STATS 330

Handout 5
Deviance as a goodness-of-fit statistic

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 a 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.

This handout introduces how we can use the deviance to assess a model's goodness-of-fit.

Goodness-of-fit

Poisson regression: *Macrorhabdus ornithogaster* chicken analysis

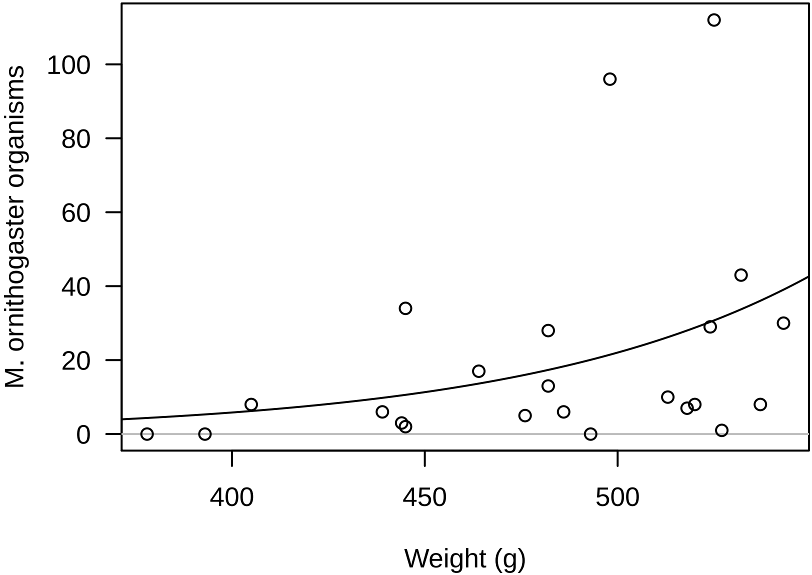
Let's summarise our *Macrorhabdus ornithogaster* chicken analysis. In Handout 1 we proposed fitting a Poisson regression model with a log link function to

1. Ensure the expected value was greater than 0 for all observations,
2. Account for nonconstant variance, and
3. Assume a discrete distribution for a discrete response.

However, we never addressed the appropriateness of the Poisson distribution assumption.

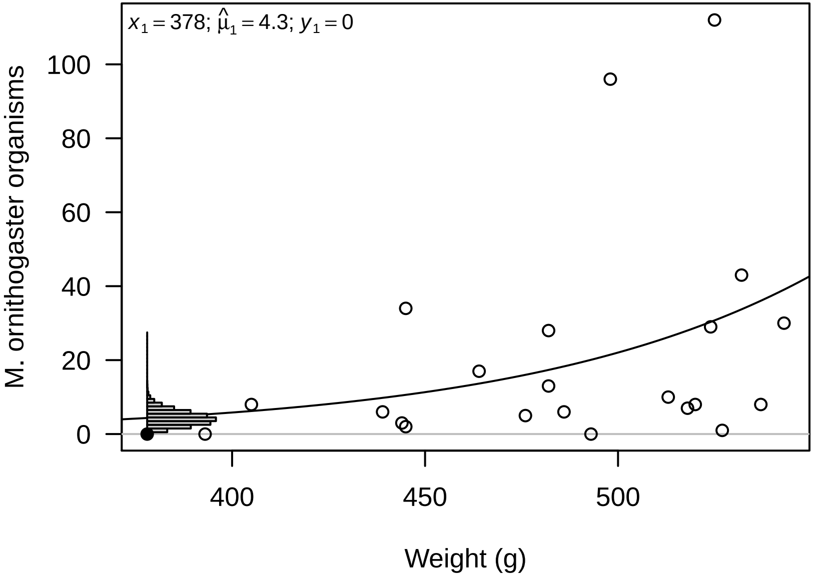
Goodness-of-fit

Poisson regression: *Macrorhabdus ornithogaster* chicken analysis



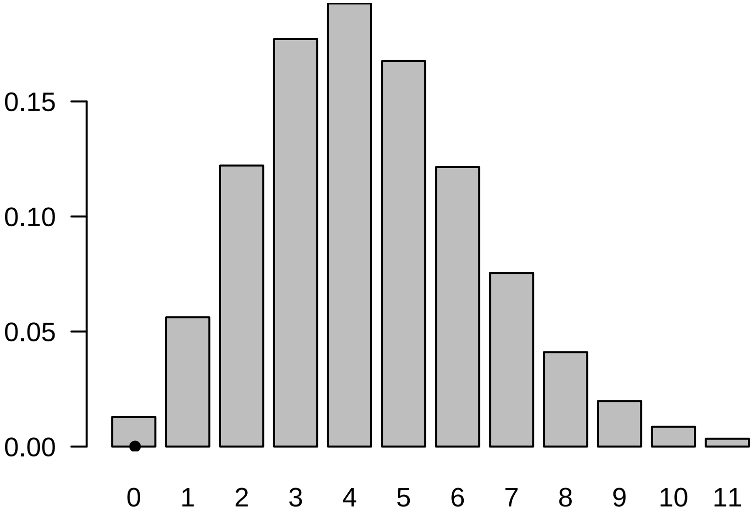
Goodness-of-fit

Poisson regression: *Macrorhabdus ornithogaster* chicken analysis



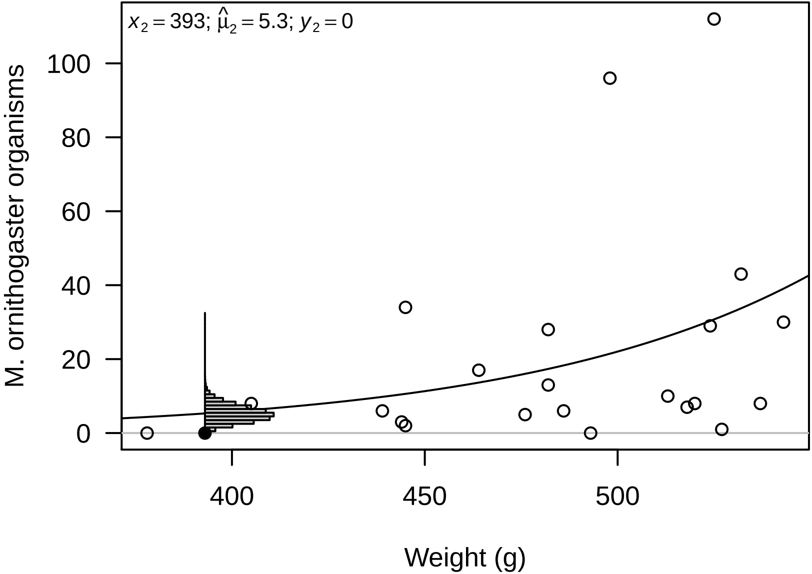
Goodness-of-fit

Poisson regression: *Macrorhabdus ornithogaster* chicken analysis



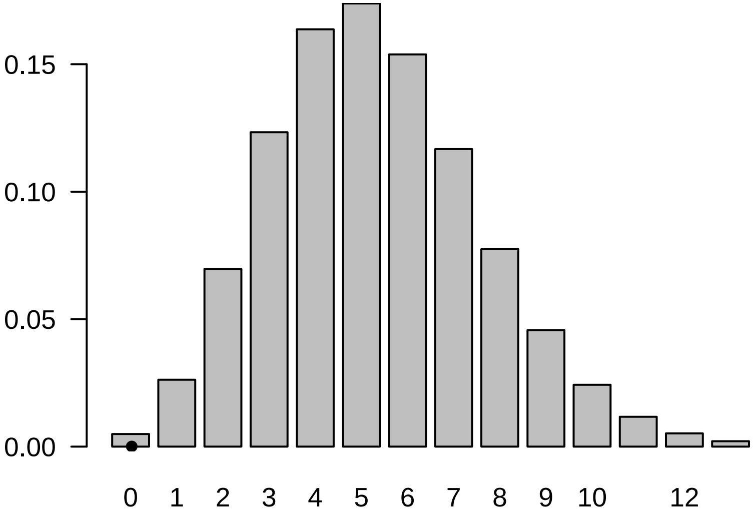
Goodness-of-fit

Poisson regression: *Macrorhabdus ornithogaster* chicken analysis



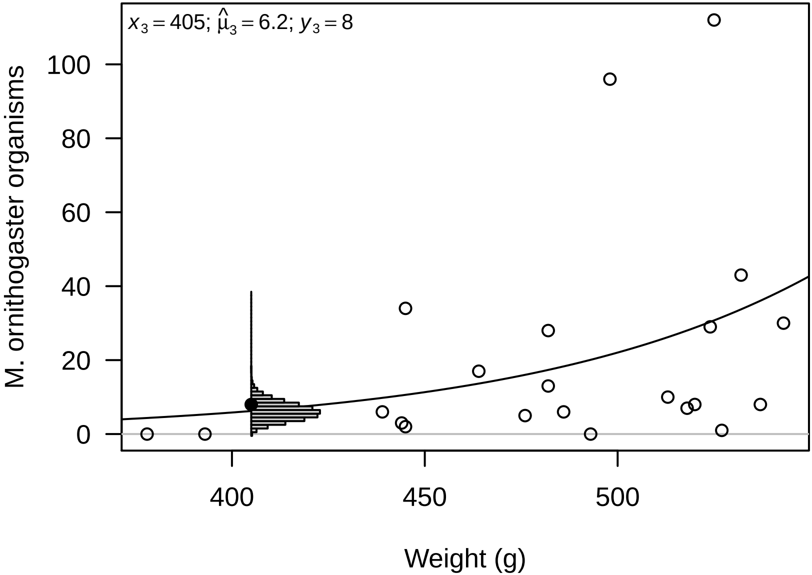
Goodness-of-fit

Poisson regression: *Macrorhabdus ornithogaster* chicken analysis



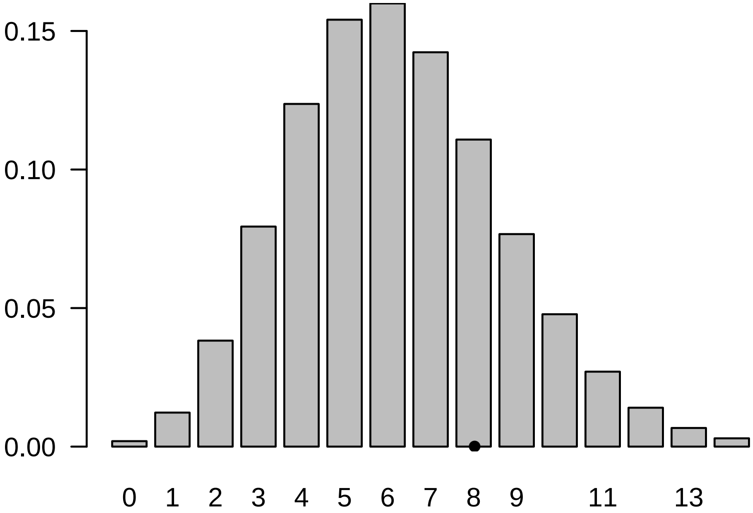
Goodness-of-fit

Poisson regression: *Macrorhabdus ornithogaster* chicken analysis



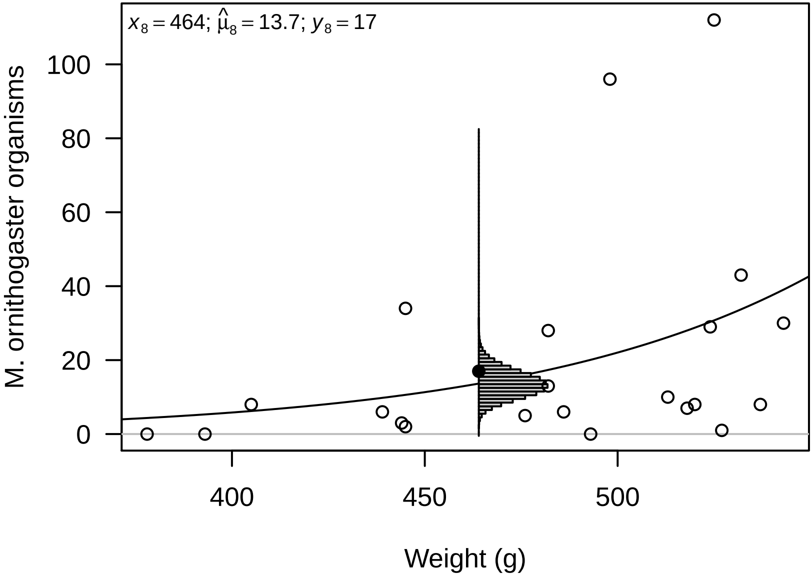
Goodness-of-fit

Poisson regression: *Macrorhabdus ornithogaster* chicken analysis



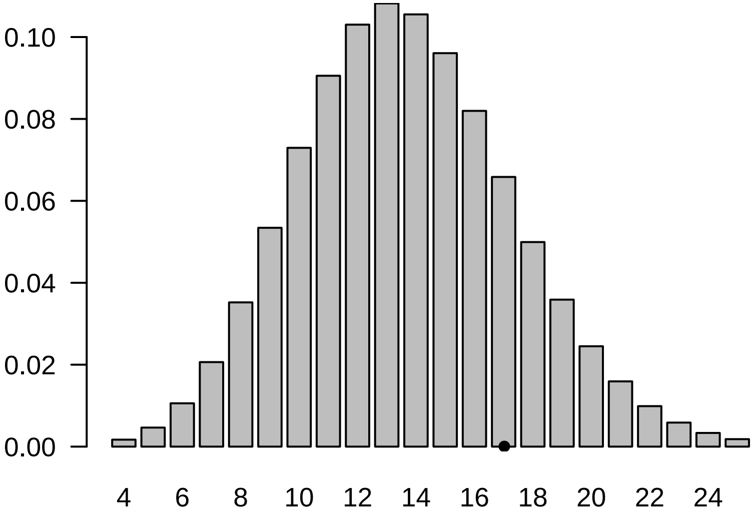
Goodness-of-fit

Poisson regression: *Macrorhabdus ornithogaster* chicken analysis



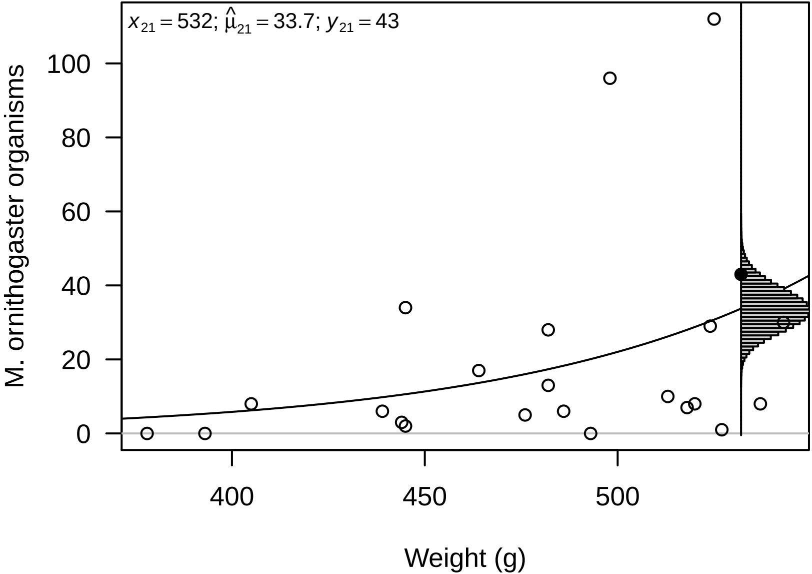
Goodness-of-fit

Poisson regression: *Macrorhabdus ornithogaster* chicken analysis



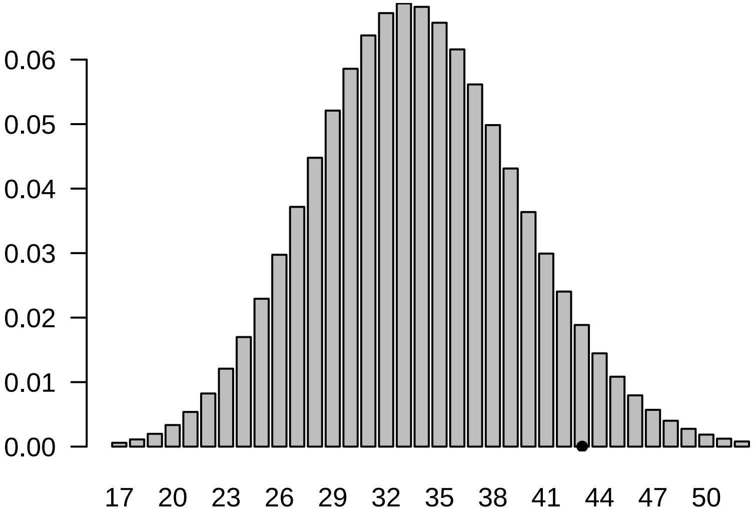
Goodness-of-fit

Poisson regression: *Macrorhabdus ornithogaster* chicken analysis



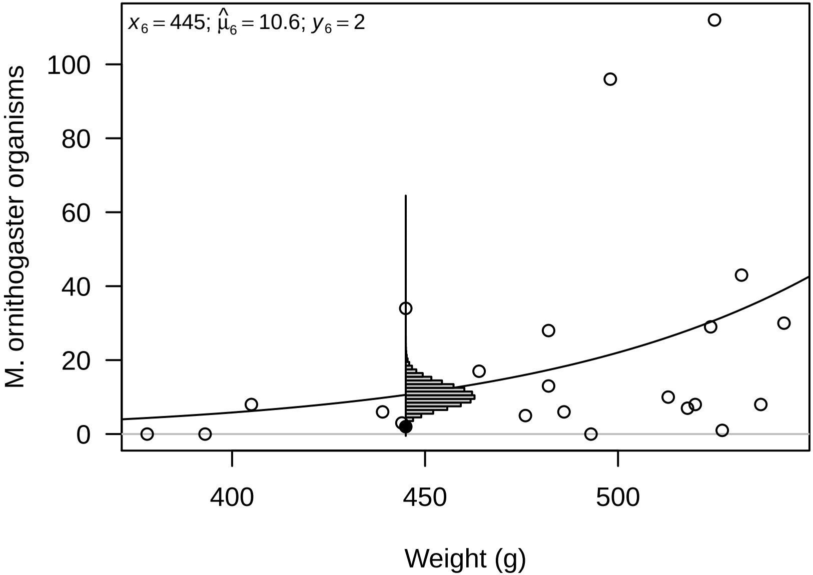
Goodness-of-fit

Poisson regression: *Macrorhabdus ornithogaster* chicken analysis



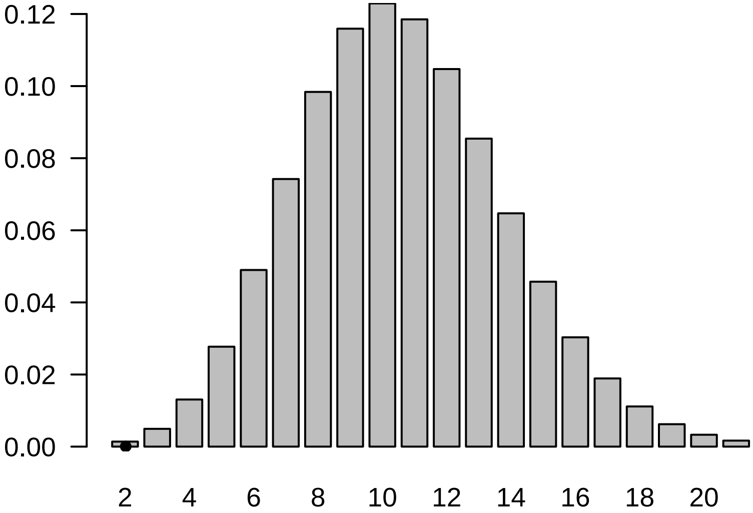
Goodness-of-fit

Poisson regression: *Macrorhabdus ornithogaster* chicken analysis



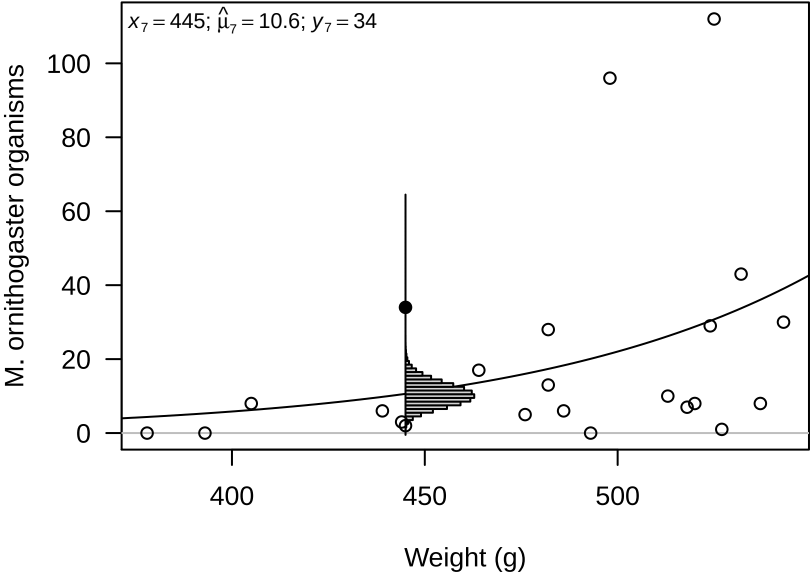
Goodness-of-fit

Poisson regression: *Macrorhabdus ornithogaster* chicken analysis



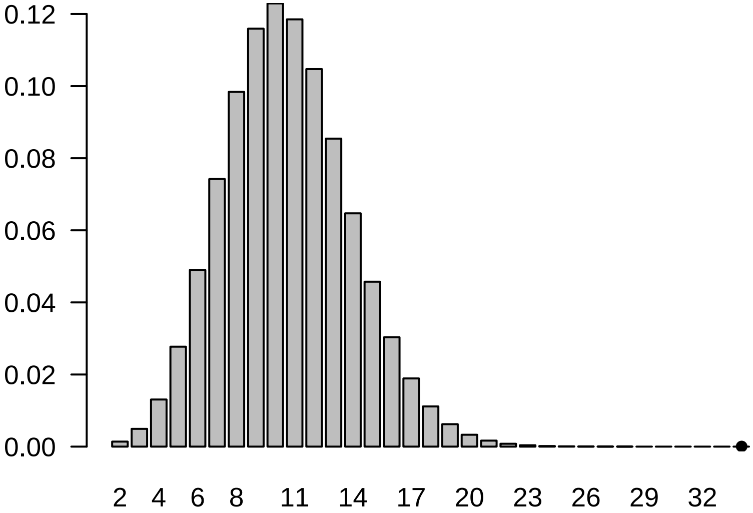
Goodness-of-fit

Poisson regression: *Macrorhabdus ornithogaster* chicken analysis



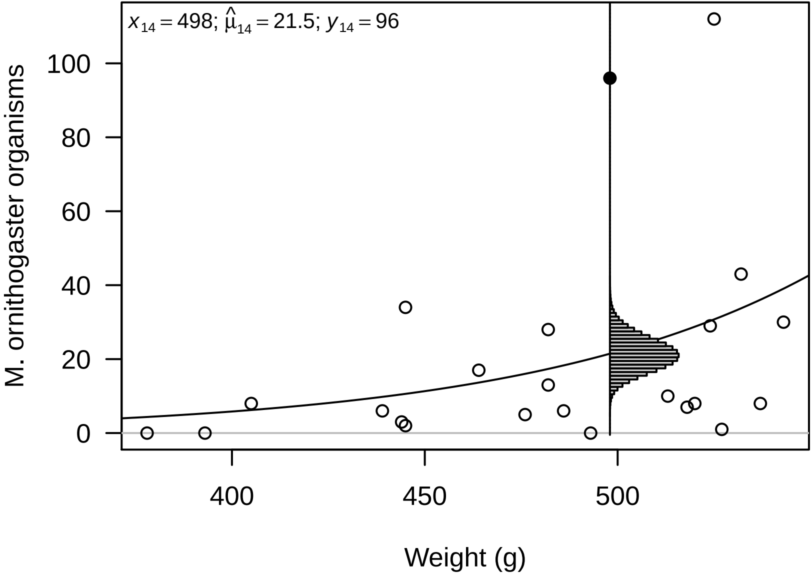
Goodness-of-fit

Poisson regression: *Macrorhabdus ornithogaster* chicken analysis



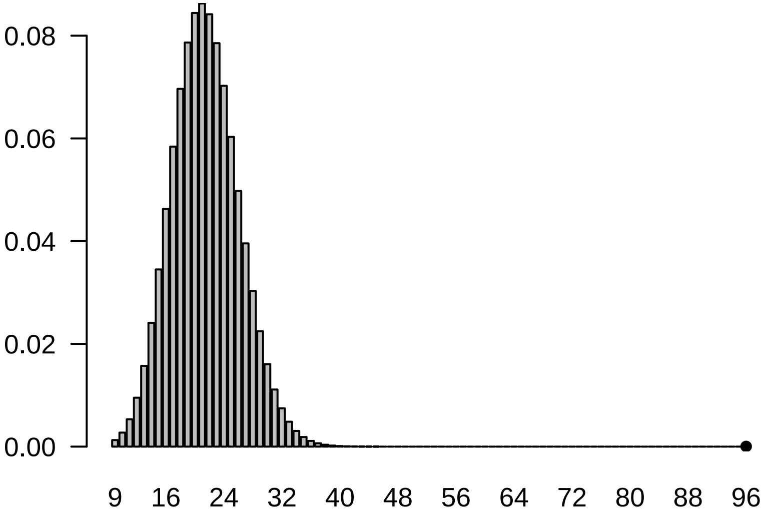
Goodness-of-fit

Poisson regression: *Macrorhabdus ornithogaster* chicken analysis



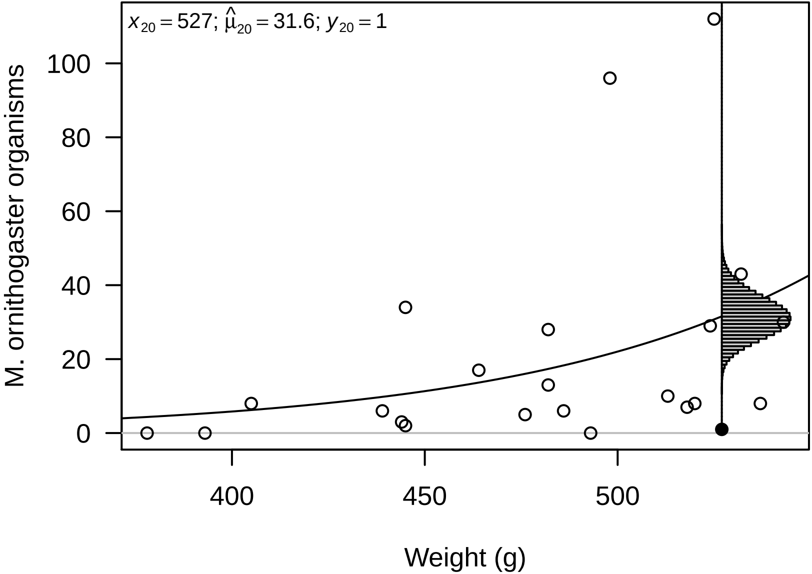
Goodness-of-fit

Poisson regression: *Macrorhabdus ornithogaster* chicken analysis



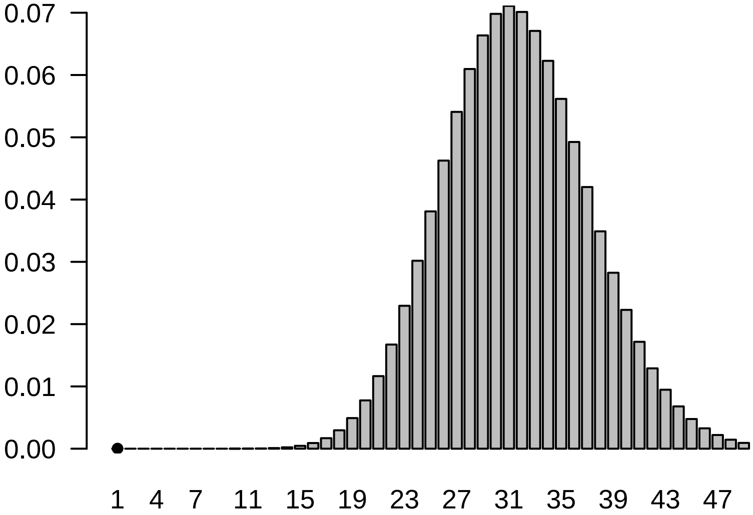
Goodness-of-fit

Poisson regression: *Macrorhabdus ornithogaster* chicken analysis



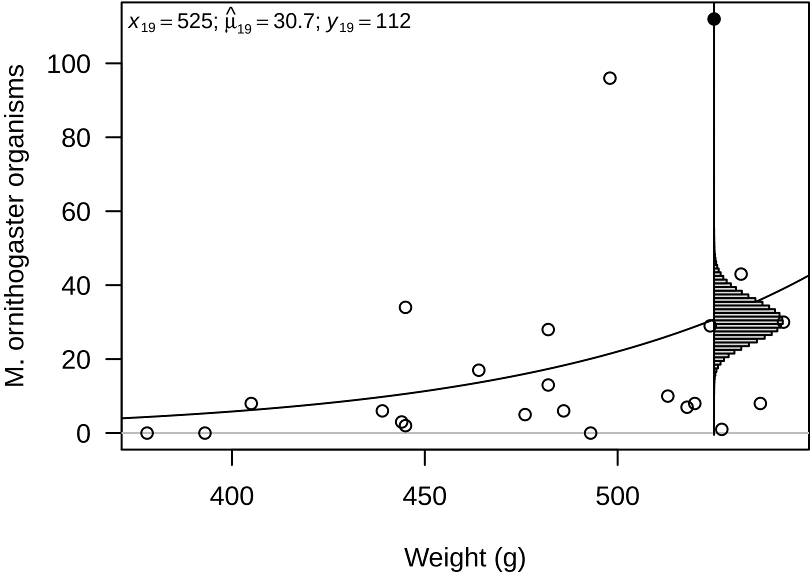
Goodness-of-fit

Poisson regression: *Macrorhabdus ornithogaster* chicken analysis



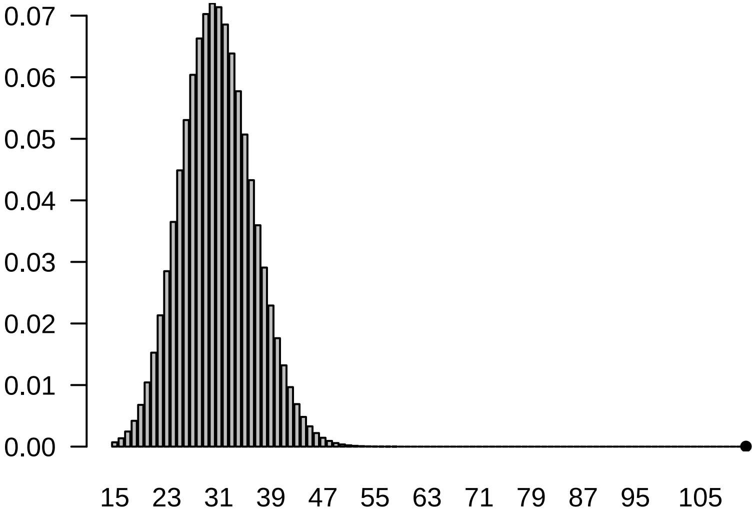
Goodness-of-fit

Poisson regression: *Macrorhabdus ornithogaster* chicken analysis



Goodness-of-fit

Poisson regression: *Macrorhabdus ornithogaster* chicken analysis



Goodness-of-fit

Poisson regression: *Macrorhabdus ornithogaster* chicken analysis

So in Handout 1 we cherry-picked observations:

- Some observations are consistent with the fitted model...
- ... But there are many observations with responses that are virtually impossible under the model's assumptions!

It seems as though the response variable has much more variance than we assume under our Poisson regression model. Our model does not fit the data well.

How can we determine this objectively, without making subjective decisions from the plots?

- By using the deviance!

As a goodness-of-fit statistic

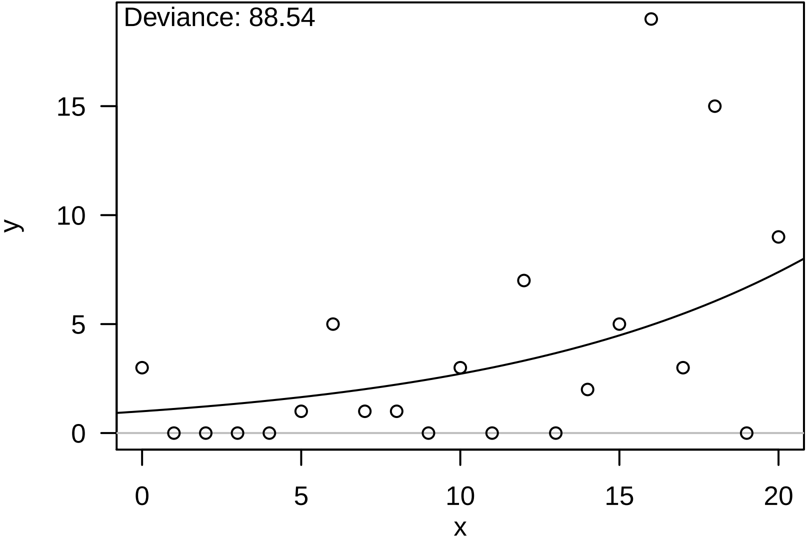
A model that fits poorly will have a large deviance. This may happen because

1. Our model is too simple and does not have the right explanatory terms, or
2. The response variable has more variance than assumed under a Poisson distribution.

Consider the following few plots, each depicting a fitted Poisson regression model.

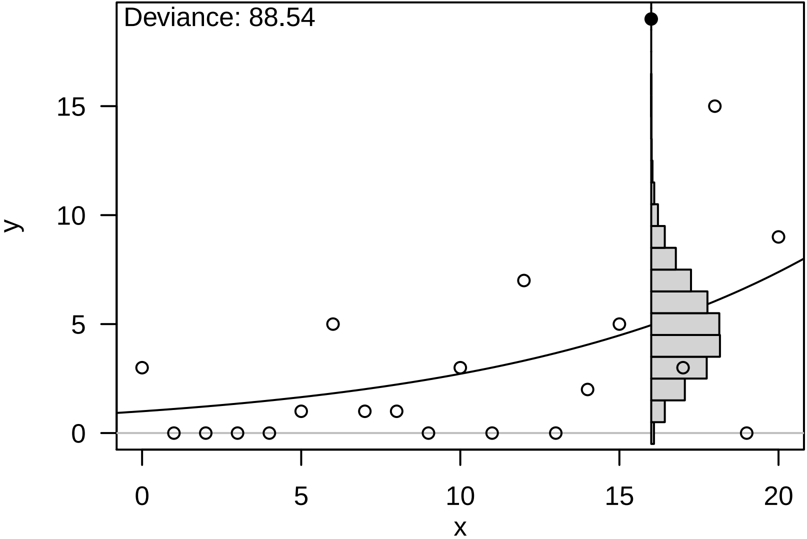
Deviance as a goodness-of-fit statistic

Variance too high: Large deviance



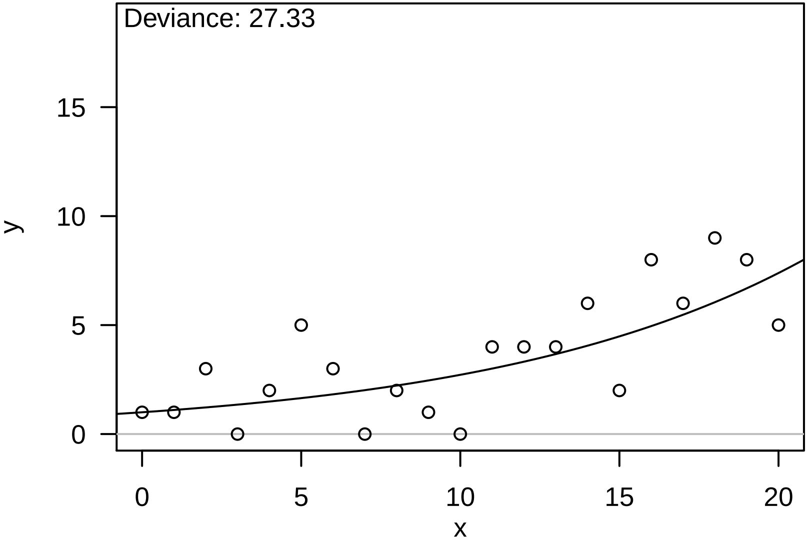
Deviance as a goodness-of-fit statistic

Variance too high: Large deviance



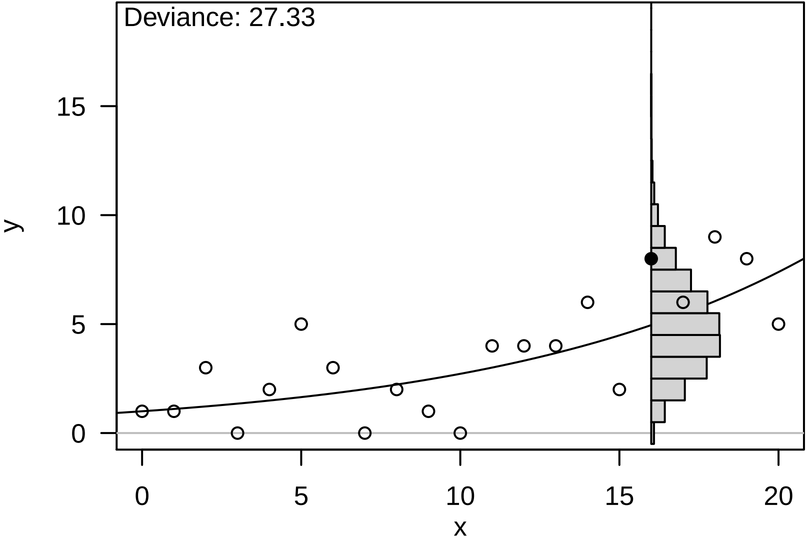
Deviance as a goodness-of-fit statistic

Variance about right: Small deviance



Deviance as a goodness-of-fit statistic

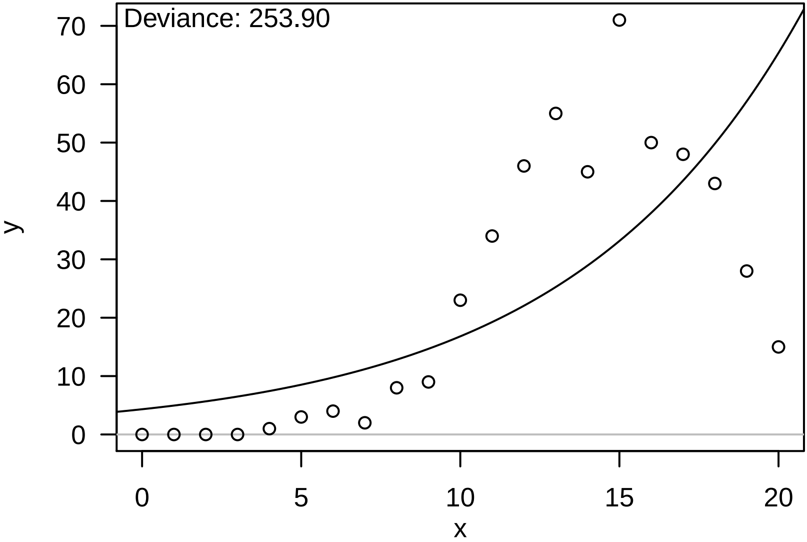
Variance about right: Small deviance



Deviance as a goodness-of-fit statistic

Insufficient explanatory terms: Large deviance

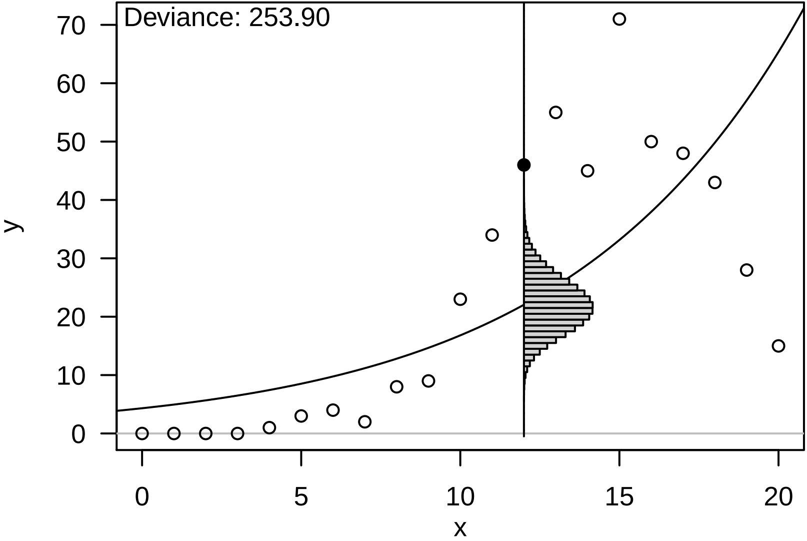
```
example.fit <- glm(y ~ x, family = "poisson")
```



Deviance as a goodness-of-fit statistic

Insufficient explanatory terms: Large deviance

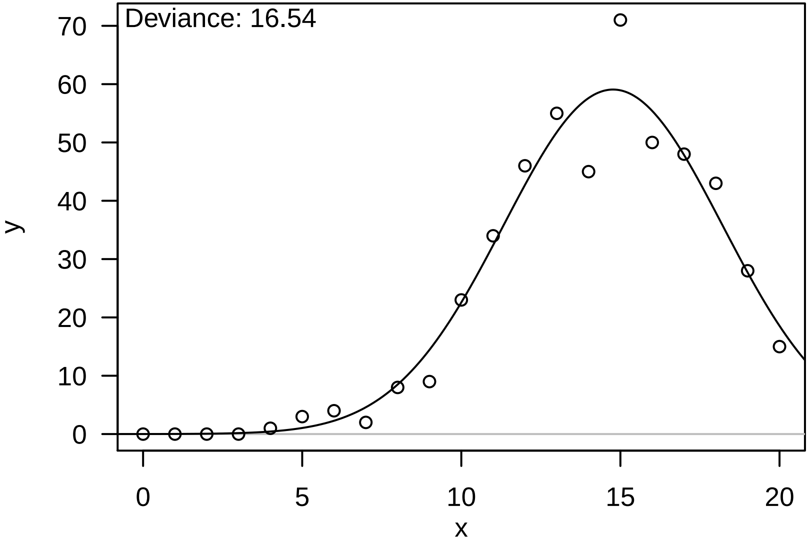
```
example.fit <- glm(y ~ x, family = "poisson")
```



Deviance as a goodness-of-fit statistic

Better selection of explanatory terms: Small deviance

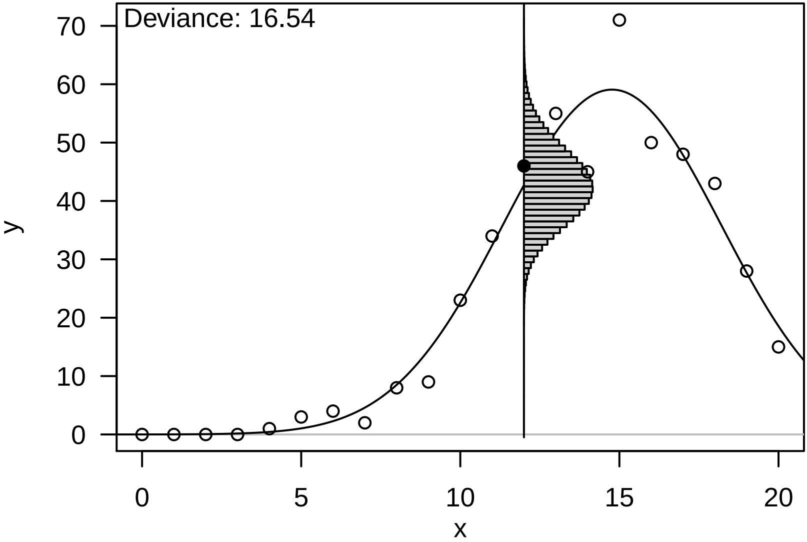
```
example2.fit <- glm(y ~ x + I(x^2), family = "poisson")
```



Deviance as a goodness-of-fit statistic

Better selection of explanatory terms: Small deviance

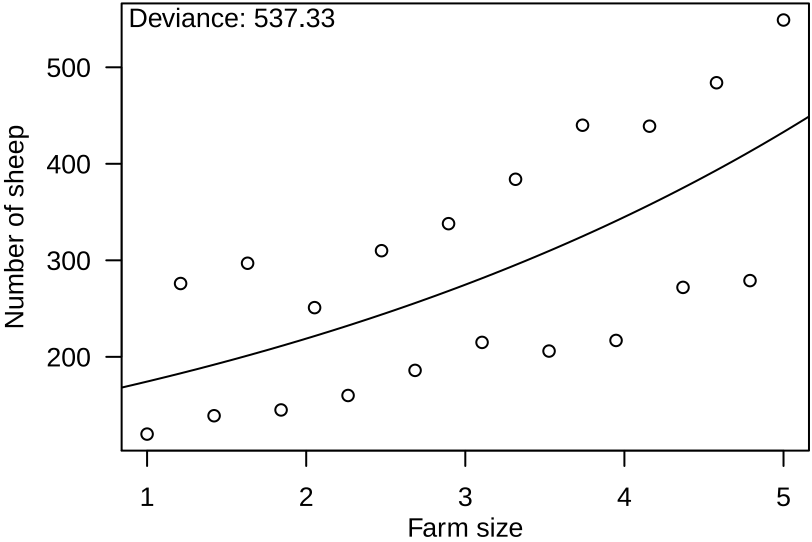
```r
example2.fit <- glm(y ~ x + I(x^2), family = "poisson")
```



Deviance as a goodness-of-fit statistic

Variance too high? Large deviance

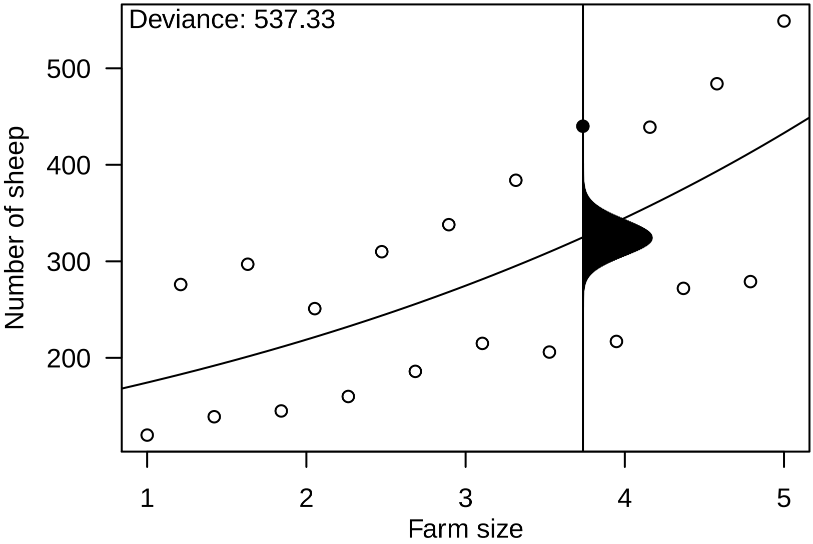
```
sheep.fit <- glm(sheep ~ farm.size, family = "poisson")
```



Deviance as a goodness-of-fit statistic

Variance too high? Large deviance

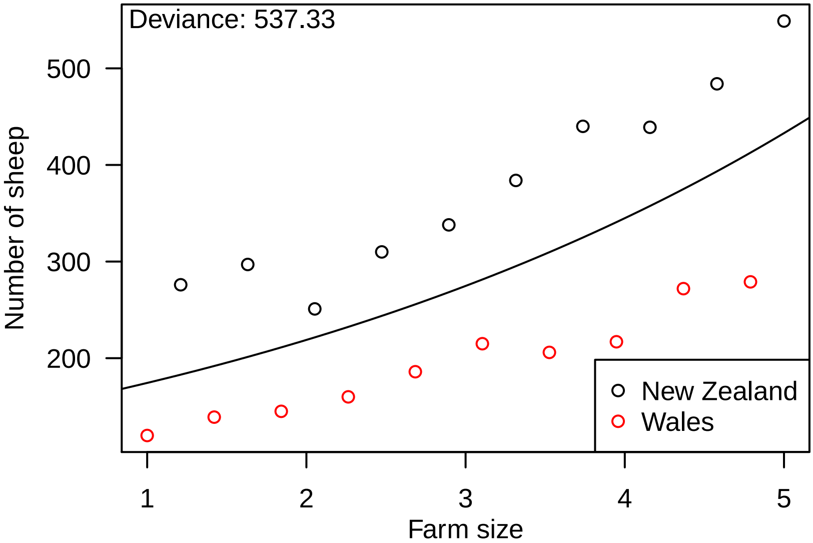
```
sheep.fit <- glm(sheep ~ farm.size, family = "poisson")
```



Deviance as a goodness-of-fit statistic

Insufficient explanatory terms! Large deviance

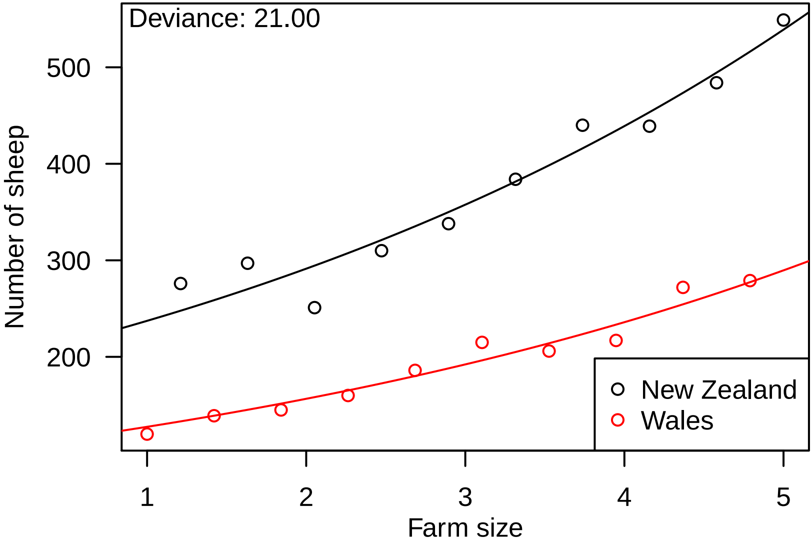
```
sheep.fit <- glm(sheep ~ farm.size, family = "poisson")
```



Deviance as a goodness-of-fit statistic

Better selection of explanatory terms: Small deviance

```r
sheep2.fit <- glm(sheep ~ farm.size + country, family = "poisson")
```



Deviance as a goodness-of-fit statistic

So a large deviance could indicate lack-of-fit. How large is too large?

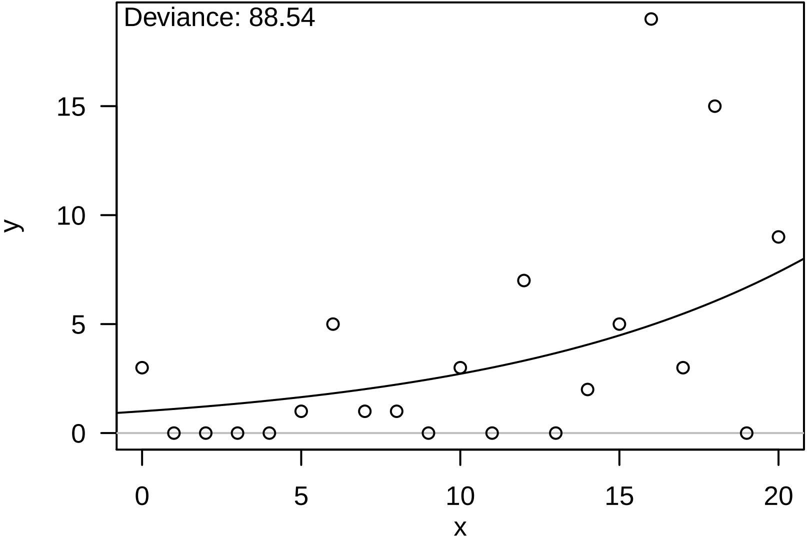
- If the model is correct, then, under certain conditions[1], the deviance comes from a chi-squared distribution with $n - k$ degrees of freedom, where n is the number of observations and k is the number of coefficients.
- If it is plausible that the deviance could have come from this distribution, we have no evidence against the hypothesis that our model is correct.
- If it is not plausible that the deviance could have come from this distribution, we do have evidence against the hypothesis that the model is correct.

Let's revisit the deviance for two of the previous examples.

[1] More on this later! For now we'll assume these conditions are met.

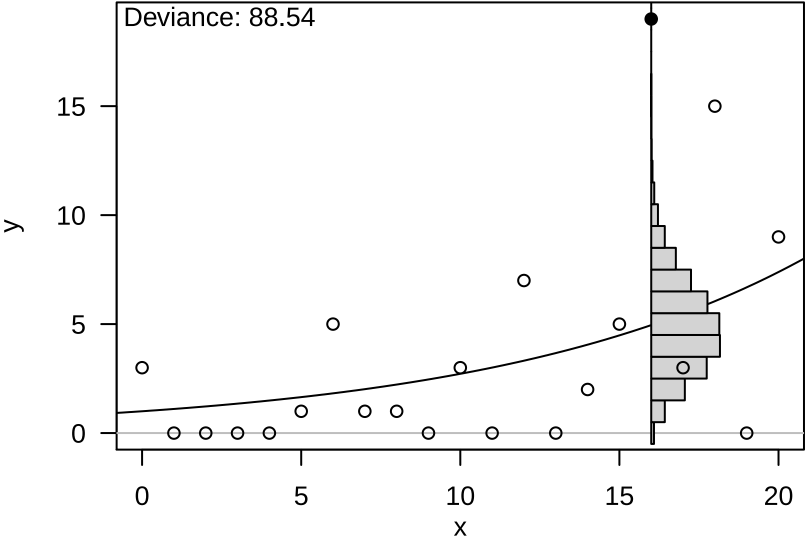
Deviance as a goodness-of-fit statistic

Variance too high: Large deviance



Deviance as a goodness-of-fit statistic

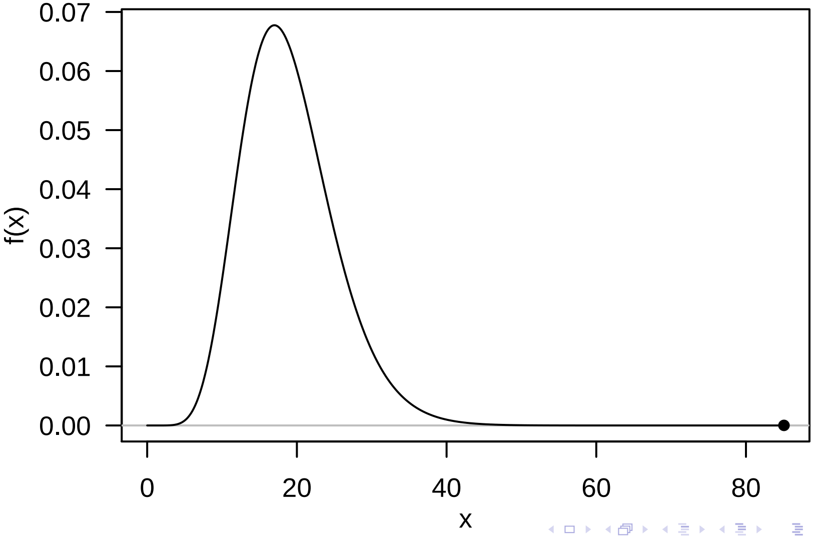
Variance too high: Large deviance



Deviance as a goodness-of-fit statistic

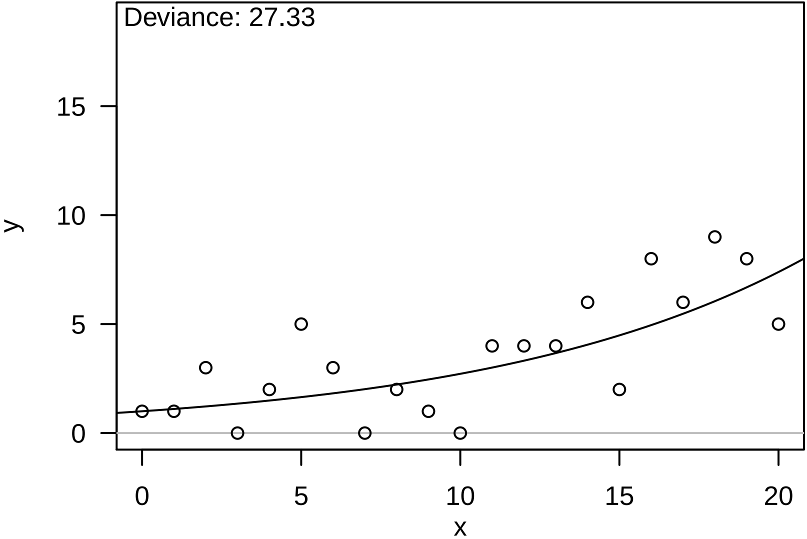
Variance too high: Large deviance

If the model is correct, the deviance comes from a chi-squared distribution with $n - k = 19$ degrees of freedom.



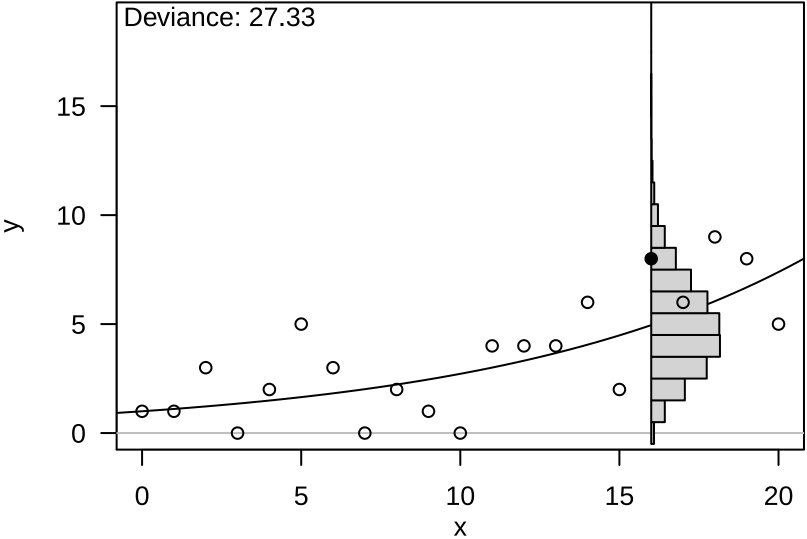
Deviance as a goodness-of-fit statistic

Variance about right: Small deviance



Deviance as a goodness-of-fit statistic

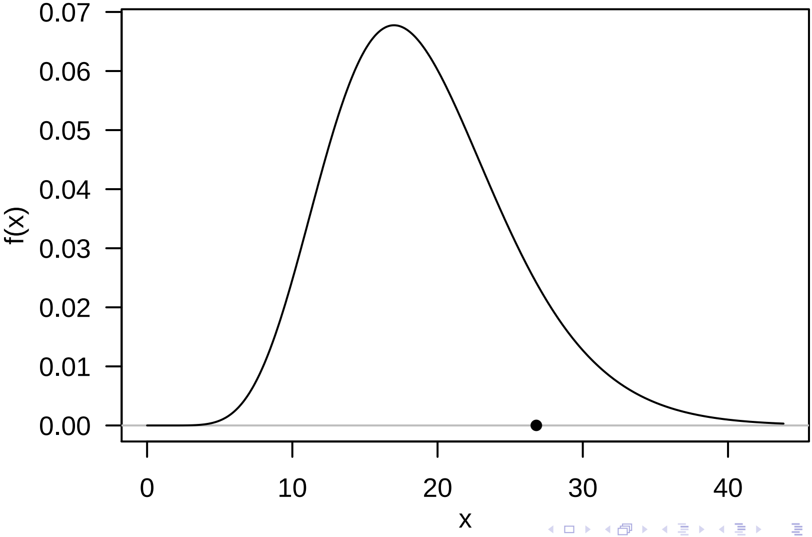
Variance about right: Small deviance



Deviance as a goodness-of-fit statistic

Variance about right: small deviance

If the model is correct, the deviance comes from a chi-squared distribution with $n - k = 19$ degrees of freedom.



Deviance as a goodness-of-fit statistic

Formally, we test the null hypothesis that the model is correct by calculating a *p*-value using

$$p = \Pr(\chi^2_{n-k} > D)$$

In other words, the *p*-value is the area beneath the chi-squared probability density function to the right of our observed deviance.

Recall that the area under the entire probability density function is equal to 1.

In R, we use

```r
1 - pchisq(deviance, df)
```

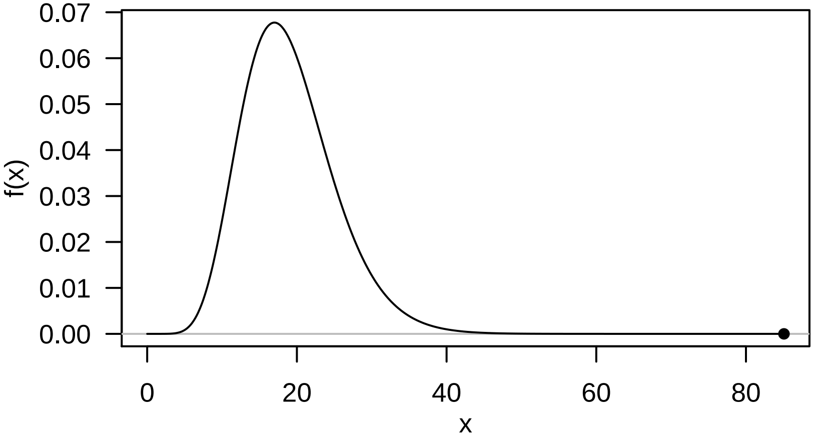
Deviance as a goodness-of-fit statistic

Variance too high: Large deviance

High deviance rejects the null hypothesis that the model is correct.

```
1 - pchisq(88.54, 19)


## [1] 6.011658e-11
```



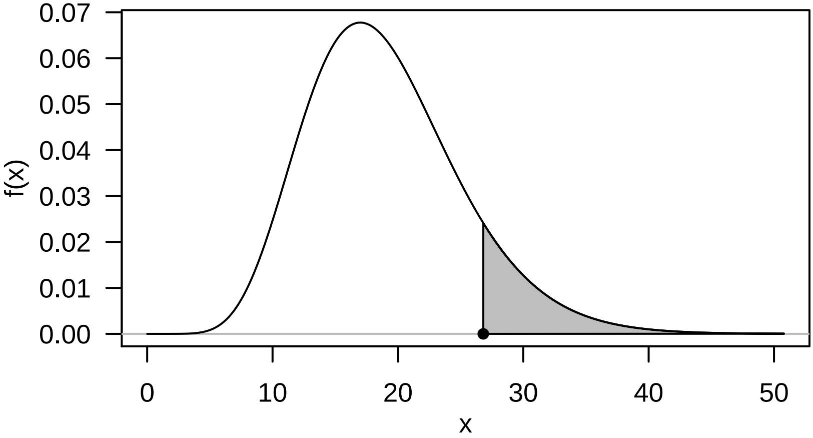
Deviance as a goodness-of-fit statistic

Variance about right: Small deviance

Small deviance does not reject the null hypothesis.

```r
1 - pchisq(27.33, 19)

## [1] 0.0971982
```



Deviance as a goodness-of-fit statistic

Poisson regression: *Macrorhabdus ornithogaster* chicken analysis

Consider our full *Macrorhabdus ornithogaster* chicken analysis, the Poisson regression containing weight of chicken and dose of Amphotericin B as explanatory variables:

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


## Call:
## glm(formula = mo ~ dose + weight, family = "poisson")
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.578171   1.015763  -2.538 0.011144 *
## doseHigh    -1.743623   0.127181 -13.710  < 2e-16 ***
## doseLow     -0.604736   0.177085  -3.415 0.000638 ***
## weight       0.012670   0.001964   6.450 1.12e-10 ***
## ---
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 683.02  on 22  degrees of freedom
## Residual deviance: 310.61  on 19  degrees of freedom
```

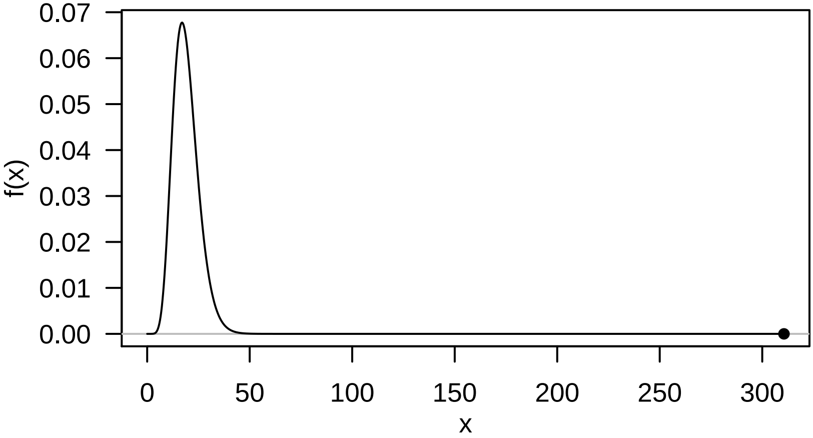
Deviance as a goodness-of-fit statistic

Poisson regression: *Macrorhabdus ornithogaster* chicken analysis

We have strong evidence to suggest lack-of-fit.

```
1 - pchisq(310.61, 19)
```

```
## [1] 0
```



Deviance as a goodness-of-fit statistic

Logistic regression: Coronary heart disease analysis

Consider our coronary heart disease analysis, the logistic regression:

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

## Call:
## glm(formula = cbind(y, n - y) ~ age, family = "binomial")
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.27844    1.13053  -4.669 3.03e-06 ***
## age          0.11032    0.02402   4.593 4.36e-06 ***
## ---
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 63.958  on 42  degrees of freedom
## Residual deviance: 34.976  on 41  degrees of freedom
```

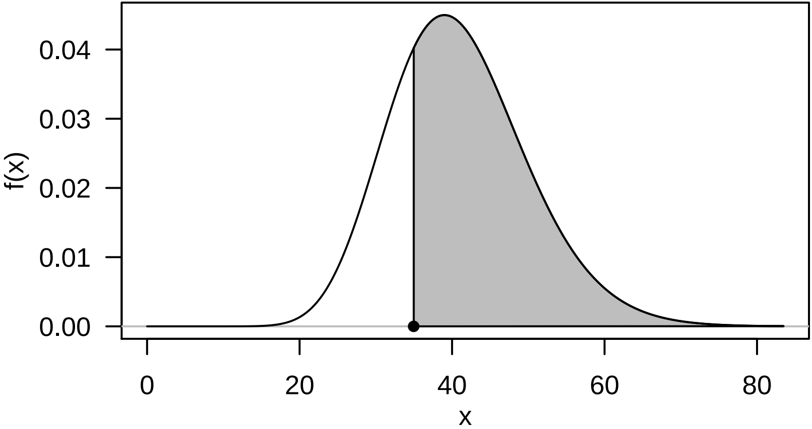
Deviance as a goodness-of-fit statistic

Logistic regression: Coronary heart disease analysis

We have no evidence to suggest lack-of-fit.

```
1 - pchisq(34.98, 41)
```

```
## [1] 0.7342761
```



Deviance as a goodness-of-fit statistic

Conditions of the chi-squared approximation

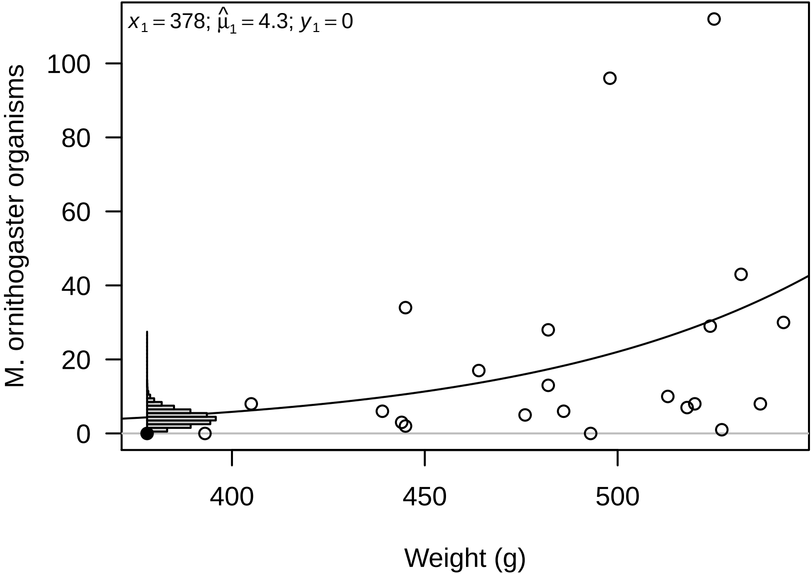
The distribution of the deviance under the null hypothesis is only approximately chi-squared if the response of each observation is well approximated by a normal distribution:

- This holds for Poisson random variables with $\mu_i \geq 5$
- This holds for binomial random variables if the number of trials, n_i, is large enough.
  - When p_i is close to 0.5, $n_i \geq 5$ is probably sufficient.
  - But if p_i is close to 0 or 1, n_i must be much larger.

The approximation is probably sound for our chicken analysis, but not for our coronary heart disease analysis.

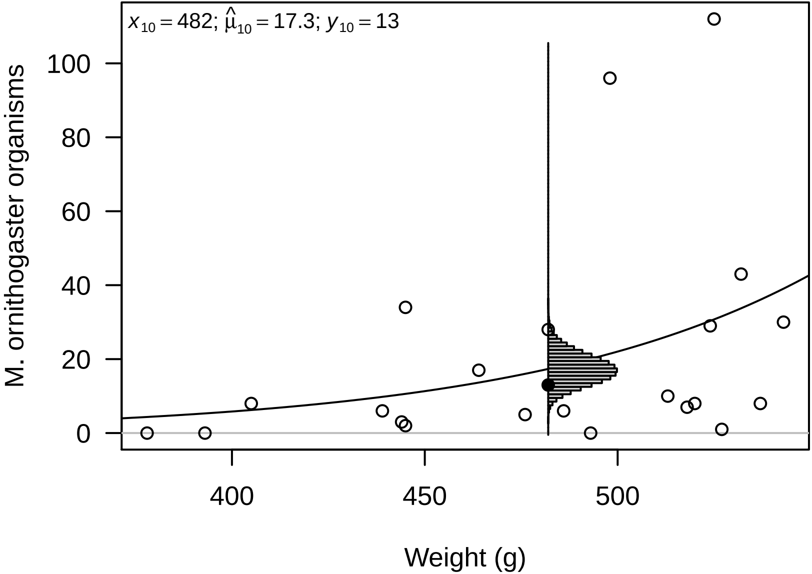
Deviance as a goodness-of-fit statistic

Conditions of the chi-squared approximation: *Macrorhabdus ornithogaster* chicken analysis



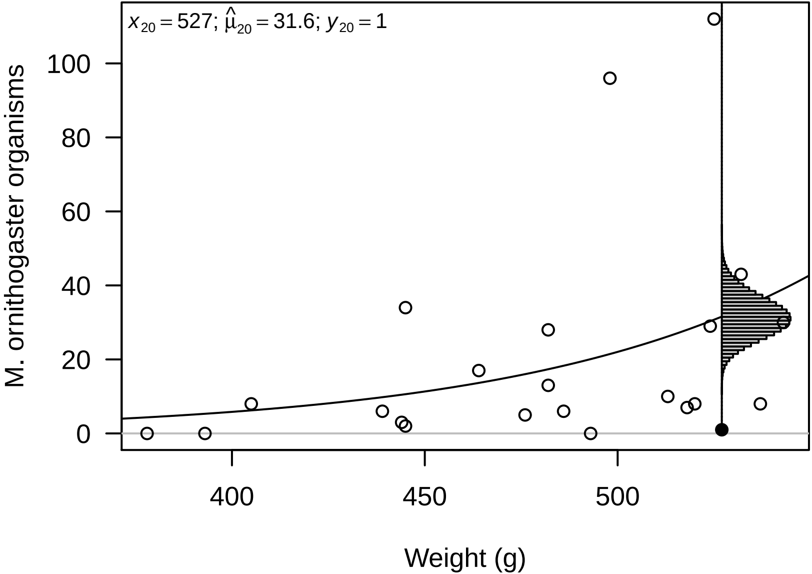
Deviance as a goodness-of-fit statistic

Conditions of the chi-squared approximation: *Macrorhabdus ornithogaster* chicken analysis



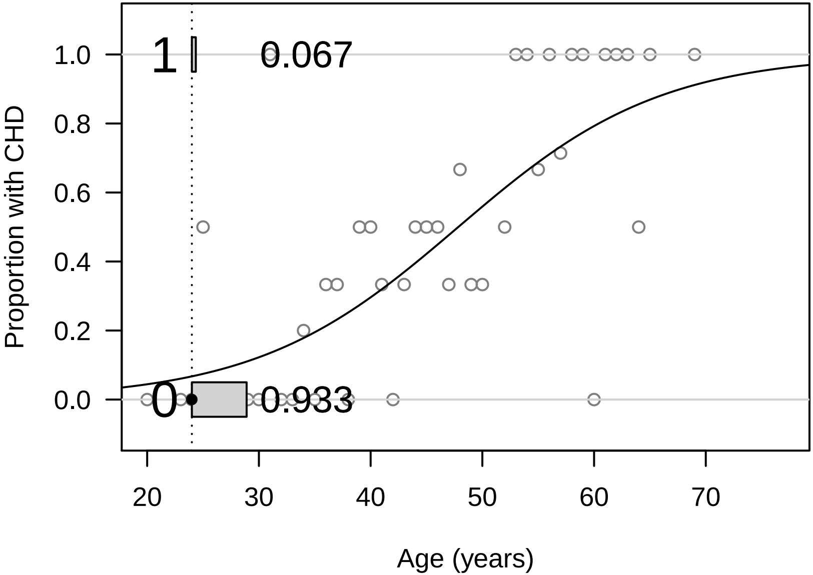
Deviance as a goodness-of-fit statistic

Conditions of the chi-squared approximation: *Macrorhabdus ornithogaster* chicken analysis



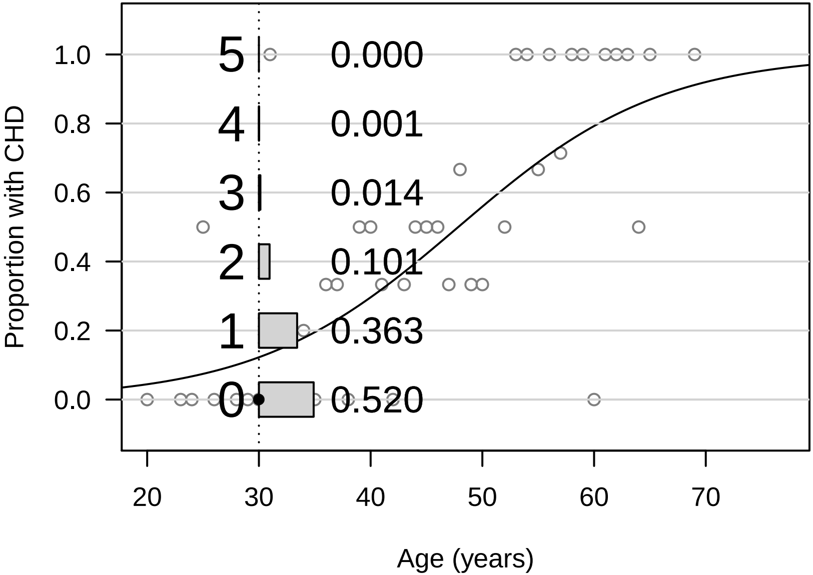
Deviance as a goodness-of-fit statistic

Conditions of the chi-squared approximation: Coronary heart disease analysis



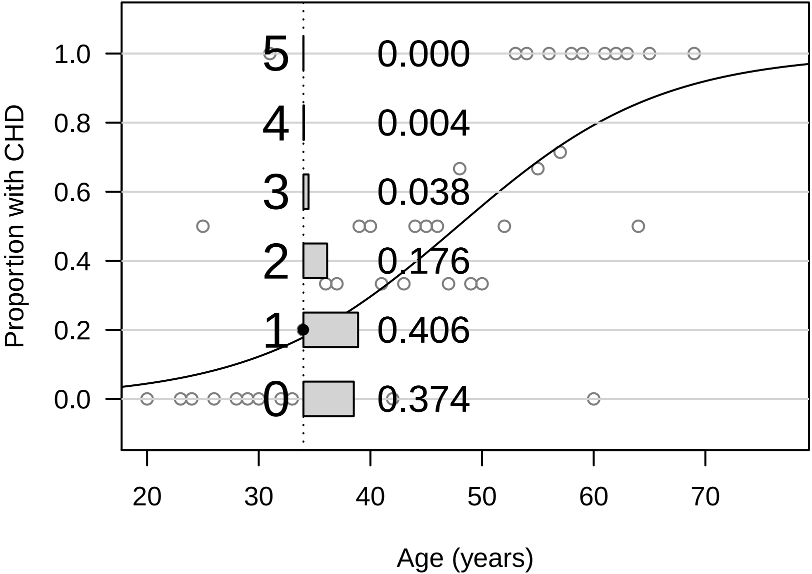
Deviance as a goodness-of-fit statistic

Conditions of the chi-squared approximation: Coronary heart disease analysis



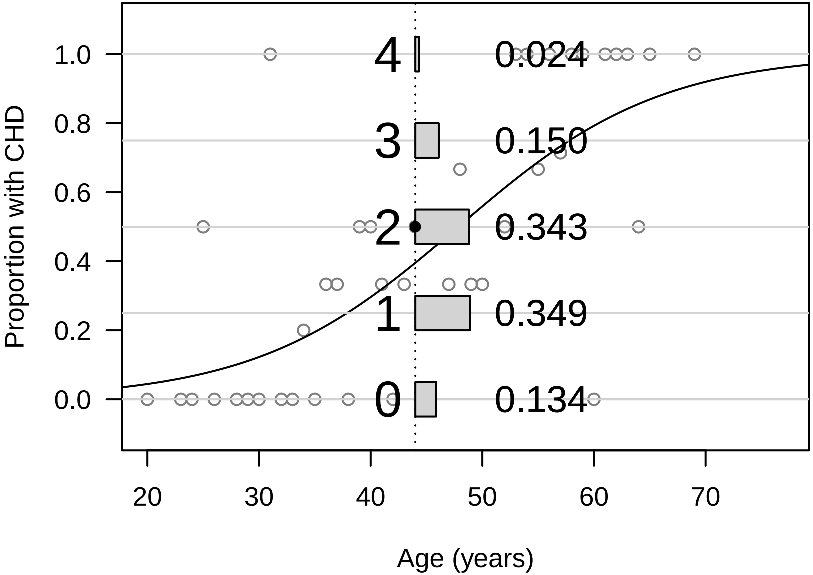
Deviance as a goodness-of-fit statistic

Conditions of the chi-squared approximation: Coronary heart disease analysis



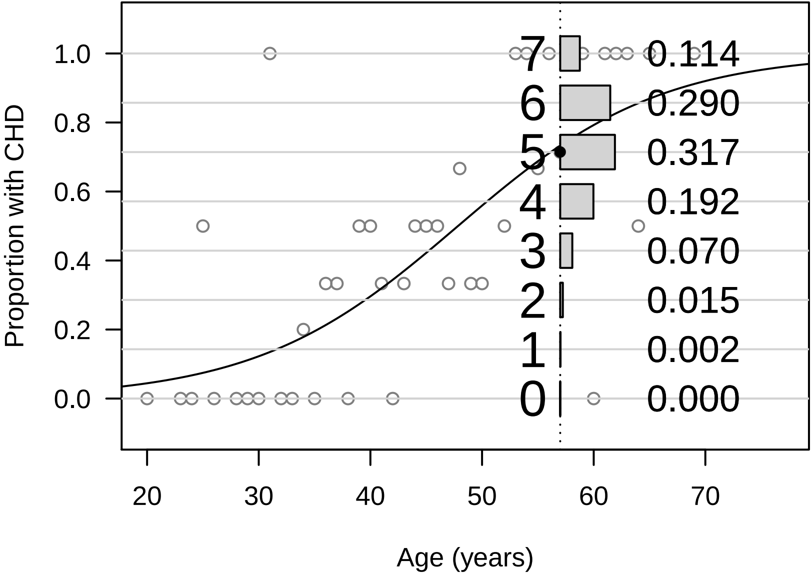
Deviance as a goodness-of-fit statistic

Conditions of the chi-squared approximation: Coronary heart disease analysis



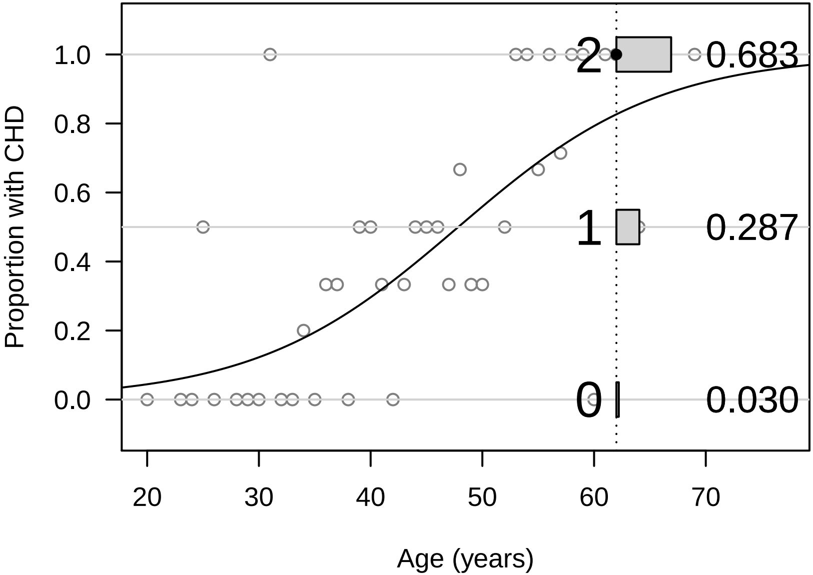
Deviance as a goodness-of-fit statistic

Conditions of the chi-squared approximation: Coronary heart disease analysis



Deviance as a goodness-of-fit statistic

Conditions of the chi-squared approximation: Coronary heart disease analysis



Deviance as a goodness-of-fit statistic

A summary

A large deviance suggests the model does not fit the data. We can sometimes compare the deviance to a chi-squared distribution.

- We have strong evidence to suggest that our Poisson regression model is not appropriate for the *Macrorhabdus ornithogaster* chicken data.

However, if the conditions of the chi-squared approximation are not met, then which distribution do we compare our deviance to?

- We can use simulation to find out! See a handout later in the course...