STATS 330

Handout 4
Deviance

Department of Statistics, University of Auckland

Goodness-of-fit

Statistical analysis does not simply end once we have fitted a model. We must determine whether or not our model seems appropriate.

'Goodness-of-fit' is a property that describes how well the data appear to fit the model's assumptions. For a generalised linear model these assumptions are

- ▶ The observations are independent.
- ▶ $g(\boldsymbol{\theta}) = \boldsymbol{X}\boldsymbol{\beta}$; after applying the link function, the parameter of interest is a linear combination of the explanatory terms.
- ▶ Each response comes from the assumed distribution.

Goodness-of-fit

This handout introduces the concept of deviance, a popular goodness-of-fit statistic for generalised linear models. We will cover the following topics:

1. The null model
2. The saturated model
3. Deviance
4. Null deviance

Model complexity, residual sum of squares, and likelihood

Consider a simple GLM:

$g(\theta_i) = \beta_0 + \beta_1 x_i$

We can add all sorts of additional explanatory terms:

Model complexity, residual sum of squares, and likelihood

Consider a simple GLM:

$$g(\theta_i) = \beta_0 + \beta_1 x_i + \beta_2 z_i$$

We can add all sorts of additional explanatory terms:

- ▶ Terms for additional numeric explanatory variables

Model complexity, residual sum of squares, and likelihood

Consider a simple GLM:

$$g(\theta_i) = \beta_0 + \beta_1 x_i + \beta_2 a_i + \beta_3 b_i$$

We can add all sorts of additional explanatory terms:

- ▶ Terms for additional numeric explanatory variables
- ▶ Terms for factors, via dummy variables

Model complexity, residual sum of squares, and likelihood

Consider a simple GLM:

$$g(\theta_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

We can add all sorts of additional explanatory terms:

- ▶ Terms for additional numeric explanatory variables
- ▶ Terms for factors, via dummy variables
- ▶ Terms for a nonlinear relationships

Model complexity, residual sum of squares, and likelihood

Consider a simple GLM:

$$g(\theta_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3$$

We can add all sorts of additional explanatory terms:

- ▶ Terms for additional numeric explanatory variables
- ▶ Terms for factors, via dummy variables
- ▶ Terms for a nonlinear relationships

Model complexity, residual sum of squares, and likelihood

Consider a simple GLM:

$$g(\theta_i) = \beta_0 + \beta_1 x_i + \beta_4 x_i a_i + \beta_5 x_i b_i$$

We can add all sorts of additional explanatory terms:

- ▶ Terms for additional numeric explanatory variables
- ▶ Terms for factors, via dummy variables
- ▶ Terms for a nonlinear relationships
- ▶ Terms for interactions

Model complexity, residual sum of squares, and likelihood

Consider a complicated GLM:

$$g(\theta_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 z_i \beta_5 a_i + \beta_6 b_i + \beta_6 x_i a_i + \beta_7 x_i b_i$$

We can add all sorts of additional explanatory terms:

- ▶ Terms for additional numeric explanatory variables
- ▶ Terms for factors, via dummy variables
- ▶ Terms for a nonlinear relationships
- ▶ Terms for interactions
- ▶ All of the above...

Model complexity, residual sum of squares, and likelihood

Consider a complicated GLM:

$$g(\theta_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 z_i \beta_5 a_i + \beta_6 b_i + \beta_6 x_i a_i + \beta_7 x_i b_i$$

We can add all sorts of additional explanatory terms:

- ▶ Terms for additional numeric explanatory variables
- ▶ Terms for factors, via dummy variables
- ▶ Terms for a nonlinear relationships
- ▶ Terms for interactions
- ▶ All of the above...

Adding explanatory terms to a model is sometimes said to make it more 'complex' or increase its 'complexity'

Model complexity, residual sum of squares, and likelihood

Increasing a model's complexity will almost always

- ▶ decrease the residual sum of squares (linear regression), or
- ▶ increase the maximised log-likelihood (GLMs)

(In rare circumstances, adding explanatory terms may lead to no change in these statistics, but only if the estimated coefficients of the additional terms are all exactly zero.)

Increasing complexity will never

- ▶ increase the residual sum of squares (linear regression), or
- ▶ decrease the maximised log-likelihood (GLMs)

Let's consider adding successive polynomial terms to our working models...

Model complexity, residual sum of squares, and likelihood

Linear regression: LSD effect on maths analysis


```
lm(score ~ lsd)
```

Predicted LSD tissue concentration (nanograms per ml)

Model complexity, residual sum of squares, and likelihood

Linear regression: LSD effect on maths analysis

```
lm(score ~ lsd + I(lsd^2)))
```


Test performance score vs Predicted LSD tissue concentration (nanograms per ml). RSS: 251.09
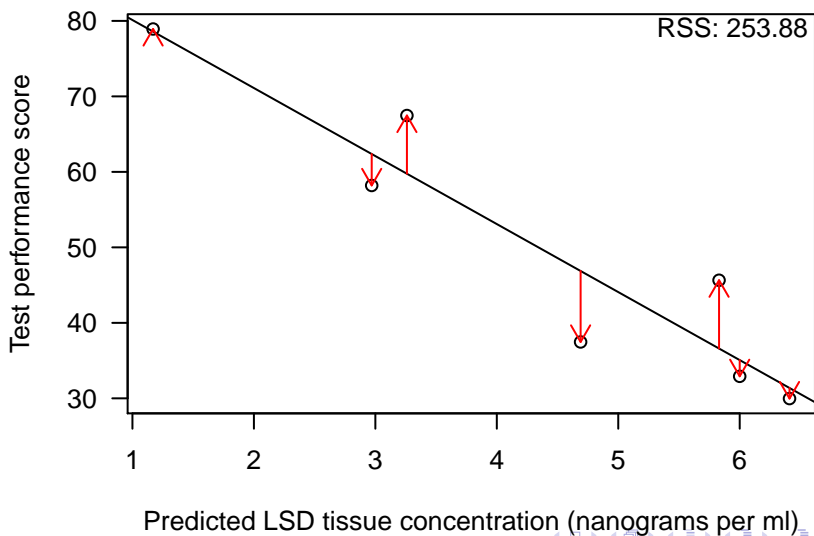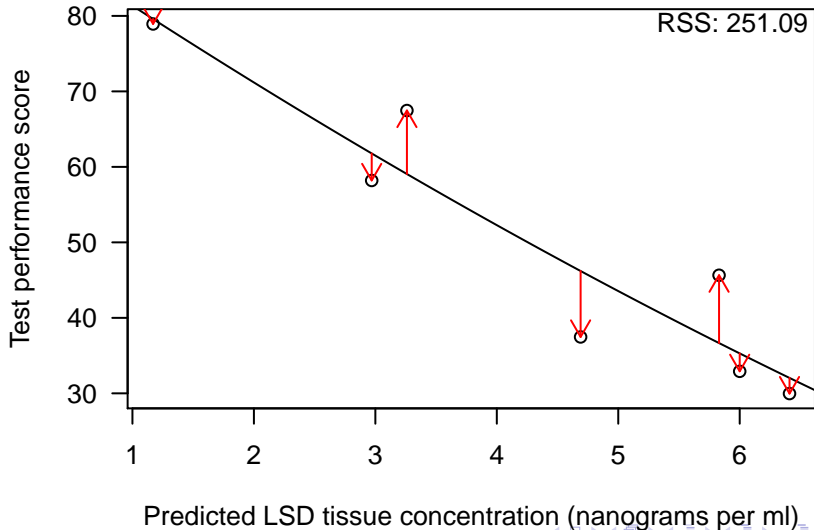
Model complexity, residual sum of squares, and likelihood

Linear regression: LSD effect on maths analysis

```r
lm(score ~ poly(lsd, 2, raw = FALSE)) ## A shortcut
```
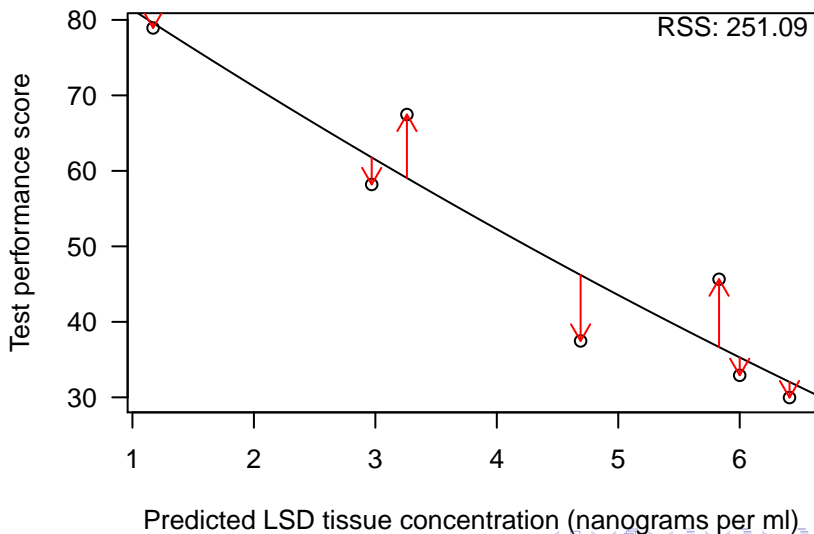

RSS: 251.09

Test performance score (y-axis)

Predicted LSD tissue concentration (nanograms per ml)

Model complexity, residual sum of squares, and likelihood

Linear regression: LSD effect on maths analysis
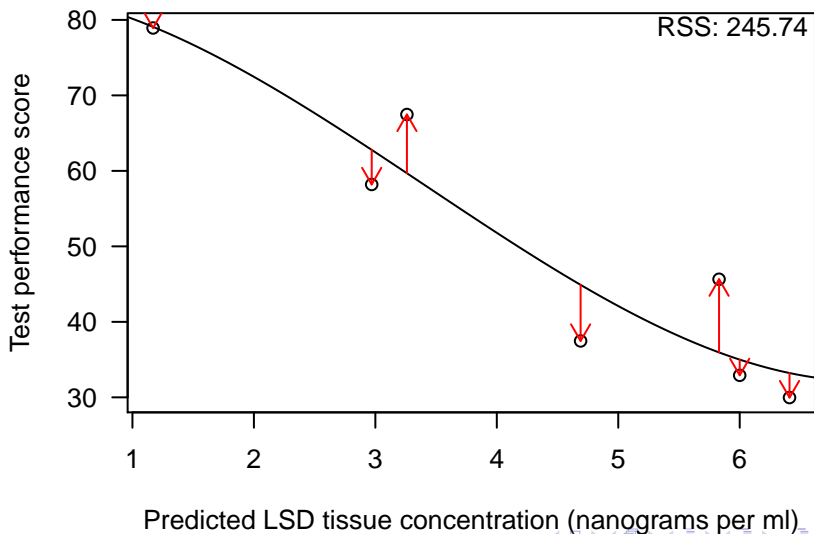
Model complexity, residual sum of squares, and likelihood

Linear regression: LSD effect on maths analysis

```
lm(score ~ poly(lsd, 4, raw = TRUE))
```
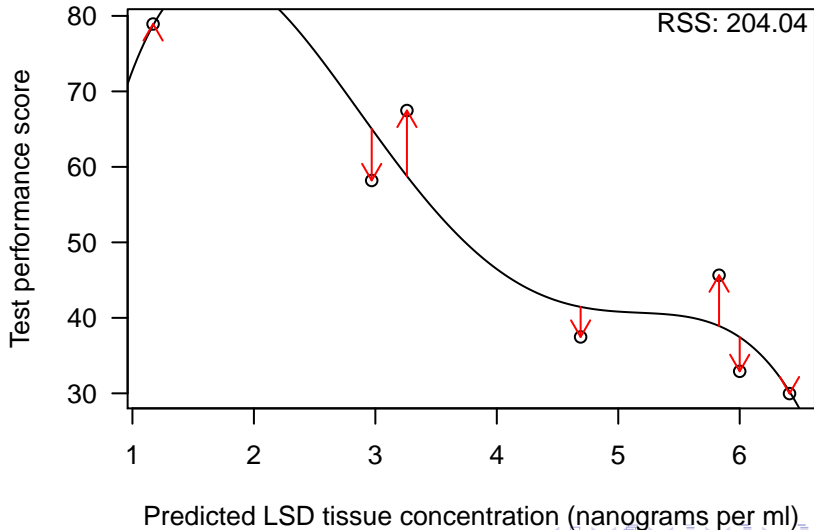

RSS: 204.04

Test performance score (y-axis)

Predicted LSD tissue concentration (nanograms per ml)

Model complexity, residual sum of squares, and likelihood

Linear regression: LSD effect on maths analysis


```
lm(score ~ poly(lsd, 5, raw = TRUE))
```
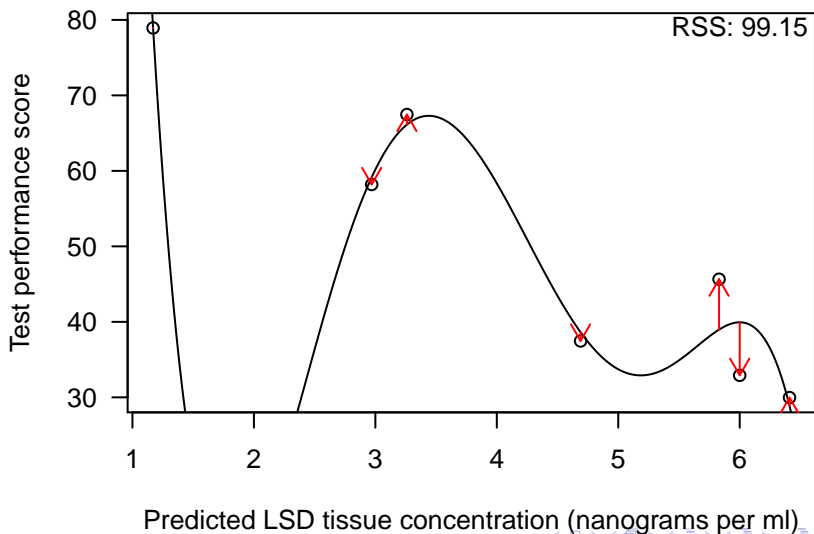
Model complexity, residual sum of squares, and likelihood

Linear regression: LSD effect on maths analysis

```
lm(score ~ poly(lsd, 6, raw = TRUE))
```


Predicted LSD tissue concentration (nanograms per ml)
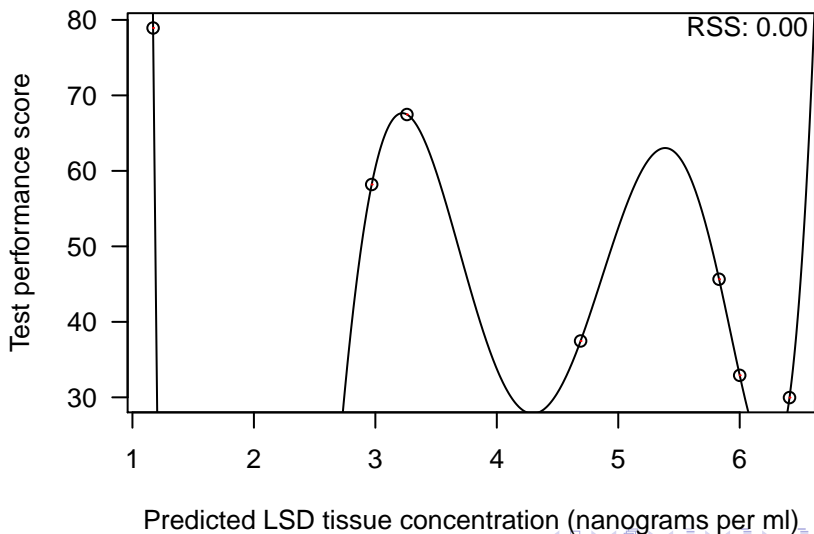
The saturated model

Linear regression

From the plots:

- ▶ As we increase the number of model parameters, the fitted line becomes more 'wiggly'.
- ▶ The more 'wiggly', the lower the RSS.
- ▶ Eventually it is 'wiggly' enough to go through every point, and the RSS is zero.

A model with as many parameters as there are observations is called a saturated model:

- ▶ This is the most complex model you can fit.
- ▶ You cannot add any further explanatory terms.
- ▶ The regression line will go through every point.
- ▶ The RSS will always be equal to zero.

The saturated model

An aside

In almost all cases, the saturated model is not a sensible model:

- ▶ If we were to take another sample from the population, their observations would not exhibit exactly the same 'wiggliness'
- ▶ The exact pattern of 'wiggliness' we see therefore reflects random white noise inherent in our particular sample, and not some underlying relationship that exists in the population

Overcomplicated models, such as the saturated model, are said to be 'overfitted':

- ▶ We should try to avoid overfitting models, so that they only describe underlying trends in the population, and not random noise from our sample.
- ▶ More on model selection coming later in the course!

The null model

The null model has no explanatory variables whatsoever:

$$g(\theta_i) = \beta_0$$

We fit can fit the null model in R as follows:

```
lsd.null.fit <- lm(score ~ 1)
```

As long as we have an intercept term in our models, the null model's RSS will be at least as large as that of any model with explanatory variables, so

$$\text{RSS}_S \leq \text{RSS} \leq \text{RSS}_N,$$

where

- $\text{RSS}_S = 0$ is the RSS of the saturated model,
- RSS is the RSS of a model with any choice of explanatory terms, and
- RSS_N is the RSS of the null model.

The null model

Linear regression: LSD effect on maths analysis
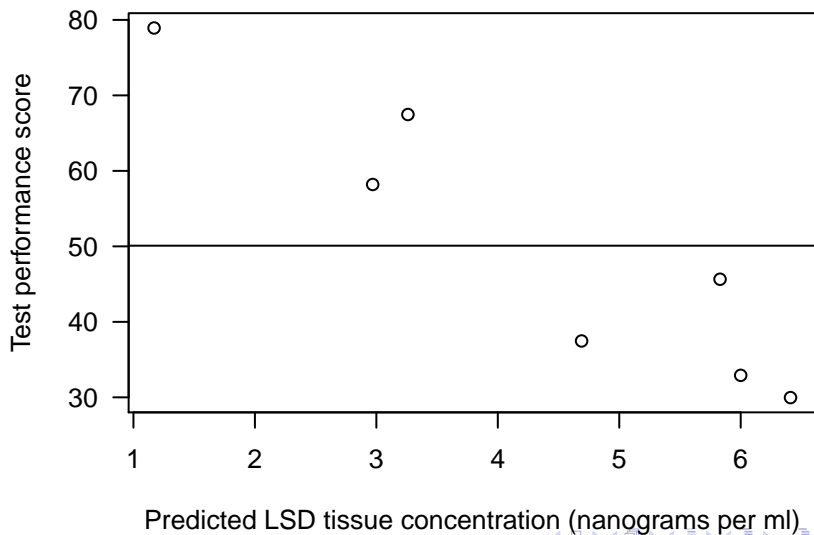
```
lm(score ~ 1)
```


Predicted LSD tissue concentration (nanograms per ml)

The null model

Linear regression: LSD effect on maths analysis


```
lm(score ~ 1)
```

Test performance score (y-axis)

RSS: 2078.18

Predicted LSD tissue concentration (nanograms per ml)

Increasing model complexity

Poisson regression: *Macrorhabdus ornithogaster* chicken analysis


```
glm(mo ~ weight, family = "poisson")
```

Log–likelihood: –320.67

M. ornithogaster organisms vs Weight (g)
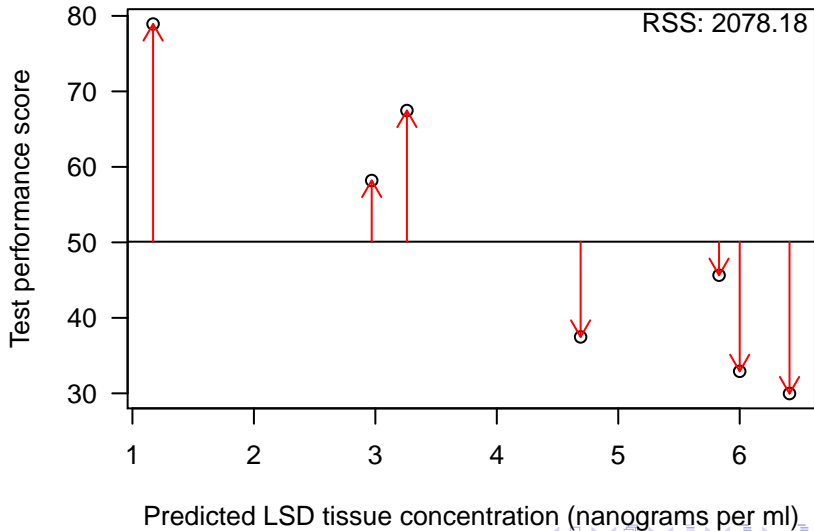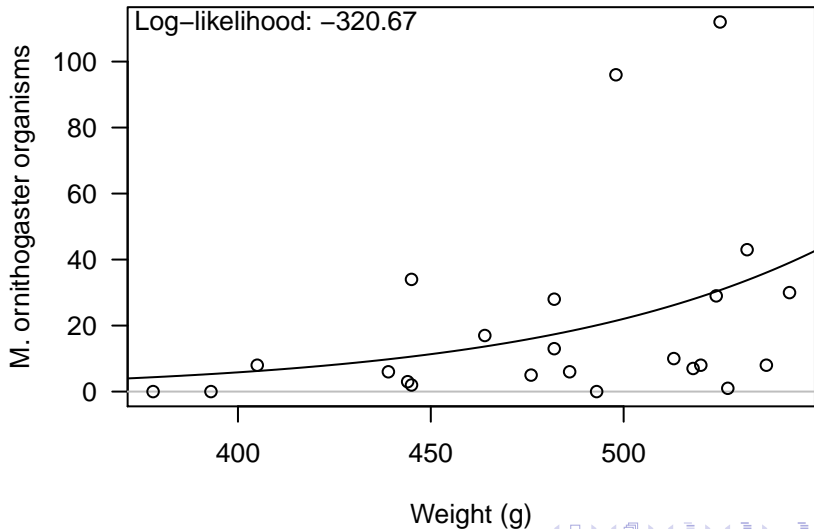
Increasing model complexity

Poisson regression: *Macrorhabdus ornithogaster* chicken analysis

```
glm(mo ~ poly(weight, 2, raw = TRUE), family = "poisson")
```
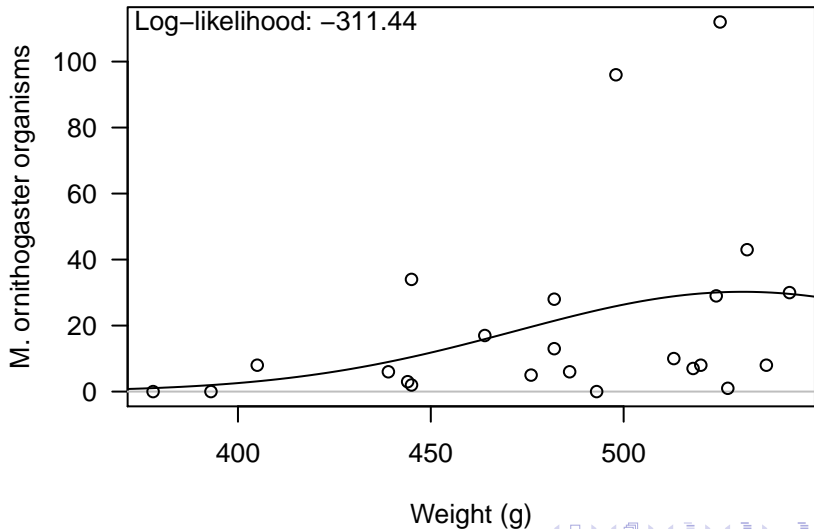
Increasing model complexity

Poisson regression: *Macrorhabdus ornithogaster* chicken analysis

```
glm(mo ~ poly(weight, 3, raw = TRUE), family = "poisson")
```
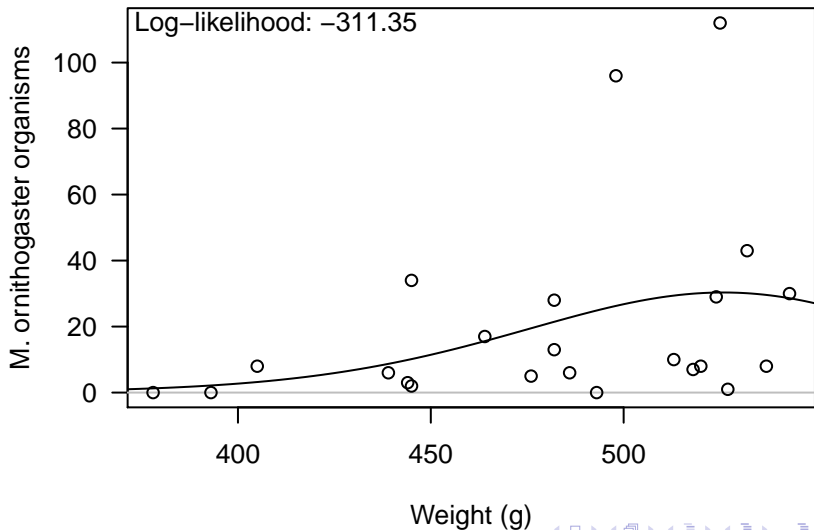

Log–likelihood: –311.35

Increasing model complexity

Poisson regression: *Macrorhabdus ornithogaster* chicken analysis

```
glm(mo ~ poly(weight, 4, raw = TRUE), family = "poisson")
```
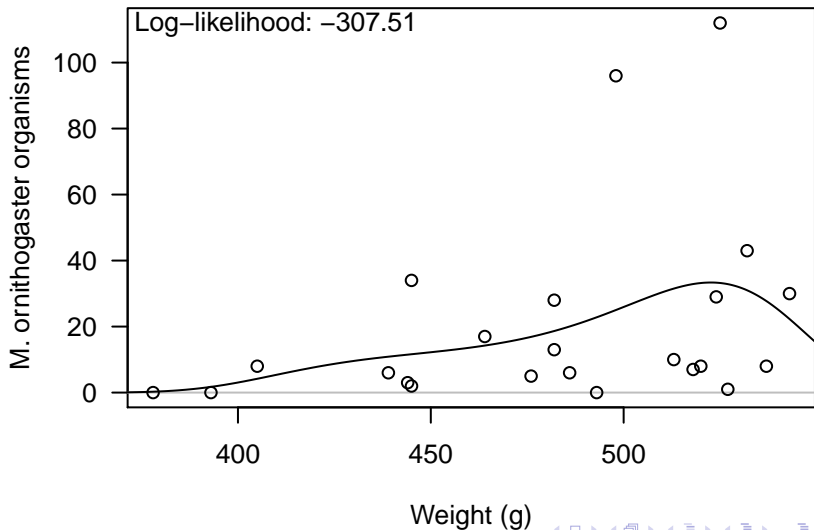
Increasing model complexity

Poisson regression: *Macrorhabdus ornithogaster* chicken analysis

```
glm(mo ~ poly(weight, 5, raw = TRUE), family = "poisson")
```
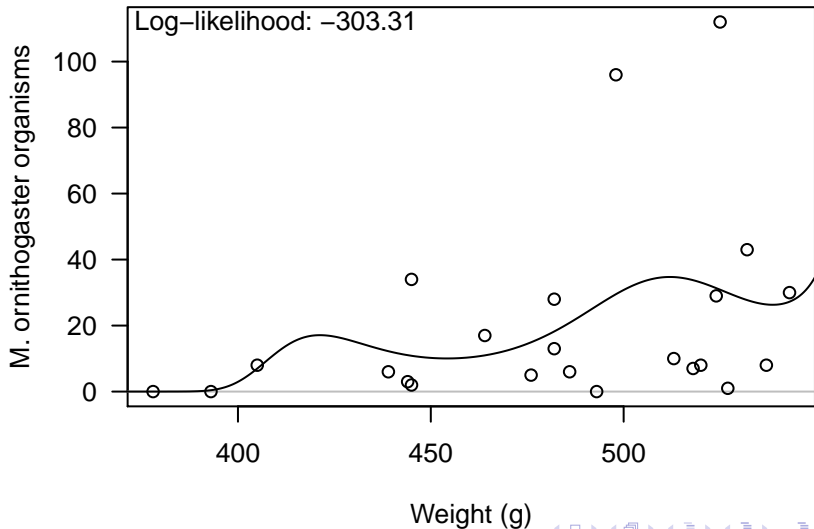
The saturated model

Poisson regression: *Macrorhabdus ornithogaster* chicken analysis

SATURATED!

Log–likelihood: –43.42
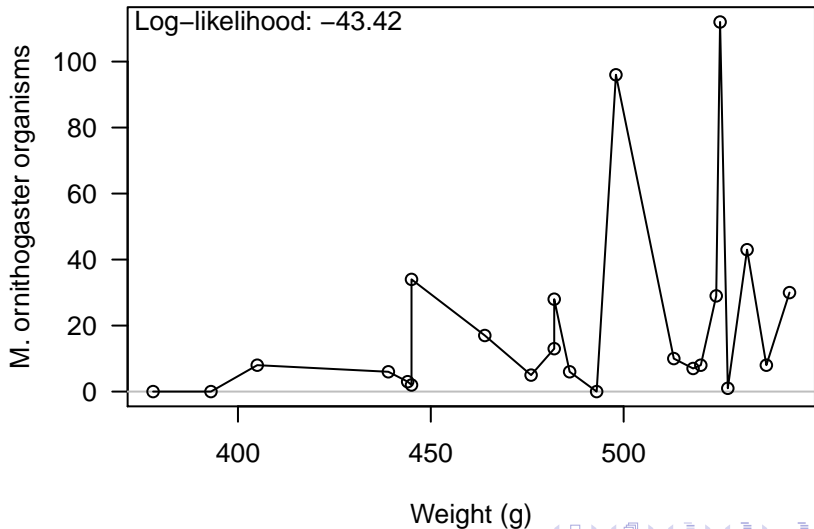
The saturated model
GLMs

From the plots:

- ▶ As we increase the number of model parameters, the fitted line becomes more 'wiggly'.
- ▶ The more 'wiggly', the higher the log-likelihood.
- ▶ Eventually it is 'wiggly' enough to go through every point.

A model with as many parameters as there are observations is called a saturated model:

- ▶ This is the most complex model you can fit.
- ▶ You cannot add any further explanatory terms.
- ▶ The regression line will go through every point.
- ▶ Its maximised log-likelihood is at least as large as any unsaturated model.

Deviance

Deviance is twice the difference between the maximised log-likelihood of the saturated model and the log-likelihood of the fitted model:

$$D = 2(\ell_S - \ell),$$

where ℓ_S is the log-likelihood of the saturated model.

So, from the plots on the previous slides, for our model

```
chickens.fit <- glm(mo ~ weight, family = "poisson")
```

we have

$$D = 2(-43.421 - -320.667)$$
$$= 554.492.$$

```
deviance(chickens.fit)
```
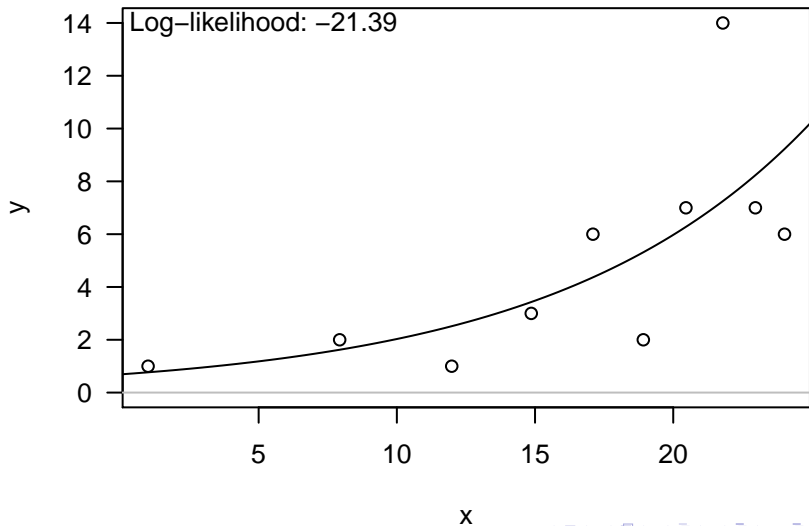
```
## [1] 554.492
```

Deviance

What does it mean?
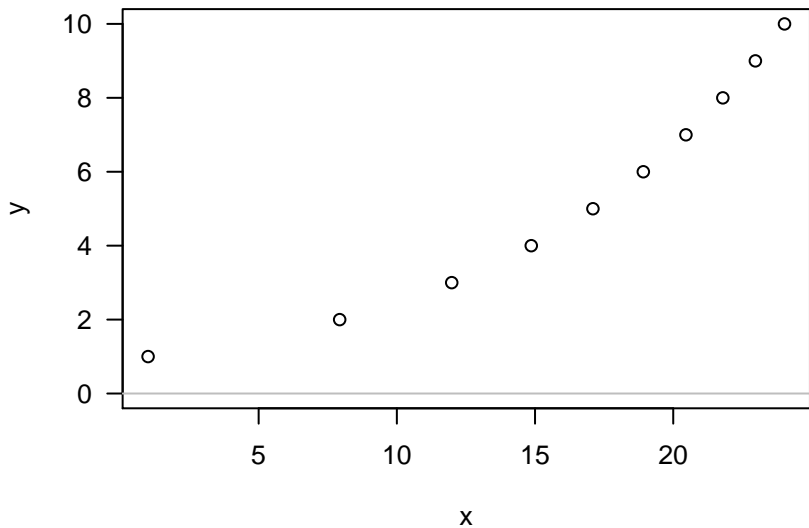
Deviance

What does it mean?

```
example.fit <- glm(y ~ x, family = "poisson")
```

Deviance
What does it mean?

```
## Saturated
```
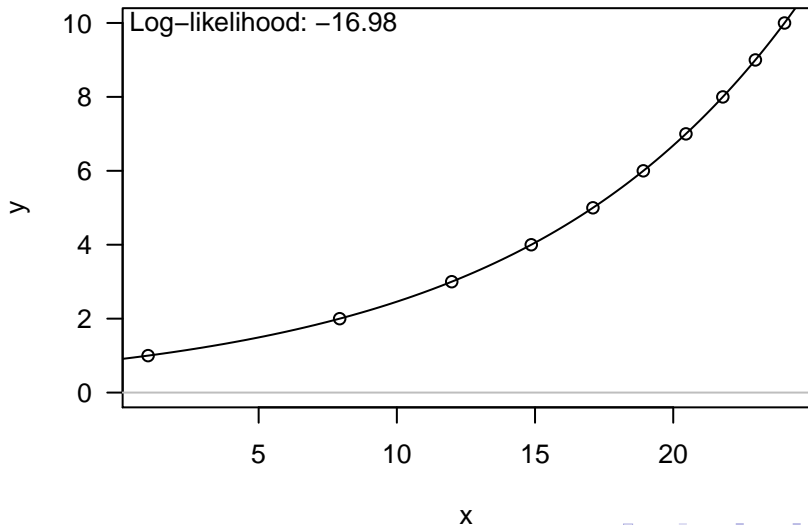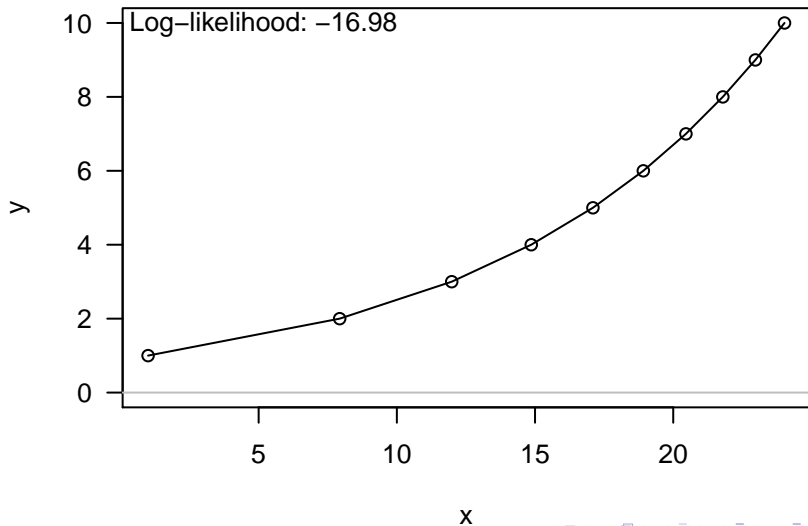
Deviance

When the observations all lie exactly on the line, the fitted values of our model all coincide with the fitted values of the saturated model.

This means that the maximised log-likelihood of the fitted model is the same as that of the saturated model. So the deviance of our model is equal to zero:

$$D = 2(-16.983 - -16.983)$$
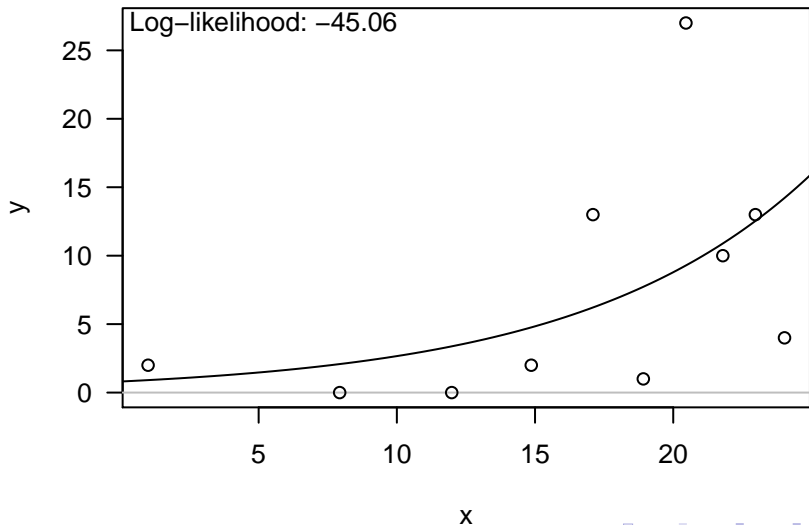$$= 0.000.$$
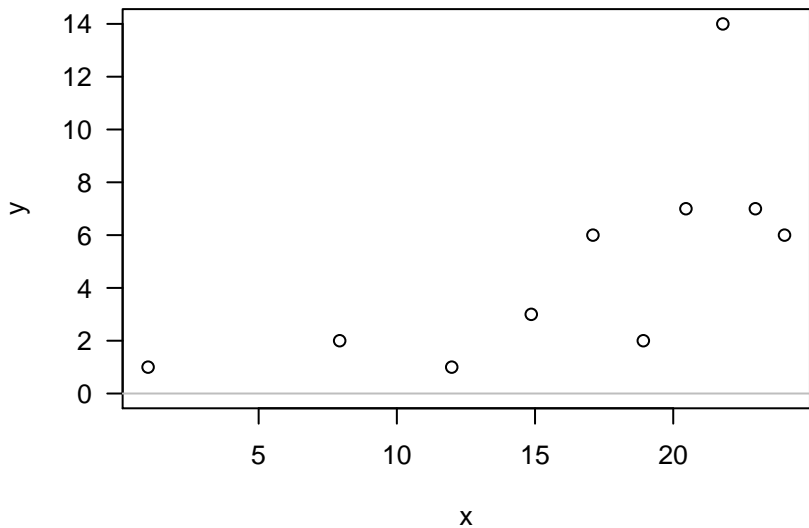
Deviance
What does it mean?

Deviance

What does it mean?

```
example.fit <- glm(y ~ x, family = "poisson")
```

Deviance

What does it mean?

```
## Saturated
```
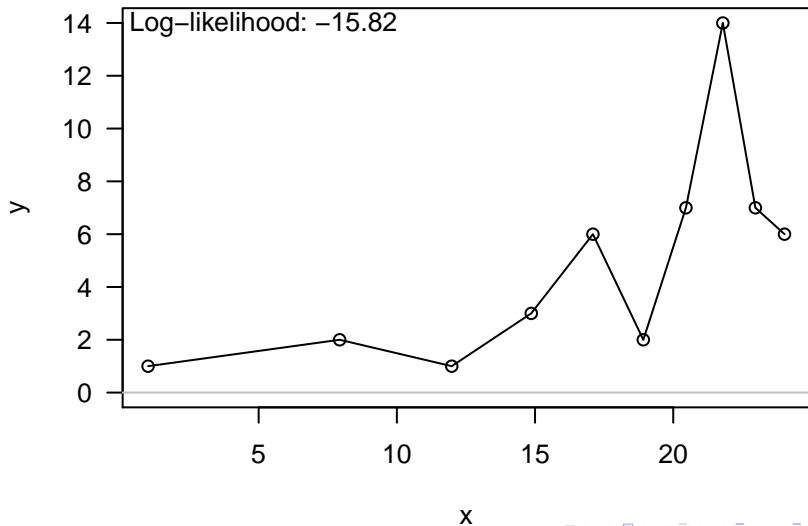
Deviance

When the observations do not all fit exactly on the line, the fitted values of our model are different to the fitted values of the saturated model.

This means that the maximised log-likelihood of the fitted model is different to that of the saturated model, and so the deviance is positive:

$$D = 2(-15.819 - -21.390)$$
$$= 11.143.$$

```
deviance(example.fit)

## [1] 11.14279
```

Deviance
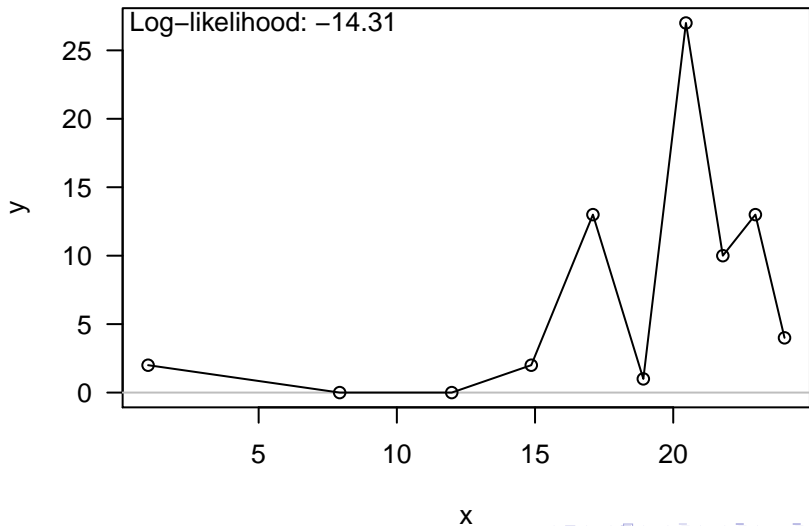
What does it mean?
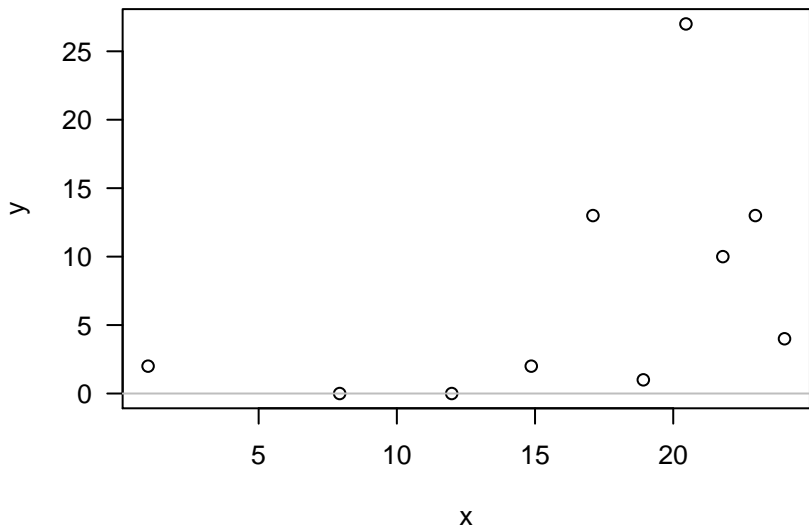
Deviance

What does it mean?

```
example.fit <- glm(y ~ x, family = "poisson")
```

Deviance
What does it mean?

`## Saturated`

Deviance
What does it mean?

The further the observations get from the line, the larger the discrepancy between the fitted values of our model and the fitted values of the saturated model.

The larger this discrepency, the larger the deviance:

$$D = 2(-14.311 - -45.064)$$
$$= 61.506.$$

```
deviance(example.fit)

## [1] 61.50621
```

Deviance

What does it mean?
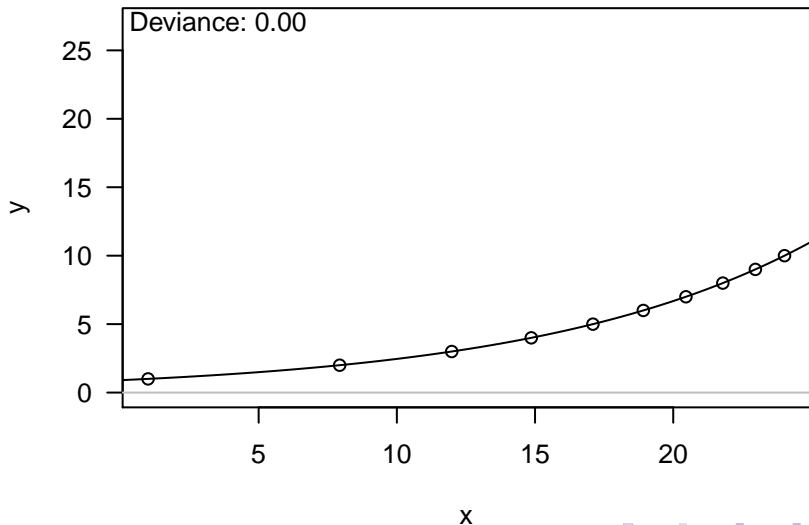
```
example.fit <- glm(y ~ x, family = "poisson")
```
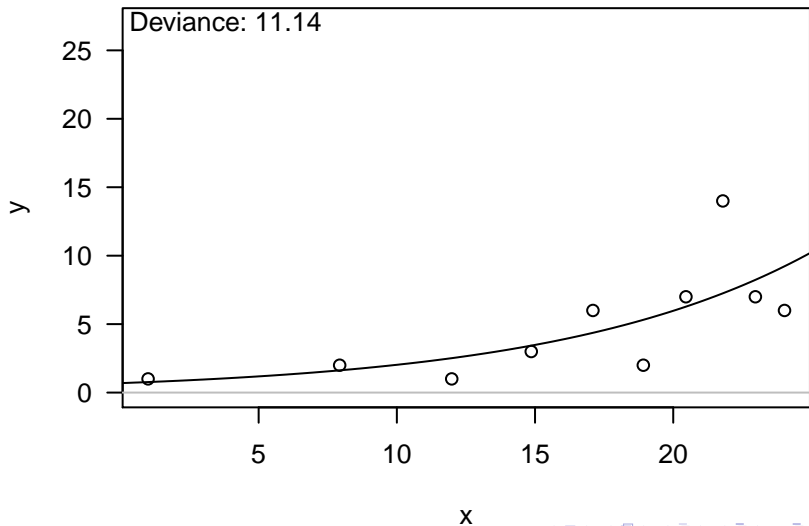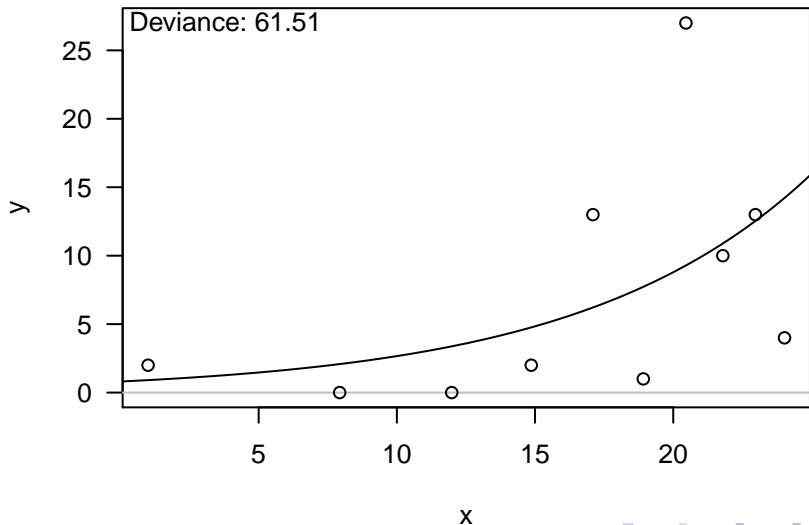
Deviance

What does it mean?

```
example.fit <- glm(y ~ x, family = "poisson")
```

Deviance
What does it mean?

```
example.fit <- glm(y ~ x, family = "poisson")
```


Deviance: 61.51

Deviance
The null model

Null deviance is the deviance of the null model:

$$D_N = 2(\ell_S - \ell_N),$$

where ℓ_N is the maximised log-likelihood of the null model.

The null model's maximised log-likelihood will be at least as small as that of any model with explanatory variables.

So the null deviance will be at least as large as that of any model with explanatory variables.

Deviance

The null model: *Macrorhabdus ornithogaster* chicken analysis
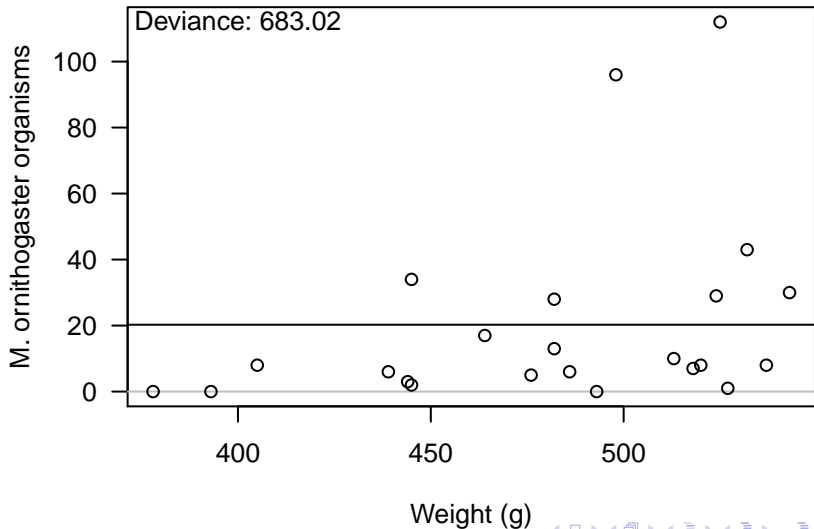
```
glm(mo ~ weight, family = "poisson")
```
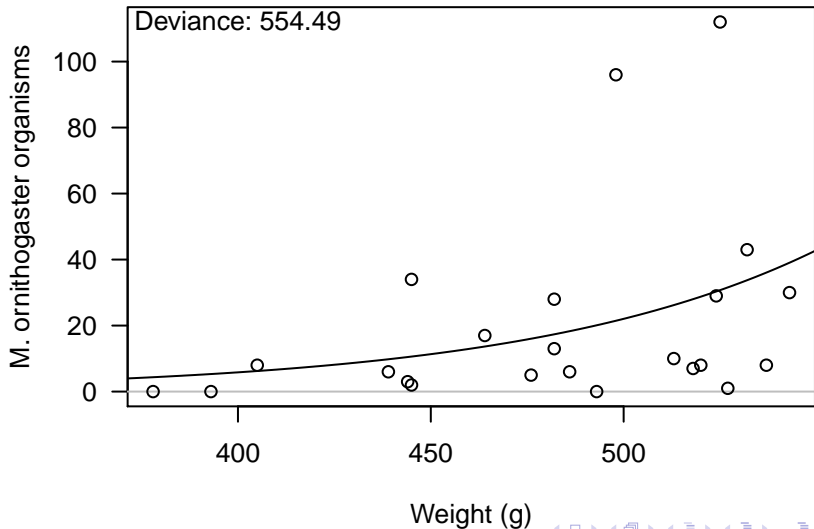
Deviance

The null model: *Macrorhabdus ornithogaster* chicken analysis

```
glm(mo ~ 1, family = "poisson")
```

Deviance
Some results

By definition, the deviance of the saturated model is zero, because it compares the log-likelihood of the saturated model with itself:

$$D_S = 2(\ell_S - \ell_S)$$
$$= 0$$

And so

$$D_S \leq D \leq D_N$$

The model parameters are estimated by maximising the log-likelihood.

▶ The larger the log-likelihood, the smaller the deviance.
▶ So, equivalently, the model parameters are estimated by minimising the deviance.

Deviance

Also, recall that increasing a model's complexity will never decrease its maximised log-likelihood. It follows that increasing a model's complexity will never increase its deviance.

Let Model A, with log-likelihood ℓ_A, be a more complex version of Model B, with log likelihood ℓ_B.

$$\ell_A \geq \ell_B,$$

because increasing complexity increases the log-likelihood

$$D_A = 2(\ell_S - \ell_A)$$
$$D_B = 2(\ell_S - \ell_B)$$

and so

$$D_A \leq D_B$$

Deviance

For a generalised linear model:

- ▶ Parameter estimation is achieved by minimising the deviance.

- ▶ The more variance, the larger the deviance.

- ▶ The saturated model has deviance equal to zero.

- ▶ Increasing a model's complexity will decrease the deviance.

- ▶ $D_S \leq D \leq D_N$

Sound familiar?

For a linear model:

- ▶ Parameter estimation is achieved by minimising the RSS.

- ▶ The more variance, the larger the RSS.

- ▶ The saturated model has RSS equal to zero.

- ▶ Increasing a model's complexity will decrease the RSS.

- ▶ $RSS_S \leq D \leq RSS_N$

Deviance

A summary

Everything we have seen so far about deviance for Poisson regression also holds for logistic regression:

- $D = 2(\ell_S - \ell)$
- Parameter estimation is achieved by minimising the deviance.
- The more variance, the larger the deviance.
- The saturated model has deviance equal to zero.
- Increasing a model's complexity will decrease the deviance.
- $D_S \leq D \leq D_N$

So, in general, deviance is analogous to residual sum of squares. We will see this idea resurface over the next few lectures.
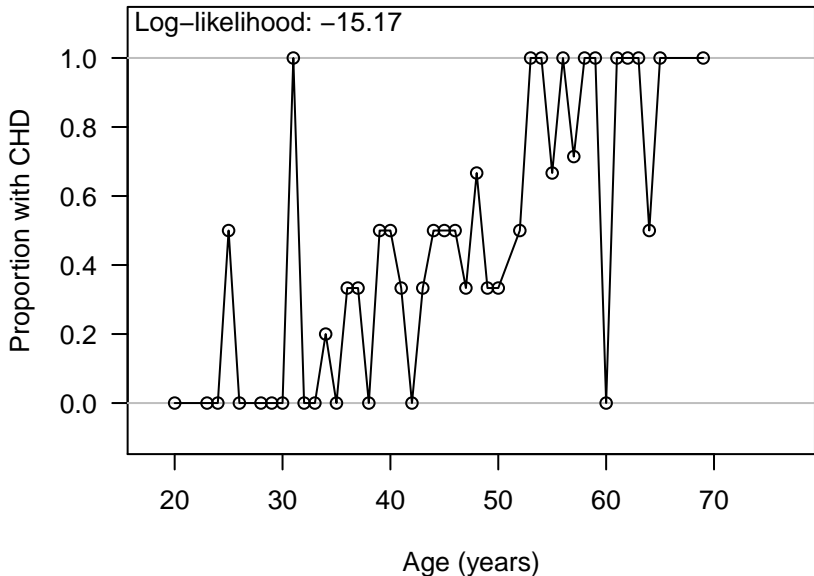
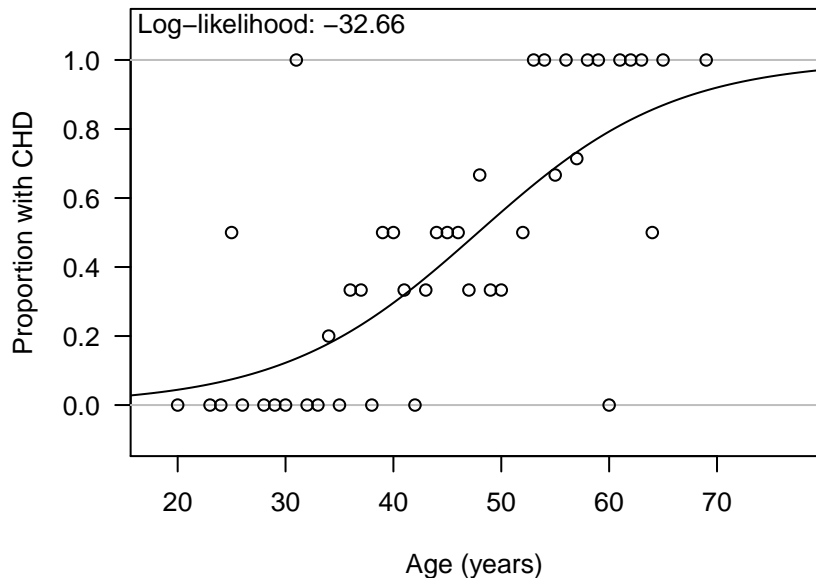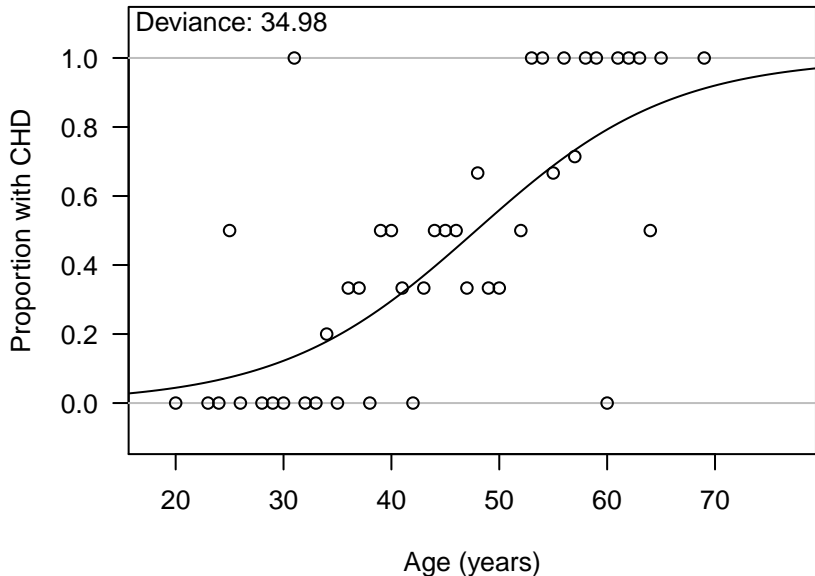# Deviance

Logistic regression: Coronary heart disease analysis
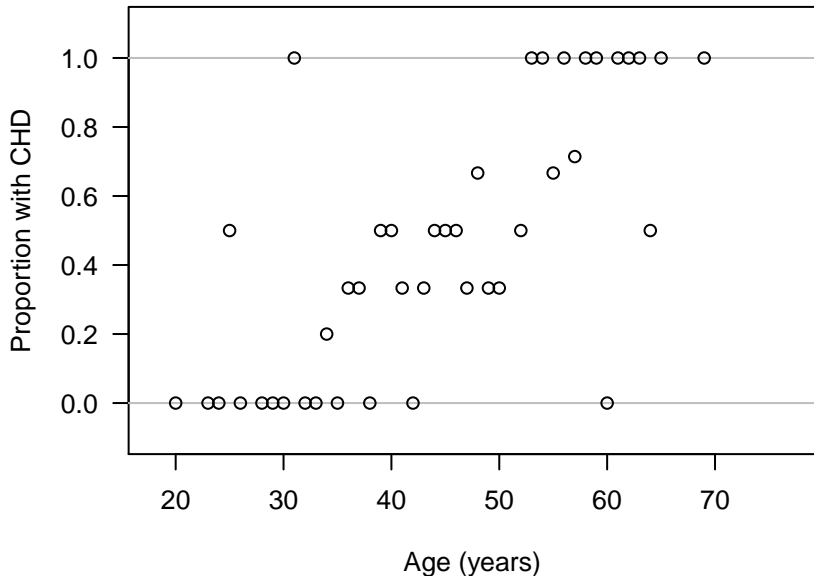
Deviance

Logistic regression: Coronary heart disease analysis

Deviance

Logistic regression: Coronary heart disease analysis

Log-likelihood: −32.66

Deviance

Logistic regression: Coronary heart disease analysis

Deviance

Deviance is twice the difference between the maximised log-likelihood of the saturated model and the log-likelihood of the fitted model:

$$D = 2(\ell_S - \ell),$$

where ℓ_S is the log-likelihood of the saturated model.

So, from the plots on the previous slides, for our model

```r
chd.fit <- glm(cbind(y, n - y) ~ age, family = "binomial")
```
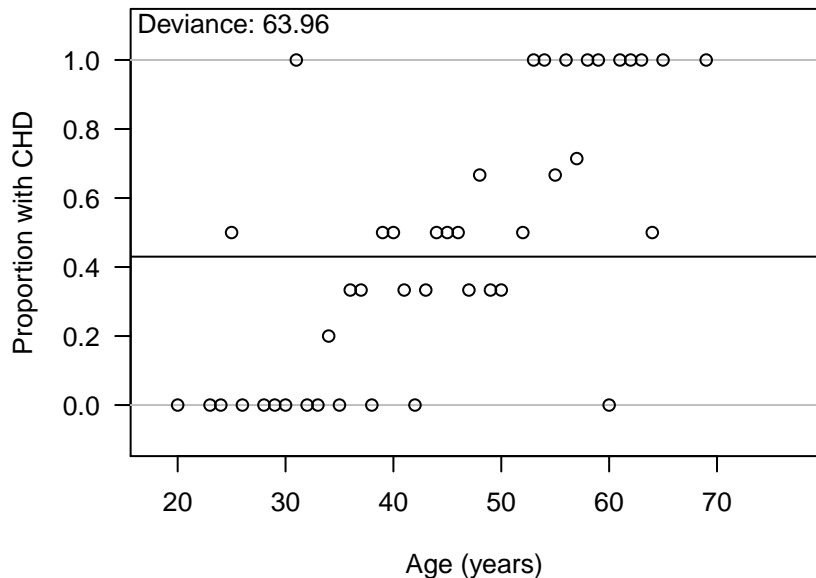
we have

$$D = 2(-15.17 - -32.66)$$
$$= 34.98.$$

```r
deviance(chd.fit)
```
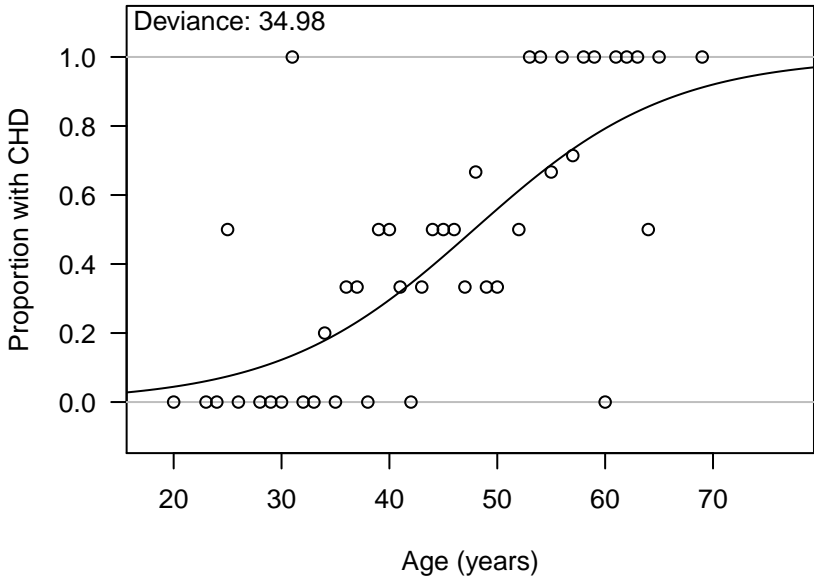
```
## [1] 34.97572
```

Deviance

The null model: Coronary heart disease analysis

Deviance

The fitted model: Coronary heart disease analysis

Deviance

The fitted model: Coronary heart disease analysis