrong conclusion. In general, p-value approach appears more useful than the other two approaches. Most statistical software packages tend to focus on the p-value approach although they produce the confidence interval as well.

Let us focus on the p-value approach. Statistical packages compute p-values for you. Once you get p-values, it is a piece of cake to test the null hypothesis. P-values alone provide sufficient information for decision-making. Of course, computing p-values is not easy in general and horrible in some cases. It is less smart if you try the test statistic approach and then look for a critical value when p-values are in your hand; why are you trying to take the worst approach?

| Step | P-value Approach | T-test for an Individual Parameter | F test for All Parameters at a Time |
|---|---|---|---|
| 1 | Determine $H_0$ and $H_a$ | $H_0: \beta_k = 0$<br>$H_a: \beta_k \neq 0$ | $H_0: \beta_0 = \beta_1 = ...\beta_k = 0$<br>$H_a:$ at least one parameter is not 0 |
| 2 | Determine the test size α | α=.05 | α=.05 |
| 3 | The test statistic and the p-value | t and its p-value from SPSS output | F and its p-value from SPSS output |
| 4 | Reject $H_0$ if p-value < α<br>Otherwise, do not reject $H_0$ | | |
| 5 | Substantive interpretation | $\beta_k = 0$ (not reject $H_0$)<br>$\beta_k \neq 0$ (reject $H_0$) | $\beta_0 = \beta_1 = ...\beta_k = 0$ (not reject $H_0$)<br>At least one parameter is not 0 (reject $H_0$) |

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, $X_3$ 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.

Answer: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

**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.

Answer: $R^2$ is .052=SSM/SST=5.219/99.432. This regression model can explain only 5.2 percent of total variation of Y. In other word, this model cannot explain 94.8 percent of total variation of Y. Therefore, this poor goodness-of-fit indicates that this model does not fit the data well. Student income and housing type (on/off campus) do not appear to explain student's car ownership well; other variables or chance can explain 94.8 percent of variation of Y.

**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?

Answer: F=12.021, p-value<.000.

1. $H_0: \beta_0 = \beta_1 = \beta_2 = 0$, alternative hypothesis: at least one parameter is not zero.
2. alpha=.05
3. F=12.021, p<.000
4. Since p<.05, reject the null hypothesis at the .05 level.
5. At least one parameter is not zero (p<.000).

This regression model appears to fit the data well since rejection of the null hypothesis indicates at least one “statistically significant” regressor in the model.

**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?

Answer: Not consistent. Since there are only two regressors, $R^2$ conveys good information about goodness-of-fit in terms of the proportion of variation of Y that the model can explain. The F test examines all regressors (independent variables) at a time to see if there is any statistically significant regressor. The F test tells us there is at least one regressor that can significantly explain the variation of Y. If you are interested in a particular regressor like `offcamp`, the F test will be more appealing than $R^2$. If you just want to know the extent of variation of Y that the model can explain, $R^2$ will give you the answer.

**Q1.5 (5 points)** Report $b_2$ (estimator of parameter $\beta_2$) 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?

Answer: $b_2$=-.024

1. $H_0: \beta_2 = 0$, $H_a: \beta_2 \neq 0$
2. alpha=.05
3. T=-.190=-.024/.125, p<.850
4. Since p>.05, do not reject the null hypothesis at the .05 level.
5. $\beta_2 = 0$ (p<.850).

$X_2$ (`income`) is not a good indicator of student’s car ownership since its impact on the ownership is zero, $\beta_2 = 0$. Then, what happened in $b_2$=-.024? We can say that sampling and other types of errors mistakenly (by chance) report some, although small in magnitude, effect of student’s income on the dependent variable (car ownership). Therefore, you should not be fooled by these errors.

**Q1.6 (10 points)** Interpret $b_2$ substantively. (In fact, if $\beta_2$ 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.

Answer: For $1,000.00 increase in student’s monthly income, the probability that a student will own his/her car will decrease by 2.4 percent, holding all other variables constant (p<.850). Yes, it is counterintuitive. The large p-value indicates that such interpretation is not reliable (too risky since you have 85% chance of being wrong).

**Q1.7 (5 points) )** Report $b_1$ (estimator of parameter $\beta_1$) 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?

Answer: $b_1$=.669

1. $H_0 : \beta_1 = 0$, $H_a : \beta_1 \neq 0$
2. alpha=.05
3. T=4.903=.669/.136, p<.000
4. Since p<.05, reject the null hypothesis at the .05 level.
5. $\beta_1 \neq 0$ (p<.000).

$X_1$ (`offcamp`) is a good indicator of student's car ownership. The positive sign indicates that students who live off campus have a higher likelihood of owning their cars than those who live on campus. Yes, it is quite intuitive.

**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.

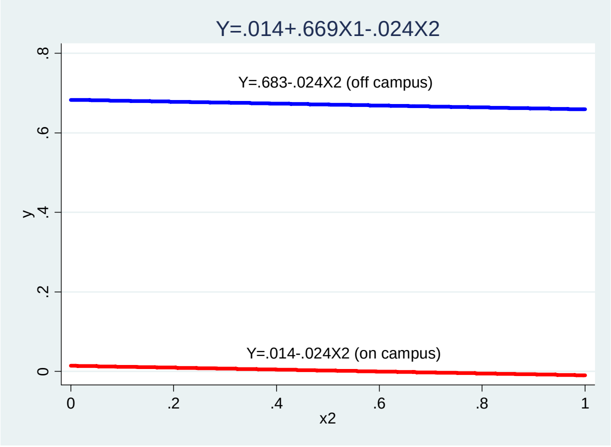
Answer: regression equation is $Y = .014 + .669X_1 - .024X_2$

$Y = .014 + .669 \times 1 - .024X_2 = .683 - .024X_2$ for male ($X_1$=1)

$Y = .014 + .669 \times 0 - .024X_2 = .014 - .024X_2$ for female ($X_1$=0)

**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.

Answer:



**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.

Answer: A student who lives off campus is 66.6 percent more likely to own a car than a student who lives on campus, holding all other variables constant.

**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.

Answer: Yes. The distance of .666 is significantly large; two groups have obviously different intercepts (p<.000).

**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*?

Answer: Despite the poor $R^2$, this model gives us some useful information about the relationship between housing type (on/off campus) and car ownership. However, you would better look for good independent variables, for example gas price and insurance premium, which can explain more variation of Y. I would say this model may or may not be a lemon; definitely, this model is not a 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.

Answer: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

**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).

Answer: Model 1 has $R^2$ of .052, while Model 2 has .061. Model 2 has a slightly (.009=.061-.052) larger $R^2$. However, adding a regressor does not improve $R^2$ substantially. I would stick to Model 1, although Model 1 is not a peach.

**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?

Answer: Model 1 and 2 respectively have adjusted $R^2$ of .048 and .054. Again, adding a regressor makes a small difference. There is not a big benefit of adding the variable `male`. I prefer Model 1.

**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.

Answer: Model 1 and 2 have F statistics of 12.021 and 9.313, respectively. SPSS report p-value of .000 in both models, but the p-value is not numerically zero, but virtually zero. The p-value in Model 2 appear to be slightly larger than that in Model 1 (12.021 is less likely than 9.313), but their difference is negligible. There is no big difference between two models. Model 1 looks better in my opinion.

**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?

Answer: $b_1$ is .669 in Model 1 and .655 in Model 2. $b_2$ is -.024 in Model 1 and -.024 in Model 2. The p-value for $b_1$ is virtually (not numerically) zero in both model. The p-value for $b_2$ is .850 in Model 1 and .847 in Model 2. There is no big difference in parameter estimators and their p-values between two models.

**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.

Answer: SSE is 94.213 in Model 1 and 93.406 in Model 2. The difference is -.807 = 93.406 - 94.213. Adding the variable $X_3$ reduces the variation of Y that the model cannot explain by .807. It is not substantial reduction compared to the total variation.

**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).

Answer: $df_{SSE}$ is 434 (=N-K=437-3) in Model 1 and 433 (=N-K=437-4) in Model 2. Note that K is the number of parameters to be estimated (or the number of regressors plus 1). $df_{SSE}$ of Model 1 is larger than that of Model 2. Adding a regressor reduces the degrees of freedom by 1, adding two regressor reduces the degrees of freedom by 2, and so forth.

**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?

Answer: SSM is 5.219 in Model 1 and 6.027 in Model 2. The difference is .808 = 6.027-5.219. Adding the variable $X_3$ increases the variation of Y that the model can explain by .808. Due to the rounding error, two differences (.807 and 808) look slightly different but they are numerically same. Adding a regressor reduces (-) SSE and in turn increases (+) SSM by the same amount. Total variation of Y (SST) remains unchanged.

**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 how adding a regressor changes the partition of variance components.

Answer: SST is 99.432 in both models. There is no difference. Adding a regressor changes variation components (SSE and SSM) but does not influence the variation of Y (SST).

**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.

Answer: Adding a regressor $X_3$ (`male`) reduces SSE and increase SSM by .807, which is negligible in terms of total variation of Y. This result means that Model 2 can explain the variation of Y more than Model 1 but does not make a big difference. In other word, adding a regressor $X_3$ (`male`) is not that valuable. I would take the parsimonious model, Model 1.