Chapter 4
Multiple Regression Analysis

Chapter 4
Multiple Regression Analysis

LEARNING OBJECTIVES

Upon completing this chapter, you should be able to do the following:

- Determine appropriate is is the ing a problem.
- Understand how regressi make predictions using the leas oncept.
- Use dummy variables with an understanding of their interpretation.
- Be aware of the assumptions underlying regression analysis and how to assess them.

Chapter 4
Multiple Regression Analysis

LEARNING OBJECTIVES continued . . .

Upon completing this chapter, you should be able to do the following:

- Select an e explain the difference b multaneous regression.
- Interpret the results of regression.
- Apply the diagnostic procedures necessary to assess "influential" observations.

Multiple Regression Defined

Multiple regression analysis . . . is a statistic used to analyze n a single dependent (criterion) v several independent (predictor) variables.

Multiple Regression

$$Y' = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n + e$$

Y = Dependent Variable = # of credit cards

b_0 = intercept (constant) = constant number of credit cards indepen

b_1 = change in th a unit change in family size (negr t).

b_2 = change in # of credit cards associated with a unit change in income (regression coefficient).

X_1 = family size

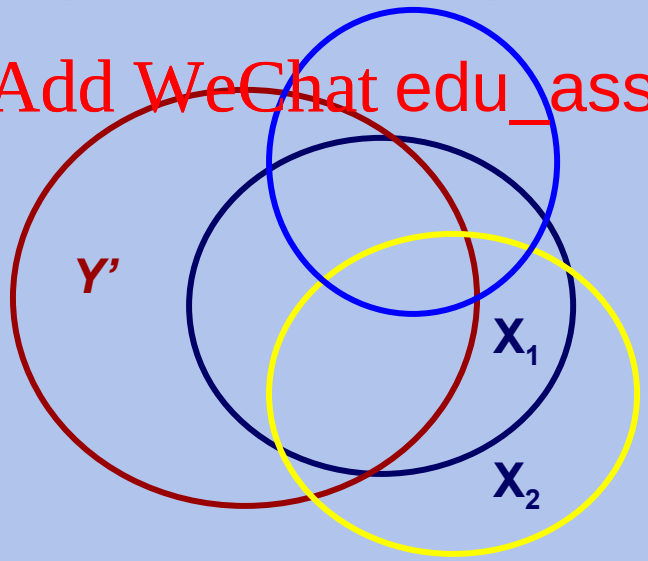
X_2 = income

e = prediction error (residual)

Variate (Y') = $X_1b_1 + X_2b_2 + \ldots + X_nb_n$

A variate value (Y') is calculated for each respondent.

The Y' value is a linear combination of the entire set of variables th al objective.



Multiple Regression Decision Process

Stage 1: Objectives of Multiple Regression

Stage 2: Research Design of Multiple Regression

Stage 3: Assumptions in Multiple Regression Analysis

Stage 4: Esti Model and As

Stage 5: Interpreting the Re ariate

Stage 6: Validation of the Results

Stage 1: Objectives of Multiple Regression

In selecting suitable applications of multiple regression, the researcher must consider three primary issues:

1. the appro rch problem,
2. specification of a statisti ship, and
3. selection of the depend ependent variables.

Selection of Dependent and Independent Variables

The researcher should always consider three issues that can affect any decision about variabl

- The the variables,
- Measurement error, e the dependent variable, and
- Specification error.

Measurement Error in Regression

Measurement error that is problematic can be addressed through either of two approaches:

- Sum
- Structural equation procedures.

Rules of Thumb 4–1

Meeting Multiple Regression Objectives

- Only structural equation modeling (SEM) can directly accommodate measurement error, but using summated scales can mitigate it when using mu
- When in doubt, include irrelevant variables (as they can o interpretation) rather than possibly omitting a relevant variable (which can bias all regression estimates).

Stage 2: Research Design of a Multiple Regression Analysis

Issues to consider . . .

- Sample size,
- Unique e nce relationship – can use iables as independents, and

- Nature of independent variables – can be both fixed and random.

Rules of Thumb 4–2

Sample Size Considerations

- Simple regression can be effective with a sample size of 20, but maintaining power at .80 in multiple regression requires a minimum sample of 50 and preferably 10 research situations.
- The minimum ratio of observations to variables is 5 to 1, but the preferred ratio is 15 to 1, and this should increase when stepwise estimation is used.
- Maximizing the degrees of freedom improves generalizability and addresses both model parsimony and sample size concerns.

Rules of Thumb 4–3

Variable Transformations

- Nonmetric variables can only be included in a regression analysis by creating dummy variables.
- Dummy variables can only be interpreted in relation to their reference category.
- Adding an addi epresents another inflection point nship.
- Quadratic and cubic polynomi erally sufficient to represent most curvilinear r s.
- Assessing the significance of a polynomial or interaction term is accomplished by evaluating incremental R2, not the significance of individual coefficients, due to high multicollinearity.

Stage 3: Assumptions in Multiple Regression Analysis

- Linearity of the phenomena measured.
- Constant terms.
- Independence.
- Normality of the error distribution.

Rules of Thumb 4–4

Assessing Statistical Assumptions

- Testing assumptions must be done not only for each dependent and independent variable, but for the variate as well.
- Graphical residual pl ression plots, the most widely used mat ty plots) are assumptions for the variate. essing
- Remedies for problems found in the variate must be accomplished by modifying one or more independent variables as described in Chapter 2.

Stage 4: Estimating the Regression Model and Assessing Overall Model Fit

In Stage 4, the researcher must accomplish three basic tasks:

1. Select a me regression model to be

2. Assess the statistical signi the overall model in predicting the dependent variable, and

3. Determine whether any of the observations exert an undue influence on the results.

Variable Selection Approaches

- Confirmatory (Simultaneous)
- Sequential Search Methods:
  - ✓ Step
    includ
    ved once
    on).
  - ✓ Forward Inclusion
    Elimination.
  - ✓ Hierarchical.
- Combinatorial (All-Possible-Subsets)

Description of HBAT Primary Database Variables

	Variable Description	Variable Type
Data Warehouse Classification Variables		
X1	Customer Type	nonmetric
X2	Industry Type	nonmetric
X3	Firm Size	nonmetric
X4	Region	nonmetric
X5	Distribution System	nonmetric
Performance Perceptions Variables		
X6	Product Quality	metric
X7	E-Commerce Activities/Website	metric
X8	Technical Support	metric
X9	Complaint	metric
X10	Advertisin	metric
X11	Product Li	metric
X12	Salesforce Image	metric
X13	Competitive Pricing	metric
X14	Warranty & Claims	metric
X15	New Products	metric
X16	Ordering & Billing	metric
X17	Price Flexibility	metric
X18	Delivery Speed	metric
Outcome/Relationship Measures		
X19	Satisfaction	metric
X20	Likelihood of Recommendation	metric
X21	Likelihood of Future Purchase	metric
X22	Current Purchase/Usage Level	metric
X23	Consider Strategic Alliance/Partnership in Future	nonmetric

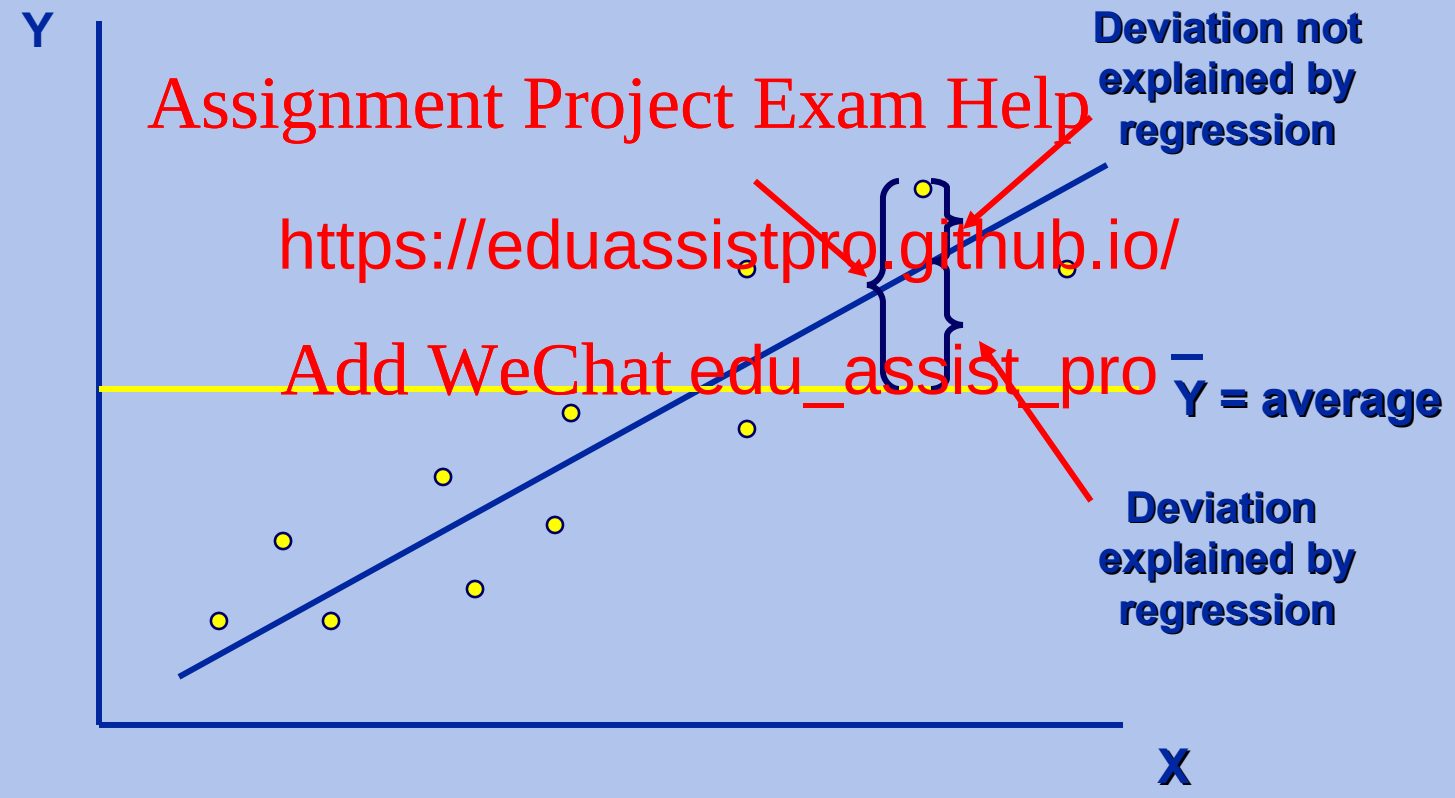
Regression Analysis Terms

- Explained variance = R^2 (coefficient of determination).
- Unexplained variance = residuals (error).
- Adjusted R-Square = reduces the R^2 by taking into account the sample size and the number of independent on model (It becomes sma observations per independent variable).
- Standard Error of the Estima a measure of the accuracy of the regression predictions. It estimates the variation of the dependent variable values around the regression line. It should get smaller as we add more independent variables, if they predict well.

Regression Analysis Terms Continued . . .

- Total Sum of Squares (SS_T) = total amount of variation that exists to be explained by the independent variables. TSS = the sum of SSE and SSR.
- Sum of Squared Errors (SS_E) = the variance in the dependent variable not acc model = residual. The objective is to obtain the smal sum of squared errors as a measure of prediction
- Sum of Squares Regression (SS_R) = the amount of improvement in explanation of the dependent variable attributable to the independent variables.

Least Squares Regression Line



Statistical vs. Practical Significance?

The F statistic is used to determine if the overall regression model is statistically significant. If the F statistic is significant, it means it is unlikely your sample will produce a large R^2 when the population R^2 is actually zero. To be considered statistically significant, a rule of thumb is there must be <.05 probability the results are due to chance.

If the R^2 is statistically significant, we then evaluate the strength of the linear association between the dependent variable and the several independent variables. R^2, also called the coefficient of determination, is used to measure the stren It represents the amount of variation in the de in all of the independent variables considered together (it also is a measure of the goodness of fit). R^2 ranges from 0 to 1. ts the amount of the dependent variable "explained" by the i riables combined. A large R^2 indicates the straight line works well while a small R^2 indicates it does not work well.

Even though an R^2 is statistically significant, it does not mean it is practically significant. We also must ask whether the results are meaningful. For example, is the value of knowing you have explained 4 percent of the variation worth the cost of collecting and analyzing the data?

Rules of Thumb 4–5

Estimation Techniques

- No matter which estimation technique is chosen, theory must be a guiding factor in evaluating the final regression model because:
  - ✓ Confirmatory Specification, the only method to allow direct testing of a pre-specified model, is also the most complex from the perspectives of specification error, model parsimony and achieving maximum predictive accuracy.
  - ✓ Sequential searc izing predictive accuracy, repres approach to model estimation, leaving the researcher a ol over the final model specification.
  - ✓ Combinatorial estimation, while considering all possible models, still removes control from the researcher in terms of final model specification even though the researcher can view the set of roughly equivalent models in terms of predictive accuracy.
- No single method is “Best” and the prudent strategy is to use a combination of approaches to capitalize on the strengths of each to reflect the theoretical basis of the research question.

Regression Coefficient Questions

Three questions about the statistical significance of any regression coefficient:

1) Was statistical significance established?
2) How d e into play?
3) Does i nce in addition to statistical s?

Rules of Thumb 4–6

Statistical Significance and Influential Observations

- Always ensure practical significance when using large sample sizes, as the model results and regression coefficients could be deemed irrelevant even when statistically significant due just to the statistical power arising from large sample sizes.
- Use the adjuste rall model predictive accur
- Statistical significance is required ship to have validity, but statistical significanc retical support does not support validity.
- While outliers may be easily identifiable, the other forms of influential observations requiring more specialized diagnostic methods can be equal to or even more impactful on the results.

Types of Influential Observations

Influential observations . . . include all observations that have a disproportionate effect on the regression results. There are three basic types based upon the nature of their impact on the regression results:

- Outliers are observations that have large residual values and can be identifie ific regression model.

- Leverage points are observation tinct from the remaining observations based o ndent variable values.

- Influential observations are the broadest category, including all observations that have a disproportionate effect on the regression results. Influential observations potentially include outliers and leverage points but may include other observations as well.

Corrective Actions for Influentials

Influentials, outliers, and leverage points are based on one of four conditions, each of which has a specific course of corrective action:

1. An error in observations or data entry – remedy by correcting the data or deleting the case,
2. A valid but exceptional observation that is explainable by an extraordinary situation – remedy by deletion of the case unless variables reflecti n are included in the regression equa
3. An exceptional observation with n nation – presents a special problem because there is deleting the case, but its inclusion cannot be justified either, suggesting analyses with and without the observations to make a complete assessment, and
4. An ordinary observation in its individual characteristics but exceptional in its combination of characteristics – indicates modifications to the conceptual basis of the regression model and should be retained.

Assessing Multicollinearity

The researcher's task is to . . .

- Assess the degree of multicollinearity,
- Deter sults, and
- Apply s if needed.

Multicollinearity Diagnostics

- Variance Inflation Factor (VIF) – measures how much the variance of the regression coefficients is inflated by multicollinearity problems. If VIF equals 0, there is no correlation between the independent measures. A VIF measure of 1 is an indication of some association between predictor variables, but generally not enough to cause problems. A maximum acceptable VIF value would be 10; anything higher would indic earity.

- Tolerance – the amount of variabl endent variable that is not explained by the other indepe es. If the other variables explain a lot of the variance of a particular independent variable we have a problem with multicollinearity. Thus, small values for tolerance indicate problems of multicollinearity. The minimum cutoff value for tolerance is typically .10. That is, the tolerance value must be smaller than .10 to indicate a problem of multicollinearity.

Interpretation of Regression Results

- Coefficient of Determination
- Regression Coefficients
  (Unstandardized – bivariate)
- Bet dized)
- Variables Entered
- Multicollinearity ??

Rules of Thumb 4–7

Interpreting the Regression Variate

- Interpret the impact of each independent variable relative to the other variables in the model, as model respecification can have a profound effect on the remaining variables:
  - ✓ Use beta weights when comparing relative importance among independent variables.
  - ✓ Regression co in the dependent variable, but c cross independent variables if the response format
- Multicollinearity may be considered it reveals a suppressor effect, but generally it is viewed as harmful since increases in multicollinearity:
  - ✓ reduce the overall R^2 that can be achieved,
  - ✓ confound estimation of the regression coefficients, and
  - ✓ negatively affect the statistical significance tests of coefficients.

Rules of Thumb 4–7 continued . . .

Interpreting the Regression Variate

- Generally accepted levels of multicollinearity (tolerance values up to .10, corresponding to a VIF of 10) almost always indicate problems with multicollinearity, but these problems may be seen at much lower le llinearity.
  - ✓ Bivariate cor result in problems, a ay be problematic if they are higher than the corre en the dependent and independent variables.
  - ✓ Values much lower than the suggested thresholds (VIF values of even 3 to 5) may result in interpretation or estimation problems, particularly when the relationships with the dependent variable are weaker.

Residuals Plots

- Histogram of standardized residuals – enables you to determine if the errors are normally distributed.
- Normal probability plot – enables you to determine if the errors are normally distributed. It compares the observed (sample) standardized re standardized residuals from
- ScatterPlot of residuals – can b t regression assumptions. It compares the st redicted values of the dependent variable against the standardized residuals from the regression equation. If the plot exhibits a random pattern then this indicates no identifiable violations of the assumptions underlying regression analysis.

Stage 6: Validation of the Results

- Additional or Split Samples
- Calculating the PRESS Statistic
- C odels
- Forecasting wit l

Description of HBAT Primary Database Variables

	Variable Description	Variable Type
Data Warehouse Classification Variables		
X1	Customer Type	nonmetric
X2	Industry Type	nonmetric
X3	Firm Size	nonmetric
X4	Region	nonmetric
X5	Distribution System	nonmetric
Performance Perceptions Variables		
X6	Product Quality	metric
X7	E-Commerce Activities/Website	metric
X8	Technical Support	metric
X9	Complaint	metric
X10	Advertisin	metric
X11	Product Li	metric
X12	Salesforce Image	metric
X13	Competitive Pricing	metric
X14	Warranty & Claims	metric
X15	New Products	metric
X16	Ordering & Billing	metric
X17	Price Flexibility	metric
X18	Delivery Speed	metric
Outcome/Relationship Measures		
X19	Satisfaction	metric
X20	Likelihood of Recommendation	metric
X21	Likelihood of Future Purchase	metric
X22	Current Purchase/Usage Level	metric
X23	Consider Strategic Alliance/Partnership in Future	nonmetric

Multiple Regression Learning Checkpoint

1. When should multiple regression be used?
2. Why should multiple regression be used?
3. What le nce and R^2 woul egression?
4. How do you use regre cients?