3 Generalized Linear Models

3.1 Binary Response

Most models in statistics

1. relate explanatory and response variables
2. describe pattern of association
3. have parameters which describe nature/strength of association
4. permit to test for association: based on sample data while controlling for confounding variables
5. can be use for prediction

GLMs are defined by their random and systemic component, and their link function.

- Random Component: It identifies that response variable Y and its distribution. The n observations $Y_1, Y_2, \ldots, Y_n$ are assumed to be independent. Here the Ys have categorical outcomes.

  Binary data: 2 possible outcomes (Success or Failure coded 1 and 0 in raw data)
  Example: $Y = 1$ or $Y = 0$ (Y is Bernoulli or Binomial with $n = 1$)

  Count data: Counts are outcomes
  Example: Y = number of successes in a given time interval or a specified region of space (Y is Poisson)

- Systematic Component: It specifies how explanatory variables are entered into the model. Here usually with a linear predictor $\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$

- Link Functions: It links the mean of the random component and the systematic components using a function $g(\mu)$, where μ is the mean of Y.

Example 1
HSB (High School and Beyond):
A sample of 600 high school seniors, taken from the Large scale Longitudinal Study conducted by National Opinion Research Center (1980) under contract with National Center for Education Statistics. It includes the following variables.

ID	Student Identification Number (3 digits)
SEX	1-Male 2-Female
RACE	(Race or Ethnicity) 1-Hispanic 2-Asian 3 - Black 4- White
SES	Socioeconomic Status 1-Low 2-Medium 3-High
SCTYP	School Type 1-Public 2-Private
HSP	High School Program 1-General 2-Academic Preparatory 3- Vocational/technical
LOCUS	Locus of Control Standardized to mean 0 and st dev 1
CONCEPT	Self Concept Standardized to mean 0 and st dev 1
MOTIVATION	Motivation Average of 3 motivation items
CAR	Career Choice 1-Clerical 2-Craftsperson 3- Farmer 4-Homemaker 5-Laborer 6-Manager 7-Mi
READING	Reading T-score standardized to mean 50 and SD 10
WRITING	Writing T-score standardized to mean 50 and SD 10
MATH	Math T-score standardized to mean 50 and SD 10
SCIENCE	Science T-score standardized to mean 50 and SD 10
CIVICS	Civics T-score standardized to mean 50 and SD 10
x	Total of RDG, WRTG, MATH, SCI, CIV
program	dummy for academic program yes=0, no=1

Below find the SPSS commands to estimate the linear equation for the GLM of interest, and discuss and interpret results. In SPSS terminology, we reference the non-academic programs, as our level of most interest is academic programs, and we wish $\pi(x)$, the probability of success to be the probability of attending an academic program. We will look at two possible ways to input SPSS commands to obtain information of interest.
Note: It is important to choose the correct reference category for the chosen dependent/response variable.
Please observe how the SPSS user interface follows the logic of choosing link function, random and systematic component by moving through the tabs.

For the Linear Probability Model:

1. Analyze>Generalized Linear Models>Generalized Linear Models

2. Type of Model Tab:
   Choose Custom
   Distribution: Binomial
   Link Function: Identity

3. Response Tab:
   Dependent Variable: program (Binary reference category: Last highest value (*category we are not interested in*)

4. Predictor Tab:
   Covariate: x (*because it is a numerical predictor*)

5. Model Tab:
   x as a main effect into Model box. check Include intercept in model

6. OK

Iteration History

Iteration	Update Type	Number of Step-halvings	Log Likelihood[a]	Parameter			Number of Invalid Cases[c]
				(Intercept)	x	(Scale)	
0	Initial	0	-327.270	-.968450	.005700	1	4
1	Scoring	0	-326.932	-.848214	.005253	1	
2	Newton	0	-326.789	-.897262	.005432	1	2
3	Newton	1	-326.788	-.893459	.005419	1	2
4	Newton	10	-326.788[b]	-.893459	.005419	1	2

Redundant parameters are not displayed. Their values are always zero in all iterations.
Dependent Variable: program
Model: (Intercept), x

a. The full log likelihood function is displayed.

b. The maximum number of step-halvings was reached but the log likelihood value cannot be further improved.

c. These cases are invalid because there are errors in computing the inverse identity link function, the log-likelihood, the gradient, or the Hessian matrix in the iterative process. Invalid cases are excluded from iterations in which they cause computational errors.

The warning is saying that the iteration did not converge, and no final model could be reached, because there are still invalid cases (probabilities outside the interval from 0 to 1.

For the Logistic Regression Model:

1. Analyze>Generalized Linear Models>Generalized Linear Models

2. Type of Model Tab:
   Choose Custom
   Distribution: Binomial
   Link Function: Logit

3. Response Tab:
   Dependent Variable: program (Binary reference category: Last highest value (*category we are not interested in*)

4. Predictor Tab:
   Covariate: x (*because it is a numerical predictor*)

5. Model Tab:
   x as a main effect into Model box. check Include intercept in model

6. OK

Model Information

Dependent Variable	program[a]
Probability Distribution	Binomial
Link Function	Logit

a. The procedure models .00 as the response, treating 1.00 as the reference category.

Case Processing Summary

	N	Percent
Included	600	99.8%
Excluded	1	0.2%
Total	601	100.0%

Categorical Variable Information

			N	Percent
Dependent Variable	program	.00	308	51.3%
		1.00	292	48.7%
		Total	600	100.0%

Continuous Variable Information

		N	Minimum	Maximum	Mean	Std. Deviation
Covariate	x	600	164.70	350.00	259.9447	40.44849

Goodness of Fit[a]

	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[b]	-325.509		
Akaike's Information Criterion (AIC)	655.019		
Finite Sample Corrected AIC (AICC)	655.039		
Bayesian Information Criterion (BIC)	663.813		
Consistent AIC (CAIC)	665.813		

Dependent Variable: program
Model: (Intercept), x

a. Information criteria are in smaller-is-better form.

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

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.

The first output tables include information about the model, some data description for the data included, goodness of fit table, and the omnibus test.
The Deviance/df is below 2.5 indicating acceptable model fit.
The omnibus test gives likelihood ratio test statistic and p-value for testing H_0 : all slope parameters are 0, which in this case can reject at $\alpha = 0.05$.

Parameter Estimates

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

Dependent Variable: program
Model: (Intercept), x

a. Fixed at the displayed value.

The last table on the parameter estimates, gives the point estimates, standard error of the estimate, the Wald Chi-Square statistic and P-value for the relevant parameter being 0.
For the example the estimated regression function is: $logit(\hat{\pi}(x)) = -7.055 + 0.027x$. To help with the interpretation of the equation, request SPSS to include the exponentiated parameter estimates by checking in the **Statistics Tab: Include exponential parameter estimates** .

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

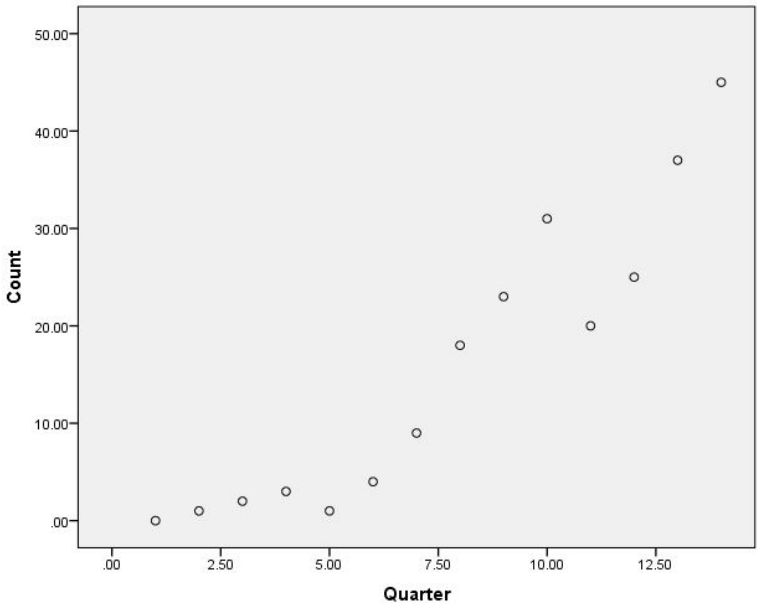
a. Fixed at the displayed value.

$exp(\hat{\beta}_1) = 1.028$, for every 1 point increase in the total exam score the odds for being in an academic program increases by 2.8%. The 95% confidence interval indicates that we can be 95% confident that the increase falls between 2.2% and 3.3%.

3.2 Count Response

Example 2
Will see how to analyse the data on death due to AIDS in Australia in 3 month periods from January 1983 to June 1986.
The data are in AIDS.sav. First get a scatteplot of the data, by following the SPSS commands below.

1. Graph>Legacy Dialogs>Scatter/Dot>SimpleScat
2. Y Axis: Count
3. X Axis: Quarter
4. OK



The data exhibit an exponential pattern, which means that a log linear model can be appropriate.

SPSS commands To fit a Loglinear Model to the data:

1. Analyze>Generalized Linear Models>Generalized Linear Models
2. Type of Model Tab:
   Choose Custom
   Distribution: Poisson
   Link Function: Log

3. Response Tab:
   Dependent Variable: count

4. Predictor Tab:
   Covariate: Quarter (*because it is a numerical predictor*)

5. Model Tab:
   Quarter as a main effect into Model box. check Include intercept in model

6. Statistics Tab:
   check Include exponential parameter estimates

7. OK

Goodness of Fit[a]

	Value	df	Value/df
Deviance	29.654	12	2.471
Scaled Deviance	29.654	12	
Pearson Chi-Square	28.847	12	2.404
Scaled Pearson Chi-Square	28.847	12	
Log Likelihood[b]	-41.290		
Akaike's Information Criterion (AIC)	86.581		
Finite Sample Corrected AIC (AICC)	87.672		
Bayesian Information Criterion (BIC)	87.859		
Consistent AIC (CAIC)	89.859		

Dependent Variable: Count
Model: (Intercept), Quarter

a. Information criteria are in smaller-is-better form.

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

Omnibus Test[a]

Likelihood Ratio Chi-Square	df	Sig.
177.619	1	.000

Dependent Variable: Count
Model: (Intercept), Quarter

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

Tests of Model Effects

Source	Type III Wald Chi-Square	df	Sig.
(Intercept)	1.828	1	.176
Quarter	135.477	1	.000

Dependent Variable: Count
Model: (Intercept), Quarter

Deviance/df=2.471 indicates acceptable model fit. The omnibus test informs that at significance level of 5% the slope for time(Quarter) is different from 0 ($\chi^2(1) = 177.619$, p-value<0.001).

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval Lower	95% Wald Confidence Interval Upper	Hypothesis Test Wald Chi-Square	Hypothesis Test df	Hypothesis Test Sig.	Exp(B)	95% Wald Confidence Interval for Exp(B) Lower	95% Wald Confidence Interval for Exp(B) Upper
(Intercept)	.340	.2512	-.153	.832	1.828	1	.176	1.404	.858	2.298
Quarter	.257	.0220	.213	.300	135.477	1	.000	1.292	1.238	1.349
(Scale)	1[a]									

Dependent Variable: Count
Model: (Intercept), Quarter

a. Fixed at the displayed value.

From the table we find the estimated model equation is $log(\hat{\mu})$= .340 + .257 Quarter, which is equivalent to $\mu = \exp(.340)\exp(.257)^{Quarter} = 1.404(1.292)^{Quarter}$. The number of AIDS deaths in

Australia over a 3 month period was in average 1.29 times higher than in the previous 3 month period. On average in every quarter the mean number of death by AIDS in Australia in the year 1983 to 1986 went up by 29%. With a 95% confidence interval ranging from 23.8% to 34.9%.