SEMATECH

1997 Statistical Methods Symposium

Austin

Regression Models for a Binary Response Using EXCEL and JMP

David C. Trindade, Ph.D.

STAT-TECH

Consulting and Training in Applied Statistics

San Jose, CA

Topics

- Practical Examples
- Properties of a Binary Response
- Linear Regression Models for Binary Responses
  - Simple Straight Line
  - Weighted Least Squares
- Regression in EXCEL and JMP
- Logistic Response Function
- Logistic Regression
  - Repeated Observations (Grouped Data)
  - Individual Observations
- Logit Analysis in EXCEL and JMP
- Conclusion

Practical Examples: Binary Responses

Consider the following situations:

- A weatherman would like to understand if the probability of a rainy day occurring depends on atmospheric pressure, temperature, or relative humidity
- A doctor wants to estimate the chance of a stroke incident as a function of blood pressure or weight
- An engineer is interested in the likelihood of a device failing functionality based on specific parametric readings

More Practical Examples

- The corrections department is trying to learn if the number of inmate training hours affects the probability of released prisoners returning to jail (recidivism)
- The military is interested in the probability of a missile destroying an incoming target as a function of the speed of the target
- A real estate agency is concerned with measuring the likelihood of selling property given the income of various clients
- An equipment manufacturer is investigating reliability after six months of operation using different spin rates or temperature settings

Binary Responses

- In all these examples, the dependent variable is a binary indicator response, taking on the values of either 0 or 1, depending on which of of two categories the response falls into: success-failure, yes-no, rainy-dry, target hit-target missed, etc.
- We are interested in determining the role of explanatory or regressor variables X_1, X_2, … on the binary response for purposes of prediction.

Simple Linear Regression

Consider the simple linear regression model for a binary response:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where the indicator variable Y_i = 0, 1.
Since $E(\varepsilon_i) = 0$, the mean response is

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

Interpretation of Binary Response

- Since Y_i can take on only the values 0 and 1, we choose the Bernoulli distribution for the probability model.
- Thus, the probability that $Y_i = 1$ is the mean p_i and the probability that $Y_i = 0$ is 1- p_i.
- The mean response

$$E(Y_i) = 1 \times p_i + 0 \times (1 - p_i) = p_i$$

is thus interpreted as the **probability** that $Y_i = 1$ when the regressor variable is X_i.

Model Considerations

Consider the variance of Y_i for a given X_i :

$$V\left(Y_i | X_i\right) = V\left(\beta_0 + \beta_1 X_i + \varepsilon_i | X_i\right) = V\left(\varepsilon_i | X_i\right)$$
$$= p_i\left(1 - p_i\right) = \left(\beta_0 + \beta_1 X_i\right)\left(1 - \beta_0 - \beta_1 X_i\right)$$

We see the **variance is not constant** since it depends on the value of X_i. This is a violation of basic regression assumptions.

- Solution: Use **weighted least squares regression** in which the weights selected are inversely proportional to the variance of Y_i, where

$$\mathrm{Var}(Y_i) = \hat{Y}_i\left(1 - \hat{Y}_i\right)$$

Distribution of Errors

- Note also that the errors cannot be normally distributed since there are only two possible values (0 or 1) for ε_i at each regressor level.

- Fitted model should have the property that the predicted responses lie between 0 and 1 for all X_i within the range of original data. No guarantee that the simple linear model will have this behavior.

Example 1: Missile Test Data*

The table shows the results of test-firing 25 ground to air missiles at targets of various speeds. A “1” is a hit and a “0” is a miss.

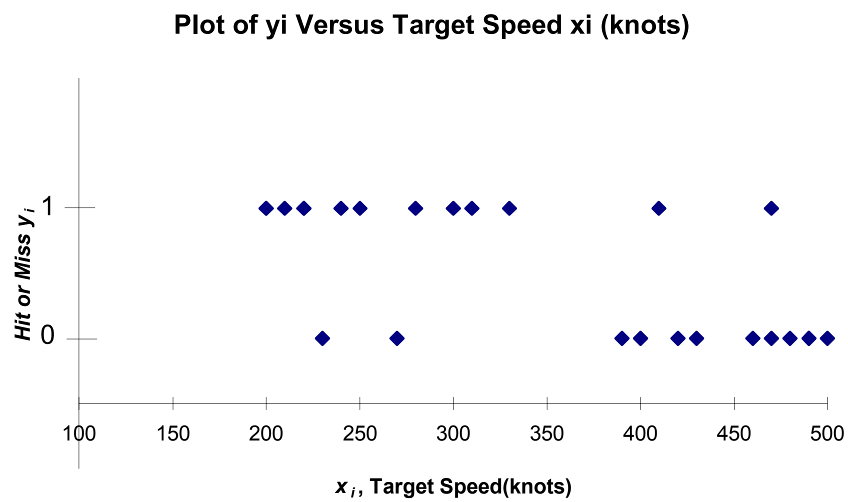
Test Firing I	Target Speed (knots) xi	Hit or Miss yi
1	400	0
2	220	1
3	490	0
4	410	1
5	500	0
6	270	0
7	200	1
8	470	0
9	480	0
10	310	1
11	240	1
12	490	0
13	420	0
14	330	1
15	280	1
16	210	1
17	300	1
18	470	1
19	230	0
20	430	0
21	460	0
22	220	1
23	250	1
24	200	1
25	390	0

* Example from Montgomery & Peck, *Introduction to Linear Regression Analysis*, 2nd Ed. Table 6.4

EXCEL Plot of Data

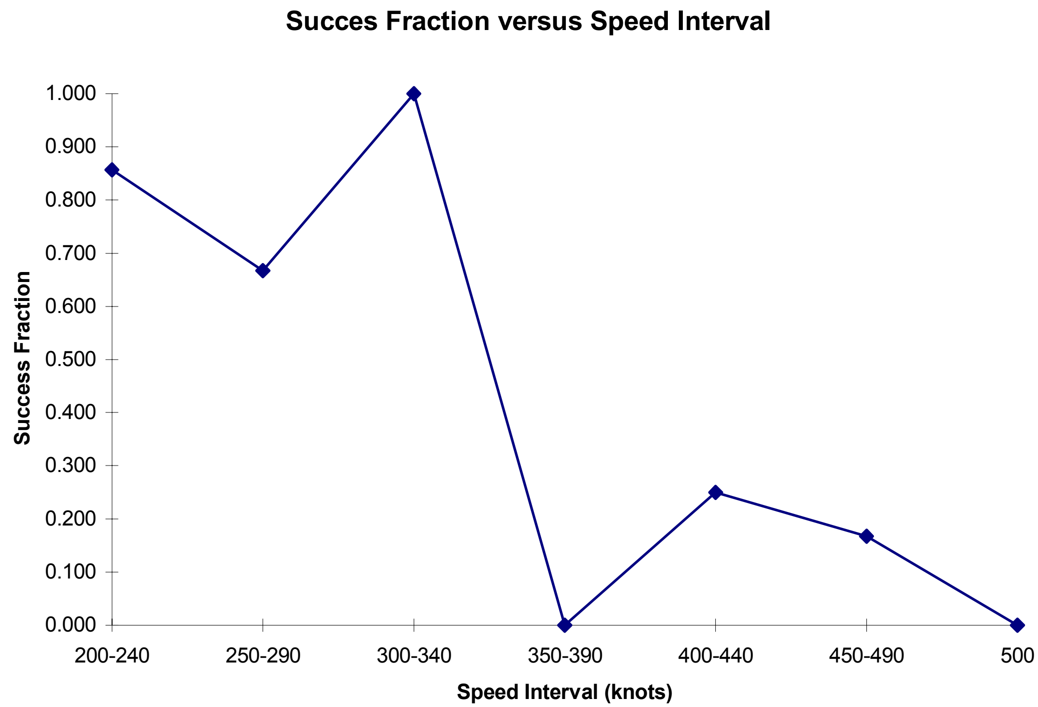
There appears to be a tendency for misses to increase with increasing target speed.

Let us group the data to reveal the association better.



Grouped Data

Speed Interval	Number of Attempts	Number of Hits	Fraction Success
200-240	7	6	0.857
250-290	3	2	0.667
300-340	3	3	1.000
350-390	1	0	0.000
400-440	4	1	0.250
450-490	6	1	0.167
500	1	0	0.000
Sum	25	13	



Clearly, the probability of a hit seems to decrease with speed. We will fit a straight-line model to the data using weighted least squares.

Weighted Least Squares

- We will use the inverse of the variance of Y_i for the weights w_i . **Problem**: these are not known because they are a function of the unknown parameters β_0, β_1 in the regression model. That is, the weights w_i are:

$$w_i = \frac{1}{V(Y_i | X_i)} = \frac{1}{p_i(1-p_i)} = \frac{1}{(\beta_0 + \beta_1 X_i)(1-\beta_0 - \beta_1 X_i)}$$

- **Solution**: We can initially estimate β_0, β_1 using ordinary (unweighted) LS. Then, we calculate the weights with these estimates and solve for the weighted LS coefficients. One iteration usually suffices.

Simple Linear Regression in EXCEL

Several methods exist:

- Use “Regression” macro in “**Data Analysis Tools**.”
- Use “**Function**” button to pull up “Slope” and “Intercept” under “Statistical” listings. Sort data first by regressor variable.
- Click on data points in plot of Y_i vs. X_i, select menubar “Insert” followed by “**Trendline**”. In dialog box, select options tab and choose “Display equation on chart.”
- Use EXCEL **array tools** (transpose, minverse, and mmult) to define and manipulate matrices. (Requires Cntrl-Shift-Enter for array entry.)

EXCEL Data Analysis Tools

Output:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.64673
R Square	0.41826
Adjusted R Square	0.39296
Standard Error	0.39728
Observations	25

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	2.60991	2.60991	16.53624	0.0004769
Residual	23	3.63009	0.15783		
Total	24	6.24			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1.56228	0.26834	5.82194	0.00001	1.00717	2.11739	1.00717	2.11739
Target Speed (knot	-0.00301	0.00074	-4.06648	0.00048	-0.00453	-0.00148	-0.00453	-0.00148

Can also display residuals and various plots.

EXCEL Functions

Sorted data.

Target Speed (knots) xi	Hit or Miss yi
200	1
200	1
210	1
220	1
220	1
230	0
240	1
250	1
270	0
280	1
300	1
310	1
330	1
390	0
400	0
410	1
420	0
430	0
460	0
470	0
470	1
480	0
490	0
490	0
500	0

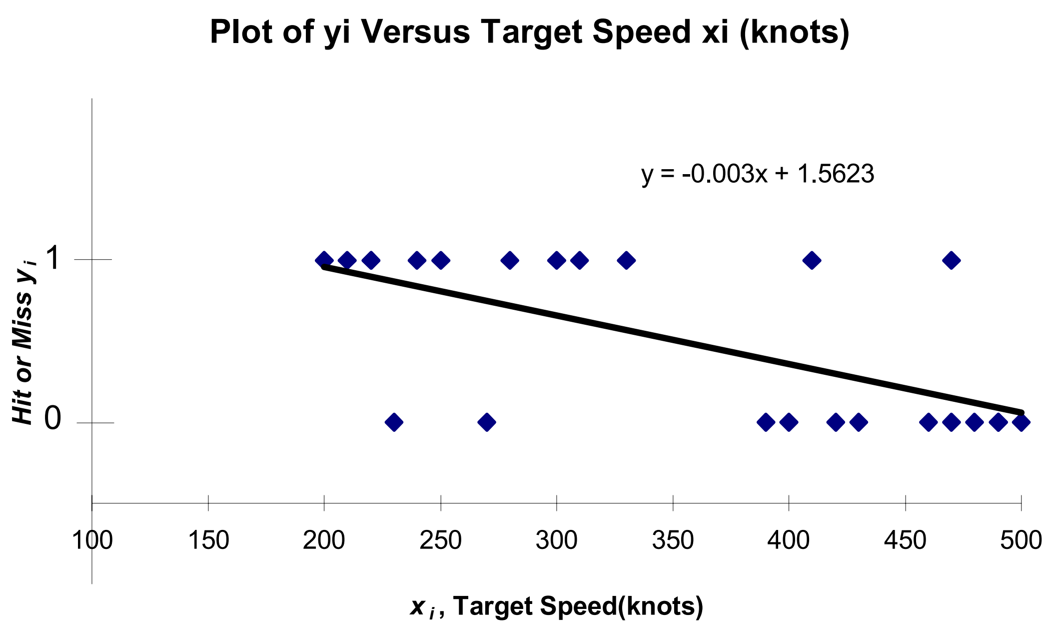
=intercept(*y* column, *x* column)

=slope(*y* column, *x* column)

Output:

Intercept	1.562282
Slope	-0.00301

EXCEL Equation on Chart



EXCEL Array Functions

Three key functions:

=transpose(range)

=mmult(range1, range2)

=minverse(range)

Requires Cntrl-Shift-Enter each time.

EXCEL Matrix Manipulation

Define the design matrix X by adding a column of “1”s for the constant in the model.

Then, progressively calculate:

- the transpose X’
- the product X’X
- the inverse of X’X
- the product X’Y
- the LS regression **coefficients** = $(X'X)^{-1}(X'Y)$

The standard errors of the coefficients can be obtained from the square root of the diagonal elements of the variance-covariance matrix: MSE x $(X'X)^{-1}$. Find MSE from the residuals SS and df.

EXCEL Matrix Example

X

1	200
1	200
1	210
1	220
1	220
1	230
1	240
1	250
1	270
1	280
1	300
1	310
1	330
1	390
1	400
1	410
1	420
1	430
1	460
1	470
1	470
1	480
1	490
1	490
1	500

Y

1
1
1
1
1
0
1
1
0
1
1
1
1
0
0
1
0
0
0
0
1
0
0
0
0

X'X

25	8670
8670	3295700

X'Y

13
3640

$[X'X]^{-1}$

0.456241	-0.0012
-0.0012	3.46E-06

Coefficients = $[X'X]^{-1}$ X'Y

β_0	1.562282
β_1	-0.00301

X'

1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
200	200	210	220	220	230	240	250	270	280	300	310	330	390	400	410	420	430

EXCEL Matrix Example
Standard Errors

Target Speed (knots) xi	Hit or Miss yi	Y pred	Residuals
200	1	0.9612	0.0388
200	1	0.9612	0.0388
210	1	0.9311	0.0689
220	1	0.9011	0.0989
220	1	0.9011	0.0989
230	0	0.8710	-0.8710
240	1	0.8410	0.1590
250	1	0.8109	0.1891
270	0	0.7508	-0.7508
280	1	0.7208	0.2792
300	1	0.6607	0.3393
310	1	0.6306	0.3694
330	1	0.5705	0.4295
390	0	0.3902	-0.3902
400	0	0.3601	-0.3601
410	1	0.3301	0.6699
420	0	0.3000	-0.3000
430	0	0.2699	-0.2699
460	0	0.1798	-0.1798
470	0	0.1497	-0.1497
470	1	0.1497	0.8503
480	0	0.1197	-0.1197
490	0	0.0896	-0.0896
490	0	0.0896	-0.0896
500	0	0.0596	-0.0596
			SS Residuals
			3.630087
			DF
			23
			MSE
			0.15783

$[X'X]^{-1}$

0.456241	-0.0012
-0.0012	3.46E-06

MSE x $[X'X]^{-1}$

0.072008	-0.00019
-0.00019	5.46E-07

Standard Errors of Coefficients

β_0	0.268344
β_1	0.000739

Fitted model appears adequate since all *Y* predictions are between 0 and 1. If not, would need non-linear model.

Simple Linear Regression in JMP

- Specify number of rows for data
- Set up X column
- Set up Y column
- Select under “Analyze” “Fit Y by X”
- For multiple regression, select under “Analyze” “Fit Model”

Data Table in JMP

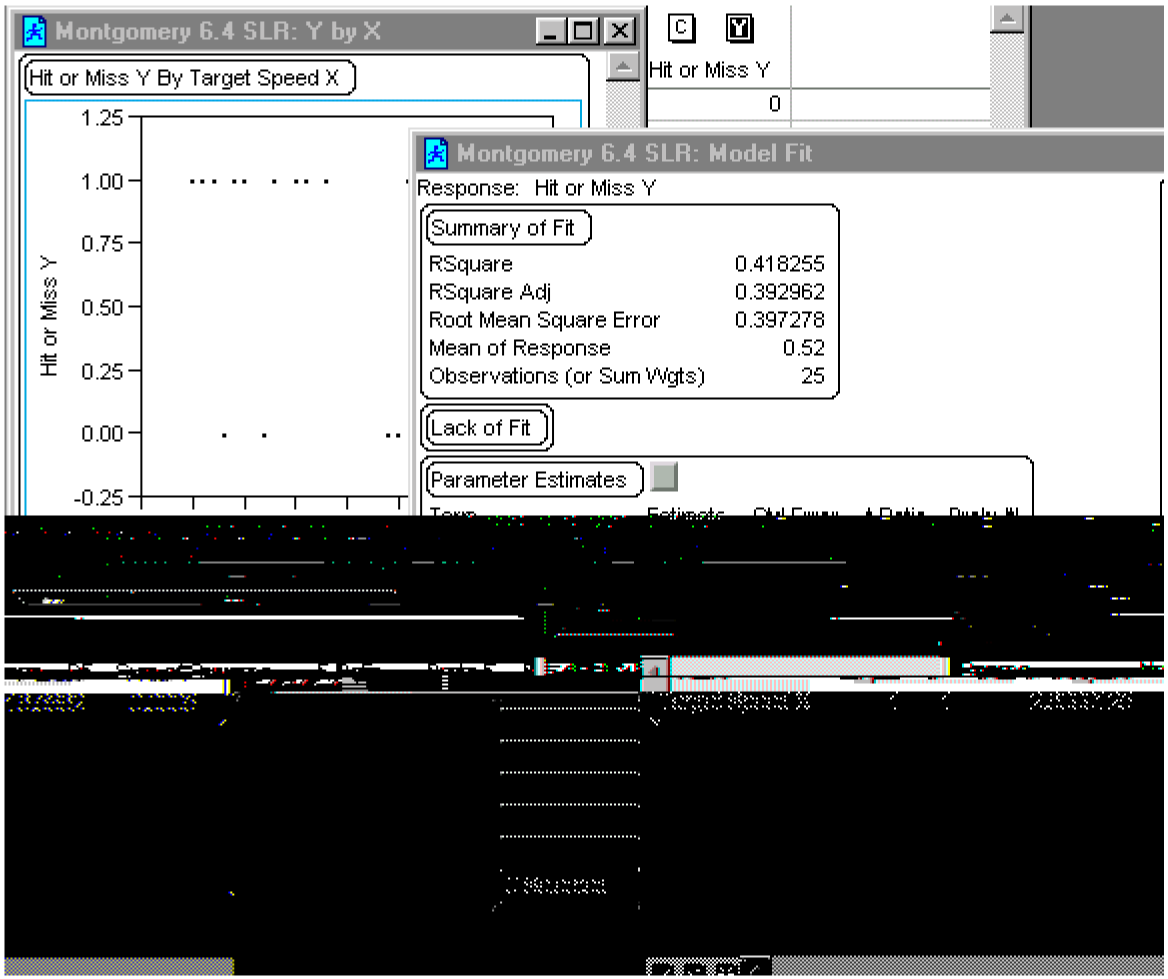
Note that Y is specified "C" for continuous at this point.

Montgomery 6.4 SLR

2 Cols / 25 Rows	C X Target Speed X	C Y Hit or Miss Y
1	400	0
2	220	1
3	490	0
4	410	1
5	500	0
6	270	0
7	200	1
8	470	0
9	480	0
10	310	1
11	240	1
12	490	0
13	420	0
14	330	1
15	280	1
16	210	1
17	300	1
18	470	1
19	230	0
20	430	0
21	460	0
22	220	1
23	250	1
24	200	1

0 Selected

Fit Model in JMP



Weighted Least Squares Regression

In weighted least squares regression, the squared deviation between the observed and predicted value (that is, the squared residual) is multiplied by weights w_i that are inversely proportional to Y_i. We then minimize the following function with respect to the coefficients β_0, β_1 :

$$SS_w = \sum_{i=1}^{n} w_i \left(Y_i - \beta_0 - \beta_1 X_i\right)^2$$

Weighted LS Regression in EXCEL

Several methods exist:

- Transform all variables, including constant. Use "Regression" macro in "**Data Analysis Tools**" with **no intercept**
- Use "**Solver**" routine on sum of squares of weighted residuals
- Use EXCEL **array tools** (transpose, minverse, and mmult) to define and manipulate matrices. (Requires Cntrl-Shift-Enter for array entry.)

Transform Method for Weighted Least Squares

Transform the variables by dividing each term in the model by the square root of the variance of Y_i.

$$SS_w = \sum_{i=1}^{n} w_i \left(Y_i - \beta_0 - \beta_1 X_i\right)^2$$

$$= \sum_{i=1}^{n} \left(\frac{Y_i}{\sqrt{var\ Y_i}} - \beta_0 \frac{1}{\sqrt{var\ Y_i}} - \beta_1 \frac{X_i}{\sqrt{var\ Y_i}} \right)^2$$

$$= \sum_{i=1}^{n} \left(Y_i' - \beta_0 Z_i - \beta_1 X_i'\right)^2$$

Transformed Variables

The expression below can be solved using ordinary LS multiple regression with the intercept (constant term) equal to zero.

$$SS_w = \sum_{i=1}^{n} \left(Y_i^{'} - \beta_0 Z_i - \beta_1 X_i^{'} \right)^2$$

Transforming Variables

							Transformed Factors		
Test Firing I	Constant	Target Speed (knots) xi	Hit or Miss yi	Y_i Pred	Var(Y) = $(Y_i)*(1-Y_i)$	T = 1/sqrt[Var(y)]	Constant T*Cnst	X = T*X	Y = T*yi
1	1	400	0	0.3601	0.2304	2.083	2.0832	833.3	0.0000
2	1	220	1	0.9011	0.0891	3.350	3.3496	736.9	3.3496
3	1	490	0	0.0896	0.0816	3.501	3.5009	1715.4	0.0000
4	1	410	1	0.3301	0.2211	2.127	2.1266	871.9	2.1266
5	1	500	0	0.0596	0.0560	4.225	4.2250	2112.5	0.0000
6	1	270	0	0.7508	0.1871	2.312	2.3119	624.2	0.0000
7	1	200	1	0.9612	0.0373	5.178	5.1780	1035.6	5.1780
8	1	470	0	0.1497	0.1273	2.803	2.8026	1317.2	0.0000
9	1	480	0	0.1197	0.1054	3.081	3.0809	1478.8	0.0000
10	1	310	1	0.6306	0.2329	2.072	2.0719	642.3	2.0719
11	1	240	1	0.8410	0.1337	2.735	2.7345	656.3	2.7345
12	1	490	0	0.0896	0.0816	3.501	3.5009	1715.4	0.0000
13	1	420	0	0.3000	0.2100	2.182	2.1822	916.5	0.0000
14	1	330	1	0.5705	0.2450	2.020	2.0202	666.7	2.0202
15	1	280	1	0.7208	0.2013	2.229	2.2290	624.1	2.2290
16	1	210	1	0.9311	0.0641	3.949	3.9493	829.3	3.9493
17	1	300	1	0.6607	0.2242	2.112	2.1120	633.6	2.1120
18	1	470	1	0.1497	0.1273	2.803	2.8026	1317.2	2.8026
19	1	230	0	0.8710	0.1123	2.984	2.9836	686.2	0.0000
20	1	430	0	0.2699	0.1971	2.253	2.2526	968.6	0.0000
21	1	460	0	0.1798	0.1475	2.604	2.6041	1197.9	0.0000
22	1	220	1	0.9011	0.0891	3.350	3.3496	736.9	3.3496
23	1	250	1	0.8109	0.1533	2.554	2.5538	638.5	2.5538
24	1	200	1	0.9612	0.0373	5.178	5.1780	1035.6	5.1780
25	1	390	0	0.3902	0.2379	2.050	2.0501	799.5	0.0000
			LS Coeff						
			b0	1.56228					
			b1	-0.00301					

EXCEL Data Analysis Regression on Transformed Factors (Intercept =0)

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8252
R Square	0.6809
Adjusted R Square	0.6235
Standard Error	1.0060
Observations	25

ANOVA

	df	SS	MS	F	Significance F
Regression	2	49.6625	24.8312	24.5376	2.4989E-06
Residual	23	23.2753	1.0120		
Total	25	72.9377			

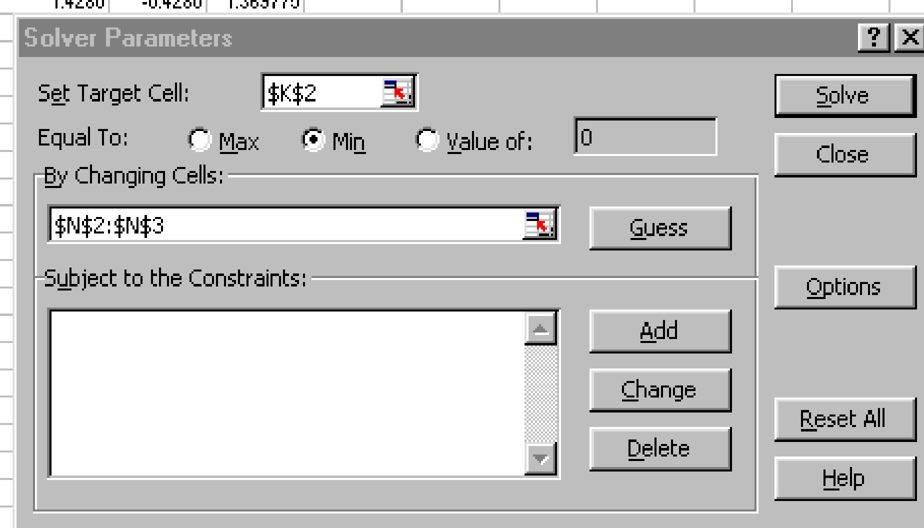
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Constant T*Cnst	1.586687	0.185892	8.535539	1.39E-08	1.20214	1.97123	1.20214	1.97123
X = T*X	-0.003091	0.000533	-5.79882	6.58E-06	-0.00419	-0.00199	-0.00419	-0.00199

Weighted Least Squares Analysis Using Solver

- Use the unweighted LS coefficients to predict *Y*.
- Calculate the variance of Y_i based on predicted *Y* in equation $Y_i(1- Y_i)$
- Calculate the weights w_i as the reciprocal variance of *Y*
- Using trial settings for the coefficients for weighted LS regression, calculate the sum of the squared residuals (= observed minus predicted response) weighted by $w_{i.}$
- Apply solver to minimize this sum by changing the weighted coefficients

Solver Routine

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Test Firing I	Constant	Target Speed (knots) xi	Hit or Miss yi	LS Pred Y	Var(Y) = (Yi)*(1-Yi)	Wi = 1/[(Yi)*(1-Yi)]	Y Wgt Pred	Residuals	Wgt*Res^2	SSQ Wres	SS Res^2		
2	1	1	400	0	0.3601	0.2304	4.339691	1.3800	-1.3800	8.264508	210.5340	24.92761	b0 WLS	1.5
3	2	1	220	1	0.9011	0.0891	11.21978	1.4340	-0.4340	2.113312	DF	DF	b1 WLS	-0.0003
4	3	1	490	0	0.0896	0.0816	12.25631	1.3530	-1.3530	22.43651	23	23		
5	4	1	410	1	0.3301	0.2211	4.522444	1.3770	-0.3770	0.64277	MSE	MSE		
6	5	1	500	0	0.0596	0.0560	17.8507	1.3500	-1.3500	32.5329	9.153651	1.083809		
7	6	1	270	0	0.7508	0.1871	5.344994	1.4190	-1.4190	10.76247				
8	7	1	200	1	0.9612	0.0373	26.81133	1.4400	-0.4400	5.190673				
9	8	1	470	0	0.1497	0.1273	7.854723	1.3590	-1.3590	14.50674				
10	9	1	480	0	0.1197	0.1054	9.49176	1.3560	-1.3560	17.45284				
11	10	1	310	1	0.6306	0.2329	4.292882	1.4070	-0.4070	0.711112				
12	11	1	240	1	0.8410	0.1337	7.47759	1.4280	-0.4280	1.369775				
13	12	1	490	0	0.0896	0.0816	12.25631							
14	13	1	420	0	0.3000	0.2100	4.76188							
15	14	1	330	1	0.5705	0.2450	4.081116							
16	15	1	280	1	0.7208	0.2013	4.9686							
17	16	1	210	1	0.9311	0.0641	15.59667							
18	17	1	300	1	0.6607	0.2242	4.460496							
19	18	1	470	1	0.1497	0.1273	7.854723							
20	19	1	230	0	0.8710	0.1123	8.902033							
21	20	1	430	0	0.2699	0.1971	5.074177							
22	21	1	460	0	0.1798	0.1475	6.781371							
23	22	1	220	1	0.9011	0.0891	11.21978							
24	23	1	250	1	0.8109	0.1533	6.522074							
25	24	1	200	1	0.9612	0.0373	26.81133							
26	25	1	390	0	0.3902	0.2379	4.202804							
27														
28				LS Coeff										
29				b0	1.56228									
30				b1	-0.00301									
31														
32														



Solver Solution

SSQ Wres	SS Res^2		
23.2753	3.63288	b0 WLS	1.586687
DF	DF	b1 WLS	-0.00309
23	23		
MSE	MSE		
1.01197	0.157951		

EXCEL Matrix Manipulation

Define the design matrix X by adding a column of "1"s for the constant in the model. Define the diagonal weight matrix V with variances along diagonal.

The standard error of the weighted LS coefficients can be obtained from:

$$\mathrm{Var}\,\beta^* = \left(X'V^{-1}X\right)^{-1}$$

Then, progressively calculate:

- the inverse V^{-1}
- the product $V^{-1}X$
- the transpose X'
- the product $X' V^{-1} X$
- the inverse of $X' V^{-1} X$
- the product $V^{-1} Y$
- the product $X' V^{-1} Y$
- the coefficients = $(X' V^{-1} X)^{-1}(X' V^{-1} Y)$

Weighted Matrix Results

$(X'V^{-1}X)^{-1}$			$X'V^{-1}y$		$(X'V^{-1}X)^{-1}X'V^{-1}y$	
0.034147	-9.16E-05		135.8388		1.586687	
-9.16E-05	2.81E-07		33314.51		-0.003091	
Std Error						
b0	0.184789					
b1	0.00053					

Weighted LS in JMP

- Set up a column for predicted Y using ordinary LS coefficients (Requires use of formula calculator in JMP)
- Set up column for weights as reciprocal variance of Y using formula calculator
- Label this column as weights and select “Fit Model”

Weighted LS Data Table in JMP

Montgomery Table 6.4

4 Cols / 25 Rows	Target Speed X	Hit or Miss Y	Predicted Y	Weights
1	400	0	0.360282	4.338792
2	220	1	0.901182	11.22927
3	490	0	0.089832	12.23059
4	410	1	0.330232	4.521228
5	500	0	0.059782	17.79103
6	270	0	0.750932	5.346646
7	200	1	0.961282	26.86806
8	470	0	0.149932	7.846067
9	480	0	0.119882	9.477747
10	310	1	0.630732	4.29352
11	240	1	0.841082	7.481498
12	490	0	0.089832	12.23059
13	420	0	0.300182	4.760255
14	330	1	0.570632	4.081447
15	280	1	0.720882	4.969904
16	210	1	0.931232	15.61549
17	300	1	0.660782	4.461315
18	470	1	0.149932	7.846067
19	230	0	0.871132	8.90781
20	430	0	0.270132	5.072005
21	460	0	0.179982	6.775597
22	220	1	0.901182	11.22927
23	250	1	0.811032	6.524898
24	200	1	0.961282	26.86806
25	390	0	0.390332	4.202159

0 Selected

Fit Model for Weighted LS in JMP

Montgomery Table 6.4: Model Fit

Response: Hit or Miss Y

Summary of Fit

RSquare	0.59373
RSquare Adj	0.576066
Root Mean Square Error	1.00592
Mean of Response	0.578757
Observations (or Sum Wgts)	234.9693

Lack of Fit

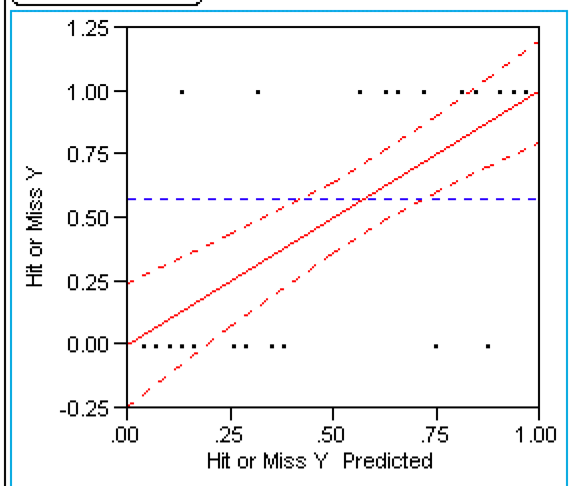
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	1.5866907	0.185826	8.54	<.0001
Target Speed X	-0.003091	0.000533	-5.80	<.0001

Effect Test

Source	Nparm	DF	Sum of Squares	F Ratio	Prob>F
Target Speed X	1	1	34.011759	33.6126	<.0001

Whole-Model Test



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	34.011759	34.0118	33.6126
Error	23	23.273148	1.0119	Prob>F
C Total	24	57.284907		<.0001

Logistic Regression, A Non-Linear Model

- The linear model constrains the response to have either a zero probability or a probability of one at large or small values of the regressor. This model may be unreasonable.
- Instead, we propose a model in which the probabilities of zero and one are reached asymptotically.
- Frequently we find that the response function is S shaped, similar to the CDF of a normal distribution. In fact, probit analysis involves modeling the response with a normal CDF.

Logistic Function Model

We attempt to model the indicator variable response using the logistic function (logit analysis):

$$E(Y \mid X) = p = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$$

$$= \frac{1}{1 + \exp(-\beta_0 - \beta_1 X)}$$

Linearizing Logistic Function

Consider the logit transformation of the probability p:

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

p^* is called the logit mean response. The logit response function is a linear model.

Fitting the Logit Response Model

Two Possibilities:

1. If we have repeat observations on Y at each level of X, we can estimate the probabilities using the proportion of "1"s at each X. Then, we fit the logit response function using weighted least squares.

2. If we have only a few or no repeat observations at the various X values, we cannot use proportions. We then estimate the logit response function from individual Y observations using maximum likelihood methods.

Weighted LS for Fitting Logit Response Model

- The observed proportion at each X level is

$$\overline{p}_i = \frac{\left(\#\text{ of 1's at } X_i\right)}{\left(\#\text{of observations at } X_i\right)}$$

- If the number of observations at each level of X is large, the variance of the transformed proportion

$$\overline{p}_i^* = \ln\left(\frac{\overline{p}_i}{1-\overline{p}_i}\right)$$

is

$$V\left(\overline{p}_i^*\right) = \frac{1}{n_i \overline{p}_i \left(1-\overline{p}_i\right)}$$

Weights for LS Regression

We use the appropriate weights

$$w_i = n_i \overline{p}_i \left(1 - \overline{p}_i\right)$$

and solve using weighted LS methods previously shown using EXCEL or JMP. Then transform p* to the original units p using logistic function.

$$\hat{p} = \frac{e^{\hat{p}^*}}{1 + e^{\hat{p}^*}}$$

Weighted LS Logit Regression

- We need to set following columns:
  – X
  – N (number of observations at each X)
  – Y (number of 1's at each X)
  – p_i (proportion)
  – p^*_i (transformed proportion)
  – w_i (weights)
- At this point, may want to consider MLE methods in JMP.

Maximum Likelihood Estimation for Logistic Grouped Data in JMP

- Data table is easy to set up.
  - Column with each X value sequentially repeated
  - Y column with alternating 0's and 1's
  - Frequency column for counts of 0's and 1's
- Label X column C and "X", Y column N (nominal) and Y, and Frequency column C and "F"
- Then run "Fit Y by X"

Caution: A JMP “Feature”

- JMP will model the lowest value of the binary response as the “success” and the alternative as the failure.
- Thus, “0” will be treated as success and “1” as failure. Similarly, “no” will be viewed as success and “yes” as failure, since “n” comes before “y” in the alphabet.
- Consequently, the function you expect to be monotonically increasing will appear as decreasing and vice versa unless you flip the indicator values.
- In the examples that follow, I have listed the tables as they appear in texts but displayed the graphs by interchanging 1’s and 0’s for analysis (Fit Y by X)

MLE Table for Grouped Data

NWK Table 16.1

3 Cols, 10 Rows

	Price Reduction X	Y	Count
1	5	1	168
2	5	0	32
3	10	1	149
4	10	0	51
5	15	1	130
6	15	0	70
7	20	1	97
8	20	0	103
9	30	1	52
10	30	0	148

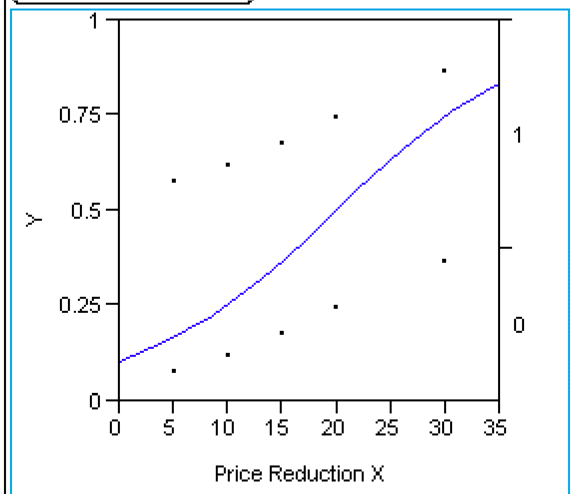
0 Selected

Example from *Applied Linear Statistical* Models by Neter, Wasserman, and Kutner, Table 16.1

Fit Y by X

NWK Table 16.1: Y by X

Y By Price Reduction X



Converged by Gradient

Whole-Model Test

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-2.1854787	0.1646661	176.15	<.0001
Price Reduction X	0.10871782	0.0088429	151.15	<.0001

Logistic Regression in Jump Individual Values

- We can use JMP's MLE to fit a model to the data.
- The data table entry is simple:
  - Column for X
  - Column for Y or 1's and 0's
- Label X column C and X
- Label Y column N and Y
- Fit Y by X

Data Table for Logistic MLE

NWK Table 16.2

2 Cols, 25 Rows

	Mths Exp X	Task Succ Y
1	14	0
2	29	0
3	6	0
4	25	1
5	18	1
6	4	0
7	18	0
8	12	0
9	22	1
10	6	0
11	30	1
12	11	0
13	30	1
14	5	0
15	20	1
16	13	0
17	9	0
18	32	1
19	24	0
20	13	1
21	19	0
22	4	0
23	28	1
24	22	1
25	8	1

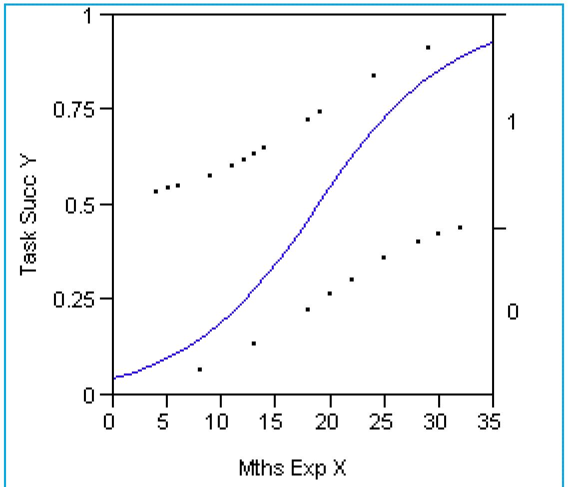
0 Selected

Example from *Applied Linear Statistical* Models by Neter, Wasserman, and Kutner, Table 16.2

Fit Y by X MLE Output

NWK Table 16.2: Y by X

Task Succ Y By Mths Exp X



Converged by Gradient

Whole-Model Test

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-3.0596954	1.2593518	5.90	0.0151
Mths Exp X	0.1614859	0.0649801	6.18	0.0129

Multiple Logistic Regression

Here's an example from the *JMP In* training manual that comes with the student version of JMP:

A weatherman is trying to predict the precipitation probability by looking at the morning temperature and the barometric pressure. He generates a table for 30 days in April. If the precipitation was greater than 0.02 inches, the day was called rainy. If below, then dry.

Spring.JMP Data

Partial Table:

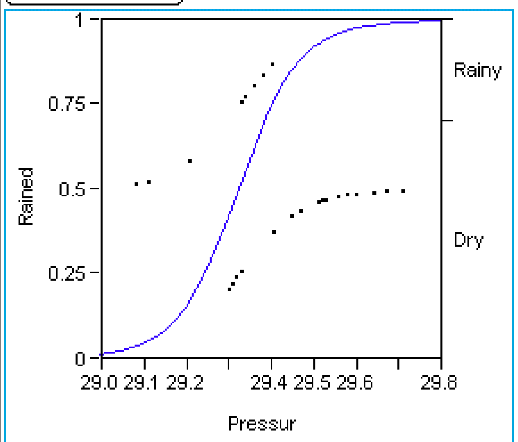
Spring

14 Cols / 30 Rows	mon#	month	date	Temp	April	Humid1:PM	Humid4:PM	Precip	Pressur	wrDir1:PM	wDir4:PM	wrSpeed	SkyCover	Rained
3	04	APR	04/03	50	3	53	86	0.27	29.4	7	86	6.8	8	Rainy
4	04	APR	04/04	43	4	47	42	0.05	29.21	25	42	14.2	10	Rainy
5	04	APR	04/05	42	5	44	48	0.001	29.32	27	48	9.7	8	Dry
6	04	APR	04/06	46	6	42	41	0.001	29.33	29	41	10.5	7	Dry
7	04	APR	04/07	52	7	40	37	0	29.3	27	37	8.8	3	Dry
8	04	APR	04/08	59	8	36	23	0	29.31	30	23	6.7	3	Dry
9	04	APR	04/09	56	9	31	25	0	29.33	34	25	3.3	3	Dry
10	04	APR	04/10	57	10	27	23	0	29.41	22	23	1.1	0	Dry
11	04	APR	04/11	62	11	42	43	0.08	29.36	17	43	6.3	8	Rainy
12	04	APR	04/12	66	12	45	66	0.23	29.38	22	66	6.8	8	Rainy

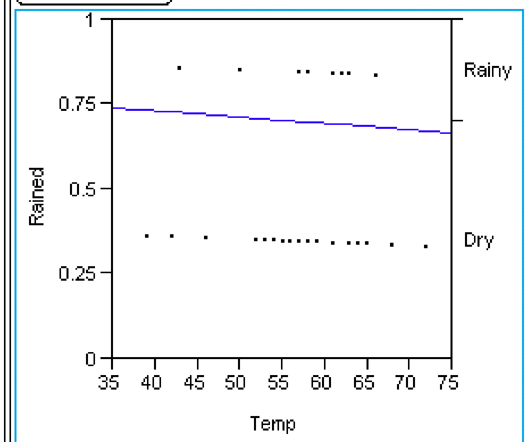
JMP Logistic Analysis: Fit Y by X

Spring: Y by X

Rained By Pressur



Rained By Temp



Converged by Objective

Whole-Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	6.250851	1	12.5017	0.0004
Full	12.075078			
Reduced	18.325929			

RSquare (U)	0.3411
Observations (or Sum Wgts)	30

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-405.36267	169.29517	5.73	0.0166
Pressur	13.8233881	5.7651316	5.75	0.0165

Converged by Gradient

Whole-Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	0.013333	1	0.026666	0.8703
Full	18.312596			
Reduced	18.325929			

RSquare (U)	0.0007
Observations (or Sum Wgts)	30

Parameter Estimates

Term		Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	Unstable	1.33837326	3.0616295	0.19	0.6620
Temp	Unstable	-0.0085936	0.0529775	0.03	0.8711

Multiple Logistic Regression in JMP

- Fit Y by X
  - Generates a separate logistic regression for each predictor column X_i

$$E(Y \mid X_i) = \frac{1}{1 + \exp(-\beta_0 - \beta_1 X_i)}$$

- Fit Model
  - Fits an overall logistic regression model for specified predictor columns X's and interactions

$$E(Y \mid X_1, X_2) = \frac{1}{1 + \exp(-\beta_0 - \beta_1 X_1 - \beta_2 X_2 - \beta_{12} X_1 X_2)}$$

Conclusion

- Binary response data occurs in many important applications.
- The simple linear regression model has constraints that may affect its adequacy.
- The logistic model has many desirable properties for modeling indicator variables.
- EXCEL and JMP have excellent capabilities for analysis and modeling of binary data.
- For logistic regression modeling, JMP's MLE routines are easy to apply and very useful.