K300 (4392) Statistical Techniques (Fall 2007)

Assignment 7: Estimating Linear Regression (345 points, Due November 28)

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

Please first read the following instructions and questions carefully.

- Download the SPSS data set `assignment7.asv` from OnCourse CL or the course web page at http://www.masil.org/method/statistics.html
- **Write down answers on the SPSS output.** Use separate sheets if you really need.
- Hand in this assignment **by Wednesday, November 28**.
- You may ask your classmates about using SPSS, but **you MAY NOT discuss with other classmates** in any circumstance when answering questions. Remember the IUPUI Student Code of Conduct and SPEA policies.
- If you have any problem with any of the questions, just talk to me.

1. (20 points) Visit OnCourse CL or course web page. Download both Powerpoint slides of correlation coefficient and linear regression model, and then print them out. Once loading them in Powerpoint, click File→Print. Choose "Handouts" in **Print what:** option at the left bottom to save papers. You may print on both sides by changing relevant options. If you have any technical problem, get help from the UITS support center at ICTC131 and BS3000. Finally, attach them to your assignment. Keep in mind that you may not be able to answer following questions successfully without these slides.
http://www.masil.org/teach/k300/Topic12_Correlation.ppt
http://www.masil.org/teach/k300/Topic13_OLS.ppt

2. (80 points) Karl Pearson bivariate correlation coefficient *r* measures a linear relationship between two interval or ratio variables. Note that Chi-square test examines the relationship (independence) of two categorical (nominal or ordinal) variables. *r* ranges from -1 (negative relationship) to 1 (positive relationship). Zero means independence or no relationship between two variables. Assignment7.sav data set includes two variables `absence` (x) and `grade` (y). See question 33 on page 542 for details. Two observations were added so that we have a total of 8 observations. See Powerpoint slides for necessary formulae. Use the significance level of .05.

1) (10 points) Launch SPSS and load the data set `assignment7.sav` you downloaded. Click Analyze→Correlate→Bivariate… Choose variables `absence` and `grade` consecutively (DO NOT switch the order) and move them to the right-hand side box labeled "Variables:" Click **Options…** button at the bottom and check **Means and standard deviations** and **Cross-product deviations and covariances**. Click **Continue** button to get back and click **OK** to conduct the bivariate correlation coefficient. Print out the output and close the Output window.
2) (5 points) On the first table, circle mean, standard deviation, and N. Report means of two variables. If `grade` appears first, go back to 1) and print again.

3) (5 points) On the second table, read the first line of the first row labeled as "absence." Circle numbers under labels of "absence" and "grade." The first number is a correlation coefficient of `absence` and itself: the coefficient should be 1. The second number is the correlation coefficient of `absence` and `grade`; report this number.
4) (5 points) Read the second line labeled "Sig. (2-tailed)." Report the number in the last column. This is the p-value of the correlation of coefficient.
5) (10 points) Now, describe the relationship between variables `absence` and `grade`. Is it positive or negative? How strong is the relationship?
6) (5 points) Read the third line of the first row. The first number is the sum of squares of `absence`, SS_{xx}, while the second number is the sum of products of `absence` and `grade`, SP_{xy}. Circle these two numbers. Read the third line of the second row labeled as "grade." Circle the second number, which is the sum of squares of `grade`, SS_{yy}. Note that the bivariate correlation coefficient matrix is symmetric with a diagonal element of 1.
7) (5 points) Compute the correlation coefficient of the two variables, *r*, using information you obtained in 6). Look at the p-value you obtained in 4). Would you conclude, on the based of this p-value, that variables `absence` and `grade` are independent?
8) (15 points) Let us test if ρ is zero. Note that the test statistic follows the t distribution with *n-2* degrees of freedom. (1) state the null and alternative hypothesis; (2) find out the critical value for the .05 significance level from the t distribution table on page 635. (3) Compute t statistic using the formulae provided in the slides; (4) compare the t statistic with the critical value. Would you like to reject the null hypothesis?; and (5) interpret the output substantively. You need to re-paraphrase the null hypothesis. Is the number of absences for a student related to his/her final grade?
9) (20 points) Compute the correlation coefficient of the two variables by constructing the following talbe.

absence (x)	grade (y)	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
10	70					
…	…					

(1) Copy all data points from `assignment7.sav`. You should have 8 observations.
(2) Compute means of two variables $\bar{x}$ and $\bar{y}$.
(3) Compute deviations of data points from means, $(x_i - \bar{x})$ and $(y_i - \bar{y})$.
(4) Compute sums of squares of `absence` and `grade`, $\sum (x_i - \bar{x})^2$ and $\sum (y_i - \bar{y})^2$.
(5) Compute the sum of product of `absence` and `grade`, $\sum (x_i - \bar{x})(y_i - \bar{y})$.
(6) Compute the correlation coefficient of the two variables, *r*.

3. (40 points) Let us draw a scatterplot to visualize the relationship between two variables.

1) (10 points) Launch SPSS and load the data set `assignment7.sav`. Click Graph→Legacy Dialogs→Scatter/Dot…. Choose the first plot of **Simple Scatter** and then click **Define** button. On the **Simple Scatterplot** window, choose `absence` and move it to a right-hand side box labeled "X Axis" and `grade` to the box labeled "Y Axis." Click **OK** button to produce a scatter plot of the two variables. Print out the output and close the Output window.
2) (5 points) Circle "absence" on the X axis and "grade" on the Y axis. Circle all data points on the plot. You should have 8 observations.
3) (10 points) Draw an imaginary line (**not a curve**) on the plot that you think fits these data points well. Note that "best fit" means the sum of squared deviations of data points from the imaginary line is minimized. There is no single answer but your line should be consistent with the correlation coefficient *r* you got in Question 1. Write down *r* on the middle of the line to indicate the correlation coefficient.
4) (5 points) How do you describe the relationship of two variables? Can you expect a high grade if you miss many classes?
5) (10 points) Circle label 6 (6 absences) on X axis and draw a **vertical line** from the number to the imaginary line. Draw a **horizontal line** from the point, where the vertical line intersects the imaginary line, to the Y axis. Read (roughly) the grade that a student, who missed class six times, is expected to get according to the imaginary line and *r*.

4. (125 points) Let us use the same data to *regress the final grade on the number of absences*. See Powerpoint slides for necessary formulae. Use the significance level of .05.

1) (10 points) Launch SPSS and load the data set `assignment7.sav`. Click Analyze→Regression→Linear…. On the **Linear Regression** window, choose `absence` and move it to a right-hand side box labeled "Independent(s):" and `grade` to the box labeled "Dependent:." Click **OK** button to regress the final grade on the number of absences. Print out the output and close the Output window. Your linear regression model is $Y= \beta_0 + \beta_1 X + \varepsilon$, where Y (dependent variable) and X (independent variable or regressor) are `grade` and `absence`, respectively. You MAY NOT switch Y and X; this is a reverse regression that regresses X on Y (In general, it is awkward to imagine that the final grade influences the number of absences).
2) (5 points) Take a look at the fourth table titled as "Coefficients." Read the first line labeled as "(Constant)." This is the line for the intercept β_0. Circle first two numbers labeled as "B" and "Std. Error", respectively. They are the estimator of parameter β_0 and its standard error, standard deviation of b_0. (Ignore the third column labeled as "Beta." Like "Sig." of the last column, "Beta" is misleading since the statistic is not beta (β_0). In general, we never know true values of β_0 and β_1. So, do not be fooled by "stupid" SPSS). How to interpret the estimator of β_0? See the slides if you have no idea.
3) (15 points) Read the t statistic and p-value. Let us test if b_0 is reliable by the p-value approach; you do not need to get the critical value in this case since SPSS gives you the p-value. (1) state the null and alternative hypothesis; DO NOT use

English alphabets in a hypothesis; (2) report the significance level. (3) report the t statistic. **Show how to compute this statistic** using information you obtained in 2). Circle and report the p-value SPSS gave you. (4) compare the p-value with the significance level. Would you like to reject the null hypothesis?; and (5) interpret the output. Do you think the intercept is zero?

4) (5 points) Read the second line for the independent variable absence. Circle the parameter estimator b_1, its standard error (deviation), t statistic, and p-value.

5) (15 points) Conduct the t-test for b_1 as you did in 3). (1) state the null and alternative hypothesis; (2) report the significance level. (3) report the t statistic. **Show how to compute this statistic** using information you obtained in 4). Circle and report the p-value. (4) compare the p-value with the significance level. Would you like to reject the null hypothesis?; and (5) interpret the output.

6) (10 points) How do you interpret b_1. Yes, b_1 is the slope of the regression line. See the slides if you are stuck.

7) (10 points) Now, take a look at third table titled as "ANOVA." This is the ANOVA table for ordinary least squares (OLS). Circle SSM, SSE, SST on the first column labeled as "Sum of Squares." Remember the structure of an ANOVA table. Note that "Residual" is equivalent to "error." Make sure SST=SSM+SSE (plug in numbers).

8) (5 points) Read the second column and circle three types of degrees of freedom, df_{model}, df_{error}, and df_{total}. Show how these degrees of freedom are computed.

9) (5 points) Read the third column and circle MSM and MSE. Show how these statistics are computed.

10) (5 points) Read the fourth column and circle F the statistic. Circle the p-value in the last column. Again, "Sig." should be replaced by "p-value." How do you compute the F statistic?

11) (15 points) Now, conduct F-test for goodness-of-fit of this regression model. Since the p-value is provided, let us take the p-value approach. (1) state the null and alternative hypothesis; (2) report the significance level. (3) report the F statistic and the p-value you obtained in 10); (4) and compare the p-value with the significance level. Would you like to reject the null hypothesis?; and (5) interpret the output. Do you think your regression model fits the data well?

12) (10 points) Take a look at the second table titled as "Model Summary." Read the first line to get first three numbers; ignore the last one. The first number labeled as "R" is the correlation coefficient, while the second is the R square or R_2, the coefficient of determination. Circle the two statistics and show the relationship between *r* and R_2.

13) (10 points) Show me how to compute R_2 using the ANOVA table. How do you interpret R_2? Do you think your regression model is good or bad?

14) (5 points) Circle the third statistic labeled as "Adjusted R square." Show how the adjusted R^2 is computed using the formulae provided in the slides. Is there a big difference between R^2 and adjusted R^2?

5. (80 points) It is time to estimate parameter estimators b_0 and b_1 by hand. OLS estimators are computed by the least square method, a linear algebraic solution. Use

separate sheets to answer. See Powerpoint slides for necessary formulae. (If you have more than one regressor, the solution should be different.)

1) (15 points) You need to compute SP_{xy}, and SS_{xx}. Construct a proper table. See page 13 of the slides if you have no idea. Of course, you may borrow some information from Question 2.9. Copy the formula for b_1. Estimate b_1 and report this estimator of the parameter β_1. This is the least squares method.

2) (5 points) Copy the formula for b_0. It is much easier to compute than b_1. Show me how you got b_0.

3) (5 points) Write down your linear regression function in the form of "$Y = b_0 + b_1X$." Simply replace two parameter estimators with proper numbers you obtained in 1) and 2).

4) (5 points) Using the linear regression function you got in 3), compute the predicted value of the final grade when the number of absences is 6. You just need to plug the value in X. Compare your prediction with one you guessed in Question 3.5. Are they similar?

5) (10 points) Now, let us compute the standard errors (deviations) of parameter estimators in order to conduct the t-test. Construct a table for computation, which is similar to one on page 24 of the slide.

6) (5 points) Copy the formula for the estimated variance of error (residual), $\hat{\sigma}_\varepsilon^2 = s_\varepsilon^2$, from the slides. Show me how you computed s_ε^2 and make sure it is the same as MSE on the ANOVA table.

7) (5 points) Copy the formula for variance of b_1, $Var(b_1)$. Note ***b_1 is a random variable***. Show how to compute $Var(b_1)$. Take a square root of $Var(b_1)$ and make sure it is the standard error of b_1 shown in the fourth table (see Question 4.4).

8) (5 points) Copy the formula for variance of b_0, $Var(b_0)$. Show how to compute $Var(b_0)$. Take a square root of $Var(b_0)$ and make sure it is the standard error of b_0 shown in the fourth table (see Question 4.2).

9) (5 points) Let us move on to SSM, SSE, and SST to construct the ANOVA table. Show how to get SSE using information (MSE) given in 6) above. Simple manipulation will do; do not try to compute individual residuals (errors) again.

10) (5 points) Report SST by providing the proper formula and a proper number from Q2.9. Do not try to compute the whole sum of squares again.

11) (5 points) Show me how to get SSM using SST=SSM+SSE. Plug in proper numbers that you obtained in 9) and 10). Do not try to compute the whole sum of squares again.

12) (10 points) Construct the ANOVA table using SSM, SSE, and SST you obtained in 9) through 11).

Checklist (you should submit)

- SPSS output for correlation coefficient (Question 2), the scatterplot (Question 3), and the linear regression model (Question 4)
- Separate sheets (needed for computation)
- Powerpoint slides of correlation coefficient (Question 1)
- Powerpoint slides of linear regression model (Question 1)

End of assignment 7. Good luck.