

## Lecture 21 Logistic Regression STAT102

- **Logistic Regression** provides analysis resembling usual regression (ordinary or multiple), but for cases in which the  $Y$  variable is categorical (nominal) rather than continuous.
- We will study only the most common setting, in which the  $Y$  variable is binary, and hence takes only two values.
  - They may be labeled by category names such as “dead” & “alive” or “good” & “bad”, etc., Or
  - They may be notated by the numbers “0” and “1” acting as labels for the two categories.
- As with regression we begin with situations having one continuous  $X$ -variable that can be used to predict or “explain” the dependent  $Y$ -variable.
- This topic is presented in *Dielman*, Section 10.3 (p383-388).

## The Dependent Variable, $Y$

- Can take either of only two values – Say  $\mathbf{0}$  or  $\mathbf{1}$ .
- It is assumed that the probability that each  $Y = \mathbf{0}$  depends only on the corresponding value,  $x$ , of the independent variable (or on the values of the independent variables  $x_1, \dots, x_p$  in a multiple regression situation).
- Its distribution at a given value of  $x$  is described by the probability  $p(x)$  that it takes this specified value, say  $\mathbf{0}$ .
- Thus, we can then write  $p(x) = \Pr\{Y = \mathbf{0}|x\}$ .
- The values of  $Y$  are assumed to be independent of each other.

## Example: “Loan10”

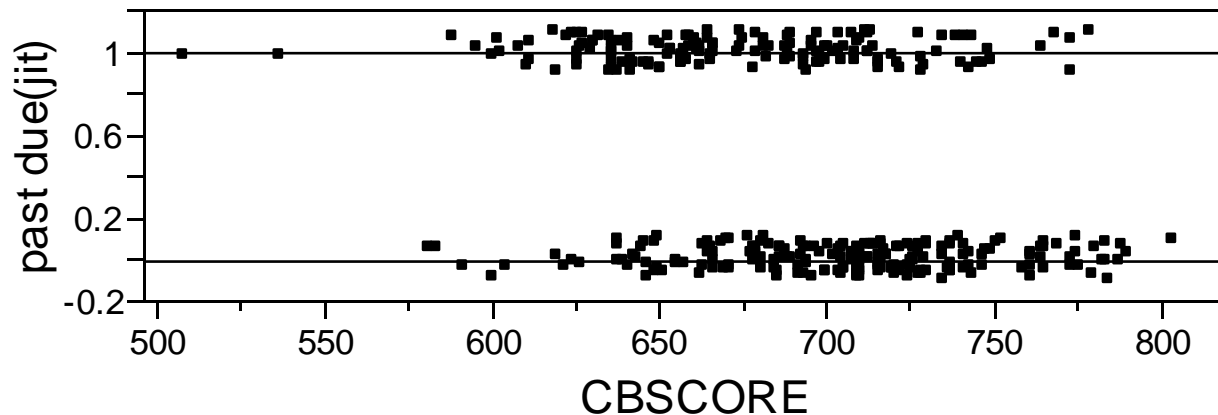
- This data is taken from *Dielman*, problem 10.6.
- It involves the outcome of a sample of a class of consumer loans.
- The borrowers are ranked on a standard “CBScore” credit score before receiving their loan.
- $x$  = the value of this score. (Higher values describe better credit risks.)
- The outcome variable,  $Y$ , describes whether the customer is past due after 2 years.
  - $Y=1$  means the customer’s payments are past due (“Bad”)
  - $Y=0$  means the customer’s payments are on time (“Good”)
- We want to study how accurately the value of  $x$  describes the probability that the customer’s payments will be on time.

## A Note About JMP's Slightly Nutty Notation

- $Y$  has 2 categories.
- The function  $p(\mathbf{x})$  denotes the probability of occurrence of a particular category.
- JMP needs to decide which category  $p(\mathbf{x})$  relates to.
- If the labels for  $Y$  are “0” and “1” then JMP sets
$$p(\mathbf{x}) = \Pr\{Y = \mathbf{0} | x\}.$$
- If the labels are alphabetic names then JMP organizes them in alphabetic order (*ie*, “Bad” then “Good”), and  $p(\mathbf{x})$  refers to the first of these (*ie*,  $p(\mathbf{x}) = \Pr\{Y = \mathbf{Bad} | x\}$ ).
- Nearly every other author and statistical software makes the opposite choice; BUT we'll stick with JMP's system, so that the JMP output will be easier to read.

## Plotting the Data

- Scatterplots with  $Y=0/1$  are less satisfactory than for continuous  $Y$ .
- Here is a possible version of such a plot for our data. The  $Y$  values have been jittered so they can be seen individually.



- It is fairly clear here that as CBScore increases there is an increasing probability of scores of **0** relative to scores of **1**.
- This corresponds to  $p(x) = \Pr\{Y = 0|x\}$  being an **increasing function** of CBScore,  $x$ .

## The Logistic Function

- If  $\theta$  is any number the logistic function at  $\theta$

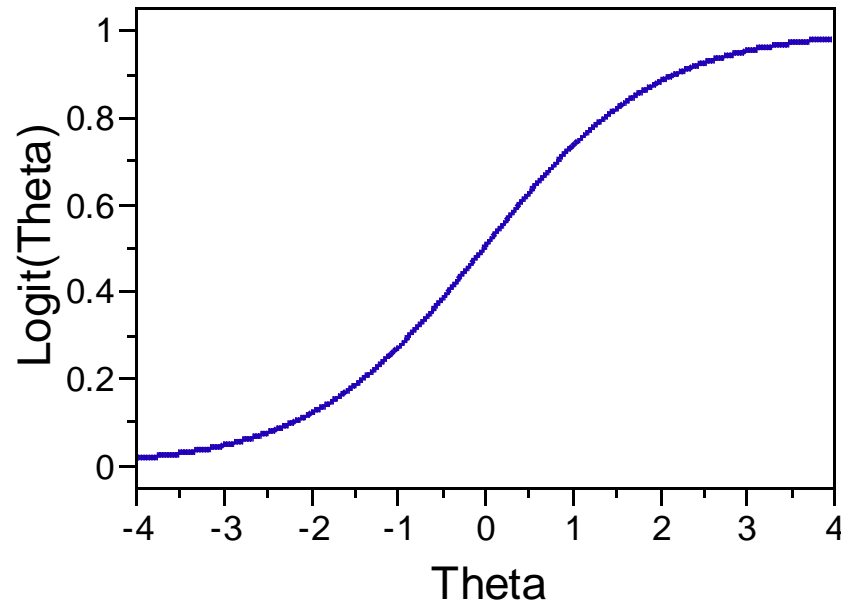
$$\ell(\theta) \triangleq \text{Logit}(\theta) = \frac{e^\theta}{1 + e^\theta}$$

- In the case of ordinary regression a function like  $p(x)$  was modeled as a linear function of  $x$ , or occasionally as a quadratic, etc.
- For binary outcomes,  $Y$ , we will instead write  $p$  as the logit of a linear function of  $x$ .
- Thus,  $p = \ell(\theta)$  where  $\theta = \beta_0 + \beta_1 x$ , i.e.,

$$p(x) = \Pr(Y=0|x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

## Properties of the Logistic $p$ Function

- Here is a plot of the logit function  $\ell(\theta) = e^\theta / (1 + e^\theta)$



- Note:  $0 < \ell(\theta) < 1$ . Also,  $1/2 = \ell(0)$ ,
- Hence:  $p = \ell(\beta_0 + \beta_1 x) = 1/2$  when  $\beta_0 + \beta_1 x = 0$ , etc.
- $\ell(\theta)$  is increasing func. If  $\beta_1 > 0$  then  $p$  is increasing in  $x$ .

## Why Do We Use the Logistic Link?

- We know that  $0 \leq p(x) \leq 1$ . [WHY?]
  1. Hence a linear function for  $p(x)$  is less natural, since it may fall outside this range, whereas the logit link  $p = \ell(\beta_0 + \beta_1 x)$  always satisfies  $0 \leq p \leq 1$
  2. A linear function of  $x$  itself for  $p$  is also slightly less convenient mathematically than is the logit link function.  
*[This isn't evident until one studies the detailed mathematics of the situation, which we won't do.]*
  3. **It's the conventional thing to do.**  
*[Any other modeling choice may require special justification.]*
- So we'll always use the logit link function when  $Y$  is a binary response variable

## Mathematical Background for Ordinary Logistic Regression

- An “ordinary” logistic regression fits the logit argument,  $\theta$ , of the probability,  $p$ , of an “event” as a linear function of one variable:

$$p = \text{Prob}(\text{“Event(0)” given } x)$$

- The “logit” argument,  $\theta$ , is

$$\theta = \theta(x) \equiv \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

Solving for  $p$  gives

$$p = P(Y = \mathbf{0} | x) = \frac{e^\theta}{1 + e^\theta} = \frac{e^{(\beta_0 + \beta_1 x)}}{1 + e^{(\beta_0 + \beta_1 x)}}.$$

## Statistical Model

- The observed  $Y$ 's are assumed to be INDEPENDENT Bernoulli( $p$ ) distributed

$$p = P(Y = \mathbf{0} | x) = \frac{e^\theta}{1 + e^\theta} = \frac{e^{(\beta_0 + \beta_1 x)}}{1 + e^{(\beta_0 + \beta_1 x)}}$$

$$P(Y = \mathbf{1} | x) = 1 - p = 1 - \frac{e^\theta}{1 + e^\theta} = \frac{1}{1 + e^{(\beta_0 + \beta_1 x)}}.$$

## Estimation of $\beta_0$ and $\beta_1$

Method: Maximum Likelihood

The “Likelihood Function” of  $\beta_0$  and  $\beta_1$  given all the **0** or **1** observations for the  $Y^s$  is proportional to the joint probability of those  $Y$ -values:

$$\Lambda_{\beta_0, \beta_1} \approx \prod_{i=1}^n \left( \frac{e^{(\beta_0 + \beta_1 x_i)}}{1 + e^{(\beta_0 + \beta_1 x_i)}} \right)^{1-y_i} \left( \frac{1}{1 + e^{(\beta_0 + \beta_1 x_i)}} \right)^{y_i} .$$

Maximum Likelihood Estimators  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are the maximizers of the above likelihood function.

Note: Usually the maximum likelihood estimators are found through numerical (computer) calculations by maximizing the log likelihood functions.

Hypotheses tests of  $H_0: \beta_1 = 0$

There are (at least) 2 different standard tests of  $H_0: \beta_1 = 0$ :

1. The likelihood ratio  $\chi^2$  test

Test statistic:  $\chi^2 = 2 (\log \Lambda_{FullModel} - \log \Lambda_{H_0})$

$$\log \Lambda_{FullModel} = \sup_{\beta_0, \beta_1} \left\{ \log \prod_{i=1}^n \left( \frac{e^{(\beta_0 + \beta_1 x)}}{1 + e^{(\beta_0 + \beta_1 x)}} \right)^{1-y_i} \left( \frac{1}{1 + e^{(\beta_0 + \beta_1 x)}} \right)^{y_i} \right\}.$$

Under  $H_0: \beta_1 = 0$  all the values of  $p(x)$  must be equal. Hence

$$\log \Lambda_{H_0} = n \log [\hat{P}_0], \text{ where } \hat{P}_0 = \frac{\text{Number of Events}}{n}.$$

**Reject  $H_0$**  if  $\chi^2$  is larger than the critical value from the  $\chi^2$  table with 1 df. P-values for this test are also drawn from the  $\chi^2$  table with 1 df.

2. The “Wald”  $\chi^2$  test:

Testing statistics: 
$$\chi^2 = \sum \frac{(\textit{observed} - \textit{expected})^2}{\textit{expected}}.$$

Here *observed* are either 0 or 1 (The  $Y$  values.)

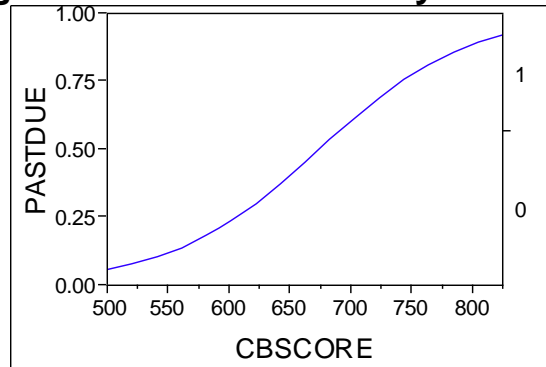
*expected* is  $p$ , calculated at the maximum likelihood estimate,  $\hat{p}$

**Reject  $H_0$**  if the statistic is larger than the critical value from the  $\chi^2$  table with 1 df.

**NOTE:** The test statistics and P-values from these 2 statistics are usually very similar (but not identical). You can use either (or both). And it is embarrassing if the conclusions from these 2 don't agree.

# JMP Data Analysis (from Fit Y by X)

## Logistic Fit of PASTDUE By CBSCORE



### Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	20.53	1	<b>2(LogLik Ratio)</b> 41.06	<.0001
Full	217.37			
Reduced	237.89			

RSquare (U)	(usually not very large)	0.0863
Observations		348

### Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-10.91	1.90	33.12	<.0001
CBSCORE	0.02	0.00	<b>(Wald)</b> 34.82	<.0001

For log odds of 0/1

## NOTES:

1. The plot shows the predicted curve  $\hat{p}(x)$ .
2. The tables give the 2 Log Likelihood Ratio and the Wald test statistics, as we've noted **in blue**. Note that these two test statistics are not quite the equal.
3. The legend **in red** reminds you that  $\hat{p}(x)$  is the estimate of the probability that  $Y = \mathbf{0}$ .
4.  $R^2$  is derived from the Whole Model Test table in the same general way that  $R^2$  is found in a conventional Linear Model analysis. The values of this  $R^2$  statistic are not usually as large as those you're used to in conventional Linear Models.
5. The term **Difference** in the whole model test is what was called **Model** in an ANOVA table and **Full**  $\equiv$  **Error** and **Reduced**  $\equiv$  **Total**.
6. You can get the same information in slightly different format from the Fit Model platform.

## Prediction

- The Fit Model Platform also provides values of  $\hat{p}(x)$ .

(These are under the “Save Probability Formula” button.)

- Look in the newly added Data column, “Prob[0]”

For example, the loan in row 1 has CBScore = 711, and the table says  $\hat{p}(711) = .6466$ .

- It also lists the most likely outcome for this loan as **0**, since  $.6466 > .5$ .
- This loan actually had a PASTDUE score of **0**. So someone who predicted that PASTDUE=**0** whenever  $\hat{p}(x) \geq 1/2$  would have received a “**true positive**” outcome on this prediction.

## ROC Curve and Table

- Often a cutoff value of  $\hat{p}(x) \geq 1/2$  is not the best way to achieve a high observed proportion of

$$\text{"True Positive Rate" } \triangleq \text{"Sensitivity" } \triangleq \frac{\#\{\mathbf{0} \text{ classed as } 0\}}{\#\{\mathbf{0} \text{ in the data}\}} \text{ vs}$$

$$\text{"False Positive rate" } \triangleq \text{"1-Specificity" } \triangleq \frac{\#\{\mathbf{1} \text{ classed as } 0\}}{\#\{\mathbf{1} \text{ in the data}\}}$$

- The ROC curve and table give the available values of these quantities for various cutoff points. See next page for part of the table.
- Note the \* in the table. This shows that the cutoff point  $\hat{p}(x) \geq .4811$  maximizes Sensitivity+Specificity.

(PS: you could also use  $\hat{p}(x) \geq .4609$  which gives you the max correct classifications)

### Portion of ROC Table

X	Prob	1-Specificity	Sensitivity	Sens+Spec	True Pos	True Neg	False Pos	False Neg
675	0.5053	0.4933	0.7828	0.2895	155	76	74	43
674	0.5013	0.5000	0.7828	0.2828	155	75	75	43
672	0.4932	0.5000	0.7980	0.2980	158	75	75	40
671	0.4892	0.5000	0.8030	0.3030	159	75	75	39
670	0.4851	0.5000	0.8081	0.3081	160	75	75	38
669	0.4811	0.5000	0.8131	0.3131 *	161	75	75	37
667	0.4730	0.5200	0.8283	0.3083	164	72	78	34
666	0.4690	0.5333	0.8333	0.3000	165	70	80	33
665	0.4649	0.5467	0.8434	0.2968	167	68	82	31
664	0.4609	0.5467	0.8535	0.3069	169	68	82	29
663	0.4569	0.5667	0.8586	0.2919	170	65	85	28

As an example, the entry at X=671 says  $\hat{p}(671) = .4892$ . The sensitivity there is

$$\text{Sensitivity}(671) = \frac{159}{198} = .8030. \text{ Also,}$$

$$1\text{-Specificity}(671) = \frac{75}{150} = \frac{\# \text{TrueNeg}}{\text{Total\# Neg}} = .5000.$$

The value \* is actually the point where (Sensitivity + 1 – Specificity) is a max.