Logistic regression

Methods 1, E2021 - Lecture 10
Tuesday 16/11/2021
Fabio Trecca


QUIZ
TIME

Quiz time (1)

- What is the difference between A, B, and C?

A

ID	Cond	Acc	RT
Tim	1	0	5.34
Joan	2	0	4.98
Kim	1	1	8.23
Sam	2	1	7.12
Ann	1	1	5.12
Bob	2	0	3.34
Ross	1	1	8.18

B

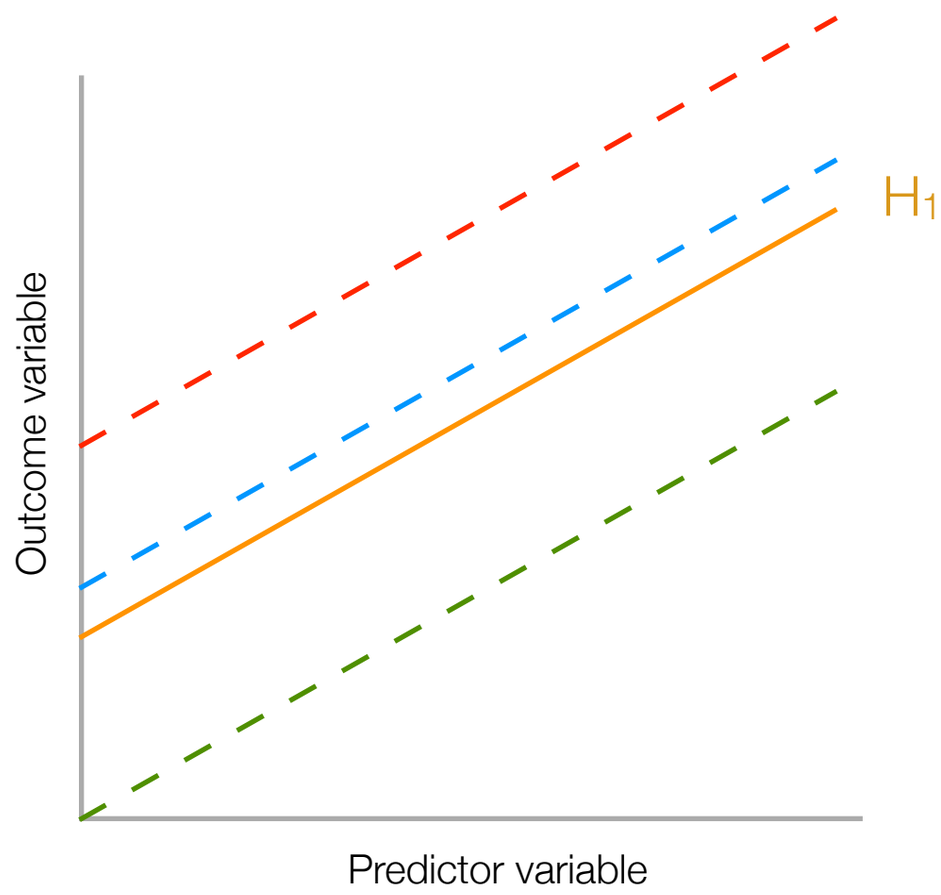
ID	Con	Trial	Acc	RT
Tim	1	1	0	5.34
Tim	1	2	0	6.44
Tim	1	3	1	8.56
Joan	2	1	1	2.34
Joan	2	2	1	3.11
Joan	2	3	0	4.56
Kim	1	1	1	6.53
Kim	1	2	1	7.34
Kim	1	3	1	7.65

C

ID	Cond	Acc	RT
Tim	1	0	5.34
Tim	2	0	5.44
Tim	3	1	5.56
Joan	1	1	6.34
Joan	2	1	6.11
Joan	3	0	6.56
Kim	1	1	3.53
Kim	2	1	3.34
Kim	3	1	3.65

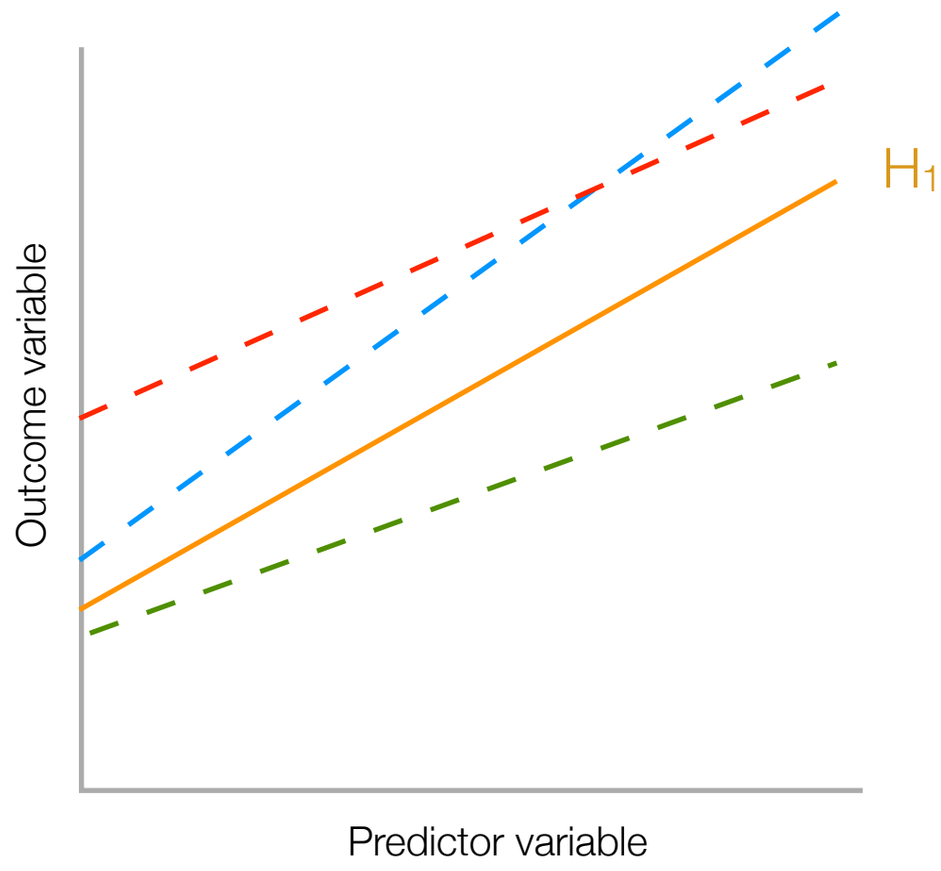
Quiz time (2)

- `lme4::lmer()`
- `summary(lmer(reading_time ~ condition + (1 | ID), data)`



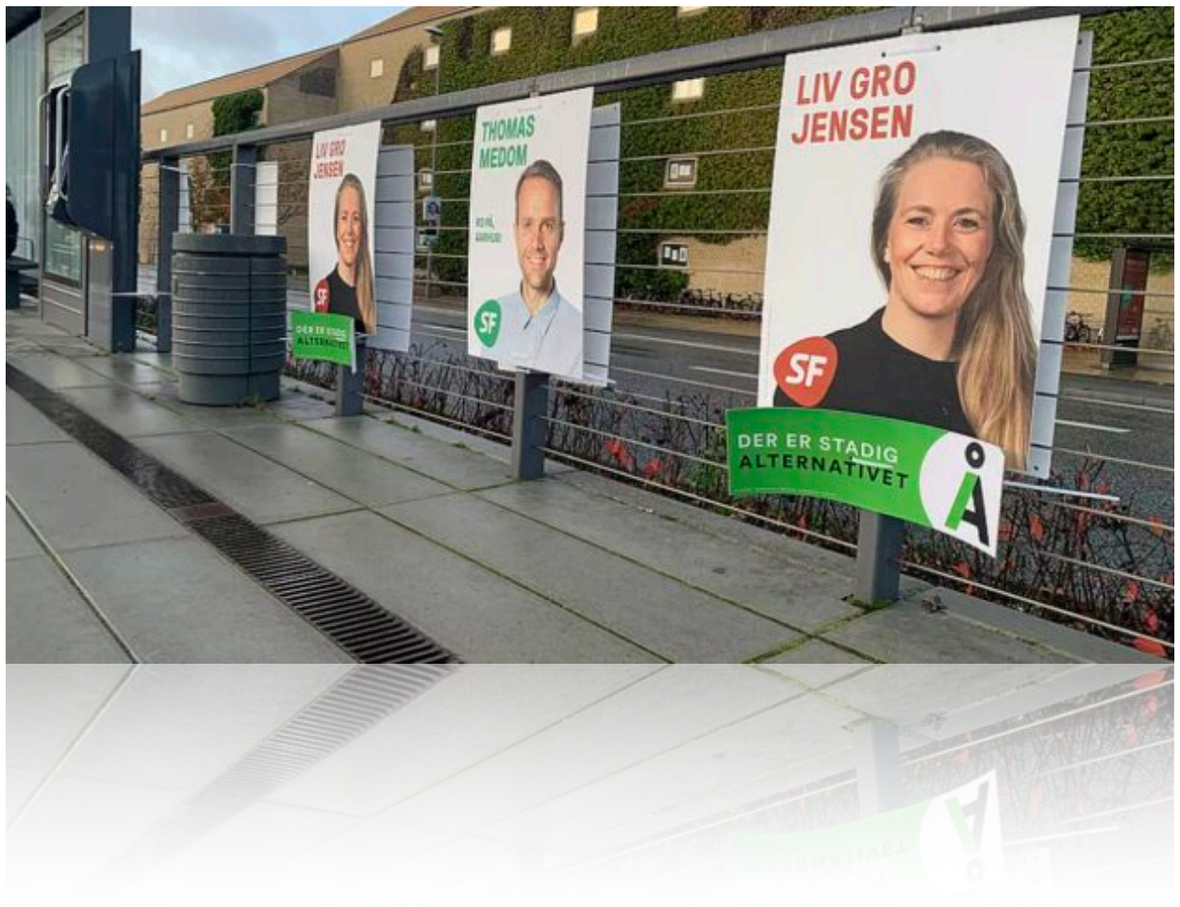
Quiz time (3)

- `lme4::lmer()`
- `summary(lmer(reading_time ~ condition + (1 + trial | ID), data)`

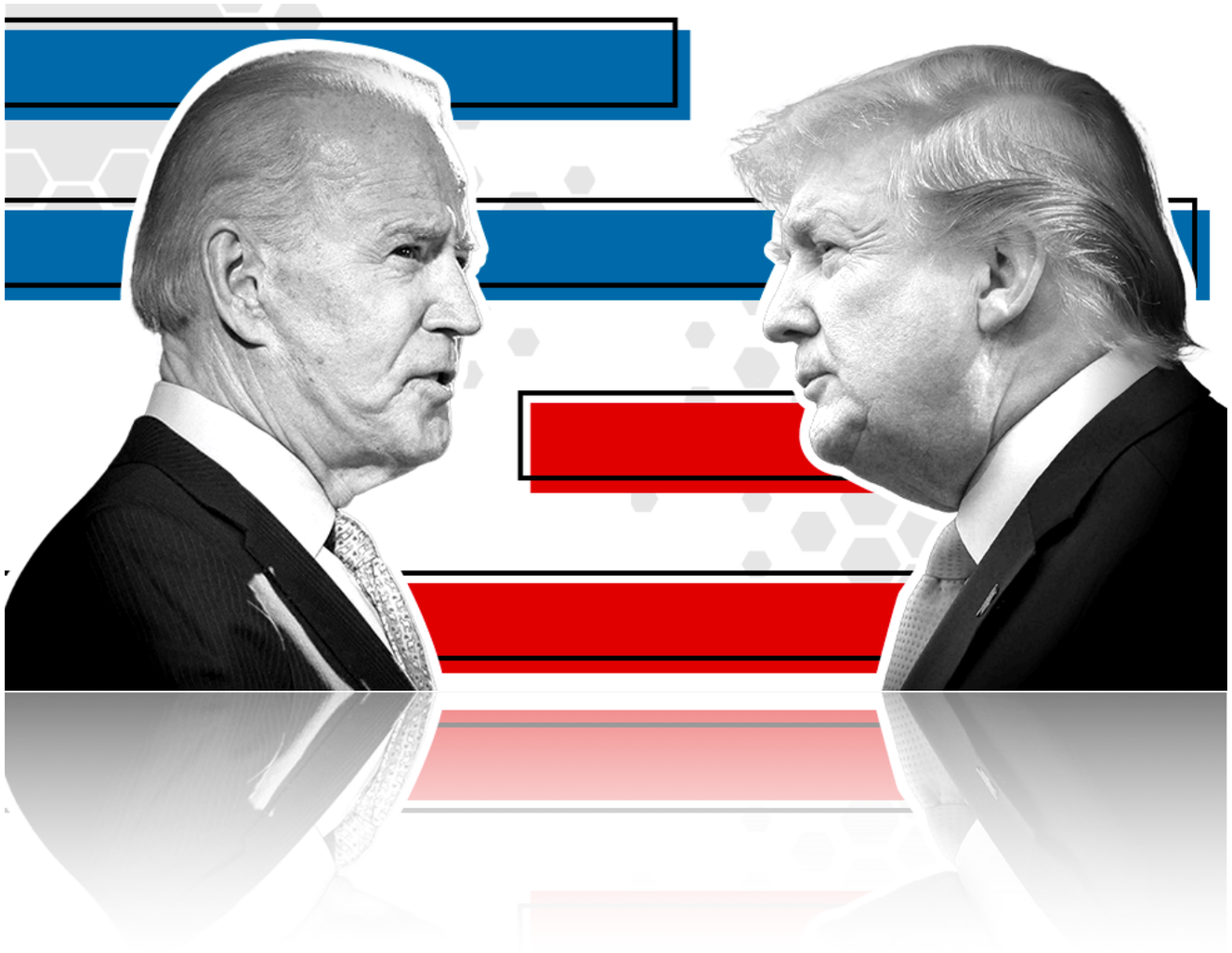


Not all outcome variables are continuous

- Linear models assume that the outcome is continuously distributed
- How about categorical outcome variables?
- A very topical example: *Kommunalvalg 2021*



A prototypical binary outcome — now 1 year old

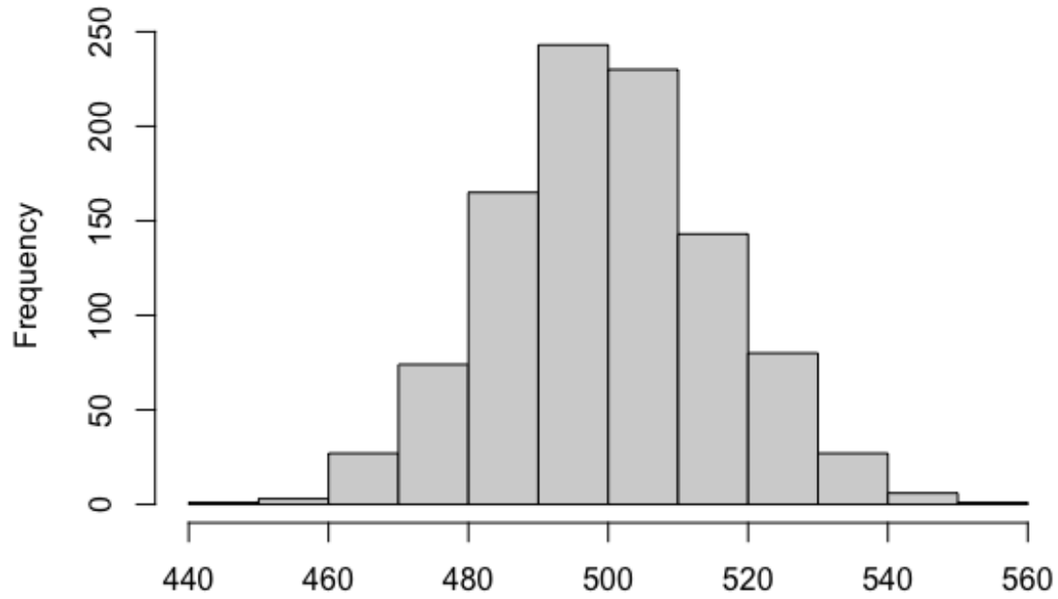


Binary outcomes are frequent in cognitive science

- Predicting accuracy on a task (correct vs. incorrect)
- Predicting preferences for one of two options
- Predicting whether a diagnosis is relevant or not
- Predicting sentiment (positive vs. negative)
- Predict whether something is TRUE or FALSE (logical variable)
- Predicting whether a child knows a specific word or not
- Predicting gender from a specific score
- … basically anything that has to do with **classification**

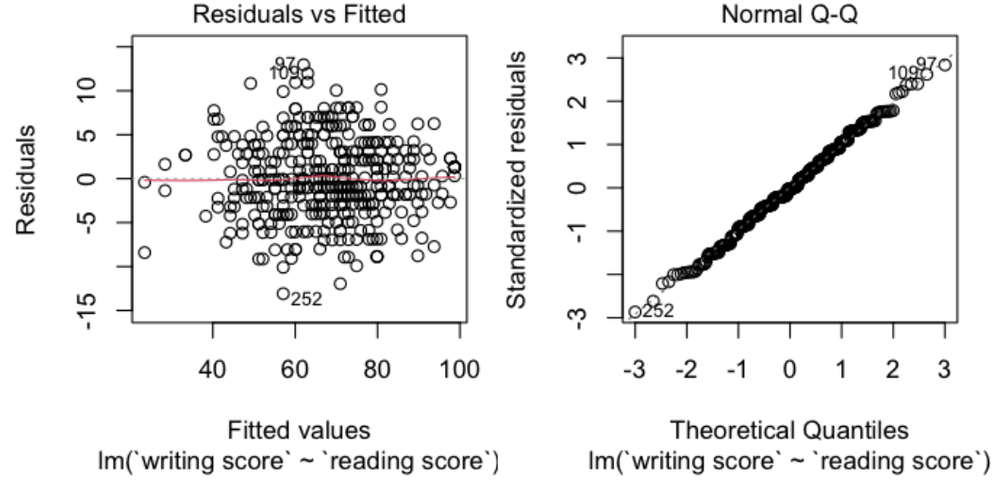
Bernoulli trials and binomial distribution

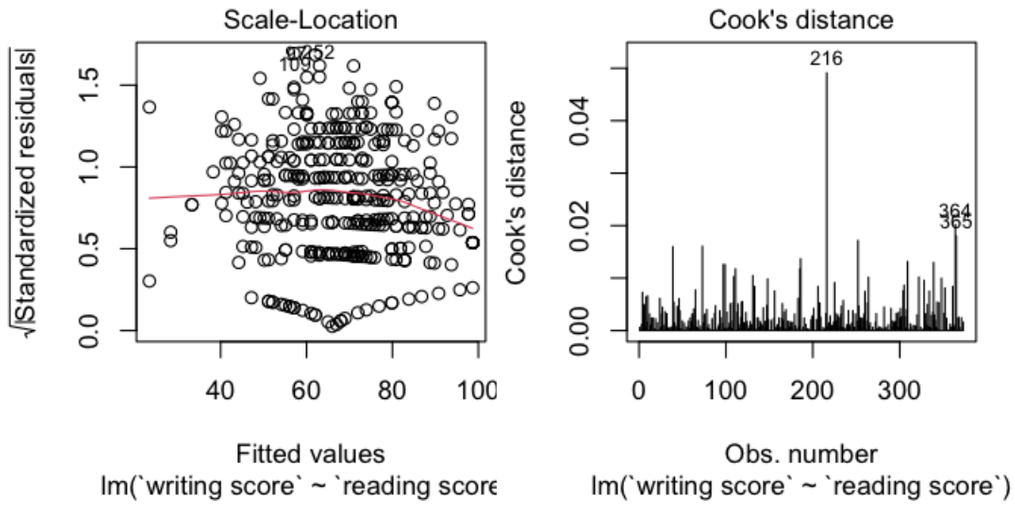
- $p = 1 - q, q = 1 - p, p + q = 1$
- $odds = \frac{p}{1 - p}$
- Eg. heads or tails
- Eg. Trump vs Biden in a randomly selected sample
- Eg. P of drawing red vs blue marble with replacement
- $P_x = \binom{n}{k} p^x q^{n-x}$
- How do we model this type of data?



Assumptions of linear regression

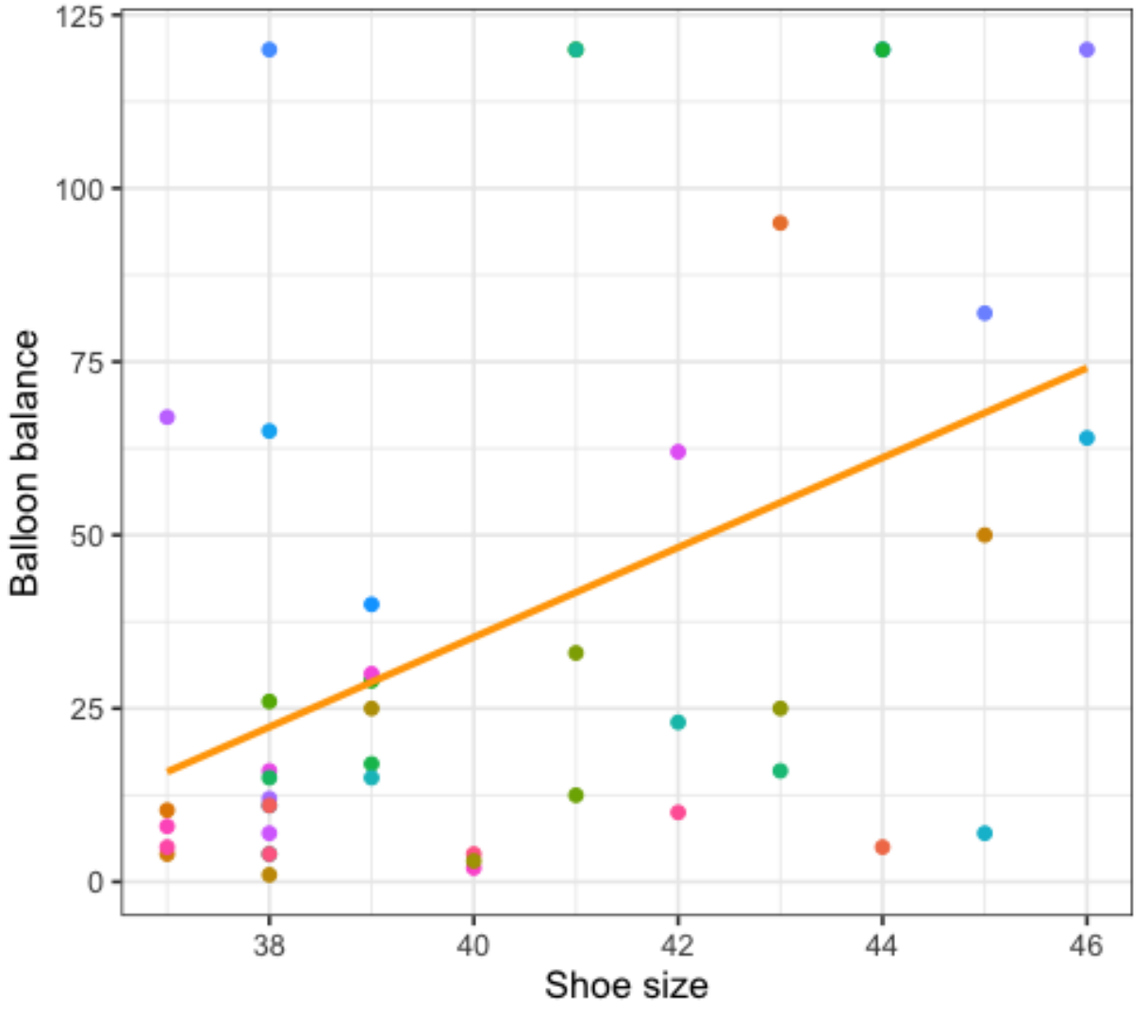
- Outcome variable must be continuous (at least at the interval level)
- Absence of multicollinearity: no perfect linear relationship between two or more predictors (NB: multiple regression only)
  - → `vif(model)` (largest value should be < 10)
- Linearity of residuals
  - → `plot(model, 1)`
- Normality of residuals: residuals are random and normally distributed with a mean of 0
  - → `plot(model, 2)`
- Homoschedasticity
  - → `plot(model, 3)`
  - → `car::leveneTest(outcome ~ predictor, data)`
- No influential cases
  - → `plot(model, 4)`
- Independence of residuals:
  - → `car::durbinWatsonTest(model)`

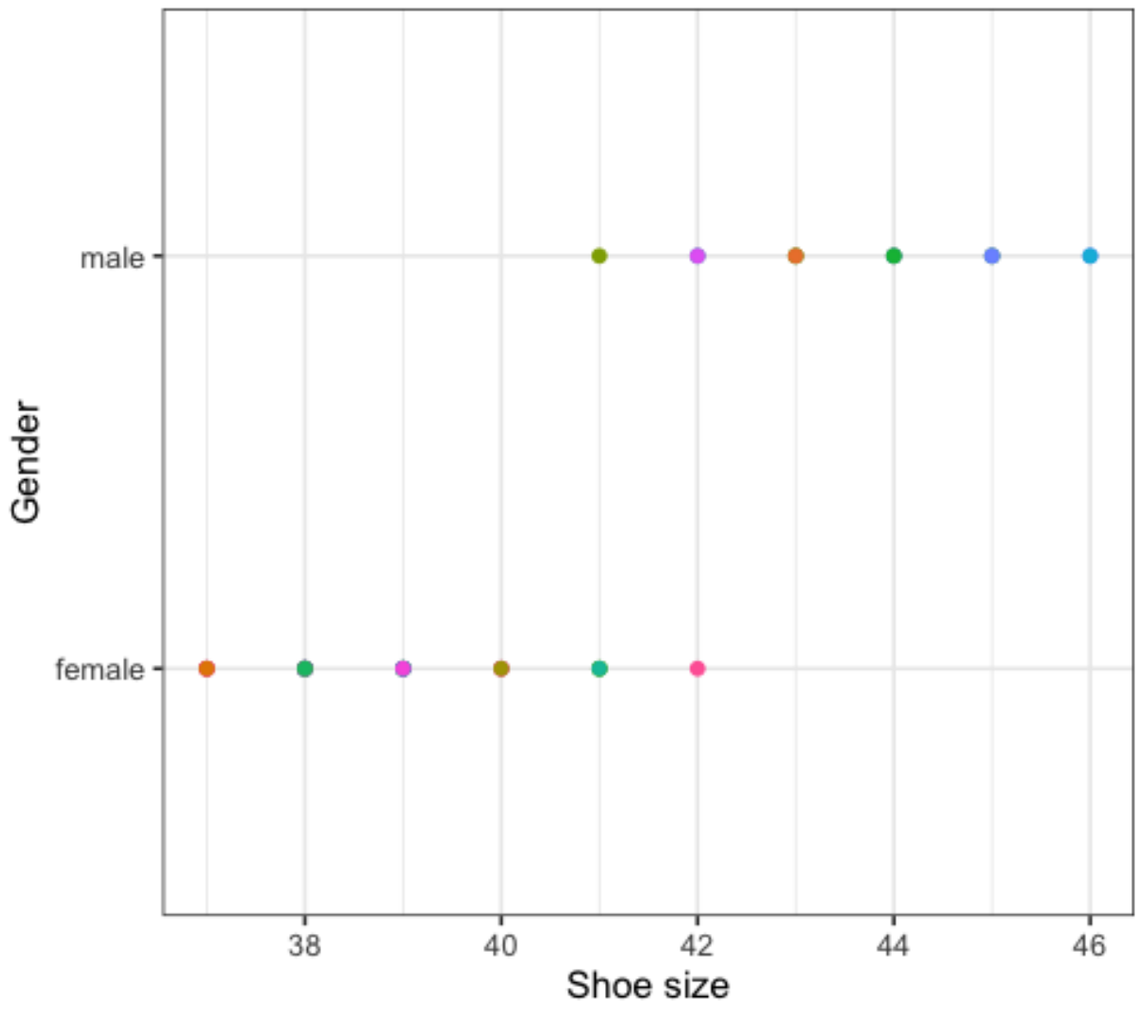




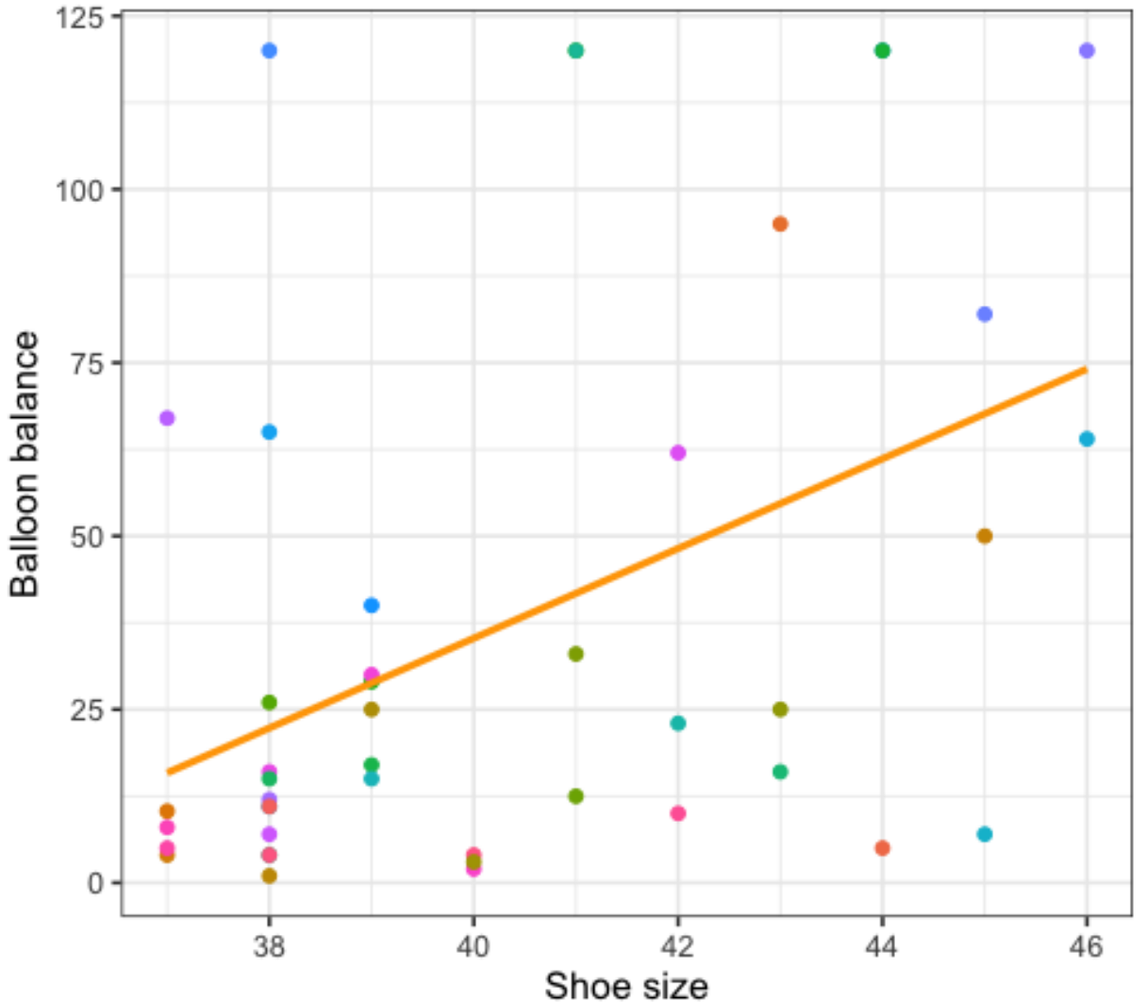

```
> car::durbinWatsonTest(model)

 lag Autocorrelation D-W Statistic p-value
   1      0.03589982      1.923831   0.484
 Alternative hypothesis: rho != 0
```

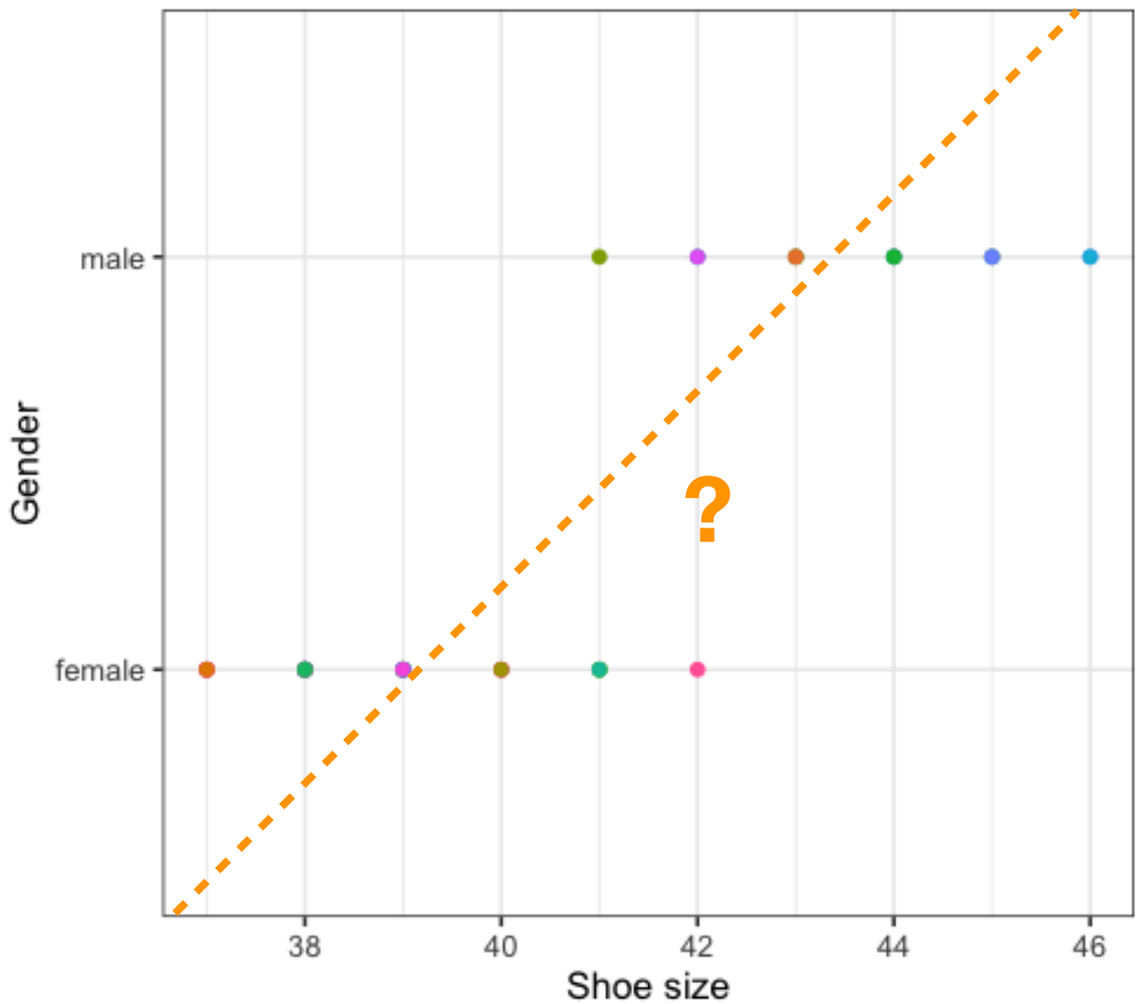
Continuous vs binary outcomes



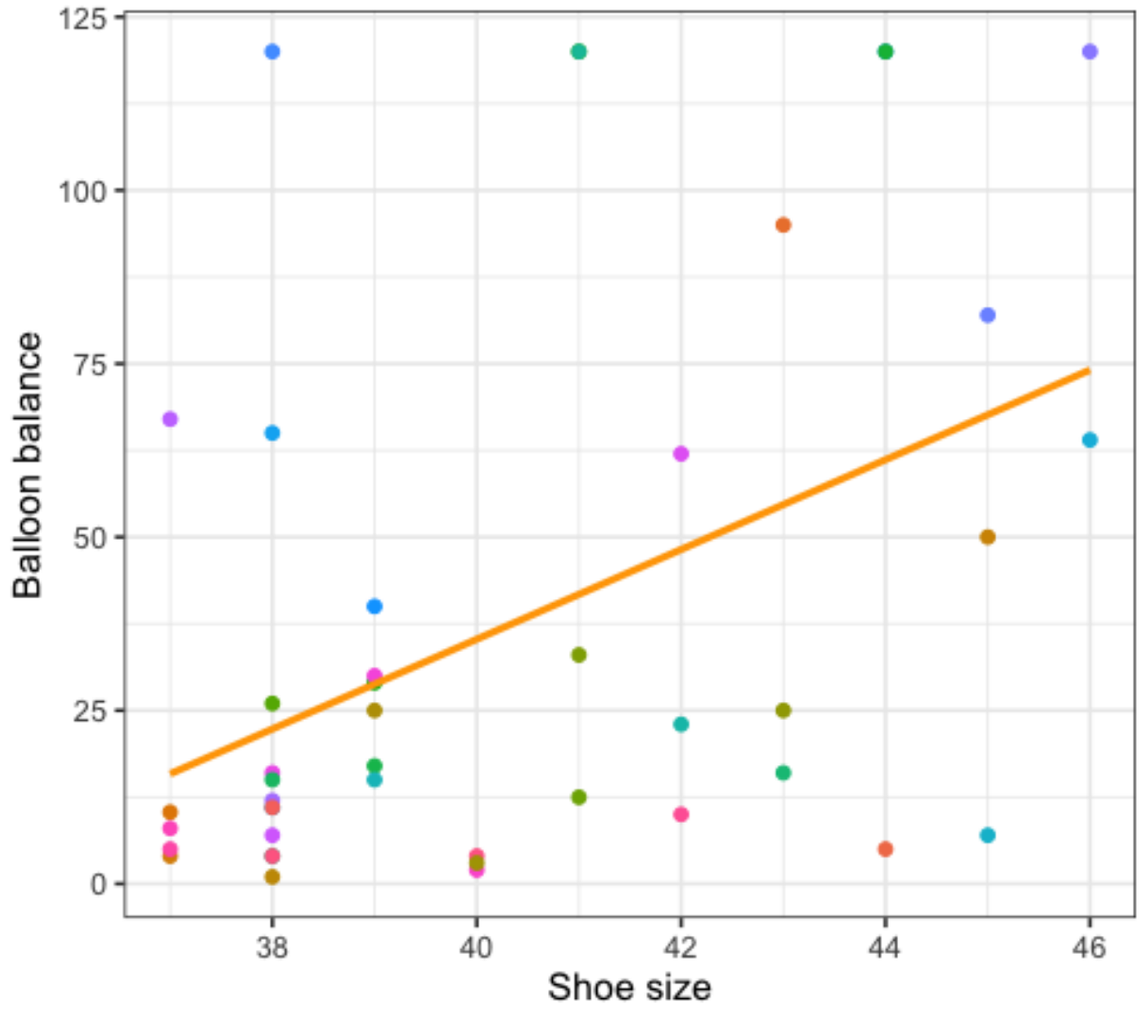


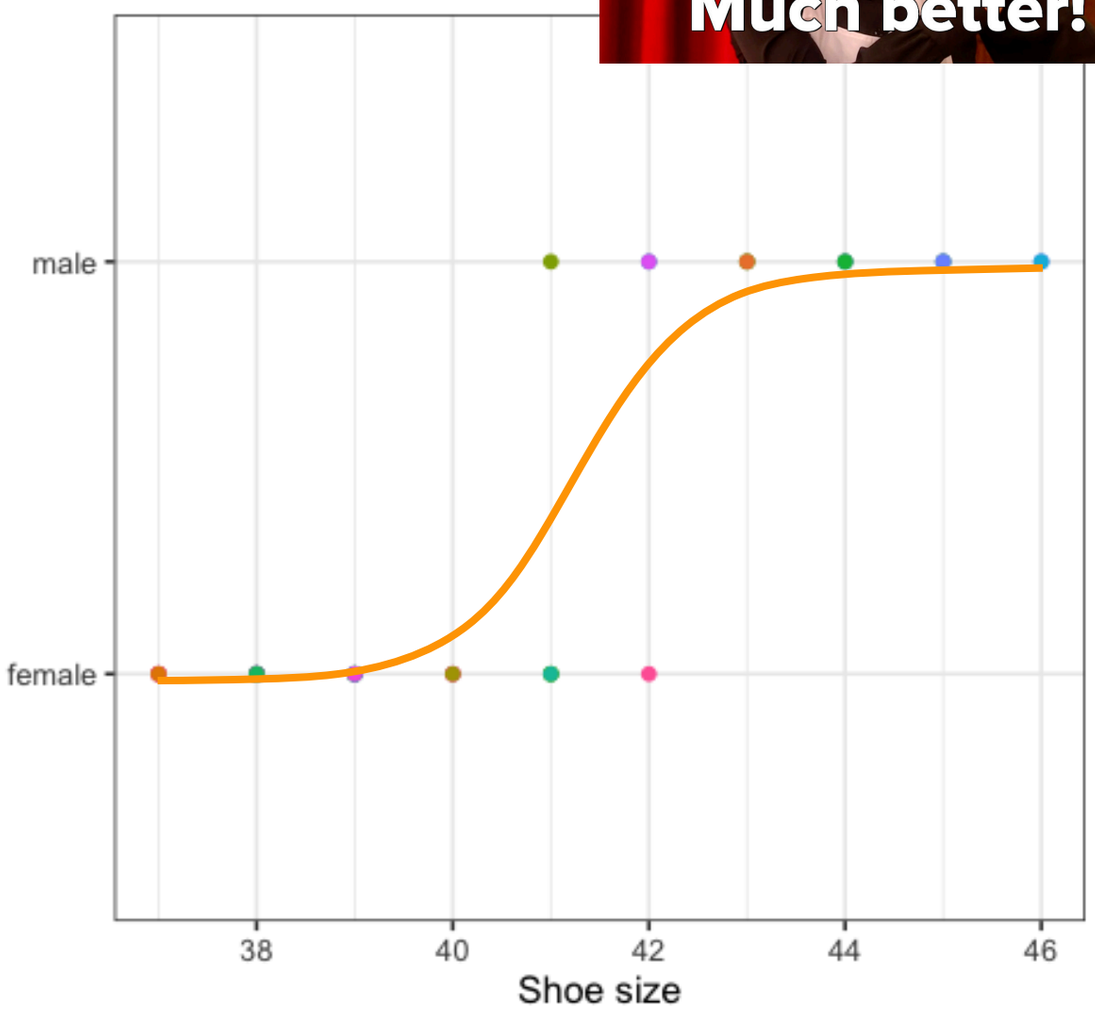
Continuous vs binary outcomes





Continuous vs binary outcomes





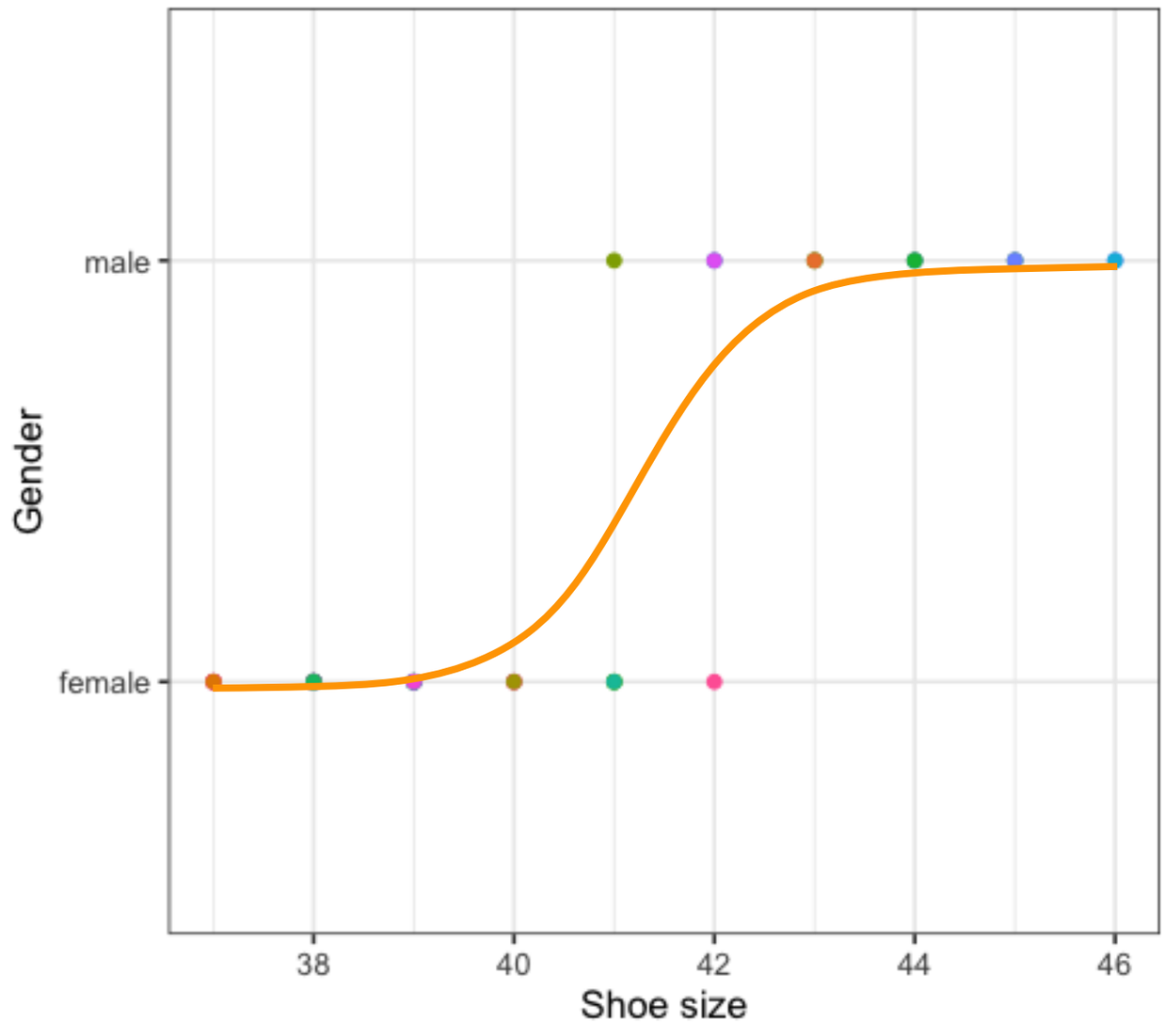
Logistic regression

- Generalized (= expanded) regression model that accommodates binomial (= binary) outcomes

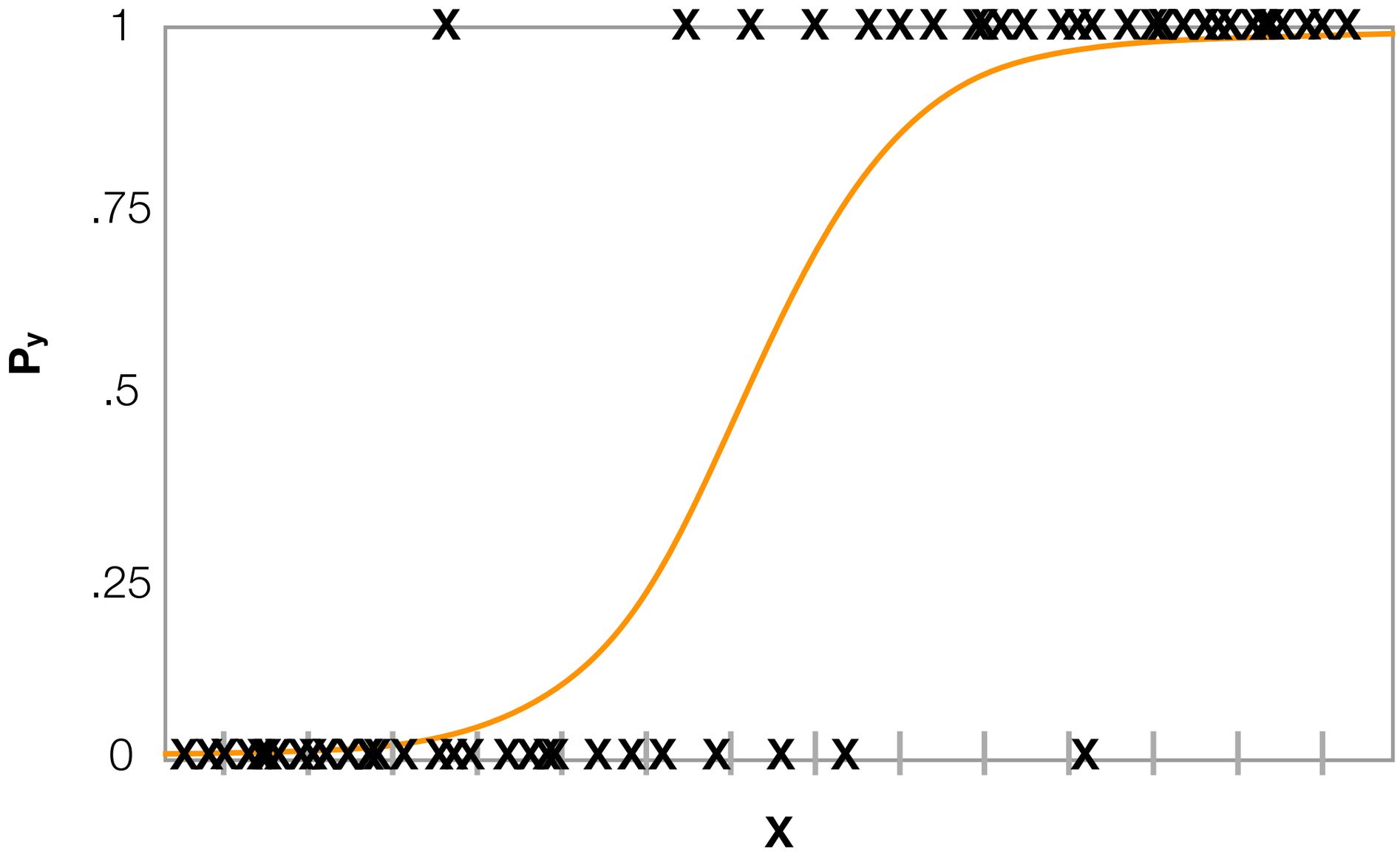
- We predict a binary outcome from continuous or categorical predictors

- Fit a squiggly line (sigmoid function)

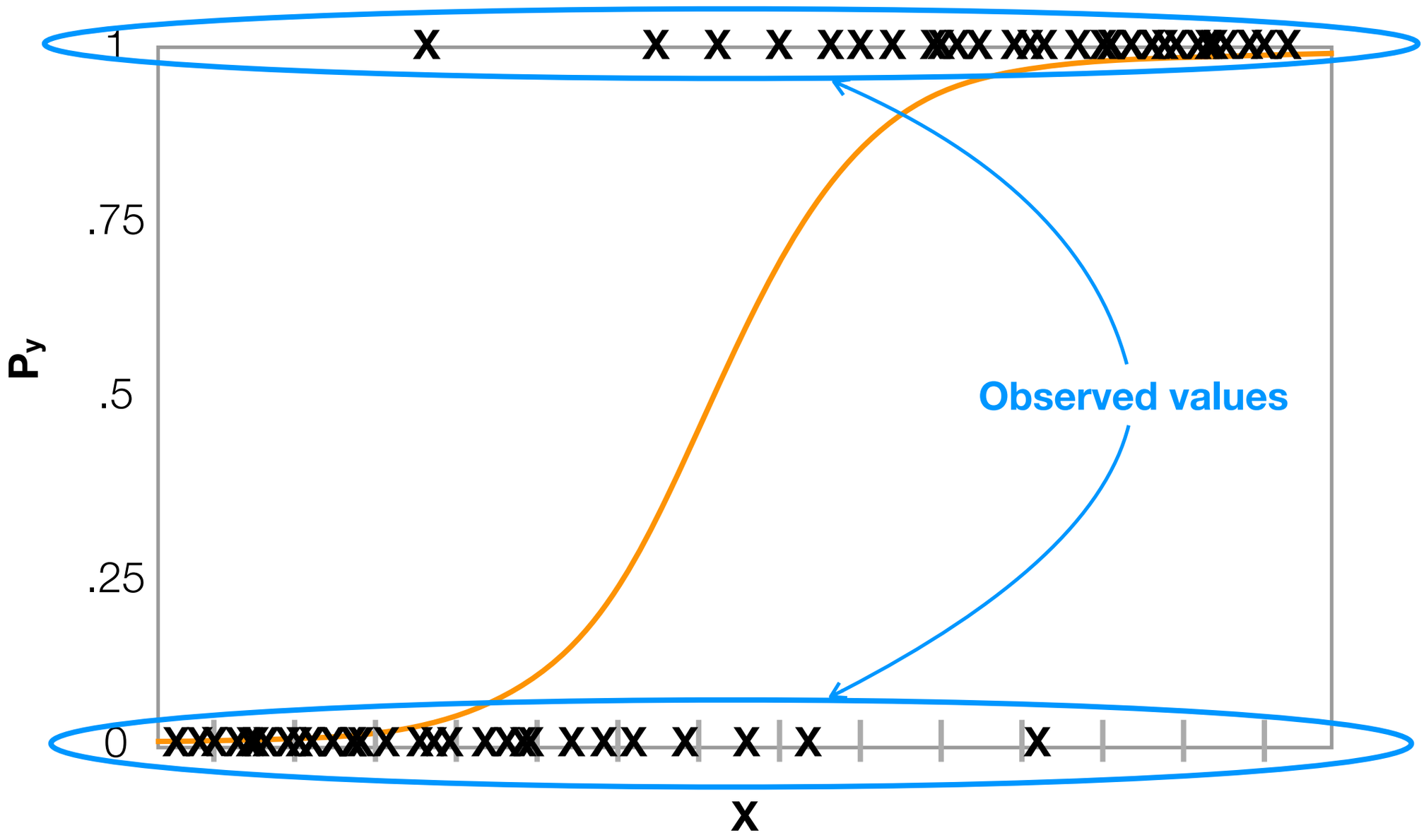
- What is the **probability** of y given different levels of x?



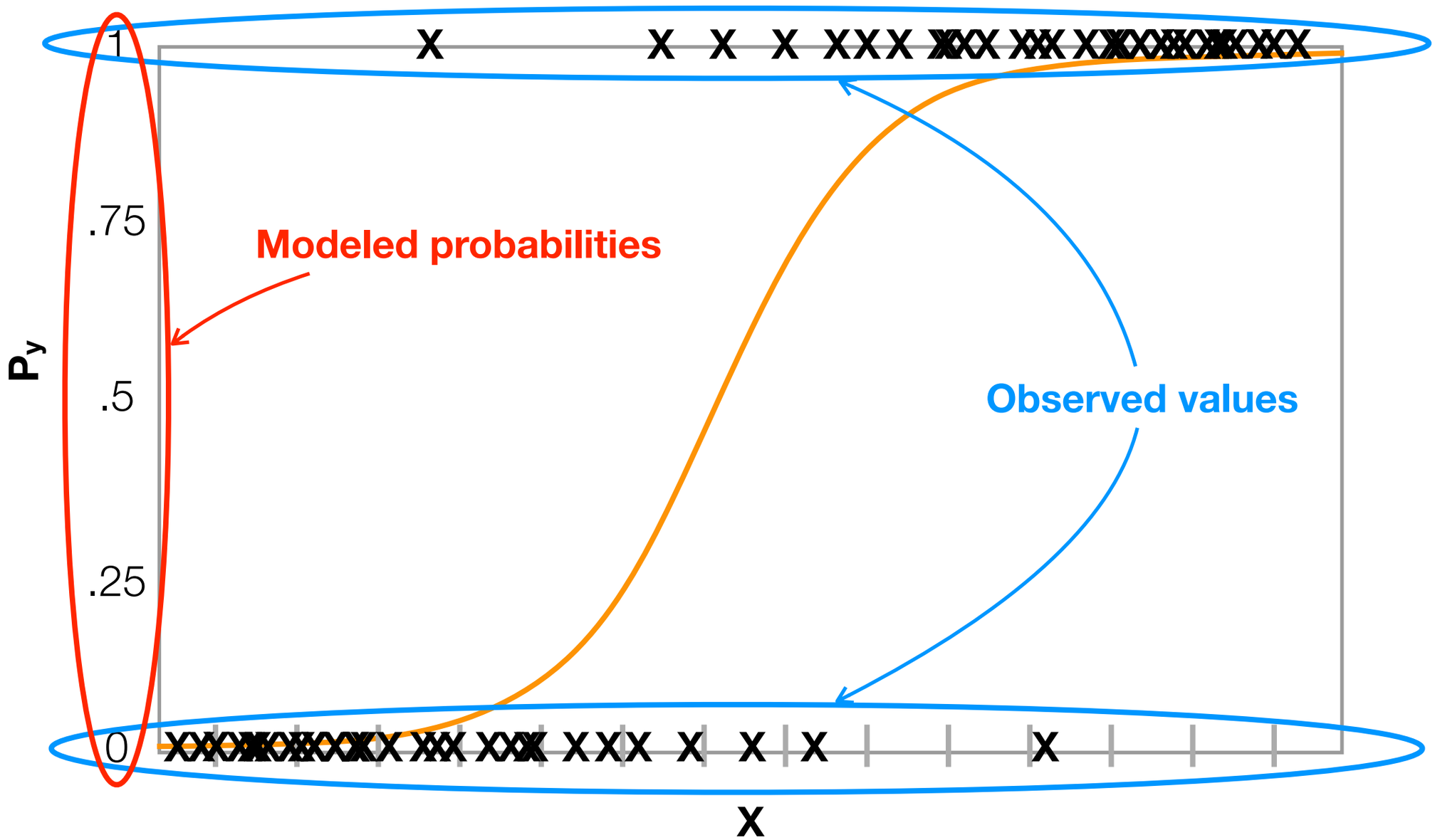
Output as probability (0,1)



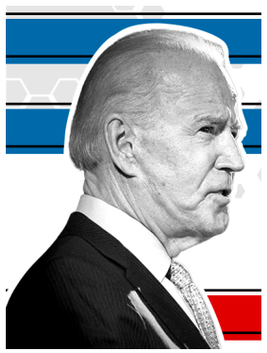
Output as probability (0,1)

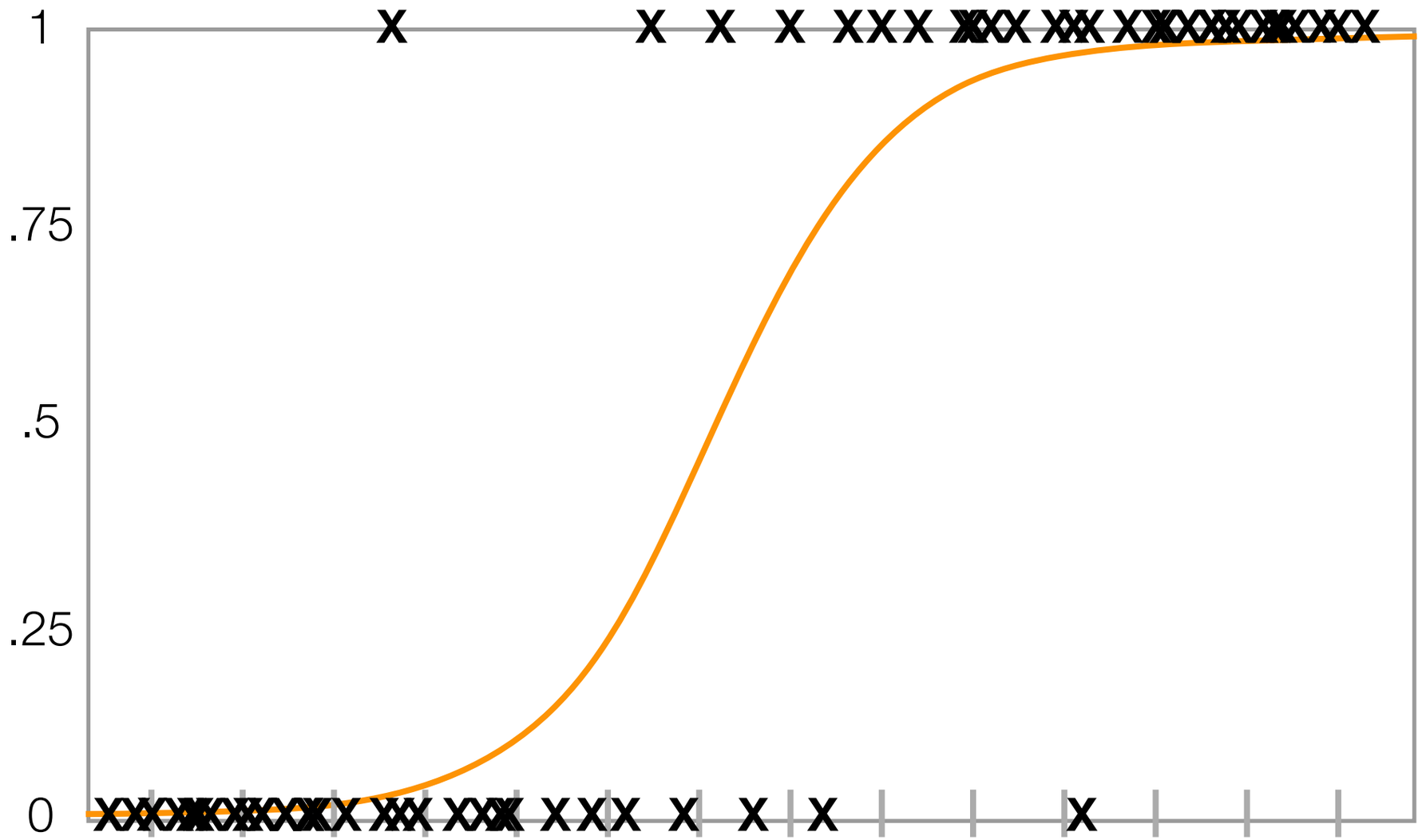


Output as probability (0,1)

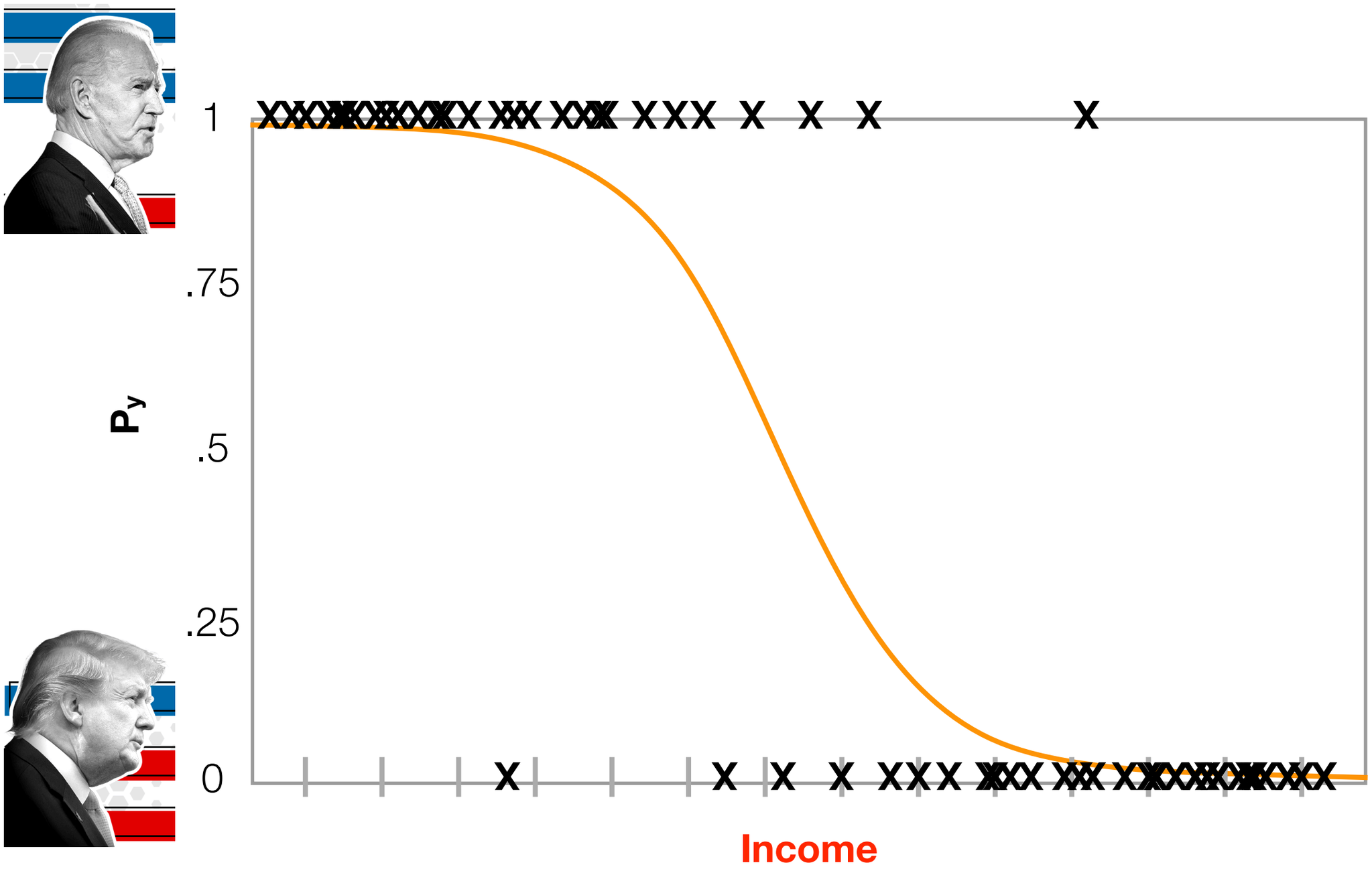


Output as probability (0,1)



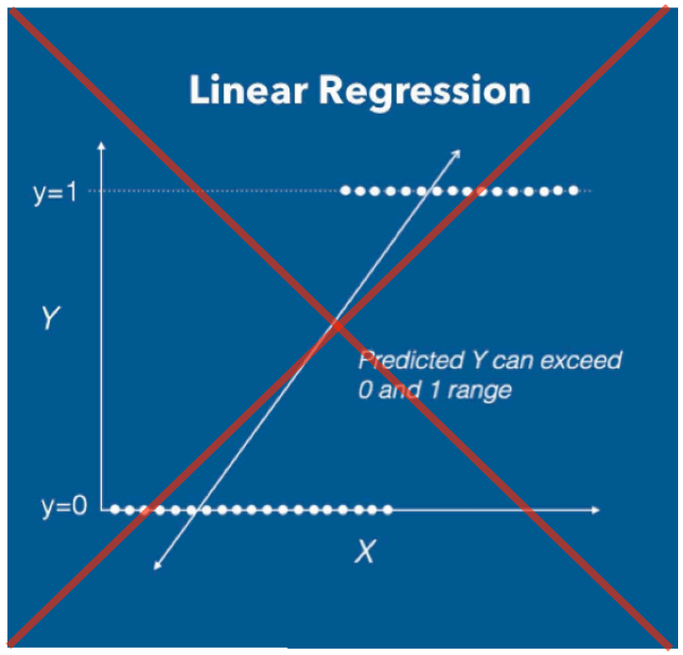


Output as probability (0,1)

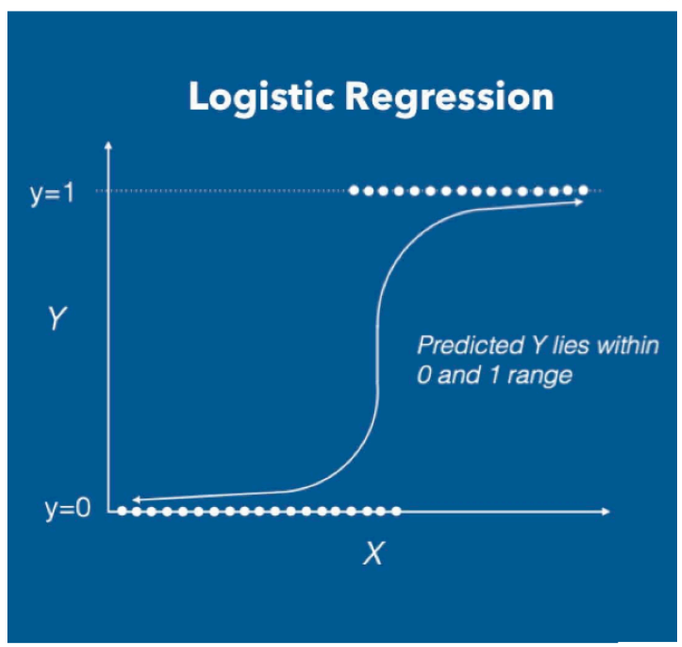


Linear vs logistic regression

- **Linear regression**:
  - gives the predicted mean value of an outcome variable at a particular value of a predictor variable



- **Logistic regression**:
  - gives the conditional probability that an outcome variable equals one at a particular value of a predictor variable
  - conditional probability = $P(Y_i \mid X_i)$



R syntax for simple logistic regression (1)

- `summary(glm(outcome ~ predictor, family = binomial(link = logit), data = data)`
- `ElectionOutcome ~ Education, data`
- `ElectionOutcome ~ Income, data`
- `gender ~ shoesize, data`

- Note: the `link = logit` parameter is default and will work even though you don’t specify it explicitly

R syntax for simple logistic regression (2)

```
> summary(glm(gender_N ~ shoesize, family = binomial(link = logit), data))

Call:
glm(formula = gender_N ~ shoesize, family = binomial(link = logit),
    data = data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.66609  -0.05431  -0.01526   0.01807   1.81759

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -105.570     49.470  -2.134   0.0328 *
shoesize       2.540      1.192   2.131   0.0331 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 55.0432  on 43  degrees of freedom
Residual deviance:  8.8056  on 42  degrees of freedom
AIC: 12.806

Number of Fisher Scoring iterations: 8
```

Gender_N:

Female = 0

Male = 1

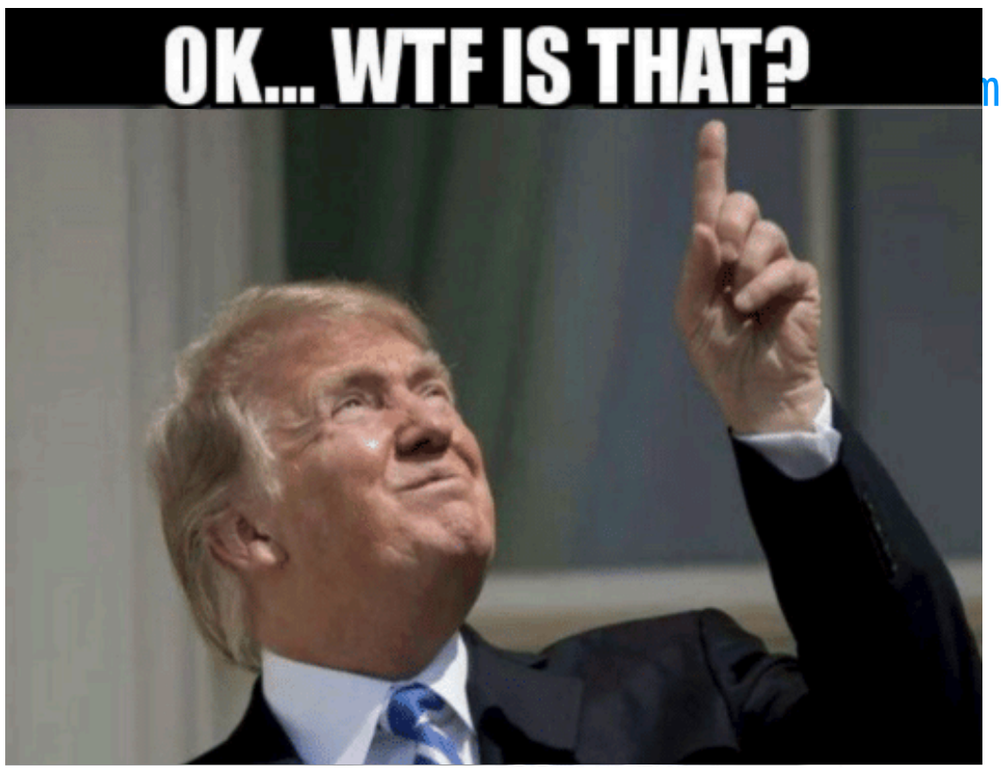
R syntax for repeated-measures logistic regression

- `lme4::glmer(..formula.., family = binomial(link = logit), data)`

- `ReadingTime_above_1s ~ condition + (1 + trial | ID), family, data)`

R syntax for repeated-measures logistic regression

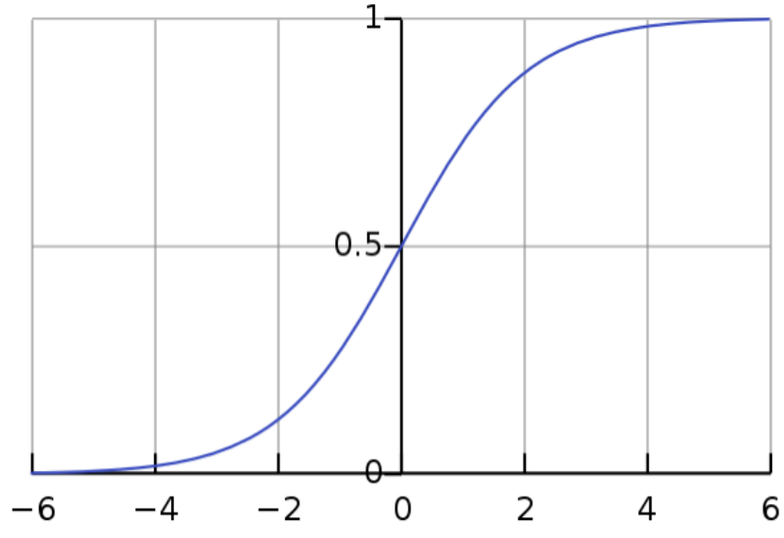
- `lme4::glmer(..formula.., family = binomial(link = logit), data)`
- `ReadingTime_above_1s ~ ... mily, data)`



How do we go from linear to logistic? (1)

- To predict the conditional probability, we use the **logistic function**

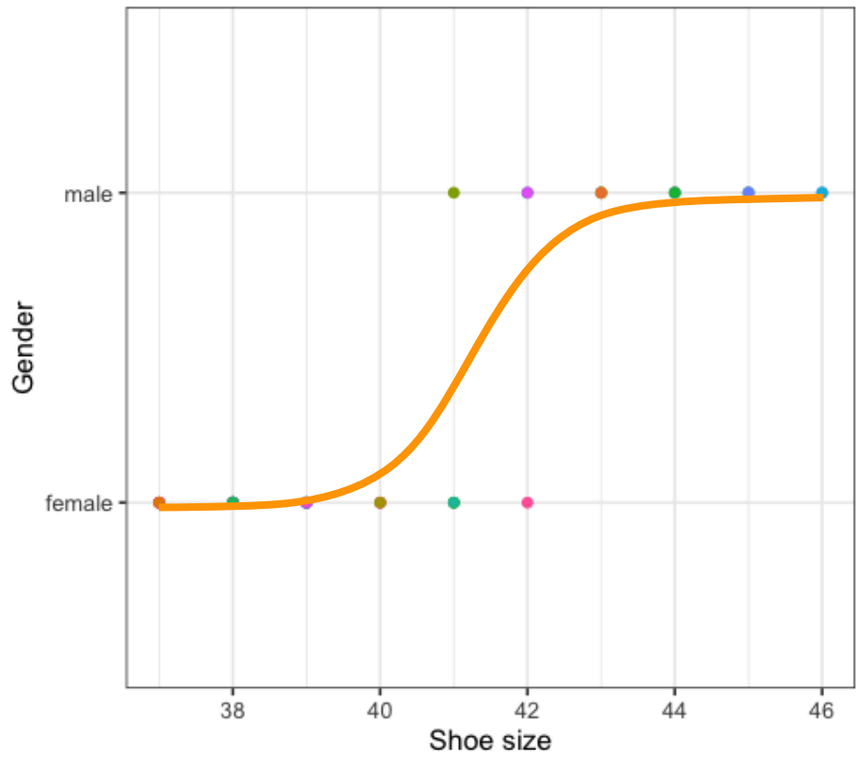
- Logistic function = sigmoid function (“squiggly line”)



- Describes the relationship between X_i and $P(Y_i = 1)$

- $$P(Y_i = 1) = \frac{e^{(\beta_0+\beta_1 X_i+\varepsilon_i)}}{1 + e^{(\beta_0+\beta_1 X_i+\varepsilon_i)}}$$

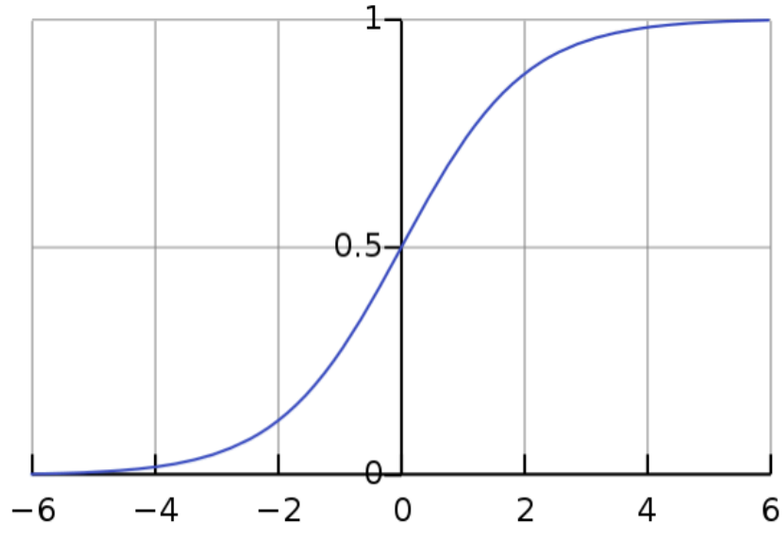
- where $e^{(\beta_0+\beta_1 X_i+\varepsilon_i)}$ is the linear regression equation expressed in the logit scale



How do we go from linear to logistic? (1)

- To predict the conditional probability, we use the **logistic function**

- Logistic function = sigmoid function ("squiggly line")



- Describes the relationship between X_i and $P(Y_i = 1)$

- $$P(Y_i = 1) = \frac{e^{(\beta_0+\beta_1 X_i+\varepsilon_i)}}{1 + e^{(\beta_0+\beta_1 X_i+\varepsilon_i)}}$$

- where $e^{(\beta_0+\beta_1 X_i+\varepsilon_i)}$ is the linear regress equation expressed in the logit scale

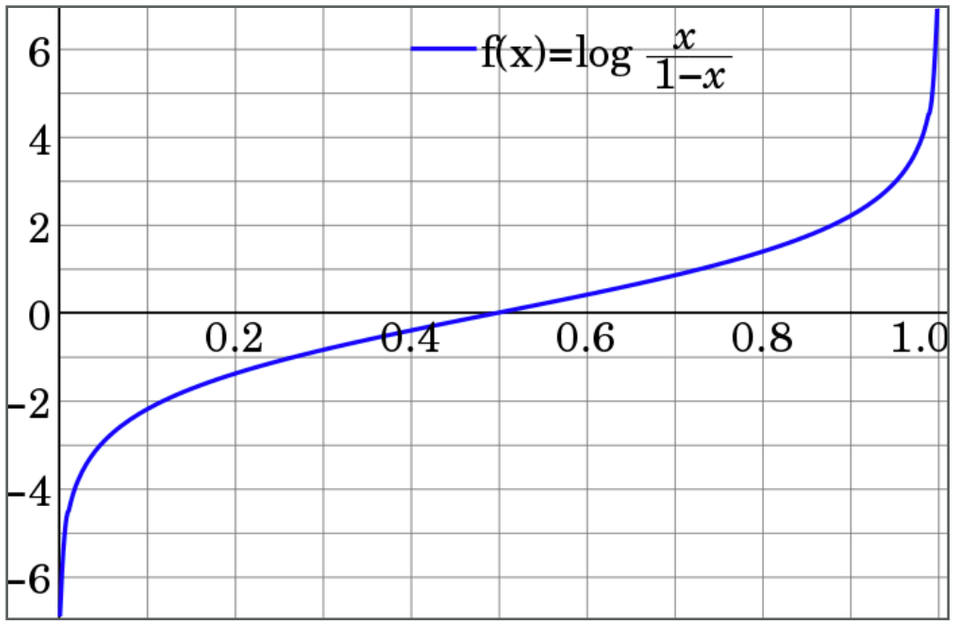
How do we go from linear to logistic? (2)

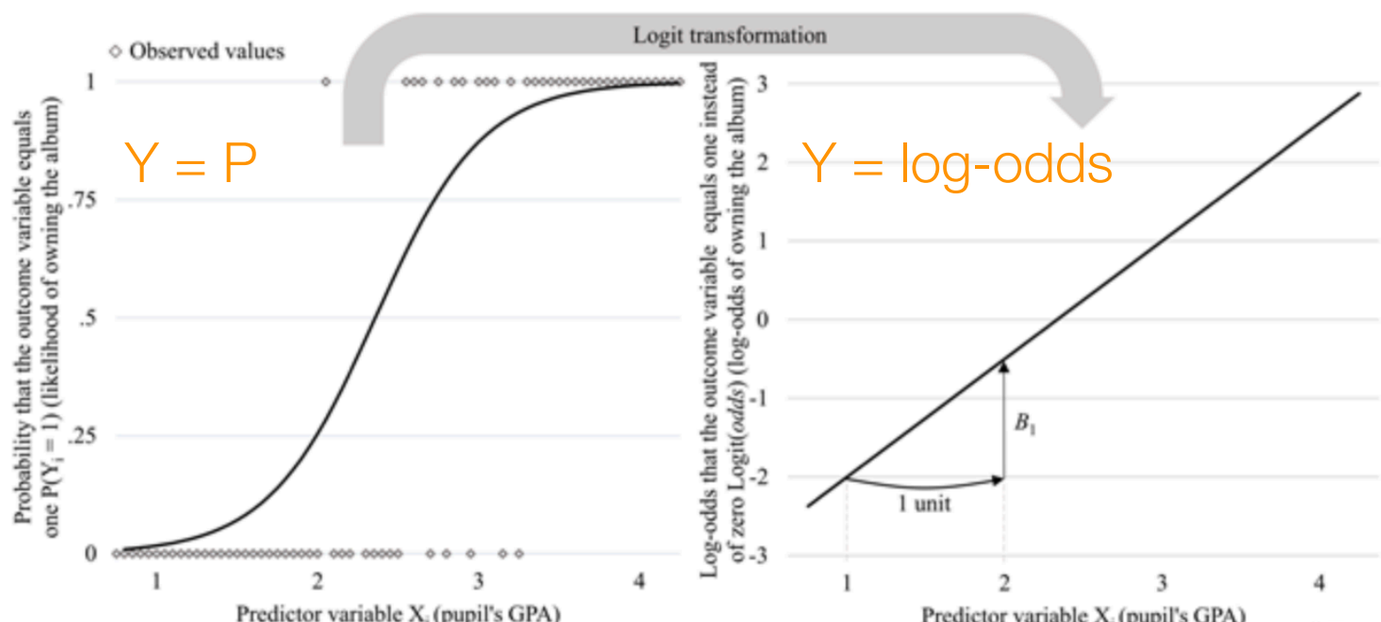
- Problem: logistic function violates assumption of linearity in linear regression

- **Logit transformation** → convert sigmoid into straight line

- **Logit-link function**: links values on predictor scale to values of P

- logit(p) =
  = log-odds(p) =
  $= \log_e(\frac{p}{1-p}) =$
  $= \beta_0 + \beta_1 X_1$





TRULY MINDBLOWING

How do we go from linear to logistic? (3)

- We assume a linear relationship between the predictor variable and the probability (0,1) of the event that Y = 1

- We assume a linear relationship between the predictor variable and the log-odds of the event that Y = 1

Binomial logistic regression

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ (simple linear regression)

- Logistic regression with one predictor:

- $P(Y_i) = \frac{e^{(\beta_0+\beta_1 X_i+\varepsilon_i)}}{1 + e^{(\beta_0+\beta_1 X_i+\varepsilon_i)}}$

Binomial logistic regression

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ (simple linear regression)
- Logistic regression with one predictor:
- $$P(Y_i) = \frac{e^{(\beta_0+\beta_1 X_i+\varepsilon_i)}}{1 + e^{(\beta_0+\beta_1 X_i+\varepsilon_i)}}$$

Binomial logistic regression

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ (simple linear regression)

- Logistic regression with one predictor:

- $$P(Y_i) = \frac{e^{(\beta_0+\beta_1 X_i+\varepsilon_i)}}{1 + e^{(\beta_0+\beta_1 X_i+\varepsilon_i)}}$$

Linear regression formula

Base of natural logarithms (constant)

Multiple binomial logistic regression

- With multiple predictors:

- $$P(Y_i) = \frac{e^{(\beta_0+\beta_1 X_i+\beta_2 X_i+...+\beta_0+\beta_n X_i+\varepsilon_i)}}{1+e^{(\beta_0+\beta_1 X_i+\beta_2 X_i+...+\beta_0+\beta_n X_i+\varepsilon_i)}}$$

Multiple binomial logistic regression

- With multiple predictors:

- $$P(Y_i) = \frac{e^{(\beta_0+\beta_1 X_i+\beta_2 X_i+\ldots+\beta_0+\beta_n X_i+\varepsilon_i)}}{1 + e^{(\beta_0+\beta_1 X_i+\beta_2 X_i+\ldots+\beta_0+\beta_n X_i+\varepsilon_i)}}$$

Multiple regression

Interpreting the output of logistic regression (1)

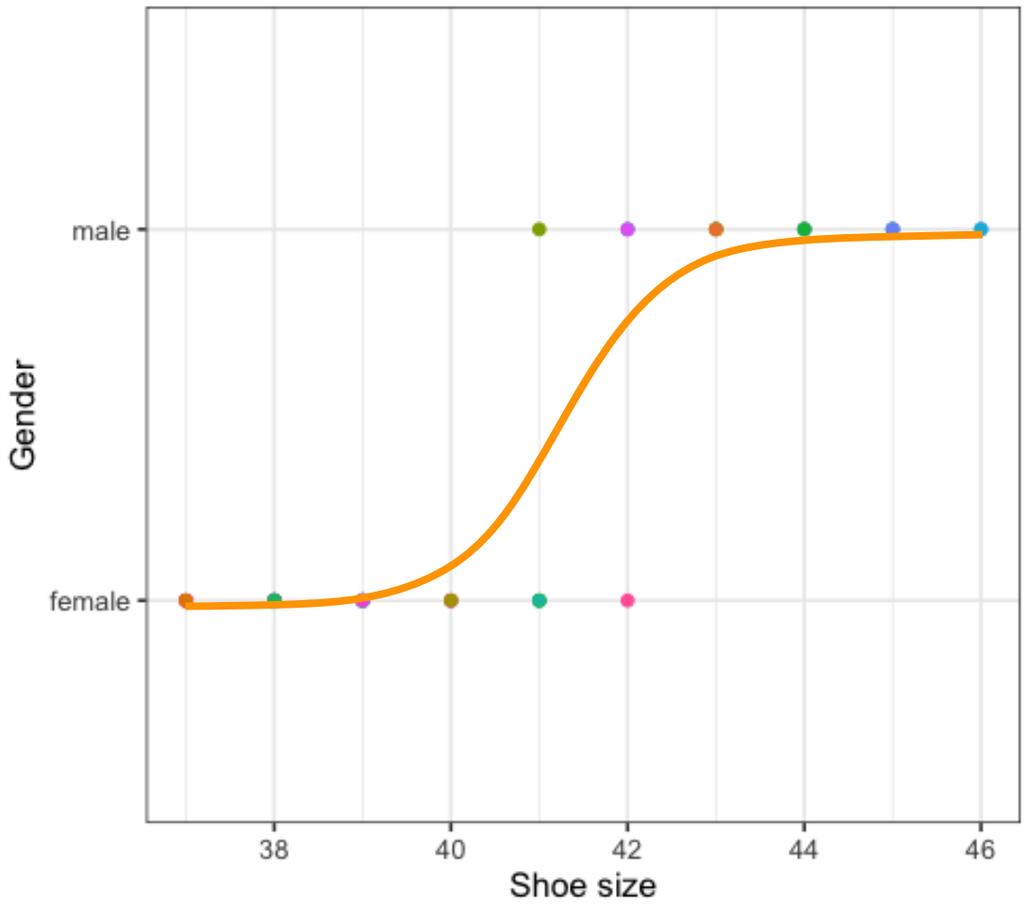
- Estimates in logistic regression are on the log scale (log-odds)
- Odds $= \frac{p}{1-p}$
  - Probability ranges from 0 to 1
  - Odds range from 0 to ∞
  - Log Odds range from −∞ to ∞
- **Odds ratio: change in the odds resulting from a unit change in the predictor**
  - Positive odds = increase in odds
  - Negative odds = decrease in odds
- `boot::inv.logit(x)` brings estimates back to probabilities

```
> summary(glm(gender_N ~ shoesize, family = binomial(link = logit),
data))

Call:
glm(formula = gender_N ~ shoesize, family = binomial(link = logit),
    data = data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.66609  -0.05431  -0.01526   0.01807   1.81759

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -105.570     49.470  -2.134   0.0328 *
shoesize       2.540      1.192   2.131   0.0331 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 55.0432  on 43  degrees of freedom
Residual deviance:  8.8056  on 42  degrees of freedom
AIC: 12.806

Number of Fisher Scoring iterations: 8
```

Interpreting the output of logistic regression (2)

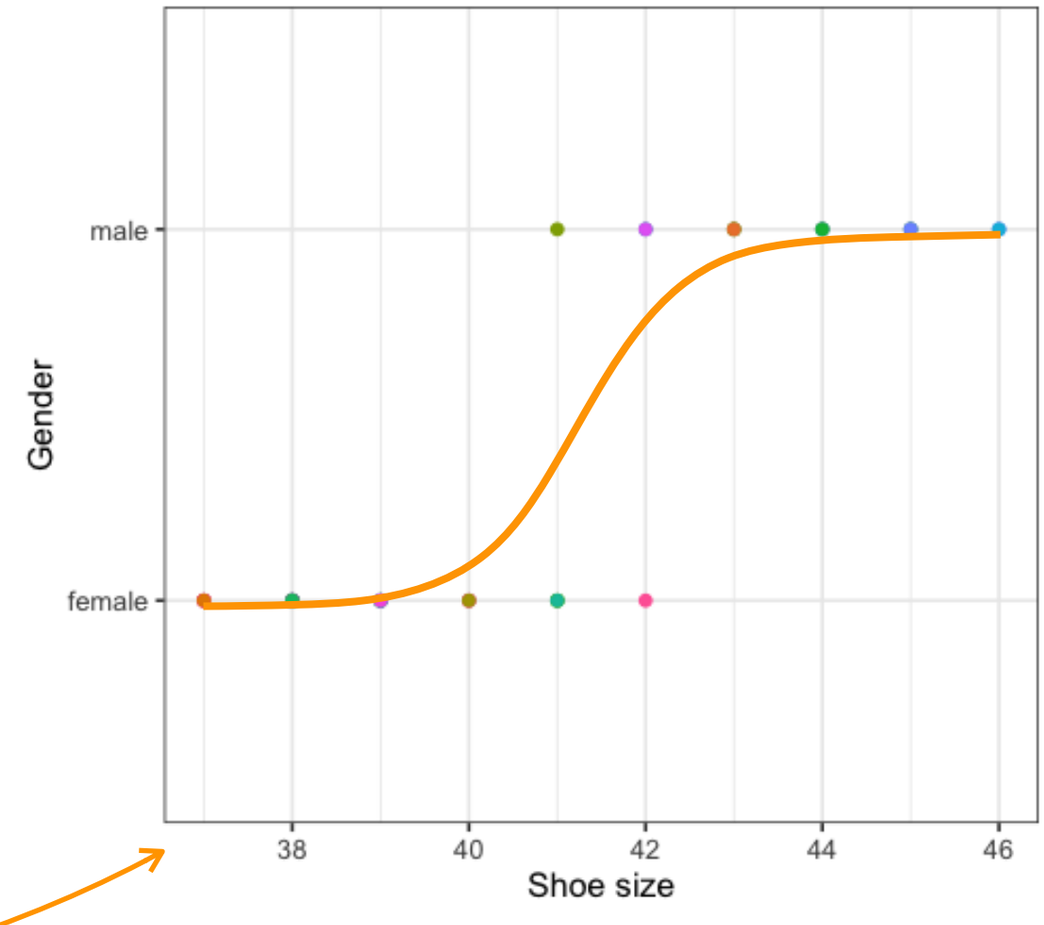
```
> summary(glm(gender_N ~ shoesize, family = binomial(link = logit),
data))

Call:
glm(formula = gender_N ~ shoesize, family = binomial(link = logit),
    data = data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.66609  -0.05431  -0.01526   0.01807   1.81759

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -105.570     49.470  -2.134   0.0328 *
shoesize       2.540      1.192   2.131   0.0331 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 55.0432  on 43  degrees of freedom
Residual deviance:  8.8056  on 42  degrees of freedom
AIC: 12.806

Number of Fisher Scoring iterations: 8
```

```
> boot::inv.logit(2.54)
[1] 0.9268988

> boot::inv.logit(105.57)
[1] 0
```



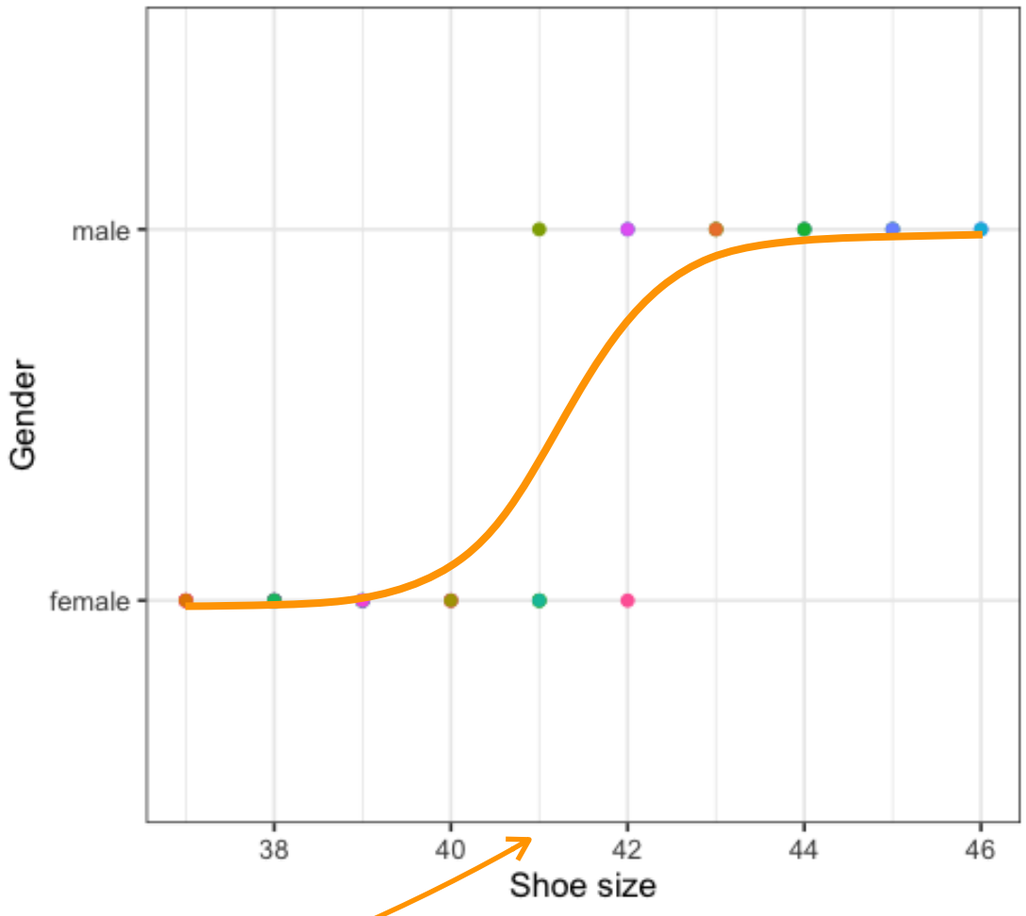
Interpreting the output of logistic regression (2)

```
> summary(glm(gender_N ~ shoesize, family = binomial(link = logit),
data))

Call:
glm(formula = gender_N ~ shoesize, family = binomial(link = logit),
    data = data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.66609  -0.05431  -0.01526   0.01807   1.81759

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -105.570     49.470  -2.134   0.0328 *
shoesize       2.540      1.192   2.131   0.0331 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 55.0432  on 43  degrees of freedom
Residual deviance:  8.8056  on 42  degrees of freedom
AIC: 12.806

Number of Fisher Scoring iterations: 8


> boot::inv.logit(-105.57+(36*2.54))
[1] 0

> boot::inv.logit(-105.57+(41*2.54))
[1] 0.1930987

> boot::inv.logit(-105.57+(43*2.54))
[1] 0.9746673

> boot::inv.logit(-105.57+(46*2.54))
[1] 0.9999873
```



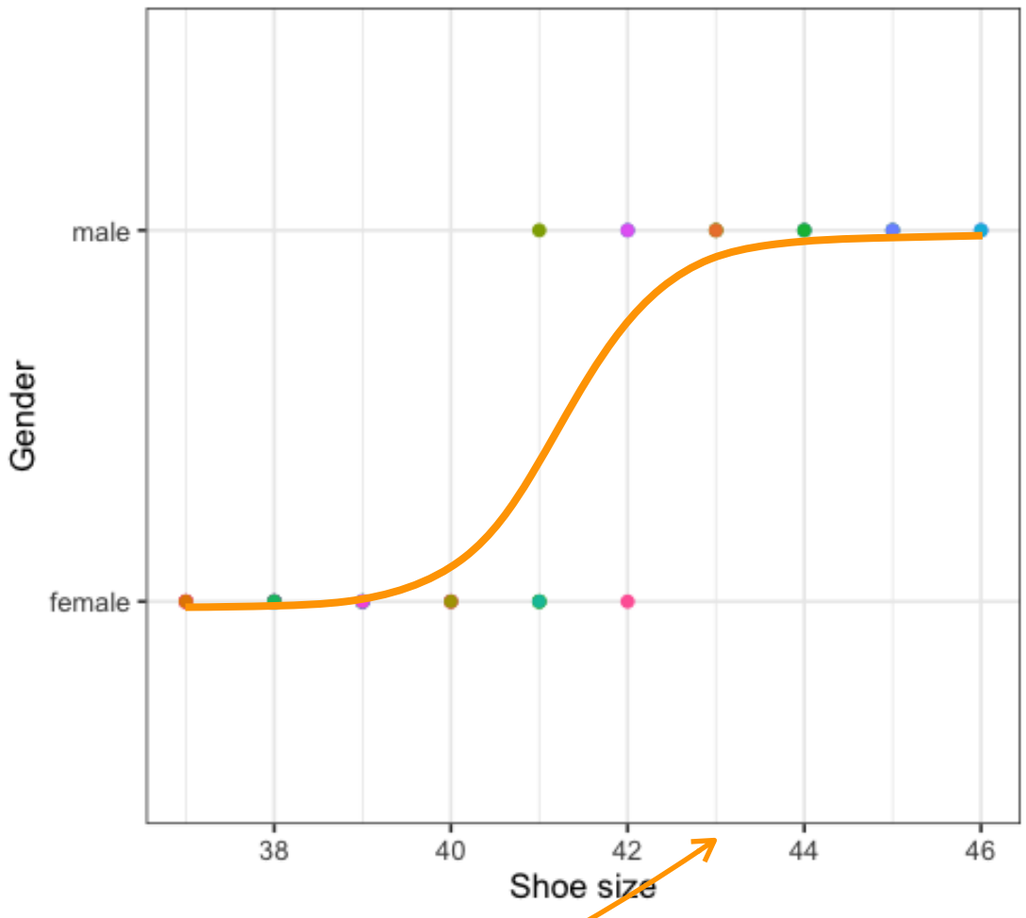
Interpreting the output of logistic regression (2)

```
> summary(glm(gender_N ~ shoesize, family = binomial(link = logit),
data))

Call:
glm(formula = gender_N ~ shoesize, family = binomial(link = logit),
    data = data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.66609  -0.05431  -0.01526   0.01807   1.81759

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -105.570     49.470  -2.134   0.0328 *
shoesize       2.540      1.192   2.131   0.0331 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 55.0432  on 43  degrees of freedom
Residual deviance:  8.8056  on 42  degrees of freedom
AIC: 12.806

Number of Fisher Scoring iterations: 8


> boot::inv.logit(-105.57+(36*2.54))
[1] 0

> boot::inv.logit(-105.57+(41*2.54))
[1] 0.1930987

> boot::inv.logit(-105.57+(43*2.54))
[1] 0.9746673

> boot::inv.logit(-105.57+(46*2.54))
[1] 0.9999873
```



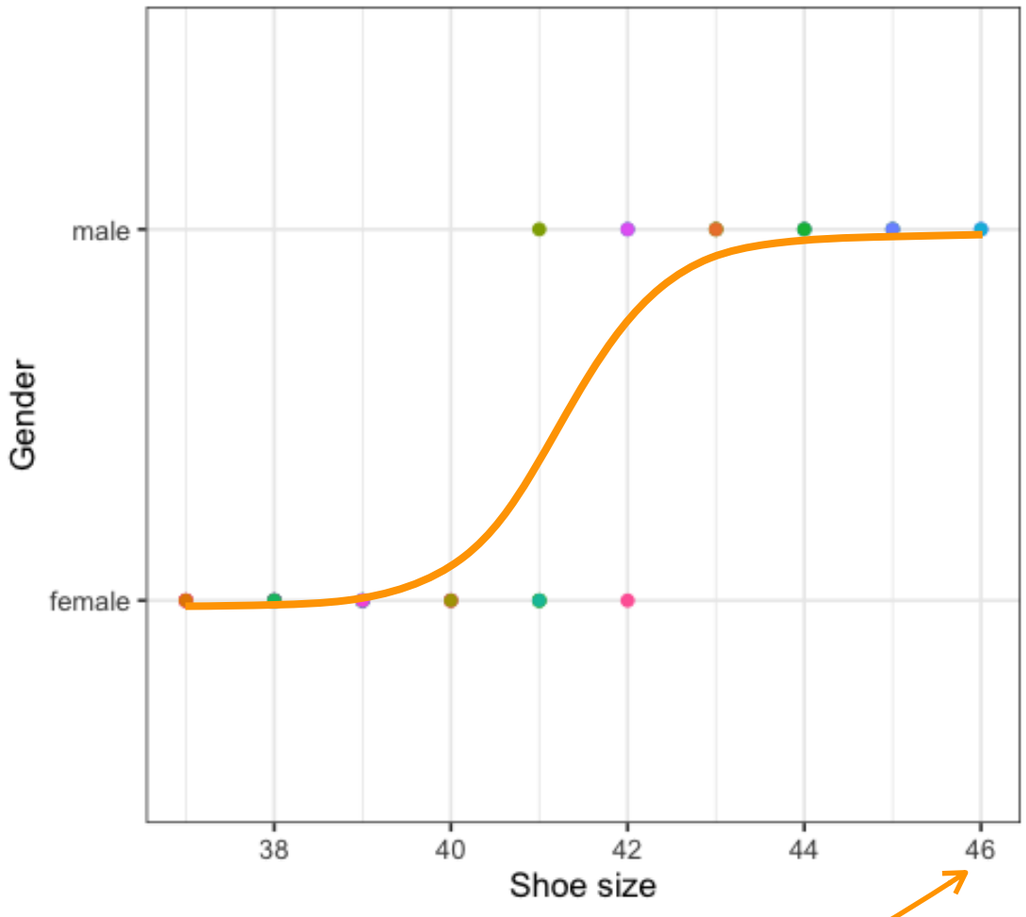
Interpreting the output of logistic regression (2)

```
> summary(glm(gender_N ~ shoesize, family = binomial(link = logit),
data))

Call:
glm(formula = gender_N ~ shoesize, family = binomial(link = logit),
    data = data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.66609  -0.05431  -0.01526   0.01807   1.81759

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -105.570     49.470  -2.134   0.0328 *
shoesize       2.540      1.192   2.131   0.0331 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 55.0432  on 43  degrees of freedom
Residual deviance:  8.8056  on 42  degrees of freedom
AIC: 12.806

Number of Fisher Scoring iterations: 8


> boot::inv.logit(-105.57+(36*2.54))
[1] 0

> boot::inv.logit(-105.57+(41*2.54))
[1] 0.1930987

> boot::inv.logit(-105.57+(43*2.54))
[1] 0.9746673

> boot::inv.logit(-105.57+(46*2.54))
[1] 0.9999873
```



Interpreting the output of logistic regression (2)

```
> summary(glm(gender_N ~ shoesize, family = binomial(link = logit),
data))

Call:
glm(formula = gender_N ~ shoesize, family = binomial(link = logit),
    data = data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.66609  -0.05431  -0.01526   0.01807   1.81759

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -105.570     49.470  -2.134   0.0328 *
shoesize       2.540      1.192   2.131   0.0331 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 55.0432  on 43  degrees of freedom
Residual deviance:  8.8056  on 42  degrees of freedom
AIC: 12.806

Number of Fisher Scoring iterations: 8


> boot::inv.logit(-105.57+(36*2.54))
[1] 0

> boot::inv.logit(-105.57+(41*2.54))
[1] 0.1930987

> boot::inv.logit(-105.57+(43*2.54))
[1] 0.9746873

> boot::inv.logit(-105.57+(46*2.54))
[1] 0.9999873
```



Assessing fit of the model (1)

- **Log-likelihood statistic:**
  - how much unexplained information left after model is fitted
  - large values = poor fit
- **Deviance (–2 × log-likelihood):**
  - Higher number = worse fit
  - Follows a chi-square distribution
  - χ^2 = deviance(H_0) – deviance(H_1)
  - equivalent to F-ratio in linear regression

```
> summary(glm(gender_N ~ shoesize, family = binomial(link = logit),
data))

Call:
glm(formula = gender_N ~ shoesize, family = binomial(link = logit),
    data = data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.66609  -0.05431  -0.01526   0.01807   1.81759

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -105.570     49.470  -2.134   0.0328 *
shoesize       2.540      1.192   2.131   0.0331 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 55.0432  on 43  degrees of freedom
Residual deviance:  8.8056  on 42  degrees of freedom
AIC: 12.806

Number of Fisher Scoring iterations: 8
```

```
> logLik(model_name)
'log Lik.' -4.402805 (df=2)
```

Assessing fit of the model (2)

```
> summary(lme4::glmer(Gender ~ ShoeSize + (1 | ID), data = df, family = binomial(link = logit)))
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: binomial  ( logit )
Formula: Gender ~ ShoeSize + (1 | ID)
   Data: df

     AIC      BIC   logLik deviance df.resid
    19.5     25.9     -6.7     13.5       59

Scaled residuals:
      Min        1Q    Median        3Q       Max
-0.076428 -0.000004  0.000000  0.000000  0.060368

Random effects:
 Groups Name        Variance Std.Dev.
 ID     (Intercept) 5055     71.1
Number of obs: 62, groups:  ID, 62

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -505.443     47.451  -10.65   <2e-16 ***
ShoeSize      12.318      1.152   10.69   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
         (Intr)
ShoeSize -0.998
```

Assessing fit of the model (3)

- Log-likelihood, deviance and AIC are hard to interpret
- We can derive a pseudo-R^2 (easier to interpret)
- $$R^2_{McFadden} = 1 - \frac{L_c}{L_{null}}$$

```
mod1 <- glm(y~x, family="binomial")
mod0 <- glm(y~1, family="binomial")
1-logLik(mod1)/logLik(mod0)
```

```
> mod1 <- glm(gender_N ~ shoesize, family = binomial(link = logit), data)
> mod0 <- glm(gender_N ~ 1, family = binomial(link = logit), data)
> 1-logLik(mod1)/logLik(mod0)
'log Lik.' 0.8400238 (df=2)
```

Assumptions

- Independence of residuals
- Linearity of residuals (in the logit-transformed data)
- Absence of multicollinearity
- Lack of strongly influential outliers

This was boring and I didn't listen-summary

- Logistic regression is for categorical (binary) outcome variables
- Predicts P(Y) given values of X's
- Expands upon linear regression (generalized linear model)
- Same output as linear model, but estimates are in the log-odd scale
- Very common classification algorithm in machine learning

Thursday

- Quick experiment on sound symbolism
    - the Bouba-Kiki effect
- Analysis using logistic regression