COSC 3337 : Data Science I

N. Rizk

College of Natural and Applied Sciences

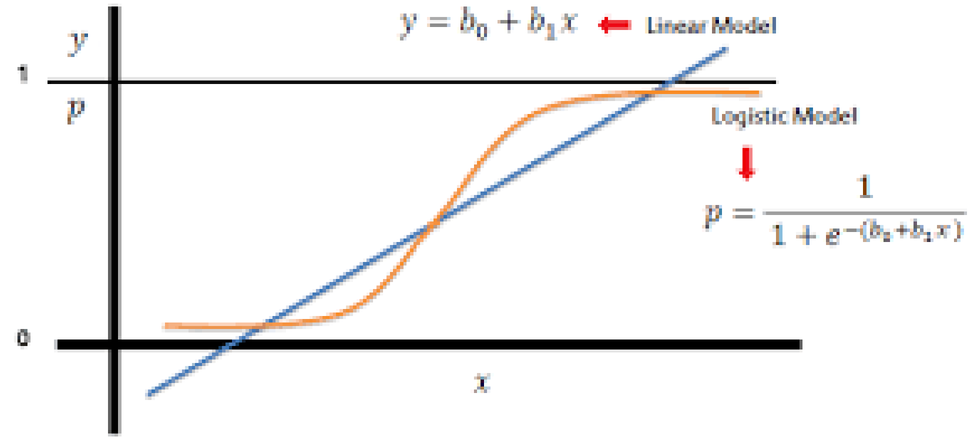
Department of Computer Science

University of Houston

Linear vs Logistic Regression

Linear regression is used to predict the continuous dependent variable using a given set of independent variables.

Logistic Regression is used to predict the categorical dependent variable using a given set of independent variables



Regression

- A form of statistical modeling that attempts to evaluate the relationship between one variable (termed the dependent variable) and one or more other variables (termed the independent variables). It is a form of global analysis as it only produces a single equation for the relationship.
- A model for predicting one variable from another.

Linear Regression Review

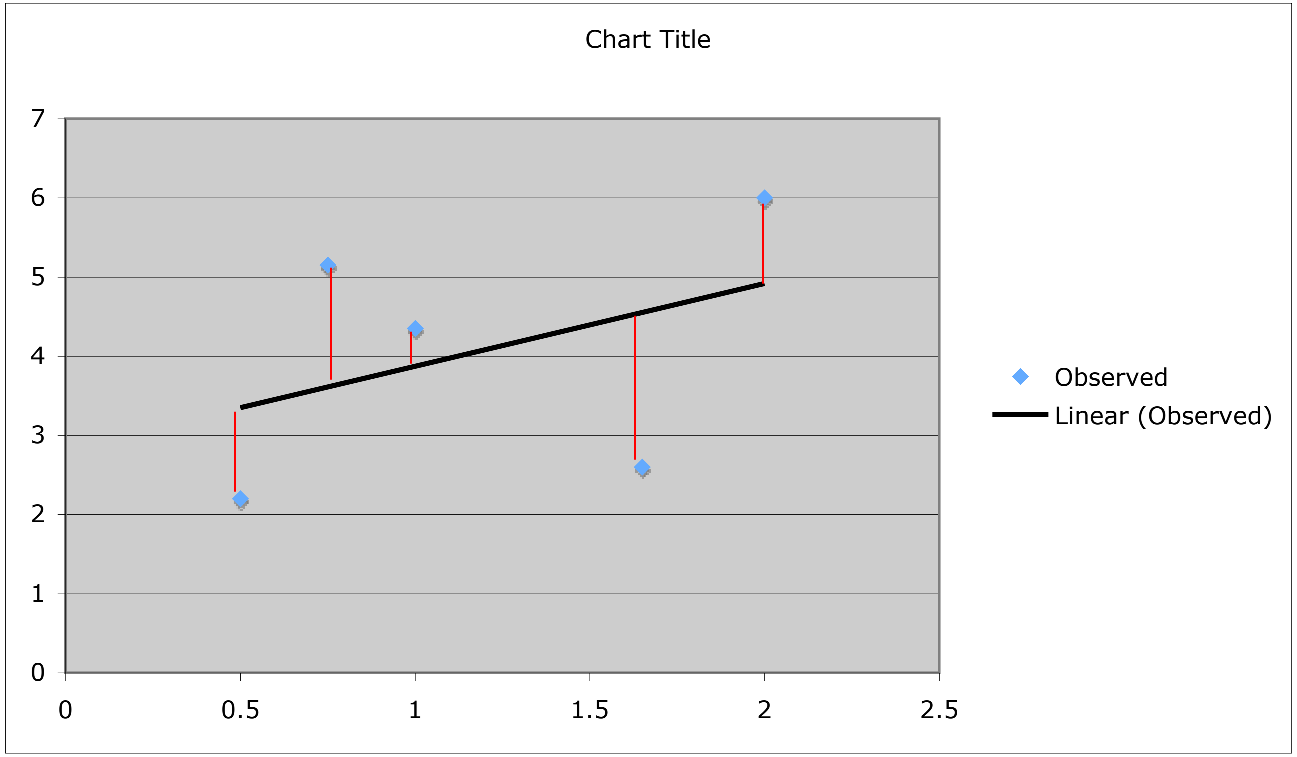
- Regression used to fit a linear model to data where the dependent variable is continuous:

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n + \epsilon$$

- Given a set of points (Xi,Yi), we wish to find a linear function (or line in 2 dimensions) that "goes through" these points.
- In general, the points are not exactly aligned:
  - Find line that best fits the points

Residue

- Error or residue:
  - Observed value - Predicted value (black line)



Sum-squared Error (SSE)

$$SSE = \sum_{y} (y_{observed} - y_{predicted})^2$$

$$TSS = \sum_{y} (y_{observed} - \bar{y}_{observed})^2$$

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

What is Best Fit?

- The smaller the SSE, the better the fit
- Hence,
  - Linear regression attempts to minimize SSE (or similarly to maximize R2)
- Assume 2 dimensions

$$Y = b_0 + b_1 X$$

Analytical Solution

$$b_0 = \frac{\sum y - b_1 \sum x}{n}$$

$$b_1 = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - \left(\sum x\right)^2}$$

Example (I)

x	y	x^2	xy
1.20	4.00	1.44	4.80
2.30	5.60	5.29	12.88
3.10	7.90	9.61	24.49
3.40	8.00	11.56	27.20
4.00	10.10	16.00	40.40
4.60	10.40	21.16	47.84
5.50	12.00	30.25	66.00
24.10	**58.00**	**95.31**	**223.61**

Target: y=2x+1.5

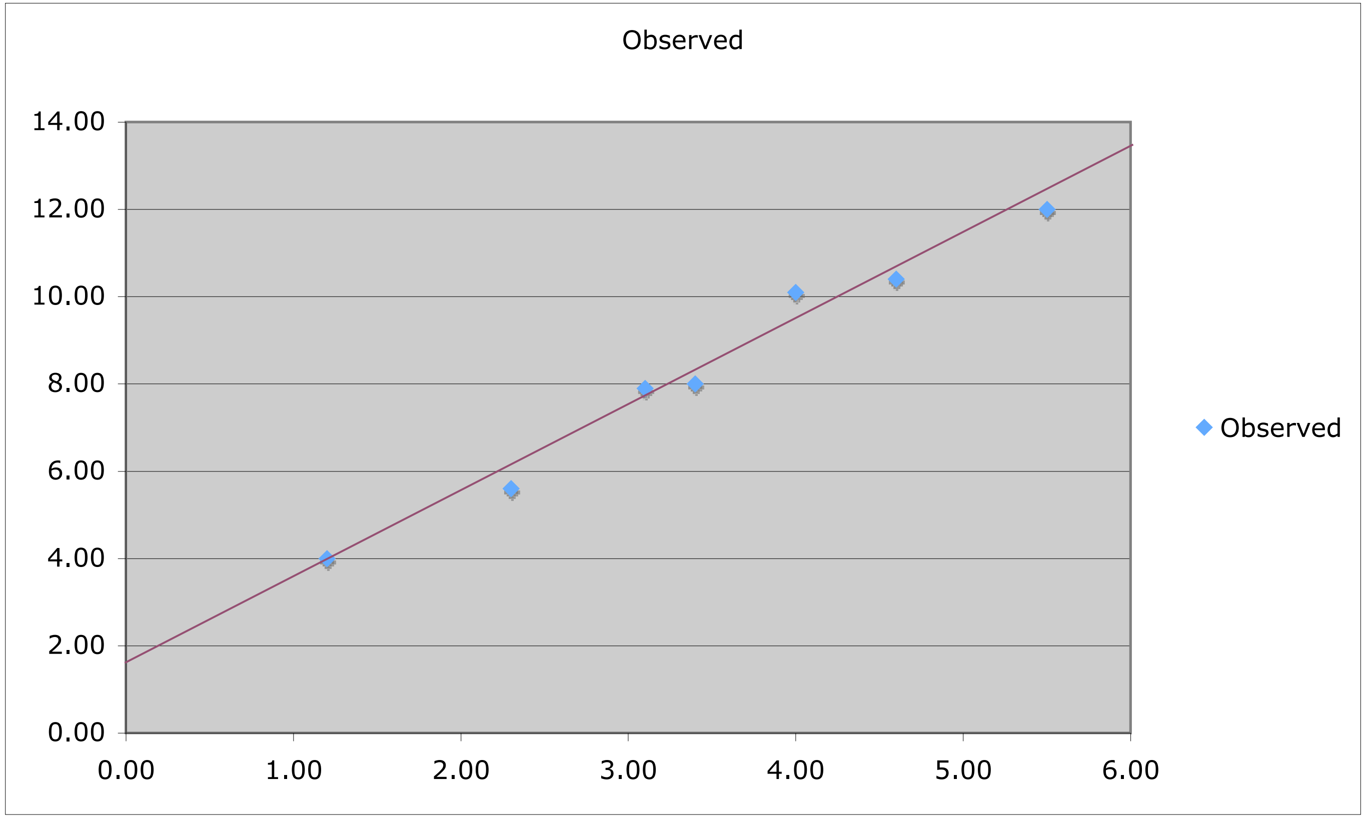
$$\begin{aligned} b_1 &= \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2} \\ &= \frac{7 \times 223.61 - 24.10 \times 58.00}{7 \times 95.31 - 24.10^2} \\ &= \frac{1565.27 - 1397.80}{667.17 - 580.81} \\ &= \frac{167.47}{86.36} = \underline{\underline{1.94}} \end{aligned}$$

$$\begin{aligned} b_0 &= \frac{\sum y - b_1 \sum x}{n} \\ &= \frac{58.00 - 1.94 \times 24.10}{7} \\ &= \frac{11.27}{7} = \underline{\underline{1.61}} \end{aligned}$$

Example (II)

y=1.94x+1.61

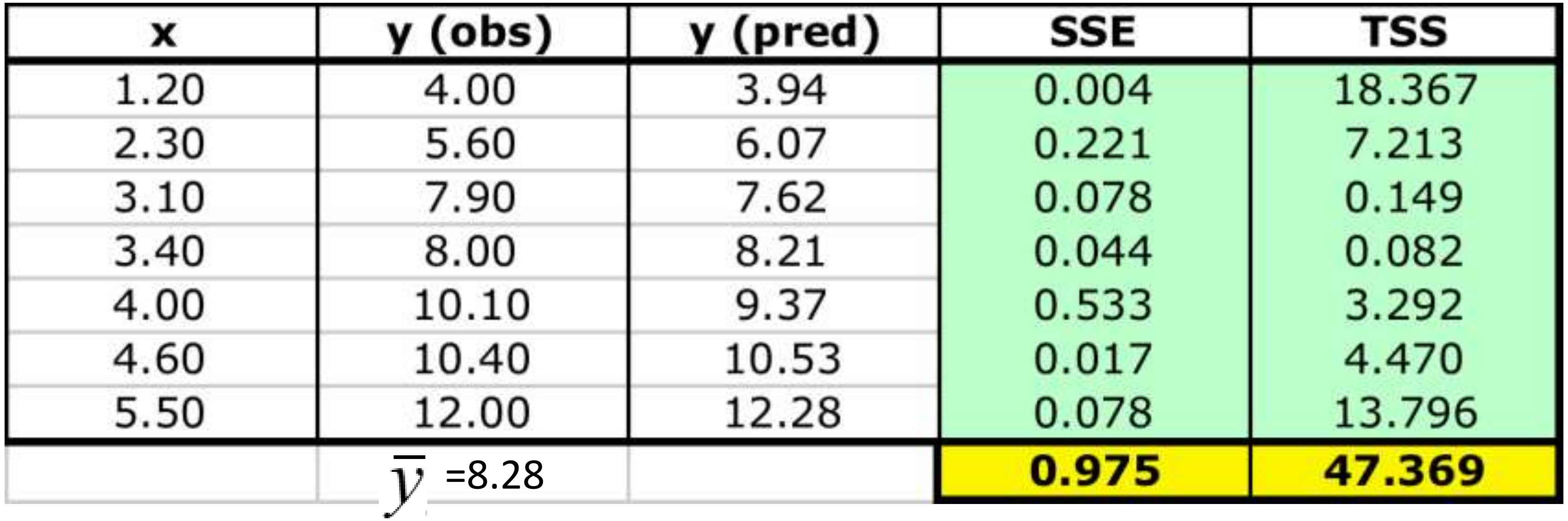
Redline



Example (III)

$$SSE = \sum_{y} (y_{observed} - y_{predicted})^2$$

$$TSS = \sum_{y} (y_{observed} - \bar{y}_{observed})^2$$



x	y (obs)	y (pred)	SSE	TSS
1.20	4.00	3.94	0.004	18.367
2.30	5.60	6.07	0.221	7.213
3.10	7.90	7.62	0.078	0.149
3.40	8.00	8.21	0.044	0.082
4.00	10.10	9.37	0.533	3.292
4.60	10.40	10.53	0.017	4.470
5.50	12.00	12.28	0.078	13.796
	$\bar{y}$ =8.28		**0.975**	**47.369**

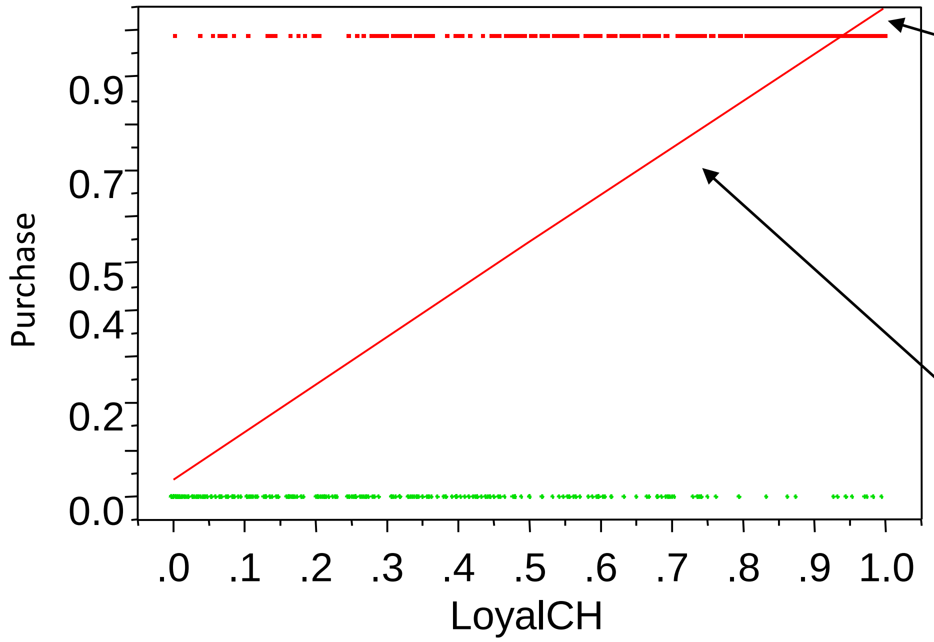
$$R^2 = 1 - \frac{SSE}{TSS} = 1 - \frac{0.975}{47.369} = 0.98$$

Case 1: Brand Preference for Orange Juice

- We would like to predict what customers prefer to buy(NO/YES): Citrus Hill or Minute Maid orange juice?
- The Y (Purchase) variable is categorical: 0 or 1
- The X (LoyalCH) variable is a numerical value (between 0 and 1) which specifies the how much the customers are loyal to the Citrus Hill (CH) orange juice
- Can we use Linear Regression when Y is categorical?

Why not Linear Regression?

- When Y only takes on values of 0 and 1, why standard linear regression in inappropriate?



How do we interpret values greater than 1?

How do we interpret values of Y between 0 and 1?

Problems

- The regression line $\beta_0+\beta_1X$ can take on any value between negative and positive infinity

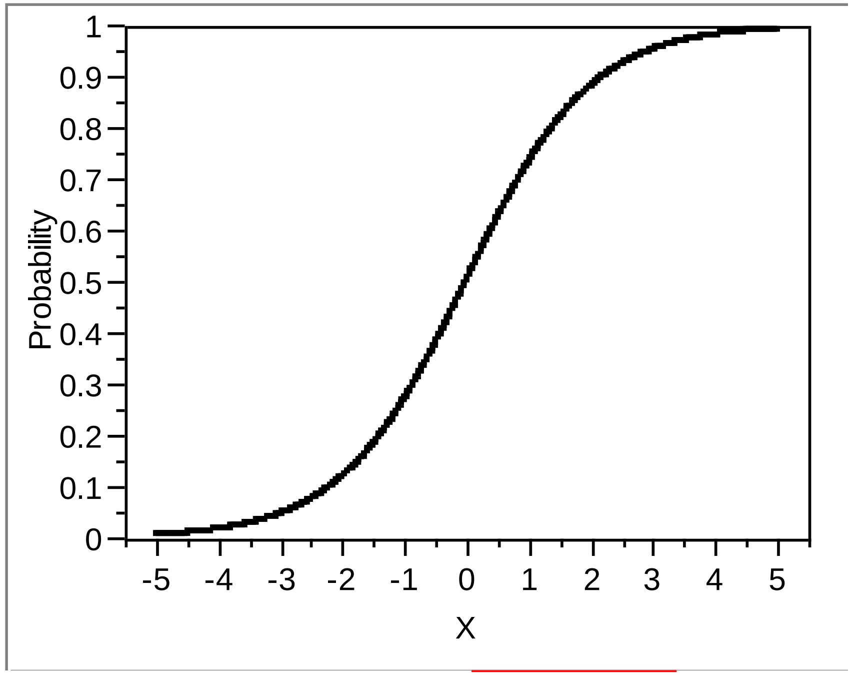
- In the orange juice classification problem, Y can only take on two possible values: 0 or 1.

- Therefore the regression line almost always predicts the wrong value for Y in classification problems

Solution: Use Logistic Function

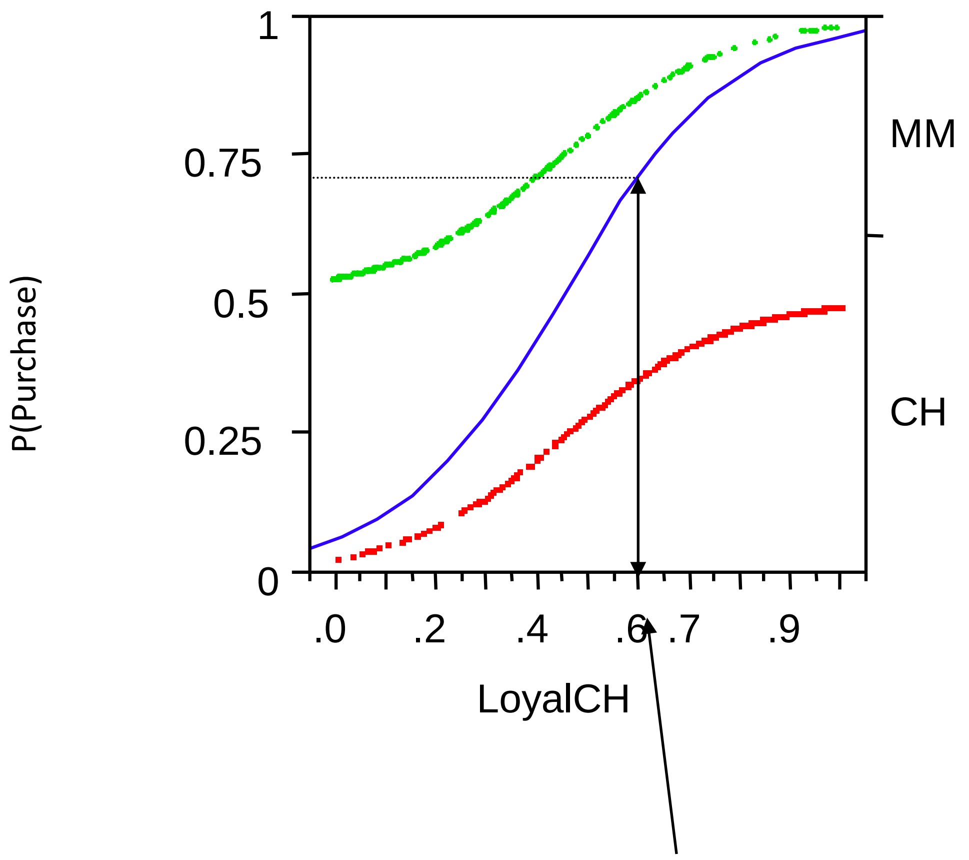
- Instead of trying to predict Y, let’s try to predict P(Y = 1), i.e., the probability a customer buys Citrus Hill (CH) juice.
- Thus, we can model P(Y = 1) using a function that gives outputs between 0 and 1.
- We can use the logistic function
- Logistic Regression!

$$\boxed{p = P(Y = 1) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}}$$



Logistic Regression

- Logistic regression is very similar to linear regression
- We come up with b_0 and b_1 to estimate β_0 and β_1.
- We have similar problems and questions as in linear regression
  - e.g. Is β_1 equal to 0? How sure are we about our guesses for β_0 and β_1?



If LoyalCH is about .6 then Pr(CH) ≈ .7.

Logistic Regression

- Regression used to fit a curve to data in which the dependent variable is binary, or dichotomous
- Typical application: Medicine
  - We might want to predict response to treatment, where we might code survivors as 1 and those who don't survive as 0

Example

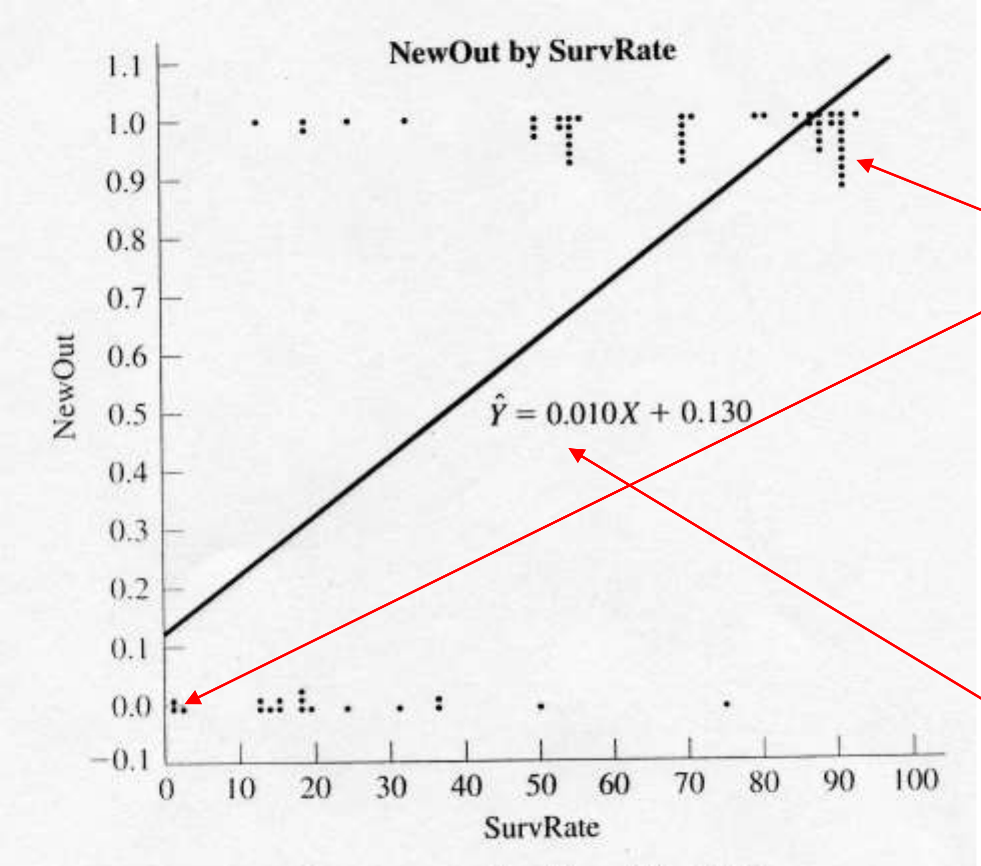


FIGURE 15·7 *Outcome as a function of SurvRate*

Observations: For each value of SurvRate, the number of dots is the number of patients with that value of NewOut

Regression: Standard linear regression

Problem: extending the regression line a few units left or right along the X axis produces predicted probabilities that fall outside of [0,1]

A Better Solution

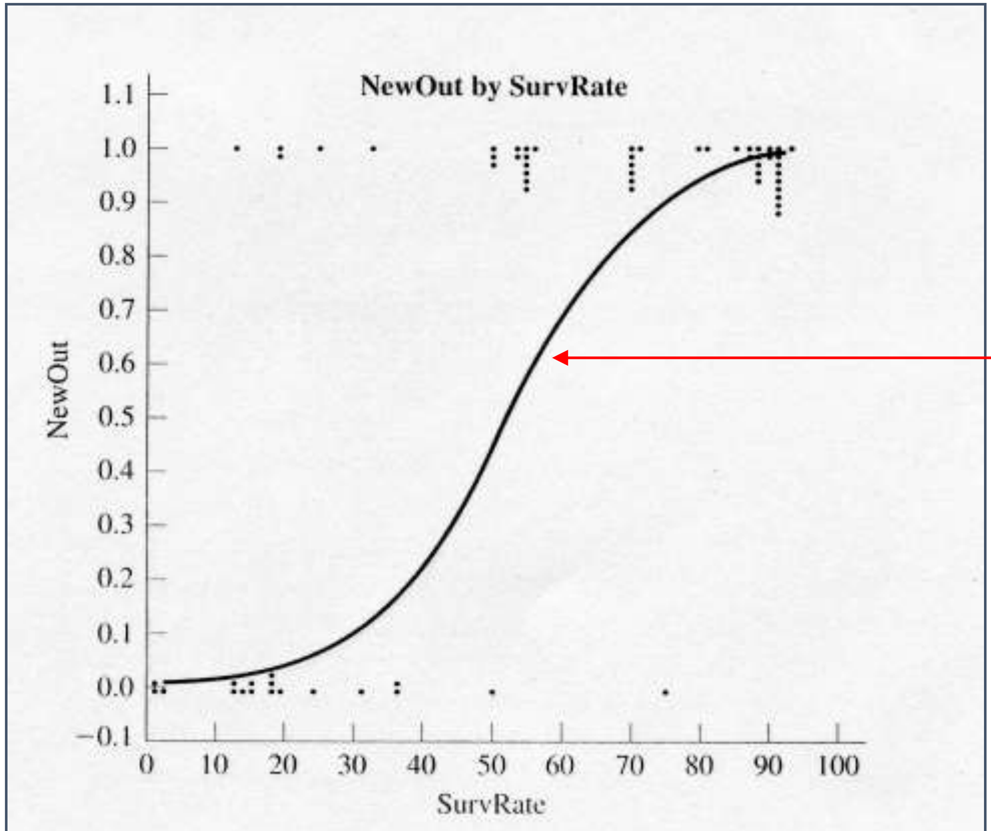


FIGURE 15.8 *More appropriate regression line for predicting outcome*

<u>Regression Curve</u>:
Sigmoid function!

(bounded by asymptotes *y*=0 and *y*=1)

Odds

- Given some event with probability *p* of being 1, the odds of that event are given by:

$$\text{odds} = p\,/\,(1-p)$$

- Consider the following data

		Delinquent		
		Yes	**No**	**Total**
Testosterone	**Normal**	402	3614	4016
	High	101	345	446
		503	3959	4462

- The odds of being delinquent if you are in the Normal group are:

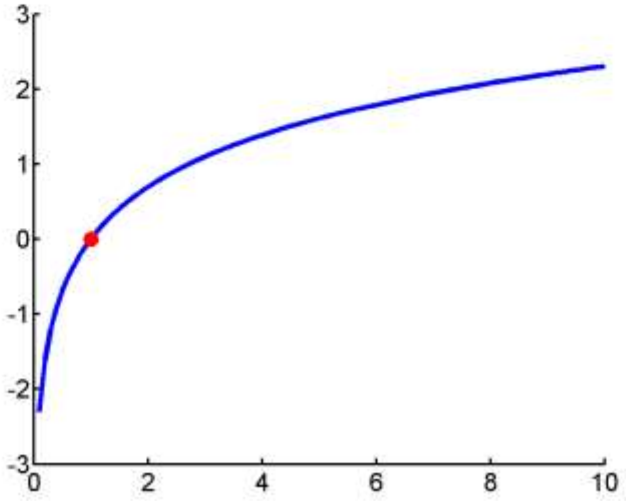
pdelinquent/(1–pdelinquent) = (402/4016) / (1 - (402/4016)) = 0.1001 / 0.8889 = 0.111

Odds Ratio

- The odds of being not delinquent in the Normal group is the reciprocal of this:
  - 0.8999/0.1001 = 8.99
- Now, for the High testosterone group
  - odds(delinquent) = 101/345 = 0.293
  - odds(not delinquent) = 345/101 = 3.416
- When we go from Normal to High, the odds of being delinquent nearly triple:
  - Odds ratio: 0.293/0.111 = 2.64
  - 2.64 times more likely to be delinquent with high testosterone levels

Logit Transform

- The logit is the natural log of the odds



- logit(p) = ln(odds) = ln (p/(1-p))

Logistic Regression

- In logistic regression, we seek a model:

$$\text{logit}(p) = b_0 + b_1 X$$

- That is, the log odds (logit) is assumed to be linearly related to the independent variable X
- So, now we can focus on solving an ordinary (linear) regression!

Recovering Probabilities

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

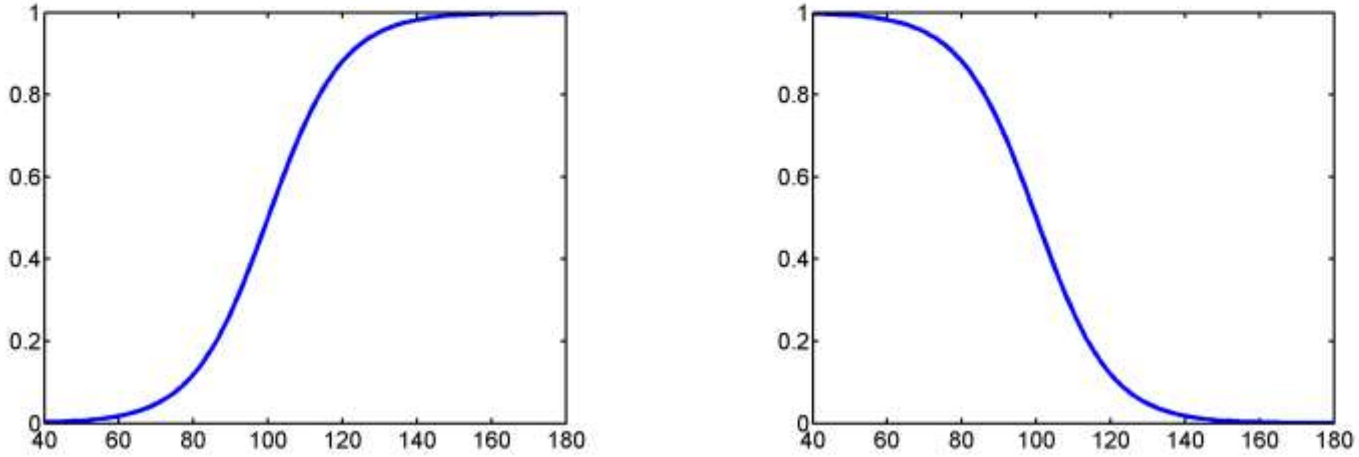
$$\Leftrightarrow \frac{p}{1-p} = e^{\beta_0 + \beta_1 X}$$

$$\Leftrightarrow p = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

which gives *p* as a sigmoid function!

Logistic Response Function

- When the response variable is binary, the shape of the response function is often sigmoidal:



Interpretation of β1

- Let:
  - odds1 = odds for value X (p/(1–p))
  - odds2 = odds for value X + 1 unit
- Then:

$$\frac{odds2}{odds1} = \frac{e^{b_0 + b_1(X+1)}}{e^{b_0 + b_1 X}}$$

$$= \frac{e^{(b_0 + b_1 X) + b_1}}{e^{b_0 + b_1 X}} = \frac{e^{(b_0 + b_1 X)} e^{b_1}}{e^{b_0 + b_1 X}} = e^{b_1}$$

- Hence, the exponent of the slope describes the proportionate rate at which the predicted odds ratio changes with each successive unit of X

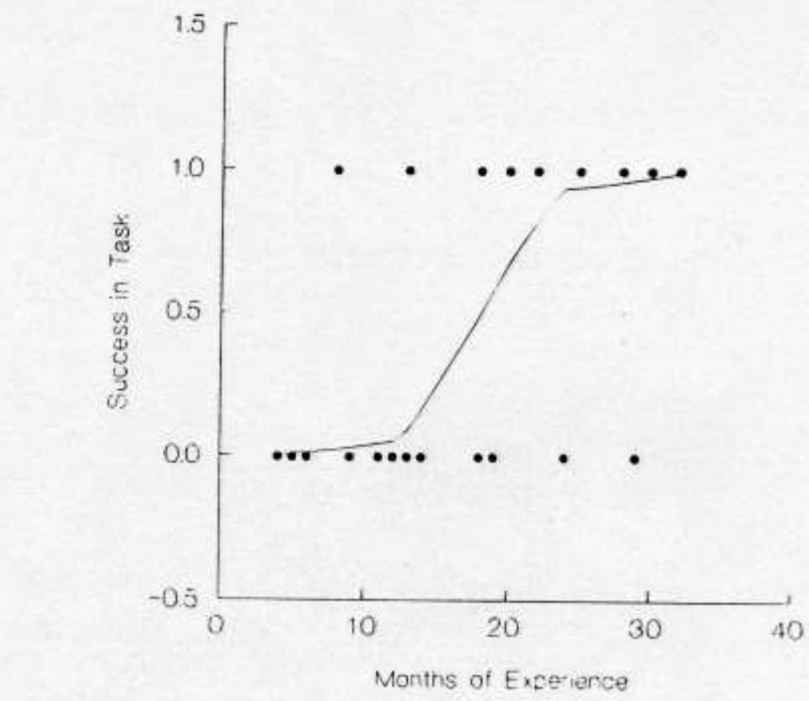
Sample Calculations

- Suppose a cancer study yields:
  - log odds = –2.6837 + 0.0812 SurvRate
- Consider a patient with SurvRate = 40
  - log odds = –2.6837 + 0.0812(40) = 0.5643
  - odds = $e^{0.5643}$ = 1.758
  - patient is 1.758 times more likely to be improved than not
- Consider another patient with SurvRate = 41
  - log odds = –2.6837 + 0.0812(41) = 0.6455
  - odds = $e^{0.6455}$ = 1.907
  - patient's odds are 1.907/1.758 = 1.0846 times (or 8.5%) better than those of the previous patient
- Using probabilities
  - p40 = 0.6374 and p41 = 0.6560
  - Improvements appear different with odds and with *p*

Example 1 (I)

- A systems analyst studied the effect of computer programming experience on ability to complete a task within a specified time
- Twenty-five persons selected for the study, with varying amounts of computer experience (in months)
- Results are coded in binary fashion: *Y* = 1 if task completed successfully; *Y* = 0, otherwise

Scatter Plot and Loess Curve—Programming Task Example.



Loess: form of local regression

Example 1 (II)

- Results from a standard package give:
  - $\beta 0 = -3.0597$ and $\beta 1 = 0.1615$
- Estimated logistic regression function:

$$p = \frac{1}{1 + e^{3.0597 - 0.1615X}}$$

- For example, the fitted value for X = 14 is:

$$p = \frac{1}{1 + e^{3.0597 - 0.1615(14)}} = 0.31$$

(Estimated probability that a person with 14 months experience will successfully complete the task)

Example 1 (III)

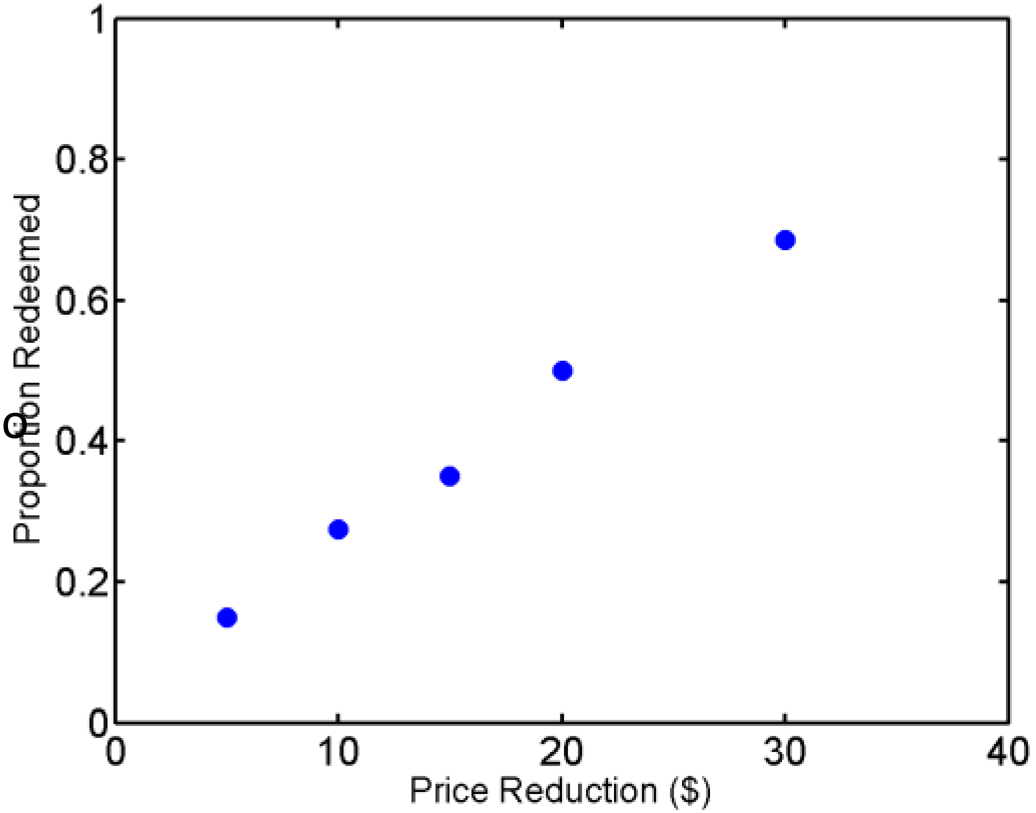
- We know that the probability of success increases sharply with experience
  - Odds ratio: exp(β1) = e0.1615 = 1.175
  - Odds increase by 17.5% with each additional month of experience
- A unit increase of one month is quite small, and we might want to know the change in odds for a longer difference in time
  - For c units of X: exp(cβ1)

Example 1 (IV)

- Suppose we want to compare individuals with relatively little experience to those with extensive experience, say 10 months versus 25 months (c = 15)
  - Odds ratio: e15x0.1615 = 11.3
  - Odds of completing the task increase 11-fold!

Example 2 (I)

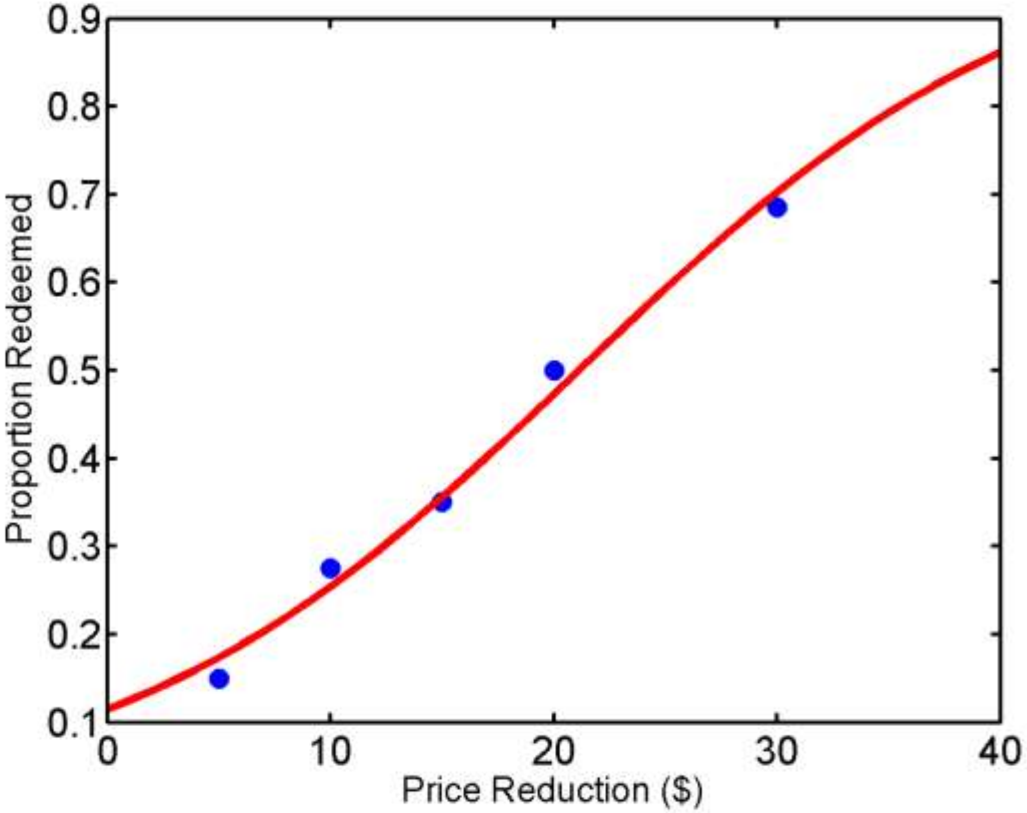
- In a study of the effectiveness of coupons offering a price reduction, 1,000 homes were selected and coupons mailed
- Coupon price reductions: 5, 10, 15, 20, and 30 dollars
- 200 homes assigned at random to each coupon value
- *X*: amount of price reduction
- *Y*: binary variable indicating whether or not coupon was redeemed



Example 2 (II)

- Fitted response function
  - β_0 = -2.04 and β_1 = 0.097

- Odds ratio: $\exp(\beta_1) = e^{0.097} = 1.102$
- Odds of a coupon being redeemed are estimated to increase by 10.2% with each $1 increase in the coupon value (i.e., $1 in price reduction)



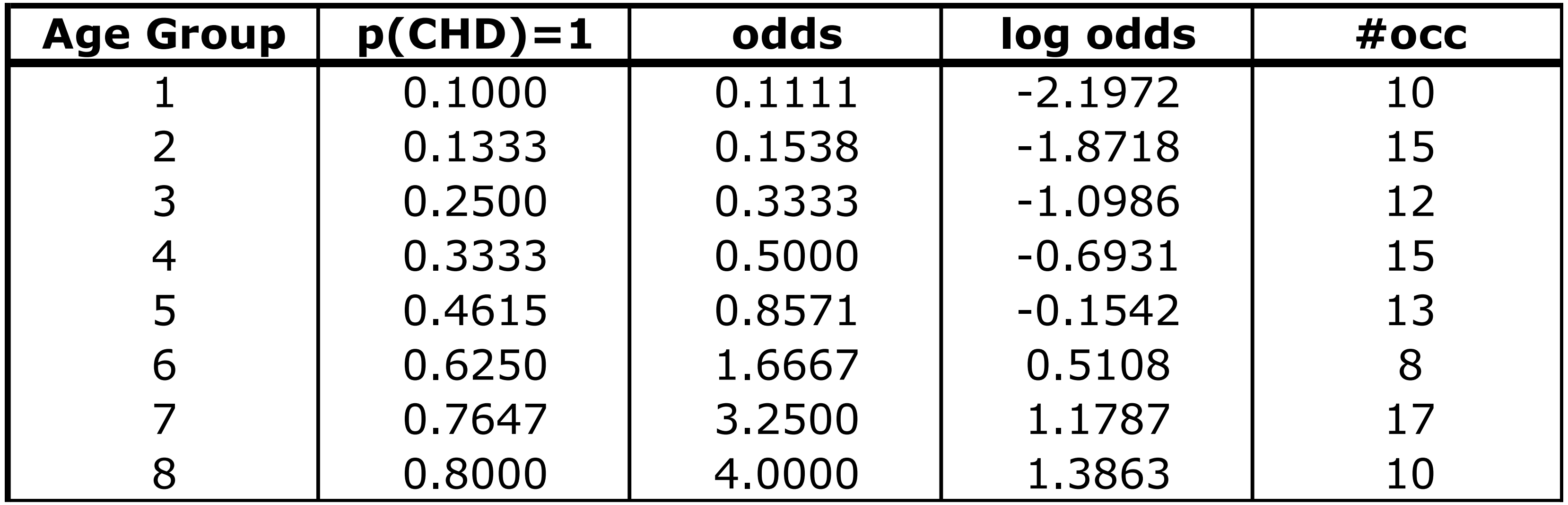
Putting it to Work

- For each value of X, you may not have probability but rather a number of <x,y> pairs from which you can extract frequencies and hence probabilities
  - Raw data: <12,0>, <12,1>, <14,0>, <12,1>, <14,1>, <14,1>, <12,0>, <12,0>
  - Probability data (p=1, 3rd entry is number of occurrences in raw data): <12, 0.4, 5>, <14, 0.66, 3>
  - Odds ratio data…

Coronary Heart Disease (I)

Age Group	Coronary Heart Disease		Total	
	No	Yes		
1	9	1	10	(20-29)
2	13	2	15	(30-34)
3	9	3	12	(35-39)
4	10	5	15	(40-44)
5	7	6	13	(45-49)
6	3	5	8	(50-54)
7	4	13	17	(55-59)
8	2	8	10	(60-69)
Total	57	43	100	

Coronary Heart Disease (II)



Age Group	p(CHD)=1	odds	log odds	#occ
1	0.1000	0.1111	-2.1972	10
2	0.1333	0.1538	-1.8718	15
3	0.2500	0.3333	-1.0986	12
4	0.3333	0.5000	-0.6931	15
5	0.4615	0.8571	-0.1542	13
6	0.6250	1.6667	0.5108	8
7	0.7647	3.2500	1.1787	17
8	0.8000	4.0000	1.3863	10

Coronary Heart Disease (III)

X (AG)	Y (log odds)	X^2	XY	#occ
1	-2.1972	1.0000	-2.1972	10
2	-1.8718	4.0000	-3.7436	15
3	-1.0986	9.0000	-3.2958	12
4	-0.6931	16.0000	-2.7726	15
5	-0.1542	25.0000	-0.7708	13
6	0.5108	36.0000	3.0650	8
7	1.1787	49.0000	8.2506	17
8	1.3863	64.0000	11.0904	10
448	**-37.6471**	**2504.0000**	**106.3981**	**100**

Note: the sums reflect the number of occurrences (Sum(X) = X1.#occ(X1)+…+X8.#occ(X8), etc.)

Coronary Heart Disease (IV)

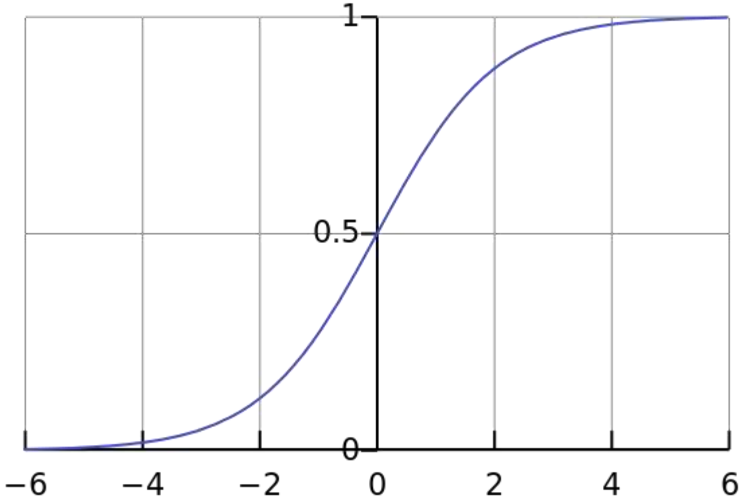
- Results from regression:
  - $\beta 0$ = -2.856 and $\beta 1$ = 0.5535

Age Group	p(CHD)=1	est. p
1	0.1000	0.0909
2	0.1333	0.1482
3	0.2500	0.2323
4	0.3333	0.3448
5	0.4615	0.4778
6	0.6250	0.6142
7	0.7647	0.7346
8	0.8000	0.8280

SSE	**0.0028**
TSS	**0.5265**
R2	**0.9946**

What is the model?

$$\log \frac{p(y=1)}{1-p(y=1)} = \beta \cdot \mathbf{x}$$



- p is the probability of the outcome.
- We compose the output of a linear model with a function which has range between -1 and 1.
- Beta defines a decision boundary in the variable space (y=1, y=0).

$$p_\beta(\mathbf{x}) = \frac{1}{1+\exp(-\beta^T \cdot \mathbf{x})}$$

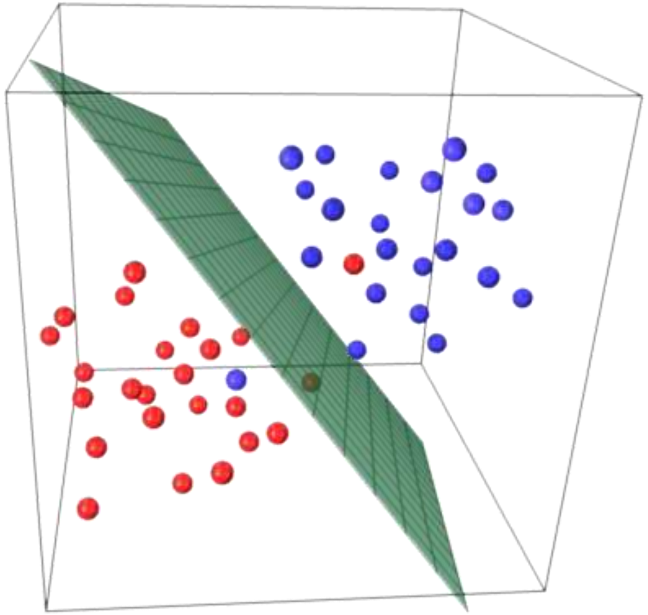
Decision boundary for Logistic Regression

$$p_\beta(\mathbf{x}) = \frac{1}{1 + \exp(-\beta^T \cdot \mathbf{x})}$$

- This defines a decision boundary where:

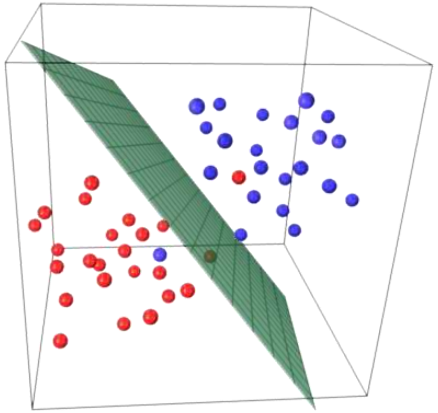
$\beta^T \cdot \mathbf{x}$

- Large and positive implies class 1
- Large and negative implies class 0



Decision boundary for Logistic Regression

$$p_\beta(\mathbf{x}) = \frac{1}{1 + \exp(-\beta^T \cdot \mathbf{x})}$$

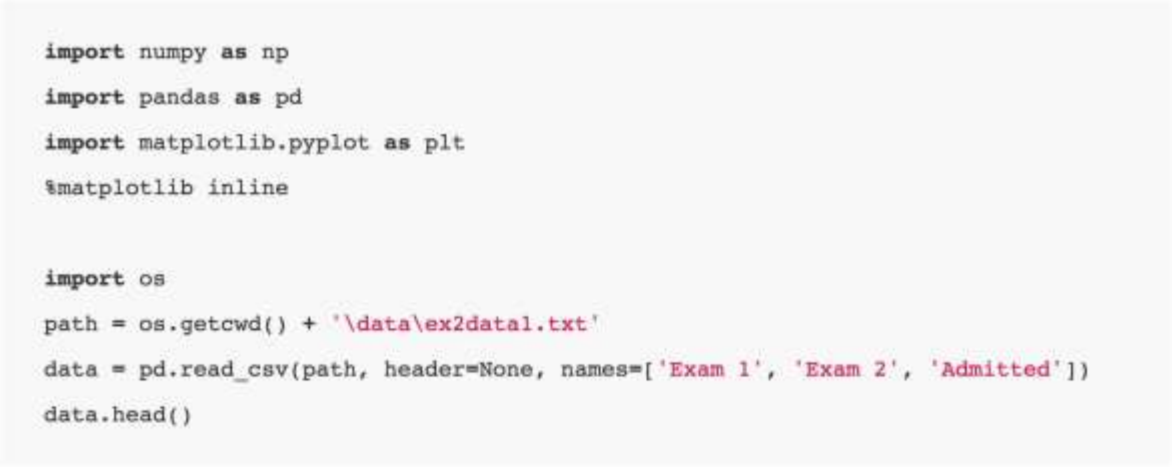


Let R denote the red class and B the blue class. Then

$$\text{If } p_\beta(\mathbf{x}) > 0.5 \text{ then } \mathbf{x} \in R$$
$$\text{If } p_\beta(\mathbf{x}) \leq 0.5 \text{ then } \mathbf{x} \in B$$

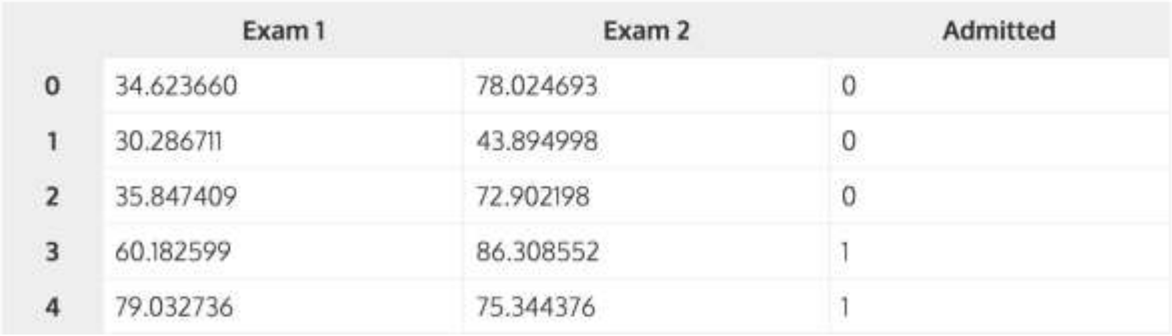
Note: The choice of 0.5 here is arbitrary.

Example - Admission to a program



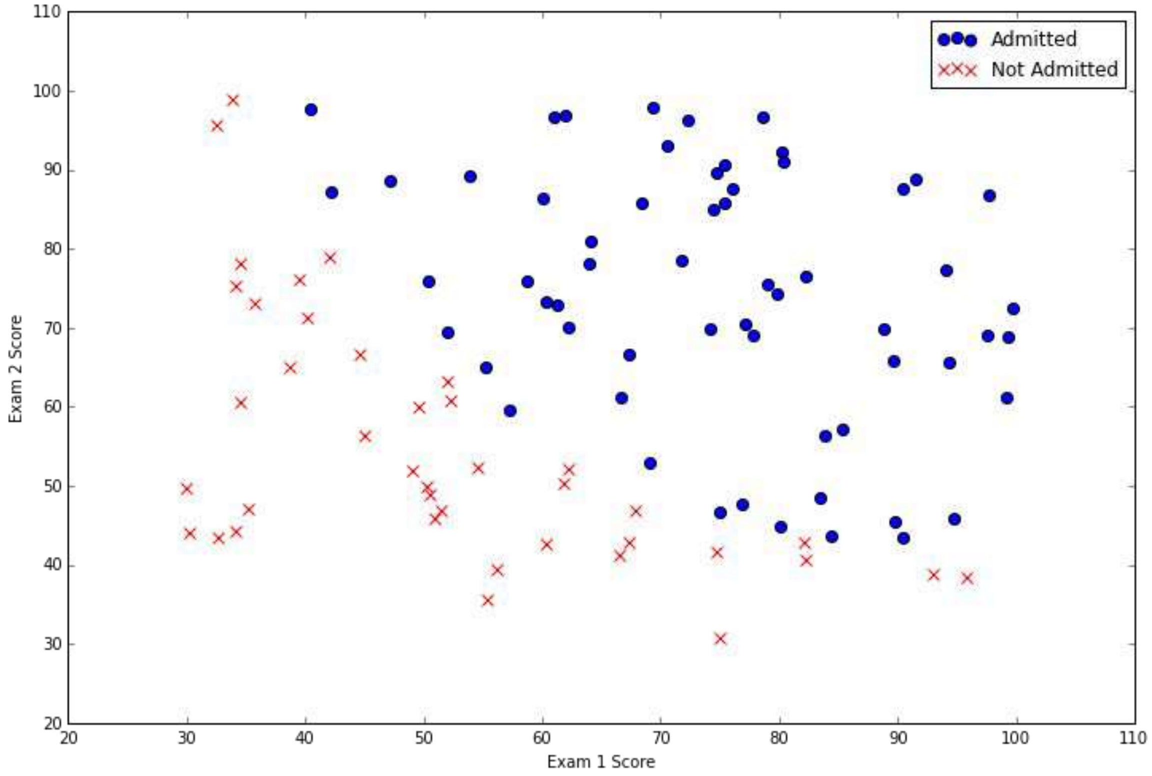
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline

import os
path = os.getcwd() + '\data\ex2data1.txt'
data = pd.read_csv(path, header=None, names=['Exam 1', 'Exam 2', 'Admitted'])
data.head()
```

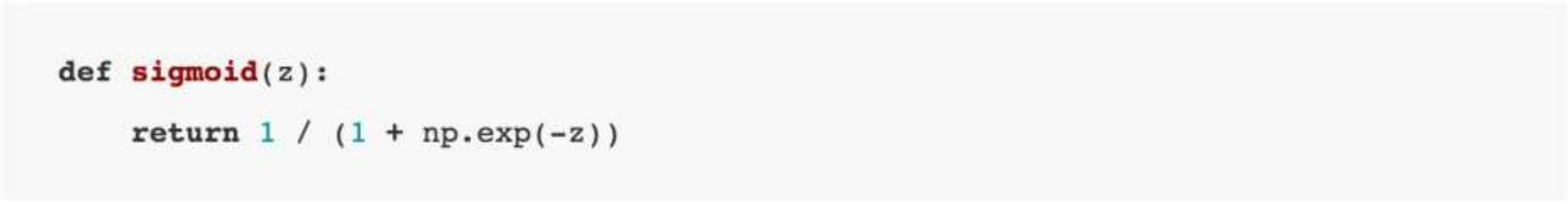


	Exam 1	Exam 2	Admitted
0	34.623660	78.024693	0
1	30.286711	43.894998	0
2	35.847409	72.902198	0
3	60.182599	86.308552	1
4	79.032736	75.344376	1

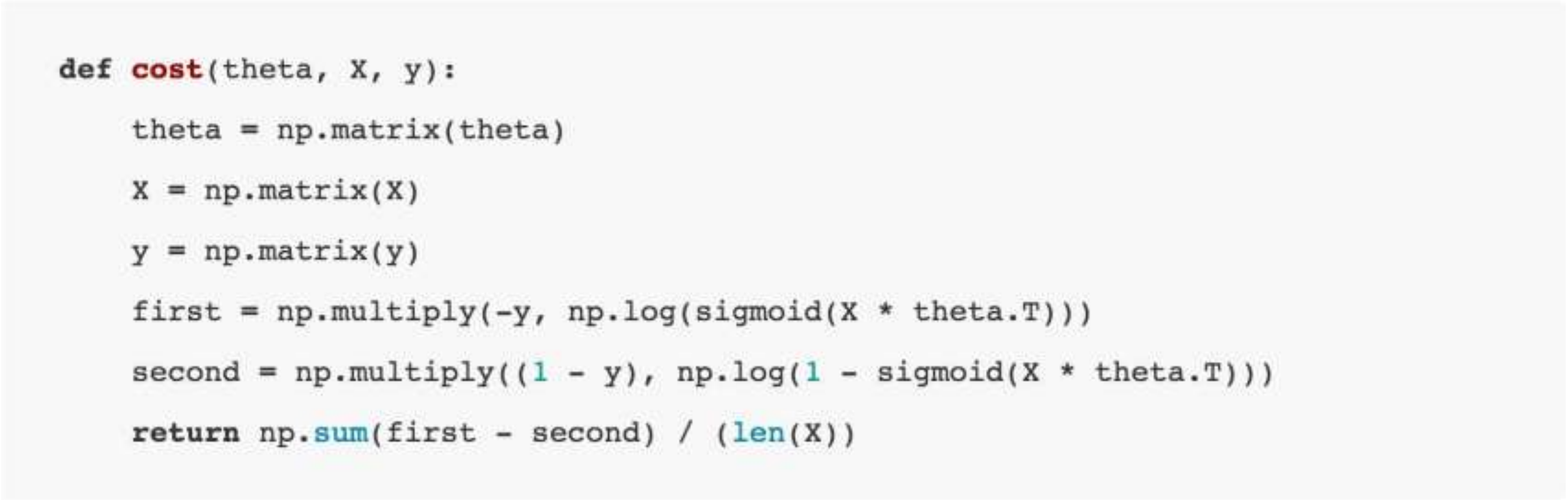
Scatter plot of admission results



Python code



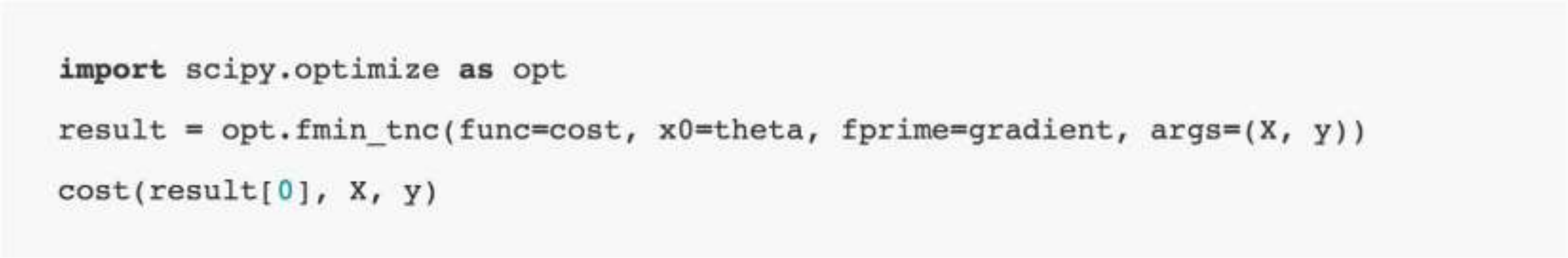
```python
def sigmoid(z):
    return 1 / (1 + np.exp(-z))
```



```python
def cost(theta, X, y):
    theta = np.matrix(theta)
    X = np.matrix(X)
    y = np.matrix(y)
    first = np.multiply(-y, np.log(sigmoid(X * theta.T)))
    second = np.multiply((1 - y), np.log(1 - sigmoid(X * theta.T)))
    return np.sum(first - second) / (len(X))
```

Minimizing Cost in Python

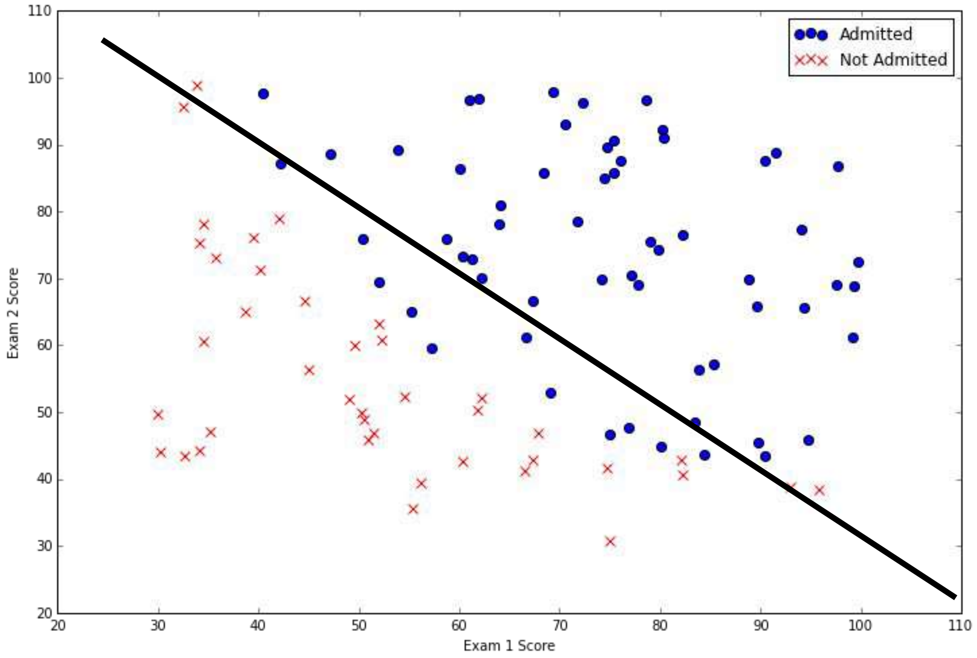
$$\text{Cost}_2(p_\beta(\mathbf{x}), y) = \begin{cases} -\log p_\beta(\mathbf{x}) & \text{if } y \text{ is } 1 \\ -\log(1 - p_\beta(\mathbf{x})) & \text{if } y \text{ is } 0 \end{cases}$$



```
import scipy.optimize as opt
result = opt.fmin_tnc(func=cost, x0=theta, fprime=gradient, args=(X, y))
cost(result[0], X, y)
```

```
0.20357134412164668
```

Final Decision Boundary



$$p_\beta(\mathbf{x}) = \frac{1}{1 + \exp(-\beta^T \cdot \mathbf{x})}$$

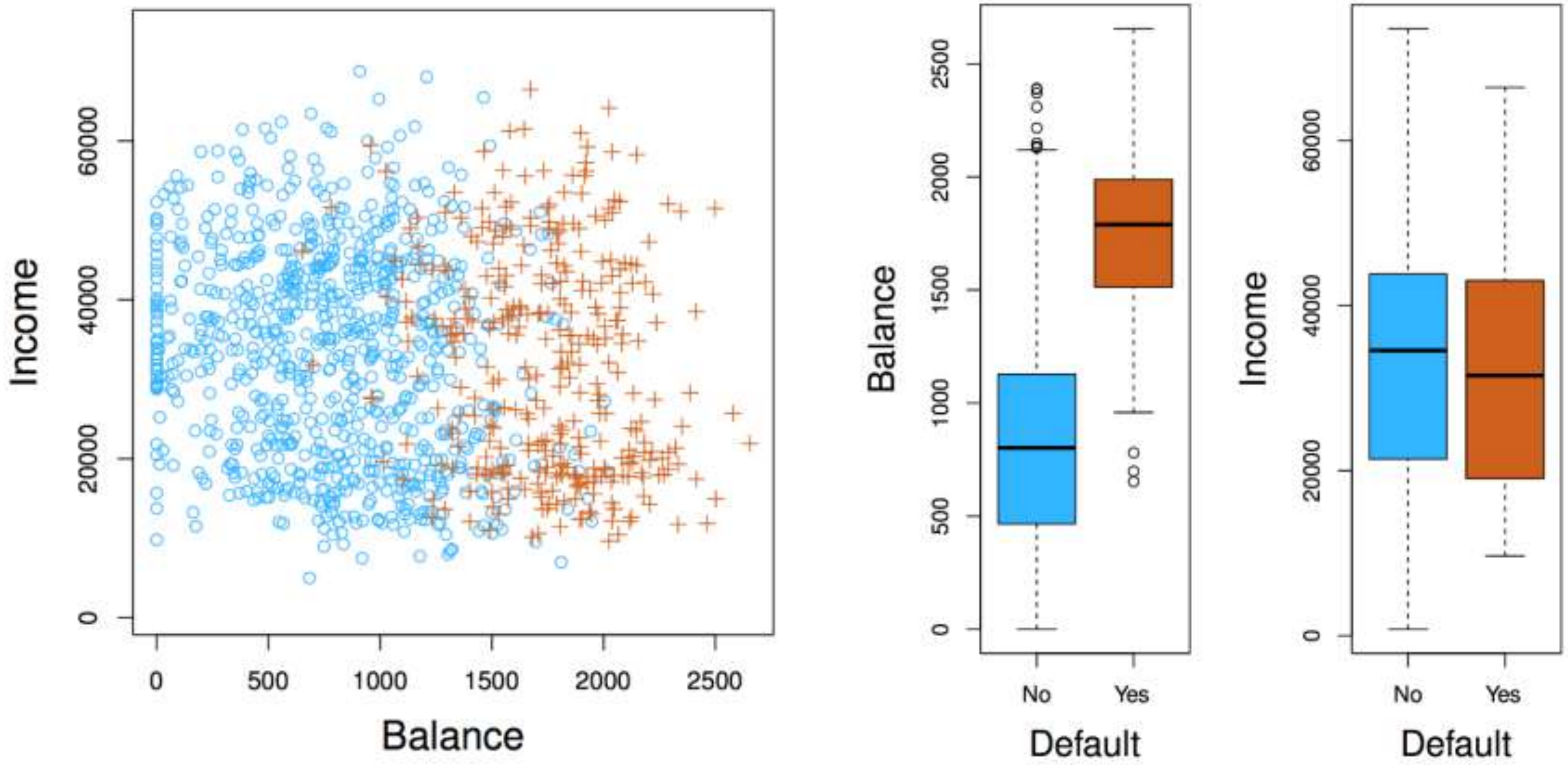
It provides prediction

It offers insight on the relative power of each variable

Case 2: Credit Card Default Data
Many independent variables

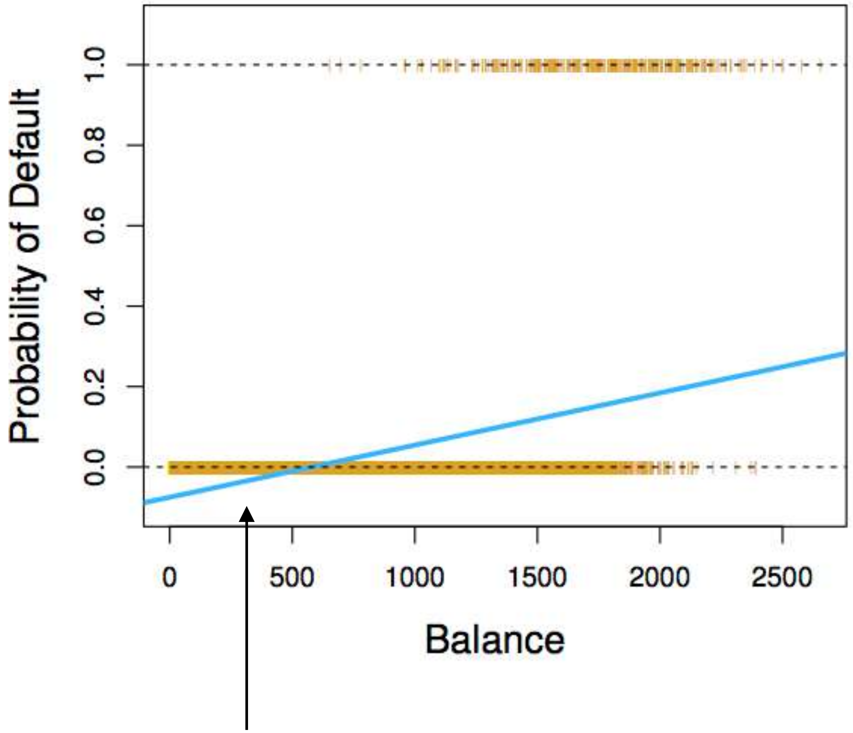
- We would like to be able to predict customers that are likely to default (to reach a negative balance)
- Possible X variables are:
  - Annual Income
  - Monthly credit card balance
- The Y variable (Default) is <u>categorical</u>: Yes or No
- How do we check the relationship between Y and X?

The Default Dataset



Why not Linear Regression?

- If we fit a linear regression to the Default data, then for very low balances we predict a negative probability, and for high balances we predict a probability above 1!

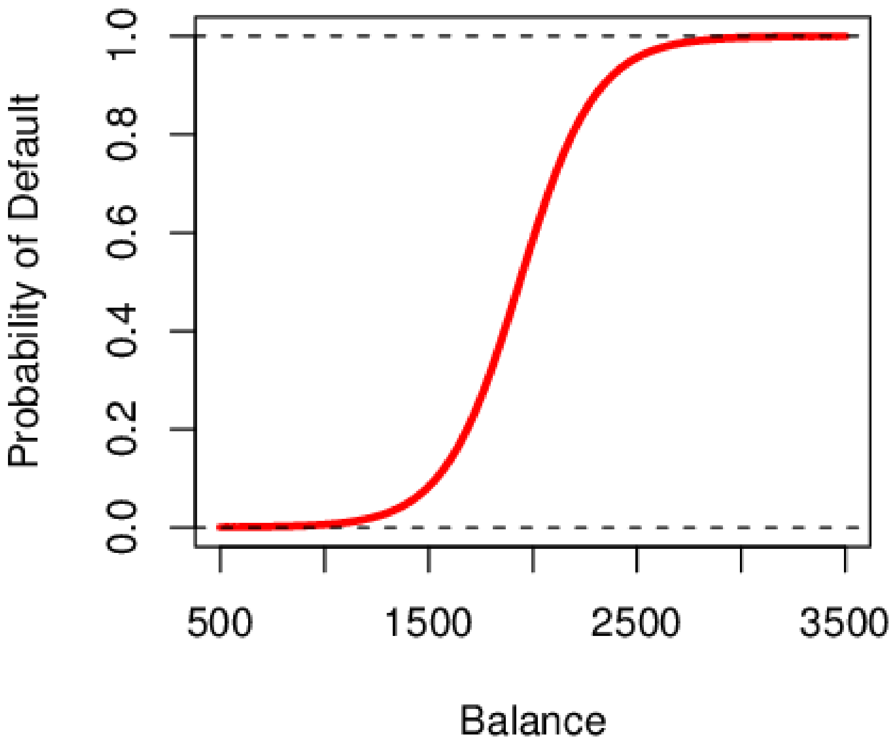


When Balance < 500,

Pr(default) is negative!

Logistic Function on Default Data

- Now the probability of default is close to, but not less than zero for low balances. And close to but not above 1 for high balances

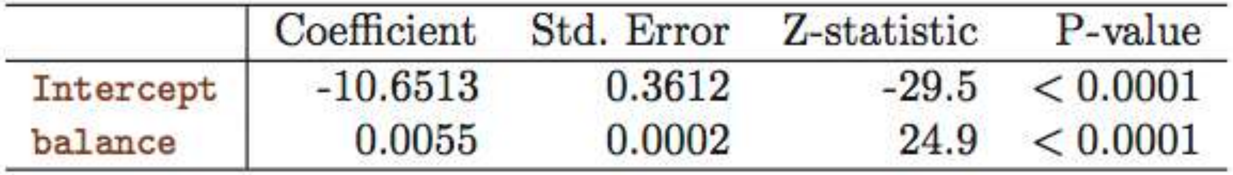


Interpreting β_1

- Interpreting what β_1 means is not very easy with logistic regression, simply because we are predicting P(Y) and not Y.
- If β_1 =0, this means that there is no relationship between Y and X.
- If β_1 >0, this means that when X gets larger so does the probability that Y = 1.
- If β_1 <0, this means that when X gets larger, the probability that Y = 1 gets smaller.
- But how much bigger or smaller depends on where we are on the slope

Are the coefficients significant?

- We still want to perform a hypothesis test to see whether we can be sure that are β_0 and β_1 significantly different from zero.
- We use a Z test instead of a T test, but of course that doesn't change the way we interpret the p-value
- Here the p-value for balance is very small, and b_1 is positive, so we are sure that if the balance increase, then the probability of default will increase as well.



	Coefficient	Std. Error	Z-statistic	P-value
`Intercept`	-10.6513	0.3612	-29.5	< 0.0001
`balance`	0.0055	0.0002	24.9	< 0.0001

Making Prediction

- Suppose an individual has an average balance of $1000. What is their probability of default?

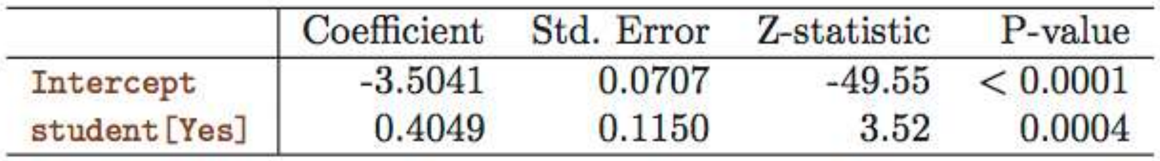
$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.00576$$

- The predicted probability of default for an individual with a balance of $1000 is less than 1%.

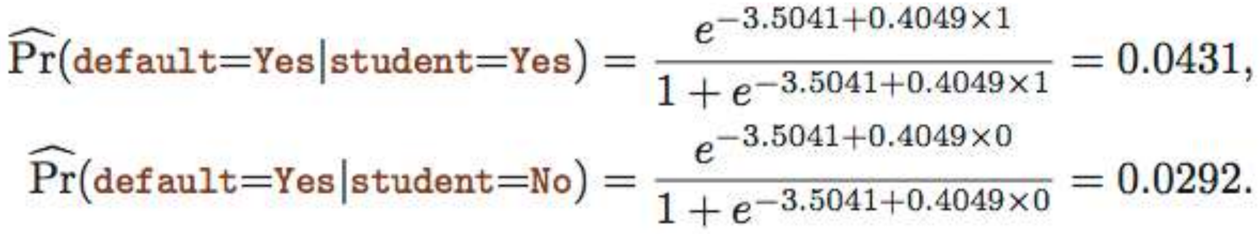
- For a balance of $2000, the probability is much higher, and equals to 0.586 (58.6%).

Qualitative Predictors in Logistic Regression

- We can predict if an individual default by checking if she is a student or not. Thus we can use a qualitative variable “Student” coded as (Student = 1, Non-student =0).
- b_1 is positive: This indicates students tend to have higher default probabilities than non-students



	Coefficient	Std. Error	Z-statistic	P-value
`Intercept`	-3.5041	0.0707	-49.55	< 0.0001
`student[Yes]`	0.4049	0.1150	3.52	0.0004



$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431,$$

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292.$$

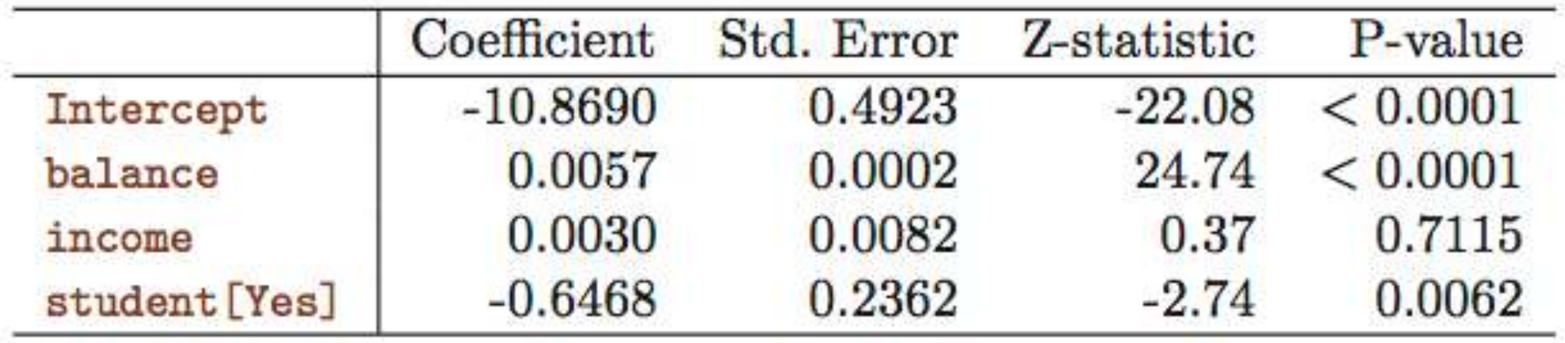
Multiple Logistic Regression

- We can fit multiple logistic just like regular regression

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}.$$

Multiple Logistic Regression- Default Data

- Predict Default using:
  - Balance (quantitative)
  - Income (quantitative)
  - Student (qualitative)



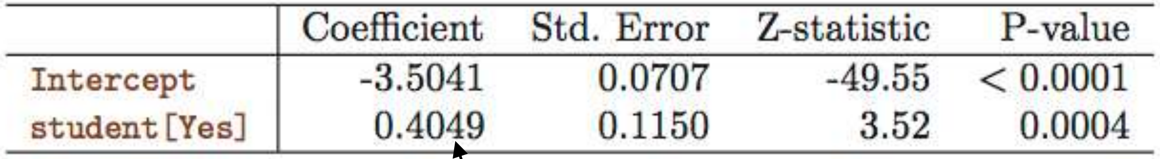
	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.8690	0.4923	-22.08	< 0.0001
balance	0.0057	0.0002	24.74	< 0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	-0.6468	0.2362	-2.74	0.0062

Predictions

- A student with a credit card balance of $1,500 and an income of $40,000 has an estimated probability of default

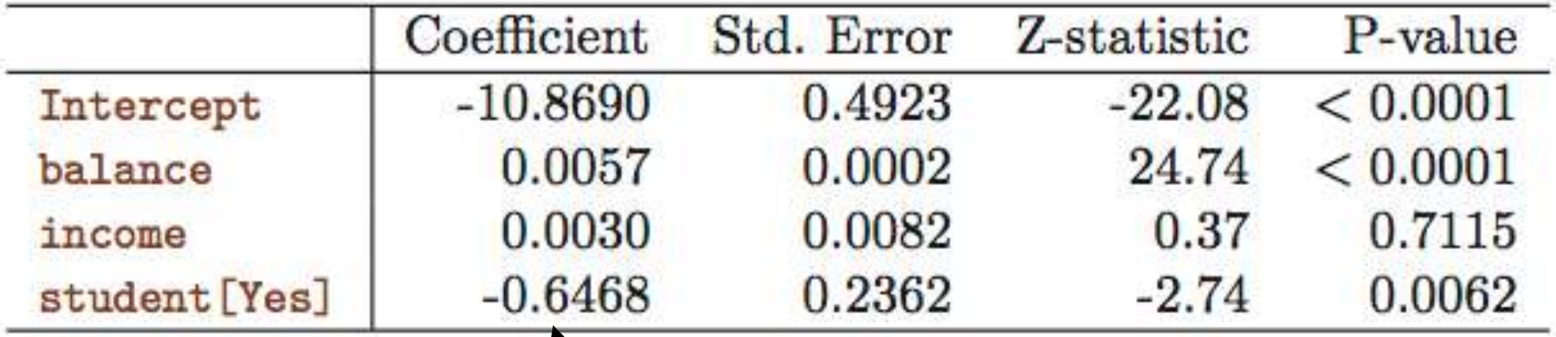
$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times 1500+0.003\times 40-0.6468\times 1}}{1+e^{-10.869+0.00574\times 1500+0.003\times 40-0.6468\times 1}} = 0.058.$$

An Apparent Contradiction!



	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-3.5041	0.0707	-49.55	< 0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

Positive

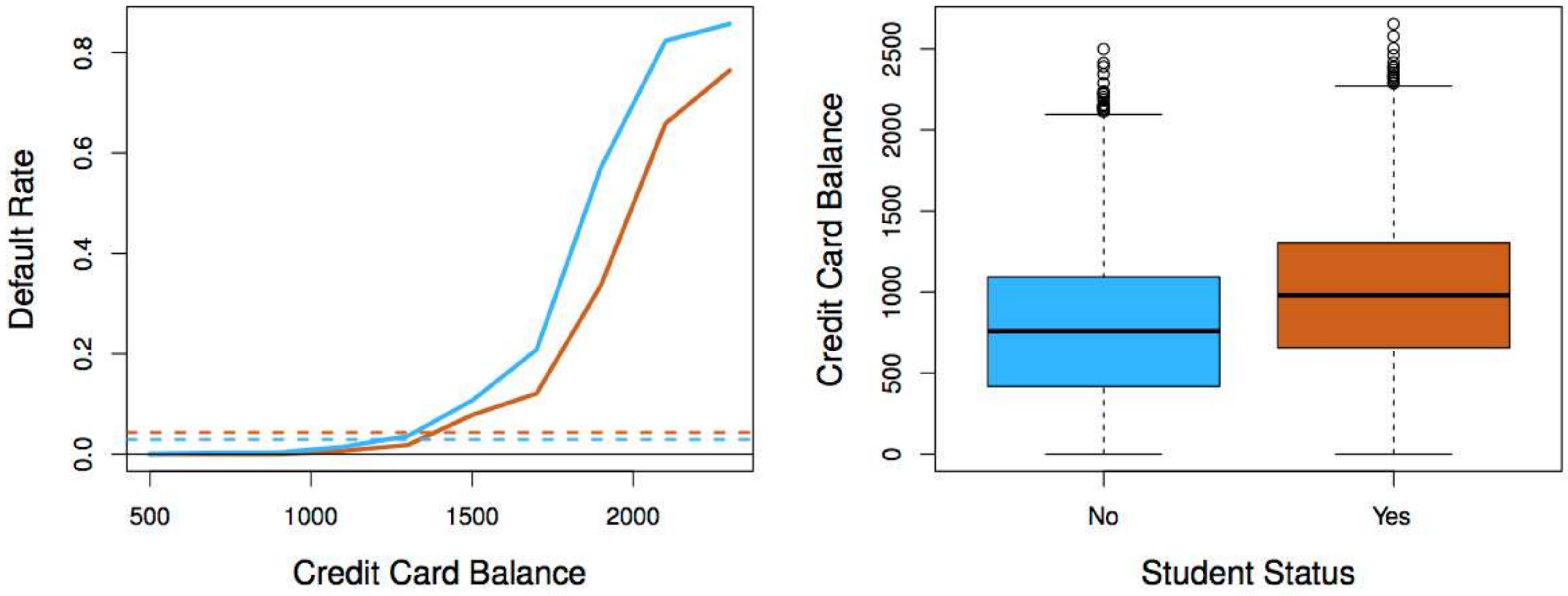


	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.8690	0.4923	-22.08	< 0.0001
balance	0.0057	0.0002	24.74	< 0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	-0.6468	0.2362	-2.74	0.0062

Negative

Students (Orange) vs. Non-students (Blue)

A student is risker than non students if no information about the credit card balance is available



However, that student is less risky than a non student with the same credit card balance!