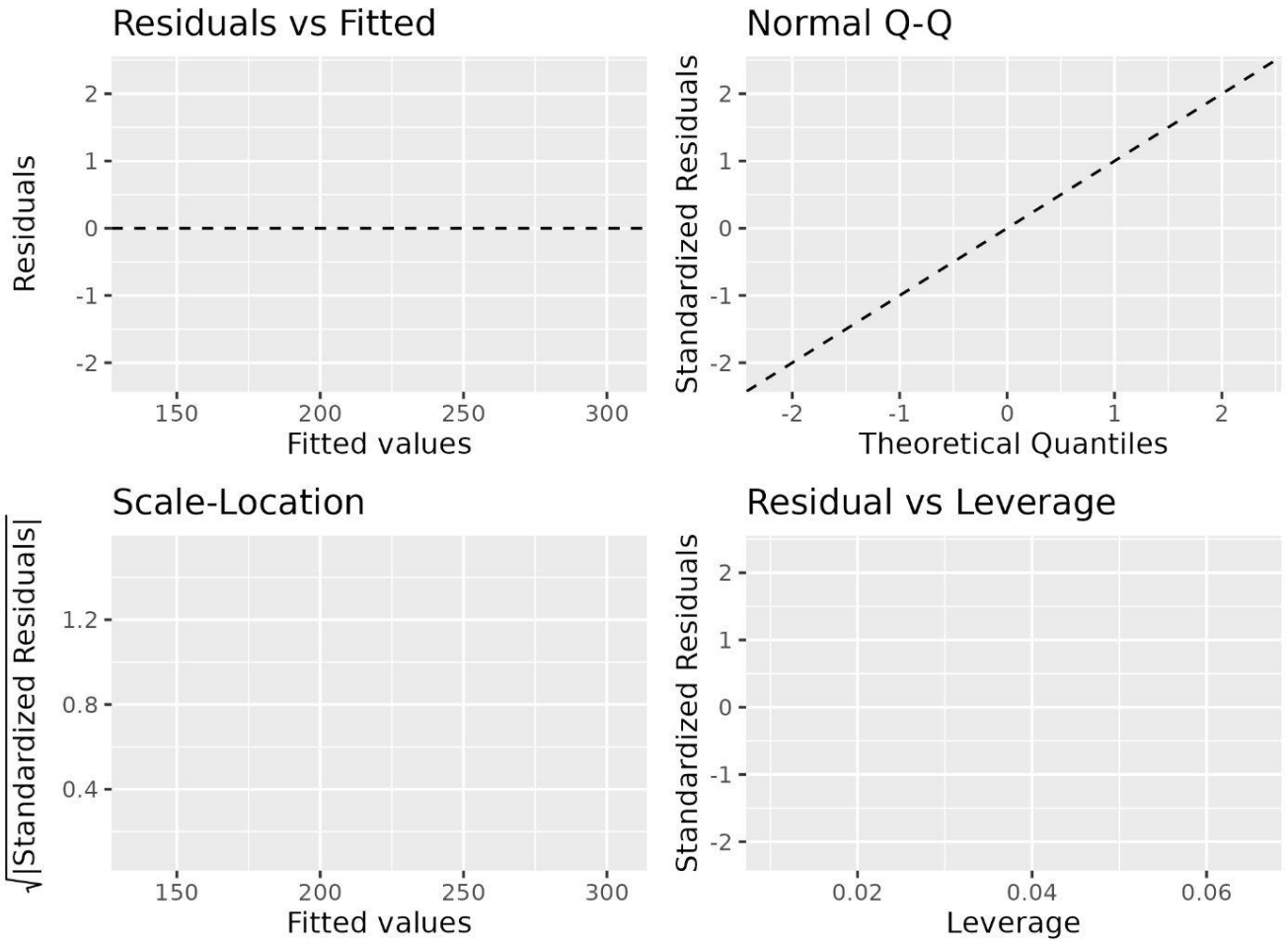
GLM model diagnostics

It’s still all about residuals

Recap: What do we look for in a linear model?



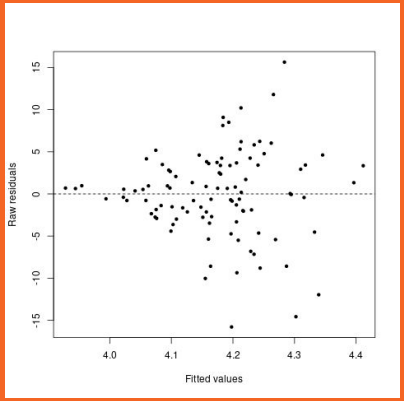
Residuals for GLMs

Response residuals are the conventional residual on the response level. That is, the fitted residuals are transformed by taking the inverse of the link function. Think back to linear models where the link function was set as the identity.

Pearson residuals are calculated by normalizing the raw residuals (i.e., expected - estimate) by the square root of the estimate.

Deviance residuals represent the contributions of individual samples to the deviance.

Residuals for GLM



Raw residuals for a Poisson regression…

But we expect the variance to increase with the fitted values under this model…

One solution is to divide these residuals by the expected variance under the assumed model (i.e., scaled/normalized) to give us **Pearson residuals**

Residuals for GLM

Quantile residuals are based on the idea of inverting the estimated distribution function for each observation to obtain exactly standard normal residuals (*definition beyond this course*)

- `statmod::qresiduals()`
- `DHARMa::simulateResiduals()`

Extra reading? http://www.statsci.org/smyth/pubs/residual.html

Deviance (using the chi-squared approach)

The **deviance** is a measure of how well the model fits the data.

That is, if the model fits well, the observed values will be close to their predicted means and so the deviance will be small. The flip side of this is that a large deviance indicates a bad fitting model.

For a fitted Poisson regression the deviance, D, is

$$D = 2\sum_{i=1}^{n}\{Y_i \log(Y_i/\mu_i) - (Y_i - \mu_i)\}$$

where if $Y_i = 0$, the $Y_i \log(Y_i/\mu_i)$ term is taken to be zero, and $\mu_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_p X_p)$ is the predicted mean for observation i based on the estimated model parameters.

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

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

*under some assumptions (e.g., for Poisson mu > 5, for binomial '*big*" n)

Deviance (via simulation)

Let **deviance** / **df** = **dispersion** statistic

Under an appropriate model the expected **dispersion** statistic = **1** (i.e., the ratio of the residual deviance and the residual degrees of freedom. Below is an excerpt from Module 4 (the bird abundance model). We have dispersion statistic = 1621/78 = 20.8! But could this have occurred by chance?

```
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2234.6  on 79  degrees of freedom
## Residual deviance: 1621.9  on 78  degrees of freedom
```

Deviance (via simulation)

Let **deviance** / **df** = **dispersion** statistic

Under an appropriate model the expected **dispersion** statistic = **1** (i.e., the ratio of the residual deviance and the residual degrees of freedom. Below is an excerpt from Module 4 (the bird abundance model). We have dispersion statistic = 1621/78 = 20.8! But could this have occurred by chance? Let's formulate a hypothesis test…

1. For each observation generate a random Poisson count under the fitted model
2. Fit a model and compute the dispersion statistic.
3. Repeat 1 and 2 for a chosen number of times
4. Compare the observed statistic to the created sampling distribution

What is this an example of??

See also `DHARMa::simulateResiduals()` and `DHARMa::testDispersion()`