Handout: Binary Outcomes - Logistic Regression

High school and Beyond (from C.J. Anderson, University of Idaho, introduced in a course on categorical data analysis)

- Question: Is the proportion of students attending an academic program related to academic achievement (x).
- Data from high school seniors (N=600) - random sample from 1983 in US.
- Response: Did students attend an academic high school program type or a non-academic program type (Y).
- Explanatory variables: Scores on 5 standardized achievement tests (Reading, Writing, Math, Science, and Civics), x being the total of the scores.

Logistic Model
The logistic model is an example for a GLM, with random component is a Bernoulli distribution, the linear predictor for the systematic component, and the logit link function, therefore, with π being the probability for being in an academic program, the model is given by

$$ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x$$

Fitting the data to this model results in:

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test			Exp(B)	95% Wald Confidence Interval for Exp(B)	
			Lower	Upper	Wald Chi-Square	df	Sig.		Lower	Upper
(Intercept)	-7.055	.6948	-8.417	-5.693	103.104	1	.000	.001	.000	.003
x	.027	.0027	.022	.033	106.411	1	.000	1.028	1.022	1.033
(Scale)	1[a]									

Dependent Variable: program
Model: (Intercept), x
a. Fixed at the displayed value.

The estimated model equation is

$$\log(\hat{odds}) = \ln\left(\frac{\hat{\pi}(x)}{1-\hat{\pi}(x)}\right) = -7.055 + .027x$$

With the $EL_{50} = 7.055/0.027 = 261.30$ and the proportion at $x = 260$ estimated to be

$$\hat{\pi}(x) = \frac{\exp(-7.055 + .027(260))}{1 + \exp(-7.055 + .027(260))} = 0.491$$

This result is very sensitive to how many significant digits you carry during your calculations.

A one point increase in the test score multiplies the odds for a student to be in an academic program by an estimated $e^{0.027} = 1.02$, saying that the odds for being in the program increases by 2% for a one unit increase in the total score.
The 95% Wald confidence interval for β is given by (see output):[0.022, 0.033] which is giving a 95% confidence interval of e^{β}: [1.022, 1.033]
We can be 95% confident that the odds for being on an academic program increase between 2.2% and 3.3% when increasing x by one point.

Omnibus Test[a]

Likelihood Ratio Chi-Square	df	Sig.
139.082	1	.000

Dependent Variable: program
Model: (Intercept), x

a. Compares the fitted model against the intercept-only model.

With the omnibus test we test if at least one of the slope parameters in the model is different from 0. Since in this example we have only one slope, in this example the test result gives the results for $H_0 : \beta = 0$.
Wald and likelihood ratio test for $\beta = 0$:

1. Hypotheses:

   $H_0 : \beta = 0$ versus $H_a : \beta \neq 0$

   $\alpha = 0.05$.

2. Assumptions: random samples, large sample size, both are met.

3. Test Statistic: $z_0 = \sqrt{106.411} = 10.32$ and $\chi_0^2 = -2(L_0 - L_1) = 139.082, \quad df = 1$

4. For both tests P-value < 0.001

5. Reject H_0

6. At significance level of 5% the data provide sufficient evidence that the odds for being on an academic program relate to the test results.

For $x = 350$ a 95% confidence interval for the true probability $\pi(350)$ is given as [.883, .954]. This has been found running SPSS and have it save the confidence intervals in the spreadsheet.

The odds ratio for being on the program for students with scores of $x = 300$ versus $x = 200$ equals $\theta(300, 200) = (e^{\hat{\beta}})^{300-200} = 1.02^{100} = 7.24$ The odds for being on an academic program are 7.24 times higher for students with a score of 300 than for students with a score of 200.

Goodness of Fit[b]

	Value	df	Value/df
Deviance	612.542	520	1.178
Scaled Deviance	612.542	520	
Pearson Chi-Square	530.969	520	1.021
Scaled Pearson Chi-Square	530.969	520	
Log Likelihood[a]	-346.134		
Akaike's Information Criterion (AIC)	696.268		
Finite Sample Corrected AIC (AICC)	696.291		
Bayesian Information Criterion (BIC)	704.783		
Consistent AIC (CAIC)	706.783		

Dependent Variable: program
Model: (Intercept), x

a. The full log likelihood function is displayed and used in computing information criteria.

b. Information criteria are in small-is-better form.

The deviance/df = 1.178 indicates excellent model fit.
The graph illustrates how close the data points are to the estimated model function. Small residuals. The data points are based on grouping data for values of x, and finding the proportion of students in the academic program for these groups.