Module 4.
Linear and Logistic Regression

Methodology Training
EXL Decision Analytics

Year 2013

EXL Decision Analytics Methodology Snapshot

We apply a set of highly effective tools, techniques and best practices for the end-to-end model development cycle

Stage 1	Preliminary Data Exploration	Univariate Analysis (EDD*) Modeling and Validation Split Bivariate Analysis
Stage 2	Data Preparation	Outlier Treatment Missing Imputation Roll Ups and Data Merge
Stage 3	Variable Creation	Dummy Variable Creation Binning and Banding Transformations Interactions and Groupings
Stage 4	Variable Reduction	Variable Clustering Inter-Correlation Analysis Variance Inflation Factor Test
Stage 5	Modeling	Modeling Technique Selection Model Improvements Ensemble
Stage 6	Validation and Stabilization	In-Sample Validation Out-of-Time Validation Bootstrapping Coefficient Blasting

These stages demand lot of manual effort in analyzing and understanding each and every variable

These stages require business sense and out-of-box thinking for brainstorming on creating hypothesis-based variables and dropping redundant features

These stages require good knowledge of statistical techniques for providing high-end quality solutions

* Extended Data Dictionary

© ExlService Holdings, Inc. 2006-2013

Objectives and Scope

Course Goals

- To provide a structured overview of linear and logistic regression modeling concepts used during application of EXL DA methodology
- To introduce trainees to SAS syntax for implementation of traditional model development techniques
- To explain interpretation of key SAS output
- Hands on exercises on real life data to practice respective modeling steps during the training course
- To provide helpful “tricks of the trade”

Beyond the Scope of this Training

- Comprehensive coaching on model building
- Derivation of statistical formulas or terms (unless required as part of methodology explanation)
- Extensions / Advanced Modeling Techniques (GLM, Multinomial Logistic Regression, Machine Learning Techniques)

Self Study Goals

- Linear and Logistic Regression model development practice on hypothetical data
- In-depth research on advanced modeling concepts
- Discussion on advanced concepts can be taken up offline

© ExlService Holdings, Inc. 2006-2013 Confidential

Motivation Behind Predictive Modeling

Two Crucial Uses of Predictive Modeling

1. **Prediction**
   - It is important when the objective is to estimate a score for each record
   - For example: Scorecard Development

2. **Explanation**
   - It is important when the objective is to identify and interpret the contributing predictors
   - For example: Key Driver Analysis

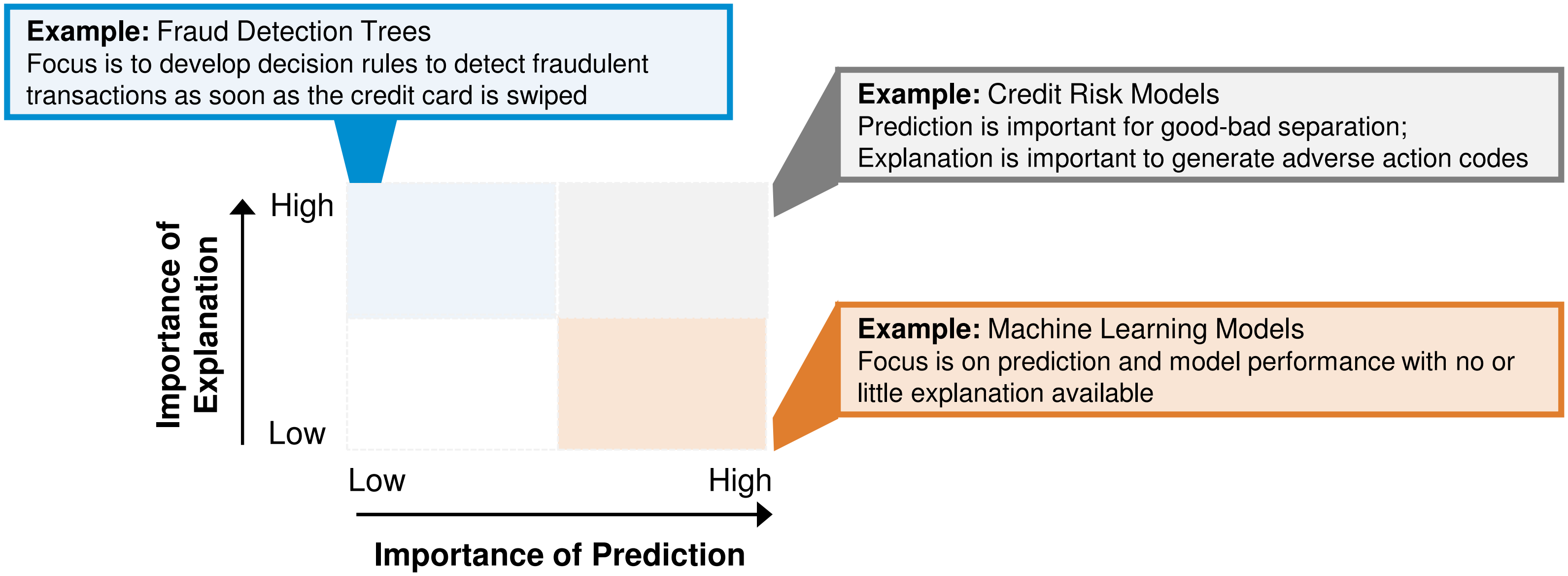


Table of Contents

© ExlService Holdings, Inc. 2006-2013

© ExlService Holdings, Inc. 2006-2013

Chapter 1: Linear Regression

1.1 What is Regression?

1.1.1. Regression Analysis

Meaning

- Study of statistical dependence of a target variable (also known as dependent variable) on one or more predictors (also known as independent variables)

Objective

- To estimate and/or predict the mean value of the dependent variable on the basis of the known values of the independent variables

Synonyms of Dependent and Independent Variables

Dependent Variable is also known as 'Target Variable' or 'Response Variable' or 'Regressand' or 'Outcome'

Independent Variable is also known as 'Predictor' or 'Explanatory Variable' or 'Regressor' or 'Covariate'

Regression Equation

$$\underbrace{Y}_{\text{Dependent Variable}} = f\underbrace{(X_1, X_2 \ldots X_k)}_{k \text{ Independent Variables}} + \underbrace{\varepsilon}_{\text{Stochastic Error Term}}$$

Examples

- Average hourly wage depends on education and occupational domain (industry)
- Price of car depends on car weight, fuel efficiency and manufacturing place among other things

© ExlService Holdings, Inc. 2006-2013 Confidential

1.1.2. Linear Regression

Usage

- Linear Regression technique may be used to study the relation between a dependent and one or more independent variables, when the dependent variable is *continuous*

Simple Linear Regression vs. Multiple Linear Regression

	Simple Linear Regression	Multiple Linear Regression
Definition	Linear regression in which the dependent variable is related to a single explanatory variable	Linear regression in which the dependent variable is related to two or more explanatory variables
Equation	$Y = \beta_0 + \beta_1 X_1 + \varepsilon$	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 ... \beta_k X_k + \varepsilon$
Example	Personal consumption expenditure (Y) depends on disposable income (X_1)	Crop yield (Y) depends on temperature (X_1), rainfall (X_2), sunshine (X_3) and fertilizer (X_4)

1.2 Meaning of Linearity

The term '**linear**' can be interpreted in two ways:

- Linearity in the Variables
- Linearity in the Parameters

Examples of Linear Regression Model

Scenario 1: Y is linear in variables (X_1 and X_2) as well as linear in parameters (β_1 and β_2)

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

Scenario 2: Y is non-linear in variables (X_1 and X_2), but linear in parameters (β_1 and β_2)

$$Y = \beta_0 + \beta_1 X_1^2 + \beta_2 X_2^3 + \varepsilon$$

Examples of Non-Linear Regression Model

Scenario 3: Y is linear in variables (X_1 and X_2), but non-linear in parameters (β_1 and β_2)

$$Y = \beta_0 + \beta_1^2 X_1 + \beta_2^3 X_2 + \varepsilon$$

Scenario 4: Y is non-linear in variables (X_1 and X_2) as well as non-linear in parameters (β_1 and β_2)

$$Y = \beta_0 + \beta_1^2 X_1^2 + \beta_2^3 X_2^3 + \varepsilon$$

The term '**linear**' regression means a regression that is linear in the parameters (β's). It may or may not be linear in the explanatory variables (X's). **Only scenario 1 and 2 correspond to Linear Regression.**

© ExlService Holdings, Inc. 2006-2013

Exercise

Exercise 1. Which of the following are cases of Linear Regression Model? Further categorize them into Simple Linear Regression Models and Multiple Linear Regression Models.

a. *Default Amount* $= \alpha + \beta^2$ *(FICO Score)* $+ \gamma^3$ *(Income)* $+ \varepsilon$

b. *CAT Score* $= \alpha + \beta$ *(# Attempts)* $+ \gamma$ *(Educational Background)* $+ \varepsilon$

c. *Consumption* $= \alpha + \beta$ *(Disposable Income)* $+ \varepsilon$

d. *Demand Price* $= \alpha + \beta$ *(Quantity Demanded)* $+ \varepsilon$

1.3 Stochastic Disturbance vs. Residual

1.3.1. Population Regression Function (PRF)

A linear PRF states that the expected value of the distribution of Y given X_i is functionally related to X_i such that it is linear in parameters

$$E(Y \mid X_i) = \beta_0 + \beta_1 X_i$$

Where β_0 and β_1 are unknown but fixed parameters known as the regression coefficients

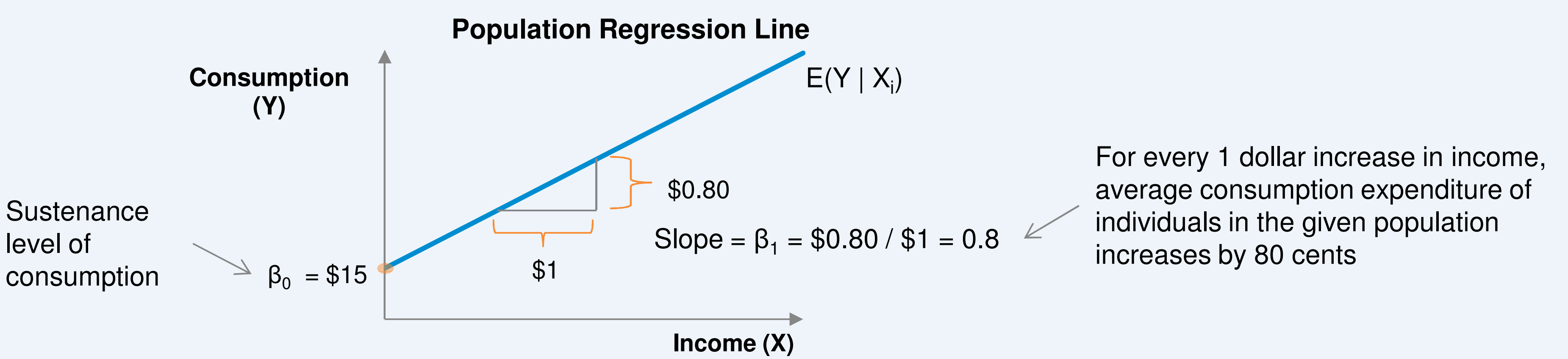
Example: Consumption = \$15 + 0.8 (Income)

Things to Remember

β_0 is known as intercept

β_1 is known as slope

Box 1 *Optional for Interested Readers*



Stochastic Specification of PRF

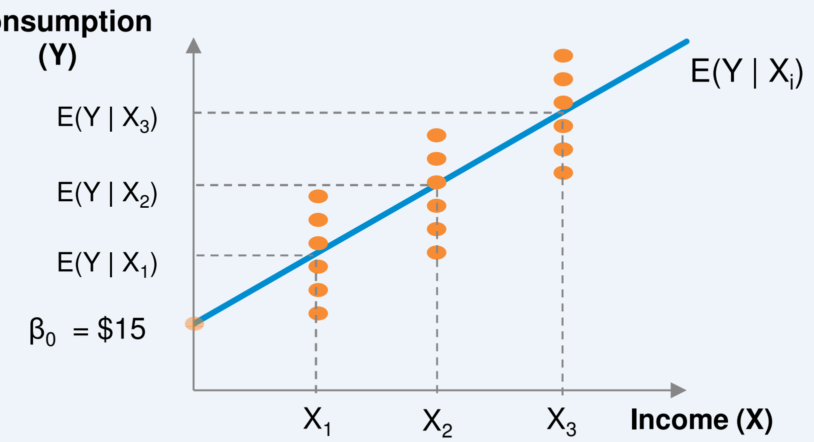
An individual's consumption expenditure (for a given income level) is sum of two components:

- $E(Y \mid X_i)$: Average consumption expenditure, which is Systematic or Deterministic Component
- ε_i : Non-systematic Component (known as Stochastic Disturbance or Random Error)

$$
\begin{aligned}
Y_i &= E(Y \mid X_i) + \varepsilon_i \\
&= \beta_0 + \beta_1 X_i + \varepsilon_i
\end{aligned}
$$

Box 2 *Optional for Interested Readers*

Continuing with illustration from Box 1, for a given level of income X_i, an **individual's consumption expenditure** is clustered around **average consumption** of all individuals at that X_i (i.e. around its conditional expectation)



1.3.2. Significance of Stochastic Disturbance Term

Stochastic disturbance term ε_i is a proxy for all those variables that are omitted from the model, but that collectively affect the dependent variable Y

Box 3 *Optional for Interested Readers*

Why do we need a stochastic disturbance term? Why don't we use all variables affecting Y?

- Such variables may be unknown due to vagueness of theory (lack of knowledge about the exact hypothesis)
- Even if they are known, quantitative data may not be available
- At least some part of variation in Y may be purely due to intrinsic randomness in human behavior. Even quantitative data may not be sufficient to explain these variations
- To keep model equation reasonably simple, it makes sense to retain only significant and stable predictors and to let the random disturbance term represent all other variables
- Even if all relevant variables affecting Y are readily available and are retained in the model, the correct form of functional relationship between target Y and predictors may be unknown

1.3.3. Sample Regression Function (SRF)

- PRF is an idealized concept. In practice, one rarely has access to the entire population
- In general, only a sample of observations from the population is used
- Sample Regression Function (SRF) is used to estimate the PRF

SRF is expressed as: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

Where

$\hat{Y}_i =$ estimator of $E(Y \mid X_i)$

$\hat{\beta}_0 =$ estimator of β_0

$\hat{\beta}_1 =$ estimator of β_1

Stochastic Specification of SRF

SRF in stochastic form is expressed as: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\varepsilon}_i$

Where

$\hat{\varepsilon}_i$ is the Residual term and can be regarded as an estimate of stochastic disturbance ε_i

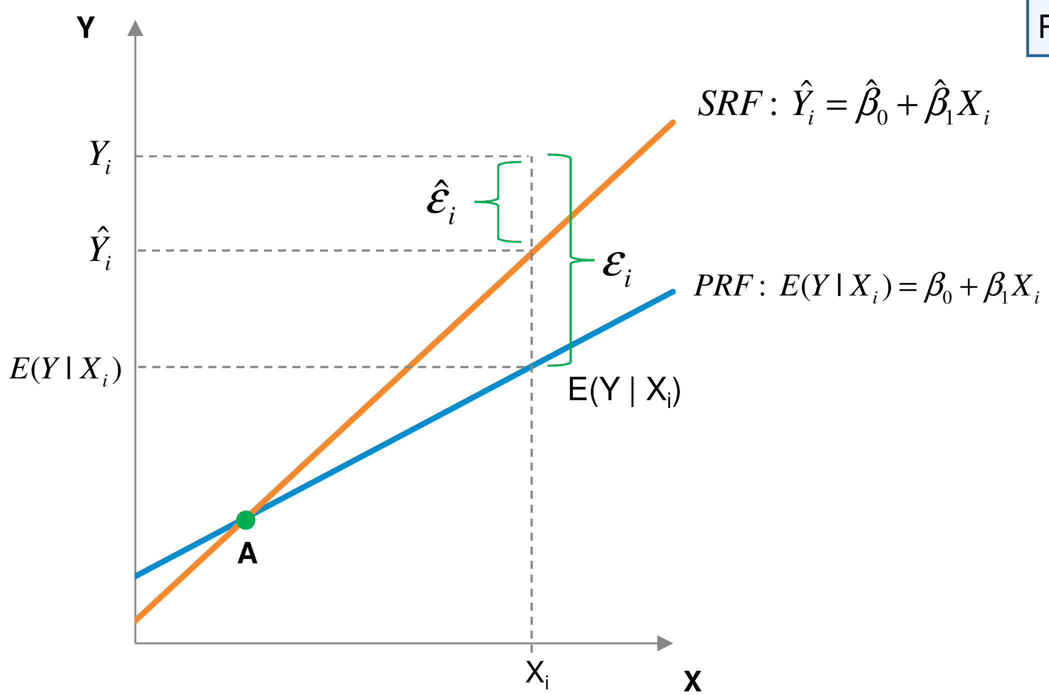
© ExlService Holdings, Inc. 2006-2013 Confidential

1.3.4. Graphical Representation: Stochastic Disturbance vs. Residual

Things to Remember

Stochastic Disturbance: $\varepsilon_i = Y_i - E(Y \mid X_i)$

Residual: $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$

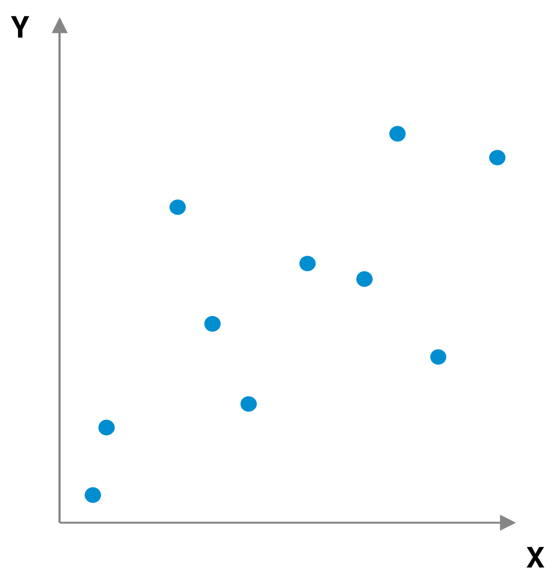


- For any X_i to the right of point A, SRF *overestimates* the true PRF
- For any X_i to the left of point A, SRF *underestimates* the true PRF
- Such over-estimation and under-estimation is inevitable due to sampling fluctuations

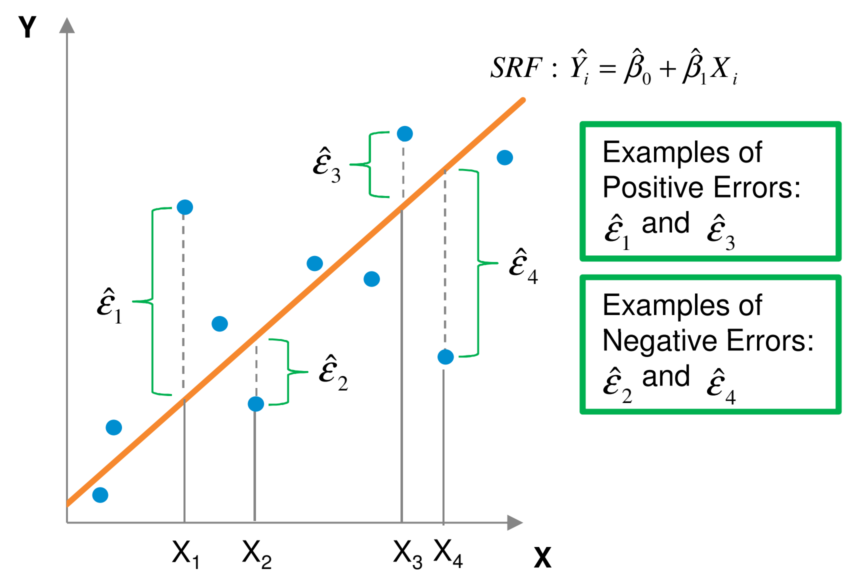
© ExlService Holdings, Inc. 2006-2013

1.4 Estimation Method: OLS

1.4.1. Ordinary Least Square (OLS) Criterion



Actual values of Y for given values of X



Least squares method chooses SRF by estimating β_i's such that $\sum \hat{\varepsilon}_i^2$ is minimum

Things to Remember

- Sum of errors is not minimized. Positive and negative errors offset each other. It is the **absolute value of errors** that matters.
- Sum of absolute errors is not minimized. The **magnitude of errors** matters. By squaring errors, the error itself is used as a weight. In other words, more weight is given to bigger error terms. Hence, the sum of square of errors is minimized.

© ExlService Holdings, Inc. 2006-2013 Confidential

1.4.2. BLUE: Characteristics of OLS Estimator

An OLS estimator $\hat{\beta}_i$ is said to be **B**est **L**inear **U**nbiased **E**stimator (BLUE) of β_i

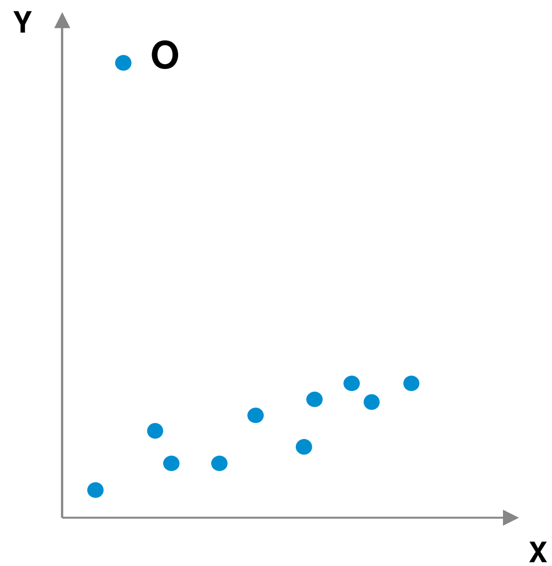
Linear	Unbiased	Best
The estimator is a linear function of dependent variable Y in the regression model	Average or expected value of the estimator is equal to the true value $E(\hat{\beta}_i) = \beta_i$	The estimator has minimum variance in the class of all such linear unbiased estimators

1.4.3. A Note on Significance Tests

- Hypothesis tests are performed during model build to test significance. For example:
  - F-test for overall significance of linear regression model
  - t-test for individual variable significance in linear regression model
- Standard error of an estimate is an important component of test statistic
- **Caution:** Due to violation of any OLS assumption affecting standard errors of estimates, the significance tests may become invalid

Exercise

Exercise 2. What does point O (in the graph below) signify? Should a modeler go ahead with linear regression model fit without any intermediate action?



[Hint: Recall

1. Steps taken at data preparation stage
2. Objective of OLS method is to minimize sum of squares of errors]

© ExlService Holdings, Inc. 2006-2013 Confidential

1.5 Linear Regression: Key Assumptions

1.5.1. Assumption 1: Predictor X is non-stochastic

Interpretation

- Values taken by the regressor X are considered fixed in repeated samples. That is, X is non-random

Violation Implication

- No serious implication as long as predictor X and disturbance ε are uncorrelated, which is yet another assumption of Classical Linear Regression Model (Refer to Assumption 6 in Section 1.5.6)

1.5.2. Assumption 2: Variability in X values

Interpretation

- X values in a given sample must not all be the same
- Example: Suppose the modeling data corresponds to a particular year (say, 2012). The 'year' variable would take single unique value '2012' for all records. Such a variable won't add any value in making any prediction.

Violation Implication

- No estimation possible for β coefficient

Things to Remember
From the list of predictors, drop all variables that take single unique value

1.5.3. Assumption 3: Zero mean value of disturbance ε

Interpretation

- Assumption that $E(\varepsilon_i \mid X_i) = 0$ implies that the positive ε_i values cancel out the negative ε_i values so that their average or mean effect on Y is zero
- $E(\varepsilon_i \mid X_i) = 0$ also implies that $E(Y_i \mid X_i) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$ (given that $Y_i = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon_i$)

Violation Implication

- No impact on the properties of slope coefficients ($\beta_1, \beta_2, \ldots, \beta_k$)
- If $E(\varepsilon_i \mid X_i)$ is a non-zero constant, we get a biased estimate of intercept β_0

Box 4 *Optional for Interested Readers*

Why do we get a biased estimate of intercept if mean value of disturbance is a non-zero constant?

Consider k - variable linear regression model :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

Assume $E(\varepsilon \mid X_1, X_2, \ldots, X_k) = \lambda$, where λ is a constant

$$\begin{aligned} E(Y \mid X_1, X_2, \ldots, X_k) &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \lambda \\ &= (\beta_0 + \lambda) + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \\ &= \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \end{aligned}$$

Apparently, $\alpha = (\beta_0 + \lambda)$ is a biased estimate of β_0

1.5.4. Assumption 4: Homoscedasticity

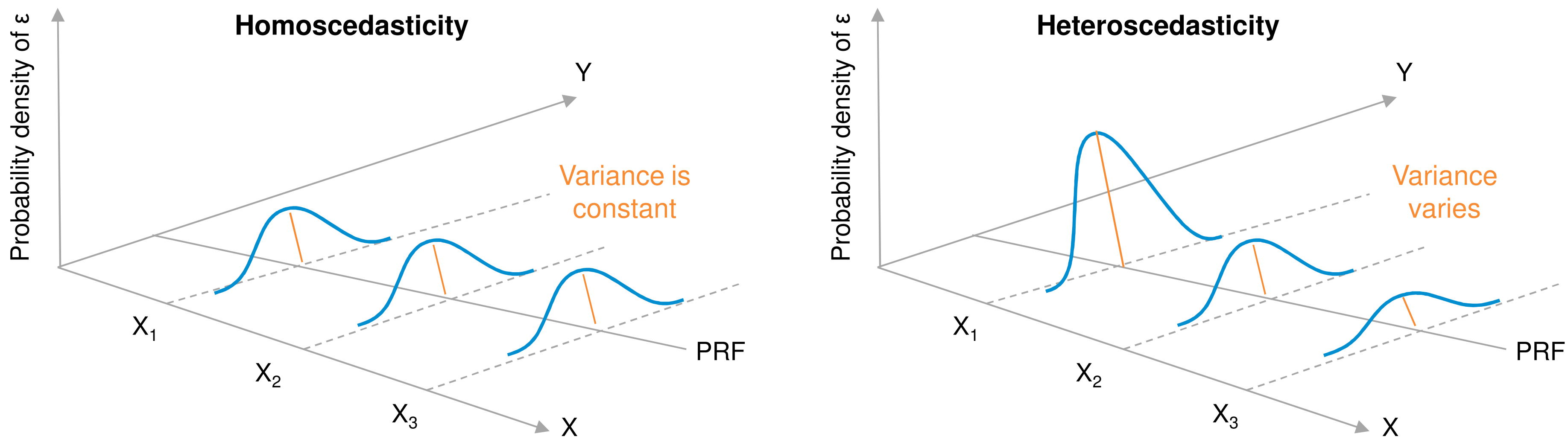
Interpretation

- Given the value of X, the variance of disturbances ε_i is the same for all observations

$$\text{variance}(\varepsilon_i \mid X_i) = \sigma^2$$

Violation Implication

- Absence of homoscedasticity implies presence of heteroscedasticity. OLS estimates remain unbiased.
- But OLS estimates no longer remain efficient (i.e. there are alternative methods of estimation such as WLS with smaller standard errors) and hence significance tests may not be valid



1.5.5. Assumption 5: No autocorrelation between disturbances

Interpretation

- Given any two X values, X_i and X_j ($i \neq j$), the correlation between ε_i and ε_j ($i \neq j$) is zero
- This assumption is more likely to get violated in case of time-series data. Usually, generalized least square (GLS) models are used to tackle this problem

Violation Implication

- OLS estimates remain unbiased
- But OLS estimates no longer remain efficient and hence significance tests may not be valid

1.5.6. Assumption 6: Zero covariance between ε and X

Interpretation

- X and ε are assumed to be uncorrelated, as the definition of PRF requires that X and ε have separate (and additive) influence on Y

Violation Implication

- OLS estimates not only become biased, but also inconsistent (i.e. as the sample size increases indefinitely, the estimators do no converge to their true population values)

1.5.7. Assumption 7: $n > k + 1$

Interpretation

- Number of observations (n) must be greater than the number of parameters to be estimated (k + 1)
  where k = Number of Independent Variables (X_1, X_2, …, X_k)
  Parameters to be estimated include k slope coefficients (β_1, β_2, …, β_k) plus 1 intercept coefficient (β_0)

Violation Implication

- Regression coefficients can't be estimated

1.5.8. Assumption 8: No perfect multicollinearity

Interpretation

- There are no perfect linear relationships among the explanatory variables

Violation Implication

- Perfect Multicollinearity Case
  - Coefficients are indeterminate and standard errors are not defined
- High Multicollinearity Case
  - Estimation of regression coefficients is possible, but standard errors tend to be large
  - Individual variable contribution tends to be less precise as predictors are highly correlated
  - Multicollinearity leads to model over-fitting. The overall measure of goodness of fit can be very high, but the t-ratio of one or more variables may be statistically insignificant.

Things to Remember

Inter-correlation analysis and VIF test are popular methods of detecting multicollinearity

1.5.9. Assumption 9: Normality of ε

Interpretation

- ε_i follow the normal distribution

Violation Implication

- Estimates remain BLUE
- But they are no longer asymptotically efficient (i.e. as sample size grows, estimates are not optimal)

Note: Assumptions 3, 4, 5 and 9 together imply that $\varepsilon \sim NID(0, \sigma^2)$, which means ε is normally and independently distributed with mean 0 and constant variance σ^2

Box 5 *Optional for Interested Readers*

Why the Normality Assumption?

Central Limit Theorem (CLT) provides the theoretical justification for the normality assumption

- Recall from Section 1.3.2. that ε represents the combined influence of a large number of independent variables that are not an explicit part of regression model
- Influence of such omitted or neglected variables is expected to be small and random
- By **Central Limit Theorem (CLT)**, if there are large number of independent and identically distributed random variables, then the distribution of their sum tends to a normal distribution as the number of such variables increase indefinitely

1.6 SAS Implementation

1.6.1. REG Procedure: SAS Syntax

Below is the syntax for PROC REG with frequently used options[1]

Syntax	Description
PROC REG DATA = <*modeling dataset*> ;	Specify name of modeling dataset for regression
MODEL <*dependent*> = <*regressors*>	
/ SELECTION = <*selection method*>	Specify variable selection method
SLE = <*SLE criterion*> SLS = <*SLS criterion*>	Specify significance level of entry and stay
COLLIN	This option produces collinearity analysis
VIF	This option computes variance-inflation factors
R	This option produces analysis of residuals
STB	This option displays standardized parameter estimates
TOL ;	This option displays tolerance values for parameter estimates
OUTPUT OUT = <*output dataset*> P = <name of predicted value variable> ;	
ODS OUTPUT PARAMETERESTIMATES = <*parameter estimates output dataset*> ;	
QUIT ;	

[1] For exhaustive list of options, refer to *SAS OnlineDoc™: Chapter 55: The REG Procedure (http://www.math.wpi.edu/saspdf/stat/chap55.pdf)*

Selection Methods

- **NONE**
  The complete model specified in the MODEL statement is used to fit the model

- **FORWARD** (i.e. Forward Selection)
  This technique begins with no variable in the model and then the variables are added one by one to the model based on their F statistics
  - For each independent variable, F statistics are computed (reflecting variable contribution to the model)
  - Variable with the largest F statistic is added to the model if its p-value < SLE= value
  - Process is repeated until there is no independent variable whose F statistic is more significant than SLE= value
  - Once a variable is in the model, it stays

- **BACKWARD** (i.e. Backward Elimination)
  This technique begins with all variables in the model and then the variables are deleted one by one from the model based on their F statistics
  - For each independent variable, F statistics are computed (reflecting variable contribution to the model)
  - Variable with the smallest F statistic is deleted from the model if its p-value > SLS= value
  - Process is repeated until all the variables in the model produce F statistic significant at SLS= value
  - Once a variable is removed from the model, it is never re-considered for inclusion

© ExlService Holdings, Inc. 2006-2013

STEPWISE

This technique is similar to the FORWARD selection technique except that the variables already in the model do not necessarily stay there

- Variables are added one by one to the model and the F statistic for a variable to be added must be significant at the SLE= value
- Once a variable is added, stepwise method looks at all the variables in the model and deletes any variable that does not produce an F statistic significant at SLS= value
- Variables are thus entered into and removed from the model in such a way that each forward selection step may be followed by one or more backward elimination steps
- Stepwise process terminates
  - If no further variable can be added to the model for the specified SLE criterion and no further variable can be deleted from the model for the specified SLS criterion;
  - Or if the variable to be added to the model is the one just deleted from it

SLE and SLS Values

- **SLE:** SLE refers to a variable's significance level of entry into the model
- **SLS:** SLS refers to a variable's significance level of stay within the model

Low SLE, SLS Values <=> Highly Significant variables are selected
<=> Fewer variables are selected
<=> Stricter Approach for Variable Selection

Commonly Used Values: 0.01, 0.05 and 0.10

As a rule of thumb, SLE= 0.05 and SLS= 0.05 are used in general

Significance Level	Confidence Level
0.01	99%
0.05	95%
0.10	90%

1.6.2. Output Interpretation

Illustration: Objective is to predict monthly sales amount for each store of a retail company

```
                    The REG Procedure
                      Model: MODEL1
   Dependent Variable: SALES Sales amount in thousand dollars

          Number of Observations Read         5000
          Number of Observations Used         5000


                   Analysis of Variance

                                Sum of          Mean
Source              DF         Squares        Square    F Value    Pr > F

Model                4      2037609054     509402263    3440.68    <.0001
Error             4995       739523290        148053
Corrected Total   4999      2777132344

      Root MSE             384.77618    R-Square     0.7337
      Dependent Mean      2506.84820    Adj R-Sq     0.7335
      Coeff Var             15.34900
```

Dependent variable

Number of observations in the dataset

Low p-value indicates that overall model is significant

Few important validation metrics – To be covered in Validation training module

Things to Remember

'Number of Observations Used' may be less than 'Number of Observations Read' if there are missing values for any variable. Make sure to do missing value imputation before modeling.

Output Interpretation

… Continued

$$Condition\ Index = \sqrt{\frac{Maximum\ Eigen\ Value}{Eigen\ Value}}$$

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation: Intercept	NUM_TRAN_6M	NUM_ COMPLAINTS	IND_SALES_ GW_GT_10PCT	NUM_CUSTOMER_ VISITS_1WEEK
1	3.98403	1.00000	0.00258	0.00870	0.00812	0.01310	0.00341
2	0.70851	2.37131	0.00183	0.01083	0.18251	0.15859	0.00016407
3	0.18948	4.58544	0.00181	0.28843	0.17261	0.77165	0.01324
4	0.09064	6.62973	0.06329	0.69032	0.27650	0.05419	0.22017
5	0.02735	12.07006	0.93049	0.00171	0.36026	0.00247	0.76302

5 principal components based on 5 inputs (one intercept plus 4 predictors)

Proportion of the variance of the estimate accounted for by each principal component

A collinearity problem occurs when a component associated with a high condition index contributes strongly (variance proportion greater than about 0.5) to the variance of two or more variables

© ExlService Holdings, Inc. 2006-2013

Output Interpretation … Continued

Parameter Estimates

Low p-value (<0.05) for a variable indicates that the variable is significant

Tolerance = 1 / VIF

	Model	Dependent	Variable	DF	Estimate	StdErr	tValue	Probt	StandardizedEst	Tolerance	VarianceInflation	Label
1	MODEL1	SALES	Intercept	1	1429.90184	26.21584	54.5434	0	0	.	0	Intercept
2	MODEL1	SALES	NUM_TRAN_6M	1	0.02487	0.000706	35.2159	7.30232E-243	0.31941	0.64802	1.54315	Number of transactions in last 6 months
3	MODEL1	SALES	NUM_COMPLAINTS	1	-164.99541	6.11000	-27.004	4.95145E-150	-0.24644	0.64011	1.56222	Number of complaints registered by customers in last month
4	MODEL1	SALES	IND_SALES_GW_GT_10PCT	1	295.98996	13.79458	21.457	9.038663E-98	0.19856	0.62256	1.60628	Takes value 1 if sales growth in previous year is greater than 10%
5	MODEL1	SALES	NUM_CUSTOMER_VISITS_1WEEK	1	12.75678	0.36448	34.9997	3.23408E-240	0.32230	0.62868	1.59063	Number of customers visited store in last week

VIF values are low, indicating no issue of multicollinearity

Model Equation

Predicted Sales = 1429.90184
+ 0.02487 * NUM_TRAN_6M
- 164.99541 * NUM_COMPLAINTS
+ 295.98996 * IND_SALES_GW_GT_10PCT
+ 12.75678 * NUM_CUSTOMER_VISITS_1WEEK

Things to Remember

VIF measures the inflation in the variance of the parameter estimate due to collinearity that exists among the predictors

Model Interpretation

- Number of transactions in last 6 months, high sales growth in previous year and customer visits in last 1 week have positive impact on sales
- Number of customer complaints in last month has negative impact on sales

© ExlService Holdings, Inc. 2006-2013

Output Interpretation

… Continued

Variable Contribution Computation

	A	B	C	D	E
1	**Variable**	**Estimate**	**Standardized Estimate**	**Abs. Std. Estimate D = ABS(C)**	**Contribution E = D / Σ(D)**
2	NUM_TRAN_6M	0.02487	0.31941	0.31941	29.4%
3	NUM_COMPLAINTS	-164.99541	-0.24644	0.24644	22.7%
4	IND_SALES_GW_GT_10PCT	295.98996	0.19856	0.19856	18.3%
5	NUM_CUSTOMER_VISITS_1WEEK	12.75678	0.3223	0.3223	29.7%
6	**Total**			**Σ(D) = 1.08671**	**Σ(E) = 100.0%**

Interpretation

- Variable contribution is well distributed across all variables
- Transaction volume and customer visits are the top predictors of a store’s monthly sales amount

Things to Remember

A standardized regression coefficient is computed by dividing a parameter estimate by the ratio of the sample standard deviation of the dependent variable to the sample standard deviation of the regressor

© ExlService Holdings, Inc. 2006-2013

Exercise

Exercise 3. Credit Line Increase Model (Line Assignment Model)

A credit card issuing company has identified the list of charge card customers eligible for line increase. It wants to predict the amount of line increase for each card holder

Location : …\methodology\module_4
Train Data : train_sample_1 (Number of Observations: 35,525)

	Variable	Type	Label
1	ID	Num	Card-holder identification number
2	LI_AMT	Num	Credit line increase amount
3	SPEND_ALL_CARDS_12M	Num	Spend on all cards in last 12 months
4	SPEND_CH_CARDS_3M	Num	Spend on charge cards in last 3 months
5	NUM_30DPD_3M_ANY_ACCT	Num	Number of months in which any charge account was 30 DPD or more in last quarter
6	IND_SPEND_GW_GT_20PCT	Num	Takes value 1 if spend growth in previous year is greater than 20%
7	IND_UTIL_CH_1M_GT_150PCT	Num	Takes value 1 if utilization on charge cards in last month is greater than 150%
8	IND_UTIL_CH_3M_GT_50PCT	Num	Takes value 1 if utilization on charge cards in last quarter is greater than 50%
9	FICO	Num	FICO score of the card-holder
10	INCOME_GE_100K	Num	Takes value 1 if per annum income of card-holder is greater than or equal to 100K
11	IND_GRADE_A	Num	Takes value 1 if the card-holder belongs to high value customer group

Build a linear regression model (target variable: LI_AMT)

Try out selection methods ‘NONE’ and ‘BACKWARD’ and notice the difference in results

© ExlService Holdings, Inc. 2006-2013 Confidential

Chapter 2: Logistic Regression

© ExlService Holdings, Inc. 2006-2013 Confidential

2.1 What is Logistic Regression?

2.1.1. Logistic Regression

Usage

- Logistic Regression is a type of regression technique used to study the relation between a dependent and one or more independent variables, when the dependent variable is *categorical*

Type

- Two types of Logistic Regression comprise: *Binary* and *Multinomial*

Types of Logistic Regression

- **Binary Logistic Regression**[1] (Dependent variable is binary)
- **Multinomial Logistic Regression**[2] (Dependent variable has multiple categories)
  - **Ordered Multinomial Logistic Regression** (Dependent variable has ordered categories)
  - **Nominal Multinomial Logistic Regression** (Dependent variable has unordered categories)

[1] Binary Logistic Regression is popularly referred to as 'Logistic Regression' and is the focus of this chapter
[2] Multinomial Logistic Regression is beyond the scope of this training module

2.1.2. Why Logistic Regression?

What if OLS Linear Regression technique is used to model a BINARY Dependent Variable?

Linear Probability Model is defined as :

$$p_i = \beta_0 + \beta_1 X_i$$

where p_i = probability of occurrence of event

Box 6 *Optional for Interested Readers*

Linear Probability Model

Consider a Linear Regression Model :

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

$$=> E(Y_i) = \beta_0 + \beta_1 X_i \quad \ldots(1)$$

In case of binary dependent variable, Y_i takes only two values : 0 and 1

$$\therefore \quad E(Y_i) = 1 \times \Pr ob(Y_i = 1) + 0 \times \Pr ob(Y_i = 0)$$

$$=> E(Y_i) = \Pr(Y_i = 1)$$

$$=> E(Y_i) = p_i \qquad [\text{Let } \Pr(Y_i = 1) = p_i] \qquad \ldots(2)$$

From (1) and (2), Linear Probability Model can be written as :

$$p_i = \beta_0 + \beta_1 X_i$$

Two key reasons why OLS Linear Regression does not work with a binary target

- **Technical Issue: Violation of Assumptions**

  A binary (i.e. dichotomous) dependent variable in a linear regression model violates assumptions of
  - Homoscedasticity
  - Normality of the Error Term

- **Fundamental Issue: Bounded Probabilities**

  Linear Probability Model: $p_i = \beta_0 + \beta_1 X_i$
  - If X has no upper or lower bound, then for any value of β there are values of X for which either $p_i > 1$ or $p_i < 0$
  - This is contradictory, as the true values of probabilities should lie within (0, 1) interval

© ExlService Holdings, Inc. 2006-2013

Solution to Bounded Probabilities

- **Step 1: Use Odds instead of Probability of Event**

  Odds is defined as:

  $$\text{Odds} = \frac{p_i}{1-p_i} = \frac{\text{probability of event}}{\text{probability of non-event}}$$

  – As probability of event ranges from 0 to 1, odds ranges from 0 to ∞
  – Transforming probabilities to odds removes the upper bound

- **Step 2: Take Natural Logarithm of Odds**

  Logistic Regression Model: $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots \beta_k X_k$

  This is called 'logit' or 'log-odds'. It ranges from -∞ to +∞

© ExlService Holdings, Inc. 2006-2013

2.1.3. Sigmoid Function

Logistic Regression Model

$$Z = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots \beta_k X_k$$

Probability of Event is therefore estimated from logit ('model score') by following transformation:

$$p = \frac{e^Z}{1+e^Z} = \frac{1}{1+e^{-Z}}$$

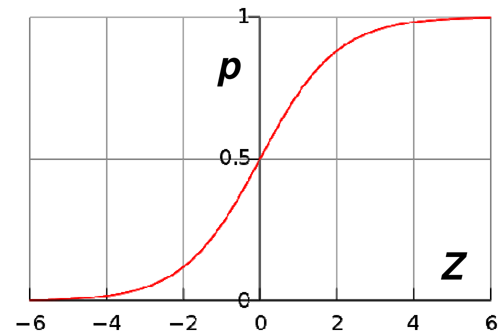
Sigmoid Function or Logistic Function

where 'Z' varies from $-\infty$ to $+\infty$

'p' varies from 0 to 1

Sigmoid or Logistic Curve

- An 'S' shaped curve
- Shows an early exponential growth
- Slows to linear growth in the middle
- Approaches p = 1 with an exponentially decaying gap



2.2 Estimation Method: MLE

Maximum Likelihood Estimation (MLE)

1. Construct Likelihood Function, expressing the likelihood of observing values of dependent variable Y for all n observations

2. Create log likelihood function to simplify the equation

3. Choose values of β's to maximize log likelihood function

Box 7 *Optional for Interested Readers*

Likelihood Function

$$L = \prod_{i=1}^{n} \left(\frac{p_i}{1 - p_i} \right)^{Y_i} (1 - p_i)$$

Derivation: See Appendix A.1

Log Likelihood Function

Taking Log of Likelihood Function

$$\log L = \sum_{i=1}^{n} Y_i \log\left(\frac{p_i}{1 - p_i} \right) + \sum_{i=1}^{n} \log(1 - p_i)$$

Substituting values from Sigmoid Function (Section 2.1.3)

$$\log L = \sum_{i=1}^{n} Y_i Z_i - \sum_{i=1}^{n} \log(1 + e^{Z_i})$$

2.3 Logistic Regression: Key Assumptions

2.3.1. Logistic Regression Assumptions

- ✓ Dependent variable has to be categorical (dichotomous for binary logistic regression)
- ✓ P(Y=1) is the probability of occurrence of event
  - ✓ Dependent variable is to be coded accordingly
  - ✓ For a binary logistic regression, the class 1 of the dependent variable should represent the desired outcome
- ✓ Error terms need to be independent. Logistic regression requires each observation to be independent.
- ✓ Model should have little or no multicollinearity
- ✓ Logistic regression assumes linearity of independent variables and log odds
- ✓ Sample size should be large enough

Note: Maximum likelihood estimates are less powerful than ordinary least squares. As a rule of thumb, while OLS needs at least 5 cases per independent variable, ML needs at least 10 cases per independent variable. Some statisticians even recommend at least 30 cases for each parameter to be estimated in Logistic Regression.

2.3.2. Conditions not required for Logistic Regression

- ✗ Linear relationship between the dependent and independent variables is not necessary
- ✗ Error terms (residuals) do not need to be normally distributed
- ✗ Homoscedasticity is not needed

2.4 Odds Ratio

2.4.1. Definition

Definition 1

- Odds ratio for a predictor is defined as the relative amount by which the odds of the outcome increase (Odds Ratio > 1) or decrease (Odds Ratio < 1) when the value of the predictor variable is increased by 1 unit

$$\text{Odds Ratio for predictor } X_1 = \frac{\left.\left(\frac{p}{1-p}\right)\right|_{X_1=1}}{\left.\left(\frac{p}{1-p}\right)\right|_{X_1=0}}$$

where p is the probability of occurrence of event

Definition 2

- Odds ratio for a predictor is defined as the exponential of its estimated coefficient

$$\text{Odds Ratio for predictor } X_1 = e^{\beta_1}$$

Proof: See Appendix A.2

© ExlService Holdings, Inc. 2006-2013

2.4.2. Interpretation

Interpretation of odds ratio depends on the type of predictor: binary or continuous

$$\text{Odds Ratio} > 1 \Rightarrow \left(\frac{p}{1-p}\right)\Bigg|_{X_1=1} > \left(\frac{p}{1-p}\right)\Bigg|_{X_1=0}$$

When X_1 is binary

Relative probability of event to non-event is higher when X_1 is present vis-à-vis when X_1 is absent

When X_1 is continuous

Relative probability of event to non-event is higher when X_1 increases by 1 unit

---

$$\text{Odds Ratio} < 1 \Rightarrow \left(\frac{p}{1-p}\right)\Bigg|_{X_1=1} < \left(\frac{p}{1-p}\right)\Bigg|_{X_1=0}$$

When X_1 is binary

Relative probability of event to non-event is lower when X_1 is present vis-à-vis when X_1 is absent

When X_1 is continuous

Relative probability of event to non-event is lower when X_1 increases by 1 unit

© ExlService Holdings, Inc. 2006-2013 Confidential

2.5 Frequently Encountered Problems

2.5.1. Complete Separation Problem

Meaning

- Complete separation implies that there is some linear combination of the predictors that perfectly predicts the dependent variable

Illustration

step01_create_data.sas

```
data outlib.comp_sep;
input x y;
datalines;
1 0
2 0
3 0
4 1
5 1
6 1
;
```

Whenever x > 3.5, y = 1
Whenever x < 3.5, y = 0
It is a case of complete separation

comp_sep.sas7bdat

	x	Y
1	1	0
2	2	0
3	3	0
4	4	1
5	5	1
6	6	1

step02_run_logistic_regression.sas

```
proc logistic data = outlib.comp_sep descending;
model y = x;
run;
```

step02_run_logistic_regression.log

WARNING: There is a complete separation of data points. The maximum likelihood estimate does not exist.

2.5.2. Quasi-Complete Separation Problem

Meaning

- Quasi-complete separation problem exists whenever there is complete separation except for at least a single value of the predictor for which both values of the dependent variable occur

Illustration

step01_create_data.sas

```
data outlib.quasi_sep;
input x y;
datalines;
1 0
2 0
3 0
4 0
4 1
5 1
6 1
;
```

For x > 4, y = 1
For x < 4, y = 0
For x = 4, there exist one record with y = 0 and another with y =1
It is a case of quasi-complete separation

quasi_sep.sas7bdat

	x	Y
1	1	0
2	2	0
3	3	0
4	4	0
5	4	1
6	5	1
7	6	1

step02_run_logistic_regression.sas

```
proc logistic data = outlib.quasi_sep descending;
model y = x;
run;
```

step02_run_logistic_regression.log

WARNING: There is possibly a quasi-complete separation of data points. The maximum likelihood estimate may not exist.

2.5.3. Remedies

> **Tip**
>
> In general, categorical (particularly binary) predictors cause separation problems

Problem Detection

- Check warnings in log file
- Identify problematic variables
  - Check cross tab frequencies of categorical independent variables with the dependent variable
  - Look out for cells with zero frequency

Resolution

- Omit problematic variables (**Recommended Solution**)
- Redefine problematic variables (if it makes sense)

Illustration:

quasi_sep.sas7bdat

	x	Y
1	1	0
2	2	0
3	3	0
4	4	0
5	4	1
6	5	1
7	6	1

redefine_x.sas

```
data new;
set quasi_sep;
x_new = x;
if x = 4 then x_new = 1;
run;
```

Creating a new variable "x_new" from "x", assuming that values 1 and 4 have similar meaning and can be clubbed together

new.sas7bdat

	x	Y	x_new
1	1	0	1
2	2	0	2
3	3	0	3
4	4	0	1
5	4	1	1
6	5	1	5
7	6	1	6

2.6 SAS Implementation

2.6.1. LOGISTIC Procedure: SAS Syntax

Below is the syntax for PROC LOGISTIC with frequently used options[1]

PROC LOGISTIC DATA = <*modeling dataset*>	Specify name of modeling (train) dataset for regression
NAMELEN = 32	This option does not let variable name length get truncated to 20
DESCENDING ;	This option reverses the sorting order for the levels of dependent variable
MODEL <*dependent*> = <*regressors*>	
/ SELECTION = <*selection method*>	Specify variable selection method
SLE = <*SLE criterion*>	Specify significance level of entry and stay
SLS = <*SLS criterion*>	
STB ;	This option displays standardized estimates
OUTPUT OUT = <*train predictions*>	Specify name of train scored output dataset
P = P_1 ;	This option requests for score variable name. For example, specify P_1
SCORE DATA = <*test dataset*>	Specify name of validation (test) dataset for scoring
OUT = <*test predictions*> ;	Specify name of test scored output dataset
ODS OUTPUT PARAMETERESTIMATES = <*parameter estimates output dataset*> ;	
RUN ;	

[1] For exhaustive list of options, refer to *SAS OnlineDoc™: Chapter 39: The LOGISTIC Procedure (http://www.math.wpi.edu/saspdf/stat/chap39.pdf)*

2.6.2. Output Interpretation

Illustration: Objective is to predict the probability of a student to score more than 80% marks in the final exam

```
                         The LOGISTIC Procedure

                           Model Information

Data Set                        OUTLIB.ILLUSTRATION_LOG_REG_SAMPLE
Response Variable               IND_MARKS_GT_80PCT
Number of Response Levels       2
Model                           binary logit
Optimization Technique          Fisher's scoring

                            Model Information

      Takes value 1 if a student scores more than 80% marks


             Number of Observations Read        500
             Number of Observations Used        500

                            Response Profile

             Ordered      IND_MARKS_              Total
               Value      GT_80PCT            Frequency

                   1      1                          57
                   2      0                         443

            Probability modeled is IND_MARKS_GT_80PCT='1'.
```

Modeling Dataset

Dependent Variable

Definition of dependent variable

Number of observations in the modeling dataset

Distribution of dependent variable, indicating count of events and non-events

Things to Remember

'Number of Observations Used' may be less than 'Number of Observations Read' if there are missing values for any variable. Make sure to do missing value imputation before modeling.

Output Interpretation

… Continued

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	356.797	192.024
SC	361.012	208.883
-2 Log L	354.797	184.024

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	170.7731	3	<.0001
Score	148.3349	3	<.0001
Wald	62.9089	3	<.0001

Akaike Information Criterion (AIC) and Schwarz Criterion (SC) penalize for number of predictors and can be used to compare different models. The models with smaller values are better.

-2 Log Likelihood is deviance statistic. The lower, the better.

These are three tests to check overall significance of the model.

Low p-values indicate that at least one of the predictors' regression coefficient is not equal to zero in the model, that is the overall model is significant

© ExlService Holdings, Inc. 2006-2013 Confidential

Output Interpretation

… Continued

```
NOTE: No (additional) effects met the 0.05 significance level for removal from the model.
```

Summary of Backward Elimination

Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq	Variable Label
1	ATTENDANCE	1	3	3.1871	0.0742	Attendance (in percentage)

Variable 'ATTENDANCE' got eliminated due to high p-value (>0.05)

© ExlService Holdings, Inc. 2006-2013
Confidential

Output Interpretation ... Continued

Parameter Estimates

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-9.1153	1.4784	38.0165	<.0001	
AVG_MARKS_PREV_5_TESTS	1	0.0953	0.0175	29.4848	<.0001	1.3988
IND_DEC_MARKS_PREV_TEST	1	-1.4206	0.3805	13.9375	0.0002	-0.3632
IND_EXT_GUIDE	1	1.0746	0.4688	5.2547	0.0219	0.2728

Low p-value (<0.05) for a variable indicates that the variable is significant

Model Equation

Probability of Scoring More Than 80% Marks

P_1 = $1 / (1 + e^{-Z})$

where Z = - 9.1153
+ 0.0953 * AVG_MARKS_PREV_5_TESTS
- 1.4206 * IND_DEC_MARKS_PREV_TEST
+ 1.0746 * IND_EXT_GUIDE

Model Interpretation

- Average marks of previous 5 tests and external guidance (tuition) have positive impact on scoring > 80% in final exam
- A declining trend in marks in last test has negative impact on scoring > 80% in final exam

© ExlService Holdings, Inc. 2006-2013

Output Interpretation

... Continued

Variable Contribution Computation[1]

Method 1: Based on Standardized Estimates

	A	B	C	D	E	F	G
1	**Variable**	**Estimate**	**Standardized Estimate**	**Wald Chi Square**	**Abs. Std. Estimate E = ABS(C)**	**Contribution F = E / Σ(E)**	**Contribution G = D / Σ(D)**
2	AVG_MARKS_PREV_5_TESTS	0.0953	1.3988	29.4848	1.3988	68.7%	60.6%
3	IND_DEC_MARKS_PREV_TEST	-1.4206	- 0.3632	13.9375	0.3632	17.8%	28.6%
4	IND_EXT_GUIDE	1.0746	0.2728	5.2547	0.2728	13.4%	10.8%
5	**Total**			**Σ(D) = 48.6770**	**Σ(E) = 2.0348**	**Σ(F) = 100.0%**	**Σ(G) = 100.0%**

Method 2: Based on Wald Chi Sq. values

Interpretation

- Avg. marks scored in previous 5 tests is the key driver for scoring 80% plus marks in final exam

[1] Another way (Method 3) to compute variable contribute is to check loss in log likelihood by removing one predictor at a time and refitting the model

© ExlService Holdings, Inc. 2006-2013 Confidential

Output Interpretation

… Continued

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
AVG_MARKS_PREV_5_TESTS	1.100	1.063	1.138
IND_DEC_MARKS_PREV_TEST	0.242	0.115	0.509
IND_EXT_GUIDE	2.929	1.169	7.341

- Likelihood of scoring more than 80% marks increases by **10%** when average marks in previous 5 tests increases by 1 unit
- Students with a decline in score in the most recent are **75.8%** less likely to score >80% marks than other students
- Students taking external guidance are **192.9%** more likely to score >80% marks than other students

The LOGISTIC Procedure

Association of Predicted Probabilities and Observed Responses

Percent Concordant	92.5	Somers' D	0.861
Percent Discordant	6.4	Gamma	0.871
Percent Tied	1.1	Tau-a	0.174
Pairs	25251	c	0.931

Few important validation metrics – To be covered in Validation training module

Exercise

Exercise 4. VIF values for a Logistic Regression Model

Continue with logistic regression model illustration where the objective is to predict the probability of scoring more than 80% marks in the final exam

Location : …\methodology\module_4
Train Data : train_sample_2 (Number of Observations: 500)

	Variable	Type	Label
1	ROLL_NO	Num	Student roll number
2	IND_MARKS_GT_80PCT	Num	Takes value 1 if a student scores more than 80% marks
3	ATTENDANCE	Num	Attendance (in percentage)
4	AVG_MARKS_PREV_5_TESTS	Num	Average marks scored in previous 5 tests
5	IND_DEC_MARKS_PREV_TEST	Num	Takes value 1 if there was a decline in score in the last test
6	IND_EXT_GUIDE	Num	Takes value 1 if student enrolled for external guidance (tuition)

a. Using PROC LOGISTIC, build a Logistic Regression model (target variable: IND_MARKS_GT_80PCT) and tally your output with illustrative output in Section 2.6.2

b. Report VIF values for final model variables

[Hint: PROC LOGISTIC does not support VIF option. Use PROC REG for generating VIF values]

Chapter 3: Model Improvements

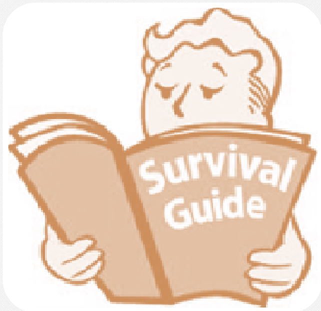
3.1 Choice of Modeling Technique

3.1.1. Is Current Technique Appropriate?

- **DO NOT** apply OLS Linear Regression technique if the dependent variable is categorical (e.g. binary)
- **DO NOT** blindly apply OLS Linear Regression technique, just because the dependent variable is not categorical

- **DO** look at the distribution of dependent variable
- **DO** residual plot analysis

3.1.2. What are the Alternatives[1]?



Count Data Models

- Poisson
- Negative Binomial
- Zero Inflated

Tools: SAS, R

Decision Trees

- Classification Tree
- Regression Tree

Tools: CART, R, SAS E-Miner

Machine Learning

- Neural Network
- Bayesian Network
- Support Vector Machines

Tools: R

Survival Analysis

- Kaplan Meier
- Life Table
- Cox Regression
- Discrete Time Logistic

Tools: SAS, R

Time Series Forecasting

- Holt-Winters
- ARIMA
- ARCH
- GARCH

Tools: SAS, R

[1] This is not to be considered as the exhaustive list of modeling scenarios and techniques

© ExlService Holdings, Inc. 2006-2013 Confidential

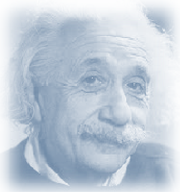
3.2 Variable Innovation

Variable creation in general should precede model development. However, in practice, they go hand-in-hand as model development is an iterative process.

- Create as many useful variables as possible
- Innovate to add value

Variable innovation may be triggered by hypothesis creation or automation need

	Variable Creation	Variable Innovation
Hypothesis Driven	From monthly income and expense information, create income trend and expenditure trend variables	Create expense to income ratio and its trend variable
Automation Driven	**Example 1**: Mathematical transforms like square, cube, square root, cube root, log and inverse	Create all mathematical transforms and retain the best transform for each predictor
	Example 2: Interaction (Variable 1 x Variable 2)	To maximize coverage, create all possible two-way interactions from a given list of predictors and retain the ones that add value and can be interpreted



I am thankful to all those who said no to me. It's because of them I did it myself.

– Albert Einstein

© ExlService Holdings, Inc. 2006-2013

3.3 Oversampling

3.3.1. When, Why and How?

Oversampling is a technique to adjust the class distribution of target variable

1. When

When distribution of dependent variable categories is highly skewed

For instance, event rate < 5%

2. Why

When event rate is low, the oversampling of events

- Reduces the class biasness between events and non-events
- Improves model performance

3. How

Two Common Approaches

- No change in non-events but increase the number of events by randomly replicating existing events
- No change in events but downsize the number of non-events by random sampling of non-events

3.3.2. Intercept Adjustment

Oversampling Implications

- Oversampling has no impact on slope coefficients and hence no impact on rank ordering
- Only the intercept term is to be adjusted to obtain correct probabilities

If β_0 is the intercept term,
Corrected Intercept = β_0 + *Offset*

$$Offset = -\log\left[\left(\frac{1-\lambda}{\lambda}\right)\left(\frac{\pi}{1-\pi}\right)\right]$$

where $\log =$ Natural Logarithm
$\lambda =$ True Event Rate
$\pi =$ Sample Event Rate

Tip

If objective is only to identify top deciles, only rank ordering matters and therefore there is no need for intercept adjustment

Illustration: Offset Calculation

Number of Events *: 10K*
Number of Non-Events *: 200K*
True Event Rate *: 5%*

Suppose 10K non-events are randomly selected from 200K non-events

Number of Events *: 10K*
Number of Non-Events *: 10K*
Sample Event Rate *: 50%*

$$\begin{aligned} Offset &= -\log\left[\left(\frac{1-0.05}{0.05}\right)\left(\frac{0.50}{1-0.50}\right)\right] \\ &= -\log(19) \\ &= -2.9444 \end{aligned}$$

Things to Remember

True Event Rate < Sample Event Rate => Offset < 0

True Event Rate > Sample Event Rate => Offset > 0

3.4 Ensemble

3.4.1. What is Ensemble?

- Ensemble means combining several models into one prediction
- An ensemble model works better than the best individual model component of that ensemble
- Ensemble technique is most effective when individual model components are diverse
- Diversity in individual model components can be attained through
  - Usage of diverse modeling techniques to build individual models
  - Usage of diverse variables in individual models

© ExlService Holdings, Inc. 2006-2013

3.4.2. Common Ensemble Methods

List of Common Ensemble Methods

- **Majority Voting**
  - Simple Majority Voting
  - Weighted Majority Voting
- **Algebraic Combiners**
  - Min Rule
  - Max Rule
  - Product Rule
  - Sum Rule
  - Median Rule
  - Mean Rule
  - Weighted Average Rule
- **Advanced Methods[1]**
  - Boosting
  - Bootstrap Aggregation (BAGGING)
  - Random Forest

[1] Advanced methods are beyond the scope of this training module

3.4.3. Algebraic Combiners and Majority Voting Illustration

Assume target variable has two classes C_1 and C_2 and there are 5 models to be considered for ensemble

	Model 1		Model 2		Model 3		Model 4		Model 5	
Model Weights →	0.25		0.20		0.10		0.15		0.30	
	C_1	C_2	C_1	C_2	C_1	C_2	C_1	C_2	C_1	C_2
Predicted Probabilities →	0.85	0.15	0.30	0.70	0.20	0.80	0.60	0.40	0.40	0.60

Ensemble Rule	Class: C_1	Class: C_2
Min Rule	**MIN**(0.85,0.30,0.20,0.60,0.40) = **0.20**	**MIN**(0.15,0.70,0.80,0.40,0.60) = **0.15**
Max Rule	**MAX**(0.85,0.30,0.20,0.60,0.40) = **0.85**	**MAX**(0.15,0.70,0.80,0.40,0.60) = **0.80**
Product Rule	**PRODUCT**(0.85,0.30,0.20,0.60,0.40) = **0.012**	**PRODUCT**(0.15,0.70,0.80,0.40,0.60) = **0.020**
Sum Rule	**SUM**(0.85,0.30,0.20,0.60,0.40) = **2.35**	**SUM**(0.15,0.70,0.80,0.40,0.60) = **2.65**
Median Rule	**MEDIAN**(0.85,0.30,0.20,0.60,0.40) = **0.40**	**MEDIAN**(0.15,0.70,0.80,0.40,0.60) = **0.60**
Mean Rule	**AVERAGE**(0.85,0.30,0.20,0.60,0.40) = **0.47**	**AVERAGE**(0.15,0.70,0.80,0.40,0.60) = **0.53**
Weighted Average Rule	25%(0.85) + 20%(0.30) + 10%(0.20) + 15%(0.60) + 30%(0.40) = **0.5025**	25%(0.15) + 20%(0.70) + 10%(0.80) + 15%(0.40) + 30%(0.60) = **0.4975**
Simple Majority Voting	**2** Votes (Given by Models 1 and 4)	**3** Votes (Given by Models 2,3 and 5)
Weighted Majority Voting	Sum of Weights of Models 1 and 4 = **0.40**	Sum of Weights of Models 2,3 and 5 = **0.60**

© ExlService Holdings, Inc. 2006-2013

Exercise

Exercise 5. Target variable has three categories: C_1, C_2 and C_3. Ensemble following 5 models using Algebraic Combiners and Majority Voting techniques.

Model Weights →		0.30			0.25			0.20			0.10			0.15	
		Model 1			Model 2			Model 3			Model 4			Model 5	
	C_1	C_2	C_3	C_1	C_2	C_3	C_1	C_2	C_3	C_1	C_2	C_3	C_1	C_2	C_3
Predicted Probabilities →	0.85	0.01	0.14	0.30	0.50	0.20	0.20	0.60	0.20	0.10	0.70	0.20	0.10	0.10	0.80

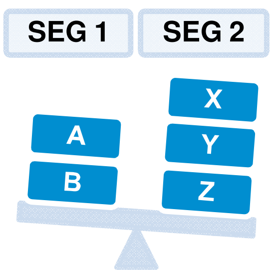
Note: Use the following table to tally answers

Approach	Ensemble Rule	Class: C_1	Class: C_2	Class: C_3
Algebraic Combiners	Min Rule	0.10	0.01	0.14
	Max Rule	0.85	0.70	0.80
	Product Rule	0.00051	0.00021	0.00090
	Sum Rule	1.55	1.91	1.54
	Median Rule	0.20	0.50	0.20
	Mean Rule	0.310	0.382	0.308
	Weighted Average Rule	0.395	0.333	0.272
Majority Voting	Simple Majority Voting	1 Vote	3 Votes	1 Vote
	Weighted Majority Voting	0.30	0.55	0.15

© ExlService Holdings, Inc. 2006-2013

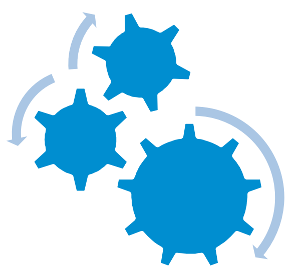
3.5 Segmentation

3.5.1. Need for Segmentation



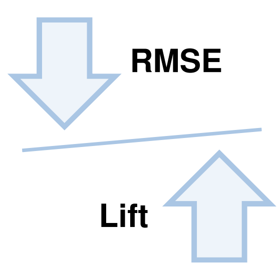
Different portions of data may be driven by different factors

- Variables A and B may be the key drivers of Segment 1
- Variables X,Y and Z may be more relevant for Segment 2



Possibility of Interaction between a binary predictor and other independent variables

- A key predictor with binary values puts a case for different patterns across the two classes
- Segmented models are one way of capturing multiple interactions



Segmentation strategies may boost model performance

- Segmented models can be combined
- Lift of logistic regression models and RMSE of linear regression models show reasonable improvements in most cases

3.5.2. Segmentation Strategies

Business Sense

A

- When the modeler has a fair idea of general patterns at a high level and/or has the required business sense for the purpose of practical application

Dominant Binary Contributor

B

- When there is an extremely high contribution of a binary variable in the base model

Flipping Correlation Sign

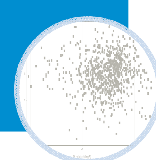
C

- When there are instances of dependent variable correlation coefficient signs getting flipped across two subsets of entire modeling population

Patterns in Error Term

D

- When it is possible to identify some patterns in the error terms of the base model



References

1. **Basic Econometrics**, 4th Ed.
   by Damodar N. Gujarati

2. **Chapter 39: The LOGISTIC Procedure** (http://www.math.wpi.edu/saspdf/stat/chap39.pdf)
   SAS OnlineDoc™

3. **Chapter 55: The REG Procedure** (http://www.math.wpi.edu/saspdf/stat/chap55.pdf)
   SAS OnlineDoc™

4. **Econometric Analysis**, 5th Ed.
   by William H. Greene

5. **Logistic Regression Using SAS**, Theory and Application
   by Paul D. Allison

6. **Should I Build a Segmented Model? A Practitioner's Perspective**
   by Krishna K. Mehta and Varun Aggarwal
   Presented by Krishna Mehta at NYASUG Conference (Jan 14, 2010), Pace University, NY (US)
   Presented by Varun Aggarwal at 2nd IIMA International Conference (Jan 8-9, 2011), Ahmadabad (India)

7. **Talk on Ensemble Methods** (EXL Decision Analytics: Internal BDA Forum)
   by Rahul Lath and Rohit Gupta

8. **Wikipedia** (http://www.wikipedia.org)

© ExlService Holdings, Inc. 2006-2013

Appendix

A.1 Logistic Regression Likelihood Function

Derivation

Back to Main Slide

Likelihood function expresses the likelihood of observing values of dependent variable Y for all n observations

$$L = \Pr(Y_1, Y_2, \ldots, Y_n)$$

$$= \Pr(Y_1)\Pr(Y_2)\ldots\Pr(Y_n)$$ (because observations are assumed to be independent of each other)

$$= \prod_{i=1}^{n} \Pr(Y_i)$$ where Π indicates repeated multiplication

$$= \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{(1-Y_i)} \qquad \because \left.\begin{matrix} \Pr(Y_i = 1) = p_i \\ \Pr(Y_i = 0) = 1 - p_i \end{matrix}\right\} \Rightarrow \Pr(Y_i) = p_i^{Y_i}(1-p_i)^{(1-Y_i)}$$

$$= \prod_{i=1}^{n} \left(\frac{p_i}{1-p_i}\right)^{Y_i}(1-p_i)$$

© ExlService Holdings, Inc. 2006-2013

A.2 Odds Ratio Proof

Proof

Back to Main Slide

Consider a k - variable logistic regression model:

$$Z = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 ... \beta_k X_k$$

where p is the probability of occurrence of event

$$\log(\text{Odds Ratio for predictor } X_1) = \log\left[\frac{\left.\left(\frac{p}{1-p}\right)\right|_{X_1=1}}{\left.\left(\frac{p}{1-p}\right)\right|_{X_1=0}}\right]$$

[Refer to Definition 1 in Section 2.4.1]

$$= \log\left.\left(\frac{p}{1-p}\right)\right|_{X_1=1} - \log\left.\left(\frac{p}{1-p}\right)\right|_{X_1=0}$$

$$= \left.(\beta_0 + \beta_1 X_1 + \beta_2 X_2 ... \beta_k X_k)\right|_{X_1=1} - \left.(\beta_0 + \beta_1 X_1 + \beta_2 X_2 ... \beta_k X_k)\right|_{X_1=0}$$

$$= (\beta_0 + \beta_1 + \beta_2 X_2 ... \beta_k X_k) - (\beta_0 + \beta_2 X_2 ... \beta_k X_k)$$

$$= \beta_1$$

$\Rightarrow$ Odds Ratio for predictor X_1 $= e^{\beta_1}$

© ExlService Holdings, Inc. 2006-2013

Thanks

For queries, contact Varun Aggarwal at Varun.Aggarwal@exlservice.com

© ExlService Holdings, Inc. 2006-2013 Confidential