Supervised Tools for Machine Learning and Data Mining

Logistic Regression

Dr. Abbas Rammal

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

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

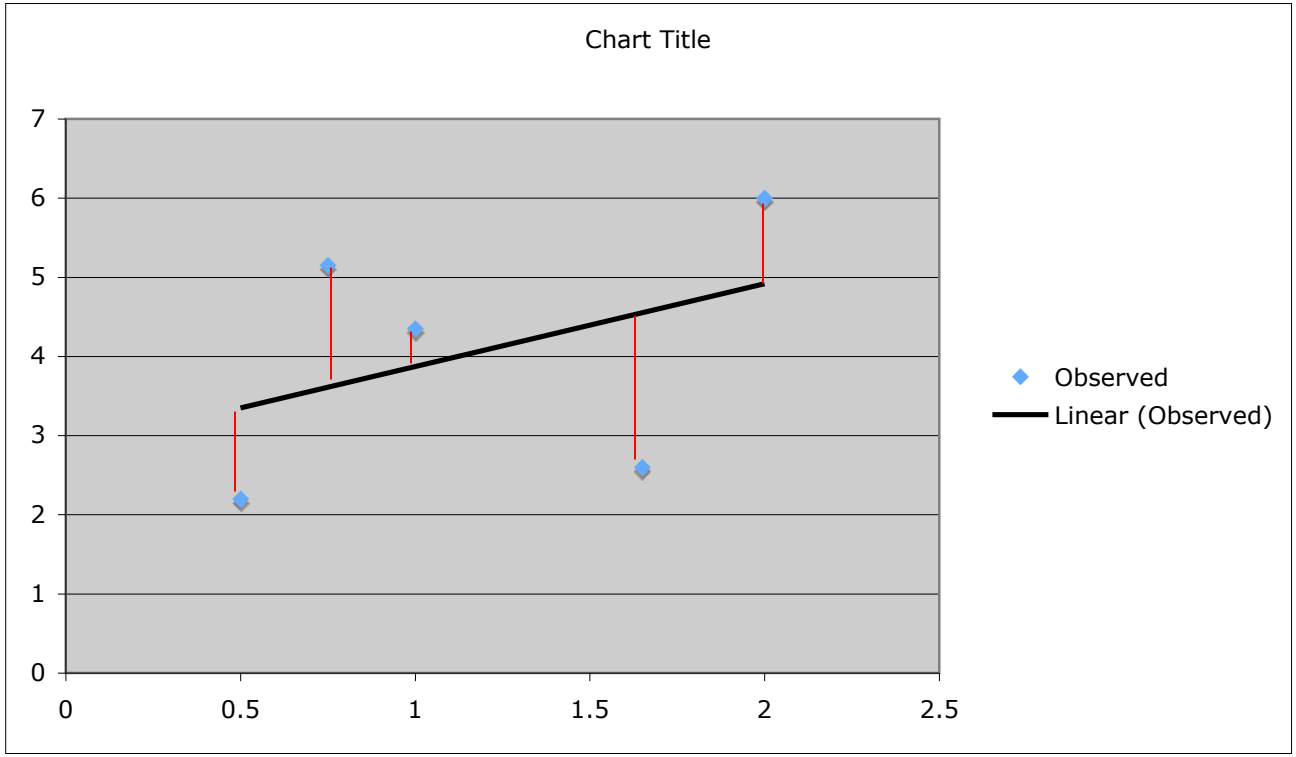
$$Y = b_0 + b_1 X_1 + b_2 X_2 + \square + b_n X_n + \epsilon$$

- 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

DML
BYU DATA MINING LAB

Residue

- Error or residue:
  - Observed value - Predicted value



DML
BYU DATA MINING LAB

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: $y=2x+1.5$

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

$$= \frac{7 \times 223.61 - 24.10 \times 58.00}{7 \times 95.31 - 24.10^2}$$

$$= \frac{1565.27 - 1397.80}{667.17 - 580.81}$$

$$= \frac{167.47}{86.36} = \underline{\underline{1.94}}$$

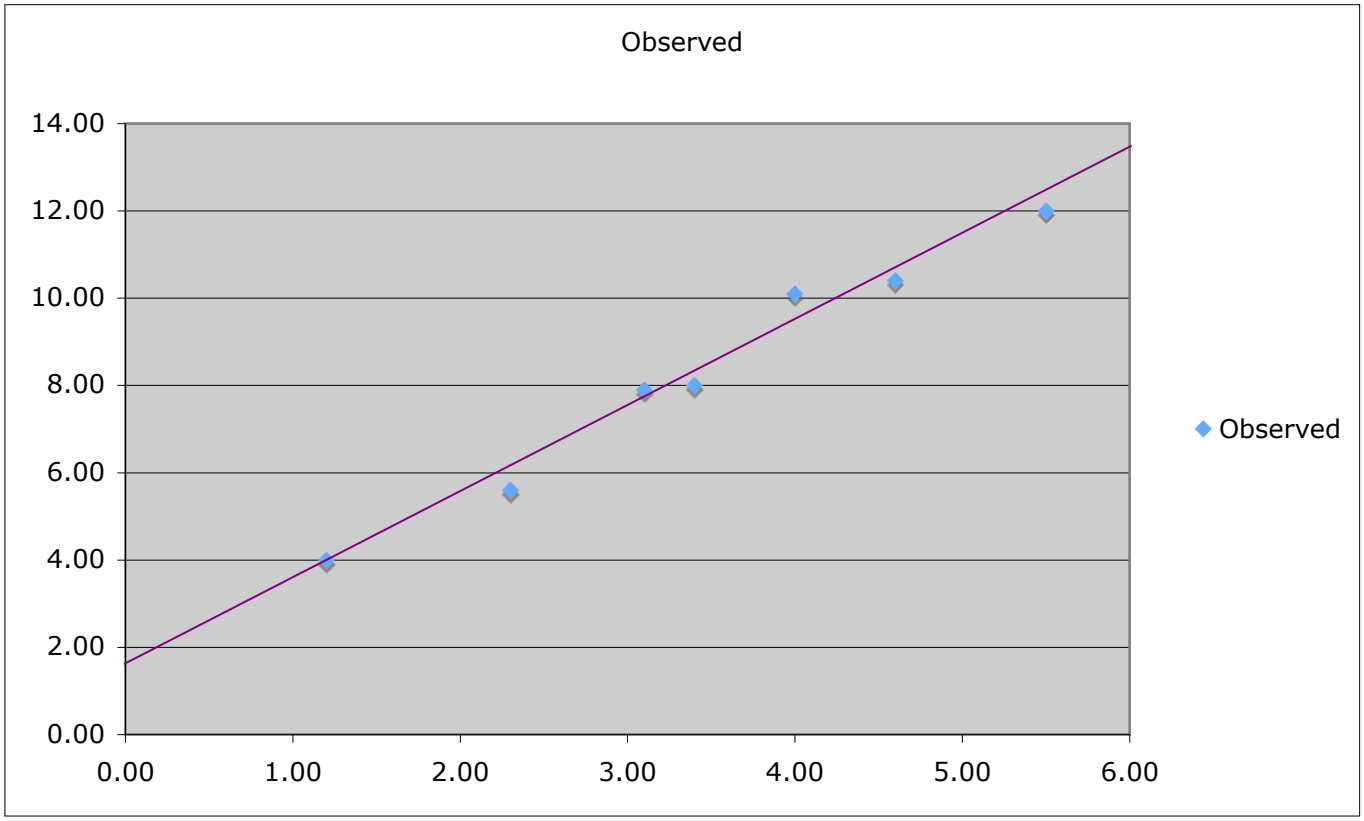
$$b_0 = \frac{\sum y - b_1 \sum x}{n}$$

$$= \frac{58.00 - 1.94 \times 24.10}{7}$$

$$= \frac{11.27}{7} = \underline{\underline{1.61}}$$

DML
BYU DATA MINING LAB

Example (II)



DML
BYU DATA MINING LAB

Example (III)

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
			0.975	**47.369**

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

DML
BYU DATA MINING LAB

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

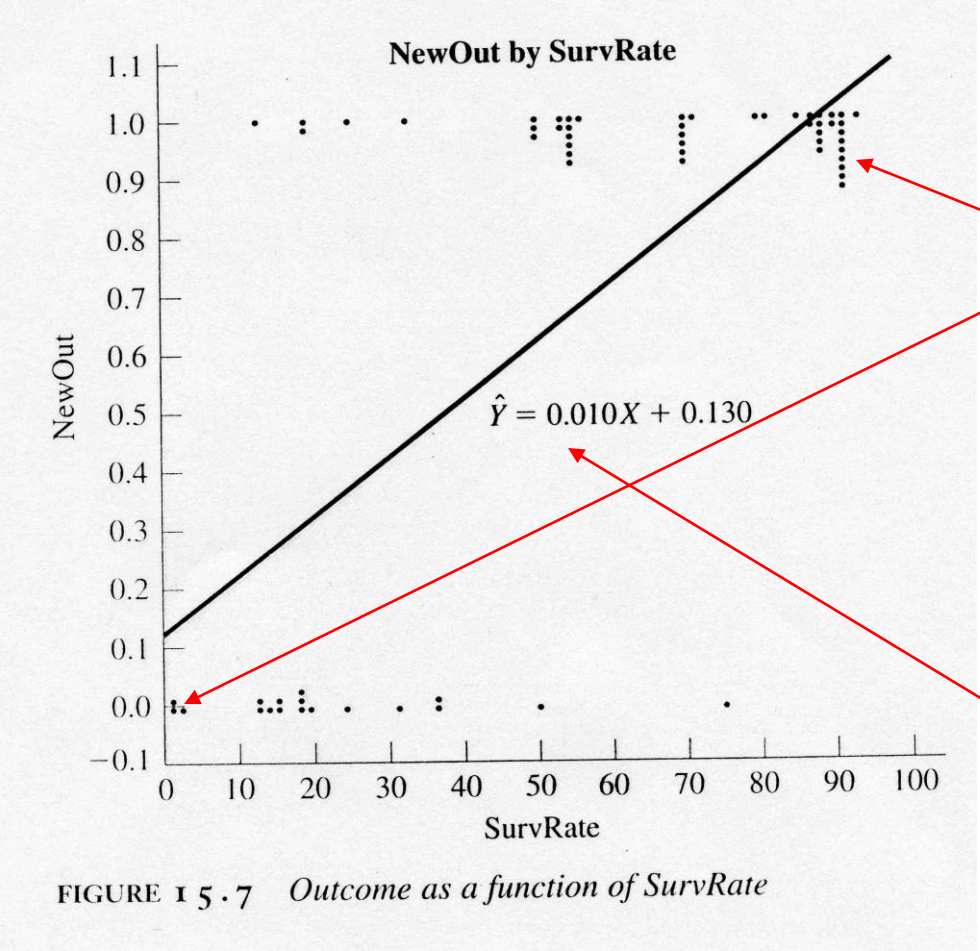


FIGURE 15.7 *Outcome as a function of SurvRate*

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]

DML
BYU DATA MINING LAB

A Better Solution

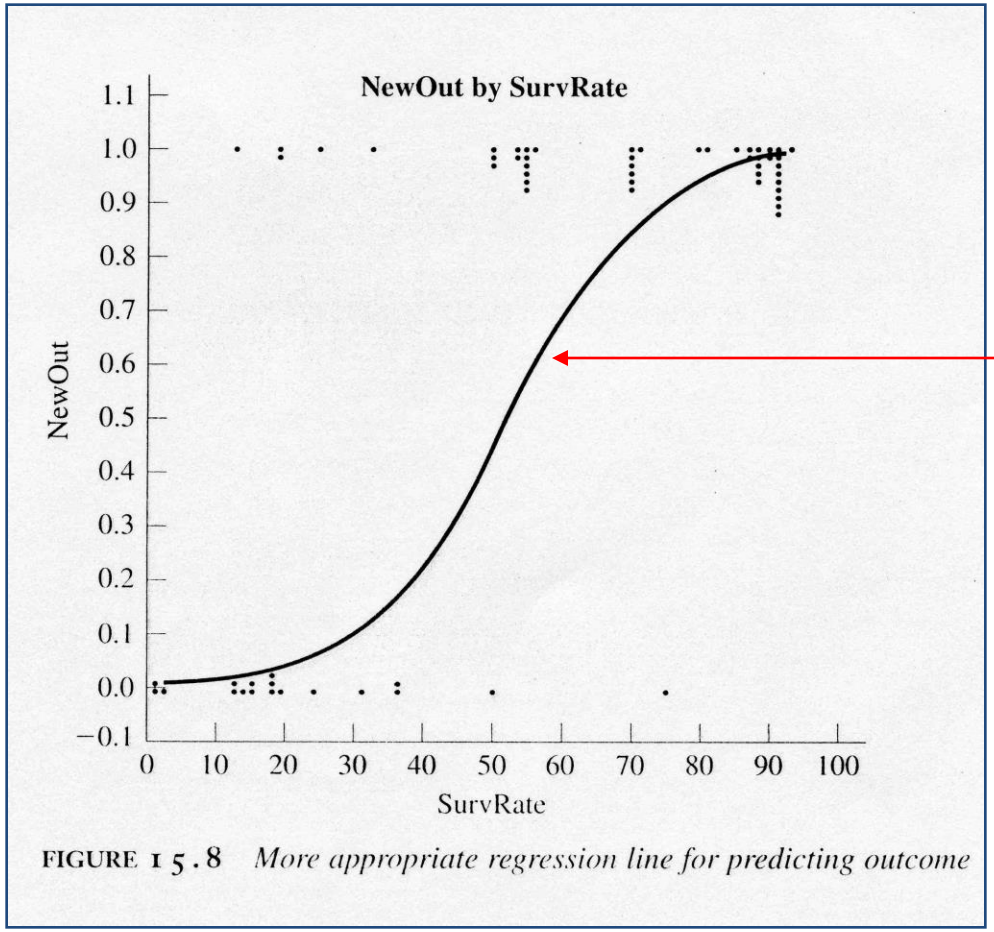


FIGURE 15.8 *More appropriate regression line for predicting outcome*

Regression Curve: Sigmoid function!

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

DML BYU DATA MINING LAB

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

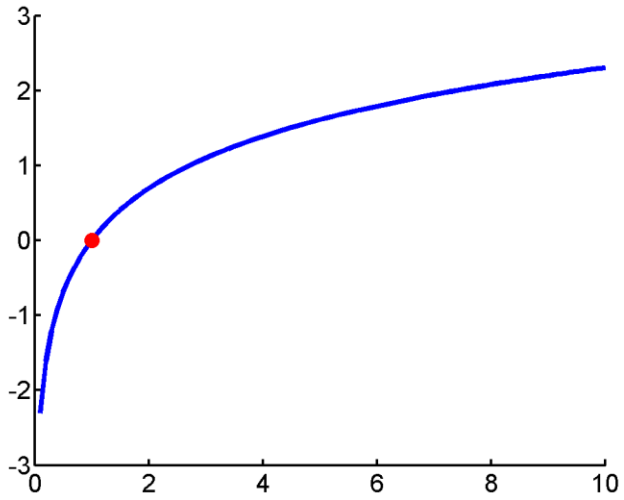
- The odds of being delinquent if you are in the Normal group are:

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

DML BYU DATA MINING LAB

Logit Transform

- The logit is the natural log of the odds



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

DML
BYU DATA MINING LAB

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

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

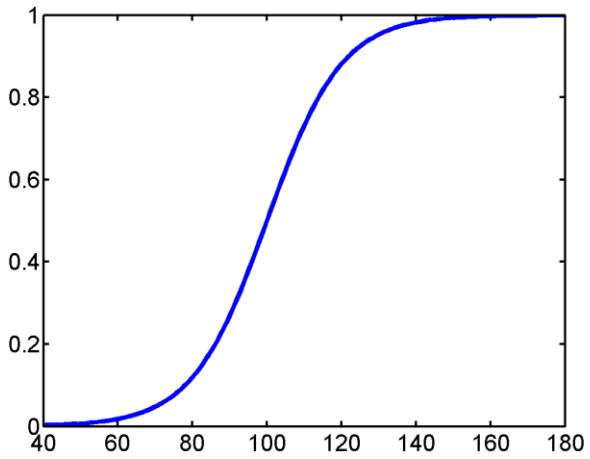
$$\Leftrightarrow \frac{p}{1-p} = e^{\beta_0 + \beta_1 X}$$

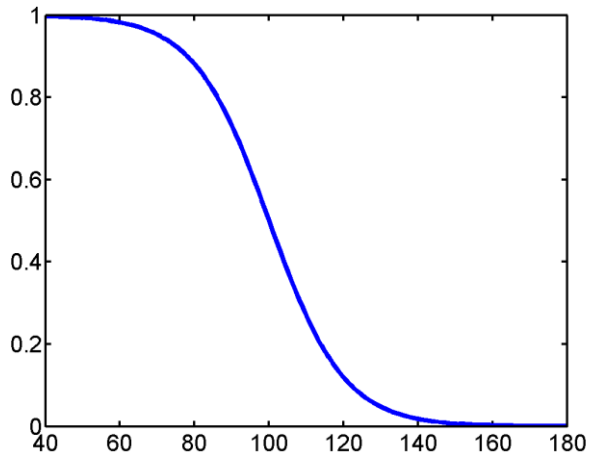
$$\Leftrightarrow p = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

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:





DML
BYU DATA MINING LAB

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*

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.)

DML
BYU DATA MINING LAB

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**

DML
BYU DATA MINING LAB

17.31 Analysis of a reduction in force. To meet competition or cope with economic slowdowns, corporations sometimes undertake a "reduction in force" (RIF), in which substantial numbers of employees are terminated. Federal and various state laws require that employees be treated equally regardless of their age. In particular, employees over the age of 40 years are in a "protected" class, and many allegations of discrimination focus on comparing employees over 40 with their younger coworkers. Here are the data for a recent RIF:

	Over 40	
Terminated	**No**	**Yes**
Yes	17	71
No	564	835

(a) Write the logistic regression model for this problem using the log odds of a termination as the response variable and an indicator for over and under 40 years of age as the explanatory variable.
(b) Explain the assumption concerning binomial distributions in terms of the variables in this exercise. To what extent do you think that these assumptions are reasonable?
(c) Software gives the estimated slope $b_1 = 1.0371$ and its standard error $SE_{b_1} = 0.2755$. Transform the results to the odds scale. Summarize the results and write a short conclusion.
(d) If additional explanatory variables were available, for example, a performance evaluation, how would you use this information to study the RIF?

17.31 **(a)** $\log(\text{odds}) = -3.5017 + 1.0369x$. **(b)** The binomial distribution assumes that each employee's termination is independent from one another's and the probability of being terminated is the same for each employee. Certainly the latter is not true because an individual's performance is likely different and largely determines whether or not they are terminated. **(c)** odds = 2.82, with 95% the confidence interval is (1.644, 4.840). Because the interval does not contain 1, the results are significant at the 5% level. Employees over 40 are 2.82 times more likely to be terminated than those under 40. **(d)** We could use the additional variables in the logistic regression model to account for their effects before assessing if age has an effect.

Summary

- Regression is a powerful data mining technique
  - It provides prediction
  - It offers insight on the relative power of each variable
- We have focused on the case of a single independent variable
  - What about the general case?