Introduction to ANOVA, Regression, and Logistic Regression, Course Notes

Predictive Modeling Using Logistic Regression

1. Introduction

Cases(observations, examples)
Input variables(predictors, explanatory)
Target variable(outcome, response)

In **supervised classification**, the target is a class label. A predictive model assigns, to each case, a score that measures the propensity that the case belongs to a particular class. With two classes, the target is binary and usually represents the occurrence of an event. The term **supervised** is used when the class label is known for each case.

The prediction model is used on new cases where the input values are known but the class labels are unknown. The aim is **generalization**, which is predicting the outcome on novel cases.

Traditional statistical analysis like hypothesis test are common inferential tools. Predictive modelers like logistic regression are used to infer how input variable affect the target.

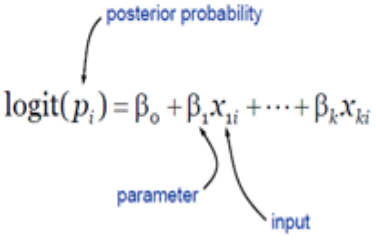
Analytical challenges: Opportunistic data, Mixed Measurement Scales, High Dimensionality, Rare Target Event, Nonlinearities and Interactions, Model selection(overfitting or using too complex a model is common, might be too sensitive to peculiarities in the sample data set and not generalize well to new data. Underfitting disregards the true features.)

2. Fitting the Model

2.1 The Model

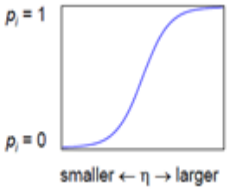
We are always modelling the outcome log(p/(1-p), we define the function logit(p)=log(p/(1-p)) and use the name logit for convenience.

Functional Form

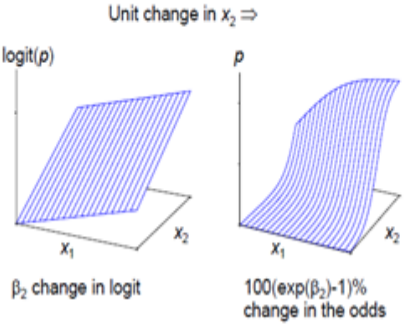


The Logit Link Function

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \eta \quad \Leftrightarrow \quad p_i = \frac{1}{1+e^{-\eta}}$$

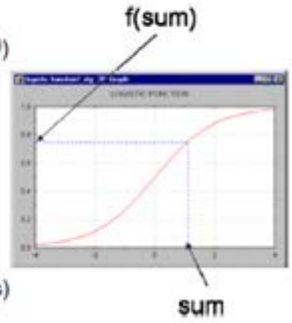


Interpretation



Activation Function f

- logistic (sigmoid)
  - f(sum)=1/(1+e(-sum))
  - between 0 and 1
- hyperbolic tangent
  - f(sum)=(e(sum)-e(-sum))/(e(sum)+e(-sum))
  - between -1 and 1
- linear
  - f(sum)=sum
  - between -∞ and +∞
- Exponential
  - f(sum)=e^sum
  - between 0 and +∞
- Radial Basis Function (RBF Networks)
  - Gaussian activation functions



Logistic regression is a special case of the generalized linear model $g(E(y \mid \mathbf{x})) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$. Each input variable affects the logit linearly. The coefficients are the slopes.

The method of maximum likelihood (ML) is usually used to estimate the unknown parameters in the logistic regression model. The likelihood function is the joint probability density function of the data treated as a function of the parameters. The maximum likelihood estimates are the values of the parameters that maximize the probability of obtaining the sample data. Many SAS procedures can be used; most notable are the LOGISTIC, GENMOD, CATMOD, and DMREG procedures (SAS Enterprise Miner).

Interpreting the parameters(Q60,calculate probability of default with SAS):
logit(Pi)=log(odds)=Beta0+Beta1X1=log(Pi/1-Pi)=default(eg.)
→**odds**=Pi/1-Pi**=exp(default)**
Pi=odds/(1+odds)
Pi=exp(default)/1+exp(default)**=1/(1+exp(-default)**

Introduction to the LOGISTIC Procedure

The LOGISTIC procedure fits a binary logistic regression model. The seven input variables included in the model were selected arbitrarily. The DES (short for descending) option is used to reverse the sorting order for the levels of the response variable Ins. **The CLASS statement names the classification variables to be used in the analysis. The CLASS statement must precede the MODEL statement(Q63, logistic model statements, CLASS statement for categorical predictors). The PARAM option in the CLASS statement specifies the parameterization method for the classification variable or variables and the REF option specifies the reference level(Qextra, codes for LOGISTIC procedure with param and ref).** In this example, **the parameterization method is reference cell coding and the reference level is S(Q56, testing the estimated logit for the reference, eg. 0.1519 in the MLE output).**

The STB option displays the standardized estimates for the parameters for the continuous input variables. The UNITS statement enables you to obtain an odds ratio estimate for a specified change in an input variable. In this example, the UNITS statement enables you to estimate the change in odds for a 1000-unit change in DDABal and DepAmt.

```
proc logistic data=develop des;
   class res (param=ref ref='S');
   model ins = dda ddabal dep depamt
               cashbk checks res
               / stb;
   units ddabal=1000 depamt=1000;
run;
```

The results tables: Class Level Information, Model Fit Statistics(AIC and SCB for goodness-of-fit), Testing Null Hypothesis: Beta=0, **Analysis of Maximum Likelihood Estimates(Q64, output of LOGISTIC procedure),** odds Ratio Estimates(The odds ratio measures the effect of the input variable on the target adjusted for the effect of the other input variables, eg, the odds of DDA is 0.379 times of non-DDA if the point estimate of DDA is 0.379).

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept		1	0.1519	0.0304	24.8969	<.0001	
DDA		1	-0.9699	0.0385	633.6092	<.0001	-0.2074
DDABal		1	0.000072	3.39E-6	448.5968	<.0001	0.2883
Dep		1	-0.0714	0.0109	42.8222	<.0001	-0.0678
DepAmt		1	0.000018	3.225E-6	30.6227	<.0001	0.0660
CashBk		1	-0.5629	0.1145	24.1615	<.0001	-0.0408
Checks		1	-0.00402	0.00317	1.6168	0.2035	-0.0114
Res	R	1	-0.0467	0.0316	2.1907	0.1388	
Res	U	1	-0.0379	0.0280	1.8375	0.1752	

The parameter estimates measure the rate of change in the logit (log odds) corresponding to a one-unit change in input variable(Q35, indicate relationship from the graph of predictor vs. log(odds)), adjusted for the effects of the other inputs. The parameter estimates are difficult to compare because they depend on the units in which the variables are measured. The standardized estimates convert them to standard deviation units. **The absolute value of the standardized estimates can be used to give an approximate ranking of the relative importance of the input variables on the fitted logistic model(Q64)(In this case, Checks is not significant, DDABal is the greatest relative important, CashBk is the least relative important) .** The variable Res has no standardized estimate because it is a class variable.

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
DDA		0.379	0.352	0.409
DDABal		1.000	1.000	1.000
Dep		0.931	0.911	0.951
DepAmt		1.000	1.000	1.000
CashBk		0.570	0.455	0.713
Checks		0.996	0.990	1.002
Res	R vs S	0.954	0.897	1.015
Res	U vs S	0.963	0.911	1.017

The odds ratio measures the effect of the input variable on the target adjusted for the effect of the other input variables. **For example, the odds ratio for a 1000-unit change in DDABal is 1.074(Q51, the interpretation of odds ratio).** Consequently, **the odds** of acquiring the insurance product increases **7.4%**

(calculated as 100(1.074−1)) for every thousand-dollar increase in the checking balance, assuming that the other variables do not change. By default, PROC LOGISTIC reports the 95% Wald confidence interval.

Association of Predicted Probabilities and Observed Responses

Percent Concordant	66.4	Somers' D	0.350
Percent Discordant	31.4	Gamma	0.358
Percent Tied	2.2	Tau-a	0.158
Pairs	235669575	c	0.675

The Association of Predicted Probabilities and Observed Responses table lists several measures that assess the predictive ability of the model. For all pairs of observations with different values of the target variable, **a pair is concordant if the observation with the outcome has a higher predicted outcome probability (based on the model) than the observation without the outcome. A pair is discordant if the observation with the outcome has a lower predicted outcome probability than the observation without the outcome(Q57, the meaning of concordant and discordant).**

The four rank correlation indexes (Somer's D, Gamma, Tau-a, and c) are computed from the numbers of concordant and discordant pairs of observations. In general, a model with higher values for these indexes (the maximum value is 1) has better predictive ability than a model with lower values for these indexes.

How many individuals and responders in the sample have a predicted response rate greater than 0.025?

a. 19323 individuals and 4833 responders
b. 8161 individuals and 2579 responders
c. 2240 individuals and 830 responders
d. 1256 individuals and 547 responders

If the value for the offset is 3.2567, then the model corrected for oversampling will have

a. an intercept that is 3.2567 lower in value compared to the model fitted to the biased sample.
b. probabilities that are 3.2567% higher than the probabilities from the model fitted to the biased sample.
c. probabilities that are 3.2567% lower than the probabilities from the model fitted to the biased sample.
d. parameter estimates that are 3.2567% lower than the parameter estimates from the model fitted to the biased sample.

Scoring new Cases (P38)

The SCORE Statement(Q53,comparing with the SCORE procedure)
The SCORE statement in the LOGISTIC procedure applies the model to a new data set. The DATA= option names the data set to be scored**, and the OUT= option names the resulting scored data set(Q62).** The **predicted probability** that Ins is 1 is named P_1.
The SCORE procedure multiplies values from two SAS data sets, one containing coefficients (SCORE=) and the other containing the data to be scored (DATA=). **The linear combination produced by PROC SCORE (the variable Ins) estimates the logit, not the posterior probability**. The logistic function (inverse of the logit) needs to be applied to compute the posterior probability.

```
proc logistic data=develop des;
   model ins=dda ddabal dep depamt cashbk checks;
   score data = pmlr.new out=scored;
run;

proc print data=scored(obs=20);
   var P_1 dda ddabal dep depamt cashbk checks;
run;
```

OUTEST= Option and the SCORE Procedure

The LOGISTIC procedure outputs the final parameter estimates to a data set using **the OUTEST= option**.

```
proc logistic data=develop des outest=betas1;
   model ins=dda ddabal dep depamt cashbk checks;
run;

proc print data=betas1;
run;
```

The SCORE procedure multiplies values from two SAS data sets, one containing coefficients (SCORE=) and the other containing the data to be scored (DATA=). **Typically, the data set to be scored would not have a target variable(Q52, Augment the training data set with new observations and return the LOGISTIC procedure is NOT an appropriate way to score new observations, because it would return a new model, set their responses to missing is OK). The OUT= option(Q55, the difference of OUTEST and OUT options)** specifies the name of the scored data set created by PROC SCORE. The TYPE=PARMS option is required for scoring regression models.

```
proc score data=pmlr.new
           out=scored
           score=betas1
           type=parms;
   var dda ddabal dep depamt cashbk checks;
run;
```

The linear combination produced by PROC SCORE (the variable Ins) estimates the logit, not the posterior probability. The logistic function (inverse of the logit) needs to be applied to compute the posterior probability.

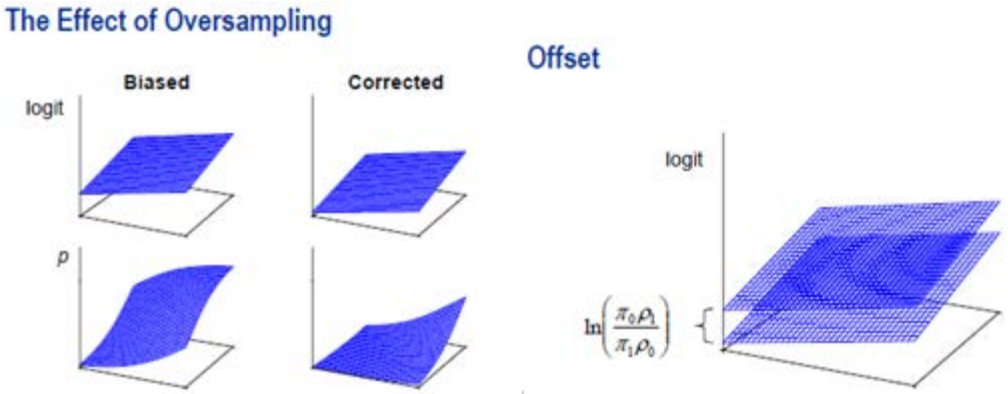
```
data scored;
   set scored;
   p=1/(1+exp(-ins));
run;

proc print data=scored(obs=20);
   var p ins dda ddabal dep depamt cashbk checks;
run;
```

Data can also be scored directly in PROC LOGISTIC using the OUTPUT statement. This has several disadvantages over using PROC SCORE: it does not scale well with large data sets, it requires a target variable (or some proxy), and the adjustments for oversampling, discussed in the next section, are not automatically applied.

2.2 Adjustments for oversampling

In joint (mixture) sampling, the input-target pairs are randomly selected from their joint distribution. In separate sampling, the inputs are randomly selected from their distributions within each target class. Separate sampling is standard practice in supervised classification. When the target event is rare, it is common to oversample the rare event, that is, take a disproportionately large number of event cases. Oversampling rare events is generally believed to lead to better predictions. Separate sampling is also known as case-control sampling, choice-based sampling, biased sampling, outcome-dependent sampling, oversampling.

The priors, π0 and π1, represent the population proportions of class 0 and 1, respectively. The proportions of the target classes in the sample are denoted ρ0 and ρ1. In separate sampling (nonproportional) π0 ≠ ρ0 and π1 ≠ ρ1. The adjustments for oversampling require the priors be known a priori.



The maximum likelihood estimates were derived under the assumption that yi have independent Bernoulli distributions. This assumption is appropriate for joint sampling but not for separate sampling. However, the effects of violating this assumption can be easily corrected. **In logistic regression, only the estimate of the intercept, β0, is affected by using Bernoulli ML on data from a separate sampling design.** If the standard model $\text{logit}(p_i) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ is appropriate for joint sampling, then ML estimates of the parameters under separate sampling can be determined by fitting the pseudo model $\text{logit}(p_i^*) = \ln\left(\frac{\rho_1\pi_0}{\rho_0\pi_1}\right) + \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ where p* is the posterior probability corresponding to the biased sample. Consequently, the effect of oversampling is to shift the logits by a constant amount – the offset $\ln\left(\frac{\rho_1\pi_0}{\rho_0\pi_1}\right)$ When rare events have been oversampled π0 > ρ0 and π1 < ρ1, the offset is positive; that is, the logit is too large. This vertical shift of the logit affects the posterior probability in a corresponding fashion.

The pseudo model can be fitted directly by incorporating the offset into the model. Alternatively, the offset could be applied after the standard model is fitted. Subtracting the offset from the predicted values and solving for the posterior probability gives $\hat{p}_i = \frac{\hat{p}_i^*\rho_0\pi_1}{(1-\hat{p}_i^*)\rho_1\pi_0 + \hat{p}_i^*\rho_0\pi_1}$ where $\hat{p}_i^*$ is the unadjusted estimate of the posterior probability. Both approaches give identical results. For both types

of adjustments, the population priors, π0 and π1, need to be known a priori while the sample priors, ρ0 and ρ1, can be estimated from the data.

Because only the intercept is affected, the adjustments may not be necessary. If the goal of the analysis is to understand the relationships between the inputs and the target, or to rank order the population, then the adjustment is not critical. If the predicted probabilities are important, and not just necessary for rank ordering or classification, then the correction for oversampling is necessary.

Correcting for Oversampling

Separate sampling was used to create the INS data set. The proportion of the target event in the population was .02, not .346 as appears in the sample. The %LET statement defines the macro variable pi1 for the population prior for class 1 (Ins=1).

```
%let pi1=.02;          /* supply the prior for class 1 */
```

The SQL procedure can be used to create macro variables as well. The following code is equivalent to %let rho1 = 0.346361;.

```
proc SQL noprint;
  select mean(INS) into :rho1 from develop;
quit;
```

The SCORE statement in the LOGISTIC procedure will correct predicted probabilities back to the population scale. The option to do this is PRIOREVENT=(Q15, first part, priorevent=probability of events in population).

```
proc logistic data=develop des;
   model ins=dda ddabal dep depamt cashbk checks;
   score data = pmlr.new out=scored priorevent=&pi1;
run;

proc print data=scored(obs=20);
   var P 1 dda ddabal dep depamt cashbk checks;
run;
```

Correcting the Intercept in the Automatic Score Code Generator (Self-Study) P49

Because the correction for oversampling is simply an adjustment to the intercept, you could build that correction into the score code generator(Q59, predicting rare events, only the intercept estimate is biased). The code is similar to the initial score code generator, but the intercept term is corrected by the amount of the offset.

The standard logistic regression model assumes that the logit of the posterior probability is a linear combination of the input variables. The logit transformation is used to constrain the posterior probability to be between zero and one. The parameter estimates are estimated using the method of maximum likelihood. This method finds the parameter estimates that are most likely given the data. When you exponentiate the slope estimates, you obtain the odds ratio, which compares the odds of the event in one group to the odds of the event in another group.

There are many possibilities for scoring new data, from the SCORE statement in PROC LOGISTIC to DATA step code.
When you oversample rare events, you can use the OFFSET option to adjust the model so that the posterior probabilities reflect the population.

General form of the LOGISTIC procedure:

```
PROC LOGISTIC DATA=SAS-data-set <options>;
      CLASS variables </option>;
      MODEL response=predictors </options>;
      UNITS predictor1=list1 </option>;
      SCORE <options>;
RUN;
```

General form of the SCORE procedure:

```
PROC SCORE DATA=SAS-data-set <options>;
      VAR variables;
RUN;
```

3. Preparing the Input Variables

3.1 Missing Values

The default method for treating missing values in most SAS modeling procedures (including the LOGISTIC procedure) is **complete-case analysis in which only the cases without any missing values are used in the analysis(Q37,Q65).**

One reasonable strategy for handling **missing values in predictive modeling is to do the following steps(Q37).**

1. Create missing indicators and treat them as new input variables in the analysis.

$$MI_j = \begin{cases} 1 & \text{if } x_j \text{ is missing} \\ 0 & \text{otherwise} \end{cases}$$

2. Use median imputation. Fill the missing value of xj with the median of the complete cases for that variable.
3. Create a new level representing missing (unknown) for categorical inputs.

The **STDIZE(Q28)** procedure with the REPONLY option can be used to replace missing values. Only numeric input variables should be used in PROC STDIZE.

Mean-imputation uses the unconditional mean of the variable. An attractive extension would be to use the mean conditional on the other inputs. This is referred to as regression imputation. Regression imputation would usually give better estimates of the missing values.
(**Q40+, Given a regression model, one of the predictor is missing, you can't predict the value because it's missing**)

Which of the following statements is false regarding missing values in predictive modeling applications?

a. In complete case analysis, there can be an enormous loss of data when the missing values are spread across many variables.
b. Observations with missing values will not be scored in PROC LOGISTIC.
c. Missing indicator variables can be used to capture the relationship between the target variable and missing inputs.
d. The missing completely at random assumption is valid in most predictive modeling applications.

3.2 Categorical inputs

Including categorical inputs in the model can cause **quasi-complete separation**. Quasi-complete separation occurs when a level of the categorical input has a **target event rate of 0 or 100%.** The coefficient of a dummy variable represents the difference in the logits between that level and the reference level. When quasi-complete separation occurs, one of the logits will be infinite. The likelihood will not have a maximum in at least one dimension, so the ML estimate of that coefficient will be infinite. If the zero-cell category is the reference level, then all the coefficients for the dummy variables will be infinite.

Quasi-complete separation complicates model interpretation. **It can also affect the convergence of the estimation algorithm(Q31)**. Furthermore, it might lead to incorrect decisions regarding variable selection.

The most common cause of quasi-complete separation in predictive modeling is categorical inputs with rare categories(Q34). The best remedy for sparseness is collapsing levels of the categorical variable.

A simple data-driven method for collapsing levels of contingency tables was developed by **Greenacre**. The levels (rows) are
hierarchically clustered based on the reduction in the chi-squared test of association between the categorical variable and the target. **At each step, the two levels that give the least reduction in the chisquared statistic are merged**. The process is continued until the reduction in chi-squared drops below some threshold (for example, 99%). **This method will quickly throw rare categories in with other categories that have similar marginal response rates(Q32). While this method is simple and effective, there is a potential loss of information because only univariate associations are considered.**

Clustering levels of categorical inputs can be clustered using Greenacre's method in the CLUSTER procedure. PROC CLUSTER was designed for general clustering applications, but with some simple pre-processing of the data, it can be made to cluster levels of categorical variables.

Using the FREQ statement and METHOD=WARD in PROC CLUSTER gives identical results to Greenacre's method. The OUTTREE= option creates an output data set that can be used by the TREE procedure to draw a tree diagram.

The column labeled RSQ in the output is equivalent to the proportion of chi-squared in the 19×2 contingency table remaining after the levels are collapsed. At each step, the levels that give the smallest decrease in chi-squared are merged. The change in chi-squared is listed in the SPRSQ column. The rows in the summary represent the results after the listed clusters were merged. The number of clusters is reduced from 18 to 1. When previously collapsed levels are merged, they are denoted using the CL as the prefix and the number of resulting clusters as the suffix.

To calculate the optimum number of clusters, the chi-square statistic and the associated p-value needs to be computed for each collapsed contingency table. This information can be obtained by multiplying the chi-square statistic from the 19×2 contingency table with the proportion of chi-squared remaining after the levels are collapsed.

Which of the following statements is false regarding Greenacre's method for collapsing levels of contingency tables?

a. The levels are hierarchically clustered based on the reduction in the chi-squared test of association.
b. Levels with similar marginal response rates are merged.
c. The method accounts for the sample size in each level.
d. The method is appropriate for any categorical input.

3.3 Variable clustering and screening

Including redundant inputs can degrade the analysis by (Q33)
- destabilizing the parameter estimates
- increasing the risk of overfitting
- confounding interpretation
- increasing computation time
- increasing scoring effort
- increasing the cost of data collection and augmentation.

Redundancy is an unsupervised concept. It does not involve the target variable. In contrast, irrelevant inputs are not substantially related to the target. In high-dimensional data sets, identifying irrelevant inputs is more difficult than identifying redundant inputs. A good strategy is to first reduce redundancy and then tackle irrelevancy in a lower dimension space.

Principal components analysis can be used for reducing redundant dimensions. The principal components (PCs) are linear combinations of the k variables constructed to be jointly uncorrelated and to explain the total variability among the original (standardized) variables.

The principal components are produced by an eigen-decomposition of the correlation matrix, which is the covariance matrix of the standardized variables. The eigenvalues are the variances of the PCs; they sum to the number of variables(each standardized variable has a variance equal to one). The first PC corresponds to the first eigenvalue and explains the largest proportion of the variability. (OUTEST= Data set, contains parameter estimates and, if requested, estimate of the covariance of the parameter

estimates. OUTS= Data set, contains the estimate of the covariance matrix of the residuals across equations.)

Variable clustering as implemented in the **VARCLUS(Q28)** procedure is an alternative method for eliminating redundant dimensions that is closely related to principal components analysis. The result of clustering k variables is a set of ≤ k cluster components. Like PCs, the cluster components are linear combinations of the original variables. Unlike the PCs, the cluster components are not uncorrelated and do not explain all the variability in the original variables.

Variable Screening
Even after variable clustering, some further variable reduction may be needed prior to using the variable selection techniques in the LOGISTIC procedure. Because some of the variable selection techniques use the full model, eliminating clearly irrelevant variables (for example, p-values greater than .50) will stabilize the full model and may improve the variable selection technique without much risk of eliminating important input variables. Keep in mind that univariate screening can give misleading results when there are partial associations. This problem is minimized because the screening is done after PROC VARCLUS, and is used in eliminating clearly irrelevant variables, rather than searching for the best predictors.

The CORR procedure can be used for univariate screening. The **SPEARMAN** option requests Spearman correlation statistics, **which is a correlation of the ranks of the input variables with the binary target(Q38)**. **The Spearman correlation statistic was used rather than the Pearson correlation statistic because Spearman is less sensitive to nonlinearities and outliers**. However, when variables are not monotonically related to each other, the Spearman correlation statistic can miss important associations. A general and robust similarity measure is Hoeffding's D (requested by the **HOEFFDING option) which will detect a wide variety of associations between two variables. Hoeffding's D statistic has values between –0.5 to 1**, but if there are many ties, smaller values may result. The RANK option prints the correlation coefficients for each variable in order from highest to lowest. **The OUTP= option is used to define a SAS data set that contains Pearson correlation coefficients(Q36).**


```
TITLE 'Example 16.3 How to save a correlation matrix and what to do with it';

PROC CORR DATA=mydata NOPRINT OUTP=outcorr;
     VAR age gpa critical satv;
RUN;

PROC PRINT DATA=outcorr;
RUN;
```

A useful table (or plot) would compare the rank order of the Spearman correlation statistic to the rank order of the Hoeffding's D statistic. **If the Spearman rank is high but the Hoeffding's D rank is low, then the association is probably not monotonic.** Empirical logit plots could be used to investigate this type of relationship.

Univariate Smoothing. In regression analysis, it is standard practice to examine scatter plots of the target versus each input variable. When the target is binary, these plots are not very enlightening. A useful plot to detect nonlinear relationships is a plot of the empirical logits. These logits use a minimax estimate of the proportion of events in each bin. The number of bins determines the amount of smoothing (for example, **the fewer bins, the more smoothing**). One large bin would give a constant logit. For very large data sets and intervally scaled inputs, 100 bins often works well.

Empirical Logit Plots

To create a plot of the empirical logits versus a continuous input variable, the input variable first needs to be binned. Use the RANK procedure with the GROUPS= option to bin variables automatically. The bins will be equal size (quantiles) except when the number of tied values exceeds the bin size, in which case the bin will be enlarged to contain all the tied values. The VAR statement lists the variable in the DATA= data set to be grouped. The RANKS statement names the variable representing the groups in the OUT= data set.

```
%let var=DDABal;

proc rank data=imputed groups=100 out=out;
   var &var;
   ranks bin;
run;
```

To compute the empirical logit, the number of target event cases (**ins=1**) and total cases in each bin needs to be computed. **The empirical logits are plotted against the mean of the input variable in each bin(Q29. Screening for non-linearity in binary logistic regression can be achieved by a trend plot of empirical logit vs a predictor variable).** This needs to be computed as well. Both tasks can be done in the MEANS procedure using the CLASS statement. The appropriate statistics (SUM and MEAN) need to be specified in the OUTPUT statement.

```
proc means data=out noprint nway;
   class bin;
   var ins &var;
   output out=bins sum(ins)=ins mean(&var)=&var;
run;
```

The variable *ins* contains the number of events while the automatic variable _FREQ_ contains the bin size.

```
data bins;
   set bins;
   elogit=log((ins+(sqrt(_FREQ_)/2))/
          (_FREQ_-ins+(sqrt(_FREQ_)/2)));
run;
```

The empirical logits can be plotted using the GPLOT procedure. The statement PLOT Y * X requests that the variable Y be plotted on the vertical axis and that the variable X be plotted on the horizontal axis.

```
proc gplot data = bins;
   title "Empirical Logit against &var";
   plot elogit * &var;
run; quit;
```

Accommodating Nonlinearities

3.4 Subset Selection

Variable selection methods in regression are concerned with finding subsets of the inputs that are jointly important in predicting the target. The most thorough search would consider all possible subsets. This can be prohibitively expensive when the number of inputs, k, is large, as there are 2k possible subsets to consider.

Stepwise selection

Stepwise variable selection is a much-maligned yet heavily used subset selection method. Stepwise selection first searches the 1-input models and selects the best. It then searches the 2-input models that contain the input selected in the first step and selects the best. The model is incrementally built in this

fashion until no improvement is made. There is also a **backward** portion of the algorithm where at each step, the variables in the current model can be removed if they have become unimportant. The usual criterion used for entry and removal from the model is the p-value from a significance test that the coefficient is zero, although other criteria can also be used.

Backward variable selection starts with all the candidate variables in the model simultaneously. At each step, the least important input variable is removed (as determined by the p-value). Backward elimination is less inclined to exclude important inputs or include spurious inputs than forward (stepwise) methods. However, it is considered more computationally expensive than stepwise because more steps are usually required and they involve larger models.

All-subsets selection is executed in PROC LOGISTIC with the SELECTION=SCORE option. When combined with the FAST option, backward variable selection requires only a single logistic regression. Fast backward elimination had the best overall performance, a linear increase in time as the number of inputs increased. **The SLSTAY option specifies the significance level for a variable to stay in the model in a backward elimination step (Q45, sls option, values must be between 0 and 1).**

```
proc logistic data=imputed des;
   class res;
   model ins=&screened res / selection=backward fast
         slstay=.001;
run;
```

SLENTRY=value SLE=value(Q44)
specifies the significance level for entry into the model used in the FORWARD and STEPWISE methods. The defaults are 0.50 for FORWARD and 0.15 for STEPWISE.
SLSTAY=value SLS=value(Q45)
specifies the significance level for staying in the model for the BACKWARD and STEPWISE methods. The defaults are 0.10 for BACKWARD and 0.15 for STEPWISE.

The SELECTION=SCORE option finds the best subsets of each model size. The number of models printed of each size is controlled by the BEST= option. Because the best subsets method does not support class variables, dummy variables for Res are created in a DATA step.

```
data imputed;
   set imputed;
   resr=(res='R');
   resu=(res='U');
run;

proc logistic data=imputed des;
   model ins=&screened resr resu
   / selection=score best=1;
run;
```

The score test statistic increases with model size. The Schwarz Bayes criterion (SBC) is often used for model selection. **The SBC** is essentially the –2 log likelihood plus a penalty term that **increases as the model gets bigger**. The penalty term for SBC is (k+1)·ln(n), where k is the number of variables in the model and n is the sample size. **Smaller values of SBC are preferable**. The score test statistic is asymptotically equivalent to the likelihood ratio statistic. The –2 log likelihood is a constant minus the likelihood ratio statistic. Thus, an SBC type statistic could be computed from the score statistic as –(score) + (k+1)·ln(n), where smaller values would be preferable.

Which of the following statements is true regarding best subsets selection?

a. The models are ranked by the likelihood ratio chi-square.
b. The method uses a forward selection to find the best model.
c. The method is relatively efficient for a small number of variables.
d. None of the above

Which model had the highest c statistic from the models generated in PROC LOGISTIC and PROC REG in Exercises 3 and 4?

a. Logistic Regression model with the FAST BACKWARD method using only the selected cluster representatives as inputs.
b. Logistic Regression model with the STEPWISE method using all numeric variables and 3 categorical variables.
c. Logistic Regression model with the FAST BACKWARD method using all numeric variables and 3 categorical variables.
d. Logistic Regression model with the lowest SBC selected by the SCORE method using all numeric variables and 3 categorical variables.
e. Linear Regression model with the STEPWISE method using all numeric variables and the dummy codes for the categorical variables.

3.5 Chapter summary

Preparing the data for predictive modeling can be laborious. First, missing values need to be replaced with reasonable values. Missing indicator variables are also needed if missingness is related to the target. **If there are nominal input variables with numerous levels, the levels should be collapsed to reduce the likelihood of quasi-complete separation and to reduce the redundancy among the levels.** Furthermore, if there are numerous input variables, variable clustering should be performed to reduce the redundancy among the variables. Additionally, there are several selection methods in the LOGISTIC procedure to select a subset of variables.

To assist in identifying nonlinear associations, the Hoeffding's D statistic can be used. **A variable with a low rank in the Spearman correlation statistic but with a high rank in the Hoeffding's D statistic may indicate that the association with the target is nonlinear(but monotonic)**.

General form of the STDIZE procedure:

```
PROC STDIZE DATA=SAS-data-set <options>;
     VAR variables;
RUN;
```

General form of the CLUSTER procedure:

```
PROC CLUSTER DATA=SAS-data-set <options>;
     FREQ variable;
     VAR variable;
     ID variable;
RUN;
```

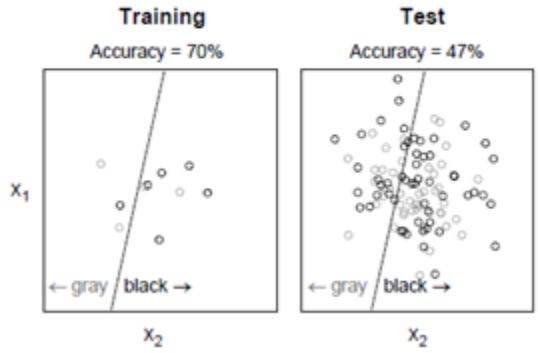
General form of the VARCLUS procedure:

```
PROC VARCLUS DATA=SAS-data-set <options>;
     VAR variables;
RUN;
```

4. Measuring Classifier Performance

4.1 Honest Assessment

The Optimism Principle

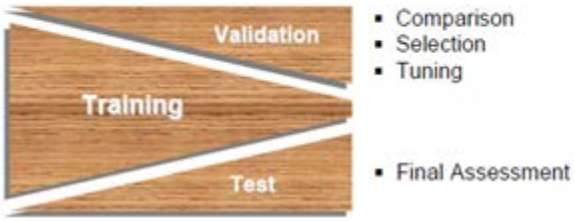


Evaluating the performance of a classifier on the same data used to train the classifier usually leads to an optimistically biased assessment.
For example, the above classifier was fit (or more properly overfit) to a 10-case data set. It correctly classified 70% of the cases. However, when the same classification rule was applied to 100 new cases from the same distribution, only 47% were correctly classified. This is called **overfitting**. The model was overly sensitive to peculiarities of the particular training data, in addition to true features of their joint distribution.

The more flexible the underlying model and less plentiful the data, the more that overfitting is a problem. When a relatively inflexible model like (linear) logistic regression is fitted to massive amounts of data, overfitting may not be a problem. However, the chance of overfitting is increased by variable selection methods and supervised input preparation (such as collapsing levels of nominal variables based on associations with the target). It is prudent to assume overfitting until proven otherwise.

Large differences between the performance on the training and test sets usually indicate overfitting(Qextra).

Data Splitting



The simplest strategy for correcting the optimistic bias is to holdout a portion of the development data for assessment. The model is fit to the remainder (**training data set**) and performance is evaluated on the holdout portion (**test data set**). **Usually from one-fourth to one-half of the development data is used as a test set**. After assessment, it is common practice to refit the final model on the entire undivided data set.

When the holdout data is used for comparing, selecting, and tuning models and the chosen model is assessed on the same data set that was used for comparison, then the optimism principle again applies. In this situation, the holdout sample is more correctly called a **validation data set**, not a test set. **The test set is used for a final assessment of a fully specified classifier(Q16, the purpose of the test data set)**. If model tuning and a final assessment are both needed, then the data should be split three ways into training, validation, and test sets.

Data splitting is a simple but costly technique. When data is scarce, it is inefficient to use only a portion for training. Furthermore, when the test set is small, the performance measures may be unreliable because of high variability. For small and moderate data sets, v-fold cross-validation is a better strategy. In 5-fold cross-validation, for instance, the data would be split into five equal sets. The entire modeling process would be redone on each four-fifths of the data using the remaining one-fifth for assessment. The five assessments would then be averaged. In this way, all the data is used for both training and assessment.

Honest Model Assessment

The objective is to split the data into training and validation sets(Q11, Honest assessment must have training data and validation data, testing data which is final assessment could be missing). The model and all the input preparation steps then need to be redone on the training set. **The validation data will be used for assessment(Q18).** Consequently, it needs to be treated as if it were truly new data where the target is unknown. **The results of the analysis on the training data need to be applied to the validation data, not recalculated(Q2).**

Several input-preparation steps can be done before the data is split. Creating missing indicators should be done on the full development data because the results will not change. The rho1 macro variable is also created before the data is split to get the best estimate of the proportion of events.

The SURVEYSELECT procedure can be used to select the records for the training and validation data sets. To create a stratified sample, the data must be sorted by the stratum variable. The SAMPRATE= option specifies what proportion of the develop data set should be selected(Q5). The default behavior of PROC SURVEYSELECT is to output the sample, not the entire data set, **so the OUTALL option can be used to return the initial data set augmented by a flag to indicate selection in the sample.**

Of course, this flag indicates membership in the training and validation data sets in this context. The FREQ procedure verifies the stratification. The SEED= option enables the user to control what series of pseudo-random numbers is generated to do the partitioning. A particular number, greater than zero, will produce the same split each time the SURVEYSELECT procedure was run. **If the seed were zero, then the data would be split differently each time the procedure was run. (Qextra, in SURVEYSELECT, sampsize= option)**

```
proc sort data=develop out=develop;
   by ins;
run;

proc surveyselect noprint
                  data = develop
                  samprate=.6667
                  out=develop
                  seed=44444
                  outall;
   strata ins;
run;

proc freq data = develop;
   tables ins*selected;
run;
```

The rest of the steps P144-166

To assess the model generalization performance, the validation data need to be prepared for scoring the same way that the training data was prepared for model building. Missing values need to be imputed, new inputs need to be created, and any transformations need to be applied. The MEANS procedure can be used to see which inputs need imputation on this valid data set. **In the validation data set, missing values should be replaced with the medians from the training data set.**

The STDIZE procedure can be used to output a data set that contains the relevant information about the imputed values for every input. In addition, the STDIZE procedure can also be used to take that information and use it to impute the training data set values in the validation data set.

As before, use the STDIZE procedure with the REPONLY option to impute missing values on the training data. In addition, specify the OUTSTAT= option to save the imputed values in a separate data set, called med.

```
proc stdize data = train out=train2
            method=median reponly
            OUTSTAT=med;
   var &inputs;
run;
```

After the med data set has been created, it can be used to impute for missing values in a different data set. The code below creates a data set, valid2, based on the unimputed valid data, with the medians from the training data imputed. The option to specify the data set with the median information is METHOD=IN(data-set-name).

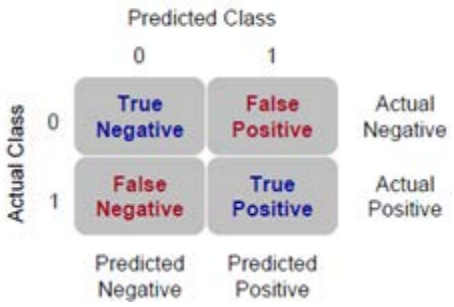
```
proc stdize data=valid out=valid2
            reponly method=in(med);
   var &inputs;
run;
```

To see that the values imputed for AcctAge, Phone, Inv, and CCBal are the same in valid1 (created above) as in valid2 (created here) you can use the COMPARE procedure.

```
proc compare base= valid1 compare=valid2;
   var acctage phone inv ccbal;
run;
```

4.2 Misclassification

Confusion Matrix



The confusion matrix(contingency table)(Q17, calculation of the statistics, eg. Sensitivity and specificity) is a cross tabulation of the actual and predicted classes. It quantifies the confusion of the classifier. The event of interest, whether it is unfavorable (like fraud, churn, or default) or favorable (like response to offer), is often called a positive, although this convention is arbitrary. The simplest performance statistics are:
accuracy (true positives and negatives) / (total cases) (Qextra)
error rate (false positives and negatives) / (total cases) (Qextra).

Two specialized measures of classifier performance are:
Sensitivity (true positives) / (total actual positives) (Q8, SAS codes for sensitivity)
positive predicted value (PV+) (true positives) / (total predicted positives).
The analogues to these measures for true negatives are
Specificity (true negatives) / (total actual negatives)
negative predicted value (PV–) (true negatives) / (total predicted negatives).

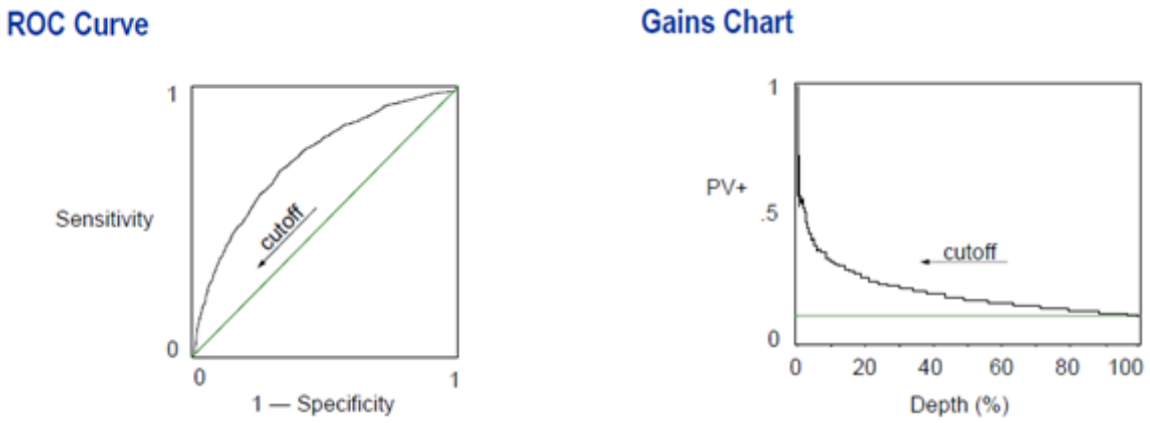
Large sensitivities do not necessarily correspond to large values of PV+. Ideally, one would like large values of all these statistics. The context of the problem determines which of these measures is the primary concern. For example, a database marketer might be most concerned with PV+ because it gives the response rate for the customers that receive an offer. In contrast, a fraud investigator might be most concerned with sensitivity because it gives the proportion of frauds that would be detected.

```
title 'Sensitivity';
proc freq data=FatComp;
   where Response=1;
   weight Count;
   tables Test / binomial(level="1");
   exact binomial;
   run;

title 'Specificity';
proc freq data=FatComp;
   where Response=0;
   weight Count;
   tables Test / binomial(level="0");
   exact binomial;
   run;

title 'Positive predictive value';
proc freq data=FatComp;
   where Test=1;
   weight Count;
   tables Response / binomial(level="1");
   exact binomial;
   run;

title 'Negative predictive value';
proc freq data=FatComp;
   where Test=0;
   weight Count;
   tables Response / binomial(level="0");
   exact binomial;
   run;

title 'False Positive Probability (Col)';
proc freq data=FatComp;
   where Response=0;
   weight Count;
   tables Test / binomial(level="1");
   exact binomial;
   run;
```

Radar detector setting	Percent of German planes detected (sensitivity)	Percent of geese flocks correctly identified (specificity)	Percent of geese flocks incorrectly identified (1- specificity)
Off	0	100	0
Setting 1	35	93	7
Setting 2	60	85	15
Setting 3	85	70	30
Setting 4	92	30	70
Full	100	0	100

ROC Curve were initially developed in World War II, by the British. They were building the "Chain Home" series of radar detectors to identify incoming German planes. But the radar detectors would also detect flocks of birds and other "false positive" signals. (the table above)

Why "sensitivity vs 1 - specificity" and not "sensitivity vs specificity"? **Well, you could do that, but because the area under the curve for sensitivity(True positive rate) vs (1-specificity)(False positive rate) has special meaning, whereas it does not for sensitivity vs specificity, we choose the former(Q58, reading ROC curve and get sensitivity given 1-specificity).**

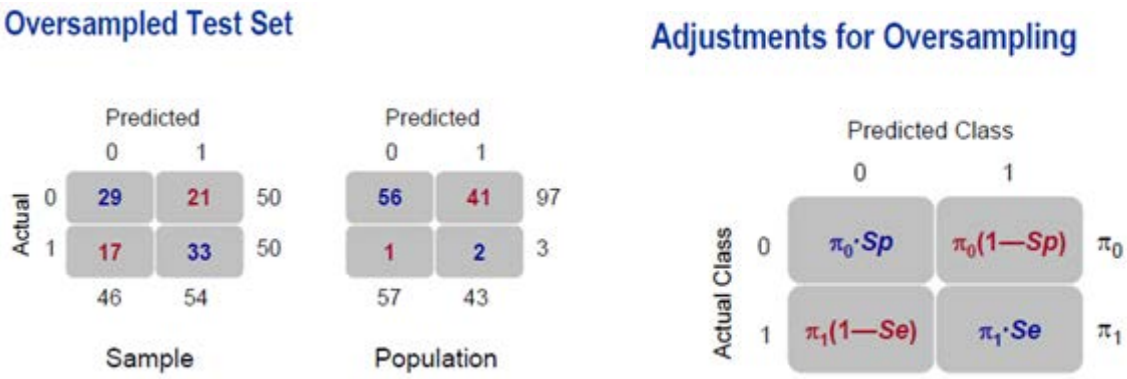


The **receiver operating characteristic (ROC) curve** was adapted from signal detection theory for the assessment of classifiers. **The ROC curve displays the sensitivity and specificity for the entire range of cutoff values(Q1, what changes as you move along the curve). As the cutoff decreases, more and more cases are allocated to class 1; hence, the sensitivity increases and specificity decreases.** As the cutoff increases, more and more cases are allocated to class 0, hence the sensitivity decreases and specificity increases. Consequently, the ROC curve intersects (0,0) and (1,1). If the posterior probabilities were arbitrarily assigned to the cases, then the ratio of false positives to true positives would be the same as the ratio of the total actual negatives to the total actual positives. Consequently, the baseline (random) model is a 45° angle going through the origin. As the ROC curve bows above the diagonal, the predictive power increases. **A perfect model would reach the (0,1) point where both sensitivity and specificity equal 1.**

The depth of a classification rule is the total proportion of cases that were allocated to class 1. The (cumulative) gains chart displays the positive predicted value and depth for a range of cutoff values. As the cutoff decreases, more and more cases are allocated to class 1; hence, the depth increases and the PV+ approaches the marginal event rate. **When the cutoff is minimum, then 100% of the cases are selected and the response rate is ρ1**. As the cutoff increases the depth decreases. A model with good predictive power would have increasing PV+ (response rate) as the depth decreases. If the posterior probabilities were arbitrarily assigned to the cases, then the gains chart would be a horizontal line at ρ1.

The gains chart is widely used in database marketing to decide how deep in a database to go with a promotion. The simplest way to construct this curve is to sort and bin the predicted posterior probabilities (for example, deciles). The gains chart is easily augmented with revenue and cost information. **The lift is PV+/ρ1, so for a given depth, there are (lift)× more responders targeted by the model than by random chance(Q6, the interpretation of lift).**

A plot of sensitivity versus depth is sometimes called a Lorentz curve, concentration curve, or a lift curve (although lift value is not explicitly displayed). This plot and the ROC curve are very similar because depth and 1–specificity are monotonically related.



If the holdout data was obtained by splitting oversampled data, then it is oversampled as well. If the proper adjustments were made when the model was fitted, then the predicted posterior probabilities are correct. However, the confusion matrices would be incorrect (with regard to the population) because the event cases are over-represented. Consequently, PV+ (response rate) might be badly overestimated**. Sensitivity and specificity, however, are not affected by separate sampling because they do not depend on the proportion of each class in the sample(Q14, which are not affected by separate sampling).**

Knowing sensitivity, specificity, and the priors is sufficient for adjusting the confusion matrix for oversampling. For example, if the sample represented the population, then nπ1 cases are in class 1. The proportion of those that were allocated to class 1 is Se. Thus, there are nπ1·Se true positives. Note that these adjustments are equivalent to multiplying the cell counts by their sample weights, for example

$$\mathrm{TP}_{\mathrm{sample}} \cdot wt_1 = \mathrm{TP}_{\mathrm{sample}} \frac{\pi_1}{\rho_1} = \mathrm{TP}_{\mathrm{sample}} \cdot \pi_1 \frac{n}{\mathrm{Tot\ Pos}_{\mathrm{sample}}} = n \cdot \pi_1 \cdot Se$$

where TP is the proportion of true positives, and sample weights are defined as πi / ρi for Class i.

Accessing Classifier Performance

The performance measures need to be calculated on the validation data. One approach would be to use the SCORE procedure to score the validation data and then use DATA steps and the FREQ procedure to calculate misclassification measures for different cutoffs. An easier approach is to score the validation data inside the LOGISTIC procedure and use the **OUTROC= data set, which contains many of the statistics necessary for assessment. The SCORE statement allows for you to output this data set.**

The OUTROC= option creates an output data set with sensitivity (_SENSIT_) and one minus specificity (_1MSPEC_) calculated for a full range of cutoff probabilities (_PROB_)(Q3, which statement and option generate sensitivity and specificity). The other statistics in the OUTROC= data set are not useful when the data is oversampled. The two variables _SENSIT_ and _1MSPEC_ in the OUTROC= data set are correct whether or not the validation data is oversampled. The variable _PROB_ is correct, provided the PRIOREVENT= was set to π1**. If they were not corrected, then _PROB_ needs to be adjusted using the formula** $\hat{p}_i = \dfrac{\hat{p}_i^* \rho_0 \pi_1}{(1-\hat{p}_i^*)\rho_1 \pi_0 + \hat{p}_i^* \rho_0 \pi_1}$ where $\hat{p}_i^*$ is the unadjusted estimate of the posterior probability (_PROB_). The Scoval (scored validation) data set will be used later.

```
proc logistic data=train1 des;
   model ins=&selected;
   score data=valid1 out=scoval
         priorevent=&pi1 outroc=roc;
run;
proc print data=roc(obs=25);
   var _prob_ _sensit_ _1mspec_;
run;
```

Each row in the OUTROC= data set corresponds to a cutoff (_PROB_). The selected cutoffs occur where values of the estimated posterior probability change, provided the posterior probabilities are more than .0001 apart, otherwise they are grouped. Consequently, the maximum number of rows in the OUTROC=data set is 9999, but it is usually much less. The grouping can be made coarser by using the ROCEPS=option in the MODEL statement.

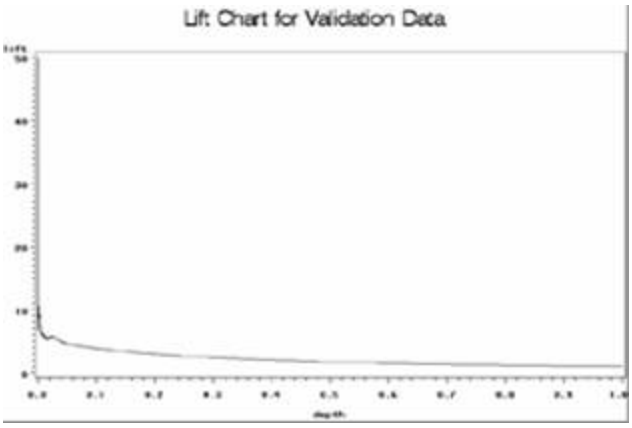
ROC curves can be created by plotting _SENSIT_ versus _1MSPEC_. **Gains charts can be created by plotting PosPV versus Depth or Lift versus Depth.** Often, overlying many statistics versus Depth is informative. These plots follow.

To generate the ROC curve, use the GPLOT procedure. The SYMBOL statement defines the line that will connect the points: join the points (I=JOIN), use no symbol to plot each point (V=NONE), and color the line black.

```
symbol i=join v=none c=black;
proc gplot data = roc;
   title "ROC Curve for the Validation Data Set";
   plot _SENSIT_*_1MSPEC_;
run; quit;
```

To create a lift chart, plot Lift against Depth.

```
symbol i=join v=none c=black;
proc gplot data=roc;
   title "Lift Chart for Validation Data";
   plot lift*depth;
run; quit;
```



To improve this graph, consider adding a reference line at the base line (a lift of 1) and restricting the focus to the region where depth is greater than 0.5% (less erratic) and less than 50%. The final TITLE statement, with no options, clears the title. The VREF= option in the PLOT statement puts a reference line on the vertical axis.

The Meaning of a reference line(Q7+): it represents the expected number of positives we would predict if we did not have a model but simply selected cases at random. It provides a benchmark against which we can see performance of the model.

```
symbol i=join v=none c=black;
proc gplot data=roc;
   where 0.005 < depth < 0.50;
   title "Lift Chart for Validation Data";
   plot lift*depth / vref=1;
run; quit;
title;
```

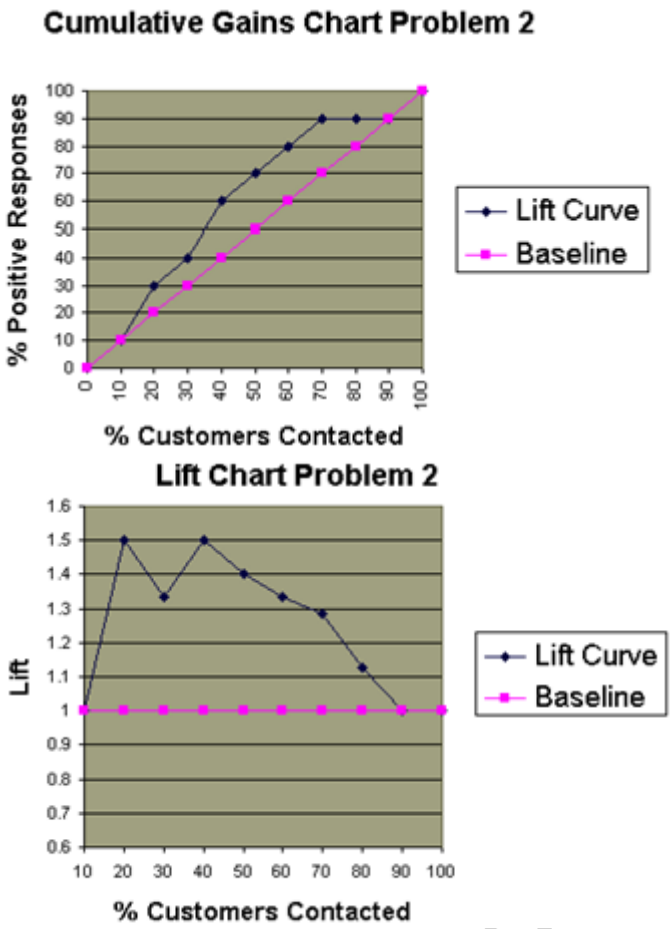
•Lift is a measure of the effectiveness of a predictive model calculated as the ratio between the results obtained with and without the predictive model.
•Cumulative gains and lift charts are visual aids for measuring model performance
•Both charts consist of a lift curve and a baseline
•The greater the area between the lift curve and the baseline, the better the model(Q12, meaning of area under ROC curve)

Example,
Using the response model P(x)=100-AGE(x) for customer x and the data table shown below, construct the cumulative gains and lift charts. Ties in ranking should be arbitrarily broken by assigning a higher rank to who appears first in the table.

Customer Name	*Height*	*Age*	*Actual Response*
Alan	70	39	N
Bob	72	21	Y
Jessica	65	25	Y
Elizabeth	62	30	Y
Hilary	67	19	Y
Fred	69	48	N
Alex	65	12	Y
Margot	63	51	N
Sean	71	65	Y
Chris	73	42	N
Philip	75	20	Y
Catherine	70	23	N
Amy	69	13	N
Erin	68	35	Y
Trent	72	55	N
Preston	68	25	N
John	64	76	N
Nancy	64	24	Y
Kim	72	31	N
Laura	62	29	Y

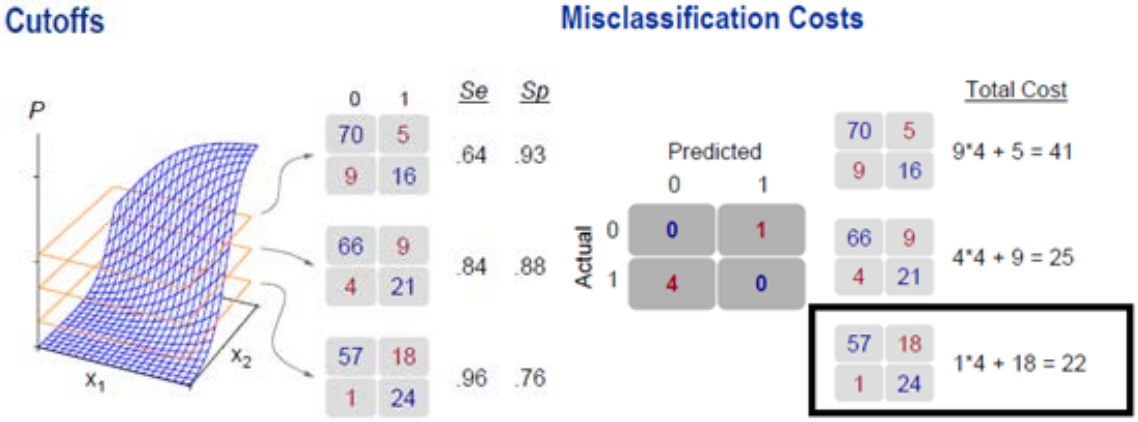
Customer Name	*P(x)*	*Actual Response*
Alex	88	Y
Amy	87	N
Hilary	81	Y
Philip	80	Y
Bob	79	Y
Catherine	77	N
Nancy	76	Y
Jessica	75	Y
Preston	75	N
Laura	71	Y
Elizabeth	70	Y
Kim	69	N
Erin	65	Y
Alan	61	N
Chris	58	N
Fred	52	N
Margot	49	N
Trent	45	N
Sean	35	Y
John	24	N

Total Customers Contacted	Number of Responses	Response Rate
2	1	10%
4	3	30%
6	4	40%
8	6	60%
10	7	70%
12	8	80%
14	9	90%
16	9	90%
18	9	90%
20	10	100%



For contacting 20% of customers, using no model we should get 20% of responders and using the given model we should get 30% of responders. The y-value of the lift curve at 20% is 30 / 20 = 1.5.

4.3 Allocation Rules



Bayes Rule

Allocate to class 1 if

$$p_i > \frac{1}{1+\left(\frac{\text{cost}_{FN}}{\text{cost}_{FP}}\right)}$$

Allocate to class 0, otherwise.

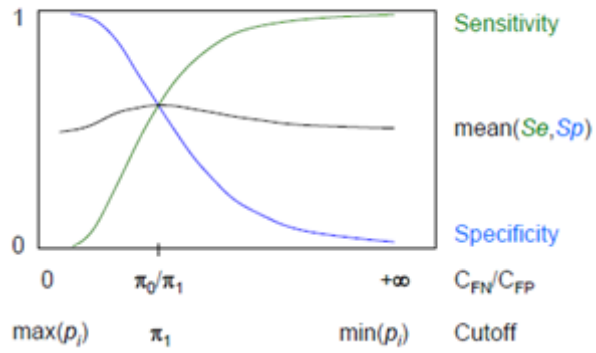
Different cutoffs produce different allocations and different confusion matrices. To determine the optimal cutoff, a performance criterion needs to be defined. If the goal were to increase the sensitivity of the classifier, then the optimal classifier would allocate all cases to class 1. If the goal were to increase specificity, then the optimal classifier would be to allocate all cases to class 0. For realistic data, there is a
trade-off between sensitivity(Se) and specificity(Sp). Higher cutoffs decrease sensitivity and increase specificity. Lower cutoffs decrease specificity and increase sensitivity.

A formal approach to determining the optimal cutoff uses statistical decision theory. The decision-theoretic approach starts by assigning misclassification costs (losses) to each type of error (false positives and false negatives). **The optimal decision rule minimizes the total expected cost (risk).**

The Bayes rule is the decision rule that minimizes the expected cost. In the two-class situation, the Bayes rule can be determined analytically. If you classify a case into class 1, then the cost is $(1-p)\,\text{cost}_{FP}$, where p is the true posterior probability that a case belongs to class 1. If you classify a case into class 0, then the cost is $p \cdot \text{cost}_{FN}$, Therefore, the optimal rule allocates a case to class 1 if $(1-p)\,\text{cost}_{FP} < p \cdot \text{cost}_{FN}$ otherwise allocate the case to class 0. Solving for p gives the optimal cutoff probability. Because p must be estimated from the data, **the plug-in Bayes rule is used in practice** $\hat{p} > \frac{1}{1+\left(\frac{\text{cost}_{FN}}{\text{cost}_{FP}}\right)}$ **(Q15, second part, P_1>?)**Consequently, the plug-in Bayes rule may not achieve the minimum cost if the estimate of the posterior probability is poorly estimated. Note that the Bayes rule only depends on the ratio of the costs, not on their actual values.

If the misclassification costs are equal, then the Bayes rule corresponds to a cutoff of 0.5. The expected cost (risk) equals $\text{cost}_{FP}\,\pi_0(1-Sp) + \text{cost}_{FN}\,\pi_1(1-Se)$ When the cost ratio equals one, the expected cost is proportional to the error rate. A .5 cutoff tends to minimize error rate (maximize accuracy). Hand commented that The use of error rate often suggests insufficiently careful thought about the real objectives…. **When the target event is rare, the cost of a false negative is usually greater than the cost of a false positive.** The cost of not soliciting a responder is greater than the cost of sending a promotion to someone who does not respond**. The cost of accepting an applicant who will default is greater than the cost of rejecting someone who would pay off the loan.** The cost of approving a fraudulent transaction is greater than the cost of denying a legitimate one. Such considerations dictate cutoffs that are less (often much less) than .5.

Cost Ratio



In many situations, it is difficult to precisely quantify the **cost ratio, costFN/costFP**. Examining the performance of a classifier over a range of cost ratios can be useful. A pertinent range of cost ratios is

$$p > \frac{1}{1+\left(\frac{\pi_0}{\pi_1}\right)} = \pi_1$$

usually centered on the ratio of priors π0/π1. This corresponds to a cutoff of π1
For instance, if the nonevent cases were nine times more prevalent than the event cases, then a false negative would be nine times more costly than a false positive and the cutoff would be 0.1. When the cost ratio equals π0/π1, the expected cost is equivalent to the negative sum of sensitivity and specificity. The central cutoff, π1, tends to maximize the mean of sensitivity and specificity. Because increasing sensitivity usually corresponds to decreasing specificity, the central cutoff tends to equalize sensitivity and specificity.

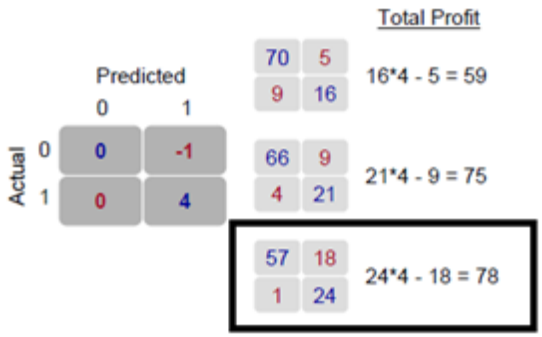
If separate samples were taken with equal allocation (50% events and 50% nonevents), then using the unadjusted cutoff of 0.5 on the biased sample is equivalent to using the central cutoff, π1, on the population.

If the cost of false negatives is 9 times higher than the cost of false positives, then according to Bayes rule what is the cutoff that minimizes the expected cost?

a. 0.90
b. 0.09
c. 1/9
d. 0.10

4.4 Overall Predictive Power

Profit Matrix



Defining a **profit matrix** (instead of a cost matrix, as in the last section) will not lead to a different classification rule. It does point to a useful statistic for measuring classifier performance. The model yields posterior probabilities, and those probabilities (in conjunction with a profit or cost matrix) classify individuals into likely positives and likely negatives. On the validation data, the behavior of these individuals is known; hence, it is feasible to calculate each individual's expected profit, and hence it is also feasible to calculate a total profit. This total profit can be used as a model selection and assessment criterion.

Using Profit to Access Fit

		Predicted Class	
		0	1
Actual Class	0	$0	-$1
	1	$0	$99

This profit matrix is consistent with a marketing effort that costs $1, and, when successful, garners revenue of $100. Hence, the profit for soliciting a non-responder is -$1, and the profit for soliciting a responder is $100-$1 = $99. Given that each individual has a posterior probability pi, you can resort either to the Bayes rule or simple algebra to find the optimum cutoff(Q13, understanding and constructing a profit matrix, may be more complicated).

A typical decision rule would be: Solicit if the expected profit for soliciting, given the posterior probability, is higher than the expected profit for ignoring the customer.

$$\begin{aligned} E(\text{Profit} \mid p_i, \text{solicit}) &> E(\text{Profit} \mid p_i, \text{do not solicit}) \\ p_i*99 + (1-p_i)*(-1) &> p_i*0 + (1-p_i)*(0) \\ 99*p_i - 1 + p_i &> 0 \\ 100*p_i - 1 &> 0 \\ \text{Solicit if:} \quad p_i &> 0.01. \end{aligned}$$

(Q10, profit calculations given all the conditions: Profit=(P_R>0.5)*Purch*200+(P_R>0.5)*(1-Purch)*(-10))

This cutoff of 0.01 can be used to calculate the expected profit of using this rule with the current model. In order to calculate total and average profit comparable to what would be achieved in the population, weights must be calculated. The decision variable is created as a flag indicating whether the predicted

probability is greater than the cutoff, 0.01. Using the information about decision and response, the profit per individual is calculated. These profits are summed and averaged by the MEANS procedure.

```
data scoval;
   set scoval;
   sampwt = (&pi1/&rho1)*(INS)
          + ((1-&pi1)/(1-&rho1))*(1-INS);
   decision = (p 1 > 0.01);
   profit = decision*INS*99
          - decision*(1-INS)*1;
run;
```

```
proc means data=scoval sum mean;
   weight sampwt;
   var profit;
run;
```

The MEANS Procedure

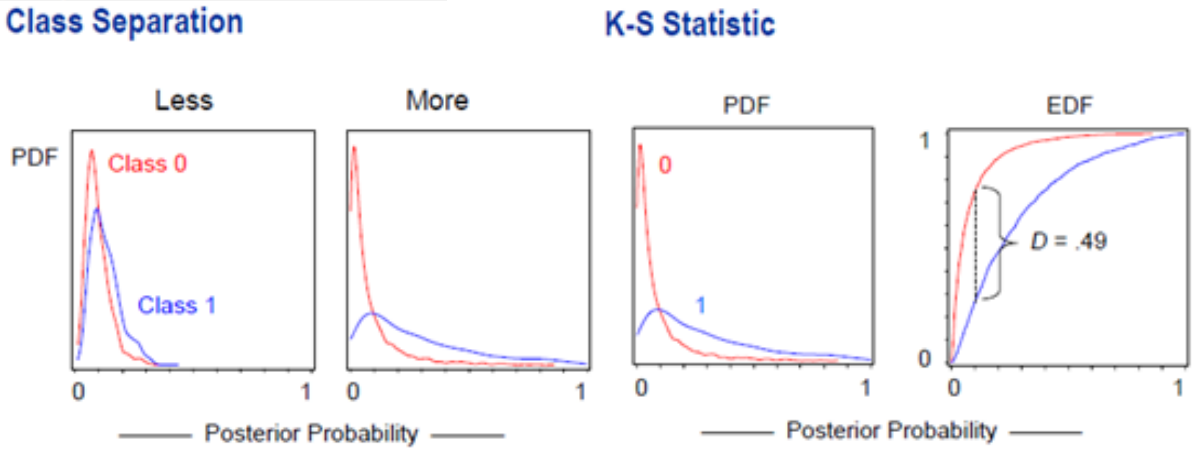
Analysis Variable : profit

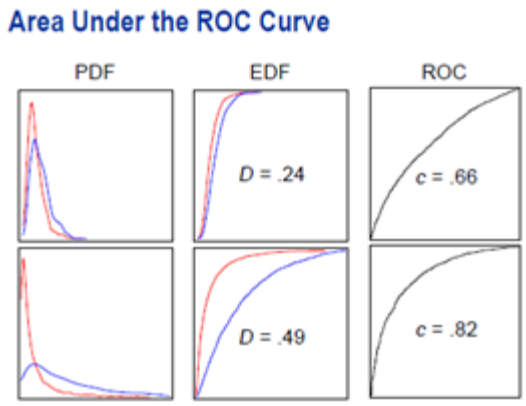
Sum	Mean
13269.60	1.2341397

Using this model to score a population of, say, 1,000,000 individuals and soliciting only those with pi greater than 0.01 would yield a total expected profit of $1,234,139.70. With the above profit matrix and π1=0.02, the "solicit everyone" rule generates a profit of 1,000,000*0.02*99-1,000,000*.98*1= $1,000,000. Using the model and some elementary decision theory leads to better decisions and more profit. Other models can be compared to this current model with this statistic.

Using the true positive and false positive rates calculated earlier, you can calculate the average profit for each of the cutoffs considered in the ROC data set. A false positive (fp) is an individual who is solicited but does not respond. Hence, the cost of soliciting the false positives, on average, is the $1 solicitation cost times the false positive rate. Likewise, the profit associated with individuals who are solicited and respond is $99 times the true positive rate, tp. **The difference of these two terms is the average profit.**

```
data roc;
   set roc;
   AveProf = 99*tp - 1*fp;
run;
```



Statistics such as sensitivity, positive predictive value, and risk depend on the choice of cutoff value. Statistics that summarize the performance of a classifier across a range of cutoffs can also be useful for assessing global discriminatory power. One approach is to measure the separation between the predicted posterior probabilities for each class. **The more that the distributions overlap, the weaker the model.**

The simplest statistics are based on the difference between the means of the two distributions. In credit scoring, the divergence statistic is a scaled difference between the means. Hand discusses several summary measures based on the difference between the means.

The well-known t-test for comparing two distributions is based on the difference between the means. The t-test has many optimal properties when the two distributions are symmetric with equal variance (and have light tails). However, the distributions of the predicted posterior probabilities are typically asymmetric with unequal variance. Many other two-sample tests have been devised for nonnormal distributions.

The Kolmogorov-Smirnov two-sample test is based on the distance between the empirical distribution functions. The test statistic, D, is the maximum vertical difference between the cumulative distributions. If D equals zero, the distributions are everywhere identical. If D>0, then there is some posterior probability where the distributions differ. The maximum value of the K-S statistic, 1, occurs when the distributions are perfectly separated. Use of the K-S statistic for comparing predictive models is popular in database marketing.

The Kolmogorov-Smirnov two-sample test is sensitive to all types of differences between the distributions – location, scale, and shape. In the predictive modeling context, it could be argued that location differences are paramount. Because of its generality, the K-S test is not particularly powerful at detecting location differences. **The most powerful nonparametric two-sample test is the Wilcoxon-Mann-Whitney test. Remarkably, the Wilcoxon-Mann-Whitney test statistic is also equivalent to the area under the ROC curve (the sample question below).**

The Wilcoxon version of this popular two-sample test is based on the ranks of the data. In the predictive modeling context, the predicted posterior probabilities would be ranked from smallest to largest. The test statistic is based on the sum of the ranks in the classes. **The area under the ROC curve, c, can be**

$$c = \frac{\sum_{\{i|y=1\}}^{n_1} R_i - \frac{1}{2} n_1 (n_1 + 1)}{n_1 \cdot n_0}$$

determined from the rank-sum in class 1. The first term in the numerator is the sum of the ranks in class 1.

A perfect ROC curve would be a horizontal line at one – that is, sensitivity and specificity would both equal one for all cutoffs. In this case, the c statistic would equal one. The c statistic technically ranges from zero to one, but in practice, it should not get much lower than one-half. A perfectly random model, where the posterior probabilities were assigned arbitrarily, would give a 45° angle straight ROC curve that intersects the origin; hence, it would give a c statistic of 0.5.

Oversampling does not affect the area under the ROC curve because sensitivity and specificity are unaffected. The area under the ROC curve is also equivalent to the Gini coefficient, which is used to summarize the performance of a Lorentz curve.

Calculating the K-S and c Statistics

The K-S statistic can be computed in the NPAR1WAY procedure using the scored validation data set. The EDF and WILCOXON options in PROC NPAR1WAY request the Kolmogorov-Smirnov and Wilcoxon tests, respectively. The tests compare the values of the variable listed in the VAR statement between the groups listed in the CLASS statement.

```
proc npar1way edf wilcoxon data=scoval;
   class ins;
   var p_1;
run;
```

4.5 Chapter Summary

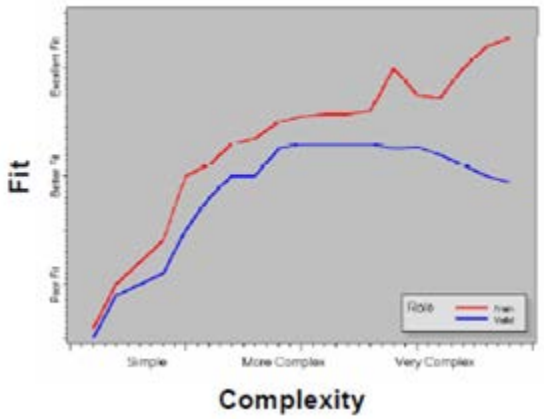
Which of the following statements is false regarding the K-S statistic?

a. The test statistic, D, is the maximum vertical distance between the cumulative distributions.
b. An oversampled validation data set does not affect the test statistic D.
c. The K-S statistic is statistically equivalent to the area under the ROC curve.
d. The K-S statistic is not as powerful at detecting location differences as the Wilcoxon-Mann-Whitney test.

5. Generating and Evaluating Many Models

5.1 Model Selection Plots

Fit versus Complexity

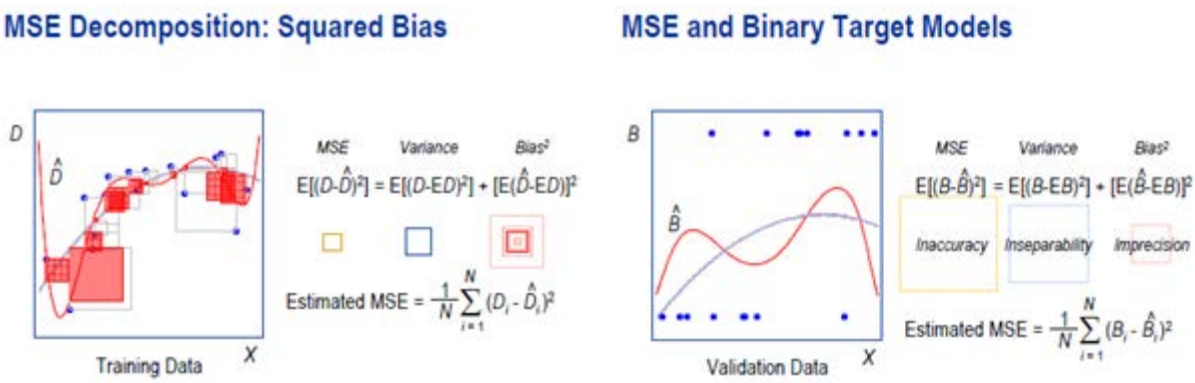


To compare many models, an appropriate fit statistic (or statistics) must be selected. For statistics like average profit, c, and Kolmogorov-Smirnov's D, **higher values mean better fitting models.** Because the goal of most predictive modeling efforts is a model that generalizes well, these statistics are typically measured on the validation data set. For a series of models, which could be generated by an automatic selection routine, it is conceivable to plot a fit measure against some index of complexity. For standard logistic regression models, this index is likely equivalent to the degrees of freedom in the model.

Typically, model performance follows a fairly straightforward trend. As the complexity increases (that is, as terms are added) the fit on the training data gets better. After a point, the fit may plateau, but on the training data, the fit gets better as model complexity increases. Some of this increase is attributable to the model capturing relevant trends in the data. Detecting these trends is the goal of modeling. Some of the increase, however, is due to the model identifying vagaries of the training data set. This behavior has been called overfitting. Because these vagaries are not likely to be repeated, in the validation data or in future observations, it is reasonable to want to eliminate those models. Hence, the model fit on the validation data, for models of varying complexity, is also plotted. The typical behavior of the validation fit line is an increase (as more complex models detect more usable patterns) followed by a plateau, which may finally result in a decline1 in performance. The decline in performance is due to overfitting. The plateau just indicates more complicated models that have no fit-based arguments for their use. A reasonable rule would be to select the model associated with the complexity that has the highest validation fit statistic.

Plotting the training and validation results together permits a further assessment of the model's generalizing power. **Typically, the performance will deteriorate from the training data to the validation data.** This phenomenon, sometimes known as **shrinkage,** is an additional statistic that some modelers use to get a measure of the generalizing power of the selected model. For example, the rule used to select a model from the many different models plotted might be: **"Choose the simplest model that has the highest validation fit measure, with no more than 10% shrinkage from the training to the validation results(Q9, the accuracy of training data and validation data, validation accuracy is more important, overfitting could happen to the training data."**

If the measure of model fit is some sort of error rate, then the plot looks like the above but flipped about the horizontal axis. In the absence of profit or cost information, **the Mean Squared Error (MSE) is one such fitness statistic that measures how poorly a model fits; that is, smaller is better(Q61, usage of MSE for prediction of a binary target variable).**



The first of these components is the residual variance of the target variable. This term quantifies the theoretical limit of prediction accuracy and the absolute lower bound for the MSE. The variance component is independent of any fitted model.

The second MSE component is the average prediction bias squared. This term quantifies the difference between the predicted and actual expected value of the target.

As always, you must be careful to obtain an unbiased estimate of MSE. MSE estimates obtained from the data used to fit the model will almost certainly be overly optimistic. Estimates of MSE from an independent validation data set allow for an honest assessment of model performance.

While MSE is an obvious choice for comparing interval target models, it is also useful for assessing **binary target models**. The estimated MSE can be thought of as measuring the overall inaccuracy of model prediction. This inaccuracy estimate can be decomposed into a term related to the inseparability of the two-target level (corresponding to the variance component) plus a term related to the imprecision of the model estimate (corresponding to the bias-squared component). **In this way, the model with the smallest estimated MSE will also be the least imprecise.**

Whether you choose the MSE, the c statistic, the average profit, Kolmogorov-Smirnov's D or any other measure of model fit, the model selection plot should help point you toward the appropriate level of complexity to achieve good validation generalization.

Likewise, there are many ways to consider generating a series of models of different complexity. Consider any of the automatic variable selection routines described in Chapter 3. Backward elimination of variables creates a series of models of decreasing complexity. Stepwise selection, in general, generates a series of models of increasing complexity. All subsets selection techniques likewise create a series of increasingly complex models.

Which of the following statements is false regarding model assessment statistics?

a. The shrinkage statistic is a measurement of the difference between the training model results and the validation model results.
b. The model with the smallest estimated mean squared error will also be the least imprecise.
c. The variance component of the mean squared error is independent of any fitted model.
d. The bias component of the mean squared error is independent of any fitted model.

All Subsets Regression (P213-224)

5.2 Chapter Summary

Using the validation data set for honest assessment, a series of models can be compared according to their fitness to this task. **In the absence of profit information, the MSE (or, equivalently, ASE) can be used as a model fitness metric to detect models with low bias, or lack-of-fit.**

Intro to ANOVA, Regression, and Logistic Regression

2. ANOVA

2.1 One-Way ANOVA: Two Populations

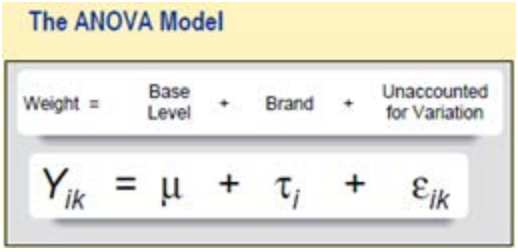
Descriptive Statistics across Groups

```
/* c2demo02 */
goptions reset=all;

proc univariate data=sasuser.b cereal;
   class brand;
   var weight;
   probplot weight / normal
                     (mu=est sigma=est color=blue w=1);
   title 'Univariate Analysis of the Cereal Data';
run;

proc sort data=sasuser.b cereal out=b cereal;
   by brand;
run;

proc boxplot data=b cereal;
   plot weight*brand / cboxes=black boxstyle=schematic;
run;
```

Sorting is not required for the CLASS statement in PROC UNIVARIATE. For the BOXPLOT procedure, however, you must sort the data in ascending order by the second variable in the PLOT statement.

The assumptions for ANOVA are

- independent observations
- normally distributed error terms for each treatment
- approximately equal error variances for each treatment.



Yik: the kth value of the response variable for the ith treatment.
μ: the overall population mean of the response, for instance cereal weight.
τi: the difference between the population mean of the ith treatment and the overall mean, μ. This is referred to as the effect of treatment i.
εik: the difference between the observed value of the kth observation in the ith group and the mean of the ith group. This is called the error term.

As its name implies, analysis of variance analyzes the variances of the data to determine whether there is a difference between the group means.

Between Group Variation — the sum of the squared differences between the mean for each group and the overall mean, Σni(τi)^2. (SSA or SSM)

Within Group Variation — the sum of the squared differences between each observed value and the mean for its group, ΣΣ(Yij−(μ+τi))^2. (SSE)

Total Variation — the sum of the squared differences between each observed value and the overall mean, ΣΣ(Yij−μ)^2. (SST)

Source of Variation	d.f.	SS	MS	F_0
Factor A (between groups)	a-1	$SSA = \sum_{i=1}^{a} n_i (\bar{y}_{i.} - \bar{y}_{..})^2$	$MSA = \frac{SSA}{(a-1)}$	$\frac{MSA}{MSE}$
Error (within groups)	N-a	$SSE = SST - SSA$	$MSE = \frac{SSE}{(N-a)}$	
Total	N-1	$SST = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{..})^2$		

• Model DF is the number of treatments minus 1.
• Corrected total DF is the sample size minus 1(Q43+, the sample size from the REG procedure output).
Source: Model- is the variability explained by your model.
Error- is the variability unexplained by your model.
• Mean square for error (MSE) is an estimate of σ^2, the constant variance assumed for all treatments. Mean square is the ratio of the sum of squares and the degrees of freedom. This value corresponds to the amount of variability associated with each degree of freedom for each source of variation.
• If μ1 = μ2, the mean square for the model (MSM) is also an estimate of σ2.
• If μ1 ≠ μ2, MSM estimates σ2 plus a positive constant.
•F = MSM/MSE . This ratio compares the variability explained by the regression line to the variability unexplained by the regression line.
Based on the above, if the F statistic is significantly larger than 1, it supports rejecting the null hypothesis, concluding that the treatment means are not equal.

R-Square	Coeff Var	Root MSE	weight Mean
0.395427	0.162394	0.024386	15.01649

The coefficient of determination, R^2, denoted in this table as R-Square, is a measure of the proportion of variability explained by the independent variables in the analysis. This statistic is calculated as $R^2 = \frac{SSM}{SST}$ (Q23, Q41, R-Square calculation)

The R^2 always increases as you include more terms in the model. However, choosing the "best" model is not as simple as just making the R2 as large as possible (Q39, what happens if a non-contributing predictor is eliminated from the multiple linear regression model).

The coefficient of variation (denoted Coeff Var) expresses the root MSE (the estimate of the standard deviation for all treatments) as a percent of the mean. It is a unitless measure that is useful in comparing the variability of two sets of data with different units of measure.

The value of R^2 is between 0 and 1. The value is
• close to 0 if the independent variables do not explain much variability in the data
• close to 1 if the independent variables explain a relatively large proportion of variability in the data.
Although values of R^2 closer to 1 are preferred, judging the magnitude of R^2 depends on the context of the problem.

One assumption of ANOVA is approximately equal error variances for each treatment. Although you can get an idea about the equality of variances by looking at the descriptive statistics and plots of the data, you should also consider a formal test for homogeneity of variances. **The GLM procedure has a homogeneity of variance test option (HOVTEST).**

HOVTEST performs Levene's test for homogeneity (equality) of variances. The null hypothesis for this test is that the variances are equal. Levene's test is the default)(Q21, fill in the syntax to test the homogeneity of variance assumption in the GLM procedure).

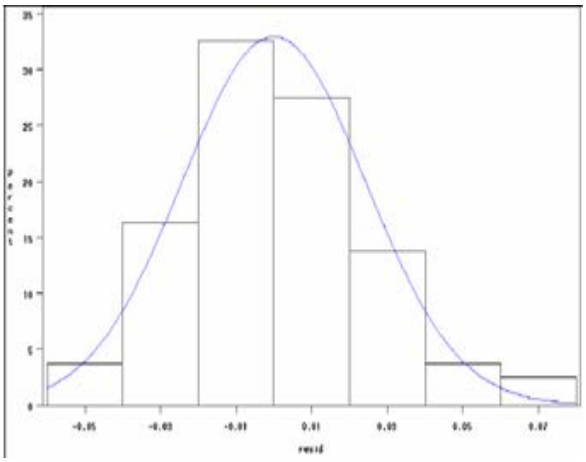
Qextra: filling the blanks
PROC GLM
MEANS A /____=LEVENE
Levene's test is widely considered to be the standard homogeneity of variance test (the HOVTEST=LEVENE option). Levene's test is of the dispersion-variable-ANOVA form discussed previously, where the dispersion variable is either of the following:

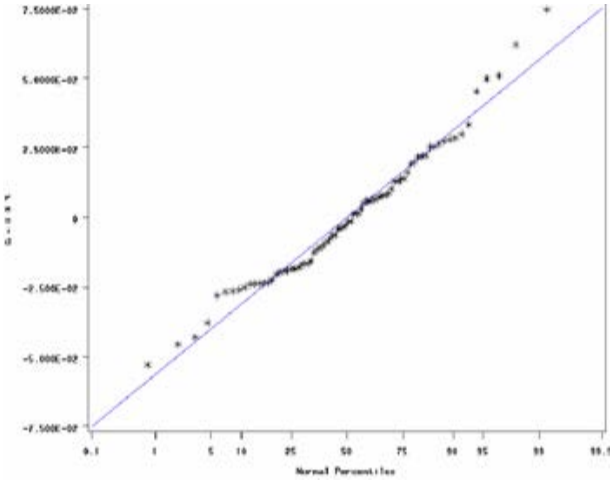
$z_{ij}^2 = (y_{ij} - \bar{y}_i)^2$ (TYPE=SQUARE, the default)
$z_{ij} = |y_{ij} - \bar{y}_i|$ (TYPE=ABS)

If at this point you determined that the variances were not equal, you would add the WELCH option to the MEANS statement. This requests Welch's variance-weighted one-way ANOVA. This alternative to the usual ANOVA is robust to the assumption of equal variances. This is similar to the unequal variance t-test for two populations.

The residuals from the ANOVA are calculated as (the actual value – the predicted value). **These residuals can be examined with PROC UNIVARIATE to determine normality.** With a reasonably sized sample, only severe departures from normality are considered a problem.

These plots provide evidence that the residuals are normally distributed (Q24, checking the assumptions of normality and constant variance).





In ANOVA with more than one predictor variable, the HOVTEST option is unavailable. In those circumstances, you can plot the residuals against their predicted values to verify that the variances are equal. The result will be a set of vertical lines equal to the number of groups. If the lines are approximately the same height, the variances are approximately equal. Descriptive statistics can also be used to determine whether the variances are equal.

The GLM Procedure

General form of the GLM procedure:

```
PROC GLM DATA=SAS-data-set;
     CLASS variables;
     MODEL dependents=independents </ options>;
     MEANS effects </ options>;
     LSMEANS effects </ options>;
     OUTPUT OUT=SAS-data-set keyword=variable...;
RUN;
QUIT;
```

CLASS specifies classification variables for the analysis(Q26, two ways of finding P-value for comparing group means, ttest, CLASS statement and VAR) .

MODEL specifies dependent and independent variables for the analysis(Q42, codes for proc reg, no VAR statement in it).

SOLUTION produces a solution to the normal equations (parameter estimates). PROC GLM displays a solution by default when your model involves no classification variables, so you need this option only if you want to see the solution for models with classification effects(Q49, with categorical predictor C1

```
Proc glm data=sasuser.mlr;
     class c1;
     model y= c1 x1-x3 /solution;
run;)
```

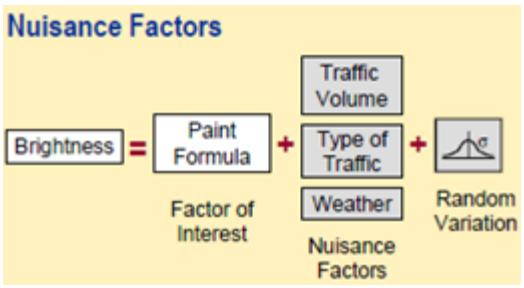
MEANS computes unadjusted means of the dependent variable for each value of the specified effect. LSMEANS produces adjusted means for the outcome variable, broken out by the variable specified and adjusting for any other explanatory variables included in the MODEL statement.

2.2 ANOVA with More than Two Populations

Objectives

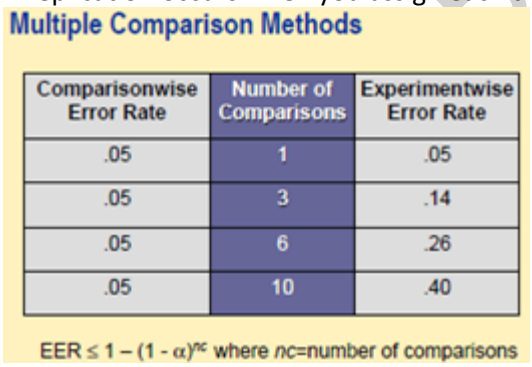
- Recognize the difference between a completely randomized design and a randomized block design.
- Differentiate between observed data and designed experiments.
- Analyze data from the different types of designs using the GLM procedure.

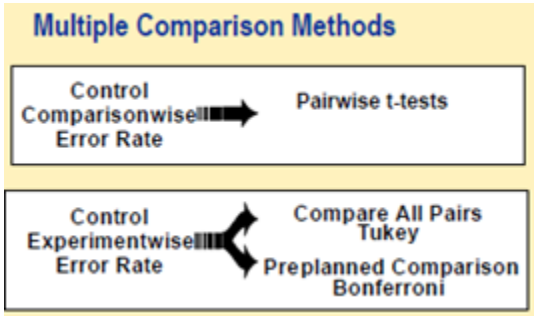
ANOVA is a good example of a situation where often a nonsignificant test is actually useful.
– Suppose we are comparing a new drug to several standard drugs already used
– Suppose also that the new drug is less expensive to produce
– In this case, mostly what we'd like to show is that the new drug is at least effective as the other standard drugs used
– So in this situation, a non-significant ANOVA is a great result!



Factors that can affect the outcome but are not of interest in the experiment are called nuisance factors. The variation due to nuisance factors becomes part of the random variation.

A replication occurs when you assign each treatment to more than one experimental unit.





All of these multiple comparison methods are requested with options in the LSMEANS statement of PROC GLM.
This course addresses these options:
Comparisonwise Control ADJUST=T
Experimentwise Control ADJUST=TUKEY or ADJUST=BONFERRONI

```
/* c2demo07 */
proc glm data=sasuser.b roads;
   class paint;
   model bright=paint;
   lsmeans paint / pdiff=all adjust=t;
   title 'Paint Data: Multiple Comparisons';
run;
quit;
```

PDIFF= requests p-values for the differences, the probability of seeing a difference between two means that is as large as the observed means or larger if the two population means are actually the same. You can request to compare all means using PDIFF=ALL. You can also specify which means to compare.

ADJUST= specifies the adjustment method for multiple comparisons. If no adjustment method is specified, the Tukey method is used by default. The T option asks that no adjustment be made for multiple comparisons. The TUKEY option uses Tukey's adjustment method. The BON option uses the Bonferroni method.

Bonferroni's Method

Bonferroni's multiple comparison method

- is used only for preplanned comparisons
- adjusts for multiple comparisons by dividing the alpha level by the number of comparisons made
- ensures an experimentwise error rate less than or equal to alpha
- is the most conservative method.

Bonferroni's method is not generally considered appropriate for comparisons made after looking at the data, because the adjustment is made based on the number of comparisons you intend to do. If you look at the data to determine how many and what comparisons to make, you are using the data to determine the adjustment.

While Bonferroni's method can be used for all pairwise comparisons (instead of examine the dif. b/t two treatment means, examine all possible combination of two treatment means), Tukey's method is generally less conservative and more appropriate.

Tukey's Multiple Comparison Method

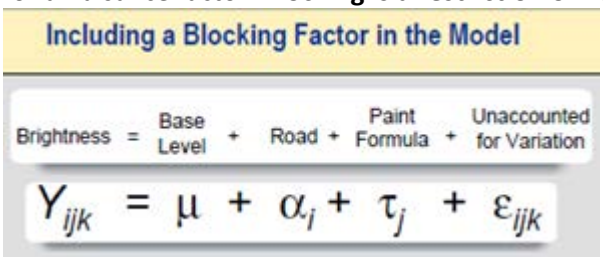
This method is appropriate when considering pairwise comparisons only.
The experimentwise error rate is

- equal to alpha when all pairwise comparisons are considered
- less than alpha when fewer than all pairwise comparisons are considered.

Tukey's multiple comparison adjustment is based on conducting all pairwise comparisons and guarantees the Type I experimentwise error rate is equal to alpha for this situation. If you choose to do fewer than all pairwise comparisons, then this method is more conservative.

• The Tukey and Dunnett tests are only used as follow up tests to ANOVA. They cannot be used to analyze a stack of P values.
• The Tukey test compares every mean with every other mean.
• The Dunnett test compares every mean to a control mean (Q25, Dunnett Adjustment, tell the control group and whether the other groups are significant).

Eg: An experimental design like this is often referred to as a randomized block design, where road is the blocking factor. The variable road is included in the model, but you are not interested in the effect of road, only in controlling the variation it represents. By including road in the model, you could account for a nuisance factor. **Blocking is a restriction on randomization.**



Including a Blocking Factor in the Model
Additional assumptions are as follows:

- Treatments are randomly assigned within each block.
- The effects of the treatment factor are constant across the levels of the blocking factor.

If the effects of the treatment factor are not constant across the levels of the blocking factor, then this condition is called **interaction**. In most randomized block designs, the blocking factor is treated as a **random effect**. Treating an effect as random changes how standard errors are calculated and can give different answers from treating it as a fixed effect.

When treatment groups are going to be compared to each other (in other words, not to 0 or some other specified value), the results from treating the block as a fixed or random effect are exactly the same.

A model that includes both random and fixed effects is called a mixed model and can be analyzed with the MIXED procedure.

In determining the usefulness of having a blocking factor (road) included in the model, you can consider the F value for the block. Some statisticians suggest that if this ratio is greater than 1, then the blocking factor is useful. But if the ratio is less than 1, then adding the variable is detrimental to the analysis. If you find that including the blocking factor is detrimental to the analysis, then you can exclude it from future studies, but it must be included in ANOVA models calculated with the sample that you have already collected.

2.3 Two-Way ANOVA with Interactions

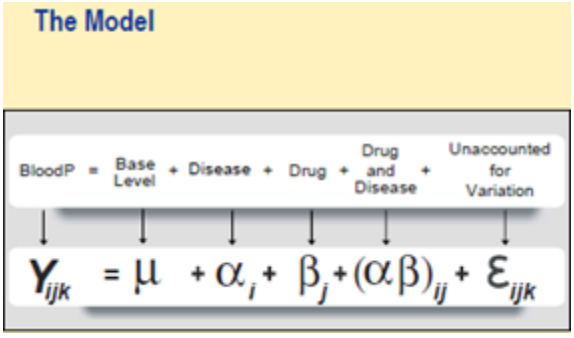
Objectives

- Fit a two-way ANOVA model.
- Detect interactions between factors.
- Analyze the treatments when there is a significant interaction.

In the previous section, you considered the case where you had one categorical predictor and a blocking variable. In this section, consider a case with two categorical predictors. In general, any time you have more than one categorical predictor variable and a continuous response variable, it is called n-way ANOVA. The n can be replaced with the number of categorical predictor variables.

The analysis for a randomized complete block design is actually a special type of n-way ANOVA.

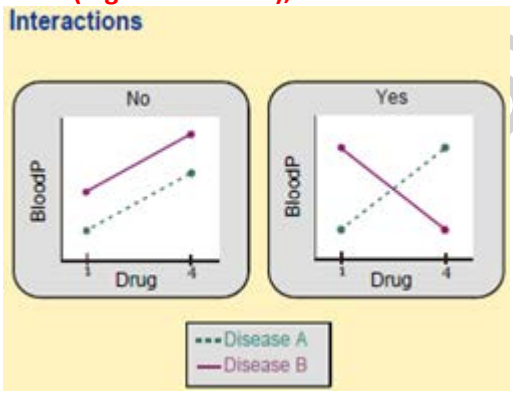
Data was collected in an effort to determine whether different levels of a given drug have an effect on blood pressure for people with a given disease.



Yijk the observed BloodP for each subject
μ the overall population mean of the response, BloodP
αi the effect of the ith Disease
βj the effect of the jth Drug
(αβ)ij the effect of the interaction between the ith Disease and the jth Drug
εijk error term, or residual

In the model it is assumed that the
• observations are independent
• data is normal for each treatment
• variances are approximately equal for each treatment.

(Q48, calculating the number of parameters with quadratic and interaction terms)
An interaction occurs when changing the level of one factor results in changing the difference between levels of the other factor(Q20, if there's significant interaction between Region and Value(high or medium), then the difference between median and high value depends on the region).



The plots displayed above are called means plots. The average blood pressure over different levels of the drug were plotted and then connected for disease A and B. In the right plot, however, as the drug level increases, disease A average blood pressure increases and disease B decreases. **This indicates an interaction between the variables Drug and Disease(Qextra, interaction plot. Income vs. value(Low Medium High)).**

When you analyze an n-way ANOVA, first look at the test for interaction in the analysis of variance output to decide whether there is interaction between the factors. If there is no interaction between the factors, the tests for the individual factor effects can be considered in the output to determine the significance/nonsignificance of these factors.

Neter, Kutner, Wasserman, and Nachtsheim suggest the following guidelines for when to delete the interaction from the model:
• there are fewer than 5 degrees of freedom for the error, and
• the mean square for the interaction divided by the error mean square is less than 2.

Two-Way ANOVA with Interactions(P114)

The data set sasuser.b_drug contains the following variables: Drug- level of drug, Disease- disease category, BloodP- blood pressure

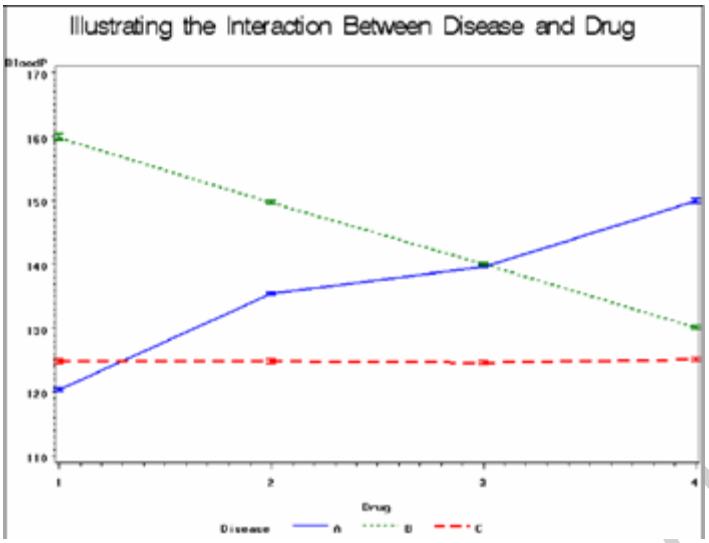
```
options nodate nonumber;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=sasuser.b drug;
   class drug;
   var bloodp;
   histogram / normal;
   probplot /  normal(mu=est sigma=est color=red w=2);
   title 'explore b drug, CLASS drug';
run;

proc univariate data=sasuser.b drug;
   class disease;
   var bloodp;
   histogram / normal;
   probplot /  normal(mu=est sigma=est color=red w=2);
   title 'explore b drug, CLASS disease';
run;
```

Presume that the initial data exploration was completed and that no particular concerns were noted about unusual data values or the distribution of the data. During this exploration, you determine that the sample sizes for all treatments are equal.

```
/* c2demo11 */
proc means data=sasuser.b drug
           mean var std;
   class disease drug;
   var bloodp;
   title 'Selected Descriptive Statistics for sasuser.b drug';
run;
```

To further explore the numerous treatments, examine the PROC MEANS output. The variable BloodP might increase, decrease, or stay the same for the four levels of the variable Drug as seen above.
A means plot can help illustrate these relationships.

```
proc gplot data=sasuser.b drug;
   symbol c=blue w=2 interpol=std1mtj line=1;
   symbol2 c=green w=2 interpol=std1mtj line=2;
   symbol3 c=red w=2 interpol=std1mtj line=3;
   plot bloodp*drug=disease;
   title 'Illustrating the Interaction Between Disease and Drug';
run;
quit;
```



From the graph, the relationship is clearer. For disease type A, blood pressure rises as the drug level increases. For disease type B, blood pressure falls as the drug level increases. For disease type C, blood pressure is relatively unchanged for different drug levels.

```
/* c2demo12 */
proc glm data=sasuser.b drug;
   class disease drug;
   model bloodp=disease drug disease*drug;
   title 'Analyze the Effects of Drug and Disease';
   title2 'Including Interaction';
run;
quit;
```

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Disease	2	8133.949263	4066.974632	4718.18	<.0001
Drug	3	65.146099	21.715366	25.19	<.0001
Disease*Drug	6	9322.836990	1553.806165	1802.60	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Disease	2	8133.949263	4066.974632	4718.18	<.0001
Drug	3	65.146099	21.715366	25.19	<.0001
Disease*Drug	6	9322.836990	1553.806165	1802.60	<.0001

The sums of squares are used to test the null hypothesis that the effect of the individual terms in the model is insignificant. You should consider the test for the interaction first. **The p-value is <0.0001. Presuming an alpha of 0.05, you reject the null hypothesis. You have sufficient evidence to conclude that there is an interaction between the two factors (Q27, detecting interactions).** As shown in the graph, the effect of the level of drug changes for different disease types.

Because of the interaction, you do not want to test the factors separately for differences between the means. Instead, specify that differences across treatment groups are supposed to be tested for both factors simultaneously by specifying them both in the LSMEANS statement separated by an asterisk.

```
proc glm data=sasuser.b drug;
   class disease drug;
   model bloodp=drug disease drug*disease;
   lsmeans disease*drug / adjust=tukey pdiff=all;
   title 'Multiple Comparisons Tests for Drug and Disease';
run;
quit;
```

2.4 Chapter Summary

An analysis of variance (ANOVA) is used to determine whether the means of a continuous measurement for two or more groups are equal. The response variable, or dependent variable, is of primary interest and is a continuous variable. The predictor variable, or independent variable, is a categorical variable. A oneway ANOVA has one independent, or grouping, variable.

Three analyses were discussed: completely randomized, randomized block, and two-way ANOVA. If the result of an analysis of variance is to reject the null hypothesis and conclude that there are differences between the population group means, then multiple comparison tests are used to determine which pairs of means are different. The least significant difference test controls only the Comparisonwise error rate. There are many multiple comparison techniques that control the experimentwise error rate.

The assumptions of an analysis of variance are
• observations are independent.
• pooled residuals are approximately normal.
• all groups have approximately equal response variances.
These assumptions can be verified using a combination of statements and options from the GLM and GPLOT (or PLOT) procedures.
• Examine the residuals plot. Look for a random scatter within each group.
• Examine the distribution of the residuals using PROC UNIVARIATE output. Look for values for skewness and kurtosis close to zero, a symmetric box-and-whisker plot, nonsignificant measures for the normality statistics, and a normal appearing normal probability plot.
• Use the MEANS statement HOVTEST option in PROC GLM, and compare the p-value with alpha; the null hypothesis for this test is that the variances are approximately equal. If you reject the null hypothesis, then you have sufficient evidence to conclude that the variances are not equal.

3. Regression

You use correlation analysis to examine and describe the relationship between two continuous variables. However, before you use correlation analysis, it is important to view the relationship between two continuous variables using a scatter plot.

Two variables are correlated if there is a linear relationship between them. If not, the variables are uncorrelated.

A common correlation statistic used for continuous variables is the Pearson correlation coefficient.
Values of correlation statistics are
• between −1 and 1
• closer to either extreme if there is a high degree of linear relationship between the two variables
• close to 0 if there is no linear relationship between the two variables
• greater than 0 if there is a positive linear relationship
• less than 0 if there is a negative linear relationship.

Generating Correlation Coefficients

Use PROC CORR to produce a Pearson correlation coefficient for Oxygen_Consumption with the other continuous predictor variables.

```
/* c3demo03 */
proc corr data=sasuser.b fitness rank;
   var runtime age weight run pulse rest pulse maximum pulse
       performance;
   with oxygen consumption;
   title 'PROC CORR: oxygen consumption with predictor variables';
run;
```

WITH produces correlations for each variable in the VAR statement with all variables in the WITH statement. The WITH statement specifies the row variables in the correlation matrix.
RANK orders the correlations from highest to lowest in absolute value.

Simple Linear Regression Model

The meaning of Intercept(β0): the predicted value of the response when all predictors=0(Q46).
The method of least squares produces parameter estimates with certain optimum properties. If the assumptions of simple linear regression are valid, the least squares estimates are unbiased estimates of the population parameters and have minimum variance. The least squares estimators are often called BLUE (Best Linear Unbiased Estimators). The term best is used because of the minimum variance property.
Model Hypothesis Test
Null Hypothesis:

- The simple linear regression model does not fit the data better than the baseline model.
- $\beta 1 = 0$

Alternative Hypothesis:

- The simple linear regression model does fit the data better than the baseline model.
- $\beta 1 \neq 0$

If the estimated simple linear regression model **does** fit the data better than the baseline model, you reject the null hypothesis. Thus, you **do** have enough evidence to say that the slope of the regression line in the population is **not** 0 and that the predictor variable explains a significant amount of variability in the response variable.

Performing Simple Linear Regression

The Number of Observations Read and the Number of Observations Used are the same, **indicating that no missing values were detected** for Oxygen_Consumption and Performance.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	635.34150	635.34150	85.22	<.0001
Error	29	216.21305	7.45562		
Corrected Total	30	851.55455			

The F value is testing whether the slope of the predictor variable is equal to 0. The p-value is small (less than .05), so you have enough evidence at the .05 significance level to reject the null hypothesis. Thus, you can conclude that the simple linear regression model fits the data better than the baseline model. In other words, Performance explains a significant amount of variability of Oxygen_Consumption.

The second part of the output provides summary measures of fit for the model.

Root MSE	2.73050	R-Square	0.7461
Dependent Mean	47.37581	Adj R-Sq	0.7373
Coeff Var	5.76349		

Coeff Var- the coefficient of variation is the size of the standard deviation relative to the mean. The coefficient of variation is

- calculated as $\left(\frac{RootMSE}{\bar{Y}}\right) * 100$
- a unitless measure, so it can be used to compare data that has different units of measurement or different magnitudes of measurement.

Adj R-Sq(Adjusted R Square**)- the adjusted R^2 is the R^2 that is adjusted for the number of parameters in the model or it takes account the number of terms in the model(Q43, which statistic measures a better model fit).** This statistic is useful in multiple regression and is discussed in a later section.

Eg.The adjusted R^2 for this model is 0.7313, smaller than the adjusted R^2 of 0.7373 for the Performance only model. This strongly suggests that the variable Runtime does not explain the oxygen consumption capacity if you know Performance.

The R^2 and adjusted R^2 are the same as calculated during the model selection program. If there are missing values in the data set, however, this might not be true(Q40+, if there are missing values in the predictors, the error terms are unpredictable).

Producing Predicted Values

Example: Produce predicted values of Oxygen_Consumption when Performance is 0, 3, 6, 9, and 12.

```
/* c3demo06 */
data need predictions;
   input performance @@;
   datalines;
0 3 6 9 12
;
run;

data predoxy;
   set sasuser.b fitness
       need predictions;
run;

proc reg data=predoxy;
   model oxygen consumption=performance / p;
   id performance;
   title 'Oxygen Consumption=Performance with Predicted Values';
run;
quit;
```

ID- specifies a variable to label observations in the output produced by certain MODEL statement options.
P- prints the values of the response variable, the predicted values, and the residual values.

Producing Confidence Intervals

```
/* c3demo07 */
options ps=50 ls=76;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;

proc reg data=predoxy;
   model oxygen consumption=performance / clm cli alpha=.05;
   id name performance;
   plot oxygen consumption*performance / conf pred;
   symbol1 c=red v=dot;
   symbol2 c=red;
   symbol3 c=blue;
   symbol4 c=blue;
   symbol5 c=green;
   symbol6 c=green;
   title;
run;
quit;
```

Concepts of Multiple Regression

Multiple Linear Regression with Two Variables

The two-variable model

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

where

Y	is the dependent variable.
X_1 and X_2	are the independent or predictor variables.
ε	is the error term.
β_0, β_1, and β_2	are unknown parameters.

If there is a relationship among Y and X1 and X2, the model is a sloping plane passing through three points:

- (Y = β0, X1 = 0, X2 = 0)
- (Y = β0 + β1, X1 = 1, X2 = 0)
- (Y = β0 + β2, X1 = 0, X2 = 1)

Model Hypothesis Test
Null Hypothesis:

- The regression model does not fit the data better than the baseline model.
- $\beta 1 = \beta 2 = \ldots = \beta k = 0$

Alternative Hypothesis:

- The regression model does fit the data better than the baseline model.
- Not all βis equal zero.

Assumptions for Linear Regression:

- The mean of the Ys is accurately modeled by a linear function of the Xs.
- The random error term, ε, is assumed to have a normal distribution with a mean of zero.
- The random error term, ε, is assumed to have a constant variance, σ2.
- The errors are independent.

Fitting a Multiple Linear Regression Model

```
/* c3demo08 */
proc reg data=sasuser.b fitness;
   model oxygen consumption=performance runtime;
   title 'Multiple Linear Regression for b fitness Data';
run;
quit;
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	55.37940	33.79380	1.64	0.1125
Performance	1	0.85780	1.06475	0.81	0.4272
Runtime	1	-1.40429	2.39427	-0.59	0.5622

Using the estimates for β0, β1, and β2 above, this model can be written as Oxygen_Consumption = 55.3794 + 0.8578*Performance − 1.40429*Runtime
Both the p-values for Performance and Runtime are large, which suggests that neither slope is significantly different from 0.

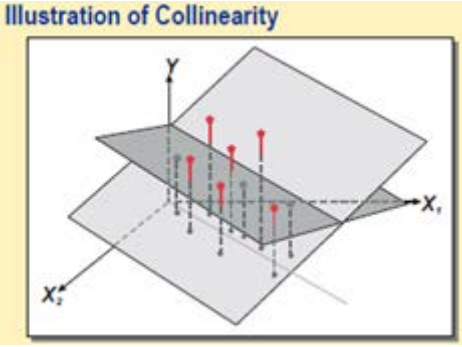
The reason is that the test for βi=0 is conditioned on the other terms in the model. So the test for β1=0 is conditional on or adjusted for X2 (Runtime). Similarly, the test for β2=0 is conditional on X1 (Performance).

Performance was significant when it was the only term in the model, but is not significant when Runtime is included. **This implies that the variables are correlated with each other.**

The significance level of the test does not depend on the order in which you list the independent variables in the MODEL statement, but it does depend upon the variables included in the MODEL statement.

Four common problems with regression are

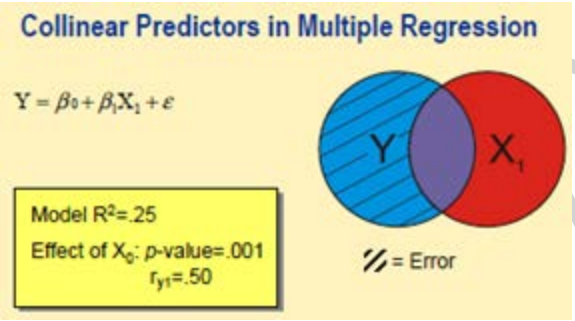
- nonconstant variance
- correlated errors
- influential observations
- **collinearity**(not a violation of assumptions of multiple regression)

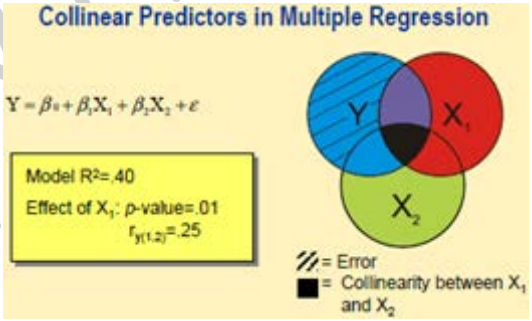


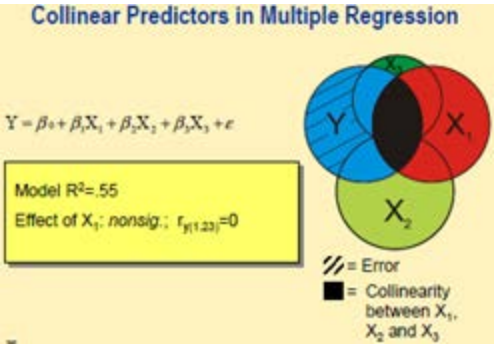
X1 and X2 almost follow a straight line X1 = X2 in the (X1, X2) plane. Consequently, one variable provides nearly as much information as the other does. They are redundant.

Why is this a problem? Two reasons exist.
1. Neither can appear to be significant when both are in the model; however, both can be significant when only one is in the model. Thus, collinearity can hide significant variables.
2. Collinearity also increases the variance of the parameter estimates and consequently increases prediction error.

When collinearity is a problem, the estimates of the coefficients are unstable. This means that they have a large variance. Consequently, the true relationship between Y and the Xs might be quite different from that suggested by the magnitude and sign of the coefficients.







The Venn diagram shows the variability of X and Y, and the extent to which variation in X explains variation in Y. The coefficient ry1 represents the correlation between Y and X1. Consider that the simple linear regression of Y on X1. X1 accounts for 25% of the variance in Y, as shown by the dark blue area of overlap.

You suspect that X2 is associated with Y and add it to the multiple regression model. However, X1 and X2 are correlated with one another. The coefficient ry(1.2) reflects the correlation of Y with X1, controlling for the variable X2. R2 increases when X2 is added to the model, but the individual effects of

X1 and X2 appear smaller because the effects tests are based on partial correlation. In other words, only the unique variance accounted for by each variable is reflected in the effect tests.

Add one more variable to the model, X3, that is correlated with X1. The coefficient ry(1.23) reflects the correlation between Y and X1 controlling for the variables X2 and X3. Notice that the independent effect of X1 is no longer statistically significant, as all the variance in Y accounted for by X1 is also accounted for by other predictors in the model. The R2 for this model has increased with each new term in the model, but the individual effects have decreased as terms are added to the model.

Model Building and Interpretation

The SELECTION= option in the MODEL statement of PROC REG supports these model selection techniques:

- All-possible regressions ranked using RSQUARE, ADJRSQ, or CP
- Stepwise selection methods STEPWISE, FORWARD, or BACKWARD

In the b_fitness data set, there are 7 possible independent variables. Therefore, there are 27 − 1=127 possible regression models. There are 7 possible one-variable models, 21 possible two-variable models, 35 possible three-variable models, and so on.

Mallows' C_p

- Mallows' C_p is a simple indicator of model bias. Models with a large C_p are biased.
- Look for models with $C_p \leq p$, where p equals the number of parameters in the model, including the intercept.

Mallows recommends choosing the first model where C_p approaches p.

Mallow's C_p is estimated by $$C_p = p + \frac{(\text{MSE}_p - \text{MSE}_{\text{full}})(n-p)}{\text{MSE}_{\text{full}}}$$

Hocking's Criterion

Hocking (1976) suggests selecting a model based on the following:

- $C_p \leq p$ for prediction
- $C_p \leq 2p - p_{\text{full}} + 1$ for parameter estimation

Automatic Model Selection (P187-190)

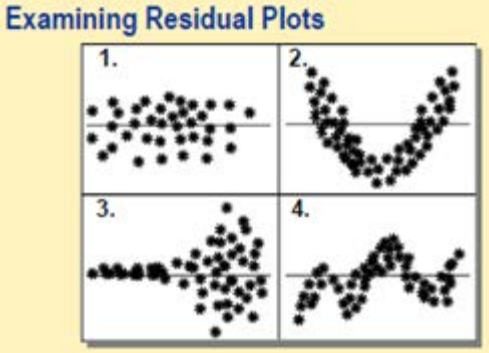
PROC REG also offers the following stepwise **SELECTION=** options:
FORWARD- first selects the best one-variable model. Then it selects the best two variables among those that contain the first selected variable. FORWARD continues this process, but stops when it reaches the point where no additional variables have a p-value level < 0.50.
BACKWARD- starts with the full model. Next, the variable that is least significant, given the other variables, is removed from the model. BACKWARD continues this process until all of the remaining variables have a p-value < 0.10.
STEPWISE- works like a combination of the two. The default entry p-value is 0.15 and the default stay p-value is also 0.15.

4. Regression Diagnostics

4.1 Examining Residuals

linear regression has the form Y=β0 + β1X + ε. When you perform a regression analysis, several assumptions about the error terms must be met to provide valid tests of hypothesis and confidence intervals. **The assumptions are that the error terms(Q40)**

- have a mean of 0 at each value of the predictor variable
- are normally distributed at each value of the predictor variable
- have the same variance at each value of the predictor variable
- are independent.

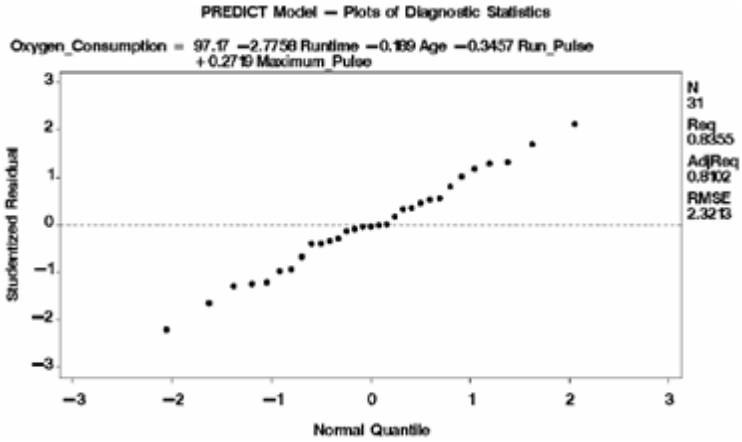


1. The model form appears to be adequate because the residuals are randomly scattered about a reference line at 0 and no patterns appear in the residual values.
2. The model form is incorrect. The plot indicates that the model should take into account curvature in the data. One possible solution is to add a quadratic term as one of the predictor variables.
3. The variance is not constant. As you move from left to right, the variance increases. One possible solution is to transform your dependent variable.
4. The observations are not independent. For this graph, the residuals tend to be followed by residuals with the same sign, which is called autocorrelation. This problem can occur when you have observations that have been collected over time. A possible solution is to use the AUTOREG procedure in SAS/ETS software.

Besides verifying assumptions, it is also important to check for outliers. Observations that are outliers are far away from the bulk of your data. These observations are often data errors or reflect unusual circumstances. In either case, it is good statistical practice to detect these outliers and find out why they have occurred.

One way to check for outliers is to use the studentized residuals. These are calculated by dividing the residual values by their standard errors. For a model that fits the data well and has no outliers, most of the studentized residuals should be close to 0. In general, studentized residuals that have an absolute value less than 2.0 could have easily occurred by chance. Studentized residuals that are between an absolute value of 2.0 to 3.0 occur infrequently and could be outliers. Studentized residuals that are larger than an absolute value of 3.0 occur rarely by chance alone and should be investigated

The plot of the normal quantiles versus the student residuals is shown below. The plot is obtained by plotting the student residuals against their expected quantiles if the residuals come from a normal distribution. If the residuals are normally distributed, the plot should appear to be a straight line with a slope of about 1. If the plot deviates substantially from the ideal, then there is evidence against normality.



You can use the NORMAL option in the UNIVARIATE procedure to generate a hypothesis test on whether the residuals are normally distributed. This could be necessary if you feel the plot above shows a violation of the normality assumption. First you must create an output data set with the residuals in PROC REG using an OUTPUT statement (as shown in Chapter 2 with an OUTPUT statement in the GLM procedure) or the Output Delivery System. Then use that data set as the input data set in PROC UNIVARIATE.

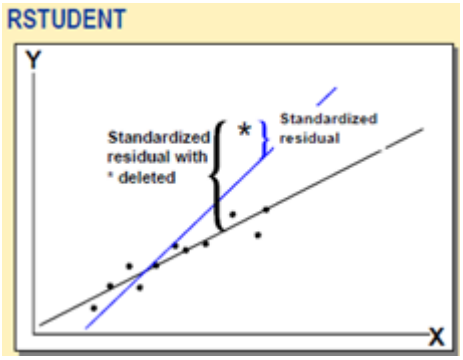
4.2 Influential Observations

Four diagnostic statistics that help identify influential observations are

- STUDENT residual
- **Cook's D(Q50, option C, Qextra, used to identify possible influential outliers)**
- RSTUDENT residual
- DFFITS.

The R option in the MODEL statement prints the first two statistics, as well as several others discussed previously. The INFLUENCE option in the MODEL statement prints the RSTUDENT and DFFITS statistics, as well as several others that are not discussed, such as the Hat Diagonal, Covariance Ratio, and DFBETAS.

Cook's D statistic is a measure of the simultaneous change in the parameter estimates when an observation is deleted from the analysis. A suggested cutoff is $D_i > \frac{4}{n}$, where n is the sample size. If the above condition is true, then the observation might have an adverse effect on the analysis.



Recall that STUDENT residuals are the ordinary residuals divided by their standard errors. The RSTUDENT residuals are similar to the STUDENT residuals except that they are calculated after deleting the ith

observation. In other words, the RSTUDENT is the difference between the observed Y and the predicted value of Y excluding this observation from the regression.

If the RSTUDENT is different from the STUDENT residual for a specific observation, that observation is likely to be influential.

DFFITS

DFFITS$_i$ measures the impact that the i^{th} observation has on the predicted value.

$$\text{DFFITS}_i = \frac{\hat{Y}_i - \hat{Y}_{(i)}}{s(\hat{Y}_i)}$$

$\hat{Y}_i$ is the i^{th} predicted value.

$\hat{Y}_{(i)}$ is the i^{th} predicted value when the i^{th} observation is deleted.

$s(\hat{Y}_i)$ is the standard error of the i^{th} predicted value.

Belsey, Kuh, and Welsch provide this suggested cutoff: $|\text{DFFITS}_i| > 2\sqrt{\frac{p}{n}}$, where p is the number of terms in the current model, including the intercept, and n is the sample size.

```
/* c4demo02a */
goptions reset=all;
proc reg data=sasuser.b fitness;
   PREDICT: model oxygen consumption
                    =runtime age run pulse maximum pulse
                    / r influence;
   id name;
   output out=ck4outliers
          rstudent=rstud dffits=dfits cookd=cooksd;
   title;
run;
quit;
```

```
Sum of Residuals                         0
Sum of Squared Residuals           140.10368
Predicted Residual SS (PRESS)      190.90531
```

The PRESS statistic is the sum of the PRESS residuals. These measure the deviation of the ith observation about the regression line formed when that observation is deleted from the analysis. In other words, it measures how well the regression model predicts the ith observation as though it were a new observation.

When the PRESS statistic is large compared to the Sum of the Squared Residuals, it indicates the presence of influential observations. The PRESS statistic is most useful when comparing several candidate models, such as comparing the PREDICT and EXPLAIN models that were examined earlier.

How to Handle Influential Observations

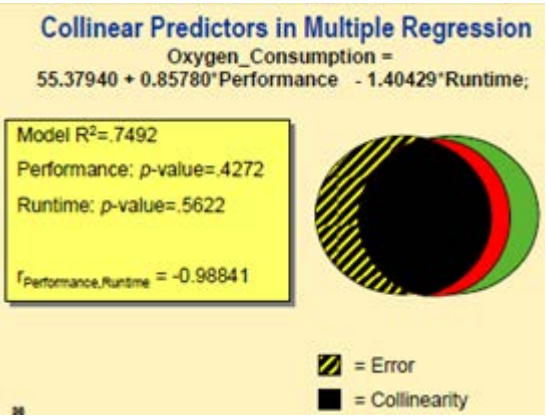
1. Recheck the data to ensure that no transcription or data entry errors have occurred.
2. If the data is valid, one possible explanation is that the model is not adequate.
   - A model with higher order terms, such as polynomials and interactions between the variables, might be necessary to fit the data well.

If more than one percent of the observations are identified as influential observations, it is possible that you do not have an adequate model; you may want to add higher-level terms, such as polynomial and interaction terms. **In general, do not exclude data.**

4.3 Collinearity

Objectives

- Determine if collinearity exists in a model.
- Generate output to evaluate the strength of the collinearity and what variables are involved in the collinearity.
- Determine methods to minimize collinearity in a model.



Collinearity can cause these problems in your model:

• truly significant terms can be hidden

• the variances of the coefficients are increased, which results in less precise estimates of the parameters and the predicted values

Collinearity is not a violation of the assumptions.

Example of Collinearity

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	722.03251	103.14750	18.32	<.0001
Error	23	129.52204	5.63139		
Corrected Total	30	851.55455			

Root MSE	2.37306	R-Square	0.8479
Dependent Mean	47.37581	Adj R-Sq	0.8016
Coeff Var	5.00900		

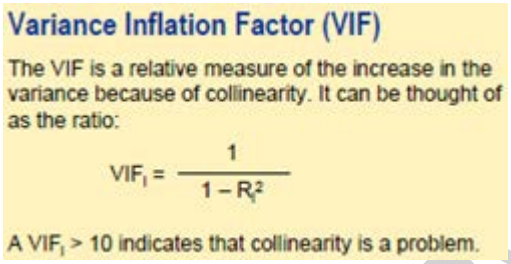
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	93.33753	36.49782	2.56	0.0176
Performance	1	0.25756	1.02373	0.25	0.8036
Runtime	1	-2.08804	2.22856	-0.94	0.3585
Age	1	-0.21066	0.10519	-2.00	0.0571
Weight	1	-0.07741	0.05681	-1.36	0.1862
Run_Pulse	1	-0.36618	0.12299	-2.98	0.0067
Rest_Pulse	1	-0.01389	0.07114	-0.20	0.8469
Maximum Pulse	1	0.30490	0.13990	2.18	0.0398

The Model F is highly significant and the R2 is large. These statistics suggest that the model fits the data well.
However, when you examine the p-values of the parameters, only Run_Pulse and Maximum_Pulse are statistically significant.
Recall that the PREDICT model included Runtime; however, in the full model, this same variable is not statistically significant (p-value=0.3585).
Including all the terms in the model hid at least one significant term.
When you have a significant Model F but no highly significant terms, collinearity is a likely problem.

Collinearity Diagnostics
PROC REG offers these tools that help quantify the magnitude of the collinearity problems and identify the subset of Xs that is collinear:

- **VIF: provides a measure of the magnitude of the collinearity (Variance Inflation Factor)(Q50, statements for detecting collinearity).**
- COLLIN: includes the intercept vector when analyzing the X'X matrix for collinearity. Condition index values between 10 and 30 suggest weak dependencies, between 30 and 100 indicate moderate dependencies, greater than 100 indicate strong collinearity.
- COLLINOINT: excludes the intercept vector.



You can calculate a VIF for each term in the model.
Marquardt suggests that a VIF > 10 indicates the presence of strong collinearity in the model (Qextra, given the output, decide how to tell collinearity).
VIFi = 1/(1 – Ri^2), where Ri^ 2 is the R2 of Xi, regressed on all the other Xs in the model.
For example, if the model is Y = X1 X2 X3 X4, i = 1 to 4. To calculate the R2 for X3, fit the model X3 = X1 X2 X4. Take the R2 from the model with X3 as the dependent variable and replace it in the formula VIF3 = 1/(1 – R3^2). If VIF3 is greater than 10, X3 is possibly involved in collinearity.

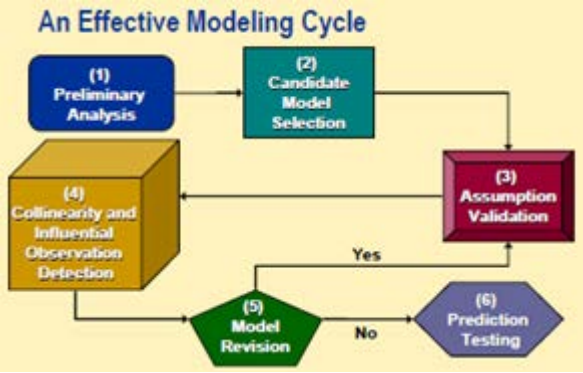
The COLLIN and COLLINOINT options calculate these types of statistics:
• eigenvalues: also called characteristic roots. Eigenvalues near zero indicate strong collinearity. A value λ is called an eigenvalue if there exists a nonzero vector z such that $(X'X)z = \lambda z$. The condition index, ηi, is the square root of the largest eigenvalue divided by λi.
• condition indices
• variance proportions. used in combination with the condition index can be used to identify the sets of Xs that are collinear. **Variance proportions greater than 0.50 indicate which terms are correlated.**
Variance proportions are calculated for each term in the model. The variance proportions for each term sum to 1.

```
/* c4demo04 */
proc reg data=sasuser.b fitness;
   FULLMODL:
   model oxygen consumption
                  = performance runtime age weight
                    run pulse rest pulse maximum pulse
                  / vif collin collinoint;
   title 'Collinearity -- Full Model';
run;
quit;
```

Guidelines for Eliminating Terms

1. Determine the set of Xs involved in collinearity using the variance proportions associated with the largest condition index (if it is greater than 100).
2. Drop the variable among the set with the largest p-value that also has a large VIF.
3. Rerun the regression and repeat, if necessary.

There are other approaches to dealing with collinearity. Two techniques are ridge regression and principle components regression. In addition, recentering the predictor variables can sometimes eliminate collinearity problems, especially in a polynomial regression.



(1) Preliminary Analysis This step includes the use of descriptive statistics, graphs, and correlation analysis.
(2) Candidate Model Selection This step uses the numerous selection options in PROC REG to identify one or more candidate models.
(3) Assumption Validation This step includes the plots of residuals and graphs of the residuals versus the predicted values. It also includes a test for equal variances.
(4) Collinearity and Influential Observation Detection. The former includes the use of the VIF statistic, condition indices, and variation proportions; the latter includes the examination of Rstudent residuals, Cook's D statistic, and DFFITS statistics.
(5) Model Revision. If steps (3) and (4) indicate the need for model revision, generate a new model by returning to these two steps.
(6) Prediction Testing. If possible, validate the model with data not used to build the model.

5. Categorical Data Analysis

5.1 Describing Categorical Data

	Type of Predictors		
Type of Response	**Categorical**	**Continuous**	**Categorical and Continuous**
Continuous	Analysis of Variance	Linear Regression	Analysis of Covariance (Regression with dummy variables)
Categorical	Logistic Regression or Contingency Tables	Logistic Regression	Logistic Regression

Nominal variables have values with no logical ordering. Gender is a nominal variable.
Ordinal variables have values with a logical order. However, the relative distances between the values are not clear. Income is an ordinal variable. Binary variables can also be considered ordinal variables.

By examining the distribution of categorical variables, you can

- screen for unusual data values
- determine the frequency of data values
- recognize possible associations among variables.

General form of the FREQ procedure;

```
PROC FREQ DATA=SAS-data-set;
    TABLES table-requests </ options>;
RUN;
```

If a continuous variable does not have a lot of values, as is the case with age in this data, then it is acceptable to use PROC FREQ. However, if age had numerous values, it would be better to use the UNIVARIATE procedure to explore this variable.

5.2 Tests of Association

Objectives

- Perform a chi-square test for association.
- Examine the strength of the association.
- Produce exact p-values for the chi-square test for association.
- Perform a Mantel-Haenszel chi-square test.(appropriate test for ordinal associations)

Chi-Square Test

NO ASSOCIATION

observed frequencies = expected frequencies

ASSOCIATION

observed frequencies ≠ expected frequencies

A commonly used test that examines whether there is an association between two categorical variables is the Pearson chi-square test. The chi-square test measures the difference between the observed cell frequencies and the cell frequencies that are expected if there is no association between the variables. If you have a significant chi-square statistic, there is strong evidence that an association exists between your variables.

p-Value for Chi-Square Test

The p-value is the

- probability of observing a chi-square statistic at least as large as the one actually observed, given that there is no association between the variables
- probability of the association you observe in the data occurring by chance.

In general, the larger the chi-square values, the smaller the p-value, which means that you have more evidence against the null hypothesis.

If you double the size of your sample by duplicating each observation, you double the chi-square statistic even though the strength of the association does not change.

One measure of the strength of the association between two nominal variables is Cramer's V statistic. It is in the range of –1 to 1 for 2-by-2 tables and 0 to 1 for larger tables. Values further away from 0 indicate the presence of a relatively strong association. Cramer's V statistic is derived from the Pearson chi-square statistic.

```
/* c5demo03 */
proc freq data=sasuser.b sales inc;
   tables gender*purchase
          / chisq expected cellchi2 nocol nopercent;
   format purchase purfmt.;
   title1 'Association between GENDER and PURCHASE';
run;
```

There are times when the chi-square test might not be appropriate. In fact, when more than 20% of the cells have expected cell frequencies of less than 5, the chi-square test might not be valid. This is because the p-values are based on the assumption that the test statistic follows a particular distribution when the sample size is sufficiently large. Therefore, when the sample sizes are small, the asymptotic (large sample) p-values might not be valid.

Observed versus Expected Values

Observed Values			Expected Values		
1	5	8	3.43	4.57	6.00
5	6	7	4.41	5.88	7.71
6	5	6	4.16	5.55	7.29

The criterion for the chi-square test is based on the expected values, not the observed values. In the slide above, 1 out of 9, or 11% of the cells, have observed values less than 5. However, 4 out of 9, or 44%, of the cells have expected values less than 5. Therefore, the chi-square test might not be valid.

The EXACT statement provides exact p-values for many tests in the FREQ procedure. Exact p-values are useful **when the sample size is small**, in which case the asymptotic p-values might not be useful. However, large data sets (in terms of sample size, number of rows, and number of columns) can require a prohibitive amount of time and memory for computing exact p-values. **For large data sets**, consider whether exact p-values are needed or whether **asymptotic p-values** might be quite close to the exact p-values. Exact p-values tend to be larger than asymptotic p-values because the exact tests are more conservative.

```
/* c5demo04 */
proc freq data=sasuser.b exact;
   tables a*b;
   exact pchi;
run;
```

Spearman versus Pearson (Qextra, the difference between Spearman and Pearson)

- The Spearman correlation uses ranks of the data.
- The Pearson correlation uses the observed values when the variable is numeric.

The Spearman statistic can be interpreted as the Pearson correlation between the ranks on variable X and the ranks on variable Y.

(Qextra, how to choose between Pearson and Spearman corrlelation)

If you want to explore your data it is best to compute both, since the relation between the Spearman (S) and Pearson (P) correlations will give some information. Briefly, **Spearman** is computed on ranks and so depicts **monotonic** relationships while **Pearson** is on true values and depicts **linear** relationships**(Qextra, Pearson correlation usage).**

As an example, if you set:

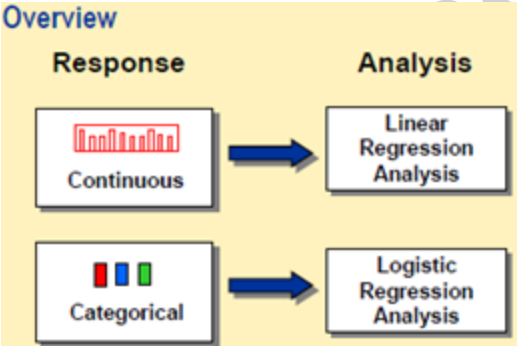
```
x=(1:100);
y=exp(x);                    % then,
corr(x,y,'type','Spearman');     % will equal 1, and
corr(x,y,'type','Pearson');      % will be about equal to 0.25
```

This is because y increases monotonically with x so the Spearman correlation is perfect, but not linearly, so the Pearson correlation is imperfect.

```
corr(x,log(y),'type','Pearson');  % will equal 1
```

Doing both is interesting because if you have S > P, that means that you have a correlation that is monotonic but not linear. Since it is good to have linearity in statistics (it is easier) you can try to apply a transformation on (such a log).

5.3 Introduction to Logistic Regression



Effect Coding: Two Levels

Class	Value	Design Variables 1
gender	Female	1
	Male	-1

Effect Coding: Three Levels

Class	Value	Label	Design Variables 1	2
inclevel	1	Low Income	1	0
	2	Medium Income	0	1
	3	High Income	-1	-1

Effect Coding: An Example

$$\text{logit}(p) = \beta_0 + \beta_1 * D_{\text{Low income}} + \beta_2 * D_{\text{Medium income}}$$

- β_0 = the average value of the logit across all categories
- β_1 = the difference between the logit for Low income and the average logit
- β_2 = the difference between the logit for Medium income and the average logit
- $-(\beta_1+\beta_2)$ = the difference between the average logit and the logit for High income

For effect coding (also called deviation from the mean coding), the number of design variables created is the number of levels of the CLASS variable minus 1. For example, because the variable inclevel has three levels, two design variables were created. For the last level of the CLASS variable (High), all the design variables have a value of –1. Parameter estimates of the CLASS main effects using this coding scheme estimate the difference between the effect of each level and the average effect over all levels. Because the sum of the deviations around the mean must equal zero, the effect for High income must be the negative of the sum of the effects for Low and Medium income.

Reference Cell Coding: Two Levels

Class	Value	Design Variables 1
gender	Female	1
	Male	0

Reference Cell Coding: Three Levels

Class	Value	Label	Design Variables 1	2
inclevel	1	Low Income	1	0
	2	Medium Income	0	1
	3	High Income	0	0

Reference Cell Coding: An Example

$$\text{logit}(p) = \beta_0 + \beta_1 * D_{\text{Low income}} + \beta_2 * D_{\text{Medium income}}$$

- β_0 = the value of the logit when income is High
- β_1 = the difference between the logits for Low and High income
- β_2 = the difference between the logits for Medium and High income

For reference cell coding, parameter estimates of the CLASS main effects estimate the difference between the effect of each level and the last level, called the reference level. For example, the effect for the level Low estimates the difference between Low and High. You can choose the reference level in the CLASS statement.

Binary Logistic Regression (P294)

Select purchase as the outcome variable and gender as the predictor variable. Specify reference cell coding and specify Male as the reference group. Also use the EVENT= option to model the probability of spending 100 dollars or more and request profile likelihood confidence intervals around the estimated odds ratios. FIRST designates the first ordered category as the event. LAST designates the last ordered category as the event. PARAM= specifies the parameterization method for the classification variable or variables.

```
/* c5demo06 */
proc logistic data=sasuser.b_sales_inc;
   class gender (param=ref ref='Male');
   model purchase(event='1')=gender / clodds=pl;
   title1 'LOGISTIC MODEL (1): purchase=gender';
run;
```

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	572.649	569.951
SC	576.715	578.084
-2 Log L	570.649	565.951

The Model Fit Statistics provides three tests: AIC is Akaike's 'A' information criterion, SC is the Schwarz criterion, and −2Log L is the −2 log likelihood. AIC and SC are goodness-of-fit measures you can use to compare one model to another. **Lower values indicate a more desirable model(Q47, which model is the champion model based on AIC or SBC statistic).** AIC adjusts for the number of predictor variables, and SCs adjust for the number of predictor variables and the number of observations. SC uses a bigger penalty for extra variables and therefore favors more parsimonious models.

Odds Ratio Calculation from the Current Logistic Regression Model

Logistic regression model:

$$\text{logit}(\hat{p}) = \log(odds) = \beta_0 + \beta_1 * (\text{gender})$$

Odds ratio (females to males):

$$\text{odds}_{\text{female}} = e^{\beta_0+\beta_1}$$

$$\text{odds}_{\text{male}} = e^{\beta_0}$$

$$\text{odds ratio} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

5.4 Multiple Logistic Regression

If you have a large number of variables, you might need to try a variable reduction method such as variable clustering. One way to eliminate unnecessary terms in a model is the backward elimination method. The significance level you choose depends on how much evidence you need in the significance of the predictor variables. In other words, the smaller your significance level, the smaller the p-value has to be to keep the predictor variable.

One major difference between a model with one predictor variable and a model with more than one predictor variable is that the reported odds ratios are now adjusted odds ratios.

Adjusted odds ratios measure the effect between a predictor variable and a response variable while holding all the other predictor variables constant.
For example, the odds ratio for the variable gender would measure the effect of gender on purchase while holding income and age constant.

The assumption is that the odds ratio for gender is the same regardless of the level of income or age. If that assumption is not true, you have an interaction.

Multiple Logistic Regression (P311)

When you use the backward elimination method with interactions in the model, PROC LOGISTIC begins by fitting the full model with all the main effects and interactions. PROC LOGISTIC then eliminates the nonsignificant interactions one at a time, starting with the least significant interaction (the one with the largest p-value). Next, PROC LOGISTIC eliminates the nonsignificant main effects not involved in any significant interactions. The final model should only have significant interactions, the main effects involved in the interactions, and the significant main effects.

For any effect that is in a model, all effects contained by that effect must also be in the model. This requirement is called model hierarchy. For example, if the interaction gender*income is in the model, then the main effects gender and income must also be in the model. This ensures that you have a hierarchically well-formulated model.

Qextra: knowing what Hierarchy principle is and how to use =single) For a more customized analysis, the HIERARCHY= option specifies whether hierarchy is maintained and whether a single effect or multiple effects are allowed to enter or leave the model in one step for forward, backward, and stepwise selection. The default is HIERARCHY=SINGLE, indicates that only one effect can enter or leave the model at one time, subject to the model hierarchy requirement. For example, supposed that you specify the main effects A and B and the interaction of A*B in the model. In the first step of the selection process, either A or B can enter the model. In the second step, the other main effect can enter the model. The interaction effect can enter the model only when both main effect can enter the model. Also, before A or B can be removed from the model, the A*B interaction must first be removed. All effects(CLASS and continuous variables) are subject to the hierarchy requirement.
Which of the following follows Hierarchy principle=single rule (Answer: A,D)
A.x+y
B.x*y
C.x+x*y
D.x+y+x*y

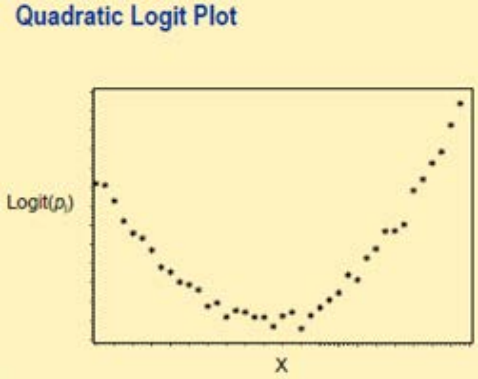
Multiple Logistic Regression with Interactions (P321, with backward elimination)
Example: Fit a multiple logistic regression model using the backward elimination method. In the MODEL statement, specify all the main effects and the two-factor interactions.

```
/* c5demo08 */
proc logistic data=sasuser.b sales inc;
   class gender (param=ref ref='Male')
         income (param=ref ref='Low');
   model purchase(event='1')=gender|age|income @2/ selection=backward;
   title1 'LOGISTIC MODEL (3): main effects and 2-way interactions';
   title2 '/ sel=backward';
run;
```

The bar notation with the @2 constructs a model with all the main effects and the two-factor interactions(Q54, model statements, | and @ go together, or none, @1 means no interactions). If you increased it to @3, then you would construct a model with all of the main effects, the two-factor

interactions, and the three-factor interaction. However, the three-factor interaction might be more difficult to interpret.

Comparing the goodness-of-fit statistics and the statistics that assess the predictive ability of the full model and the final model shows that the **full model has better predictive ability (because of the higher c statistic) and the final model has better goodness-of-fit statistics (because of the lower AIC and SC statistics).**



The logit plot can also show serious nonlinearities between the outcome variable and the predictor variable. The above graph reveals a quadratic relationship between the outcome and predictor variables. Adding a polynomial term or binning the predictor variable into three groups (two dummy variables would model the quadratic relationship) and treating it as a classification variable can improve the model fit.

5.5 Chapter Summary

Categorical data analysis deals with the analysis of categorical response variables, regardless of whether the explanatory variables are categorical or continuous. The scale of measurement of the variables is an important consideration when you decide the appropriate statistic to use. When you have two nominal variables, the Pearson chi-square statistic is appropriate. The strength of the association can be measured by Cramer's V. Because the Pearson chi-square statistic requires a large sample size, Fisher's exact test should be used to detect an association when you have a small sample size. When you have two ordinal variables, the Mantel-Haenszel chi-square statistic should be used to detect an ordinal association. The strength of the association can be measured by the Spearman correlation statistic.

(Q30, creating dummy variable)

Method 1: Method2: Method3:

```
data dummy_test1 (drop=i);
   set test;
   array inc(*) Inc_Group1 - Inc_Group5;
    do i = 1 to 5;
      inc(i)= (Inc_Group=i);
   end;
end;
```

```
data dummy_test1;
   set test;
   Inc_Group1=(Inc_Group1);
   Inc_Group2=(Inc_Group2);
   Inc_Group3=(Inc_Group3);
   Inc_Group4=(Inc_Group4);
   Inc_Group5=(Inc_Group5);
run;
```

```
data dummy_test1;
   set test;
   if Inc_Group1=1 then Inc_Group1=1;
    else Inc_Group1=0;
   if Inc_Group2=1 then Inc_Group2=1;
    else Inc_Group2=0;
   if Inc_Group3=1 then Inc_Group3=1;
    else Inc_Group3=0;
   if Inc_Group4=1 then Inc_Group4=1;
    else Inc_Group4=0;
   if Inc_Group5=1 then Inc_Group5=1;
    else Inc_Group5=0;
run;
```

The questions not accounted for
Q4, Q19 can't locate, Q22

The questions posted as extra: Qextra