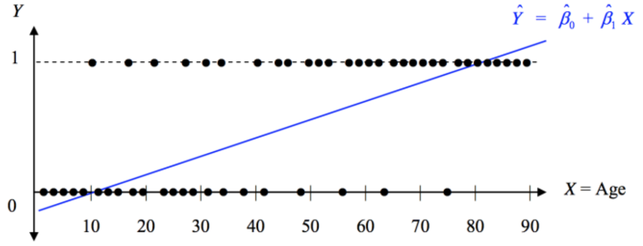
Chapter 15. Generalized Linear Models (GLM)

15.1. Motivation: Why GLM?

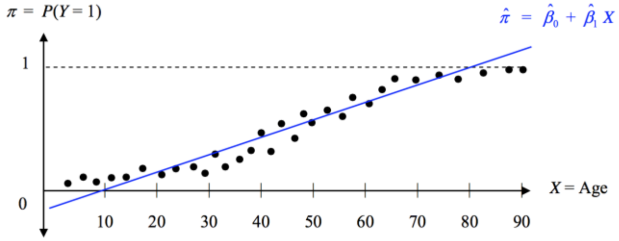
Example: *"If you live long enough, you will need a surgery."*

X = Age

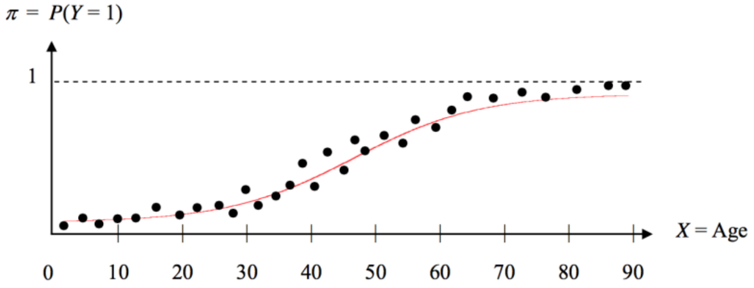
Y = Ever had a major surgery (1 = Yes, 0 = No)



Simple linear regression: Little predictive value for the response (either 0 or 1)



Modeling the probability of Y: Restricted to the finite interval / Violation of assumptions



Transform the probability π:

$$g(\pi) = \log\left(\frac{\pi}{1-\pi}\right) \in (-\infty, +\infty)$$

15.2. Generalized Linear Model (GLM)

- Framework to *generalize* the methods in linear models to the wide class of distributions
- Model functions of the mean
- Components

Component	Description
Random	Response variable Y with independent observations (Y_1, Y_2, …, Y_n) forms a distribution in a natural exponential family. $f(y;\theta) = h(y)\exp[T(y)\,b(\theta) - A(\theta)]$ e.g. Poisson, binomial, normal
Systematic	Systematic component involves the explanatory variables $x_1, x_2, \dots, x_p$ as linear predictors. $g(\mu) = \eta = \sum_{j=1}^{p} \beta_j x_j = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ where $E(Y_i) = \mu_i, i = 1, 2, \dots, n$.
Link	Link function $g(\cdot)$ describes the relationship between the random and systematic components. $g(\mu) = \eta$ e.g. $g(\mu) = \mu$: Identity link

- Types of GLM

Random	Support	Link		Model
Normal	$(-\infty, +\infty)$	Identity	$g(\mu) = \mu = X\beta$	Linear-response regression
Exponential Gamma	$(0, +\infty)$	Inverse	$g(\mu) = \frac{1}{\mu} = X\beta$	Exponential-response regression
Poisson	{0, 1, 2, …}	Log	$g(\mu) = \log(\mu) = X\beta$	Log-linear regression

Bernoulli
Binomial
	{0, 1}
	{0, 1, 2, … N}
		Logit	$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right) = X\beta$	Logistic regression

Multinomial	K outcomes	Logit	$\log\left(\frac{\Pr(Y=k)}{\Pr(Y=K)}\right) = \beta_k\, X$ $k = 1, 2, \dots, K-1$	Multinomial logistic regression

- In case of over-dispersion, consider negative binomial distribution instead of Poisson.
- Multinomial distribution with orders: Ordinal logistic regression
- Predictors (X) can take on any form: Binary, categorical, and/or continuous
- Log: Natural log (i.e. *ln*)

15.3. PROC GENMOD

General Syntax

```
proc genmod data=dataset;
     class categorical-variable(ref="Reference");
     model dependent-variable = list-of-independent-variables
           / dist = distribution link = link-function;
     lsmeans categorical-variable / <options>;
run;
```

- More flexible than PROC GLM with a choice of link functions
- CLASS: Specify categorical variables and their reference category.
- (Distribution) DIST = normal (default), poisson, bin, negbin
- (Link function) LINK = identity (default), log, logit, probit, cloglog
- LSMEANS: Compute least squares means corresponding to the specified effects.

Option	Description
ALPHA = *n*	Specify the level for the confidence limits. Between 0 (100% confidence) and 1 (0% confidence). Default is 0.05 (95% confidence limits).
CL	Request the confidence limits for each of the LS-means.
CORR [COV]	Request the estimated correlation [covariance] matrix of the LS-means.

- PROC HPGENSELECT: Conduct model selection

15.4. Log-linear Regression

- Random component

$$Y_i \mid X \sim Poisson(\lambda_i),\ \ E(Y_i \mid X) = \lambda_i,\ \ i = 1, 2, \dots, n$$

- Systematic component: Linear predictor $(x_1, x_2, \dots, x_p)$

$$\eta_i = \sum_{j=1}^{p} \beta_j x_{ij} = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

- Link function (log)

$$g(\lambda_i) = \log(\lambda_i) \in (-\infty, +\infty)$$

- Log-linear regression

$$g(\lambda_i) = \log(\lambda_i) = \sum_{j=1}^{p} \beta_j x_{ij} = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

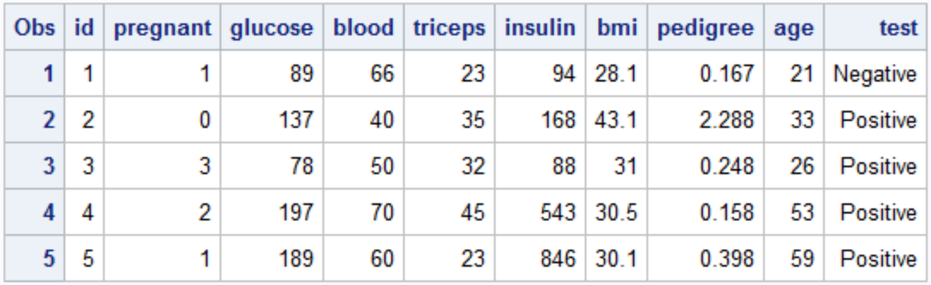
- SAS: PROC GENMOD

General Syntax

```
proc genmod data=dataset;
     class categorical-variable(ref="Reference");
     model dependent-variable = list-of-independent-variables
           / dist = poisson link = log;
run;
```

Example: Log-linear regression

Raw Data

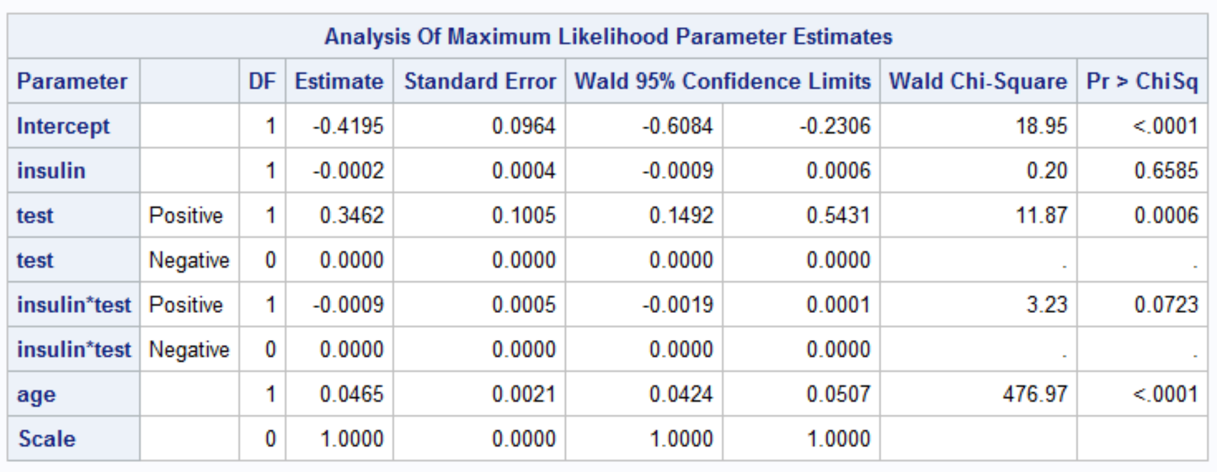


Obs	id	pregnant	glucose	blood	triceps	insulin	bmi	pedigree	age	test
1	1	1	89	66	23	94	28.1	0.167	21	Negative
2	2	0	137	40	35	168	43.1	2.288	33	Positive
3	3	3	78	50	32	88	31	0.248	26	Positive
4	4	2	197	70	45	543	30.5	0.158	53	Positive
5	5	1	189	60	23	846	30.1	0.398	59	Positive

SAS Code

```
* Poisson distribution / Log link;
proc genmod data=pima;
     class test(ref="Negative");
     model pregnant = insulin|test age / dist = poisson link = log;
run;
```

Output



Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.4195	0.0964	-0.6084	-0.2306	18.95	<.0001
insulin		1	-0.0002	0.0004	-0.0009	0.0006	0.20	0.6585
test	Positive	1	0.3462	0.1005	0.1492	0.5431	11.87	0.0006
test	Negative	0	0.0000	0.0000	0.0000	0.0000	.	.
insulin*test	Positive	1	-0.0009	0.0005	-0.0019	0.0001	3.23	0.0723
insulin*test	Negative	0	0.0000	0.0000	0.0000	0.0000	.	.
age		1	0.0465	0.0021	0.0424	0.0507	476.97	<.0001
Scale		0	1.0000	0.0000	1.0000	1.0000		

15.5. Logistic Regression

- Random component

$$Y_i \mid X \sim Binomial(n_i, p_i), \;\; E(Y_i/n_i \mid X) = \; p_i, \;\; i = 1, 2, \dots, n$$

- Systematic component: Linear predictor $(x_1, x_2, \dots, x_p)$

$$\eta_i = \sum\nolimits_{j=1}^{p} \beta_j x_{ij} = \beta_1 x_{i1} + \; \beta_2 x_{i2} + \cdots + \; \beta_p x_{ip}$$

- Link function (Logit)

$$g(p_i) = logit(p_i) = \log\left(\frac{p_i}{1 - \; p_i}\right) \in (-\infty, +\infty)$$

- Logistic regression

$$g(p_i) = \mathrm{logit}(p_i) = \sum\nolimits_{j=1}^{p} \beta_j x_{ij} = \beta_1 x_{i1} + \; \beta_2 x_{i2} + \cdots + \; \beta_p x_{ip}$$

- SAS: PROC GENMOD

General Syntax

```
proc genmod data=dataset;
     class categorical-variable(ref="Reference");
     model dependent-variable = list-of-independent-variables
           / dist = bin link = logit;
run;
```

- SAS: PROC LOGISTIC

General Syntax

```
proc logistic data=dataset descending;
     class categorical-variable(ref="Reference") / param = ref;
     model dependent-variable = list-of-independent-variables / lackfit;
run;
```

- DESCENDING: Sort the response variable from highest to lowest.
- By default, SAS models the probability of the lower category.
- PARAM = REF: Use the specified reference values for modeling.
- LACKFIT: Provide the Hosmer-Lemeshow for goodness-of-fit test

  H_0: The logistic regression fits well.

- Interpretation

  - The sign of β determines whether the log odds of Y is increasing or decreasing.

  - If $\beta = 0$, then there is no linear relationship between the *log odds* of Y and X.

  - Odds ratio (OR) = e^{β}

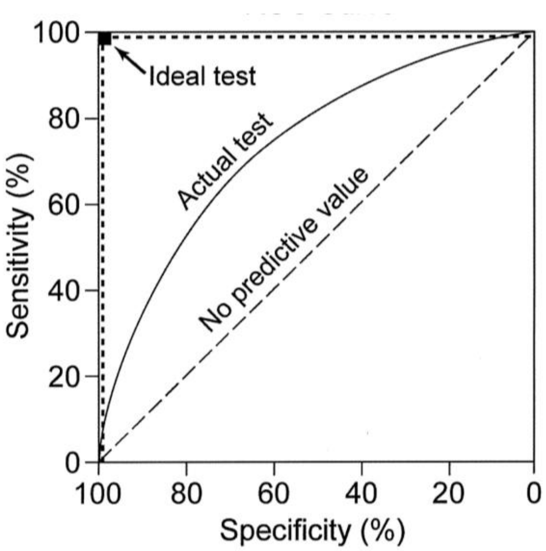
    1) Ratio of the probability of success (group 1) and that of failure (group 2)

    2) $OR \in [0, +\infty)$

    3) OR = 1: There is no difference between the groups compared.

    4) OR > 1: Group 1 has a greater probability than group 2.

- Receiver operating characteristic (ROC) curve
  - Sensitivity (True positive rate) / Specificity (True negative rate)
  - A model with high discrimination ability will have high sensitivity and specificity simultaneously, leading to the ROC curve getting close to the top left corner of the plot.
  - Area under the curve (AUC): Provide the probability that a randomly selected pair of subjects (one truly positive and one truly negative) will be correctly ordered by the test.
  - AUC ∈ [0.5 (No discrimination), 1 (Perfect discrimination)]

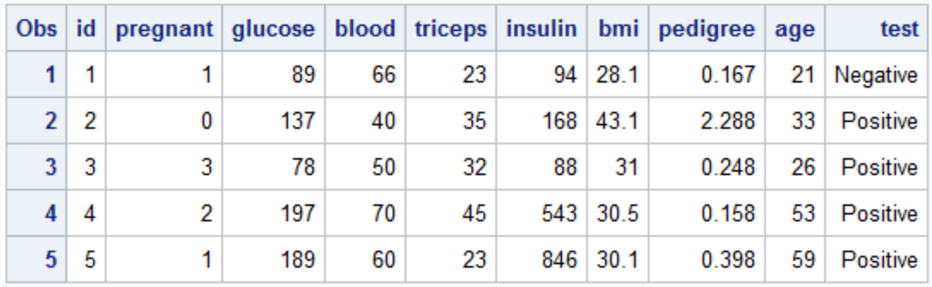


15.6. Comparison between Procedures

Procedure	Description
PROC REG	Perform a linear regression with diagnostic tests.
PROC GLM	Perform a simple/multiple/polynomial/weighted regression. Provide a wide range of options for analysis with limited model-checking capacity.
PROC LOGISTIC	Perform logistic regression with diagnostic tests.
PROC GENMOD	Fit a generalized linear model using MLE.

Example: Logistic regression

Raw Data



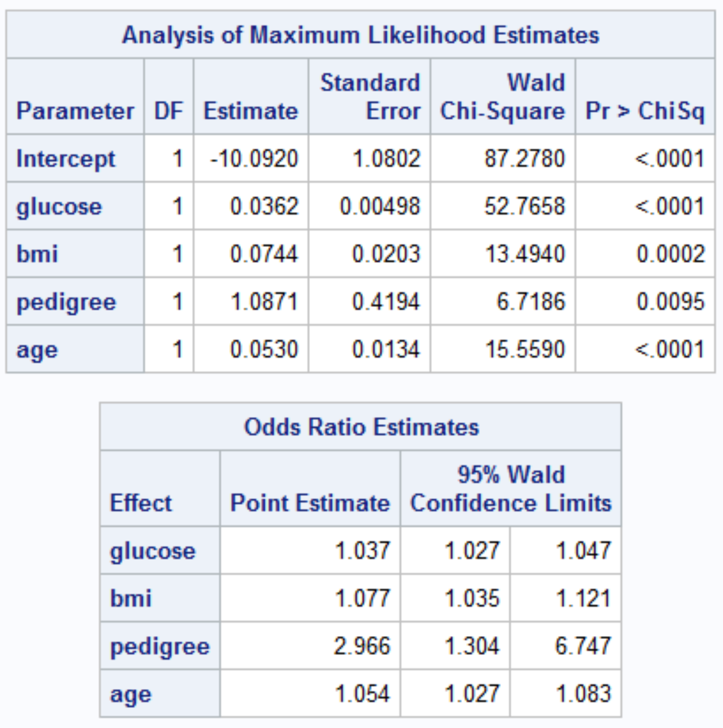
Obs	id	pregnant	glucose	blood	triceps	insulin	bmi	pedigree	age	test
1	1	1	89	66	23	94	28.1	0.167	21	Negative
2	2	0	137	40	35	168	43.1	2.288	33	Positive
3	3	3	78	50	32	88	31	0.248	26	Positive
4	4	2	197	70	45	543	30.5	0.158	53	Positive
5	5	1	189	60	23	846	30.1	0.398	59	Positive

SAS Code

```
* Binomial distribution / Logit link;
proc genmod data=pima descending;
     model test = glucose bmi pedigree age / dist = bin link = logit;
run;

* PROC LOGISTIC;
proc logistic data=pima plots(only)=(roc effect);
     class test (ref="Negative") / param=ref;
     model test = glucose bmi pedigree age / lackfit outroc=roc;
run;
```

Output

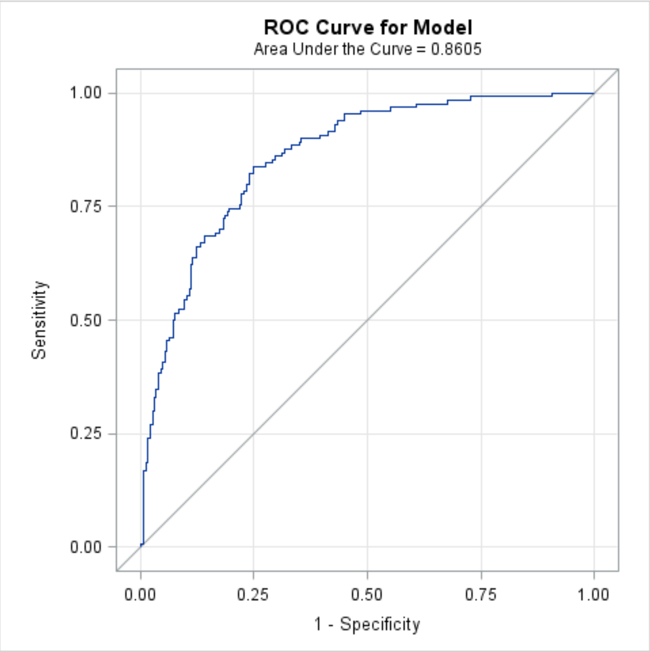


Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-10.0920	1.0802	87.2780	<.0001
glucose	1	0.0362	0.00498	52.7658	<.0001
bmi	1	0.0744	0.0203	13.4940	0.0002
pedigree	1	1.0871	0.4194	6.7186	0.0095
age	1	0.0530	0.0134	15.5590	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
glucose	1.037	1.027	1.047
bmi	1.077	1.035	1.121
pedigree	2.966	1.304	6.747
age	1.054	1.027	1.083


ROC Curve for Model
Area Under the Curve = 0.8605
Sensitivity
1 - Specificity