Study

My study notes will draw heavily from the required texts and multimedia. I will also draw from external sources that I find to be adept at explaining a particular topic. If something is referenced here, it is because I found it to be very useful in understanding a topic.

Standard Mathematical and Statistical Notation

Notes below are from the following sources; [Bhatti 2011a].

Vector and Matrix Notation

- A *scalar* is a number. *Scalars* are represented by lower case letters from the beginning of the alphabet such as a, b, c etc.
- A *vector* is a $n \times 1$ array defined with the mathematical operations of addition and multiplication. The standard convention is for all vectors to be column vectors, i.e. they are 'long' with n rows and 1 column. Vectors are represented as **bold** faced lower case letters frequently from the end of the alphabet, such as $\mathbf{x}, \mathbf{u}$, and $\mathbf{v}$. the ith entry of a vector $\mathbf{u}$ is denoted by $\mathbf{u}[i] = u_i$.
- A *matrix* is a $n \times m$ array defined with the mathematical operations of addition and multiplication. Matricies are represented by a bold face upper cased letter such as $\mathbf{A}, \mathbf{W}, \mathbf{X}$, etc. The (i, j)th entry of a matrix $\mathbf{A}$ is denoted by $\mathbf{A}[i, j] = a_{ij}$.
- The transpose of a $(n \times 1)$ column vector $\mathbf{a}$ is the $(1 \times n)$ row vector $\mathbf{a}^T = [a_1 \dots a_n]$. Sometimes the transpose $\mathbf{a}^T$ is denoted by $\mathbf{a}'$.
- The transpose of a $(n \times m)$ matrix $\mathbf{A}$ is the $(m \times n)$ matrix $\mathbf{A}^T$ where $\mathbf{A}[i, j] = \mathbf{A}^T[j, i]$. When a matrix is transposed, the rows become the columns and the columns become the rows.
- It is preferred to use the T notation $\mathbf{a}^T$ instead of the "prime notation" $\mathbf{a}'$.

Random Variable Notation

Random variables are how you develop calculus based probability theorey.Random variables are the unkown statistical experiement that generate data. Statistical theory is based upon the concept of random sampling.

- Random variables are denoted by capital letters from the end of the alphabet such as U, V, X, Y, or Z.
- The *observed value* of a random variable is denoted by the lower cased counterpart such as u, v, x, y, or z.
- When we have a *random sample* of independent and identically distributed (iid) random variables, we will index the variables in a set such as $X_1, X_2, \dots, X_n$ for the random variables and $x_1, x_2, \dots, x_n$ for the observed values.
- Random variables are used to devleop statistical estimators. Observes values of random variables are used to compute statistical estimates.
- Random variable notation can become convoluded when we move to multivariate random variables. Pay attention to how an author presents these concepts in text.

'Distribution'

The term *distribution* is used throughout all statistical applications and discussions. Loosly speaking, the term distribution is meant to describe how a group of values are related to either each other or to the range of values on which they are defined (their *support*).

The term *distribution* is used rather sloppily. If you don't understand the context you wont understand the use. The term *distribution* is mapped to many related concepts. In general the term *distribution* is related to the characterization of a random variable, or data generated by a random variable.

There are many mathematical notations for characterizing a statistical distribution. The choice of characterization will depend on the context and the existence of the characterization. A random variable can be characterized by any of the following functions.

- The *cumulative distribution function* (cdf), denoted by $F(x) = Pr(X \leq x)$. the cdf will exist for all random variables, and in general is why we use the term "distribution" so looslely throughout statistics. cdf exists for all random variables. From data you can always estimate a distribution function.
- The *probability density function* (pdf) for continuous random variables, denoted by $f(x)$, or the *probability mass function* (pmf) for discrete random variables, denoted by $p(x)$. Note that neither of these functions are guaranteed to exist. A random variable that can be described with a cdf will not always possess a pdf or pmf.
- Transformation functions such as the moment generating function $m(t) = \mathbb{E}[\exp(tX)]$ and the characteristic function $\phi(t) = \mathbb{E}[\exp(itX)]$. Transformation functions are not used when working with data, they may be used to develop conceptual underpinnings of modeling.
- Specialized representations for particular applications such as the *hazard fucntion* $h(t) = \frac{f(t)}{S(t)}$ and the *survival function* $S(t) = 1 - F(t)$ used in Survival Analysis. Survival function is a simple map of the cdf. The hazard function allows you to get a generic representation of a survival function.
- In data analysis distributions can be analyzed using the empirical cdf, the histogram, the Quantile-Quantile plot, and the Kolmogorov-Smirnov test.
- If you need to assess the distribution of residuals in linear regression and compare that to the assumption that they are normally distributed.

Mathematical Expectation

Mathematical Expertation is the theoretical averaging of a random variable with respect to its distribution function. In this sense the pdf of pmf act as a weight function that allows you to find the "center" of the distribution.

For a continuous random variable X with pdf function $f(x)$, the mathematical expectation of X can be computed by

$$\mathbb{E}[X] = \int x f(x) dx$$

For a discrete random variable X with pmf function $p(x) = Pr(X = x)$, the mathematical expectation of X can be computed by

$$\mathbb{E}[X] = \sum_x x p(x)$$

$\mathbb{E}[X]$ is also referred to as the first moment of X.

Expectation, Variance, and Covariance as Mathematical Operators

Let X denote a random variable. Consider the affine transformation $aX + b$.

- $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$
- $\text{Var}[aX + b] = a^2\text{Var}[X]$

Let X and Y be random variables with a joint distribution function. (In the continuous case we would denote this joint distribution function by the join density function $f(x, y)$.) Consider the linera transformtions aX and bY.

- $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$
- $\text{Var}[aX + bY] = a^2\text{Var}[X] + b^2\text{Var}[Y] + ab\text{Cov}[X, Y]$

Here the reader should note that in general $\text{Cov}[aX + b, cY + d] = ac\text{Cov}[X, Y]$. If X and Y are independent random variables, then $\text{Cov}[X, Y] = 0$. The converse of this statement is not true except when both X and Y are normally distributed. In general $\text{Cov}[X, Y] = 0$ does not imply that X and Y are indepdented random variables.

Statistical Assumptions for Ordinary Least Squares Regression

Notes below are from the following sources; [Bhatti 2011b].

- In Ordinary Lease Squares (OLS) regression we wish to model a continuous random variable Y (the response variable) given a set of *predictor variables* $X_1, X_2, \ldots, X_k$.
- While we require that the response variable Y will be continuous, or approximately continuous. The *predictor variables* $X_1, X_2, \ldots, X_k$ can be either continuous or discrete.
- It is fairly standard notation to reserve k for the number of predictor variables in the regression model, and p for the number of parameters (regression coefficients or βs).
- When formulating a regression model, we want to explain the variation in the response variable by the variation in the predictor variable.

Statistical Assumptions for OLS Regression

There are two primary assumptions for OLS regression:

1. The regression model can be expressed in the form

   $$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \epsilon$$

   Notice that the model formulation specifies error term ϵ to be additive, and that the model parameters (βs) enter the modeling linearly, that is, β_i represents the change in Y for a one unit increase in X_i when X_i is a continuous predictor variable. Any statistical model in which the parameters enter the model linearly is referred to as a *linear model*.

2. The response variable Y is assumed to come from an independent and identically distributed (iid) random sample from a $N(\mathbf{X}\beta, \sigma^2)$ distribution where the variance of σ^2 is a fixed but unkown quantity. The statistical notation for this assumption is $Y \ N(\mathbf{X}\beta, \sigma^2)$.

Linear Versus Nonlinear Regression

Remember that a *linear model* is linear in the parameters, not the predictor variables.

- The following regression models are all linear regression models:

  $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \epsilon$$

  $$Y = \beta_0 \beta_1 \ln(X_1) + \epsilon$$

- The following regression models are all nonlinear regression models:

  $$Y = \beta_0 \exp(\beta_1 X_1) + \epsilon$$

  $$Y = \beta_0 + \beta_2 \sin(\beta_1 X_1) + \epsilon$$

- If you know a little calculus, then there is an easy mathematical definition of a nonlinear regression model. In a nonlinear regression model at least one of the partial derivatives will be dependent on a model parameter.
- Any quantity that has a β in front of it counts as a degree of freedom used, and subsequently counts as a predictor variable.
- A hint to identify a nonlinear model is when a parameter is within a function, specifically a nonlinear function.

Distributional Assumptions for OLS Regression

The assumption $Y\ N(\mathbf{X}\beta, \sigma^2)$ can also be presented in terms of the error term ϵ. Most introductory bookos present the distributional assumption in terms of the error term ϵ, but more advanced books will use the standard Generalized Linear Model (GLM) presentation in terms of the response variable Y.

In terms of the error term ϵ the distributional assumption can also be presented as:

- The error term $\epsilon\ N(0, \sigma^2)$. Since $Y\ N(\mathbf{X}\beta, \sigma^2)$, then $\epsilon = Y - \mathbf{X}\beta$ has a $N(0, \sigma^2)$.

Distributional Assumptions in Terms of the Error

1. The errors are normally distributed.
2. The errors are mean zero.
3. The errors are independent and identically distributed (iid).
4. The errors are *homoscedastic*, i.e. the errors do not have any correlation "in time or space".

When we build statistical models, we will check the assumptions about the errors by assessing the model *residuals*, which are our estimates of the error term.

Homoscedasticity: a sequence or vector of random variables is *homoscedastic* if all random variables in the sequence or vector have the same finite variance. This is also known as the *homogeneity of variance*. [Wikipedia 2015a]

You'd need a pretty gross violation of *homoscedastic* in the kind of problems that we work with today.

Further Notation and Details

When we estimate an OLS regression model, we will be working with a random sample of response variables $Y_1, Y_2, \ldots, Y_n$, each with a vector of predictor variables $[X_{1i}, X_{2i}, \ldots, X_{ki}]$. In matrix notation we will denote the regression problem by

$$Y_{(n\times 1)} - X_{(n\times p)}\beta_{(p\times 1)} + \epsilon_{(n\times 1)}$$

where the matrix size is denoted by the subscript. Note that $X = [1, X_1, X_2, \ldots, X_k]$ and $\beta = [\beta_0, \beta_1, \beta_2, \ldots, \beta_k]$.

- When we want to express the regression in terms of a single observation, the new typically use the i subscript notation

  $$Y_i = \mathbf{X_i}\beta + \epsilon_i$$

  or simply

  $$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \epsilon_i$$

Estimation and Inference for Ordinary Least Squares Regression

Notes below are from the following sources; [Bhatti 2011c].

It's important to understand some aspects of estimation and inference for every statistical method that is used.

Estimation - Simple Linear Regression

- A *simple linear regression* is the special case of an OLS regression model with a single predictor variable.

$$Y = \beta_0 + \beta 1 X + \epsilon$$

- For the ith observation we will denote the regression model by

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- For the random sample $Y_1, Y_2, \ldots, Y_n$ we can estimate the parameters β_0 and β_1 by minimizing the sum of the squared errors,

$$\min \sum_{i=1}^{n} \epsilon_i^2$$

  which is equivalent to minimizing

$$\min \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Estimators and Estimates for Simple Linear Regression

- The estimators for β_0 and β_1 can be computed analytically and are given by

$$\hat{\beta_1} = \frac{\sum (Y_i - \bar{Y})(X_i - \bar{X})}{(X_i - \bar{X})^2} = \frac{\text{Cov}(Y, X)}{\text{Var}(X)}$$

  and

$$\hat{\beta_0} = \bar{Y} - \hat{\beta_1} \bar{X}$$

- The regression line always goes through the centroid $(\bar{X}, \bar{Y})$.
- We refer to the formulas for $\hat{\beta_0}$ and $\hat{\beta_1}$ as estimators and the values that these formulas can take for a given random sample as the estimates.
- In statistics we put hats on all estimators and estimates.
- Given $\hat{\beta_0}$ and $\hat{\beta_1}$ the predicted value or fitted value is given by

$$\hat{Y} = \hat{\beta_0} + \hat{\beta_1} X$$

Estimation - The General Case

- We seldom build regression models with a single predictor variable. Typically we have multiple predictor variables denoted by $X_1, X_2, \ldots, X_k$, and hence the standard regression case is sometimes referred to as *multiple regression* in introductory regression texts.
- We can still think about the estimation of $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ in the same manner as the sime linear regression case

$$\min \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} - \ldots - \beta_k X_{ki})^2$$

  but the computations will be performed as matrix computations.

General Estimation - Matrix Notation

Before we set up the matrix formulation for the OLS model, let's begin by defining some matrix notation.

- The error vector $\epsilon = [\epsilon_1, \ldots, \epsilon_n]^T$.
- The response vector $Y = [Y_1, \ldots, Y_n]^T$.
- The design matrix or predictor matrix $X = [1, X_1, X_2, \ldots, X_k]$.
- The parameter vector $\beta = [\beta_0, \beta_1, \beta_2, \ldots, \beta_k]^T$.
- All vectors are column vectors, and the superscript T denotes the vector or matrix *transpose*.

General Estimation - Matrix Computations

- We minimize the sum of the squared error by minimizing $S(\beta) = \epsilon^T \epsilon$ which can be re-expressed as
$$S(\beta) = (Y - X\beta)^T (Y - X\beta)$$
- Taking the matrix derivative of $S(\beta)$, we get
$$S_\beta(\hat{\beta}) = -2X^T Y + 2X^T X\hat{\beta}$$
- Setting the matrix derivative to zero, we can write the expression for the least squares *normal equations*
$$X^T X\hat{\beta} = X^T Y$$
, which yield the estimator
$$\hat{\beta} = (X^T X)^{-1} X^T Y$$
- The estimator form $\hat{\beta} = (X^T X)^{-1} X^T Y$ assumes that the inverse matrix $(X^T X)^{-1}$ exists and can be computed. In practice your statistical software will directly solve the normal equations using a QR Factorization.

normal equations: projection of a linear space into a subspace to ensure that a solution exists.

QR Factorization: or QR decompositionof a matrix is a decomposition of a matrix A into a product $A = QR$ of an orthogonal matrix Q and an upper triangular matrix R [Wikipedia 2015b].

Statistical Inference with the t-Test

- In OLS regression the statistical inference for the individual regression coefficients can be performed using a t-test.

t-test: any statistical test using a t-statistic to derive the test and the p-value for the test. Alternatively, any statistical test that uses a t-statistic as the decision variable.

statistical test: have a null and alternative hypothesis, and a test statistic with a known distribution.

- When performing a t-test there are three primary components: (1) stating the null and alternative hypotheses, (2) Computing the value of the test statistic, and (3) deriving a statistical conclusion based on a desired significance level.
- Step 1: The null and alternate hypotheses for β_i are given by
$$H_0 : \beta_i = 0 \text{ versus } H_1 : \beta_i \neq 0$$
- Step 2: The t statistic for β_i is computed by
$$t_i = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$
and has a degrees of freedom equal to the sample size minus the number of model parameters, i.e. $df = n - dim(model)$. For example if you had a regression model with two predictor variables and an intercept estimated on a sample of size 50, then the t statistic would have 47 degrees of freedom.
- Step 3: Reject the H_0 or Fail to Reject H_0 based on the value of your t statistic and your significance level. This decision can be made by using the p-value of your t statistic or by using the critical value for your significance level.

Confidence Intervals for Parameter Estimates

An alternative to performing a formal hypothesis test is to use a confidence interval for your parameter estimate. There is a duality between confidence intervals and formal hypothesis testing for regression parameters.

- The confidence interval for $\hat{\beta}_i$ is given by

$$\hat{\beta}_i \pm t(df, \frac{\alpha}{2}) \times SE(\hat{\beta}_i)$$

  where $t(df, \frac{\alpha}{2})$ is a t value from a theoretical t distribution, not a t statistic value.

- If the confidence interval does not contain zero, then this is the equivalent to rejecting the null hypothesis $H_0 : \beta_i = 0$.

Statistical Intervals for Predicted Values

The phrase *predicted value* is used in statistics to refer to the in-sample *fitted values* from the estimated model or to refer to the out-of-sample *forecasted values*. The dual use of this phrase can be confusing. A better habit is to use the phrase *in-sample fitted values* and in the *out-of-sample predicted values* to clearly reference these different values.

inference is an in-sample activity, measuring the quality of the model based on in-sample performance. *predictive modeling* is an out-of-sample activity, measuring the quality of the model based on out-of-sample.

- Given $\hat{\beta} = (X^T X)^{-1} X^T Y$ the vector of fitted values can be computed by $\hat{y} = X\hat{\beta} = HY$, where $H = X(X^T X)^{-1} X^T$. The matrix H is called the *hat matrix* since it puts the hat on Y.
- The point estimate $\hat{Y_0}$ at the point x_0 can be computed by $\hat{Y_0} = x_0^T \hat{\beta}$.
- The confidence interval for an in-sample point x_0 on the estimated regression function is given by

$$x_0^T \hat{\beta} \pm \hat{\sigma} \sqrt{x_0^T (X^T X)^{-1} x_0}$$

- The prediction interval for the point estimator $\hat{Y_0}$ for an out-of-sample x_0 is given by

$$x_0^T \hat{\beta} \pm \hat{\sigma} \sqrt{1 + x_0^T (X^T X)^{-1} x_0}$$

- Note that the out-of-sample prediction interval is always wider than the in-sample confidence interval.

Further Notation and Details

In order to compute the t statistic you need the standard error of the parameter estimate. Most statistical software packages should provide this estimate and compute this t statistic for you However, it is always a good idea to know from where this number comes. Here are the details needed to compute the standard error for $\hat{\beta}_i$.

- The estimated parameter vector $\hat{\beta}$ has the covariance matrix given by

$$\text{Cov}(\hat{\beta}) = \hat{\sigma}^2 X^T X$$

  where

$$\hat{\sigma}^2 = \frac{SSE}{n - k - 1}$$

- The variance of $\hat{\beta}_i$ is the ith diagonal element of the covariance matrix

$$\text{Var}(\hat{\beta}_i) = \hat{\sigma}^2 (X^T X)_{ii}$$

Analysis of Variance and Related Topics for Ordinary Least Squares Regression

Notes below are from the following sources; [Bhatti 2011d].

The ANOVA Table for OLS Regression

The Analysis of Variance or ANOVA Table is a fundamental output from a fitted OLS regression model. The output from the ANOVA table is sued for a number of purposes:

- Show the decomposition of the total variation
- Compute the R-Squared and Adjusted R-Squared metrics
- Perform the Overall F-test for a regression effect
- Perform a F-test for nested models as commonly used in forward, back-ward, and stepwise variable selection

Decomposing the Sample Variation

- The Total Sum of Squares is the total variation in the sample
- The Regression Sum of Squares is the variation in the sample that has been explained by the regression model
- The Error Sum of Squares is the variation in the sample that cannot be explained

SST	$\sum_i^n (Y_i - \bar{Y})^2$	Total Sum of Squares
SSR	$\sum_i^n (\hat{Y}_i - \bar{Y})^2$	Regression Sum of Squares
SSE	$\sum_i^n (Y_i - \hat{Y})^2$	Error Sum of Squares

Metrics for Goodness-Of-Fit in OLS Regression

The Coefficient of Determination - R-Squared

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- The Coefficient of Determintation R^2 will take values $0 \leq R^2 \leq 1$ and represents the proportion of the variance explained by the regression model.
- Implicitly, R^2 is a function of the number of parameters in the model. For a nested subset of predictor variables $p_0 < p_1$, i.e. p_1 contains the original p_0 predictor variables and some new predictor variables, R^2 will have a monotonic relationship such athat $R^2(p_0) \leq R^2(p_1)$.

Adjusted R-Squared

$$R^2_{ADJ} = 1 - \frac{\frac{SSE}{(n-k-1)}}{\frac{SST}{(n-1)}} = 1 - \frac{\frac{SSE}{(n-p)}}{\frac{SST}{n-1}}$$

- Note that the standard regression notation uses k for the number of predictor variables included in the regression model and p for the total number of parameters in the model. When the model includes an intercept term, then $p = k + 1$. When the model does not include an intercept term, then $p = k$.
- The Adjusted R-Squared metric accounts for the model complexity of the regression model allowing for models of different sizes to be compared.
- The Adjusted R-Squared metric will not be monotonic in the number of model parameters.
- The Adjusted R-Squared metric will increase until you reach an optimal model, then it will flatten out and likely decrease.

The Overall F-Test for a Regression Effect

Consider the regression model

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

The Overall F-Test for a regression effect is a joint hypothesis test that at least one of the predictor variables has a non-zero coefficient.

- The null and alternate hypotheses are given by

  $$H_0 : \beta_1 = \ldots = \beta_k = 0 \text{ versus } H_1 : \beta_i \neq 0$$

  for some $i \in 1, \ldots, k$.

- The test statistic for the Overall F-test is given by

  $$F_0 = \frac{\frac{SSR}{k}}{\frac{SSE}{(n-p)}}$$

  which has a F-distribution with $(k, n-p)$ degrees-of-freedom for a regression model with k predictor variables and p total parameters. When the regression model includes and intercept, then $p = k + 1$. If the regression model does not include an intercept, then $p = k$.

- In some cases this can be very useful, such as if we had a categorical variable that has segmentation, the F-test can be useful. It is less likely that continuous variables will all have a zero coefficient.

The F-Test for Nested Models

For our discussion of nested models, let's consider two concrete examples which we will refer to as the *full model* (FM)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

and a *reduced model* (RM)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Notice that the predictor variables in the reduced model are a subset of the predictor variables in the full model, i.e. $RM \subset FM$.

- In this notation we say that the FM *nests* the RM, or the RM is *nested by* the FM.
- We only use the terms *full model* and *reduced model* in the context of nested models.
- We can use a F-test for nested models to decide whether or not to include an additional predictor variable in the final model.

Given a *full model* and a *reduced model* we can perform a F-test for nested models for the exclusion of a single predictor variable or multiple predictor variables.

In the context of our example, we could test either of these null hypotheses:

- Example 1: Test a Single Predictor Variable

  $$H_0 : \beta_3 = 0 \text{ versus } H_1 : \beta_3 \neq 0$$

- Example 2: Test Multiple Predictor Variables

  $$H_0 : \beta_2 = \beta_3 = 0 \text{ versus } H_1 : \beta_i \neq 0$$

  for some $i \in 2, 3$.

The test statistic for the F-test for nested models will always have this form in terms of the FM and RM.

- Test Statistic for the Nested F-Test
$$F_0 = \frac{\frac{[SSE(RM) - SSE(FM)]}{(dim(FM) - dim(RM))}}{\frac{SSE(FM)}{[n - dim(FM)]}}$$
- The test statistic is based on the reduction in the SSE obtained from adding additional predictor variables. Note that $SSE(FM)$ is always less than $SSE(RM)$.
- The *dimension* of a statistical model is the number of parameters.

Connection to Forward Variable Selection

The F-test for nested models is a the standard statistical test implemented in most statistical software packages for performing forward and backward, and hence stepwise, variable selection.

Forward Variable Selection

- Given the model $Y = \beta_0 + \beta_1 X_1$ and a set of candidate predictor variables $Z_1, \ldots, Z_s$, how do we select the best Z_i to include in our model as X_2?
- In forward variable selection the FM will be $Y = \beta_0 + \beta_1 X_1 + \beta_2 Z_i$ and the RM will be $Y = \beta_0 + \beta_1 X_1$. The forward variable selection algorithm will select the Z_i with the largest F-statistic that is statistically significant at a predetermined level. The algotihm will continue to add predictor variables until there are no predictor variables that are statistically significant to the predetermined level.

Connection to Backward Variable Selection

Backward Variable Selection

- Given the model
$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_s X_s$$
how do we eliminate the predictor variables whose effects are not statistically significant?
- In backward variable selection the FM will be $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_s X_s$ and the RM will be $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{s-1} X_{s-1}$, for notional convenience. The backward variable selection algorithm will drop the X_i with the smallest F-statistic that is not statistically significant at a predetermined level. The algorithm will continue to drop predictor variables until there are no predictor variables that aren't statistically significant to the predetermined level.
- Note that both the forward and backward variable selection procedures consider only one variable at each iteration.

Statistical Inference Versus Predictive Modeling in OLS Regression

Notes below are from the following sources; [Bhatti 2011e]

- There are two reasons to build statistical models: (1) for inference, and (2) for prediction.
- Statistical inference is focused on a set of formal hypotheses, denoted by H_0 for the *null hypothesis* and H_1 for the *alternate hypothesis*, and a test statistic with a known sampling distribution. A test statistic will have a specified distribution, e.g. the t-statistic for an OLS regression parameter has a t-distribution with the degrees-of-freedom equal to $n - p$ where p is the number of model parameters for the dimension of the model.
- Predictive modeling is focused on accurately producing an estimated value for the primary quantity of interest or assigning an observation to the correct class (group). Typically, when we use the term 'predictive', we are referring to the model's ability to predict future or out-of-sample values, not in-sample values.

The Standard Modeling Process

0. Data Quality Check
1. Exploratory Data Analysis: How do our predictor variables relate to the response variable?
2. Model Identification: Which predictor variables should be included in our model?
3. Model Validation: Should we trust our models and the conclusions that we wish to derive from our model?

How we perform the Model Validation step is determined on the prescribed use of the model. Is the model to be used for statistical inference or is it to be used for predictive modeling?

Model Validation for Statistical Inference

- Model validation when the model is to be used for statistical inference is generally referred to as the *assessment of goodness-of-fit*.
- When we fit a statistical model, we have underlying assumptions about the probabilistic structures for that model. All of our statistical inference is derived from those probabilistic assumptions. Hence, if our estimated model, which is dependent upon the sample data, does not conform to these probabilistic assumptions, then our inference will be incorrect.
- When we validate a statistical model to be used for statistical inference, we are validating that the estimated model conforms to these probabilistic assumptions.
- For example in OLS regression we examine the residuals to make sure that they have a normal probability distribution and that they are homoscedastic.

Model Validation for Predictive Modeling

- Model validation when the model is to be used for predictive modeling is generally referred to as the *assessment of predictive accuracy*.
- When we fit a statistical model for predictive modeling, we can be much more tolerant of violations of the underlying probabilistic assumptions.
- Our primary interest in predictive modeling is estimating the response variable Y as 'accurately' as possible. When validating a predictive model, we tend to focus on summary statistics based on the quantity $(Y_i - \hat{Y}_i)$. Examples include the Mean Absolute Error (MAE) and the Mean Squared Error (MSE).
- The evaluation of predictive models is typically performed through a form of *cross-validation* where the sample is split into a *training sample* and a *test sample*. In this mdel validation, the model is estimated on the *training sample* and then evaluated out-of-sample on the *testing sample*.

Goodness-Of-Fit Versus Predictive Accuracy

- Goodness-Of-Fit
  - Goodness-Of-Fit (GOF) is assessed in-sample
  - The objective is to confirm the model assumptions
  - In OLS regression the GOF is typically assessed using graphical procedures (scatterplots) for the model residuals $e_i = Y_i - \hat{Y}_i$.
- Predictive Accuracy
  - Predictive Accuract (PA) is assessed out-of-sample
  - The objective is to measure the error of the predicted values
  - In OLS regression PA is typically assessed using error based metrics: Mean Square Error, Root Mean Square Error, and Mean Absolute Error.

Assessing the Goodness-Of-Fit in OLS Regression

- Validate the normality assumption: produce a Quantile-Quantile plot (QQ-Plot) of the residuals to compare their distribution to a nromal distribution.
- Validate the homoscedasticity assumption (equal variance): produce a scatterplot of the residuals against each predictor variable. If there is any structure in this plot, then the model will need a transformation of the predictor variable or an additional predictor variable added to the model.
- Interpret the R-Squared measure for your model. Applications tend to have typical ranges for "good" R-Squared values. If Model 1 has R-Squared of 0.23 and Model 2 has R-Squared of 0.54, then Model 2 should be preferred to Model 1, provided that Model 2 satisfies the other GOF conditions.
- By itself R-Square is not an exclusive measure of GOF. It's a measure of GOF provided everything else is satisfied.

Statistical Inference in OLS Regression

If our Analysis of Goodness-Of-Fit for our OLS regression does not uncover any major violations of the underlying probabilistic assumptions, then we can feel confident in our use of the two primary forms of statistical inference in OLS regression.

- The t-test for the individual model coefficients:

  $$H_0 : \beta_i = 0 \text{ versus } H_1 : \beta_i \neq 0$$

  for model coefficient i.
- The test statistic for the corresponding t-test is given by

  $$t_i = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

  where t_i has degrees of freedom equal to the sample size minus the number of model parameters, i.e. $df = n - dim(Model)$.

In addition to the 'local' tests of a regression effect for the individual predictor variables, we also have a 'global' test for a regression effect.

- The Overall F-test for a regression effect:

  $$H_0 : \beta_1 = \beta_2 = \ldots = 0 \text{ versus } H_1 : \beta_i \neq 0$$

  for some i, i.e. at least one of the predictor variables has an estimated coefficient that is statistically different from zero.

- The test statistic for the Overall F-test is given by:

$$F_0 = \frac{\frac{SSR}{k}}{\frac{SSE}{(n-p)}}$$

  which has a F-distribution with $(k, n-p)$ degrees-of-freedom for a regression model with k predictor variables and p total parameters. When the regression model includes an intercept, then $p = k+1$. If the regression model does not include an intercept, then $p = k$.

Predictive Accuracy in OLS Regression

The two primary metrics for assessing statistical models for out-of-sample predictive accuracy are Mean Square Error and Mean Absolute Error.

- Mean Square Error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

  Root Mean Square Error (RMSE) is the square root of the MSE. There is no statistical reason to prefer one measure over the other. However, the RMSE can be used for presentation purposes when the MSE is very small or very large as the square root transformation will increase the small numbers and decrease the large numbers.

- Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$$

The Bias-Variance Trade-Off

An interesting and useful property of Mean Square Error (MSE) is that it can be decomposed into two components: the prediction variance and the square of the prediction bias. This decomposition is referred to as the *Bias-Variance Trade-Off*, and it is referenced throughout predictive modeling, especially in the presentation of concepts from statistical and machine learning.

- Throughout these notes we have been using the *empirical* Mean Square Error for the predictved values $\hat{Y}_i$.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

- The *Bias-Variance Trade-Off* is presented from the *theoretical Mean Square Error*

$$MSE = \mathbb{E}(Y_i - \hat{Y}_i)^2$$

  where $\mathbb{E}[X]$ denotes the mathematical expectation of X.

Final Comments on the Bias-Variance Trade-Off

The crux of the *Bias-Variance* Decomposition is to note that both terms of the decomposition are non-negative. Hence, we can choose to minimize either the Variance or the bias.

- The variance of the predicted value is a measure of the spread of the predicted value from its mean.
- The bias of the predicted value is a measure of the distance from the mean of the predicted value to the target value.

Both of these components are functions of *model complexity*, i.e. the nubmer of parameters in the model. Ideally, you would want to have your prediction to be accurate (low bias) and precise (low variance). Bias will decline and variance will increase as the model complexity increases.

Further Notation and Details

The Mean Square Error of the predicted values $\hat{Y}_i MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2\$$ should not be confused with the estimate or variance parameter σ^2 in an OLS regression model with the Error Sum of Squares denoted by SEE and p parameters,

$$\sigma^2 = \frac{SSE}{n-p}$$

which is frequently referred to as the *mean square error* of the regression or the *mean square* of the residuals, but is not denoted by MSR as to not be confused with the *mean square of the regression* ($MSR = \frac{SSR}{k}$).

If you are in the context of a fitted OLS regression model, then the term MSE is referring to the estimate $\hat{\sigma^2}$.

Introduction to Principle Components Analysis

Notes below are from the following sources; [Bhatti 2011f]

- *Statistical Interpretation* - PCA is a transformation of a set of correlated random variables to a set of uncorrelated (or orthogonal) random variables
- *Linear Algebra Interpretation* - PCA is a rotation of the coordinate system to the canonical coordinate system, i.e. the natural coordinate system defined by the variation in the data.
- *Numerical Linear Algebra Interpretation* - PCA is a reduced rank matrix approximation that facilitates dimension reduction.

Facts and Caveats

- PCA does not require any statistical assumptions, e.g. the data are not assumed to have a multivariate normal distribution.
- PCA is a (numerical) linear algebra technique, i.e. it relies on a matrix factorization (the Spectral Decomposition or Singular Value Decomposition).
- PCA is sensitive to the scale of the data. Most of the time the data should be standardized, i.e. the variables should have a (0,1) distribution. When the data re standardized our covariance matrix and correlation matrix are the same matrix.
- If the data are 'standardized' to a common scale that is not (0,1), then it should not be standardized to a (0,1) distribution.

Why do we use PCA?

- PCA can be used in its own right to understand the covariance structure in multivariate data with respect to the measured basis.
- PCA can be used as a method to create a reduced rank approximation to the covariance structure, i.e. PCA can be used to approximate the variation in p predictor variables using $k < p$ principal components. This property is typically referred to as dimension reduction.
- PCA can be used as a means of creating a set of orthogonal predictor variables from a set of raw predictor variables. Since the principal components created from the original predictor variables are orthogonal, we can use PCA as a remedy for multicollinearity in regression problems or as a preconditioner to cluster analysis.

How do we compute the principal components?

- Consider the $n \times p$ data matrix of predictor variables $X = [X_1, \ldots, X_p]$.
- Depending on your software the data may need to be standardized before the principal components are computed. This is typically true if you use a software to compute eigenvalues and eigenvectors. Statistical software designed to perform PCA, such as princomp in SAS, will typically internally standardize the data for you.
- Compute the eigenvalue-eigenvector pairs $(\lambda_1, e_1), \ldots, (\lambda_p, e_p)$ of the square matrix $X^T X$ where the eigenvalues are ordered largest to smallest such that $\lambda_i > \lambda_j$ for $i > j$.
- Your software will compute the eigenvalue-eigenvector pairs using a matrix factorization called Singular Value Decomposition or SVD.
- Compute the principal components $Z_1, \ldots, Z_p$ using the eigenvalues as the component loadings
- In vector format we can compute each component individuall

$$Z_i = X \times e_i$$

or we can compute all of the principal components using one matrix computation

$$[Z_1, \ldots, Z_p] = X \times [e_1, \ldots, e_p]$$

.

Study Questions for Ordinary Least Squares Regression

Question: 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'?

Response from [Montgomery et al. 2012] pages 2 and 4.

$$y = \beta_0 + \beta_1 x + \varepsilon$$

The equation abovce is called a **linear regression model**. Customarily x is called the independent variable and y is called the dependent variable. However, this often causes confustion with the concept of statistical independence, so we refer to x as the **predictor** or **regressor**and y as the **response** variable. Because the equation above involves only one **regressor** variable, it is called a **simple linear regression model**.

In general the response variable y may be related to k regressors, $x_1, x_2, \ldots, x_k$, so that:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_x + \varepsilon$$

This is called a **multiple linear regression model** because more than one regressor is involved. The adjetive linear is employed to indicate that the model is linear in the parameters $\beta_0, \beta_1, \ldots, \beta_k$, not because the y is a linear function of the x's.

Question: 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'?

Response from [Montgomery et al. 2012] pages 1.

Regression analysis is a _statistical technique **for investigating and** modeling the relationship between variables___. The goal of regression analysis is to determine the values of parameters for a function that causes the function to best fit a set of data. Regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed [Wikipedia 2015c].

Question: What do we mean by 'linear regression'? represent a linear regression? Which equations represent a linear regression?

(a) $y = \beta_0 + \beta_1 x_1$
(b) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2$
(c) $y = \beta_0 + \exp(\beta_1) x_1$

Both (a) and (c) are representations of linear regression because they only have a single **regressor**. (b), with the presence of x_2^2 has two **regressors**.

Question: 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?

Question: In the simple linear regression model what is the relationship between R-squared and the correlation coefficient rho?

Question: How do we interpret a regression coefficient in OLS regression?

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

Question: 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? 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?

Question: How are the parameters, i.e. the model coefficients, estimated in OLS regression? How does this relate to maximum likelihood estimation? How do you show the relationship between OLS regression and maximum likelihood estimation?

Question: What is the overall F-test? What is the null hypothesis and what is the alternate hypothesis? The overall F-test is also called the 'test for a regression effect'. Why is it called this?

Question: What is the difference between R-squared and adjusted R-squared? How is each measure computed, and which measure should we prefer? How does the interpretation of R-squared change as we move from the simple linear regression model to the multiple regression model?

Question: The simple linear regression model $Y = b_0 + b_1 * X_1$ has three parameters. Two of the parameters are b_0 and b_1. What is the third parameter?

Question: 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?

Question: 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?

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

Question: What is the Analysis of Variance Table for a regression model? How do we interpret it and what statistical tests and quantities can be computed from it?

Question: When the intercept is excluded in a regression model, how does the computation and the interpretation of R-squared change? Fit a no intercept model in SAS and check the SAS output for any noted differences.

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

Question: Why do we plot each predictor variable against the residual as a model diagnostic?

Question: Why do we perform transformations in the construction of regression models? Name at least two reasons.

Question: What is multicollinearity and how does it affect the parameter estimates in OLS regression? How do we diagnose multicollinearity?

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

Question: Given a set of predictor variables $X_1, \ldots, X_n$, 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?

Question: Variable Selection: How does forward variable selection work? How does backward variable selection work? How does stepwise variable selection work?

Study Questions for Multivariate Analysis

Principle Components Analysis

Question: Principal Components Analysis (PCA): What is principal components analysis? How does PCA eliminate the problem of multicollinearity? What does it mean for X1 and X2 to be orthogonal? In order to better understand orthogonality, take the building prices data set and perform these steps:

(a) Perform a PROC CORR on X1-X9.
(b) Create nine orthogonal predictor variables using PCA. Call these variables Z1-Z9.
(c) Perform a PROC CORR on Z1-Z9.

Question: Principal Components Analysis is described as a method of 'dimension reduction'. How does PCA reduce the dimension of a statistical problem? How do you select the reduced dimension for your problem.

Factor Analysis

Question: Are the factor scores always orthogonal? Are they orthogonal after a rotation?

Question: If two analysts perform a factor analysis, are they likely to arrive at the same result? If the same two analysts perform a principal components analysis, are they likely to get the same result?

Question: What is the first step in performing a factor analysis?

Question: In the context of factor analysis, what is the communality of factors?

Cluster Analysis

Question: What is the difference between hierarchical and non-hierarchical clustering?

Question: What is linkage? What types of linkage are there?

Question: How do we examine the goodness-of-fit of a cluster analysis or two comparative cluster analyses?

Question: Do the data need to be treated before we perform a cluster analysis?

References

Bhatti, D.C. 2011a. Statistical preliminaries and mathematical notation. http://nwuniversity.adobeconnect.com/p2u4z1zop3a/.

Bhatti, D.C. 2011b. Statistical assumptions for ordinary least squares regression. http://nwuniversity.adobeconnect.com/p14gl2rughs/.

Bhatti, D.C. 2011c. Estimation and inference for ordinary least squares regression. http://nwuniversity.adobeconnect.com/p9rbxcf6431/.

Bhatti, D.C. 2011d. Analysis of variance and related topics for ordinary least squares regression. http://nwuniversity.adobeconnect.com/p5hhosjmkbh/.

Bhatti, D.C. 2011e. Statistical inference versus predictive modeling in oLS regression. http://nwuniversity.adobeconnect.com/p5vo4o95h0r/.

Bhatti, D.C. 2011f. Introduction to principal components analysis. https://nwuniversity.adobeconnect.com/_a799312996/p32n9izw7y3.

Montgomery, D.C., Peck, E.A., and Vining, G.G. 2012. *Introduction to linear regression analysis*. John Wiley & Sons.

Wikipedia. 2015a. Homoscedasticity — wikipedia, the free encyclopedia. /url\protect\T1\textbraceleftthttp://en.wikipedia.org/w/index.php?title=Homoscedasticity&oldid=650997386\protect\T1\textbraceright.

Wikipedia. 2015b. QR decomposition — wikipedia, the free encyclopedia. /url\protect\T1\textbraceleftthttp://en.wikipedia.org/w/index.php?title=QR_decomposition&oldid=655839188\protect\T1\textbraceright.

Wikipedia. 2015c. Regression analysis — wikipedia, the free encyclopedia. /url\protect\T1\textbraceleftthttp://en.wikipedia.org/w/index.php?title=Regression_analysis&oldid=647603059\protect\T1\textbraceright.