Statistics Review
Week 3 – Wednesday
Mathematical Modeling (MATH 420/620)

Paul J. Hurtado

Wednesday, 13 Sept, 2017

Announcements

Math Club Meeting: 4pm Thursday

Random Variables & Probability Distributions

What does it mean for X to be a random variable?

1. X is the outcome of an experiment; a place-holder for a random number.

Random Variables & Probability Distributions

What does it mean for X to be a random variable?

1. X is the outcome of an experiment; a place-holder for a random number.
2. X has a *distribution* associated with it.

Random Variables & Probability Distributions

What does it mean for X to have a distribution?

1. Distributions describe the propensity for some outcomes to occur more often, or with greater likelihood, than others.

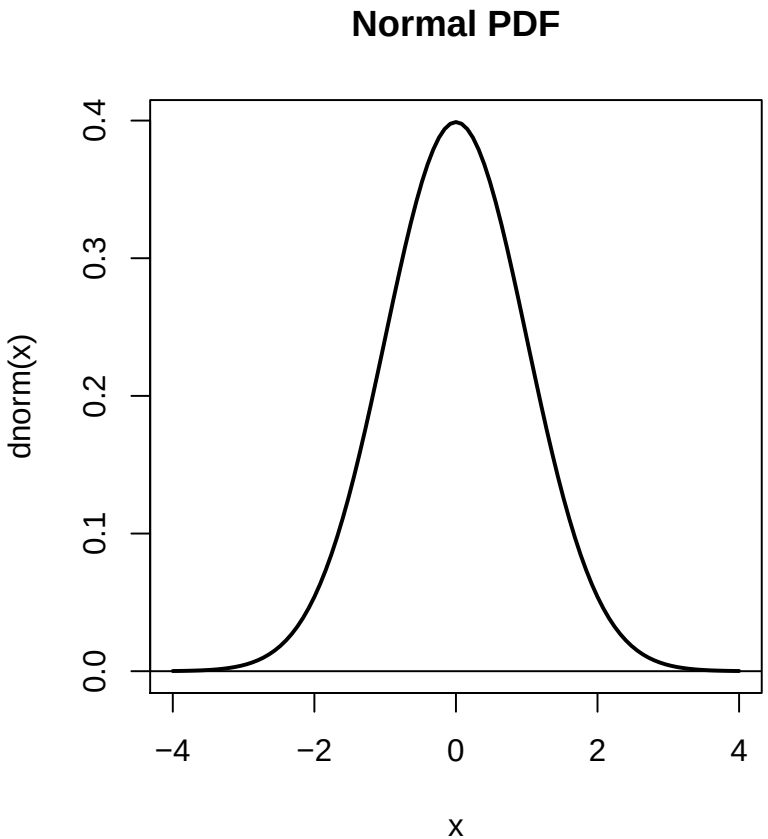
Random Variables & Probability Distributions

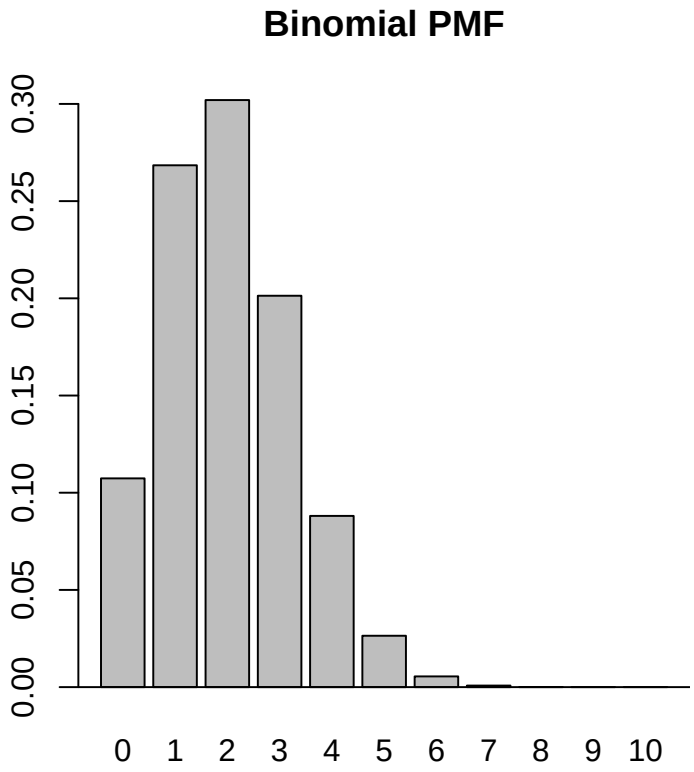
What does it mean for X to have a distribution?

1. Distributions describe the propensity for some outcomes to occur more often, or with greater likelihood, than others.
2. When we refer to the distribution, we are referring to a few different, but equivalent, functions!

Random Variables & Probability Distributions

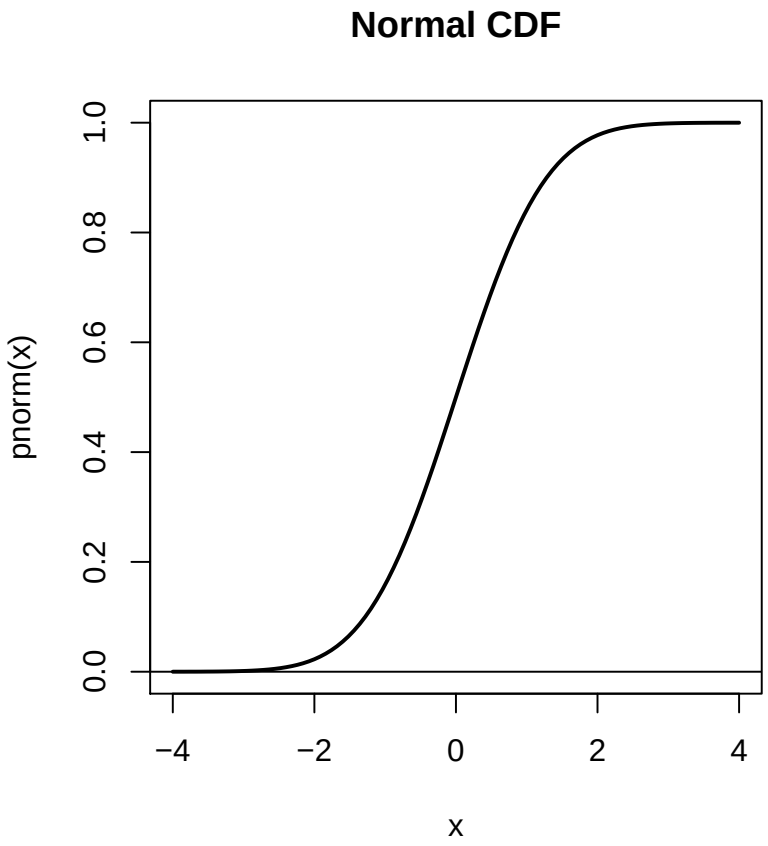
Probabilities of events are calculated from either the PDF (continuous) or PMF (discrete):

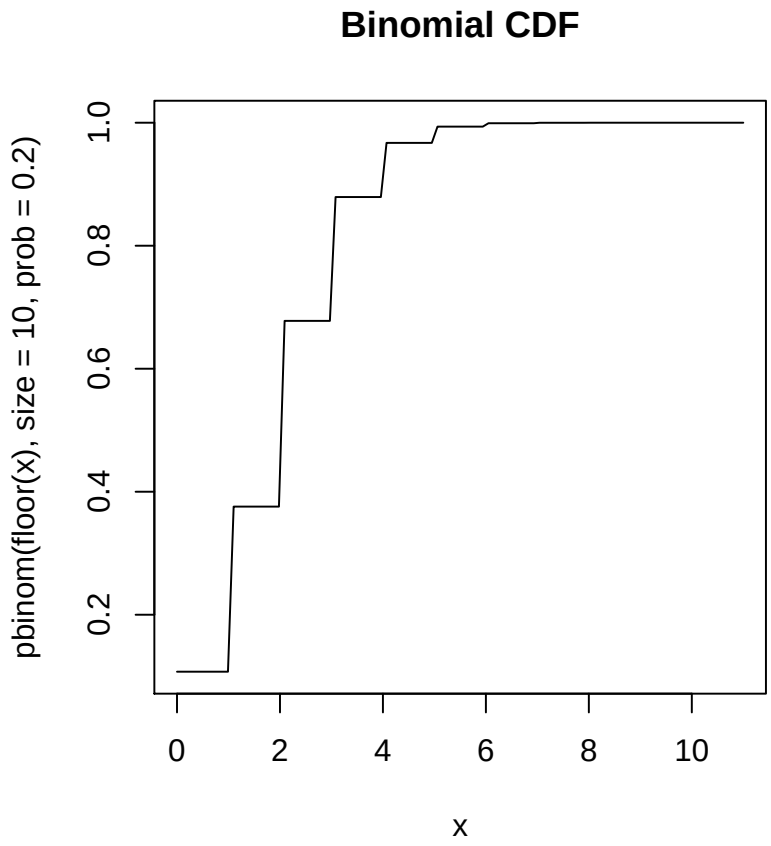




Random Variables & Probability Distributions

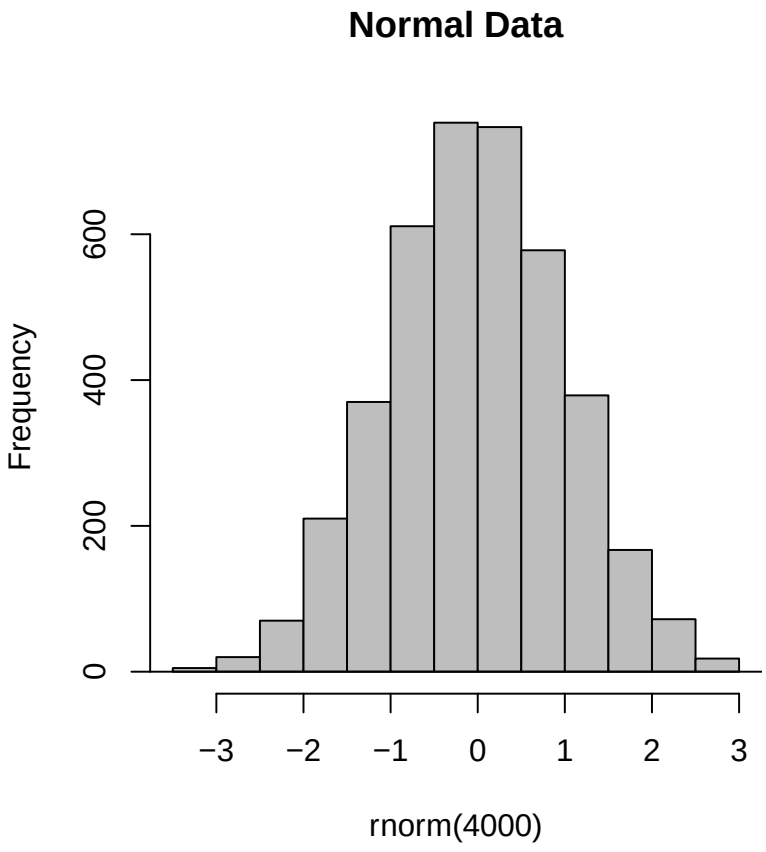
If you know the CDF, you know the PDF/PMF, and *vice versa*.

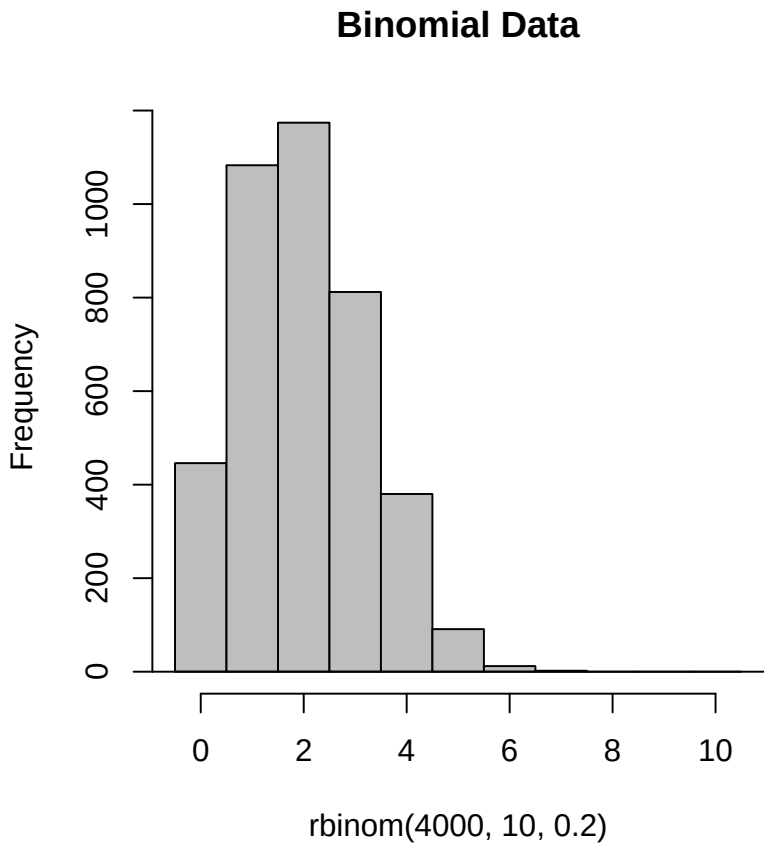




Random Variables & Probability Distributions

Histograms of large random samples look like the PDF/PMF!





Estimates vs Estimators?

1. **Estimators** are *functions of random variables*, and thus are themselves random variables. Rules for calculating...

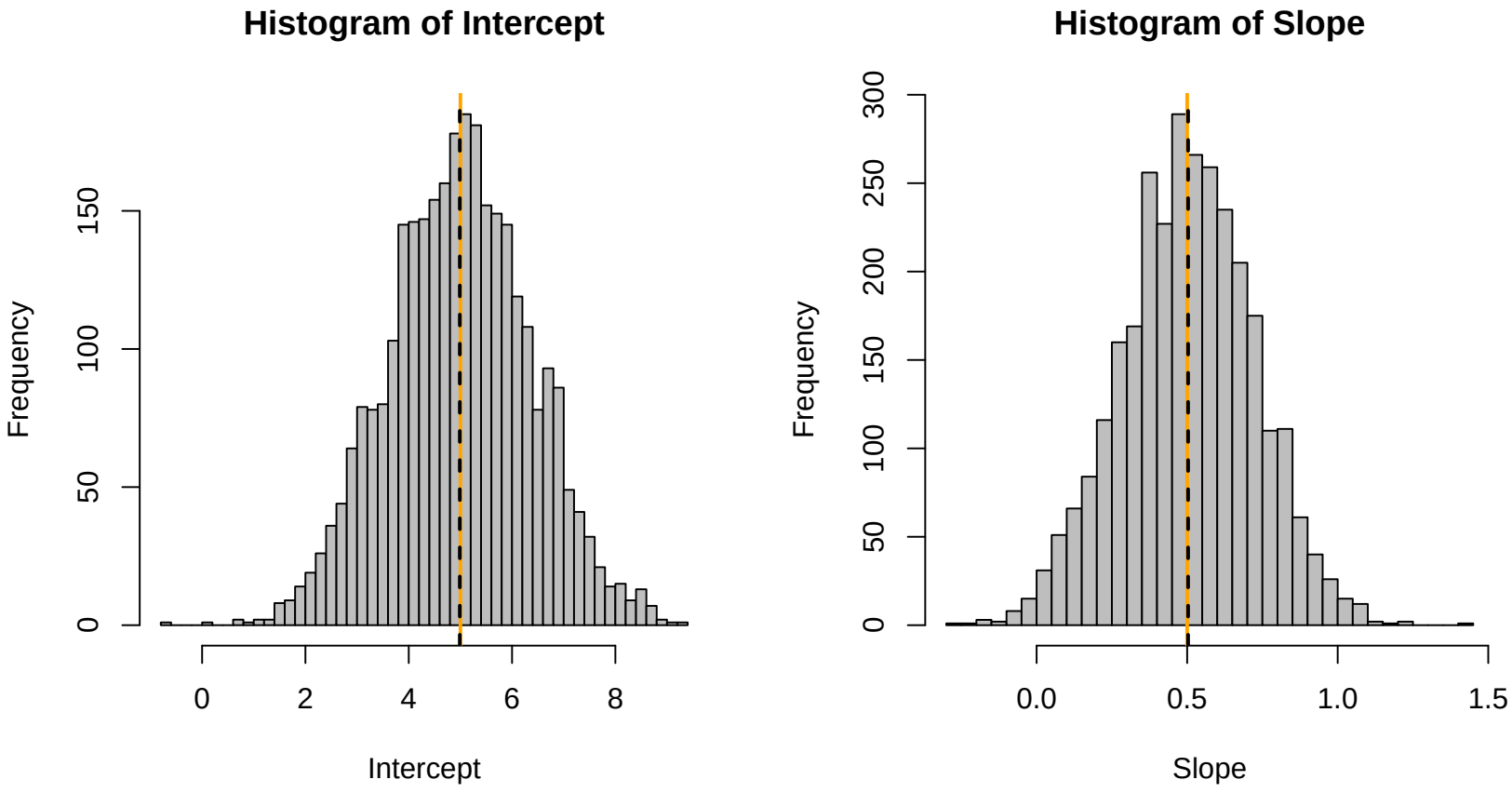
Estimates vs Estimators?

1. **Estimators** are *functions of random variables*, and thus are themselves random variables. Rules for calculating...
2. **Estimates**, which are *single numbers*.

Unfortunately, you must rely on context for which $\hat{\beta}$ refers to!

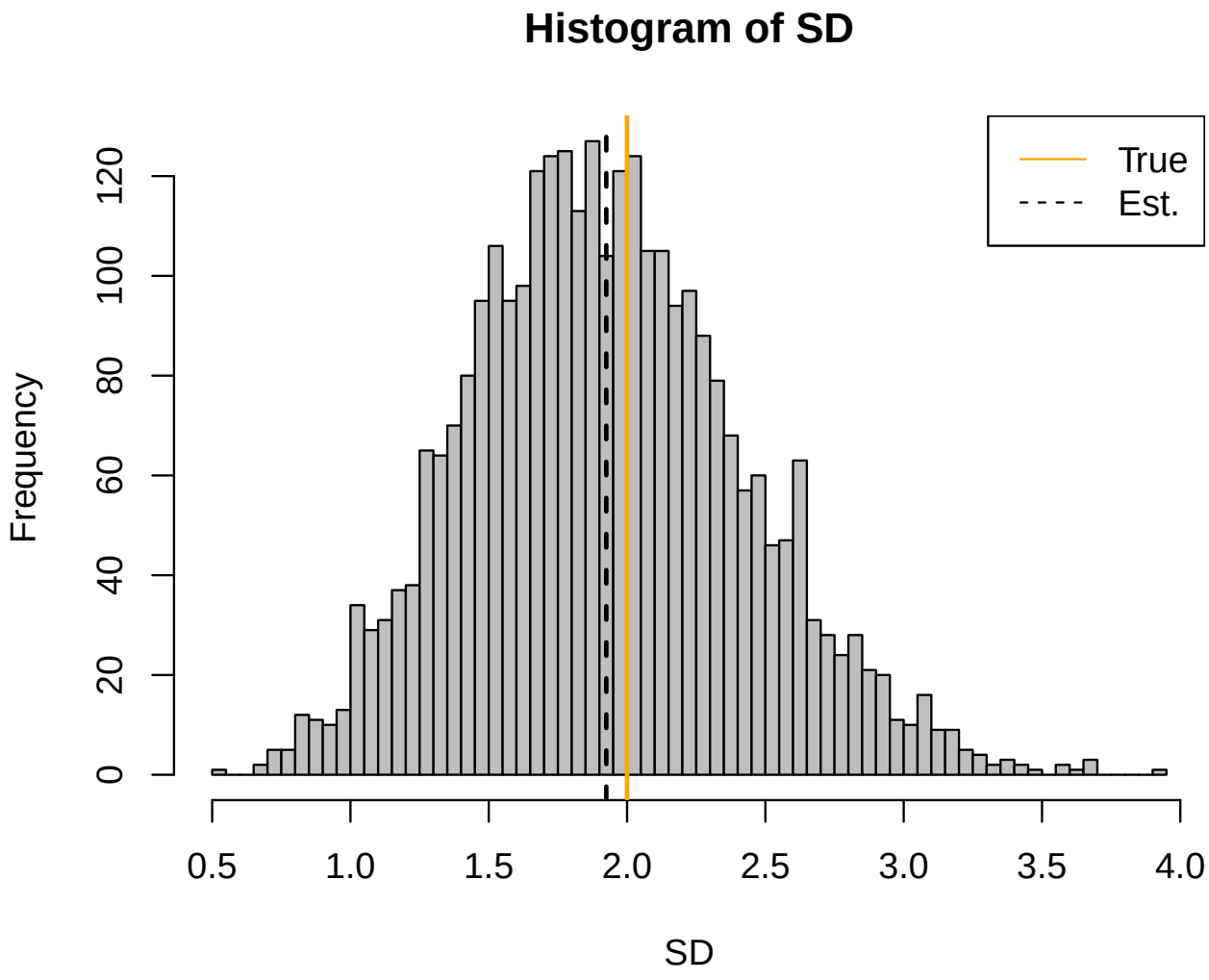
Estimators as random variables

3000 Replicated SLR Estimates (N=10)



Estimators as random variables

3000 Replicated SLR Estimates (N=10)



Recall the SLR Model:

Estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma^2}$ yield the *best fit model*

$$Y_i|X = x_i \sim \text{Normal}(\text{mean} = \hat{\beta}_0 + \hat{\beta}_1 x_i, \ \text{var} = \hat{\sigma}^2)$$

or, alternatively stated

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \epsilon_i$$

where ϵ_i is Normal error with mean 0, variance $\hat{\sigma^2}$.

Parameter Estimation via Optimization

Recall $\widehat{\beta_0}$ and $\widehat{\beta_1}$ were obtained by finding the slope (b_1) and intercept (b_0) that **minimize** the *RSS*

$$RSS = \sum_{i=1}^{n} (y_i - \overbrace{(b_0 + b_1 x_i)}^{\widehat{y_i}})^2$$

This is an example of an **optimization** problem: Finding function arguments (i.e., input values) that *minimize* or *maximize* an **objective function**.

Optimization in Practice

Suppose we aim to minimize the objective function

$$G(\theta) = G(a, b, c).$$

Typically, two approaches are used:

1. **Analytical:** Solve $\nabla G(\theta) = 0$, i.e.,

$$\frac{\partial G}{\partial a} = 0, \qquad \frac{\partial G}{\partial b} = 0, \qquad \frac{\partial G}{\partial c} = 0.$$

2. **Computational:** Use minimization algorithms (see `optimize()`, `optimx()` in **R**)

Analytical Example

$$\frac{\partial RSS}{\partial b_0} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0$$

$$\frac{\partial RSS}{\partial b_1} = -2 \sum_{i=1}^{n} x_i \, (y_i - b_0 - b_1 x_i) = 0$$

which rearranges to two linear equations in b_0 and b_1:

$$\sum_{i=1}^{n} y_i = b_0 \, n + b_1 \sum_{i=1}^{n} x_i; \qquad \sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2$$

Solving yields the estimates $\widehat{\beta}_1 = S_{XY}/S_{XX}$ and $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$

Computational

```
library(optimx)
# Minimize f(a,b)=a^2+b^2
f <- function(ps){ return(ps[1]^2+ps[2]^2) }
opt=optimx(c(a=1,b=1),f)
opt
```

```
##                          a             b        value fevals gevals niter
## Nelder-Mead  3.754010e-05  5.179101e-05 4.091568e-09     63     NA    NA
## BFGS        -4.263536e-16 -4.263536e-16 9.087931e-30      8      3    NA
##             convcode kkt1 kkt2 xtimes
## Nelder-Mead        0 TRUE TRUE      0
## BFGS               0 TRUE TRUE      0
```

How close to the analytical optimum (0,0) are these estimates?

$$a = -4.2635361 \times 10^{-16} \qquad b = -4.2635361 \times 10^{-16}$$

Exercise

Edit the following code to compare estimates of the slope and intercept obtained from `optimx()` versus `lm()`.

```
library(optimx)
# Simulated data set
set.seed(757)
x=1:20
y=rnorm(length(x),11+1.2*x,sd=pi)

# Minimize obj()=RSS
obj <- function(ps){
   return( sum( (???)^2 ) )
}
p.initial=c(b0=0,b1=0)
opt=optimx(p.initial,obj)
opt

# lm() gives...
summary(lm(y~x))
```

Prediction & Confidence Intervals

Recall that our estimators $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are Normal r.v.s

Properties of Normal Distributions:

Suppose $Y \sim Normal(\mu, \sigma)$ and $a \in \mathbb{R}$.

1. If $Z = a + Y$ then $Z \sim Normal(\mu + a, \sigma)$.
2. If $Z = Y/a$ then $Z \sim Normal(\mu/a, \sigma/a)$.
3. (Standard Normal) If $Z = (X - \mu)/\sigma$ then $Z \sim Normal(0, 1)$.

Note $\bar{X}$ is $Normal(\mu, \sigma/\sqrt{n}$

Prediction & Confidence Intervals

To characterize uncertainty in estimates of β_0 and β_1 (or predictions of $Y|X = x$), use the distributions of $\widehat{\beta_0}$ and $\widehat{\beta_1}$ (or $\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1}\, x_i$) to compute **confidence intervals** (or **prediction intervals**). Compare to **credible interval**.

1. A $100(1-\alpha)\%$ **confidence interval** is the interval (a, b) that will contain the expected value of the given distribution approximately $100(1-\alpha)\%$ of the time.
2. If *the complete process of data collection and analysis* were repeated, a $100(1-\alpha)\%$ **prediction interval** will contain the next observation of $Y|X = x$ approximately $100(1-\alpha)\%$ of the time.

Confidence Interval

```
x=1:20
B0=11
B1=1.2
Nreps=1000
CIdat=data.frame(L0=rep(NA,Nreps),U0=NA,B0.in.CI=NA,L1=NA,U1=NA,B1.in.CI=NA)
for(i in 1:Nreps) {
  y=rnorm(length(x),B0+B1*x,sd=pi)
  M=confint(lm(y~x),level = 0.95)
  CIdat$L0[i] = M[1,1];   CIdat$U0[i] = M[1,2]
  CIdat$L1[i] = M[2,1];   CIdat$U1[i] = M[2,2]
  CIdat$B0.in.CI[i] = ( M[1,1]<B0 & B0<M[1,2] )
  CIdat$B1.in.CI[i] = ( M[2,1]<B1 & B1<M[2,2] )
}
sum(CIdat$B0.in.CI)/Nreps

## [1] 0.964

sum(CIdat$B1.in.CI)/Nreps

## [1] 0.954
```

Prediction Interval (SLR)

```
x=1:20
B0=11
B1=1.2
Nreps=1000
PIdat=data.frame(L=rep(NA,Nreps),U=NA,Y.in.PI=NA)
for(i in 1:Nreps) {
  y=rnorm(length(x),B0+B1*x,sd=pi)
  PI=predict(lm(y~x),data.frame(x=12),interval="prediction",level=0.95)
  PIdat$L[i] = PI[2]
  PIdat$U[i] = PI[3]
  Y=rnorm(1,B0+B1*12,sd=pi)
  PIdat$Y.in.PI[i] = ( PI[2]<Y & Y<PI[3] )
}
sum(PIdat$Y.in.PI)/Nreps

## [1] 0.958
```

Credible Intervals

The difference between *credible intervals* and *confidence intervals* is mostly philosophical: the former arising in Bayesian frameworks, the latter in Frequentist frameworks. *The two can differ substantially* in more complex models, but it's reassuring that for many models (e.g., linear with Normal errors) *they are often indistinguishable*.

For now, it suffices to know (1) that they're different concepts, and (2) what *exactly* defines a confidence interval.

1. The **100$(1-\alpha)$% confidence interval** is an interval calculated from a single data set that for 100$(1-\alpha)$% of such data sets includes the true value in question.
2. The 100$(1-\alpha)$**% credible interval** is the interval that with 100$(1-\alpha)$% probability contains the true value.

For more information, see comparisons in applications, e.g.,

> *Lu, Ye, and Hill. 2012.* **Analysis of regression confidence intervals and Bayesian credible intervals for uncertainty quantification.** *URL: people.sc.fsu.edu/ mye/pdf/paper31.pdf*

Exercise

1. See `?qt`. Modify the **Confidence Interval** code above so that instead of using `confint()` you calculate upper and lower limits using `qt()` and the formulas in Ch. 2.
2. Modify the code resulting from the exercise above to instead (erroneously!) use the Normal distribution, i.e., assume we can use the mean for the expected value and the sample standard deviation for the population standard deviation. *Does the t distribution or the Normal distribution give the broader Confidence Interval?*

Recap: Confidence Intervals

Intercept CI: The distribution of the intercept estimator $\widehat{\beta_0}$ can be computed by *un-standardizing* the following r.v., which follows a *Student's t* distribution with $n-1$ d.f.

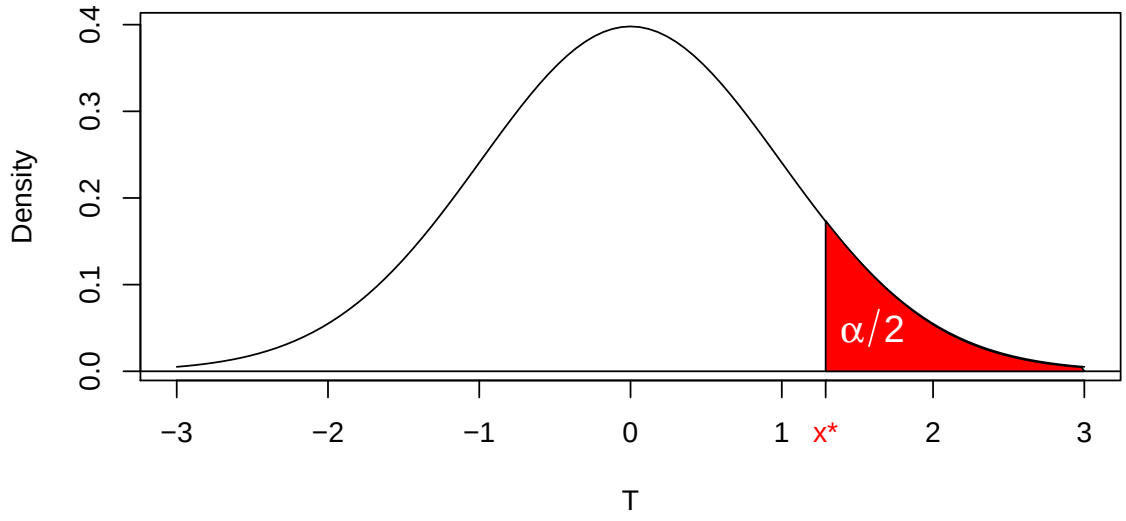
$$T = \frac{\widehat{\beta_0} - \beta_0}{\text{se}(\widehat{\beta_0})}$$

Recall $S = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$ and $S_{XX} = \sum_{i=1}^{n}(x_i - \bar{x})^2$, and that the standard error of $\widehat{\beta_0}$ is $\text{se}(\widehat{\beta_0}) = S\sqrt{\frac{1}{n}\frac{\bar{x}^2}{S_{XX}}}$.

CI: 100(1-α)% of the time, *parameter* β_0 is in

$$\left[\widehat{\beta_0} + qt(\tfrac{\alpha}{2}, n-2)\text{se}(\widehat{\beta_0}),\ \ \widehat{\beta_0} + qt(1-\tfrac{\alpha}{2}, n-2)\text{se}(\widehat{\beta_0})\right]$$

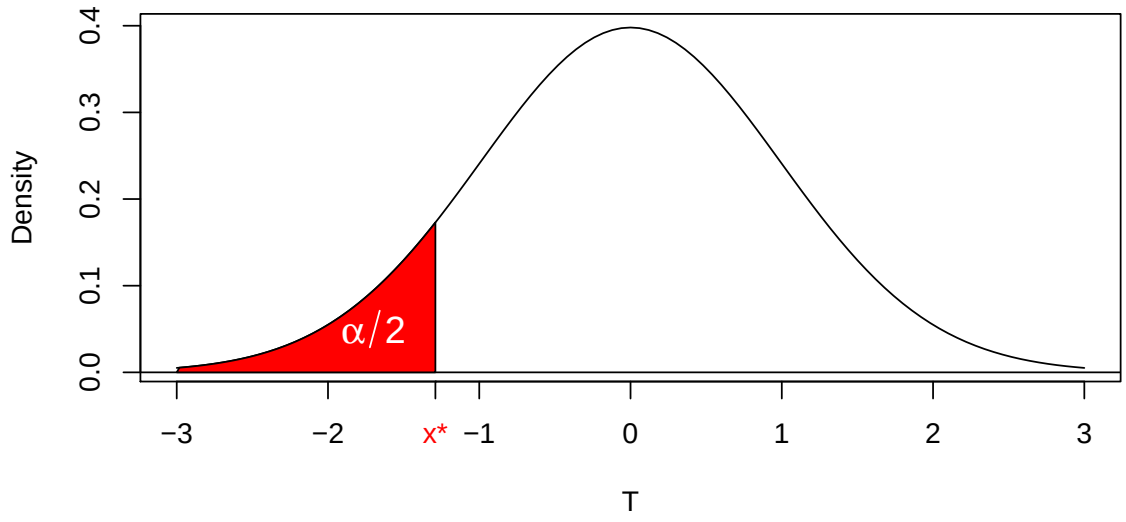
Note: t() vs qt()



The T-score (and Z-score) tables found in classical textbooks tell you the *random variable value* x_* that satisfy

$$P(X > x_*) = \alpha/2$$

Our textbook uses $x_* = t(\alpha/2, n-2)$ to denote these "upper tail" values. **However, in R...**



... we compute these values as **quantiles**. For example, the 25% quantile, x_*, satisfies $P(X \leq x_*) = 0.25$.
Thus, to clarify typical **textbook** vs **R** notation:

$$-t(\alpha/2, n-2) = \texttt{qt}(\alpha/2, n-2)$$
$$t(\alpha/2, n-2) = \texttt{qt}(1-\alpha/2, n-2)$$

Recap: Confidence Intervals

Slope CI: The distribution of the slope estimator $\widehat{\beta}_1$ can be computed by *un-standardizing* the t_{n-1} distributed r.v.

$$T = \frac{\widehat{\beta}_1 - \beta_1}{\text{se}(\widehat{\beta}_1)}$$

Recall $S = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$ and $S_{XX} = \sum_{i=1}^{n}(x_i - \bar{x})^2$, and that the standard error of $\widehat{\beta}_1$ is $\text{se}(\widehat{\beta}_1) = S/\sqrt{S_{XX}}$.

CI: 100(1-α)% of the time, *parameter* β_1 is in

$$\left[\widehat{\beta}_1 + qt(\tfrac{\alpha}{2}, n-2)\text{se}(\widehat{\beta}_1),\ \ \widehat{\beta}_1 + qt(1-\tfrac{\alpha}{2}, n-2)\text{se}(\widehat{\beta}_1)\right]$$

Recap: Confidence Intervals

Regression Line CI: The distribution of $\widehat{y_*} = \widehat{\beta_0} + \widehat{\beta_1}\, x_*$ (or, $\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1}\, x_i$) can be computed by *un-standardizing* the t_{n-1} distributed r.v.

$$T = \frac{\widehat{y_*} - (\beta_0 + \beta_1\, x_*)}{S\sqrt{\frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{XX}}}}$$

Thus,

CI: 100(1-α)% of the time, $E(\hat{y}_*) = \beta_0 + \beta_1\, x_*$ is in

$$\widehat{y_*} \pm qt(1 - \tfrac{\alpha}{2}, n-2) S\sqrt{\frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{XX}}}$$

Recap: Prediction Intervals

Prediction Interval: The distribution of $\widehat{Y_*} = \widehat{\beta_0} + \widehat{\beta_1}\, x_*$ (or, $\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1}\, x_i$) can be computed by *un-standardizing* the t_{n-1} distributed r.v.

$$T = \frac{\widehat{y_*} - (\beta_0 + \beta_1\, x_*)}{S\sqrt{1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{XX}}}}$$

Note the CI for $E(\widehat{y_*})$ uses se$(\widehat{y_*})$ while the prediction interval uses se$(\widehat{Y_* - y_*})$. Thus,

CI: 100(1-α)% of the time, $E(\hat{y}_*) = \beta_0 + \beta_1\, x_*$ is in

$$\widehat{y_*} \pm qt(1 - \tfrac{\alpha}{2}, n-2) S\sqrt{1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{XX}}}$$

Analysis of Variance

Observe that, if $\beta_1 = 0$ then the SLR model becomes

$$Y = \beta_0 + \epsilon \sim Normal(\beta_0, \sigma)$$

and so $\widehat{\beta_0} = \widehat{y} = \bar{y}$. To test for a significant linear relation ship, we test against this null hypothesis ($H_0 : \beta_1 = 0$) using

$$T = \frac{\widehat{\beta_1} - 0}{\text{se}(\widehat{\beta_1})} \sim t_{n-2}$$

In multiple regression, however, we need to generalize. This leads us to a different test statistic...

Analysis of Variance

Q: How much of the variation in y_i values comes from the linear component?

$$SST = S_{YY} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SS_{reg} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

It can be shown (see next slide) that

$$\overbrace{SST}^{\text{Total variation in Y}} = \overbrace{SS_{reg}}^{\text{SS explained by regression}} + \overbrace{RSS}^{\text{residuals}}$$

Proof that $SST = SS_{reg} + RSS$:

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\overbrace{y_i - \widehat{y}_i}^{e_i} + \widehat{y}_i - \bar{y})^2$$

$$= \sum_{i=1}^{n} e_i^2 + (\widehat{y}_i - \bar{y})^2 + 2\, e_i(\widehat{y}_i - \bar{y}) = SS_{reg} + RSS + 2\sum_{i=1}^{n} e_i(\widehat{y}_i - \bar{y})$$

However, from our derivation of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ (see textbook pg. 18), recall that $\sum_{i=1}^{n} e_i = 0$ and $\sum_{i=1}^{n} x_i\, e_i = 0$. Thus we see that

$$\sum_{i=1}^{n} e_i\,(\widehat{y}_i - \bar{y}) = \sum_{i=1}^{n} \widehat{y}_i e_i \; - \bar{y} e_i \; = \sum_{i=1}^{n}(\widehat{\beta}_0 + \widehat{\beta}_1\, x_i)\, e_i - \bar{y}\sum_{i=1}^{n} e_i$$

$$= \widehat{\beta}_0 \sum_{i=1}^{n} e_i + \widehat{\beta}_1 \sum_{i=1}^{n} x_i\, e_i - \bar{y}\sum_{i=1}^{n} e_i = 0.$$

Analysis of Variance

Generalizing the t test above, we can also get a p-value for the hypothesis test

$$H_0 \ : \ \beta_1 = 0 \quad vs \quad H_A \ : \ \beta_1 \neq 0$$

by using a more general test statistic that quantifies how much variation in Y results from the linear trend relative to the random variation determined by the magnitude of σ:

$$F = \frac{SS_{reg}/1}{RSS/(n-2)}$$

F has an $F_{1,n-2}$ distribution, and will be revisited in Ch. 5.

Reading Data

More at:
- www.r-tutor.com/r-introduction/data-frame/data-import

- www.datacamp.com/community/tutorials/r-data-import-tutorial

and of course...
cran.r-project.org/doc/manuals/r-release/R-data.html

Example:

Download: anascombe-xl.R and anascombe.xlsx

```
## Error in
setwd("C:/Users/MrStandardUser/Dropbox/UNR2015/MathStat-757-applied-regression/
cannot change working directory

## Based on http://blog.rstudio.org/2015/04/15/readxl-0-1-0/
library(readxl) # make sure to install.packages("readxl") if needed!
# Anascombe's data from the textbook. Each data set in it's own sheet:
excel_sheets("anascombe.xlsx")

## Error:  'anascombe.xlsx' does not exist in current working directory
('C:/Users/phurtado/Dropbox/UNR2015/Teaching/Hurtado-Math420-FA17/slides').

# Load sheet1
xydat=read_excel("anascombe.xlsx") # defaults to sheet=1

## Error:  'anascombe.xlsx' does not exist in current working directory
('C:/Users/phurtado/Dropbox/UNR2015/Teaching/Hurtado-Math420-FA17/slides').

class(xydat)

## Error in eval(expr, envir, enclos):  object 'xydat' not found
```

Recall SLR Assumptions

By assuming the SLRM, you assume...

1. *All* data follow $Y|X = x_i \sim N(\beta_0 + \beta_1 x_i,\ \sigma)$, hence $E(Y|X = x_i) = \beta_0 + \beta_1 x_i$
2. Normal errors: $e_i \sim N(0, \sigma)$
3. Independent errors e_i
4. $Var(Y|X = x_i) = Var(e_i) = \sigma^2$

Do Those Assumptions Hold?

Test using...

1. Residuals and Standardized Residuals
2. Leverage
3. Outliers
4. Correlations, etc...

Re

```
## Error in plot(xydat$x, xydat$y - 3 - 0.5 * xydat$x, xlab = "x", ylab
= "Residuals"):  object 'xydat' not found
```

Leverage

We can quantify **leverage** with h_{ii}, where

$$\text{mean}(h_{ii}) = \frac{2}{n}$$

where a high leverage point is 2x that mean, i.e. $> 4/n$.

Checking Assumptions

Remember: Estimates, confidence intervals, p-values, etc. **are all meaningless** if you're using the wrong model!

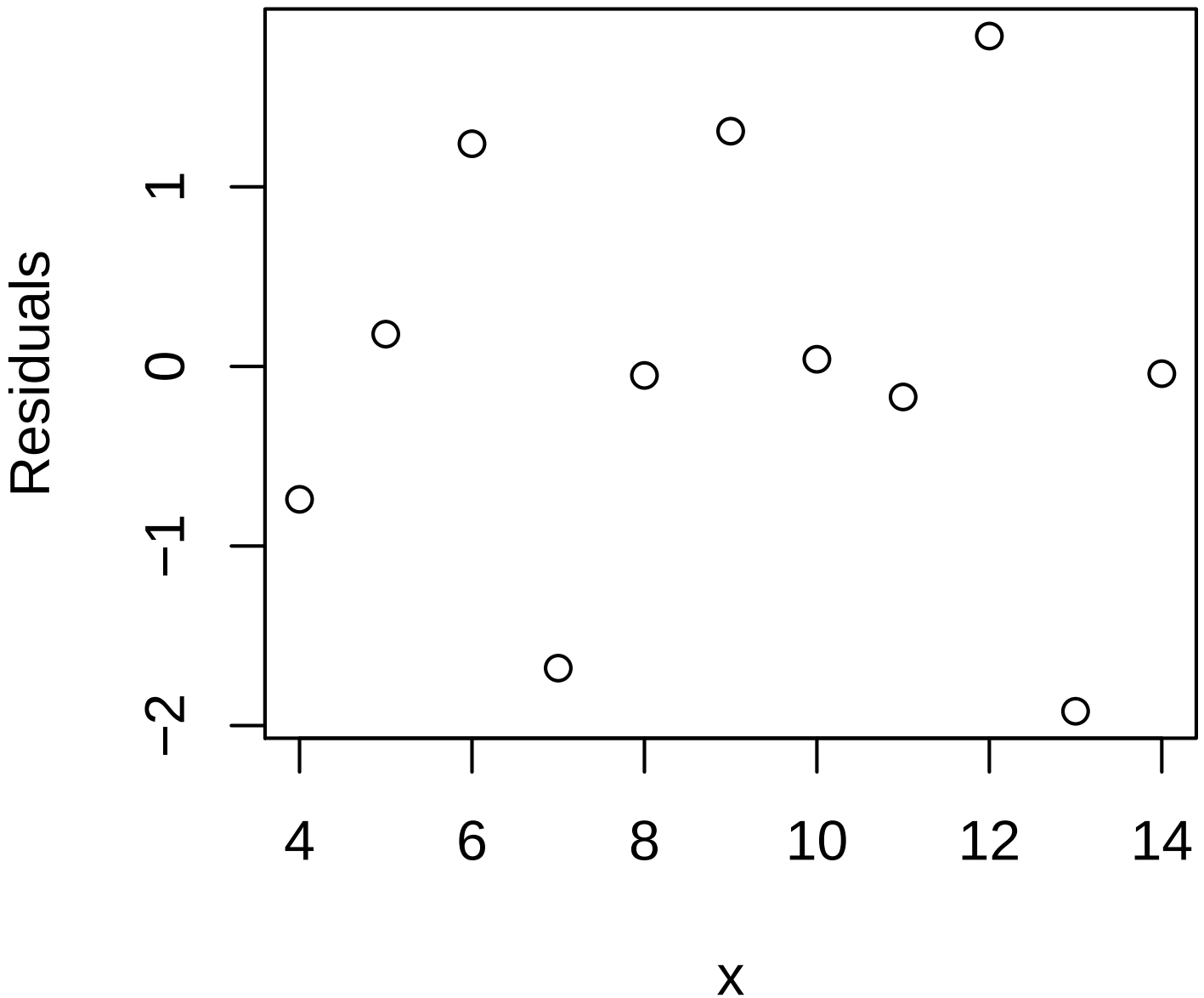
Diagnostics help identify violations of your model assumptions.

SLR Model Assumptions:

1. *All* data follow $Y|X = x_i \sim N(\beta_0 + \beta_1 x_i,\ \sigma)$, hence $E(Y|X = x_i) = \beta_0 + \beta_1 x_i$
2. Normal errors: $e_i \sim N(0, \sigma)$
3. Independent errors e_i
4. $Var(Y|X = x_i) = Var(e_i) = \sigma^2$

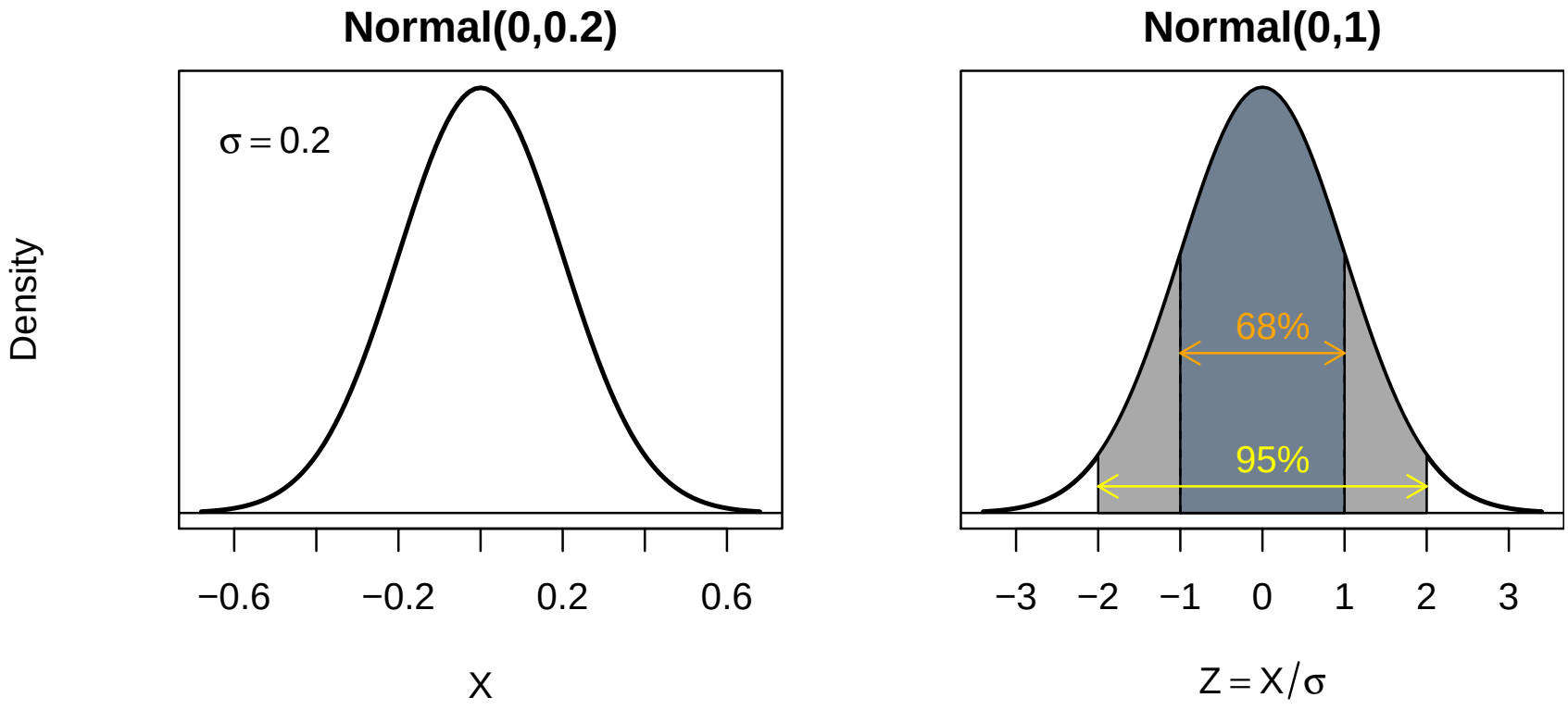
Many problems lead to **outliers** and **high leverage** points.

Residuals $\widehat{e_i} = y_i - \widehat{y_i} \approx e_i$



Standardized Residuals

Recall that $e_i \sim N(0, \sigma)$, which means that standardizing $z_i = e_i/\sigma$ (by dividing by the standard deviation) would yield values that follow a Normal(0,1) distribution (if we knew σ!):



Leverage & "Hat" values (h_{ij})

Observe that

$$\widehat{y_i} = \sum_{j=1}^{n} h_{ij} y_j$$

where

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{k=1}^{n}(x_k - \bar{x})^2},$$

Leverage & "Hat" values (h_{ij})

Observe that

$$\widehat{y}_i = \sum_{j=1}^{n} h_{ij} y_j$$

where

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{k=1}^{n}(x_k - \bar{x})^2}, \quad \sum_{j=1}^{n} h_{ij} = 1,$$

Leverage & "Hat" values (h_{ij})

Observe that

$$\widehat{y}_i = \sum_{j=1}^{n} h_{ij} y_j$$

where

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{k=1}^{n}(x_k - \bar{x})^2}, \quad \sum_{j=1}^{n} h_{ij} = 1, \quad \text{and} \quad \sum_{i=1}^{n} h_{ii} = 2$$

Leverage & "Hat" values (h_{ij})

Observe that

$$\widehat{y}_i = \sum_{j=1}^{n} h_{ij} y_j$$

where

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{k=1}^{n}(x_k - \bar{x})^2}, \quad \sum_{j=1}^{n} h_{ij} = 1, \quad \text{and} \quad \sum_{i=1}^{n} h_{ii} = 2$$

We call h_{ii} the **leverage** of the i^{th} data point.

Note $\overline{h_{ii}} = \frac{2}{n}$. A **high leverage** point is 2x that mean: $h_{ii} > \frac{4}{n}$.

"Hat" values (h_{ij})

Side Note: These "hat" values form a matrix **H** which gives

$$\widehat{\mathbf{y}} = \mathbf{H}\,\mathbf{y}$$

and these values show up in many places!

- $\text{Var}(\widehat{y}_i) = \sigma^2 h_{ii}$
- Alternative definition: $h_{ij} = \frac{cov(\widehat{y}_i, y_j)}{var(y_j)}$
- Residuals, in matrix notation: $\mathbf{r} = (I - \mathbf{H})\mathbf{y}$
- Properties: $\mathbf{H}$ is symmetric, $\mathbf{H}^2 = \mathbf{H}$, $\mathbf{HX} = \mathbf{X}$
- Similar **H** matrices for other models may not have all these properties.

Want more? See online resources and publications such as
Hoaglin and Welsch. 1978. The Hat Matrix in Regression and ANOVA.
http://www.stat.ucla.edu/~cocteau/stat201b/handout/hat.pdf

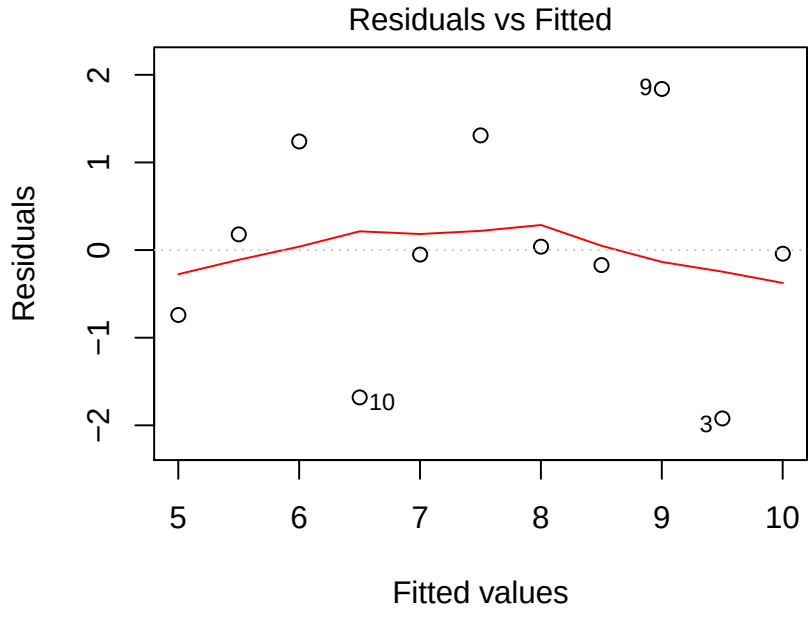
Standardized Residuals

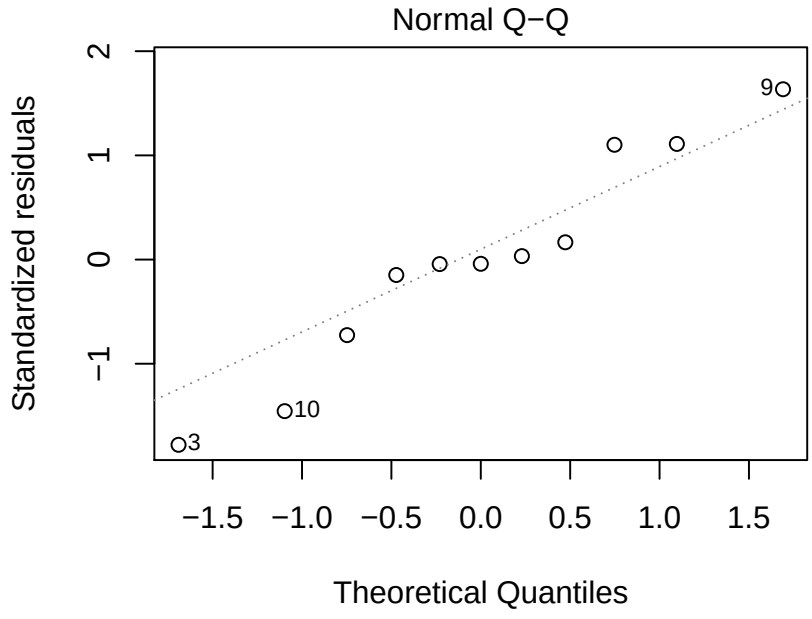
Recall $e_i/\sigma \sim \text{Normal}(0,1)$, **BUT** we don't know σ!

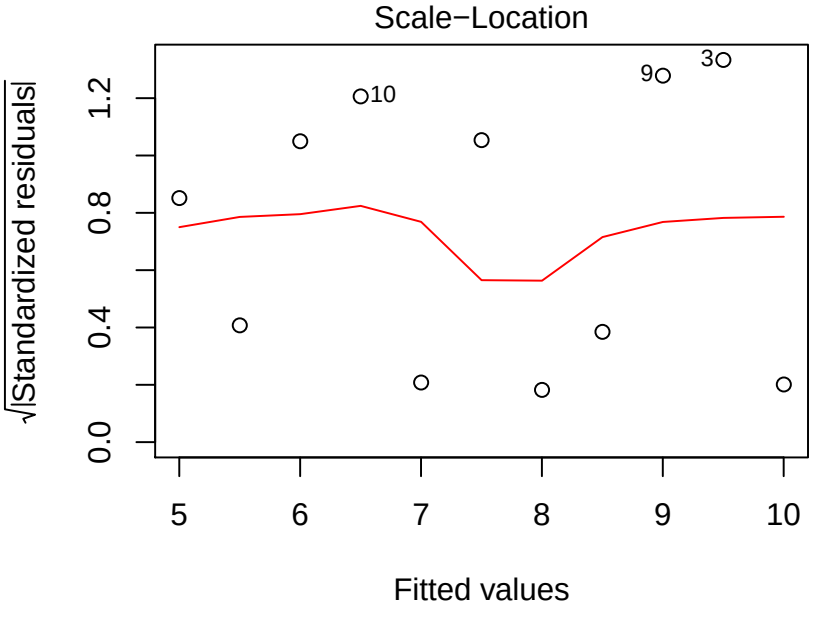
Using our estimate, S, in it's place (and some algebra to show that $\text{Var}(\widehat{e}_i) = \sigma^2(1 - h_{ii})$) yields **standardized residuals** r_i:

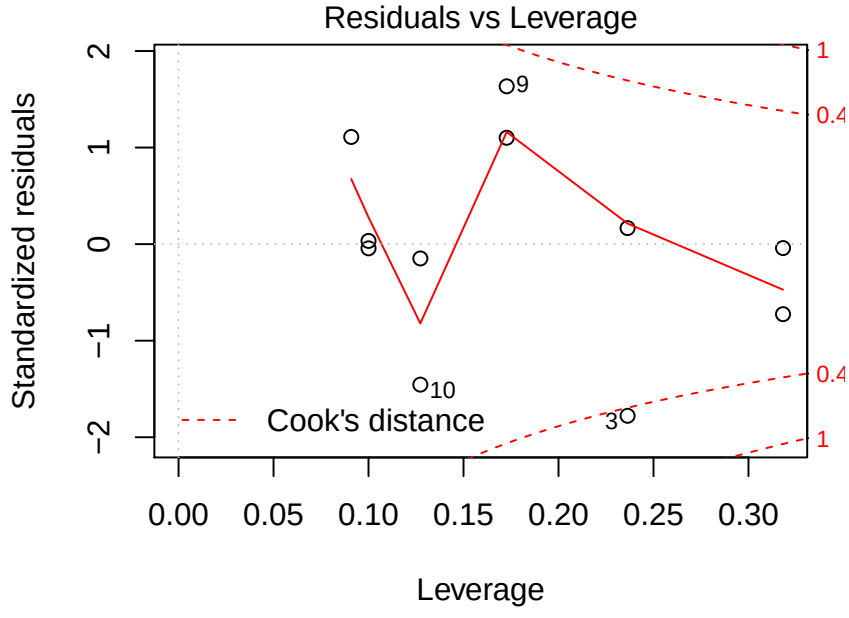
$$r_i = \frac{\widehat{e}_i}{S\sqrt{1 - h_{ii}}}$$

These can be more informative to look at than residual plots, especially if high leverage points exist.


Residuals vs Fitted
Residuals
Fitted values
9
10
3


Normal Q−Q
Standardized residuals
Theoretical Quantiles
9
10
3

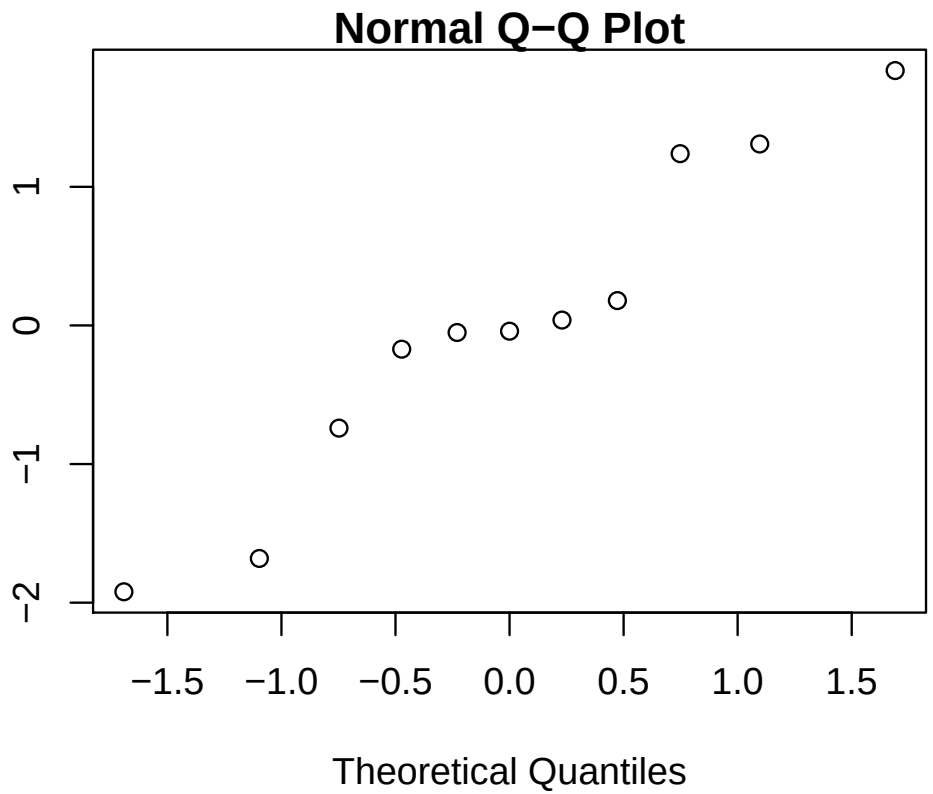

Scale−Location
√|Standardized residuals|
Fitted values
9
10
3

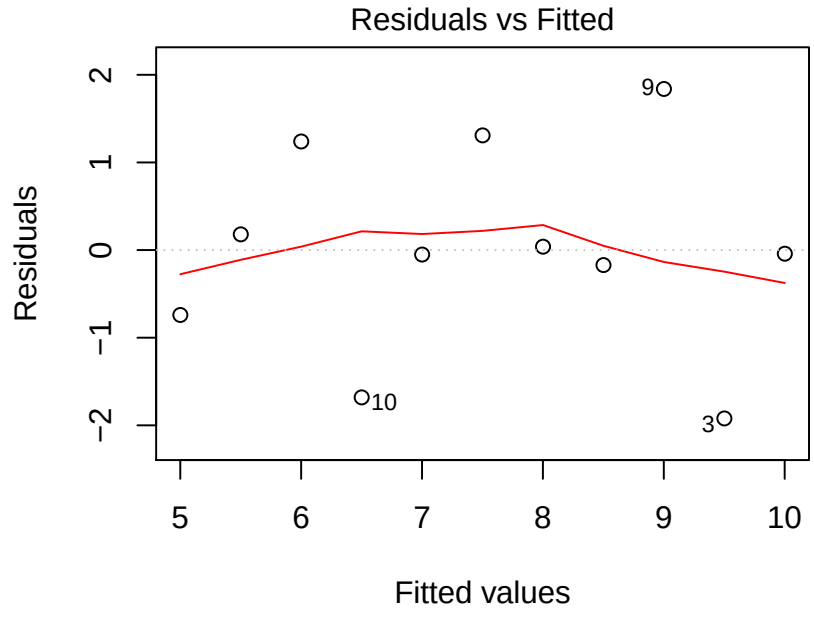

Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
9
10
3

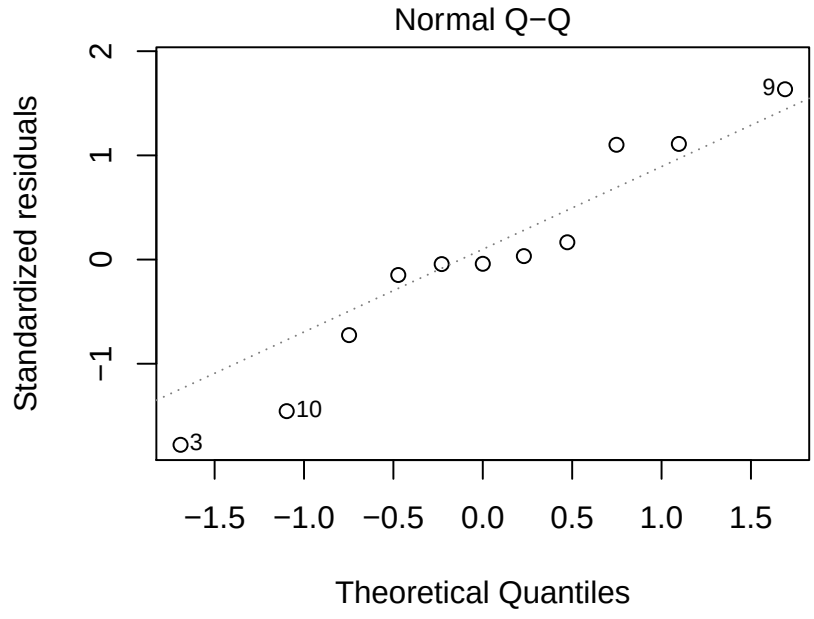
Normal Quantile-Quantile Plots

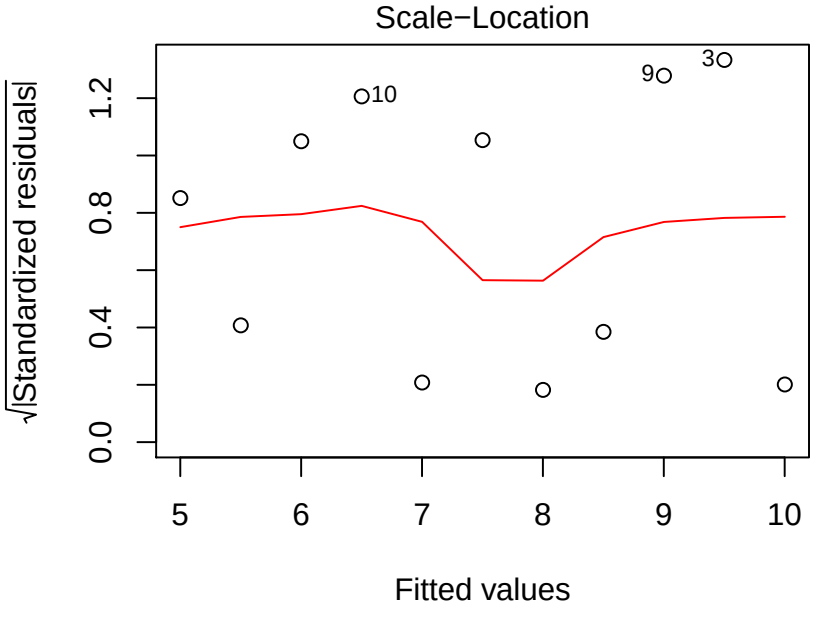
In place of a **Shapiro-Wilk** test, plot Standardized Residuals versus the Expected Values of the Order Statistics for a Normal(0,1) distribution. See `shapiro.test()` & `qqnorm()`.

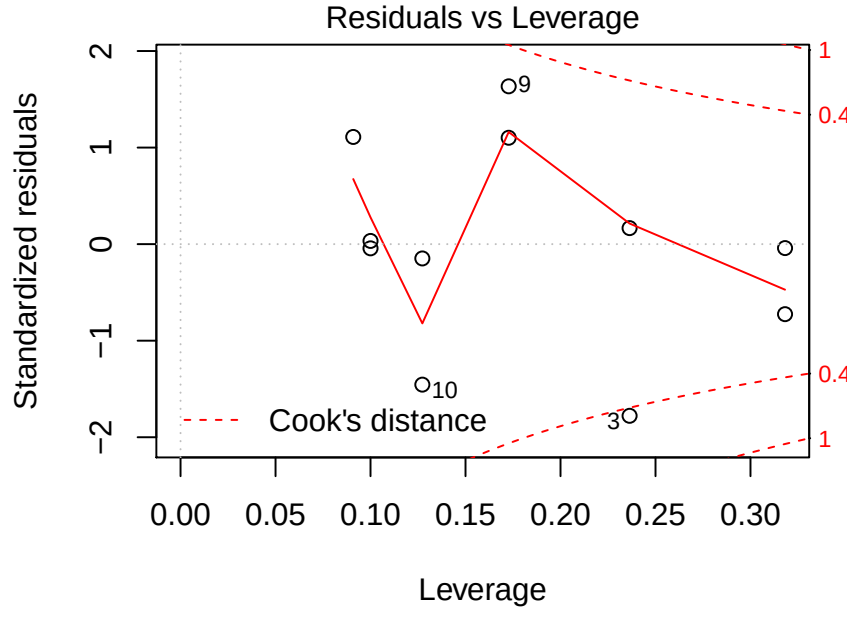
```
qqnorm(fit1$residuals)
```




Residuals vs Fitted
Residuals
Fitted values
9
10
3


Normal Q−Q
Standardized residuals
Theoretical Quantiles
9
10
3


Scale−Location
√|Standardized residuals|
Fitted values
9
10
3


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
9
10
3

Leave-one-out Diagnostics

Another approach to identifying problem data points (with problematic *influence*) is to compare estimates with and without them. For example, if $\widehat{y_{j(i)}}$ is the estimate of $\widehat{y_j}$ with the j^{th} data point removed...

Cook's Distance:

$$D_i = \frac{\sum_{i=1}^{n}(\widehat{y_{j(i)}} - \widehat{y_j})^2}{2S^2} = \cdots = \frac{r_i^2}{2}\frac{h_{ii}}{1-h_{ii}}$$

Roughly speaking, scrutinize points with $D_i > \frac{4}{n-2}$ or values that deviate markedly from the other distances.

Summary Remark

"Bad" leverage points are **high leverage** points that are also **outliers** – they signal a problem with your model!

The two main approaches to fixing that problem:

1. **Omit the data point** from the data set, or
2. Redo your analysis using a **more appropriate model**. This is often the preferred approach.