Statistics and Data Analysis, Fall 2016
Pre-lecture Problems 8

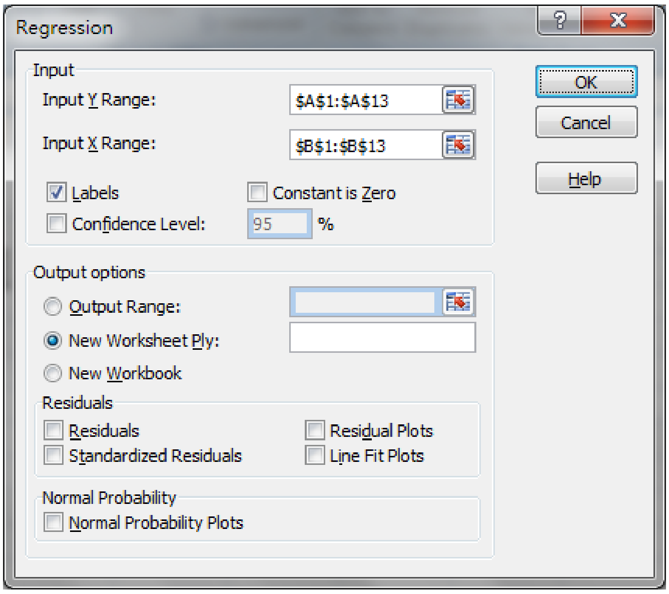
Instructor: Ling-Chieh Kung
Department of Information Management
National Taiwan University

Note 1. The deadline of submitting the pre-lecture problem is ***14:30, November 16, 2016***. Please submit a hard copy of your work to the instructor in class. Late submissions will not be accepted. Each student must submit her/his individual work. Submit ONLY the problem that counts for grades.

Note 2. Please make your answer as clear (i.e., easy to read) as possible. We reserve the right to take away points when the correctness cannot be easily determined (e.g., when the writing is messy and cannot be easily understood).

Before you start, please read the document "SDA-Fa16_dataAnalysis.pdf" in the "Handouts" section on the course website and install the add-in "Data Analysis" (or something equivalent) in your MS Excel. We will teach you how to use Data Analysis in MS Excel to do regression. If you want to use something else, please teach yourself the steps of conducting a regression study.

1. (0 point) Consider the MS Excel file "SDA-Fa16_08_regression1_pl_data.xlsx," in which the sizes (in m^2), numbers of bedrooms, ages (in years), and prices (in \$1000) of 12 houses are recorded in the spreed sheet "House." In this problem, we will demonstrate how to use MS Excel to construct a regression model based on sizes and prices. First, we open the Data Analysis add-in and select "Regression." We then input the following things into the window (cf. Figure 1):



Regression

Input
Input Y Range: $A$1:$A$13
Input X Range: $B$1:$B$13
☑ Labels
□ Constant is Zero
□ Confidence Level: 95 %

Output options
○ Output Range:
● New Worksheet Ply:
○ New Workbook

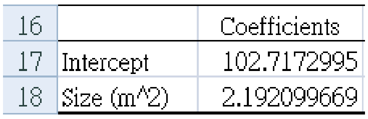
Residuals
□ Residuals
□ Residual Plots
□ Standardized Residuals
□ Line Fit Plots

Normal Probability
□ Normal Probability Plots

Figure 1: Regression input for the size-price model

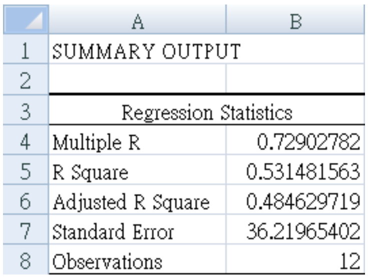
- For "Input Y Range," select the cells containing the prices.
- For "Input X Range," select the cells containing the sizes.
- Select the "Labels" option to indicate that you have included variable labels in your selected input data.

Do not care about anything else, press "OK." You will see the regression report in a new spreed sheet. While the report contains a lot of information, note that the estimated $\hat{\beta}_0 = 102.717$ and $\hat{\beta}_1 = 2.192$ are included (cf. Figure 2). Finally, note that the regression report also contains R^2



16		Coefficients
17	Intercept	102.7172995
18	Size (m^2)	2.192099669

Figure 2: Regression coefficients for the size-price model



	A	B
1	SUMMARY OUTPUT	
2		
3	Regression Statistics	
4	Multiple R	0.72902782
5	R Square	0.531481563
6	Adjusted R Square	0.484629719
7	Standard Error	36.21965402
8	Observations	12

Figure 3: R^2 and R^2_{adj} for the size-price model

and R^2_{adj} (cf. Figure 3). The other values shown in Figure 3 are relevant but not important in this course.

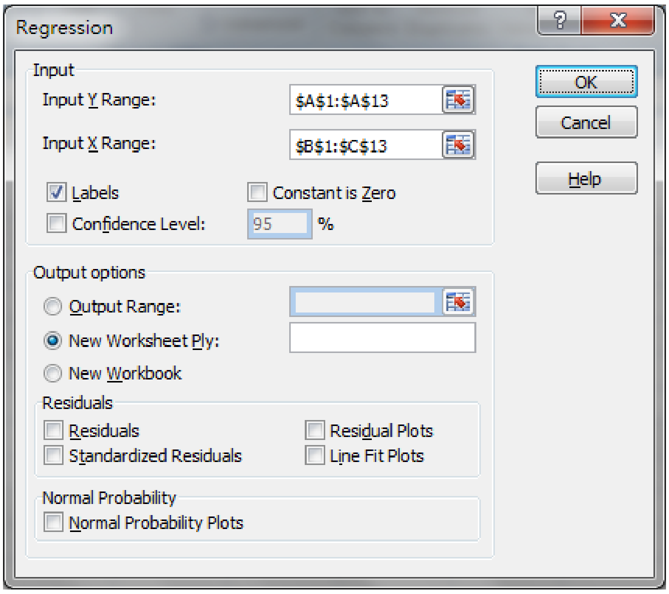
(a) Try to use the number of bedrooms as the only independent variable and show that the estimated model will be

$$y = 205.087 + 32.543x_2,$$

where y and x_2 are the price and number of bedrooms, respectively. Interpret the model.

(b) Verify that for the bedroom-price model, we have $R^2 = 0.290$ and $R^2_{\text{adj}} = 0.219$.

2. (0 point) To use MS Excel to do the regression analysis based on sizes, numbers of bedrooms, and prices, we open the Data Analysis add-in and select "Regression." We then input the following things into the window (cf. Figure 4):



Regression

Input

Input Y Range: $A$1:$A$13

Input X Range: $B$1:$C$13

☑ Labels ☐ Constant is Zero

☐ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

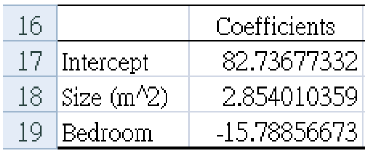
☐ Normal Probability Plots

Figure 4: Regression input for the size-bedroom-price model

- For "Input Y Range," select the cells containing the prices.
- For "Input X Range," select the cells containing the sizes and numbers of bedrooms.

- Select the “Labels” option to indicate that you have included variable labels in your selected input data.

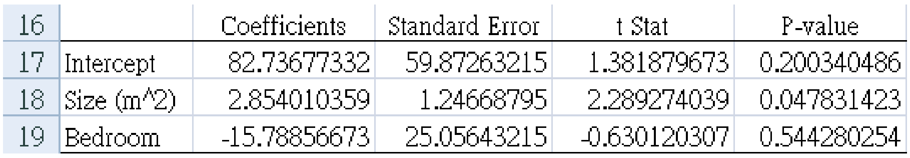
Do not care about anything else, press “OK.” You will see the regression report in a new spreed sheet. While the report contains a lot of information, note that the estimated $\hat{\beta}_0 = 82.737$, $\hat{\beta}_1 = 2.854$, and $\hat{\beta}_2 = -15.786$ are included (cf. Figure 5).



16		Coefficients
17	Intercept	82.73677332
18	Size (m^2)	2.854010359
19	Bedroom	-15.78856673

Figure 5: Regression coefficients for the size-bedroom-price model

The regression report also contains information about the significance of variables. For this regression model, we can see the p-values of testing $\beta_1 \neq 0$ and $\beta_2 \neq 0$ as 0.048 and 0.544, respectively (cf. Figure 6). This shows that size is a good predictor of price but number of bedrooms is not (at this in this model).



16		Coefficients	Standard Error	t Stat	P-value
17	Intercept	82.73677332	59.87263215	1.381879673	0.200340486
18	Size (m^2)	2.854010359	1.24668795	2.289274039	0.047831423
19	Bedroom	-15.78856673	25.05643215	-0.630120307	0.544280254

Figure 6: Variable significance for the size-bedroom-price model

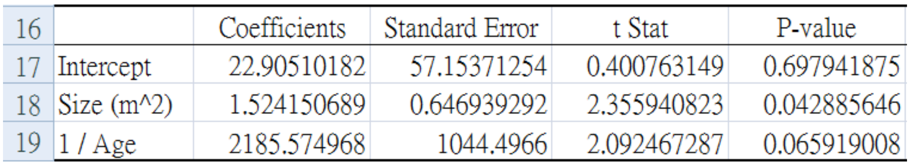
(a) Try to use the number of bedrooms and age as the independent variables and show that the estimated model will be

$$y = 383.612 + 12.473x_2 - 8.099x_3,$$

where y, x_2, and x_3 are the price, number of bedrooms, and age, respectively. Interpret the model.

(b) Verify that for this model, the p-values for the number of bedrooms and age are 0.478 and 0.065. Convince yourself that age is a good predictor of price but number of bedrooms is not.

3. (0 point) To use size and $\frac{1}{\text{age}}$ as the independent variables, we need to manually prepare a new column containing $\frac{1}{\text{age}}$. This is done in the spreed sheet “House2.” We may then do regression with the first three columns and obtain the regression report (cf. Figure 7). We can see that both variables are significant (at different significance levels).



16		Coefficients	Standard Error	t Stat	P-value
17	Intercept	22.90510182	57.15371254	0.400763149	0.697941875
18	Size (m^2)	1.524150689	0.646939292	2.355940823	0.042885646
19	1 / Age	2185.574968	1044.4966	2.092467287	0.065919008

Figure 7: Variable significance for size and the reciprocal of age

(a) Use size, age, and the square of age to be the independent variables and show that the estimated model will be

$$y = 250.746 + 1.537x_1 - 5.113x_3 - 0.032x_3^2,$$

where y, x_1, and x_3 are the price, size, and age, respectively. Interpret the model.

(b) Verify that $R^2 = 0.696$ and $R^2_{\text{adj}} = 0.583$.

(c) Verify that the p-values of size, age, and age square are 0.052, 0.878, and 0.970, respectively. Convince yourself that size is a good predictor of price but age and age square are not.

4. (10 points) Continue from the previous problems. Now use size, the square of size, age, and the square of age as the independent variables.

(a) (3 points) Find the estimated model.

(b) (3 points) Find R^2 and R^2_{adj}. Explain why R^2 must be larger compared to the model in Problem 3. How about R^2_{adj}?

(c) (4 points) Find the p-values for all the four independent variables. Give interpretations based on these p-values.