Logistic Regression with SAS

Summer Semester 2018/2019

Adam Korczyński

The Collegium of Economic Analysis of Warsaw School of Economics

Institute of Statistics and Demography

Event History and Multilevel Analysis Unit

e-mail: akorczy@sgh.waw.pl

Contingency table and measures of association

Agenda

- Introductory notes: scope of the class, schedule and requirements
- Basic notions
- Contingency table
- Comparing two proportions - measures in the analysis of categorical data
- Types of studies
- Calculating risk, risk ratio, odds and odds ratio
- Proc Freq in SAS
- Logistic regression
- Basic logistic regression models

Basic notions (Agresti, 2002)

- Categorical variable: scale consisting of a set of categories
- Response *Y* – Explanatory variable *X*
- Scales:
  - **Nominal**
  - **Ordinal**
  - Invterval
  - Ratio

Contingency table

- *Y* and *X* – two categorical variables with *k*,*l* categories respectively
- We can present the distribution of the two variables using a table with *k* rows of *Y* and *l* columns for *X* categories
- The cells represent possible outcomes
- The table with frequency counts of outcomes for a sample, the table is called contingency table (by Karl Pearson, 1904).
- 2x2 table:

	X=x_1	X=x_2
Y=y_1	n_{11}	n_{12}
Y=y_2	n_{21}	n_{22}

E.g. *X*- tax regimen (A, B); *Y*- survival of a company at time *t*

Logistic Regression with SAS ©AK, SGH 2019

Comparing two proportions - measures in the analysis of categorical data

$$Y = \begin{cases} 1 & company\ is\ on\ the\ market\ at\ time\ t \\ 0 & company\ got\ out\ of\ the\ market\ prior\ to\ t \end{cases}$$

$$X = \begin{cases} 1 & tax\ regimen\ A \\ 0 & tax\ regimen\ B \end{cases}$$

$$Risk\ (conditional\ probability){:}\ R_{x_j} = P(Y = i|x = j)$$

$$Relative\ risk{:}\ \ RR_{10} = \frac{P(Y = i|x = 1)}{P(Y = i|x = 0)}$$

$$Odds\ (comparing\ the\ probability\ with\ complementary\ probability){:}\ \ O_{x_j} = \frac{P(Y = i|x = j)}{1 - P(Y = i|x = j)}$$

$$Odds\ ratio{:}\ OR_{x_{10}} = \frac{O_{x_1}}{O_{x_0}}$$

2x2 contingency table
Calculating R, RR, O, OR

	$X=1$	$X=0$	$n_{i.}$
$Y=1$	n_{11}	n_{12}	$n_{1.}$
$Y=0$	n_{21}	n_{22}	$n_{2.}$
$n_{.j}$	$n_{.1}$	$n_{.2}$	n

$R_1 = P(Y=1|X=1) =$

$R_0 = P(Y=1|X=0) =$

$RR_{10} = \frac{P(Y=1|X=1)}{P(Y=1|X=0)} =$

$O_1 = \frac{P(Y=1|X=1)}{1-P(Y=1|X=1)} =$

$O_0 = \frac{P(Y=1|X=0)}{1-P(Y=1|X=0)} =$

$OR_{10} = \frac{P(Y=1|X=1)}{1-P(Y=1|X=1)} / \frac{P(Y=1|X=0)}{1-P(Y=1|X=0)} =$

Logistic Regression with SAS ©AK, SGH 2019

Types of studies

- Follow-up study (prospective)
  - Risk factor → Observation → Outcome
  - Assigning tax regimen A or B →Observation → Survived or not
- Case-control study (retrospective)
  - Risk factor ← Observation ← Outcome
  - what was the tax regimen under which companies were operating? ← Survived or not
- Cross-sectional
  - Snapshot of data at a given time point. Allows assessing coexistence of features.

Prospective study – experimental design

- Set a target sample of subjects (e.g. companies) to be in the study
- Randomize subjects into groups of interest, e.g. tax regimens *x*
- Then observe the companies at scheduled time points: 6 months, 1 year, 2 years and determine the companies that are still operating *Y*
- In this setting *Y* is a random variable and *x* is fixed. Therefore we can estimate P(Y=i|x=j) the probability of being on the market at time *t* given that the company was operating under tax regimen *j*

Retrospective studies

- Research organization decides to study the status of the Polish companies in a specific industry in years 2015-2016. There were n_1 companies which went bankrupt in this period. For the purpose of the comparison n_2 companies operating throughout the whole period have been added to the sample. The tax regimens under which the companies were operating were identified afterwards.
- In this setting *X* is a random variable and *Y* is set *post hoc*. We do identify the status of survival at first and then identify the characteristics. As a consequence we can estimate P(X=j|y=i). This would mean probability of being in a tax regimen A or B given that a company survived/not survived. This does not address the research question though. The counts for Y are not collected through a random process and therefore P(Y=i|X=j) is not good estimate of survival conditional on tax regimen.

Numeric example – comparison of measures of association

Population					Survival				
	Regimen A	Regimen B	ni.			RISKS	RR	ODDS	OR
Survived	478	220	698		Regimen A	0.080282	**1.47646**	0.08729	**1.51805**
Not survived	5476	3826	9302		Regimen B	0.054375		0.0575	
n.j	5954	4046	10000						
Prospective sample									
	Regimen A	Regimen B	ni.			RISKS	RR	ODDS	OR
Survived	95	44	139		Regimen A	0.079832	**1.46963**	0.08676	**1.51038**
Not survived	1095	766	1861		Regimen B	0.054321		0.05744	
n.j	1190	810	2000						
Retrospective sample									
	Regimen A	Regimen B	ni.			RISKS	RR	ODDS	OR
Survived	616	284	900		Regimen A	0.537522	**1.23781**	1.16226	**1.51422**
Not survived	530	370	900		Regimen B	0.434251		0.76757	
n.j	1146	654	1800						

The chance of survival within the studied time period is 50% greater in companies operating under tax regimen A.

Proc Freq in SAS

PROC FREQ < options > ;
    BY variables ;
    EXACT statistic-options < / computation-options > ;
    OUTPUT <OUT=*SAS-data-set* > output-options ;
    TABLES requests < / options > ;
    TEST options ;
    WEIGHT variable < / option > ;

BY	Provides separate analyses for each BY group
EXACT	Requests exact tests
OUTPUT	Requests an output data set
TABLES	Specifies tables and requests analyses
TEST	Requests tests for measures of association and agreement
WEIGHT	Identifies a weight variable

Source:
http://support.sas.com/documentation/cdl/en/procstat/66703/HTML/default/viewer.htm#procstat_freq_syntax.htm

Exercise

- Assess the risk of lung cancer among smokers numerically in order to answer the following question: What is the risk of having LC while being a smoker as compared to not being a smoker?
- Use Excel and Proc Freq in SAS

	Smoker	
Lung Cancer	Yes	No
Yes	688	21
No	650	59

Data source: Agresti, 2002, Table 2.5.

SAS code

```
data a01;
input y$ x$ n;
cards;
L S 688
L X 21
N S 650
N X 59
;
run;

proc freq data=a01;
      tables y*x / relrisk;
      weight n;
run;
```

Logistic regression

The response variable is given by Bernoulli distribution, e.g.:

$$Y = \begin{cases} 1 & company\ is\ on\ the\ market\ at\ time\ t \\ 0 & company\ got\ out\ of\ the\ market\ prior\ to\ t \end{cases}$$

- The purpose of the analysis is to estimate the probability of observing an event *Y*=*y* conditionally on *X*=*x* which is P(*Y*=*y*|*X*=*x*). The explanatory variables in the model can be:
  - discrete
  - continuous
- Example of research questions: finding factors affecting company survival, identifying variables having an impact on decision making process e.g. in marketing to say why customer chooses to buy a product.

Notation in the logistic model

Each observation on *Y* has Bernoulli distribution with parameters:

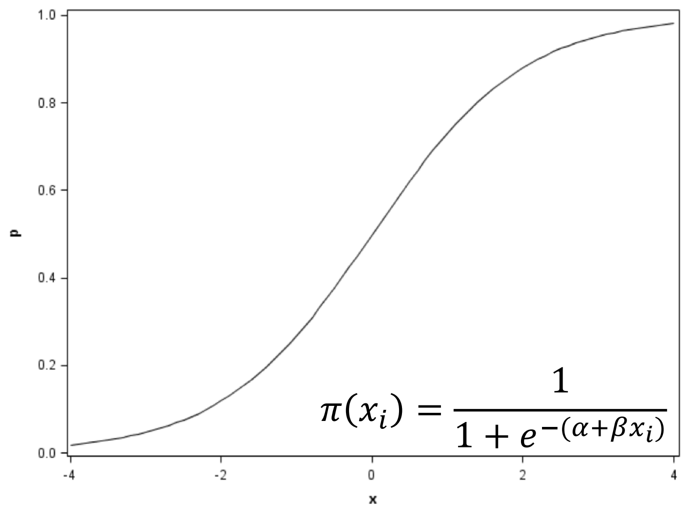
$$P(Y_i = 1) = \pi_i$$
$$P(Y_i = 0) = 1 - \pi_i$$

$$\pi(x_i) = P(Y = 1|X = x_i) = \frac{exp(\alpha + \beta x_i)}{1 + exp(\alpha + \beta x_i)} = \frac{1}{1 + e^{-(\alpha + \beta x_i)}}$$

Through logit transformation we get:

$$logit[\pi(x_i)] = \ln \frac{\pi(x_i)}{1 - \pi(x_i)} = \alpha + \beta x_i$$

Logistic curve



$$\pi(x_i) = \frac{1}{1 + e^{-(\alpha + \beta x_i)}}$$

Logistic regression models to be discussed

- Binary(i=2)

$$Y = \begin{cases} 1 & company\ is\ on\ the\ market\ at\ time\ t \\ 0 & company\ got\ out\ of\ the\ market\ prior\ to\ t \end{cases}$$

- Ordinal (i>2)

$$Y = \begin{cases} 0 & Good \\ 1 & Moderate \\ 2 & Poor \end{cases}$$

- Multinomial (i>2)

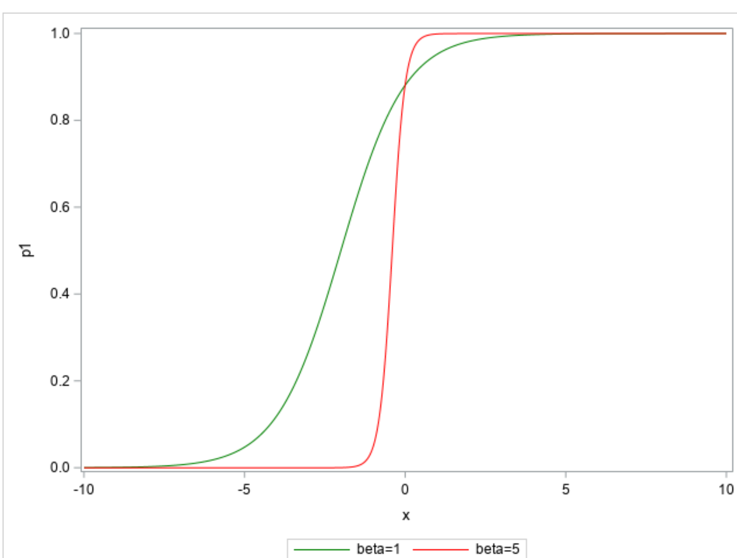
$$Y = \begin{cases} 0 & Standard\ phone \\ 1 & Android \\ 2 & iPhone \end{cases}$$

Logistic curve on an example – program code

```
%let beta1a=1;
%let beta1b=5;

data a01;
 beta0=2;
 beta1a=&beta1a.;
 beta1b=&beta1b.;
 do x=-10 to 10 by 0.01;
  p1=1/(1+exp(-1*(beta0+beta1a*x)));
  p2=1/(1+exp(-1*(beta0+beta1b*x)));
  output;
 end;
run;

proc sgplot data=a01;
     series x=x y=p1 / lineattrs=(color=green) legendlabel="beta=&beta1a.";
    series x=x y=p2 / lineattrs=(color=red) legendlabel="beta=&beta1b.";
run;
```

19

Logistic curve on an example – output

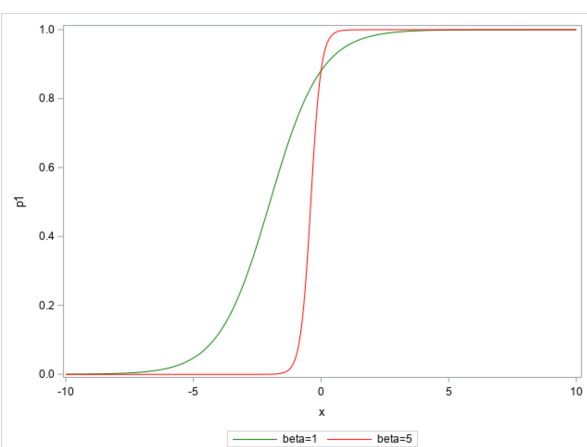


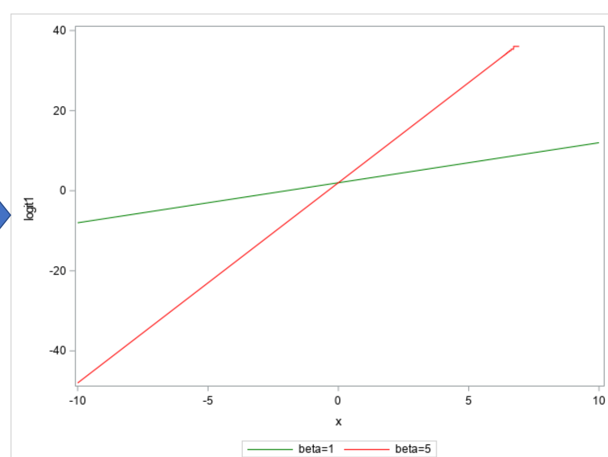
20

Logit transformation on an example – program code

```
%let beta1a=1;
%let beta1b=5;

data a01;
 beta0=2;
 beta1a=&beta1a.;
 beta1b=&beta1b.;
 do x=-10 to 10 by 0.01;
  p1=1/(1+exp(-1*(beta0+beta1a*x)));
  logit1=log(p1/(1-p1));
  p2=1/(1+exp(-1*(beta0+beta1b*x)));
  logit2=log(p2/(1-p2));
 output;
 end;
run;

proc sgplot data=a01;
 series x=x y=logit1 / lineattrs=(color=green) legendlabel="beta=&beta1a.";
series x=x y=logit2 / lineattrs=(color=red) legendlabel="beta=&beta1b.";
run;
```


Logit transformation on an example – output





Multivariate binary logistic regression model

$$logit[\pi(x_i)] = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$$

- $logit[\pi(x_i)]$ - logit of the probability of the event
- β_0 - intercept of the regression equation
- β_k - parameter estimate of the *k*th predictor variable

PROC LOGISTIC in SAS

- The basic syntax of the LOGISTIC procedure is as follows:

```
PROC LOGISTIC <options> ;
 CLASS variable <(options)><variable <(options)> ></ options> ;
 <label:> MODEL events/trials=<effects></ options> ;
 ODDSRATIO <'label'> variable </ options> ;
 OUTPUT <OUT=SAS-data-set><keyword=name <keyword=name >></ option> ;
RUN;
```

SAS OnLineDoc:
https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#logistic_toc.htm

Basic model assessment: comparing pairs $\pi(x_i)\rho\ \pi(x_j)$?

- To find concordant, discordant and tied pairs we compare subjects that had the outcome of interest against subjects that did not

Event = 0

$\hat{p}_i = ?$

Event = 1

$\hat{p}_j = ?$

Concordant pair

- Outcome in agreement with probabilities

Event = 0

$\hat{p}_i = 0.03$

Event = 1

$\hat{p}_j = 0.40$

Discordant pair

- Outcome in disagreement with probabilities

Event = 0

$\hat{p}_i = 0.80$

Event = 1

$\hat{p}_j = 0.60$

Tied pair

- The model cannot distinguish between the two subjects

Event = 0

$\hat{p}_i = 0.60$

Event = 1

$\hat{p}_j = 0.60$

german.sas

Exercise 1
Fit the logistic regression model using dataset German with Default as response variable and housing type as categorical variable.

1. Write the logistic model equation.
2. Assess the statistical significance of the parameter estimates.
3. Assess the risk of not paying off the loan using odds ratio estimates from the model.
4. Assess the effect of AGE, DURATION on risk of being default.

```
*Model 1;
proc logistic data=german;
class housing (param=ref ref='own');
model default(event='1')= housing;
output out=p predicted=p;
run;
```


The SAS System

The LOGISTIC Procedure

Model Information	
Data Set	WORK.GERMAN
Response Variable	default
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	1000
Number of Observations Used	1000

Response Profile		
Ordered Value	**default**	**Total Frequency**
1	0	700
2	1	300

Probability modeled is default=1.

Logistic Regression with SAS ©AK, SGH 2019

Class Level Information			
Class	Value	Design Variables	
housing	for free	1	0
	own	0	0
	rent	0	1

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1223.729	1210.049
SC	1228.636	1224.772
-2 Log L	1221.729	1204.049

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	17.6797	2	0.0001
Score	18.1998	2	0.0001
Wald	17.9443	2	0.0001

31

Logistic Regression with SAS ©AK, SGH 2019

The LOGISTIC Procedure

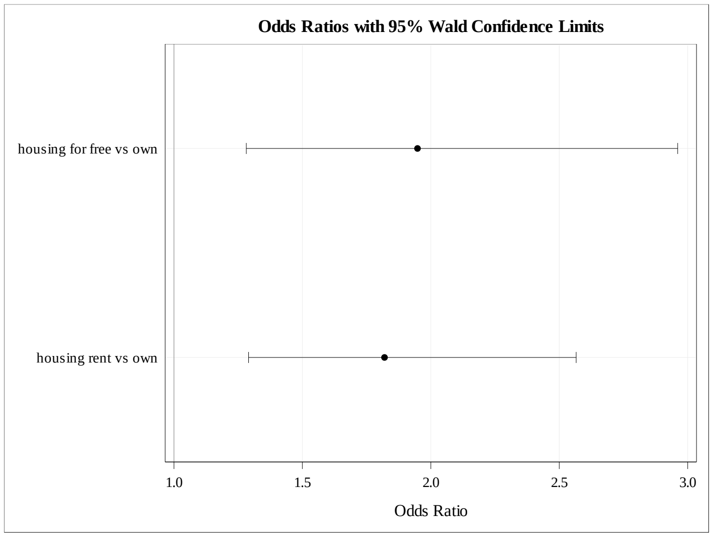
Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
housing	2	17.9443	0.0001

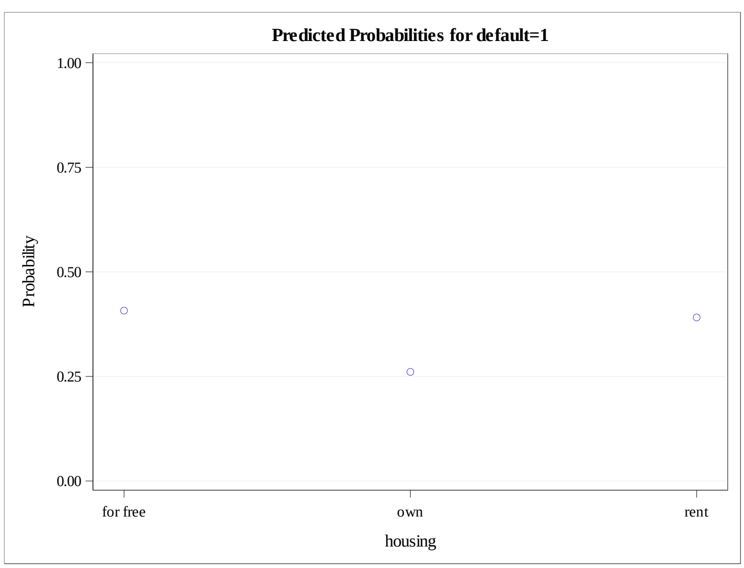
Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.0414	0.0853	149.1094	<.0001
housing	for free	1	0.6669	0.2136	9.7473	0.0018
housing	rent	1	0.5987	0.1753	11.6620	0.0006

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
housing for free vs own	1.948	1.282	2.961
housing rent vs own	1.820	1.291	2.566

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	30.9	Somers' D	0.134
Percent Discordant	17.5	Gamma	0.278
Percent Tied	51.7	Tau-a	0.056
Pairs	210000	c	0.567

32


Odds Ratios with 95% Wald Confidence Limits
housing for free vs own
housing rent vs own
1.0
1.5
2.0
2.5
3.0
Odds Ratio


Predicted Probabilities for default=1
1.00
0.75
0.50
0.25
0.00
Probability
for free
own
rent
housing

Binary logistic regression model

Logistic regression

The response variable is given by Bernoulli distribution, e.g.:

$$Y = \begin{cases} 1 & company\ is\ on\ the\ market\ at\ time\ t \\ 0 & company\ got\ out\ of\ the\ market\ prior\ to\ t \end{cases}$$

- The purpose of the analysis is to estimate the probability of observing an event *Y*=*y* conditionally on *X*=*x* which is P(*Y*=*y*|*X*=*x*). The explanatory variables in the model can be:
  - discrete
  - continuous

- Example of research questions: finding factors affecting company survival, identifying variables having an impact on decision making process e.g. in marketing to say why customer chooses to buy a product.

Specification of the logistic regression model

- Each observation on *Y* has Bernoulli distribution:

$$P(Y_i = 1) = \pi_i$$
$$P(Y_i = 0) = 1 - \pi_i \quad (0 < \pi_i < 1)$$

- In logistic regression model the probabilities are conditional on X_k (k=1,…, p):

$$\pi(x_i) = P(Y = 1 \mid X = x_i) = \frac{\exp(\alpha + \sum_{k=1}^{p} \beta_k x_{ki})}{1 + \exp(\alpha + \sum_{k=1}^{p} \beta_k x_{ki})} = \frac{1}{1 + e^{-(\alpha + \sum_{k=1}^{p} \beta_k x_{ki})}}$$

- Logit transformation provides:

$$\text{logit}\,[\pi(x_i)] = \ln \frac{\pi(x_i)}{1 - \pi(x_i)} = \alpha + \sum_{k=1}^{p} \beta_k x_{ki}$$

Maximum likelihood estimation

- The method of maximum likelihood yields values for the unknown parameters that maximize the probability of obtaining the observed set of data (Hosmer et al., 2013, section 1.2)
- Setup the *likelihood function* describing the process under analysis:

p(*y*) for observed set of outcomes (Y_1, Y_2, …, Y_n)

- (Y_1, Y_2, …, Y_n)→sample→(y_1, y_2, …, y_n)
- Assuming that individual *observations are independent* the probability:

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n) = P(Y_1 = y_1) P(Y_2 = y_2) \ldots P(Y_n = y_n) = \prod_{i=1}^{n} p(y_i)$$

- Likelihood function:

$$L(y_1, y_2, \ldots, y_n \mid \theta_1, \ldots, \theta_r) = \prod_{i=1}^{n} p(y_i \mid \theta_1, \ldots, \theta_r) \longrightarrow \underset{\theta_1, \ldots, \theta_r}{\max}$$

to get ML estimates: $\hat{\theta}_1, \ldots, \hat{\theta}_r$

Estimation in logistic regression model (1)

- For *n* independent observations from Bernoulli distribution likelihood function is:

$$L(y_1, y_2, ..., y_n \mid \theta_1, ..., \theta_r) = \prod_{i=1}^{n} p(y_i \mid \theta_1, ..., \theta_r)$$

$$= \pi(x_1)^{y_1}(1-\pi(x_1))^{(1-y_1)} \ldots \pi(x_n)^{y_n}(1-\pi(x_n))^{(1-y_n)} = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1-\pi(x_i))^{(1-y_i)}$$

- Natural logarithm of likelihood function gives:

$$\ln L(y_1, y_2, ..., y_n \mid \theta_1, ..., \theta_r) = \ln \prod_{i=1}^{n} \pi(x_i)^{y_i}(1-\pi(x_i))^{(1-y_i)}$$

$$= \sum_{i=1}^{n} \left\{ y_i \ln[\pi(x_i)] + (1-y_i)\ln[1-\pi(x_i)] \right\}$$

$$= \sum_{i=1}^{n} y_i \ln \frac{\pi(x_i)}{1-\pi(x_i)} + \sum_{i=1}^{n} \ln[1-\pi(x_i)]$$

$$\boxed{= \sum_{i=1}^{n} y_i \left[\alpha + \sum_{k=1}^{p} \beta_k x_{ki} \right] - \sum_{i=1}^{n} \ln\left[1 + \exp(\alpha + \sum_{k=1}^{p} \beta_k x_{ki}) \right]}$$

Estimation in logistic regression model (2)

- In order to get ML estimates the task is to find values of parameters that maximize the likelihood function. For a model with one explanatory variable the *likelihood equations* are found by:

$$\frac{\partial \ln L(y_1, y_2, ..., y_n \mid \theta_1, ..., \theta_r)}{\partial \alpha} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}$$

$$\frac{\partial \ln L(y_1, y_2, ..., y_n \mid \theta_1, ..., \theta_r)}{\partial \beta} = \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \frac{x_i \exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}$$

$$\left\{ \begin{array}{l} \dfrac{\partial \ln L(y_1, y_2, ..., y_n \mid \theta_1, ..., \theta_r)}{\partial \alpha} = 0 \\ \dfrac{\partial \ln L(y_1, y_2, ..., y_n \mid \theta_1, ..., \theta_r)}{\partial \beta} = 0 \end{array} \right. \longrightarrow \hat{\alpha}, \hat{\beta}$$

Estimates of the parameters in logistic regression model

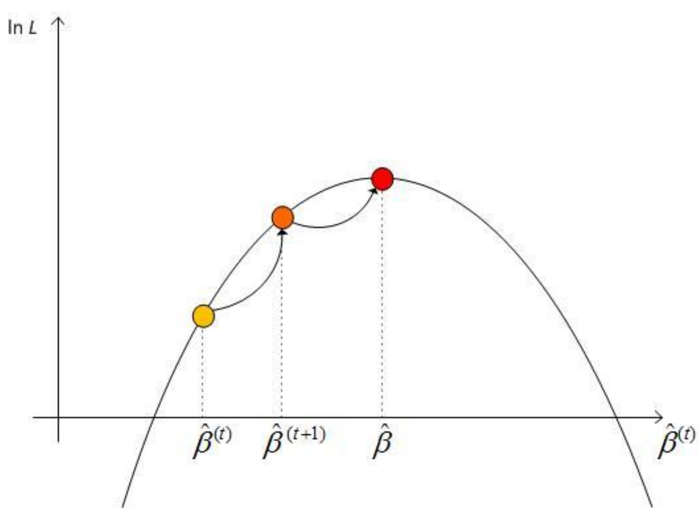
Newton-Raphson Algorithm

(Agresti, 2002, p. 143-145)

Fisher Scoring

(Agresti, 2002, p. 145-146)

Algorithms for model fitting visually



Newton-Raphson Algorithm numerically

- For *t* representing the iteration of the algorithm the procedure is:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{t} - \left(\mathbf{H}^{(t)}\right)^{-1}\mathbf{u}^{(t)}$$

$$\mathbf{u} = \begin{bmatrix} \dfrac{\partial L(\boldsymbol{\beta})}{\partial \beta_1} \\ \dfrac{\partial L(\boldsymbol{\beta})}{\partial \beta_2} \\ \vdots \\ \dfrac{\partial L(\boldsymbol{\beta})}{\partial \beta_p} \end{bmatrix}, \mathbf{H} = \begin{bmatrix} \dfrac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_1^2} & \dfrac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_1 \partial \beta_2} & \cdots & \dfrac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_1 \partial \beta_p} \\ \dfrac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_2 \partial \beta_1} & \dfrac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_2^2} & \cdots & \dfrac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_2 \partial \beta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_p \partial \beta_1} & \dfrac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_p \partial \beta_2} & \cdots & \dfrac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_p^2} \end{bmatrix}$$

- **H** is called the observed information matrix in the setting in Newton-Raphson algorithm.

Newton-Raphson Algorithm

Logistic regression (LR) model with one explanatory variable

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{t} - \left(\mathbf{H}^{(t)}\right)^{-1}\mathbf{u}^{(t)}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

$$\mathbf{u} = \begin{bmatrix} \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \dfrac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)} \\ \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \dfrac{x_i \exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)} \end{bmatrix}$$

$$\mathbf{H} = \begin{bmatrix} -\sum_{i=1}^{n} \dfrac{e^{\alpha + \beta x_i}}{\left(1 + e^{\alpha + \beta x_i}\right)^2} & -\sum_{i=1}^{n} \dfrac{x_i e^{\alpha + \beta x_i}}{\left(1 + e^{\alpha + \beta x_i}\right)^2} \\ -\sum_{i=1}^{n} \dfrac{x_i e^{\alpha + \beta x_i}}{\left(1 + e^{\alpha + \beta x_i}\right)^2} & -\sum_{i=1}^{n} \dfrac{x_i^2 e^{\alpha + \beta x_i}}{\left(1 + e^{\alpha + \beta x_i}\right)^2} \end{bmatrix}$$

Newton-Raphson Algorithm

Example, program code

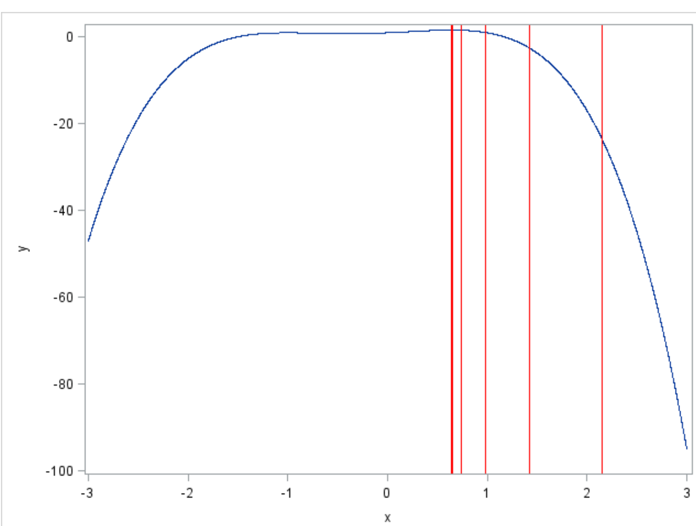
Code301.sas

```
*Function: y = -x.^4 -x.^3 + x.^2 + x + 1
Task find the maximum using Newton-Raphson algorithm;
data a01;
 do x=-3 to 3 by 0.001;
     y=-x**4-x**3+x**2+x+1;
       output;
    end;
run;

data a02;
    x=5;
    do until(conv<0.01);
        dy  = -4*x**3-3*x**2+2*x + 1;
        ddy = -12*x**2-6*x + 2;
        xt=x;
        x=x-dy/ddy;
        conv = abs(xt - x);
        output;
    end;
run;
```

```
*Plot the function and iterations:
proc sql noprint;
 select x into :a separated by ' ' from a02;
quit;

proc sgplot data=a01;
    series x=x y=y;
    xaxis min=-3 max=3;
    refline &a. / axis=x lineattrs=(color=red);
run;
```


Newton-Raphson Algorithm

Example, Results



Homework:

1. What if the starting point is different, e.g. -2? What does this tell about properties of Newton-Raphson algorithm?
2. Write a program to estimate the parameters of logistic regression model with one explanatory variable using Newton-Raphson algorithm.

Fisher scoring numerically

- Fisher scoring method is similar to Newton-Raphson with the exception that the former uses **J** matrix of expected information rather than matrix of observed information **H**. In For *t* representing the iteration of the algorithm the procedure is:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{t} - \left(\mathbf{J}^{(t)}\right)^{-1}\mathbf{u}^{(t)}$$

$$\mathbf{u} = \begin{bmatrix} \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_1} \\ \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_2} \\ \vdots \\ \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_p} \end{bmatrix}, \mathbf{J} = \begin{bmatrix} -E\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_1^2}\right) & -E\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_1 \partial \beta_2}\right) & \cdots & -E\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_1 \partial \beta_p}\right) \\ -E\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_2 \partial \beta_1}\right) & -E\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_2^2}\right) & \cdots & -E\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_2 \partial \beta_p}\right) \\ \vdots & \vdots & \ddots & \vdots \\ -E\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_p \partial \beta_1}\right) & -E\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_p \partial \beta_2}\right) & \cdots & -E\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_p^2}\right) \end{bmatrix}$$

For large samples **J** and **H** are approximately equal. For details on expected information matrix see: Agresti, 2002, p. 135-139.

Plotting log likelihood (1)

Model (0)

$$Y = \begin{cases} 1 & default \\ 0 & non-default \end{cases}$$

$$X = \begin{cases} 1 & having\ telephone \\ 0 & not\ having\ telephone \end{cases}$$

$$\pi(x_i) = \frac{1}{1 + e^{-(\alpha + \beta x_i)}}$$

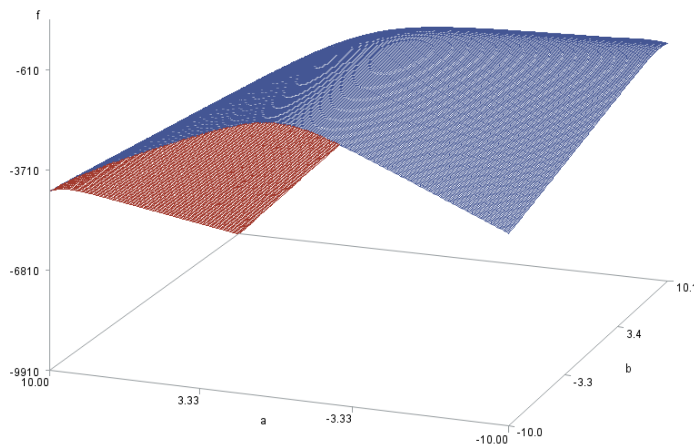
Log-likelihood for model (0):

$$\ln L(y_1, y_2, ..., y_n \mid \alpha, \beta) =$$
$$= \sum_{i=1}^{n} y_i[\alpha + \beta x_i] - \sum_{i=1}^{n} \ln[1 + \exp(\alpha + \beta x_i)]$$
$$= \alpha \sum_{i=1}^{n_{y=1}} y_i + \beta \sum_{i=1}^{n_{y=1;x=1}} y_i x_i - \sum_{i=1}^{n_{x=1}} \ln[1 + \exp(\alpha + \beta)] - \sum_{i=1}^{n_{x=0}} \ln[1 + \exp(\alpha)]$$
$$= n_{y=1}\alpha + n_{y=1;x=1}\beta - n_{x=1}\ln[1 + \exp(\alpha + \beta)] - n_{x=0}\ln[1 + \exp(\alpha)]$$

Plotting log likelihood for model (0)

plotting_lnL.sas

$$\ln L(y_1, y_2, \ldots, y_n \mid \alpha, \beta) =$$
$$= n_{y=1}\alpha + n_{y=1;x=1}\beta - n_{x=1}\ln\left[1+\exp(\alpha+\beta)\right] - n_{x=0}\ln\left[1+\exp(\alpha)\right]$$
$$= 300\alpha + 113\beta - 404\ln[1+\exp(\alpha+\beta)] - 596\ln\left[1+\exp(\alpha)\right]$$



	y	x	COUNT
1	0	0	409
2	0	1	291
3	1	0	187
4	1	1	113

```
goptions device=activex;
proc g3d data=p01;
   plot a*b=f;
run;
quit;
```


Model (1)

- Response *Y* (Customer status) and explanatory *X* (Housing status) variables are:

$$Y = \begin{cases} 1 & default \\ 0 & non-default \end{cases}$$

$$X = \begin{cases} 1 & Own \\ 2 & Rent \\ 3 & For\ Free \end{cases}$$

- Estimate the model assessing the risk of customer being default given their housing status:

$$\pi(x_i) = \frac{1}{1 + e^{-(\alpha + \beta_1 x_{1i} + \beta_2 x_{2i})}}$$

Estimation of logistic regression in SAS

Code302.sas

Choosing estimating algorithm

- The option to choose the estimating algorithm is in model statement:

```
technique=Fisher | Newton
```

- Fisher scoring is default

```
*Binary logistic regression model;
proc logistic data=class3.german;
    class housing (param=ref ref='own');
    model default(event='1')= housing / technique=Newton;
run;
```

Model Information	
Data Set	CLASS3.GERMAN
Response Variable	default
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Model Information	
Data Set	CLASS3.GERMAN
Response Variable	default
Number of Response Levels	2
Model	binary logit
Optimization Technique	Newton-Raphson

Alternative estimating procedure for LR model

```
proc genmod data=class3.german descending;
 class housing (param=ref ref='own');
 model default= housing / dist = bin link = logit;
run;
```

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.0415	0.0853	-1.2086	-0.8743	149.11	<.0001
housing	for free	1	0.6668	0.2136	0.2481	1.0854	9.74	0.0018
housing	rent	1	0.5986	0.1753	0.2550	0.9422	11.66	0.0006
Scale		0	1.0000	0.0000	1.0000	1.0000		

Parametrization of the logistic model

Parametrization (see SAS Online Doc for other types of parametrizations):

- Reference - Parameter estimates of CLASS main effects, using the reference coding scheme, estimate the difference in the effect of each nonreference level compared to the effect of the reference level.
- Effect - Parameter estimates of CLASS main effects, using the effect coding scheme, estimate the difference in the effect of each nonreference level compared to the average effect over all levels.

Class Level Information			
Class	Value	Design Variables	
housing	for free	1	0
	own	0	0
	rent	0	1

Class Level Information			
Class	Value	Design Variables	
housing	for free	1	0
	own	-1	-1
	rent	0	1

Logistic Regression with SAS ©AK, SGH 2019

Reference vs. effect parametrization

- In model assessing relationship between a customer being default and their housing status the model have three structural parameters, intercept and effects related to the housing status. The individual probabilities are calculated using:

$$\pi(x_i) = \frac{1}{1 + e^{-(\alpha + \beta_1 x_{1i} + \beta_2 x_{2i})}}$$

Reference coding

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.0414	0.0853	149.1094	<.0001
housing	for free	1	0.6669	0.2136	9.7473	0.0018
housing	rent	1	0.5987	0.1753	11.6620	0.0006

$$\pi(x_i) = \frac{1}{1 + e^{-(-1.0414 + 0.6669\ x_{1i} + 0.5987\ x_{2i})}}$$

Effect coding

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.6196	0.0876	50.0112	<.0001
housing	for free	1	0.2450	0.1430	2.9344	0.0867
housing	rent	1	0.1768	0.1245	2.0178	0.1555

$$\pi(x_i) = \frac{1}{1 + e^{-(-0.6195 + 0.2450\ x_{1i} + 0.1768\ x_{2i})}}$$

Reference parametrization

$$\pi(x_i) = \frac{1}{1+e^{1.0414 - 0.6669 x_{1i} - 0.5987 x_{2i}}}$$

- Calculating risk of being default comparing those renting with customers owning flats [model (1)]:

$$\pi(x_i = rent) = \frac{1}{1+e^{1.0414 - 0.5987}} \approx 0.3911$$

$$\pi(x_i = own) = \frac{1}{1+e^{1.0414}} \approx 0.2610$$

Class Level Information			
Class	Value	Design Variables	
housing	for free	1	0
	own	0	0
	rent	0	1

$$O(x_i = rent) = \frac{\pi(x_i)}{1-\pi(x_i)} \approx 0.6423$$

$$O(x_i = own) = \frac{\pi(x_i)}{1-\pi(x_i)} \approx 0.3530$$

$$OR(rent \;\; vs \;\; own) = \frac{O(x_i = rent)}{O(x_i = own)} \approx 1.8198$$

Effect parametrization

$$\pi(x_i) = \frac{1}{1+e^{0.6195 - 0.2450 x_{1i} - 0.1768 x_{2i}}}$$

- Calculating risk of being default comparing those renting with customers owning flats [model (1)]:

$$\pi(x_i = rent) = \frac{1}{1+e^{0.6195 - 0.1768}} \approx 0.3911$$

$$\pi(x_i = own) = \frac{1}{1+e^{0.6195 + 0.2450 + 0.1768}} \approx 0.2609$$

Class Level Information			
Class	Value	Design Variables	
housing	for free	1	0
	own	-1	-1
	rent	0	1

$$O(x_i = rent) = \frac{\pi(x_i)}{1-\pi(x_i)} \approx 0.6423$$

$$O(x_i = own) = \frac{\pi(x_i)}{1-\pi(x_i)} \approx 0.3530$$

$$OR(rent \;\; vs \;\; own) = \frac{O(x_i = rent)}{O(x_i = own)} \approx 1.8196$$

Parametrization programmatically

- Reference:

```
proc logistic data=class3.german;
    class housing (param=ref ref='own');
    model default(event='1')= housing;
run;
```

- Effect:

```
*Binary logistic regression model;
proc logistic data=class3.german;
    class housing (param=effect ref='own');
    model default(event='1')= housing;
run;
```


Selecting variables for the model

- Theoretically sensible
- Frequencies analysis to check the distributions – sparse categories will affect the model estimation
- Selection method (e.g. forward, backward, stepwise)
- Statistical significance of the effects

Variable selection methods in SAS

- SAS OnlineDoc:
- When SELECTION=FORWARD, PROC LOGISTIC first estimates parameters for effects forced into the model. These effects are the intercepts and the first *n* explanatory effects in the MODEL statement, where *n* is the number specified by the START= or INCLUDE= option in the MODEL statement (*n* is zero by default). Next, the procedure computes the score chi-square statistic for each effect not in the model and examines the largest of these statistics. If it is significant at the SLENTRY= level, the corresponding effect is added to the model. Once an effect is entered in the model, it is never removed from the model.
- When SELECTION=BACKWARD, estimate model including all variables. Results of the Wald test for individual parameters are examined. The least significant effect that does not meet the SLSTAY= level for staying in the model is removed. Once an effect is removed from the model, it remains excluded. The process is repeated until no other effect in the model meets the specified level for removal or until the STOP= value is reached.
- The SELECTION=STEPWISE option is similar to the SELECTION=FORWARD option except that effects already in the model do not necessarily remain. Effects are entered into and removed from the model in such a way that each forward selection step can be followed by one or more backward elimination steps. The stepwise selection process terminates if no further effect can be added to the model or if the current model is identical to a previously visited model.

Logistic Regression with SAS ©AK, SGH 2019

Exploring variables through SAS selection method

Code308.sas

- Ex 1. Estimate the model assessing risk of customer being default adding new explanatory variables using three selection methods. Compare the results.

```
*Exploring variables using selection procedure;
proc logistic data=german;
   class housing personal_status debtors telephone / param=ref;
   model default(event='1')= duration credit_amt age housing
          personal_status debtors telephone
      / selection=forward /*Forward | Backward | Stepwise*/;
run;
```

Assessing goodness-of-fit

Code304.sas

Parameter		DF	Estimate
Intercept		1	-1.0414
housing	for free	1	0.6669
housing	rent	1	0.5987

$$\ln L(y_1, y_2, \dots, y_n \mid \theta_1, \dots, \theta_r)$$
$$= \sum_{i=1}^{n} y_i \left[\alpha + \sum_{k=1}^{p} \beta_k x_{ki} \right] - \sum_{i=1}^{n} \ln \left[1 + \exp(\alpha + \sum_{k=1}^{p} \beta_k x_{ki}) \right]$$

In model (1):

$$\ln L(y_1, y_2, \dots, y_n \mid \alpha, \beta_1, \beta_2)$$
$$= \sum_{i=1}^{n} y_i [\alpha + \beta_1 x_{1i} + \beta_2 x_{2i}] - \sum_{i=1}^{n} \ln [1 + \exp(\alpha + \beta_1 x_{1i} + \beta_2 x_{2i})]$$

$$AIC = -2 \ln L + 2p$$

$$SC = -2 \ln L + p \ln n$$

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1223.729	1210.049
SC	1228.636	1224.772
-2 Log L	1221.729	1204.049

p-number of parameters [model (1) has 3 parameters]
n- sample size

Logistic Regression with SAS ©AK, SGH 2019

Testing significance of effects

- Testing if at least one explanatory variable in the model is statistically significant ie. the model allows to identify at least one effect which affects the probability under analysis

$$H_0 : \boldsymbol{\beta} = \mathbf{0}$$
$$H_1 : \bigvee_k \beta_k \neq 0$$

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	17.6797	2	0.0001
Score	18.1998	2	0.0001
Wald	17.9443	2	0.0001

Likelihood ratio test (Agresti, 2002, p. 12)

- The idea is to compare *ln*L in the null model with the model with explanatory variable(s):
  The test statistic is as follows:

$$-2\ln\Lambda = -2\ln\left(\frac{l_0}{l_1}\right) = -2\left(\ln L_0 - \ln L_1\right)$$

- For model (1) we get:

$$-2\ln\Lambda = -2\ln L_0 + 2\ln L_1 = 1221.729 - 1204.049 = 17.68$$

- -2lnΛ has a limiting chi-square distribution with degrees of freedom equal to the difference in the dimension of the parameter spaces between the compared models

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1223.729	1210.049
SC	1228.636	1224.772
-2 Log L	1221.729	1204.049

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	17.6797	2	0.0001
Score	18.1998	2	0.0001
Wald	17.9443	2	0.0001

Wald test (Agresti, 2002, p. 11)

Code305.sas

- The idea is to compare the estimates with their standard errors. The test statistic is as follows:

$$W = \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)^T \left[\mathrm{cov}(\hat{\boldsymbol{\beta}})\right]^{-1} \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)$$

- For model (1) we get:

$$W = \left[\hat{\beta}_1 \quad \hat{\beta}_2\right] \begin{bmatrix} \hat{\sigma}^2_{\hat{\beta}_1} & \hat{\sigma}_{\hat{\beta}_1\hat{\beta}_2} \\ \hat{\sigma}_{\hat{\beta}_1\hat{\beta}_2} & \hat{\sigma}^2_{\hat{\beta}_2} \end{bmatrix}^{-1} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}$$

- W has a limiting chi-square distribution with degrees of freedom equal to the rank of covariance matrix.

Parameter		DF	Estimate
Intercept		1	-1.0414
housing	for free	1	0.6669
housing	rent	1	0.5987

Estimated Covariance Matrix			
Parameter	Intercept	housingfor_free	housingrent
Intercept	0.007274	-0.00727	-0.00727
housingfor_free	-0.00727	0.045625	0.007274
housingrent	-0.00727	0.007274	0.030733

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	17.6797	2	0.0001
Score	18.1998	2	0.0001
Wald	17.9443	2	0.0001

Score test (Agresti, 2002, p. 12-13)

- The idea is to compare the slope and expected curvature of the log-likelihood function at β_0. The test statistic is as follows:

$$S = \mathbf{u}^T \mathbf{J}^{-1} \mathbf{u}$$

- S has approximate chi-square distribution with *r* degrees of freedom where *r* is the number of restrictions imposed on $\boldsymbol{\beta}$. See details in OnlineDoc.

Score test (2)

Under H_0 the assumption is that regression coefficients are equal to 0 and therefore the logit equals:

$$\operatorname{logit}\left[\pi(x_i)\right] = \ln \frac{\pi(x_i)}{1-\pi(x_i)} = \alpha \qquad \pi(x_i) = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Estimate of P(Y=1)

$$H_0 : \left[\alpha \quad \beta_1 \quad \dots \quad \beta_p\right]^T = \left[\ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) \quad 0 \quad \dots \quad 0\right]^T$$

$$H_1 : \underset{k}{\vee}\, \beta_k \neq 0$$

The test statistic is:

$$S = \mathbf{u}^T \mathbf{J}^{-1} \mathbf{u}$$

$$\mathbf{u} = \begin{bmatrix} \dfrac{\partial L(\boldsymbol{\beta})}{\partial \beta_1} \\ \dfrac{\partial L(\boldsymbol{\beta})}{\partial \beta_1} \\ \vdots \\ \dfrac{\partial L(\boldsymbol{\beta})}{\partial \beta_p} \end{bmatrix}, \mathbf{J} = \begin{bmatrix} -E\left(\dfrac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_1^2}\right) & -E\left(\dfrac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_1 \partial \beta_2}\right) & \cdots & -E\left(\dfrac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_1 \partial \beta_p}\right) \\ -E\left(\dfrac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_2 \partial \beta_1}\right) & -E\left(\dfrac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_2^2}\right) & \cdots & -E\left(\dfrac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_2 \partial \beta_p}\right) \\ \vdots & \vdots & \ddots & \vdots \\ -E\left(\dfrac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_p \partial \beta_1}\right) & -E\left(\dfrac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_p \partial \beta_2}\right) & \cdots & -E\left(\dfrac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_p^2}\right) \end{bmatrix}$$

Score test example

In model (0) we get:

$$\ln L(y_1, y_2, \ldots, y_n \mid \alpha, \beta) = n_{y=1}\alpha + n_{y=1;x=1}\beta - n_{x=1}\ln\left[1+\exp(\alpha+\beta)\right] - n_{x=0}\ln\left[1+\exp(\alpha)\right]$$

$$\mathbf{u} = \begin{bmatrix} \dfrac{\partial L(\boldsymbol{\beta})}{\partial \alpha} \\ \dfrac{\partial L(\boldsymbol{\beta})}{\partial \beta} \end{bmatrix} = \begin{bmatrix} n_{y=1} - n_{x=1}\dfrac{1}{1+\exp(\alpha+\beta)} \\ n_{y=1;x=1} - n_{x=1}\dfrac{\exp(\alpha+\beta)}{\left[1+\exp(\alpha+\beta)\right]^2} \end{bmatrix}, \quad \mathbf{J} = \begin{bmatrix} -E\left(\dfrac{\partial^2 L(\boldsymbol{\beta})}{\partial \alpha^2}\right) & -E\left(\dfrac{\partial^2 L(\boldsymbol{\beta})}{\partial \alpha \partial \beta}\right) \\ -E\left(\dfrac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta \partial \alpha}\right) & -E\left(\dfrac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta^2}\right) \end{bmatrix}$$

$$\mathbf{J} = \begin{bmatrix} E\left(-\sum_{i=1}^{n}\dfrac{e^{\alpha+\beta x_i}}{\left(1+e^{\alpha+\beta x_i}\right)^2}\right) & E\left(-\sum_{i=1}^{n}\dfrac{x_i e^{\alpha+\beta x_i}}{\left(1+e^{\alpha+\beta x_i}\right)^2}\right) \\ E\left(-\sum_{i=1}^{n}\dfrac{x_i e^{\alpha+\beta x_i}}{\left(1+e^{\alpha+\beta x_i}\right)^2}\right) & E\left(-\sum_{i=1}^{n}\dfrac{x_i^2 e^{\alpha+\beta x_i}}{\left(1+e^{\alpha+\beta x_i}\right)^2}\right) \end{bmatrix} = \begin{bmatrix} -n_{x=1}\dfrac{e^{\alpha+\beta}}{\left(1+e^{\alpha+\beta}\right)^2} - n_{x=0}\dfrac{e^{\alpha}}{\left(1+e^{\alpha}\right)^2} & -n_{x=1}\dfrac{e^{\alpha+\beta}}{\left(1+e^{\alpha+\beta}\right)^2} \\ -n_{x=1}\dfrac{e^{\alpha+\beta}}{\left(1+e^{\alpha+\beta}\right)^2} & -n_{x=1}\dfrac{e^{\alpha+\beta}}{\left(1+e^{\alpha+\beta}\right)^2} \end{bmatrix}$$

where $n_{y=1} = 300,\ n_{y=0} = 700,\ n_{x=0} = 596,\ n_{x=1} = 404,\ n_{x=1;y=1} = 113$

Under $H_0: \alpha = \ln\left(\dfrac{0.3}{0.7}\right) = -0.8473,\ \beta = 0$

$S = \mathbf{u}^T \mathbf{J}^{-1}\mathbf{u} = 1.329784$

	y	x	COUNT
1	0	0	409
2	0	1	291
3	1	0	187
4	1	1	113

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1.3359	1	0.2478
Score	$\boxed{1.3298}$	1	0.2488
Wald	1.3286	1	0.2491

Wald test for specific variable

Estimated Covariance Matrix

Parameter	Intercept	housingfor_free	housingrent	age
Intercept	0.063276	0.006936	-0.01486	-0.00163
housingfor_free	0.006936	0.049934	0.00542	-0.00042
housingrent	-0.01486	0.00542	0.032048	0.000219
age	-0.00163	-0.00042	0.000219	$\boxed{0.000047}$

- For model (1) with additional effect of age included we get:

$se(\beta_3) = \hat{\sigma}_{\beta_3}$

$$W = \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)^T \left[\mathrm{cov}(\hat{\boldsymbol{\beta}})\right]^{-1}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)$$

$$= \frac{\hat{\beta}_3^2}{\hat{\sigma}_{\beta_3}^2} = \frac{-0.02130406\quad 7}{0.00004728\quad 34} \approx 9.5988$$

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
housing	2	19.5400	<.0001
age	1	$\boxed{9.5988}$	$\boxed{0.0019}$

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.2959	0.2515	1.3841	0.2394
housing	for free	1	0.8480	0.2235	14.4022	0.0001
housing	rent	1	0.4941	0.1790	7.6173	0.0058
age		1	-0.0213	$\boxed{0.00688}$	9.5988	0.019

- W has a limiting chi-square distribution with degrees of freedom equal to the rank of covariance matrix.

Measures of Goodness-of-Fit,

- Assessing fit of the model - we would like to know whether the probabilities produced by the model accurately reflect the true outcome experience in the data (Hosmer et al., 2007).
- Deviance and Pearson χ^2: the statistics compare **saturated model** with the estimated model. Saturated model reflects ideal fit – probabilities from the model equal to the estimated probabilities for each **profile** (k=1,2,…,n_{lk})

n_{11}	n_{12}	…	n_{1k}
n_{21}	n_{22}	…	n_{2k}
…	…	…	…
n_{l1}	n_{l2}	…	n_{lk}

Saturated Model | Estimated Model

$$\frac{n_{ij}}{n} \quad vs. \quad \hat{p}_{ij} = \frac{\hat{n}_{ij}}{n}$$

Logistic Regression with SAS ©AK, SGH 2019

Saturated model

Code306.sas

- Let us consider model (1) with TELEPHONE being included as explanatory variable

```
proc freq data=german;
   tables housing*telephone /
      out=f01;
run;

  proc logistic data=german;
     class housing (param=ref ref='own')
  telephone (param=ref ref='yes');
     model default(event='1')= housing
  telephone / aggregate scale=none  ;
     output out=out predprobs=(i);
  run;
```

Counts (n_{ij}) in saturated model

	default	telephone	housing	COUNT	PERCENT
1	0	no	rent	66	6.6
2	0	no	own	319	31.9
3	0	no	for free	24	2.4
4	0	yes	rent	43	4.3
5	0	yes	own	208	20.8
6	0	yes	for free	40	4
7	1	no	rent	50	5
8	1	no	own	114	11.4
9	1	no	for free	23	2.3
10	1	yes	rent	20	2
11	1	yes	own	72	7.2
12	1	yes	for free	21	2.1

Probabilities in saturated model

Deviance and Pearson χ^2

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	$\boxed{2.7211}$	2	1.3605	0.2565
Pearson	$\boxed{2.7166}$	2	1.3583	0.2571

- Deviance:

$$D = 2\sum_{i=1}^{l}\sum_{j=1}^{k} n_{ij} ln\left(\frac{n_{ij}}{\hat{n}_{ij}}\right)$$

- Pearson χ^2:

$$\chi^2 = \sum_{i=1}^{l}\sum_{j=1}^{k}\frac{\left(n_{ij}-\hat{n}_{ij}\right)^2}{\hat{n}_{ij}}$$

n_{ij} - observed counts
$\hat{n}_{ij}$ – counts estimated in the model E(X)=nπ (by property of binomial distribution)
D and χ^2 follow chi-square distribution with degrees of freedom equal to the difference between the number of profiles (parameters estimated in the saturated model) and estimated parameters.

Example

- Test goodness of fit of model (1) with TELEPHONE being included as explanatory variable using Deviance and Pearson χ^2. Calculate Deviance and Pearson χ^2 statistics.

										Expected	Observed		D	Chi-square
							0	no	rent	68.73725	66	6.6	-2.68201	0.1090029
Count		IP_0	IP_1				0	no	own	313.67	319	31.9	5.37501	0.09056869
116	11.6	0.592563	0.407437	68.73725	47.26275		0	no	for free	26.58863	24	2.4	-2.45832	0.25202516
433	43.3	0.724411	0.275589	313.67	119.33		0	yes	rent	40.25858	43	4.3	2.832714	0.18667833
47	4.7	0.565716	0.434284	26.58863	20.41137		0	yes	own	213.3264	208	20.8	-5.25938	0.13299282
63	6.3	0.639025	0.360975	40.25858	22.74142		0	yes	for free	37.40772	40	4	2.680094	0.1796398
280	28	0.76188	0.23812	213.3264	66.67357		1	no	rent	47.26275	50	5	2.815032	0.15852994
61	6.1	0.613241	0.386759	37.40772	23.59228		1	no	own	119.33	114	11.4	-5.20913	0.23806829
							1	no	for free	20.41137	23	2.3	2.746248	0.32829759
							1	yes	rent	22.74142	20	2	-2.56912	0.33047201
							1	yes	for free	66.67357	72	7.2	5.533743	0.42551922
							1	yes	for	23.59228	21	2.1	-2.44434	0.28483535
													2.721091	2.7166301

Deviance and Pearson χ^2 practical aspects

- If logistic regression model satisfies either of the following conditions:
  - includes many explanatory variables
  - explanatory variables have considerable number of categories
  - model includes continuous variables

Deviance and Pearson χ^2 do not follow χ^2 distribution

Practically contingency table includes sparse categories – small or zero counts. A test which allows to measure the goodness of fit in case of a model satisfying either of above conditions is **Hosmer-Lemeshow** (HL) test.

Hosmer-Lemeshow test

- Hosmer Lemeshow statistic is calculated as follows:
- Estimate the individual probabilities.
- Sort within response and divide set into 10 groups with similar number of observations (k=1,2,3,…,10)
- Using the groups calculate expected counts and compare with observed counts using Pearson χ^2 statistic, which has approximately χ^2 distribution with v=10-2=8 degrees of freedom.

```
proc logistic data=german;
   class housing (param=ref ref='own') telephone (param=ref ref='yes');
   model default(event='1')= housing telephone / aggregate scale=none
          lackfit;
   output out=out predprobs=(i);
run;
```

Example

Code307.sas

- Test goodness of fit of model (1) with AGE being included as explanatory variable using Hosmer-Lemeshow statistic.

[SAS OnlineDoc]

Partition for the Hosmer and Lemeshow Test

Group	Total	default = 1 Observed	default = 1 Expected	default = 0 Observed	default = 0 Expected
1	99	18	18.21	81	80.79
2	108	26	24.69	82	83.31
3	89	17	22.40	72	66.60
4	88	22	23.30	66	64.70
5	94	24	26.13	70	67.87
6	105	29	30.60	76	74.40
7	102	36	30.93	66	71.07
8	100	38	32.87	62	67.13
9	97	44	38.47	53	58.53
10	118	46	52.41	72	65.59

0.002428	0.000547
0.069971	0.020732
1.302721	0.43821
0.072579	0.026138
0.173811	0.066921
0.083389	0.034293
0.831501	0.361852
0.801296	0.392317
0.796441	0.523366
0.784196	0.626636
	7.409344

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
7.4093	8	0.4932