Diagnostics for linear models

Analysis of Ecological and Environmental Data

QERM 514

Mark Scheuerell

13 April 2020

Goals for today

- Recognize that diagnostic checks are necessary for any model
- Learn how to check for constant variance, normally distributed errors, and autocorrelation
- Learn how to check for outlying or influential observations

Model diagnostics

We have seen how to fit models, estimate parameters with uncertainty, and conduct hypothesis tests

All of these rely on a number of assumptions about

- our model (its structure is correct)
- the errors (independent, equal variance, normally distributed)
- observations and predictors (no undue influence)

Model structure

Our focus here is on linear models, and we saw previously that we can use linear models to approximate nonlinear functions

The specific form of the model should reflect our understanding of the system and any particular hypotheses we’d like to test

Checking error assumptions

So far our models have assumed the errors to be *independent and identically distributed* (IID)

What exactly does this mean?

Checking error assumptions

Constant variance

Let’s begin with the notion of “identically distributed”, which suggests no change in the variance across the model space

For example, if our errors are assumed to be normally distributed, such that

$$\epsilon_i \sim \text{N}(0, \sigma^2) ~ \Rightarrow ~ \boldsymbol{\epsilon} \sim \text{MVN}(\mathbf{0}, \sigma^2 \mathbf{I})$$

then we expect no difference in σ^2 among any of the ϵ_i.

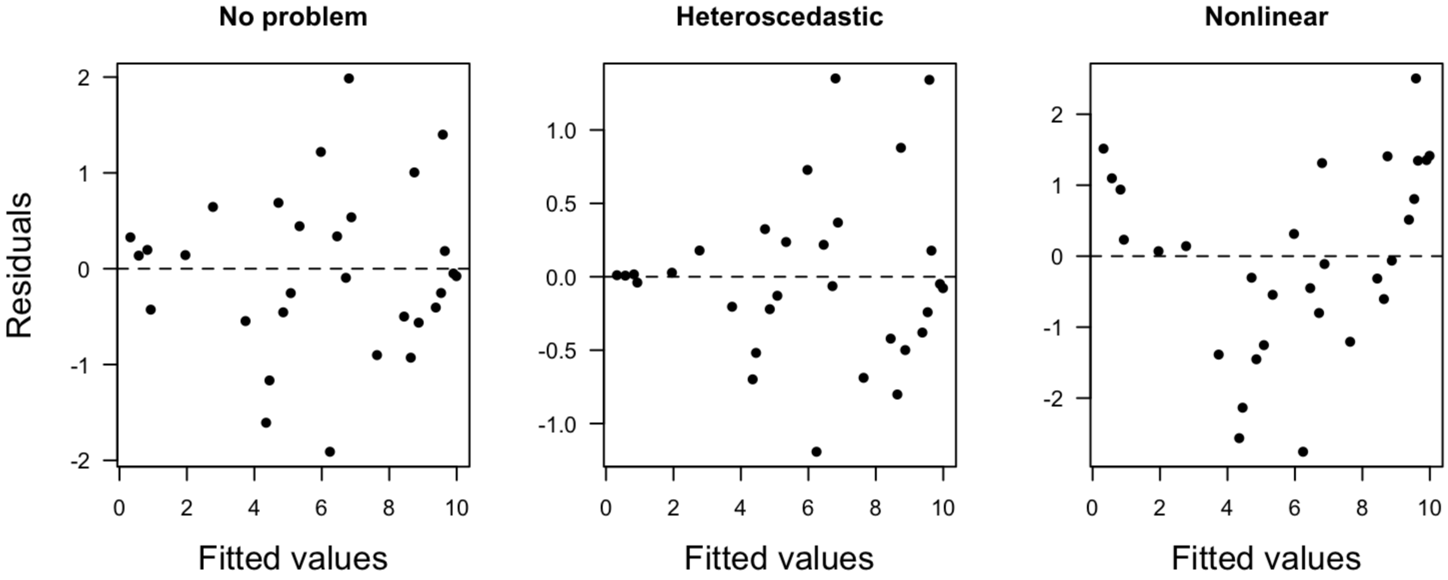
Checking error assumptions

Constant variance

To check this assumption, we can plot our estimates $\hat{\epsilon}_i = e_i = y - \hat{y}$ against our fitted values $\hat{y}_i$ and look for any patterns

Checking error assumptions

Constant variance



Checking error assumptions

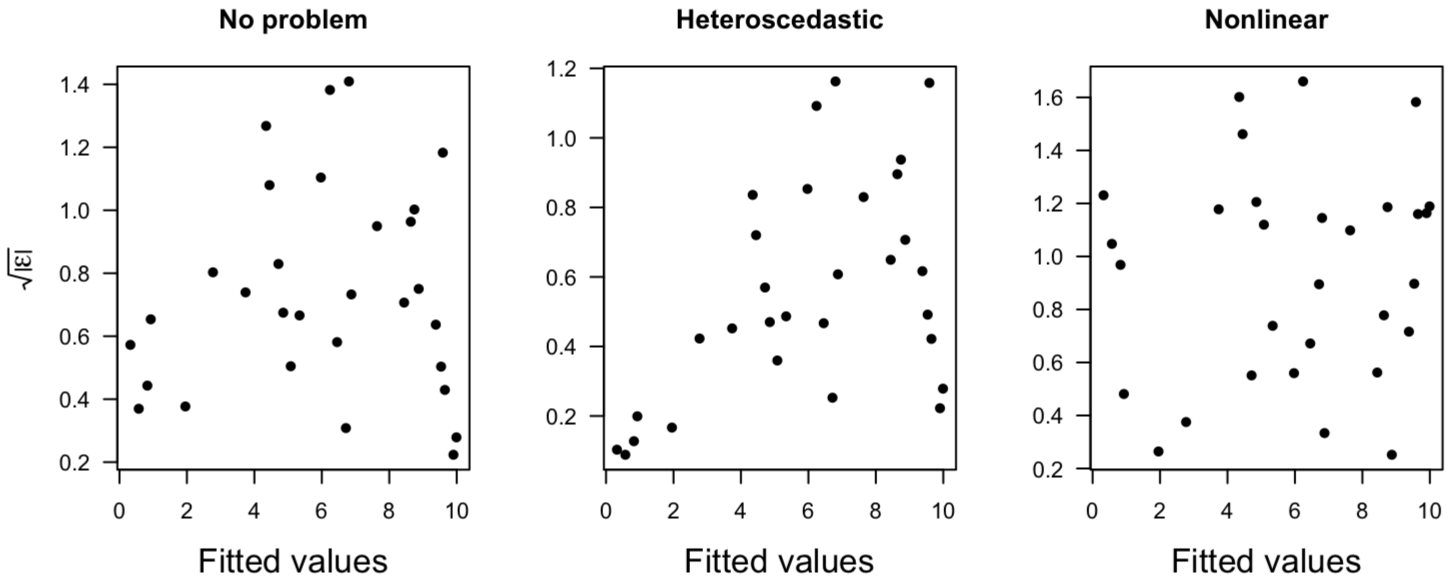
Constant variance

For a finer resolution, we can also plot $\sqrt{|\hat{\epsilon}_i|}$ against our fitted values $\hat{y}_i$ and look for any patterns

The distribution of $|\hat{\epsilon}_i|$ is a skewed half-normal on the positive interval; the square-root transformation makes them less skewed

Checking error assumptions

Constant variance



Checking error assumptions

Constant variance

We can formally test the assumption of homogeneous variance via *Levene's Test*, which compares the absolute values of the residuals among j groups of data

$$Z_{ij} = \left| y_{ij} - \hat{y}_j \right|$$

Levene's test is a one-way ANOVA of the residuals

Checking error assumptions

Constant variance

The statistic for *Levene's Test* is

$$W = \frac{(n-k)}{(k-1)} \cdot \frac{\sum_{j=1}^{k} n_j \left(Z_j - \bar{Z}\right)^2}{\sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(Z_{ij} - \bar{Z}_i\right)^2}$$

The test statistic W is approximately F-distributed with $k-1$ and $N-k$ degrees of freedom

Checking error assumptions

Levene’s Test is easy to compute in **R**

```r
## split residuals (ee) into 2 groups
g1 <- ee[ee < median(ee)]
g2 <- ee[ee > median(ee)]
## Levene's Test
var.test(g1, g2)
```

```
##
##  F test to compare two variances
##
## data:  g1 and g2
## F = 0.90486, num df = 14, denom df = 14, p-value = 0.8543
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3037877 2.6951999
## sample estimates:
## ratio of variances
##          0.9048584
```

Checking error assumptions

Constant variance

What can we do if we find evidence of heteroscedasticity?

Try a transformation or weighted least squares, which we will see later this week

Checking error assumptions

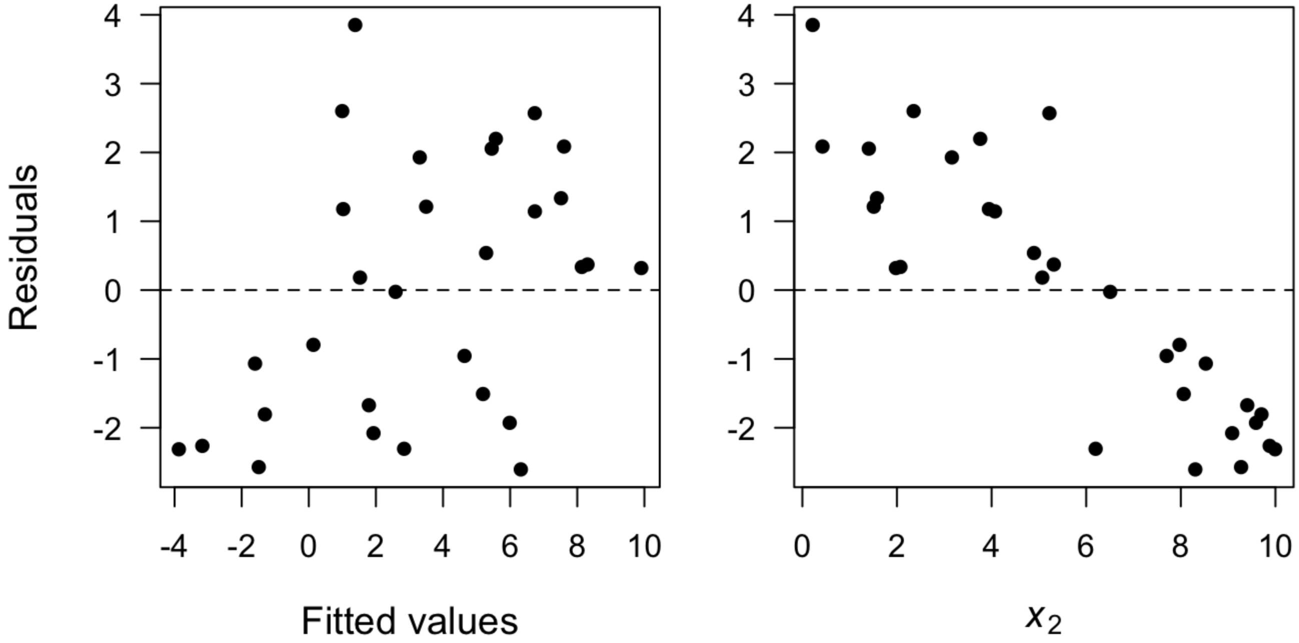
Residuals vs other predictors

We can also plot the residuals against any potential predictors that were *not* included in the model

If we see a (linear) pattern, then consider including that predictor in a new model

Checking error assumptions

Residuals vs other predictors for $y_i = \alpha + \beta x_{1,i} + e_i$



Checking error assumptions

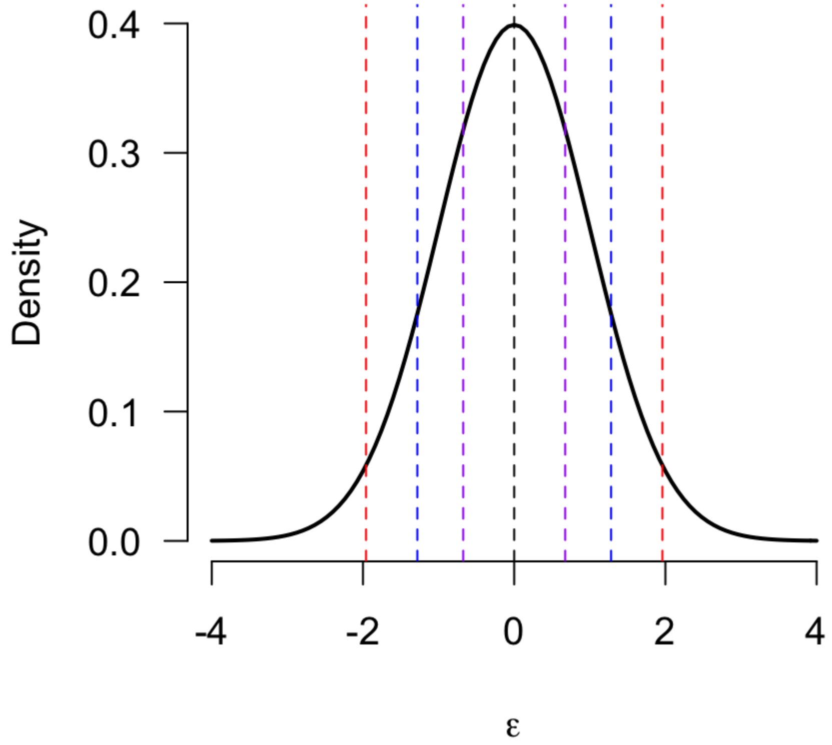
Normality

We seek a method for assessing whether our residuals are indeed normally distributed

The easiest way is via a so-called Q-Q plot (for quantile-quantile)

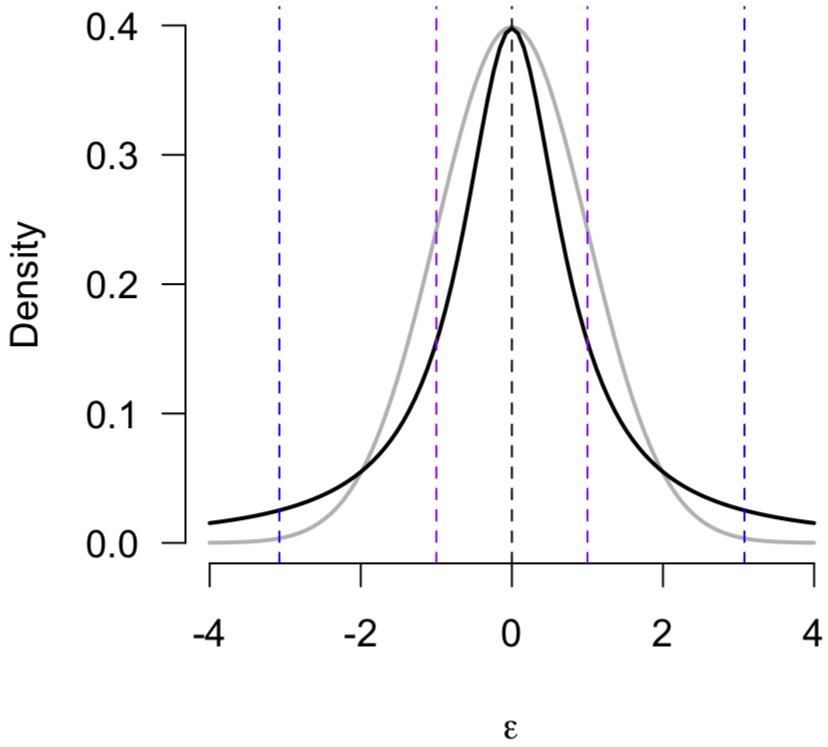
Checking error assumptions

Expected quantiles for $\epsilon \sim \mathrm{N}(0, 1)$



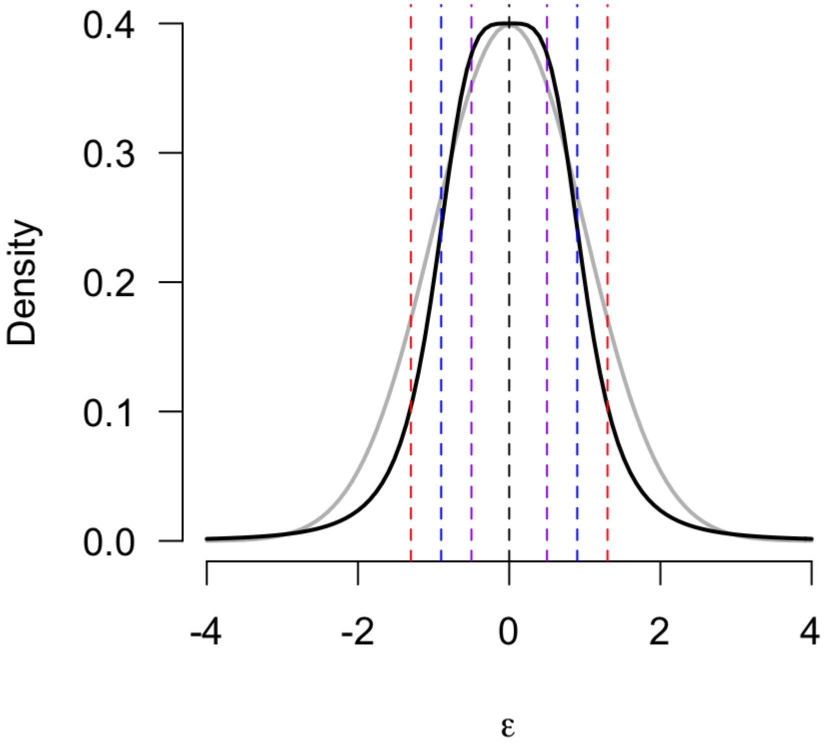
Checking error assumptions

Heavy-tailed (*leptokurtic*)



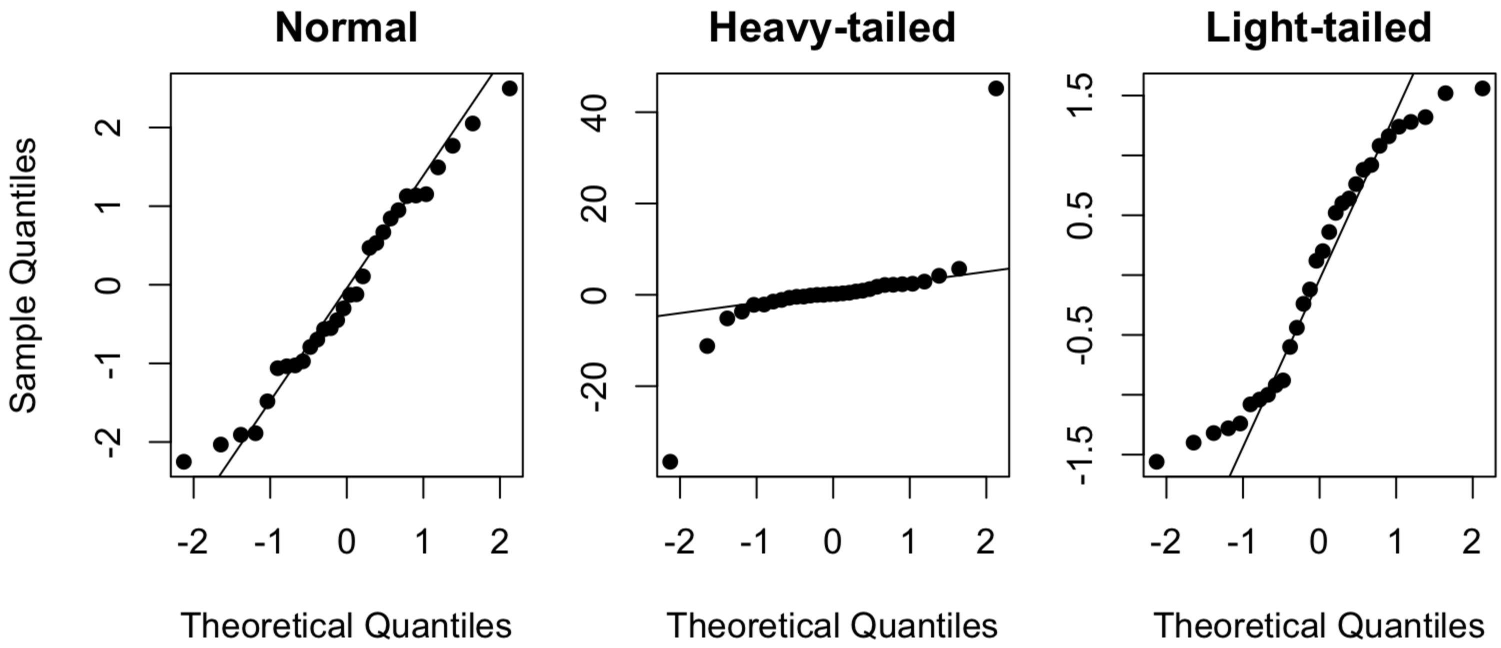
Checking error assumptions

Short-tailed (*platykurtic*)



Checking error assumptions

Q-Q plots via `qqnorm(x)` in **R**



Correlated errors

One component of *IID* errors is “independent”

This means we expect no correlation among any of the errors

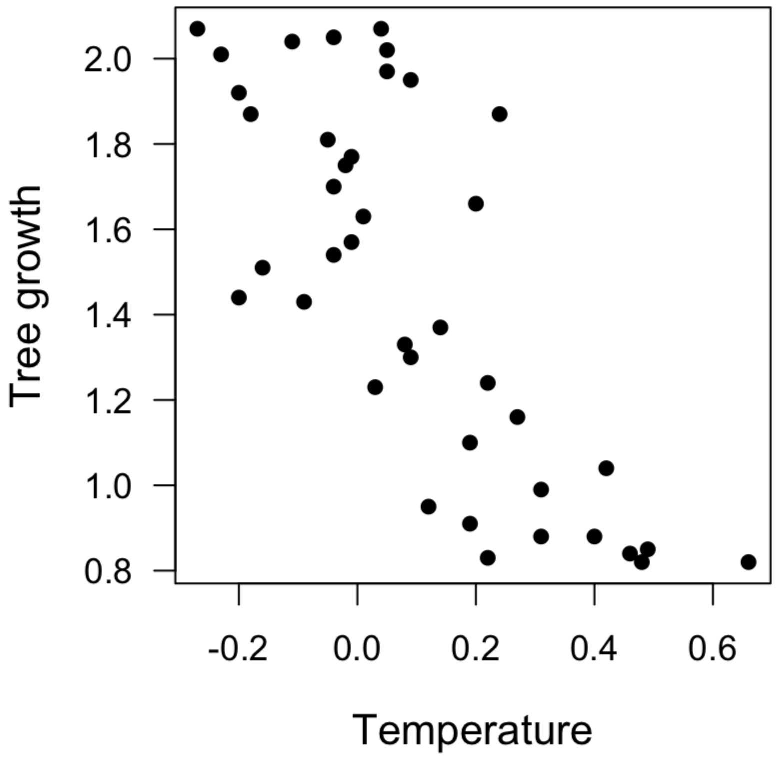
Correlated errors

We might expect to find correlated errors when working with

- Temporal data
- Spatial data
- Blocked data

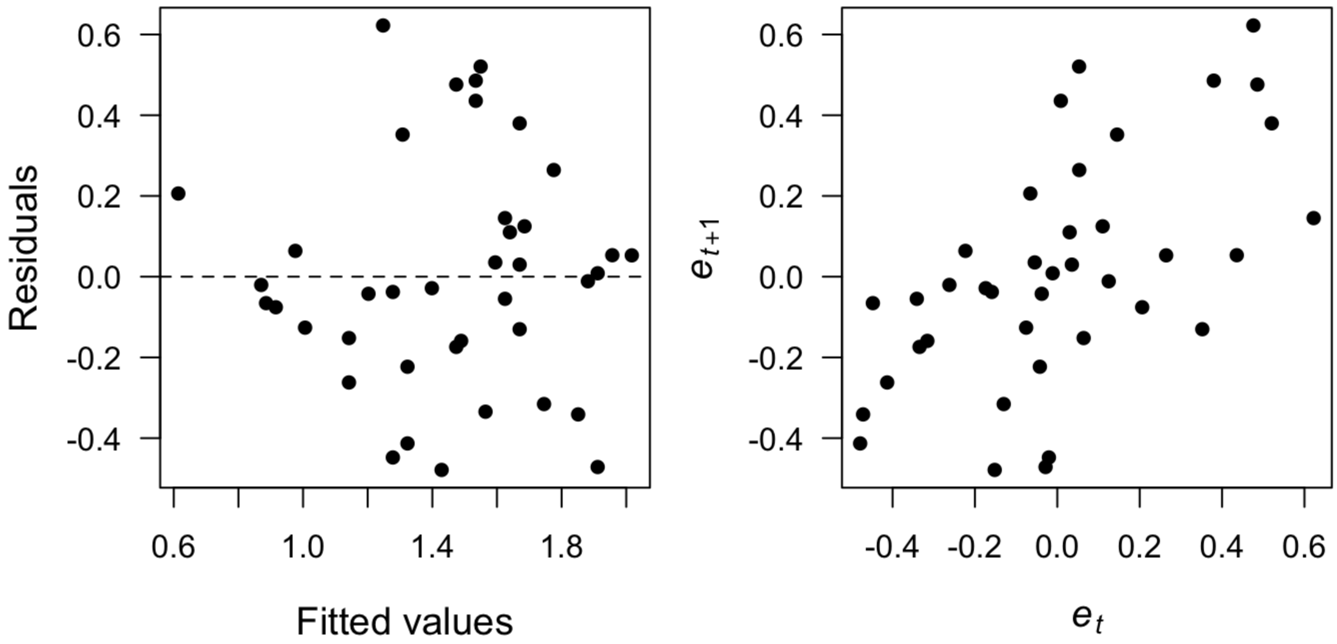
Correlated errors

Consider a model for tree growth as a function of temperature



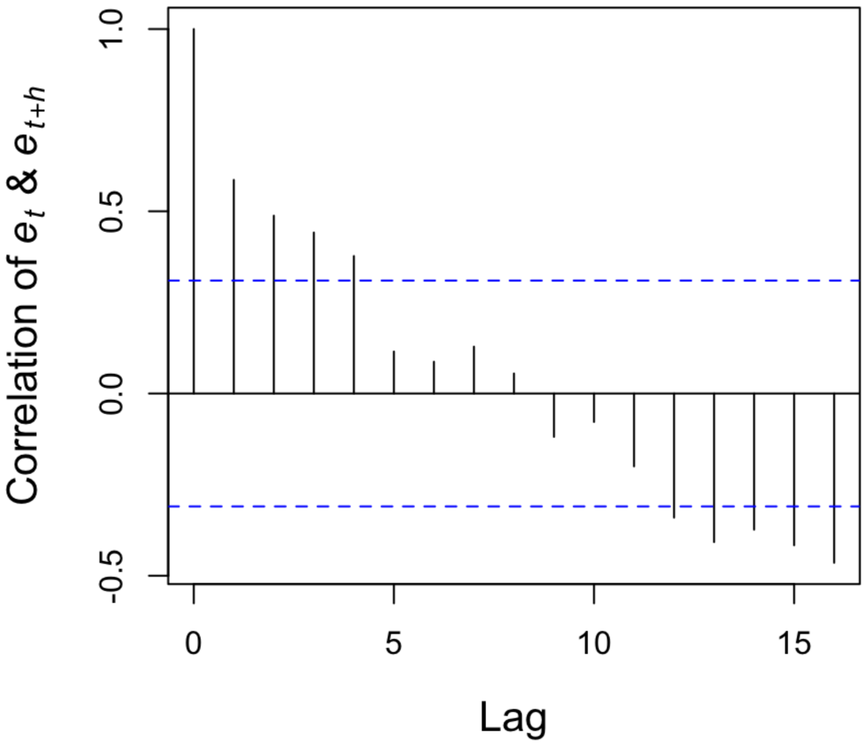
Correlated errors

Closer examination of the residuals reveals a problem



Correlated errors

We can estimate the *autocorrelation function* in **R** with `acf()`



QUESTIONS

Unusual observations

Outliers

It is often the case that one or more data points do not fit our model well

We refer to these as *outliers*

Unusual observations

Influence

Some outliers affect the fit of the model

We refer to these as *influential* observations

Unusual observations

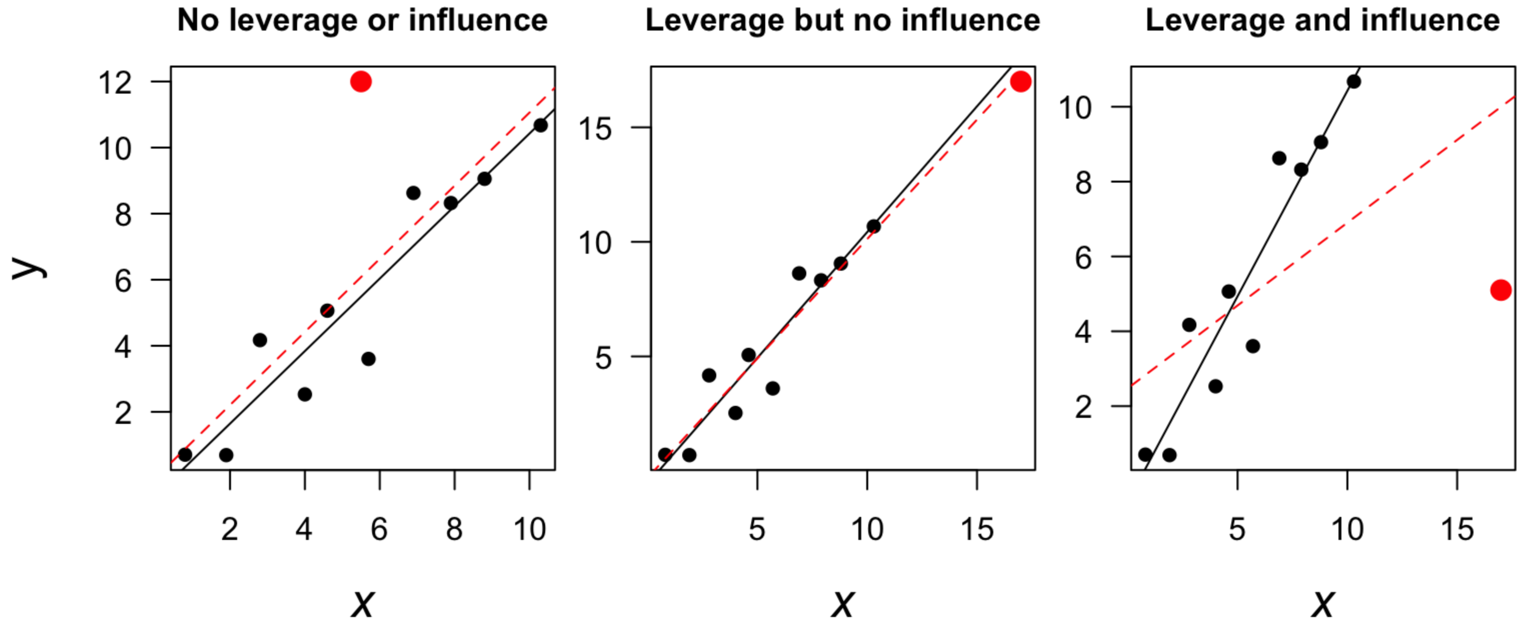
Leverage points

Leverage points are extreme in the predictor (X) space

They may or may not affect model fit

Unusual observations

Examples



Unusual observations

Identifying leverage points

Remember the “hat matrix” $(\mathbf{H})$?

The values along the diagonal $h_i = \mathbf{H}_{ii}$ are the leverages

Unusual observations

Identifying leverage points

Also recall that

$$\text{Var}(\hat{\epsilon}_i) = \sigma^2(1 - h_i)$$

Large h_i lead to small variances of ϵ_i & hence $\hat{y}_i$ tends to y_i

Unusual observations

Identifying leverage points

$\mathbf{H}$ has dimensions $n \times n$ and $\text{trace}(\mathbf{H}) = \sum_{i=1}^{n} h_i = k$

Thus, on average we should expect that $\bar{h}_i = \frac{k}{n}$

Any $h_i > 2\frac{k}{n}$ deserve closer inspection

Unusual observations

Identifying leverage points

We can easily compute the h_i in **R** via the function `hatvalues()`

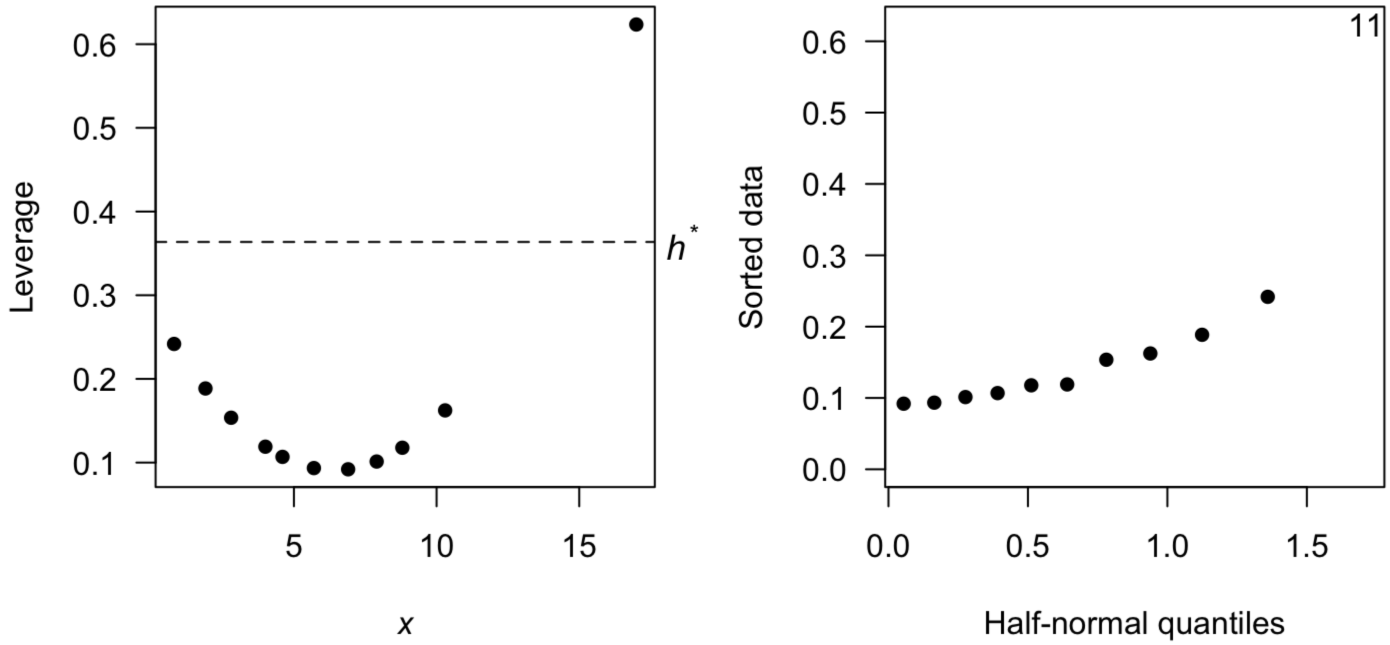
```
## leverages of points in middle plot on slide 30
hv <- hatvalues(m2)
## threshold value for h_i ~= 0.36
th <- 2 * (2 / length(hv))
## are any h_i > Eh?
hv > th
```

```
##     1     2     3     4     5     6     7     8     9    10    11 
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
```

Unusual observations

Identifying leverage points

We can also identify high leverage via a half-normal plot (R)



Using leverage to standardize residuals

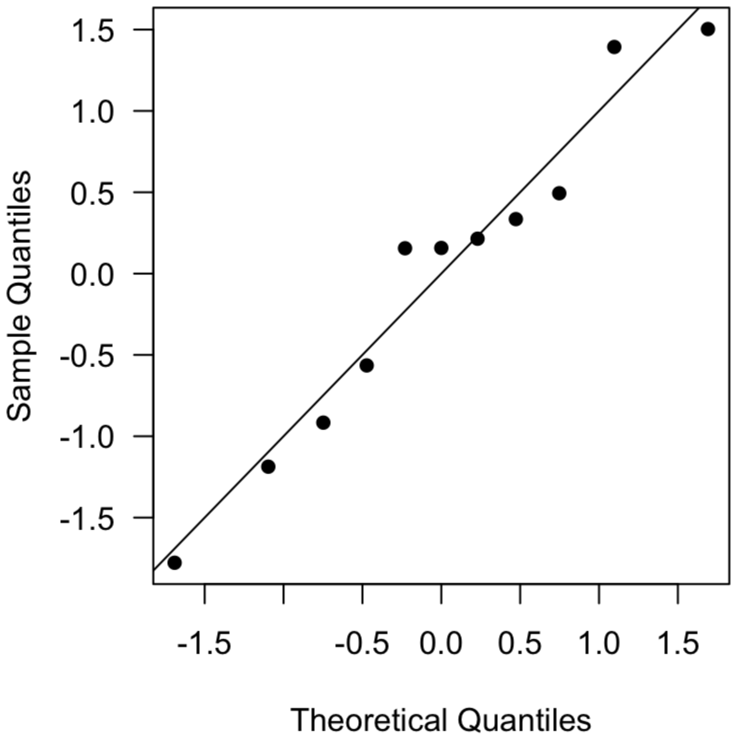
We can use the leverages to scale the residuals so their variance is 1

$$r_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1 - h_i}}$$

Doing so allows for easy examination via Q-Q plots as values should lie on the 1:1 line

Using leverage to standardize residuals

Standardized residuals from the high leverage example



Unusual observations

Identifying outliers

One way to detect outliers is to estimate n different models where we exclude one data point from each model

More formally we have

$$\hat{\mathbf{y}}_{(i)} = \mathbf{X}_{(i)}\hat{\boldsymbol{\beta}}_{(i)}$$

where (i) indicates that the i^{th} datum has been omitted

If $y_i - \hat{y}_{(i)}$ is large, then observation i is an outlier

Unusual observations

Identifying outliers

To evaluate the size of particular outlier we need to scale the residuals

This is similar to scaling a parameter estimate by its standard deviation to test model hypotheses, with

$$t_i = \frac{\beta_i}{\text{SE}\left(\beta_i\right)}$$

and we compare it to a t-distribution with $n - k$ degrees of freedom

Unusual observations

Identifying outliers

It turns out that the variance of the difference $y_i - \hat{y}_{(i)}$ is just like that for a prediction interval

$$\widehat{\text{Var}}\left(y_i - \hat{y}_{(i)}\right) = \hat{\sigma}^2_{(i)}\left(1 + \mathbf{X}_i^\top\left(\mathbf{X}_{(i)}^\top\mathbf{X}_{(i)}\right)^{-1}\mathbf{X}_i\right)$$

Unusual observations

Identifying outliers

We can now compute the “studentized” (scaled) residuals as

$$t_i = \frac{y_i - \hat{y}_{(i)}}{\hat{\sigma}_{(i)} \sqrt{1 + \mathbf{X}_i^\top \left(\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)} \right)^{-1} \mathbf{X}_i}}$$

which are distributed as a t distribution with $n - k - 1$ df

Unusual observations

Identifying outliers

There is an easer way to do this without fitting n different models, where

$$t_i = \frac{y_i - \hat{y}_{(i)}}{\hat{\sigma}_{(i)}\sqrt{1 - h_i}} = e_i \sqrt{\frac{n - k - 1}{n - k - e_i^2}}$$

and r_i is the residual for the i^{th} case based on a model that includes *all* of the data

Unusual observations

Identifying outliers

Some points to consider

- Two or more outliers next to each other can hide each other
- An outlier in one model may not be an outlier in another
- The error distribution may not be normal and so larger residuals may be expected
- Individual outliers are usually much less of a problem in larger datasets

Unusual observations

Identifying outliers

What can be done about outliers?

- Check for a data-entry error
- Examine the physical context — why did it happen?
- Exclude the point from the analysis but try reincluding it later if the model is changed
- Consider using “robust regression” (more later)
- Be wary of automatic discarding of outliers

Unusual observations

Influential observations

Influential observations might not be outliers nor have high leverage, but we want to identify them

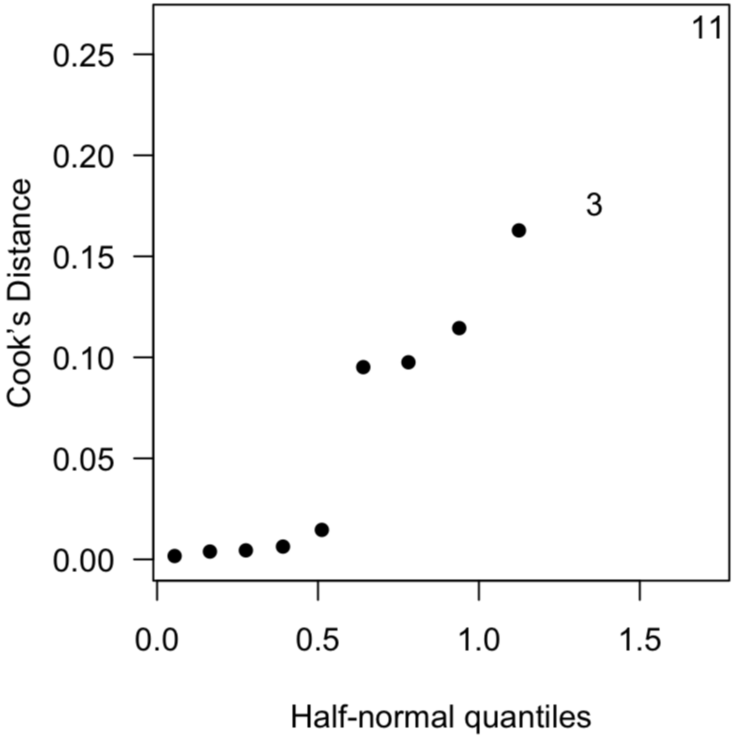
Cook's Distance (D) is a popular choice, where

$$D_i = e_i^2 \frac{1}{k} \left(\frac{h_i}{1 - h_i} \right)$$

D_i scales the errors by their df and leverage

Unusual observations

We can evaulate Cook's D with a half-normal plot



Summary

When fitting linear models via least squares we make several assumptions about our model

Summary

The importance of our assumptions can be ranked as

1. Systematic form of the model

If we get this wrong, explanations & predictions will be off

Summary

The importance of our assumptions can be ranked as

1. Systematic form of the model

2. Independence of errors

Dependence (correlation) among errors means there is less info in the data than the sample size suggests

Summary

The importance of our assumptions can be ranked as

1. Systematic form of the model

2. Independence of errors

3. Non-constant variance

This may affect inference and confidence/prediction intervals

Summary

The importance of our assumptions can be ranked as

1. Systematic form of the model
2. Independence of errors
3. Non-constant variance
4. Normality of errors

This is less of a concern as sample size increases