Introduction Logistic regression

Smart Analytics for Big Data

Iragaël Joly
Grenoble INP - UMR GAEL

2020-11-09

iragael.joly@grenoble-inp.fr

"2020-11-09"

Overview

Motivations

Regression Modelling

R and Tests for Categorical Data

Binary Logistic Regression

Logistic Regression Model

Logit Model Interpretation

Logit Model Estimation

Motivations

Motivations

Human dimension

- Demand concept in engineering, business, marketing, planning, policy making
- Operational in many fields (transportation, marketing, energy, finance, etc.)
- Choice of product, brand, mode, destination, contract type and usage, buy and sell, etc.
- Need for behavioral theories, quantitative methods, operational mathematical models

Typical choice questions

McFadden (2014) merges the three elements of consumption choices :

Three choice questions

"Which" : Product choice : option choice
mode choice, itinary choice, etc.

"How many" : Product quantity
Number of trips, motorisation rate, travelled distance
travel time

"When" : Moment or length of a consumption
Car replacement, departure time, duration between events

Statistical and Econometric Models

- How much & When : Duration model (survival model)
- How many : Count models
- Which : Discrete Choice Models

Four broad frameworks in categorical variables analysis

Binary Choice : A choice between a pair of options

- taking or not an action
- decision be between two distinctly different choices, (public or private transportation).
- the 0/1 outcome is a label for "no/yes"

Multinomial Choice : A choice among more than two choices

- the observed response is simply a label (a brand, a place, the travel mode)
- numerical assignments are not meaningful

Four broad frameworks in categorical variables analysis

Can be extended to

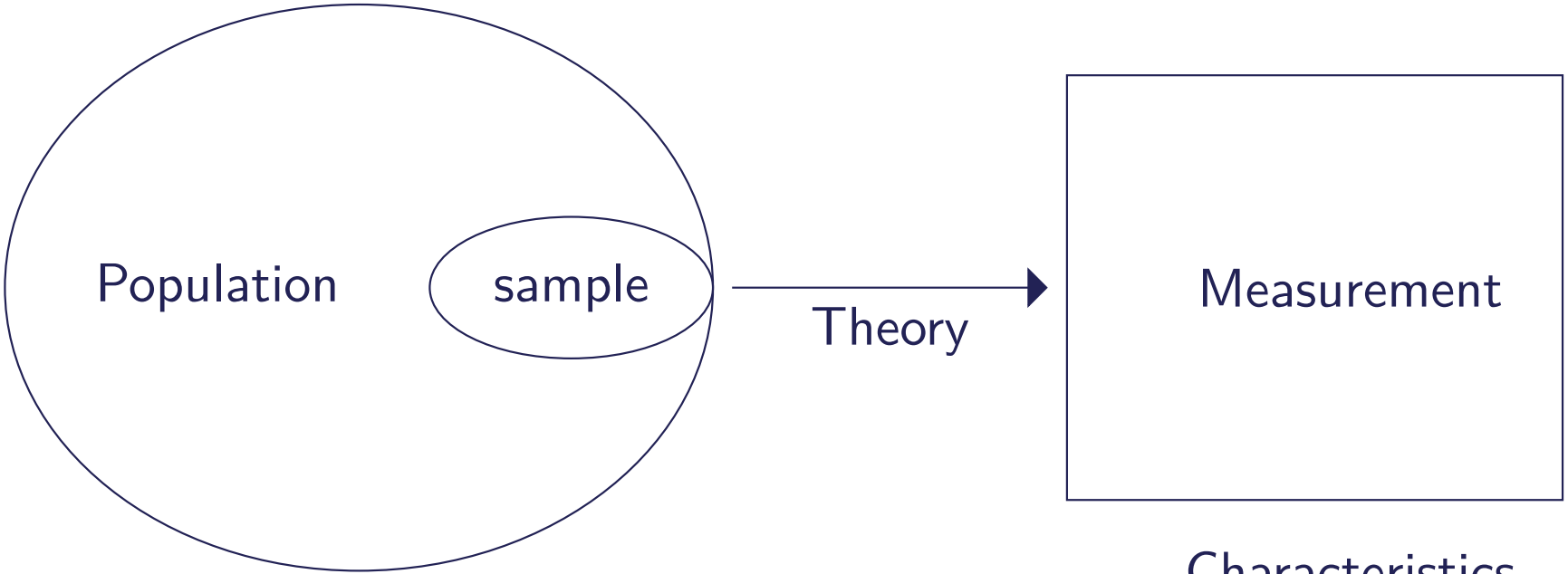
Ordered Choice : The individual reveals the strength of his or her preferences with respect to a single outcome.

- Familiar cases : strength of feelings about a commodity or self-assessed well-being.
- Opinions are given meaningful numeric values, usually 0, 1, . . . , J for some upper limit, J .
- numerical values are only a ranking, not a quantitative measure
- Thus a "1" is greater than a "0" in a qualitative sense, but not by one unit,
- The difference between a "2" and a "1" is not the

Event Counts : The observed outcome is a count of the number of occurrences.

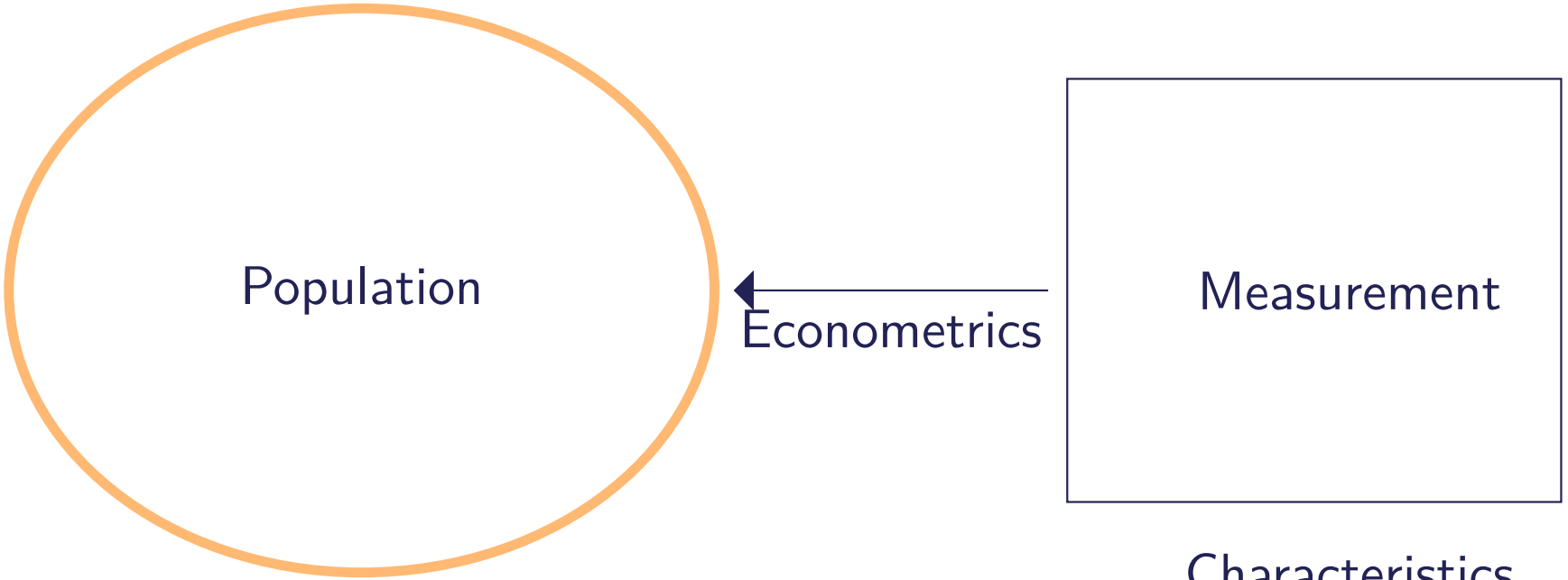
- Similar to the preceding three settings : "dependent variable" measures an individual choice (number of visits...)
- The event count might be the outcome of some natural process, such as incidence of a disease in a population (or the number of defects per unit of time in a production process).
- The models will still be constructed to accommodate the discrete nature of the observed response variable.

Sample and Measurement



Characteristics
Behavior
Patterns
Choices

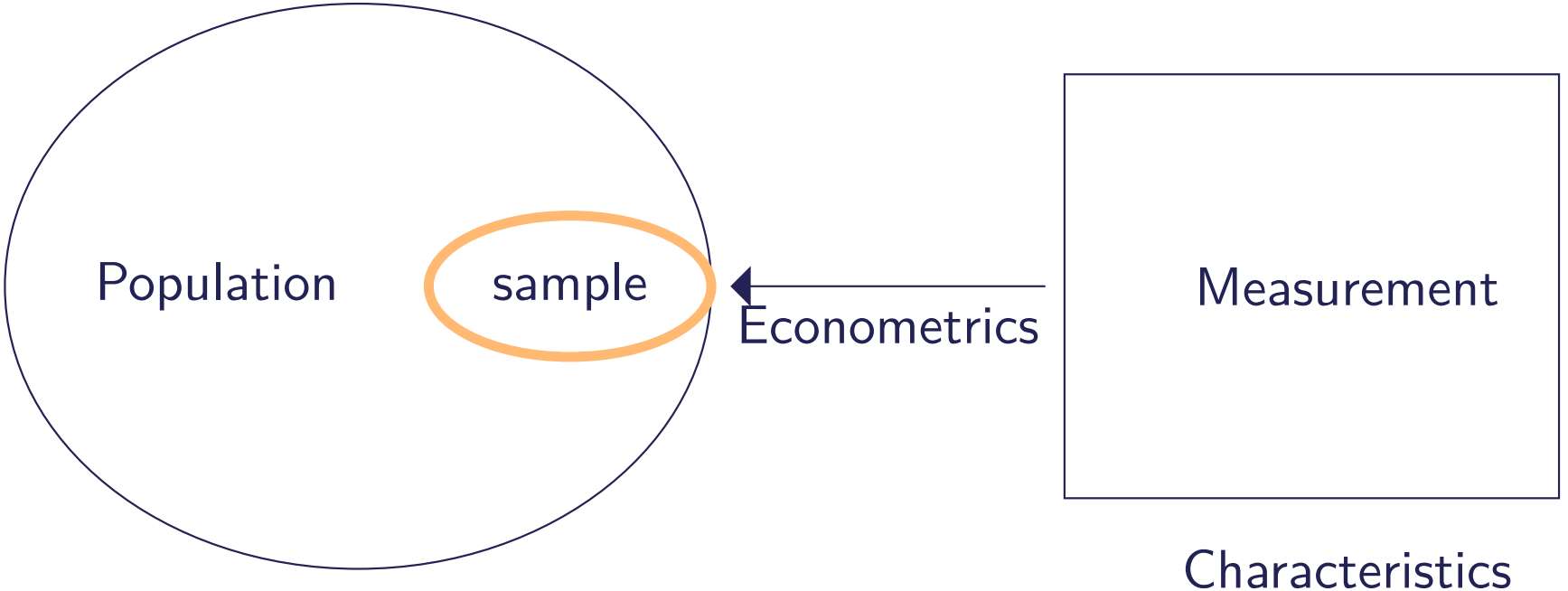
Classical Inference



Imprecise inference
sampling theory
and asymptotics

Characteristics
Behavior
Patterns
Choices

Bayesian Inference



'exact' inference
about only the sample
the 'posterior' density

Characteristics
Behavior
Patterns
Choices

Issues in Model Building and Regression Basics

- Model : the conditional mean function
  - modeling the mean or modeling the probabilities for DCM
- Estimation
  - Coefficients
  - Interesting Partial Effects
- Functional Form and Specification
- Statistical Inference
- Prediction
  - Individuals or Aggregates
- Model Assessment and Evaluation

Regression Modelling

Hypothesis Testing, Estimation, and Prediction

Statistics comprises :

- *Hypothesis testing*
- *Statistical estimation*
- *Prediction*

Hypothesis testing

- **Null hypothesis** is the absence of some effect
- Hypothesis testing can be done **without** regression model
- even with **non parametric** test and p-values
- but
  - easily be done within the context of a statistical model
  - non parametric test produces hardly magnitudes of effects
  - incorporate complexities (cluster sampling ; repeated measurements)
  - with a model : carry out many different statistical tests

"the dataset was too small to allow modeling, so we just did hypothesis tests"

- It is not widely recognized that multivariable modeling is extremely **valuable* even in well-designed randomized experiments
- But to be able to estimate absolute effects one must develop a **multivariable** model of the response variable.

Statistical estimation is usually model-based

But

- Inaccurate estimates can result from *incorrectly assumed linearity* assumption (not on ly in linera reg. model)
- Too *many "confounding" variables* leads to "over-adjusted" estimates (*spurious* associations)
- Reasonable multivariable predictive model : hypothesis testing and estimation of effects are byproducts of the fitted model

Prediction

"Superset of hypothesis testing and estimation"

Prediction vs. Classification

- Classification is **still inferior** to probability modeling
  - for predictive instrument ; for estimation ; hypothesis testing
- Better to use the full information in the data
  - to develop a *probability model*,
  - then develop *classification rules* based on estimated probabilities
- At the least, this forces the analyst to use a proper accuracy score
- In many cases, best option : to refuse to make a decision or to obtain more data

A gray zone can be helpful, and predictions include gray zones automatically.

Utility function (or loss or cost function) depends on variables

- that are not predictive of outcome
- are not collected (e.g., subjects' preferences)
- that are available only at the decision point

Planning for Modeling

"Will this model actually be used ?"

Models are often developed using a "convenience sample," that is, a dataset that was not collected with such predictions in mind.

- predictor or response variables may not have been collected,
- subjects are not representative of the population
- Key variables are missing in large numbers of subjects.
- Data are not missing at random
- Operational definitions of some of the key variables were never made
- Reliability of measurements is unknown, as measurement errors factors

Predictive model will be more accurate, as well as useful, when data collection is planned prospectively

Many things can go wrong in statistical modeling, including the following.

1. The process generating the data is not stable
2. The model is misspecified with regard to nonlinearities or interactions, or there are predictors missing
3. The model is misspecified in terms of the transformation of the response variable or the model's distributional assumptions
4. The model contains discontinuities (e.g., by categorizing continuous predictors or fitting regression shapes with sudden changes) that can be gamed by users
5. Correlations among subjects are not specified, or the correlation structure is misspecified, resulting in inefficient parameter estimates and overconfident inference
6. The model is overfitted, resulting in predictions that are too extreme or positive associations that are false
7. The user of the model relies on predictions obtained by extrapolating to combinations of predictor values well outside

Emphasizing Continuous Variables

- Some categorical variables are *subjective and hard* to **standardize**,
- On the average they **do not contain the same amount of statistical information** as continuous variables
- **Unwise to categorize naturally continuous** variables during data collection

Choice of the Model

Where do probability models come from ?

- Method by which an underlying statistical model should be chosen by the analyst is not well developed
- Knowledge exists to pre-specify
  - a model (e.g., Weibull or log-normal survival model),
  - a transformation for the response variable,
  - a structure for how predictors appear in the model
- Question : **Does the notion of a true model even exist ?**

Choice of the Model

- Few general guidelines choosing the statistical model :
  - The model must use the data efficiently
    - ex : Proba to live 5 years after diag : Logistic vs survival model
    - ordered appreciation : MNL vs Ordered MNL
  - Choose a model that fits overall structures likely to be present in the data
    - constant vs non constant hazard
  - Choose a model that is robust to problems in the data that are difficult to check
  - Choose a model whose mathematical form is appropriate for the response being modeled
    - minimizing numbers of interactions terms
  - Choose a model that is readily extendable

Choice of the Model

Ameen [1] stated that a good model is

a) satisfactory in performance relative to the stated objective,
b) logically sound,
c) representative,
d) questionable and subject to on-line interrogation,
e) able to accommodate external or expert information and
f) able to convey information."

Many authors point :

> *Severe problems that result from treating an empirically derived model as if it were pre-specified and as if it were the correct model.*

1. cited in "C. Chatfield. Model uncertainty, data mining and statistical inference (with discussion). J Roy Stat Soc A, 158 :419–466, 1995."

R and Tests for Categorical Data

Proportion Tests - Motivation

- Theory of test and application with R
- Statistical inference and indicators of interest : confidence interval, p-value
- Proportion and risk and odd ratio

Inference for a (Single) Proportion

Proportion Test

- $H_0 : \pi = 0.5$ is tested against the two-sided alternative $H_1 : \pi \neq 0.5$
- a 95 % confidence interval for π is calculated
- both the test and the CI incorporate a continuity correction.

Example

Half of the students have experienced inferencial test with `R` ?

Our sample : $n = 20$, `R` users :$n_1 = 8$

```
prop.test(8,20,p=.5,alternative="two.sided",
    conf.level=0.95,correct=TRUE)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  8 out of 20, null probability 0.5
## X-squared = 0.45, df = 1, p-value = 0.5023
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.1997709 0.6358833
## sample estimates:
##   p
## 0.4
```

One-way Proportion Test - Exercise

Explain this new test and conclude ?

```
prop.test(8,20,p=.4,alternative="greater",
          conf.level=0.99,correct=FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  8 out of 20, null probability 0.4
## X-squared = 0, df = 1, p-value = 0.5
## alternative hypothesis: true p is greater than 0.4
## 99 percent confidence interval:
##  0.194216 1.000000
## sample estimates:
##   p
## 0.4
```

Two-Way Contingency Tables

The table as Matrix

```
Ruser <- matrix(c(435,147,375,134),nrow=2,byrow=TRUE)
dimnames(Ruser) <- list(c("Female","Male"),c("Yes","No"))
names(dimnames(Ruser)) <- c("Gender","R.User")
```

```
round(Ruser / sum(Ruser), 3)
rowtot <- apply(Ruser,1,sum)
coltot <- apply(Ruser,2,sum)
sweep(Ruser,1,rowtot,"/") #apply "/" to Ruser by row
sweep(Ruser,2,coltot,"/")
```

```
##          R.User
## Gender      Yes    No
##   Female 0.399 0.135
##   Male   0.344 0.123
```

Two-Way Contingency Tables

```
## proportion by row - Gender

##         R.User
## Gender         Yes        No
##   Female 0.7474227 0.2525773
##   Male   0.7367387 0.2632613

## proportion by column - R-Use

##         R.User
## Gender        Yes        No
##   Female 0.537037 0.5231317
##   Male   0.462963 0.4768683
```

Independance test

```
prop.test(Ruser)
```

```
##
##  2-sample test for equality of proportions with continuity correctio
##
## data:  Ruser
## X-squared = 0.11103, df = 1, p-value = 0.739
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.04321353  0.06458148
## sample estimates:
##    prop 1    prop 2
## 0.7474227 0.7367387
```

Comparing Proportions

Aim

Identify *significant* factor effect on a categorical variable (or discrete choice)

Let take an example and define parameters for comparing groups difference, relative risk, odds ratio

	Disease	Status	
Risk Factor	Yes (1)	No (0)	Total
Yes (1)	50	20	70
No (0)	100	130	230
Total	150	150	300

- p_1 : probability of success (disease) in row 1
- $1 - p_1$: probability of failure (no disease) in row 1
- p_2 : probability of success (disease) in row 2
- $1 - p_2$: probability of failure (no disease) in row 2

Difference of proportions

The difference of proportions of successes : $p_1 - p_2$

(basic comparison of the two rows)

Comparison on failures is equivalent : $(1 - p_1) - (1 - p_2) = p_1 - p_2$

- between -1 and $+1$
- equals zero when identical conditional distribution : Y is independant
- may have greater importance when both p_i are close to 0 or 1
- difference between 0.010 and 0.001 is more noteworthy than the difference between 0.410 and 0.401

Relative Risk

The relative risk is the ratio of probabilities

$$RR = \frac{p_1}{p_2}$$

- nonnegative real number
- A relative risk of 1.0 corresponds to independence
- $(1 - p_1)/(1 - p_2)$ gives a different relative risk

Odds Ratio

Let define the odds

The odds of getting disease for the people who were exposed to the risk factor (Risk factor = Yes (1)) :

	Disease	Status	
Risk Factor	Yes (1)	No (0)	Total
Yes (1)	50	20	70
No (0)	100	130	230
Total	150	150	300

p_1 : probability of success (disease) in row 1

$1 - p_1$: probability of failure (no disease) in row 1

$$O_X = \frac{p_1}{1 - p_1} \sim \frac{\hat{p_1}}{1 - \hat{p_1}} = \frac{50/70}{20/70} = 2.5$$

Odds Ratio

The odds of getting disease for the people who were not exposed to the risk factor (Risk factor = No (0))

	Disease	Status	
Risk Factor	Yes (1)	No (0)	Total
Yes (1)	50	20	70
No (0)	100	130	230
Total	150	150	300

p_2 : probability of success (disease) in row 2

$1 - p_2$: probability of failure (no disease) in row 2

$$O_{NoX} = \frac{p_2}{1 - p_2} \sim \frac{\hat{p}_2}{1 - \hat{p}_2} = \frac{100/230}{130/230} = 0.77$$

The Odds-Ratio

The Odds Ratio of having disease for people who were exposed to the risk factor versus not exposed

$$\theta = OR = \frac{O_X}{O_{NoX}} \sim \frac{\frac{\hat{p}_1}{1-\hat{p}_1}}{\frac{\hat{p}_2}{1-\hat{p}_2}} = \frac{50 \times 130}{20 \times 100} = 3.25$$

Interpretation : The odds of having disease are 3.25 times higher for those who were exposed to the risk factor than those who were not exposed to the risk factor.

The Odds-Ratio

- If $\theta > 1$, then the odds of success are higher for exposed individuals than for not exposed
- If $\theta < 1$, then the odds of success are lower for exposed individuals than for not exposed
- If $\theta = 1$, then the odds of success are equal for exposed individuals than for not exposed
- For inference : independence corresponds to $log(\theta) = 0$
- The log odds ratio is symmetric about this value

OR, RR, DR ?

Limits of DR :

- DR cannot be constant
- DR depend on the initial risk

Limits of RR :

- RR have limited range of risk
- ex. $RR = 2$ can only apply to risks below .5
- above that point the RR must diminish
- Risk ratios are similar to odds ratios if the risk is small

OR, RR, DR ?

OR is a natural description of an effect in a probability model since an OR can be constant.

Without Risk Factor		With Risk Factor	
Probability p_1	Odds	Odds	Probability p_2
Risk Factor	Yes (1)	No (0)	Total
.2	.25	.5	.33
.5	1	2	.67
.8	4	8	.89
.9	9	18	.95
.98	49	98	.99

Effect of an odds ratio of 2 on various risks

OR, RR, DR ?

- OR : unlimited range $\rightarrow$ can describe an effect over the entire range of risk
- positive OR still yield valid probability
- constant OR describe the effect independently of other covariables

OR, RR, DR ?

Let X_1 a binary factor and $A = \{X_2, \ldots, X_p\}$ other factors (independant from X_1)

The estimate of $DR = P(Y = 1|X_1 = 1, A) - P(Y = 1|X_1 = 0, A)$ is :

$$
\begin{aligned}
DR = & \frac{1}{1 + exp\{-(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 X_2 + ...)\}} \\
& - \frac{1}{1 + exp\{-(\hat{\beta}_0 + \hat{\beta}_2 X_2 + ...)\}} \\
= & \frac{1}{1 + \left(\frac{1-\hat{R}}{\hat{R}}\right) exp(-\hat{\beta}_1)} - \hat{R}
\end{aligned}
$$

OR, RR, DR ?

The DR estimate can be plotted against $\hat{R}$ or against levels of A to display absolute risk increase against overall risk.

Fig. 1 depicts the relationship between risk of a subject without the risk factor and the increase in risk for a variety of relative increases (odds ratios). It demonstrates how absolute risk increase is a function of the baseline risk (Harrell (2013))

OR, RR, DR ?

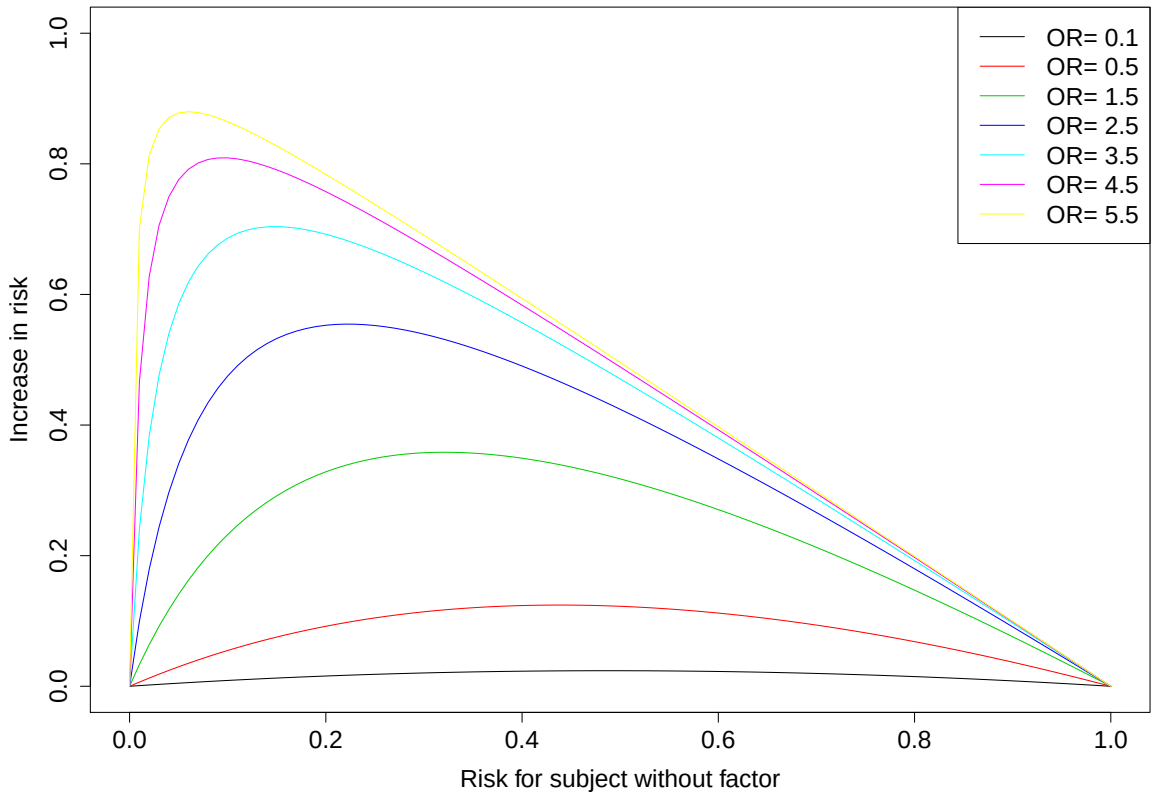


Figure 1 – Absolute benefit as a function of the baseline risk. The odds ratios are given for each curve

The Odds-Ratio with R

Go back on our example of R users

```
Ruser
```

```
##          R.User
## Gender    Yes  No
##   Female 435 147
##   Male   375 134
```

```
R.test <- prop.test(Ruser)
names(R.test)
```

```
## [1] "statistic"   "parameter"   "p.value"     "estimate"    "null.va
## [6] "conf.int"    "alternative" "method"      "data.name"
```

The Odds-Ratio with R

Proportion of "Yes"

```
R.test$estimate
```

```
##    prop 1    prop 2
## 0.7474227 0.7367387
```

Odds of "Yes" over "No"

```
odds <- R.test$estimate/(1-R.test$estimate)
names(odds) <- c("Odds1: F","Odds2: M")  ; odds
```

```
## Odds1: F Odds2: M
## 2.959184 2.798507
```

Odds are almost equivalent

The Odds-Ratio with R

Odds ratio

```
## Odds of "F" over "M"
OR <- odds[1]/odds[2]
names(OR) <- c("OR")  ; OR
```

```
##       OR
## 1.057415
```

The odds of *being R-user* are "greater" by 6% for exposed individuals (women) than for not exposed (men)

Confidence Interval for Odds Ratio

For large sample, the log of odds ratio, $ln(\hat{\theta})$, follows asymptotically a normal distribution.

The $(1-\alpha)100$ % confidence interval estimate for the Log Odds Ratio is

$$ln(\hat{\theta}) \pm z_{\alpha/2}\hat{s}$$

The $(1-\alpha)100$ % confidence interval estimate for the Odds Ratio is

$$\left(e^{ln(\hat{\theta})-z_{\alpha/2}\hat{s}}, e^{ln(\hat{\theta})+z_{\alpha/2}\hat{s}}\right)$$

where $\hat{s} = \sqrt{\sum(1/n_{ij})}$ (from Delta Method)

and $z_{\alpha/2}$ the normal quantile

Confidence Interval for Odds Ratio

CI for Odds Ratio

```
theta <- odds[1]/odds[2]
ASE <- sqrt(sum(1/Ruser))
# ASE
ASE
logtheta.CI <- log(theta) + c(-1,1)*1.96*ASE
# IC log(theta)
logtheta.CI
# IC(OR)
exp(logtheta.CI)
```

```
## ASE:
```

Confidence Interval for Odds Ratio

```
## [1] 0.1386756

## CI of log theta

## [1] -0.2159770  0.3276314

## exp(CI log theta)

## [1] 0.8057539 1.3876774
```

Interpretation ?

The Odds-Ratio

Automatic calculus with function `oddsratio.R`

```
source(file = "../fichiers_aux/data/oddsratio.R")
odds.ratio(Ruser)
```

```
## $estimator
## [1] 1.057415
##
## $ASE
## [1] 0.1386756
##
## $conf.interval
## [1] 0.8057579 1.3876705
##
## $conf.level
## [1] 0.95
```

The Odds-Ratio - Exercice 1

Seat-Belt Use and Traffic Deaths (Agresti (2013))

Fatality results for children under age 18 who were passengers in auto accidents in Florida in 2008 according to whether the child was wearing a seat belt

	Injury	Outcome	
Seat-Belt Use	Fatal	Nonfatal	Total
No	54	10,325	10,379
Yes	25	51,790	51,815

Calculte OR and its CI. Conclude on the benefit of waering the seatbelt ?

Solution

- The sample odds ratio $\hat{\theta} = 10.83$
- Estimated standard error $\hat{s}$ of $log(\hat{\theta}) = 2.383$ is $\hat{s}(log(\hat{\theta})) = 0.242$
- A 95 % confidence interval for $log(\hat{\theta})$ is $2.383 \pm 1.96(0.242)$, or $(1.908, 2.857)$.
- interval fore is $[exp(1.908), exp(2.857)]$ or $(6.74, 17.42)$
- There is a very strong association
- Even though the overall sample size is extremely large, the estimate of the true odds ratio is rather imprecise because of the relatively small number of fatalities

Solution to Agresti exercise

```
SeatBelt <- c("No","No","Yes","Yes")
Accident <- c("Fatal","Nonfatal","Fatal","Nonfatal")
Count <- c(54 , 10325, 25 , 51790)
Accdf <- data.frame(SeatBelt,Accident,Count)
Accdf
```

SeatBelt	Accident	Count
No	Fatal	54
No	Nonfatal	10325
Yes	Fatal	25
Yes	Nonfatal	51790

Solution to Agresti exercise

```
rm(SeatBelt,Accident,Count)
names(dimnames(Accdf)) <- c("SeatBelt","Accident")
tab1 <- tapply(Accdf$Count,list(Accdf$SeatBelt,
Accdf$Accident),c);tab1
```

```
##     Fatal Nonfatal
## No     54    10325
## Yes    25    51790
```

```
R.test <- prop.test(tab1)
odds <- R.test$estimate/(1-R.test$estimate)
names(odds) <- c("Odds1: F","Odds2: NF") ; odds
```

Solution to Agresti exercise

```
##      Odds1: F     Odds2: NF
## 0.0052300242 0.0004827187
```

```
OR <- odds[1]/odds[2]
names(OR) <- c("OR") ; OR
```

```
##        OR
## 10.83452
```

```
theta <- odds[1]/odds[2]
ASE <- sqrt(sum(1/tab1))
# ASE
ASE
```

```
## [1] 0.242146
```

Solution to Agresti exercise

```
logtheta.CI <- log(theta) + c(-1,1)*1.96*ASE
# IC log(theta)
logtheta.CI
```

```
## [1] 1.908131 2.857343
```

```
# IC(OR)
exp(logtheta.CI)
```

```
## [1]  6.740479 17.415198
```

```
source(file = "../fichiers_aux/data/oddsratio.R")
odds.ratio(tab1)
```

Solution to Agresti exercise

```
## $estimator
## [1] 10.83452
##
## $ASE
## [1] 0.242146
##
## $conf.interval
## [1]  6.740538 17.415047
##
## $conf.level
## [1] 0.95
```

Binary Logistic Regression

Introduction

Context

- Assume we observe the output from a binary choice
- We have to analyse a binary variable (0 or 1)
- Are the classic linear regression a suitable tool to analyse binary choice ?

Linear Probability Model

Regression Models with Binary Outcome Variable

Simple Linear Regression Model

- $Y_i = \alpha + \beta X_i + u_i$
- assuming $E(u_i) = 0$
- $E(Y_i|X_i) = \alpha + \beta X_i$
- The hypothetical model is simple linear regression

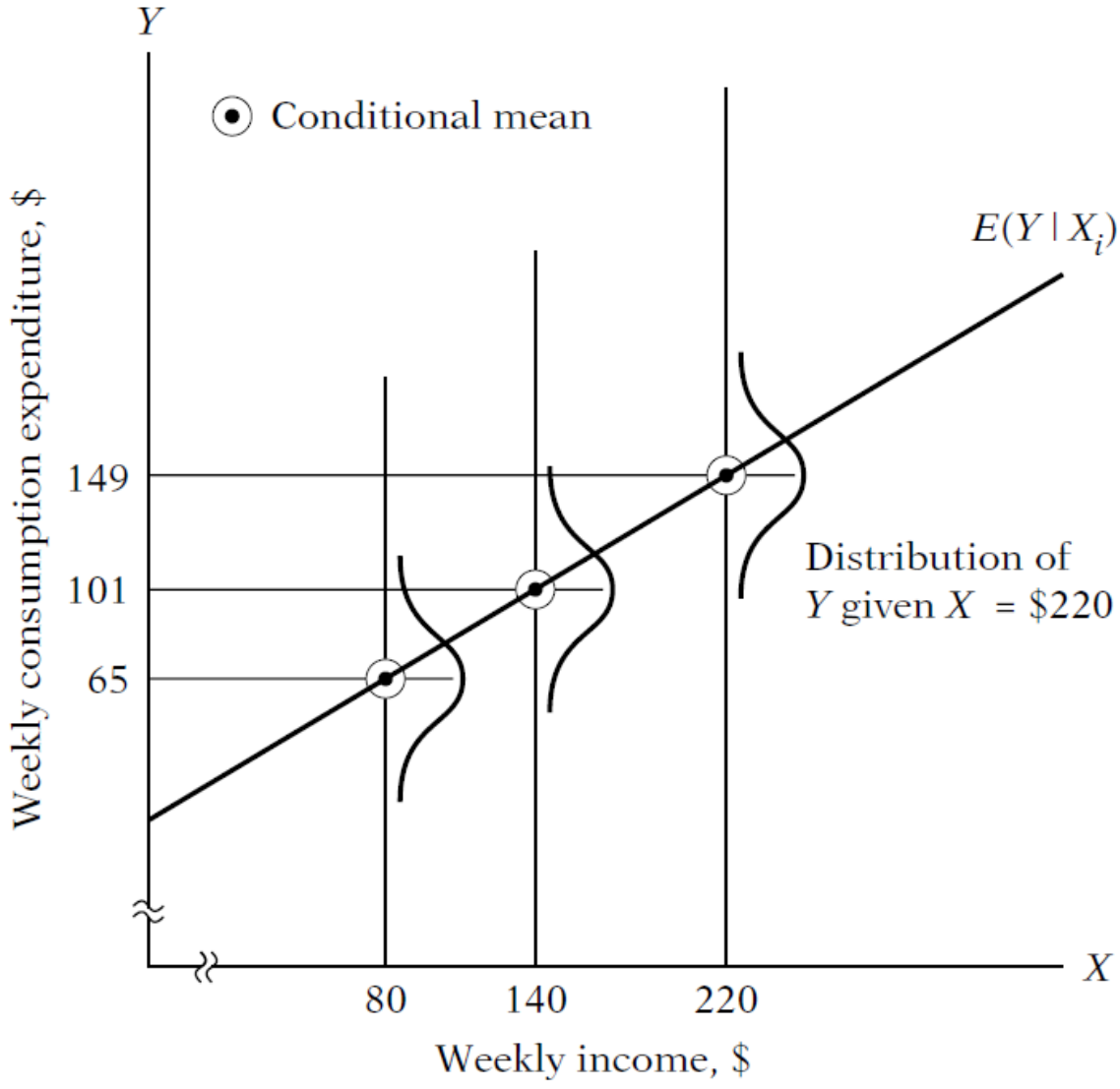


Figure 2 – Simple Linear Regression Gujarati (2003)

Linear Probability Model

Introductory example **Simple Linear Regression Model**

Hypothetical Data on Home Ownership and Income

FAMILY : 40 families

Y : Home Ownership where

1 :Owns a House

0 : Does Not Own a House

X :Family Income, Thousands of $

1. **Estimate** Y over X
2. **Interpret** this regression (intercept, slope, predicted probabilities, residuals)

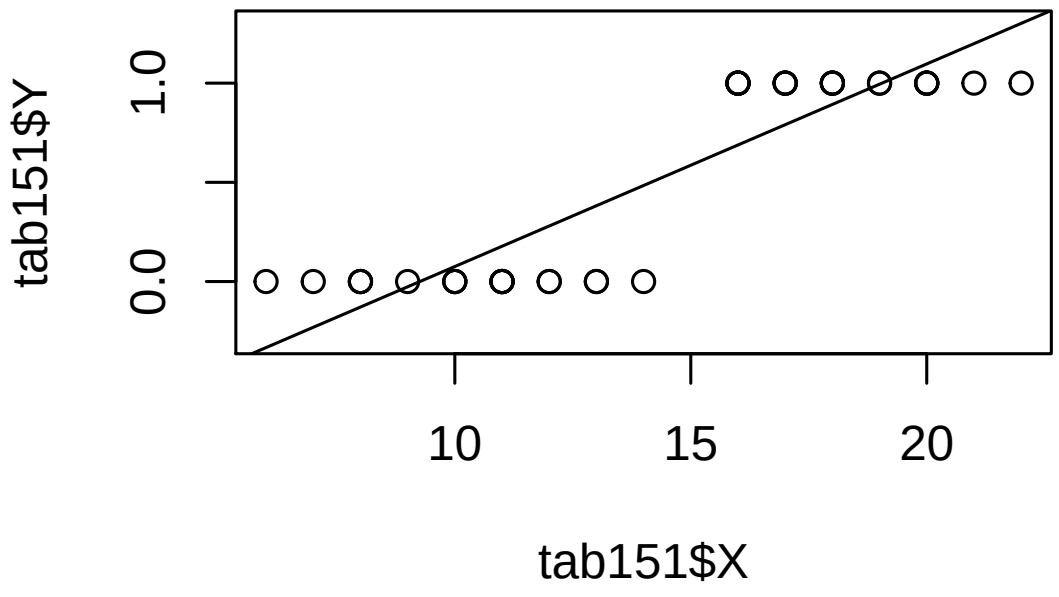
```r
tab151 <- read.table("../fichiers_aux/data/Tab151.txt", header = T)
head(tab151)
lm(data=tab151, Y~X)
```

FAMILY	Y	X
1	0	8
2	1	16
3	1	18
4	0	11
5	0	12
6	1	19

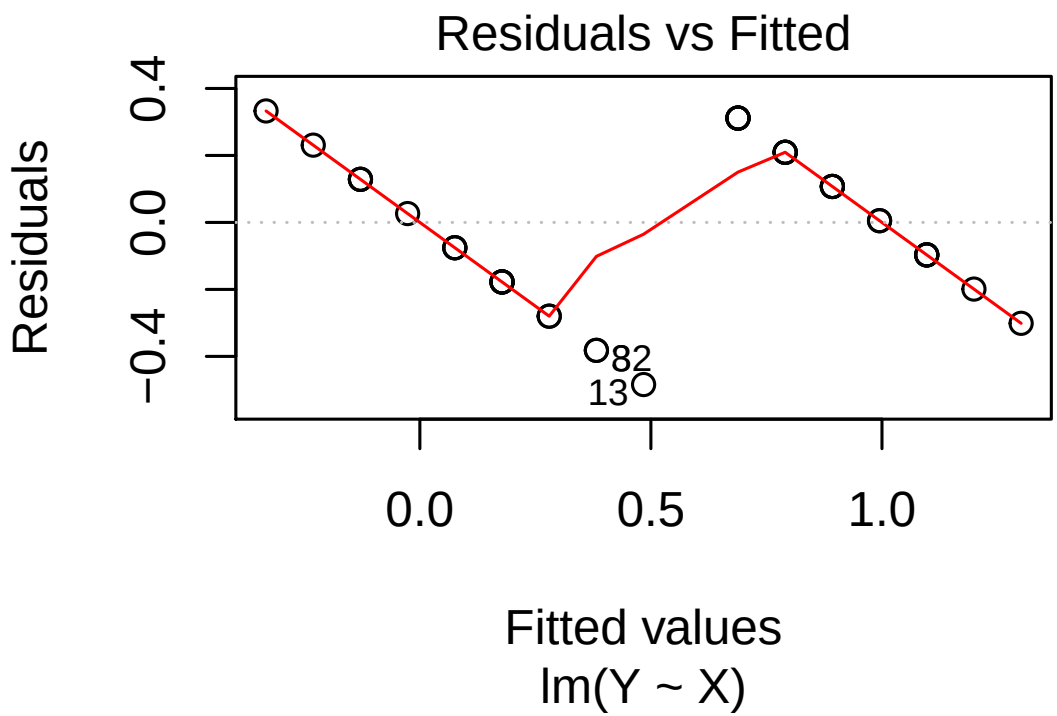
```
##
## Call:
## lm(formula = Y ~ X, data = tab151)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4842 -0.1777  0.0052  0.2095  0.3329
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.94569    0.12284  -7.698 2.85e-09 ***
## X            0.10213    0.00816  12.515 4.73e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 0.2264 on 38 degrees of freedom
## Multiple R-squared:  0.8048, Adjusted R-squared:  0.7996
## F-statistic: 156.6 on 1 and 38 DF,  p-value: 4.726e-15
```

Linear Probability Model Problems

Interpretation ?



Interpretation ?



Non normality of the Disturbances u_i

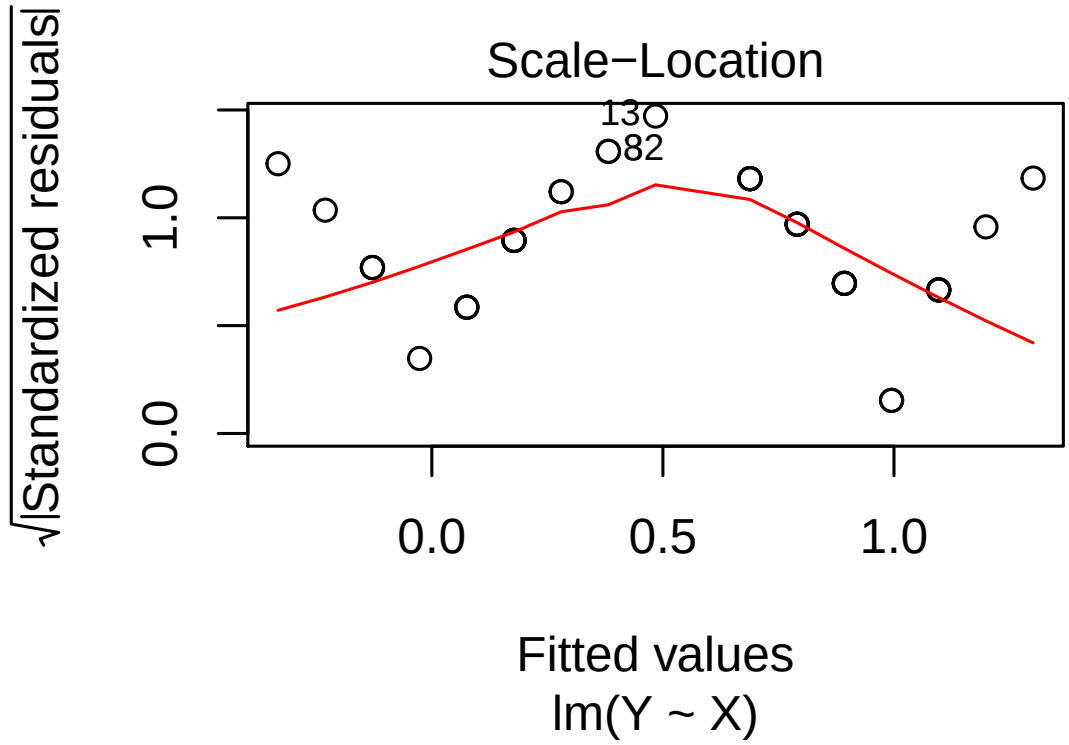
The probability distribution of u_i is given by $u_i = Y_i - \alpha - \beta X_i$

	u_i	Prob.
When $Y_i = 1$	$1 - \alpha - \beta X_i$	P_i
When $Y_i = 0$	$\alpha - \beta X_i$	$(1 - P_i)$

Heteroscedastic Variances of the Disturbances

- Even if $E(u_i) = 0$ and $cov(u_i, u_j) = 0$
- The variance of the u_i is

$$var(u_i) = P_i(1 - P_i)$$



Problems with LPM

LPM is plagued with several problems

- non-normality of u_i
- heteroscedasticity of u_i
- possibility of $\hat{Y}_i$ lying outside $[0, 1]$
- generally lower R^2

That can be surmountable...

- assumes $P_i = E(Y = 1|X)$ increases linearly with X : constant marginal effect of X

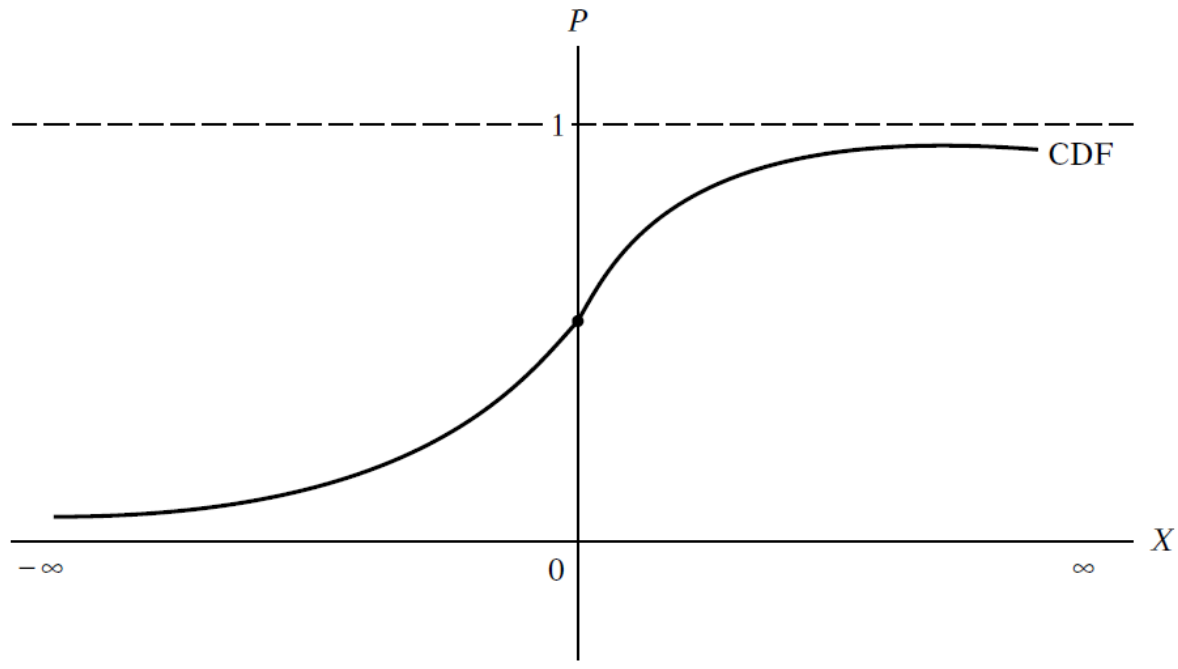


Figure 3 – A cumulative distribution function (cdf) Gujarati (2003)

Logistic Regression Model

Logit Model - v1

LPM

$$P_i = E(Y = 1|X_i) = \alpha + \beta X_i$$

Logit Model - v1

This linear form can be replaced by the logistic cdf :

$$P_i = \frac{1}{1 + e^{-Z_i}} = \frac{e^{Z_i}}{1 + e^{Z_i}}$$

where $Z_i = \alpha + \beta X_i$

- ranges between 0 and 1
- P_i is non linearly related to Z_i
- can be estimated after simple transformation : see *v2*

Logit Model - v2

The logistic model is a regression of the logit (log odds) (in favor of success at X_i) :

$$L_i = \text{logit}(P_i) = ln\left(\frac{P_i}{1-P_i}\right) = \alpha + \beta X_i$$

We can write

$$\frac{P_i}{1-P_i} = \frac{1+e^{Z_i}}{1+e^{-Z_i}} = e^{Z_i}$$

Logit Model Interpretation

Logit Model Interpretation

How can we interpret β ?

- The sign of L_i gives the relative position of P_i vs $1 - P_i$
- β measures the change in L_i for a unit change of X_i : the log-odds change !
- β sign determines whether $L_i(x)$ is increasing or decreasing as x increases.
- α is the value of the log-odds when $X_i = 0 \ldots$
- *Logit v2* : odds are an exponential function of x. The odds multiply by e^{β} for every 1-unit increase in x. In other words, e^{β} is an odds ratio, the odds at $X = x + 1$ divided by the odds at $X = x$
- For dummy variables d : e^{β} is the odds ratio for $d = 1$ vs $d = 0$.

Logit Model Interpretation

Logistic regression function has curved appearance (rather than linear) then the rate of change in $P_i(x)$ per unit change in x varies

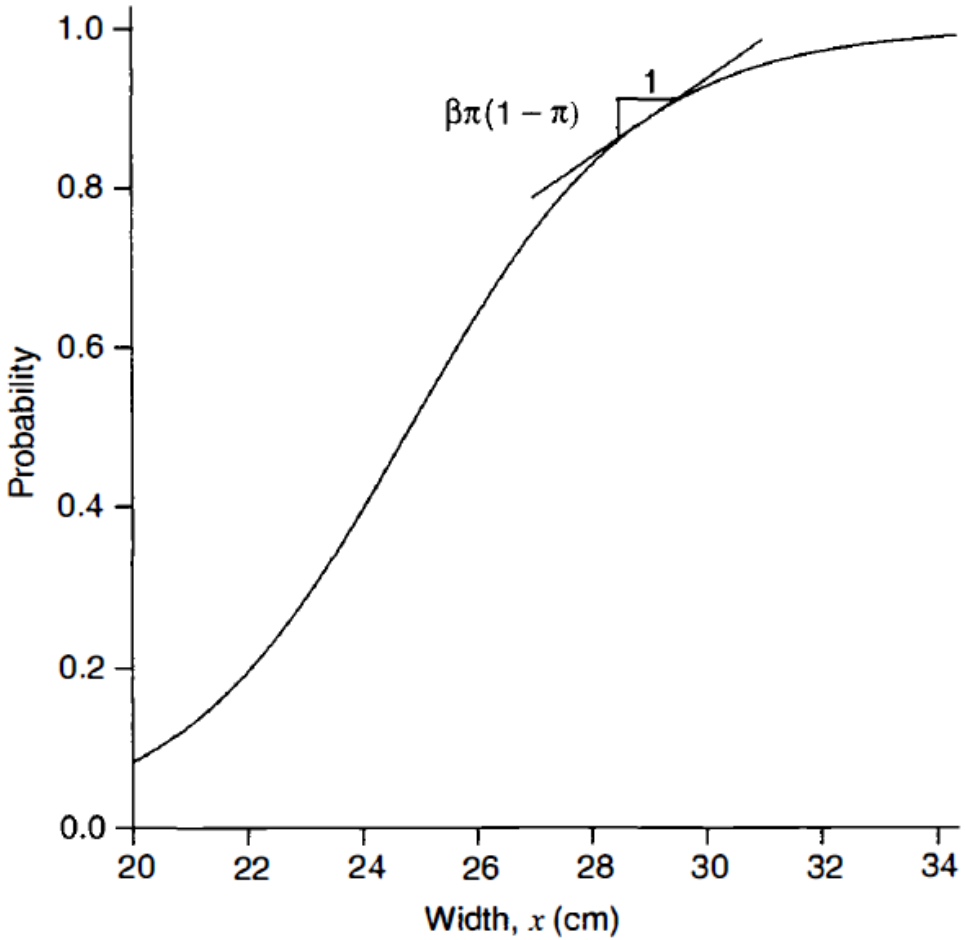


Figure 4 – Linear approximation to logistic regression curve Agresti (2013)

Logit Model Interpretation

Marginal effect (m.e.) :

- $\partial P_i(x)/\partial x$:

$$\frac{\partial P_i(x)}{\partial x_i} = \beta \cdot P_i(x) \cdot (1 - P_i(x))$$

- m.e. is large when $P_i(x) \sim 0.5$ (steepest slope)
  - i.e. $x = -\alpha/\beta \quad$ (i.e. $e^{-z_i} = 1$)
- Near x where $P_i(x) = 1/2$
  - a change in x of $1/\beta$ corresponds to a change in $P_i(x)$ of roughly $(1/\beta)(\beta/4) = 1/4$
  - that is $1/\beta$ approximates the distance between x values where $P_i(x) = 0.50$ and where $P_i(x) = 0.25$ or 0.75.

Logit Model Interpretation

Derivative of the probability

$$\begin{aligned}
\frac{\partial P_i(x)}{\partial x} &= \frac{\partial(e^{Z_i}/1+e^{Z_i})}{\partial x} \\
&= \frac{e^{Z_i}(1+e^{Z_i}) - e^{Z_i}e^{Z_i}}{(1+e^{Z_i})^2}\frac{\partial Z_i}{\partial x} \\
&= \frac{e^{Z_i}}{(1+e^{Z_i})^2}\frac{\partial Z_i}{\partial x} \\
&= \frac{\partial Z_i}{\partial x}P_i(1-P_i)
\end{aligned}$$

With $P_i(x) = \frac{e^{Z_i}}{1+e^{Z_i}}$ and $1 - P_i(x) = \frac{1}{1+e^{Z_i}}$

Logit Model Interpretation

Computing marginal effects (quantitative x)

Since the rate of change varies with x :

1. Evaluate the marginal effects at the sample means of the data ($\bar{x}$)
2. Evaluate the average partial effects : $APE = E[\frac{\partial P_i(x)}{\partial x}]$ use the sample average of the individual marginal effects ($\bar{m.e.}$)

$$\frac{\partial P_i(x)}{\partial x} = \frac{\partial F(t)}{\partial t}\frac{\partial t}{\partial x} = f(\alpha + \beta x)\beta$$

$$\hat{APE} = \frac{1}{n}\sum_{i=1}^{n} f(\hat{\alpha} + \hat{\beta} X_i)\hat{\beta}$$

with the logistic c.d.f. : $F(x) = \frac{e^x}{1+e^x}$ and $f(x) = \frac{e^x}{(1+e^x)^2}$

(or the probit density function for probit)

Logit Model Interpretation

Computing marginal effects (quantitative x)

- MEA and AME give Equivalent results in large samples
- Current practice favors averaging the individual marginal effects

Computing marginal effects dummy variables d

$$m.e. = P(Y = 1|\bar{x}_{(d)}, d = 1) - P(Y = 1|\bar{x}_{(d)}, d = 0)$$

where $\bar{x}_{(d)}$ denotes the means of all other variables.

Logit Model Estimation

Estimation in Binary Logit

- Estimation of binary models is usually based on ML.
- Each observation is a single draw from a Bernoulli distribution (binomial with one draw).
- The model (resp. Logit) with success probability $P_i = F(\alpha + \beta X_i)$ (resp. $F(\alpha + \beta X_i) = \frac{1}{1+e^{\alpha+\beta X_i}}$) and n iid observations
- leads to the joint probability (likelihood function) :

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n|X) = \prod_{i=1}^{n} F_i(Y_i)$$

$$= \prod_{y_i=1} F(\alpha + \beta X_i) \prod_{y_i=0} [1 - F(\alpha + \beta X_i)]$$

Estimation in Binary Logit

- The likelihood function for a sample of n observations

$$\mathcal{L} = f(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n|X) = \prod_{i=1}^{n} P_i^{y_i}(1 - P_i)^{1-y_i}$$

$$= \prod_{i=1}^{n} [F(\alpha + \beta X_i)]^{y_i}[1 - F(\alpha + \beta X_i)]^{1-y_i}$$

- the log likelihood function :

$$ln\mathcal{L} = lnf(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n|X) =$$

$$\sum_{i=1}^{n} y_i ln[F(\alpha + \beta X_i)] + (1 - y_i) ln[1 - F(\alpha + \beta X_i)]$$

- with the logit function : $F(\alpha + \beta X_i) = \frac{1}{1+e^{\alpha+\beta X_i}}$ for logit model

Estimation in Binary Probit

- For the normal distribution, the log-likelihood is

$$lnL = \sum_{y_i=1} \Phi(\alpha + \beta X_i) \sum_{y_i=0} ln[1 - \Phi(\alpha + \beta X_i)]$$

Likelihood equations

$$\frac{\partial ln\mathcal{L}}{\partial \beta} = \sum_{i}^{n} \left(\frac{y_i f_i}{F_i} + (1 - y_i) \frac{-fi}{1 - F_i} \right) x_i = 0$$

- where $fi = dF_i / d(\beta x_i)$
- choice of F_i leads to the empirical model

Likelihood equations

- Unless for LPM, these equation are non-linear and require an iterative solution
- if x_i contains a constant term, the first-order conditions imply that the average of the predicted probabilities must equal the proportion of ones in the sample

Optimisation method

Likelihood equations

- Second derivatives for the logit model :

$$H = \frac{\partial^2 ln\mathcal{L}}{\partial\beta\partial\beta'}$$

  do not involve the random variable y_i , so Newton's method is also the method of scoring
- Hessian is always negative definite, so the log-likelihood is globally concave (less obvious for probit)
- Newton's method will usually converge to the maximum of the log-likelihood in just a few iterations

Asymptotic covariance matrix

Likelihood equations

- Asymptotic covariance matrix for the maximum likelihood estimator can be estimated by using the inverse of the Hessian evaluated at the maximum likelihood estimates
- Hessian for the logit model does not involve y_i, so $H = E[H]$, not true for probit.

Asymptotic covariance matrix

For the logit model :

$$\mathcal{L} = \prod_i^n p_i^{y_i}(1 - p_i^{1-y_i}) = \prod_i^n \left(\frac{p_i}{1-p_i}\right)^{y_i}(1-p_i)$$

$$ln(\mathcal{L}) = \sum_i y_i ln\left(\frac{p_i}{1-p_i}\right) + \sum_i ln(1-p_i)$$

$$= \sum_i \beta x_i y_i - \sum_i ln(1 + e^{\beta x_i})$$

Asymptotic covariance matrix

For the logit model :

$$\begin{aligned}
\frac{\partial ln\mathcal{L}}{\partial \beta} &= \sum_{i=1}^{n}\left(\frac{y_i f_i}{F_i} + (1-y_i)\frac{-f_i}{1-F_i}\right)x_i = 0 \\
&= \sum_i x_i y_i - \sum_i x_i (1+e^{-\beta x_i})^{-1} \\
&= \sum_i x_i y_i - \sum_i x_i \hat{y}_i \\
\frac{\partial^2 ln\mathcal{L}}{\partial \beta \partial \beta} &= -\sum_{i}^{n} x_i x_i' \hat{y}_i (1-\hat{y}_i)
\end{aligned}$$

where $\hat{y}_i = \frac{1}{1+e^{-\beta x_i}}$

Exercise : Binary Logit and Probit

Exercise : Effect of Personalized system of instruction [2]

- Research question : *Does a new method of teaching economics (Personalized System of Instructions, PSI) influence performance in later economics courses*
- Data available
  - GRADE : Indicator of improving grade between basic and intermediate economics courses (binary)
  - GPA : Grade point average (number in range)
  - TUCE : Score on pretest : entering knowledge (number in range)
  - PSI : Exposure to new teaching method (binary)

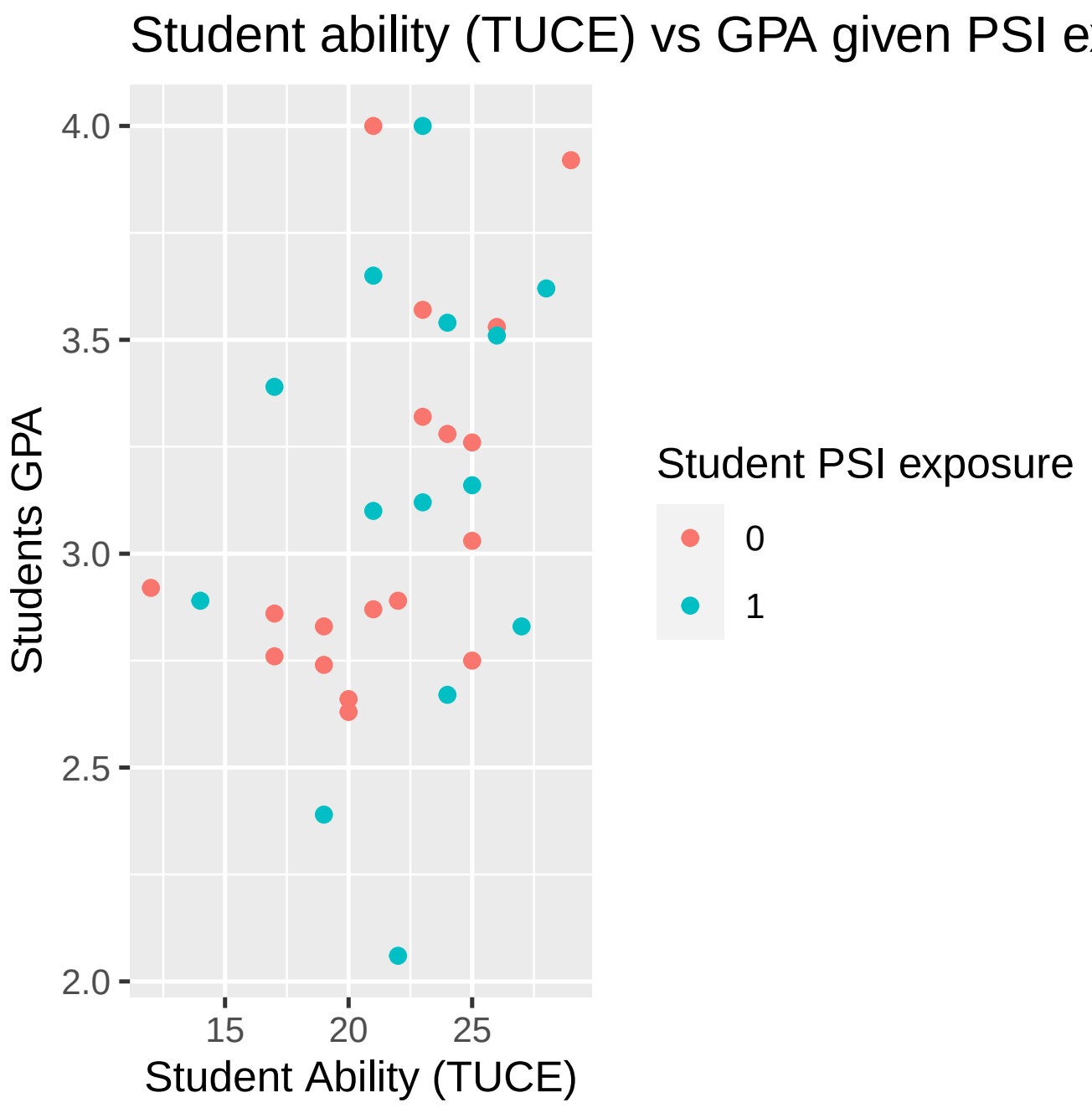
2. Example Greene (2008) p.694 : Study of Spector and Mazzeo (1980) Data:(http://people.stern.nyu.edu/wgreene/Text/Edition6/tablelist6.htm)

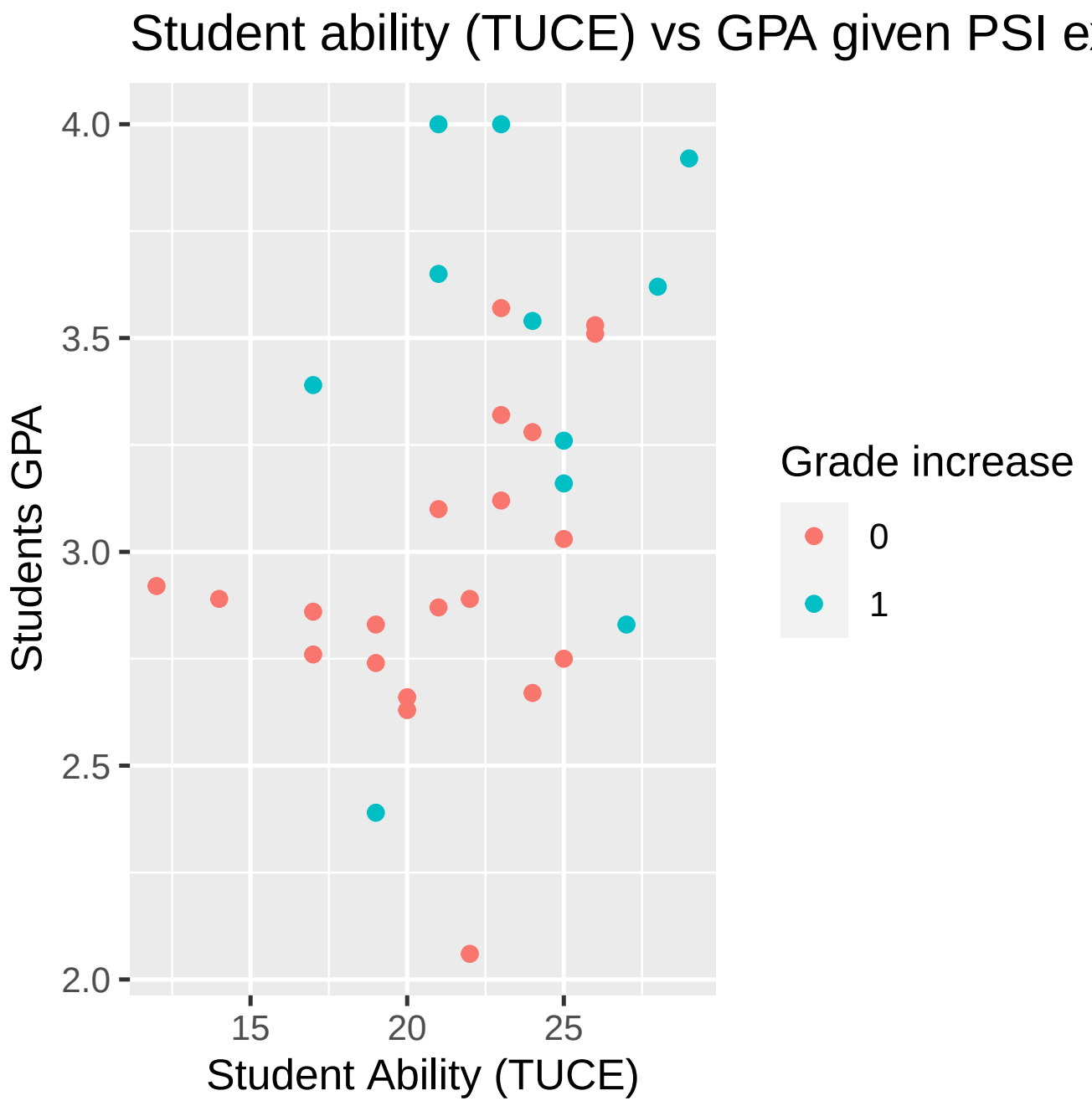
Estimating a Binary Logit

```
library(faraway)
head(spector)
spector$psi <- factor(spector$psi)
spector$grade <- factor(spector$grade)
```

grade	psi	tuce	gpa
0	0	20	2.66
0	0	22	2.89
0	0	24	3.28
0	0	12	2.92
1	0	21	4.00
0	0	17	2.86

```
ggplot(data = spector, aes(x=tuce, y=gpa)) +
   geom_point(aes(colour=psi)) +
   ggtitle("Student Ability (TUCE) vs GPA given PSI exposure") +
   xlab("Student Ability (TUCE)") + ylab("Students GPA") +
   labs(colour="Student PSI exposure")
```


Student ability (TUCE) vs GPA given PSI ex
Students GPA
4.0
3.5
3.0
2.5
2.0
15
20
25
Student Ability (TUCE)
Student PSI exposure
0
1


Student ability (TUCE) vs GPA given PSI ex
Students GPA
4.0
3.5
3.0
2.5
2.0
15
20
25
Student Ability (TUCE)
Grade increase
0
1

Estimating the corresponding Binary Logit

Exercise

Assuming :

$$L_i = ln\left(\frac{P_i}{1-P_i}\right) = \alpha + \beta_2 GPA_i + \beta_3 TUCE_i + \beta_3 PSI_i + u_i$$

- Let estimate the logit model : - Calculate effects of variables on odds-ratio - Calculate marginal effects on probabilities at the mean point

```
Logit1 <- glm(grade ~ gpa + tuce + psi, x = TRUE,
    data = spector, family = binomial(link = "logit"))
summary(Logit1)
exp(coefficients(Logit1))
```

```
##
## Call:
## glm(formula = grade ~ gpa + tuce + psi, family = binomial(link = "lo
##     data = spector, x = TRUE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9551  -0.6453  -0.2570   0.5888   2.0966
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -13.02135    4.93127  -2.641  0.00828 **
## gpa           2.82611    1.26293   2.238  0.02524 *
## tuce          0.09516    0.14155   0.672  0.50143
## psi1          2.37869    1.06456   2.234  0.02545 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 41.183  on 31  degrees of freedom
```

Odds Ratios

```
## [1] "exp(beta)"
```

```
## (Intercept)         gpa        tuce        psi1
##      0.0000     16.8797      1.0998     10.7907
```

```
# Logit  # xb*:
betas<-t(data.frame(coef(Logit1)))  ;  betas
xmean <- c(1, mean(spector$gpa), mean(spector$tuce),
           mean(as.numeric(spector$psi))-1)
xmean
print("XBetas:")
xb_logit <- sum(xmean*betas)  ;  xb_logit
# Slopes (at mean): Lambda(mean(xb))*(b)
print("Slopes:")
logit_slopes <- dlogis(xb_logit)*betas  ;  logit_slopes
```

```
##                (Intercept)      gpa       tuce     psi1
## coef.Logit1.    -13.02135 2.826113 0.09515766 2.378688

## [1]  1.000000  3.117188 21.937500  0.437500

## [1] "XBetas:"

## [1] -1.083627

## [1] "Slopes:"

##                (Intercept)       gpa       tuce      psi1
## coef.Logit1.    -2.459761 0.5338588 0.01797549 0.4493393
```

Estimating the corresponding Binary Probit

Exercise

- Let estimate the probit model
- Calculate effects of variables on odds-ratio
- Calculate marginal effects on probabilities at the mean point

```r
probit1 <- glm(grade ~ gpa + tuce + psi, x = TRUE,
    data = spector, family = binomial(link = "probit"))
summary(probit1)
```

```
##
## Call:
## glm(formula = grade ~ gpa + tuce + psi, family = binomial(link = "pr
##     data = spector, x = TRUE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9392  -0.6508  -0.2229   0.5934   2.0451
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.45231    2.57152  -2.898  0.00376 **
## gpa          1.62581    0.68973   2.357  0.01841 *
## tuce         0.05173    0.08119   0.637  0.52406
## psi1         1.42633    0.58695   2.430  0.01510 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparison of marginal effects at mean :

	X.Intercept.	gpa	tuce	psi1
coef.probit1.	-2.444734	0.5333484	0.0169695	0.4679083
coef.Logit1.	-2.459761	0.5338588	0.0179755	0.4493393

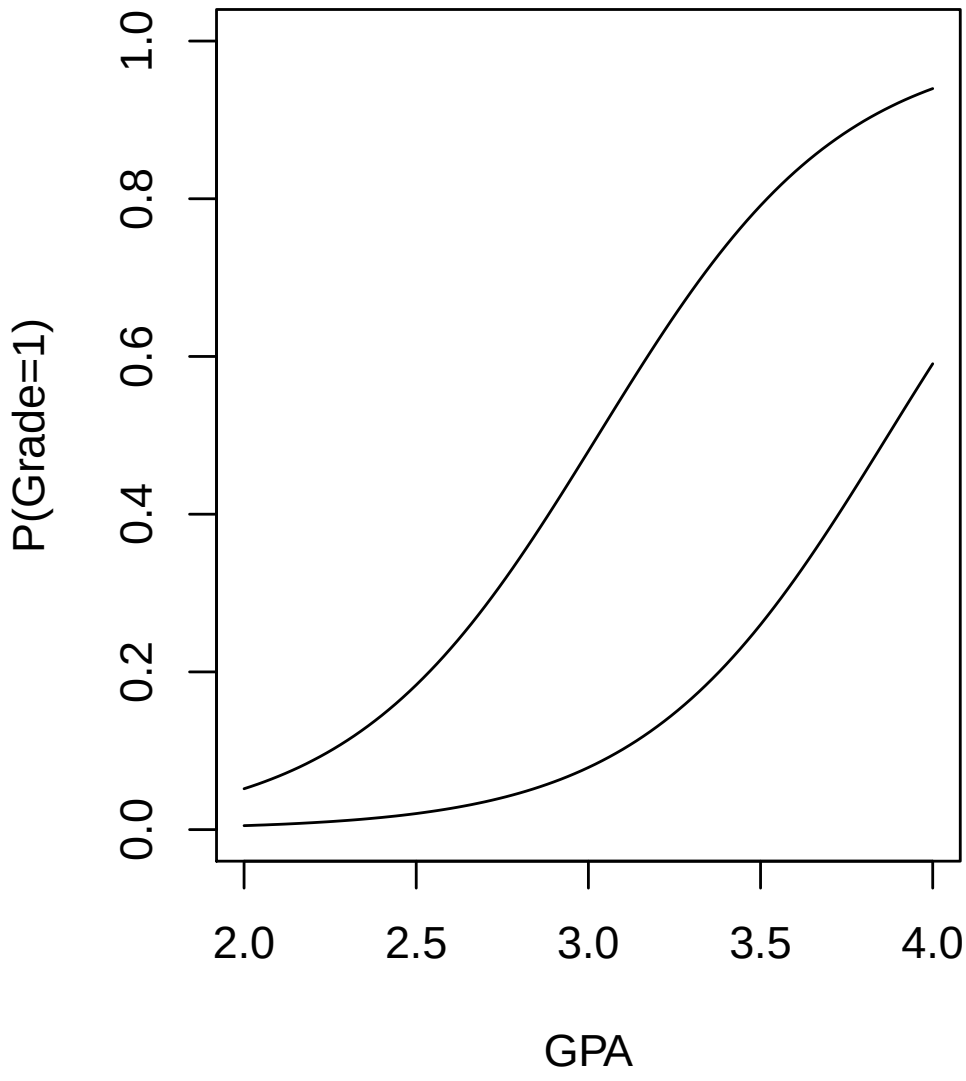
Estimating the corresponding Binary Logit

Exercise

- Draw Predicted probabilities with varying GPA, fixed TUCE (at its average), and for PSI $= 0$ or PSI $= 1$, for logit and probit

```
# Probability Curves for the Probit model
psi0 <- function(x) pnorm(-7.452 + 1.626*x  + 0.052*21.938)
psi1 <- function(x) pnorm(-7.452 + 1.626*x  + 0.052*21.938 + 1.426)
curve(psi0, xlim=c(2,4), ylim=c(0,1), main="Effect of PSI on Pred.
      Probabilities (Probit)", ylab="P(Grade=1)", xlab="GPA")
curve(psi1, add=T)

# Probability Curves for the Logit model
psi0 <- function(x) plogis(-13.021 + 2.826*x  + 0.095*21.938)
psi1 <- function(x) plogis(-13.021 + 2.826*x  + 0.095*21.938 + 2.379)
curve(psi0, xlim=c(2,4), ylim=c(0,1), main="Effect of PSI on Pred.
      Probabilities (Logit)", ylab="P(Grade=1)", xlab="GPA")
curve(psi1, add=T)
```


Effect of PSI on Pred. Probabilities (Logi
P(Grade=1)
0.0
0.2
0.4
0.6
0.8
1.0
2.0
2.5
3.0
3.5
4.0
GPA

Estimating the corresponding Binary Logit

Exercise

- Calculate Average Partial Effect, for logit and probit

Average Marginal effects

predict() returns $X\beta$

```
# Probit
fav_probit <- mean(dnorm(predict(probit1, type = "link")))
fav_probit * coef(probit1)
# Logit
fav_Logit <- mean(dlogis(predict(Logit1, type = "link")))
fav_Logit * coef(Logit1)
```

V1	X.Intercept.	gpa	tuce	psi1
probit	-1.65375669119005	0.360787010132720.011479159271291		4.31651968447309
probit	-1.6705954252256	0.362580831594724	0.012208410957018	2.30517770227781 2

- LPM, logit, and probit give qualitatively similar results
- Between logit and probit, which model is preferable ?
- In most applications the models are quite similar
- Main difference being slightly fatter tails of the logistic distribution
- Conditional probability P_i approaches 0 or 1 at a slower rate in logit than in probit
- There is no compelling reason to choose one over the other
- In practice, logit model because of its comparative mathematical simplicity

Inference in Binary Logit Model

Hypothesis tests

Three familiar procedures for conventional hypothesis tests about restrictions on the model coefficients

1.Likelihood Ratio test

- The likelihood ratio statistic is : $\lambda_{LR} = 2(lnL_1 - lnL_2)$
- where lnL_1 indicates the log-likelihood computed at the unrestricted (alternative) estimator
- lnL_0 at the restricted (null) estimator
- It follows an $\chi^2(df)$, where df are the degree of freedom = number of restrictions

2. Wald test

- Hypothesis : $r(\theta, c) = 0$
  $r(\theta, c)$ is a vector of J functionally independent restrictions on θ and c is a vector of constants.
- The Wald statistic uses the delta method to obtain an asymptotic covariance matrix for $r(\theta, c)$.
  The statistic is : $W = r(\theta, c)'[Var(r(\theta, c))]^{-1} r(\theta, c)$

Example : Wald test

```r
# Former Probit model : probit1
# a) joint significance of all regressors
library(aod)
wald.test(b = coef(probit1), Sigma = vcov(probit1),
          Terms =2:length(coef(probit1)))
# b) linear combination of coefficients
# (e.g. are the coefficients signif. different?)
restr <- cbind(0, -1, 1, 0)
wald.test(b = coef(probit1), Sigma = vcov(probit1), L = restr)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 10.6, df = 3, P(> X2) = 0.014

## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 4.8, df = 1, P(> X2) = 0.028
```

```r
# LR test
library(lmtest)
probit0 <- glm(grade ~ 1 , x = TRUE,
    data = spector, family = binomial(link = "probit"))
summary(probit0)
lrtest(probit1, probit0)
```

```
##
## Call:
## glm(formula = grade ~ 1, family = binomial(link = "probit"),
##     data = spector, x = TRUE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9178  -0.9178  -0.9178   1.4614   1.4614
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.4023     0.2282  -1.763   0.0779 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 41.183  on 31  degrees of freedom
## Residual deviance: 41.183  on 31  degrees of freedom
## AIC: 43.183
##
```

Measuring the Fit of the Model to the Data

ρ^2 or LRI

In Logit and Probit models :

- There is no R squared
- There are no residuals or sums of squares
- The model is not computed to optimize the fit of the model to the data

Measuring the Fit of the Model to the Data

ρ^2 or LRI

Considering *null* and *full* model :

- The null model or *restricted* model is : $E(Y_i) = \bar{y}_i$, (LnL_0)
- The full model or *saturated* model is : $E(Y_i) = y_i$, then the likelihood is 1 and $lnL_f = 0$

Evaluating *deviances*

- Null deviance is : $D_0 = -2ln\frac{L_0}{L_f} = 2(LnL_f - LnL_0)$
- Residual deviance is $D_x = -2ln\frac{L_x}{L_f} = 2(LL_f - LL_x)$
- Variable effect is measured by $D_x - D_0 = -2ln\frac{L_x}{L_0} = 2(LL_0 - LL_x)$

```
sum(as.numeric(spector$grade)-1 )
```

```
## [1] 11
```

```
length(spector$grade)
```

```
## [1] 32
```

```
sum(as.numeric(spector$grade)-1 )/length(spector$grade)
```

```
## [1] 0.34375
```

```
-2*(11*log(0.34375)+21*log(1-0.34375))
```

```
## [1] 41.18346
```

Measuring the Fit of the Model to the Data

- Likelihood ratio index (LRI) ou Pseudo-R^2 :

$$\rho^2 = 1 - \frac{lnL\beta^*}{lnL_0}$$

  Where $L(\beta^*)$ is the Likelihood of the estimated model and $L(0)$ the Likelihood of the *null* model with no regressors (only intercept)
- McFadden R^2 : $R^2_{MCF} = 1 - \frac{ln(L(\beta^*))}{ln(L_0)}$
- Cox and Snell R^2 is : $R^2_{C\&S} = 1 - (\frac{L_0}{L(\beta^*)})^{2/n}$

To Compare Models :

You can also use :

- $logL$
- Use information criteria to compare nonnested models (AIC, BIC,. . .)
  - $AIC = -2 \times lnL + 2 \times k$,
    where k represents the number of parameters in the fitted model
  - $BIC = -2 \times lnL + k \times log(n)$,

The Hosmer-Lemeshow goodness of fit test

- Dividing the sample up according to their predicted probabilities, or risks.
- Observations in the sample are then split into g groups according to their predicted probabilities
- With g=10, the 1rst group consists of the observations with the lowest 10% predicted probabilities, etc
- Within each interval, the expected number of events is obtained by adding up the predicted probabilities
- The expected number of non-events is obtained by subtracting the expected number of events from the number of cases in the interval

The Hosmer-Lemeshow goodness of fit test

- These expected frequencies (e_{kl}) are compared with observed frequencies (o_{kl}) by the conventional Pearson chi-square statistic
- The degrees of freedom is the number of intervals minus 2

$$\sum_{k=0,1} \sum_{l=1}^{g} \frac{(n_{kl} - e_{kl})^2}{e_{kl}}$$

```
library(ResourceSelection)
Logit1 <- glm(grade ~ gpa + tuce + psi, x = TRUE,
    data = spector, family = binomial(link = "logit"))
names(Logit1)
```

The Hosmer-Lemeshow goodness of fit test

```
##  [1] "coefficients"      "residuals"         "fitted.values"
##  [4] "effects"           "R"                 "rank"
##  [7] "qr"                "family"            "linear.predictors"
## [10] "deviance"          "aic"               "null.deviance"
## [13] "iter"              "weights"           "prior.weights"
## [16] "df.residual"       "df.null"           "y"
## [19] "converged"         "boundary"          "model"
## [22] "x"                 "call"              "formula"
## [25] "terms"             "data"              "offset"
## [28] "control"           "method"            "contrasts"
## [31] "xlevels"
```

```
hoslem.test(Logit1$y, fitted(Logit1))
```

The Hosmer-Lemeshow goodness of fit test

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  Logit1$y, fitted(Logit1)
## X-squared = 6.0948, df = 8, p-value = 0.6366
```

ROC Curves

ROC is short for Receiver Operating Characteristic

Classification tables

to generate actual predictions of whether or not $Y = 1$, we need some cutpoint value **natural choice .5**

```
classDF <- data.frame(response = Logit1$y, predicted = round(fitted(Log

xtabs(~ predicted + response, data = classDF)
```

```
##           response
## predicted  0  1
##         0 18  3
##         1  3  8
```

Classification Table

Overall proportion of predictions that are correct

```
ClassTab <- xtabs(~ predicted + response, data = classDF)


(ClassTab[1,1] + ClassTab[2,2] ) / sum(ClassTab)
```

```
## [1] 0.8125
```

Classification table

But

- Suppose that a data set has 100 events and 900 non-events.
- A no predictors model will generate predicted values that are all .10,
- all the cases would be predicted as non-events
- right 90 % of the time !

Classification table

Sensitivity

- proportion of non-events that are correctly predicted, in this case 18/21

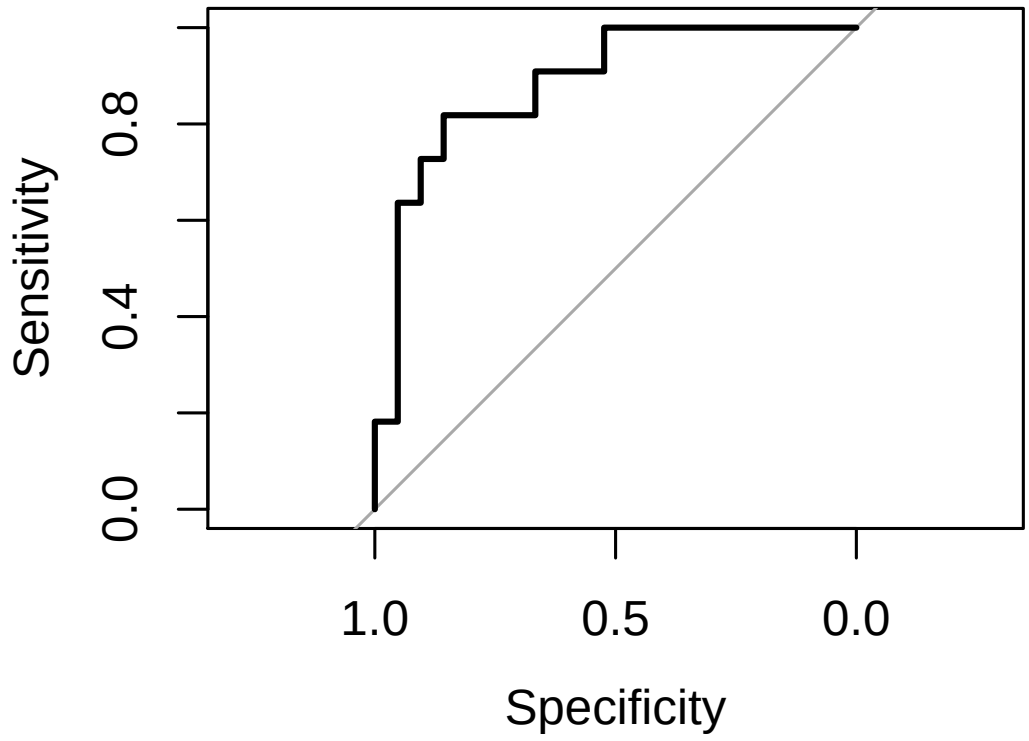
Specificity

- proportion of events that are correctly predicted, 8/11

ROC

- for each cut point$c$ sensitivity and specificity are estimated
- increase the cut-point c, fewer observations will be predicted as positive
- ROC : plot of the values of sensitivity against one minus specificity, as the value of the cut-point c is increased from 0 through to 1

```
predpr <- predict(Logit1,type=c("response"))
library(pROC)
roccurve <- roc(spector$grade ~ predpr)
plot(roccurve)
```


Sensitivity
0.0
0.4
0.8
1.0
0.5
0.0
Specificity

ROC

- The 45-degree line represents the expected ROC curve for the **null** model (no predictive power)
- The more the curve departs from the 45-degree line, the greater the predictive power.
- The standard statistic for summarizing that departure is the area under the curve

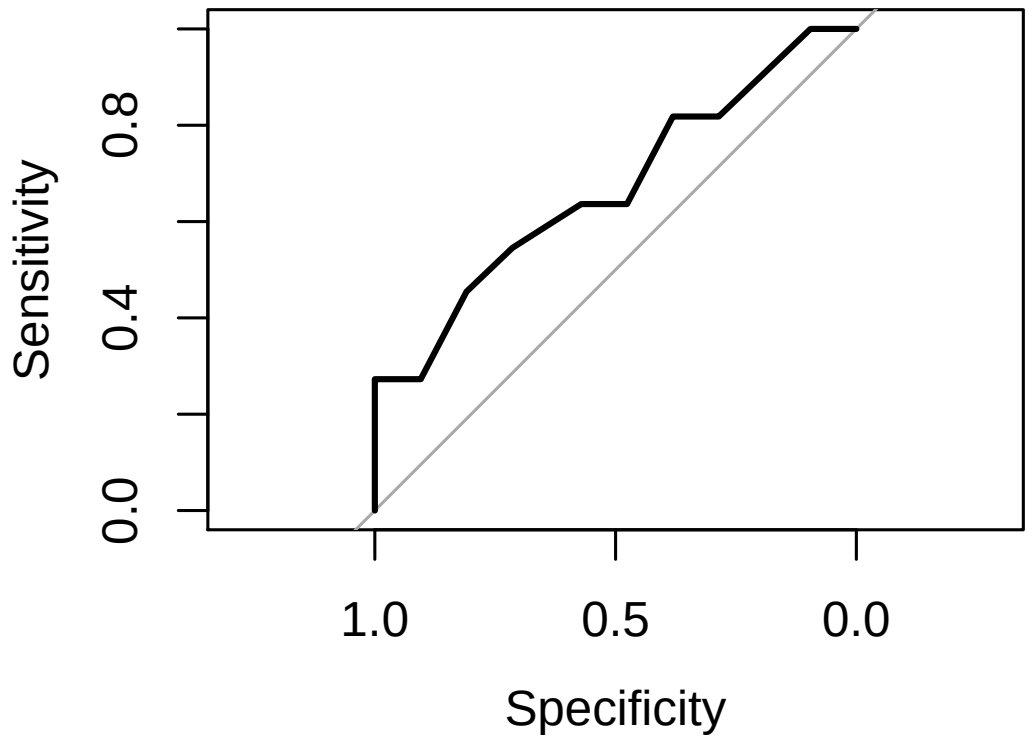
AUC - Area under the ROC

- Discrimination ability : ROC curve which goes closer to the top left hand corner
- No discrimination ability has an ROC curve close to a 45 degree line
- AUC ranges from 1, corresponding to perfect discrimination, to 0.5, corresponding to a model with no discrimination ability

```
auc(roccurve)
```

```
## Area under the curve: 0.8831
```

```
Logit2 <- glm(grade ~ tuce , x = TRUE,
    data = spector, family = binomial(link = "logit"))
predpr2 <- predict(Logit2,type=c("response"))
library(pROC)
roccurve2 <- roc(spector$grade ~ predpr2)
plot(roccurve2)
auc(roccurve2)
```




```
## Area under the curve: 0.6688
```

Multinomial Logistic Regression

Multinomial Logistic Regression

Generalize the model to J categories, with the running index $j = 1, \ldots, J$. Let p_{ij} be the probability that individual i falls into category j.

The model is then

$$L_{ij} = ln\left(\frac{P_{ij}}{P_{iJ}}\right) = \alpha_j + \beta_j X_i$$

$j = 1, \ldots, J - 1$

where - x_i is a column vector of variables describing individual i - β_j is a row vector of coefficients for category j - Note that each category is compared with the highest category J

Multinomial Logistic Regression

with $Z_{ij} = \alpha_j + \beta_j X_i$

We can write

$$P_{ij} = \frac{e^{Z_{ij}}}{1 + \sum_{k=1}^{J-1} e^{-Z_{ik}}}$$

Because the probabilities for all J categories must sum to 1, we have

$$P_{iJ} = \frac{1}{1 + \sum_{k=1}^{J-1} e^{-Z_{ik}}}$$

Interpreting the coefficients

Comparing any two alternatives j and k of the dependent variable :

$$ln\left(\frac{P_{ij}}{P_{ik}}\right) = (\beta_j - \beta_k)x_i$$

- naturally interpreted in terms of effects on contrasts between pairs of categories for the dependent variable

Problem if interpreting the probabilities

Suppose

$$ln\left(\frac{P_1}{P_3}\right) = 1.0 - 2.0x$$

$$ln\left(\frac{P_2}{P_3}\right) = 1.0 - 1.0x$$

- it's tempting to say that increases in x produce decreases in the probability of being in category 2.
- But if we graph the probabilities,

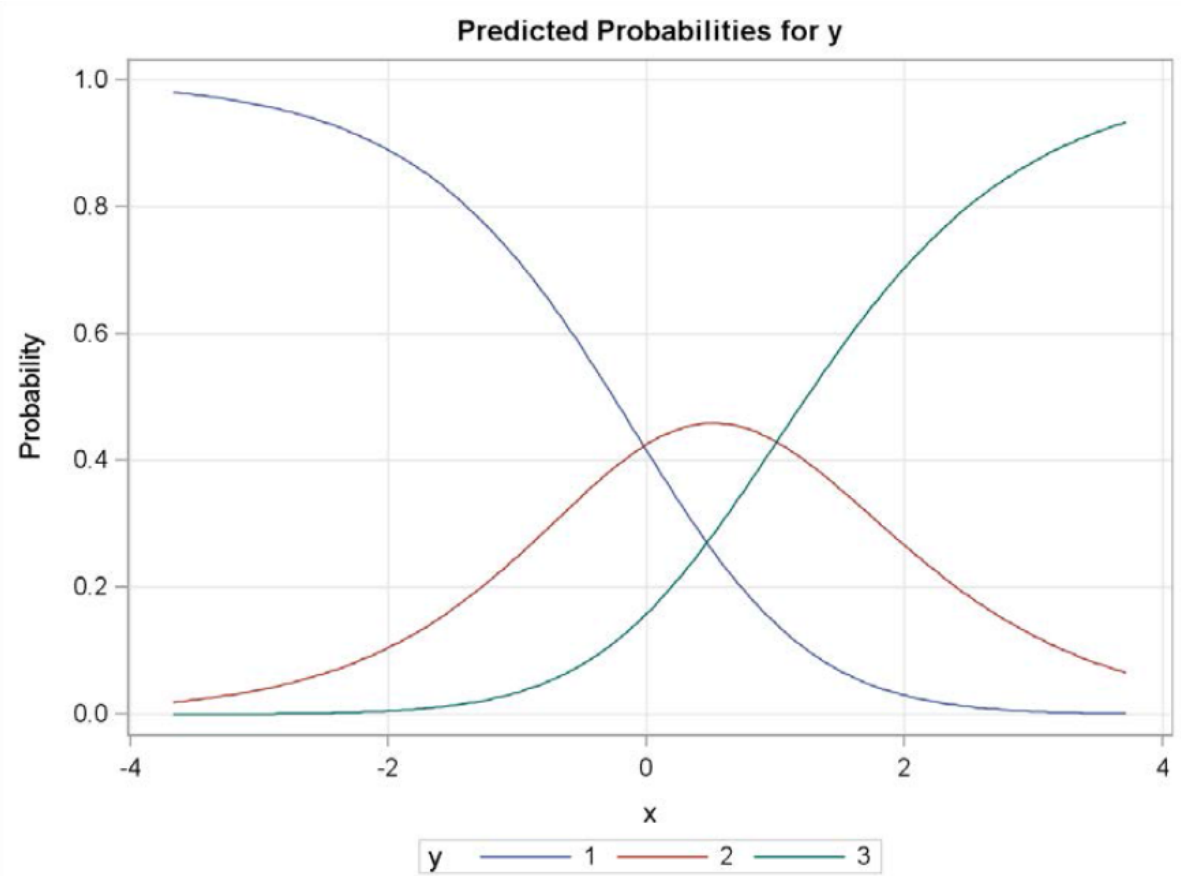


Figure 5 – Effect of x on the Probability of Being in Each of the Three Categories

- multinomial logit coefficients must always be interpreted as effects on contrasts between pairs of categories
- never on the probability of being in a particular category

Bibliography

Agresti, A. 2013. *Categorical Data Analysis*. Wiley.

Greene, W. H. 2008. *Econometric Analysis, 6th*. Prentice-HallOxford : Clarendon Press.

Gujarati, D. 2003. *Basic Econometrics*. McGraw & Hill.

Harrell, F. E. 2013. *Regression Modeling Strategies : With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer Series in Statistics. Springer New York. https://books.google.fr/books?id=7D0mBQAAQBAJ.

Bibliography

McFadden, D. 2014. "The New Science of Pleasure : Consumer Choice Behavior and the Measurement of Well-Being." In *Handbook of Choice Modelling*, edited by S. Hess and A. Daly, 7–48. UK : Edward Elgar.

Spector, L. C., and M. Mazzeo. 1980. "Probit Analysis and Economic Education." *Journal of Economic Education* 11 : 37–44.