CHAPTER 19

Visualizing Binary Predictors in Regression

CONCEPTS

- Binary Variable, Qualitative Variable, Quantitative Variable, Intercept Binary, Slope Binary, Interaction, Seasonal Binary, Median Split

OBJECTIVES

- Recognize the difference between quantitative and qualitative predictors in a regression model
- Be able to interpret the results of a fitted regression model that has either an intercept binary or a slope binary or both
- Learn how to interpret seasonal binaries for time series data
- Recognize if the data is time series or cross-sectional
- Understand how to create binary predictors from a third variable

Overview of Concepts

It is natural to think of regression modeling as a method of studying relationships among **quantitative variables** (data whose values are integers or decimal numbers). Examples would be home size (square feet), educational attainment (years), or credit card debt (dollars). However, researchers also study **qualitative variables** (attribute data whose values are non-numerical). Such data are usually encoded as **binary variables** using either 0 or 1. For example, we could create a binary variable Q_i to represent gender by defining $Q_i = 1$ if the i^{th} individual is female, or $Q_i = 0$ if the i^{th} individual is male. Similarly, a binary could denote a college graduate (yes = 1, no = 0). For time-series data, a binary could designate a recession year (recession = 1, none = 0). The choice of which is 0 and which is 1 is arbitrary, although 1 often is the condition of interest to the researcher. A special case is the **seasonal binary** that is used to represent a season (e.g., 1 if it is January, 0 otherwise). Binary predictors may be freely used in OLS regression, and cause no special problems.

A regression model *cannot* include a multi-valued attribute that is encoded numerically (for example, 1 = liberal, 2 = moderate, 3 = conservative) since the numbers themselves have no intrinsic meaning. Rather, we would define a binary for each category (e.g., Liberal = 1 if the individual is liberal, 0 otherwise; Moderate = 1 if the individual is moderate, 0 otherwise; Conservative = 1 if the individual is conservative, 0 otherwise). To avoid multicollinearity, the number of binaries allowed in the regression is always one *less* than the number of categories. Thus, gender (two categories) requires *one* binary, and political orientation (three categories) requires *two* binaries. If a qualitative predictor has c categories, we omit one so that only c–1 binaries enter the model. Although the decision of which to omit is arbitrary, the estimated coefficients and the interpretation of the results will be affected by the decision.

Binary predictors may enter the model in several ways. For an **intercept binary** the model is $Y_i = \beta_0 + \beta_1 X_i + \beta_2 Q_i$. If $Q_i = 0$ then the intercept is β_0, while if $Q_i = 1$ the intercept is $\beta_0 + \beta_2$. In other words, Q_i shifts the *intercept* of the regression line. For a **slope binary** the model is $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i Q_i$. If $Q_i = 0$, the slope of X is β_1, while if $Q_i = 1$, the slope is $\beta_1 + \beta_2$. In other words, Q_i changes the *slope* of the regression line. The term $X_i Q_i$ is called an **interaction** because the quantitative predictor X_i interacts with the qualitative predictor Q_i. We can also have both an intercept binary *and* a slope binary, using a model of the form $Y_i = \beta_0 + \beta_1 X_i + \beta_2 Q_i + \beta_3 X_i Q_i$. This amounts to a separate regression for each binary value, albeit with a common error term (not shown), i.e., when $Q_i = 0$, the model is $Y_i = \beta_0 + \beta_1 X_i$ whereas when $Q_i = 1$ the model is $Y_i = (\beta_0 + \beta_2)+ (\beta_1 + \beta_3)X_i$.

A binary (qualitative predictor) can be created from a quantitative variable Z_i. One common way to do this is a **median split**, in which we assign $Q_i = 1$ for all Z_i values above the median, and assign $Q_i = 0$ to all values of Z_i below the median. This method would be appropriate when we feel that Z_i may influence Y_i, but are unwilling to assume a continuous effect. It is also useful when Z_i has outliers. We could also create multiple quantitative variables from Z_i by choosing appropriate cutpoints (e.g., divide Z_i at its quartiles to create four qualitative predictors, then omit one of them from the model). There are four common methods for creating binary variables based on a third predictor Z_i: create binaries based on intervals of equal width (e.g., 10-20, 20-30, 30-40, etc), based on intervals of equal frequency (e.g., median split or quartiles), assigning one binary to each category (e.g., conservative, liberal, etc.), or selecting convenient cutpoints on Z_i (e.g., 10-25, 25-50, 50-100, etc.).

Illustration of Concepts

Using 1990 data for the 50 states in the U.S., we want to estimate the simple regression model $Y_i = \beta_0 + \beta_1 X_i$, where Y is a state's average score on the verbal portion of the SAT (Scholastic Aptitude Test) and X is the state's per capita income (dollars). Both X and Y are **quantitative variables**. Is the fitted regression different in the 15 Midwestern states than in the other 35 states? To investigate this question, we create a **qualitative variable** coded $Q_i = 1$ if the i^{th}state is in the Midwest and $Q_i = 0$ otherwise. Because Q_i has only two values, it is a **binary variable**. For an **intercept binary** the model is $Y_i = \beta_0 + \beta_1 X_i + \beta_2 Q_i$. If $Q_i = 0$, the intercept is β_0, while if $Q_i = 1$, the intercept is $\beta_0 + \beta_2$. In other words, Q_i shifts the *intercept* of the regression line (see Figure 1). Both fitted lines have a negative slope, indicating that higher income is associated with lower SAT scores. This somewhat unexpected result may reflect a phenomenon known as *selection bias* (most universities and colleges require the ACT rather than the SAT test, while the most selective universities and colleges, generally located in more-affluent states, require the SAT test). The two regression lines are parallel. That is, they have the same *slope* but different *intercepts*. After accounting for effects of income, the Midwestern states ($Q_i = 1$) appear to have higher SAT scores than the other states ($Q_i = 0$).

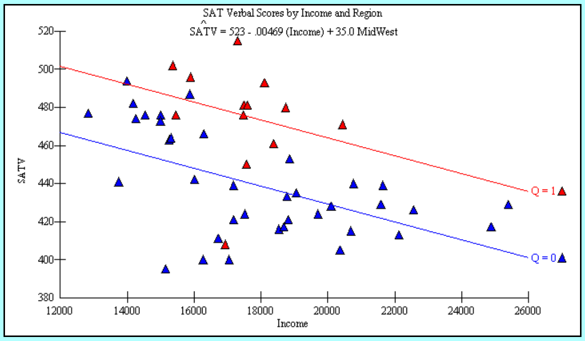


Figure 1: Model Using Intercept Binary

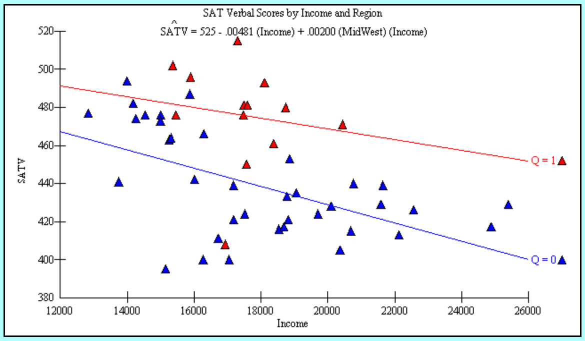


Figure 2: Model Using Slope Binary

Alternatively, for a **slope binary**, the model is $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i Q_i$. If $Q_i = 0$, the slope of X is β_1, while if $Q_i = 1$, the slope is $\beta_1 + \beta_2$. In other words, Q_i changes the *slope* of the regression line. The term $X_i Q_i$ is called an **interaction** because the quantitative predictor X_i interacts with the qualitative predictor Q_i. Figure 2 shows the results of this model. Both slopes are negative, but the slope is flatter for the Midwestern states ($Q_i = 1$) than for the non-Midwestern states ($Q_i = 0$). In all cases, the p-values (not shown) indicate that these regression coefficients differ significantly from zero. This means that both X_i and $X_i Q_i$ are significant predictors of Y_i.

We could create a binary predictor from a third variable Z (e.g., per capita spending on grades K-12). For example, a **median split** would assign $Q_i = 1$ if Z_i is above its median and $Q_i = 0$ if Z_i is below its median. This approach is appropriate when we believe that Z_i may influence Y_i, but does *not* have a *continuous* effect. For example, if the coefficient for Q is 10, we conclude that students living in states that spend above the median on K-12 education score on average 10 points higher on the SAT test. The amount above the median is not considered.

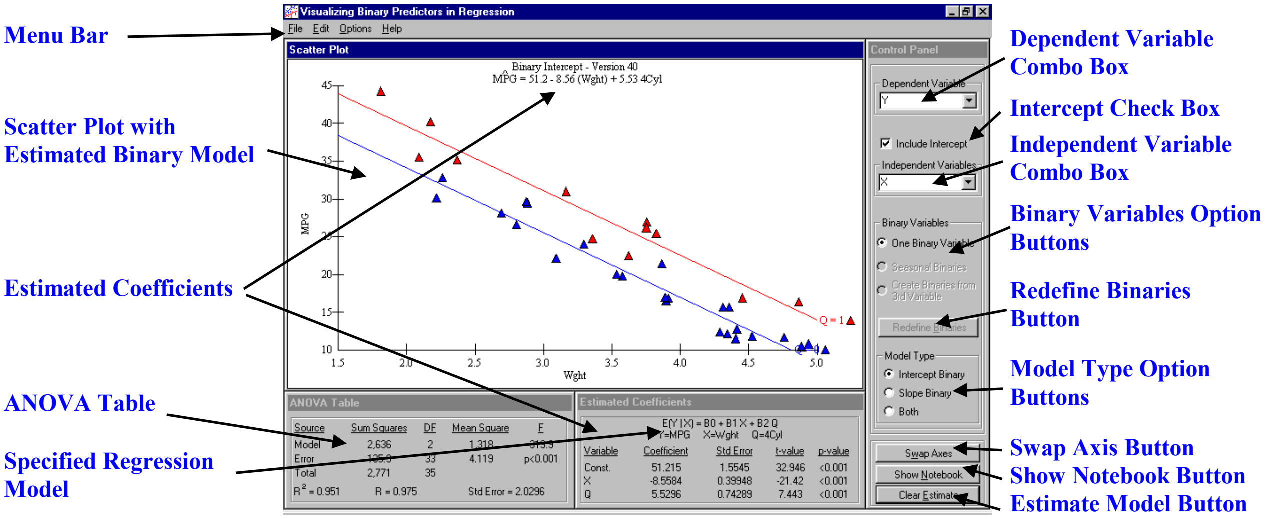
If we have time-series data (e.g., quarterly or monthly data) we can include a **seasonal binary** to test for systematic effects. For c seasons, we use c – 1 binaries. For example, with quarterly data, we could include the three binaries Qtr2, Qtr3, Qtr4, arbitrarily omitting Qtr1. The omitted season is the reference point to which the other seasons are compared.

Orientation to Basic Features

This module estimates models with binary variables. You can analyze a scenario that uses artificial data created to illustrate a concept, an example that uses real data, create your own models using one of the nine databases, or create your own model using the data editor. The choice is made in the Notebook.

1. Opening Screen

Start the module by clicking on the module's icon, title, or chapter number in the *Visual Statistics* menu and pressing the Run Module button. When the module is loaded, you will be on the introduction page of the Notebook. Read the questions and then click the Concepts tab to see the concepts that you will learn. Select the Scenarios tab, click on One Binary Variable, and select Binary Intercept 1. Read the scenario, type 40 in the Version box to see version 40 of the 99 versions of this scenario and press OK. A Hint will appear; read it and press OK. Press the flashing Estimate Model button to estimate the binary model; your computer screen should appear as below.



$\widehat{MPG}$ = 51.2 - 8.56 (Wght) + 5.53 4Cyl

ANOVA Table

Source	Sum Squares	DF	Mean Square	F
Model	2,636	2	1,318	319.9
Error	135.9	33	4.119	p<0.001
Total	2,771	35		

R^2 = 0.951 R = 0.975 Std Error = 2.0296

Estimated Coefficients

E(Y | X) = B0 + B1 X + B2 Q
Y=MPG X=Wght Q=4Cyl

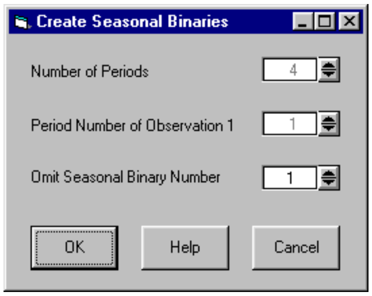
Variable	Coefficient	Std Error	t-value	p-value
Const.	51.215	1.5545	32.946	<0.001
X	-8.5584	0.39948	-21.42	<0.001
Q	5.5296	0.74289	7.443	<0.001

2. Control Panel

Use the Dependent or Independent Variable combo boxes to select a transformation of the variable (studied in Visualizing Regression Models, Chapter 18). Deselect the Intercept check box to specify a regression model without an intercept. Use the Model Type option buttons to select the type of binary variable to include in the model: intercept binary, slope binary, or both. The Swap Axis button switches the variables on the axes and the Show Notebook button returns you to the Notebook. The model to be estimated is displayed at the top of the Estimated Coefficients window (if it is too long for the display, click on the message to display the model). Press the Estimate Model button to estimate this model, display its statistics in the ANOVA and Estimated Coefficients windows, display the fitted binary model (red and blue lines) and its fitted equation on the scatter plot, and rename the button the Clear Model button. Pressing this button clears the statistics and regression lines. Use the Binary Variables option buttons (active only when the data are consistent with their operation) to select the type of binary variable. The Redefine Binaries button allows you to redefine the binary variables if you select either the Seasonal Binaries or Create Binaries from 3rd Variable options.

Seasonal Binaries

Press the Show Notebook button, select the Examples tab, click on Examples Using Time Series Data, and select Traffic Deaths. Read the example, and press OK. Since this is time series data the Seasonal Binaries option button is active (some Scenarios, and Databases also use time series data). Select the Seasonal Binaries option button and the dialog box to the right is displayed. The first spin button displays the number of seasonal periods. The second spin button shows the period number assigned to observation 1. Both spin buttons are only active if you are using your own data (from the Data Editor) since this information is known for data contained in this module. The last spin button specifies which binary to eliminate from the model. Press the OK button. Notice that the scatter plot is now color-coded. The legend shows the color assigned to each period. The legend for S0 shows the color of the period whose binary was omitted from the model.



Number of Periods 4

Period Number of Observation 1 1

Omit Seasonal Binary Number 1

4. **Options**
   a. Select Options from the menu bar and Show Y Intercept to show X = 0 on the horizontal axis (the Y intercept can now be seen). If X = 0 is already showing, it has no effect.
   b. The legend for each regression line is normally next to the line. If it is difficult to read because it overlaps another legend, select Options, Legend and Place Right of Graph.
   c. Selecting Graph Display under Options allows you to control the display. Color Printer uses different colors to differentiate the regression lines and symbols, while Black and White Printer uses different types of lines and symbols that are easier to see on a non-color printer. The default symbols are solid if the sample size is less than 150 or an outline of the symbol otherwise. Select Symbol and Solid or Outline Only to override the default. For time series data, data points can be connected by selecting Type of Plot and Time plot (select Scatter plot to remove the connecting line).
   d. Selecting Equation Display and Hide hides the equation on the scatter plot. If the estimated equation is too long, it is displayed as a label box that can be moved and resized (select Resize). To copy the label box, right click, choose Select All, right click and choose Copy. Select Use Moveable Display to display the equation as a label box.
   e. Select Show Data to display the data. It can also be copied and pasted into a spreadsheet.

5. **Copying Graphs**

   Click on the window you want to copy and select Copy from the Edit menu on the menu bar at the top of the screen. It can then be pasted into a word processor or spreadsheet program.

6. **Help**

   Click on Help on the menu bar at the top of the screen. Search for Help lets you search an index for this module, Contents shows a table of contents for this module, Using Help gives instructions on how to use Help, and About gives licensing and copyright information about this module.

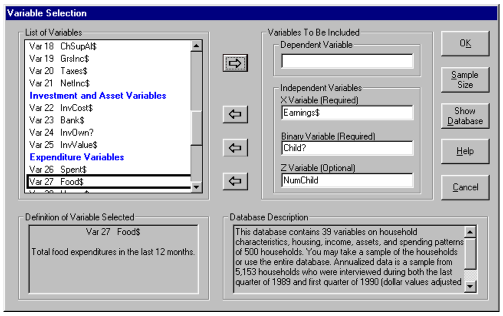
7. **Exit**

   Close the module by selecting Exit in the File menu (or click ☒ in the upper right-hand corner of the window). You will be returned to the *Visual Statistics* main menu.

Orientation to Additional Features

1. **Using Databases**

Press the Show Notebook button, select the Databases tab, click on Individual Consumer Cross Sectional Data and select Interview Data (read description). Press OK. The Variable Selection window (left below) appears. The bottom right frame describes the database. The upper left frame lists the database's variables (binary variables end in ?) grouped in categories (blue headings). The bottom left frame defines the variable selected. The upper right frame defines the model. Press the ⇨ button to write the selected variable in the adjacent box (⇨ changes to ⇦). Pressing a ⇦ button removes the variable from the box (⇦ changes to ⇨). Create the model: Food$ (Dependent Variable), Earnings$ (X Variable), Child? (Binary Variable), and NumChild (Z Variable). Press the Sample Size button, to select a portion of the database (middle below). Select the Use a portion of the database option button. The bottom frame becomes active. Select the top option. Enter 100 into the box (option reads Use first 100 observations). The second option selects a random sample. Press OK. Press the Show Database button to display the *entire* database. Highlight a portion of the table and it can be copied by pressing Ctrl-c. Close the table. Press the OK button.

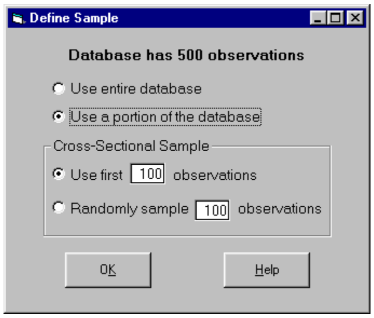


Var 18 ChSupAl$
Var 19 GrsInc$
Var 20 Taxes$
Var 21 NetInc$
Investment and Asset Variables
Var 22 InvCost$
Var 23 Bank$
Var 24 InvOwn?
Var 25 InvValue$
Expenditure Variables
Var 26 Spent$
Var 27 Food$

Dependent Variable:
Independent Variables
X Variable (Required): Earnings$
Binary Variable (Required): Child?
Z Variable (Optional): NumChild

Definition of Variable Selected
Var 27 Food$
Total food expenditures in the last 12 months.

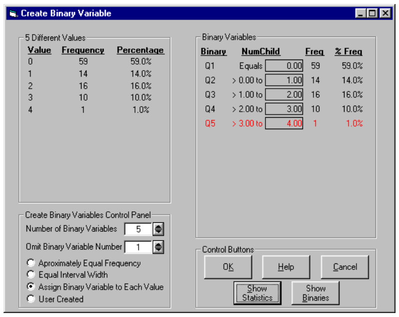
Database Description
This database contains 39 variables on household characteristics, housing, income, assets, and spending patterns of 500 households. You may take a sample of the households or use the entire database. Annualized data is a sample from 5,153 households who were interviewed during both the last quarter of 1989 and first quarter of 1990 (dollar values adjusted



Database has 500 observations

○ Use entire database
● Use a portion of the database

Cross-Sectional Sample
● Use first 100 observations
○ Randomly sample 100 observations



5 Different Values

Value	Frequency	Percentage
0	59	59.0%
1	14	14.0%
2	16	16.0%
3	10	10.0%
4	1	1.0%

Binary Variables

Binary	NumChild		Freq	% Freq
Q1	Equals	0.00	59	59.0%
Q2	> 0.00 to	1.00	14	14.0%
Q3	> 1.00 to	2.00	16	16.0%
Q4	> 2.00 to	3.00	10	10.0%
Q5	> 3.00 to	4.00	1	1.0%

Create Binary Variables Control Panel
Number of Binary Variables: 5
Omit Binary Variable Number: 1
○ Aproximately Equal Frequency
○ Equal Interval Width
● Assign Binary Variable to Each Value
○ User Created

2. **Creating Binaries from a 3rd Variable**

Select the Create Binaries from 3rd Variable option button (window above right appears). Since there are fewer than 13 values (5 in this case), a frequency table is shown. Press the Show Statistics button to display summary statistics. Four options for creating binary variables are listed; select Assign Binary Variable to Each Value. Read the message that appears and press OK. The fifth binary variable is in red (only 1 observation equals 4). Click the User Created option button. Reduce the Number of Binary Variables to 4. The last 2 groups are combined. Pressing the Show Binaries button displays a table of the created binary variables. The second spin button specifies which binary to eliminate from the model. Press the OK button. The scatter plot is now color-coded using the 4 new binaries. The legend shows the color assigned each binary. The legend for S0 shows the color of the binary omitted from the model.

3. **Data Editor**

To use your own data, click on the Show Notebook button, select the Data Editor tab, and press OK. A five-column spreadsheet appears. You can enter variable names (top row), labels (col 1), and data into this editor. The third variable (col 4) *must* be a binary variable. The fourth variable (Var3 – col 5) is optional. Data from another spreadsheet can be pasted into the data editor. To save data in *Visual Statistics* format, select File and Save or press the 🖫 button. When finished, select File and Exit Editor and Use Data or press the ☺ button. To leave the editor and not use the data, select File and Exit Editor and Discard Data or press the ☹ button.

Basic Learning Exercises Name ________________

Binary Intercept

1. In the Notebook, click on the Scenarios tab, click on One Binary Variable, and select Binary Intercept 2 scenario. Read the scenario, enter 39 in the Version box, and press OK. a) Write the model that will be estimated. b) When you estimate this model how many lines will appear? c) What relation do you expect between the lines?

2. Click on Estimate Model. a) Write down the estimated model. b) Why are there two lines? c) Why are the two lines parallel? c) What is the vertical distance between the two lines?

3. a) If your model is $Scps = \beta_0 + \beta_1 \, Tmp + \beta_2 \, Ovr40$, $\alpha = 0.05$, and your null and alternative hypotheses are H_0: $\beta_2 = 0$, H_a: $\beta_2 = 0$, would you reject or not reject H_0 and why? b) How would you interpret this result about people over 40? c) If your null and alternative hypotheses are H_0: $\beta_2 \geq 0$, H_a: $\beta_2 < 0$, would you reject or not reject H_0? d) How would you interpret this result about people over 40? e) What would be an underlying reason for using the hypotheses in a) versus c)?

Binary Slope

4. Press the Show Notebook button and select the Binary Slope scenario. Read the scenario, enter 59 in the Version box, and press OK. a) Write the model that will be estimated. b) When you estimate this model how many lines will appear? c) What relation do you expect between the lines?

5. Click on Estimate Model. Click Options and select Show Y Intercept. a) Write down the estimated model. b) Why are there two lines with the same intercept but different slopes?

6. If your model is StUn = $\beta_0 + \beta_1$ CtyUn + β_2 (Urbn)(CtyUn) + ε, $\alpha = 0.01$, and your null and alternative hypotheses are H_0: $\beta_2 \leq 0$, H_a: $\beta_2 > 0$, would you reject or not reject H_0 and why? How would you interpret this result about urban areas? Why would this be true?

Binary Intercept and Slope

7. Press the Show Notebook button and select the Binary Intercept and Slope scenario. Read the scenario, enter 39 in the Version box, and press OK. a) Write the model that will be estimated. b) When you estimate this model how many lines will appear? c) What relation do you expect between the lines?

8. Click on Estimate Model. a) Write down the estimated model. b) Why are there two lines with different intercepts and slopes? c) What does a different intercept mean? d) What does a different slope mean?

9. a) Suppose your model is Inc = $\beta_0 + \beta_1$ (Exp) + β_2 Male + β_3 (Male) (Exp) + ε, $\alpha = 0.05$, and your null and alternative hypotheses are H_0: $\beta_2 = 0$, H_a: $\beta_2 \neq 0$. Would you reject or not reject H_0 and why? b) How would you interpret this result regarding starting salaries? c) Would this result support the contention of gender discrimination? Why?

10. a) Suppose your null and alternative hypotheses are H_0: $\beta_3 = 0$, H_a: $\beta_3 \neq 0$. Would you reject or not reject H_0 and why? b) How would you interpret this result regarding starting salaries? c) Would this result support the contention of gender discrimination? Why? d) What other factors would need to be examined to confirm this result?

Intermediate Learning Exercises Name ________________

Seasonal Period Binaries

11. In the Notebook, click on the Scenarios tab, click on Seasonal Binary Variables, and select the Quarterly Seasonal Binaries scenario. Read the scenario, enter 40 in the Version box, and press OK. a) Write the model that will be estimated. b) When you estimate this model how many lines will appear? c) What relation do you expect between the lines?

12. Click on Estimate Model. Click Options on the menu bar and select Show Y Intercept. a) Write down the estimated model. b) Why are there four lines with different intercepts and slopes? c) What does a different intercept mean? d) What does a different slope mean?

13. Suppose your model is Clos = $\beta_0 + \beta_1$ (Pop) + β_2 S2 + β_3 (S2) (Pop) + β_4 S3 + β_5 (S3) (Pop) + β_6 S4 + β_7 (S4) (Pop) + ε, $\alpha = 0.05$, and your null and alternative hypotheses are H_0: $\beta_6 = 0$, H_a: $\beta_6 \neq 0$. Would you reject or not reject H_0 and why? How would you interpret this result?

14. If your null and alternative hypotheses are H_0: $\beta_7 = 0$, H_a: $\beta_7 \neq 0$. a) Would you reject or not reject H_0 and why? b) How would you interpret this result regarding clothes sales? c) Redo a) and b) for parameters β_2, β_3, β_4, and β_5.

15. Press the Redefine Binaries button and change the omitted period binary from 1 to 4. Press OK. Press the Estimate Model button. Write down the estimated model. Compare this result with question 12 above. How do the results differ because you omitted a different period?

Other "Seasonal" Binaries

16. Press the Show Notebook button. Select the Daily Binaries scenario. Read the scenario, enter 45 in the Version box, and press OK. Click Options and deselect Show Y Intercept. In this scenario, the "seasonal" or period binaries have nothing to do with seasons. However, since these period binaries are used in the same way as seasonal binaries, period binaries are usually called seasonal binaries. a) Write the model that will be estimated. b) When you estimate this model how many lines will appear? c) What relation do you expect between the lines?

17. Click on Estimate Model. a) Write down the estimated model. b) Why does each line have a different intercept? c) What does each intercept mean? d) Why do the lines have the same slope?

18. The model is Cars = $\beta_0 + \beta_1$ (Per) + β_2 S2 + β_3 S3 + β_4 S4 + β_5 S5 + ε. Many auto experts believe that sales increase during the month because dealerships are more willing to deal near the end of a month in order to earn incentive awards from the automaker. If true, what would be the null and alternative hypotheses? Test this hypothesis at $\alpha = 0.05$. Interpret this result.

19. a) Why would many experts believe that sales for this dealership should be less on Wed., Thurs., and Fri.? b) To test this, what are the three null and alternative hypotheses? c) Using $\alpha = 0.05$, do these results support the belief that dealerships should be open more hours?

20. Press the Show Notebook button, select the Data Editor tab, and press OK. Press the 🗁 button to open Ques20.vsq. Select the file and press the Open button. The data is from the dealership scenario with a binary (WThF) for Wed.-Thurs.-Fri. Press the ☺ button to use the data. The red points are WThF. Press the Estimate Model button. The model is Cars = β_0 + β_1 Period + β_2 WThF. a) What is the estimated model? b) Test H_0: $\beta_2 \geq 0$, H_a: $\beta_2 < 0$ (α = 0.05)? c) Answer question 19c. d) Why is the answer different than before?

Advanced Learning Exercises Name ________________

Median Split

21. In the Notebook, click on the Scenarios tab, click on Creating Binaries from 3rd Variables, and select the Equal Frequencies – Median Split scenario. Read the scenario, enter 40 in the Version box, and press OK. Press the Redefine Binaries button to see how the binary variable was defined. a) How many binaries were created? b) Which criterion was used to create the binaries? c) How many observations are there for each binary variable? d) How was the value 2594.5 selected? e) Why is this called a median split? f) Which binary will be omitted from this model?

22. a) What model will be estimated. b) When estimated, how many lines will appear? c) What relation do you expect between the lines?

23. Press the Estimate Model button. a) What is the estimated model? b) If you believe that the wealthy are more risk averse, what are the null and alternative hypotheses? c) Using $\alpha = 0.05$, would you reject or not reject H_0 and why?

24. Press the Redefine Binaries button. Click the down arrow on the Omit Binary Variable Number spin button to change the option to None. Press the OK button. The model to be estimated is now $Per = \beta_o + \beta_1 (Q1) (Rsk) + \beta_2 (Q2) (Rsk) + \varepsilon$. Notice that Rsk is not in the model as a variable by itself and that *both* binary variables are now in the model. a) What do β_1 and β_2 mean in this model? b) Given your answer above, what will be the estimate for β_1 and β_2? c) Press the Estimate Model button. Read the message that appears and press OK. What is the estimated model? d) Using the results from above, how is the 1.29 calculated?

25. Press the Redefine Binaries button. Click the up arrow on the Omit Binary Variable Number spin button to change the option back to 1. Press the OK button. Select the Intercept Binary option button. How does this model differ from the model with a slope binary?

26. Press the Estimate Model button. What is the estimated model? Using $\alpha = 0.10$, do wealthy investors earn a *different* rate of return on their investments?

Nonlinear Binary Models

27. Click on the Dependent Variable combo box and select Ln Y. The hypothesized model is now $\text{Ln(Per)} = \beta_0 + \beta_1 \text{(Rsk)} + \beta_2 \text{Q2} + \varepsilon$. Press the Estimate Model button. a) Seeing the estimated model, what does the model say about the relationship between percent return and risk? b) Since there is only a binary intercept (therefore the slopes are the same), why does the gap between the two lines increase as Per increases?

28. Select the Slope Binary option button. The proposed model is $\text{Ln(Per)} = \beta_0 + \beta_1 \text{(Rsk)} + \beta_2 \text{(Q2) (Rsk)} + \varepsilon$. Press the Estimate Model button. a) Using $\alpha = 0.05$, test the hypothesis H_0: $\beta_2 \leq 0$, H_a: $\beta_2 > 0$. b) Based on this test, what would you conclude about how wealthy versus the non-wealthy view the relationship between percent return and risk?

One Value per Category

29. Press the Show Notebook button and select the One Value per Category scenario. Read the scenario. a) What do the 1, 2, and 3 represent? Use version 50 and press OK. b) Press the Redefine Binaries button. How many binaries were created? c) How were they created? d) Which binary is being omitted?

30. Using categorical data is one of the most common methods to create binary variables. Press the OK button. The model being estimated is $\text{Con} = \beta_0 + \beta_1 \text{ 26 (Inc)} + \beta_2 \text{Q2} + \beta_3 \text{(Q2) (Inc)} + \beta_4 \text{Q3} + \beta_5 \text{(Q3) (Inc)} + \varepsilon$. How many lines will be drawn when the model is estimated? Press the Estimate Model button. Using $\alpha = 0.05$, is there any difference between the three types of individuals?

Individual Learning Projects

Write a report on one of the three topics listed below. Use the cut-and-paste facilities of the module to place the appropriate graphs and tables in your report.

1. The purpose of this project is to demonstrate the use and interpretation of the three types of models using a single binary variable. Select an example to investigate (not a scenario). Describe the dependent, independent and binary variable. For *each* of the three models (binary intercept, binary slope, both binary intercept and slope) present the regression model, explain how the binary variable is used in the model and the question it enables you to answer. Based on this discussion, develop a null and alternative hypothesis. Estimate the model. Provide a table of the model's statistics, an ANOVA table, and a scatterplot showing the estimated model. Discuss these results, interpreting each of the estimated coefficients. Test your hypothesis regarding the binary variable and explain what it means. Before starting this project, make sure you have completed the Basic Learning Exercises.

2. The purpose of this project is to demonstrate that you know how to use and interpret seasonal binary variables. Select either a time series example (not a scenario) or develop your own model using one of the 3 Not Seasonally Adjusted Time Series Data databases. Define your dependent, independent, and seasonal binary variables. Indicate the number of periods in a time frame and the number of time frames covered, e.g., 5 days per week over 10 weeks equals 50 observations. Develop a regression model using an intercept binary (don't forget to tell which period is omitted). Explain how binary variables are used in the model and the questions they enable you to answer. Based on this discussion, develop null and alternative hypotheses that can be tested. Estimate the model. Provide a table of the model's statistics, an ANOVA table, and a scatterplot showing the estimated model. Discuss these results, interpreting each of the estimated coefficients. Test your hypothesis regarding the binary variable and explain what it means regarding the binary variable. Repeat the exercise using either a slope binary model or a slope and intercept binary model. Before starting this project, make sure you have completed the Intermediate Learning Exercises.

3. The purpose of this project is to demonstrate that you know how to interpret a model with multiple binary variables regardless of which binary variable is omitted. You can do this project by using seasonal binaries (do the Intermediate Learning Exercises first) or create your own binaries using a 3rd variable (do the Advanced Learning Exercises first). You can obtain your data from one of the examples (not a scenario) or develop your own model using one of the databases. Your data must have at least four binary variables. Define your dependent, independent, and binary variables. Select a binary model type (intercept, slope, or both). Explain how the binary variables are used in the model and the questions they enable you to answer. Which binary was omitted and why was it selected? Based on this discussion, develop null and alternative hypotheses that can be tested. Estimate the model. Provide a table of the model's statistics, an ANOVA table, and a scatterplot showing the estimated model. Discuss these results, interpreting each of the estimated coefficients. Test your hypothesis regarding the binary variable and explain what it means regarding the binary variable. Re-estimate the model omitting a different binary variable and a third time omitting no binary variable. Show that although the overall results are the same, your estimated coefficients are different (show the relationships between them).

Team Learning Projects

Select one of the projects listed below. In each case, produce a team project that is suitable for an oral presentation using either presentation software or large poster board(s). Graphs and tables must be large enough for your audience to see. Each team member is responsible for producing some of the exhibits. Ask your instructor if a written report is also expected.

1. In this project, a team of four will use the single binary variable model (this is covered in the Basic Learning Exercises). The team should select a dependent variable, two independent variables, and two binary variables from a database. Each independent and binary variable should be expected to affect the dependent variable. Define each variable. Each team member will use the dependent variable, one independent, and one binary variable (there are four combinations) to estimate three binary models (binary intercept, binary slope, both binary intercept and slope). The focus of the presentation should be on the similarities and differences obtained when different variables are used to estimate a dependent variable. The presentation should consider interpretation issues (what questions does the model answer and how is this affected by the type of binary model), as well as statistical results (coefficients, t-statistics, summary statistics). If the scatterplot reveals a nonlinear relationship, variable transformations should be used (see exercises 27 and 28).

2. In this project, a team of three to five will use a seasonal binary model (this is covered in the Intermediate Learning Exercises). The team will model either the number of people employed or unemployment rates. The data is in one of the three databases containing Not Seasonally Adjusted data. The team should select a period of between 96 and 120 months to analyze (press the Time Period button in the Variable Selection window to set the time period). One team member will use the U.S. data, and the rest will use *different* state data. Alternatively, one team member will use the U.S. data, two will use different state data and two will use city data (selecting one city from each state). The independent variable will be the period number (Period). Select any binary variable (e.g., 4thQ?). Analyze three seasonal binary models (intercept binaries, slope binaries, and both intercept and slope binaries). If your scatterplot reveals a nonlinear relationship, variable transformations should be used (see exercises 27 and 28). You may also find that using a single binary variable (i.e., 4thQ? or Jan?) works better than the 12 monthly seasonal binaries (if so, this must be explained *and* illustrated). The focus of the presentation should be on the similarities and differences between the U.S. model and the state models (and city models for teams of five).

3. In this project, a team of four will create binary variables from a 3rd variable (this is covered in the Advanced Learning Exercises). The data must come from one of the databases. The team agrees on the number of observations to use (between 100 and 150, *don't* select the data randomly) and on a dependent and independent variable. Each team member will select a 3rd variable from which to create *at least* 3 binary variables. Each team member will create binary variables in a different way: equal frequencies, equal widths, user created, and one binary per value (use categorical data). Each team member will estimate three models (intercept binaries, slope binaries, and both intercept and slope binaries). If the scatterplot reveals a nonlinear relationship, variable transformations should be used (see exercises 27 and 28). The focus of the presentation will be on how to create binary variables and the similarities and differences in interpreting models using these types of binary variables.

Self-Evaluation Quiz

1. A binary variable
   a. has more than two values.
   b. is a quantitative variable.
   c. is a qualitative variable.
   d. cannot be used in a regression.
   e. is not important to researchers.

2. Depending on how it is used, a binary predictor in a regression could affect
   a. the intercept.
   b. the slope.
   c. the fit.
   d. the t-values.
   e. all of the above.

3. Which statement is *not* correct for an intercept binary Q_i?
   a. The form of the equation is $Y_i = \beta_0 + \beta_1 X_i + \beta_2 Q_i + u_i$.
   b. The coefficient of Q_i is added to the intercept if $Q_i = 1$.
   c. The coefficient of Q_i is subtracted from the intercept if $Q_i = 0$.
   d. The fitted regression line's intercept is shifted up or down by β_2.
   e. We may use a standard t-test for significance for the binary Q_i.

4. Which statement is *incorrect* for a slope binary X_iQ_i?
   a. The form of the equation is $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i Q_i + u_i$.
   b. The coefficient of X_i is added to the intercept if $Q_i = 1$.
   c. The coefficient of X_iQ_i is added to the slope if $Q_i = 1$.
   d. The fitted regression slope is shifted by β_2 if $Q_i = 1$.
   e. The equation is $Y_i = \beta_0 + \beta_1 X_i + u_i$ if $Q_i = 0$.

5. In a regression of the form $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i Q_i + u_i$
   a. the X_iQ_i term is called an interaction.
   b. the intercept may be shifted but not the slope.
   c. both the slope and intercept can shift.
   d. neither the slope nor the intercept will be shifted.
   e. we will see parallel lines on the graph of fitted regression lines.

6. If we see two fitted regression lines with different slopes,
   a. the researcher included only an intercept binary.
   b. the researcher used both a slope and an intercept binary.
   c. the researcher used only a slope binary.
   d. the researcher included an interaction term.
   e. the researcher's slope binary is significantly different from 0.

7. In assigning values to a binary variable, which statement is *incorrect*?
   a. The values assigned usually are 1 and 0.
   b. It is arbitrary which condition is assigned 1 and which is assigned 0.
   c. The value 1 often represents a condition of interest, but this is up to the researcher.
   d. For multiple conditions, we can use a multi-valued variable (e.g., $Q_i = 1, 2, 3, 4$).
   e. 1 represents the presence of an attribute and 0 its absence.

8. A valid set of seasonal binary predictors (each 0 or 1) to use in a regression would be
   a. Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov.
   b. Qtr1, Qtr2, Qtr3, Qtr4.
   c. Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec.
   d. Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec.
   e. Two choices are correct.

9. If you used quarterly seasonal binaries (Q1, Q2, Q3, Q4) and omitted Q3 from your model
   a. then the regression estimates would be invalid.
   b. then the coefficient of Q1 shows how quarter 1 differs from the average quarter.
   c. then the coefficient of Q2 shows how quarter 2 differs from quarter 3.
   d. then the coefficient of Q2 shows how quarter 2 differs from quarter 1.
   e. then we cannot use a slope binary.

10. To use an attribute variable with 5 categories in a regression we would
    a. create 5 binary predictors but omit one from the model.
    b. create 5 binary predictors and include them all in the model.
    c. include 5 binary predictors in the model and see which were significant.
    d. utilize a single predictor with values 1, 2, 3, 4, 5.
    e. do none of the above.

11. Use of a median split on a third variable Z_i would *not*
    a. free the researcher from having to specify a linear relationship with Z_i.
    b. require that Z_i be normally distributed with no outliers
    c. create two groups of observations of roughly equal size.
    d. permit a simplified t-test test for significance of Z_i.
    e. shift the intercept or slope, depending on how the created binary enters the model.

12. A third variable Z_i can be used
    a. to divide the observations into two groups using a median split.
    b. to divide the observations into four groups using the quartiles.
    c. to divide the observations into k groups using equal Z_i intervals.
    d. to divide the observations into k groups using equal Z_i frequencies.
    e. to divide the observations in any of the above ways.

Glossary of Terms

ANOVA table Decomposition of variance in a regression, showing total sum of squares and its sources (regression, error) along with degrees of freedom and mean squares. *Total* degrees of freedom equals n – 1, *error* degrees of freedom equals n – k – 1, and the *regression* degrees of freedom equals k, where n is the sample size and k is the number of independent variables.

Binary predictor Independent variable that has only two values, used for qualitative data (e.g., male, female). Generally, 1 denotes the presence of the attribute of interest and 0 denotes its absence, but other values may be used (e.g., 1 and 2). If the attribute has c categories, we need c – 1 binary variables. See **Intercept binary, Slope binary,** and **Seasonal binary**.

Estimated coefficient Sample statistic used to estimate a parameter of the regression model. The estimated regression coefficients are denoted $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$. See **Ordinary least squares**.

F statistic In a regression ANOVA table, the ratio of the *regression* mean square to the *error* mean square. It is used to test the overall significance of the regression. See **p-value**.

Interaction See **Slope binary**.

Intercept Value of the dependent variable when all the independent variables in the regression model are zero. Zero values of X_i may have little or no meaning for some models.

Intercept binary A simple regression model can be modified to fit *two* regression lines with different intercepts by including a binary predictor (Q = 0, Q = 1). The form of the equation is $Y_i = \beta_0 + \beta_1 X_i + \beta_2 Q_i + u_i$. When $Q_i = 0$ the equation's intercept is β_0. When $Q_i = 1$ the equation's intercept is $\beta_0 + \beta_2$. Since β_2 could be negative, the line might be shifted either up or down. See **Slope binary**.

Median split Binary predictor created from another variable Z_i by defining $Q_i = 0$ if Z_i is below its median and $Q_i = 1$ if Z_i is above its median. In this way, a researcher can investigate whether "low" and "high"values of Z_i have different effects.

Multiple binary predictors More than one qualitative predictor can be included in a model. If an attribute has c categories, we use c– 1 binaries. For example, an individual's marital status (single, married, divorced, widowed) can be coded using *three* binaries. See **Seasonal binary**.

Multiple correlation coefficient Measure of overall fit in a regression. It is the square root of R^2. It may be interpreted as the correlation between Y_{actual} and Y_{fitted} over all n observations.

Parameter Numerical constant needed to define a particular model or distribution. A regression model's parameters are the intercept and the coefficients of the k independent variables, whose true values are denoted $\beta_0, \beta_1, \beta_2, ..., \beta_k$. See **Estimated coefficient**.

P-value Probability (usually two-tailed) of committing type I error if we reject the null hypothesis that a parameter is zero (for example, a regression coefficient). A small p-value (such as 0.01) would incline us to reject the hypothesis that the true parameter is zero.

Qualitative predictor One whose value is not a number (e.g., eye color). In contrast, a *quantitative* variable is one whose value is numerical (e.g., income). See **Binary predictor**.

R-squared Ratio of the *regression* sum of squares to the *total* sum of squares. R^2 near 0 indicates the fit is poor while R^2 near 1 indicates the fit is good. Also called coefficient of determination.

Seasonal binary The number of seasonal binaries is always one less than the number of seasons. One seasonal period must be omitted. Thus, quarterly data (four periods) will require three binaries, and monthly data (12 periods) will require 11 binaries. The choice of which seasonal period to omit is arbitrary. For example, in quarterly data we could omit the first quarter:

Qtr2 = 1 if it is the 2nd quarter, 0 otherwise
Qtr3 = 1 if it is the 3rd quarter, 0 otherwise
Qtr4 = 1 if it is the 4th quarter, 0 otherwise

The omitted period (Qtr1) may be viewed as the baseline, reflected in the intercept when the other binaries are all zero. Failure to omit one seasonal binary creates perfect multicollinearity (in this example, Qtr1 = 1– Qtr2 – Qtr3 – Qtr4) which poses a dire estimation problem.

Slope Coefficient of an independent variable. If $Y_i = \beta_0 + \beta_1 X_i$ then β_1 is the change in the Y for a *one unit* change in X. If $Y_i = \beta_0 + \beta_1 \ln X_i$ then $\beta_1/100$ is the change in Y for a *one percent* change in X. If $\ln Y_i = \beta_0 + \beta_1 X_i$ then $100\beta_1$ is the *percent* change in Y for a *one unit* change in X. If $\ln Y_i = \beta_0 + \beta_1 \ln X_i$ then β_1 is the *elasticity*, or *percent* change in Y for a *one percent* change in X.

Slope binary A simple regression can be modified to fit *two* regression lines with different slopes by including an *interaction* term equal to the independent variable X_i multiplied by a qualitative predictor Q_i. The form of the equation is $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i Q_i + u_i$. When $Q_i = 0$ the slope is β_1. When $Q_i = 1$ the slope is $\beta_1 + \beta_2$. Since β_2 can be negative, the slope for $Q_i = 1$ may be either greater or smaller than β_1. See **Intercept binary**.

Standard error Estimate of the standard deviation of the stochastic disturbances, using the square root of the sum of the squared residuals, divided by $n - k - 1$, where n is the sample size and k is the number of independent variables.

Transformed variable Common transformations of the variables are the logarithmic transformation [$\ln(X_i)$ or $\ln(Y_i)$], raising to a power [X_i^c or Y_i^c], inversion [$1/X_i$ or $1/Y_i$], or standardization by subtracting the mean and dividing by the standard deviation.

t-value Ratio of an estimated coefficient in a regression model to its standard error, used to test the null hypothesis that the parameter is zero. This ratio is distributed as Student's t if the parameter is zero. A large t-value would suggest that the true parameter is not zero.

Solutions to Self-Evaluation Quiz

1. c Read the Overview of Concepts, Illustration of Concepts. Consult the Glossary.
2. e Do Exercises 1–6. Read the Overview of Concepts, Illustration of Concepts.
3. c Do Exercises 1–3. Read the Overview of Concepts, Illustration of Concepts.
4. b Do Exercises 4–6. Read the Overview of Concepts, Illustration of Concepts.
5. a Do Exercises 4–6. Read the Overview of Concepts, Illustration of Concepts.
6. d Do Exercises 1–10. Read the Overview of Concepts, Illustration of Concepts.
7. d Do Exercises 1–10. Read the Overview of Concepts, Illustration of Concepts.
8. e Do Exercises 11–15. Read the Overview of Concepts, Illustration of Concepts.
9. c Do Exercises 11, 12, 15. Read the Overview of Concepts, Illustration of Concepts.
10. a Do Exercises 29, 30. Read the Overview of Concepts, Illustration of Concepts.
11. b Do Exercises 21–26. Read the Overview of Concepts, Illustration of Concepts.
12. e Do Exercises 21–26. Read the Overview of Concepts, Illustration of Concepts.