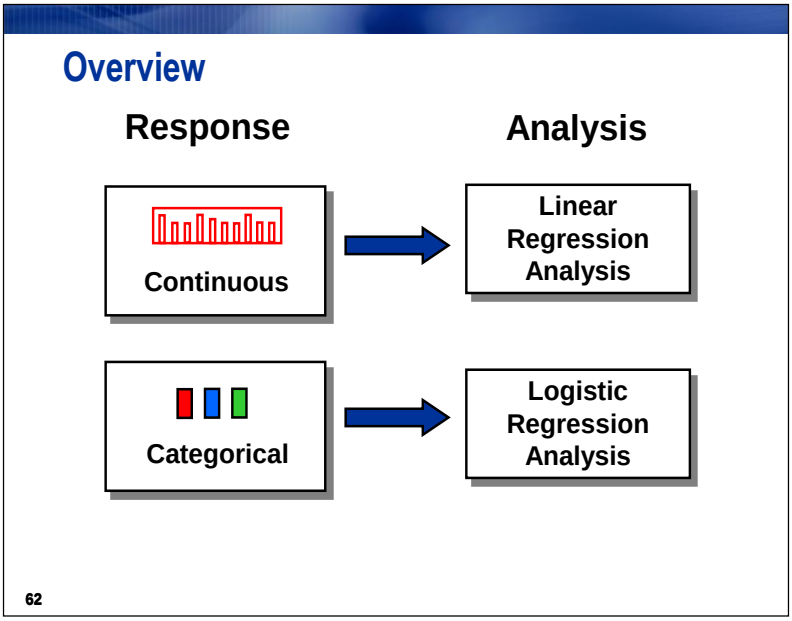
SASEG 10 - Logistic Regression

(Fall 2015)

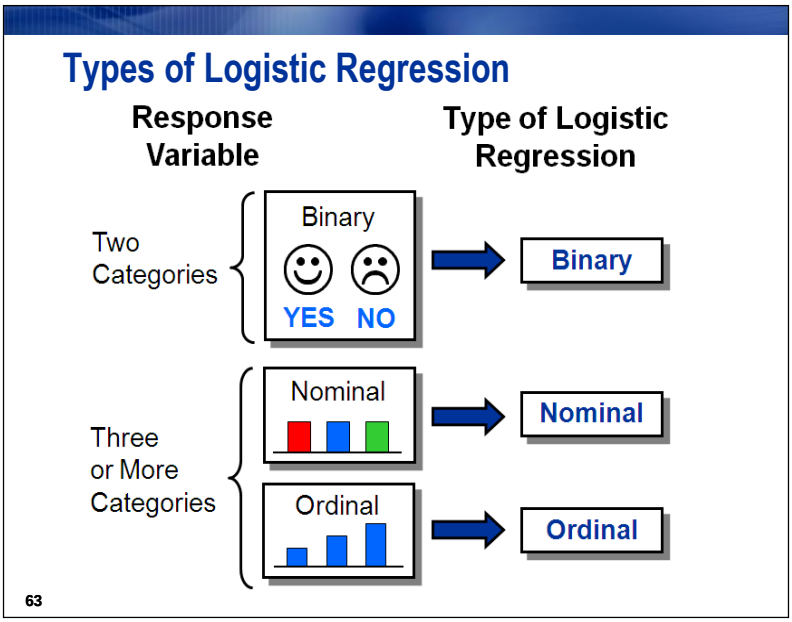
Sources (adapted with permission)-

T. P. Cronan, Jeff Mullins, Ron Freeze, and David E. Douglas Course and Classroom Notes
Enterprise Systems, Sam M. Walton College of Business, University of Arkansas, Fayetteville
Microsoft Enterprise Consortium
IBM Academic Initiative
SAS® Multivariate Statistics Course Notes & Workshop, 2010
SAS® Advanced Business Analytics Course Notes & Workshop, 2010
Microsoft® Notes
Teradata® University Network

Introduction to Logistic Regression



Regression analysis enables you to characterize the relationship between a response variable and one or more predictor variables. In linear regression, the response variable is continuous. In *logistic regression*, the response variable is categorical.



If the response variable is dichotomous (two categories), the appropriate logistic regression model is binary logistic regression.

If you have more than two categories (levels) within the response variable, then there are two possible logistic regression models:

- If the response variable is nominal, you fit a nominal logistic regression model.
- If the response variable is ordinal, you fit an ordinal logistic regression model.

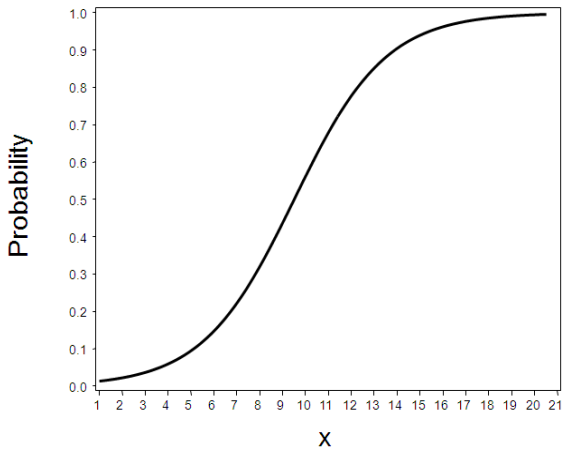
Why Not Ordinary Least Squares Regression?

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

- If the response variable is categorical, then how do you code the response numerically?
- If the response is coded (1=Yes and 0=No) and your regression equation predicts 0.5 or 1.1 or -0.4, what does that mean practically?
- If there are only two (or a few) possible response levels, is it reasonable to assume constant variance and normality?

64

Logistic Regression Curve



66

This plot shows a model of the relationship between a continuous predictor and the probability of an event or outcome. The linear model clearly will not fit if this is the true relationship between X and

the probability. In order to model this relationship directly, you must use a nonlinear function. One such function is displayed.

The parameter estimate of this curve determines the rate of increase or decrease of the estimated curve. When the parameter estimate is greater than 0, the probability of the outcome increases as the predictor variable values increase. When the parameter estimate is less than 0, the probability decreases as the predictor variable values increase. As the absolute value of the parameter estimate increases, the curve has a steeper rate of change. When the parameter estimate is equal to 0, the curve can be represented by a straight, horizontal line that shows an equal probability of the event for everyone.

The β values cannot be computed in the Linear Regression task. This is not a general linear model.

Logit Transformation

Logistic regression models transform probabilities called logits*.

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{(1-p_i)}\right)$$

where

i	indexes all cases (observations)
p_i	is the probability the event (a sale, for example) occurs in the i^{th} case
ln	is the natural log (to the base e).

* The logit is the natural log of the odds.

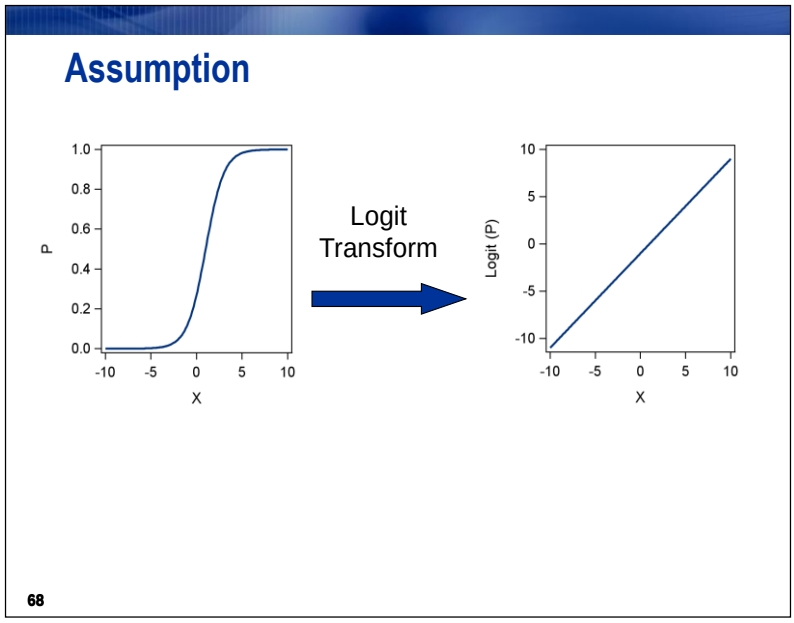
67

A logistic regression model applies a logit transformation to the probabilities.

First, deal with the problem of restricted range of the probability. What about the range of a logit? As p approaches its maximum value of 1, the value ln (p / (1 – p)) approaches infinity. As p approaches its minimum value of 0, p / (1 – p) approaches 0. The natural log of something approaching 0 is something approaching negative infinity. So, the logit has no upper or lower bounds.

If you can model the logit, then simple algebra will allow you to model the odds or the probability. The logit transformation ensures that the model generates estimated probabilities between 0 and 1.

The logit is the natural log of the odds. The odds and odds ratios were discussed in a previous section. This relationship between the odds and the logit will become important later in this section.



Assumption in logistic regression: The logit transformation of the probabilities results in a linear relationship with the predictor variables.

If the thoughts about the nature of the direct relationship between X and p are correct, then the logit will have a straight line relationship with X. In other words, a linear function of X can be used to model the logit. In that way, you can indirectly model the probability.

To verify this assumption, it would be useful to plot the logits by the predictor variable. Logit plots are illustrated in a later section.

Logistic Regression Model

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

where

$\text{logit}(p_i)$= logit of the probability of the event

β_0= intercept of the regression equation

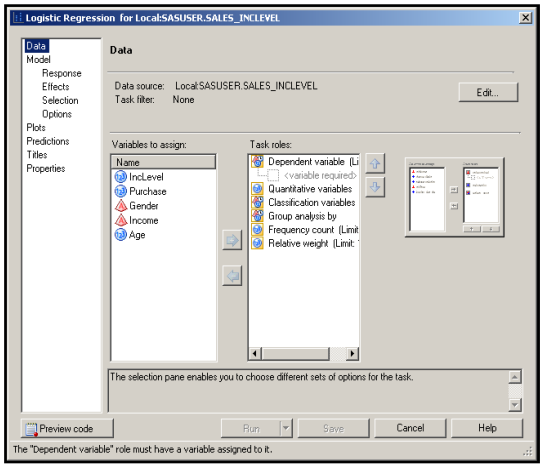
β_k= parameter estimate of the k^{th} predictor variable

69

For a binary outcome variable, the linear logistic model with one predictor variable has the form above.

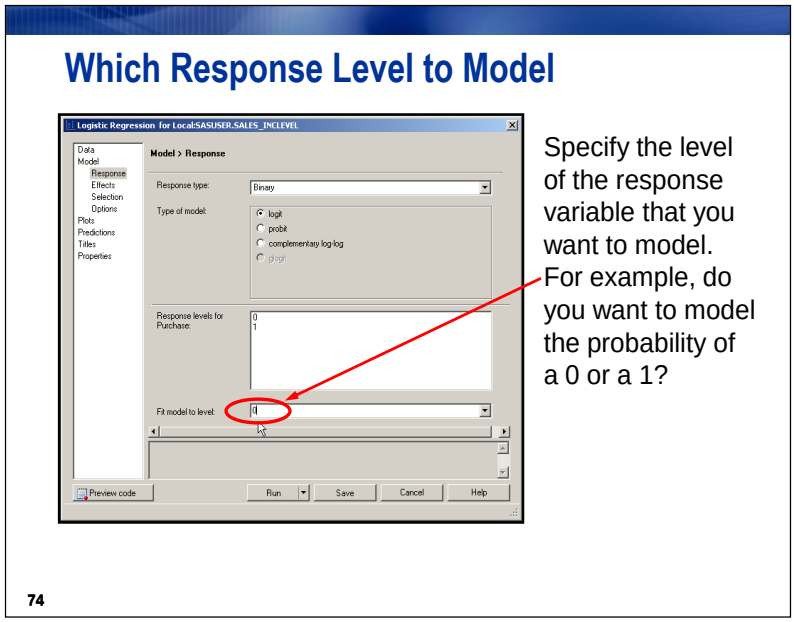
Unlike linear regression, the logit is not normally distributed and the variance is not constant. Also, logistic regression usually requires a more complex estimation method called maximum likelihood to estimate the parameters than linear regression. This method finds the parameter estimates that are most likely to occur given the data. This is accomplished by maximizing the likelihood function that expresses the probability of the observed data as a function of the unknown parameters.

The Logistic Regression Task



73

In the Logistic Regression task, you specify the proposed relationship between the categorical dependent variable and the independent variables.

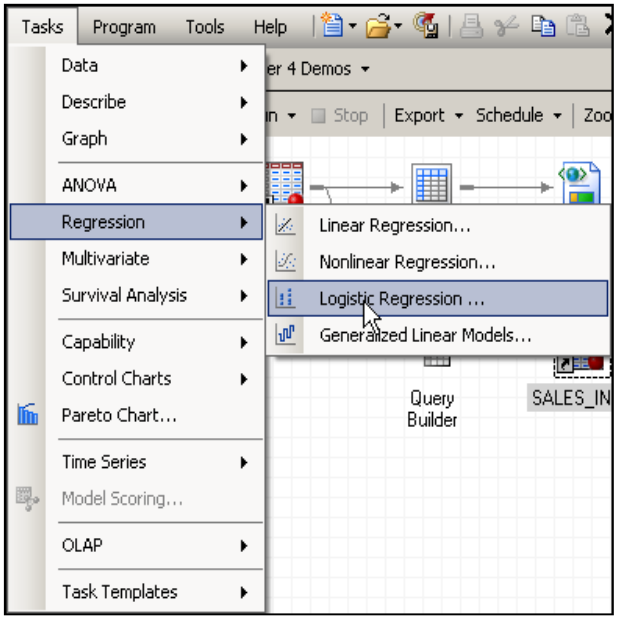


Be careful to pay attention to which level of the response variable that you would like to model. That might prevent you from accidentally reporting an effect that is exactly opposite to the one that you had thought that you were modeling. Modeling the probability of a 0 is the same as modeling the probability of **not** a 1 for a binary response variable.

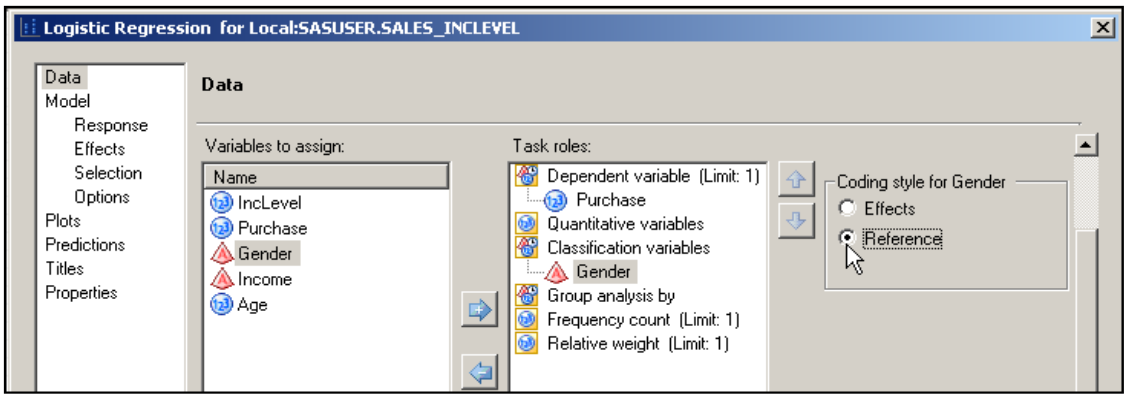
Binary Logistic Regression

Fit a binary logistic regression model. 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 `Fit model to level` **1** option to model the probability of spending 100 dollars or more and request profile likelihood confidence intervals around the estimated odds ratios.

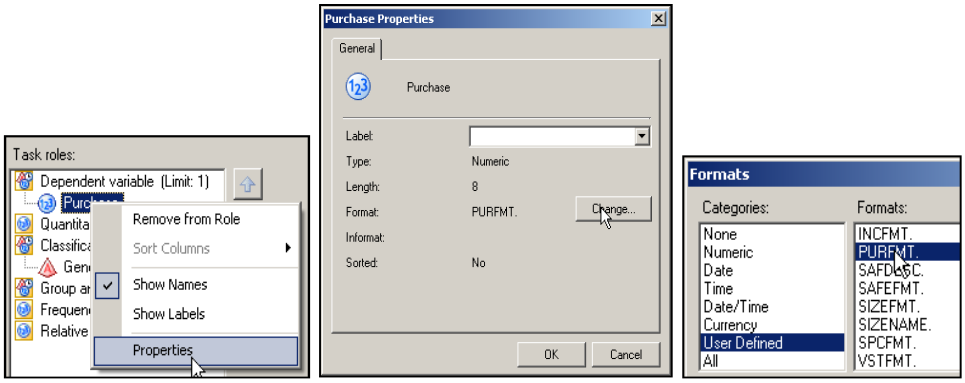
1. Open the **SALES_INCLEVEL** data set. Select **Tasks** ⇨ **Regression** ⇨ **Logistic Regression…**.



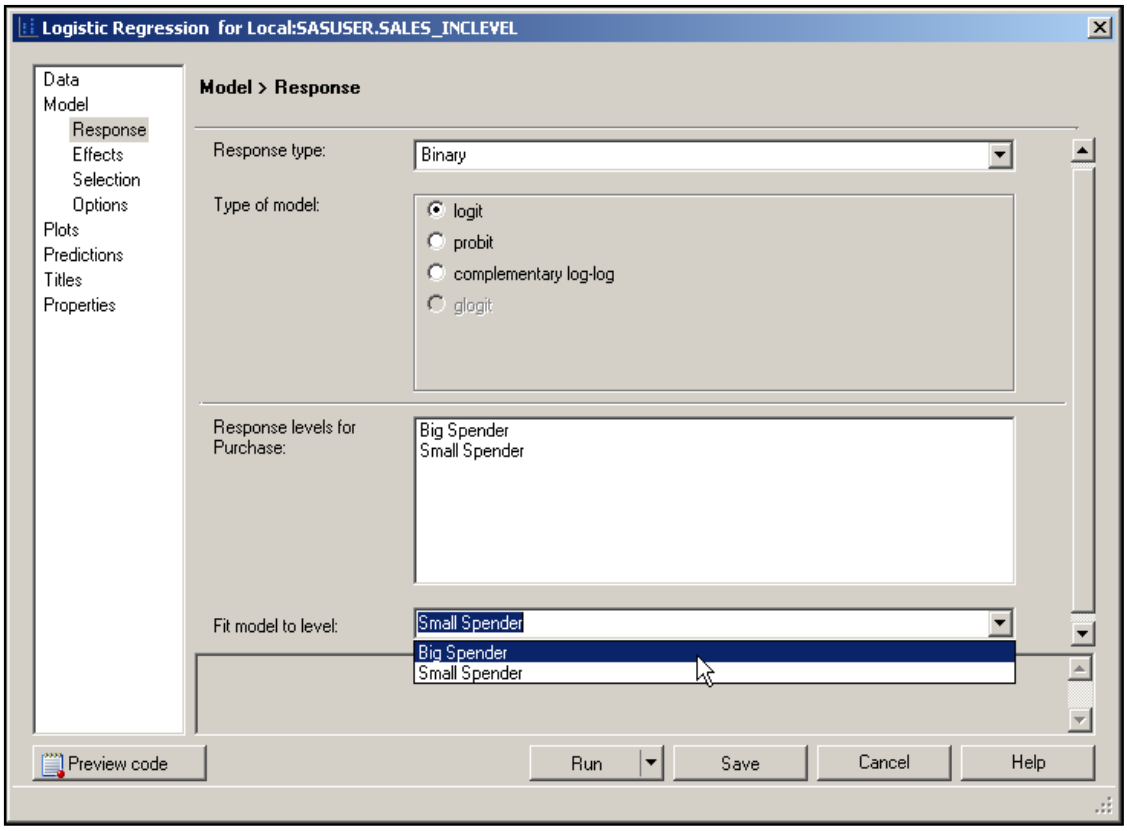
2. Assign **Purchase** to the dependent variable task role and **Gender** to the classification variables role and check select **Reference** as the coding style for **Gender**. `Male` will be the default reference level because it is the highest in alphanumeric order.



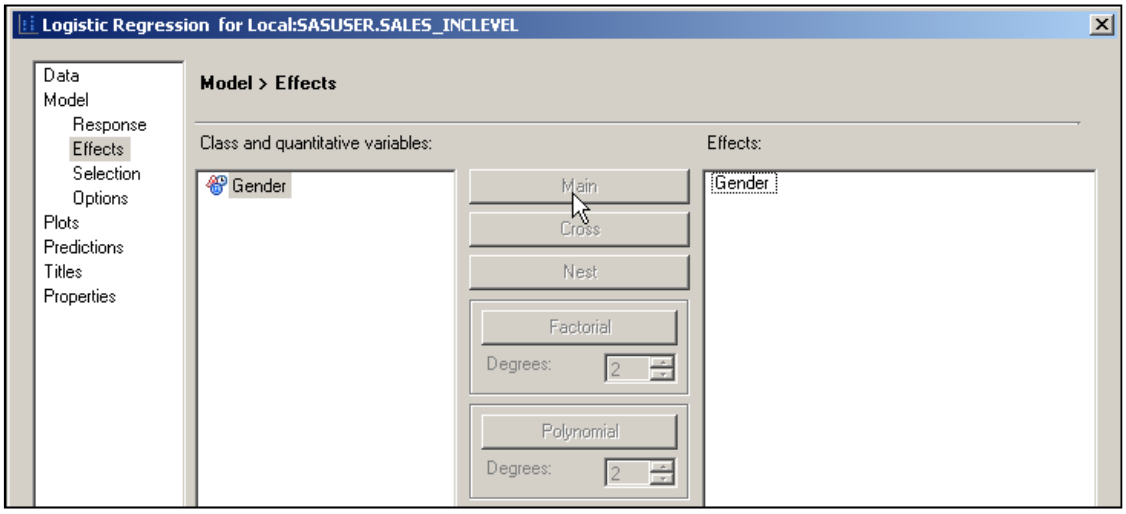
3. Right-click **Purchase** to change properties. Click the **Change...** button. In the window that appears, select User Defined from the **Categories:** section, and assign the PURFMT format from the **Formats:** section. Click **OK** twice to return to the main screen.



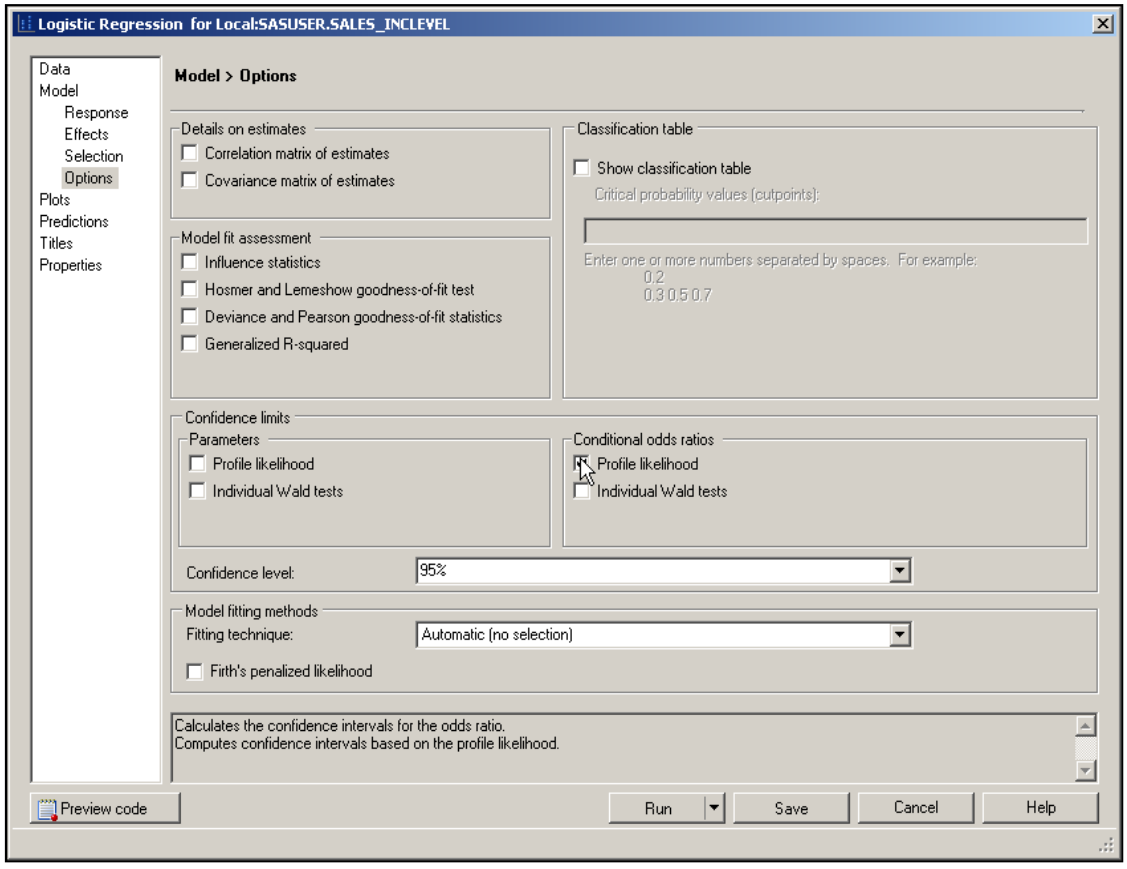
4. With **Response** selected at the left, assure that the value of `Fit model to level` is set to **`1 – Big Spender`**. Note that `Response levels for Purchase` displays the possible values in the data set to fit.



5. With **Effects** selected at the left, add **`Gender`** as a [Main] effect in the model.

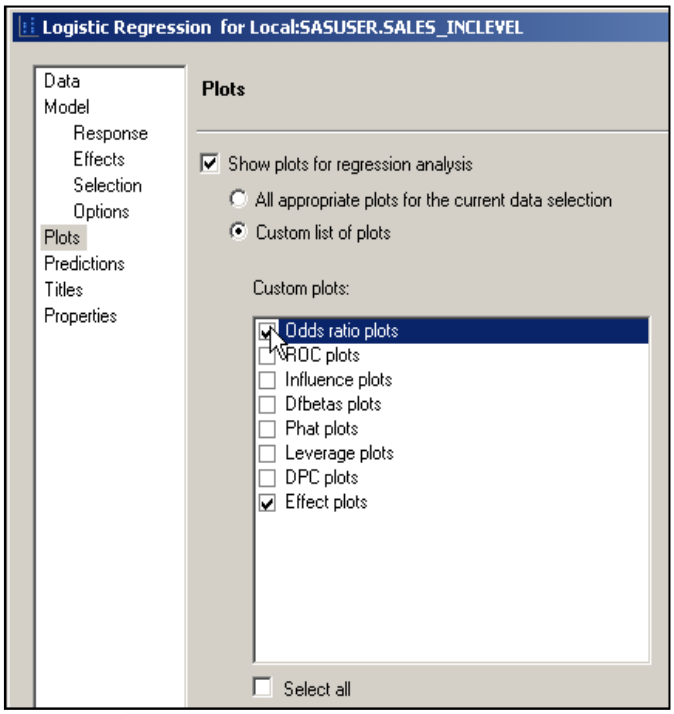


6. With **Options** selected at the left, check the box for **Profile likelihood** under `Conditional odds ratios` in the middle of the window.



7. With **Plots** selected at the left, select **Custom list of plots**. Then check only the boxes next to **Odds ratio plots** and **Effect plots** in the list.

8. Click Run.



The **Model Information** table describes the data set, the response variable, the number of response levels, the type of model, the algorithm used to obtain the parameter estimates, and the number of observations read and used.

The **Response Profile** table shows the response variable values listed according to their ordered values. By default, the Logistic Regression task orders the response variable alphanumerically so that it bases the logistic regression model on the probability of the smallest value. Because you specified it in the task window in this example, the model is based on the probability of purchasing items of 100 dollars or more (**Purchase**=1).

The **Response Profile** table also shows the value of the response variable and the frequency.

The **Class Level Information** table includes the predictor variable that was assigned as a classification variable. Because you used reference cell coding and `Male` comes last in alphabetical order, this table reflects **Gender**=`Male` as the reference level. The design variable is 1 when **Gender**=`Female` and 0 when **Gender**=`Male`.

Fisher's scoring algorithm converged to a solution. This message should always be checked before moving on. There are a number of options to control the convergence criterion, but the default is the gradient convergence criterion with a default value of 1E-8 (0.00000001).

Logistic Regression Results

The LOGISTIC Procedure

Model Information	
Data Set	WORK.SORTTEMPTABLESORTED
Response Variable	Purchase
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	431
Number of Observations Used	431

Response Profile		
Ordered Value	Purchase	Total Frequency
1	Big Spender	162
2	Small Spender	269

Probability modeled is Purchase='Big Spender'.

Class Level Information		
Class	Value	Design Variables
Gender	Female	1
	Male	-1

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

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. AIC adjusts for the number of predictor variables, and SC adjusts 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.

A reference for AIC can be found in Findley and Parzen (1995).

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

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	4.6978	1	0.0302
Score	4.6672	1	0.0307
Wald	4.6436	1	0.0312

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Gender	1	4.6436	0.0312

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.5380	0.1015	28.1144	<.0001
Gender	Female	1	0.2186	0.1015	4.6436	0.0312

The **Testing Global Null Hypothesis**: BETA=0 table provides three statistics to test the null hypothesis that all regression coefficients of the model are 0.

Using the Likelihood Ratio test, a significant *p*-value for the Likelihood Ratio test provides evidence that at least one of the regression coefficients for an explanatory variable is nonzero (in this example the *p*-value is 0.0302, which is significant at the .05 level). **This statistic is similar to the overall *F* test in linear regression.** The Score and Wald tests are also used to test whether all the regression coefficients are 0. The likelihood ratio test is the most reliable, especially for small sample sizes (Agresti 1996).

The **Type 3 Analysis of Effects** table is generated when a predictor variable is used as a classification variable. The listed effect (variable) is tested using the Wald Chi-Square statistic (in this example, 4.6436 with a *p*-value of 0.0312). This analysis is in the Linear Regression task. Because **`Gender`** is the only variable in the model, the value listed in the table will be identical to the Wald test in the Testing Global Null Hypothesis table.

The **Analysis of Maximum Likelihood Estimates** table lists the estimated model parameters, their standard errors, Wald tests, and odds ratios.

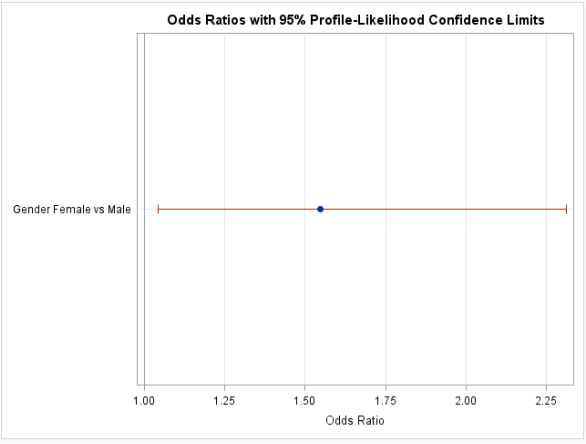
The parameter estimates are the estimated coefficients of the fitted logistic regression model. The logistic regression equation is logit($\hat{P}$) = –0.7566 + 0.4373***Gender**, for this example.

The Wald chi-square, and its associated *p*-value, tests whether the parameter estimate is significantly different from 0. For this example, both the *p*-values for the intercept and the variable **`Gender`** are significant at the 0.05 significance level. As in linear regression, hypothesis testing of the intercept term is uncommon.

The **Odds Ratio Estimates** table shows that females have odds 1.549 times those of males of making a purchase. This table will be described further.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	30.1	Somers' D	0.107
Percent Discordant	19.5	Gamma	0.215
Percent Tied	50.4	Tau-a	0.050
Pairs	43578	c	0.553

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
Gender Female vs Male	1.0000	1.549	1.043	2.312

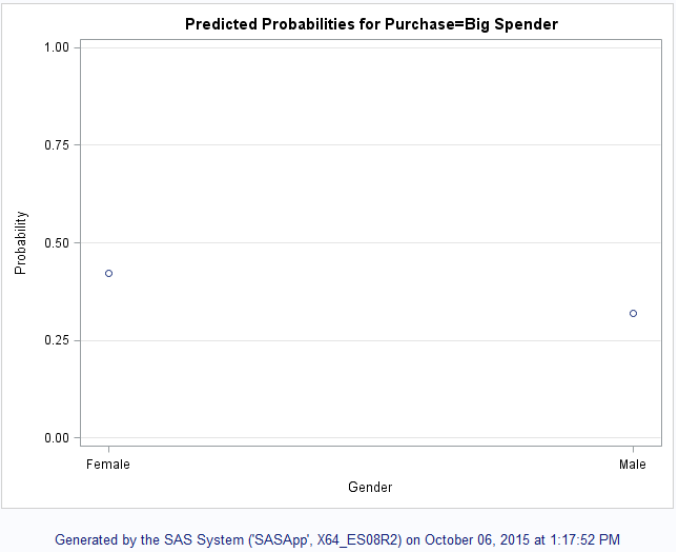


This table assesses the ability of the model (the **Gender** variable) to discriminate `Big Spenders` from `Small Spenders`. This table will be described in full detail.

The **Profile Likelihood Confidence Intervals** are often preferred to the Wald, especially for relatively small sample sizes. In this case, the differences are very small.

The **Odds Ratios** plot is for the Profile-Likelihood Confidence Limits. The 95% confidence interval does not cross the reference line at 1. That is to be expected because the gender effect is statistically significant at the 0.05 alpha level.

The **Effect** plot shows the difference between levels of the gender on the probability scale.



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):

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

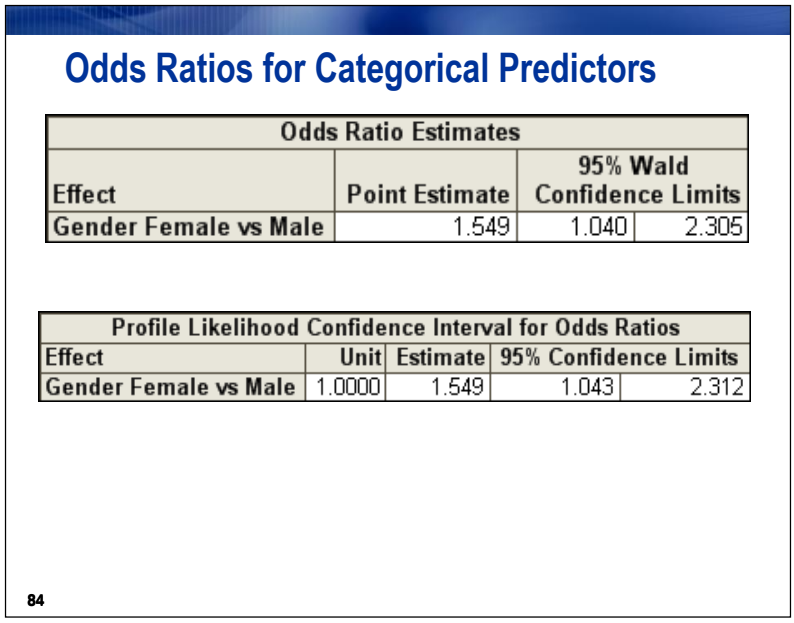
$$\text{odds}_{\text{males}} = e^{\beta_0}$$

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

83

Remember that in logistic regression you model the natural log of the odds and not the odds or probability directly. For interpretation, often the parameter estimates are converted into something more interpretable – an odds ratio. In order to understand this, write out the linear model predicting the natural log of the odds. In order to see that in terms of odds, the natural log is "undone" by exponentiation. Exponentiation of the right side of the equation must also be done to maintain equality. You thereby can look at the model in terms of odds and can estimate odds for females or males. The odds ratio is then the ratio of the odds of one group to the odds of another group.

The odds ratio reported by the Logistic Regression task is for a 1-unit difference for a variable. Because you used reference cell coding for **Gender** and used `Male` as the reference level, females are coded 1 and males are coded 0. Therefore, a 1-unit increase in **Gender** corresponds to the difference between females and males.

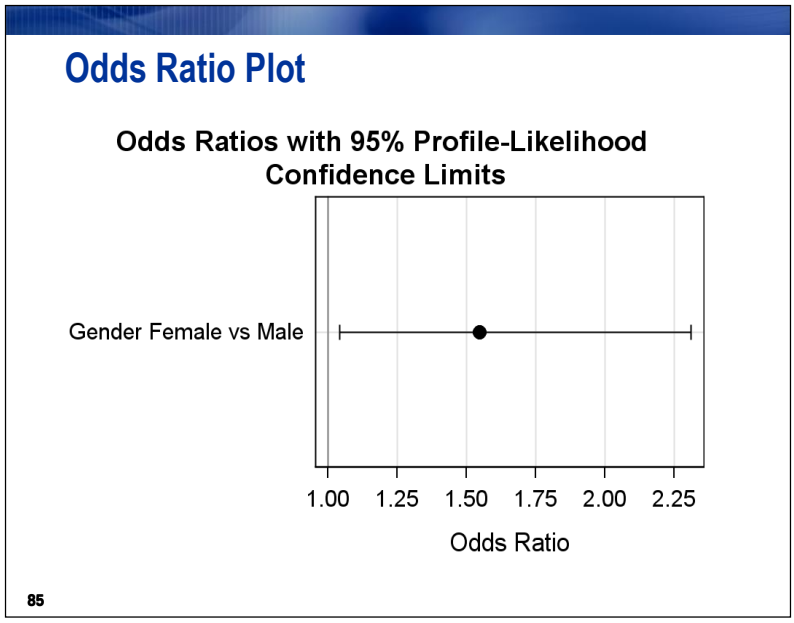


The odds ratio indicates that females have 1.549 times the odds to purchase 100 dollars or more, relative to males.

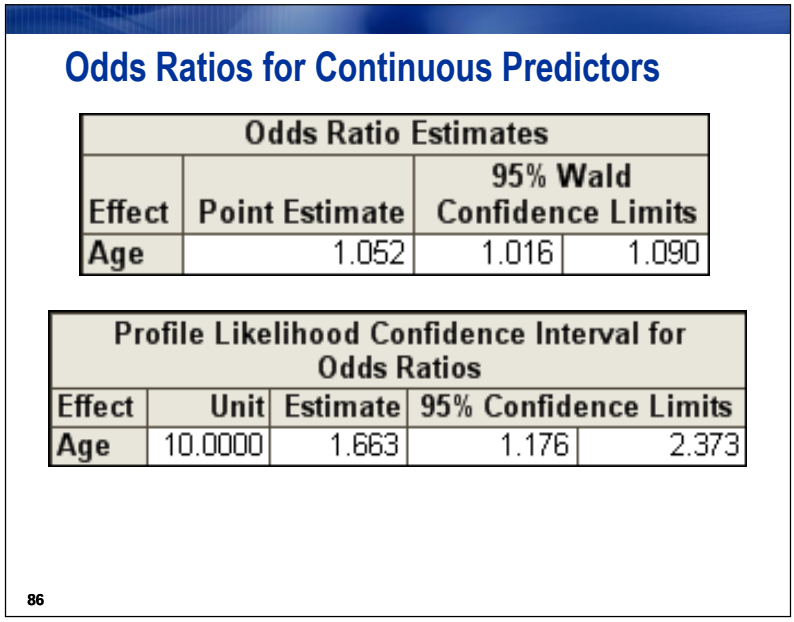
The 95% confidence limits indicate that you are 95% confident that the true odds ratio is between 1.04 and 2.31. Because the 95% confidence interval does not include 1.00, the odds ratio is significant at the .05 significance level.

If you want a different significance level for the confidence intervals, you can change the values of the confidence intervals in the options panel of the Logistic Regression task.

The profile likelihood confidence intervals are different from the Wald-based confidence intervals. This difference is because the Wald confidence intervals use a normal approximation, whereas the profile likelihood confidence intervals are based on the value of the log-likelihood. These likelihood-ratio confidence intervals require much more computation but are generally preferred to the Wald confidence intervals, especially for sample sizes less than 50 (Allison 1999).



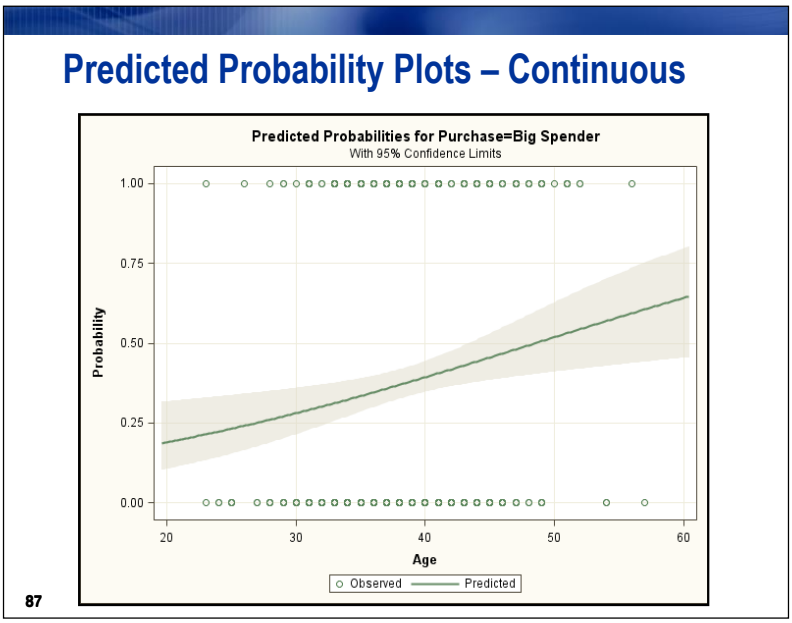
The odds ratio plot displays the odds ratio and confidence interval based on the method chosen in the Options panel.



For a continuous predictor variable, the default odds ratio table measures the increase or decrease in odds associated with a one-unit difference on the predictor variable. For example, **Age** shows an odds ratio of 1.052, which means that a person who is one year older has 5.2% ((1.052 – 1.000) * 100%) greater odds of purchasing $100 or more of items from the catalog than the younger person. The model assumes that this odds ratio is the same across all ages, so it does not matter if you compare a 21-year-old with a 20-year-old or a 35-year-old with a 34-year-old. Notice that the confidence interval for **Age** does not include 1, which corroborates the conclusion of significance from the *p*-value.

If you additionally choose a conditional odds ratio method for confidence limits in the Options panel, you might choose units other than 1 for calculating odds ratios. For example, a 10-unit difference in age is associated with a 1.663 odds ratio, meaning that for any 10-year difference in age, the odds of being a big

spender is multiplied by 1.663 for the older person, compared with the younger person. Another way of expressing that is that the odds for the 10-year-older person are 66.3% greater compared with the odds for the younger person. The 10-unit odds ratio could be calculated by hand by just raising the 1-unit odds ratio of 1.052 to the 10^{th} power. 1.052^{10} =~ 1.663 after rounding.



Where there is one continuous variable in the model, ODS Statistical Graphics can produce a plot of the modeled relationship between the continuous predictor and the response probability.

Model Assessment: Comparing Pairs

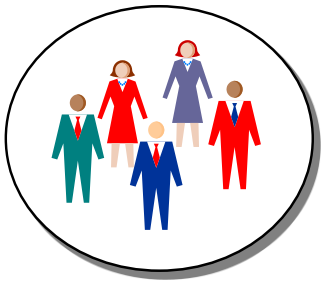
- Counting concordant, discordant, and tied pairs is a way to assess how well the model predicts its own data and therefore how well the model fits.
- In general, you want a high percentage of concordant pairs and low percentages of discordant and tied pairs.

88

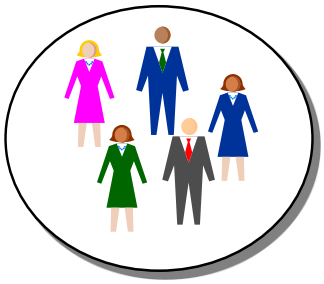
Comparing Pairs

To find concordant, discordant, and tied pairs, compare everyone who had the outcome of interest against everyone who did not.

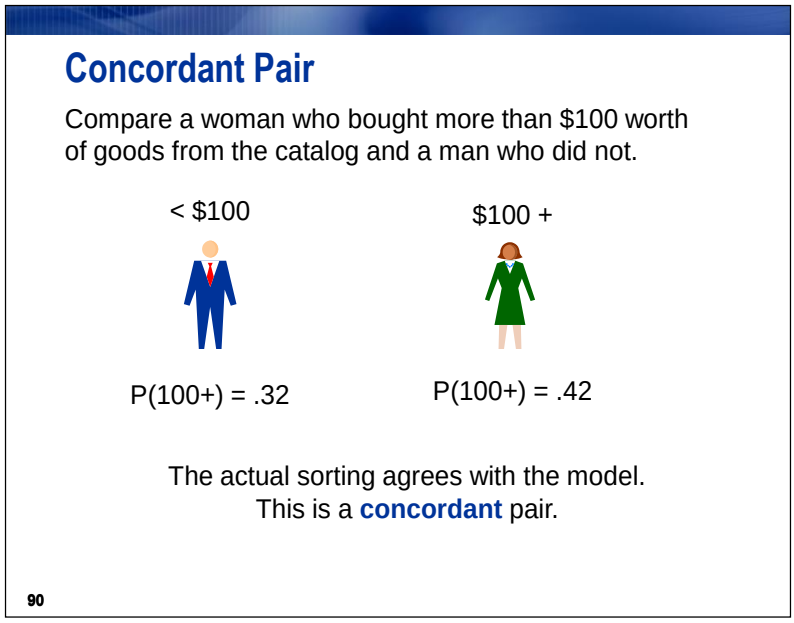
< $100



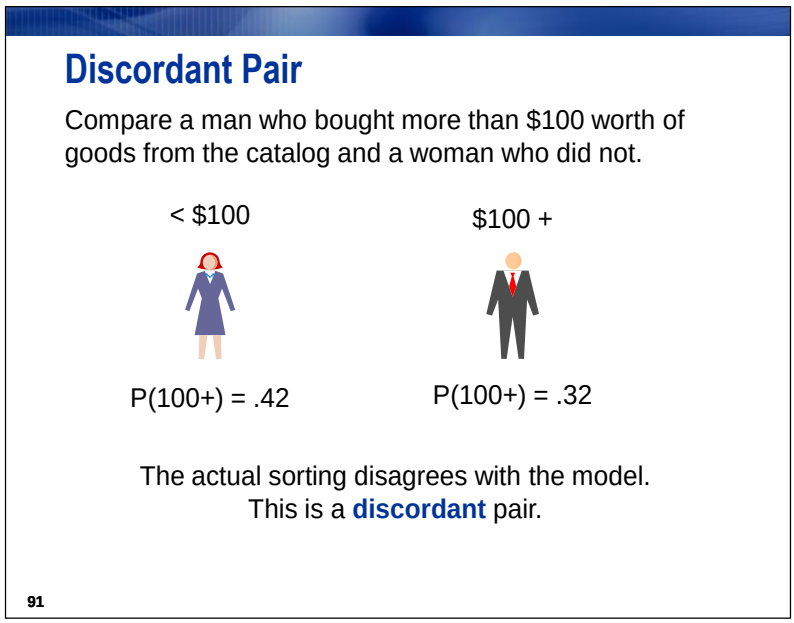
$100 +



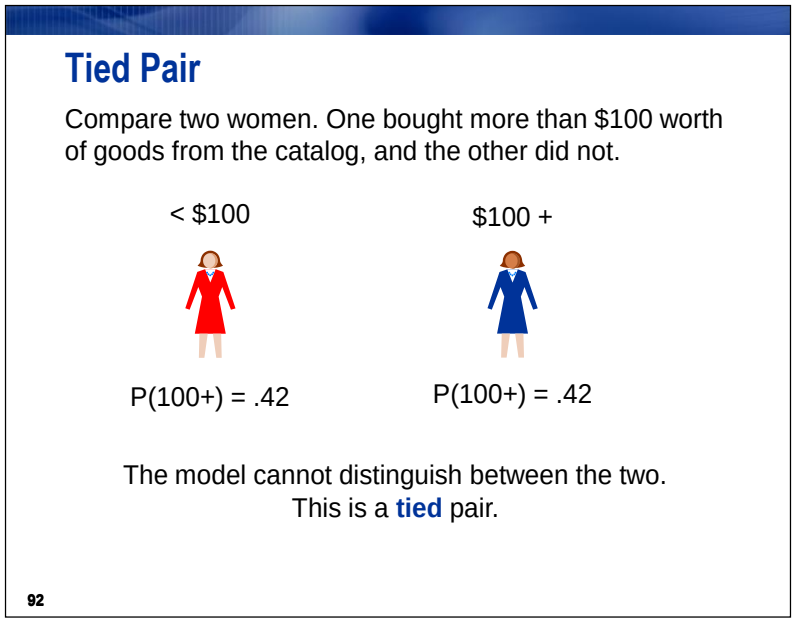
89



For all pairs of observations with different values of the response 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.



A pair is *tied* if it is neither concordant nor discordant (the probabilities are the same).