ETH zürich

# Applied Generalized Linear Models (FS 20)

## Binary outcomes: logistic regression and probit models (Practical)

Viviana Amati

# Course structure - schedule

| Date | Topic | Assignment |
|---|---|---|
| 18.02 | Introduction to the course | Ass. 1<br>released on 25.02<br>due to 19.03 |
| 25.02 | Introduction to R and review of the linear regression mode | |
| 03.03 | The general linear model: ANOVA and ANCOVA | |
| 10.03 | Practical: ANOVA and ANCOVA | |
| 17.03 | Binary outcomes: logistic regression and probit models | Ass. 2<br>released on 18.03<br>due to 23.04 |
| **24.03** | **Practical: logistic regression and probit models** | |
| 31.03 | Nominal outcomes: multinomial logistic regression | |
| 07.04 | Practical: multinomial logistic regression | |
| 21.04 | Ordinal outcomes: ordered logistic regression and probit models | Ass. 3<br>released on 25.04<br>due to 21.05 |
| 28.04 | Practical: ordered logistic regression and probit models | |
| 05.05 | Count outcomes: Poisson and negative binomial models | |
| 12.05 | Practical: Poisson and negative binomial models | |
| 19.05 | Survival models | L+P |
| 26.05 | Regular lecture: panel data model | L+P |

# A bit of (re-)organization

- Assignments:
  - A few people are working alone.
    Please let me know if you would like to team up and will put you in contact
  - Some groups forgot to attach the script and will receive an email
- Exam:
  - I am collecting the available options and publish them in moodle by the end of this week
  - I will call for an opinion poll at the beginning of the lecture next week
  - If you cannot attend the lecture send me an email by Monday 29 March
- Zoom:
  - Instead of writing in chat please unmute your mic. and ask your question
  - You can also ask questions after the lecture
  - Will be in zoom every Thursday from 5 p.m. to 6 p.m. for questions
    Join URL: https://ethz.zoom.us/j/702918320 Meeting ID: 702 918 320

# Today's agenda

- Quick overview of the ANCOVA model example
- Estimation of binary logistic regression models with R
- Interpretation of the model results
- Data and scripts are in the folder `BLR.zip` in moodle
- Commented output in lecture notes

# ANCOVA model: Data

The data set `incomeRaceEduc.dat` contains information collected on a sample of 80 adults American aged over 25. The variables in the data set are:

- inc: annual income (thousands of dollars)
- educ: number of years of education (12 = high school graduate, 16 = college graduate)
- race: racial-ethnic group (b = black, h=hispanic, w=white)

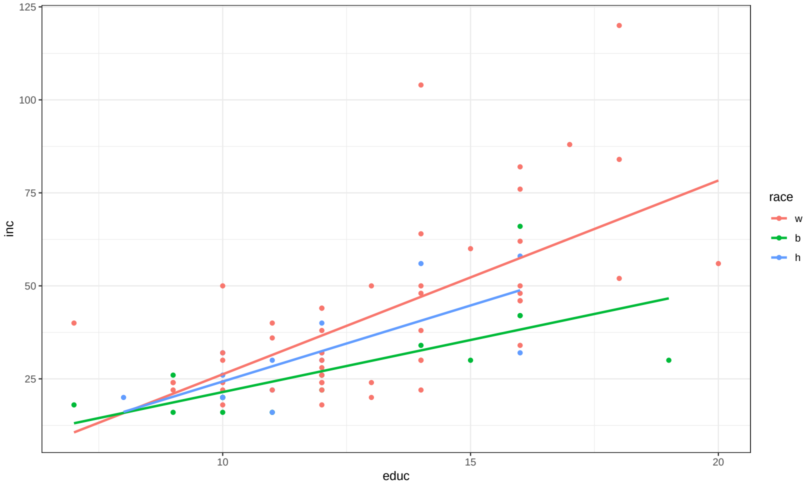
Researchers would like to test if there is a relationship between race and income while controlling for education.

# ANCOVA model

$$Y_{ij} = \mu + \alpha_j + \beta(x_{ij} - \overline{x}) + \varepsilon_{ij}$$

| | Estimate | Std. Error | t-value | Pr(>\|t\|) |
|---|---|---|---|---|
| Intercept | 35.294 | 2.033 | 17.357 | 0.000 |
| race black ($\alpha_1$) | -5.605 | 3.012 | -1.861 | 0.067 |
| race Hispanic ($\alpha_2$) | 0.336 | 3.156 | 0.106 | 0.916 |
| educCentred ($\beta$) | 4.432 | 0.619 | 7.158 | 0.000 |

# ANCOVA model with interaction



# ANCOVA model with interaction

$$Y_{ij} = \mu + \alpha_j + \beta(x_{ij} - \overline{x}) + \gamma_j(x_{ij} - \overline{x}) + \varepsilon_{ij}$$

ANCOVA Table

| | Df | Sum Sq | Mean Sq | F-value | Pr(>F) |
|---|---|---|---|---|---|
| race | 2 | 3352.470 | 1676.235 | 7.099 | 0.002 |
| educCentred | 1 | 12245.232 | 12245.232 | 51.862 | 0.000 |
| race:educCentred | 2 | 691.837 | 345.918 | 1.465 | 0.238 |
| Residuals | 74 | 17472.412 | 236.114 | | |

The data does not provide evidence for an interaction between race and education

# Logistic regression: Data

The data set `admission.csv` contains information on the admission of 400 students into a business school. The variables in the data set are:

- *admit*: binary variable taking value 1 if the student was admitted into the business school and 0 otherwise
- *gpa*: grade point average in the undergraduate institution (range 1 – 6)
- *gre*: graduate record examination score obtained in the undergraduate institution (range 0 – 1000)
- *rank*: prestige of the undergraduate institution. The variable takes on the values 1 (highest prestige) through 4 (lowest prestige).

Test the association between admit and all the other variables

# Logistic regression model

$$\log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_{\text{gre}}X_{\text{gre}} + \beta_{\text{gpa}}X_{\text{gpa}} + \beta_{\text{r2}}D_{\text{r2}} + \beta_{\text{r3}}D_{\text{r3}} + \beta_{\text{r4}}D_{\text{r4}} \quad ,$$

with $D_r$ the dummy variables for rank with reference category highest prestige (1).

# Grouped data

The titanic.csv data set

| Economic status | Age group | Gender | Survived | Died | Total |
|---|---|---|---|---|---|
| Crew | A | W | 20 | 3 | 23 |
| Crew | A | M | 192 | 670 | 862 |
| 1st | A | W | 140 | 4 | 144 |
| 1st | A | M | 57 | 118 | 175 |
| 2nd | A | W | 80 | 13 | 93 |
| 2nd | A | M | 14 | 154 | 168 |
| 3rd | A | W | 76 | 89 | 165 |
| 3rd | A | M | 75 | 387 | 462 |
| 1st | C | W | 1 | 0 | 1 |
| 1st | C | M | 5 | 0 | 5 |
| 2nd | C | W | 13 | 0 | 13 |
| 2nd | C | M | 11 | 0 | 11 |
| 3rd | C | W | 14 | 17 | 31 |
| 3rd | C | M | 13 | 35 | 48 |
| Total | | | 711 | 1490 | 2201 |