Prediction: point estimate and prediction interval

Motivation

- Besides model inference, another main use of regression analysis is prediction.
- Suppose we build a model $Y = X\beta + \epsilon$ and obtain estimates $\hat{\beta}$.
- Given a new set of predictors x_0, the predicted response is: $\hat{y}_0 = x_0^\top \hat{\beta}$.
- Besides this point estimate, one may want to further assess the uncertainty in this prediction.
  - Decision makers need more than just a point estimate to make rational choices.
  - If the prediction has a wide "confidence" interval, we need to allow for outcomes far from the point estimate.
  - For example, suppose we need to predict the high water mark of a river. We may need to construct barriers high enough to withstand floods much higher than the predicted maximum when "confidence" interval is wide.

Two types of predictions

- There are two kinds of predictions made from regression models.
  - One is a predicted mean response: $\mathrm{E}(y_0 \mid x_0) = x_0^\top \beta$.
  - Another is a prediction of a future observation: Y_{future}. Intuitively, $y_0 = \mathrm{E}(y_0 \mid x_0) + \epsilon = x_0^\top \beta + \epsilon$ (an additional mean zero random error term.)
- Example: Suppose we have built a regression model that predicts the rental price of houses in a given area based on predictors such as the number of bedrooms and closeness to a major highway. There are two kinds of predictions that can be made for a given x_0:
  - 1. Suppose we ask the question — "What would a house with characteristics x_0 rent for on average?" This selling price is $x_0^\top \beta$ and is again predicted by $x_0^T \hat{\beta}$ but now only the variance in $\hat{\beta}$ needs to be taken into account.
  - 2. Suppose a specific house comes on the market with characteristics x_0. Its rental price will be $x_0^\top \beta + \varepsilon$. Since $E\varepsilon = 0$, the predicted price is $x_0^\top \hat{\beta}$, but in assessing the variance of this prediction, we must include the variance of ε.

Most times, we consider the second case, which is called "prediction of a future value," while the first case, called "prediction of the mean response" is less commonly required.

I. First type of prediction

- The confidence interval (CI) for the mean response for given x_0, i.e., $\mathrm{E}(y_0 \mid x_0) = x_0^\top \beta$ is:

$$\hat{y}_0 \pm t_{n-p}^{(\alpha/2)} \hat{\sigma} \sqrt{x_0^T \left(X^T X\right)^{-1} x_0}$$

  - This is by

$$\frac{(\hat{y}_0 - x_0^\top \beta)/\sqrt{\mathrm{var}(\hat{y}_0 - x_0^\top \beta)}}{\hat{\sigma}/\sigma} \sim t_{n-p},$$

  - where $\hat{y}_0 - x_0^\top \beta = x_0^\top \hat{\beta} - x_0^\top \beta \sim N(0, \mathrm{var}(x_0^\top \hat{\beta}))$ and $\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}/(n-p)$
  - Variance:

$$\mathrm{var}(\hat{y}_0 - x_0^\top \beta) = \mathrm{var}\left(x_0^\top \hat{\beta}\right) = x_0^T \left(X^T X\right)^{-1} x_0 \sigma^2.$$

II. Second type of prediction

- A future observation for $y_0 = x_0^\top \beta + \epsilon$ should be predicted to be $x_0^\top \hat{\beta} + \epsilon$.
- But we do not know the future ϵ but we expect it has mean zero so the point prediction is $\hat{y}_0 = x_0^\top \hat{\beta}$.
- Uncertainty: (similar to the derivation of t-test)

$$\frac{(\hat{y}_0 - y_0)/\sqrt{\mathrm{var}(\hat{y}_0 - y_0)}}{\hat{\sigma}/\sigma} \sim t_{n-p},$$

  - where $\hat{y}_0 - y_0 = x_0^\top \hat{\beta} - (x_0^\top \beta + \epsilon) \sim N(0, \mathrm{var}(x_0^\top \hat{\beta} - \epsilon))$ and $\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}/(n-p)$
  - Variance: It is usually reasonable to assume future ϵ is independent of $\hat{\beta}$ and has variance σ^2. Then

$$\mathrm{var}(\hat{y}_0 - y_0) = \mathrm{var}\left(x_0^\top \hat{\beta} - \epsilon\right) = \mathrm{var}\left(x_0^\top \hat{\beta}\right) + \mathrm{var}(\epsilon) = \sigma^2 \left\{x_0^\top (X^\top X)^{-1} x_0 + 1\right\}.$$

    * Parameter uncertainty + Model uncertainty
- ***Prediction interval***: $100(1-\alpha)\%$ prediction interval for a single future response is:

$$\hat{y}_0 \pm t_{n-p}^{(\alpha/2)} \hat{\sigma} \sqrt{1 + x_0^T \left(X^T X\right)^{-1} x_0}.$$

  - Previous confidence intervals have been for model parameters β. Parameters are considered to be fixed but unknown — they are not random under the Frequentist approach we are using here.

  - However, a future observation is a random variable. For this reason, it is better to call this a "prediction interval" not confidence interval.
- This prediction interval is typically much wider than the CI above. Although we would like to have a narrower interval generally, we should not make the mistake of using CI when forming prediction intervals for predicted values.

Data Example: Predicting body fat

Problem background

- Measuring body fat is not simple. Muscle and bone are denser than fat so an estimate of body density can be used to estimate the proportion of fat in the body. Measuring someone's weight is easy but volume is more difficult.
- One method requires submerging the body underwater in a tank and measuring the increase in the water level. Most people would prefer not to be submerged underwater to get a measure of body fat so we would like to have an easier method.
- In order to develop such a method, researchers recorded age, weight, height, and 10 body circumference measurements for 252 men. Each man's percentage of body fat was accurately estimated by an underwater weighing technique.
- Can we predict body fat using just the easy-to-record measurements?

```
library(faraway)
data(fat,package="faraway")
head(fat, 2)
```

```
##   brozek siri density age weight height adipos  free neck chest abdom  hip
## 1   12.6 12.3  1.0708  23 154.25  67.75   23.7 134.9 36.2  93.1  85.2 94.5
## 2    6.9  6.1  1.0853  22 173.25  72.25   23.4 161.3 38.5  93.6  83.0 98.7
##   thigh knee ankle biceps forearm wrist
## 1  59.0 37.3  21.9   32.0    27.4  17.1
## 2  58.7 37.3  23.4   30.5    28.9  18.2
```

- Use `brozek` as the response (Brozek's equation estimates percent body fat from density).
- Fit a model using all thirteen predictors.

```
lmod <- lm(brozek ~ age + weight + height + neck + chest + abdom +
             hip + thigh + knee + ankle + biceps + forearm + wrist, data=fat)
```

- Let's consider predicting the typical man, exemplified by the median value of all the predictors.

```
x <- model.matrix(lmod)
(x0 <- apply(x,2,median)) #take medians of predictors
```

```
## (Intercept)         age      weight      height        neck       chest
##        1.00       43.00       176.50       70.00       38.00       99.65
##       abdom         hip       thigh        knee       ankle      biceps
##       90.95       99.30       59.00       38.50       22.80       32.05
##     forearm       wrist
##       28.70       18.30
```

```
(y0 <- sum(x0*coef(lmod))) #predicted value
```

```
## [1] 17.49322
```

```
predict(lmod,new=data.frame(t(x0))) #predicted value by R function
```

```
##        1
## 17.49322
```

- Type I: predicting the mean body fat for all men have these same characteristics

```
predict(lmod,new=data.frame(t(x0)),interval="confidence")
```

```
##        fit      lwr      upr
## 1 17.49322 16.94426 18.04219
```

- Type II: predicting the body fat for one particular man

```
predict(lmod,new=data.frame(t(x0)),interval="prediction")
```

```
##        fit     lwr      upr
## 1 17.49322 9.61783 25.36861
```

The second type returns a wider interval.

- The prediction interval ranges from 9.6% body fat up to 25.4%. This is a wide interval since there is a large practical difference between these two limits.
- One might question the value of such a model. The model has an R^2 of 0.75 but perhaps it is not sufficient for practical use.
- The confidence interval for the mean response is much narrower, indicating we can be quite sure about the average body fat of the man with the median characteristics.
- Such information might be useful from a public health perspective where we are concerned about populations rather than individuals.

Extrapolation

- Extrapolation occurs when we try to predict the response for values of the predictor which lie outside the range of the original data.
- There are two different types of extrapolation — quantitative and qualitative.
  - Quantitative extrapolation concerns x_0 that are far from the original data. Prediction intervals become wider as we move further from the original data.
    - Let's see what happens with a prediction for values at the 95th percentile of the data:

```
(x1 <- apply(x,2,function(x) quantile(x,0.95)))
```

```
## (Intercept)         age      weight      height        neck       chest
##       1.000      67.000     225.650      74.500      41.845     116.340
##       abdom         hip       thigh        knee       ankle      biceps
##     110.760     112.125      68.545      42.645      25.445      37.200
##     forearm       wrist
##      31.745      19.800
```

```
predict(lmod, new=data.frame(t(x1)), interval="confidence")
```

```
##        fit      lwr      upr
## 1 30.01804 28.07072 31.96537
```

```
predict(lmod, new=data.frame(t(x1)), interval="prediction")
```

```
##        fit      lwr      upr
## 1 30.01804 21.92407 38.11202
```

- The confidence interval for the mean response is now almost 4% wide compared with the just over 1% width seen in the middle of the data.

What Can Go Wrong with Predictions?

1. Bad model. The Statistician does a poor job of modelling the data.
2. Quantitative extrapolation. We try to predict outcomes for cases with predictor values much different from what we saw in the data. This is a practical problem in assessing the risk from low exposure to substances which are dangerous in high quantities - consider second-hand tobacco smoke, asbestos and radon.
3. Qualitative extrapolation. We try to predict outcomes for observations that come from a different population. For example, suppose we used the models above to predict body fat for women? This is a common problem because circumstances are always changing and it's hard to judge whether the new case is comparable. We prefer experimental data to observational data but sometimes experience from the laboratory does not transfer to real life.
4. Overconfidence due to overfitting. Data analysts search around for good models for the data they have and often do too good a job in finding a fit. This can lead to unrealistically small $\hat{\sigma}$. (Model selection.)
5. Black swans. Sometimes errors can appear to be normally distributed because you haven't seen enough data to be aware of `extremes`. This is of particular concern in financial applications where stock prices are characterized by mostly small changes (normally distributed) but with infrequent large changes (usually falls).

Diagnostics

Motivation

- Estimation and inference of the regression model $Y = X\beta + \epsilon$ depend on several assumptions.
- These assumptions should be checked using regression diagnostics.
- The potential problems can be divided into three categories:
  - *Error*. We have assumed that $\epsilon \sim N(0, \sigma^2 I)$ or in words, that the errors are (1) independent, have (2) equal variance and are (3) normally distributed.
  - *Unusual observations*. Sometimes just a few observations do not fit the model. These few observations might change the choice and fit of the model.
  - *Model*. We have assumed that the structural part of the model, $E(Y \mid X) = X\beta$, is correct.

1. Checking Error Assumptions

- Goal: check error assumptions on (1) independence, (2) constant variance, and (3) normality.
- Errors are not observable $\rightarrow$ examine the sample residuals $\hat{\epsilon} = Y - X\hat{\beta}$.

- Similarity: $\mathrm{E}(\hat{\epsilon} \mid X) = 0$
- But there are differences:
  - Note $\hat{\epsilon} = (I - P)Y = (I - P)(X\beta + \epsilon) = (I - P)\epsilon$.
  - $\mathrm{cov}(\epsilon \mid X) = \sigma^2 I_n$ whereas $\mathrm{cov}(\hat{\epsilon} \mid X) = \sigma^2(I - P)$.
    - Even if the errors have equal variance and are uncorrelated, the residuals do not.
    - When this impact is small, diagnostics can reasonably be applied to residuals to check the assumptions on errors.

1.1 Constant variances

- homoscedasticity: constant symmetrical variation in the verticle $\hat{\epsilon}$ direction.

- heteroscedasticity: nonconstant variance.

Plot of residuals $\hat{\epsilon}$ against fitted values $\hat{y}$

- $\text{cov}(\hat{\epsilon}, \hat{Y}) = \text{cov}(PY, (I-P)Y) = \sigma^2 P(I-P) = 0$

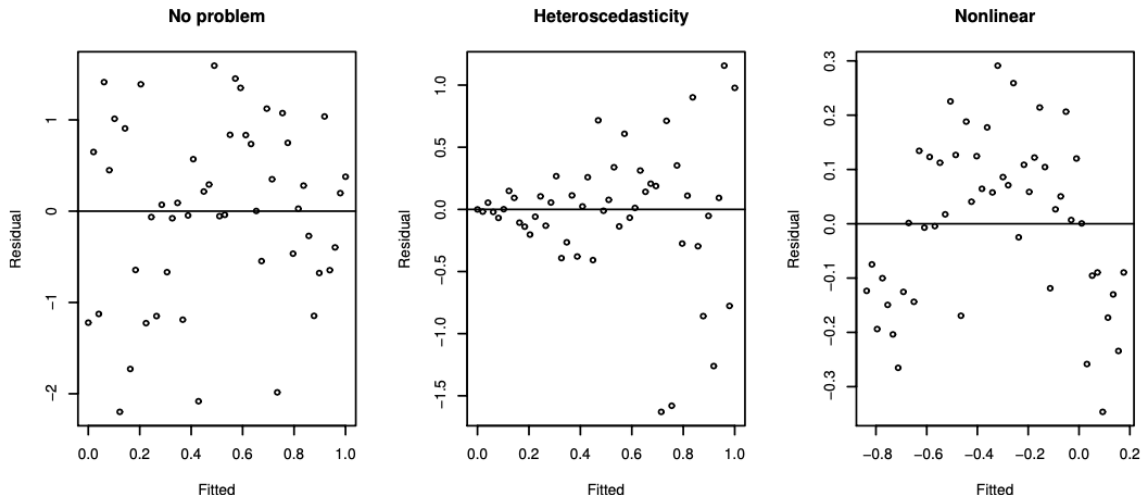
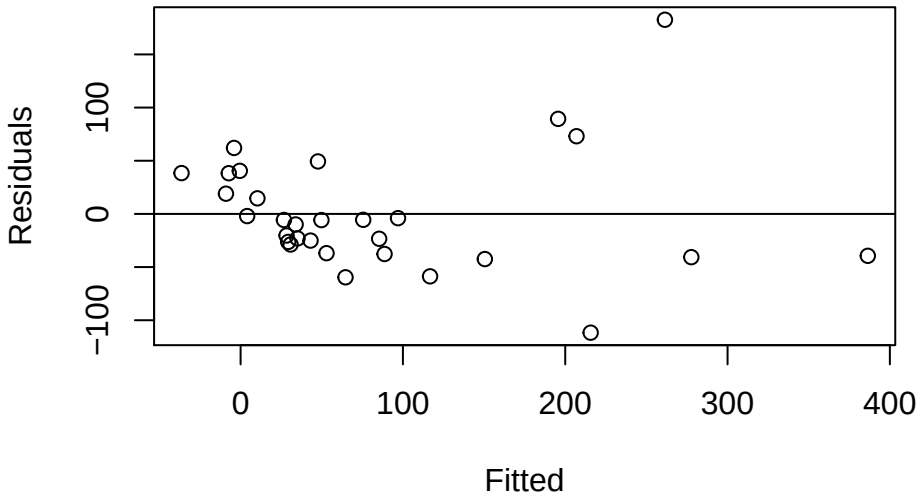


Figure 1: Residuals vs. fitted plots - the first suggests no change to the current model; the second shows nonconstant variance; the third indicates some nonlinearity, which should prompt some change in the structural form of the model.

- Ohter plots $\hat{\epsilon}$ versus x, $|\hat{\epsilon}|$ versus $\hat{y}$, or $\sqrt{|\hat{\epsilon}|}$ versus $\hat{y}$ can all be used.
- It could be hard to judge residual plots without prior experience so it is helpful to generate some artificial plots where the true relationship is known. Sample codes provided below.

```
par(mfrow=c(3,3))
for(i in 1:9) plot(1:50,rnorm(50)) #constant variance
for(i in 1:9) plot(1:50,(1:50)*rnorm(50)) #strong non-constant variance
for(i in 1:9) plot(1:50,sqrt((1:50))*rnorm(50)) #mild non-constant variance
for(i in 1:9) plot(1:50,cos((1:50)*pi/25)+rnorm(50)) #nonlinearity
par(mfrow=c(1,1))
```

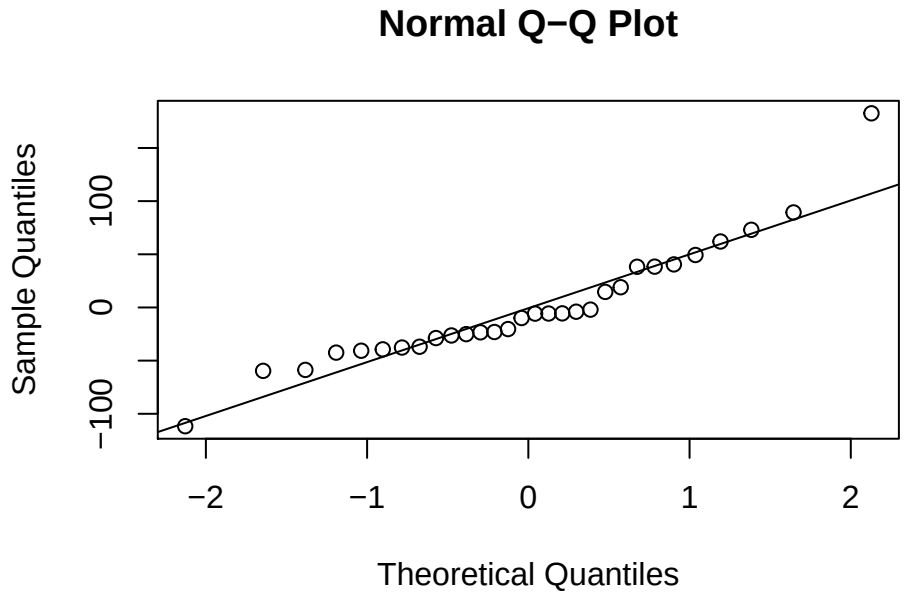
```
data(gala, package = "faraway")
lmod <- lm(Species ~ Area + Elevation + Scruz + Nearest + Adjacent, gala)
plot(fitted(lmod),residuals(lmod),xlab="Fitted",ylab="Residuals")
abline(h=0)
```



1.2 Normality

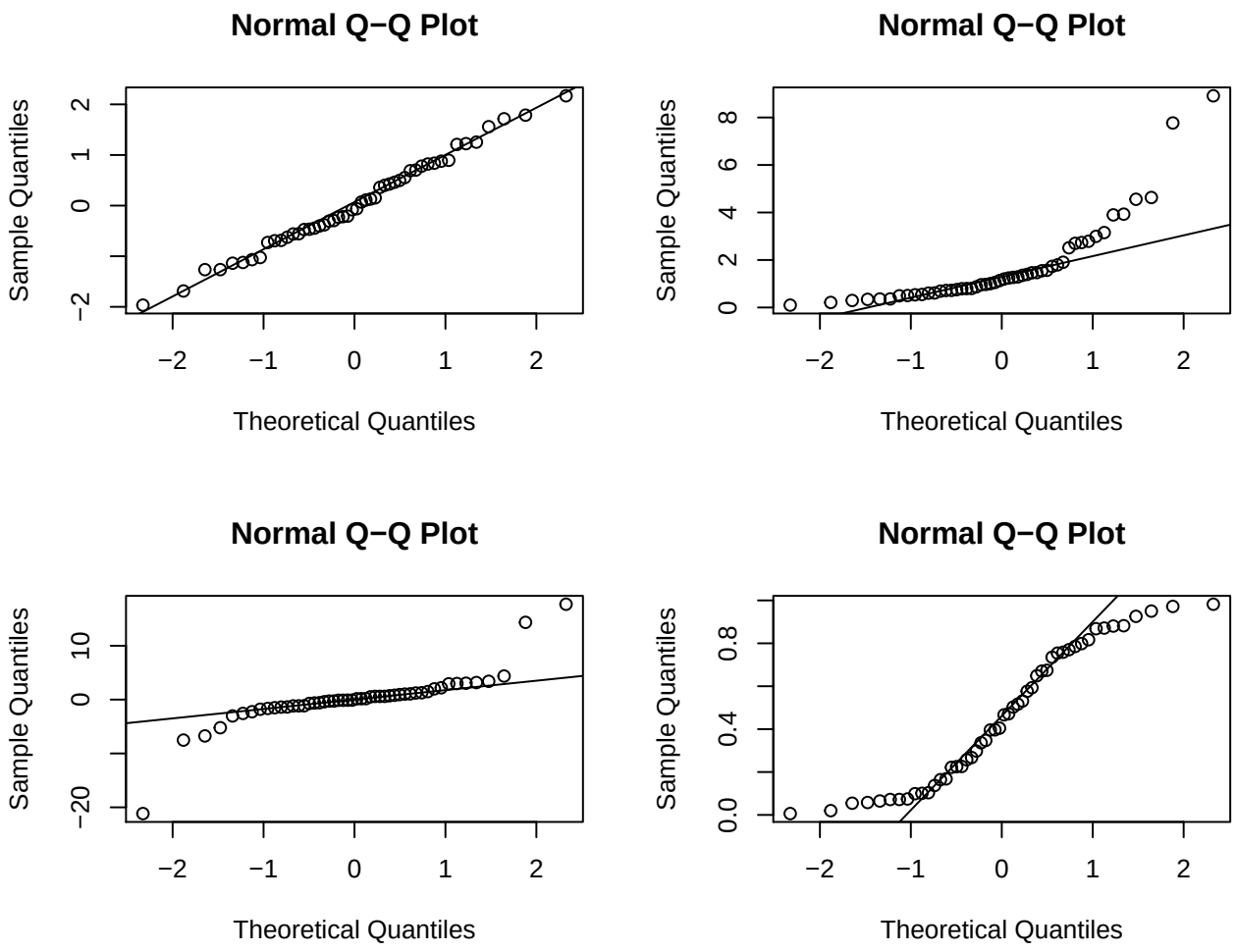
- Tests and confidence intervals are based on the assumption of normal errors.
- The normality of residuals can be assessed by a Q-Q plot. (Histogram is usually not preferred.)
  - Sort sample residuals $\hat{\epsilon}_{(1)} < \ldots < \hat{\epsilon}_{(n)}$.
  - Plot them against $\Phi^{-1}(\frac{1}{n+1}) < \ldots < \Phi^{-1}(\frac{n}{n+1})$ where $\Phi(\cdot)$ denotes the CDF (cumulative distribution fucntion)
  - Normal residuals should follow the line approximately.

```
qqnorm(residuals(lmod))
qqline(residuals(lmod)) #passes through first and third quantiles
```



We can generate sample data from different distributions.

```
par(mfrow=c(2,2))
set.seed(123)
n = 50
#1. normal
x = rnorm(n); qqnorm(x); qqline(x)
#2. log normal: an example of a skewed distribution
x = exp(rnorm(n)); qqnorm(x); qqline(x)
#3. Cauchy: an example of a long-tailed distribution
x = rcauchy(n); qqnorm(x); qqline(x)
#4. Uniform:  an example of a short-tailed distribution
x = runif(n); qqnorm(x); qqline(x)
```



```
par(mfrow=c(1,1))
```

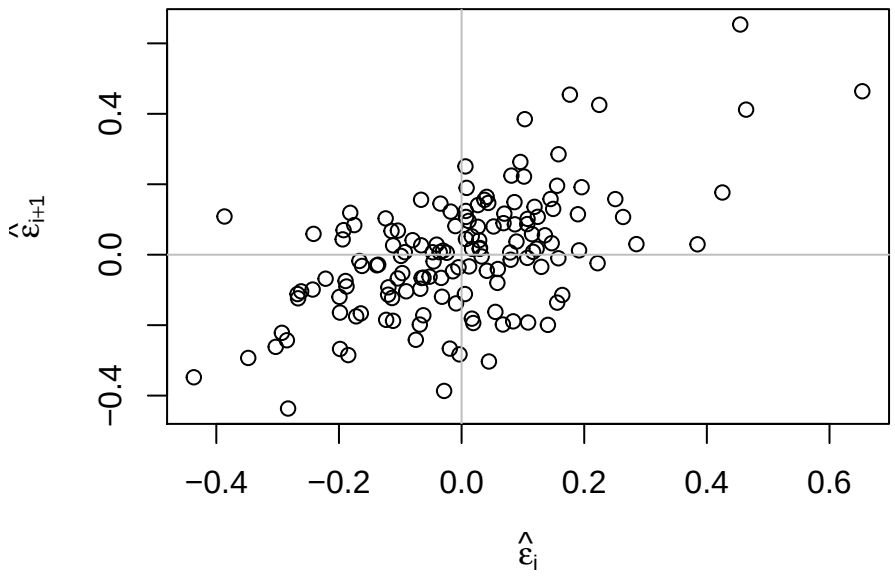
1.3 Correlated errors

- It is difficult to check for correlated errors in general because there are just too many possible patterns of correlation that may occur.
  - Data collected over time may have some correlation in successive errors.
  - Spatial data may have correlation in the errors of nearby measurements.
  - Data collected in blocks may show correlated errors within those blocks.

As an example, we consider a serial data on records of annual temperatures.

- The data contains temprature information 1856 through 2000 by Jones and Mann (2004) `Climate over past millennia`.
- We can build a linear model to predict temperature since 1856 and then subsequently use this to predict earlier temperatures based on proxy information.

```
data(globwarm,package="faraway")
lmod <- lm(nhtemp ~ wusa + jasper + westgreen + chesapeake +
tornetrask + urals + mongolia + tasman, globwarm)
n <- length(residuals(lmod))
plot(tail(residuals(lmod),n-1) ~ head(residuals(lmod),n-1), xlab=
expression(hat(epsilon)[i]),ylab=expression(hat(epsilon)[i+1]))
abline(h=0,v=0,col=grey(0.75))
```



- We can see a positive correlation again indicating positive serial correlation. If you have some doubt as to the significance of the correlation.
- We can model the observed correlation directly by linear regression:

```
summary(lm(tail(residuals(lmod),n-1) ~ head(residuals(lmod),n-1) -1))$coefficients
```

```
##                                Estimate Std. Error  t value     Pr(>|t|)
## head(residuals(lmod), n - 1) 0.5950759 0.06931205 8.585462 1.390651e-14
```

- Note that the the serial correlation is confirmed.

2. Finding unusual observations

1. Leverage: A leverage point is extreme in the predictor space.

2. Outliers: Some observations do not fit the model well.
3. Influential: Some observations change the fit of the model in a substantive manner.

2.1 Leverage

- $h_{ii} = H_{ii}$, i.e., i-th diagonal of hat matrix $H = X(X^\top X)^{-1}X^\top$ are called `leverages`.
- $\text{var}(\hat{\epsilon}_i) = \sigma^2(1 - h_{ii})$
  - a large leverage h_i, will make $\text{var}(\hat{\epsilon}_i)$ small.
  - The fit will be attracted toward y_i.
  - Large values of h_i are usually due to extreme values in the X-space.
- The value of h_{ii} depends only on X and not y.
- $\sum_{i=1}^n h_{ii} = \text{tr}(H) = p$ (See notes Sep 26)
  - Average $\sum_{i=1}^n h_{ii}/n = p/n$.
  - A rough rule is that leverages of more than $2p/n$ should be looked at more closely.

```
data(gala, package = "faraway")
lmod <- lm(Species ~ Area + Elevation + Scruz + Nearest + Adjacent, gala)
hatv <- hatvalues(lmod)
head(hatv)
```

```
##       Baltra    Bartolome     Caldwell     Champion      Coamano Daphne.Major
##   0.07871937   0.09135324   0.06231443   0.07237676   0.16878374   0.07163790
```

```r
sum(hatv) #number of parametrs in the model
```

```
## [1] 6
```

- Standardized residuals:

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

  - Standardization can only correct for the natural non-constant variance in residuals when the errors have constant variance.
  - If there is some underlying heteroscedasticity in the errors, standardization cannot correct for it.
  - When there are unusually large leverages, there could be differences between raw and standardized residuls in their plots.

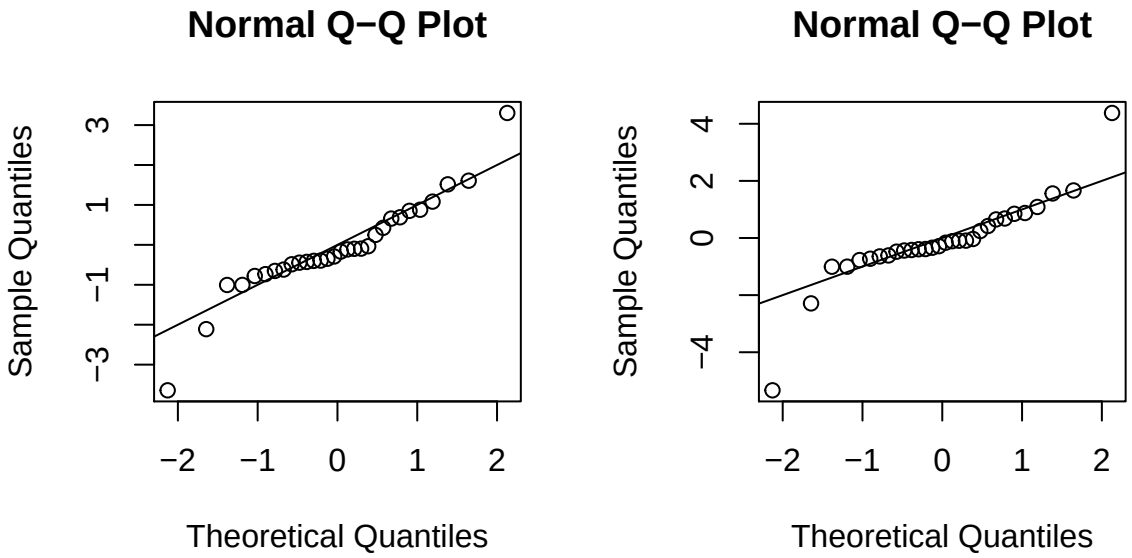
- Leave-one-out studentized residuals:

$$\tilde{r}_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_i}}$$

  - $\hat{\sigma}_{(i)}$: estimate σ without i-th observation.

```r
par(mfrow=c(1,2))
qqnorm(rstandard(lmod))
abline(0,1)

qqnorm(rstudent(lmod))
abline(0,1)
```



```r
par(mfrow=c(1,1))
```

2.2 Outliers

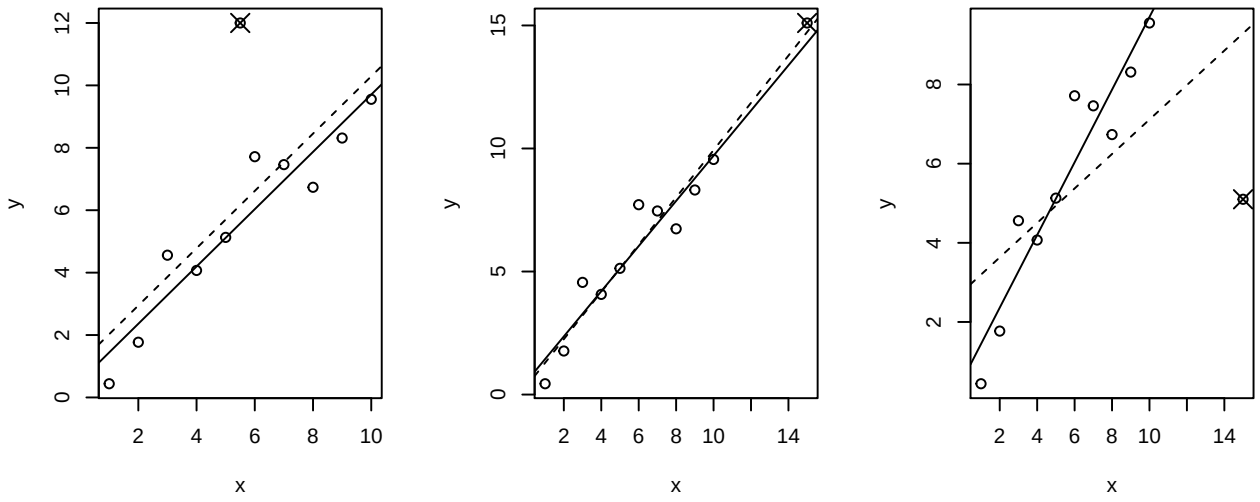
- An outlier is a point that does not fit the current model well.
- Outliers may or may not affect the fit substantially.

```r
set.seed(123)
testdata <- data.frame(x=1:10,y=1:10+rnorm(10))
lmod0 <- lm(y ~ x, testdata)
```

```r
par(mfrow=c(1,3))
p1 <- c(5.5,12)
lmod1 <- lm(y ~ x, rbind(testdata, p1))
plot(y ~ x, rbind(testdata, p1))
points(5.5,12,pch=4,cex=2)
abline(lmod0)
abline(lmod1, lty=2)

p2 <- c(15,15.1)
lmod2 <- lm(y ~ x, rbind(testdata, p2))
plot(y ~ x, rbind(testdata, p2))
points(15,15.1,pch=4,cex=2)
abline(lmod0)
abline(lmod2,lty=2)

p3 <- c(15,5.1)
lmod3 <- lm(y ~ x, rbind(testdata, p3))
plot(y ~ x, rbind(testdata, p3))
points(15,5.1,pch=4,cex=2)
abline(lmod0)
abline(lmod3,lty=2)
```




```
par(mfrow=c(1,1))
```

- A solid regression line shows the fit without the additional point marked with a cross.
- The dashed line shows the fit with the extra point.
- Left panel:
  - Added point is an outlier.
  - But it does not have large leverage or influential (on the fit).
- Middle panel:
  - Added point has large leverage. (well outside of the range of X.)
  - But is not an outlier and is not influential.
- Right panel:
  - Added point changes the fitted line substantially. (influential point.)
  - This is both an outlier and an influential point.

To detect outliers,

- we exclue the point i and recmpute the estimates to get $\hat{\beta}_{(i)}$ and $\hat{\sigma}_{(i)}$.
- Let $\hat{y}_{(i)} = x_i^T \hat{\beta}_{(i)}$
- We have

$$\text{var}\left(y_i - \hat{y}_{(i)}\right) = \sigma^2 \left(1 + x_i^T \left(X_{(i)}^T X_{(i)}\right)^{-1} x_i\right)$$

- Define
$$t_i = \frac{y_i - \hat{y}_{(i)}}{\hat{\sigma}_{(i)} \left(1 + x_i^T \left(X_{(i)}^T X_{(i)}\right)^{-1} x_i\right)^{1/2}},$$
where $X_{(i)}$ represents the design matrix deleting i-th obesrvation.
  - It can be proved $t_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}} = r_i\sqrt{\frac{(n-p-1)}{n-p-r_i^2}}$. (Easy-to-compute. Proof see Theorem 10.1 in Lee and Seber.)
  - If i-th case is not outlier, model is correct, and $\epsilon \sim N(0, \sigma^2 I_n)$, $t_i \sim t_{(n-1)-p}$, where $n-1$ is the sample size.
- Test outliers
  - Practically, $|t_i| > 3$ can imply possible outliers.
  - If we want a level α test,
    - P(all tests accept) $= 1 - P$(at least one rejects) $\geqslant 1 - \sum_i P$(test i rejects) $= 1 - n\alpha$.
    - Each test should use level α/n. (Bonferroni correction.)

2.3 Influential point

- An influential point is one whose removal from the dataset would cause a large change in the fit.
  - An influential point may or may not be an outlier,
  - and may or may not have large leverage,
  - but it will tend to have at least one of these two properties.
- Measure of influence: Cook's distance statistic
$$D_i = \frac{\left(\hat{y} - \hat{y}_{(i)}\right)^T \left(\hat{y} - \hat{y}_{(i)}\right)}{p\hat{\sigma}^2}$$
  - It can also be computed as $D_i = \frac{1}{p} r_i^2 \frac{h_{ii}}{1-h_{ii}}$, where r_i represents i-th standardized residual.

3. Checking the Structure of the Model

- Check linear structure $E(Y \mid X) = X\beta$
- Residual plots can suggest transformations of the variables which might improve the structural form of the model.
- A formal lack of test may be conducted in some cases.