Model diagnostics

Patrick Breheny

February 10

Introduction

- We have derived an F test for testing several hypotheses at once; when would one use this test in practice?
- Its primary use is in comparing a complicated model and a simpler model that can be considered to be a special case of the more complicated model (such models are said to be *nested*)

Nested models: example

- For example, consider these two models for our alcohol metabolism data:

$$\mathrm{E}(\texttt{Metabol}) = \beta_0 + \beta_1 \texttt{Gastric}$$
$$\mathrm{E}(\texttt{Metabol}) = \beta_0 + \beta_1 \texttt{Male} + \beta_2 \texttt{Gastric} + \beta_3 \texttt{MaleGastric}$$

- The first is a simple linear regression model; the second allows for separate regression lines by sex
- Note that the first is a special case of the second, if $\beta_1 = \beta_3 = 0$
- Model 1 is said to be "nested" inside Model 2, with Model 2 the "full" model and Model 1 the "reduced" model

Comparing Models 1 and 2

- Because these two models are nested, Model 2 will always be able to explain more variability than Model 1 ($\hat{\beta}_1$ and $\hat{\beta}_3$ will be specifically chosen so as to make the RSS as small as possible)
- However, that doesn't necessarily make Model 2 better
- Simply by random chance, the full model will always explain more variability than the reduced model; what the F test does is to test whether the observed reduction in variability is larger than what you would expect by chance alone

ANOVA tables

The information relevant to this test can be summarized in what is called an *ANOVA table*:

	p	RSS	q	ΔRSS	F	p
Model 1	2	69.13				
Model 2	4	40.81	2	28.32	9.7	.0006

In this case, the reduction is RSS is much larger than you would expect by chance alone

ANOVA tables (cont'd)

ANOVA tables are often used to describe the reduction in RSS that occurs as a sequence of increasingly complicated models are fit to data:

	p	RSS	q	$\Delta\mathrm{RSS}$	F	p
Intercept only	1	219.09				
`Gastric`	2	69.13	1	149.97	102.89	< 0.0001
`Gastric` + `Male`	3	51.40	1	17.73	12.16	0.0016
`Gastric` * `Male`	4	40.81	1	10.59	7.26	0.0118

ANOVA tables (cont'd)

Note, however, that ANOVA tables are order-dependent:

	p	RSS	q	ΔRSS	F	p
Intercept only	1	219.09				
`Male`	2	147.20	1	71.89	49.32	< 0.0001
`Gastric` + `Male`	3	51.40	1	95.80	65.73	< 0.0001
`Gastric` $*$ `Male`	4	40.81	1	10.59	7.26	0.0118

This idea can be extended to R^2 as well:

	p	R^2	$\Delta\mathrm{R}^2$
Intercept only	1	0	
`Gastric`	2	0.684	0.684
`Gastric` + `Male`	3	0.765	0.081
`Gastric` * `Male`	4	0.814	0.048

	p	R^2	$\Delta\mathrm{R}^2$
Intercept only	1	0	
`Male`	2	0.328	0.328
`Gastric` + `Male`	3	0.765	0.437
`Gastric` * `Male`	4	0.814	0.048

R^2 is sometimes called the *coefficient of determination* of the model, and these $\Delta\mathrm{R}^2$ values the *coefficients of partial determination*

"The" F test

- It should be noted that both SAS and R report by default an F test associated with the entire model
- This is an F test of $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$, and is sometimes called the "overall F test" or just "the" F test
- This test is sometimes used to justify the model
- However, this is a mistake

"The" F test

- Recall the assumptions of our F test derivation: that the model (2) holds
- Thus, the F test takes the model as given and cannot possibly be a test of the validity of the model
- The only thing one can conclude from a significant overall F test is that, if the model is true, some of its coefficients are nonzero
- In other words, big deal!
- Addressing the validity of a model is much more complicated than a simple overall F-test, and it is here that we now turn our attention

Graphical diagnostics

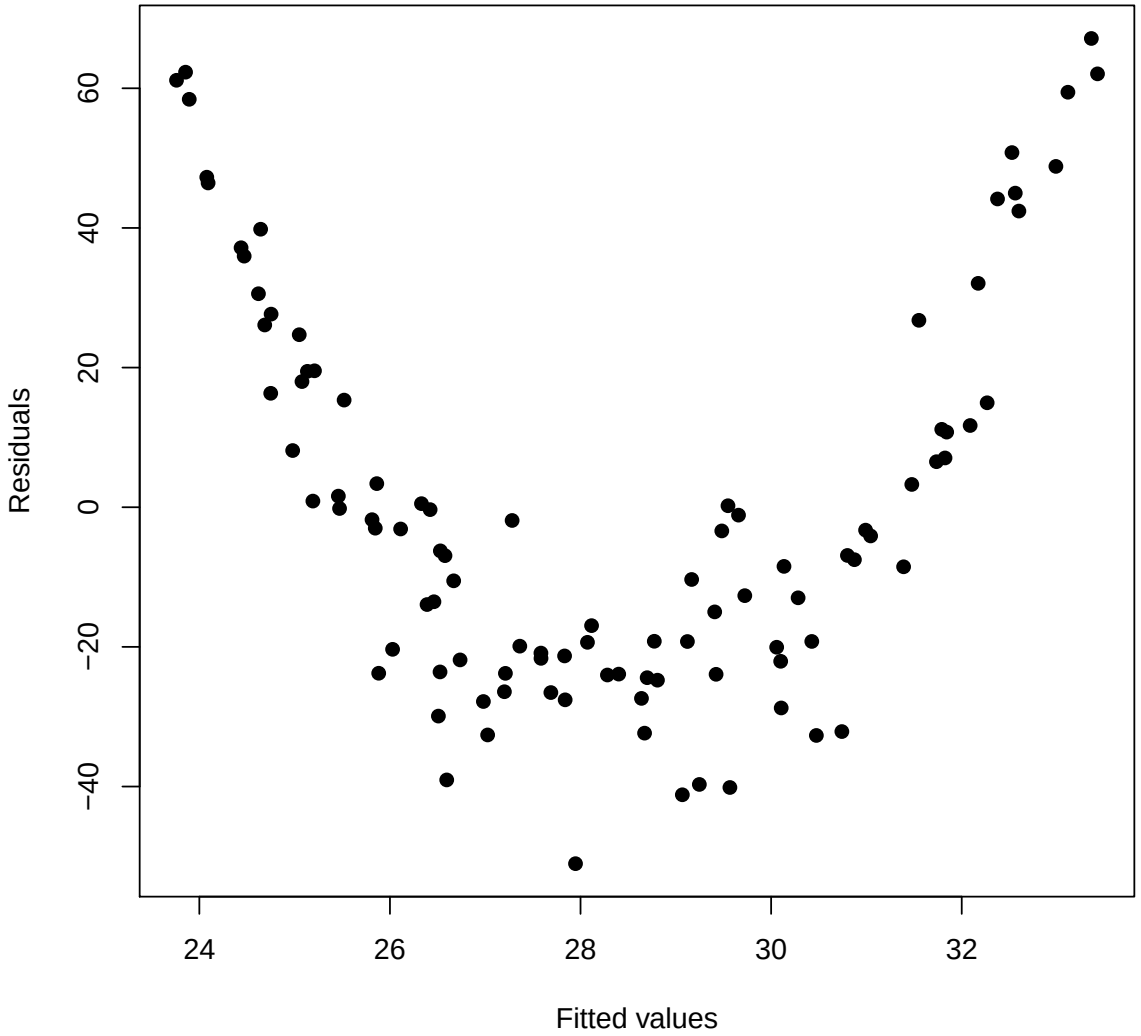
- The validity of the model's assumptions are best checked through the use of graphical diagnostics
- Each type of graph that we are about to discuss has different strengths in terms of allowing you to spot particular discrepancies between the data and the assumptions of the model

Residual plot

- One basic plot is a plot of the fitted values $\{\hat{\mu}_i\}$ versus the residuals $\{r_i\}$
- If the model accurately describes the data, this should just look like random noise
- However, any of the following scenarios can produce patterns in this plot which should give the statistician cause for concern:
  - Inadequate fit
  - Constancy of σ^2 (*homoskedasticity*)
  - Outliers

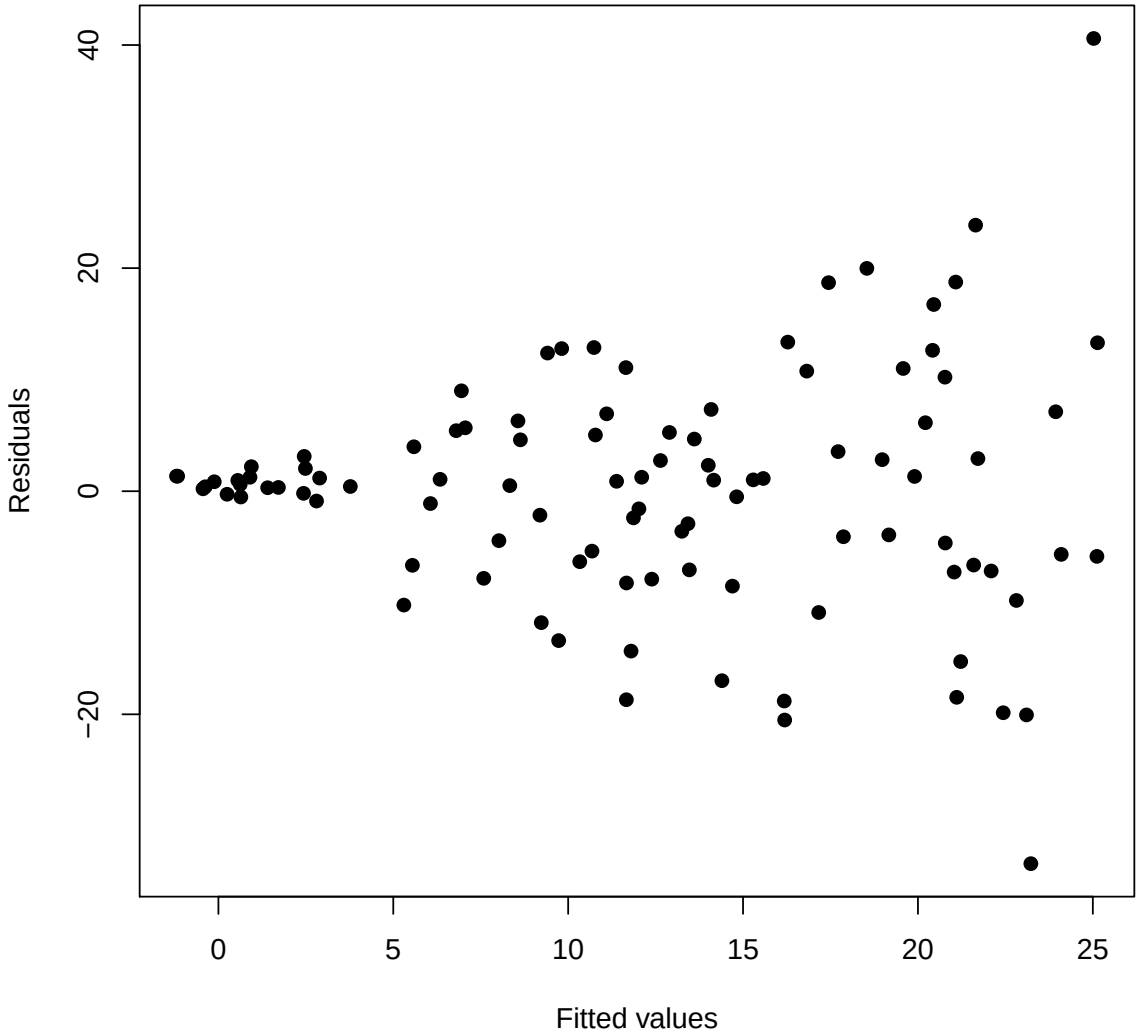
Inadequate fit

For example, suppose that we fit a linear model, but $\mathrm{E}(y) \propto x^2$:



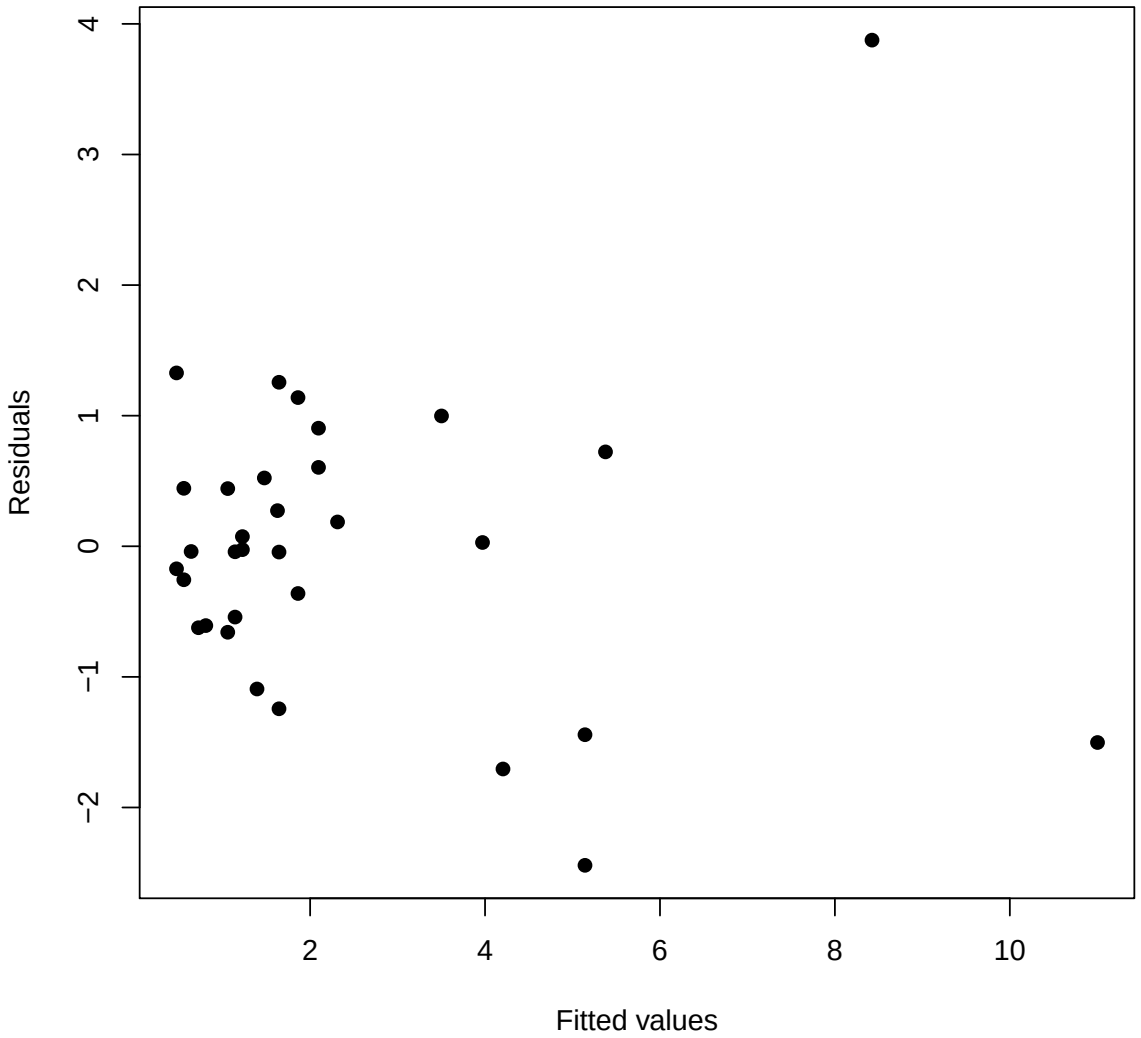
Constancy of σ^2

Suppose that $\mathrm{Var}(y) \propto \mathrm{E}(y)$:



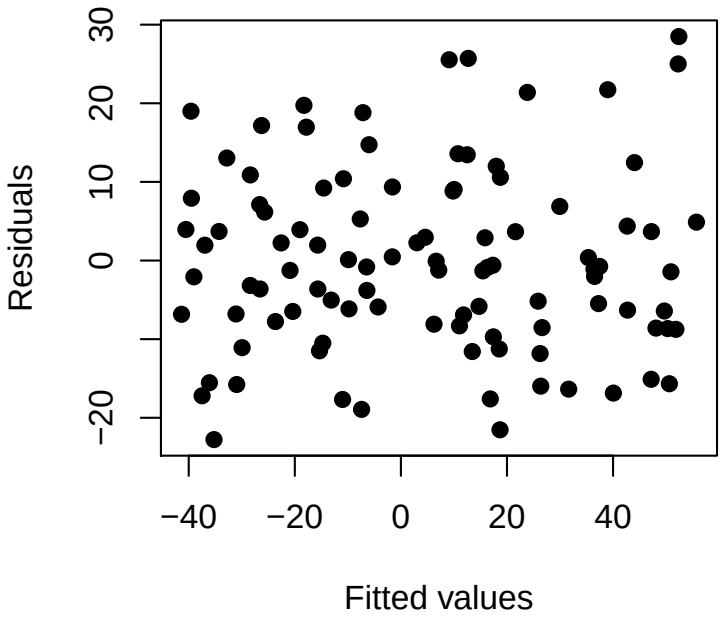
Outliers

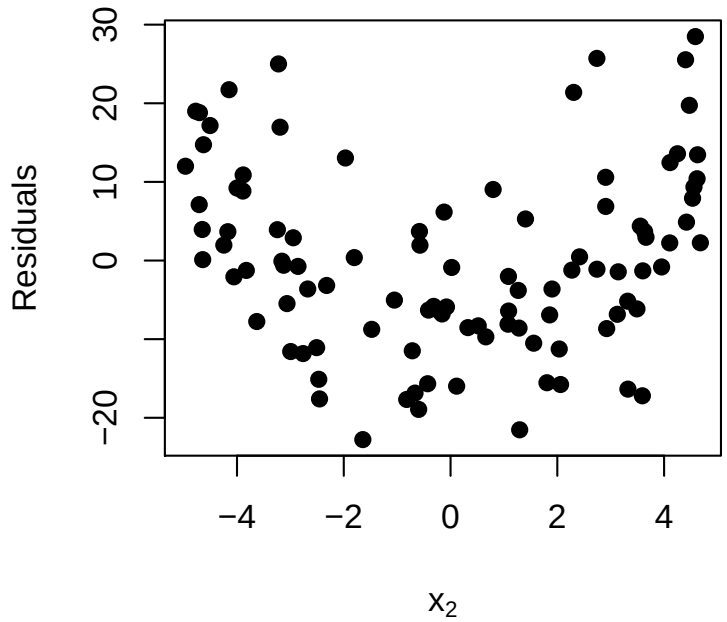
From the `alcohol` data set, using the `Gastric` $*$ `Male` model:



Residuals versus explanatory variables

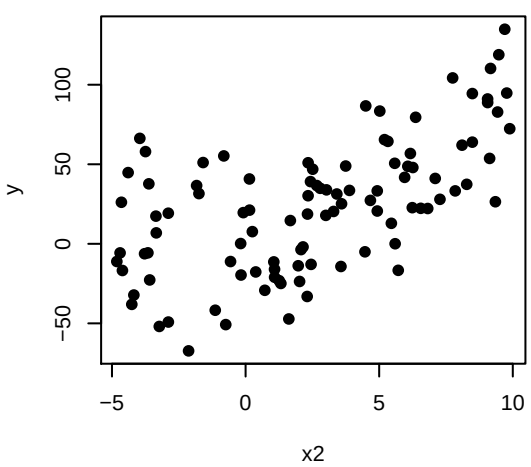
- Another useful way of plotting residuals is to plot them versus one of the explanatory variables
- For example, suppose we fit a linear model with both x_1 and x_2 as explanatory variables, but that $\mathrm{E}(y) \propto x_2^2$:

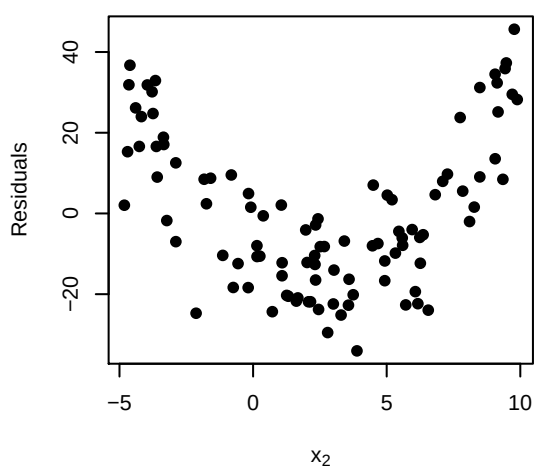


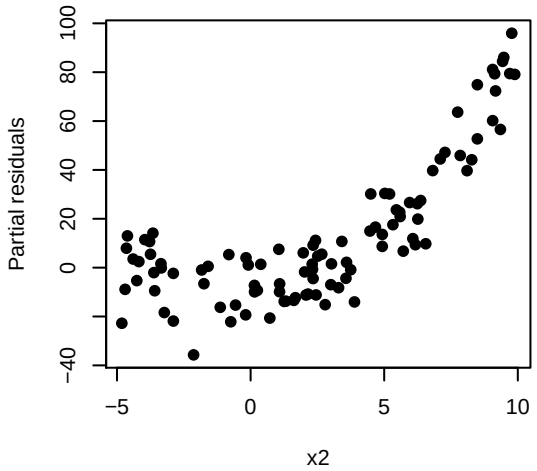


Partial residuals

- A similar idea is to fit the model without a variable, then plot the residuals of that model versus the explanatory variable that was left out
- These residuals are called the *partial residuals*
- Partial residual plots allow you to look at the relationship between the outcome and the explanatory variable while adjusting for the other explanatory variables:

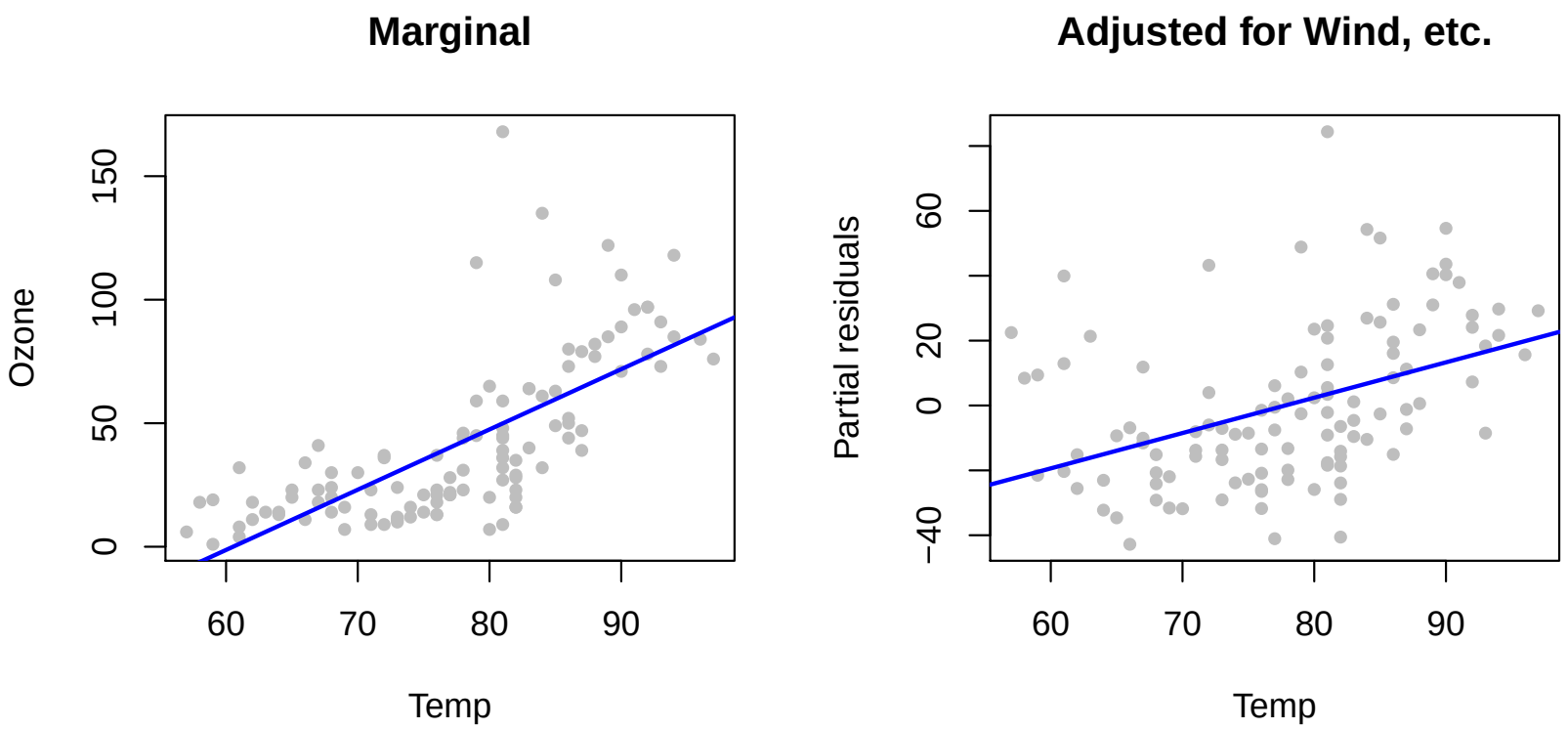






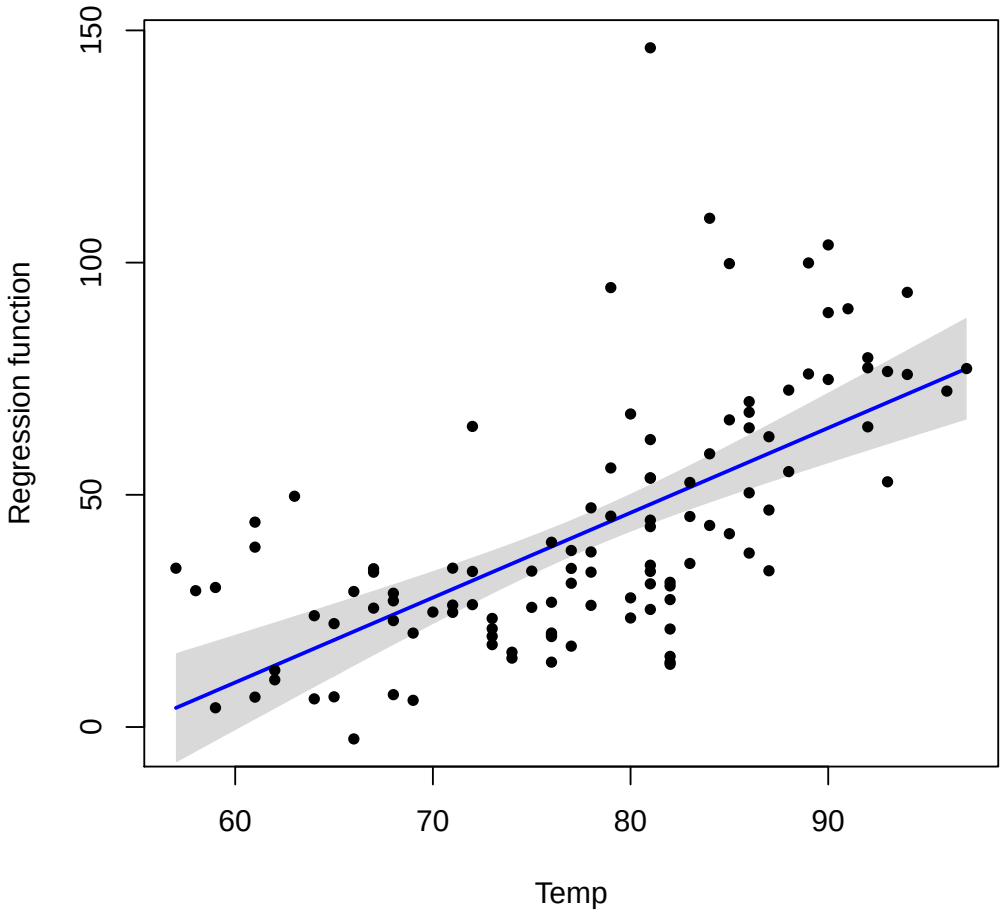
Partial residuals (cont'd)

One of the things that makes partial residuals useful to look at is that the simple linear regression fit to the partial residuals produces a line with exactly the same slope as the multiple regression model:



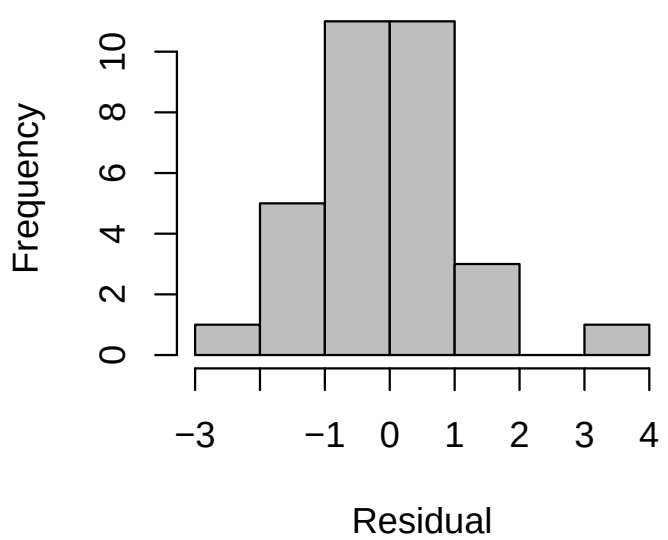
Partial residuals (cont'd)

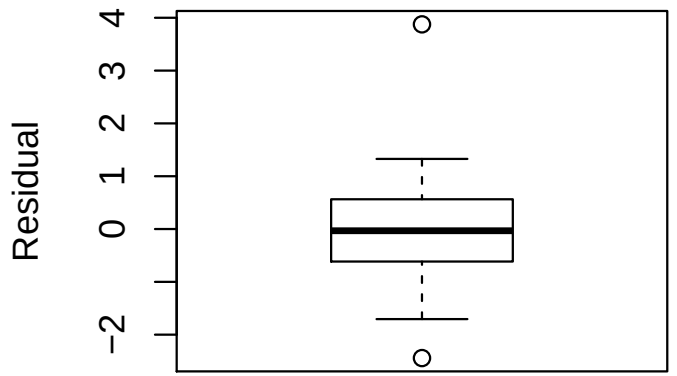
A related idea is to hold the rest of the variables fixed at their mean (or median) and examine the change in $\widehat{\mathrm{E}}(y)$ as the explanatory variable is changed:



Histogram and box plots

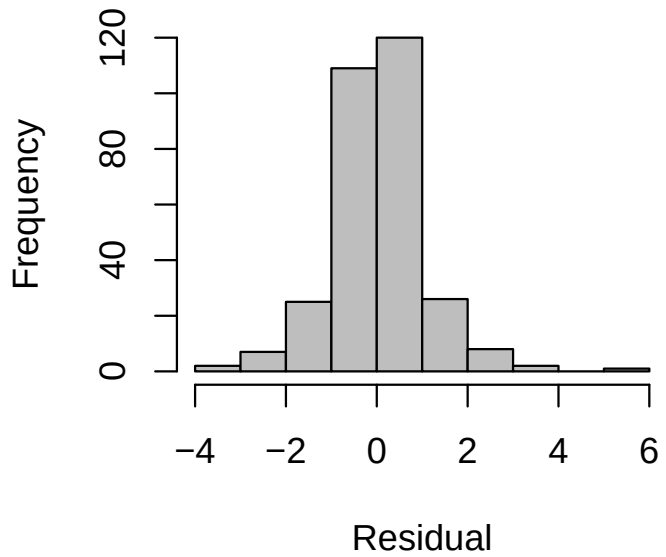
- Other plots attempt to check the assumption of normality
- One possibility is to plot the residuals with a histogram or box plot
- From the `alcohol` data set, using the `Gastric` $*$ `Male` model:

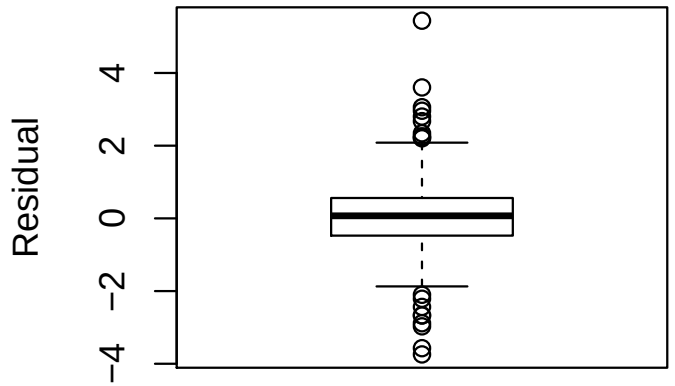




Histogram and box plots – shortcomings

- However, it can be hard to tell from a histogram or box plot whether the tails of the distribution are thicker than you would expect from the normal distribution or not
- For example:



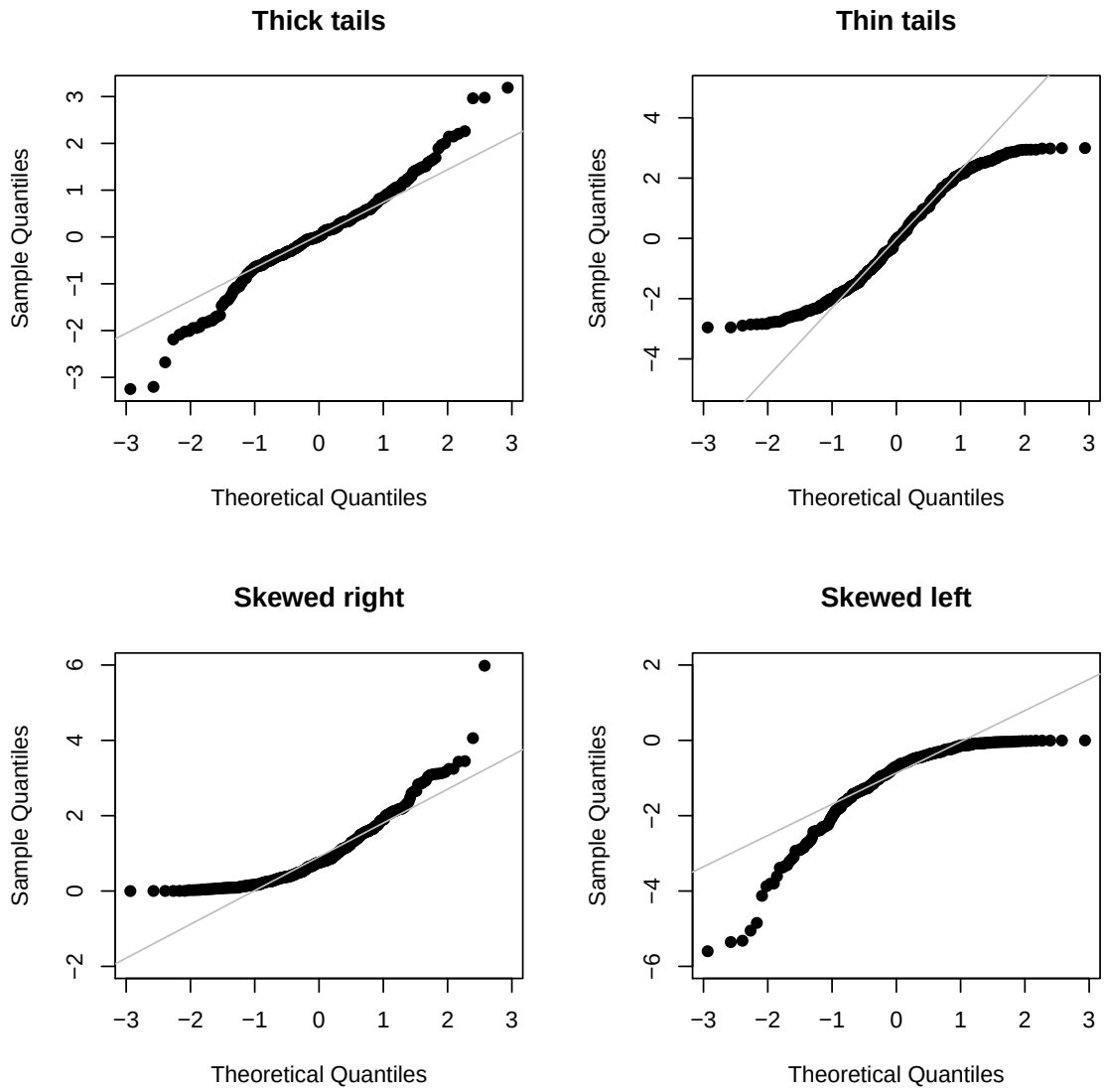


is this a problem?

Q-Q plots

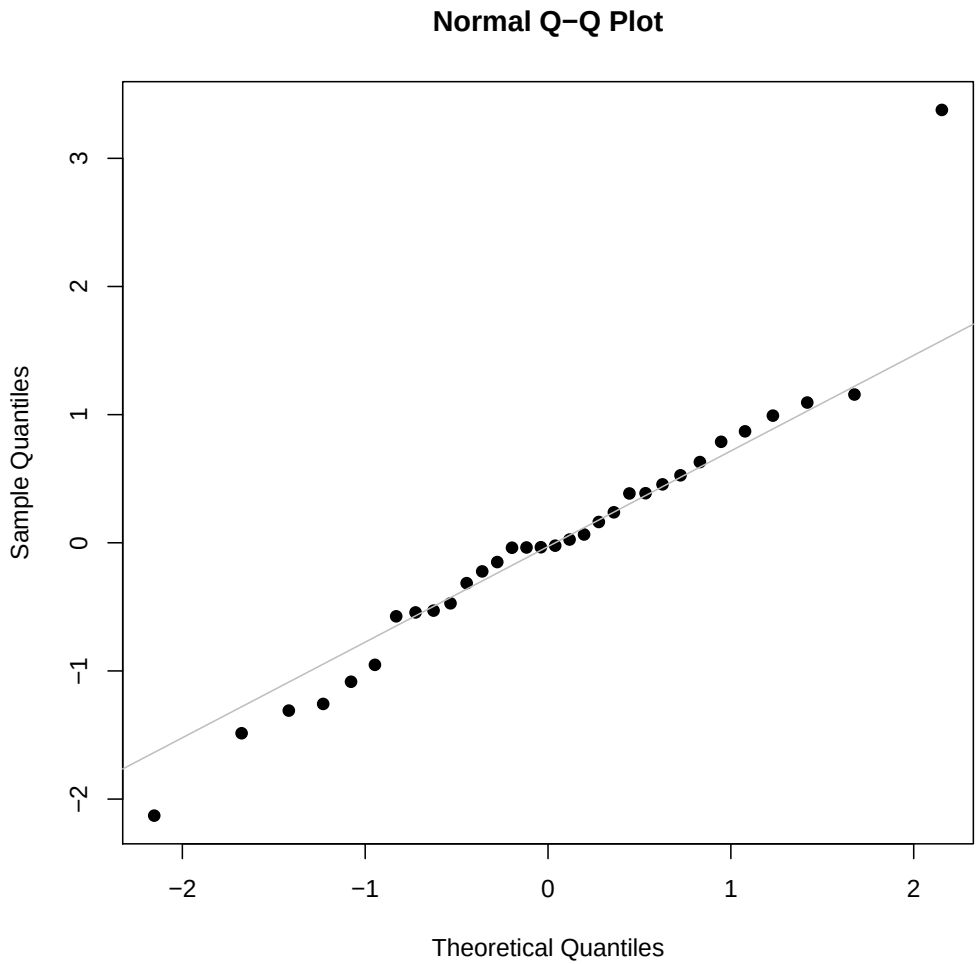
- An alternative plot that has been developed is to plot the quantiles of the normal distribution, $\left\{\Phi^{-1}\left(\frac{i}{n+1}\right)\right\}$, versus the actual quantiles
- If the data follow a normal distribution, this plot should be a straight line
- If not, various divergences from a straight line are possible

Typical patterns in Q-Q plots



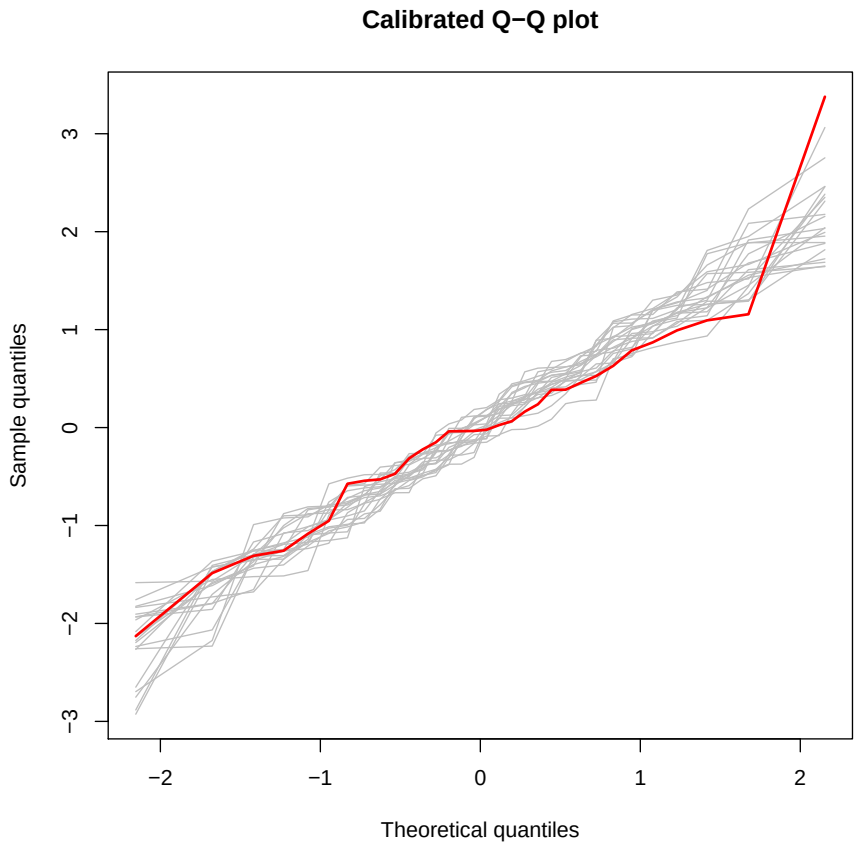
Q-Q plot for the `alcohol` data

From the `alcohol` data set, using the `Gastric` $*$ `Male` model:



Calibrated Q-Q plot

On the surface, this may seem to indicate a problem with thick tails, but this amount of divergence from a straight line is not uncommon, even with data coming from the normal distribution:



Outlier complications

- Whether or not a point is an outlier, however, is a bit tricky in regression
- Outlying points may not appear to be outliers simply because they affect the fit of the model so much that $\hat{\mu}_i$ ends up being close to y_i
- One way to account for this is to recognize that the residuals (note: residuals, not random errors) do not have equal variance:

$$\text{Var}(\mathbf{r}) = \sigma^2(\mathbf{I} - \mathbf{H})$$

- For example, in our alcohol model that we have been using, $(\mathbf{I} - \mathbf{H})_{2,2} = 0.94$, while $(\mathbf{I} - \mathbf{H})_{31,31} = 0.46$

Standardized residuals

- Thus, comparing $\{r_i\}$ to a normal distribution is misleading; we should be comparing

$$s_i = \frac{r_i}{\sqrt{\hat{\sigma}^2(\mathbf{I} - \mathbf{H})_{ii}}}$$

  to a normal distribution
- These residuals (which should be homoskedastic) are called the *standardized residuals*, or sometimes the *studentized residuals*

Deleted residuals

- An alternative approach is fit the model without point i, then see how far off y_i is from its predicted value $\hat{\mu}_{i(-i)}$
  - Note: this is now a real prediction, because we didn't use y_i to fit the model
  - Note: the notation $(-i)$ in the subscript refers to the fact that the estimate is coming from a model without point i in it
- A rather neat algebraic result is that we do not actually have to refit the model to obtain $\hat{\mu}_{i(-i)}$:

$$d_i = y_i - \hat{\mu}_{i(-i)} = \frac{r_i}{1 - H_{ii}}$$

- These residuals are sometimes called the "deleted" residuals or the "jackknifed" residuals

Studentized deleted residuals

- These two approaches can be combined to yield a type of residual called the *studentized deleted residuals*:

$$t_i = \frac{d_i}{\sqrt{\widehat{\mathrm{Var}}(d_i)}}$$

  where $\widehat{\mathrm{Var}}(d_i)$ is the estimated variance of d_i (not hard to derive, but bulky)
- The notation is intentional; under the model assumptions,

$$t_i \sim t_{n-p-1}$$

  although the $\{t_i\}$ are not independent
- These residuals are also sometimes called the "studentized residuals", which can be confusing, as the standardized residuals are also sometimes called the studentized residuals

Should we always studentize our residuals?

- One can make an argument that studentized deleted residuals should always be used instead of standard residuals in any sort of diagnostic plot, on the logic that the $\{t_i\}$ actually follow a standard distribution, while the $\{r_i\}$ do not
- This is certainly true in principle, although in practice, most residuals do not change much with studentization, and looking at the $\{r_i\}$ is usually sufficient to observe bulk trends like heterskedasticity
- However, studentized residuals should definitely be used when assessing outliers

Leverage

- A related but separate issue is the fact that some points affect the fit of the model much more than other points
- For a variety of reasons, the quantity H_{ii} serves as a useful quantifier of the influence of the ith point, and is referred to as the *leverage* of point i
- Points have high leverage when they are outliers with respect to the explanatory variables – this is a separate issue from being an outlier with respect to the outcome variable

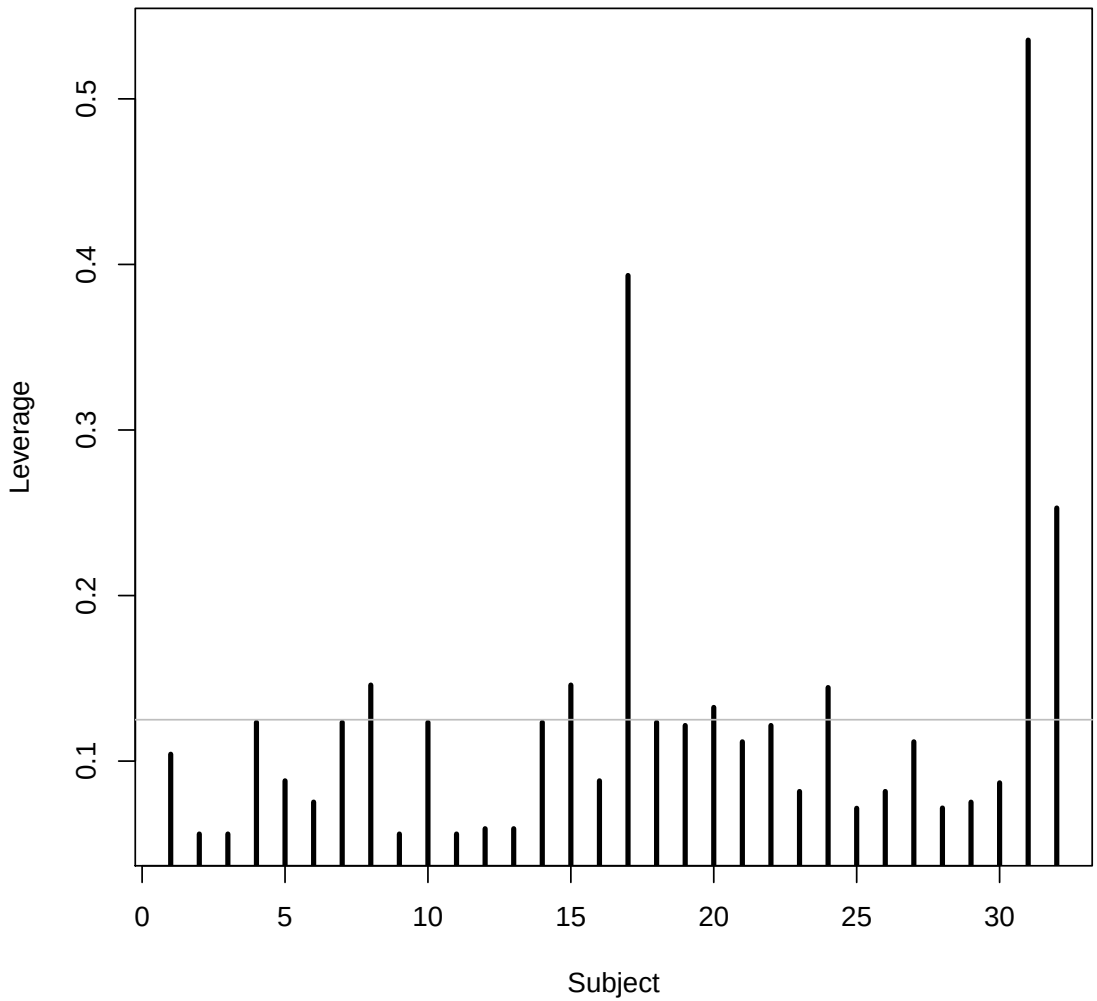
Leverage: example

- We already know that $\mathrm{tr}(\mathbf{H}) = p$, so the average leverage should be p/n
- Consider, however, two subjects in our alcohol metabolism data set and their influence in the `Gastric` $*$ `Male` model:

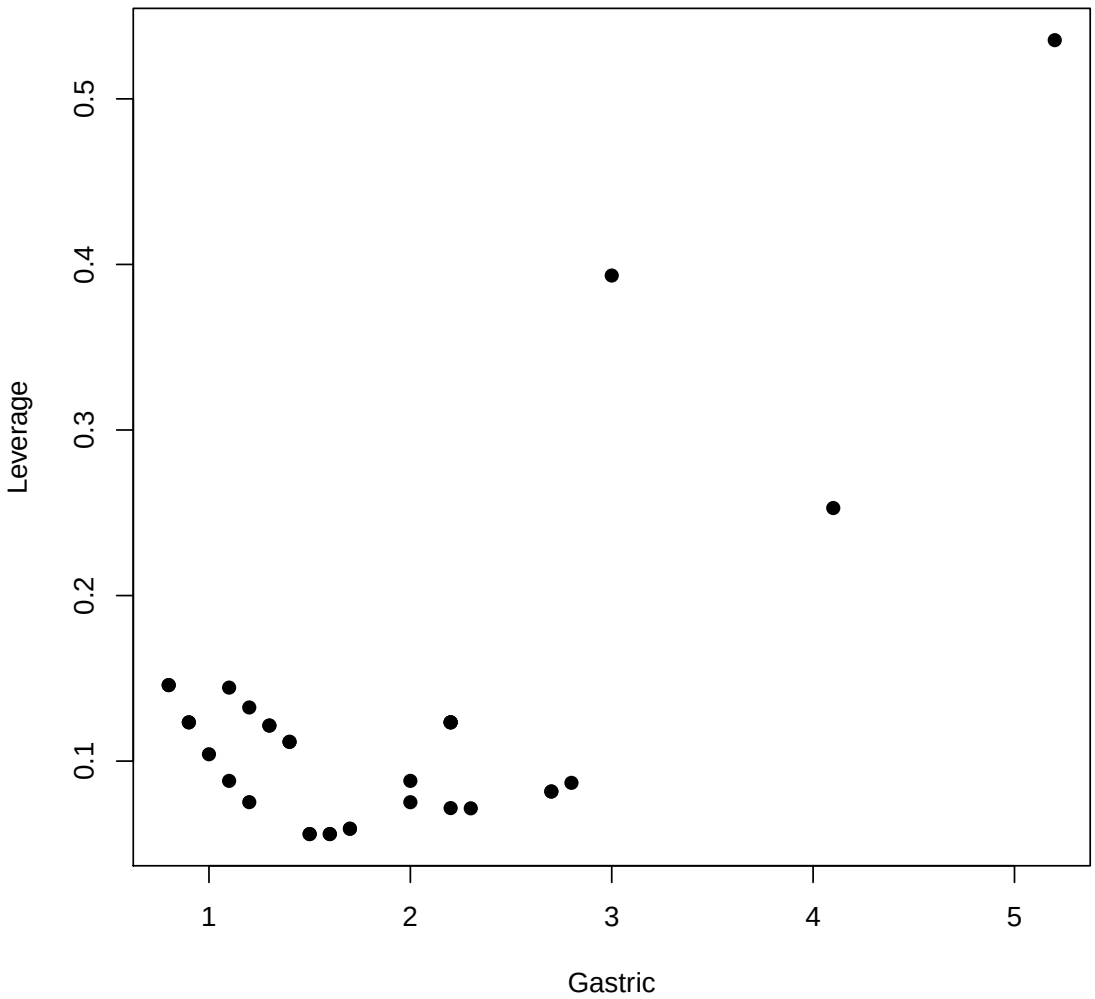
$$H_{2,2} = 0.06 \qquad H_{31,31} = 0.54,$$

  while $p/n = 4/32 = 0.125$
- The reason is that subject 2 has `Gastric` $= 1.6$, very close to the mean `Gastric` level of 1.55 in females, while subject 31 has the largest value of `Gastric` in the sample, 5.2 (two and a half standard deviations above the mean for males)

Leverage plot

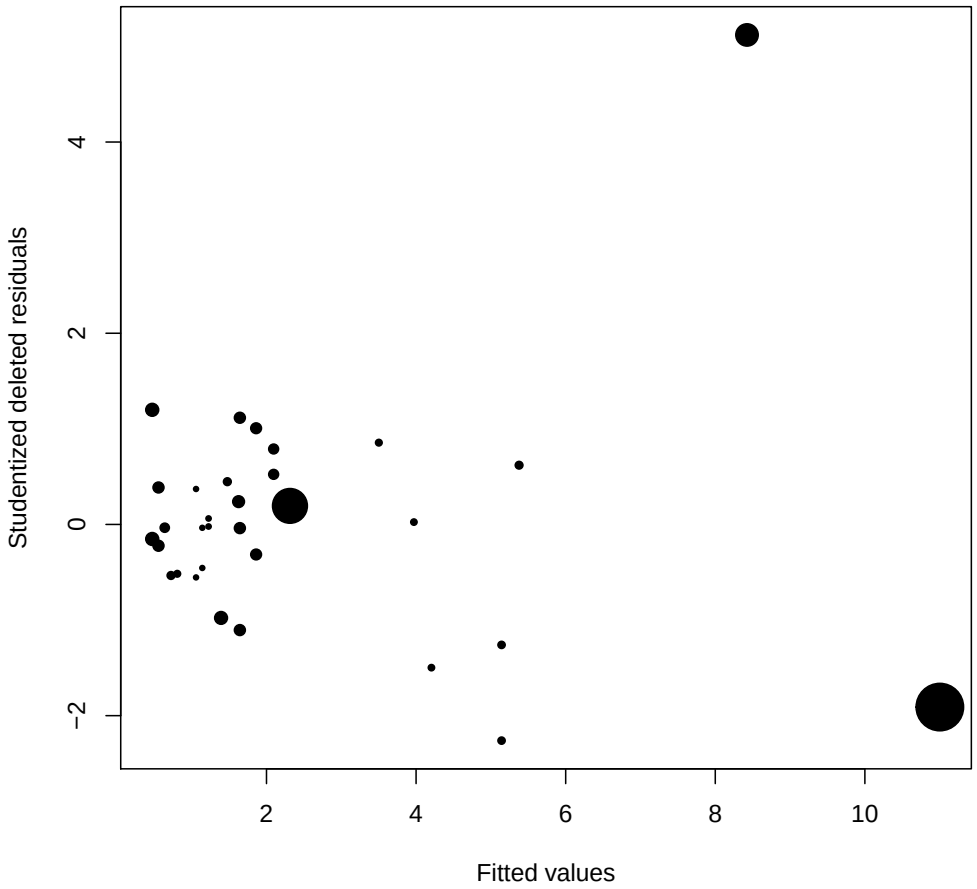


Leverage vs. Gastric



Proportional influence plot

Another useful plot is a *proportional influence plot*:

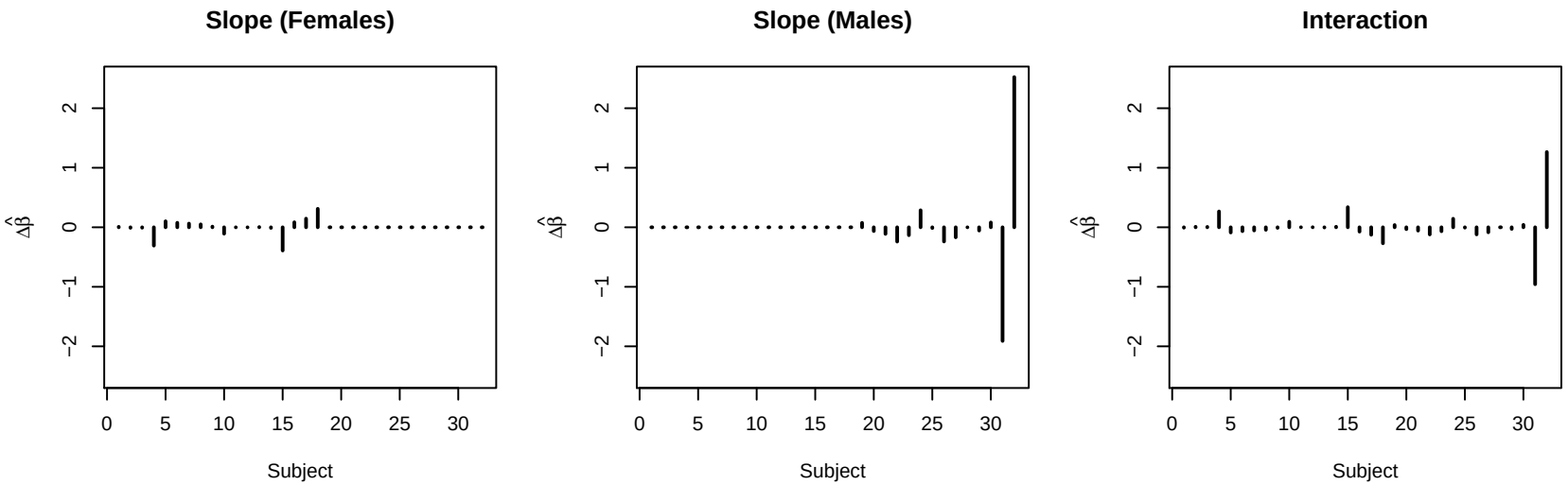


Leverage vs. residual

- As has been mentioned, leverage and residual are separate factors to consider:
  - Points with low leverage and small residual are fairly inconsequential to the fit
  - Points with low leverage and large residual do not exert a large influence on the fit of the model
  - Points with high leverage and small residual do not change the fit of the model greatly
  - Points with high leverage and high residual, on the other hand, can drastically change the model
- As you might imagine, it is desirable to have a single summary which combines influence and residual

$\Delta\hat{\beta}$ plots

One approach is to directly measure the change in the estimate of a regression parameter upon refitting a model without the ith observation, the so-called "delta-beta" plot:



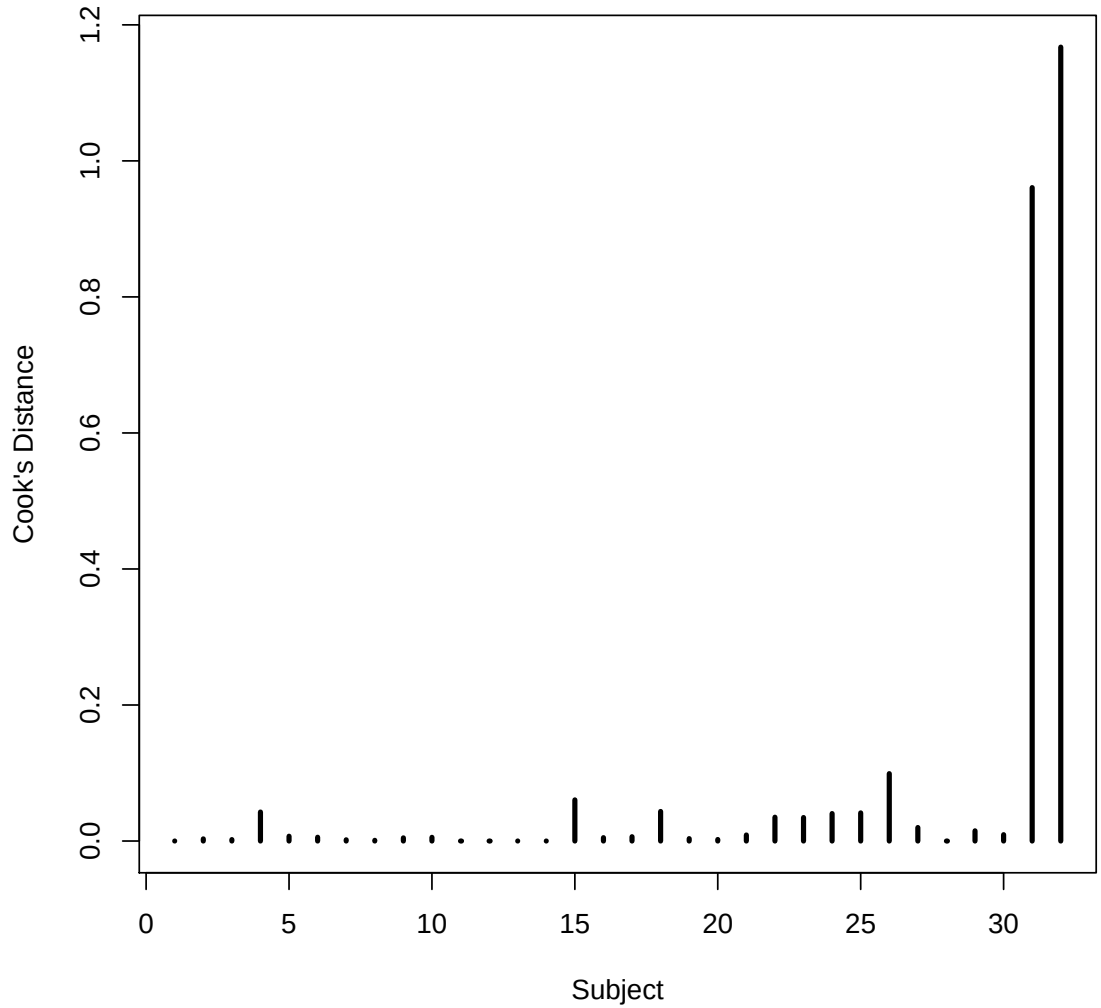
Cook's distance

- A further attempt to reduce the notion of an influential point down to just a single number was proposed by Cook in 1977:

$$D_i = \frac{\sum_j (\hat{\mu}_j - \hat{\mu}_{j(-i)})^2}{p\hat{\sigma}^2}$$

- D_i therefore measures the overall distance between the original fitted values and the fitted values you would obtain by removing the ith observation from the data set
- This measurement is referred to as *Cook's distance*

Cook's distance: Alcohol data



Tests

- Many of the diagnostics we have talked about today have tests associated with them:
  - Tests for normality
  - Tests for constancy of variance
  - Tests for outliers
  - "Goodness-of-fit" tests
- We don't really have time to talk about them in detail, but they can be helpful in terms of supplying an objective measure of whether there seems to be a problem in terms of model assumptions

Remarks on diagnostic tests

Additional remarks:

- These tests are not always helpful – just because a test of normality is not significant does not mean the data are normally distributed
- The notion of significance and a $p < .05$ cutoff is a bit dubious in such tests
- Tests provide little to no insight about the model and why and where the problems might lie, or how you might remedy them