Generalised Linear Models

Logistic regression

Francisco Rodríguez-Sánchez

https://frodriguezsanchez.net

Q: Survival of passengers on the Titanic ~ Class

Read `titanic_long.csv` dataset and fit linear model (survival ~ class).

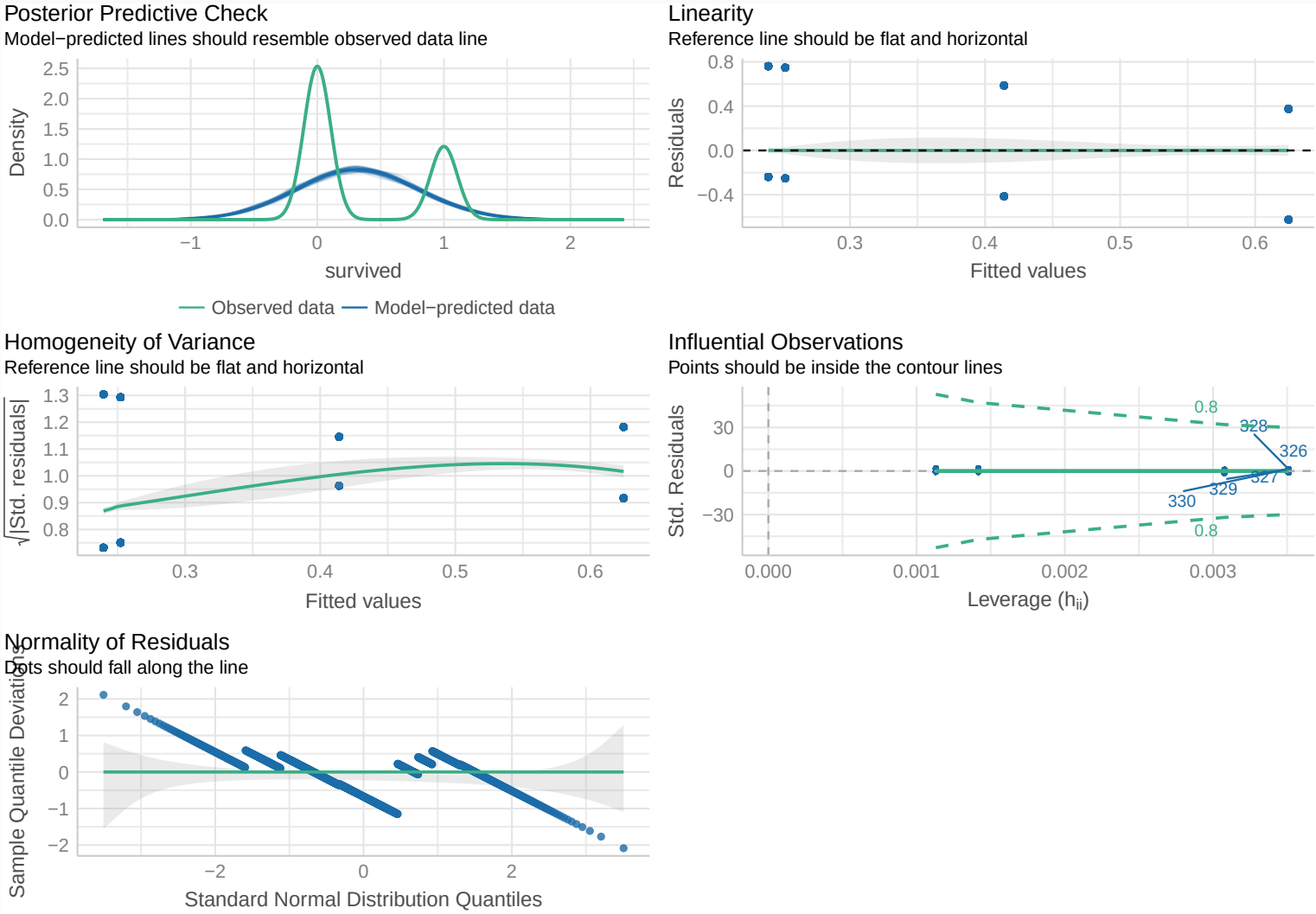
```
  class   age  sex survived
1 first adult male        1
2 first adult male        1
3 first adult male        1
4 first adult male        1
5 first adult male        1
6 first adult male        1
```

Quiz: Did passenger class influence survival?

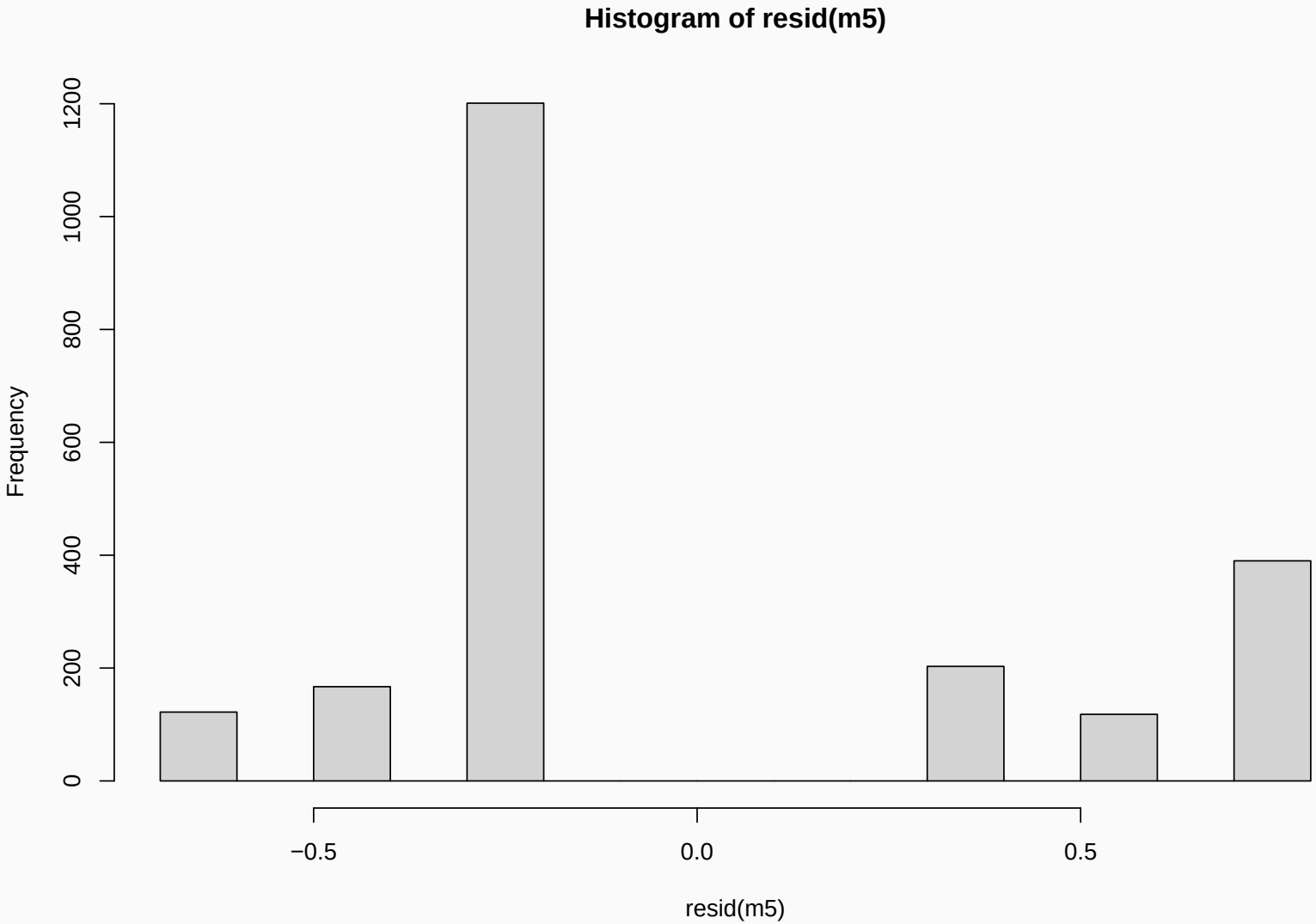
https://pollev.com/franciscorod726

Let’s check linear model:

```
m5 <- lm(survived ~ class, data = titanic)
library("easystats")
check_model(m5)
```



Weird residuals!



What if your residuals are clearly non-normal or variance not constant (heteroscedasticity)?

- Binary variables (0/1)

What if your residuals are clearly non-normal or variance not constant (heteroscedasticity)?

- Binary variables (0/1)
- Counts (0, 1, 2, 3, ...)

What if your residuals are clearly non-normal or variance not constant (heteroscedasticity)?

- Binary variables (0/1)
- Counts (0, 1, 2, 3, ...)
- Categories ("small", "medium", "large"...)

What if your residuals are clearly non-normal or variance not constant (heteroscedasticity)?

- Binary variables (0/1)
- Counts (0, 1, 2, 3, ...)
- Categories ("small", "medium", "large"...)

- **Generalised Linear Models to the rescue!**

Generalised Linear Models

1. **Response variable** - distribution `family`

Generalised Linear Models

1. **Response variable** - distribution `family`
    - Bernouilli - Binomial

Generalised Linear Models

1. **Response variable** - distribution `family`
    - Bernouilli - Binomial
    - Poisson

Generalised Linear Models

1. **Response variable** - distribution `family`
    - Bernouilli - Binomial
    - Poisson
    - Gamma

Generalised Linear Models

1. **Response variable** - distribution `family`
    - Bernouilli - Binomial
    - Poisson
    - Gamma
    - etc

Generalised Linear Models

1. **Response variable** - distribution `family`
    - Bernouilli - Binomial
    - Poisson
    - Gamma
    - etc

2. **Predictors** (continuous or categorical)

Generalised Linear Models

1. **Response variable** - distribution `family`
    - Bernouilli - Binomial
    - Poisson
    - Gamma
    - etc

2. **Predictors** (continuous or categorical)

3. **Link function**

Generalised Linear Models

1. **Response variable** - distribution `family`
    - Bernouilli - Binomial
    - Poisson
    - Gamma
    - etc

2. **Predictors** (continuous or categorical)

3. **Link function**
    - Gaussian: identity

Generalised Linear Models

1. **Response variable** - distribution `family`
    - Bernouilli - Binomial
    - Poisson
    - Gamma
    - etc

2. **Predictors** (continuous or categorical)

3. **Link function**
    - Gaussian: identity
    - Binomial: logit, probit

Generalised Linear Models

1. **Response variable** - distribution `family`
    - Bernouilli - Binomial
    - Poisson
    - Gamma
    - etc

2. **Predictors** (continuous or categorical)

3. **Link function**
    - Gaussian: identity
    - Binomial: logit, probit
    - Poisson: log...

Generalised Linear Models

1. **Response variable** - distribution `family`
    - Bernouilli - Binomial
    - Poisson
    - Gamma
    - etc

2. **Predictors** (continuous or categorical)

3. **Link function**
    - Gaussian: identity
    - Binomial: logit, probit
    - Poisson: log...
    - See `family`.

The modelling process

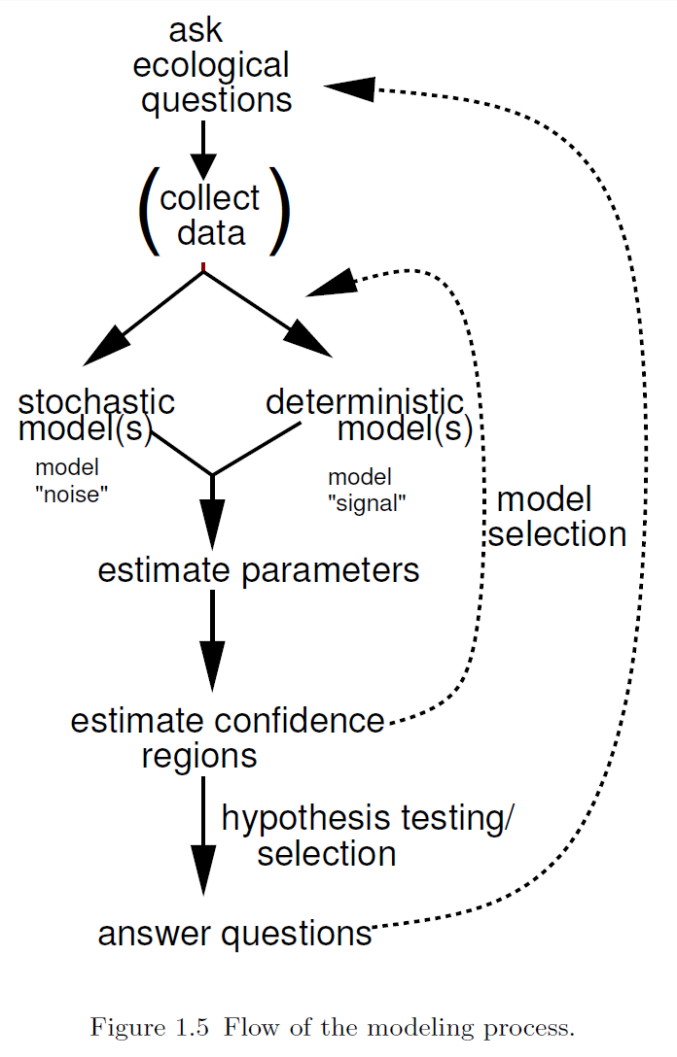


Figure 1.5 Flow of the modeling process.

Bolker 2008

Bernouilli - Binomial distribution (Logistic regression)

Response variable: **Yes/No** (e.g. survival, sex, presence/absence)

Canonical link function: `logit` (*log odds*), but others possible (see `family`)

$$logit(p) = \log\left(\frac{p}{1-p}\right)$$

Then

$$logit(P(alive)) = a + bx$$

$$P(alive) = invlogit(a + bx) = \frac{e^{a+bx}}{1 + e^{a+bx}}$$

Where is the variance?

In a Gaussian GLM

$$y \sim Normal(\mu, \sigma)$$

In a Binomial GLM

$$y \sim Binomial(n, p)$$

n = number of trials

p = probability of success

$$Var(y) = np(1 - p)$$

(maximum variance when **p** around 0.5)

Back to survival of Titanic passengers

How many survived in each class?

```
table(titanic$class, titanic$survived)
```

```
          0   1
  crew   673 212
  first  122 203
  second 167 118
  third  528 178
```

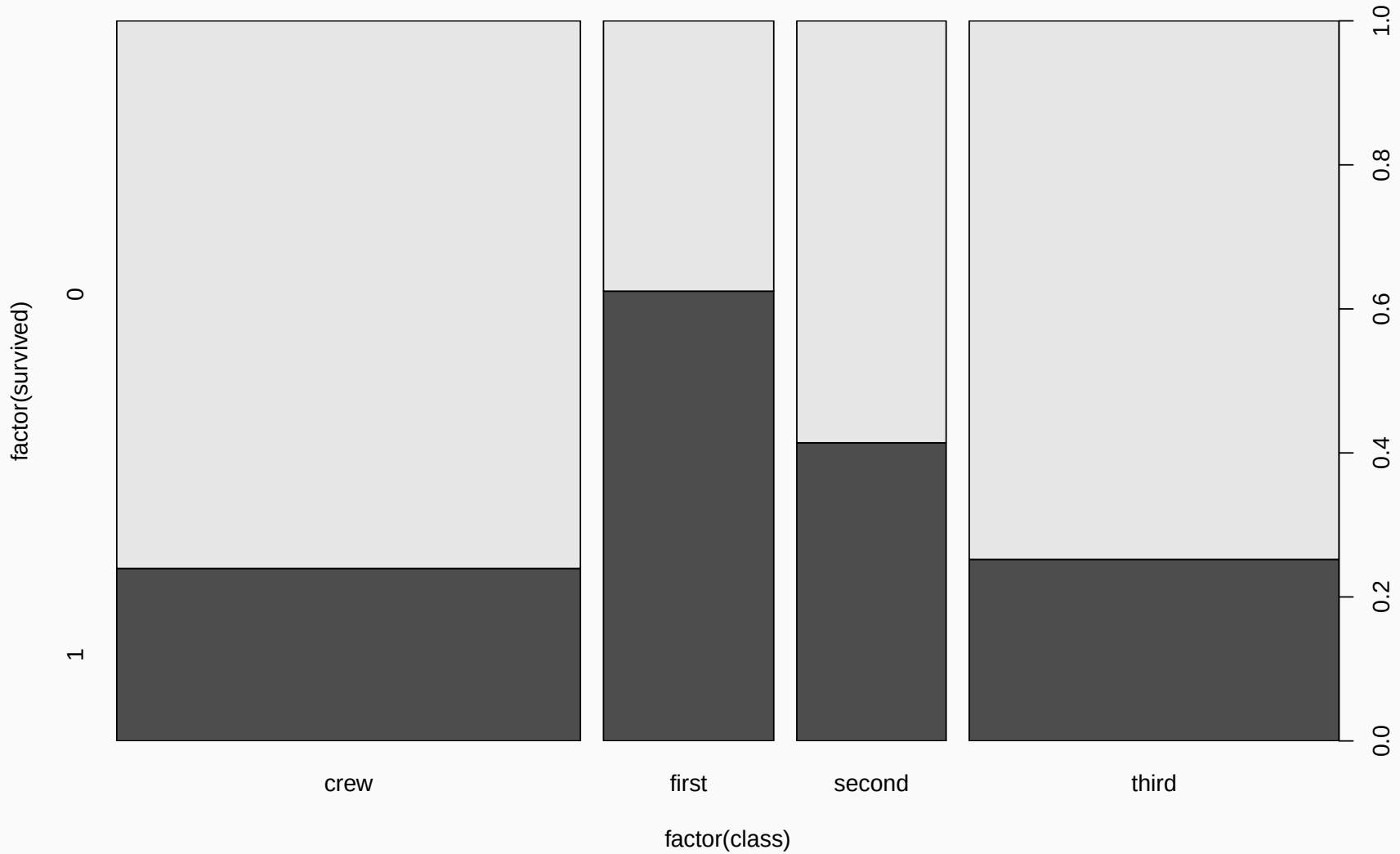
How many survived in each class? (*dplyr*)

```
titanic %>%
  group_by(class, survived) %>%
  summarise(count = n())
```

```
# A tibble: 8 x 3
# Groups:   class [4]
  class  survived count
  <chr>     <int> <int>
1 crew          0   673
2 crew          1   212
3 first         0   122
4 first         1   203
5 second        0   167
6 second        1   118
7 third         0   528
8 third         1   178
```

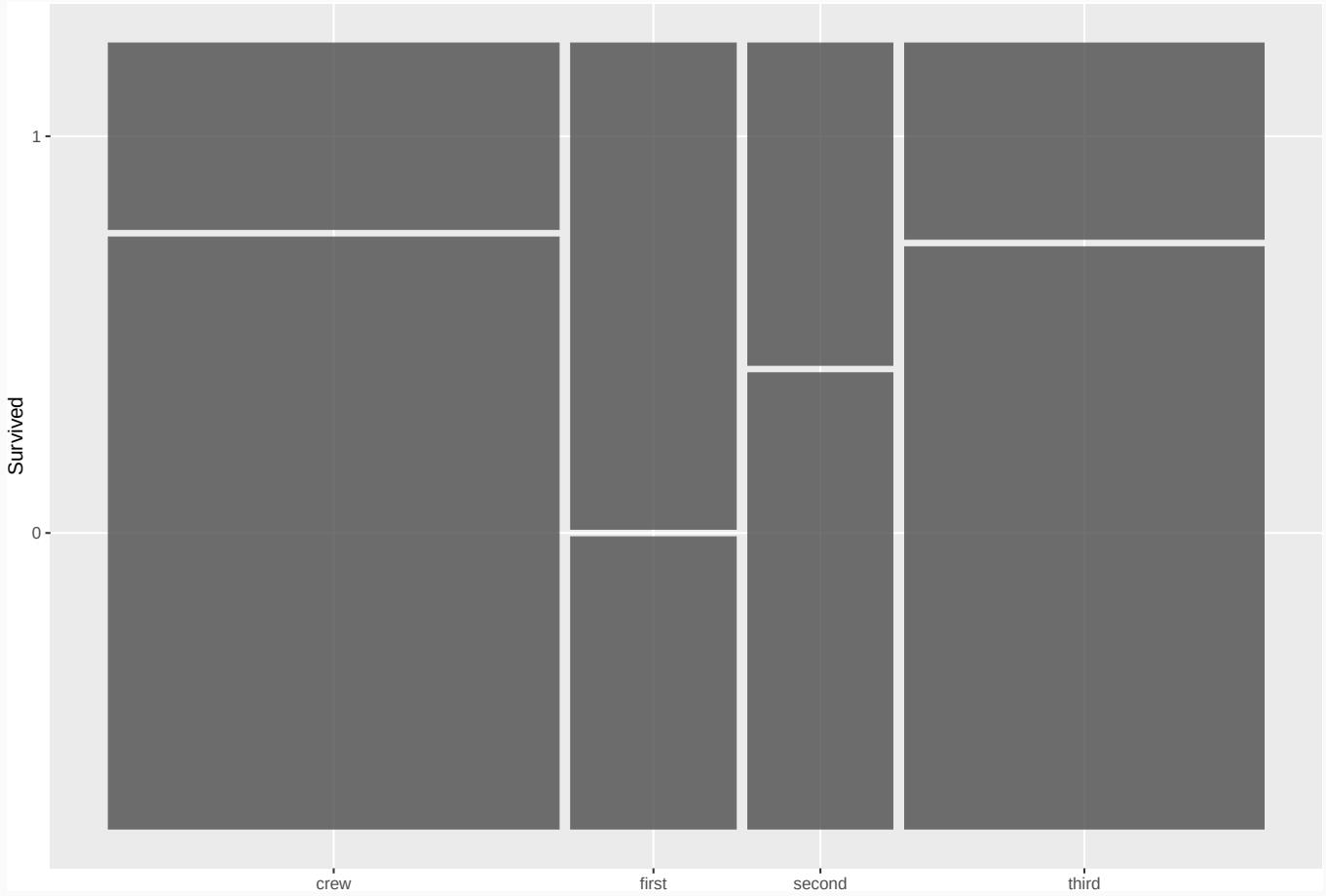
Data visualisation (mosaic plot)

```
plot(factor(survived) ~ factor(class), data = titanic)
```



Mosaic plots (ggplot2)

```
library("ggmosaic")
ggplot(titanic) +
  geom_mosaic(aes(x = product(survived, class))) +
  labs(x = "", y = "Survived")
```



Fitting GLMs in R: `glm`

```
tit.glm <- glm(survived ~ class,
               data = titanic,
               family = binomial)
```

which corresponds to

$$logit\big(P(survival)_i\big) = a + b \cdot class_i$$

$$logit\big(P(survival)_i\big) = a + b_{first} + c_{second} + d_{third}$$

Interpreting binomial GLM

```
tit.glm <- glm(survived ~ class, data = titanic, family = binomial)
```

```
Call:
glm(formula = survived ~ class, family = binomial, data = titanic)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.15516    0.07876 -14.667  < 2e-16 ***
classfirst   1.66434    0.13902  11.972  < 2e-16 ***
classsecond  0.80785    0.14375   5.620 1.91e-08 ***
classthird   0.06785    0.11711   0.579    0.562
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2769.5  on 2200  degrees of freedom
Residual deviance: 2588.6  on 2197  degrees of freedom
AIC: 2596.6

Number of Fisher Scoring iterations: 4
```

Binomial GLM estimates are in `logit` scale!

We need to **back-transform** (apply *inverse logit*):

- Manually: `plogis`

Binomial GLM estimates are in `logit` scale!

We need to **back-transform** (apply *inverse logit*):

- Manually: `plogis`
- Automatically: `easystats`, etc.

Interpreting logistic regression output (easystats)

```r
library("easystats")   # 'modelbased' pkg
estimate_means(tit.glm)
```

```
Estimated Marginal Means

class  | Probability |   SE |       95% CI
-----------------------------------------
first  |        0.62 | 0.03 | [0.57, 0.68]
second |        0.41 | 0.03 | [0.36, 0.47]
third  |        0.25 | 0.02 | [0.22, 0.29]
crew   |        0.24 | 0.01 | [0.21, 0.27]

Marginal means estimated at class
```

Analysing differences among factor levels (class)

```
estimate_contrasts(tit.glm)
```

```
Marginal Contrasts Analysis

Level1 | Level2 | Difference |        95% CI |   SE |  df |     z |      p
-------------------------------------------------------------------------
first  |   crew |       1.66 | [ 1.30, 2.03] | 0.14 | Inf | 11.97 | < .001
first  | second |       0.86 | [ 0.42, 1.29] | 0.17 | Inf |  5.16 | < .001
first  |  third |       1.60 | [ 1.22, 1.98] | 0.14 | Inf | 11.11 | < .001
second |   crew |       0.81 | [ 0.43, 1.19] | 0.14 | Inf |  5.62 | < .001
second |  third |       0.74 | [ 0.35, 1.13] | 0.15 | Inf |  4.99 | < .001
third  |   crew |       0.07 | [-0.24, 0.38] | 0.12 | Inf |  0.58 | 0.562

Marginal contrasts estimated at class
p-value adjustment method: Holm (1979)
```

Pseudo R-squared for GLMs

```
library("easystats")   # 'performance' pkg
r2(tit.glm)
```

```
# R2 for Logistic Regression
  Tjur's R2: 0.087
```

But there are caveats (e.g. see here and here)

Presenting model results

```
kable(xtable::xtable(tit.glm), digits = 2)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.16	0.08	-14.67	0.00
classfirst	1.66	0.14	11.97	0.00
classsecond	0.81	0.14	5.62	0.00
classthird	0.07	0.12	0.58	0.56

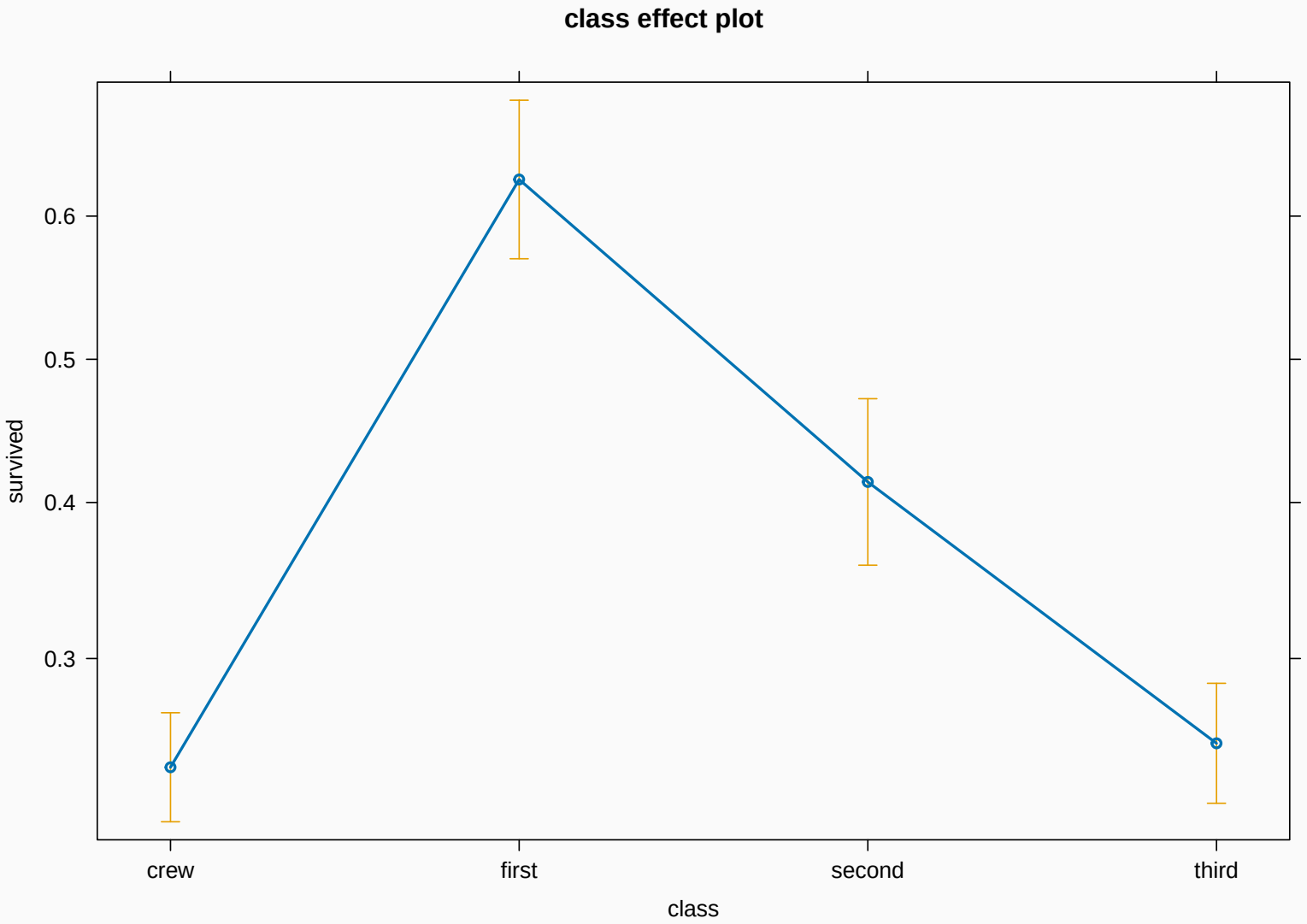
Presenting model results

```
library("modelsummary")
modelsummary(tit.glm, output = "markdown")
```

	(1)
(Intercept)	-1.155
	(0.079)
classfirst	1.664
	(0.139)
classsecond	0.808
	(0.144)
classthird	0.068
	(0.117)
Num.Obs.	2201
AIC	2596.6
BIC	2619.3
Log.Lik.	-1294.278
F	57.743
RMSE	0.45

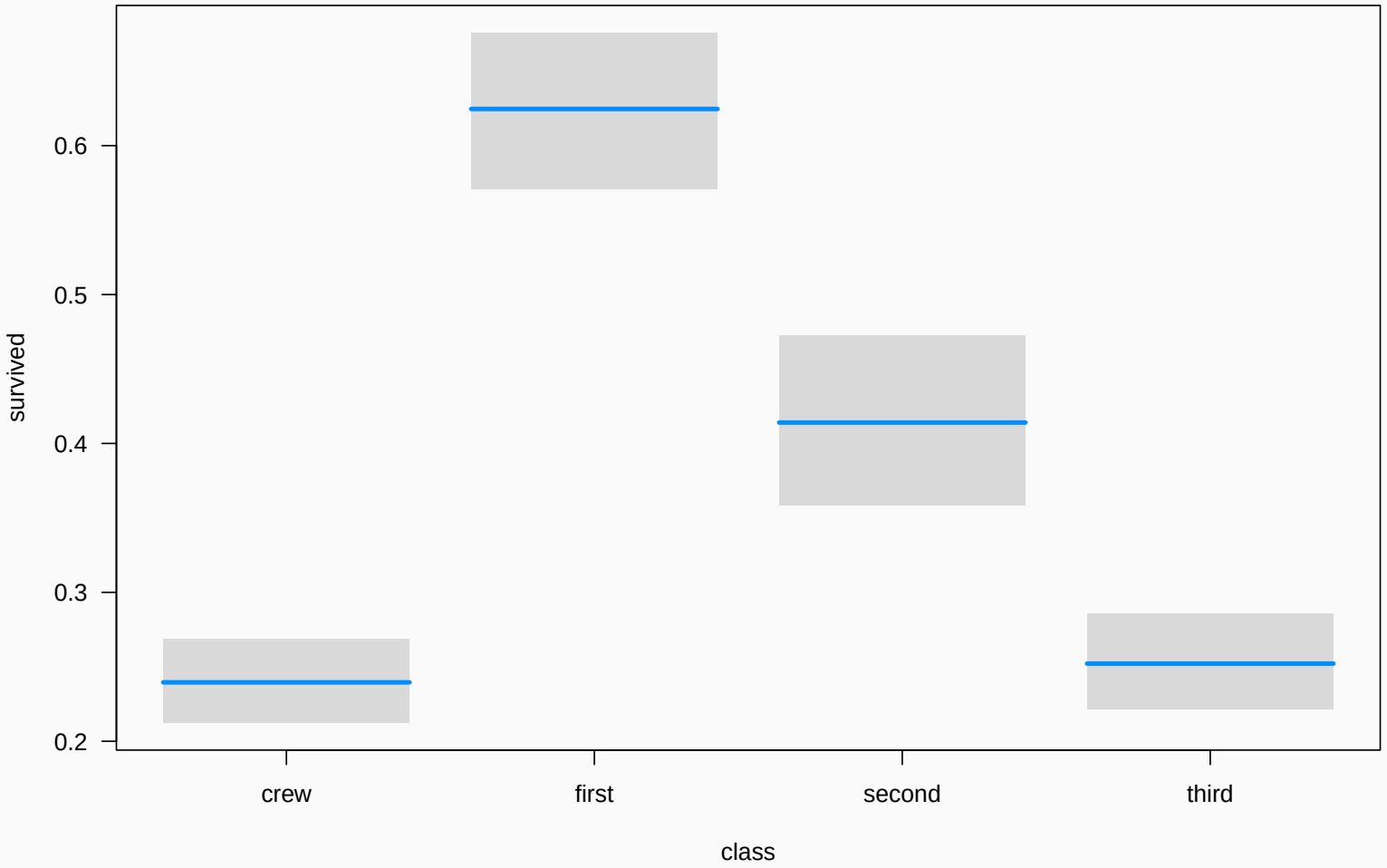
Visualising model: `effects` package

```
plot(allEffects(tit.glm))
```



Visualising model: `visreg` package

```
visreg(tit.glm, scale = "response", rug = FALSE)
```

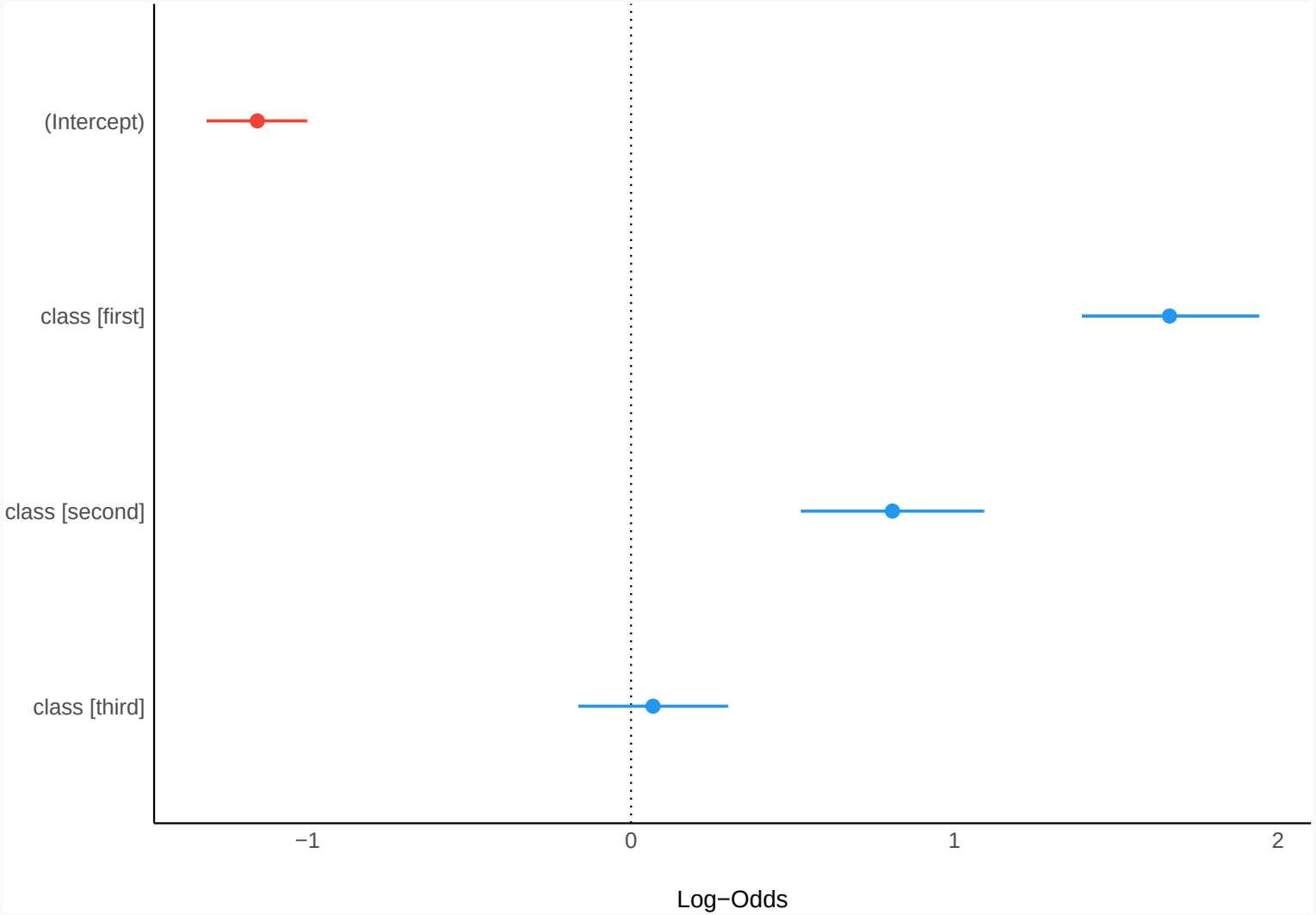


Visualising model: `sjPlot` package

```
sjPlot::plot_model(tit.glm, type = "eff")
```

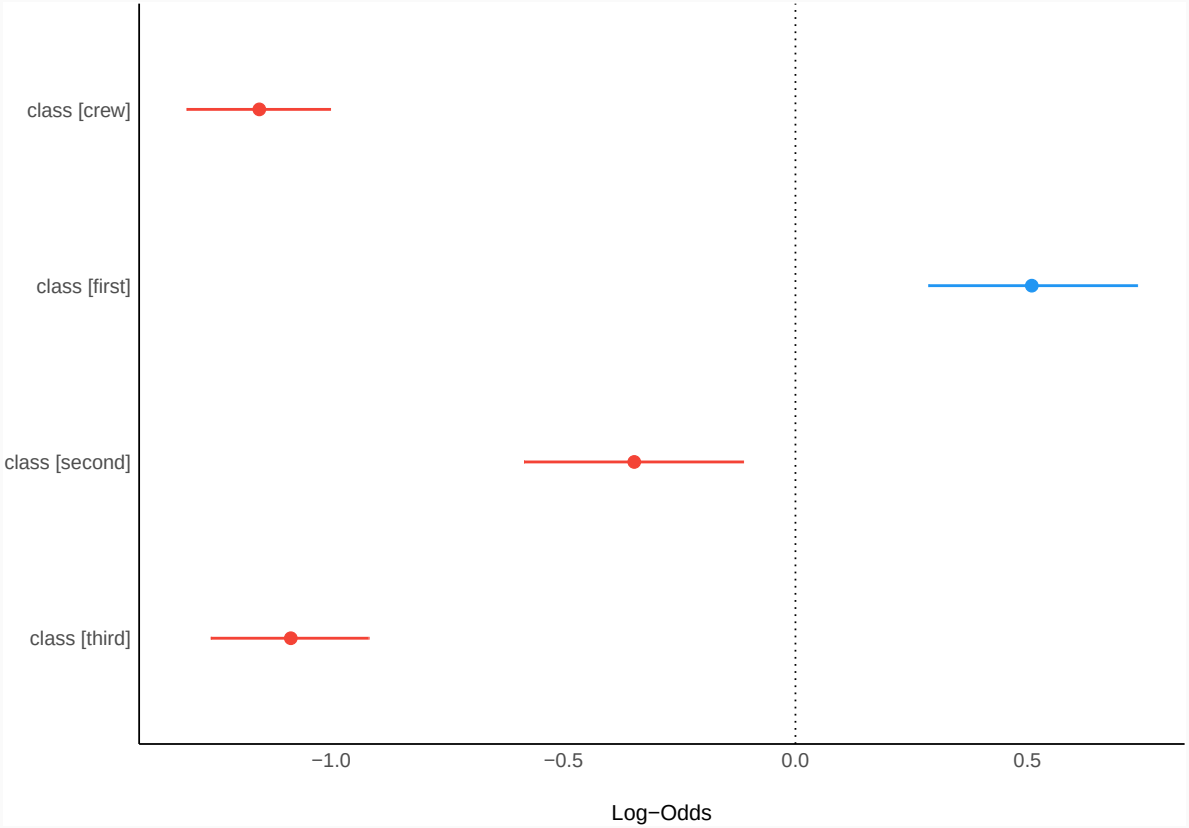
Visualising model: easystats (see package)

```
plot(parameters(tit.glm), show_intercept = TRUE)
```



Model without intercept

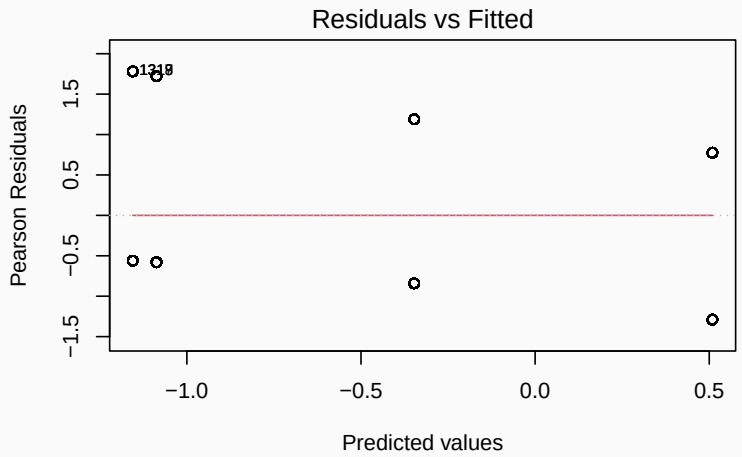
```
no.intercept <- glm(survived ~ class - 1, family = binomial, data =
plot(parameters(no.intercept))
```

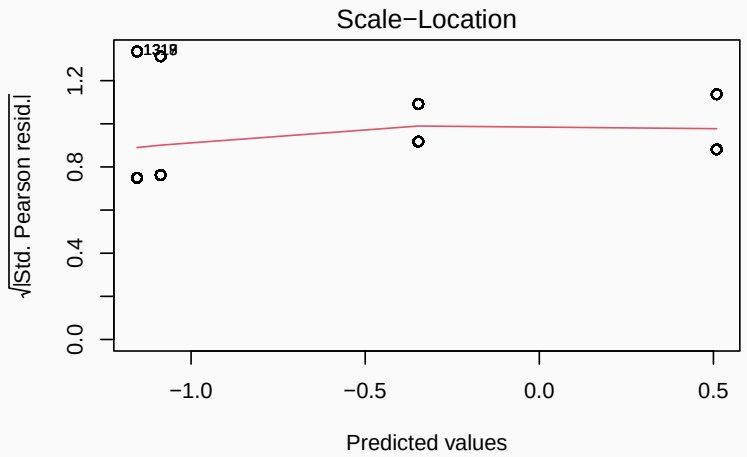


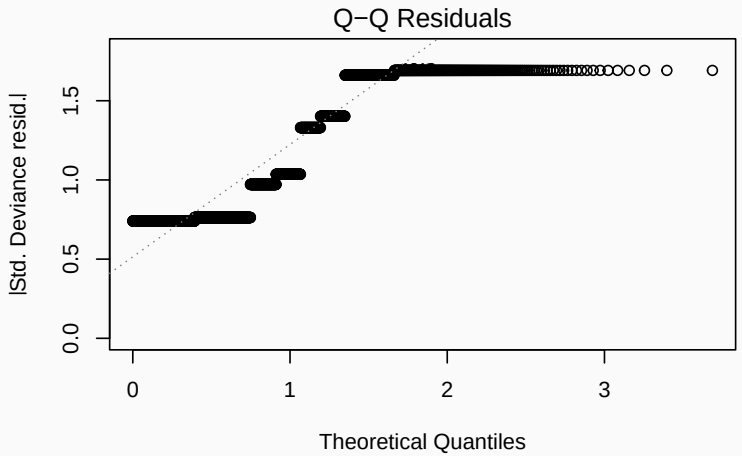
Model checking

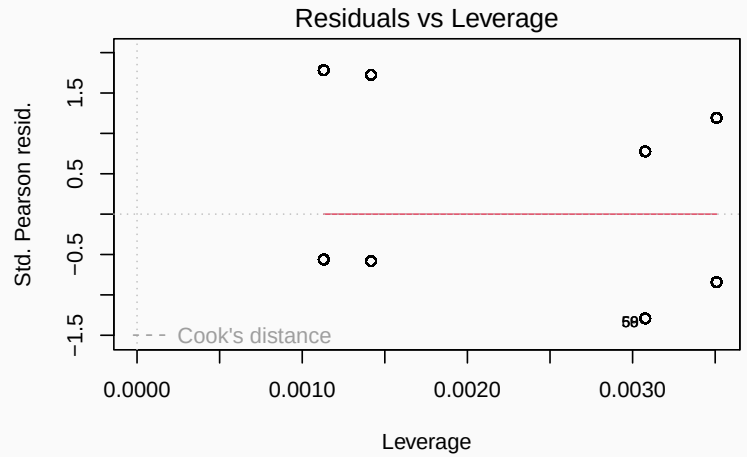
`plot(model)` not very useful with binomial GLM

```
plot(tit.glm)
```









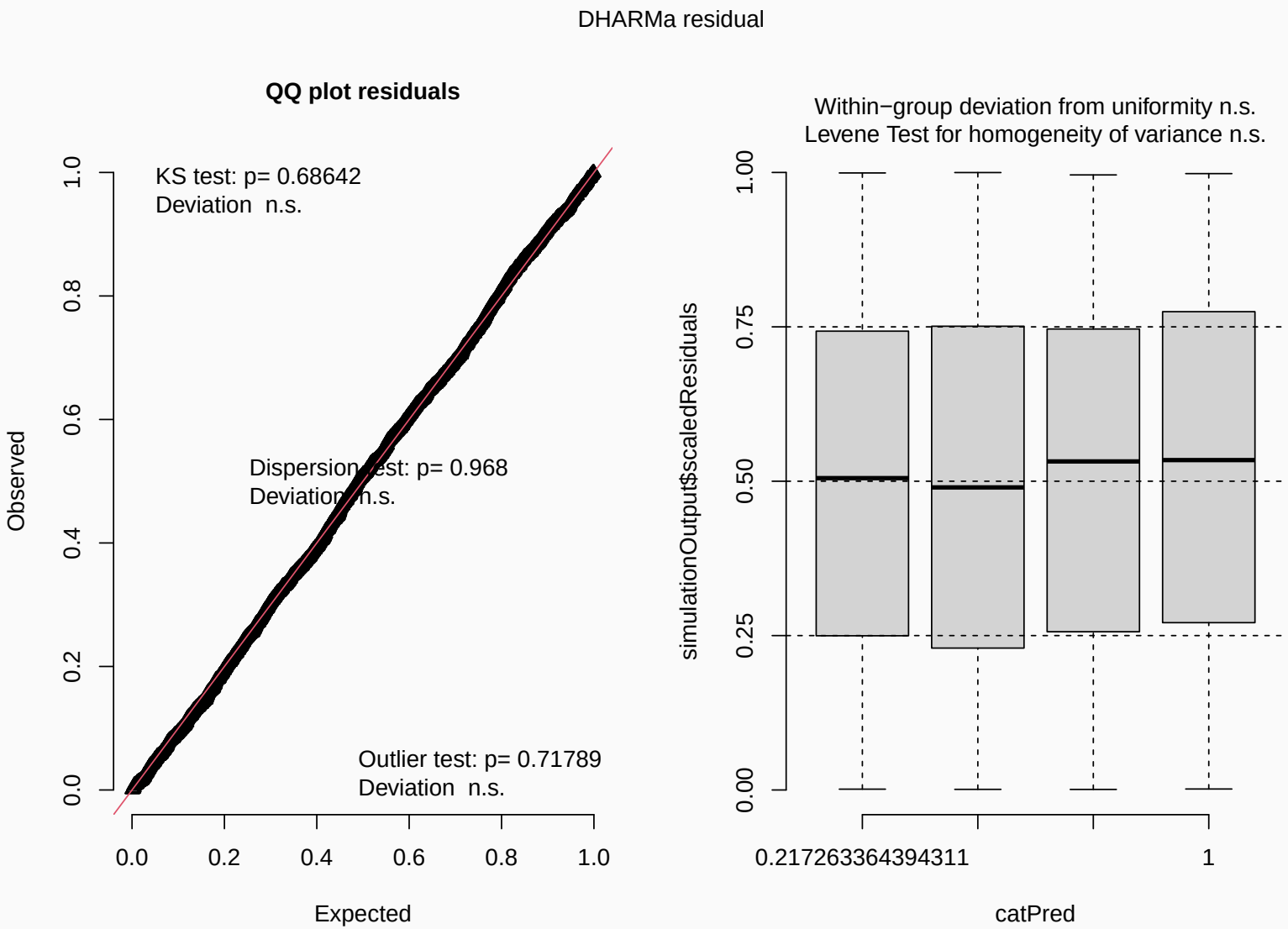

```
null device
          1
```

check_model (easystats)

```
check_model(tit.glm)
```

Residual diagnostics with DHARMa

```
library("DHARMa")
simulateResiduals(tit.glm, plot = TRUE)
```

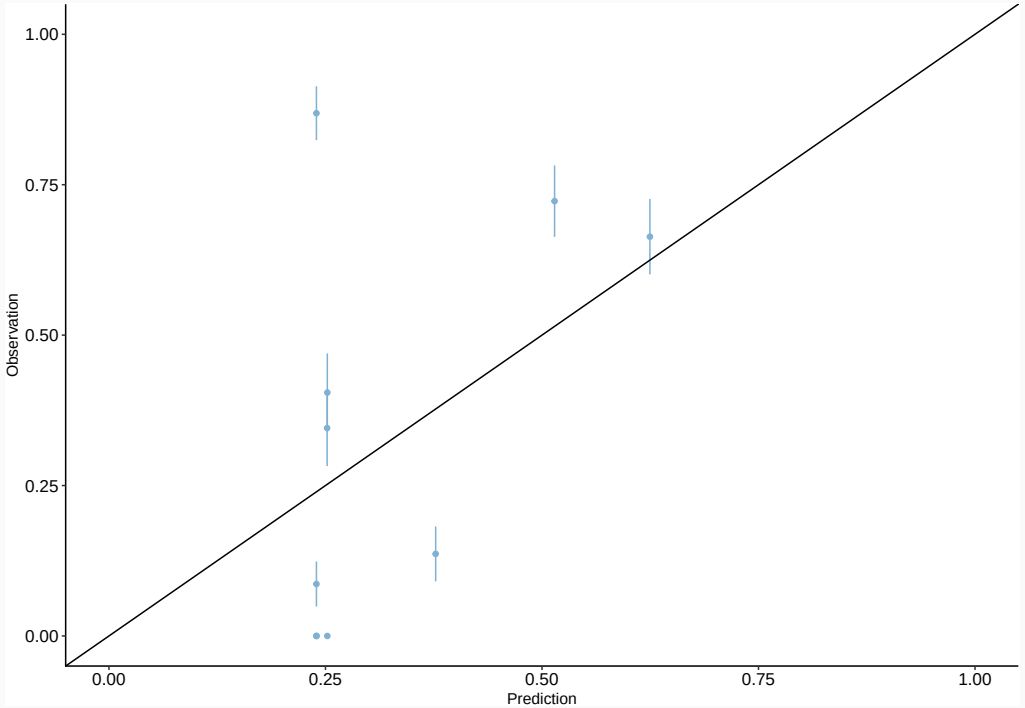


Calibration plot

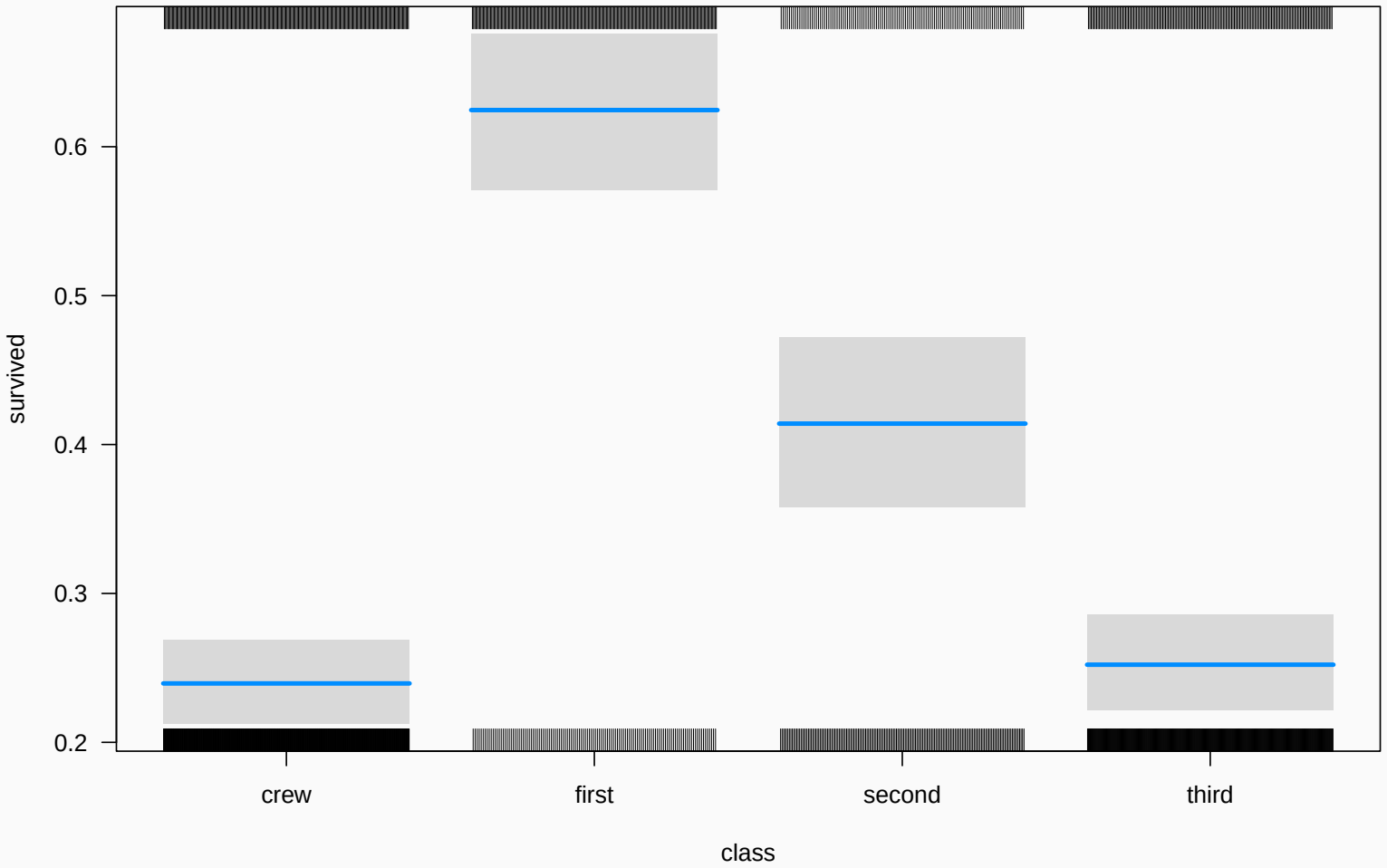
Compares predicted vs observed probabilities (grouped by quantiles)

```
library("predtools")
titanic$surv.pred <- predict(tit.glm, type = "response")
calibration_plot(data = titanic, obs = "survived", pred = "surv.pred",
                 x_lim = c(0,1), y_lim = c(0,1))
```

```
$calibration_plot
```



Passenger class was important, but lots of unexplained variation



The goal is not to test whether the model's assumptions are "true", because all models are false.

Rather, the goal is to assess exactly **how the model fails to describe the data**, as a path towards **model comprehension, revision, and improvement**.

Richard McElreath. *Statistical Rethinking*

Recapitulating

1. Visualise data

Recapitulating

1. Visualise data

2. **Fit model**: `glm`. Don't forget to specify `family`!

Recapitulating

1. Visualise data

2. **Fit model**: `glm`. Don't forget to specify `family`!

3. **Examine model**: `summary`

Recapitulating

1. **Visualise data**

2. **Fit model**: `glm`. Don't forget to specify `family`!

3. **Examine model**: `summary`

4. **Back-transform parameters** from *logit* into probability scale (`estimate_means`)

Recapitulating

1. **Visualise data**

2. **Fit model**: `glm`. Don't forget to specify `family`!

3. **Examine model**: `summary`

4. **Back-transform parameters** from *logit* into probability scale (`estimate_means`)

5. **Plot model**: `visreg`, ...

Recapitulating

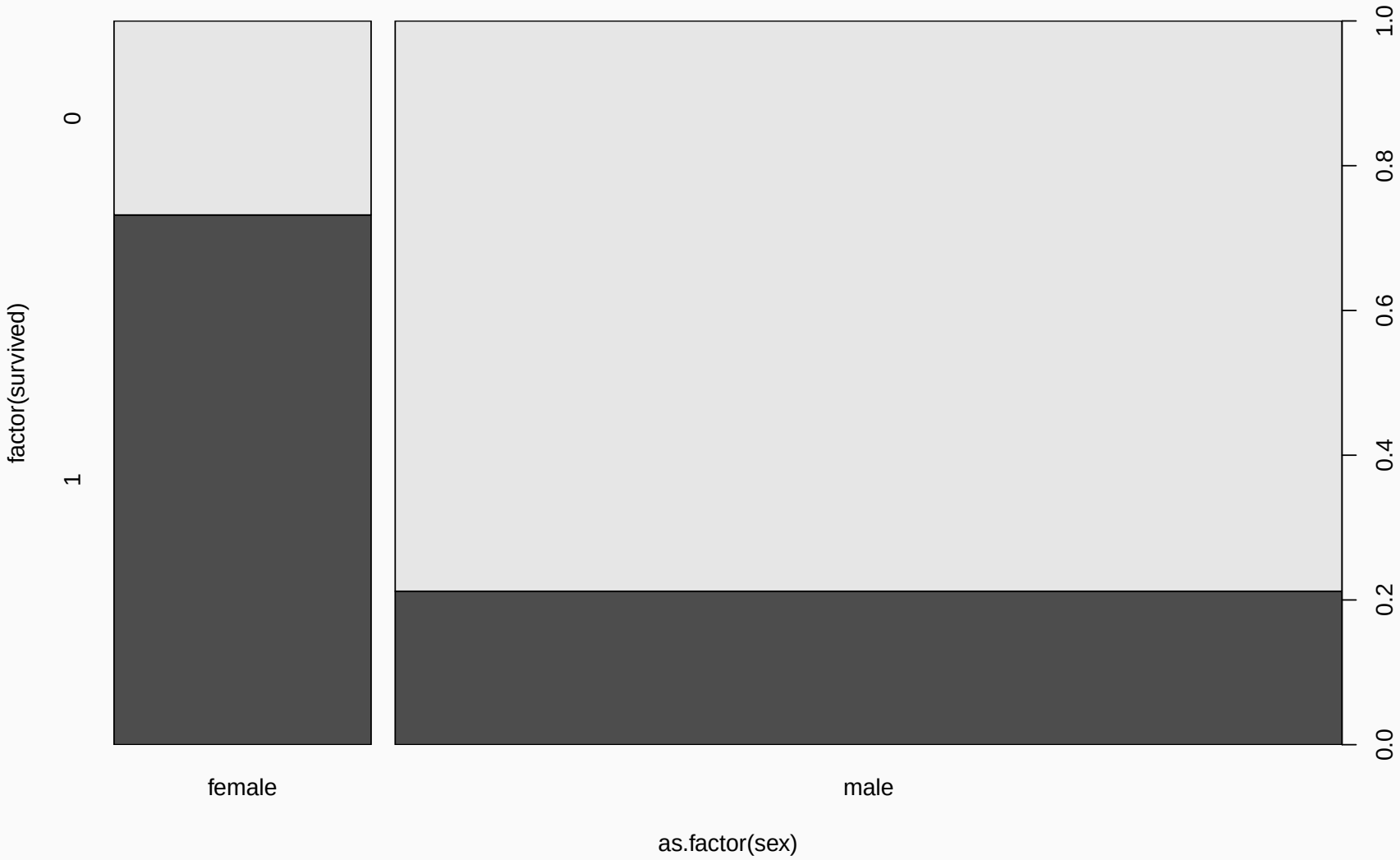
1. Visualise data
2. **Fit model**: `glm`. Don't forget to specify `family`!
3. **Examine model**: `summary`
4. **Back-transform parameters** from *logit* into probability scale (`estimate_means`)
5. **Plot model**: `visreg`, ...
6. **Check model**: `check_model`, `DHARMa::simulateResiduals`, `calibration_plot`

Q: Did men have higher survival than women?

Quiz

https://pollev.com/franciscorod726

First, visualise data



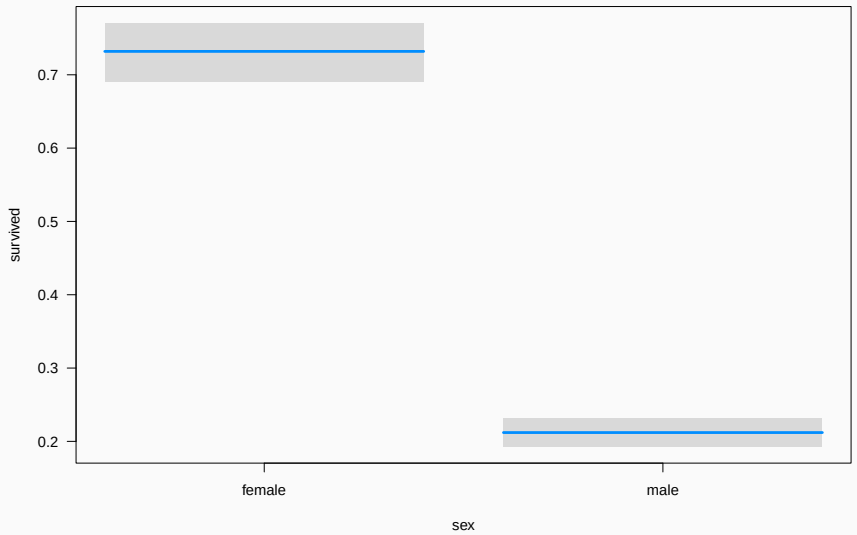
Fit model

```
Call:
glm(formula = survived ~ sex, family = binomial, data = titanic)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.0044     0.1041   9.645   <2e-16 ***
sexmale      -2.3172     0.1196 -19.376   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2769.5  on 2200  degrees of freedom
Residual deviance: 2335.0  on 2199  degrees of freedom
AIC: 2339

Number of Fisher Scoring iterations: 4
```

Model interpretation

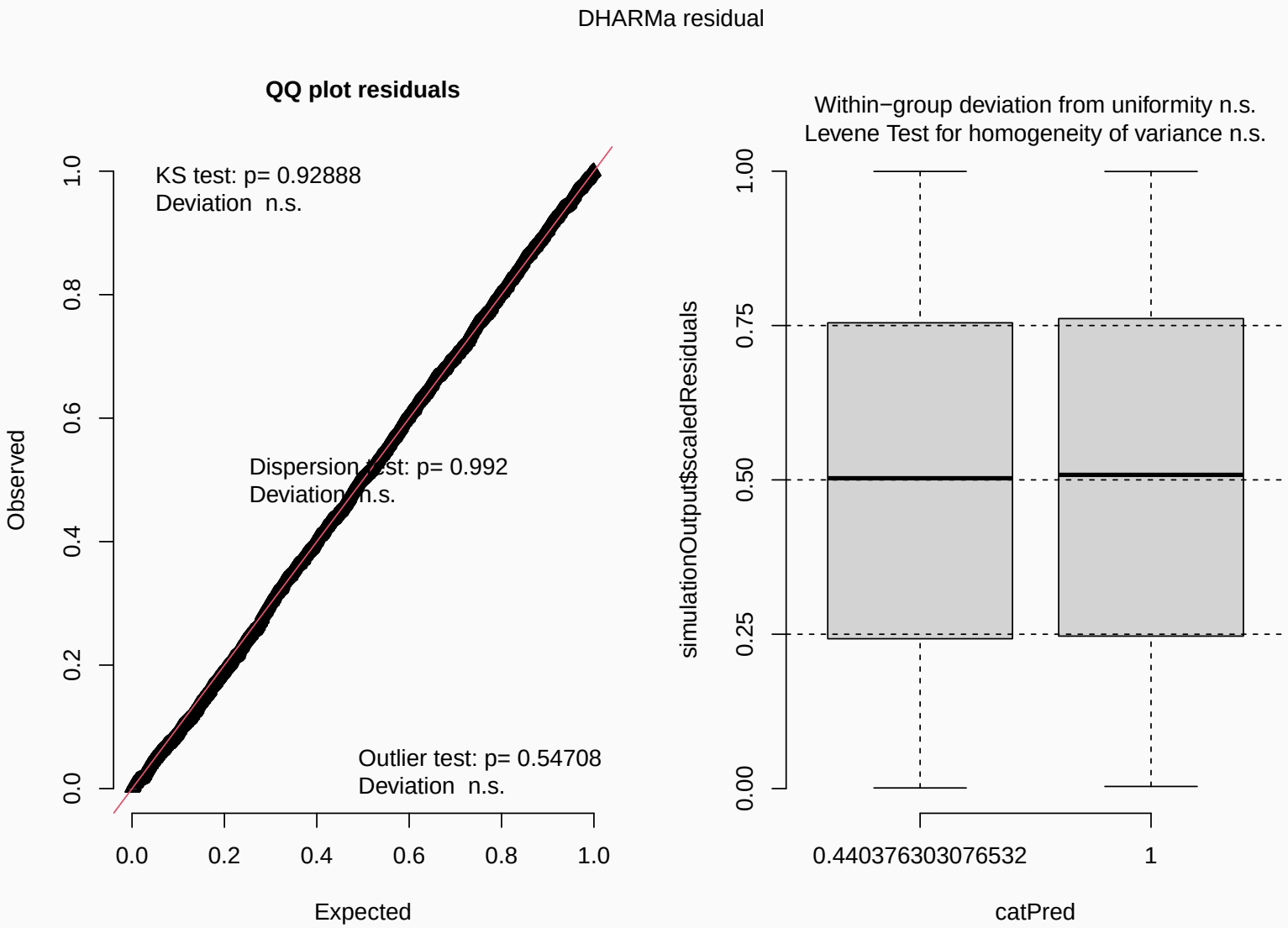
```
estimate_means(tit.sex)
```

```
Estimated Marginal Means

sex    | Probability |       SE |       95% CI
-----------------------------------------------
male   |        0.21 | 9.82e-03 | [0.19, 0.23]
female |        0.73 |     0.02 | [0.69, 0.77]

Marginal means estimated at sex
```



Model checking

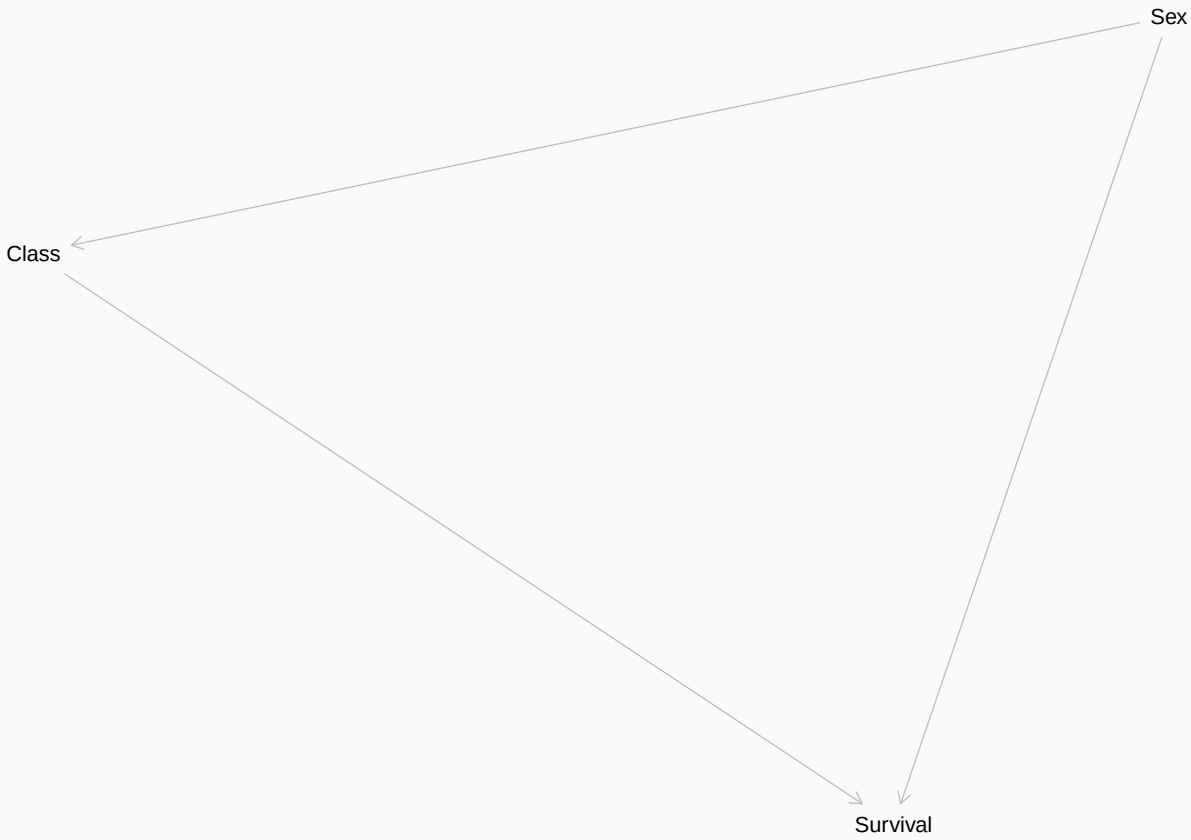
```
simulateResiduals(tit.sex, plot = TRUE)
```



Q: Did women have higher survival because they travelled more in first class?

Did women have higher survival because they travelled more in first class?

Sex is a confounder



Let’s look at the data

```
table(titanic$class, titanic$survived, titanic$sex)
```

```
, ,  = female


          0   1
  crew    3  20
  first   4 141
  second 13  93
  third 106  90

, ,  = male


          0   1
  crew  670 192
  first 118  62
  second 154  25
  third 422  88
```

Quiz

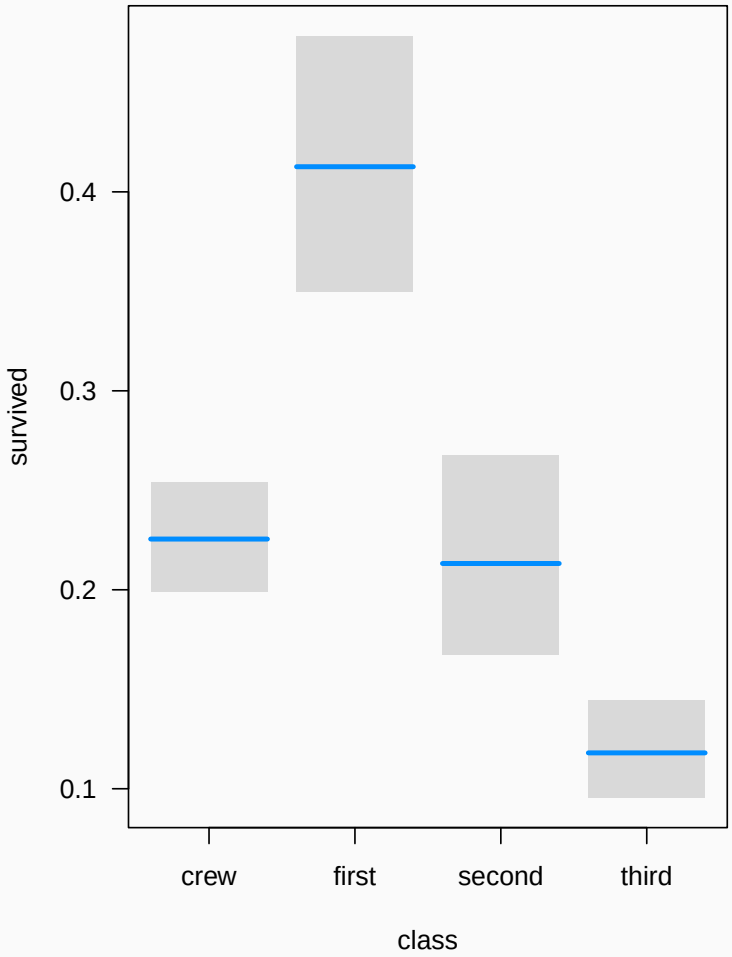
https://pollev.com/franciscorod726

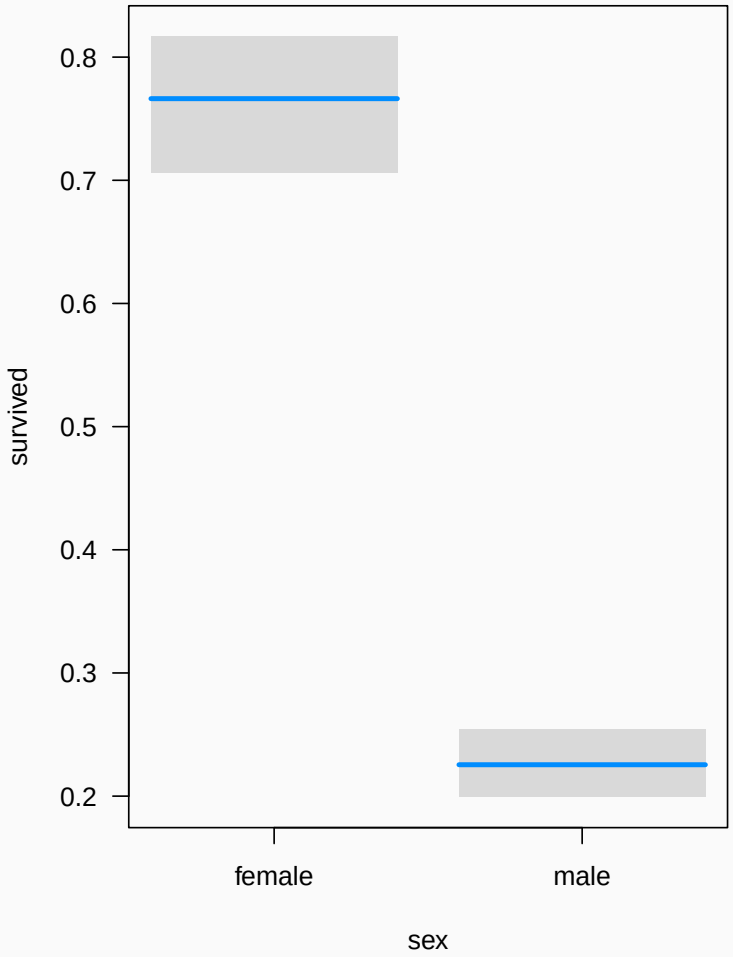
Fit additive model with both factors

```
Call:
glm(formula = survived ~ class + sex, family = binomial, data = titanic)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.18740    0.15747   7.541 4.68e-14 ***
classfirst   0.88081    0.15697   5.611 2.01e-08 ***
classsecond -0.07178    0.17093  -0.420    0.675
classthird  -0.77742    0.14231  -5.463 4.69e-08 ***
sexmale     -2.42133    0.13909 -17.408  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2769.5  on 2200  degrees of freedom
Residual deviance: 2228.9  on 2196  degrees of freedom
AIC: 2238.9

Number of Fisher Scoring iterations: 4
```

Plot additive model

```
visreg(tit.sex.class.add, scale = "response", rug = FALSE)
```






```
null device
          1
```

Fit model with the interaction of both factors

```
Call:
glm(formula = survived ~ class * sex, family = binomial, data = titanic)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.89712    0.61914   3.064  0.00218 **
classfirst          1.66535    0.80026   2.081  0.03743 *
classsecond         0.07053    0.68630   0.103  0.91815
classthird         -2.06075    0.63551  -3.243  0.00118 **
sexmale            -3.14690    0.62453  -5.039 4.68e-07 ***
classfirst:sexmale -1.05911    0.81959  -1.292  0.19627
classsecond:sexmale -0.63882    0.72402  -0.882  0.37760
classthird:sexmale  1.74286    0.65139   2.676  0.00746 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2769.5  on 2200  degrees of freedom
Residual deviance: 2163.7  on 2193  degrees of freedom
AIC: 2179.7
```

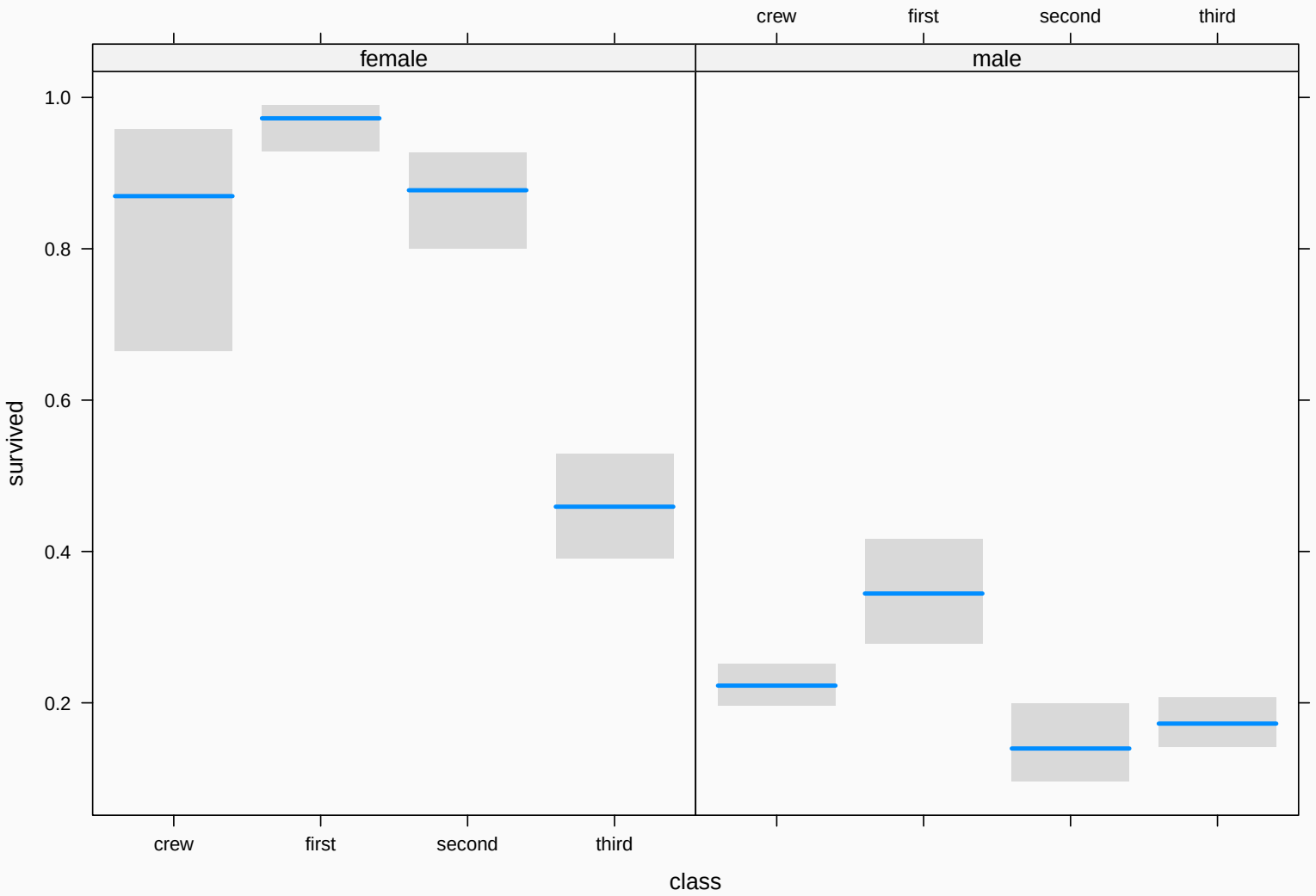
Women had higher survival than men, even within the same class

```
estimate_means(tit.sex.class.int)
```

```
Estimated Marginal Means

class  |    sex | Probability |   SE |       95% CI
-----------------------------------------------------
first  |   male |        0.34 | 0.04 | [0.28, 0.42]
second |   male |        0.14 | 0.03 | [0.10, 0.20]
third  |   male |        0.17 | 0.02 | [0.14, 0.21]
crew   |   male |        0.22 | 0.01 | [0.20, 0.25]
first  | female |        0.97 | 0.01 | [0.93, 0.99]
second | female |        0.88 | 0.03 | [0.80, 0.93]
third  | female |        0.46 | 0.04 | [0.39, 0.53]
crew   | female |        0.87 | 0.07 | [0.66, 0.96]

Marginal means estimated at class, sex
```

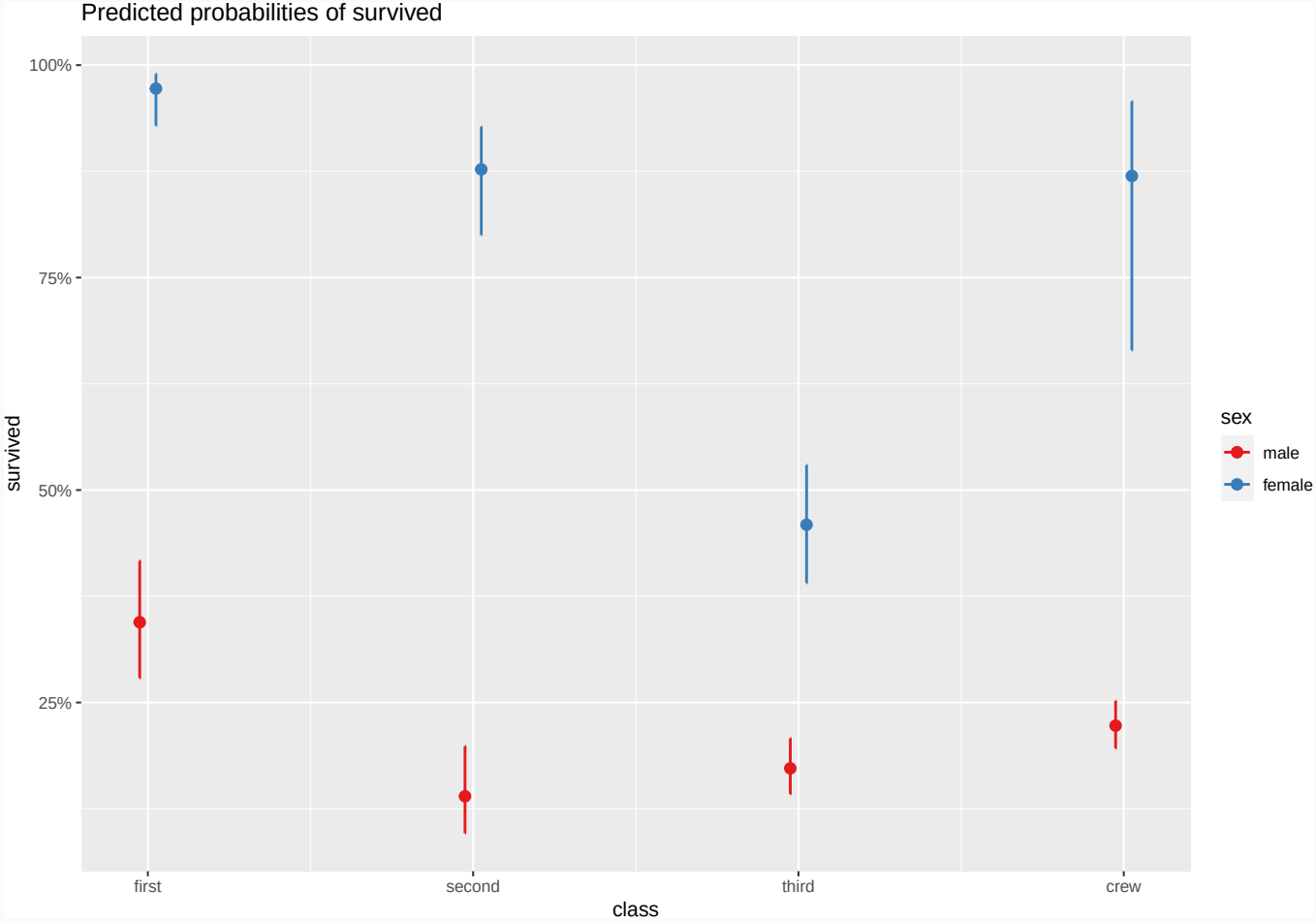
Women had higher survival than men, even within the same class

```
visreg(tit.sex.class.int, by = "sex", xvar = "class", scale = "response", rug = FAL
```



Visualising model (`sjPlot`)

```
library("sjPlot")
plot_model(tit.sex.class.int, type = "int")
```



Comparing models

```
library("easystats")   # 'performance' pkg
compare_performance(tit.sex.class.add, tit.sex.class.int)
```

```
# Comparison of Model Performance Indices

Name              | Model |  AIC (weights) | AICc (weights) |  BIC (weights) | Tjur's R2 |  RMS
-------------------------------------------------------------------------------------------
tit.sex.class.add |   glm | 2238.9 (<.001) | 2238.9 (<.001) | 2267.4 (<.001) |     0.248 | 0.40
tit.sex.class.int |   glm | 2179.7 (>.999) | 2179.8 (>.999) | 2225.3 (>.999) |     0.271 | 0.39
```

Comparing parameters

```
compare_parameters(tit.sex.class.add, tit.sex.class.int)
```

Parameter	tit.sex.class.add	tit.sex.class.int
(Intercept)	1.19 (0.88, 1.50)	1.90 (0.68, 3.11)
class (first)	0.88 (0.57, 1.19)	1.67 (0.10, 3.23)
class (second)	-0.07 (-0.41, 0.26)	0.07 (-1.27, 1.42)
class (third)	-0.78 (-1.06, -0.50)	-2.06 (-3.31, -0.82)
sex (male)	-2.42 (-2.69, -2.15)	-3.15 (-4.37, -1.92)
class (first) × sex (male)		-1.06 (-2.67, 0.55)
class (second) × sex (male)		-0.64 (-2.06, 0.78)
class (third) × sex (male)		1.74 (0.47, 3.02)
Observations	2201	2201

Extra exercises:

Is survival related to age?

Are age effects dependent on sex?

Logistic regression for proportion data

Read Titanic data in different format

Read `titanic_prop.csv` data.

```
  X Class    Sex   Age  No Yes
1 1   1st Female Adult   4 140
2 2   1st Female Child   0   1
3 3   1st   Male Adult 118  57
4 4   1st   Male Child   0   5
5 5   2nd Female Adult  13  80
6 6   2nd Female Child   0  13
```

These are the same data, but summarized (see `Freq` variable).

Use `cbind(n.success, n.failures)` as response

```
prop.glm <- glm(cbind(Yes, No) ~ Class, data = tit.prop, family = binomial)
```

```
Call:
glm(formula = cbind(Yes, No) ~ Class, family = binomial, data = tit.prop)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.5092     0.1146   4.445 8.79e-06 ***
Class2nd     -0.8565     0.1661  -5.157 2.51e-07 ***
Class3rd     -1.5965     0.1436 -11.114  < 2e-16 ***
ClassCrew    -1.6643     0.1390 -11.972  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 671.96  on 13  degrees of freedom
Residual deviance: 491.06  on 10  degrees of freedom
AIC: 545.68

Number of Fisher Scoring iterations: 4
```

Survival probability by class

```
estimate_means(prop.glm)
```

```
Estimated Marginal Means

Class | Probability |   SE |       95% CI
-----------------------------------------
1st   |        0.62 | 0.03 | [0.57, 0.68]
2nd   |        0.41 | 0.03 | [0.36, 0.47]
3rd   |        0.25 | 0.02 | [0.22, 0.29]
Crew  |        0.24 | 0.01 | [0.21, 0.27]

Marginal means estimated at Class
```

Logistic regression with continuous predictors

Example dataset: GDP and infant mortality

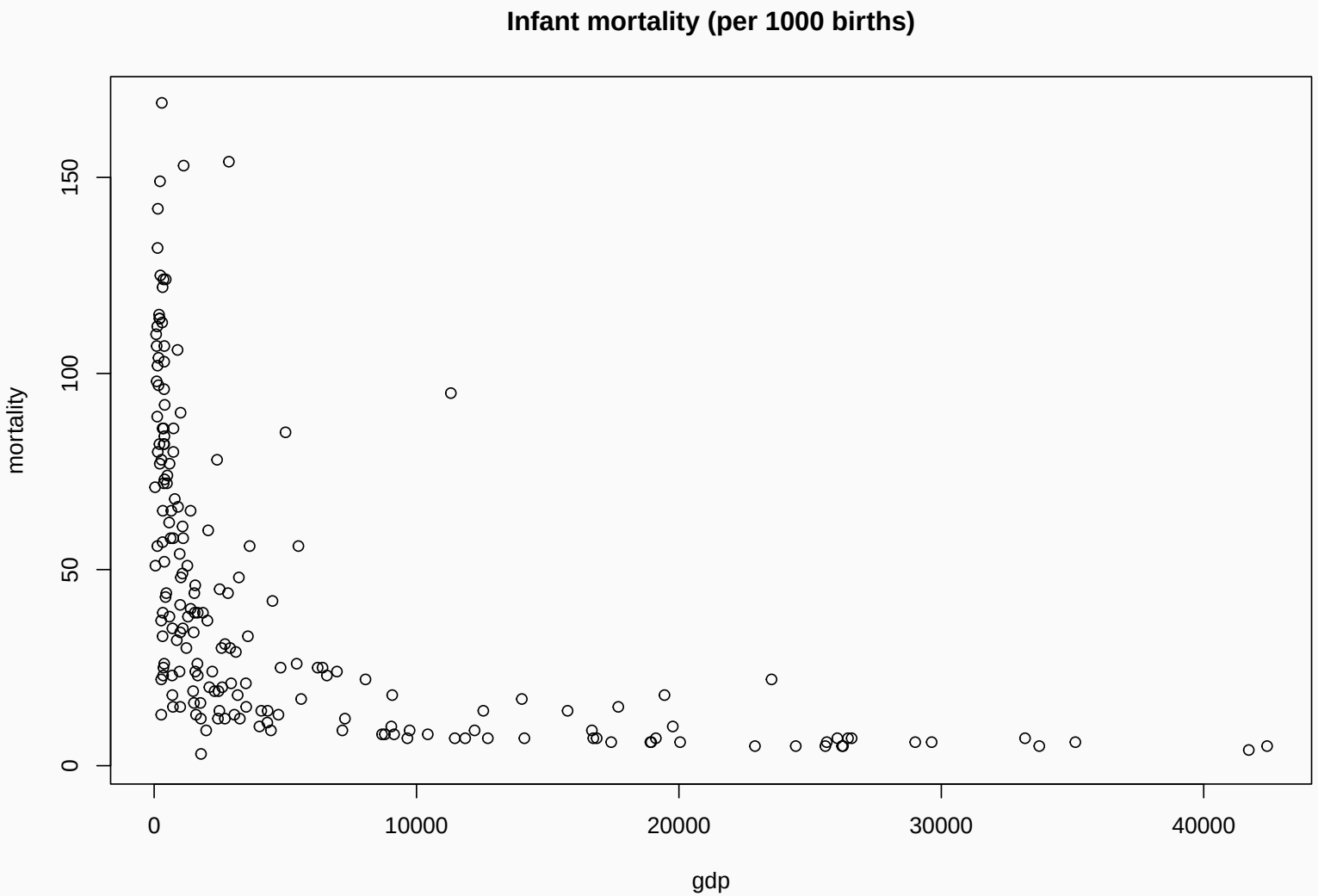
Read `UN_GDP_infantmortality.csv`.

```
   country              mortality           gdp        
 Length:207         Min.   :  2.00   Min.   :   36  
 Class :character   1st Qu.: 12.00   1st Qu.:  442  
 Mode  :character   Median : 30.00   Median : 1779  
                    Mean   : 43.48   Mean   : 6262  
                    3rd Qu.: 66.00   3rd Qu.: 7272  
                    Max.   :169.00   Max.   :42416  
                    NA's   :6        NA's   :10     
```

Q: Is infant mortality related to GDP?

https://pollev.com/franciscorod726

Visualising data



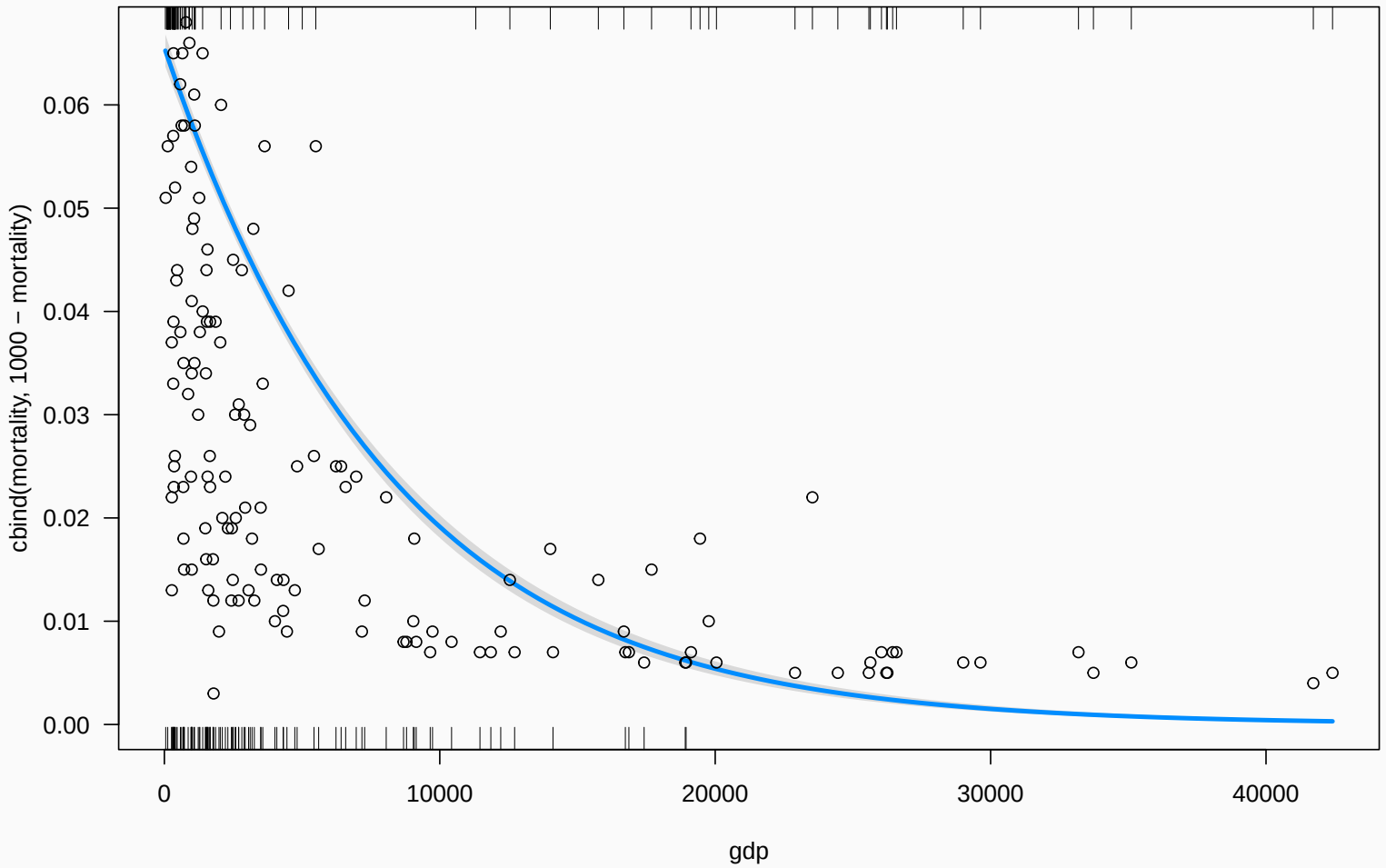
Fit model

```
gdp.glm <- glm(cbind(mortality, 1000 - mortality) ~ gdp,
               data = gdp, family = binomial)
```

```
Call:
glm(formula = cbind(mortality, 1000 - mortality) ~ gdp, family = binomial,
    data = gdp)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.657e+00  1.311e-02 -202.76   <2e-16 ***
gdp         -1.279e-04  3.458e-06  -36.98   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6430.2  on 192  degrees of freedom
Residual deviance: 3530.2  on 191  degrees of freedom
  (14 observations deleted due to missingness)
AIC: 4525.8
```

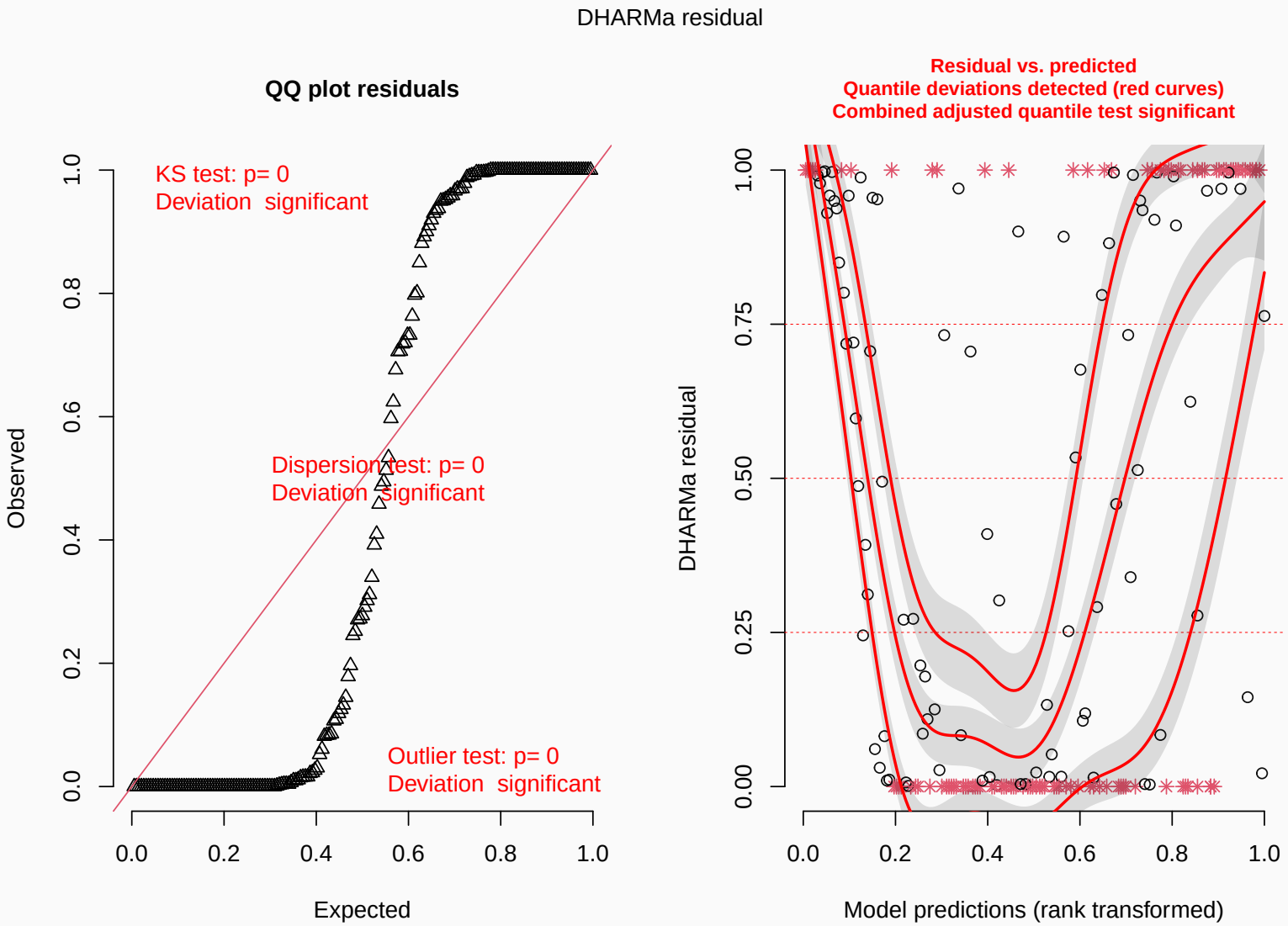
Plot model using visreg:

```
visreg(gdp.glm, scale = "response")
points(mortality/1000 ~ gdp, data = gdp)
```



Residuals diagnostics with DHARMa

```
simulateResiduals(gdp.glm, plot = TRUE)
```



Overdispersion

Overdispersion:

more variation in the data than assumed by statistical model

$$Var(y) = np(1-p)$$

Testing for overdispersion (DHARMa)

```
simres <- simulateResiduals(gdp.glm, refit = TRUE)
testDispersion(simres, plot = FALSE)
```

```
    DHARMa nonparametric dispersion test via mean deviance residual
    vs. simulated-refitted

data:  simres
dispersion = 21, p-value < 2.2e-16
alternative hypothesis: two.sided
```

`quasibinomial` allows us to model overdispersed binomial data

Overdispersion in logistic regression with proportion data

```r
gdp.overdisp <- glm(cbind(mortality, 1000 - mortality) ~ gdp,
                    data = gdp, family = quasibinomial)
```

```
Call:
glm(formula = cbind(mortality, 1000 - mortality) ~ gdp, family = quasibinomial,
    data = gdp)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.657e+00  5.977e-02 -44.465  < 2e-16 ***
gdp         -1.279e-04  1.577e-05  -8.111 5.96e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 20.7947)

    Null deviance: 6430.2  on 192  degrees of freedom
Residual deviance: 3530.2  on 191  degrees of freedom
  (14 observations deleted due to missingness)
AIC: NA
```

Mean estimates do not change after accounting for overdispersion

But standard errors (uncertainty) do!

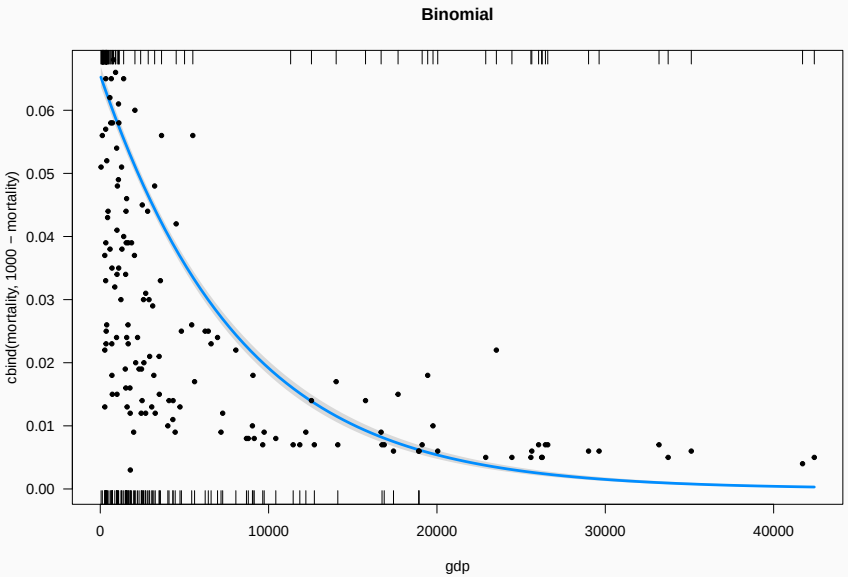
```
parameters(gdp.overdisp)
```

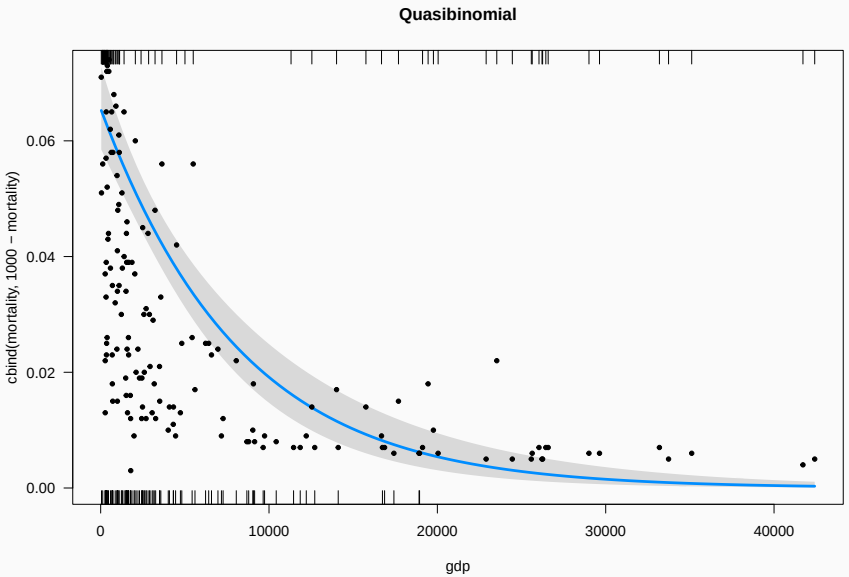
```
Parameter   |  Log-Odds |       SE |         95% CI | t(191) |      p
-------------------------------------------------------------------
(Intercept) |     -2.66 |     0.06 | [-2.78, -2.54] | -44.46 | < .001
gdp         | -1.28e-04 | 1.58e-05 | [ 0.00,  0.00] |  -8.11 | < .001
```

```
parameters(gdp.glm)
```

```
Parameter   |  Log-Odds |       SE |         95% CI |       z |      p
-------------------------------------------------------------------
(Intercept) |     -2.66 |     0.01 | [-2.68, -2.63] | -202.76 | < .001
gdp         | -1.28e-04 | 3.46e-06 | [ 0.00,  0.00] |  -36.99 | < .001
```

But standard errors (uncertainty) do!

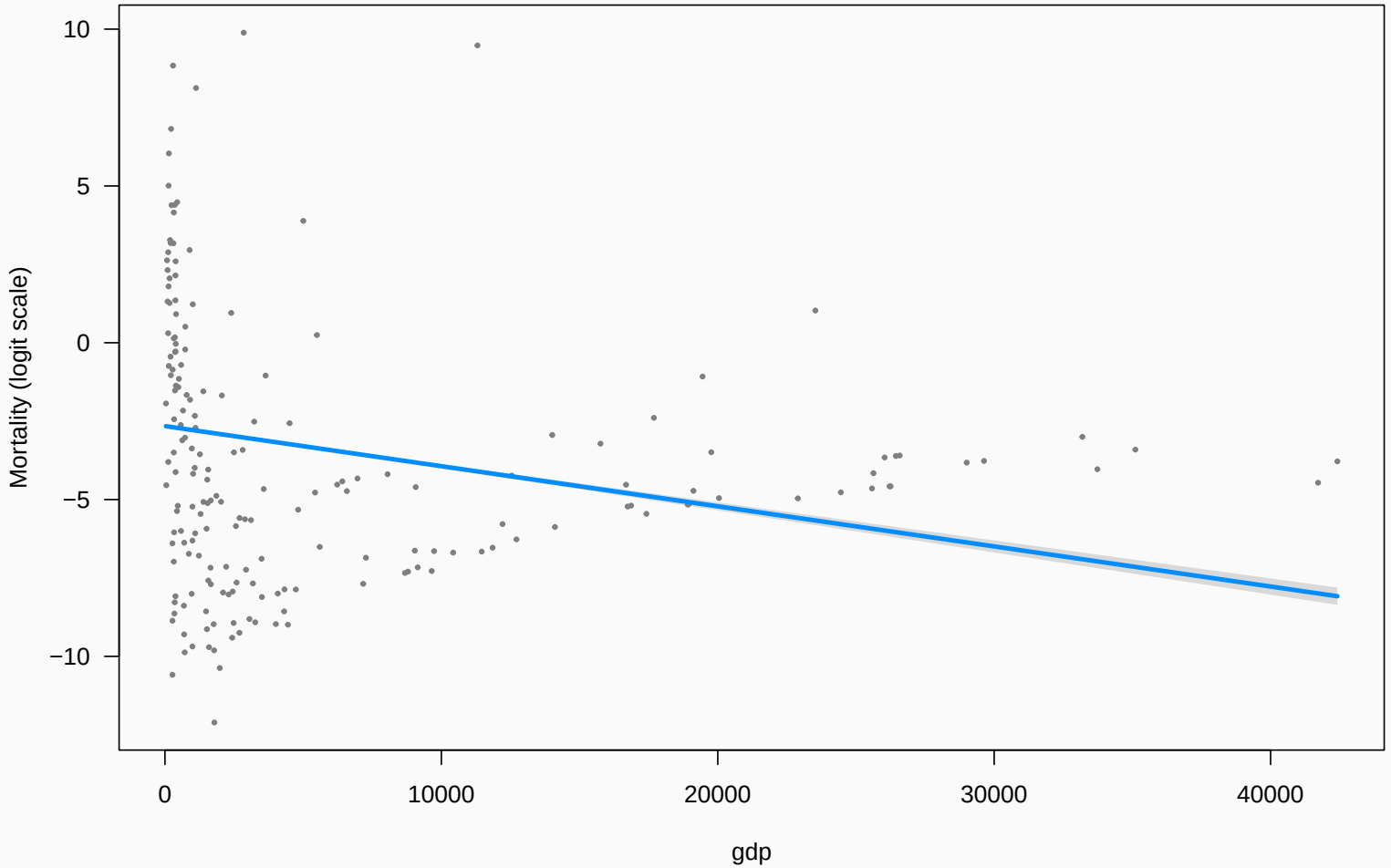




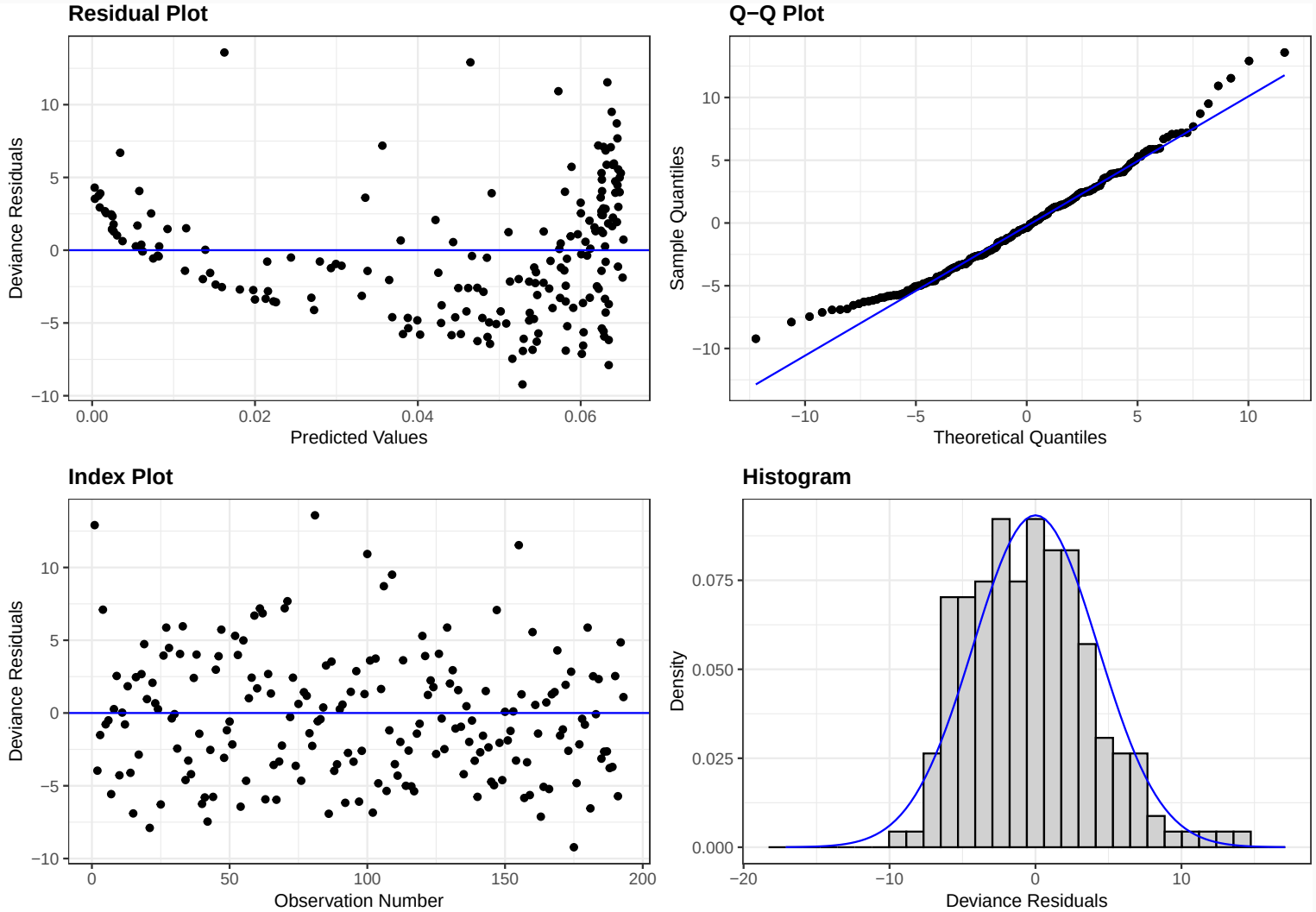
Think about the shape of relationships

Think about the shape of relationships

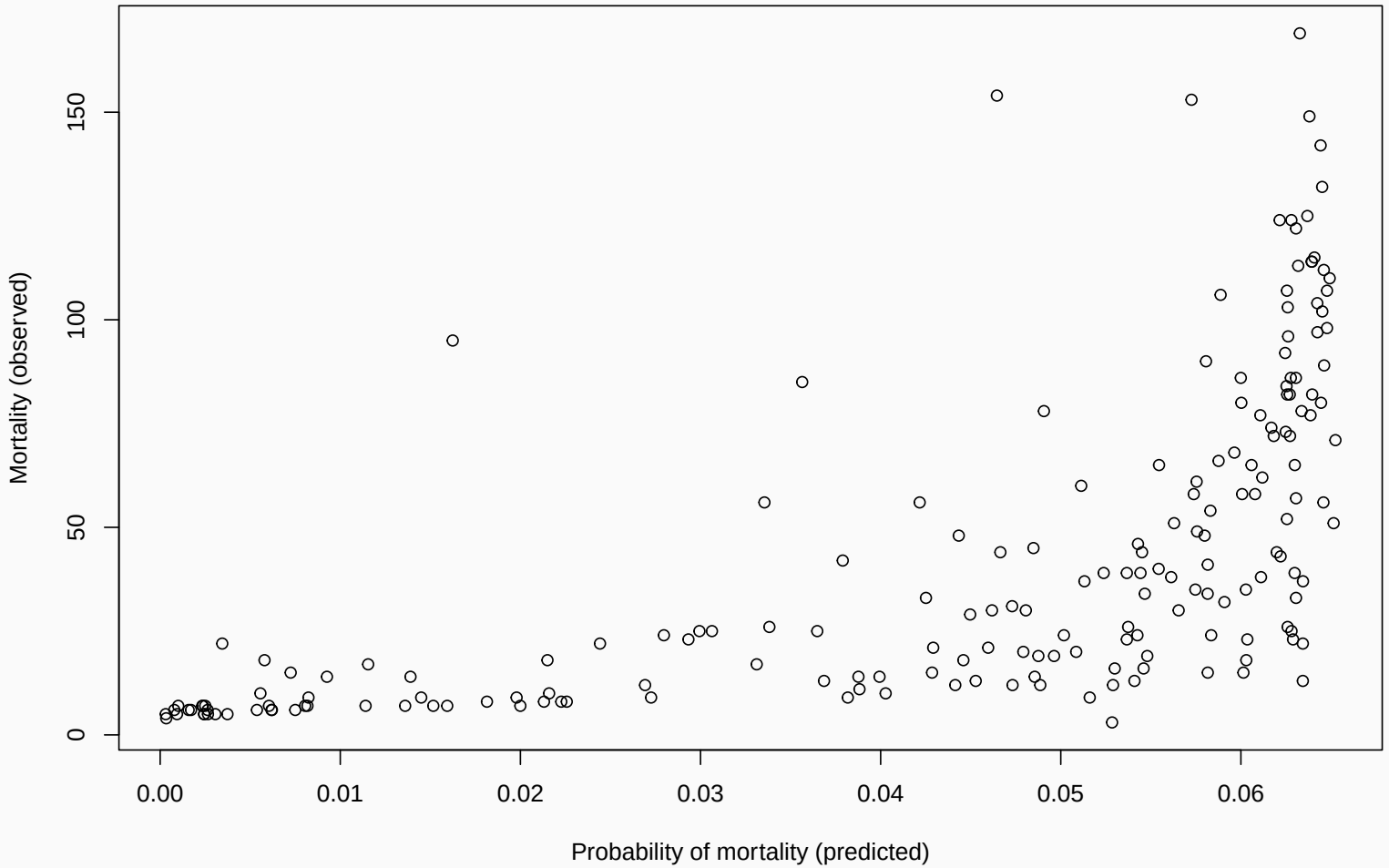
Not everything has to be linear...



Residuals show non-linear pattern

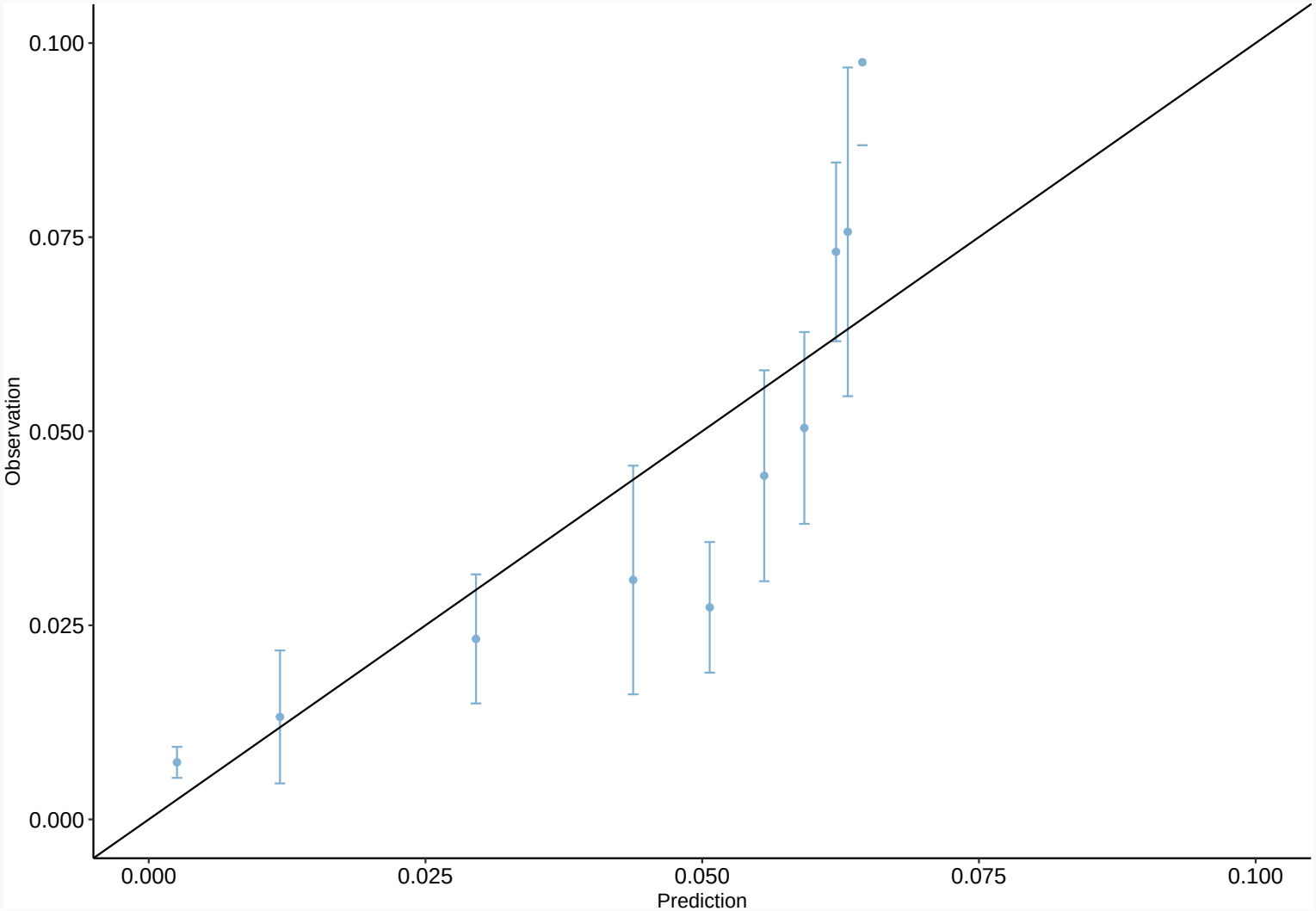


Calibration plot shows non-linear pattern

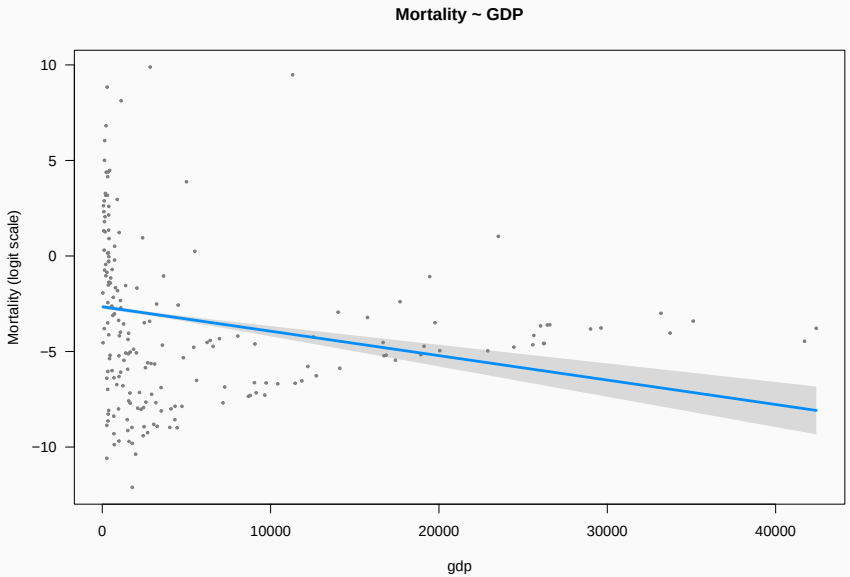


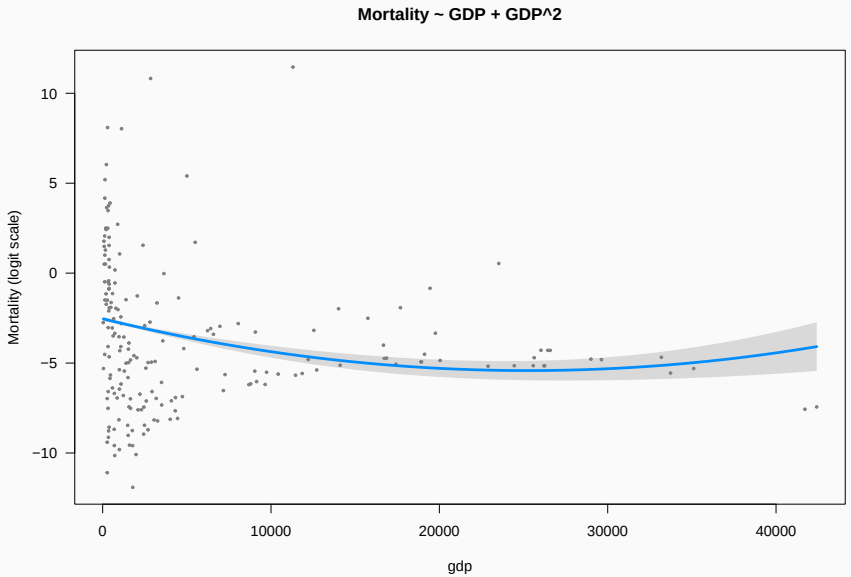
Calibration plot shows non-linear pattern

```
$calibration_plot
```

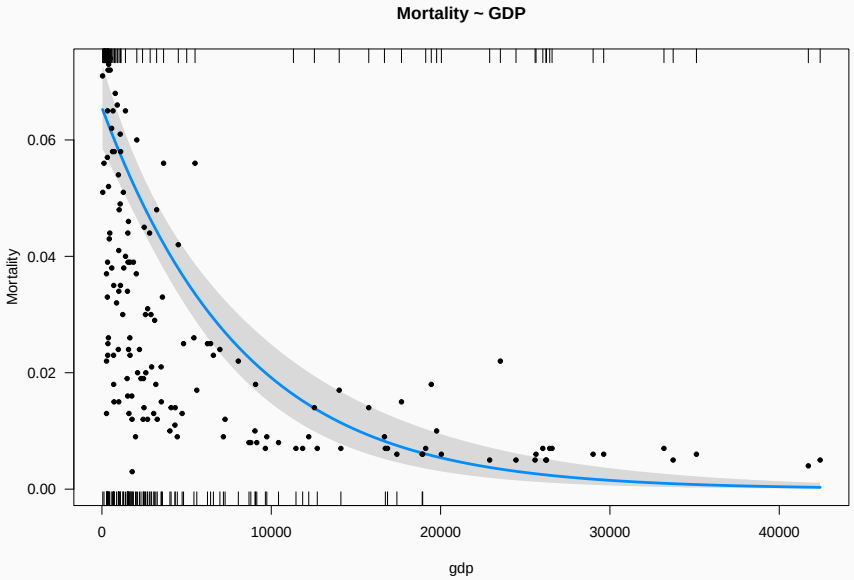


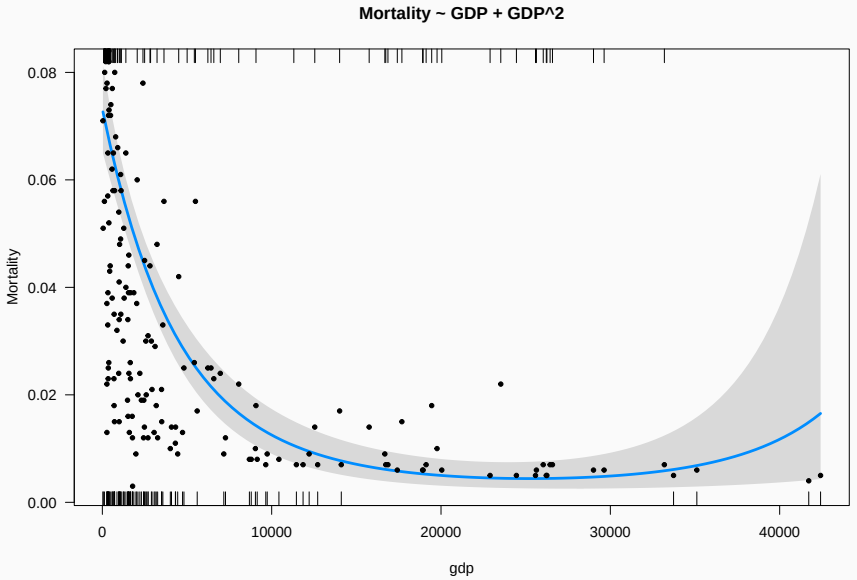
Trying polynomial predictor (GDP + GDP^2)



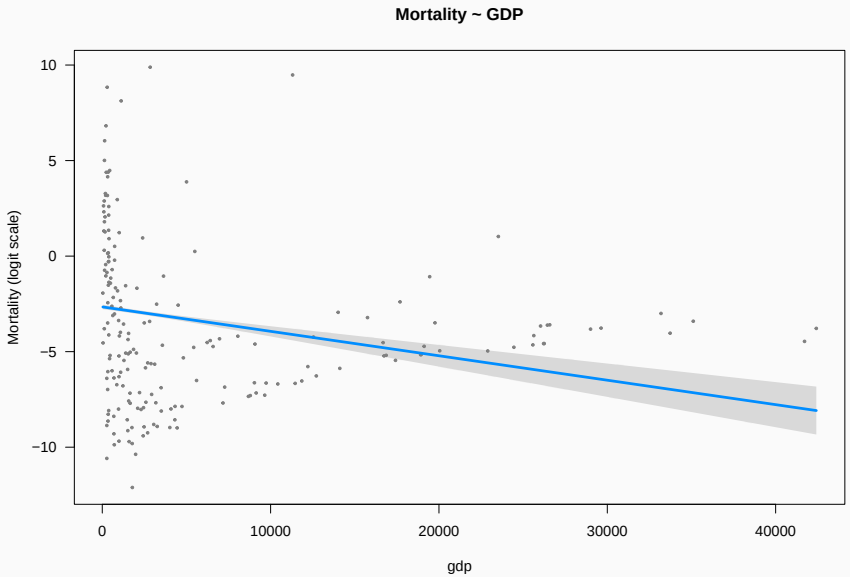


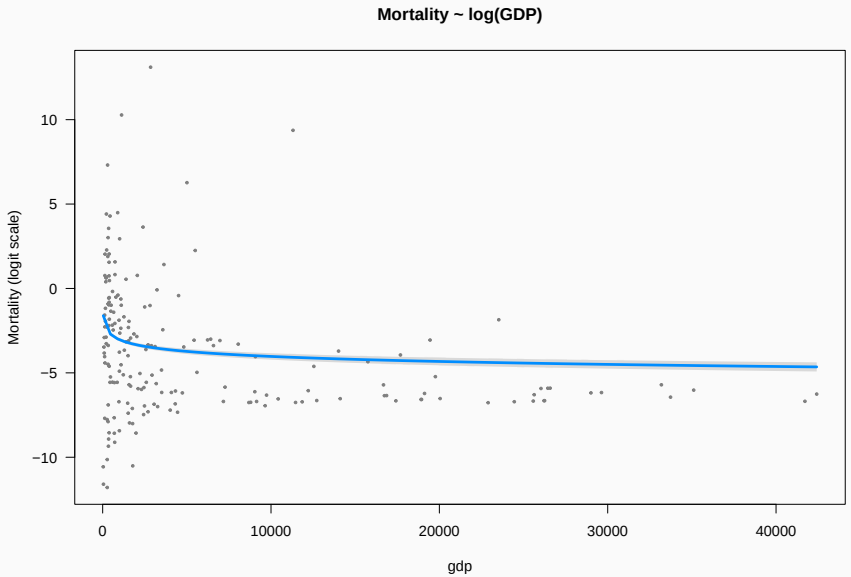
Think about the shape of relationships



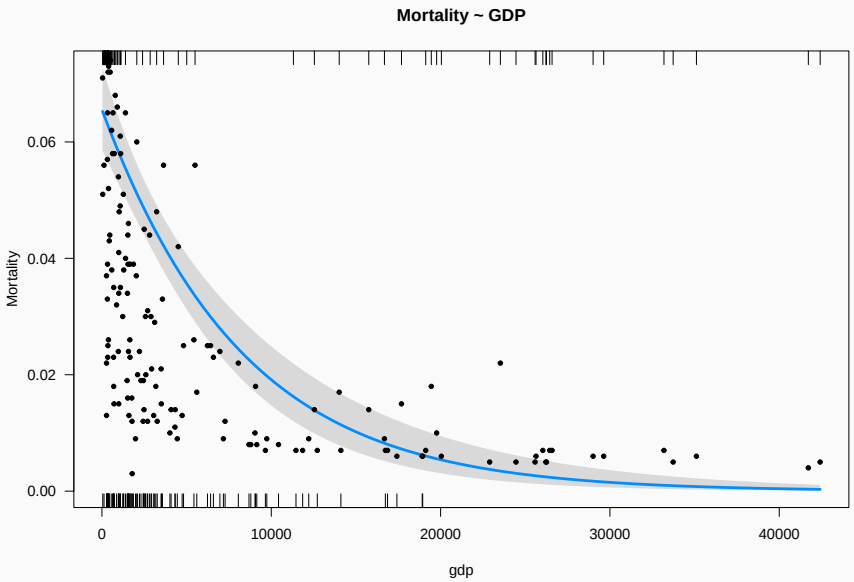


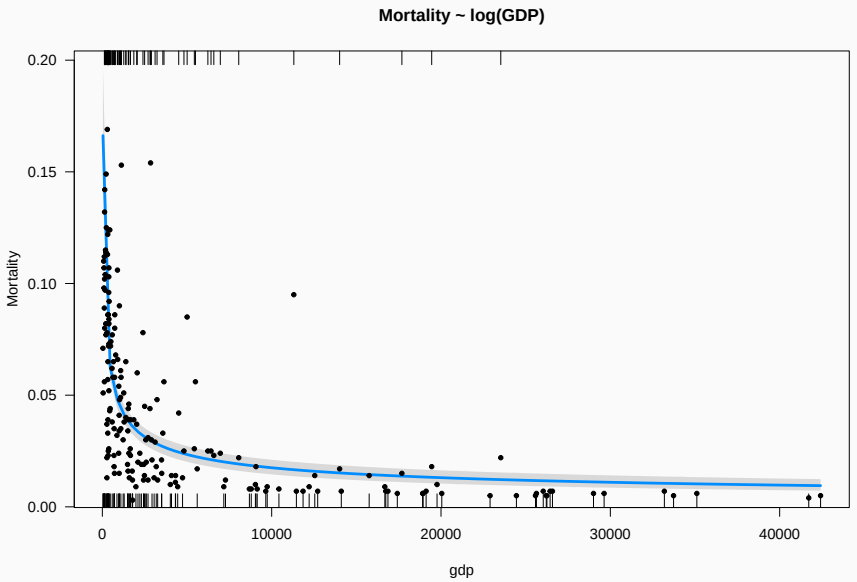
Trying log(GDP)





Trying log(GDP)





More examples

- `moth.csv`: Probability of moth predation on trunk trees depending on morph (light/dark) and distance to Liverpool (Bishop 1972)

More examples

- `moth.csv`: Probability of moth predation on trunk trees depending on morph (light/dark) and distance to Liverpool (Bishop 1972)
- `seedset.csv`: Comparing seed set among plants (Data from Harder et al. 2011)

More examples

- `moth.csv`: Probability of moth predation on trunk trees depending on morph (light/dark) and distance to Liverpool (Bishop 1972)
- `seedset.csv`: Comparing seed set among plants (Data from Harder et al. 2011)
- `soccer.csv`: Probability of scoring penalty depending on goalkeeper's team being ahead, behind or tied (Roskes et al 2011)

Moth predation

The industrial revolution and evolution of dark morphs

The data

```
moth <- read.csv("data/moth.csv")
```

```
  MORPH DISTANCE PLACED REMOVED
1 light      0.0     56      17
2  dark      0.0     56      14
3 light      7.2     80      28
4  dark      7.2     80      20
5 light     24.1     52      18
6  dark     24.1     52      22
```

Creating new variable: REMAIN

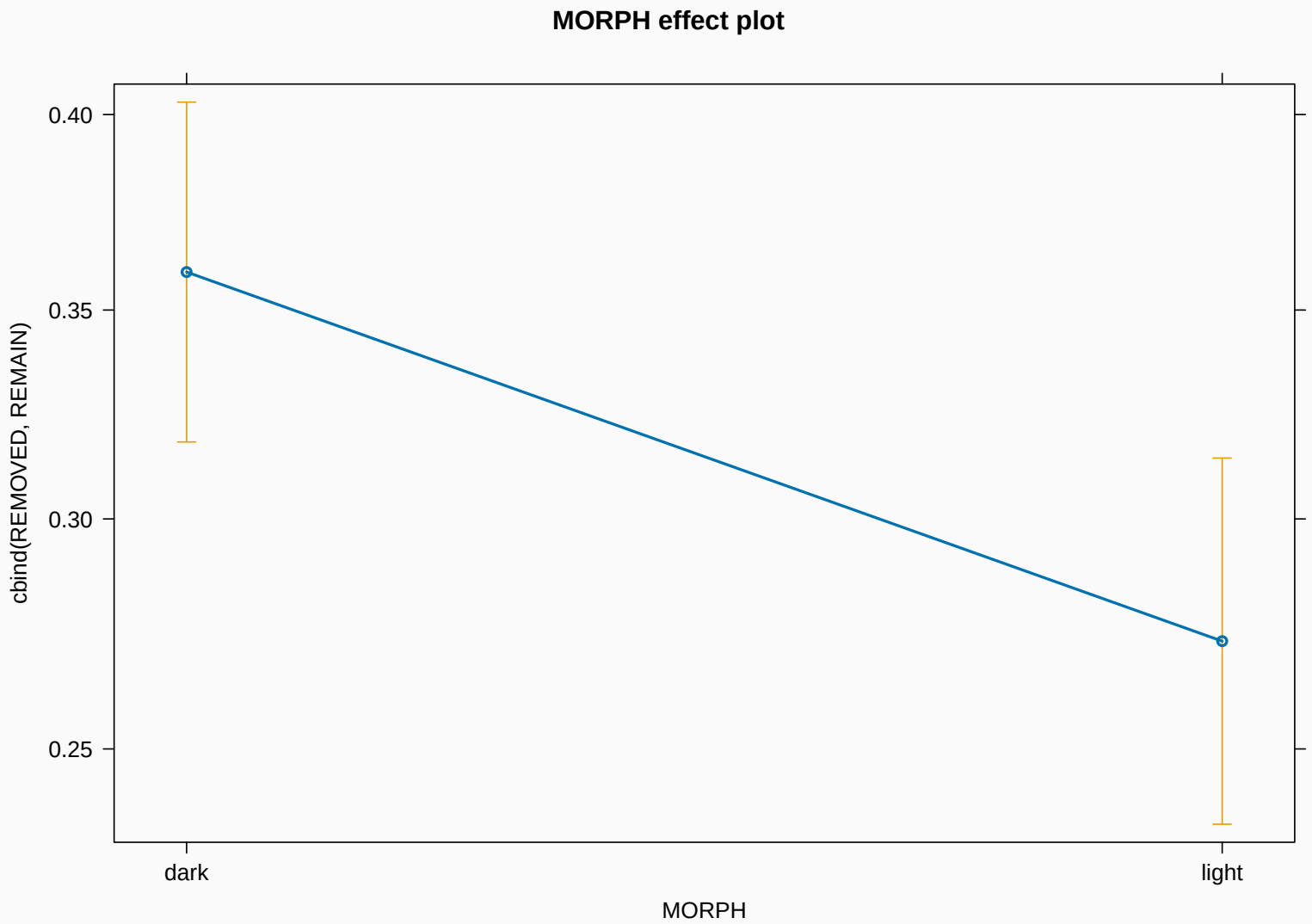
```
moth$REMAIN <- moth$PLACED - moth$REMOVED
```

```
  MORPH DISTANCE PLACED REMOVED REMAIN
1 light      0.0     56      17     39
2  dark      0.0     56      14     42
3 light      7.2     80      28     52
4  dark      7.2     80      20     60
5 light     24.1     52      18     34
6  dark     24.1     52      22     30
```

Did some morph have higher predation overall?

```
Call:
glm(formula = cbind(REMOVED, REMAIN) ~ MORPH, family = binomial,
    data = moth)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.57752    0.09473  -6.097 1.08e-09 ***
MORPHlight  -0.40331    0.13925  -2.896  0.00377 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 35.385  on 13  degrees of freedom
Residual deviance: 26.936  on 12  degrees of freedom
AIC: 93.61

Number of Fisher Scoring iterations: 4
```

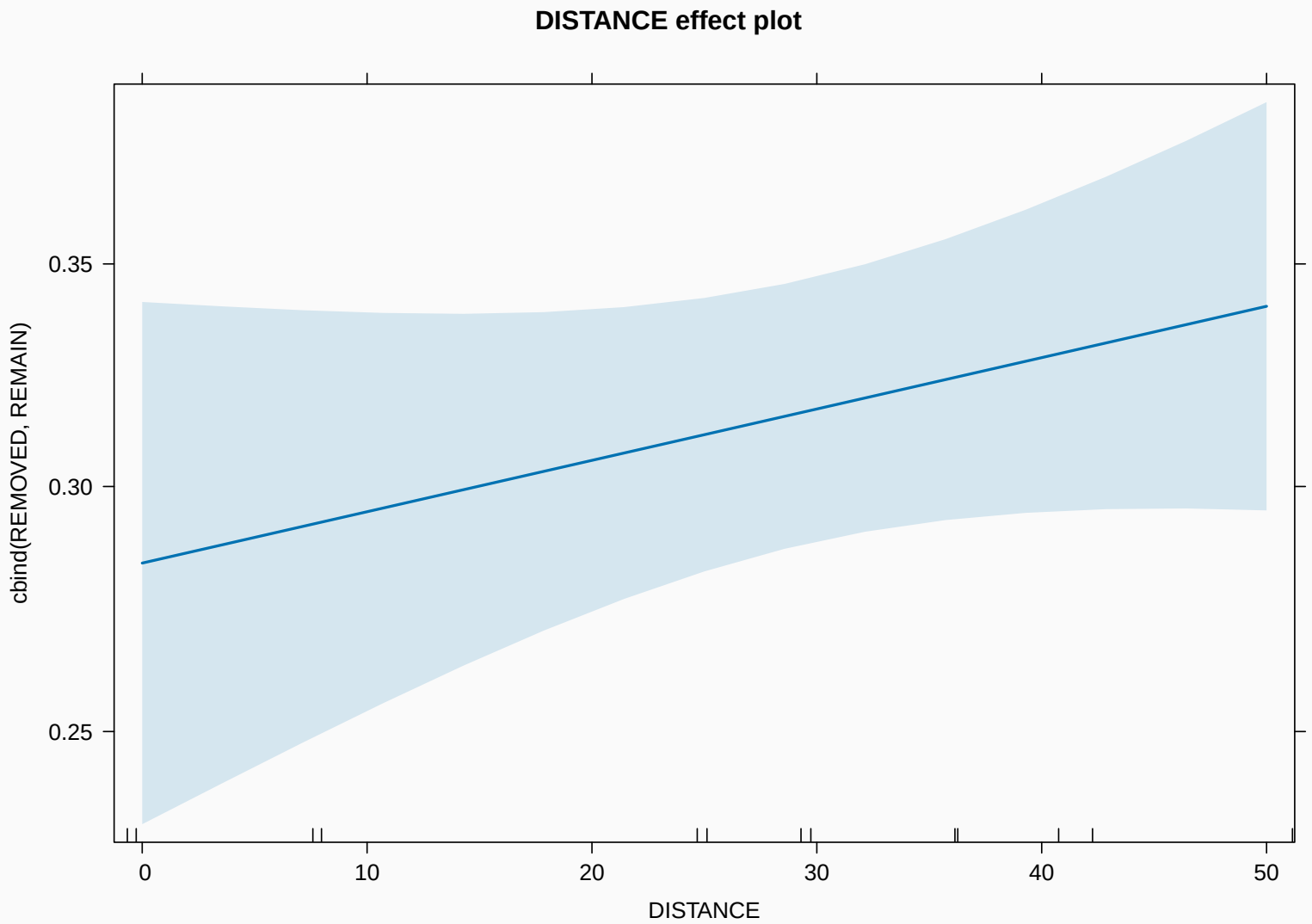
Did some morph have higher predation overall?



Did predation increase farther from city centre?

```
Call:
glm(formula = cbind(REMOVED, REMAIN) ~ DISTANCE, family = binomial,
    data = moth)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.925861   0.136634  -6.776 1.23e-11 ***
DISTANCE     0.005268   0.003984   1.322    0.186
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 35.385  on 13  degrees of freedom
Residual deviance: 33.626  on 12  degrees of freedom
AIC: 100.3

Number of Fisher Scoring iterations: 4
```

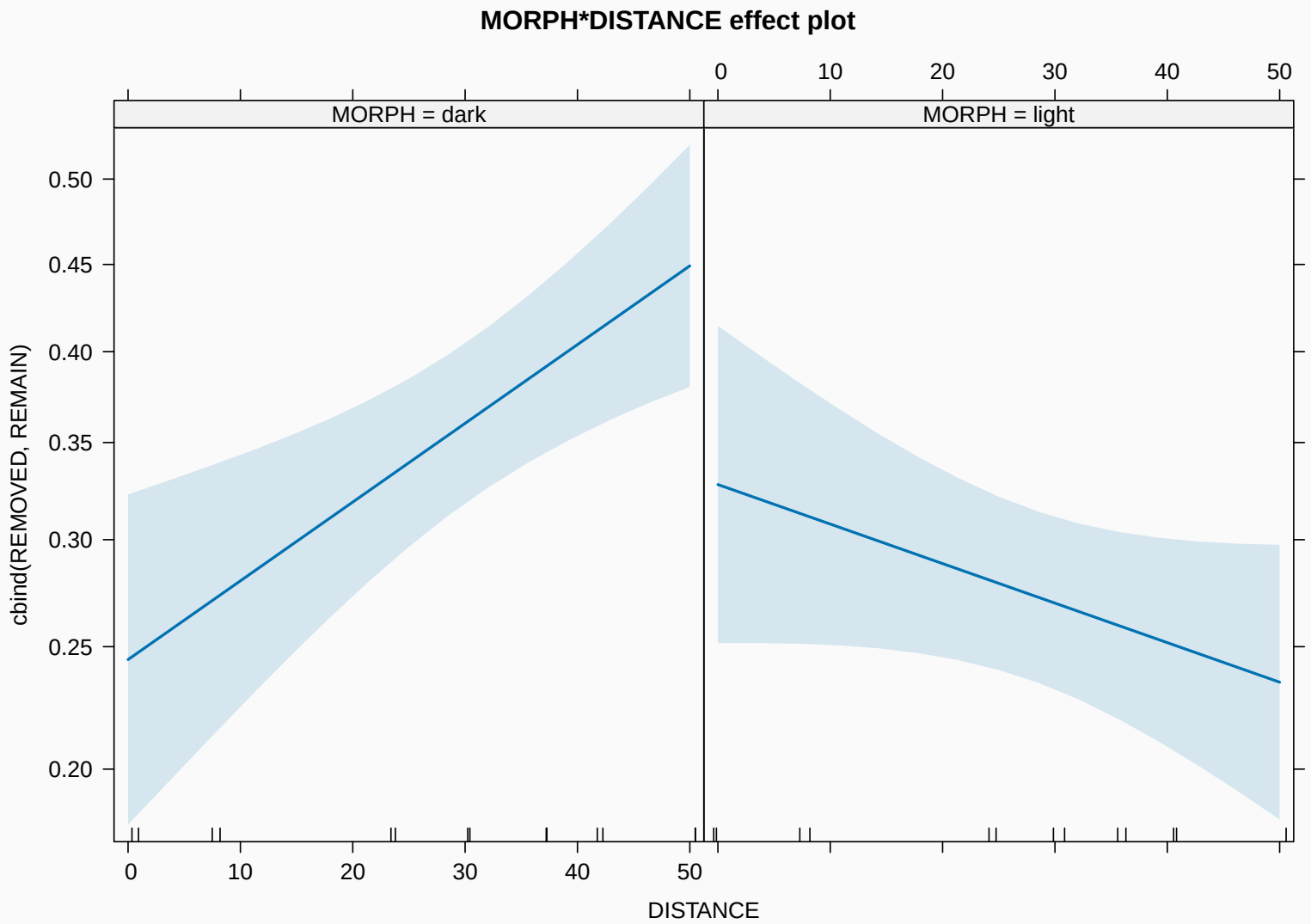
Did predation increase farther from city centre?



Did dark morph have lower predation in city & light have lower predation in countryside?

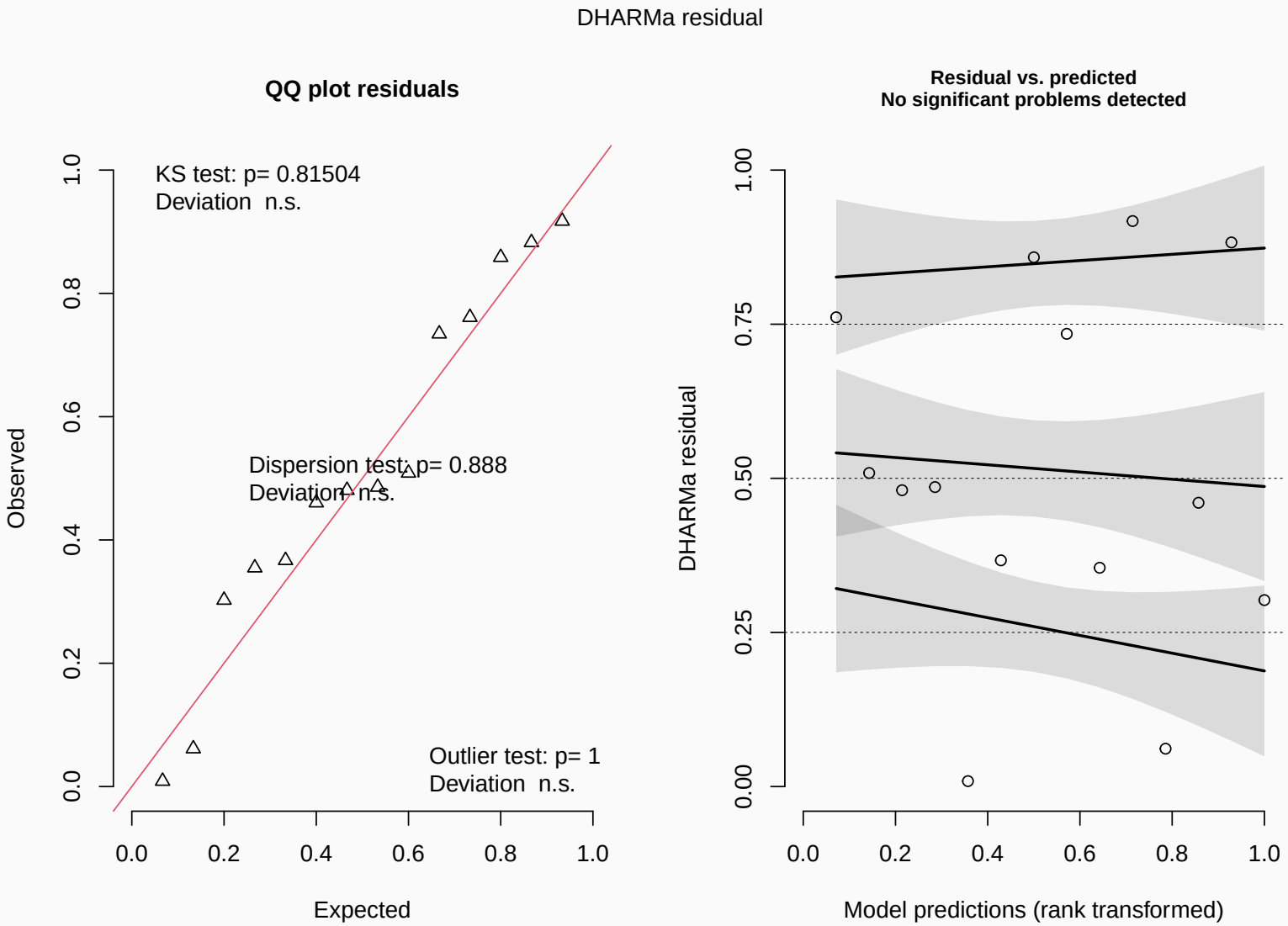
```
Call:
glm(formula = cbind(REMOVED, REMAIN) ~ MORPH * DISTANCE, family = binomial,
    data = moth)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.128987   0.197906  -5.705 1.17e-08 ***
MORPHlight           0.411257   0.274490   1.498 0.134066
DISTANCE             0.018502   0.005645   3.277 0.001048 **
MORPHlight:DISTANCE -0.027789   0.008085  -3.437 0.000588 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 35.385  on 13  degrees of freedom
Residual deviance: 13.230  on 10  degrees of freedom
AIC: 83.904
```

Did dark morph have lower predation in city & light have lower predation in countryside?



Model check

```
simulateResiduals(pred.int, plot = TRUE)
```



Seed set among plants

Seed set among plants

Seed set among plants

```
# A tibble: 6 x 6
  species    plant pcmass fertilized seeds ovulecnt
  <chr>      <dbl>  <dbl>      <dbl> <dbl>    <dbl>
1 ferruginea     2  0             70    52      330
2 ferruginea     2  0.2          321   188      461
3 ferruginea     2  0.485        351   278      435
4 ferruginea     2  0.737        386   301      430
5 ferruginea     2  1            367   342      419
6 ferruginea     3  0            185    39      470
```

Questions:

https://pollev.com/franciscorod726

- Is seed set related to proportion of outcross pollen (pcmass)?

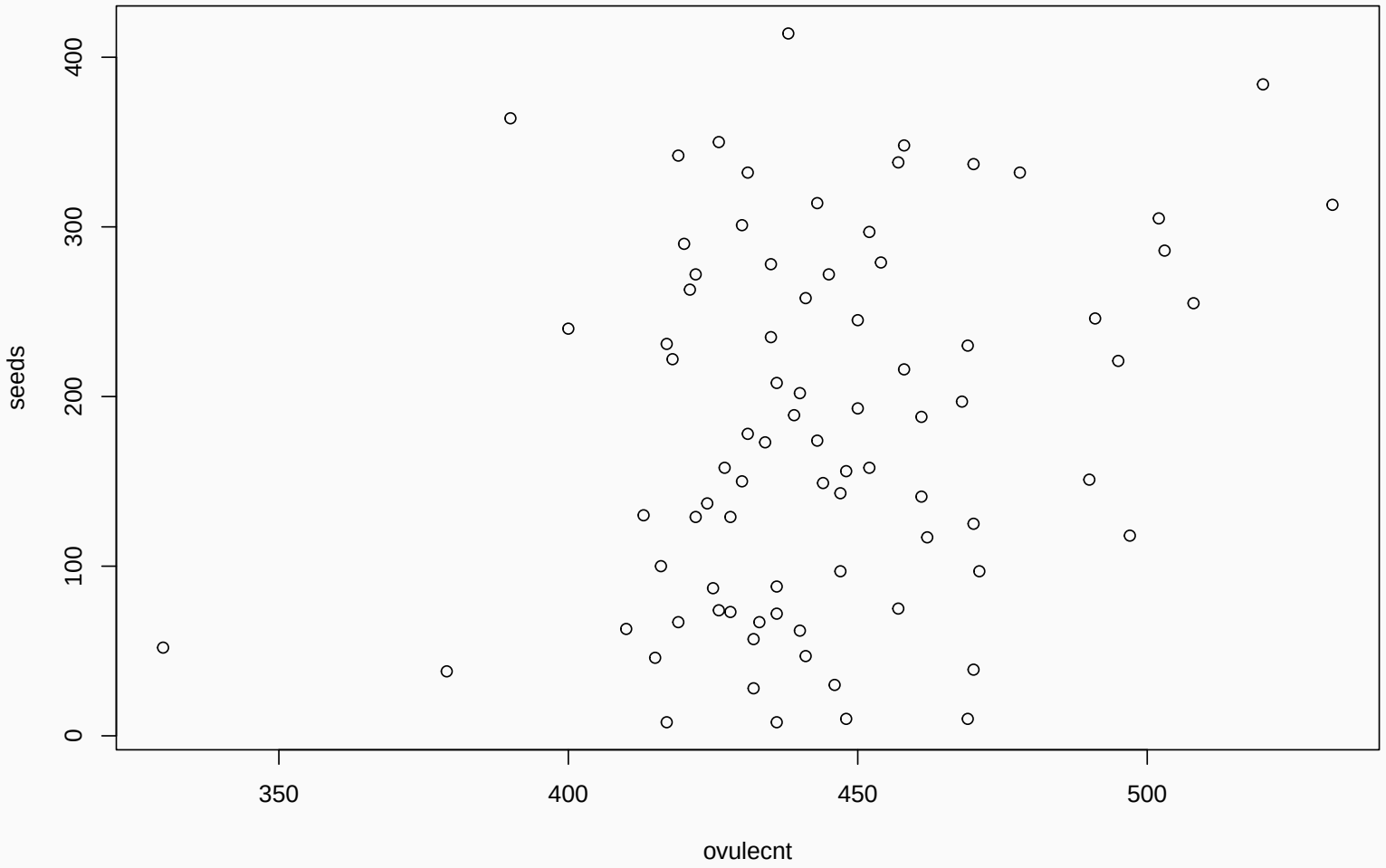
Questions:

https://pollev.com/franciscorod726

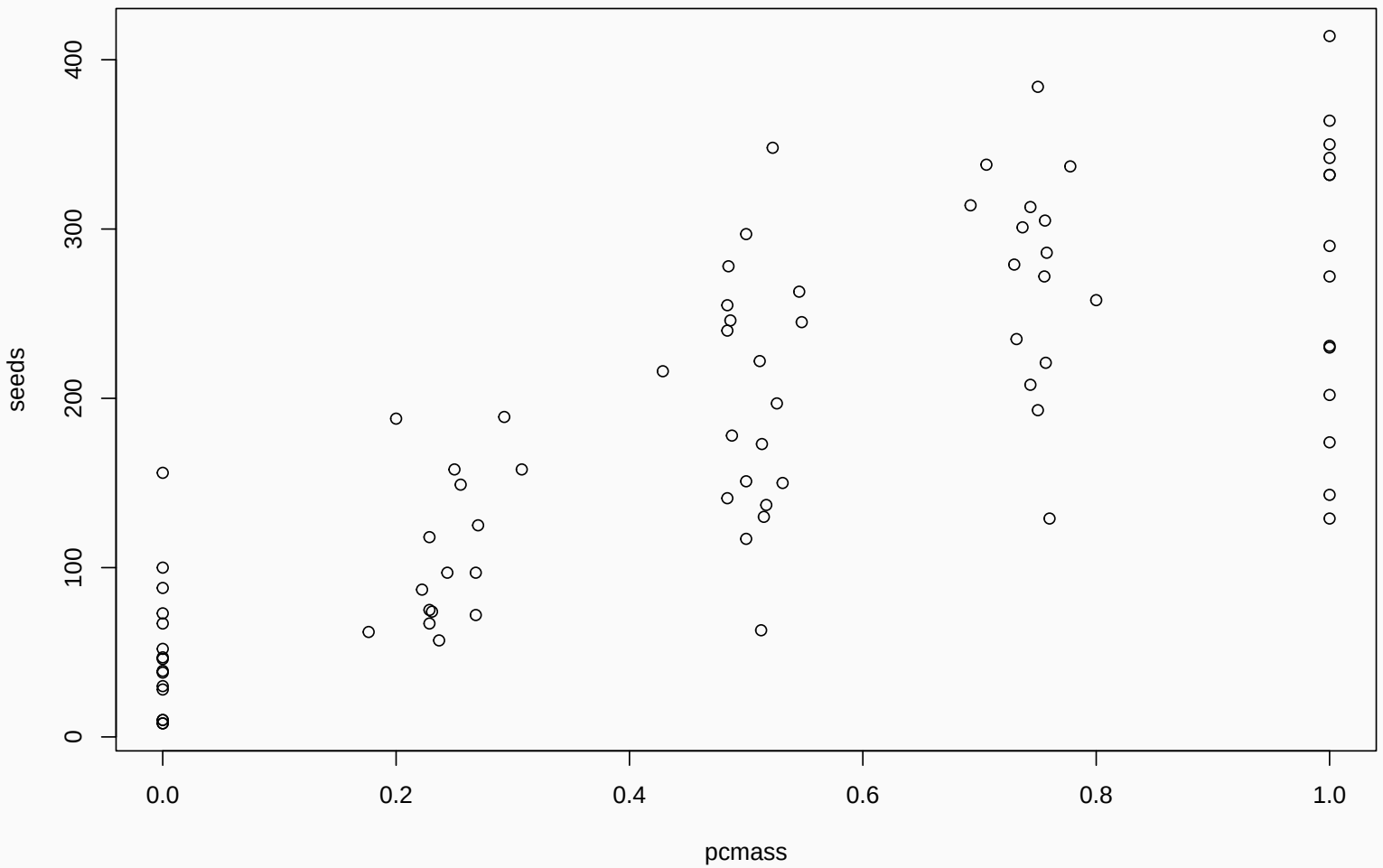
- Is seed set related to proportion of outcross pollen (pcmass)?

- Which plant had lower seed set?

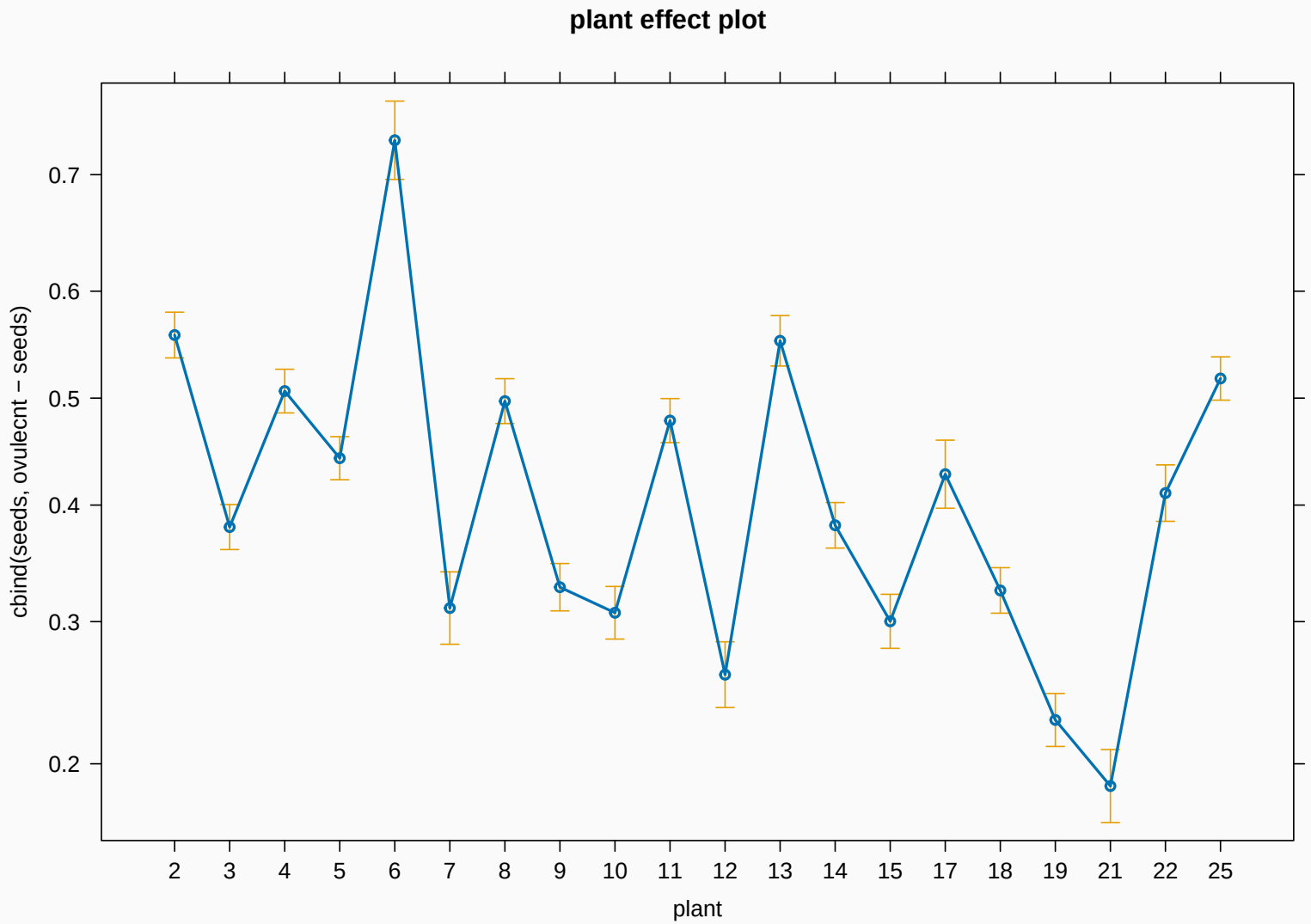
Number of seeds vs Number of ovules



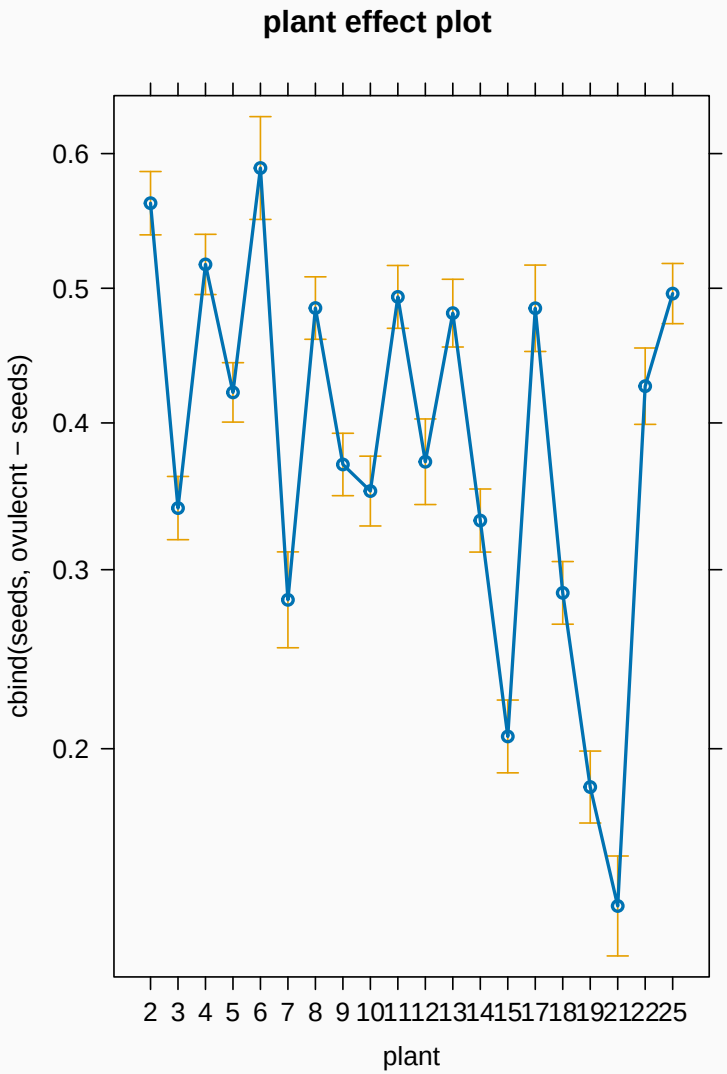
Number of seeds vs Proportion outcross pollen

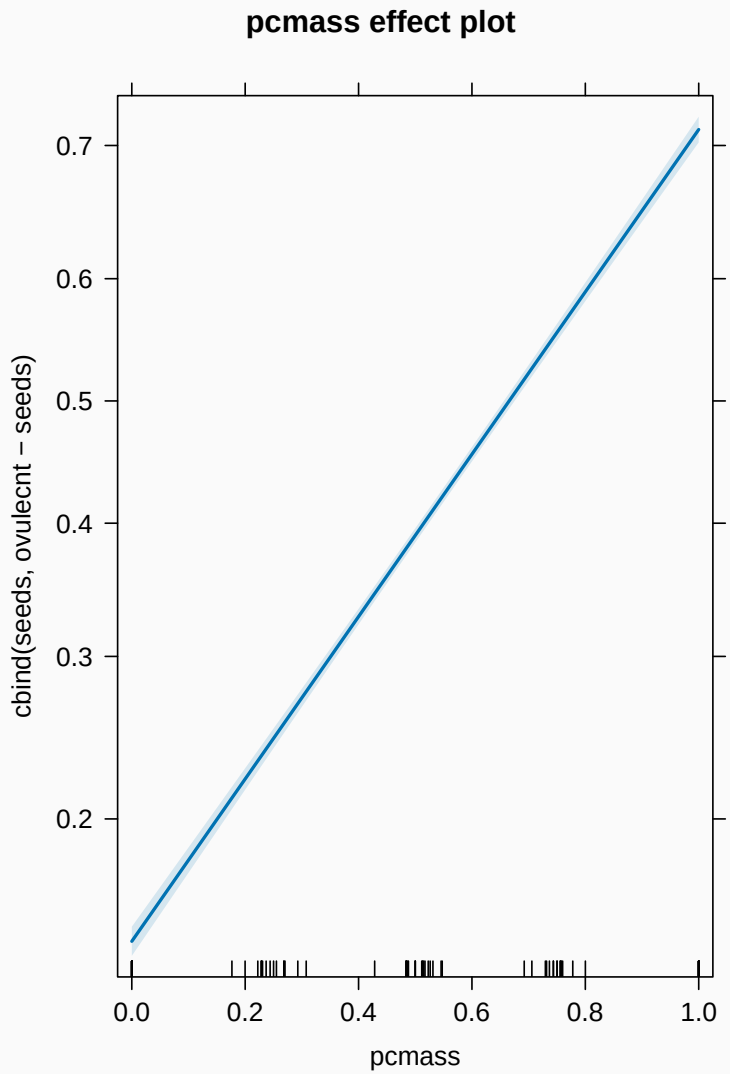


Seed set across plants



Seed set ~ outcross pollen





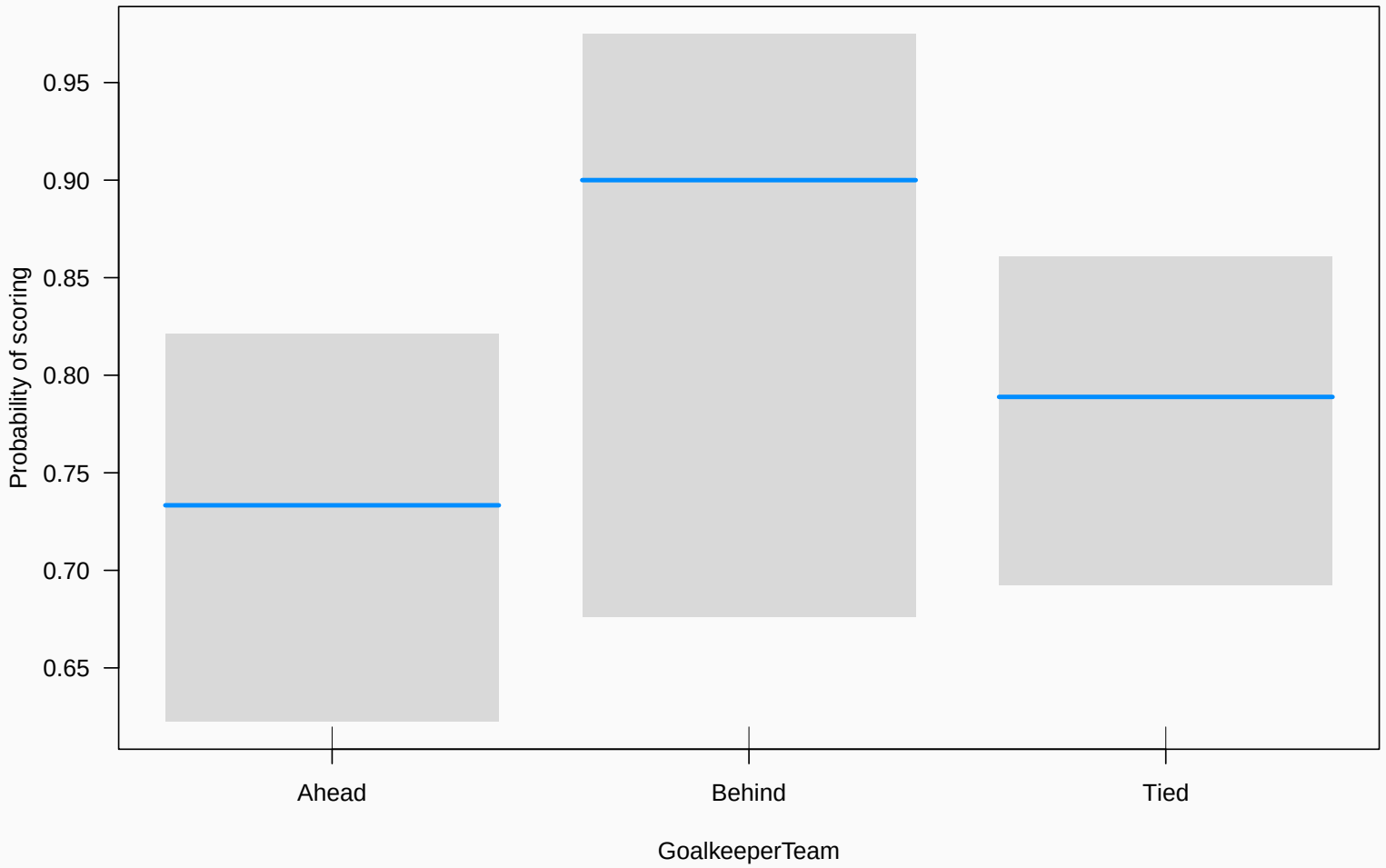
Probability of scoring penalty

Data on penalty shots

```
soccer <- read.csv("data/soccer.csv")
soccer
```

```
  GoalkeeperTeam Nshots Scored
1         Behind     20     18
2           Tied     90     71
3          Ahead     75     55
```

Does probability of scoring penalty depends on match situation?

https://pollev.com/franciscorod726

Probability of scoring depending on match situation