

## 4. Hypothesis Testing

[Test 1] (1) Likelihood function

(2) MLE under full model and constraints  $A\beta = c$

$$(3) \text{Likelihood ratio test } \Lambda = \frac{ML(\text{full model})}{ML(\text{constrained model})}$$

ML: Maximum likelihood  $A\beta = c$

$$\Lambda = \frac{L(\hat{\beta}, \hat{\sigma}^2)}{L(\hat{\beta}_H, \hat{\sigma}_H^2)} = \left( \frac{\hat{\sigma}_H^2}{\hat{\sigma}^2} \right)^{\frac{n}{2}}$$

$\Downarrow$

(4) Pivotal statistic ( $\Lambda$  after monotone transformation)

$$F\text{-stat} = \frac{n-p}{q} \left( \frac{\hat{\sigma}_H^2}{\hat{\sigma}^2} - 1 \right) \quad (1)$$

$$= \frac{(RSS_H - RSS)/q}{RSS/(n-p)} = \frac{(RSS_H - RSS) / (df.H - df.F)}{RSS / df.F} \quad (2)$$

$$= \frac{(A\hat{\beta} - c)^T \{ \text{var}(A\hat{\beta} - c) \}^{-1} (A\hat{\beta} - c) / q}{\hat{\sigma}^2 / q} \quad \text{with } \hat{\sigma}^2 = \frac{RSS}{n-p} \quad (3)$$

$\sim F_{q, n-p}$  under  $H_0$  (F-distribution with d.o.f  $q$  and  $n-p$ )

(5) Interpretations of F-test for a general linear hypothesis

(1) Pivotal statistic transformed from LRT

(2) Relative change of residual sum of squares (RSS)

(3) Quadratic form of  $A\hat{\beta} - c$  (If  $A\beta = c$ ,  $A\hat{\beta} - c$  should be small.)

## [Test 2] F-distribution

(1) Definition

(2)  $F\text{-stat} \sim F_{q, n-p}$  under  $H_0 \Rightarrow p\text{-value \& test}$

(3) R: (1) Fit full model (2) Fit constrained model (3) anova

## [Test 3] t-distribution (squared $t \Leftrightarrow F\text{-test with } q=1$ )

(1) Test one linear constraint:  $H_0: a^T \beta - c = 0$

$$T_{\text{stat}} = \frac{(a^T \hat{\beta} - c) / \sqrt{\text{var}(a^T \hat{\beta} - c)}}{\hat{\sigma} / 6} = (a^T \hat{\beta} - c) / \sqrt{\hat{\text{var}}(a^T \hat{\beta} - c)}$$

$$\left[ \hat{\text{var}}(a^T \hat{\beta} - c) = \hat{\sigma}^2 a^T (X^T X)^{-1} a \text{ (replacing } \sigma \text{ in } \text{var}(a^T \hat{\beta} - c) \text{ by } \hat{\sigma}) \right]$$

(2) Single coefficient  $H_0: \beta_i = 0$

$$T_{\text{stat}} = \frac{\hat{\beta}_i}{\text{S.E.}(\hat{\beta}_i)}$$

(3)  $t\text{-test} \Rightarrow \text{confidence intervals}$

$F\text{-test} \Rightarrow \text{confidence regions}$

## [Test 4] $R^2$ : multiple coefficient of determination

(1) 3 equivalent definitions

$$R^2 = \text{corr}^2(\hat{Y}, Y) = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{TSS} = 1 - \frac{RSS}{TSS}$$

$$\text{with } TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2, \quad RSS = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

## 12) Decomposition of variances

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

$$\begin{array}{ccccc} SS_{Total} & = & SS_{Error} & + & SS_{Regression} \\ SST & & SSE & & SSR \end{array}$$

$$(Also \text{ } RSS = \sum_{i=1}^n \hat{\epsilon}_i^2)$$

$$\begin{array}{ccc} \text{Then } R^2 = \frac{SSR}{SST} & = & 1 - \frac{SSE}{SST} \\ \downarrow & & \downarrow \\ \frac{\text{variation in regression}}{\text{total variation}} & & 1 - \frac{\text{variation in error}}{\text{total variation}} \end{array}$$

$$(3) \text{ Adjusted } R^2: \quad R^2_{Adj} = 1 - \frac{SSE / (n-p)}{SST / (n-1)}$$

(with degree corrected)

## [Test 5] Prediction intervals

- 1) Confidence intervals for mean prediction.
- 2) Prediction intervals for a future response.

## Prediction: point estimate and prediction interval

Lecture 1

①

Model inference

②

Prediction

### 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. (0 meters)

### Two types of predictions

- There are two kinds of predictions made from regression models.
  - One is a predicted mean response:  $E(y_0 | x_0) = x_0^\top \beta$ .  $\rightarrow y_0$
  - Another is a prediction of a future observation:  $Y_{\text{future}}$ . Intuitively,  $y_0 = E(y_0 | x_0) + \epsilon = x_0^\top \beta + \epsilon$  (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  $x_0^\top \hat{\beta}$  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  $x_0^\top \beta + \epsilon$ . Since  $E\epsilon = 0$ , the predicted price is  $x_0^\top \hat{\beta}$  but in assessing the variance of this prediction, we must include the variance of  $\epsilon$ .

② variation  $\hat{\beta}$

③

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

- ① The confidence interval (CI) for the mean response for given  $x_0$ , i.e.,  $E(y_0 | x_0) = x_0^\top \beta$  is:

$$\hat{y}_0 \pm t_{n-p}^{(\alpha/2)} \hat{\sigma} \sqrt{x_0^\top (X^\top X)^{-1} x_0} \quad \text{var}(x_0^\top \hat{\beta})$$

– This is by

$$\frac{(\hat{y}_0 - x_0^\top \hat{\beta}) / \sqrt{\text{var}(\hat{y}_0 - x_0^\top \hat{\beta})}}{\hat{\sigma} / \sigma} \sim t_{n-p}, \quad \Rightarrow \text{derivation of t-test for general linear combination of } \beta$$

– where  $\hat{y}_0 - x_0^\top \hat{\beta} = x_0^\top \hat{\beta} - x_0^\top \hat{\beta} \sim N(0, \text{var}(x_0^\top \hat{\beta}))$  and  $\hat{\sigma}^2 / \sigma^2 \sim \chi_{n-p}^2 / (n-p)$  (normality of  $\hat{\beta}$ )

– Variance:

$$\text{var}(\hat{y}_0 - x_0^\top \hat{\beta}) = \text{var}(x_0^\top \hat{\beta}) = x_0^\top (X^\top X)^{-1} x_0 \sigma^2. \quad \text{model}$$

## 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  $\epsilon$  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)

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

– where  $\hat{y}_0 - y_0 = x_0^\top \hat{\beta} - (x_0^\top \beta + \epsilon) \sim N(0, \text{var}(x_0^\top \hat{\beta} - \epsilon))$  and  $\hat{\sigma}^2 / \sigma^2 \sim \chi_{n-p}^2 / (n-p)$

– Variance: It is usually reasonable to assume future  $\epsilon$  is independent of  $\hat{\beta}$  and has variance  $\sigma^2$ . Then

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

\* Parameter uncertainty + Model uncertainty  $\Rightarrow \epsilon$

- ② **Prediction interval:** 100(1 -  $\alpha$ )% prediction interval for a single future response is:

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

– Previous confidence intervals have been for model parameters  $\beta$ . 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
```

$$x_0^T \hat{\beta}$$

```
## [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. *median*
- 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 3-4
```



```
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?

- ✓ Bad model. The Statistician does a poor job of modelling the data.
- ✓ 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. *95% wider than 50%*
- ✓ 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. *252 men ↓ women*
- ✓ 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). 

*robust methods  
(Quantile regression.)*

# 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,  $E(Y | X) = X\beta$  is correct.

$\epsilon_i$   
 $i=1 \dots n$

## 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  $\hat{\epsilon} = Y - X\hat{\beta}$ .  

$$\hat{\epsilon} = Y - X\hat{\beta} \quad \text{mean zero} \quad E(\hat{\epsilon}) = 0$$

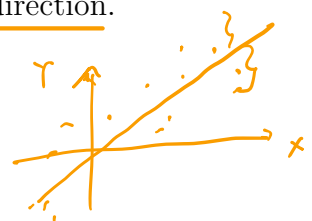
$\downarrow$  observed       $\downarrow$  fitted
- Similarity:  $E(\hat{\epsilon} | X) = 0$  (proved)
- But there are differences:  $\hat{\epsilon} = Y - \hat{Y} = Y - PY$ 
  - Note  $\hat{\epsilon} = (I - P)Y = (I - P)(X\beta + \epsilon) = (I - P)\epsilon$ .  $\hat{\epsilon} \neq \epsilon$
  - $\text{cov}(\epsilon | X) = \sigma^2 I_n$  whereas  $\text{cov}(\hat{\epsilon} | X) = \sigma^2(I - P)$ .
    - \* 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.

$$[(I - P)X = X - PX = 0]$$

$$\begin{aligned} \text{cov}(\hat{\epsilon}) &= (I - P) \text{cov}(\epsilon) (I - P)^T \\ &= (I - P) \sigma^2 I (I - P)^T \\ &= \sigma^2 (I - P) \quad (\text{by idempotent symmetry}) \end{aligned}$$

### 1.1 Constant variances

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



↗ weighted LS

- heteroscedasticity: nonconstant variance.

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

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

$$P(I - P) = P - P^2 = 0$$

(idempotent symmetric)

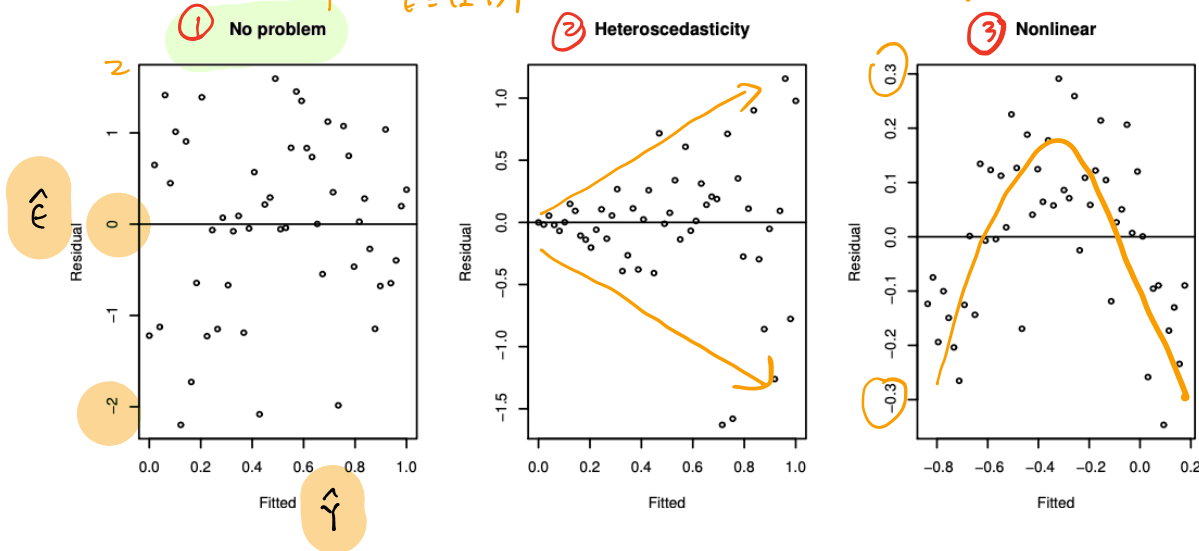
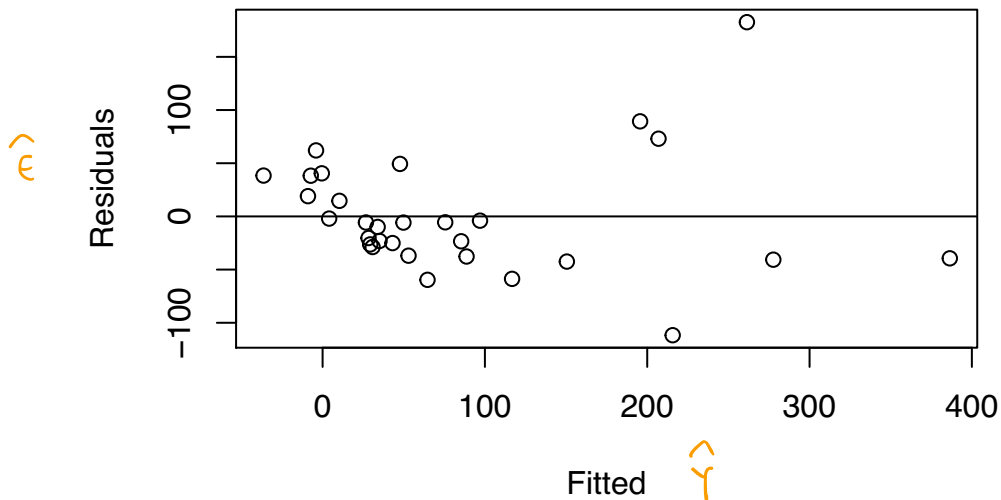


Figure 1: Residuals vs. fitted plots - the first suggests no change to the current model; the second shows nonconstant variance; the third indicates some nonlinearity, which should prompt some change in the structural form of the model. ( $\times \beta$ )

- Other 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  $\epsilon_i \sim N(0,1)$ 
for(i in 1:9) plot(1:50,(1:50)*rnorm(50)) #strong non-constant variance  $\epsilon_i = i \times r_i$   $r_i \sim N(0,1)$ 
for(i in 1:9) plot(1:50,sqrt((1:50))*rnorm(50)) #mild non-constant variance  $\epsilon_i = \sqrt{i} \times r_i$   $r_i \sim N(0,1)$ 
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 + Scrub + Nearest + Adjacent, gala)
plot(fitted(lmod), residuals(lmod), xlab="Fitted", ylab="Residuals")
abline(h=0)
```



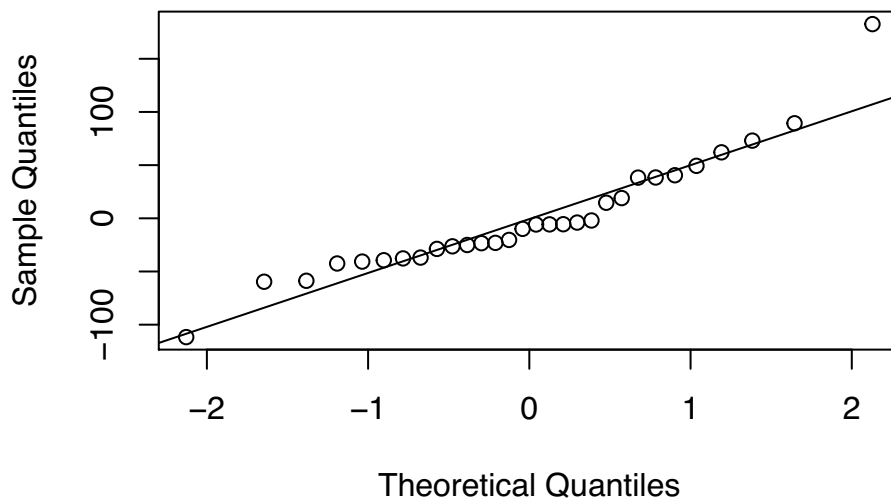
## 1.2 Normality

- 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  $\hat{\epsilon}_{(1)} < \dots < \hat{\epsilon}_{(n)}$ .
  - Plot them against  $\Phi^{-1}(\frac{1}{n+1}) < \dots < \Phi^{-1}(\frac{n}{n+1})$  where  $\Phi(\cdot)$  denotes the CDF (cumulative distribution function).
  - Normal residuals should follow the line approximately.

✓ `qqnorm(residuals(lmod))`

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

### Normal Q-Q Plot



Bell shape  
+ distribution

We can generate sample data from different distributions.

```
par(mfrow=c(2,2))
```

```
set.seed(123)
```

```
n = 50
```

```
#1. normal
```

① `x = rnorm(n); qqnorm(x); qqline(x)`

```
#2. log normal: an example of a skewed distribution
```

② `x = exp(rnorm(n)); qqnorm(x); qqline(x)`

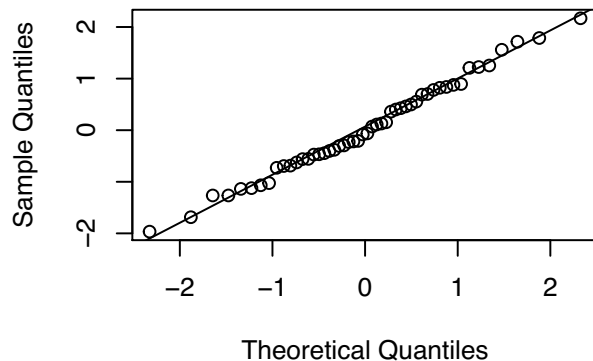
```
#3. Cauchy: an example of a long-tailed distribution
```

③ `x = rcauchy(n); qqnorm(x); qqline(x)`

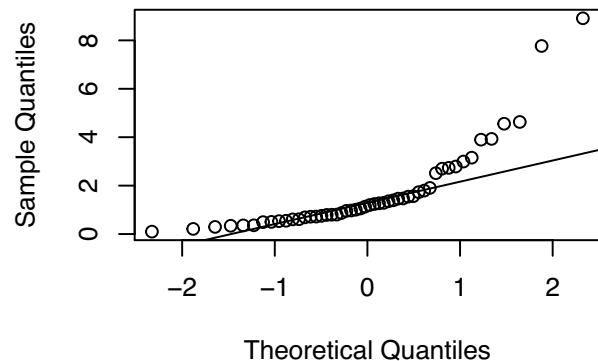
```
#4. Uniform: an example of a short-tailed distribution
```

④ `x = runif(n); qqnorm(x); qqline(x)`

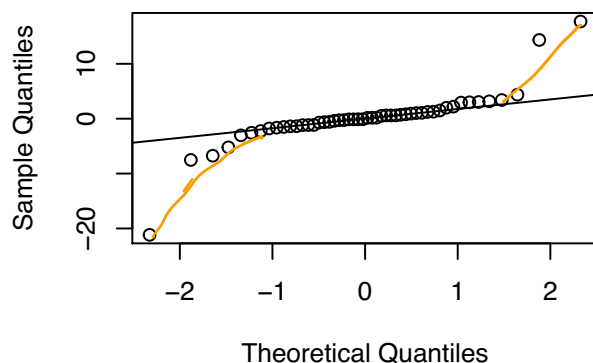
① Normal Q-Q Plot



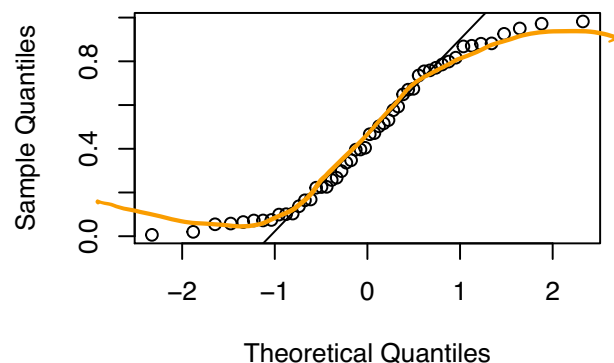
② Normal Q-Q Plot



③ Normal Q-Q Plot



④ Normal Q-Q Plot



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

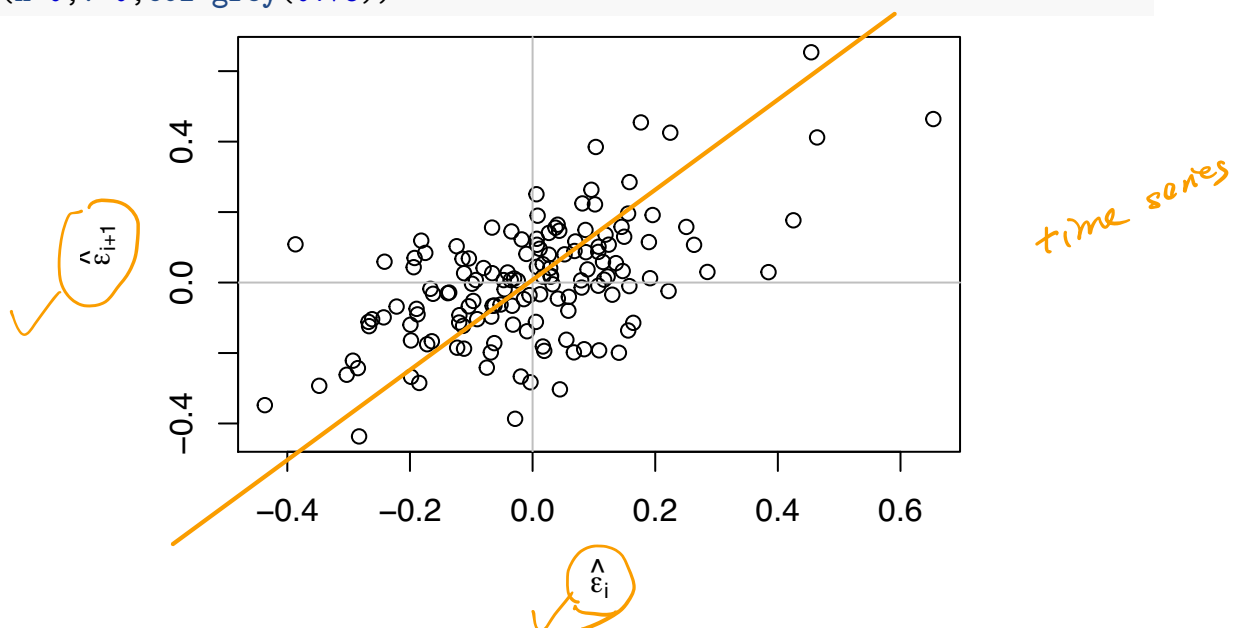
### 1.3 Correlated errors

- 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 temperature 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 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

Projection  $P = X(X^T X)^{-1} X^T$   $\hat{Y} = P Y = H Y$

- $h_{ii}$  =  $H_{ii}$ , i.e.,  $i$ -th diagonal of hat matrix  $H = X(X^T X)^{-1} X^T$  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.

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

$$= \sigma^2(I - H)$$

$$Y \sim \begin{matrix} X \\ T \\ 1\text{-dim} \end{matrix}$$

```
data(gala, package = "faraway")
```

```
lmod <- lm(Species ~ Area + Elevation + Scrub + Nearest + Adjacent, gala)
```

```
hatv <- hatvalues(lmod)
```

```
head(hatv) ⇒ first 5 hat values head(.)
```

island names

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

$$p = 5 + 1 = 6$$

✓ `sum(hatv)` #number of parameters in the model

## [1] 6

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

- Standardized residuals:

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

- 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  $\hat{\epsilon}$  and standardized residuals  $r_i$  in their plots.

- Leave-one-out studentized residuals:

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

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

subscript (i) means excluding  $i$ th observation

```
par(mfrow=c(1,2))
```

```
qqnorm(rstandard(lmod))
```

$r_i$

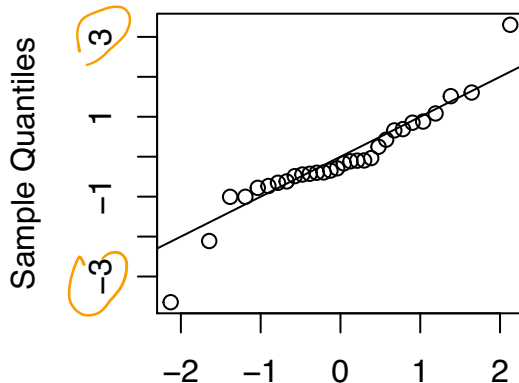
```
abline(0,1)
```

```
qqnorm(rstudent(lmod))
```

$\tilde{r}_i$

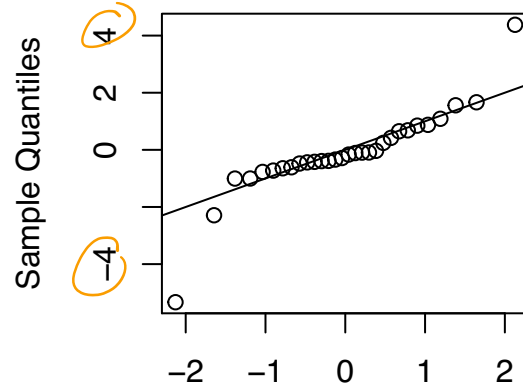
```
abline(0,1)
```

Normal Q-Q Plot



Theoretical Quantiles

Normal Q-Q Plot



Theoretical Quantiles



```
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.

```
set.seed(123)
```

```
testdata <- data.frame(x=1:10, y=1:10+rnorm(10))
```

$y = x + \epsilon$     10 observations

```
lmod0 <- lm(y ~ x, testdata)
```

```
par(mfrow=c(1,3))
```

```
p1 <- c(5.5, 12)
```

```
lmod1 <- lm(y ~ x, rbind(testdata, p1))
```

new point added

```
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)
```

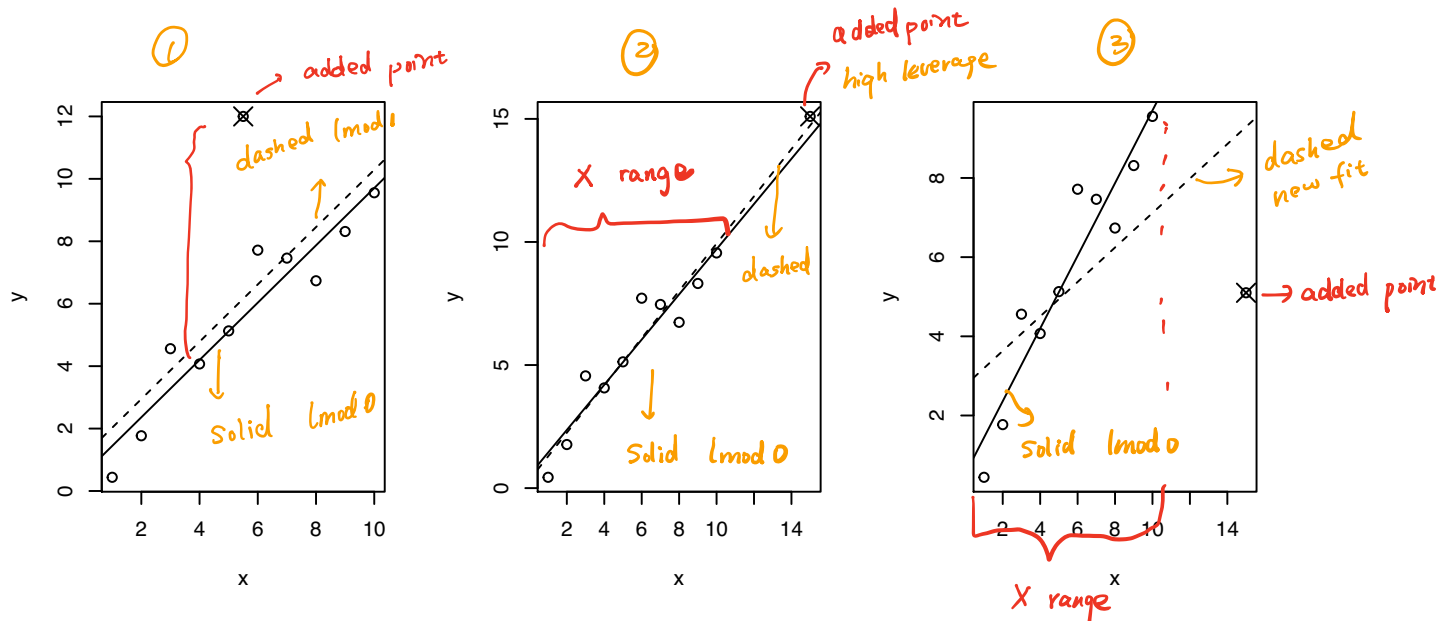
```
lmod3 <- lm(y ~ x, rbind(testdata, p3))
```

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

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

$$y_i = x_i^T \beta + \epsilon_i \quad y_i - \hat{y}_{(i)} = \epsilon_i + x_i^T (\beta - \hat{\beta}_{(i)})$$

*independent*

$$\begin{aligned} \text{var}(y_i - \hat{y}_{(i)}) &= \text{var}(\epsilon_i) + \text{var}(x_i^T (\beta - \hat{\beta}_{(i)})) \\ &= \sigma^2 + x_i^T \text{var}(\hat{\beta}_{(i)}) x_i \quad (\text{prediction type I}) \\ &= \sigma^2 \left\{ 1 + x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i \right\} \end{aligned}$$

In full observation model,  $\text{var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$   
 After leaving one, similarly,  $\text{var}(\hat{\beta}_{(i)}) = \sigma^2 (X_{(i)}^T X_{(i)})^{-1}$

- Define

$$t_i = \frac{y_i - \hat{y}_{(i)}}{\hat{\sigma}_{(i)} \left( 1 + x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i \right)^{1/2}},$$

where  $X_{(i)}$  represents the design matrix deleting  $i$ -th observation.

- 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  $\alpha$  test,
    - \*  $P(\text{all tests accept}) = 1 - P(\text{at least one rejects}) \geq 1 - \sum_i P(\text{test } i \text{ rejects}) = 1 - n\alpha$ .
    - \* Each test should use level  $\alpha/n$ . (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

$$D_i = \frac{(\hat{y} - \hat{y}_{(i)})^T (\hat{y} - \hat{y}_{(i)})}{p \hat{\sigma}^2}$$

- 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  $E(Y | X) = X\beta$
- 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.