Topic: Ordinary Least Squares Regression

Our learning format requires that you complete the assigned readings efficiently and “intelligently”. In order to help you focus your attention on important concepts in the course reading, we have constructed a list of study questions for each topic covered in PREDICT 410. You should preview each list of study questions before you begin your reading, and then answer the questions in a notebook while you are performing you reading. If you cannot answer a question, then you should look up the answer. If you cannot find the answer, then you should post a question in your Blackboard course shell.

(1) When we refer to a 'simple linear regression', to what type of model are we referring? How does a 'simple linear regression' differ from a 'multiple regression'?

A regression equation containing one predictor variable is called a simple regression equation, 17. When an equation contains more than one predictor variable it is called a multiple regression equation, 17.

(2) In statistics, and in this course, we use the term 'regression' as a general term. What do we mean by the term 'regression'? What is the objective of a 'regression model'?

Regression analysis is a set of data analytic techniques that are used to help understand the interrelationship among variables in a certain environment, 20. The objective of a regression model is to numerically describe the relationship between the response and predictor variables, 32. Such that, the predictor variable, X, accounts for the variability of the response variable, Y, 33.

(3) What do we mean by 'linear regression'? Which equations represent a linear regression?

Linear regression means that the regression parameters enter the equation linearly, 17. Linear means to create or fall along a line, weather imposed or naturally occurring, 36.

(a) Y = b0 + b1\*X1

(b) Y = b0 + b1\*X1 + b2\*(X2^2)

(c) Y = b0 + exp(b1)\*X1

(4) Before building statistical models, it is a common and preferred practice to perform an Exploratory Data Analysis (EDA). What constitutes an EDA for a simple linear regression model? Is this EDA satisfactory for a multiple regression model, or do we need to change or extend the EDA? As we move forward in this course we will also learn about logistic regression models and survival regression models, will these methods need their own EDA or is EDA general to all statistical models?

1. Definition of the problem

2. Determining the technique

3. Use of competing techniques

4. Rough comparison of efficacy

Scatter Plot

5. Comparison in terms of a precise criterion

6. Optimization in terms of precise and inadequate criterion

7. Comparison in terms of several optimization criteria

For simple linear regression, steps 1,2,3 & 4 make sense.

The EDA model can be tweaked for specific steps, but the overall concept of intenerate learning is pivotal.

(5) In the simple linear regression model what is the relationship between R-squared and the correlation coefficient rho?

The correlation coefficient is a useful quantity for measuring both the direction and the strength of the relationship between Y and X. The CorCof only measures linear relationships, and is susceptible to outlier data points that distort the relationship, 28. The correlation coefficient computes a relational score, then when that is squared the result of the squaring equals the actual percentage of predicting that is being predicted by the predictor variable.

(6) How do we interpret a regression coefficient in OLS regression?

Generally, a regression coefficient is the constant/s in an OLS regression. It is the specific parameters the regression line is plotted along, p32. It is the slope that the OLS regression follows.

(7) Frequently, as a form of EDA for OLS regression we make a scatterplot between the response variable Y and a predictor variable X. As an assumption of OLS, the response variable Y must be continuous. However, the predictor variable X could be continuous or discrete. When the predictor variable is discrete, does a scatterplot still make sense?

The word discrete I am taking to mean that the predictor variable is not fully understood visually due to too many points of information, 21 Ratner. A scatter plot still makes sense, but how the scatter plot is assembled needs to be re-worked in order to make sense of the discrete predictor variable.

If not, what type of visual EDA does make sense?

Ratner makes the case for what he calls a smoothed scatterplot, which creates rough-free scatterplot, which reveals the underlying relationship in big data plots, 21.

Does the appropriateness of the scatterplot make sense if the discrete variable takes on many discrete values (such as the set of integers, think of dollar amounts rounded to the nearest dollar) versus only a few discrete values(such as a coded categorical variable which only takes the values 1, 2, or 3)?

The point of a smoothed scatter plot is to take the averages of both the dependent variable and independent variable and put those averages within predictor neighborhoods, or ranges, Rat 23. In response to the question, taking rounded variables will allow one to see a relationship more clearly.

(8) The simple linear regression model is a special case of 'Multiple Regression' or 'Ordinary Least Squares'(OLS) regression. (We will typically use the term OLS regression.) What are the assumptions of OLS regression?

1. Linearity Assumption: The response Y to the predictors X is assumed to be linear in its regression parameters. (Meaning it conforms to a straight line), 94 Chat.
2. The errors are assumed to be independently and identically distributed normal random variables each with mean zero and a common variance. This implies four other sub-assumptions.
   1. The error has a normal distribution, this can be seen visually through graphs.
   2. The errors have mean zero.
   3. The errors have the same variance o2, known as constant variance assumption. When this is not the case, it will be discussed in Chap 7.
   4. The errors are independent of each other, thus they have no covariance. Chapter 8 covers when this assumption is not met.
3. Assumptions of the Predictors:
   1. The predictor variables are non-random, meaning they are assumed fixed or selected in advance.
   2. The values are measured without error.
   3. The predictor variables are assumed to be linearly independent of each other.
4. All observations are equally reliable and have an approximately equal role in determining the regression results, 96.

In the final step of a regression analysis we perform a 'check of model adequacy'. What model diagnostics do we use to validate our fitted model against the model assumptions of OLS regression?

Y is approximately a linear function of X, and E measures the discrepancy in that approximation. One performs an Rsquared test to see how much of the Y variable is explained by the X variable.

To measure the quality of fit there are four diagnostics that can be used to check model assumptions of OLS.

1. The larger the t-score the greater the linear relationship.
2. Graphically, on the scatter plot the straighter the line the stronger the relationship
3. Compare actual Y and Yhat, the closer to a straight line, the stronger the relationship.
4. R2 is also another indicator of how the data fits the model, 46.

(9) How are the parameters, i.e. the model coefficients, estimated in OLS regression?

In order to estimate the parameters, one must find the straight line that gives the best fit of the points in the scatter plot of the response verses the predictor variable, 33. This is done using the least squares method, which gives the line that minimizes the sum of squares of the vertical distances from each point on the line.

How does this relate to maximum likelihood estimation?

Both approaches use the iterative process, and create a model that is linear in its parameters, 338.

How do you show the relationship between OLS regression and maximum likelihood estimation?

The relationship between OLS and maximum likelihood can be demonstrated through the same use of estimated parameters and estimator of variance (http://fhayashi.fc2web.com/hayashi%20econometrics/ml.pdf ).

(10) What is the overall F-test?

The overall F-test is a procedure that verifies that the regression slope is not a result of a sampling error, Sirkin 498. Formula on page 499.

What is the null hypothesis and what is the alternate hypothesis?

The null hypothesis is stating that the slope and r are equal to 0.

The alternative hypothesis is stating the slope and r are not equal to 0.

The overall F-test is also called the 'test for a regression effect'. Why is it called this?

The F-test can be used to discern which variables have a significant regression effect on the response variable, 75.

(11) What is the difference between R-squared and adjusted R-squared?

The difference between R-squared and adjusted R-squared is adjusted R-squared can not be interpreted as the proportion of total variation in Y accounted for by the predictors, p68. Also, adjusted R-squared tries to equally show the R value for variables with different numbers of observations.

How is each measure computed, and which measure should we prefer?

Page 68 & 69. We should prefer Rsquared because it gives more information than adjusted R.

How does the interpretation of R-squared change as we move from the simple linear regression model to the multiple regression model?

R-squared applies to each predictor variable but it is assumed that the other predictor variables are constant, and that other predictor variables are not co-related or collinear.

(12) The simple linear regression model Y = b0 + b1\*X1 has three parameters. Two of the parameters are b0 and b1. What is the third parameter?

The thirds parameter is X1.

(13) What is a sampling distribution? What theoretical distribution do the parameter estimates have in OLS regression? What distribution do we use in practice? Why do we use a different distribution in practice?

(14) The final step of a regression analysis is a 'check of model adequacy'. This 'check of model adequacy' or 'goodness-of-fit' is a very important step in regression analysis. Why? Which quantities in the regression output are affected when the fitted model deviates from the underlying assumptions of OLS regression?

(15) Nested Models: Given two regression models M1 and M2, what does it mean when we say that 'M2 nests M1'?

M2 nests M1 means that M2 can be obtained from M1 as a special case, thus M1 is a subset of values to M2, 71.

(16) What is the Analysis of Variance Table for a regression model?

The ANOVA table displays the values critical computing the F-test, and the actual F value, 74.

How do we interpret it and what statistical tests and quantities can be computed from it?

Column 1 has two factors, the regression factor and the residual factor. The regression factor equals the numbers that CAN be explained by the predictor variable (S). The residual equals the numbers that CAN NOT be explained by the predictor variables. The sum of squares column gives you the actual summation of the Regression and Residual factors and that total equals SST or Sum of Squares Total. The other column info can be found on page 74 & 75. From this table the F-test can be conducted and found, which is a test that confirms the validity in a regression model verses a sampling error.

(17) When the intercept is excluded in a regression model, how does the computation and the interpretation of R-squared change?

When dealing with a no-intercept model, one only needs to scale the model, 66.

Fit a no intercept model in SAS and check the SAS output for any noted differences.

(18) How do we interpret the diagnostic plots output by the PLOTS(ONLY)=(DIAGNOSTICS) option in PROC REG in SAS?

The diagnostic plots graphically validate the assumptions needed to conduct OLS.

(19) Why do we plot each predictor variable against the residual as a model diagnostic?

We plot each predictor variable against the residual as a model diagnostic because we want to verify that the residuals are uncorrelated to each of the predictor variables. The plots should be a random scatter of points, and pattern would prove a violation for OLS, 105, 95.

(20) Why do we perform transformations in the construction of regression models? Name at least two reasons.

Transformations take place to ensure linearity, achieve normality, or stabilize the variance, 163. This is often done because the assumptions of OLS are not met.

(21) What is multicollinearity and how does it affect the parameter estimates in OLS regression? How do we diagnose multicollinearity?

Multicollinearity is the existence of strong linear relationships among the predictor variables, 234. When collinearity is present, the parameters greatly change with the when a variable is deleted. Diagnosing collinearity can be done through noticing large changes in the estimated coefficients when a variable or data point is added, deleted, or altered, 245. Metric tests like the VIF and Conditional Indices also help diagnose collinearity (248-252).

(22) What is a Variance Inflation Factor (VIF) and how does it relate to multicollinearity?

The Variance Inflation Factors (VIF’s) is a metric that quantifies collinearity between data in OLS, 250, 305. It provides an index that measures how much the [variance](http://en.wikipedia.org/wiki/Variance) (the square of the estimate's [standard deviation](http://en.wikipedia.org/wiki/Standard_deviation)) of an estimated regression coefficient is increased because of collinearity (http://en.wikipedia.org/wiki/Variance\_inflation\_factor).

(23) Given a set of predictor variables X1,..., Xn, which are determined to show a high degree of multicollinearity between some of the variables, how should we choose a subset of these predictor variables to reduce the degree of multicollinearity and improve our OLS regression performance?

First Approach: Try to break down the collinearity by deleting variables. The collinear structure present in the variables is revealed by the eigenvectors corresponding to the very small eigenvalues. Once the collinearities are identified, a set of a set of variables can then be deleted to produce a reduced noncollinear data set. One can also use ridge regression, 314.

Predictor variables that are highly correlated are called multicollinearity.

(24) Variable Selection: How does forward variable selection work? How does backward variable selection work? How does stepwise variable selection work? [Variable\_Selection\_Procedure](#Variable_Selection_Procedure)

Extra Notes:

Iterative Process: A process in which the outputs are used to diagnose, validate, criticize, and modify the inputs, 20.

Predictor variables that are highly correlated are called multicollinerarity.

Orthogonal: The complete absence of linear relationship among the predictor variables.

Reading Notes

**Session 3:**

Chapter 11- Variable Selection Procedures:

Selecting which variables to include in a regression analysis is often one of the first and most important steps of EDA.

The process of variable selection should be viewed as an intensive analysis of the correlational structure of the predictor variables and how they individually and jointly affect the response variable under study, 303.

Two main questions have to be asked:

* Which variables should be included – Do this first…
* What form should they be included – Xsquared, logX, or both? – Then this…

Rationale for variable selection:

* Delete variables with non-zero regression coefficients that have a high level of variance, thus the retained variables can be estimated with smaller variance than if the deleted variables were included.
* The gain in precision is not offset by the bias when calculations are followed.
* Take out variables that have zero coefficients, or coefficients whose magnitudes are smaller than the standard deviation of the estimates, 301.

Model building requires the balance of two principles:

* Account for as much of the variation as possible (AKA high R-score)
* Parsimony – describe the process with as few variables as possible.

When a regression equation is constructed for prediction purposes the variables are selected with the purpose towards minimizing the Mean Square Error (MSE) of prediction, 302.

There may be several subsets of variables that are adequate in forming an equation, it depends on what the purpose is of the equation.

**Criteria for Evaluating Equations:**

Residual Mean Square (RMS): The RMS basically shows how much error is in each variable, and selecting the equation with the lower RMS is preferred, especially in PA.

Mallows Cp: Consider the Mean Square Error (MSE) of the predicted value. MSE has two components, the variance of prediction arising from estimation and a bias component arising from the deletion of variables. The Cp statistic measures the performance of the variables in terms of the standardized total means square error of prediction for the observed data points irrespective of the unknown true model. This takes into account both the bias and the variance. Sets of variables corresponding to points close to the line Cp=p are the good or desirable subsets of variance to form an equation.

Akaike Information Criterion (AIC), 305: Balances the conflicting demands of accuracy (fit) and simplicity (small number of variables). Models with a smaller AIC are better, because with models that have a similar Sum of Squared Errors (SSE) AIC penalizes the model with more variables. Models with AIC not differing by 2 should be treated as equal. This model is great for non-nested models.

Bayes Information Criterion (BIC): Similar to AIC, BIC has a far more severe penalty when n>8.

Collinearity and Variable Selection:

* Step 1: Calculate the Variance Inflation Factors (VIFs), if none are greater than ten, collinearity is not a problem.
* Step 2: Calculate the eigenvalues of the correlation matrix of the predictor variables. Small eigenvalues indicate collinearity. If the condition number is greater than 15, the variables are collinear. If any of the eigenvalues are less than .01 or the sum of the reciprocals of the eigenvalues is greater than five times the number of predictor variables, than the variables are collinear.

**Variable Selection Procedures:**

Forward Selection Procedure (FS),308 : The equation starts with no predictor variable, just a constant term. The first variable included in the equation is the one which has the highest simple correlation with the response variable Y. The second variable is then entered into the equation which has the highest correlation with Y, after Y has been adjusted for the effects of the first variable. The procedure is terminated when the last variable entering the equation has an insignificant regression coefficient. This is determined by a t-test.

Backward Elimination Procedure (BE): The backward elimination procedure starts with the full equation and successively drops one variable at a time. The first variable deleted is the one with the smallest t-test in the equation. In other words, the variable that least contributes to the reduction in error sum of squares. The procedure is stopped when all the t-tests are significant.

Stepwise Method: The stepwise method is essentially a forward selection procedure but with the added proviso that at each stage the possibility of deleting a variable is considered.

[Back to Questions](#Question_24)

Session 1

Chatterjee & Hadi

Chapter 1 & 2

Chapter 1:

**Regression Analysis**: A method for investigating functional relationships among variables.

The relationship is expressed in the form of an **Equation** or **Model**. Where the response variable, also called the dependent variable or Y, is set equal to predictor variables,

Predictor Variables are also known as: **Independent Variables, covariates, regressors, factors,** and **carriers,** p2**.** Independent Variable should not be used because the variables are rarely independent of each other.

In an equation, X is the predictor variable or a number of X’s denoted by the following syntax, X1, X2, X3, …Xp, p denotes the number of predictor variables.

This is a linear regression model: Y = β0 + β1X1 + β2X2 + …+βpXp + ε

ε = the random error representing the discrepancy in the approximation. It accounts for the failure of the model to fit the data exactly.

β = the **regression parameters** or **coefficients,** are unknown constants that are determined (estimated) from the data, 2.

Steps in Regression Analysis: Statement of the Problem, Selection of Potentially Relevant Variables, Data Collection, Model Specification, Choice of Fitting Method, Model Validation and Criticism, Using the Model for the Solution of the Posed Problem, 13.

**Statement of the Problem**: This is super important because this defines the rest of the regression analysis.

**Selection of Potentially Relevant Variables:** The variables should be thought to accurately predict the response variable.

**Data Collection:** The collected data consist of observations on η subjects. Each of the η observations consists of measurements for each of the potentially relevant variables.

η = number of subjects of observations being collected.

**Column:** Represents one η subject being measured.

**Row:** Represents an observation, which is a set of ρ+1 values for a single subject (house); one value for each of the predictor variables. The notation xij refers to the ith value of the jth variable. The first subscript refers to observation number and the second number refers to variable number. See page 16 for visual explanation. Chapter 5 deals with ANOVA.

**Model Specification:** Choosing which model to use that relates best the response variable to the predictor variables is often relegated to the experts. An equation is considered linear if the regression parameters (AKA coefficients) enter the equation linearly. If the relationship between Y and the X’s can be transformed to be linear it is called a linearizable function found in chapter 6, 17. Intrinsically nonlinear functions are functions that cannot be linearized.

**Univariate Regression:** One quantitative response variable.

**Multivariate Regression:** 2 or more quantitative response variables.

Session 1

RABE

Continued Chapter 1, p18

Types of Regression Conditions

**Univariate** 1 quantitative response variable

**Multivariate** 2 or more quantitative response variables

**Simple** 1 predictor variable

**Multiple** 2 or more predictor variables

**Linear** All parameters (coefficients) enter the equation linearly, even if transformation is necessary

**Nonlinear** Not possible to make the predictors or parameters linear

**Analysis of Variance** All predictors are qualitative variables

**Analysis of Covariance** Some predictors are quantitative and others are qualitative variables.

**Logistic** The response variable is qualitative

**Method of Fitting:** Estimate the parameters (coefficients) of the model based on the collected data. AKA parameter estimation model fitting.

**Least Squares:** a method for fitting a model that produces estimators with desirable properties – this method is most popular and primarily used in the RABE along with variants of this model. In instances in which the assumptions of OLS do not hold, other methods including maximum likelihood, ridge regression, and principal components can be used.

**Model Fitting:** Estimate the regression parameters or fit the model to the collected data using the chosen estimation method (ie OLS), 19. When estimating a dumb little hat goes on the top of the estimated coefficients and response variable. The Y-hat is call the fitted value and corresponds with **actual** fitted n values in the data. The Y-hat can also be called a predicted value when it predicts other values other than actual observations.

**Model Criticism and Selection:** Assumption must be validated or explained before any conclusions are drawn from the analysis.

**Iterative Process:** A process in which the outputs are used to diagnose, validate, criticize, and modify the inputs. – Isn’t this circular reasoning?

**A satisfactory output is an estimated model that satisfies the assumptions and fits the data reasonably well, p20. The regression equation is the final product.**

Session 1

RABE

Chapter 2

Simple Linear Regression (SLR):

In SLR, one wishes to measure both the direction and strength of the relationship between X and Y.

**Covariance:** Determines the direction of the linear relationship between Y and X. The formula can be found on page 27, the point to grasp is if the covariance is negative the linear relationship is negative, and vice versa for positive.

**Correlation Coefficient (CC):** The covariance between the standardized X and Y data. The CC is standardized and not only measures the direction of the relationship, but also measures the strength of the relationship as it relates to the data being a straight line. The CC only measures *linear* relationships, if a CC is 0 it does not mean that a relationship does not exist, rather it demonstrates there is not a linear relationship present. The CC is very susceptible to outliers, and cannot be used to predict responses.

**Regression Analysis:** An attractive extension to correlation analysis as it postulates a model that can be used to measure the direction, strength, and numerically describe variables in a model.

**Simple Linear Regression Model:** Y = β0 + β1X1 + ε

β0 and β1 are the parameters or model regression coefficients, and ε is the random error.

Y is approximately a linear function of X and ε measures the discrepancy. β1 can be interpreted as the change in Y for unit change in X. β0 = the predicted value of Y when X = O

In regression analysis we are primarily concerned with the value of Y. X is evaluated on its ability to predict Y, 33.

**Parameter Estimation**: The goal is to estimate the parameters of β0 and β1 such that it gives the best fit line of the scatter plot between X and Y.

**Least Squares Method**:Gives the line that minimizes the sum of squares of the vertical distances from each point to the line. This vertical distance represents the errors in the response variable. We want the line with the Least Squares also known as the Lease Squares regression model. The vertical distance between the fitted value and the predicted value is expresses as yi – y(hat)i = ei. This vertical distance is called the **Ordinary Least Squares** **Residuals**. The sum of all these residuals for an equation equal 0.

The least squares method assumes that Y and X are linearly related, we validate that by plotting the response verses the predictors.

Using the properties of least square estimators we can better understand the relationships by conducting statistical inference procedure (confidence interval estimation, tests of hypothesis, and goodness of fit tests).

**Tests of Hypotheses**: A formal way of measuring the usefulness of X as a predictor of Y by conducting a test of hypothesis about the regression parameter β1. See Page 37 for the formula.

TIME OUT! So far we have the Correlation Coefficient that can be used to determine the usefulness of X as a predictor of Y, backed up with the scatter plot, combine that with the OLS and we can predict stuff.

The standard error as found on page 37 is a measure of how precisely the slope has been measured, and the smaller the standard error the better.

Session 2

RABE

Chapter 3

Multiple Linear Regression:

The multiple regression represents a plane or a hyper plane (when there are more than two predictor variables.)

**(3.6)** The regression coefficients need to be unit less, this is done by centering and scaling the variables. If the equation does not have a constant (β0)) just scale the variables, 64. A centered variable is obtained by subtracting from each observation the mean of all observations. A centered response variable is (Y – ybar). The centered *j*th predictor variable is Xj – xbarj. The mean of a centered variable is zero.

**Scaling Variables:** There are two types of scaling – unit-length and standardizing. The formulas can be found on page 65. The regression coefficients obtained using the standardized version of the variables are often referred to as the beat coefficients. This is very confusing please email the teacher…

**Multiple Correlation Coefficient**: The strength of the linear relationship between Y and the set of predictor variables can be assessed through an examination of the scatter plot of Y and versus Yhat. R-squared is called the multiple correlation coefficient. When the model fits the data well, it is clear that the value of R-squared is close to unity. With a good fit, the observed and predicted values will be close to each other. The sum of errors squared will be small. If the fit is not good, the best predictor then becomes the sample mean because it minimizes the sum of squared deviations. The value of R-Squared is used as a summary measure to judge the fit of the linear model to a given body of data. That being said, a more detailed analysis is necessary to ensure that the model fits the data. Adjusted R-Squared is used to compare models with an unequal number of predictor variables. Adj R-squared cannot be interpreted as the proportion of total variation in Y accounted for by the predictors. –What, why is this?

**3**.**9 Inference for Individual Regression Coefficients:** We make statistical inference regarding the regression coefficients. The rejection of H0 means that β is likely different than 0, and hence the predictor variable Xj is a statistically significant predictor of the response variable Y, 70. – Why does proving that β is statistically significant mean that the predictor variable is statistically significant, p70?

A constant should ALWAYS be included even if the term is statistically not significant.

**Tests of Hypotheses in a Linear Model**, 71: The different hypotheses can be tested using a unified approach. The null hypothesis to be tested specifies values for some of the regression coefficients. The number of parameters to be estimated in the reduced model is smaller than the number of parameters to be estimated in the full model. H0: Reduced model is adequate | H1: Full model is adequate. We want to compare the goodness-of-fit of that to the full model compared to the nested model. lf the reduced model gives as good a fit as the full model the null hypothesis is not rejected. This is analyzed by looking at the lack of fit in the data associated with the full model, which is analyzed through the squared residuals obtained when fitting the full model. Likewise, the lack of fit for the Reduced Model (RM) is analyzed through the lens of the squared residuals when the reduced data is fitted. So if we take SSE(RM) – SSE (FM) we will find the difference which could be an increase or decrease in residuals between the models. If the difference is large, the reduced model is inadequate. To determine model adequacy, we use the F-test, 72.

Session 2

RABE

Chapter 3 & 4

Simplicity of description or the principle of parsimony is one of the most important principles in regression analysis, 75. When we want to determine whether Y can be explained adequately by a subset, we use the following hypothesis tests. H0: β2 =β4=β5=β6=0. The other hypothesis is H1 which has β0, β1 β3. If the F-Value is NOT significant the H0 value is not rejected. One can use R-squared in place of SSE for the F-test and arrive at the same conclusion for F, 77. When the reduced model only one less predictor variable, the t-test will suffice. Just square t to compare, 77. In simple regression the F and t tests are equivalent,77.

**Chapter 4: Regression Diagnostics: Detection of Model Violations**

Regression analysis works only if certain assumptions are satisfied. The focus of this book is to detect violations and correct them such that a thorough analysis of the data can be satisfied.

**Standard Regression Assumptions 94**:

**1**. The model that relates the response Y to the predictors X1… is assumed to be linear in the regression parameters. This is the linearity assumption and can be validated by checking the scatter plot of Y versus X.

**2.** The errors are assumed to be independently and identically distributed. This implies four assumptions:

**i.** The errors have a normal distribution , this can be validated by examining the residuals.

**ii.** The errors have mean zero.

**iii.** The errors have the same variance, known as the constant variance assumption, it also goes by the names homogeneity, homoscedasticity assumption. When the assumption does not hold it is called heterogeneity, and heteroscedasticity, chapter 7 covers more on this.

**iv.** The errors are independent of each other, known as the independent-errors assumption. When this assumption does not hold, the issue is called auto-correlation problem. Chapter 8 deals with this.

**3.** Assumptions about the predictors:

**i.** The predictor variables are nonrandom, which follows EDA.

**ii.** The values are measured without error.

**iii.** The predictor variables are assumed to be linearly independent of each other. If not, they are collinear and Chapter 9 & 10 deals with this.

**4.** Assumptions about the observations:

**i.** Each observation is equally reliable in its ability to predict towards the conclusion.

Gross violations of the model assumptions can seriously distort conclusions. It is very important to investigate the structure of the residuals and the data pattern through graphs.

There are three types of groupings to classify graphs: Graphs for checking linearity and normality assumptions, graphs for the detection of outliers and influential observations, diagnostic plots for the effect of variables.

One can check the linearity and normality assumptions by examining the residuals after fitting a given model to the data, 105. Plots:

Normal probability plot of the standardized residuals: The normal scores, random sample of “n” from a normal distribution, are plotted against the ordered standardized residuals. If the residuals are normally

Session 2

RABE

Chapter 4 – Using Graphs

**4.7 Checking Linearity and Normality Assumptions, 105.**

**Normal probability plot of the standardized residuals:** The normal scores, random sample of “n” from a normal distribution, are plotted against the ordered standardized residuals. If the residuals are normally

distributed, the ordered residuals should be the same as the ordered normal scores. Under the normality assumption, this plot (Q-Q plot) should be resemble a nearly straight line with an intercept zero and s slope of one. (http://statmaster.sdu.dk/courses/st111/module04/computing/index.html)

**Scatter plot of the standardized residual against each of the predictor variables.** Standard Assumption –The standardized residuals are not correlated with each of the predictor variables. If this holds, the scatter plot should be a random scatter of points.

**Scatter plot of the standardized residual verses the fitted values:** Standard assumption – the standardized residuals are also uncorrelated with the fitted values, this plot should also be a random scatter. This is the residual against the predictor value in SAS,

**Index plot of the standardized residuals.** If order does not matter, this plot is not needed. If it does matter this plot validates the assumption of independence of errors. The plot should be a scattered randomly within a horizontal band around zero.

**Leverage, Influence, and Outliers:**

In OLS we need to discern which points are outliers or unduly influence the fitted model.

Outlier: A point with a standard residual larger than 2 or 3 standard deviations away from the mean.

For the X-side of the equation, Points with a pii (p97) greater than 2(p+1) /n are generally regarded as points with high leverage. A plot of the leverage values (index plot, dot plot, box plot) will reveal points with high leverage if they exist, p108. Outliers tend to pull the equation towards them. Look at the example on page 110 for further clarification.

**Cooks Distance:** Measure the difference between the regression coefficients obtained from the full data and the regression coefficients obtained by deleting the *i*th observation. Points that have a value greater then one should be flagged as influential points.

**Role of Variables in a Regression Equation**:

What are the effects of deleting or adding one of the variables from or to the model? Compute the t-test for each variable in the model. If the t-test is large in its absolute value, keep it. This method is only valid if the underlying assumptions hold. Two plots have been proposed that are used to compliment the t-test. The added variable plot and Residual plus component plot are both used to discern the roles of variables in a regression equation, see page 118 for more information.

**Effects of an Additional Predictor:** When introducing a new variable in a regression equation, these two questions should be addressed: Is the regression coefficient of the new variable significant? Does the introduction of the new variable substantially change the regression coefficients of the variables already in the regression equation.

Session 2

RABE

Chapter 5

Qualitative Variables as Predictors:

Qualitative variables can be represented as dummy variables, which means they take on a value of 0 or 1. This demonstrates that the observations belongs to one of two possible categories, and does not reflect a numerical ordering. We still assume that the response variable is continuous and is quantitative. Chapter 12 deals with qualitative response variables, 131. When using indicator variables to represent categories, the number of these variables requires is one less than the number of categories. The category that is not represented by an indicator variable is called the base category or control group, 131 has the example.

**Section 5.4 137.**

A collection of data may consist of two or more distinct subsets, which my require a separate regression equations for each subset. Cross-sectional or time series data utilizes this approach.

disciplinary

Session 3

RABE

Chapter 11 – Variable Selection Procedures

Selecting which variables to include in a regression analysis is often one of the first and most important steps of EDA.

The process of variable selection should be viewed as an intensive analysis of the correlational structure of the predictor variables and how they individually and jointly affect the response variable under study, 303.

Two main questions have to be asked:

Which variables should be included – Do this first…

What form should they be included – Xsquared, logX, or both? – Then this…

Variables should be deleted when: the coefficients equal zero and the variables do not have a linear relationship, 301.Coefficients that are not precise got to go. Deleting variables never increases the variance for the smaller model. Comparing the precision of estimates is done when comparing the Mean Square Errors (MSE)s of the fitted model to the variances of the full model, 301. The variance of a predicted value from the subset is smaller or equal to the variance of the predicted value from the full model, 301.

Rationale for variable selection:

Delete variables with non-zero regression coefficients that have a high level of variance, thus the retained variables can be estimated with smaller variance than if the deleted variables were included. The gain in precision is not offset by the bias when calculations are followed.

When deleting variables the game is all about gaining precision in estimation and prediction.

When a regression equation is used for predictive purposes, the variables are selected with an eye towards minimizing the MSE. (Other uses for regression equations are listed on page 302.) The purpose of the equations is pivotal to building the right model and selecting the right variables. There may even be multiple subsets of variables for an equation. Good variable selection will point out a few strong subsets. The process of variable selection should be viewed as an intensive analysis of the correlation structure of the predictor variables and how they individually and jointly affect the response variable under study. Take out variables that have zero coefficients, or coefficients whose magnitudes are smaller than the standard deviation of the estimates, 301.

Model building requires the balance of two principles:

Account for as much of the variation as possible (AKA high R-score)

Parsimony – describe the process with as few variables as possible.

When a regression equation is constructed for prediction purposes the variables are selected with the purpose towards minimizing the Mean Square Error (MSE) of prediction, 302.

There may be several subsets of variables that are adequate in forming an equation, it depends on what the purpose is of the equation.

**Criteria for Evaluating Equations:**

Residual Mean Square (RMS): The RMS basically shows how much error is in each variable, and selecting the equation with the lower RMS is preferred, especially in PA, 303.

Session 3

RABE

Chapter 11 – Variable Selection

Mallows Cp: Consider the Mean Square Error (MSE) of the predicted value. MSE has two components, the variance of prediction arising from estimation and a bias component arising from the deletion of variables. The Cp statistic measures the performance of the variables in terms of the standardized total means square error of prediction for the observed data points irrespective of the unknown true model. This takes into account both the bias and the variance. Sets of variables corresponding to points close to the line Cp=p are the good or desirable subsets of variance to form an equation.

Akaike Information Criterion (AIC), 305: Balances the conflicting demands of accuracy (fit) and simplicity (small number of variables). Models with a smaller AIC are better, because with models that have a similar Sum of Squared Errors (SSE) AIC penalizes the model with more variables. Models with AIC not differing by 2 should be treated as equal. This model is great for non-nested models.

Bayes Information Criterion (BIC): Similar to AIC, BIC has a far more severe penalty for *p* when n>8.

Collinearity and Variable Selection:

Step 1: Calculate the Variance Inflation Factors (VIFs), if none are greater than ten, collinearity is not a problem.

Step 2: Calculate the eigenvalues of the correlation matrix of the predictor variables. Small eigenvalues indicate collinearity. If the condition number is greater than 15, the variables are collinear. If any of the eigenvalues are less than .01 or the sum of the reciprocals of the eigenvalues is greater than five times the number of predictor variables, than the variables are collinear.

**Variable Selection Procedures:**

Forward Selection Procedure (FS),308 : The equation starts with no predictor variable, just a constant term. The first variable included in the equation is the one which has the highest simple correlation with the response variable Y. The second variable is then entered into the equation which has the highest correlation with Y, after Y has been adjusted for the effects of the first variable. The procedure is terminated when the last variable entering the equation has an insignificant regression coefficient. This is determined by a t-test.

Backward Elimination Procedure (BE): The backward elimination procedure starts with the full equation and successively drops one variable at a time. The first variable deleted is the one with the smallest t-test in the equation. In other words, the variable that least contributes to the reduction in error sum of squares. The procedure is stopped when all the t-tests are significant.

Stepwise Method: The stepwise method is essentially a forward selection procedure but with the added proviso that at each stage the possibility of deleting a variable is considered.

[Back to Questions](#Question_24)

Ratner

Session 1&2,3,4

Chapter 2, 10, 12

P 17

The correlation coefficient, denoted by r, is a measure of the strength of the straight-line or linear relationship between two variables, 17.

Data Mining: The process of revealing unexpected relationships in data.

Big Data: a sample size greater than 50,000 individuals, also it can have 50 variables or more, 8.

Scatter plots (SP) with more information are often less informal than those with less information. A SP with too much information to make discernible is called a SP in the rough. Taking the information and breaking it down to understand the relationships will require creating a smoothed scatterplot.

Chapter 10, 177

The seven step EDA cycle serves as a notable solution to variable selection in regression. EDA allows the statistician to assess whether a given variables needs transformation or reexpression, 178. Inability of construction of new variables is a serious weakness of the variables selection methodology.

Variables selection follows these three selection components:

Stats tests and significance tests (F, chi-square, and ttests)

Stat Criteria (R-squared, adj r-squared, Mallows’s Cp, and Mean Square Error.)

Statistical Stopping rules (p value flags for variable entry/deletion/staying in a model).

These components distort the original theoretical and inferential meanings when used in this setting, 179. Problems with stepwise selection: Yields confidence limits that are far too narrow.

It will not produce the best model if there are redundant predictors.

Variable selection is about finding the necessary variables among the complete set by deleting irrelevant and redundant variables. \*\*\*The F-test verifies a continuous dependent variable, 181\*\*\*

Page 182: Weaknesses in Stepwise Selection

Data mining and creating new variables from the original variables is an enhanced variables selection technique, 184. Current methods do not alter the variables to get more out of them.

His technique utilizes EDA as the variables selection technique.

**Chapter 12, 213**

The p-value validates that Xi is important for making good predictions. The statistical p-value is the probability of observing a value of the sample statistic (MSE or Bi) as extreme or more extreme than the observed value given the NH is true. If the p-value is small it reduces the prediction error and is thus significant. The p-value is an indicator of the likelihood that the variable has some predictive importance, not an indicator of how much importance. A smaller p-value implies a greater likelihood of some predictive importance, not a greater predictive importance. The p-value is affected by sample size, as sample size increases p-value decreases. Variables associated with the greatest reduction in prediction error can be declared important predictors, 217.

The common interpretation for the coefficient or bi is, the average change in the predicted Y values associated with a unit change in Xi when the other X’s are held constant. It is a measure of the linear relationship between Y and Xi when the other influences are held constant.

Ratner

Session 4

Chapter 12

When ranking predictor variables, it is very hard to pick a best based on the size of the coefficient because the units are different, 221. The correct ranking of predictor variables in terms of their effects on the prediction of Y is the ranking of the variables by the magnitude of the standardized regression coefficient (SRC). The SRC is produced by multiplying the original regression coefficient by a conversion factor. The transformation equation that converts a unit specific raw regression coefficient into a unitless standardized regression coefficient is Conversion Factor (CF) \* Raw regression coefficient for X1. The CF is defined as the ratio of a unit measure of y variation to a unit measure of Xi variation. The Standard Deviation is often the technique used for normally distributed variables.

The H-spread is defined as the difference between the 75th percentile and 25th percentile of the variable distribution. This can also be used instead of the STDV, 222.

Chapter 13

Two essential characteristics of a predictive model are its reliability and validity. Reliability refers to a model yielding consistent results, like how dependable is the model for predictive purposes. A predictive model is reliable to the extent the model is producing repeatable predictions, 225.

Model validity refers to the extent to which the model measures what it is intended to measure with respect to a given criterion (a predictive model criterion is small prediction errors). If the reliability of the model is low, so too will the validity of the model. Models need to be kept up to date in order to be kept valid. Content Validity: refers to the variables in a model in that the content of the individual variables and the ensemble of the variables are relevant to the purpose of the model. Reliability does not imply validity, 227. A reliable regression model is predicting consistently (precisely), but may not be predicting what it is intended to predict. These two concepts complement each other.

**The Average Correlation**: provides a quantitative criterion for assessing competing predictive models and the importance of the predictor variables defining a model. The average correlation is the mean of the absolute values of the pairwise correlation coefficients of either the upper of lower triangular matrix, not both, page 230 for clarification. A desirable value of the average correlation is “small positive values”, 232. This tool helps to clarify the reliability of the model.

Small values indicate the predictor variables are not highly inter-correlated, which means the variables are not collinear (collinearity). Model performance is *not* affected by the condition of multicollinearity as long as the condition of multicollinearity remains the same as when the model was initially built.

Average correlation values in the range of .35 or less are desirable. Average correlation values that are greater than .55 are not desirable as they indicate the predictor variables are excessively redundant.

Session 5

ALR

Chapter 1

Last Sync-session

Goodness of Fit equals statistical inferences and the conclusions from the P-values

ROC Curve: When being wrong is really bad – You don’t want false positives.

Lift Chart: Marketing pieces

First Sync-Session

10/06/2012

Course Details/Questions

Study Groups work well if there is an agenda and a strong leader. Dr. Bhatti has created course questions we should use the discussion questions.

Questions:

On Page 44 & 68 of RABE, the authors write that the strength of a linear relationship can be assessed through the examination of a the scatter plot of Y verses Yhat (especially in multiple regression), yet, they do not write about how to interpret the scatter plot. Where can we go to find out how to interpret the Y verses Y-hat scatter plot?

Page 63 of RABE writes about the adjustments made for multiple regression coefficients, the process described is very difficult to comprehend for a novice given that the examples in the book do not necessarily break down the regression equations, and the chart for partial residuals needs more explaining. How can we better understand this material presented in section 3.5?

3.6.1 Centering and Scaling in Intercept Models: The calculations behind unit-length scaling and standardizing are given in the formula setting and not executed in example. It is rather complicated and hard to follow, how would recommend we better understand the material? 3.6.2 also follows this model, and not seeing an example makes it very hard to comprehend.

3.7 Properties of the Least Squares Estimators: The properties use words and symbols not seen anywhere else in the book, how can we better understand this material?

On page 68, RABE says “When the model fits the data well, it is clear that the value of R-squared is close to unity.” What is meant by the word unity? Page 81 states that β1 +β3 = 1, then it is referred that the two partial regression coefficients equal unity. In my opinion, it is assumed that unity means 1.

On page 69 RABE states “… adjusted R-squared cannot be interpreted as the proportion of total variation in Y accounted for by the predictors.” This is very confusing given that R-squared is often explained as the proportion of the total variability in the response variable that can be accounted for by the predictor variables. How then should we view or interpret adjusted R-squared?

Why does proving that β is statistically significant mean that the predictor variable is statistically significant, p70?

On page 131, an example is used to demonstrate using dummy variables. In the example for education, is it assumed that all employees must have at least a high school education? The value E1 = 0 and E2 = 0 is

Questions

Session 2

Chapter 5

Continued from top page…

supposed to mean that the person has an advanced degree, but this assumption only holds if all employees must have a high school education or higher?

On page 135 of RABE, an example is being used to explain using dummy variable in regression analysis. Would you please explain why and how the last two variables enter into the model. Will we learn how to do this with SAS? Would you please explain more about how the coefficients are linear model 5.2?

On page 131 of RABE, the authors are explaining the linearity of the variables. It is clear that the management variable is not linear, but they do not explain how this violates the OLS assumption. How do we validate or get around the OLS assumption of linearity with dummy variables?

Chapter 12 Ratner

When conducting the Standardized regression coefficient, how does one find the standard deviation of Y? Page 222 Ratner.

**Comments by Professor Bhatti**

R-squared is not the poster child of analytics.  We want to develop a general sense of validating models.  Furthermore, R-Square is only a valid measure for OLS regression models and only maintains it's standard interpretation if the model contains an intercept.

Cross-Validation: The concept of cross-validation deals with how well a model will be able to estimate a prediction. Model-fit statistics, like correlation coefficient or coefficient of determination, indicate how well the model describes the data, not how well it predicts the data. The model will understand, or know, the relationship in the data between the response and predictor variables. This leads the model-fit statistics to overestimate, or over fit, the predictive skill of the model. To estimate the predictive skill of the model, a separate sample is used instead of the sample that was used to construct the model. Specifically, the testing sample should be completely distinct from the training sample.

In simple terms, cross-validation is when a model is formulated with one sample of the population and is then checked for validity by using another sample from the same population.

**Second Sync-Session:**

Each bullet is correlated to a question.

**variable.**  Use these discretize variables in Part 1(5).  The choice of the four cut-points is completely heuristic.  You want to choose cut-points to create separation between the two classes. For example if I told you that A15 took a value below 1.5 and asked you to guess which class the observation was from (Y=0 or Y=1), what would you guess?  How about for (A15 >=1.5) and (A15 < 50)?

### 5.1 Key concepts and definitions

#### Multiple Regression

* **Multiple regression:** In multiple regression analysis, we are studying the relationship between one dependent variable and several independent variables (called predictors). The regression equation takes the form
  + *Y* =*b*0+ *b1x1* + *b2x2* …+ *bp*+ *e*,
  + where *Y* is the dependent variable, the *b*'s are the regression coefficients for the corresponding *x* (independent) terms, *b*0 is a constant or intercept, and *e* is the error term reflected in the residuals. The parameters of the regression equation are estimated using the ordinary least squares method (OLS).
* **Ordinary least squares:** This method derives its name from the criterion used to draw the best-fit regression line: a line such that the sum of the squared deviations of the distances of all the points to the line is minimized.
* **Intercept**: The intercept, *b*0, is where the regression plane intersects with the Y-axis. It is equal to the estimated *Y* value when all the independents have a value of 0.
* **Regression coefficient**: Regression coefficients *b*i are the slopes of the regression plane in the direction of *x*i. Each regression coefficient represents the net effect the *i*th variable has on the dependent variable, holding the remaining *x*'s in the equation constant**.**
* **Beta weights** are the regression coefficients for standardized data. Beta is the average amount by which the dependent variable increases when the independent variable increases one standard deviation and other independent variables are held constant. The ratio of the beta weights is the ratio of the predictive importance of the independent variables.
  + **Standardized** means that for each datum the mean is subtracted and the result divided by the standard deviation. The result is that all variables have a mean of 0 and a standard deviation of 1.
* **Residuals** are the difference between the observed values and those predicted by the regression equation
* **Dummy variables**:Regression assumes interval data, but dichotomies may be considered a special case of intervalness. Nominal and ordinal categories can be transformed into sets of dichotomies, called dummy variables. To prevent perfect multicollinearity, one category must be left out**.**
  + **Interpretation of *b* for dummy variables**. For *b* coefficients for dummy variables, which have been *binary coded* (the usual 1=present, 0=not present), *b* is relative to the *reference category* (the category left out).
  + **Multiple R:** The correlation coefficient between the observed and predicted values. It ranges in value from 0 to 1. A small valueindicatesthatthere is little or no linear relationship between the dependent variable and the independent variables.
  + **Multiple *R* 2** is the percent of the variance in the dependent variable, explained by the independent variables. It is also called the coefficient of multiple determination. Mathematically, *R*2 = [ 1 - (SSE/SST) ] , where

         **SSE =** error sum of squares =  (*Y*i - *Est* *Y*i) 2 where Yi is the actual value of *Y* for the *i*th case and *Est* *Y*i is the regression prediction for the *i*th case.

         **SST** = total sum of squares = (*Y*i - Mean*Y*) 2

         **Adjusted R-Square:** When there are a large number of independent variables, it is possible that *R*2 may become artificially large, simply because some independent variables' chance variations "explain" small parts of the variance of the dependent variable. It is therefore essential to adjust the value of *R*2 as the number of independent variables increases. In the case of a few independent variables, *R*2 and adjusted *R*2 will be close. In the case of a large number of independent variables, adjusted *R*2 may be noticeably lower.

         **Multicollinearity** is the intercorrelation of the independent variables. The values of *r*2's near 1 violate the assumption of no perfect collinearity, while high *r*2's increase the standard error of the regression coefficients and make assessment of the unique role of each independent variable difficult or impossible. While simple correlations tell something about multicollinearity, the preferred method of assessing multicollinearity is to compute the determinant of the correlation matrix. Determinants near zero indicate that some or all independent variables are highly correlated.

         **Partial correlation** is the correlation of two variables while controlling for a third or more other variables. For example *r*12.34 is the correlation of variables 1 and 2, controlling for variables 3 and 4. Partial correlation *r*12.34 equal to uncontrolled correlation *r*12  No effect of control variables Partial correlation near 0  Original correlation is spurious**.**

* **Stepwise Regression:** Stepwise regression is a sequential process for fitting the least squares model, where at each step a single predictor variable is either added to or removed from the model in the next fit.

#### Multiple Classification Analysis

* **Multiple classification analysis:** Multiple Classification Analysis (MCA) is a technique for examining the interrelationship between several predictor variables and one dependent variable in the context of an additive model Independent variables may be measured on nominal or ordinal scales and the dependent variable may be interval scale or a dichotomy.
* **Additive model:** Such a model assumes that the dependent variable can be predicted from an additive combination of the independent (or predictor) variables. In other words, they assume that the average score on the dependent variable for a given set of individuals (objects or cases) is predictable by adding the effects of several predictors.

* **Eta:** Eta indicates the ability of a predictor, using the given categories, to explain variation in the dependent variable.

* **Eta square:** Eta2 is the correlation ratio and indicates the proportion of the total sum of squares, explained by the predictor.

* **MCA Beta:** This is directly analogous to Eta statistic, but is based on the adjusted means rather than the raw means. Beta is a measure of the ability of a predictor to explain variation in the dependent variable, after adjusting for the effects of all other predictors. Note that this is not in terms of percentage of variance explained.
* **Multiple correlation coefficient squared**: This coefficient indicates the proportion of variance explained in this run of the program.

* **Adjustment for degrees of freedom:** This is the factor used to correct for capitalizing on chance in fitting the model in the particular sample being analyzed**.**

* **Multiple correlation coefficient squared (Adjusted):** This coefficient estimates the proportion of variance in the dependent variable, explained by the predictor variables**.**