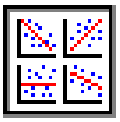
CHAPTER 17



Visualizing Multiple Regression Analysis

CONCEPTS

- Regression Model, Predictor, Parameter, Estimated Coefficient, Standard Error, t-value, p-value, Confidence Interval, Data Conditioning, Multicollinearity, Variance Inflation Factor, Residual Plots, Probability Plot, R-Squared, Adjusted R-Squared

OBJECTIVES

- Recognize and use the terminology of multiple regression
- Be able to perform significance tests and interpret confidence intervals for unknown model parameters
- Understand the importance of data conditioning and the potential effects of ill-conditioned data on a regression
- Detect multicollinearity and recognize its common symptoms
- Learn when a model may be overfitted and why that can be a problem
- Use visual displays to check the residuals for possible non-normality, autocorrelation, and heteroskedasticity

Overview of Concepts

A **regression model** proposes a relationship between a dependent variable Y and several independent variables X_1, X_2, ..., X_k. We call the independent variables **predictors** for short, since they are intended to help us predict Y. The form of the proposed regression model is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$ where ε is a random disturbance. Given a data set of n observations on each variable, the k+1 **parameters** β_0, β_1, ..., β_k can be estimated using the ordinary least squares (OLS) method. These OLS estimates (called **estimated coefficients**) are chosen so as to minimize the sum of squared differences between Y_{actual} and Y_{fitted}.

Dividing a coefficient by its estimated **standard error** gives its **t-value**. A small **p-value** would indicate that the estimated coefficient differs significantly from zero (i.e., that zero is not included in the **confidence interval** for that parameter). The overall fit of a regression model is measured by **R-squared** (or R^2) which lies between 0 and 1. When R^2 is multiplied by 100 it is called the "percent of variation explained" (where 100% would be a "perfect fit"). Since adding more predictors cannot decrease R^2, there may be a temptation to clutter the model with superfluous variables. The **adjusted R-squared** (or R^2_{adj}) includes a penalty for the number of predictors. An added predictor that improves the fit enough to offset the penalty will raise R^2_{adj}, while adding a superfluous predictor will reduce R^2_{adj}. Conversely, removing a weak predictor may raise R^2_{adj} even though R^2 stays the same or declines.

The independent variables may be related to one another (i.e., the predictors may contain redundant information). For example, a person's age, income, years of education, and gender may be related. If we use these four variables to predict the weekly number of hours the person spends watching live professional sports, we face a potential problem called **multicollinearity**. The resulting variance inflation can affect the **standard error** of each estimated coefficient and render its t-tests and p-values misleading. If the **variance inflation factor** (**VIF**) for a predictor is large (one rule of thumb says above 10) then one might consider discarding that predictor, but *only* if that can be done without damaging the logic of the model.

Other tests of a model's adequacy concern violations of certain assumptions about the disturbances (the ε's) that are supposed to be normally distributed, have constant variance, and be independent of one another. One visual test is to plot the residuals (n differences between Y_{actual} and Y_{fitted}) against each predictor to get a **residual plot**. Ideally, there is no apparent visual pattern. A "fan-shaped" or "funnel-shaped" pattern suggests *heteroskedasticity* (non-constant variance of residuals). Plotting the residuals against entry order (only in time series data) can suggest another problem known as *autocorrelation* (runs of + + + or – – – in the residuals). We can check for *non-normality* by examining the histogram of residuals, which should be bell-shaped. A more sensitive test is the **probability plot** of residuals, which should resemble a 45^o line if the residuals are normal. We look at the list of residuals (or a residual plot) to see if there are outliers (or near-outliers) which may be data errors or may suggest a left-out predictor.

Data conditioning refers to the magnitude of the data. *Ill-conditioned* variables have values that are unnecessarily small or large. For example, sales measured in dollars (e.g., $123,456,789) could be rewritten in millions and rounded to fewer significant digits (e.g., 123.457) without harm and perhaps with increased clarity. *Well-conditioned* data help avoid tiny or huge regression coefficients, and may prevent computational errors.

Illustration of Concepts

Here are some results from a **regression model** of aircraft cockpit noise against seven **predictors**: flight phase (climb, cruise, descent), airspeed, airspeed^2, altitude, and altitude^2.

Estimated Coefficients and Statistics

Variable	Est Coeff	Std Error	t	P	VIF	95% Lower	95% Upper
Intercept	83.08	8.075	10.29	0.000	---	66.89	99.27
Climb	-0.8140	0.5649	-1.44	0.155	3.4	-1.9465	0.3185
Descent	-1.661	0.5557	-2.99	0.004	3.3	-2.775	-0.547
AirSpeed	-0.04918	0.05253	-0.94	0.353	283.6	-0.15450	0.05613
Altitude	0.3134	0.1328	2.36	0.022	43.2	0.0472	0.5796
SpeedSqr	0.1867	0.07992	2.34	0.023	280.9	0.0265	0.3470
AltSqr	-.007390	0.003148	-2.35	0.023	44.5	-.013700	-.001079
		p < .01		p < .05		VIF > 10	

Figure 1: Estimated Coefficients

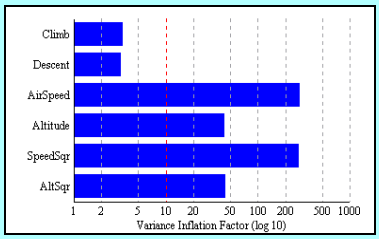


Figure 2: Variance Inflation Factors

Correlation Matrix

	NoiseLevel	Climb	Descent	AirSpeed	Altitude	SpeedSqr	AltSqr
NoiseLevel	1.0000	-0.2682	0.2383	0.9459	0.1095	0.9431	-0.0468
Climb	-0.2682	1.0000	-0.6944	-0.3187	-0.1751	-0.3507	-0.2145
Descent	0.2383	-0.6944	1.0000	0.3529	-0.2817	0.3934	-0.2737
AirSpeed	0.9459	-0.3187	0.3529	1.0000	0.0633	0.9969	-0.0867
Altitude	0.1095	-0.1751	-0.2817	0.0633	1.0000	0.0396	0.9717
SpeedSqr	0.9431	-0.3507	0.3934	0.9969	0.0396	1.0000	-0.1034
AltSqr	-0.0468	-0.2145	-0.2737	-0.0867	0.9717	-0.1034	1.0000
		p < .01		p < .05			

Figure 3: Correlation Matrix

Figure 1 shows the **estimated coefficient** and **standard error** for each predictor. From the **t-values** and **p-values** and **confidence intervals** we infer that three coefficients differ significantly from zero at $\alpha = 0.05$ (shown in green) and two at $\alpha = .01$ (shown in cyan). The **variance inflation factors** shown in Figure 2 reveal **multicollinearity** among several predictors. However, we may prefer to ignore these indications of variance inflation since they arise from squaring predictors, which was done by design. Figure 3 shows several significant correlations among the predictors (shaded green for $\alpha = 0.05$ or cyan for $\alpha = 0.01$).

ANOVA Table

Source	Sum of Squares	D.F.	Mean Square	F	P
Regression	860.47	6	143.4	103.2	0.000
Error	75.05	54	1.390		
Total	935.52	60			
R-squared	0.9198		Adj R-sqr	0.9109	
R (mult corr)	0.9591		Durbin Watson	1.21	
Std Error	1.1789		Sample size	61	

Figure 4: ANOVA Table

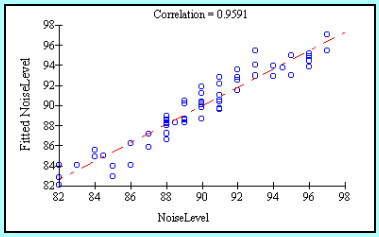


Figure 5: Actual and Fitted Y

Descriptive Statistics for All Variables

Variable	Mean	S.D.	Min	Max	Range	Type
NoiseLevel	89.68	3.949	82.0	97.0	Medium	Decimal
Climb	0.4098	0.4959	0	1	Small	Binary
Descent	0.4098	0.4959	0	1	Small	Binary
AirSpeed	332.5	48.80	230	420	Medium	Integer
Altitude	19.33	7.533	4.5	39.0	Medium	Decimal
SpeedSqr	112.9	31.91	52.900	176.400	Medium	Decimal
AltSqr	429.5	322.4	20.25	1,521.00	Medium	Decimal

Figure 6: Descriptive Statistics

The ANOVA table in Figure 4 shows that the R^2 and R^2_{adj} for this model are similar (0.92 and 0.91 respectively). The scatter plot in Figure 5 shows that Y_{actual} and Y_{fitted} are reasonably alike, a visual indication of the high R^2. The descriptive statistics in Figure 6 suggest that there are no major **data conditioning** problems.

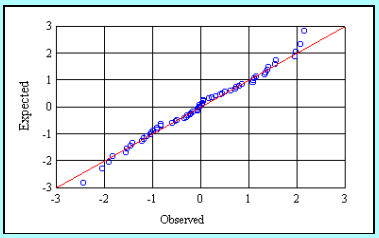


Figure 7: Residual Probability Plot

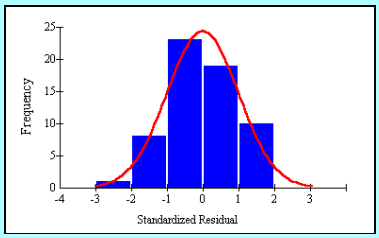


Figure 8: Residual Histogram

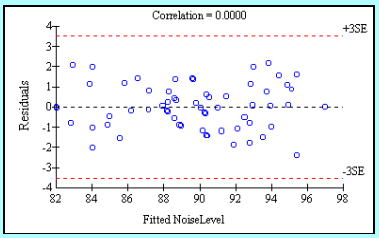


Figure 9 Residuals and Fitted Y

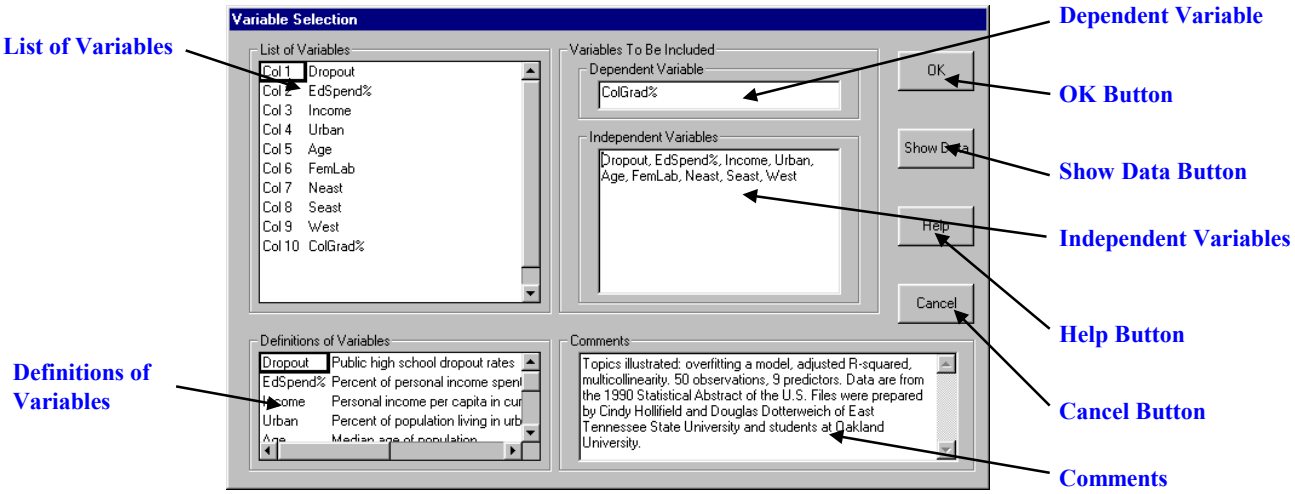
Since the **probability plot** in Figure 7 is linear except for one data point on each end, it supports the view that the residuals are normal. Since the standardized residual histogram in Figure 8 is somewhat bell-shaped and has no outliers, it supports the assumption that the disturbances are normally distributed. Since the **residual plot** against the fitted Y in Figure 9 shows no pattern, it supports the assumption that the variance is constant (homoskedastic).

Orientation to Basic Features

This module does multiple regression. You can analyze a variety of different data sets by selecting them from the Notebook or create your own using the data editor.

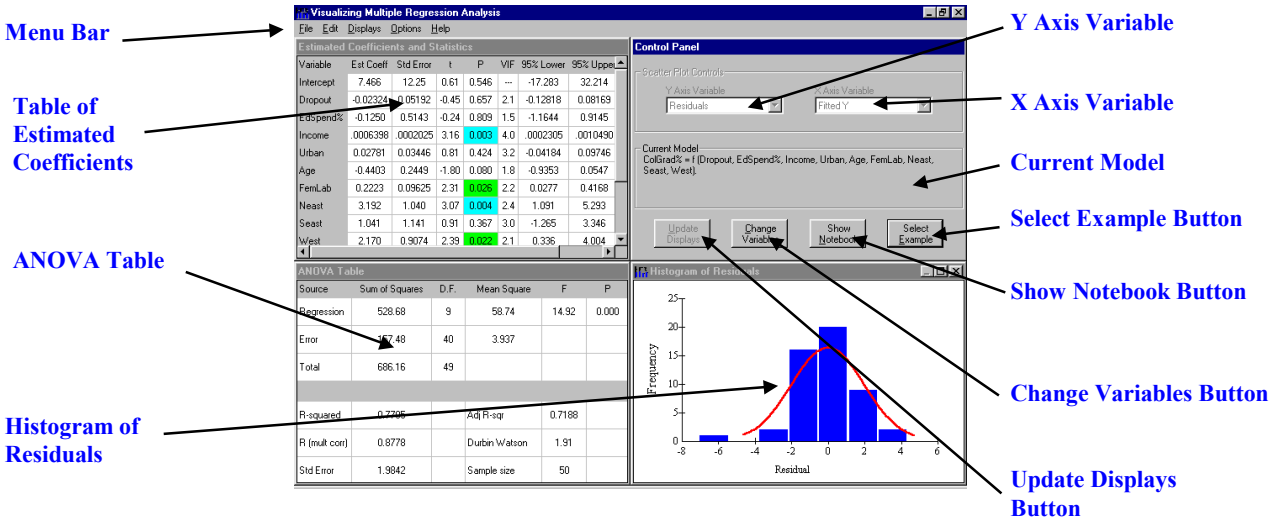
1. Opening Screen

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. Click on the Examples tab, click on Education, select College Graduates, and press OK to see the Variable Selection screen shown below. After examining this screen, click OK.



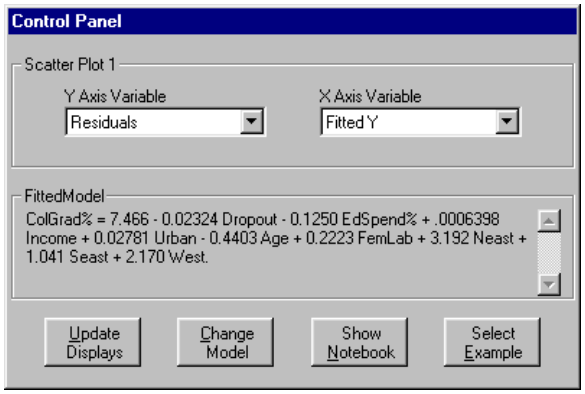
2. Main Display

A Hint appears in the middle of the display. Read it and press OK. The upper left of the screen shows a table of estimated coefficients and related statistics. The Control Panel appears on the right. On the bottom left is an ANOVA table. On the bottom right is a histogram of the residuals. Other displays may be chosen from the menu bar at the top of the screen or by right-clicking a display and using the menu that will appear.



3. **Control Panel**

   a. The control panel is only active when you are displaying one or more scatter plots. The opening screen contains no scatter plots. You can have up to three scatter plots at once. Right-click the quadrant where you want the scatter plot and then select a scatter plot from the menu that will appear. Initially, Scatter Plot 1 plots the residuals against fitted Y, Scatter Plot 2 plots Y_{fitted} against Y_{actual}, and Scatter Plot 3 plots Y_{actual} against the first predictor in your model, but you can change them. Click the scatter plot you want to change. Its title bar will be highlighted and the control panel's Y axis variable and X axis variable will display the variables displayed on the selected scatter plot. Use the variable selection combo boxes on the control panel to change the variables. The Update Displays button will flash if you have changed the Y axis variable or the X axis variable on the selected scatter plot. *The control panel only manipulates one scatter plot at a time.*



   b. Click the Change Model button to bring up the Variable Selection window again. This permits you to revise the variables in your model. Click the Select Example button to pick a new data set from the list of examples. This button's caption changes to Select Databases or Edit Data depending upon the origin of the data you are analyzing. It is a shortcut to the Notebook tab previously selected.

4. **Copying Graphs**

   Select Copy from the Edit menu (on the menu bar at the top of the screen), or the Copy option when you right click on a display, to copy the display. It can then be pasted into other applications, such as a document or spreadsheet, so it can be printed. Graphs are copied as bitmaps, and tables as tab-delimited text.

5. **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 *Visual Statistics*.

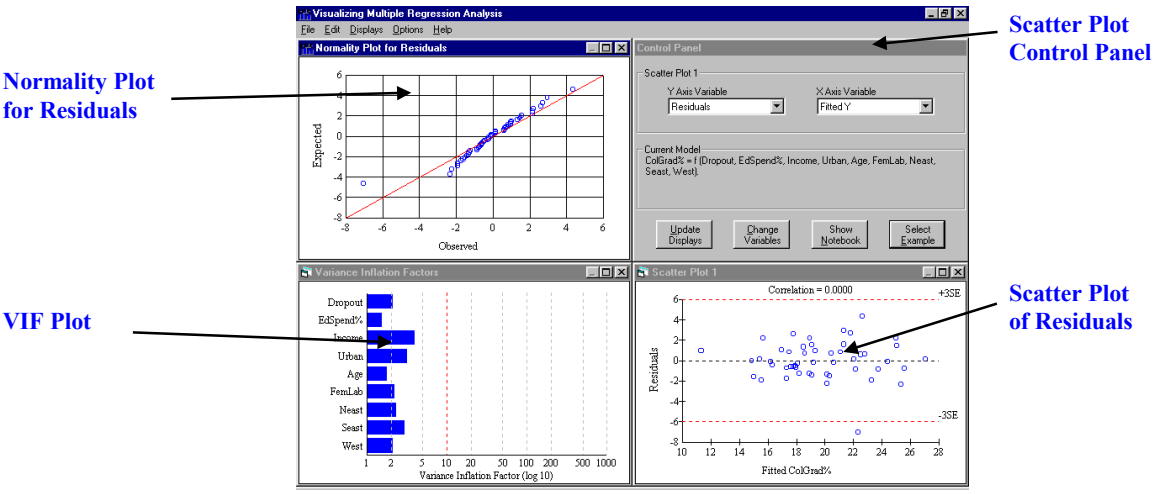
6. **Exit**

   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. Other Displays

This module offers four different types of graphs, six types of tables, and a verbal interpretation. The Hint that was displayed as the module began said, "Click the right mouse button on a quadrant to select a different display." This can also be done by selecting a quadrant and clicking on Displays in the menu bar. Only the Control Panel cannot be replaced. Under Graphs you can select a Residual Histogram, Standardized Residual Histogram, Normality Plot of Residuals, or VIF Plot. Under Tables you can select Estimated Coefficients, ANOVA Table, Residual List, Data List, Correlation Matrix, or Descriptive Statistics. Under Scatter Plots you can select up to three scatter plots. Here is a screen showing three of these displays.



2. Databases

Click the Show Notebook button to bring up the Notebook.. Select the Databases tab. There are seven databases that you can access with this module: Aircraft, Alcoholism, Body Fat, MBA Profile, Statistics Grades, States, and Time Series. Click on the database you want to examine, and you will see its description in the text window. If you use a database, you must specify your own model by selecting a dependent variable and one or more independent variables (in contrast, the Examples tab has already selected them for you). Click the Cancel button to return to the Notebook.

3. Data Editor

Click on the Show Notebook button to bring up the Notebook. To use your own data, select the Data Editor tab, and press OK. A simple multicolumn spreadsheet appears. You can enter variable names, case labels, and data directly into this editor. To title each variable, click the top cell in each column and enter the desired name. You can copy data from another spreadsheet (such as Excel or 123) and paste it into the data editor. To save the data in *Visual Statistics* format, use the File option on the menu bar. When you are finished, choose File and Exit Editor and Use Data. If you want to leave without using the data click on File and select Exit Editor and Discard Data.

Basic Learning Exercises Name _____

Estimated Regression Model

In the Notebook, click on the Examples tab, click on Education, select College Graduates, and press OK to see the Variable Selection screen.

1. Examine the variable definitions, the dependent variable, and the independent variables. Does the proposed model seem reasonable? What *a priori* sign would you anticipate for each model coefficient (i.e., would you expect a positive or negative estimated coefficient)? Explain your reasoning about causation. **Hint**: The observations are states, not persons.

Predictor	*Reasoning About Expected Sign of Estimated Coefficient*
Dropout (public high school dropout rates)	
EdSpend% (percent of personal income spent on K-12 educ)	
Income (personal income per capita in current dollars)	
Urban (percent of population living in urban areas)	
Age (median age of population)	
FemLab (labor force participation rate among females)	
Neast (1 if state is in the northeastern U.S., 0 otherwise)	
Seast (1 is state is in the southeastern U.S., 0 otherwise)	
West (1 if state is in the western U.S., 0 otherwise)	

2. Click on OK. A Hint appears in the middle of the display. Read it and press OK. Use the table of estimated coefficients and statistics to compare your expected signs (– or + or –/+) with the regression results. Use the p-values to determine significance (for example, p-values below 0.10 are significant at $\alpha = 0.10$). Describe the results in two or three sentences. **Hint**: Unless a coefficient differs significantly from zero, the actual sign is NA (that is, neither – nor +).

Predictor	*Expected Sign of Coefficient*	*Actual Sign of Coefficient*	*p-Value*	*Differs from Zero at*		
				$\alpha = 0.10$?	$\alpha = 0.05$?	$\alpha = 0.01$?
Dropout						
EdSpend%						
Income						
Urban						
Age						
FemLab						
Neast						
Seast						
West						

3. a) Write the equation for the estimated model. b) Is the magnitude of a coefficient a reliable guide to its significance (from Exercise 2)? c) How might this equation be misleading to a novice who did not look at the p-values?

Statistics of Fit

4. Record the main measures of fit from the ANOVA table and interpret them.

R^2 _____ R^2_{adj} _____ F statistic _____ p-value _____

Intermediate Learning Exercises Name _____

Residuals

5. In the Notebook, click on the Examples tab, click on Education, select College Graduates, and press OK to see the Variable Selection screen. Click OK.Describe the histogram of residuals and the fitted normal curve. Based on this histogram, are the residuals normal?

6. Right-click on the lower left quadrant, choose Graphs, and select Residual Standardized Histogram. a) Based on this histogram, are the residuals normal? Is the visual impression the same as for the residual histogram? b) What percent of the standardized residuals are within ± 2? How does this compare with a normal? c) Are there any outliers?

7. Right-click on the lower right quadrant, choose Tables, and select Residual List. Identify any outliers (residuals that exceed 3 standard errors) or unusual residuals (residuals that exceed 2 standard errors) and explain what they mean.

8. Right-click on the lower right quadrant, choose Graphs, and select Residual Normality Plot. Explain what it tells you.

Scatter Plot

9. a) On the menu bar, click Option and verify that Display Correlation on Scatterplots is selected. Then right-click on the lower right quadrant, choose Scatter Plots, and select Scatter Plot 2. What are the variables? Describe the scatter plot. b) Right-click on the lower left quadrant, choose Tables, select ANOVA Table. What relationship exists between the multiple correlation coefficient, the correlation shown on the scatter plot, and the R^2?

Correlation Matrix

10. Right-click on the lower left quadrant, choose Tables, and select Correlation Matrix. Look at the top row. Describe the p-values. Do the significance levels of the correlations between the dependent variable *ColGrads%* and the nine independent variables correspond to the predictor p-values in the table of estimated coefficients? Should they be the same?

Multicollinearity

11. In the correlation matrix, look only at the correlations among the independent variables. That is, ignore the top row and do not consider duplicates (below the diagonal). a) Among these predictors, how many correlations are significant at $\alpha = 0.05$? At $\alpha = 0.01$? What does this tell you about the data? b) How many of the correlations exceed 0.5000? c) Overall, do you see evidence of multicollinearity?

12. a) Describe the VIFs in the table of estimated coefficients. Are any VIFs large enough to cause alarm? b) What is the sum of the VIFs? Is it large enough to cause alarm? c) If there are different indications (individual VIFs versus the sum of VIFS) what does it suggest? Would removing a predictor help?

13. Right-click on the lower right quadrant, choose Graphs, and select VIF Plot. a) What is the smallest possible VIF, and what would it mean? b) Describe the VIF plot. What is the meaning of the dashed red line? c) Why is the VIF scale shown in log form? d) What is the connection between the VIF plot and the matrix of correlation coefficients?

Advanced Learning Exercises Name _____

Data Conditioning

14. In the Notebook, click on the Examples tab, click on Education, select College Graduates, and press OK to see the Variable Selection screen. Click OK.Right-click on the lower left quadrant, choose Tables, and select Descriptive Statistics. Examine the magnitude of each variable by examining its minimum, maximum, and mean. Do you see evidence of ill-conditioned data?

15. Right-click on the lower right quadrant, choose Tables, and select Data List. Do you see any evidence of unusual data values in any of the variables? If so, identify them and compare them with their respective means (see Exercise 13).

16. Right-click on the lower right quadrant and select Interpretation. Did it agree with your own analysis? Would this be helpful to a novice?

Heteroskedasticity

17. Right-click on the upper left quadrant, choose Scatter Plots, and select Scatter Plot 1. This scatter plot is called a residual plot because it plots the residuals against other variables. The default should show the residuals against fitted *ColGrad%*. As your eye moves to the right, do you see any pattern that could be called "fan-out" or "funnel-in" (heteroskedastic) or do the residuals appear to have a constant vertical dispersion (homoskedastic)?

18. Click on the Scatter Plot 1 graph to select it, then click on the control panel. The Y axis variable should be *Residuals*. Use the combo box to change the X axis variable to *Dropout* and click Update Displays. Repeat the check for heteroskedasticity (see Exercise 17). Do this systematically for all nine predictors. What do you conclude? What is unique about the binary predictors?

Model Specification

19. On the control panel, click the Change Model button. Deselect the Include Intercept option and click OK. What effect does this have on the statistics of fit in the ANOVA table? Does forcing the intercept through the origin make sense in this model?

20. On the control panel, click the Change Model button. Select the Include Intercept option. Delete the predictor with the highest p-value and click OK. Record the statistics of fit below. Repeat, deleting at each step the predictor with the highest p-value, until you have only one predictor left. Discuss the pattern of change in each statistic and try to interpret its meaning.

Model	Predictor Deleted	R^2	R^2_{adj}	F statistic
1	None	0.7705	0.7188	14.92
2				
3				
4				
5				
6				
7				
8				
9				

Time Series Data

21. Click the Show Notebook button, select Examples and Money Supply, then click OK. On the variable selection screen, also click OK. What is the main problem with this model? Plot the residuals against entry order. What does this graph tell you?

Individual Learning Projects

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

1. From the Notebook, select Time Series from the Examples tab and choose either Longley Data or Money Supply. Fit a regression model using all of the independent variables. Examine the correlation matrix and explain what it says. Click Change Model, delete the predictor with the highest VIF, and re-estimate the model. Repeat until only one predictor remains. In a table like the one below, record the variable names, estimated coefficients, t-values, and fit statistics (R^2, R^2_{adj}, standard error, and F statistic). Are the estimated coefficients stable? Discuss what this series of estimates tells you about how variance inflation, fit, and predictor significance change as correlated predictors are eliminated.

	Model 1			Model 2			...	Model k-1		
Var	Est β	t	VIF	Est β	t	VIF	...	Est β	t	VIF
Intcpt	xxx	xxx	xxx	xxx	xxx	xxx	...	xxx	xxx	xxx
X_1	xxx	xxx	xxx	xxx	xxx	xxx	...	xxx	xxx	xxx
X_2	xxx	xxx	xxx	xxx	xxx	xxx	...	xxx	xxx	xxx
:	:	:	:	:	:	...	:	:		
X_k	xxx	xxx	xxx	xxx	xxx	xxx		xxx	xxx	xxx
R^2		xxx			xxx		...		xxx	
R^2_{adj}		xxx			xxx		...		xxx	
Std err		xxx			xxx		...		xxx	
F		xxx			xxx		...		xxx	

2. From the Notebook, select Technology and choose either LCD Monitors or Turbofan Engines. Fit a regression model using all of the independent variables. Identify the weakest predictor (highest p-value), click Change Model, delete the weakest predictor, and re-estimate the model. Repeat until only one predictor remains. In a table like the one below, record the variable names, estimated coefficients, t-values, and fit statistics (R^2, R^2_{adj}, standard error, and F statistic). Discuss what this series of estimates tells you about how overall fit and predictor significance change as predictors are eliminated. Which model seems most appropriate, and why?

	Model 1		Model 2		...	Model k-1	
Var	Est β	t	Est β	t	...	Est β	t
Intcpt	xxx	xxx	xxx	xxx	...	xxx	xxx
X_1	xxx	xxx	xxx	xxx	...	xxx	xxx
X_2	xxx	xxx	xxx	xxx	...	xxx	xxx
:	:	:	:	:	...	:	:
X_k	xxx	xxx	xxx	xxx		xxx	xxx
R^2	xxx		xxx		...	xxx	
R^2_{adj}	xxx		xxx		...	xxx	
Std err	xxx		xxx		...	xxx	
F	xxx		xxx		...	xxx	

Team Learning Projects

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

1. This is a project for a team of two to four. Within one of the databases provided in this module or using another database of the team's choice, the team should agree on a dependent variable. Each team member should independently propose a regression model by choosing explanatory variables that may help explain the dependent variable (explaining the causal logic of including each predictor). Each team member should prepare exhibits that summarize his/her results (predictors used, fitted equation, fit statistics, predictor significance, variance inflation, residual normality tests, check for outliers, and data conditioning). Were the models and results similar or different? Whose model seems "best" overall? The objective of the project is to compare and contrast the choices and results made by independent researchers faced with the same task.
2. A team of three should use the time-series database, choose a dependent variable, propose a regression model (with arguments about causation for all predictors), estimate it, and do an *in-depth* analysis of the results. The first team member should analyze the estimated equation and degree of significance of predictors (estimated coefficients, t-values, p-values, confidence intervals for parameters) and fit statistics (R^2, R^2_{adj}, standard error, and F statistic). The second team member should analyze variance inflation, correlation among predictors, and data conditioning. The third team member should analyze the residual histograms, residual plots, and probability plot for evidence of non-normality, heteroskedasticity, and autocorrelation. The objective of this project is to understand the role of each aspect of model estimation, and to illustrate the special problems that confront researchers who study time series data.
3. This project is for a team of four. Two team members should use the state database, and two should use the economic time series database. Each team member should select a dependent variable, propose a multiple regression model (explaining the causal logic), and estimate it. Each team member should prepare exhibits that summarize his/her estimation results (predictors used, fitted equation, fit statistics, predictor significance, variance inflation, residual normality tests, check for outliers, and data conditioning). Compare and contrast the issues (e.g., fit, variance inflation, and residual behavior) that arise in using cross-sectional data (states) and time series data (economic time series). The objective is to understand that each data type raises its own archetypal set of issues.

Self-Evaluation Quiz

1. A multiple regression model does *not* require
   a. an intercept.
   b. a dependent variable.
   c. independent variables (predictors).
   d. assumptions about the errors.
   e. observed data for all variables.

2. When a multiple regression model is estimated, the overall fit is *not* measured by the
   a. values of the estimated coefficients.
   b. adjusted R-squared statistic.
   c. R-squared statistic.
   d. F-statistic in the ANOVA table.
   e. estimated standard error.

3. The p-value for the estimated regression coefficient for predictor j shows
   a. the probability of rejecting the null hypothesis that $\beta_j = 0$.
   b. the probability of type I error if you reject the null hypothesis that $\beta_j = 0$.
   c. the probability that the null hypothesis $\beta_j = 0$ is true.
   d. the ratio of the estimated coefficient to its standard error.
   e. all of the above.

4. The significance of an individual predictor can be assessed using
   a. the t-statistic for the predictor.
   b. the p-value for the predictor.
   c. the confidence interval for the true parameter of interest.
   d. the ratio of the estimated coefficient to its standard error.
   e. all of the above.

5. Multicollinearity
   a. is a relationship among the predictors.
   b. makes the estimated coefficients more stable.
   c. leads to small VIFs for some of the predictors.
   d. is best revealed in the p-values for the predictors.
   e. has all of these characteristics.

6. The probability plot for the residuals is *primarily* intended to offer a test for
   a. normality.
   b. heteroskedasticity.
   c. autocorrelation.
   d. ill-conditioned data.
   e. multicollinearity.

7. Which is *not* an advantage of a standardized residual histogram?
   a. Greater number of histogram bars.
   b. Easier comparison with a normal distribution.
   c. Better visual detection of outliers.
   d. Common axis scale for all regression models.
   e. Better visual check for asymmetry.

8. Evidence of possible multicollinearity would be found in
   a. several VIF that exceed 10.
   b. a sum of VIFs that is above 10.
   c. insignificant p-values for predictors that are correlated with Y.
   d. significant correlations between several predictors in the correlation matrix.
   e. all of the above.

9. Which is a characteristic of well-conditioned data?
   a. Decimal data are never rounded off to avoid spurious accuracy.
   b. Units (e.g., thousands) are set to avoid unnecessarily large or small values.
   c. Variables should have widely varying magnitudes whenever possible.
   d. Observations must be obtained from a census rather than from a sample.
   e. At least one variable must be a binary variable.

10. Data conditioning can be checked using which table display(s)?
    a. ANOVA table.
    b. Correlation matrix.
    c. Data list.
    d. Descriptive statistics.
    e. None of the above.

11. A residual plot that shows a fan-shaped pattern (more vertical dispersion moving right)
    a. suggests homoskedasticity.
    b. suggests heteroskedasticity.
    c. suggests non-normality.
    d. suggests multicollinearity.
    e. suggests of the above.

12. Overfitting a regression model refers to
    a. including too few predictors.
    b. having an excessive sample size.
    c. using ill-conditioned data.
    d. using too many independent variables.
    e. none of the above.

Glossary of Terms

Adjusted R-squared Alternate measure of the fit of a regression, based on the R^2 but with a penalty for the number of predictors. The intent is to prevent gratuitous inclusion of predictors to improve the fit. If weak predictors are added, the R^2_{adj} may decline. Conversely, R^2_{adj} may increase if weak predictors are deleted. If there are k predictors, the definition is:

$$R^2_{adj} = 1 - (1 - R^2)\left[\frac{n-1}{n-(k+1)}\right]$$

ANOVA table Summary of decomposition of variance in a regression, showing total sum of squares and its sources (regression, error) along with degrees of freedom and mean squares. See **Error sum of squares** and **Regression sum of squares**.

Autocorrelation Non-independent errors in a regression model (violation of a regression assumption). Evidence of autocorrelation may be sought by examining the residuals in a regression. Runs of residuals with the same sign (e.g., + + + – – –) would suggest *positive* autocorrelation, while runs of residuals with alternating sign (e.g., + – + – + –) would suggest *negative* autocorrelation. Autocorrelation (particularly positive) is common in time series data. See **Durbin-Watson test**.

Binary variable Variable that has only two values, used for qualitative data (e.g., male, female). Generally the values 0 and 1 are assigned, where 1 denotes the presence of the attribute of interest and 0 denotes its absence. However, other values may be used (e.g., 1 and 2). If the attribute has c categories, we need c – 1 binary variables. See **Predictor**.

Coefficient of determination See **R-squared**.

Confidence interval Upper and lower limits that are expected to enclose the true parameter (e.g., regression coefficient). For a 95% confidence interval, on average, 95 out of 100 such intervals will contain the true parameter in repeated sampling. See **Confidence level**.

Confidence level Desired probability of enclosing an unknown parameter, equal to $1 - \alpha$. Typical confidence levels are 90%, 95%, and 99%.

Correlation coefficient Measure of association between two variables, equal to the sample covariance divided by the product of the sample standard deviations of X and Y. A correlation of –1 indicates a perfect inverse relationship, 0 indicates no relationship, and +1 indicates a perfect direct relationship. The formula for the correlation coefficient is:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

Data conditioning Good statistical practice requires that each variable's units be chosen to avoid extremely dissimilar magnitudes of the variables in the data set, that decimals be chosen to avoid extremely large or small values, and that data be rounded to a reasonable number of significant digits to avoid spurious accuracy. Thus, defect rates might be rounded to 4-digit accuracy and defined as defects per 1,000,000 parts (i.e., 75.42 instead of .00007542117).

Degrees of freedom In a regression ANOVA table, *total* degrees of freedom is n – 1, *error* degrees of freedom is n – k – 1, and the *regression* degrees of freedom is k, where n is the sample size and k is the number of independent predictors in the model.

Dependent variable In a regression, the variable (denoted Y) that is placed on the left-hand side of the equation and is assumed to be affected by the independent variables X_1, X_2, ..., X_k.

Durbin-Watson test Test statistic derived from the residuals of a regression to test for first-order autocorrelation. The D-W statistic can range from 0 to 4, with a value near 2 suggesting the absence of autocorrelation.

Error sum of squares In a regression ANOVA table, the error sum of squares is the portion of the total sum of squares that is not explained by the model.

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.

Fitted model Regression equation estimated from sample data. For example, the equation ColGrad% = 7.960 + .0008861 Income + 0.02169 Urban - 0.5276 Age + 0.2023 FemLab is an example of a fitted model using data for the 50 states in the U.S.

Heteroskedasticity Non-constant variance of errors in a regression (violation of a regression assumption). Evidence of heteroskedasticity may be sought by plotting the residuals from a fitted regression against each predictor. A "fan-shaped pattern" (increasing variance as we move to the right) or a "funnel-in pattern" (decreasing variance as we move to the right) on a residual plot would suggest heteroskedasticity. See **Residual plot**.

Homoskedasticity Constant variance of errors in a regression model. If the errors are homoskedastic, there should be no discernible pattern in the regression residuals when plotted against any predictor. See **Heteroskedasticity**.

Ill-conditioned data Data set whose variables are of greatly dissimilar magnitudes, or whose values are extremely large or small. See **Data conditioning**.

Independent variable In a regression, the variables (denoted X_1, X_2, ..., X_k) that appear on the right-hand side of the equation and are thought to cause variation in the dependent variable.

Intercept Value of the dependent variable when all the independent variables in the regression model are zero. However, zero values may have little or no meaning for some predictors. Although it is often included by default, an intercept is not required in a regression model

Multicollinearity Intercorrelation among the predictors in a regression (i.e., lack of independence among the independent variables). In effect, the predictors contain redundant information, causing potential loss of accuracy in model estimation. Usually one or more predictors can be eliminated from the model without significant loss of fit. See **Variance inflation**.

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.

Non-normal errors Violation of a basic regression assumption that may affect confidence intervals and hypothesis tests. Evidence may be found in the residuals from a fitted regression. See **Probability plot**.

Ordinary Least Squares (OLS) Method of estimating a regression that guarantees the smallest possible sum of squared residuals. The residuals sum to 0 using the OLS method.

Parameter Numerical constant needed to define a particular model or distribution. A regression model's the 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**.

Predictor An independent variable in a regression model. See **Binary variable**.

Probability plot Comparison of each observed residual with the value that would be expected assuming that it came from a normal distribution. To construct a probability plot, calculate the inverse of the hypothesized normal distribution function for the i^{th} residual, and plot it against the observed i^{th} residual. This is done for all n residuals to produce a scatter plot. If the hypothesized normal distribution is correct, the scatter plot should be roughly linear along the diagonal. This is a simple, powerful visual test for normality of the sample residuals.

P-value Probability (usually two-tailed) of type II error if we reject the null hypothesis of a zero parameter. Thus, a small p-value (such as 0.01) would incline us to reject the hypothesis that the true parameter is zero.

Regression model Equation representing a relationship between a dependent variable and one or more independent variables. See **Predictor**.

Regression sum of squares In a regression ANOVA table, the regression sum of squares is the portion of the total sum of squares that is explained by the model.

Residual Difference between an actual and estimated value of the dependent variable.

Residual plot Scatter plot of the residuals against a predictor, used to check the residuals for evidence of a violation known as *heteroscedasticity* (non-constant residual variance). For k predictors we get k residual plots. To simplify matters, statisticians sometimes just look at a plot of the residuals against the fitted Y, though this method reveals less than the k plots. If the residuals are *homoskedastic*, there should be no discernible pattern. See **Heteroskedasticity**.

R-squared Also called the coefficient of determination, it is the 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.

Standard error Estimate of the standard deviation of the stochastic disturbances, using the square root of the sum of the squared residuals, divided by $n - k - 1$. It is often called the *standard error of the estimate* to distinguish it from the standard error of each regression coefficient. See **Degrees of freedom**.

Standardized residual For each observation, the residual divided by the estimated standard error of the estimate.

Sum of squares In a regression ANOVA table, the total sum of squares is decomposed into two parts: *regression* sum of squares and *error* sum of squares.

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.

Variance inflation Effects arising from interrelationships among the predictors in a model. May lead to unstable coefficient estimates that vary when predictors are added or deleted from the model, inflated standard errors, and unreliable t-values. See **Multicollinearity**.

Variance inflation factor Abbreviated VIF, it is a measure of multicollinearity for each predictor in a regression model. The VIF for predictor k is $VIF_k = [1 - R_k^2]^{-1}$ where R_k^2 is the coefficient of determination that arises when predictor k is regressed against all the other predictors. If R_k^2 is small (indicating that predictor k is not associated with the other predictors) then VIF_k will be near the ideal value of 1. A single VIF that exceeds 10 (or a sum of all VIFs that exceeds 10) is sometimes taken as an indication that multicollinearity is severe. However, data sets (especially time series data) may have VIFS of 100 or more. This can make it difficult to isolate the role of the affected predictors.

Well-conditioned data Variables whose units are chosen so that the magnitudes of the variables in the analysis are not too dissimilar, with decimals adjusted to avoid extremely large or small values, and rounded to a reasonable number of significant digits to avoid spurious accuracy. See **Data conditioning**.

Solutions to Self-Evaluation Quiz

1. a Consult the Glossary. Read the Overview of Concepts.
2. a Do Exercises 2–4. Consult the Glossary. Read the Overview of Concepts.
3. b Do Exercise 2. Consult the Glossary. Read the Overview of Concepts.
4. e Do Exercise 2. Consult the Glossary. Read the Overview of Concepts.
5. a Do Exercises 10–13 and Individual Learning Project 1. Consult the Glossary.
6. a Do Exercises 5–8. Consult the Glossary. Read the Overview of Concepts.
7. c Do Exercises 5–7. Consult the Glossary.
8. e Do Exercises 10–13 and Individual Learning Project 1. Consult the Glossary.
9. b Do Exercises 14–16. Consult the Glossary. Read the Overview of Concepts.
10. d Do Exercises 14–16. Consult the Glossary.
11. b Do Exercises 17–18. Consult the Glossary.
12. d Do Exercises 19–20. Do Individual Learning Project 2.