An Introduction to Logistic Regression

Outline

- Introduction and Description
- Some Potential Problems and Solutions
- Writing Up the Results

Introduction and Description

- Why use logistic regression?
- Estimation by maximum likelihood
- Interpreting coefficients
- Hypothesis testing
- Evaluating the performance of the model

Why use logistic regression?

- There are many important research topics for which the dependent variable is "limited."
- For example: voting, morbidity or mortality, and participation data is not continuous or distributed normally.
- Binary logistic regression is a type of regression analysis where the dependent variable is a dummy variable: coded 0 (did not vote) or 1(did vote)

The Linear Probability Model

In the OLS regression:

$Y = \gamma + \phi X + e$; where $Y = (0, 1)$

- The error terms are heteroskedastic
- e is not normally distributed because Y takes on only two values
- The predicted probabilities can be greater than 1 or less than 0

An Example: Hurricane Evacuations

Q: EVAC

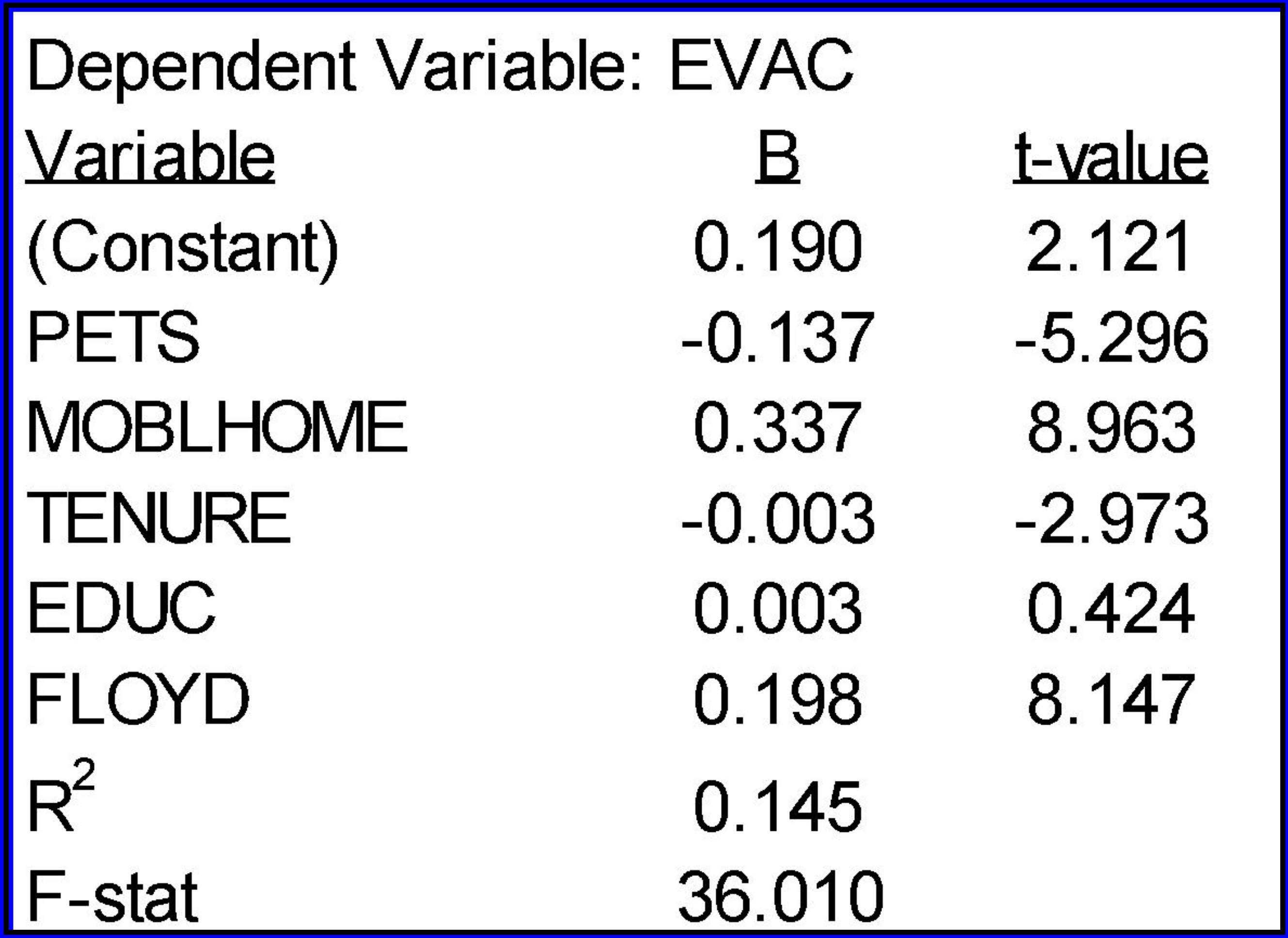
Did you evacuate your home to go someplace safer before Hurricane Dennis (Floyd) hit?

1 YES
2 NO
3 DON'T KNOW
4 REFUSED

The Data

EVAC	PETS	MOBLHOME	TENURE	EDUC
0	1	0	16	16
0	1	0	26	12
0	1	1	11	13
1	1	1	1	10
1	0	0	5	12
0	0	0	34	12
0	0	0	3	14
0	1	0	3	16
0	1	0	10	12
0	0	0	2	18
0	0	0	2	12
0	1	0	25	16
1	1	1	20	12

OLS Results

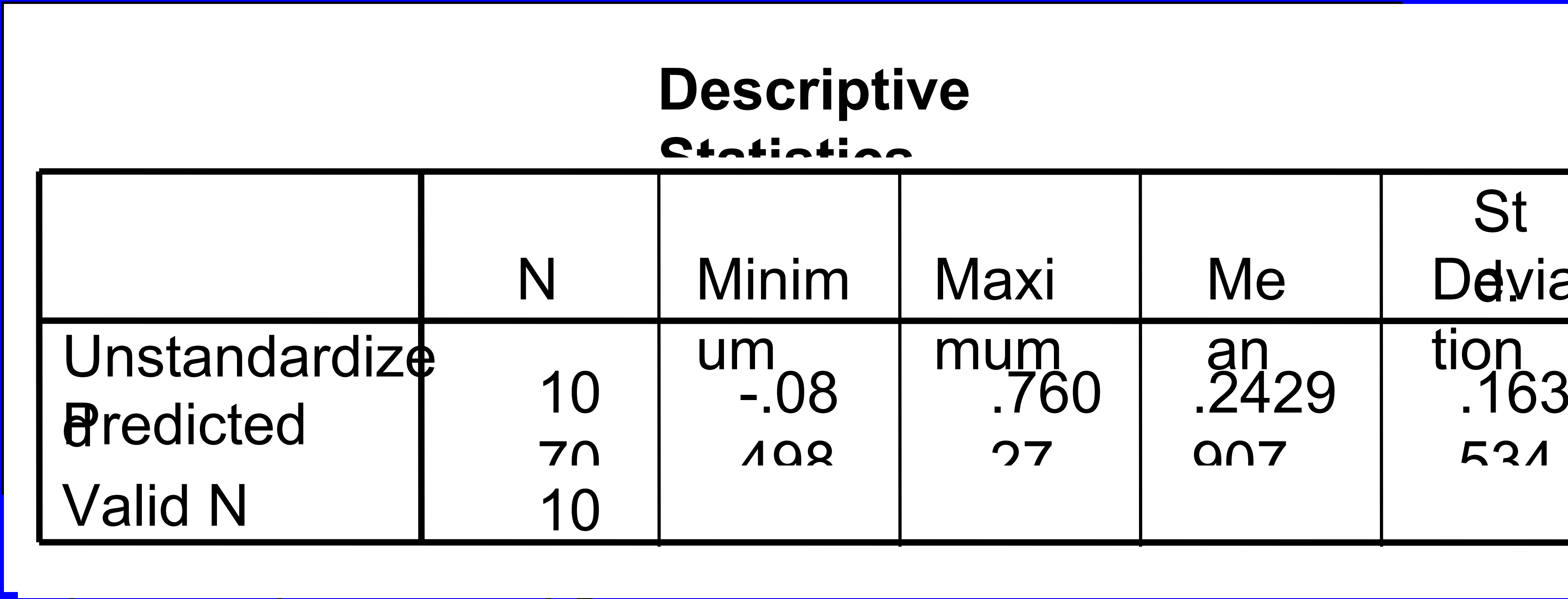


Dependent Variable: EVAC

Variable	B	t-value
(Constant)	0.190	2.121
PETS	-0.137	-5.296
MOBLHOME	0.337	8.963
TENURE	-0.003	-2.973
EDUC	0.003	0.424
FLOYD	0.198	8.147
R^2	0.145	
F-stat	36.010	

Problems:

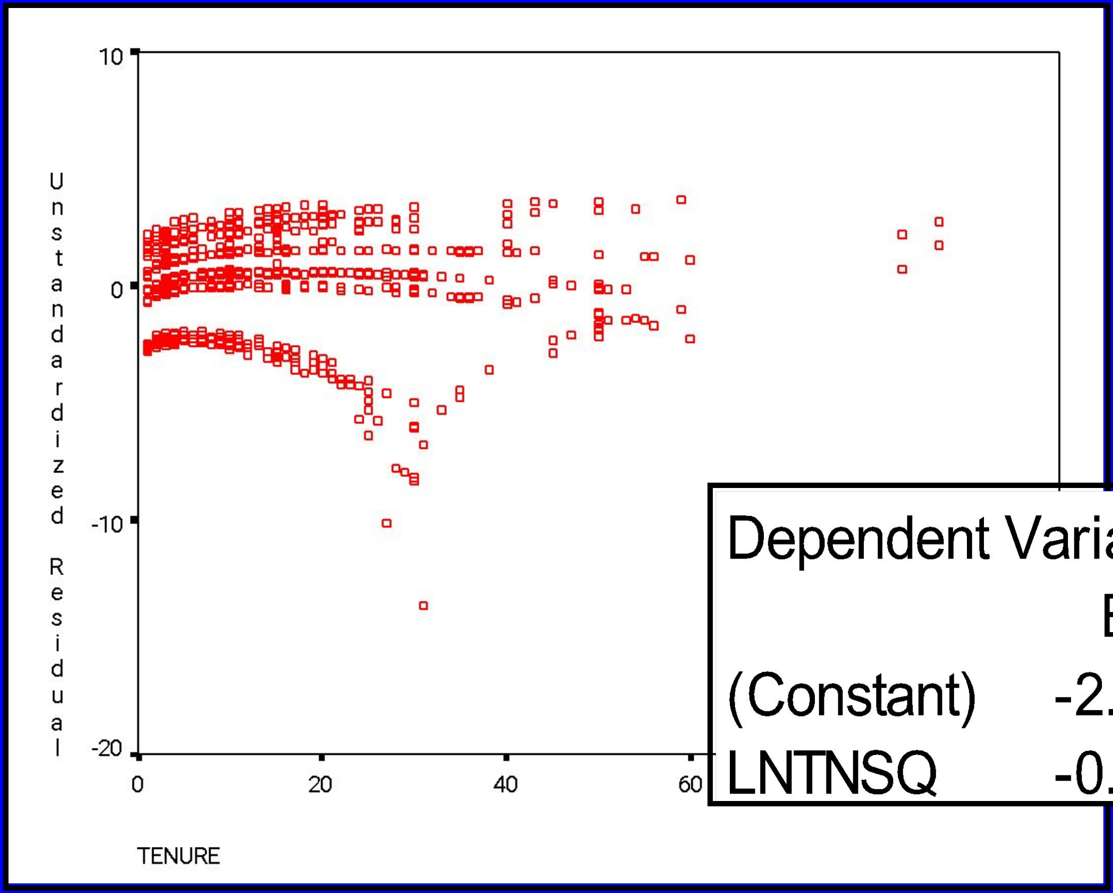
Predicted Values outside the 0,1 range



Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Unstandardized Predicted	10 70	-.08 498	.760 27	.2429 907	.163 534
Valid N	10				

Heteroskedasticity



Park Test

Dependent Variable: LNESQ

	B	t-stat
(Constant)	-2.34	-15.99
LNTNSQ	-0.20	**-6.19**

The Logistic Regression Model

The "logit" model solves these problems:

$$\ln[p/(1-p)] = \alpha + \beta X + e$$

- p is the probability that the event Y occurs, $p(Y=1)$
- $p/(1-p)$ is the "odds ratio"
- $\ln[p/(1-p)]$ is the log odds ratio, or "logit"

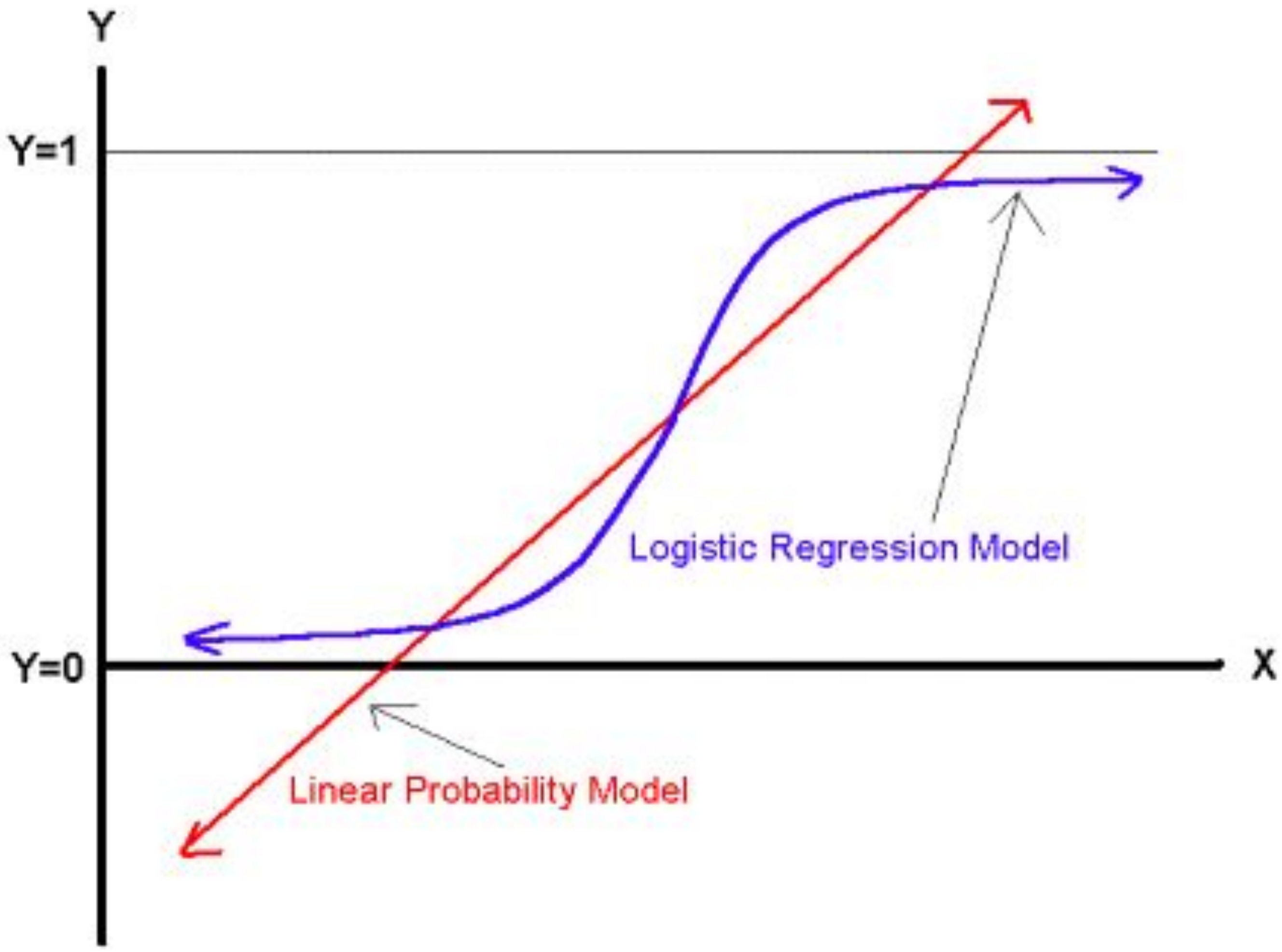
More:

- The logistic distribution constrains the estimated probabilities to lie between 0 and 1.
- The estimated probability is:

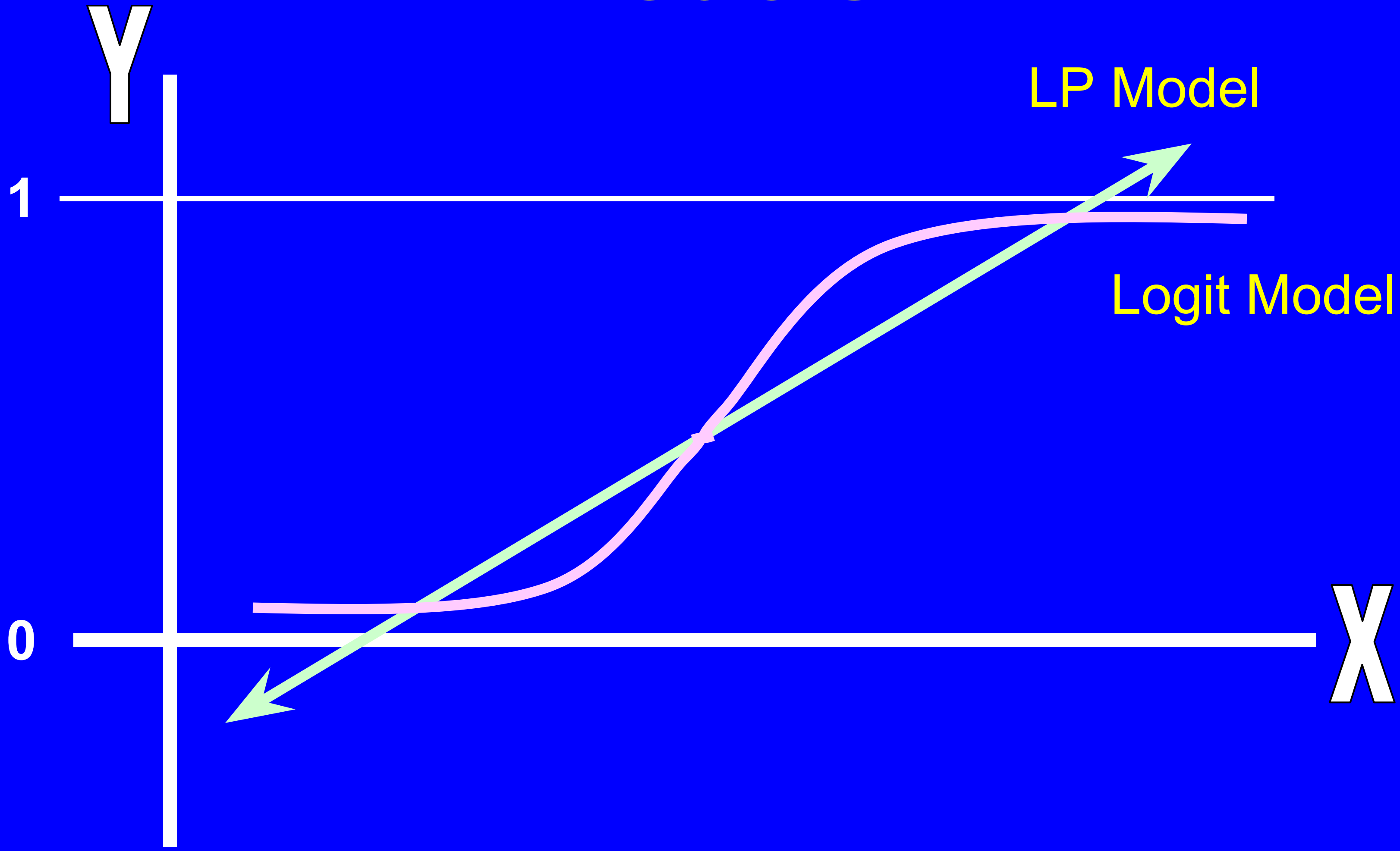
$$p = 1/[1 + \exp(-\alpha - \beta X)]$$

- if you let $\alpha + \beta X = 0$, then $p = .50$
- as $\alpha + \beta X$ gets really big, p approaches 1
- as $\alpha + \beta X$ gets really small, p approaches 0

Comparing the LP and Logit Models



Comparing LP and Logit Models



Maximum Likelihood Estimation (MLE)

- MLE is a statistical method for estimating the coefficients of a model.
- The likelihood function (L) measures the probability of observing the particular set of dependent variable values (p_1, p_2, ..., p_n) that occur in the sample:
  $L = \text{Prob}\ (p_1 * p_2 * * * p_n)$
- The higher the L, the higher the probability of observing the ps in the sample.

- MLE involves finding the coefficients (α, β) that makes the log of the likelihood function (LL < 0) as large as possible
- Or, finds the coefficients that make -2 times the log of the likelihood function (-2LL) as small as possible
- The maximum likelihood estimates solve the following condition:

$$\{Y - p(Y=1)\}X_i = 0$$

summed over all observations, i = 1,…,n

Interpreting Coefficients

- Since:

  $\ln[p/(1-p)] = \alpha + \beta X + e$

  The slope coefficient (β) is interpreted as the rate of change in the "log odds" as X changes … not very useful.

- Since:

  $p = 1/[1 + \exp(-\alpha - \beta X)]$

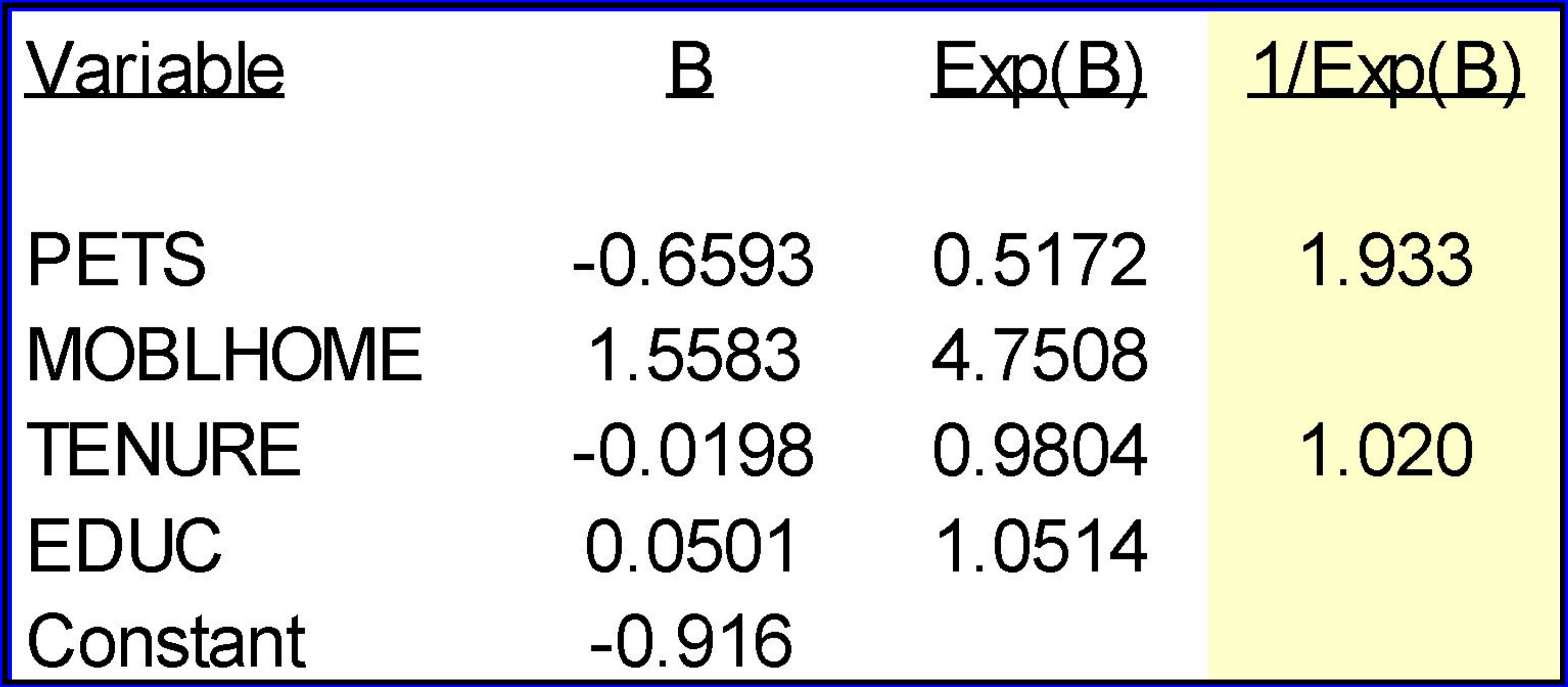
  The marginal effect of a change in X on the probability is: □p/□X = $f(\beta X)\ \beta$

- An interpretation of the logit coefficient which is usually more intuitive is the "odds ratio"
- Since:

$$[p/(1-p)] = \exp(\alpha + \beta X)$$

exp(β) is the effect of the independent variable on the "odds ratio"

From SPSS Output:



Variable	B	Exp(B)	1/Exp(B)
PETS	-0.6593	0.5172	1.933
MOBLHOME	1.5583	4.7508	
TENURE	-0.0198	0.9804	1.020
EDUC	0.0501	1.0514	
Constant	-0.916		

“Households with pets are 1.933 times more likely to evacuate than those without pets.”

Hypothesis Testing

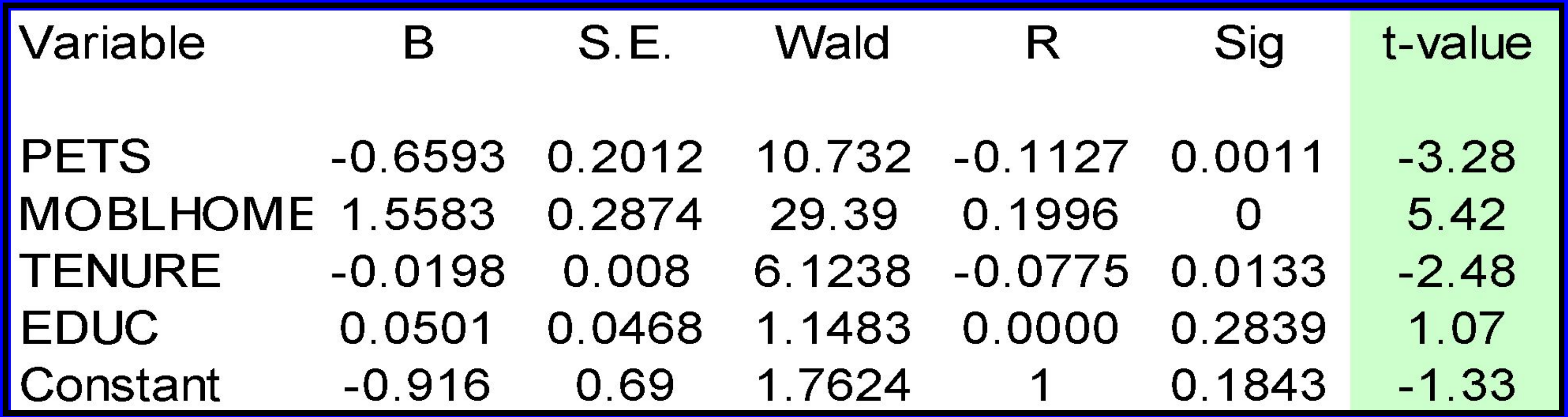
- The Wald statistic for the *β* coefficient is:

  $$\text{Wald} = [\beta\ /\text{s.e.}_{B}]^{2}$$

  which is distributed chi-square with 1 degree of freedom.
- The "Partial R" (in SPSS output) is

  $$R = \{[(\text{Wald-2})/(-2\text{LL}(\alpha)]\}^{1/2}$$

An Example:



Variable	B	S.E.	Wald	R	Sig	t-value
PETS	-0.6593	0.2012	10.732	-0.1127	0.0011	-3.28
MOBLHOME	1.5583	0.2874	29.39	0.1996	0	5.42
TENURE	-0.0198	0.008	6.1238	-0.0775	0.0133	-2.48
EDUC	0.0501	0.0468	1.1483	0.0000	0.2839	1.07
Constant	-0.916	0.69	1.7624	1	0.1843	-1.33

Evaluating the Performance of the Model

There are several statistics which can be used for comparing alternative models or evaluating the performance of a single model:

- Model Chi-Square
- Percent Correct Predictions
- Pseudo-R^2

Model Chi-Square

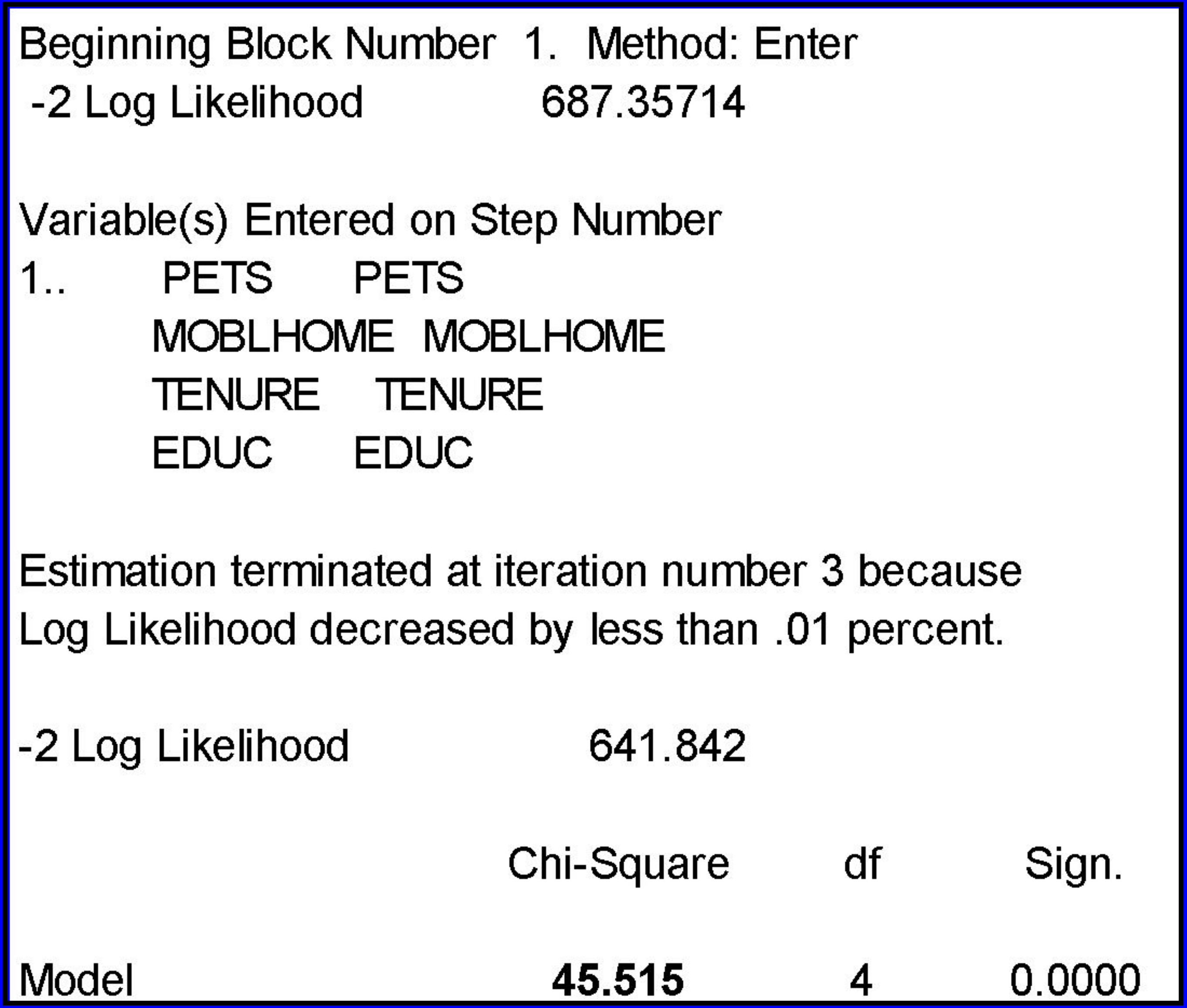
- The model likelihood ratio (LR), statistic is

  LR[i] = -2[LL(α) - LL(α, *β*)]

  {Or, as you are reading SPSS printout:

  LR[i] = [-2LL (of beginning model)] - [-2LL (of ending model)]}
- The LR statistic is distributed chi-square with i degrees of freedom, where i is the number of independent variables
- Use the “Model Chi-Square” statistic to determine if the overall model is statistically significant.

An Example:

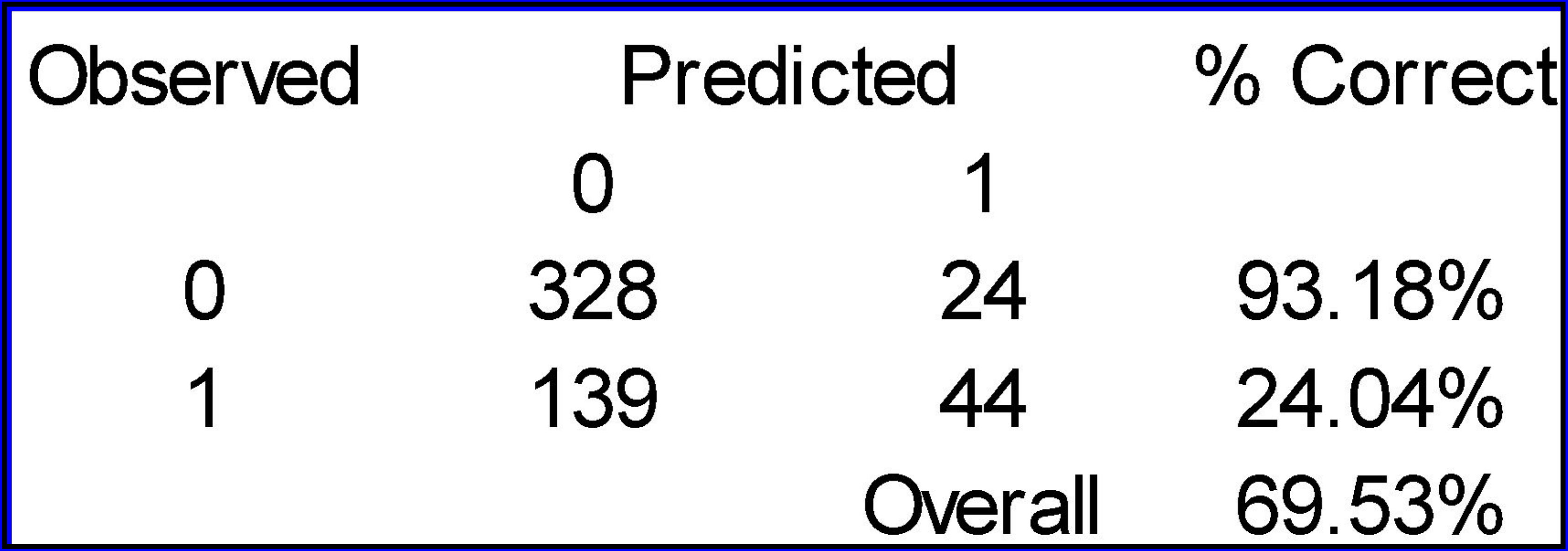


```
Beginning Block Number  1.  Method: Enter
 -2 Log Likelihood           687.35714

Variable(s) Entered on Step Number
1..       PETS       PETS
         MOBLHOME  MOBLHOME
         TENURE    TENURE
         EDUC      EDUC

Estimation terminated at iteration number 3 because
Log Likelihood decreased by less than .01 percent.

-2 Log Likelihood               641.842

                         Chi-Square      df        Sign.

Model                      45.515         4       0.0000
```

Percent Correct Predictions

- The "Percent Correct Predictions" statistic assumes that if the estimated p is greater than or equal to .5 then the event is expected to occur and not occur otherwise.
- By assigning these probabilities 0s and 1s and comparing these to the actual 0s and 1s, the % correct Yes, % correct No, and overall % correct scores are calculated.

An Example:



Observed	Predicted		% Correct
	0	1	
0	328	24	93.18%
1	139	44	24.04%
		Overall	69.53%

Pseudo-R^2

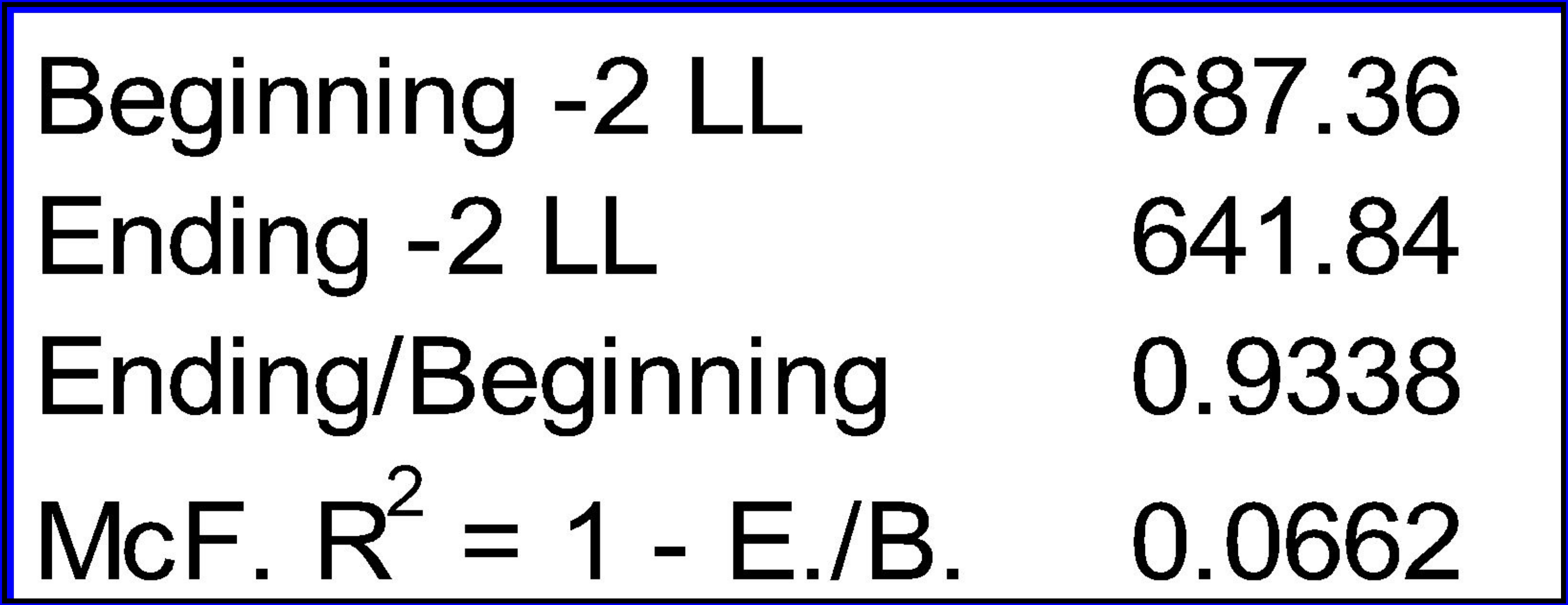
- <u>One</u> psuedo-R^2 statistic is the McFadden's-R^2 statistic:

$$\text{McFadden's-}R^2 = 1 - [LL(\alpha,\beta)/LL(\alpha)]$$

$$\{= 1 - [-2LL(\alpha, \beta)/-2LL(\alpha)] \text{ (from } \textbf{\textit{SPSS}} \text{ printout)}\}$$

- where the R^2 is a scalar measure which varies between 0 and (somewhat close to) 1 much like the R^2 in a LP model.

An Example:



Beginning -2 LL	687.36
Ending -2 LL	641.84
Ending/Beginning	0.9338
McF. R^2 = 1 - E./B.	0.0662

Some potential problems and solutions

- Omitted Variable Bias
- Irrelevant Variable Bias
- Functional Form
- Multicollinearity
- Structural Breaks

Omitted Variable Bias

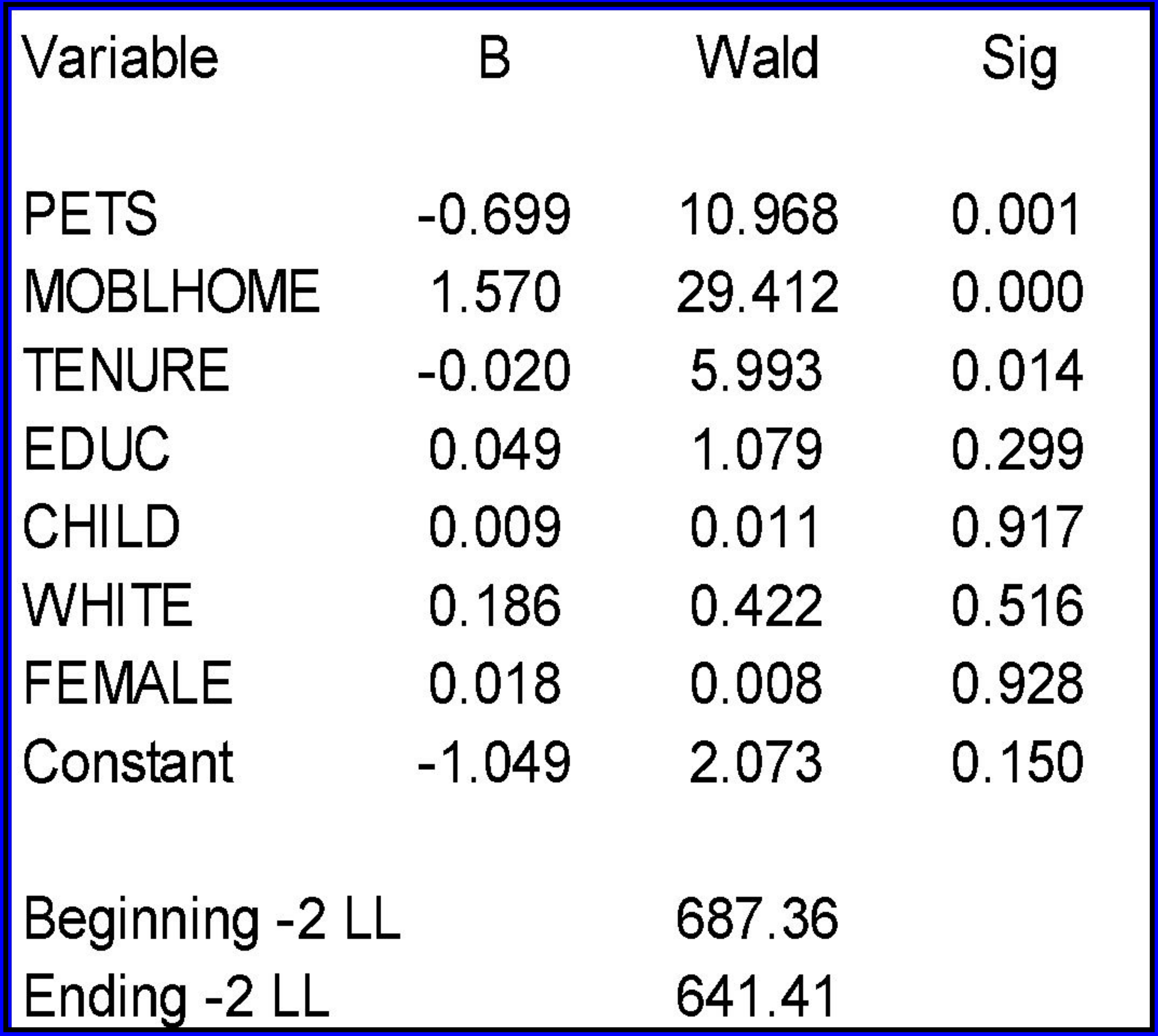
- Omitted variable(s) can result in bias in the coefficient estimates. To test for omitted variables you can conduct a likelihood ratio test:

  LR[q] = {[-2LL(constrained model, i=k-q)]

  - [-2LL(unconstrained model, i=k)]}

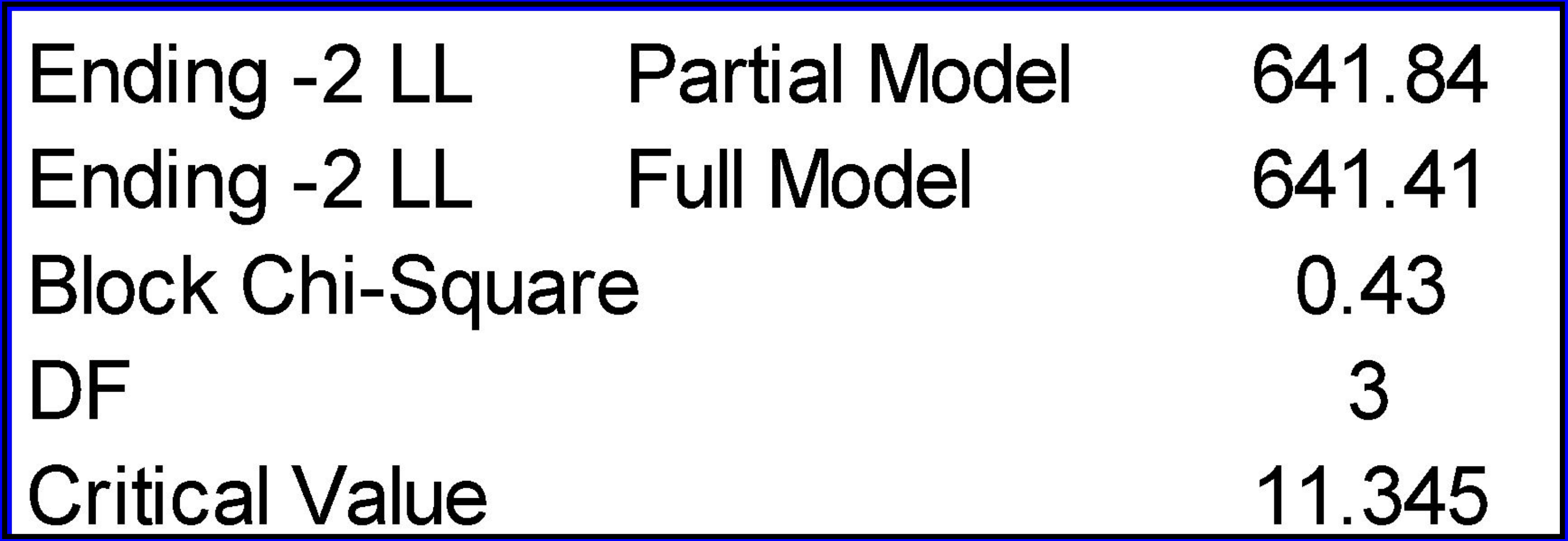
  where LR is distributed chi-square with q degrees of freedom, with q = 1 or more omitted variables
- {This test is conducted automatically by ***SPSS*** if you specify "blocks" of independent variables}

An Example:



Variable	B	Wald	Sig
PETS	-0.699	10.968	0.001
MOBLHOME	1.570	29.412	0.000
TENURE	-0.020	5.993	0.014
EDUC	0.049	1.079	0.299
CHILD	0.009	0.011	0.917
WHITE	0.186	0.422	0.516
FEMALE	0.018	0.008	0.928
Constant	-1.049	2.073	0.150
Beginning -2 LL		687.36	
Ending -2 LL		641.41	

Constructing the LR Test



Ending -2 LL	Partial Model	641.84
Ending -2 LL	Full Model	641.41
Block Chi-Square		0.43
DF		3
Critical Value		11.345

“Since the chi-squared value is less than the critical value the set of coefficients is not statistically significant. The full model is not an improvement over the partial model.”

Irrelevant Variable Bias

- The inclusion of irrelevant variable(s) can result in poor model fit.
- You can consult your Wald statistics or conduct a likelihood ratio test.

Functional Form

- Errors in functional form can result in biased coefficient estimates and poor model fit.
- You should try different functional forms by logging the independent variables, adding squared terms, etc.
- Then consult the Wald statistics and model chi-square statistics to determine which model performs best.

Multicollinearity

- The presence of multicollinearity will not lead to biased coefficients.
- But the standard errors of the coefficients will be inflated.
- If a variable which you think should be statistically significant is not, consult the correlation coefficients.
- If two variables are correlated at a rate greater than .6, .7, .8, etc. then try dropping the least theoretically important of the two.

Structural Breaks

- You may have structural breaks in your data. Pooling the data imposes the restriction that an independent variable has the same effect on the dependent variable for different groups of data when the opposite may be true.
- You can conduct a likelihood ratio test:

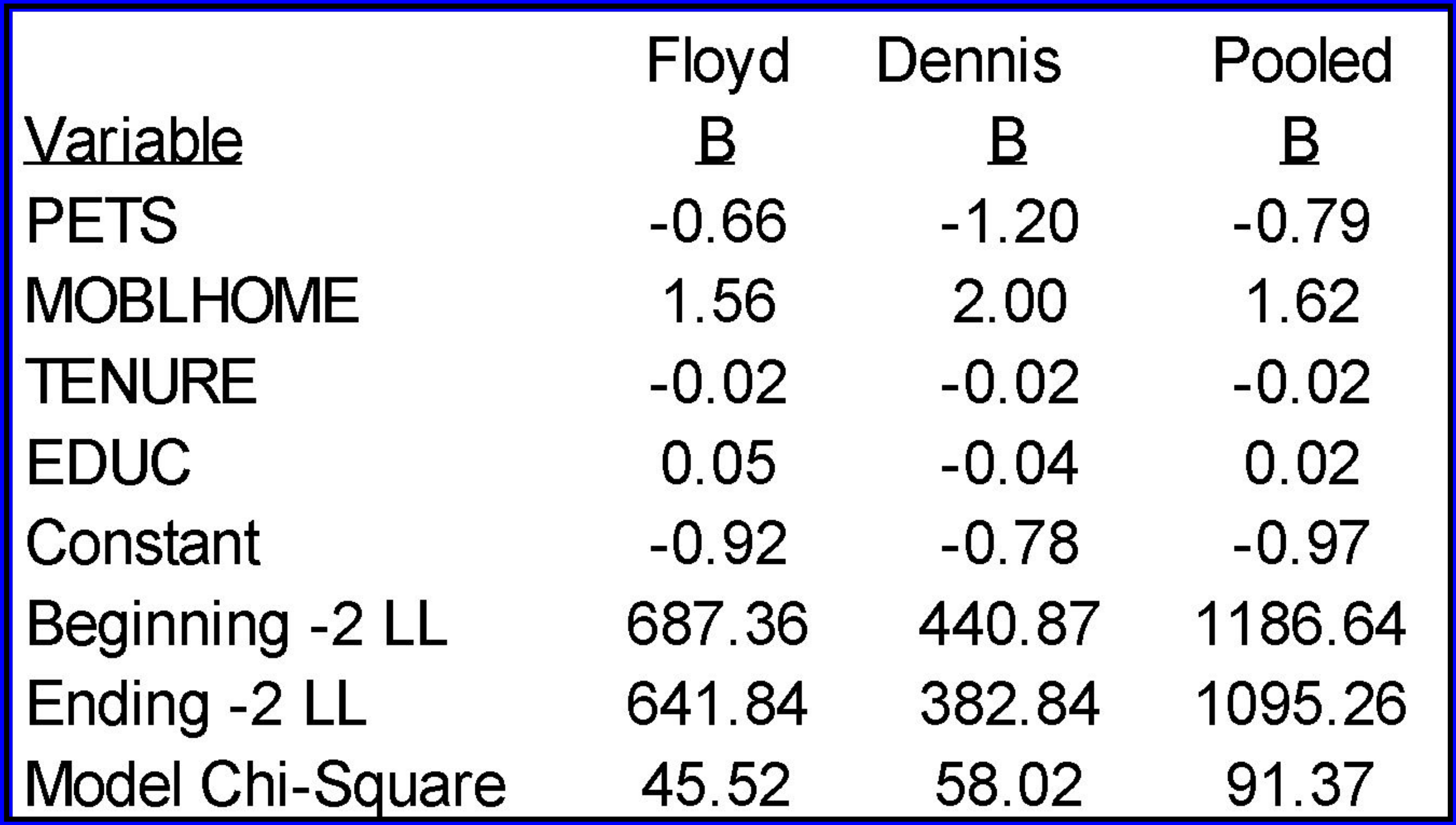
$$LR[i+1] = -2LL(\text{pooled model})$$

$$[-2LL(\text{sample } 1) + -2LL(\text{sample } 2)]$$

where samples 1 and 2 are pooled, and i is the number of dependent variables.

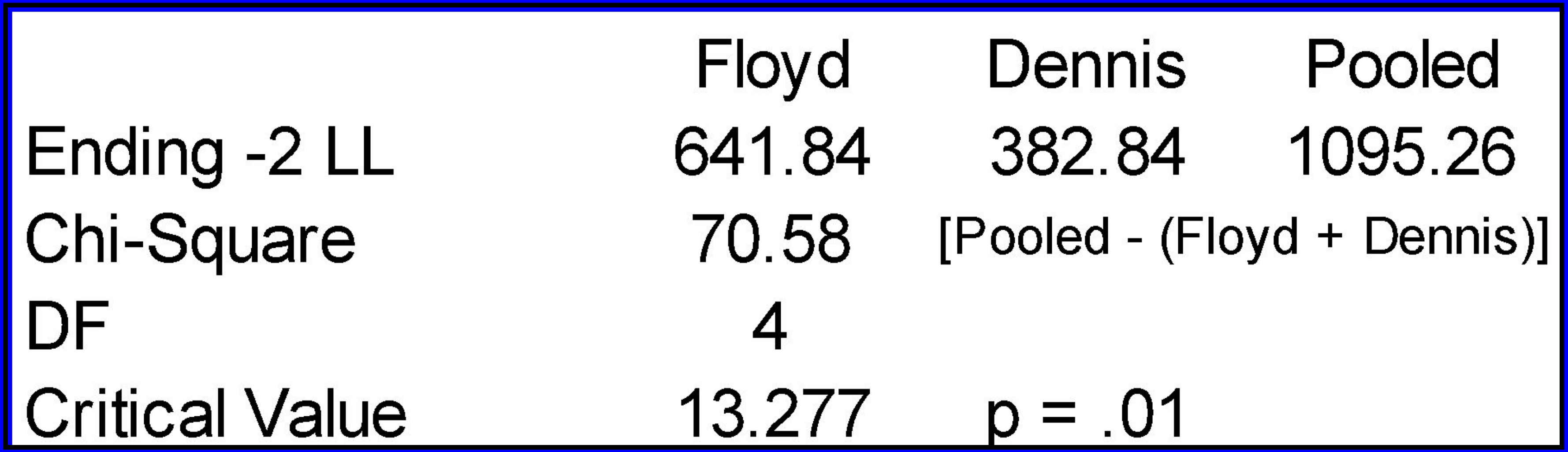
An Example

- Is the evacuation behavior from Hurricanes Dennis and Floyd statistically equivalent?



	Floyd	Dennis	Pooled
Variable	B	B	B
PETS	-0.66	-1.20	-0.79
MOBLHOME	1.56	2.00	1.62
TENURE	-0.02	-0.02	-0.02
EDUC	0.05	-0.04	0.02
Constant	-0.92	-0.78	-0.97
Beginning -2 LL	687.36	440.87	1186.64
Ending -2 LL	641.84	382.84	1095.26
Model Chi-Square	45.52	58.02	91.37

Constructing the LR Test



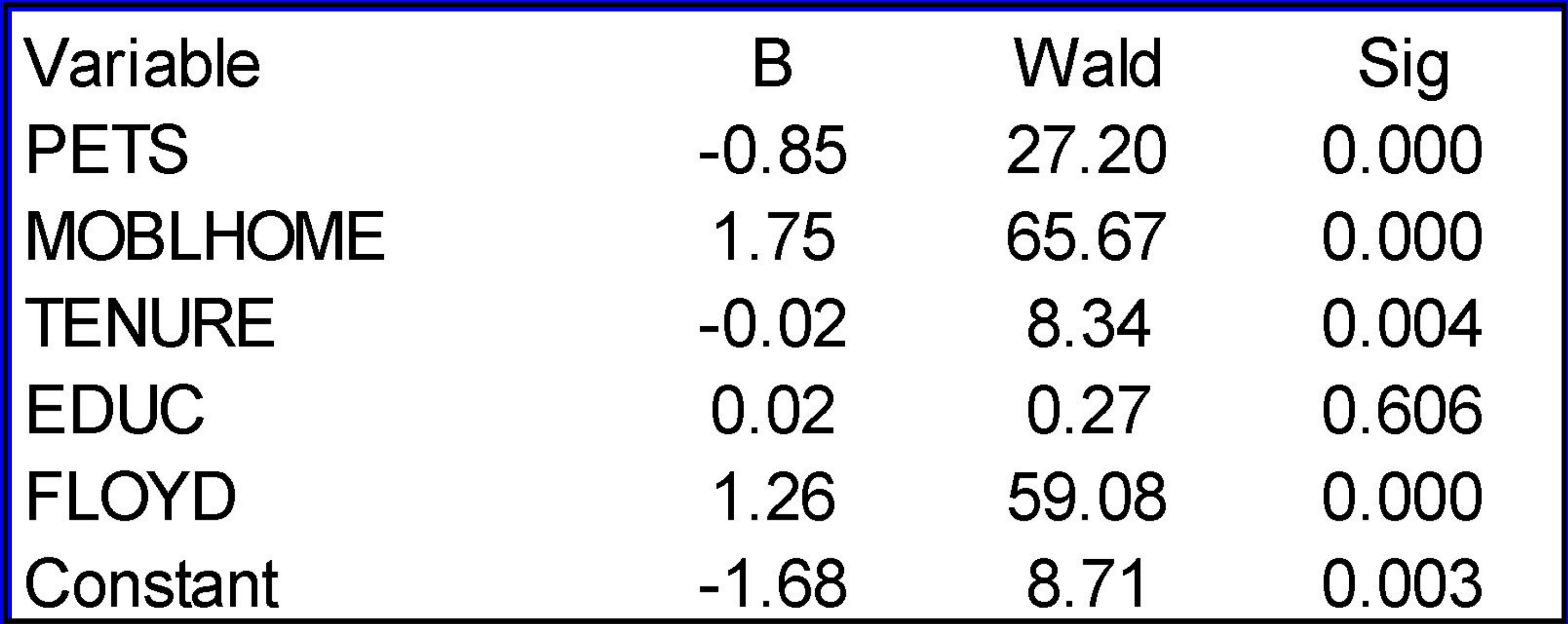
	Floyd	Dennis	Pooled
Ending -2 LL	641.84	382.84	1095.26
Chi-Square	70.58	[Pooled - (Floyd + Dennis)]	
DF	4		
Critical Value	13.277	p = .01	

Since the chi-squared value is greater than the critical value the set of coefficients are statistically different. The pooled model is inappropriate.

What should you do?

- Try adding a dummy variable:

  FLOYD = 1 if Floyd, 0 if Dennis



Variable	B	Wald	Sig
PETS	-0.85	27.20	0.000
MOBLHOME	1.75	65.67	0.000
TENURE	-0.02	8.34	0.004
EDUC	0.02	0.27	0.606
FLOYD	1.26	59.08	0.000
Constant	-1.68	8.71	0.003

Writing Up Results

- Present descriptive statistics in a table
- Make it clear that the dependent variable is discrete (0, 1) and not continuous and that you will use logistic regression.
- Logistic regression is a standard statistical procedure so you don't (necessarily) need to write out the formula for it. You also (usually) don't need to justify that you are using Logit instead of the LP model or Probit (similar to logit but based on the normal distribution [the tails are less fat]).

An Example:

"The dependent variable which measures the willingness to evacuate is EVAC. EVAC is equal to 1 if the respondent evacuated their home during Hurricanes Floyd and Dennis and 0 otherwise. The logistic regression model is used to estimate the factors which influence evacuation behavior."

Organize your regression results in a table:

- In the heading state that your dependent variable (dependent variable = EVAC) and that these are "logistic regression results.”
- Present coefficient estimates, t-statistics (or Wald, whichever you prefer), and (at least the) model chi-square statistic for overall model fit
- If you are comparing several model specifications you should also present the % correct predictions and/or Pseudo-R^2 statistics to evaluate model performance
- If you are comparing models with hypotheses about different blocks of coefficients or testing for structural breaks in the data, you could present the ending log-likelihood values.

An Example:

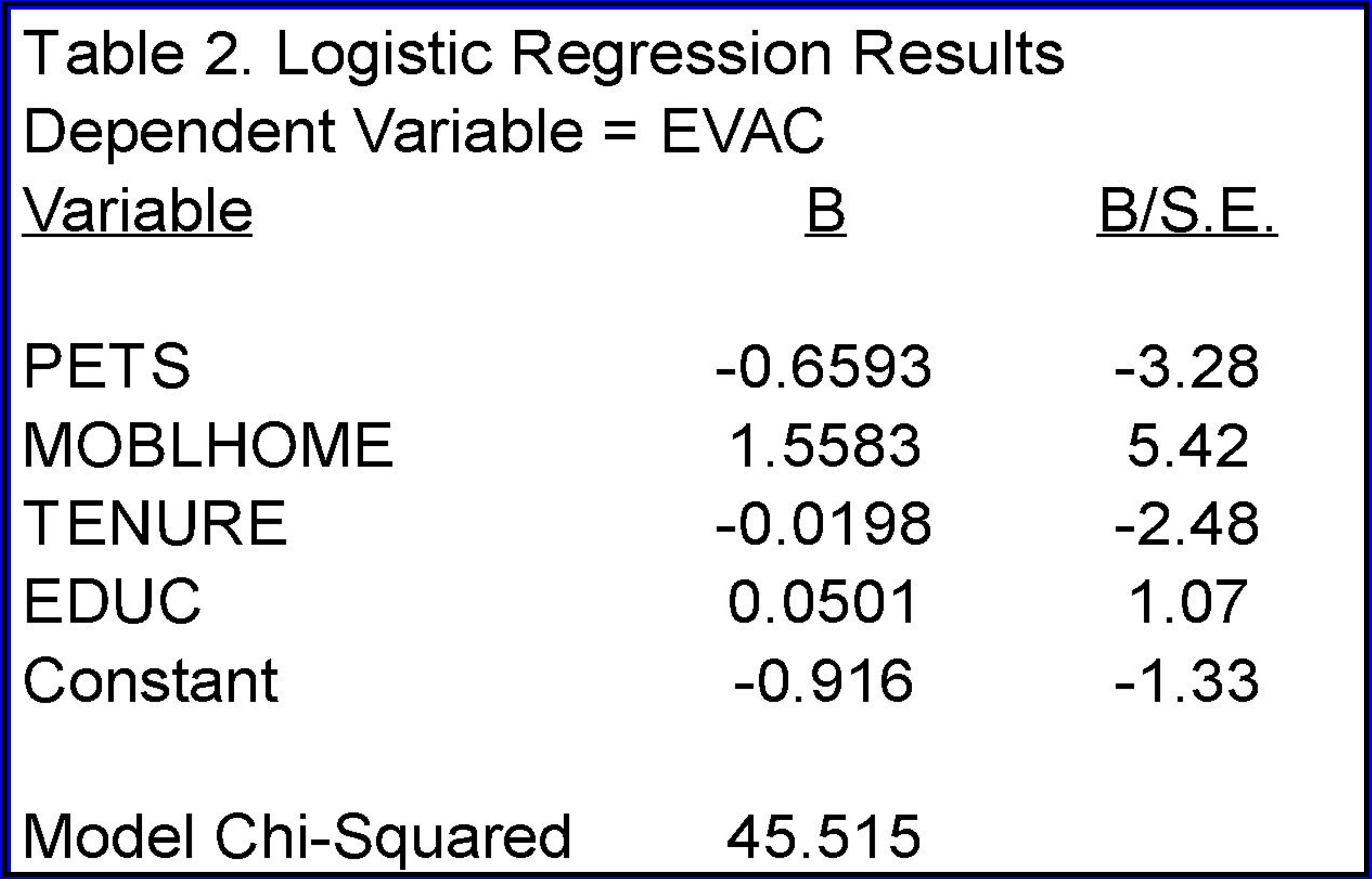


Table 2. Logistic Regression Results

Dependent Variable = EVAC

Variable	B	B/S.E.
PETS	-0.6593	-3.28
MOBLHOME	1.5583	5.42
TENURE	-0.0198	-2.48
EDUC	0.0501	1.07
Constant	-0.916	-1.33
Model Chi-Squared	45.515	

When describing the statistics in the tables, point out the highlights for the reader.

What are the statistically significant variables?

> **"The results from Model 1 indicate that coastal residents behave according to risk theory. The coefficient on the MOBLHOME variable is negative and statistically significant at the p < .01 level (t-value = 5.42). Mobile home residents are 4.75 times more likely to evacuate."**

Is the overall model statistically significant?

“The overall model is significant at the .01 level according to the Model chi-square statistic. The model predicts 69.5% of the responses correctly. The McFadden's R^2 is .066."

Which model is preferred?

"Model 2 includes three additional independent variables. According to the likelihood ratio test statistic, the partial model is superior to the full model of overall model fit. The block chi-square statistic is not statistically significant at the .01 level (critical value = 11.35 [df=3]). The coefficient on the children, gender, and race variables are not statistically significant at standard levels."

Also

- You usually don't need to discuss the magnitude of the coefficients--just the sign (+ or -) and statistical significance.
- If your audience is unfamiliar with the extensions (beyond ***SPSS*** or ***SAS*** printouts) to logistic regression, discuss the calculation of the statistics in an appendix or footnote or provide a citation.
- Always state the degrees of freedom for your likelihood-ratio (chi-square) test.

References

- http://personal.ecu.edu/whiteheadj/data/logit/
- http://personal.ecu.edu/whiteheadj/data/logit/logitpap.htm
- E-mail: WhiteheadJ@mail.ecu.edu