Logistic regression

SHP03

Hanne Oberman
h.i.oberman@uu.nl

Outline

- Today's lecture
- Introduction
- Theory
- Intermezzo
- Theory (continued)
- Case study
- Marginal effects
- Take-aways

Today's lecture

Topics

- Logistic regression
  - Probabilities, logits, and odds
  - Significance and relevance of effects
  - Marginal effects

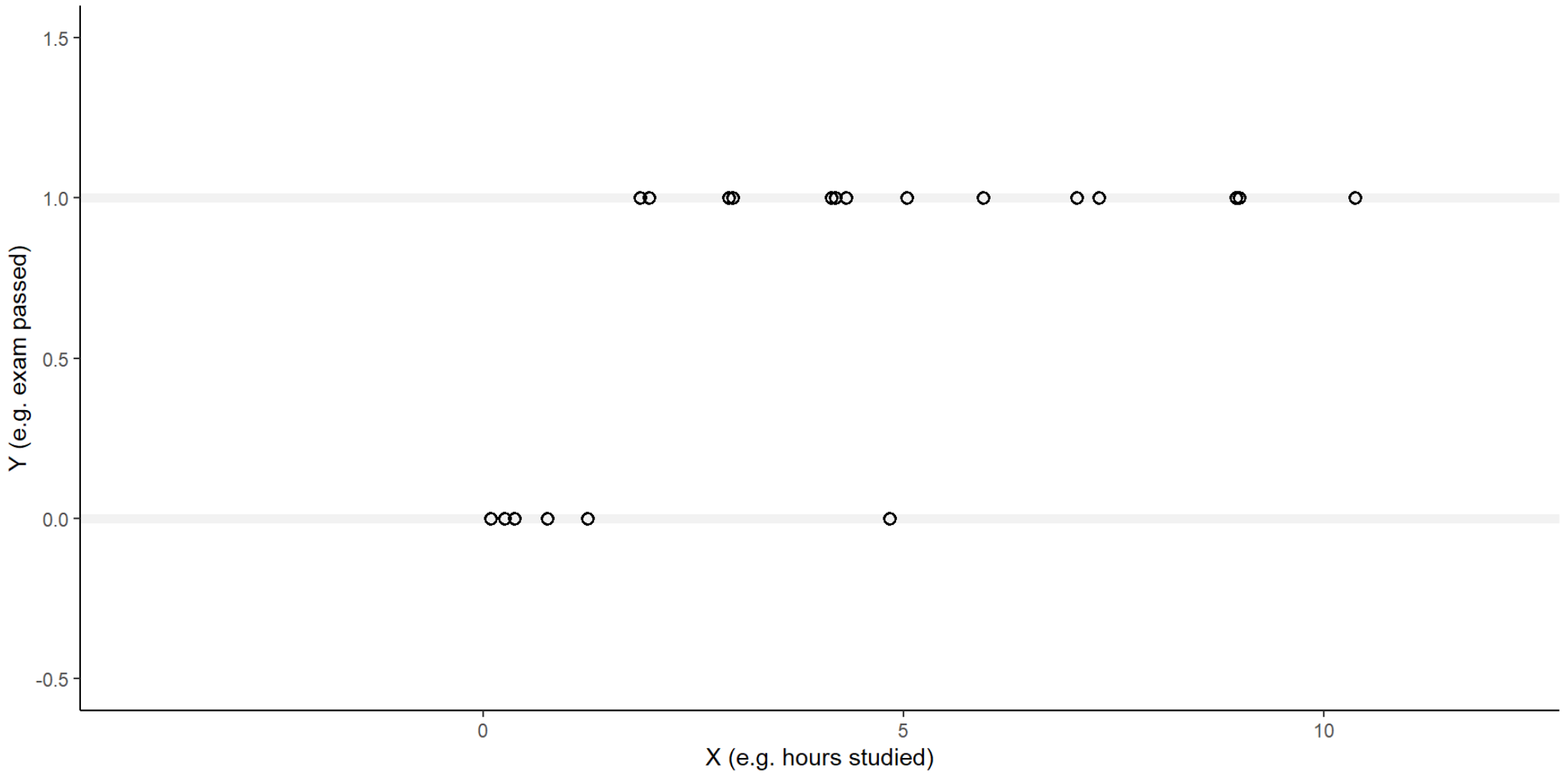
Introduction

When to use logistic regression

- Prediction of a binary dependent variable
- Binary dependent variables are categorical variables with two categories:
  - e.g. passing an exam ($Y = \text{failure}$ or $Y = \text{success}$),
  - or smoking status ($Y = \text{no}$ or $Y = \text{yes}$);
  - in general, $Y = 0$ or $Y = 1$, nothing observed in between

Binary outcome

- Code

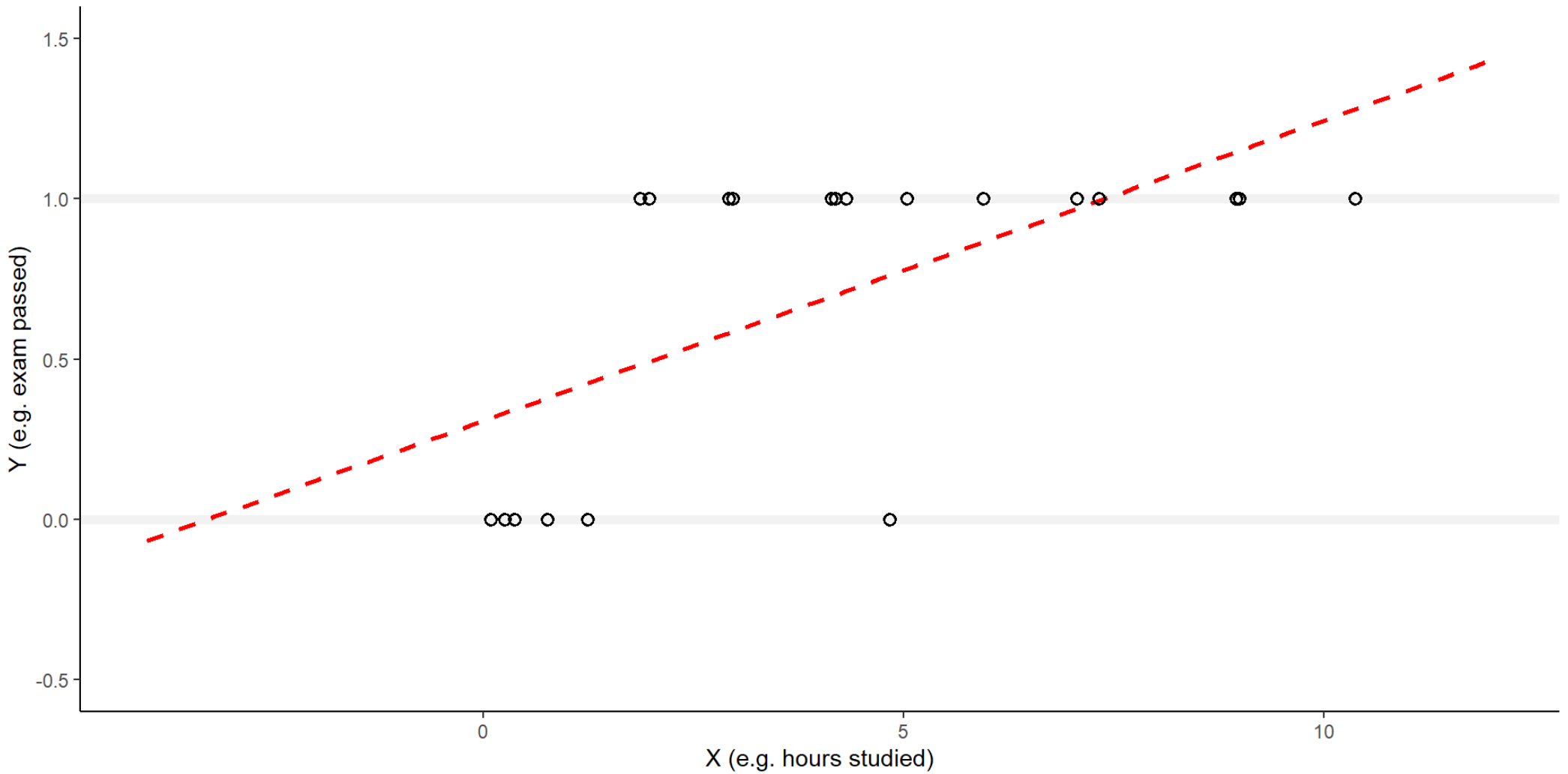


Why not use linear regression?

- Linear regression:
  - requires a continuous dependent variable
  - requires a linear relation between predictor(s) and outcome
  - predicts values outside the [0, 1] range

Linear regression

- Code

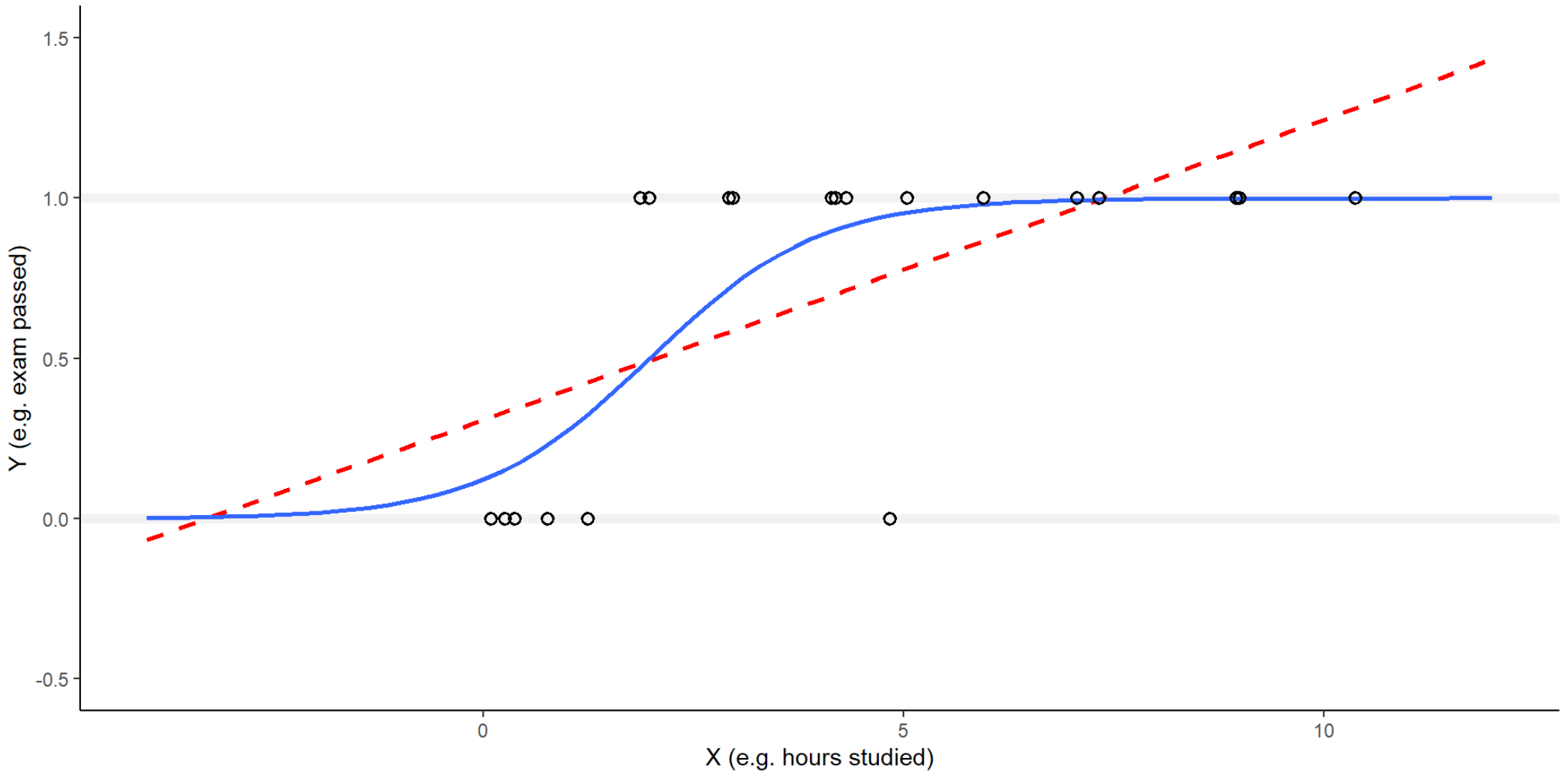


What to use instead?

- Logistic regression:
  - predicts probabilities (instead of raw outcome values)
  - models how the probability of success varies with the independent variables
  - converts continuous predictions to binary outcome with **logit link** function
  - ensures that the predicted probabilities are between 0 and 1

Logistic regression

- Code

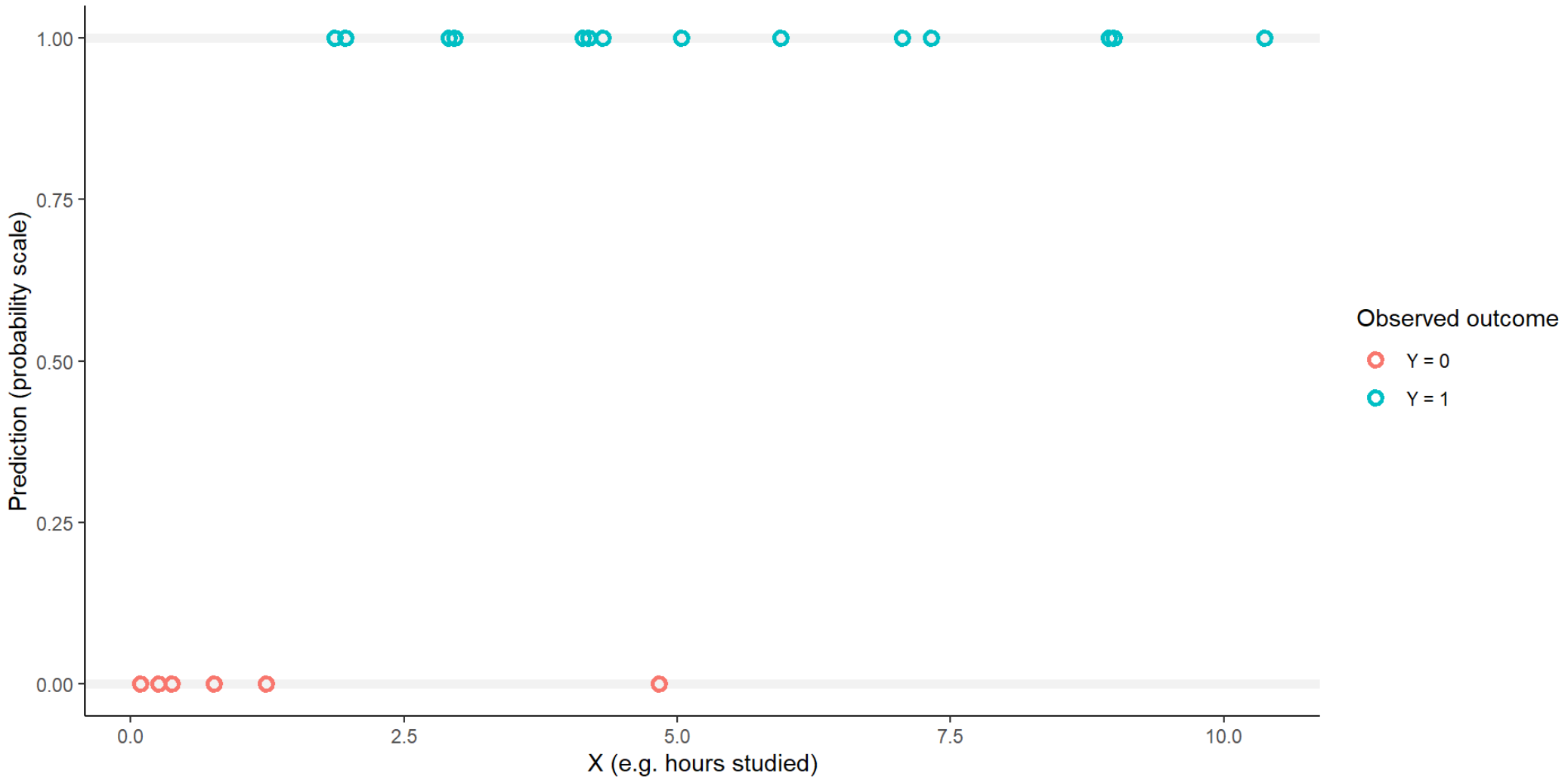


What is a logit link function?

- The logit link function maps the linear regression line to the probability scale
- The logit is the natural logarithm of the odds of the outcome being 1, where the odds are
  - a ratio of
  - the probability of the dependent variable being 1
  - divided by
  - the probability of the dependent variable being 0
- $\text{logit}(\text{Y}{=}1) = \log(\text{odds}(\text{Y}{=}1)) = \log(\frac{P(Y{=}1)}{P(Y{=}0)})$

Predicting a binary outcome

▸ Code



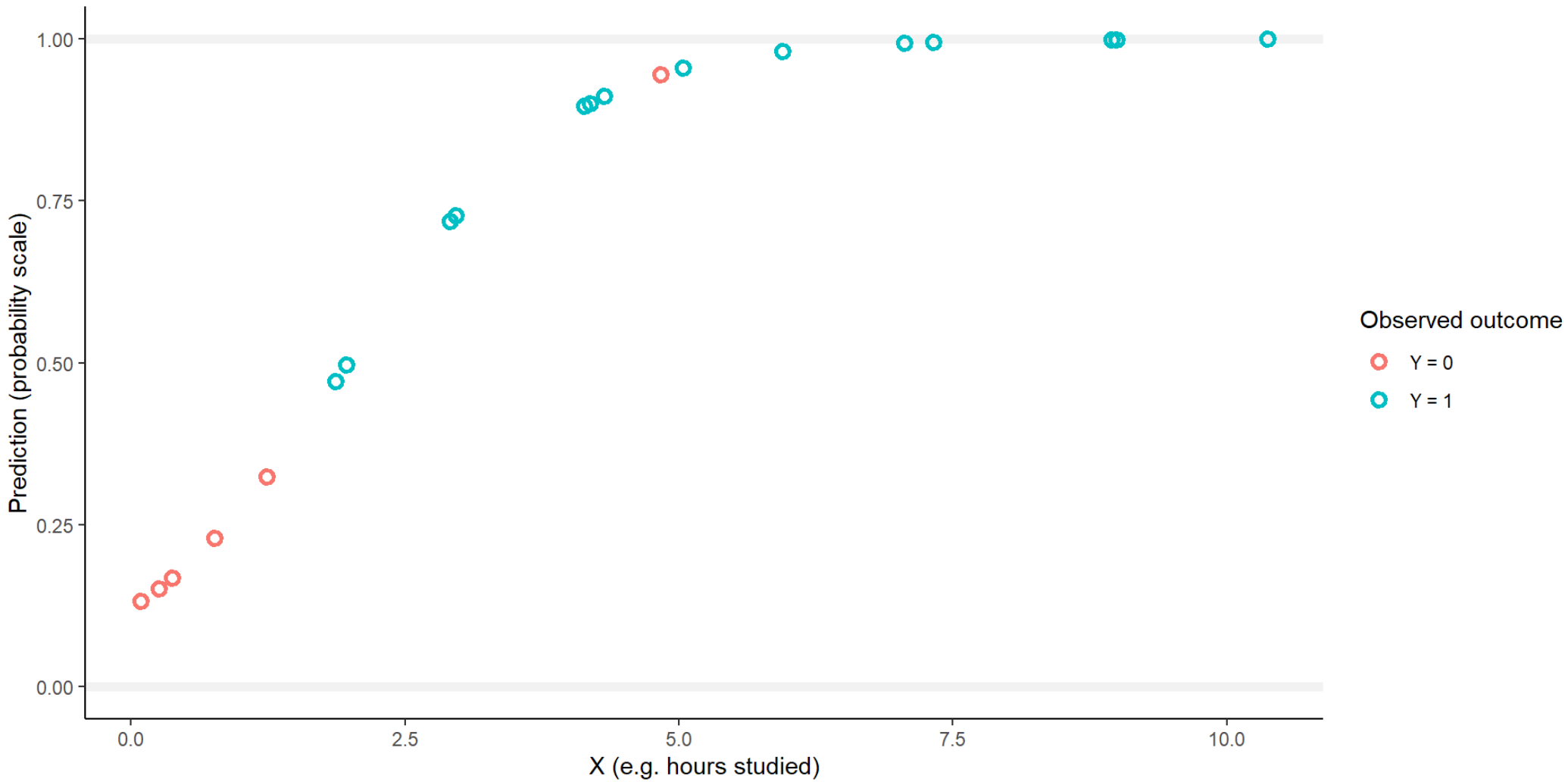
What are probabilities?

Probabilities:

- are values between 0 and 1, with midpoint 0.5
- are the likelihood of an event occurring, $P(Y = 1)$

Predicting probabilities

- Code



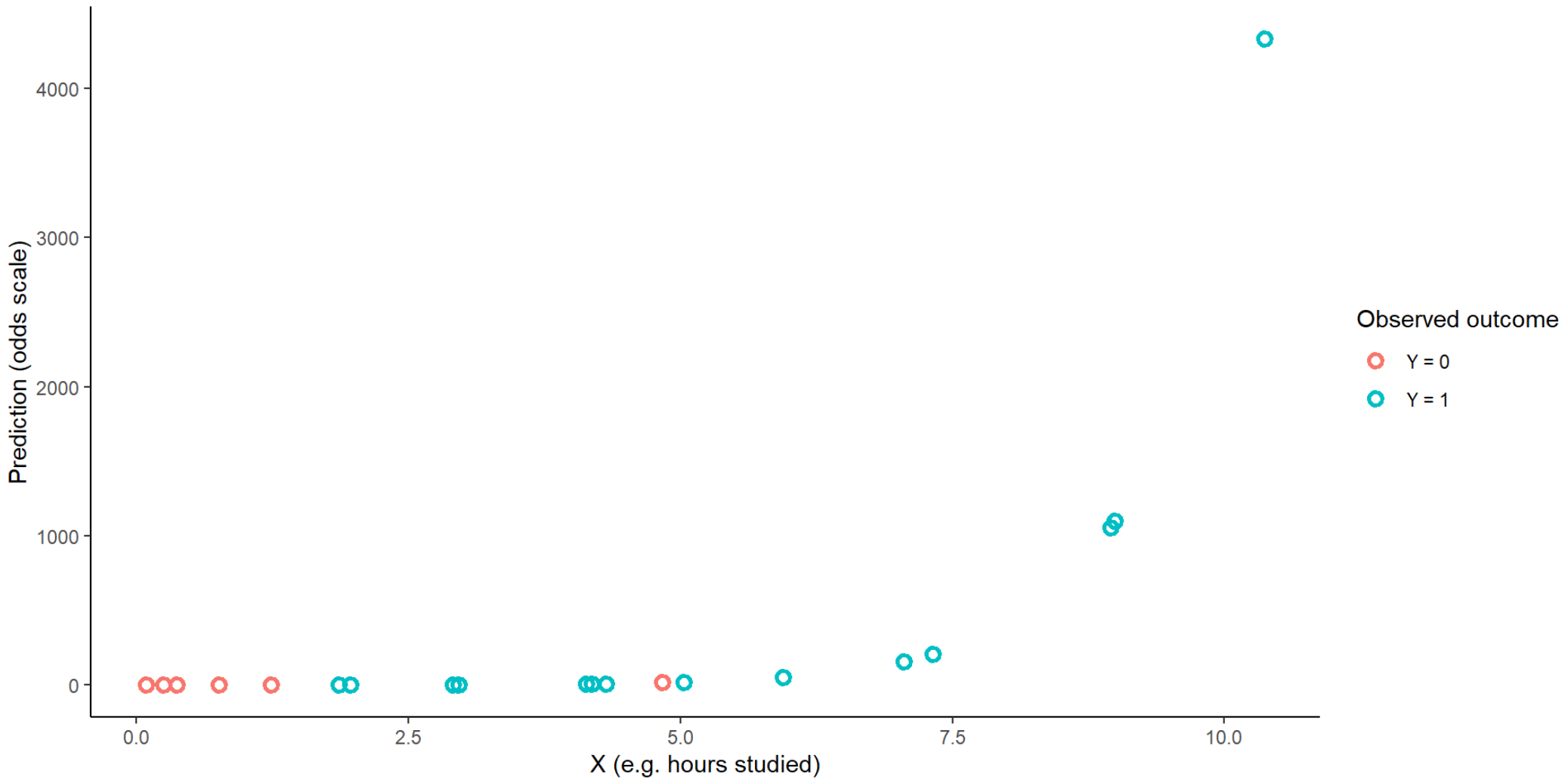
What are odds?

Odds:

- are values between 0 and $\inf$, with midpoint 1
- are a ratio of probabilities
- e.g. the probability of an event occurring relative to the probability of the event not occurring, $P(Y = 1) : P(Y = 0)$
- of a dependent variable being 1 is equal to
  - the probability of the dependent variable being 1
  - divided by
  - the probability of the dependent variable being 0
  - $\text{odds}(Y = 1) = \frac{P(Y=1)}{P(Y=0)}$

Odds scale

- Code



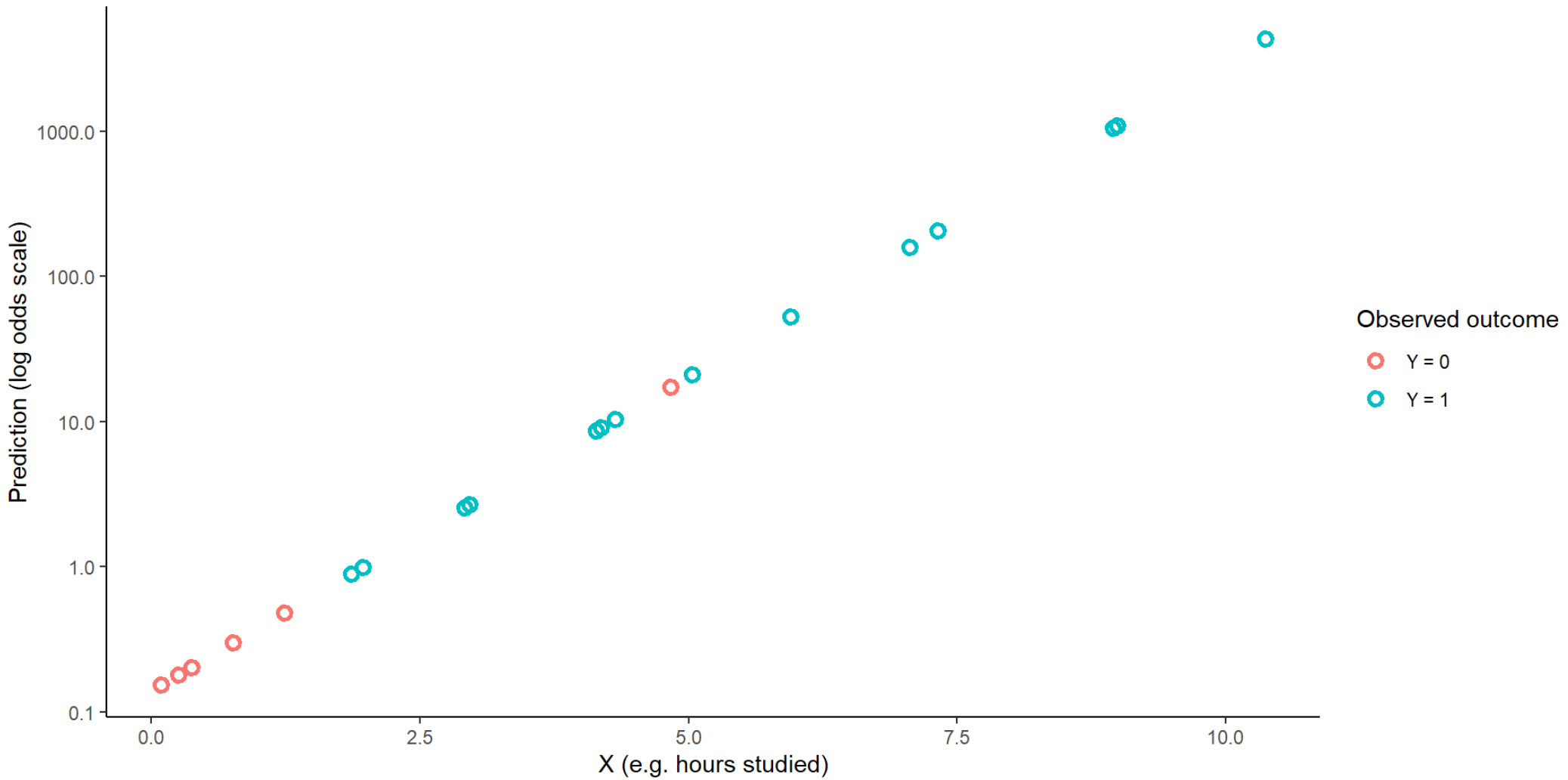
What are log odds?

Log odds:

- are odds transformed to a logarithmic scale
- mathematical ‘shortcut’

Odds scale (log transformed)

- Code



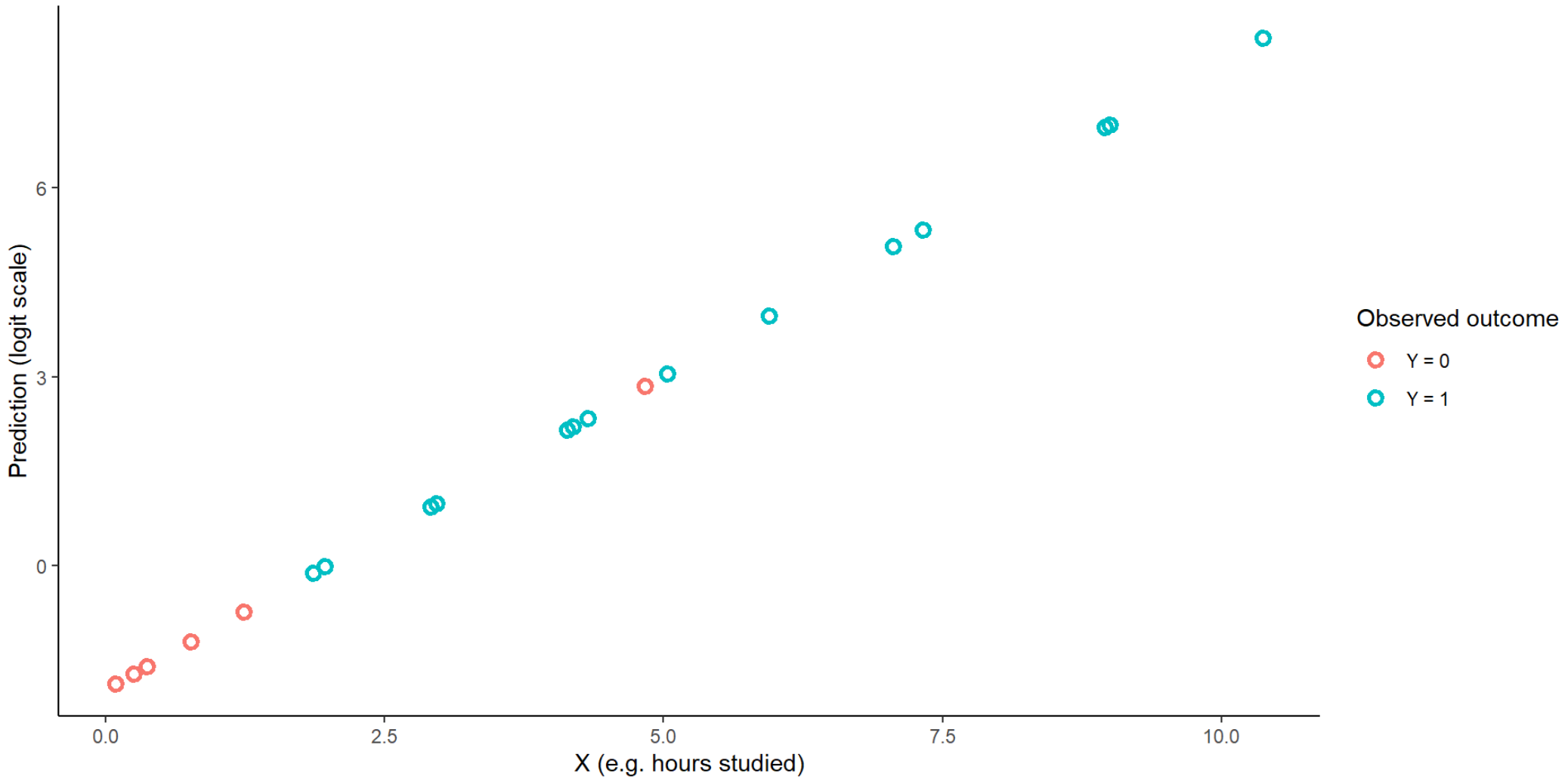
What are logits?

Logit:

- is the natural logarithm of the odds
- transforms the range from $[0, 1, \text{inf}]$ (odds) to $[-\text{inf}, 0, \text{inf}]$ (logit)
- is symmetric around 0

Logit scale

- Code



Theory

General(ized) linear model

$$\text{data} = \text{model} + \text{error}$$

$$Y \sim \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots$$

Estimation of regression coefficients

Linear regression:

$$y_i = b_0 + b_1 x_i$$

Logistic regression:

$$\text{logit}(\pi_i) = b_0 + b_1 x_i,$$

- where coefficients are estimated with respect to the logit scale, $\text{logit}(\pi_i) = \log(\text{odds}(P(y_i = 1))$

Logit model

Logistic regression is linear regression on the logit scale:

- $\text{logit}(\pi) = \log(\pi/(1-\pi)) = b_0 + b_1 X_1 + b_2 X_2 + \cdots$
- $\text{odds} = (\pi/(1-\pi)) = e^{(b_0 + b_1 X_1 + b_2 X_2 + \cdots)}$
- $\text{probability} = \pi = 1/(1 + e^{-(b_0 + b_1 X_1 + b_2 X_2 + \ldots)})$

Conversion

probability	odds	logit
π_i	$\frac{\pi_i}{1-\pi_i}$	$\log_e \frac{\pi_i}{1-\pi_i}$
0.05	.05/.95 = 0.05	-2.94
0.10	.10/.90 = 0.11	-2.20
0.30	.30/.70 = 0.43	-0.85
0.50	.50/.50 = 1	0.00
0.55	.55/.45 = 1.22	0.20
0.70	.70/.30 = 2.33	0.85
0.90	.90/.10 = 9	2.20
0.95	.95/.05 = 19	2.94

- **Probability:** between 0 and 1
- **Odds:** Not symmetric, always > 0
- **Logit:** Symmetric around 0, can take on any value

Modeling probability

Logistic regression models the probability of $Y = 1$ given a known value on X

$$\pi_i = P(Y = 1|X = x_i)$$

- To ensure that π_i stays between 0 and 1, the predictions are mapped onto the interval [0, 1] using the logit function
- To convert predictions on the logit scale back to probability, use the inverse logit function

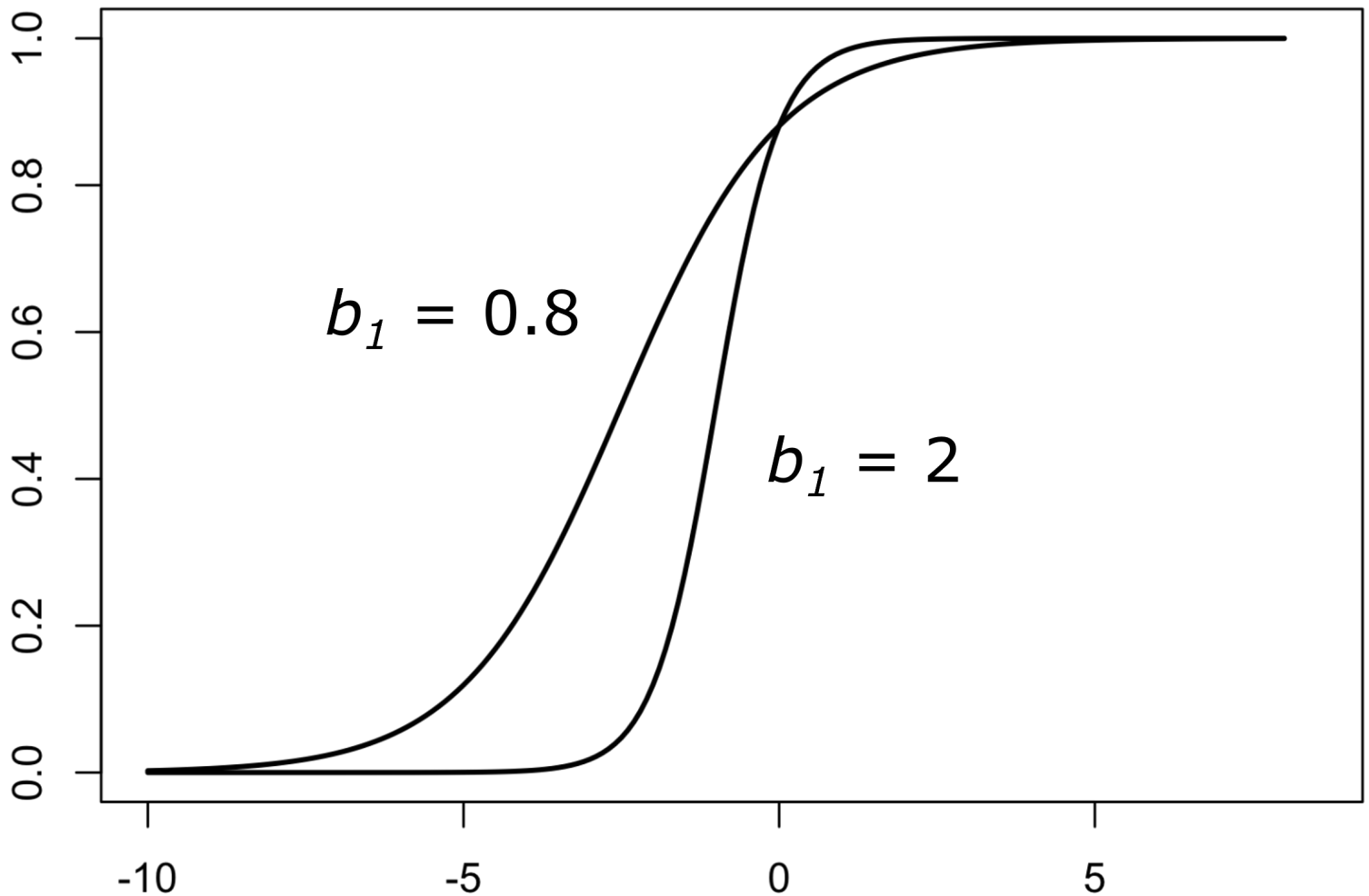
$$\pi_i = \text{logit}^{-1}(b_0 + b_1 x_i)$$

$$\pi_i = \frac{1}{1 + e^{-(b_0 + b_1 x_i)}}$$

$$\hat{y} = \frac{\exp(b_0 + b_1 x_i)}{1 + \exp(b_0 + b_1 x_i)}$$

Interpretation of slope b_1

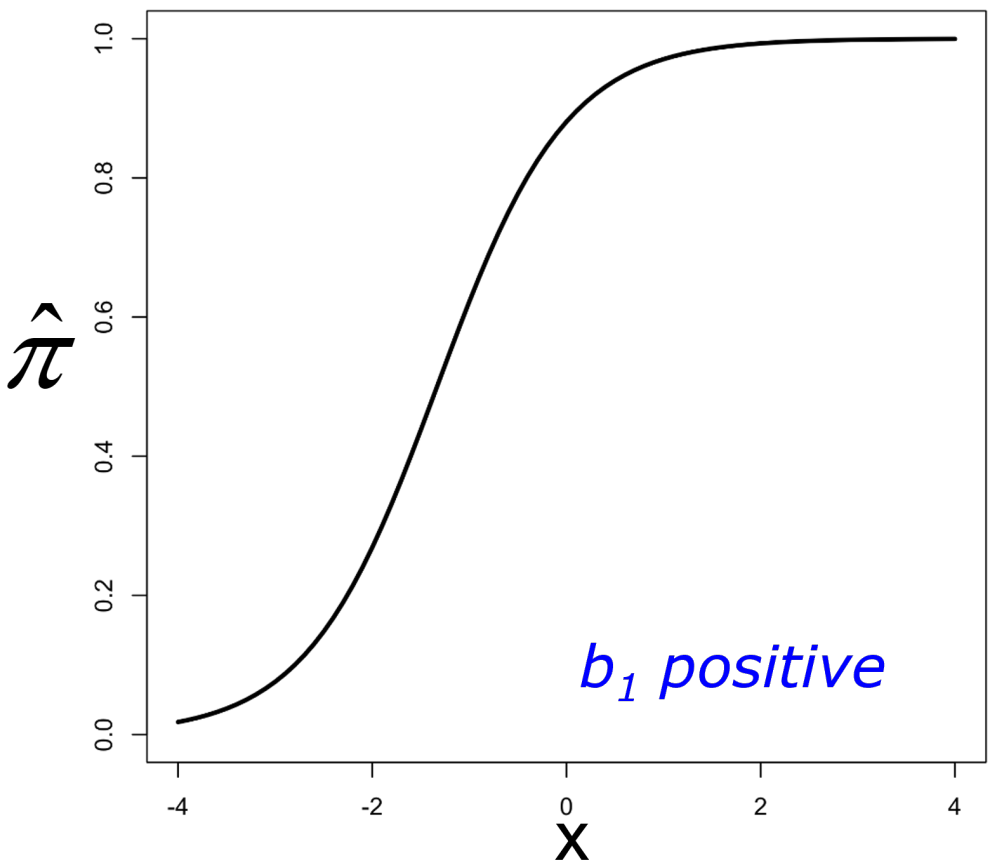
b_1 reflects the slope (steepness) of the logistic curve

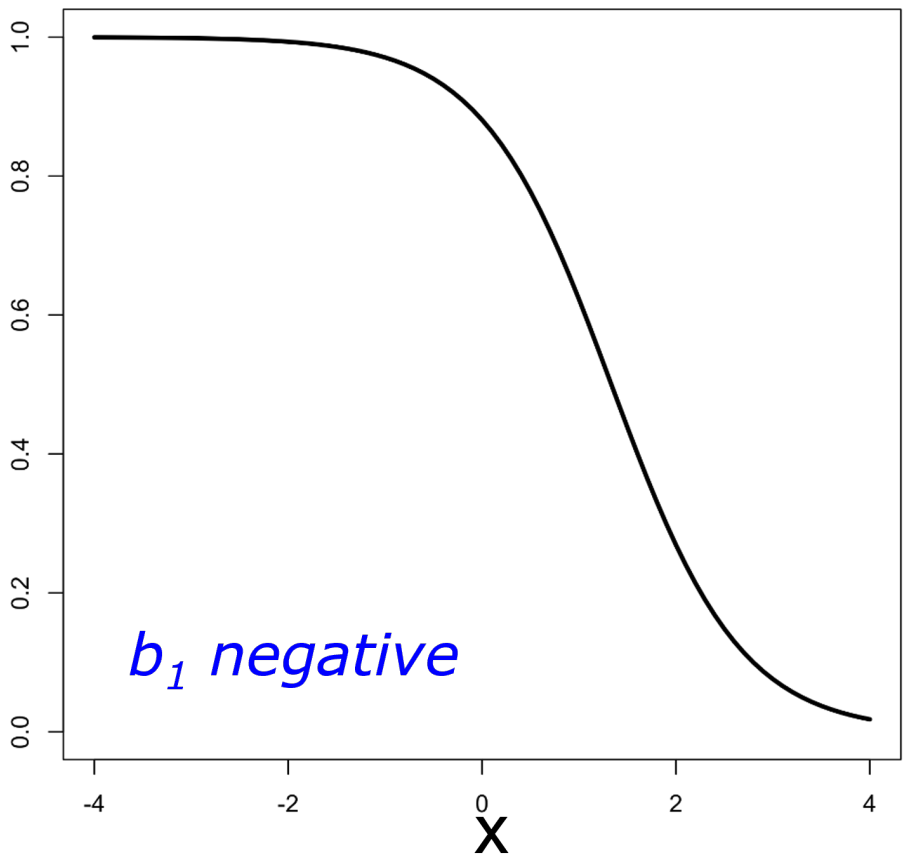


Note. Slope is nearly linear between $\hat{\pi} = .2$ and $\hat{\pi} = .8$

Interpretation of slope b_1

- b_0 (intercept) renders $\hat{\pi}$ when $X = 0$
- b_1 (slope) is positive (left) or negative (right)





Interpretation of slope b

- Linear regression (**wrong for binary outcome!**):
  - 1 unit increase in X produces a change of b in Y.
- Logistic regression:
  - 1 unit increase in X produces a change of b in the logit of Y. This is an *additive* effect.
  - 1 unit increase in X multiplies the odds of Y by e^b. This is a *multiplicative* effect.

Interpretation of the exponentiated slope e^b

- For continuous predictors:

$$e^b = \text{change in odds for an increase of } X \text{ by 1 unit}^{[1]}$$

- For categorical predictors:

$$e^b = \text{difference in odds compared to the reference category}^{[1]}$$

- In general:

$$\begin{aligned} e^b &= \exp(b) \\ &= \text{odds ratio} \\ &= \text{measure of effect size} \end{aligned}$$

[1] keeping all other predictors constant

What are odds ratios?

Odds ratios:

- are ratios of ratios, $\frac{\text{odds}}{\text{odds}}$
- are values between 0 and $\inf$, with midpoint 1
- are a measure of effect size in logistic regression

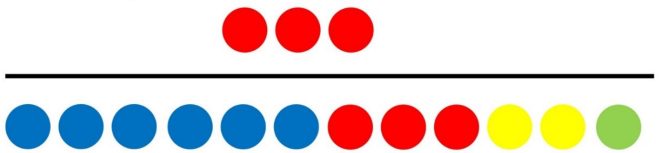
Intermezzo

Lucky Box #1

- Lucky box #1:
  - 6 blue
  - 3 red
  - 2 yellow
  - 1 green
- Red = win
- How to express chances of winning?

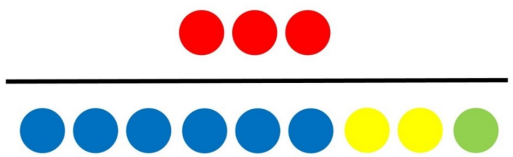
Lucky Box #1

Probability of Red



$$P(\text{red}) = \frac{n_{\text{red}}}{n_{\text{total}}} = \frac{3}{12} = \frac{1}{4}$$

Odds for Red



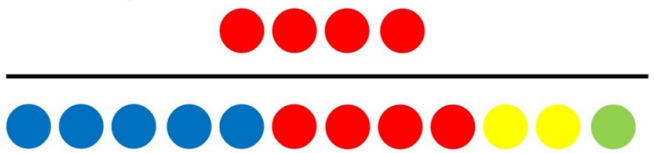
$$\text{odds}(\text{red}) = \frac{n_{\text{red}}}{n_{\text{not red}}} = \frac{3}{9} = \frac{1}{3}$$

Lucky Box #2

- Lucky box #2:
  - 5 blue
  - 4 red
  - 2 yellow
  - 1 green
- Red = win
- How to express chances of winning?

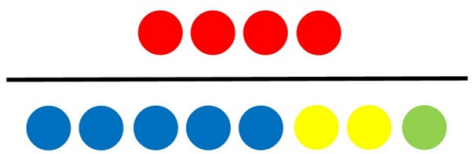
Lucky Box #2

Probability of Red



$$P(\text{red}) = \frac{n_{\text{red}}}{n_{\text{total}}} = \frac{4}{12} = \frac{1}{3}$$

Odds for Red

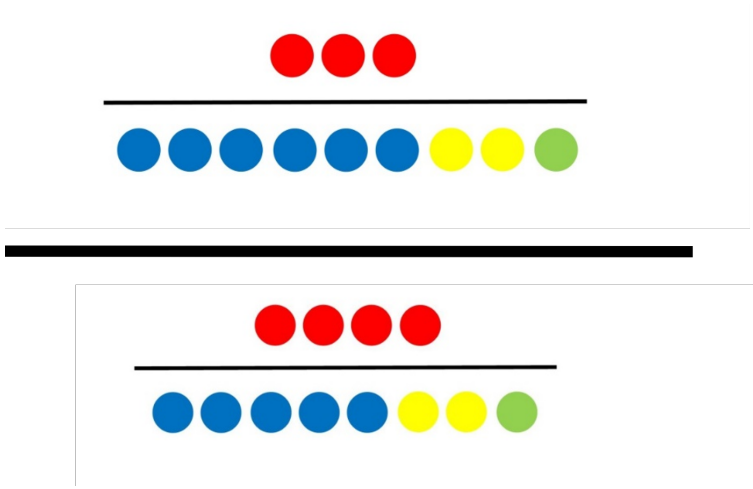


$$\text{odds}(\text{red}) = \frac{n_{\text{red}}}{n_{\text{not red}}} = \frac{4}{8} = \frac{1}{2}$$

Odds ratio


加送
100张
便签纸
一支签名笔
买就
送
50
个开口
抽奖球
幸运
大抽奖
30cm
30cm
LUCKY BOX
幸运抽奖箱
惠有好运

Odds ratio



$$\text{odds ratio} = \frac{\text{odds(red)}_{\text{Lucky Box 1}}}{\text{odds(red)}_{\text{Lucky Box 2}}} = \frac{1/3}{1/2} = \frac{2}{3}$$

Theory (continued)

Interpretation of slope b

$$
\begin{aligned}
e^b &= \exp(b) \\
&= \text{odds ratio} \\
&= \text{measure of effect size}
\end{aligned}
$$

Statistical significance of b

Wald test for coefficients:

- $H_0 : \beta = 0$
- Wald statistic: $\frac{b}{SE} \approx z$
- p-value: $P(Z > |z|)$

Model fit versus null model

Via omnibus null hypothesis test:

- Testing all regression coefficients simultaneously by specifying a null model (intercept only model)
  - $H_0 : b_1 = b_2 = b_3 = b_4 = b_5 = 0$
  - H_a : At least one b is not equal to zero

Model fit nested models

Via maximum likelihood estimation (instead of least squares):

- Compare -2 log likelihood (-2LL):
  - H_0 : -2LL $(\text{null model}) =$ -2LL $(\text{alternative model})$.
  - H_a : -2LL $(\text{null model}) >$ -2LL $(\text{alternative model})$.
- -2LL is a measure of misfit (generalization of residual sum of squares).
- Adding independent variables will decrease -2LL. Test for the significance of this decrease.
- The difference in -2LL of two nested models follows a chi-squared distribution with degrees of freedom equal to number of added bs.

Explained variance

- **Linear regression:**
  - R^2 (= proportion of variance in y explained by the model) $\rightarrow$ comparing models with R^2-change (F-distributed)
- **Logistic regression:**
  - -2log likelihood (= generalization of residual sum of squares) $\rightarrow$ comparing models with change in $-2\log L$ (χ^2-distributed) $\rightarrow$ relevance: using a classification table

Relevance

Classification table, or “confusion matrix”

	Predicted 0	**Predicted 1**
Actual 0	**true negative**	false positive
Actual 1	false negative	**true positive**

Accuracy

	Predicted 0	**Predicted 1**
Actual 0	**true negative**	false positive
Actual 1	false negative	**true positive**

Accuracy:

- $\frac{\text{true positive}+\text{true negative}}{\text{total}}$
- Proportion of correctly classified cases

Sensitivity

	Predicted 0	**Predicted 1**
Actual 0	**true negative**	false positive
Actual 1	false negative	**true positive**

Sensitivity:

- $\frac{\text{true positive}}{\text{actual positives (TP + FN)}}$
- Proportion of actual positive cases that are correctly classified

Specificity

	Predicted 0	**Predicted 1**
Actual 0	**true negative**	false positive
Actual 1	false negative	**true positive**

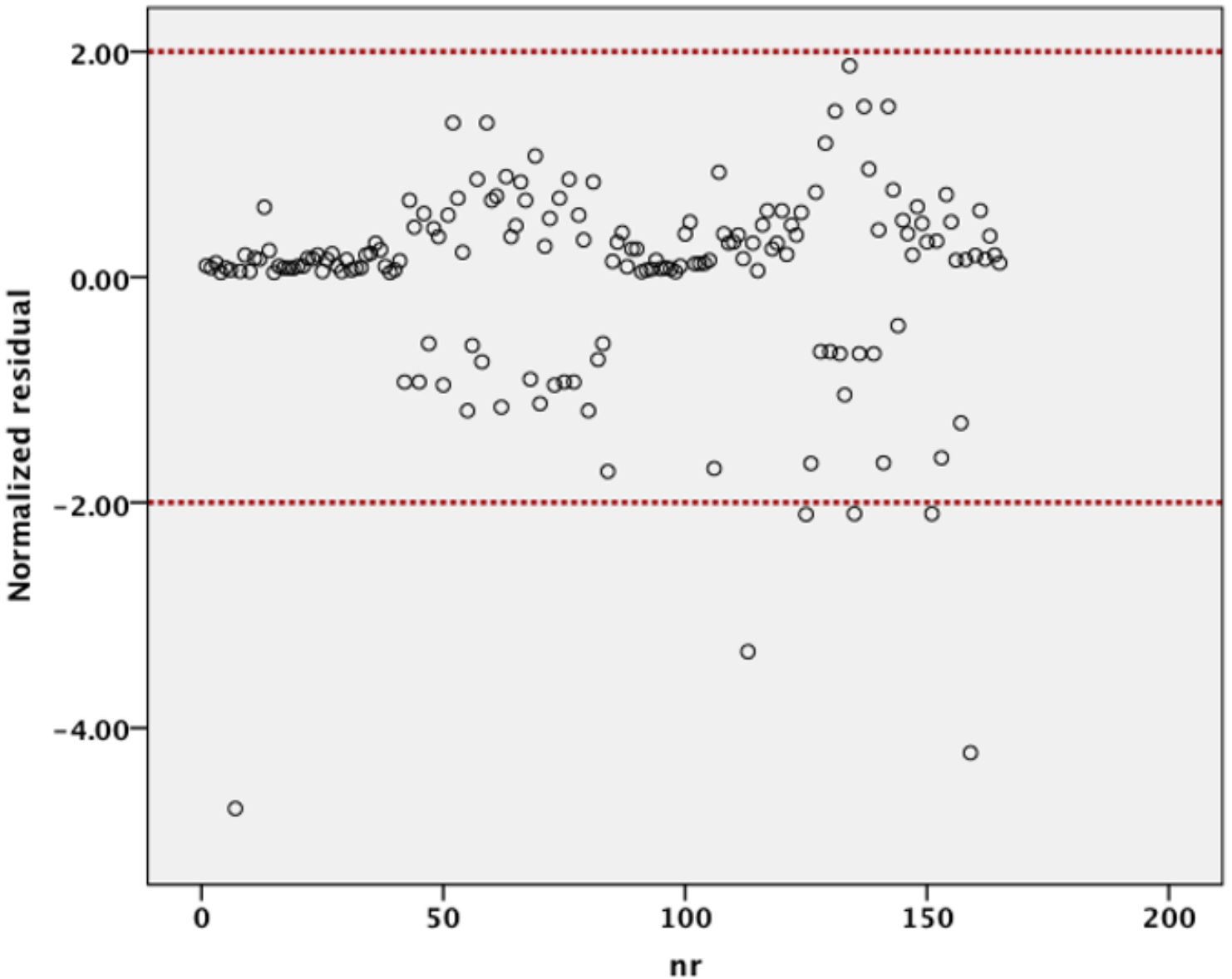
Specificity:

- $\frac{\text{true negative}}{\text{actual negatives (TN + FP)}}$
- Proportion of actual negative cases that are correctly classified

Assumptions of logistic regression

- Independence of observations.
- Measurement level of variables.
- Linearity of relations: linear relation between *logit* of dependent variable and the independent variables.
- Absence of outliers (y-, x-, xy-space).
- Absence of (multi)collinearity.
- ~~Normality of residuals~~.
- ~~Homoscedasticity~~.

Residuals



Case study

Research article

'Evaluating the morality of animal research: Effects of ethical ideology, gender, and purpose' by Wuensch and Poteat (1998), *Journal of Social Behavior and Personality*.

Evaluating the Morality of Animal Research: Effects of Ethical Ideology, Gender, and Purpose

Karl L. Wuensch
G. Michael Poteat
East Carolina University

College students (N = 315) were asked to pretend that they were serving on a university research committee hearing a complaint against animal research being conducted by a member of the university faculty. Five different research scenarios were used: Testing cosmetics, basic theory testing, agricultural (meat production) research, veterinary research, and medical research. Participants were asked to rate how justified they thought the research was and to decide whether or not the research should be halted. An ethical inventory was used to measure participants' idealism and relativism. Idealism was negatively associated and relativism positively associated with support for animal research. Women were much less accepting of animal research than were men. Support for the cosmetic, theoretical, and agricultural research projects was significantly less than that for the medical research.

During the past few years, psychologists have frequently addressed the morality of conducting research on nonhuman animals (Baldwin, 1993; Bowd & Shapiro, 1993; Ulrich, 1991). A few people may argue that such research is never morally acceptable, others may argue that it is always acceptable, and most people are of the opinion that animal research is acceptable in some circumstances but not in others. For example, three fourths of Plous' (1996a,b) respondents (psychologists and psychology majors) supported the use of animals in psychological research, but a majority opposed research which caused pain or resulted in the death of the animal. The opposition to painful or terminal research was most pronounced when the subject was a primate or a dog rather than a pigeon or a rat.

Authors' Note: We would like to thank Pamela S. George for her assistance on this project.

Author info: KARL L. WUENSCH, Department of Pyschology, East Carolina University, Greenville, NC 27858-4353; (252) 328-4102; fax (252) 328-6283, pswuensc@ecuvm.cis.ecu.edu

Journal of Social Behavior and Personality, 1998, Vol. 13, No. 1, 139–150.

Variables

Dependent variable:

- **Decision:** Whether the mock juror recommended to stop the study (0) or continue (1)

Independent variables:

- **Scenario:** Whether the study presented to the mock juror was:
  - Cosmetic (1)
  - Theory (2)
  - Meat (3)
  - Veterinary (4)
  - Medical (5; reference)

Variables

- **Gender:** Participants' gender
  - Male (0)
  - Female (1)
- **Idealism:** Participants' average score on idealism questionnaire (0-9, higher = more idealism)
- **Relativism:** Participants' average score on relativism questionnaire (0-9, higher = more relativism)

Research questions

- Is the **gender** of the mock juror **associated with** the **outcome**?
- Can the **scenario** of the study presented to the mock juror **predict** whether the **outcome** of the decision is ‘stop’ or ‘continue’?
- Does the **ethics position** of the mock juror **add predictability** of the model?

Load data

- Code

```
tibble [315 × 13] (S3: tbl_df/tbl/data.frame)
 $ decision   : dbl+lbl [1:315] 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1,
...
   ..@ format.spss: chr "F1.0"
   ..@ labels     : Named num [1:2] 0 1
   .. ..- attr(*, "names")= chr [1:2] "stop" "continue"
 $ idealism   : num [1:315] 8.2 6.8 8.2 7.4 1.7 5.6 7.2 7.8 7.8 8 ...
  ..- attr(*, "format.spss")= chr "F12.4"
 $ relatvsm   : num [1:315] 5.1 5.3 6 6.2 3.1 7.7 6.7 4 4.7 7.6 ...
  ..- attr(*, "format.spss")= chr "F12.4"
 $ gender     : dbl+lbl [1:315] 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0,
...
   ..@ format.spss: chr "F1.0"
   ..@ labels     : Named num [1:2] 0 1
   .. ..- attr(*, "names")= chr [1:2] "Female" "Male"
 $ cosmetic   : num [1:315] 1 1 1 1 1 1 1 1 1 1 ...
  ..- attr(*, "format.spss")= chr "F1.0"
```


Clean data

- Code

```
tibble [315 × 5] (S3: tbl_df/tbl/data.frame)
 $ decision  : num [1:315] 0 1 1 0 1 1 0 0 0 0 ...
 $ idealism  : num [1:315] 8.2 6.8 8.2 7.4 1.7 5.6 7.2 7.8 7.8 8 ...
 $ gender    : Factor w/ 2 levels "male","female": 1 2 1 1 1 2 1 2 1 1 ...
 $ scenario  : Factor w/ 5 levels "medical","cosmetic",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ relativism: num [1:315] 5.1 5.3 6 6.2 3.1 7.7 6.7 4 4.7 7.6 ...
```


Descriptive statistics

- Code

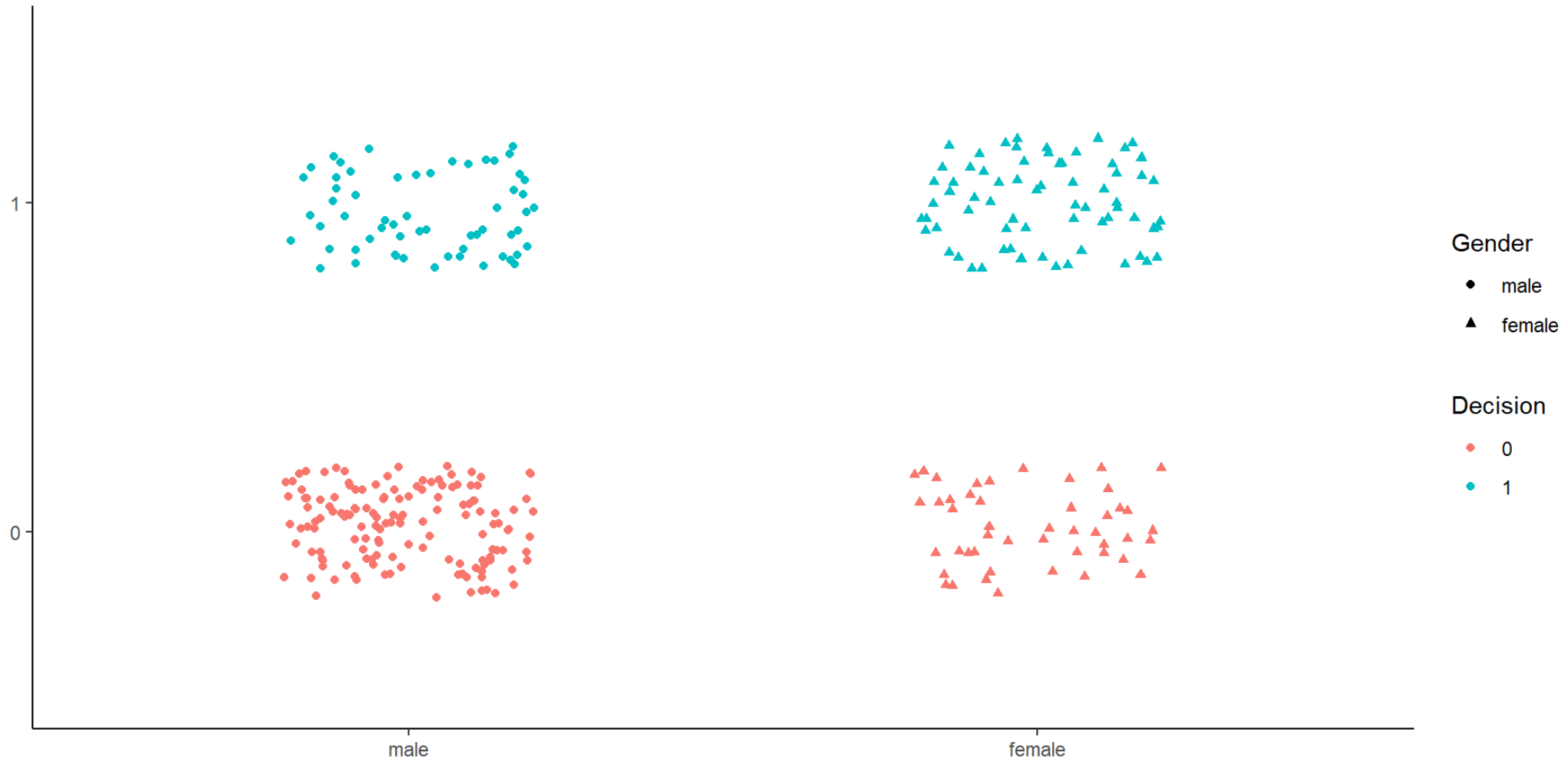
```
    decision          idealism        gender          scenario    relativism   
 Min.   :0.000   Min.   :1.70   male  :200   medical   :63   Min.   :3.10  
 1st Qu.:0.000   1st Qu.:5.60   female:115   cosmetic  :62   1st Qu.:5.40  
 Median :0.000   Median :6.50                theory    :64   Median :6.10  
 Mean   :0.406   Mean   :6.49                meat      :63   Mean   :6.05  
 3rd Qu.:1.000   3rd Qu.:7.50                veterinary:63   3rd Qu.:6.80  
 Max.   :1.000   Max.   :9.00                                Max.   :8.90  
```


Outcome variable

- Code

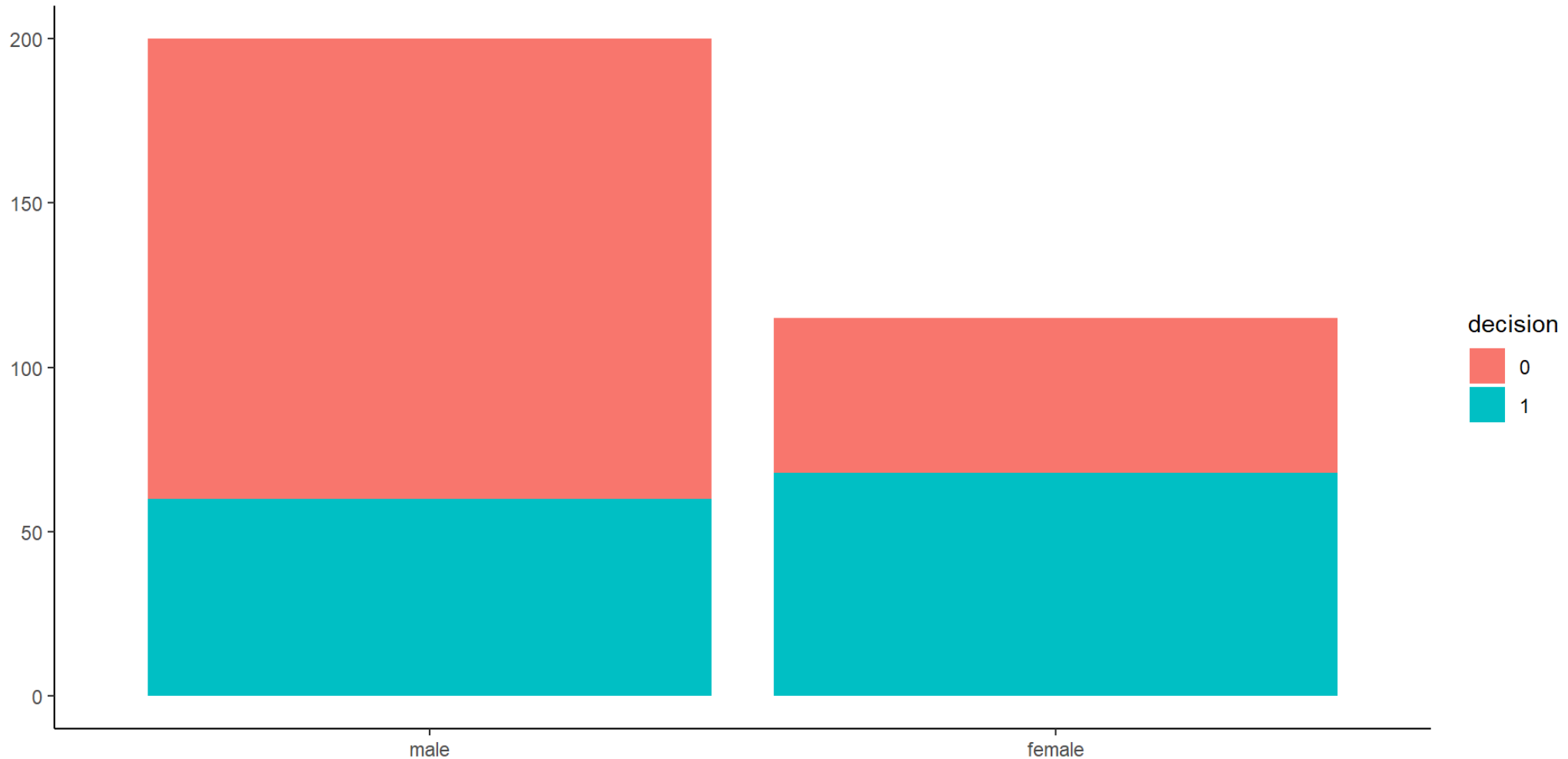
Association with IV

- Code



Association with IV

- Code



Logistic regression with 1 IV

- Code

```
Call:  glm(formula = decision ~ gender, family = "binomial", data = dat_clean)

Coefficients:
 (Intercept)  genderfemale
      -0.847         1.217

Degrees of Freedom: 314 Total (i.e. Null);  313 Residual
Null Deviance:      426
Residual Deviance: 400  AIC: 404
```


Interpretation of slope b_1

- Code

```
(Intercept) genderfemale
     -0.847        1.217
```

$$\log(\frac{\hat{\pi}\ \text{continue}}{\hat{\pi}\ \text{stop study}}) = -0.85 + 1.22 \times \text{gender}$$

$$\frac{\hat{\pi}\ \text{continue}}{\hat{\pi}\ \text{stop study}} = 0.43 \times 3.38 \times \text{gender}$$

Interpretation of slope b_1

- Code

```
(Intercept) genderfemale
     -0.847        1.217
```

- Code

```
(Intercept) genderfemale
      0.429        3.376
```


Interpretation of slope b_1

$$\log(\frac{\hat{\pi}\ \text{continue}}{\hat{\pi}\ \text{stop study}}) = -0.85 + 1.22 \times \text{gender}$$

- Coefficient b_1 for gender = 1.22
- Increasing gender by 1 unit increases the logit by 1.22
- Since gender is coded as a binary variable, increasing it by one unit means comparing it to the reference category (male)
- Hence, female mock jurors have a 1.22 higher logit than male mock jurors

Interpretation of slope b_1

$$\log(\frac{\hat{\pi}\ \text{continue}}{\hat{\pi}\ \text{stop study}}) = -0.85 + 1.22 \times \text{gender}$$

- Coefficient b_1 for gender = 1.22
- Increasing gender by 1 unit **multiplies** the odds of deciding to continue rather than stop by $e^{1.22} = \mathbf{3.39}$
- Female mock jurors have 3.39 **times** higher odds of deciding to continue the research than male mock jurors

Statistical significance of b_1

- Code

```
Call:
glm(formula = decision ~ gender, family = "binomial", data = dat_clean)

Coefficients:
             Estimate Std. Error z value   Pr(>|z|)
(Intercept)    -0.847      0.154   -5.49 0.00000004 ***
genderfemale    1.217      0.245    4.98 0.00000065 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 425.57  on 314  degrees of freedom
Residual deviance: 399.91  on 313  degrees of freedom
AIC: 403.9
```


Confidence interval of b_1

▸ Code

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.429	0.154	-5.49	0	0.315	0.577
genderfemale	3.376	0.245	4.98	0	2.100	5.482

Logistic regression with multiple IVs

- Code

```
Call:  glm(formula = decision ~ gender + scenario, family = "binomial",
    data = dat_clean)

Coefficients:
       (Intercept)         genderfemale    scenariocosmetic      scenariotheory
            -0.229                1.316               -0.796               -1.168
     scenariomeat  scenarioveterinary
            -0.804               -0.560

Degrees of Freedom: 314 Total (i.e. Null);  309 Residual
Null Deviance:      426
Residual Deviance: 390  AIC: 402
```


Interpretation of multiple bs

- Code

```
Call:
glm(formula = decision ~ gender + scenario, family = "binomial",
    data = dat_clean)

Coefficients:
                  Estimate Std. Error z value   Pr(>|z|)
(Intercept)         -0.229      0.272   -0.84     0.3988
genderfemale         1.316      0.254    5.19 0.00000021 ***
scenariocosmetic    -0.796      0.384   -2.07     0.0384 *
scenariotheory      -1.168      0.392   -2.98     0.0029 **
scenariomeat        -0.804      0.382   -2.10     0.0353 *
scenarioveterinary  -0.560      0.377   -1.49     0.1368
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```


Interpretation of multiple bs

- Code

term	estimate	std.error	statistic	p.value
(Intercept)	-0.229	0.272	-0.844	0.399
genderfemale	1.316	0.254	5.187	0.000
scenariocosmetic	-0.796	0.384	-2.070	0.038
scenariotheory	-1.168	0.392	-2.982	0.003
scenariomeat	-0.804	0.382	-2.104	0.035
scenarioveterinary	-0.560	0.377	-1.488	0.137

Interpretation of multiple bs

- Code

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.795	0.272	-0.844	0.399	0.465	1.357
genderfemale	3.730	0.254	5.187	0.000	2.282	6.182
scenariocosmetic	0.451	0.384	-2.070	0.038	0.210	0.952
scenariotheory	0.311	0.392	-2.982	0.003	0.142	0.662
scenariomeat	0.448	0.382	-2.104	0.035	0.209	0.940
scenarioveterinary	0.571	0.377	-1.488	0.137	0.271	1.190

Dummy coding

Check carefully! Important for interpretation of results.

- Code

```
Call:
glm(formula = decision ~ gender + scenario2, family = "binomial",
    data = dat_clean)

Coefficients:
                   Estimate Std. Error z value   Pr(>|z|)
(Intercept)          -1.398      0.312   -4.49 0.00000727 ***
genderfemale          1.316      0.254    5.19 0.00000021 ***
scenario2medical      1.168      0.392    2.98     0.0029 **
scenario2cosmetic     0.373      0.393    0.95     0.3425
scenario2meat         0.365      0.395    0.92     0.3562
scenario2veterinary   0.608      0.392    1.55     0.1211
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

Note. Reference category for scenario changed to 'theory'.

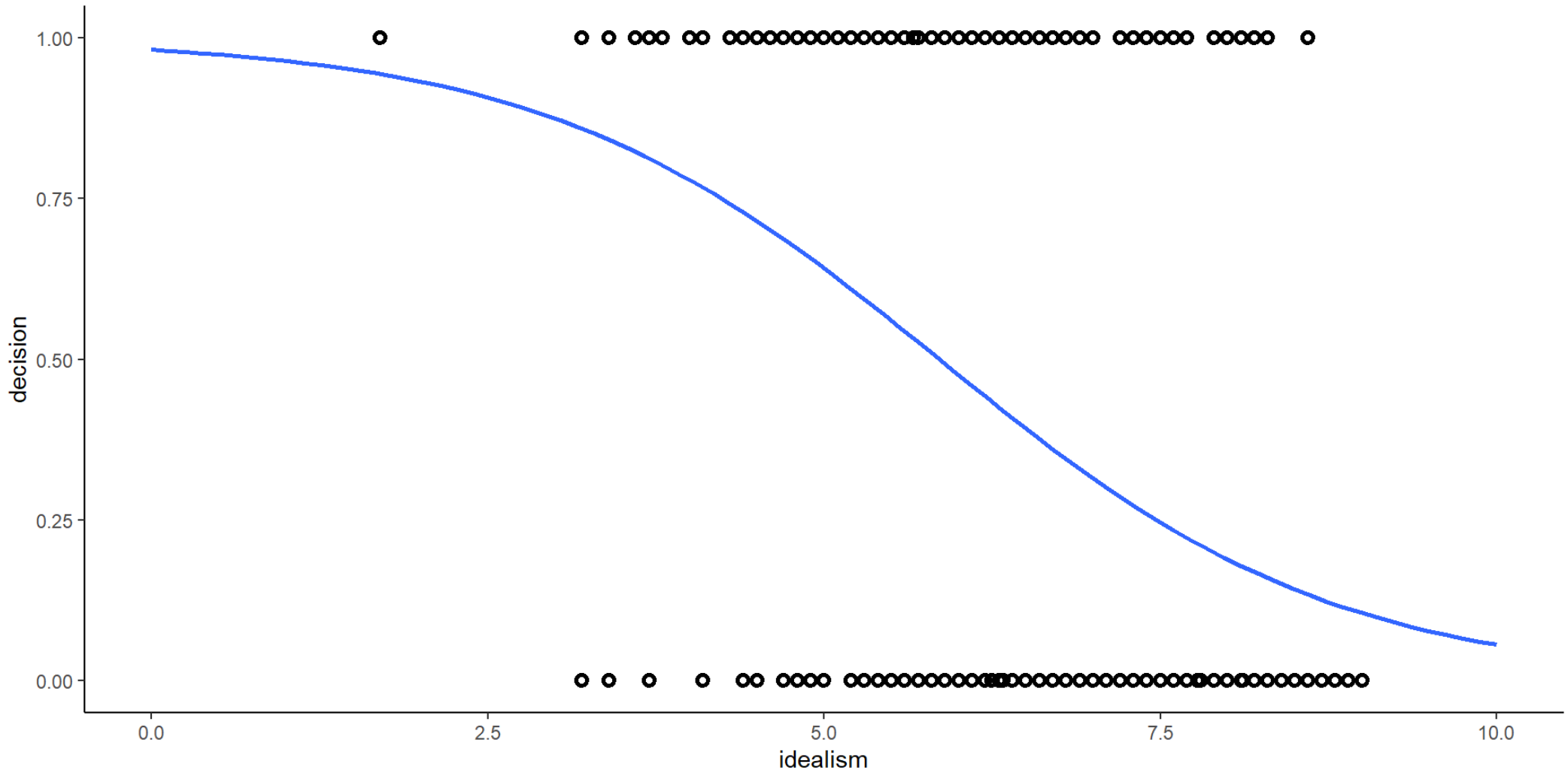
Comparing nested models

$$\hat{y} = \frac{\exp(u)}{1 + \exp(u)}$$

- **Model 0:**
  - $u = b_0$
- **Model 1:**
  - $u = b_0 + b_1 \times \text{gender}$
- **Model 2:**
  - $u = b_0 + b_1 \times \text{gender} + b_2 \times \text{idealism} + b_3 \times \text{relativism}.$

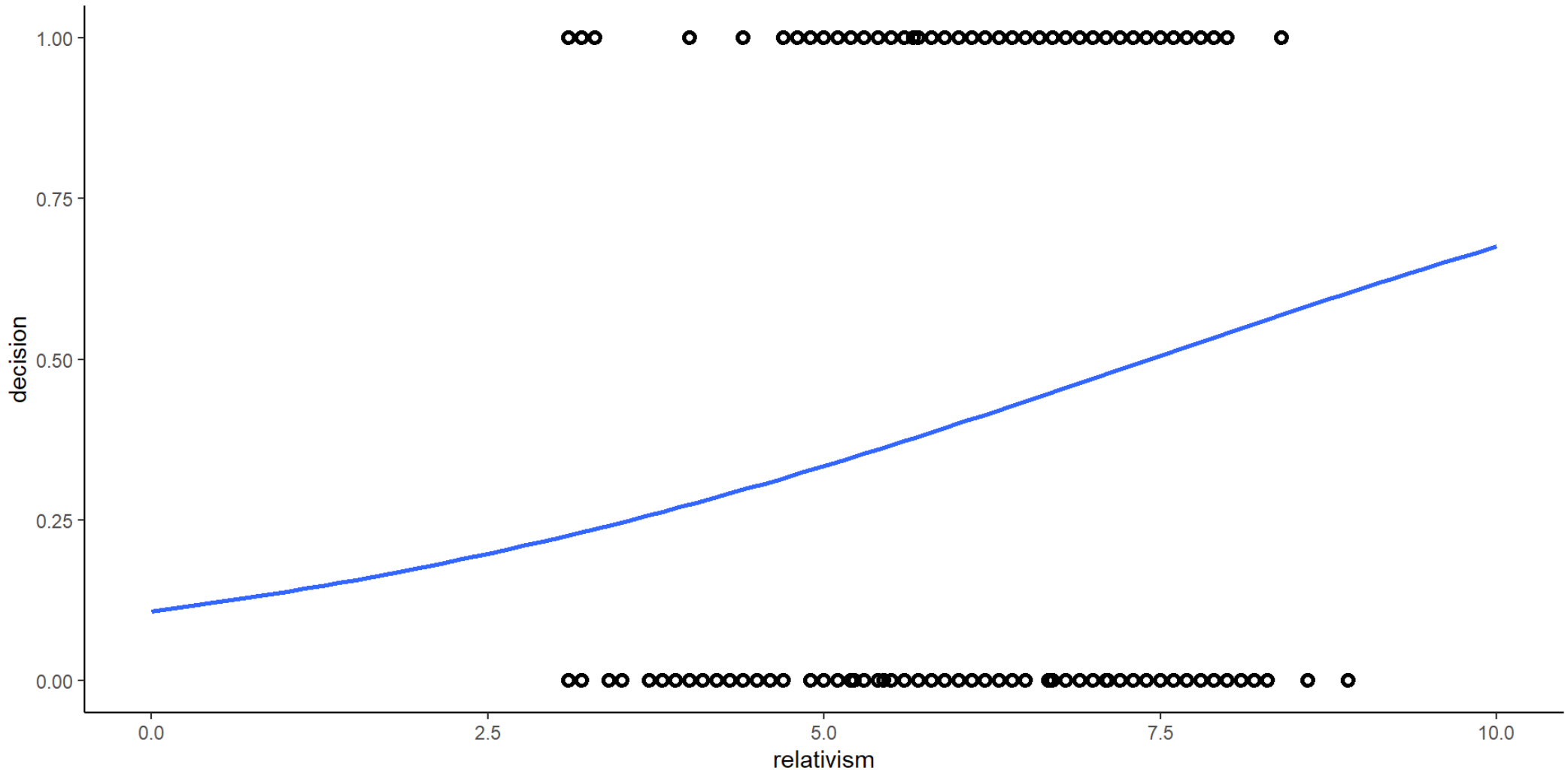
Visualization idealism

- Code



Visualization relativism

- Code



Compare coefficients

- Code

```
(Intercept)
      0.684
```

- Code

```
  (Intercept) genderfemale
        0.429        3.376
```

- Code

```
  (Intercept) genderfemale     idealism   relativism
        4.427        3.225        0.502        1.409
```


Compare -2LL

▸ Code

```
'log Lik.' -213 (df=1)
```

▸ Code

```
'log Lik.' -200 (df=2)
```

▸ Code

```
'log Lik.' -173 (df=4)
```

- Initial -2LL (null model) minus -2LL model 1: $426 - 400 = 26$.
- Model 1 -2LL minus -2LL model 2 (full model): $400 - 347 = 53$.
- The differences follow a Chi-square distribution.

Chi square test for nested models

- Model 0: no IV
- Model 1: gender
- Model 2: gender, idealism, relativism

▸ Code

```
Analysis of Deviance Table

Model 1: decision ~ 1
Model 2: decision ~ gender
Model 3: decision ~ gender + idealism + relativism
  Resid. Df Resid. Dev Df Deviance        Pr(>Chi)
1       314        426
2       313        400  1     25.7 0.0000004086164 ***
3       311        347  2     53.4 0.0000000000025 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Relevance: correct classification

▸ Code

- Model 0: 59% correctly classified
- Model 1: 66% correctly classified
- Model 2: 71% correctly classified

Relevance: other metrics

- Code

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 151  55
         1  36  73

               Accuracy : 0.711
                 95% CI : (0.658, 0.761)
    No Information Rate : 0.594
    P-Value [Acc > NIR] : 0.0000098

                  Kappa : 0.387

 Mcnemar's Test P-Value : 0.0592

            Sensitivity : 0.807
```


Interpret final model

- Code

term	estimate	std.error	statistic	p.value
(Intercept)	1.488	0.979	1.52	0.128
genderfemale	1.171	0.268	4.37	0.000
idealism	-0.689	0.112	-6.18	0.000
relativism	0.343	0.124	2.76	0.006

Interpret final model

▸ Code

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	4.427	0.979	1.52	0.128	0.663	31.15
genderfemale	3.225	0.268	4.37	0.000	1.918	5.49
idealism	0.502	0.112	-6.18	0.000	0.400	0.62
relativism	1.409	0.124	2.76	0.006	1.109	1.81

Interpret final model

- Code

```
(Intercept) genderfemale     idealism   relativism
      1.488        1.171       -0.689        0.343
```

- Female mock jurors have $e^{1.17} = 3.22$ **times** higher odds of deciding continue the study than male jurors.
- Increasing idealism by 1 unit **multiplies** the odds of deciding to continue rather than stop the study by $e^{-0.69} = 0.502$. Decreasing idealism by 1 unit **multiplies** the odds of deciding to continue rather than stop by $e^{0.69} = 2.01$.
- Increasing relativism by 1 unit **multiplies** the odds of deciding continue rather than stop by $e^{0.34} = 1.41$.

Interpret final model

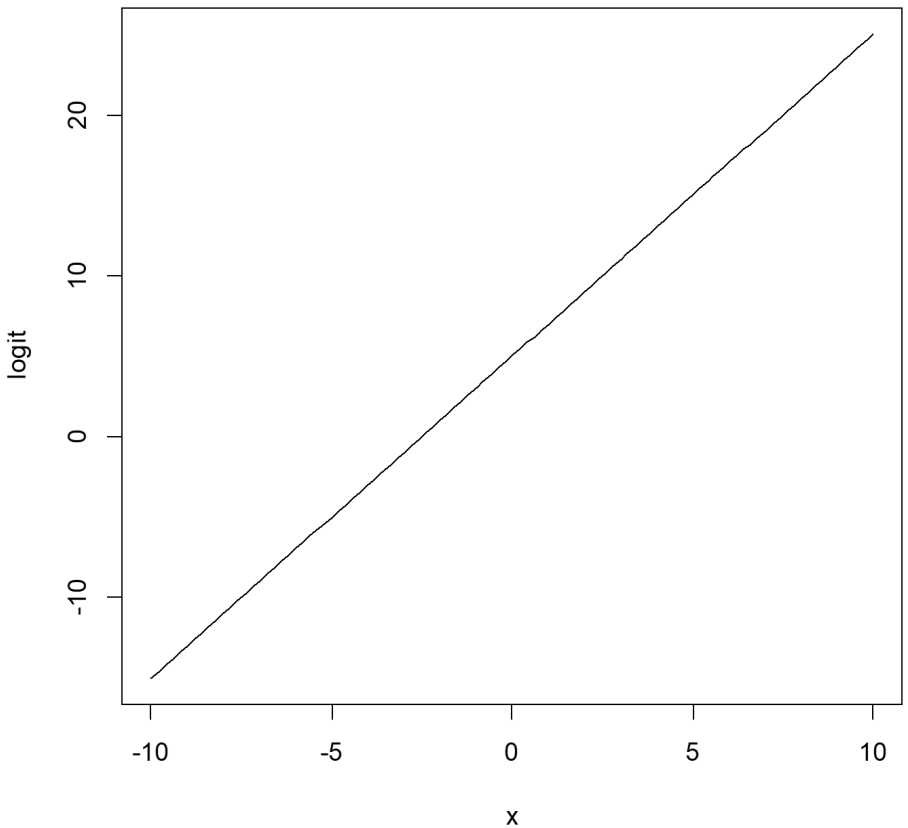
- Code

```
Nagelkerke's R2
            0.3
```

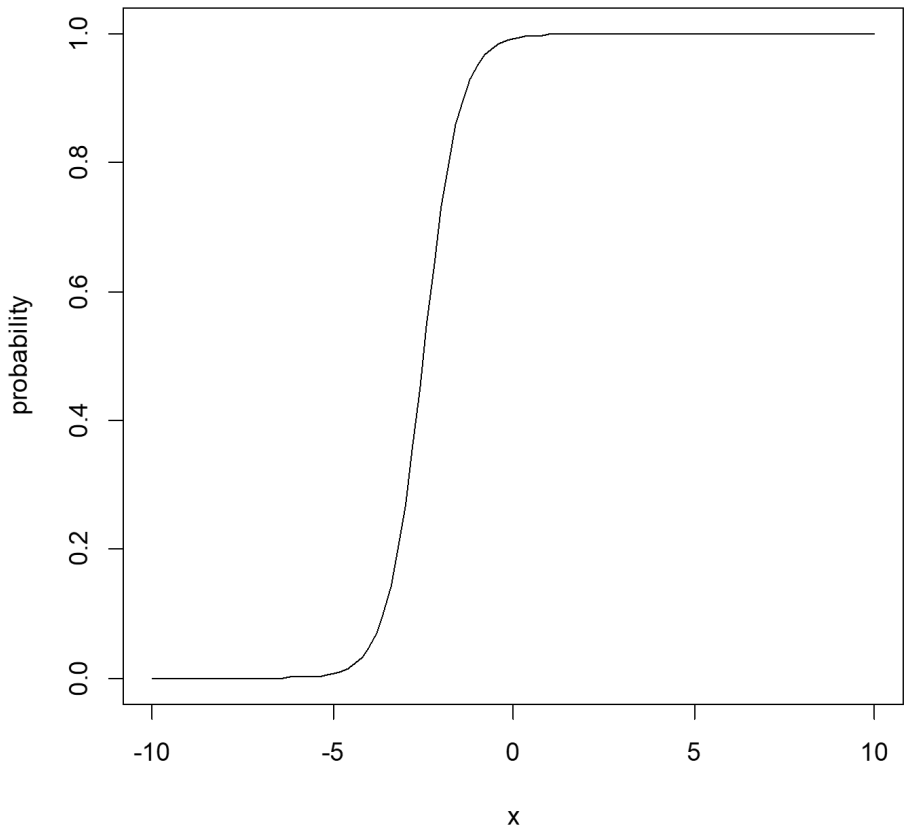
- Nagelkerke R^2 = 0.3

Marginal effects

Marginal effects



Linear relation IV and logit



Nonlinear relation IV and probability

Estimating marginal effects

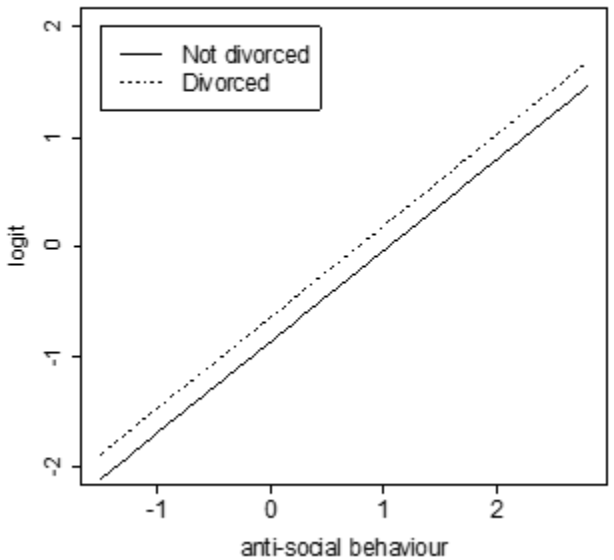
Marginal effect =$ $

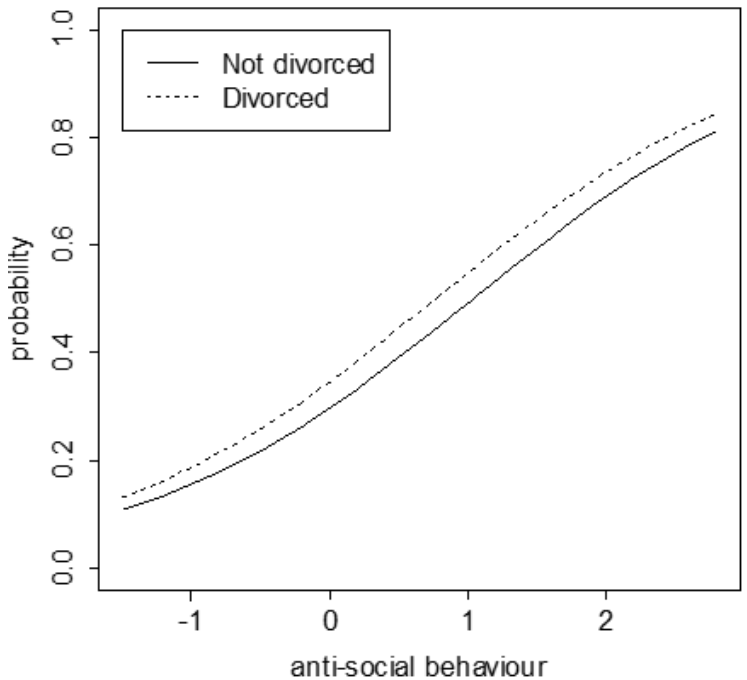
Model = full, saturated model - **Step 1:** Estimate model parameters. - **Step 2:** Estimate predicted probability for a “typical” respondent. - **Step 3:** Vary a single predictor while holding all others at typical values.

Example

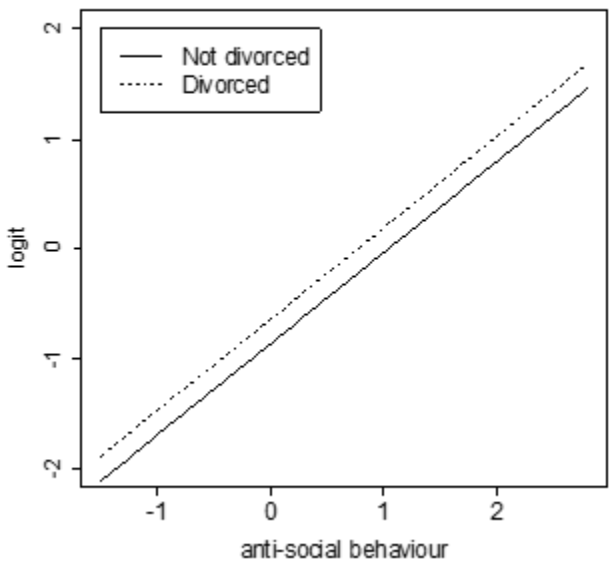
- Dependent variable:
  - sex before end of 10th school year (0 = no, 1 = yes)
- Independent variables:
  - antisocial parental behavior during the respondent's childhood (range = [-1.41, 2.78], mean = -0.14)
  - parents divorced (0 = no, 1 = yes)

Logit versus probability





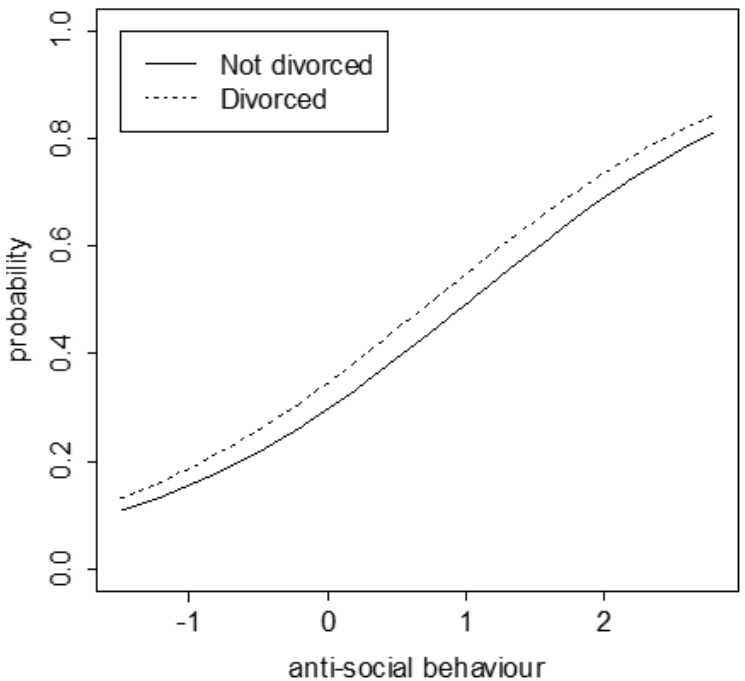
Logit scale



Effect of divorced on logit scale = 0.224

$\rightarrow$ this effect does not depend on antisocial parental behavior

Probability scale



Anti-social = -1

- $P(\text{sex}|\text{not divorced}) = 0.16$
- $P(\text{sex}|\text{divorced}) = 0.19$
- marginal effect = difference in $P = 0.03$

Anti-social = +1

- $P(\text{sex}|\text{not divorced}) = 0.49$
- $P(\text{sex}|\text{divorced}) = 0.55$
- marginal effect = difference in $P = 0.06$
- $\rightarrow$ effect of divorced on probability scale does depend on antisocial

Take-aways

Logistic regression

- Logistic regression is used to predict a binary dependent variable
  - Dichotomous outcome ($Y = 0$ or $Y = 1$, nothing observed in between)
  - Logit link function between the linear model and the logit of the probability
- The odds ratio is a measure of effect size
- The Wald test is used to test individual coefficients
- The omnibus test is used to test the model as a whole
- The -2LL is used to compare nested models
- Marginal effects are used to interpret the effect of a predictor on the probability scale

Various types of logistic regression

- Logistic regression: DV with 2 categories
- Ordinal regression: DV with 3-6 ordered categories
  - Cumulative logit models, (non-)proportional odds models
- (Multi)nominal regression: DV with unordered categories

Next meeting(s)

- Lab on Friday
- Lecture on Monday

Homework:

- Check you course manual for deadlines and what to hand in
- Assignment 2 will be online by tomorrow