Introduction to Computational BioStatistics with R: generalized linear models

Erik Spence

27 October 2022

SciNet compute • calcul CANADA

Today's slides

Today's slides can be found here. Go to the "Introduction to Computational BioStatistics with R" page, under Lectures, "Generalized Linear Models".

`https://scinet.courses/1246`

SciNet compute • calcul CANADA

Today's class

Today we will continue our adventures in data analysis.

- Verification of models.
- Generalized linear models.

As always, ask questions.

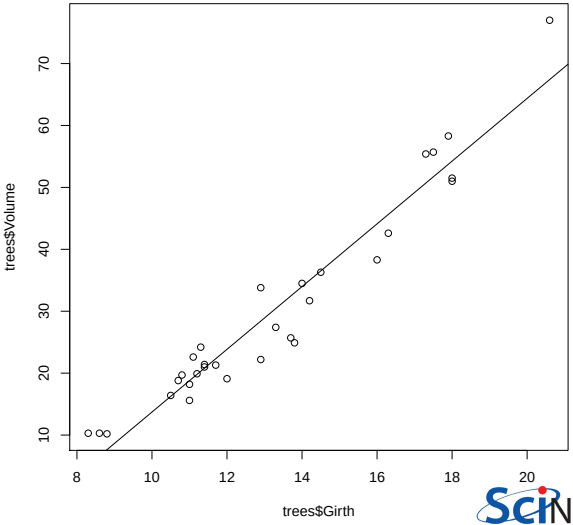
SciNet compute • calcul CANADA

Our linear model

At the end of last class we had created a linear model to describe the relationship between Girth and Volume.

```
>
> model <- lm(Volume ~ Girth,
+       data = trees)
>
> plot(trees$Girth, trees$Volume)
> abline(model)
>
```

How do we assess the quality of our model?



Our linear model, continued

As noted last class, the summary gives important information:

- Information about the null hypothesis $\beta_1 = ... = \beta_n = 0$.
- Information about the individual null hypotheses: $\beta_1 = 0$, $\beta_2 = 0$, etc.

Remember that the significance code only tells you the likelihood that $\beta_i = 0$.

```
> summary(model)
Call:
lm(formula = Volume ~ Girth, data = trees)
Residuals:
     Min       1Q   Median       3Q      Max
 -8.065   -3.107    0.152   3.495   9.587
Coefficients:
             Estimate   Std. Error   t value   Pr(>|t|)
 (Intercept)  -36.9435       3.3651    -10.98   7.62e-12   ***
 Girth          5.0659       0.2474     20.48    < 2e-16   ***
---
Signif.  codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error:  4.252 on 29 degrees of freedom
Multiple R-squared:  0.9353, Adjusted R-squared:  0.9331
F-statistic:  419.4 on 1 and 29 DF, p-value:  < 2.2e-16
>
```

That's great, but we're not done yet

It's always a good idea to do some further analysis of your model before declaring success. There are a few things in particular that should always be done.

- plot the residuals of the model, in various ways,
- examine the statistics of the residuals,
- examine the statistics of the model.

What are residuals? Residuals are the distance between the actual value, and the value predicted by the model, for each data point:

$$\boldsymbol{R_i = f(x_i) - y_i}$$

where $\boldsymbol{f}$ is the model, evaluated at data point $\boldsymbol{x_i}$, and $\boldsymbol{y_i}$ is the actual value of the dependent variable.

SciNet compute • calcul CANADA

Step 1: plot the residuals

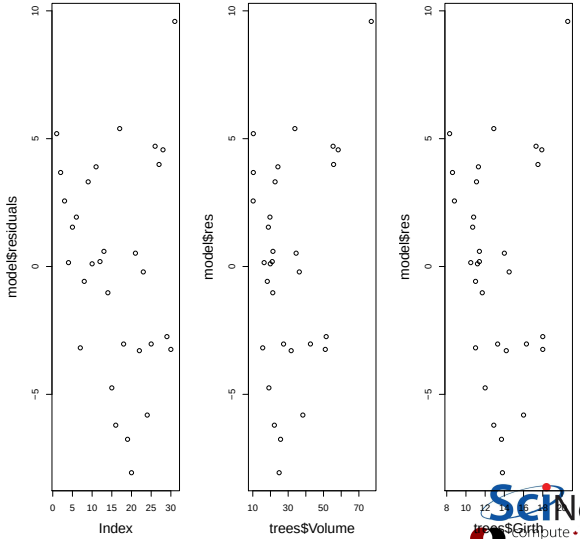
Always plot your residuals. Always.

```
> par(mfrow = c(1, 3))
>
> plot(model$residuals)
> plot(trees$Volume, model$residuals)
> plot(trees$Girth, model$residuals)
>
```

Plot your residuals against everything:

- index,
- against the dependent variables,
- against the independent variables.

You should see a snowstorm. There should be no clumps.

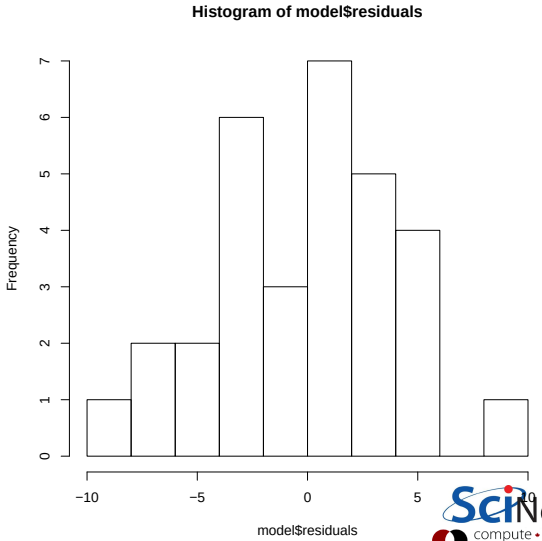


Step 2: plot the residuals via histogram

Always plot a histogram of your residuals. Things to look for:

- The mean should be zero. If your residuals are not centred on zero your model is missing something.
- The distribution should be symmetric. If it's not, it's biased (there 'structure' in the data which has not been captured by the model).
- Distribution should be a Gaussian (an assumption made as part of the fit).

```
> par(mfrow = c(1, 1))
> hist(model$residuals, breaks = 11)
```

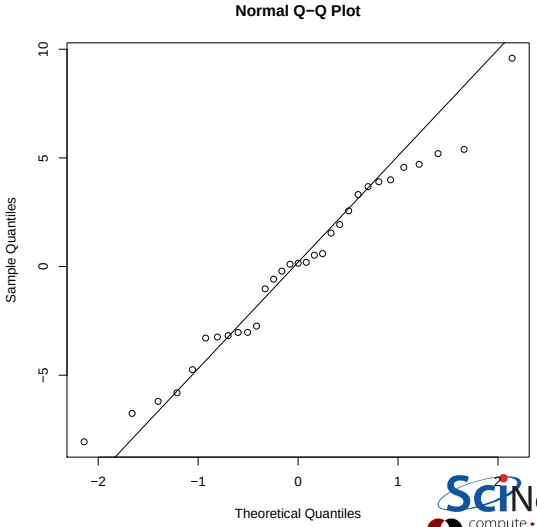


Step 3: plot the residuals via Q-Q plot

Plot your residuals on a Q-Q plot.

- A Q-Q plot graphically demonstrates how normally-distributed the residuals are.
- Ideally the residuals should be normally distributed.
- A perfect Gaussian will lie exactly on the line.

```
>
> qqnorm(model$residuals)
> qqline(model$residuals)
>
```



Anscombe's Quartet

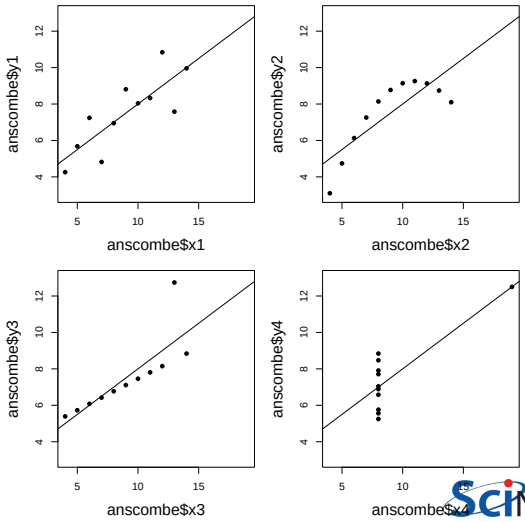
There's a strange data set called Anscombe's Quartet. This data set is good for demonstrating the utility of examining model residuals.

The data set consists of four sets of data which have essentially identical properties. It's built into R.

```
> print.stats <- function(x, y) {
+    cat(mean(x), sd(x), mean(y), sd(y),
+        cor(x,y), var(x,y), '\n') }
>
> print.stats(anscombe$x1, anscombe$y1)
9 3.316625 7.500909 2.031568 0.8164205 5.501
>
> print.stats(anscombe$x2, anscombe$y2)
9 3.316625 7.500909 2.031657 0.8162365 5.5
>
> print.stats(anscombe$x3, anscombe$y3)
9 3.316625 7.5 2.030424 0.8162867 5.497
>
> print.stats(anscombe$x4, anscombe$y4)
9 3.316625 7.500909 2.030579 0.8165214 5.499
>
```

Anscombe's Quartet, continued

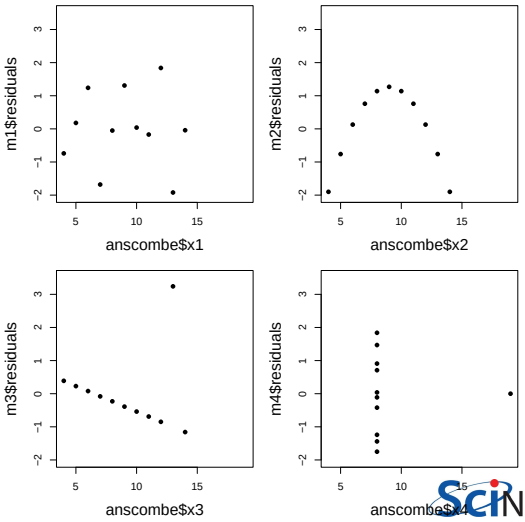
```
>
> plot.ans <- function(x, y) {
+     plot(x, y, xlim = c(4, 19),
+          ylim = c(3, 13))
+     m <- lm(y ~ x)
+     abline(m)
+ }
>
> par(mfrow = c(2, 2))
>
> plot.ans(anscombe$x1, anscombe$y1)
> plot.ans(anscombe$x2, anscombe$y2)
> plot.ans(anscombe$x3, anscombe$y3)
> plot.ans(anscombe$x4, anscombe$y4)
>
```



Anscombe's Quartet, continued more

```
> plot.res <- function(x, y) {
+     m <- lm(y ~ x)
+     plot(x, m$residuals, xlim = c(4, 19),
+          ylim = c(-2, 3.5))
+ }
>
> par(mfrow = c(2, 2))
>
> plot.res(anscombe$x1, anscombe$y1)
> plot.res(anscombe$x2, anscombe$y2)
> plot.res(anscombe$x3, anscombe$y3)
> plot.res(anscombe$x4, anscombe$y4)
>
```

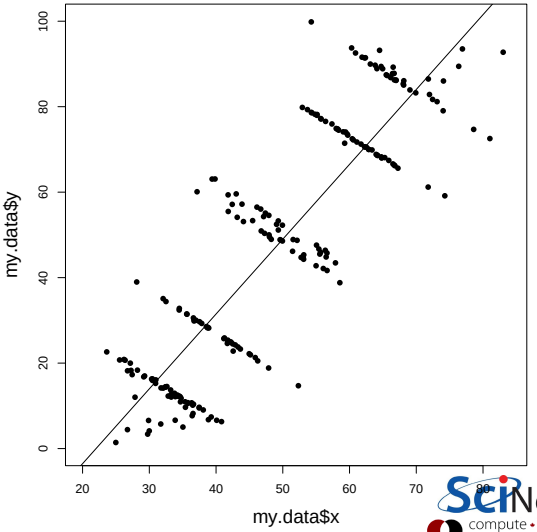
Clearly, for the later three cases, the residuals are not randomly spread.



Simpson's paradox

Simpson's paradox (also called the Yule-Simpson effect) is a phenomenon in statistics in which a trend appears in several different groups of data, but disappears or reverses when these groups are combined.

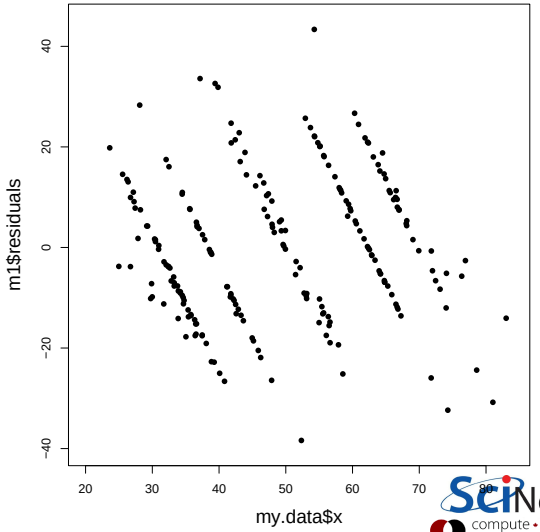
```
> library(datasauRus)
>
> sp <- simpsons_paradox
> my.data <- sp[sp$dataset == 'simpson_2',]
>
> plot(my.data$x, my.data$y)
> m1 <- lm(y ~ x, data = my.data)
> abline(m1)
```



Simpson's paradox, continued

Plotting the residuals in this case will demonstrate plenty of structure, indicating that there's something wrong with the model, or that, as in this case, the model is incomplete.

```
>
> plot(my.data$x, m1$residuals)
>
```



Using R^2

R^2 = (explained variation) / (total variation).

- Explains how much of the variance in the data can be explained by the model.
- All other variation is caused by shortcomings in the model, or noise.
- A high R^2 value is necessary, but not sufficient, for the model to be satisfactory.

```
> summary(model)
Call:
lm(formula = Volume ~ Girth, data = trees)
Residuals:
     Min       1Q   Median       3Q      Max
 -8.065   -3.107    0.152   3.495   9.587
Coefficients:
             Estimate   Std. Error   t value   Pr(>|t|)
 (Intercept)  -36.9435       3.3651    -10.98   7.62e-12   ***
 Girth          5.0659       0.2474     20.48    < 2e-16   ***
---
Signif.  codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error:  4.252 on 29 degrees of freedom
Multiple R-squared:  0.9353, Adjusted R-squared:  0.9331
F-statistic:  419.4 on 1 and 29 DF, p-value:  < 2.2e-16
>
```

Other regression models

There are other types of regression models available:

- Logistic (Logit) Regression: used to fit a categorical variable against a continuous independent variable
- Multinomial Logistic Regression: logistic regression where the dependent variable has more than two outcome categories. If the multiple categories are ordered this is called Ordinal Logistic Regression.
- Generalized Linear Models: multiple independent variables, different link functions and noise families.

We will examine Generalized Linear Models today. We will revisit Logistic Regression in few weeks.

SciNet compute • calcul CANADA

Generalized linear models

The linear model built by "lm" has some built-in assumptions:

- Normally distributed noise,
- Uncorrelated noise,
- Constant variance of the noise.

There are situations where these assumptions are dramatically violated. To deal with this, let us examine "Generalized Linear Models". These allow

- Non-normally distributed noise.
- Non-constant variance.

If you find that you have structure in your residuals, it's possible that you need to use a generalized linear model.

SciNet compute • calcul CANADA

Generalized linear models, continued

When should you use a generalized linear model?

- You know that your data should come from a non-linear, non-polynomial distribution (exponential, Poisson, etc).
- You don't know what your distribution should be, and you've got structure in your residuals.

How do generalized linear models work? Let's start with a regular linear model. Assuming the vectors of data are $(\boldsymbol{X}, \boldsymbol{Y})$, the problem is to find the vector of coefficients $\boldsymbol{\beta}$ such that

- $\boldsymbol{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$
- assuming that $\boldsymbol{Y} \sim \boldsymbol{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\sigma^2})$,

where $\boldsymbol{E}$ is the expectation value, $\boldsymbol{N}(\boldsymbol{\mu}, \boldsymbol{\sigma^2})$ is the symbol for a normal distribution centred on $\boldsymbol{\mu}$ with a standard deviation of $\boldsymbol{\sigma}$.

Generalized linear models, continued

As an example, for a log-linked Gaussian GLM, we have

- $\log\left(\boldsymbol{E}(\boldsymbol{Y})\right) = \boldsymbol{X\beta}$,
- which means that $\boldsymbol{E}(\boldsymbol{Y}) = e^{\boldsymbol{X\beta}}$,
- $\boldsymbol{Y} \sim \boldsymbol{N}(e^{\boldsymbol{X\beta}}, \boldsymbol{\sigma^2})$.

where $\boldsymbol{E}$ is the expectation value, $\boldsymbol{N}(\boldsymbol{\mu}, \boldsymbol{\sigma^2})$ is the symbol for a normal distribution centred on $\boldsymbol{\mu}$ with a standard deviation of $\boldsymbol{\sigma}$.

Generalized linear models consist of 3 parts:

- A "link" function. A function which transforms the data such that it becomes linear.
- A linear predictor ($\boldsymbol{X\beta}$).
- A probability distribution, which describes the type of noise to be expected in the dependent variable.

Generalized linear models, continued more

There are many possible link functions available. The most common ones are

- Identity: $\boldsymbol{E(Y) = X\beta}$,
- Log: $\boldsymbol{\log\left(E(Y)\right) = X\beta \quad \rightarrow \quad E(Y) = e^{X\beta}}$.
- Logit: $\boldsymbol{\log\left(\frac{E(Y)}{1-E(Y)}\right) = X\beta \quad \rightarrow \quad E(Y) = \frac{1}{1+e^{-X\beta}}}$
- Inverse: $\boldsymbol{1/E(Y) = X\beta \quad \rightarrow \quad E(Y) = 1/\left(X\beta\right)}$

The identity link function results in a standard linear regression. By performing a generalized linear model using this link function, with Gaussian noise, you will get the same result as using the "lm" function.

Generalized linear models, continued even more

Once a link function has been chosen, the type of error in the data must be chosen. The different error families have different default link functions.

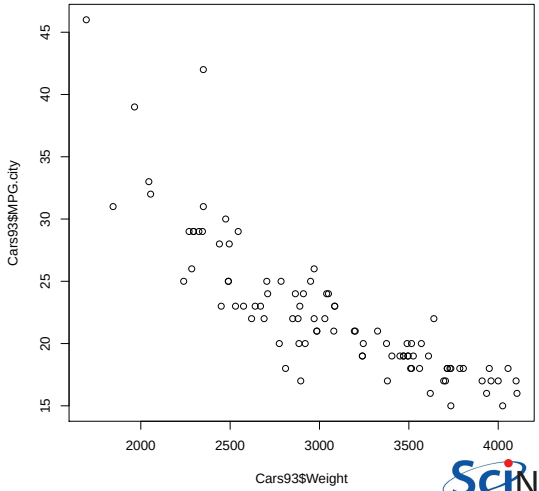
Error family	Default link	Link inverse	Use for:
gaussian	identity	1	Normally distributed error
poisson	log	exp	Counts
binomial	logit	$1/(1+e^{-x})$	Proportions or binary data
gamma	inverse	$1/x$	Continuous data with non-constant error

```
> glm(formula, family = binomial(link = log))
```

GLM example

Consider the Cars93 data set.
Plotting the MPG in the city, versus Weight, suggests a non-linear relationship.

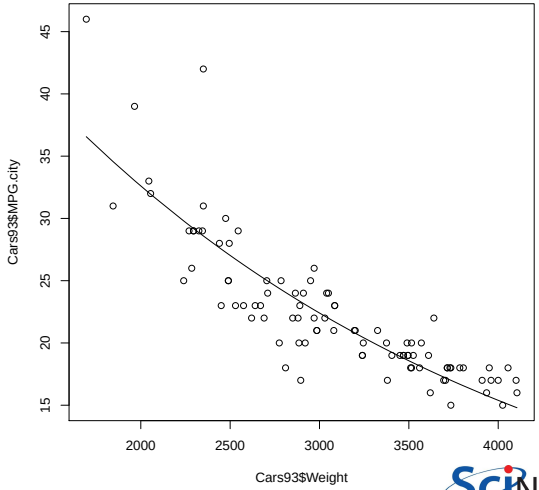
```
>
> library(MASS)
>
> plot(Cars93$Weight, Cars93$MPG.city)
>
```



GLM example, continued

Let's perform a GLM, using Gaussian noise and the log link function.

```
> sorted.weights <- sort(Cars93$Weight)
>
> glm1 <- glm(MPG.city ~ Weight,
+     data = Cars93,
+     family = gaussian(link = "log"))
>
> plot(Cars93$Weight, Cars93$MPG.city)
> lines(sorted.weights,
+     predict(glm1,
+     data.frame(Weight = sorted.weights),
+     type = "response"))
>
```



Summary

We've started looking at the quality of our linear models. Things to remember:

- Plot the residuals! There is important information in there!
  - make sure you get a snow storm!
  - make sure there is no structure, and no clumps, in your residuals.
  - make sure the spread in the data is constant, and not increasing or decreasing.
  - make sure the histogram of the residuals is Gaussian.
- If the data are not polynomial, or the residuals are not normally distributed, you may need to use a Generalized Linear Model.
- You will likely need to play around with the different noise families and link functions to find one that best works with your data.
- Other types of regression include logistic regression, for fitting categories, and multinomial regression, for multiple dependent variables.

SciNet compute • calcul CANADA