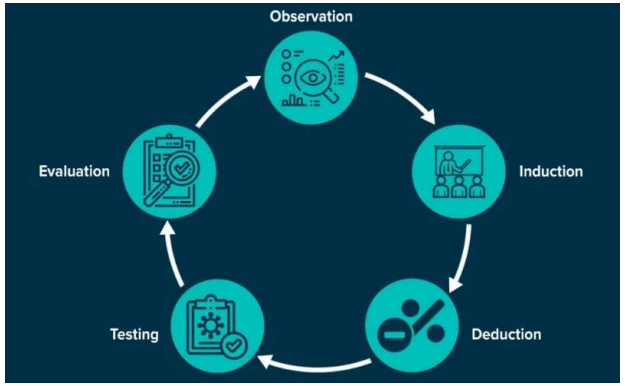
Empirical research in management and economics

Logistic regression



Thorsten Pachur

Technical University of Munich
School of Management
Chair of Behavioral Research Methods

Brief mid-term survey

https://forms.gle/bt71nwC1Q2r247hE6

(Please participate at the survey irrespective of whether you are attending in person or watching the recording.)

Exam

Please register on TUMonline for the exam!

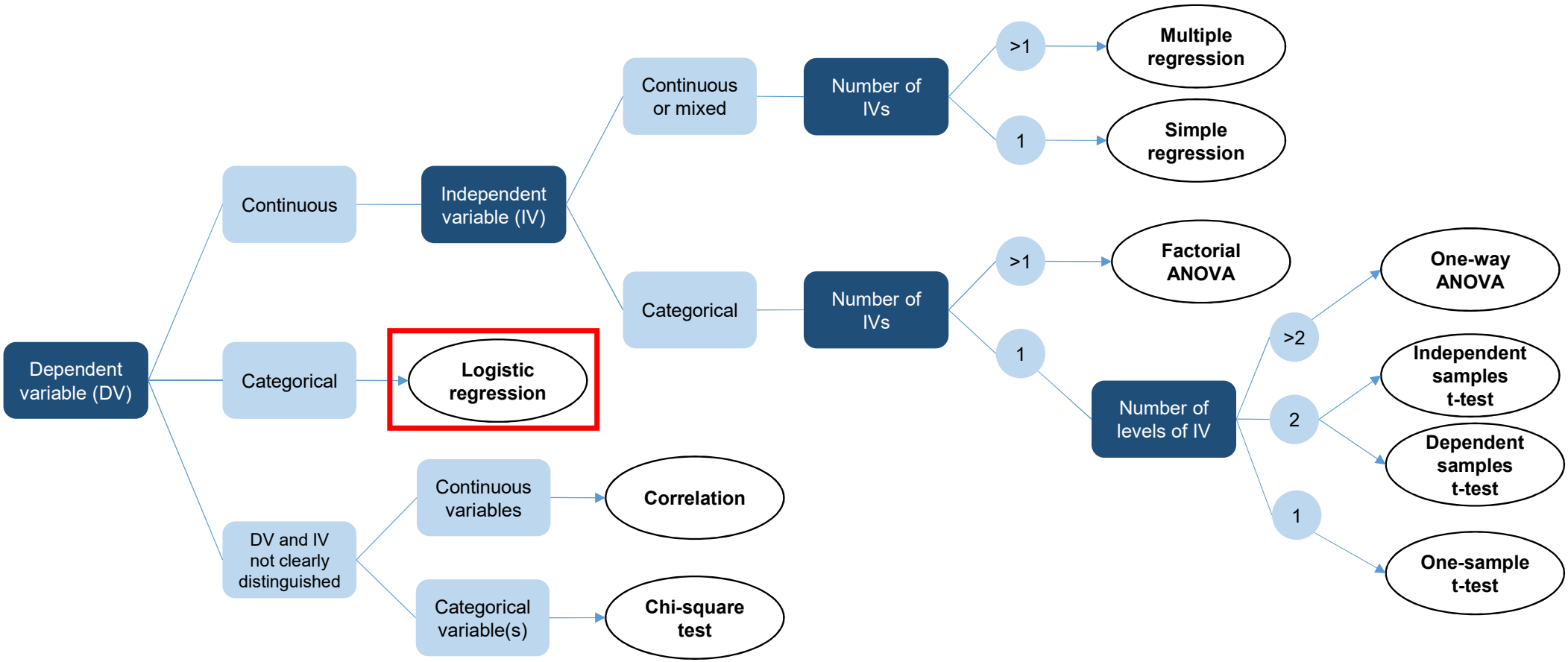
Deadline: 15 January 2026

Recap from last week

- Why can the regression coefficient for a predictor in a multiple regression differ from the regression coefficient in a simple regression?
- How can a multiple regression analysis be evaluated statistically (overall model fit, regression coefficients)?
- What is multicollinearity, when is it a problem, and how can one test for it?
- What aspects are relevant for a power analysis for multiple regression?
- How do you test whether a variable acts as a moderator in the relationship between two other variables?
- What are the steps of a mediation analysis?
- What is dummy coding and how it is used to include a categorical predictor in a regression model?

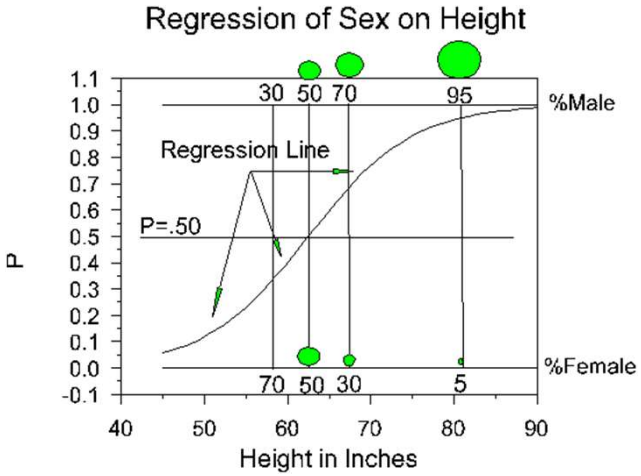
Agenda for the semester

Session	Date	Topic
1	13 October	Introduction
2	20 October	Descriptive data analysis
3	27 October	Hypothesis development and measurement
4	3 November	Inferential data analysis I
5	10 November	Inferential data analysis II
6	17 November	Simple regression
7	24 November	Multiple regression
8	**1 December**	**Logistic regression**
9	8 December	Factor analysis
10	15 December	Cluster analysis
11	12 January	Conjoint analysis
12	19 January	The replication crisis and open science
13	26 January	Summary and questions
	11 February	Exam

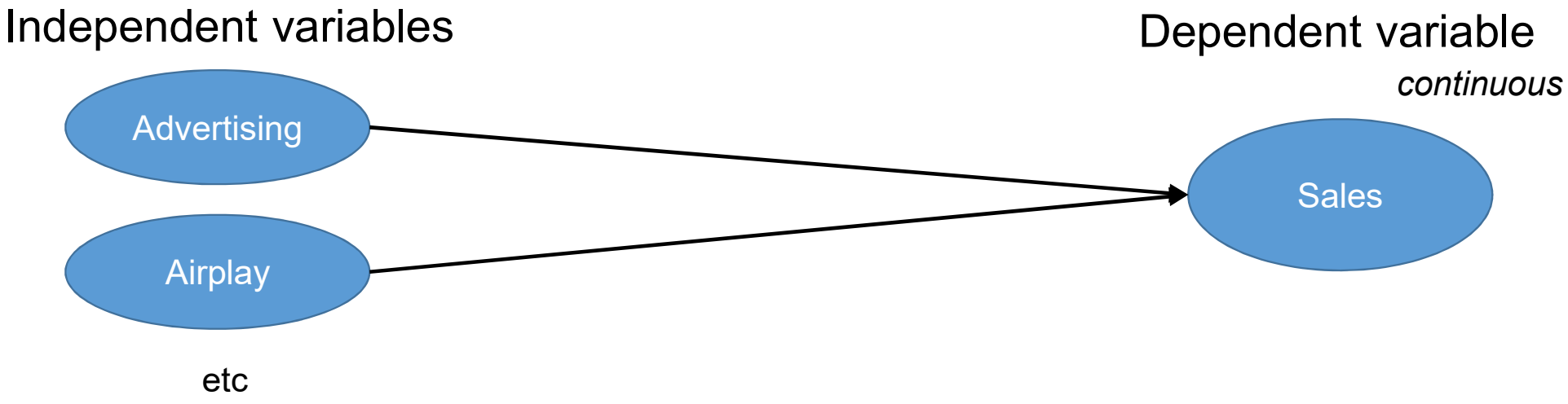

Dependent variable (DV)
Continuous
Independent variable (IV)
Continuous or mixed
Number of IVs
>1
Multiple regression
1
Simple regression
Categorical
Number of IVs
>1
Factorial ANOVA
1
Number of levels of IV
>2
One-way ANOVA
2
Independent samples t-test
Dependent samples t-test
1
One-sample t-test
Categorical
Logistic regression
DV and IV not clearly distinguished
Continuous variables
Correlation
Categorical variable(s)
Chi-square test

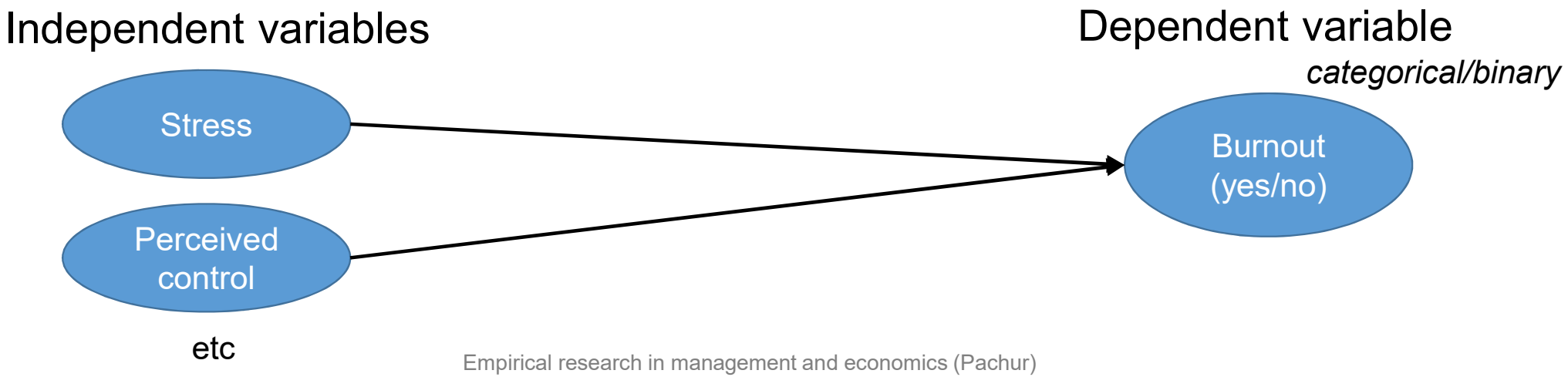
Goals for this week

- You know when to use logistic rather than linear regression
- You know the structure of the logistic regression equation
- You know the link between probability, odds, and log odds
- You understand how to interpret estimated regression coefficients in a logistic regression model
- You know how to statistically evaluate a logistic regression model
- You know the assumptions underlying logistic regression
- You know guidelines for sample-size considerations for logistic regression

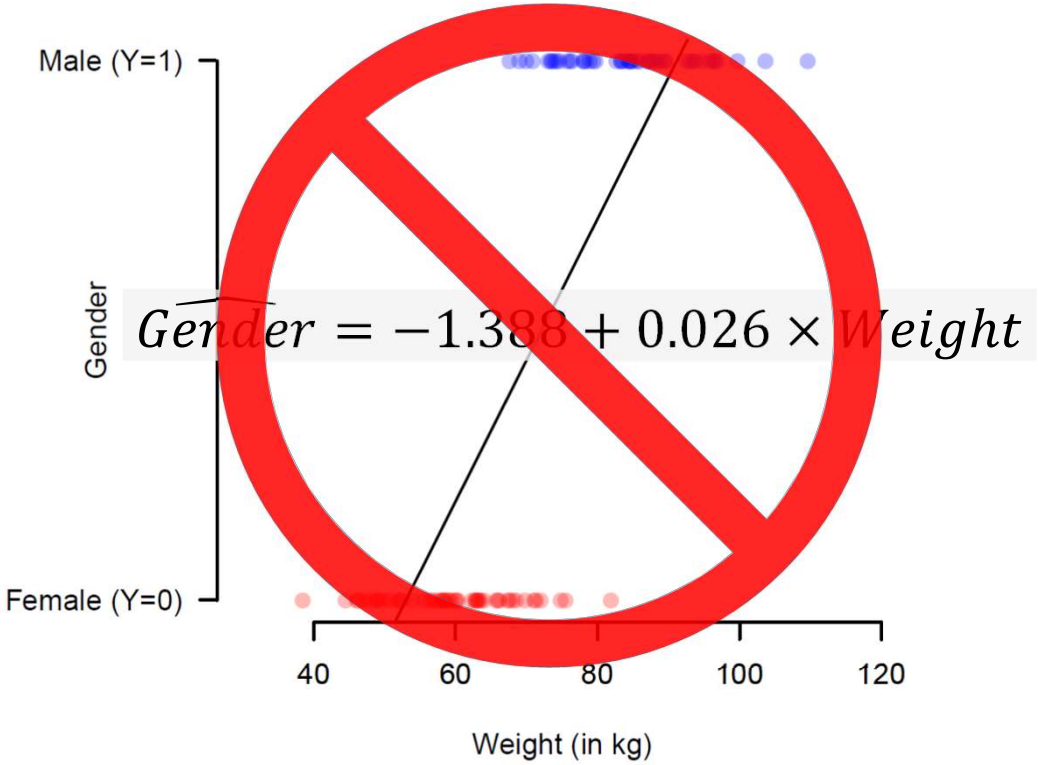


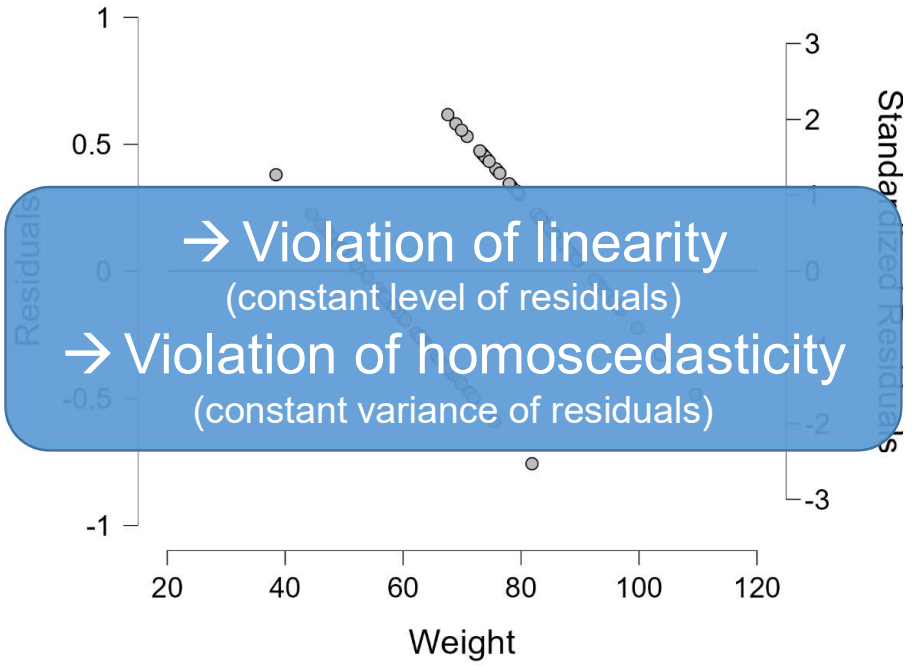
Logistic regression


Independent variables
Advertising
Airplay
etc
Dependent variable
continuous
Sales


Independent variables
Stress
Perceived control
etc
Dependent variable
categorical/binary
Burnout (yes/no)

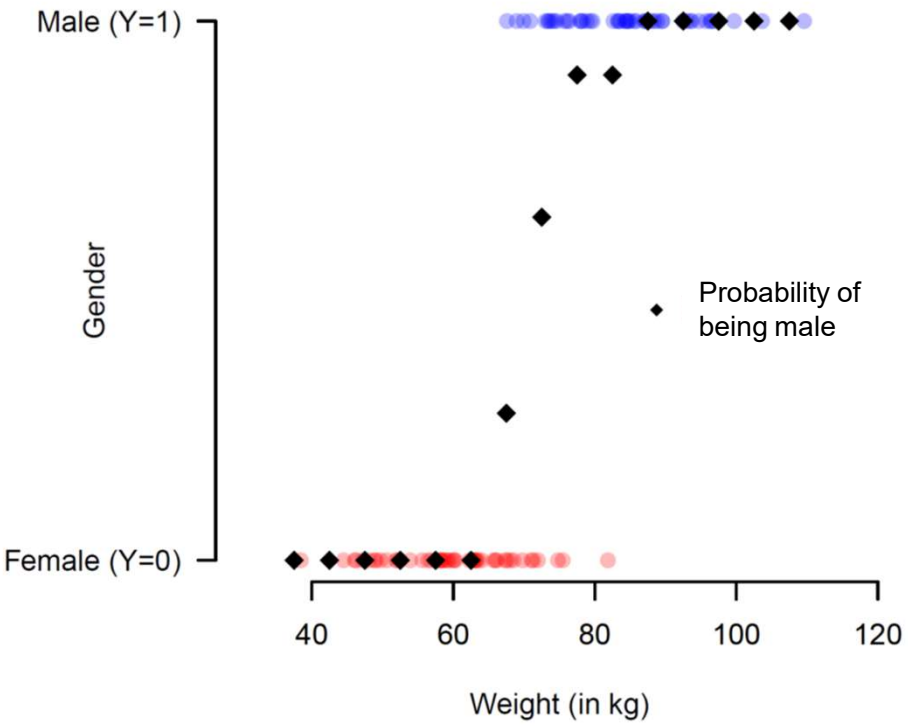
Predicting gender from weight

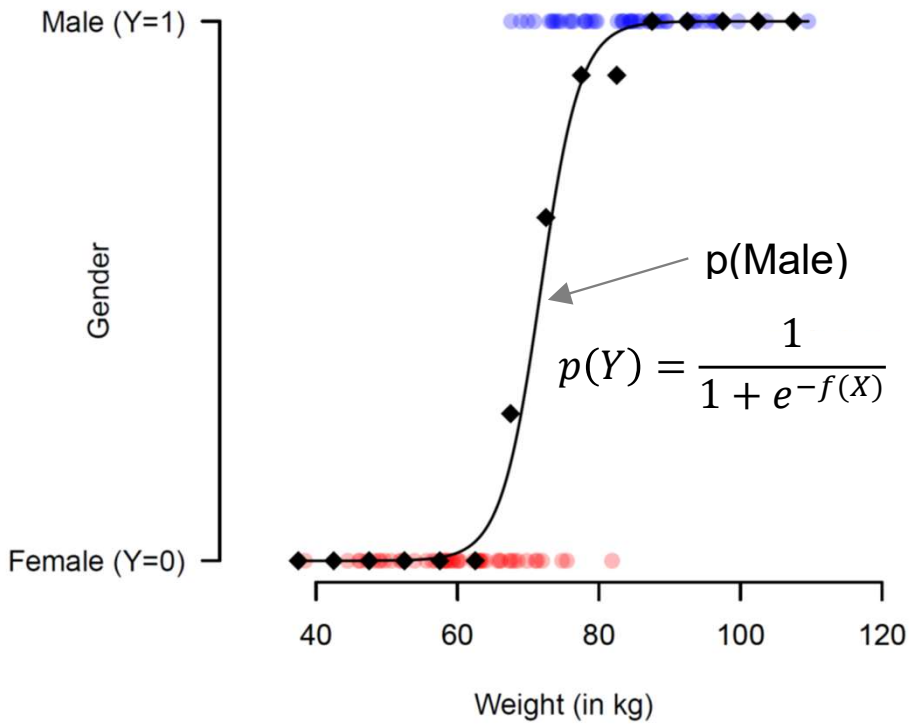




→ Violation of linearity
(constant level of residuals)
→ Violation of homoscedasticity
(constant variance of residuals)

Predicting gender from weight





Logistic function

Predicted probability of Y ($\epsilon[0,1]$)

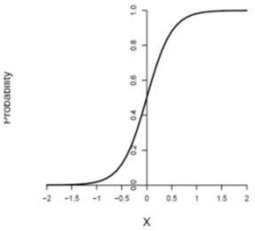
$$\widehat{p(Y)} = \frac{1}{1 + e^{-(b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n)}}$$

Regression equation

Logistic function

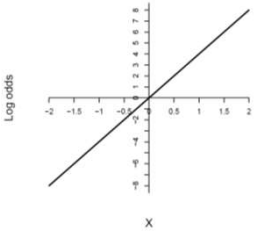
Probability

$$\widehat{p(Y)} = \frac{1}{1 + e^{-(b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n)}}$$

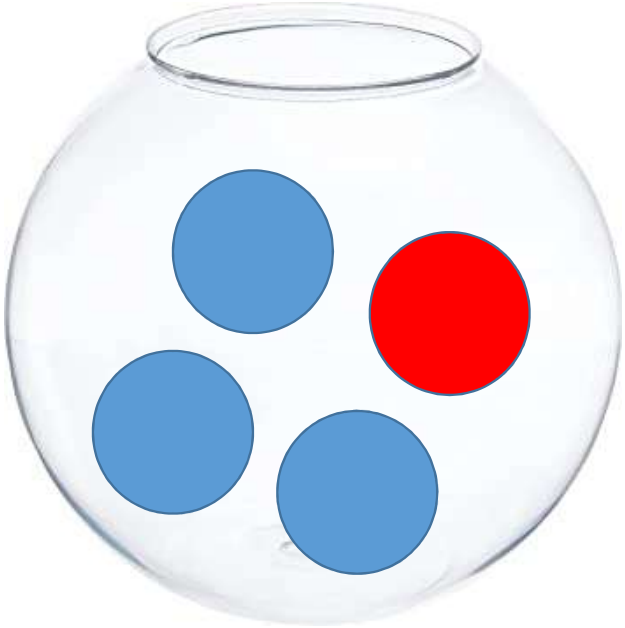


Log odds

$$\log\left(\frac{\widehat{p(Y)}}{(1 - p(Y))}\right) = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n$$



Probability, odds, and log odds



Probability $p(red) = \frac{1}{4} = .25$ $p(blue) = \frac{3}{4} = .75$

Odds $\frac{p(red)}{(1 - p(red))} = \frac{p(red)}{p(blue)} = \frac{.25}{.75} = 1:3 = .33$

From odds to probability: $p(red) = \frac{odds}{1 + odds} = \frac{.33}{1 + .33} = .25$

Log odds $\log\left(\frac{p(red)}{(1 - p(red))}\right) = \log\left(\frac{.25}{1 - .25}\right) = -1.10$

From log odds to odds: $e^{log\ odds} = odds$ $e^{-1.10} = .33$

Probability, odds, and log odds

Probability $p(red) = \frac{1}{2} = .5$ $p(blue) = \frac{1}{2} = .5$

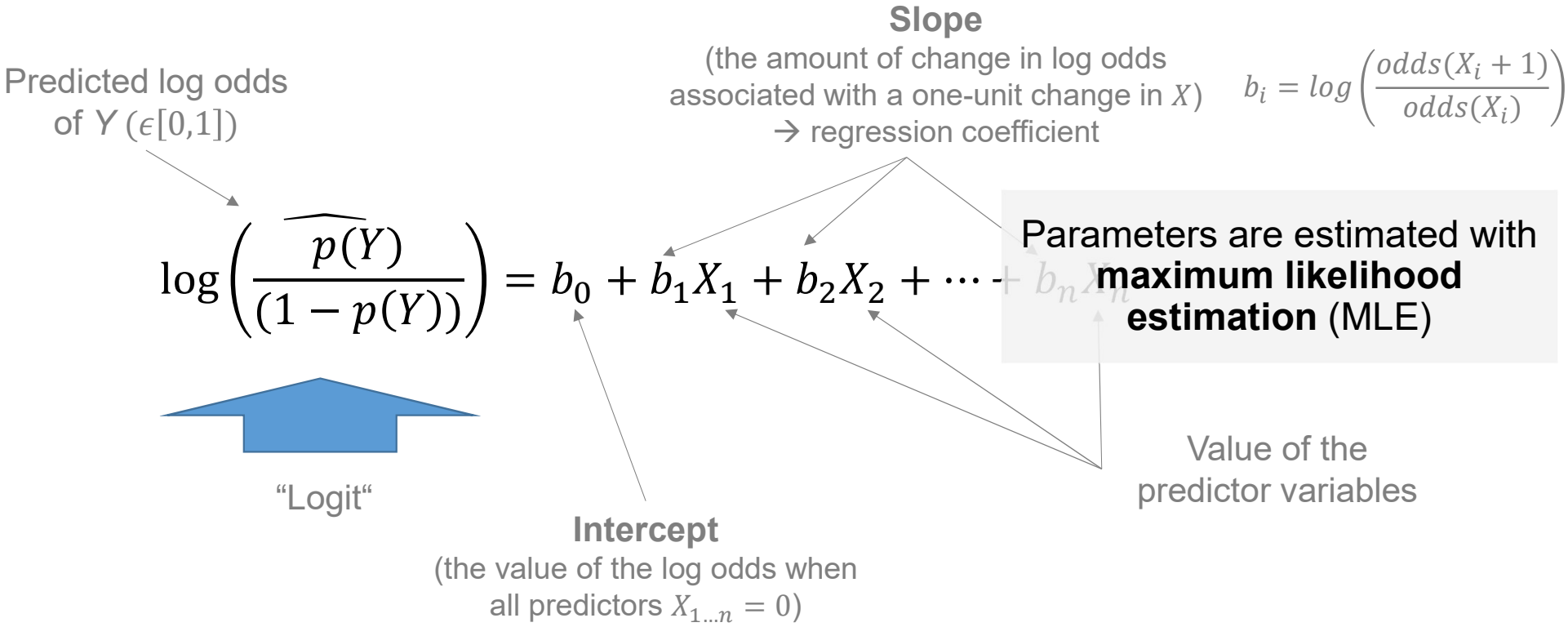
Odds $\frac{p(red)}{(1 - p(red))} = \frac{p(red)}{p(blue)} = \frac{.5}{.5} = 1{:}1 = 1$

From odds to probability: $p(red) = \frac{odds}{1 + odds} = \frac{1}{1 + 1} = .5$

Log odds $\log\left(\frac{p(red)}{(1 - p(red))}\right) = \log\left(\frac{.5}{1 - .5}\right) = 0$

From log odds to odds: $e^{log\ odds} = odds$ $e^0 = 1$

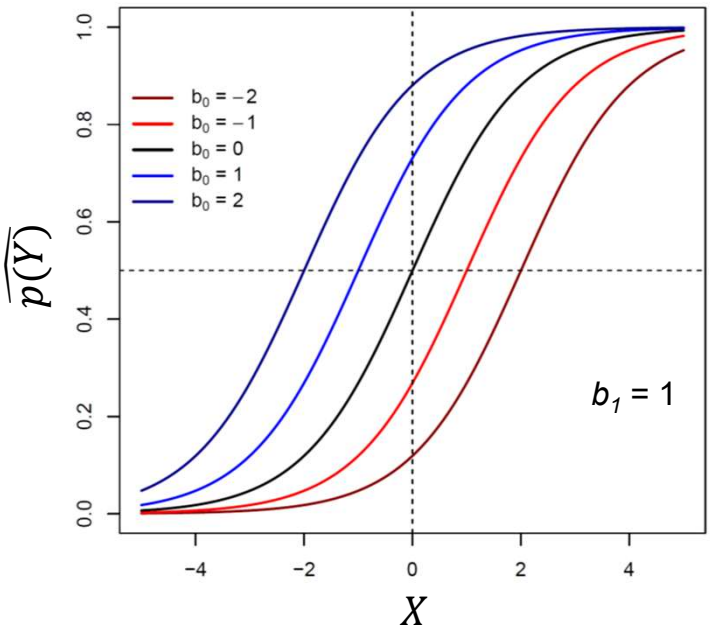
Regression equation



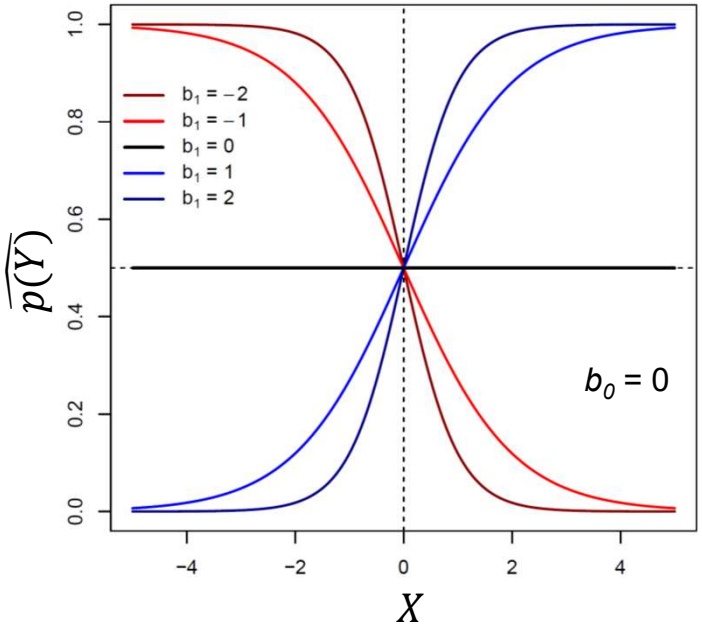
Logistic function

How do the intercept (b_0) and slope parameters (b_1) affect the logistic regression function?

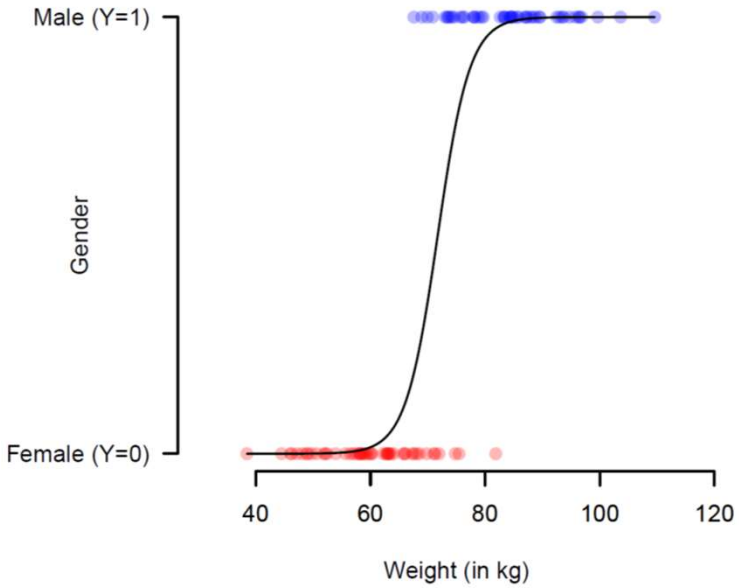
Intercept



Slope



Predicting gender from weight: Interpretation of the intercept



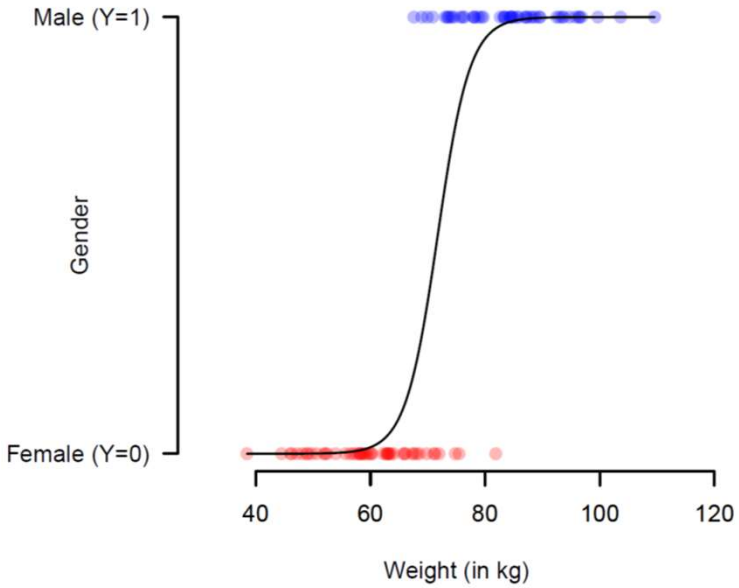
$$\log\left(\frac{\widehat{p(Male)}}{p(Female)}\right) = b_0 + b_{Weight} \times Weight$$

$$\log\left(\frac{\widehat{p(Male)}}{p(Female)}\right) = -25.65 + 0.357 \times Weight$$

$e^{-2\ .65} = 0.000000000007$

For a weightless person, the **odds** of being male (versus female) are extremely low

Predicting gender from weight: Interpretation of the intercept



$$\log\left(\frac{\widehat{p(Male)}}{p(Female)}\right) = b_0 + b_{Weight} \times Weight$$

$$\log\left(\frac{\widehat{p(Male)}}{p(Female)}\right) = 0.112 + 0.357 \times Weight_{centered}$$

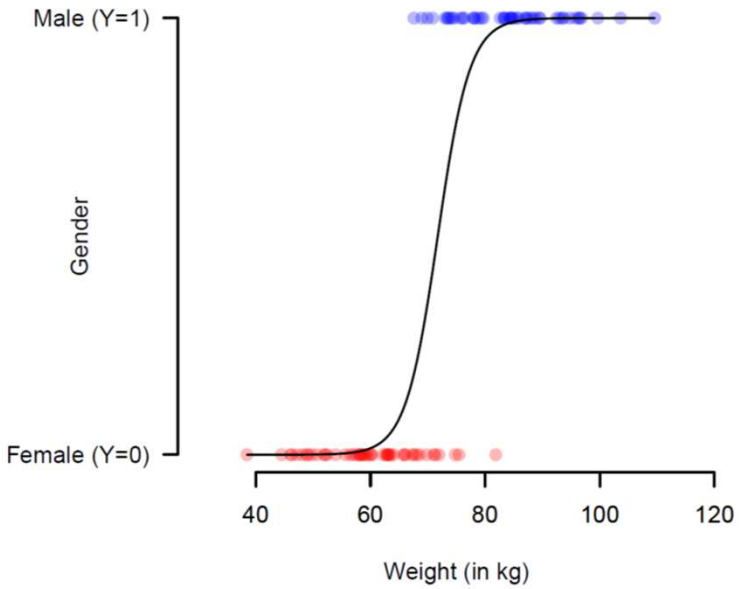
$e^{0.112} = 1.118$

For a person with the average weight in the sample, the **odds** of being male (versus female) are 1.118 to 1

$\frac{1.118}{1 + 1.118} = .528$

For a person with the average weight in the sample, the **probability** of being male is 52.8%

Predicting gender from weight: Interpretation of the slope



$$\log\left(\frac{\widehat{p(Male)}}{p(Female)}\right) = b_0 + b_{Weight} \times Weight$$

$$b_i = log\left(\frac{odds(X_i + 1)}{odds(X_i)}\right)$$

$$e^{0.357} = 1.43$$

For each additional kilogram of weight, the odds of being male increase by (OR-1)*100 = 43% (OR = 1.43).

Odds ratio (OR)

$$\log\left(\frac{\widehat{p(Male)}}{p(Female)}\right) = -25.65 + 0.357 \times Weight$$

$$\log\left(\frac{\widehat{p(Male)}_{X=72}}{p(Female)_{X=72}}\right) = -25.65 + 0.357 \times 72 = 0.054$$

$$\frac{\widehat{p(Male)}_{X=72}}{p(Female)_{X=72}} = e^{0.054} = 1.055$$

$$\log\left(\frac{\widehat{p(Male)}_{X=73}}{p(Female)_{X=73}}\right) = -25.65 + 0.357 \times 73 = 0.411$$

$$\frac{\widehat{p(Male)}_{X=73}}{p(Female)_{X=73}} = e^{0.411} = 1.508$$

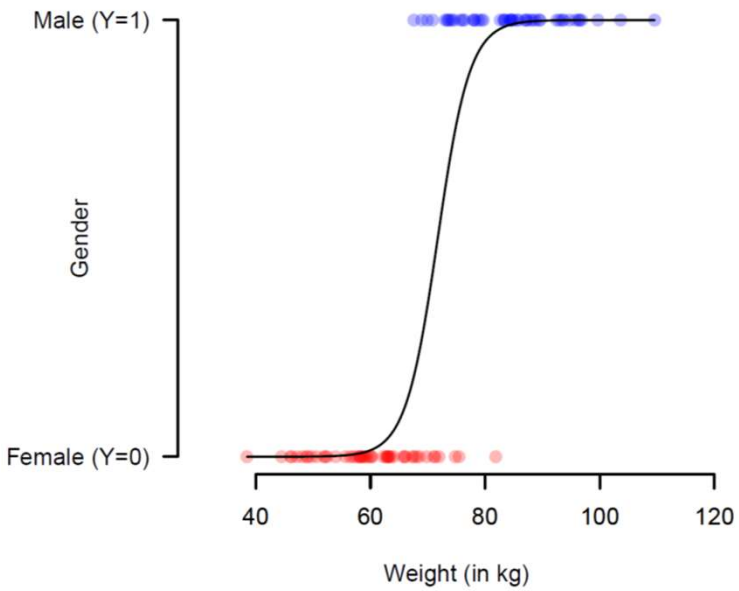
To probability from odds
(p = odds/(1+odds))

$$\widehat{p(Male)}_{X=72} = \frac{1.055}{1 + 1.055} = .513$$

1.508/1.055 = 1.43

$$\widehat{p(Male)}_{X=73} = \frac{1.508}{1 + 1.508} = .601$$

Predicting gender from weight: Interpretation of the slope



$$\log\left(\frac{\widehat{p(Male)}}{p(Female)}\right) = b_0 + b_{Weight} \times Weight$$

$$b_i = log\left(\frac{odds(X_i + 1)}{odds(X_i)}\right)$$

$$e^{0.357} = 1.43$$

For each additional kilogram of weight, the odds of being male increase by (OR-1)*100 = 43% (OR = 1.43).

Odds ratio (OR)

$$\log\left(\frac{\widehat{p(Male)}}{p(Female)}\right) = -25.65 + 0.357 \times Weight$$

$$\log\left(\frac{\widehat{p(Male)}_{X=66}}{p(Female)_{X=66}}\right) = -25.65 + 0.357 \times 66 = -2.088$$

$$\frac{\widehat{p(Male)}_{X=66}}{p(Female)_{X=66}} = e^{-2.088} = 0.123$$

$$\log\left(\frac{\widehat{p(Male)}_{X=67}}{p(Female)_{X=67}}\right) = -25.65 + 0.357 \times 67 = -1.731$$

$$\frac{\widehat{p(Male)}_{X=67}}{p(Female)_{X=67}} = e^{-1.731} = .177$$

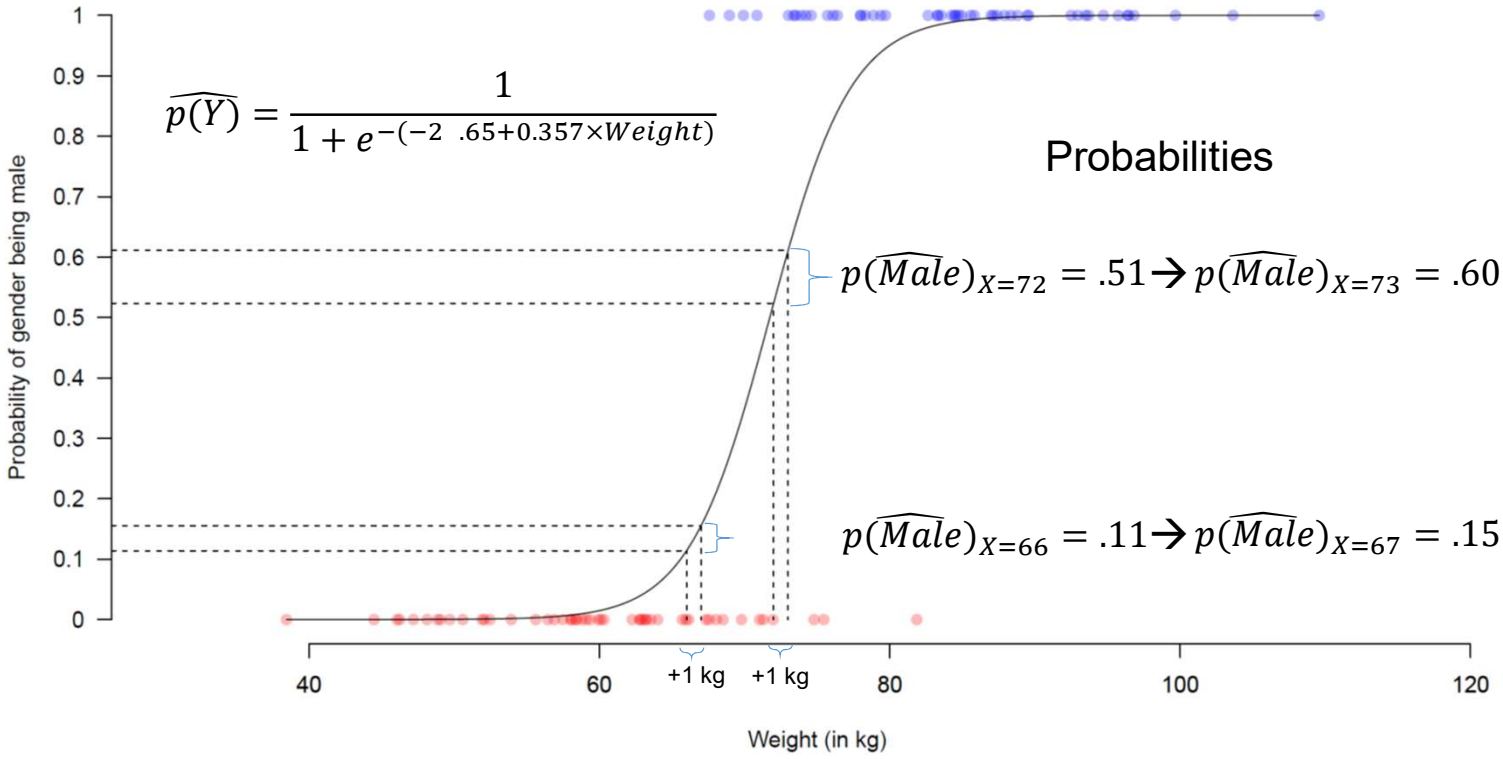
To probability from odds
(p = odds/(1+odds))

$$\widehat{p(Male)}_{X=66} = \frac{0.123}{1 + 0.123} = .11$$

177 /.123 = 1.43

$$\widehat{p(Male)}_{X=67} = \frac{.177}{1 + .177} = .15$$

Predicting gender from weight: Interpretation of the slope



Probabilities

$p(\widehat{Male})_{X=72} = .51 \rightarrow p(\widehat{Male})_{X=73} = .60$

$p(\widehat{Male})_{X=66} = .11 \rightarrow p(\widehat{Male})_{X=67} = .15$

Odds ratios

1.508/1.055

.177 /.123

= 1.43

Surviving the Titanic disaster

$$\log\left(\frac{\widehat{p(survive)}}{p(die)}\right) = b_0 + b_{Age} \times \ldots + b_{Class} \times Class$$

$e^{-2.631} = 0.072$

Being **male** decreases the odds of surviving by (OR-1)*100 = 92.8%

$$\log\left(\frac{\widehat{p(survive)}}{p(die)}\right) = 3.761 - 0.039 \times Age - 2.631 \times Gender - 1.292 \times 2ndClass - 2.521 \times 3rdClass$$

male = 1; female = 0

$$\frac{\widehat{p(survive)}}{p(die)}_{\textbf{male},40\ years,2nd\ class} = e^{3.761-0.039\times40-2.631\times1-1.292\times1-2.521\times0} = .178$$

.178/2.48 = 0.072

$$\frac{\widehat{p(survive)}}{p(die)}_{\textbf{female},40\ years,2nd\ class} = e^{3.761-0.039\times40-2.631\times0-1.292\times1-2.521\times0} = 2.48$$

From odds to probability (p = odds/(1+odds))

$$\widehat{p(survive)}_{\textbf{male},40\ years,2nd\ class} = \frac{.178}{1+.178} = .15$$

$$\widehat{p(survive)}_{\textbf{female},40\ years,2nd\ class} = \frac{2.48}{1+2.48} = .71$$

British Board of Trade (1990)

Evaluating the size of a regression coefficient

- Odds ratio

  $$OR_i = e^{b_i}$$

  Note: The size of an odds ratio depends on the scaling of the predictor (e.g., kg vs. g). Odds ratios are often used as effect size measures, but are only comparable across predictors when predictors are z-standardized.

- Standardized regression coefficient

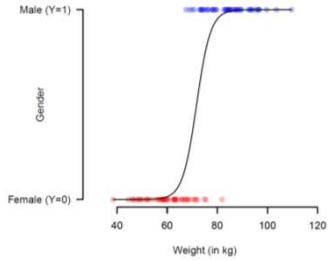
  $$\beta_i = b_i \times SD_{X_i}$$

  The standardized regression coefficient indicates the change in $\log\left(\frac{\widehat{p(Y)}}{(1-p(Y))}\right)$ for a change of predictor *X* by one standard deviation.
  It allows for a comparison of regression coefficients across continuous predictors (nominal predictors are not standardized).

Statistical evaluation: Regression coefficients

Wald statistic

$$\log\left(\frac{\widehat{p(Male)}}{p(Female)}\right) = -25.65 + 0.357 \times Weight$$



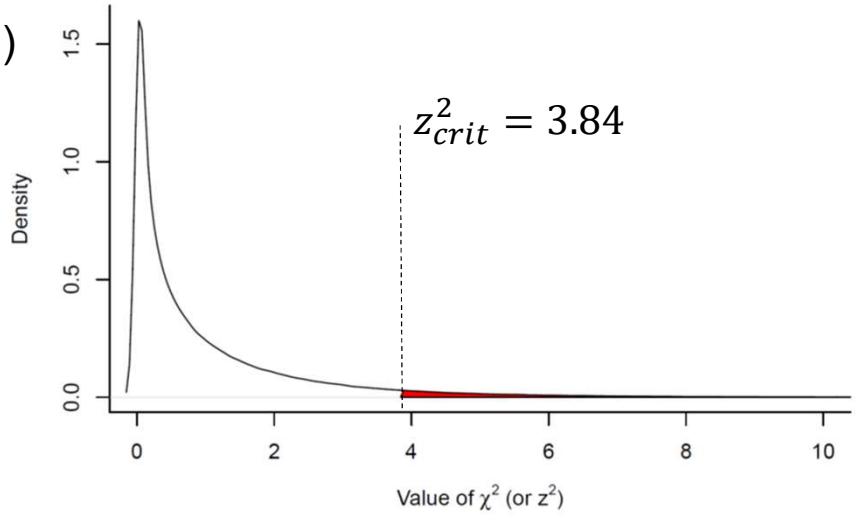
Abraham Wald

$$z^2 = \left(\frac{b}{SE_b}\right)^2 \rightarrow \chi^2\text{-distributed (with df = 1)}$$

$$SE_b = \frac{1}{\sqrt{\sum_{i=1}^{N} \hat{p}_i(1-\hat{p}_i)(X_i - \bar{X})^2}}$$

$$z^2 = \left(\frac{b}{SE_b}\right)^2 = \left(\frac{0.357}{0.084}\right)^2 = 18.1$$

$p < .0001$



Statistical evaluation: Goodness of fit

$$\text{log likelihood } (LL) = \sum_{i=1}^{N} \left[Y_i \times log\left(\widehat{p(Y_i)} \right) + (1 - Y_i) \times log\left(1 - \widehat{p(Y_i)}\right) \right]$$

$$deviance = -2 \times LL$$

$$\chi^2 = deviance_{baseline} - deviance_{model}$$

→ χ^2-distributed

df = k_{model} - $k_{baseline}$

$$\log\left(\frac{\widehat{p(Male)}}{p(Female)}\right)_i = b_0$$

$$\log\left(\frac{\widehat{p(Male)}}{p(Female)}\right)_i = b_0 + b_{Weight} \times Weight_i$$

$\chi^2 = 138.6 - 38.5 = 110.1$

df = 2-1 = 1 *p* < .0001

Statistical evaluation: Goodness of fit

Cox & Snell (1989)

$$R^2_{CS} = 1 - e^{\frac{-2LL_{model} - (-2LL_{baseline})}{n}}$$

$$R^2_{CS} = 1 - e^{\frac{38.5 - 13\ .6}{100}} \qquad n = 100$$

$$R^2_{CS} = 1 - 0.367 = 0.632$$

Nagelkerke (1991)

$$R^2_N = \frac{R^2_{CS}}{1 - e^{\frac{-(-2LL_{baseline})}{n}}}$$

$$R^2_N = \frac{0.632}{1 - e^{\frac{-(138.6)}{100}}}$$

$$R^2_N = \frac{0.632}{0.749} = 0.842$$

Statistical evaluation: Goodness of fit

Confusion matrix

Performance Diagnostics

Confusion matrix

Observed	Predicted		% Correct
	Female	Male	
Female	46	4	92.00
Male	4	46	92.00
Overall % Correct			92.00

Note. The cut-off value is set to 0.5

Statistical evaluation: Predictive accuracy

Information criteria that take **model complexity** into account

- Akaike Information Criterion (AIC)

$$AIC = -2LL + 2k$$

k: number of estimated parameters (including intercept)

- Bayesian Information Criterion (BIC)

$$BIC = -2LL + k \times \log(n)$$

n: number of observations

→ Lower values indicate a better trade-off between model fit and model complexity

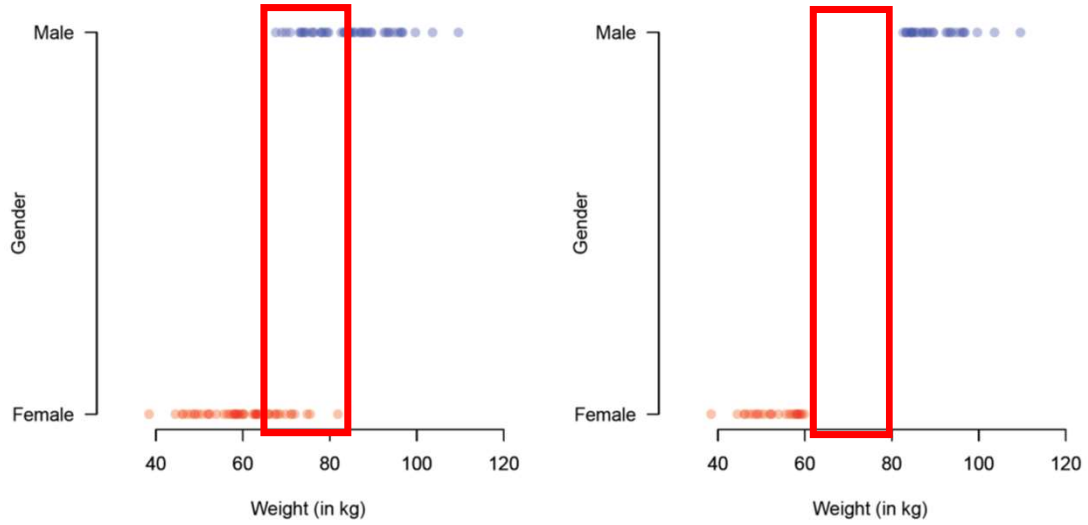
Assumptions and requirements in logistic regression

- Absense of multicollinearity
- Linearity (in log-odds space)

  → Box-Tidwell approach: Test whether the interaction between the predictor and its logarithm is nonsignificant

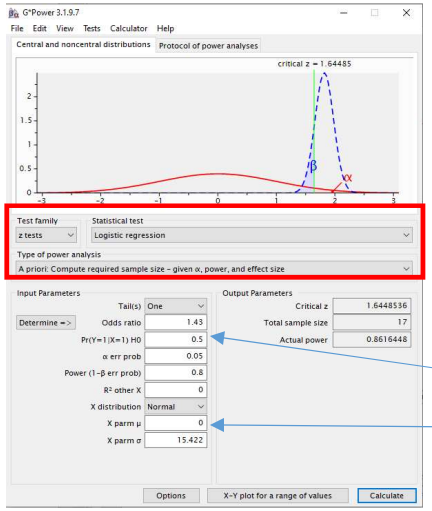
$$\log\left(\frac{\widehat{p(Y)}}{(1-p(Y))}\right) = b_0 + b_1 X_1 + b_2 X_1 \times \log(X_1)$$

- No complete separation



Sample-size considerations in logistic regression

- Power analysis with G*Power



Focus is on the test with a single (or the most important) predictor. The hypothesized effect size (odds ratio) refers to this predictor

Proportion of *Y* cases in sample

Assume that the predictor is centered

- At least 10 events per predictor ("events" are the cases in the less frequent category of the dependent variable)

  Example

  If you want to predict survival among passengers on the Titanic based on 4 predictors, you need at least 4 × 10 = 40 surviving or killed passengers, depending on which of the two is less frequent in your sample.

Self-quiz questions

- In logistic regression, what does the linear combination of intercept and predictors predict?
- How are probability, odds, and log odds related to each other?
- How do you get from an estimated slope (i.e., regression coefficient) of a predictor to the odds ratio?
- How do you test whether an estimated regression coefficient differs significantly from zero?
- How can you assess the performance of a logistic regression model—with and without taking model complexity into account?
- What are key assumptions in logistic regression?
- How can you plan the sample size for a logistic regression analysis?

Background reading for next week

Backhaus, K., Erichson, B., Gensler, S., Weiber, R., & Weiber, T. (2021). Factor analysis. In K. Backhaus, B. Erichson, S. Gensler, R. Weiber, & T. Weiber, *Multivariate analysis: An application-oriented introduction* (p. 381–450). Springer.