201ab Quantitative methods
L.13 Linear model: Desiderata and Diagnostics

ANOVA Desiderata

- Peculiar designs: Imbalance, one data point per cell
- Effect sizes
- Power

Multicolinearity in unbalanced ANOVA

	North Korea	USA
Male	67 66 64 64 68 67 69 70 65	74 83
Female	64 68	59 63 68 60 64 67 62 59 68 69

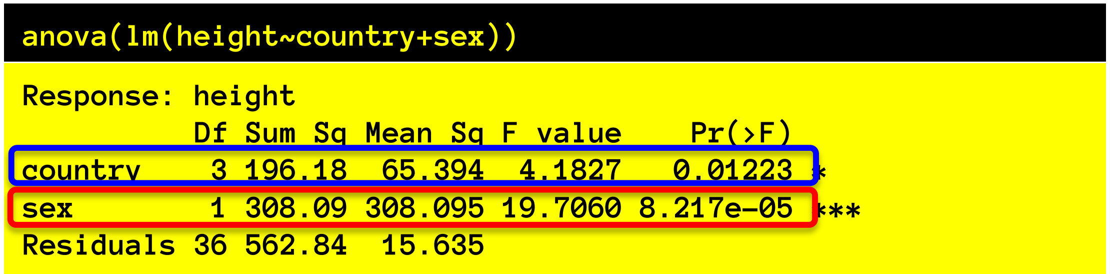
Unbalanced design: different ns in different cells, so factors are not independent, so we have multicolinearity, and a credit assignment problem.

Multicolinearity effects: Contamination across main effects, and order-dependence in sum sq. allocation.

Type I sums of squares (R default)

SS for factor 1: SSR[factor1]

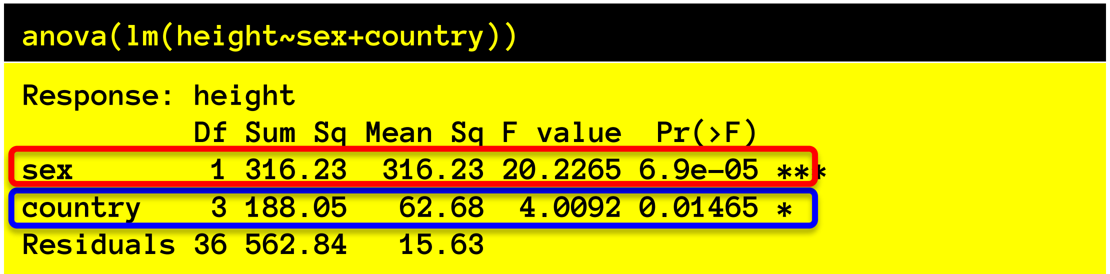
SS for factor 2: SSR[factor2 | factor 1]



```
anova(lm(height~country+sex))

Response: height
          Df Sum Sq Mean Sq F value    Pr(>F)
country    3 196.18  65.394  4.1827   0.01223 *
sex        1 308.09 308.095 19.7060 8.217e-05 ***
Residuals 36 562.84  15.635
```

SSR[country] and SSR[sex|country]



```
anova(lm(height~sex+country))

Response: height
          Df Sum Sq Mean Sq F value  Pr(>F)
sex        1 316.23  316.23 20.2265 6.9e-05 ***
country    3 188.05   62.68  4.0092 0.01465 *
Residuals 36 562.84   15.63
```

SSR[sex] and SSR[country|sex]

Type II and III sums of squares, calculate SS for a given factor controlling for other stuff. II and III do not depend on order, but also don't preserve the SST = sum(all SS). Type III is default in SPSS. They implicitly test slightly different null hypotheses.

One observation per cell.

	North Korea	USA
Male	67	74
Female	64	59

- If we have one observation per cell, the interaction *is* the error.
- Therefore, if we include interaction in the model, we have no error left over (data points do not deviate at all from cell means).
  - Also n = # of parameters… so df error is 0…
- So we can't compute any F ratios or ascertain significance.
- Solution: omit interaction term, then that variance will be error, and you can assess main effects.

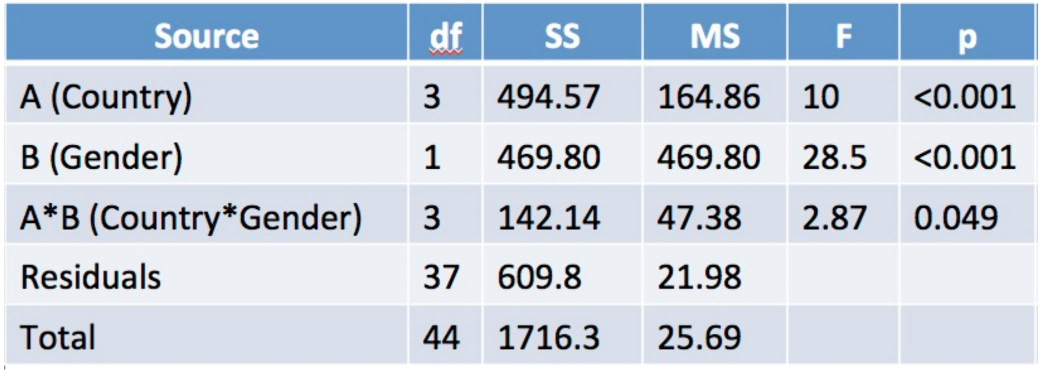
ANOVA effect size

Percent variance accounted for….

- Counterpart of R^2:
  η^2 “eta squared”

$$\eta_A^2 = \frac{SS[A]}{SST}$$

$$\eta_A^2 = \frac{494.57}{1716.3} = 0.288$$



Source	df	SS	MS	F	p
A (Country)	3	494.57	164.86	10	<0.001
B (Gender)	1	469.80	469.80	28.5	<0.001
A*B (Country*Gender)	3	142.14	47.38	2.87	0.049
Residuals	37	609.8	21.98		
Total	44	1716.3	25.69		

Note that this is equal to full-model R^2 when there is only one factor, but if there is more than one, it will be smaller.

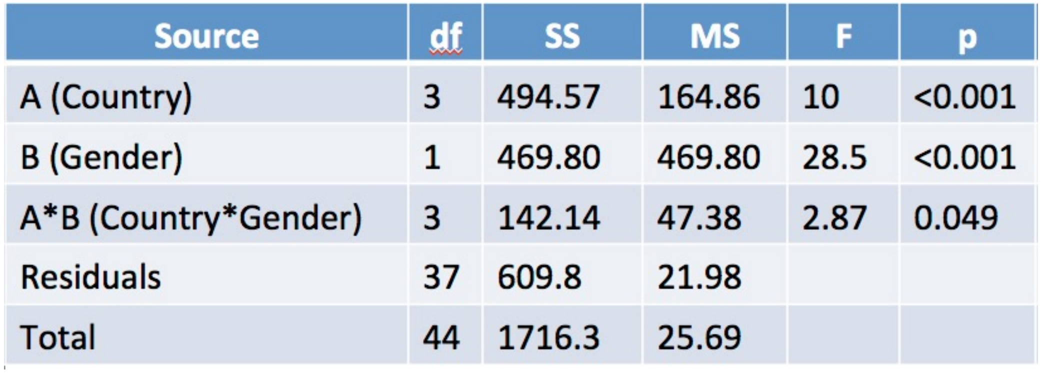
ANOVA effect size

Percent variance accounted for….

- Counterpart of R^2: η^2 "eta squared"

$$\eta_A^2 = \frac{SS[A]}{SST}$$

$$\eta_A^2 = \frac{494.57}{1716.3} = 0.288$$



Source	df	SS	MS	F	p
A (Country)	3	494.57	164.86	10	<0.001
B (Gender)	1	469.80	469.80	28.5	<0.001
A*B (Country*Gender)	3	142.14	47.38	2.87	0.049
Residuals	37	609.8	21.98		
Total	44	1716.3	25.69		

- Partial η^2 (this is like "R^2 everything else constant")

$$partial: \eta_A^2 = \frac{SS[A]}{SS[A] + SS[error]}$$

$$partial: \eta_A^2 = \frac{494.57}{494.57 + 609.8} = 0.448$$

ANOVA effect size

Percent variance accounted for….

- Counterpart of R^2: proportion of all variance
  η^2 "eta squared"

$$\eta_A^2 = \frac{SS[A]}{SST}$$

- Counterpart of partial R^2 : "R^2 everything else constant"
  Partial η^2

$$partial : \eta_A^2 = \frac{SS[A]}{SS[A] + SS[error]}$$

But these measures are not good estimates of the effect size in the population – they are biased because SS[A] includes some variance due to noise…

ANOVA effect size.

- There is a surprisingly large number of candidate effect sizes for an ANOVA, all interrelated, but with slightly different properties.
  - η^2, ω^2, f^2, f, Ψ, …
- What do we want from an effect size?
  - Quantify standardized relationship strength in population (independence from sample size)
  - …in an interpretable way
  - …that we can estimate from a sample
  - …and will allow us to predict power
  - …while generalizing across study designs

My preference: ω² (omega squared)

- Effect size: Variance of signal in population, relative to unexplained variance in population.

$$\omega^2_{Source} = \frac{\sigma^2_{Source}}{\sigma^2_{Source} + \sigma^2_{Error}}$$

- It’s like partial η^2, but is a population property
  - So to generalize across designs, it must assume that variability due to other factors was introduced by the experiment, and will not occur otherwise.
- Partial η^2 overestimates; we need a correction.

$$\hat{\omega}^2_{Source} = \frac{SS[Source] - df_{source} \cdot MS[Error]}{SS[Source] + (N - df_{source}) \cdot MS[Error]}$$

ω² and other measures

$$f_{Source}^2 = \frac{\omega_{Source}^2}{1-\omega_{Source}^2} = \frac{\sigma_{Source}^2}{\sigma_{Error}^2}$$

This is a “signal-to-noise” ratio measurement: Variance of signal divided by variance of noise.

$$f_{Source} = \sqrt{\frac{\omega_{Source}^2}{1-\omega_{Source}^2}} = \frac{\sigma_{Source}}{\sigma_{Error}}$$

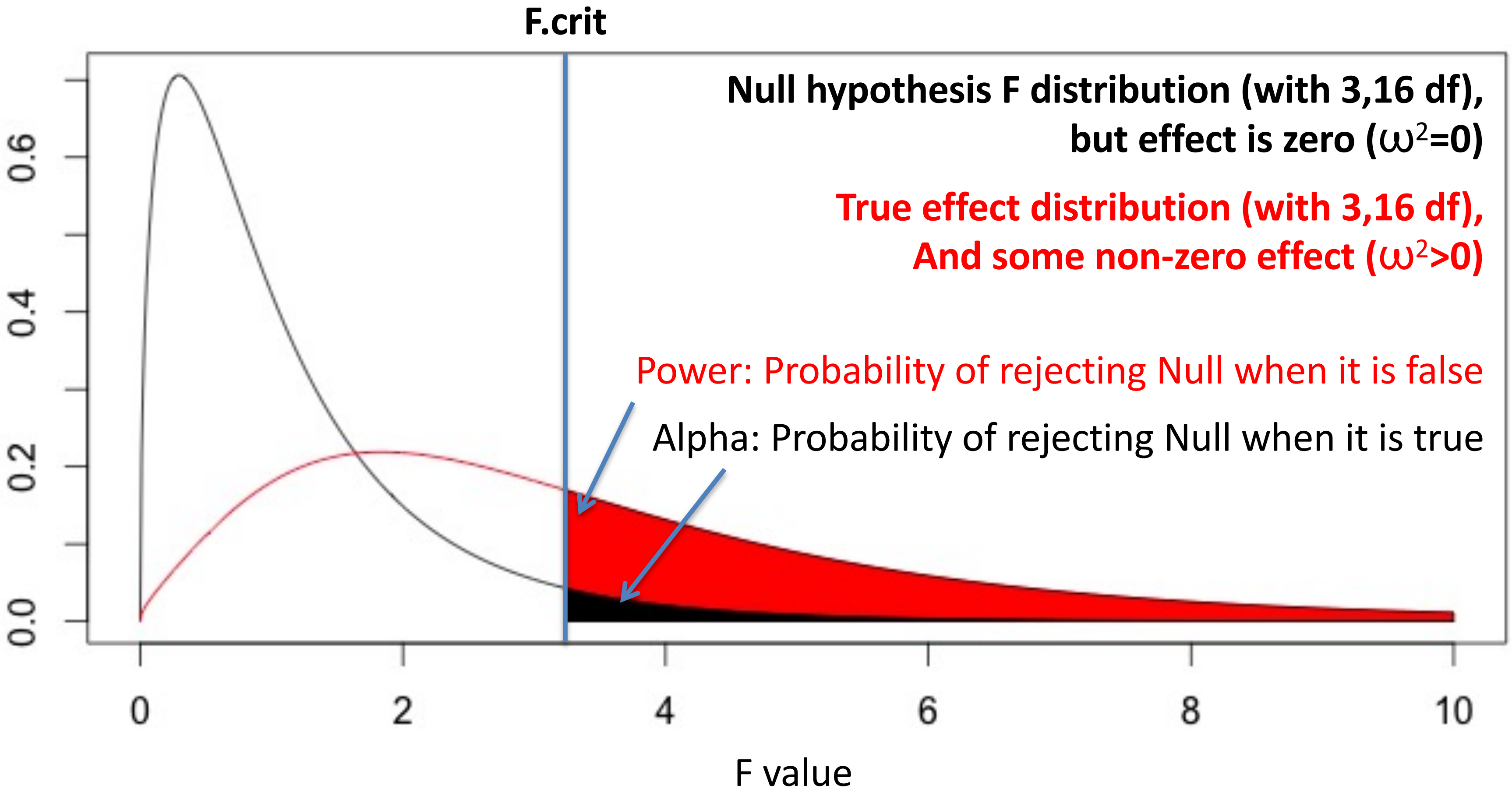
This is a “signal-to-noise” ratio measurement in original (not squared) units, thus is more analogous to Cohen’s d

$$\lambda = \mathrm{N} * f_{Source}^2 = N * \frac{\omega_{Source}^2}{1-\omega_{Source}^2}$$

This is the F distribution “non-centrality parameter” used to describe the distribution of F statistics obtained when samples come from a distribution with some real effect.

What’s a big effect? Some say ω^2=0.15 is big, 0.06 is medium, 0.01 is small.

Power for the F-test



Null hypothesis F distribution (with 3,16 df), but effect is zero (ω^2=0)

True effect distribution (with 3,16 df), And some non-zero effect (ω^2>0)

Power: Probability of rejecting Null when it is false

Alpha: Probability of rejecting Null when it is true

So, to figure out the power of an F test we need to know the sample size, alpha, and true effect.

Power for the F-test

Total number of cells

```
k=4
```

Total (balanced) sample size

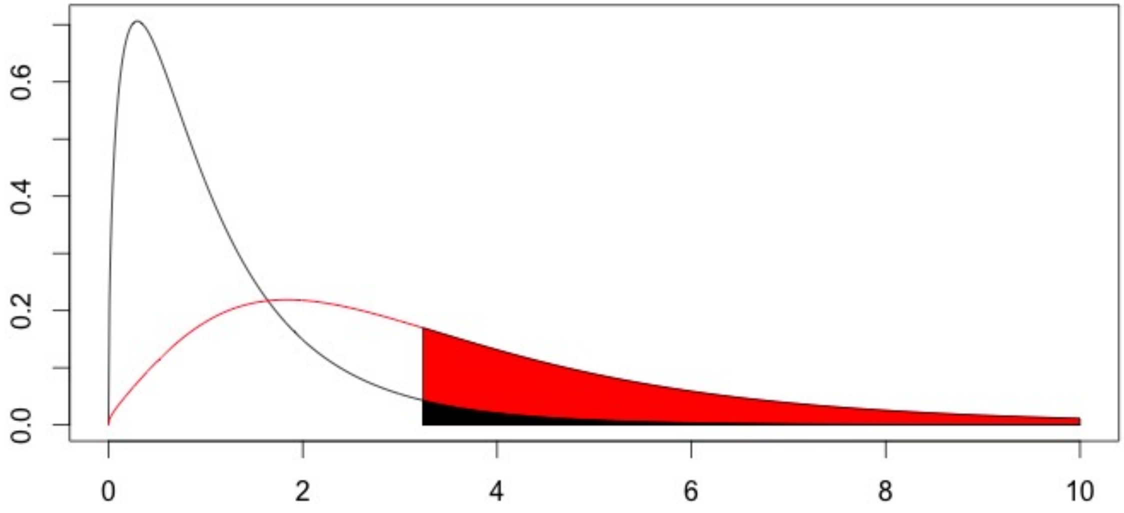
```
N = k*10
```

Effect size (ω^2)

```
w2 = 0.25
```

alpha

```
alpha = 0.05
```



Non-centrality parameter

```
lambda = N*w2/(1-w2)
```

```
[1] 13.33
```

F value at which we reject H0

```
f.crit = qf(1-alpha, k-1, N-k)
```

```
[1] 2.866266
```

Power

```
power = 1-pf(f.crit, k-1, N-k, lambda)
```

```
[1] 0.84
```

Required n for certain power

This is trickier, as changing n changes both the null distribution and the true-effect distribution

```
n = 5
power = 1-pf(qf(0.95, k-1, k*(n-1)), k-1, k*(n-1), n*k*w2/(1-w2))            [1] 0.46

n = 6
power = 1-pf(qf(0.95, k-1, k*(n-1)), k-1, k*(n-1), n*k*w2/(1-w2))            [1] 0.56

n = 7
power = 1-pf(qf(0.95, k-1, k*(n-1)), k-1, k*(n-1), n*k*w2/(1-w2))            [1] 0.65

n = 8
power = 1-pf(qf(0.95, k-1, k*(n-1)), k-1, k*(n-1), n*k*w2/(1-w2))            [1] 0.73

n = 9
power = 1-pf(qf(0.95, k-1, k*(n-1)), k-1, k*(n-1), n*k*w2/(1-w2))            [1] 0.79

n = 10
power = 1-pf(qf(0.95, k-1, k*(n-1)), k-1, k*(n-1), n*k*w2/(1-w2))            [1] 0.84

n = 11
power = 1-pf(qf(0.95, k-1, k*(n-1)), k-1, k*(n-1), n*k*w2/(1-w2))            [1] 0.88
```

So we have to solve for it numerically… I recommend using the pwr R package.

Regression safety tips.

Assumptions:

(1) Validity: Make sure your measures make sense, and map onto the substantive research questions you have.

(2) Additivity and linearity: The relationship between x and y may not be neatly linear, check scatterplots, residuals! Noise and predictors should be additive.

(3) Errors should have equal variance and be normally distributed

(4) Independence of errors: errors should not be correlated with each other, y, x, have repeated measures, etc.

(5) Most error in y, not in x. (parameter estimates biased!)

Safety tips:

(1) Don’t trust extrapolation.

(2) Check for structure in the residuals.

(3) Be careful with causal interpretations.

Assumptions (and when stuff breaks)

- Errors are independent…
  - Violated under repeated measures, sequential / temporal dependence, non-random sampling, etc.
    - Consider: mixed effects, covariates
- …identically distributed…
  - Violated if some conditions have higher variance.
    - Consider: ignoring (if not that different)
    - Consider: log transform (if errors are multiplicative)
- …and Normal.
  - Violated if measure has high skew, kurtosis, floor, ceiling effects.
    - Consider: various transformations.

Multicolinearity of predictors

- If predictors are somewhat colinear (i.e., you can predict one *reasonably well* via linear regression of other ones) you get problems from ambiguous credit assignment:
  - Marginal standard errors of coefficients are inflated (variance inflation factor)
  - Coefficient magnitudes tend to decrease
(but may increase and reverse in some circumstances more when we get to ANCOVA)
  - Coefficient values change erratically with the addition of new predictors, or new data points
(because credit assignment is resolved by the noise)

(if you have perfect colinearity, you can't do the regression at all)

Check via correlation matrices, variance inflation factors.

Looking for multicolinearity

```
summary( lm(son ~ mom+dad+mgma+pgma+mgpa+pgpa) )
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.04727   16.41549   1.709   0.0947 .
mom          0.33379    0.14885   2.243   0.0301 *
dad          0.27050    0.17126   1.580   0.1215
mgma         0.11868    0.19411   0.611   0.5442
pgma         0.21772    0.17592   1.238   0.2226
mgpa        -0.06801    0.23855  -0.285   0.7769
pgpa        -0.23884    0.22878  -1.044   0.3023
```

I designed heavily correlated fake height data for families with a son:
Mom, dad, maternal grand ma, maternal grand pa, paternal grand ma, and paternal grand pa

Calculating correlation of a data frame gives us the full correlation matrix…
So, here it seems that everything is positively correlated.
Some people would use this to drop particular variables, but that's a little silly. In this case, I would suggest making composite indexes.

```
cor(data.frame(son, mom, dad,
                mgma, mgpa, pgma, pgpa))
```

```
      son  mom  dad mgma mgpa pgma pgpa
son  1.00 0.54 0.51 0.43 0.36 0.39 0.23
mom  0.54 1.00 0.37 0.70 0.70 0.12 0.38
dad  0.51 0.37 1.00 0.31 0.35 0.65 0.52
mgma 0.43 0.70 0.31 1.00 0.53 0.09 0.37
mgpa 0.36 0.70 0.35 0.53 1.00 0.14 0.45
pgma 0.39 0.12 0.65 0.09 0.14 1.00 0.25
pgpa 0.23 0.38 0.52 0.37 0.45 0.25 1.00
```

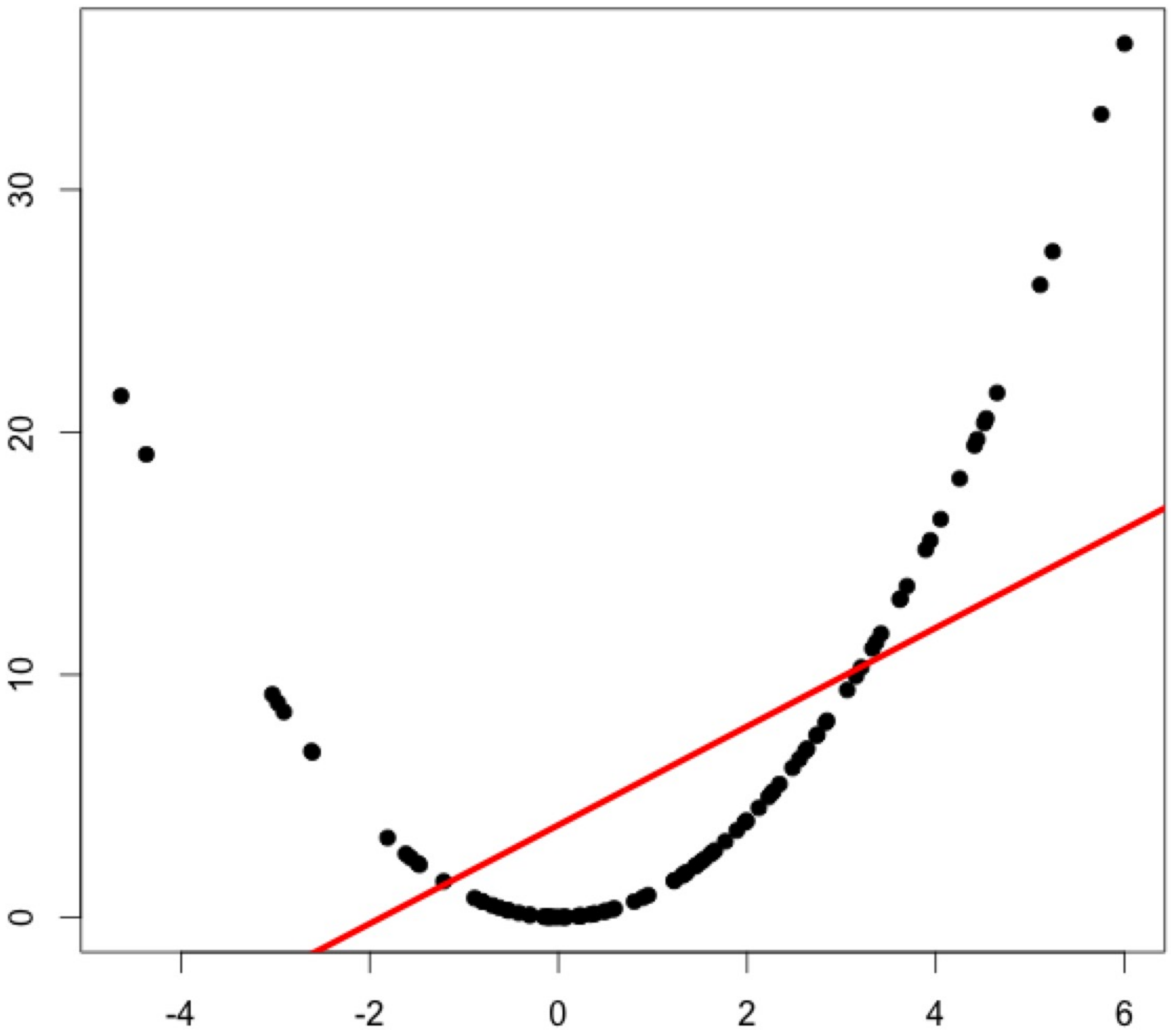
```
vif(lm(son~mom+dad+mgma+pgma+mgpa+pgpa))
```

```
  mom   dad  mgma  pgma  mgpa  pgpa
2.851 2.488 2.020 1.832 2.113 1.603
```

Variance inflation factor: How much larger is coefficient error variance than it would be, if it were independent of other predictors?

Some people advocate specific cutoffs (like vif > 4 or 5 is bad).

Look at the scatterplot!

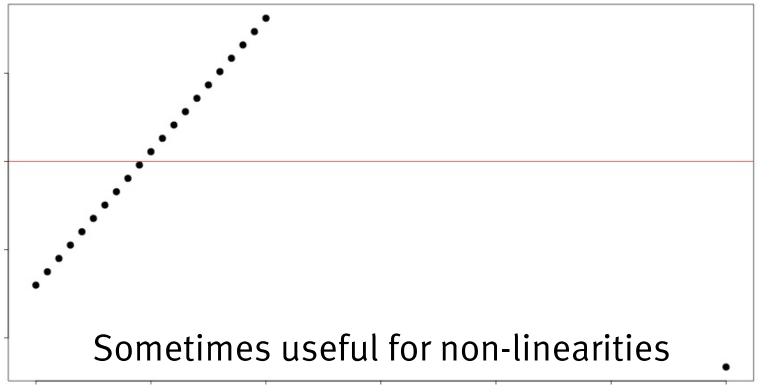


Residual plots

We may see a bit more by looking at different residual plots.

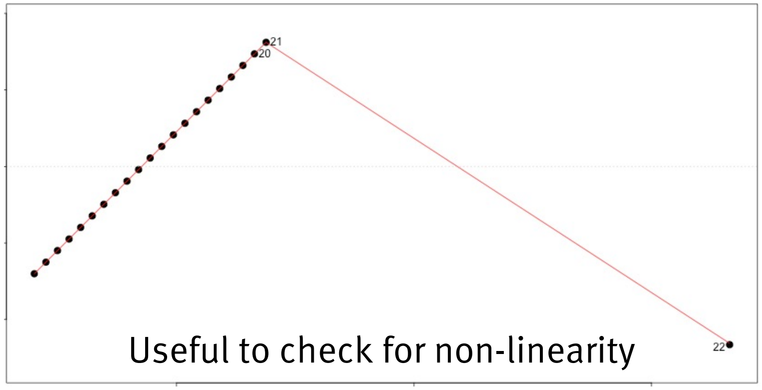
Residuals as a function of x position

(impossible with many predictors)



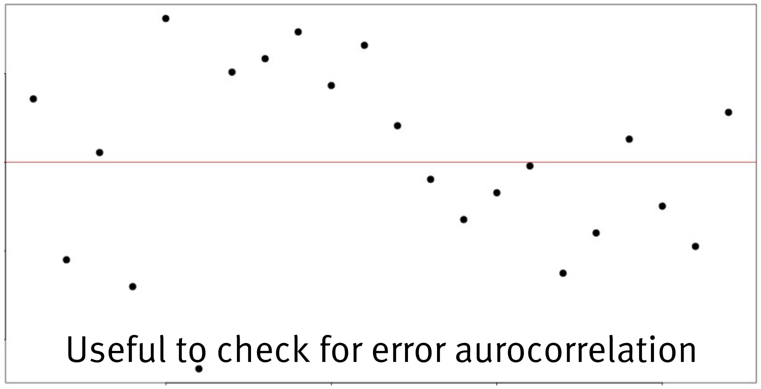
Sometimes useful for non-linearities

Residuals as a function of predicted y plot(lm, 1)



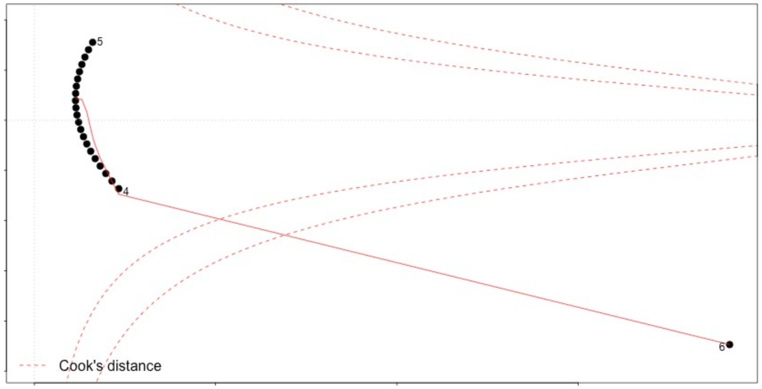
Useful to check for non-linearity

Residuals as a function of observation number



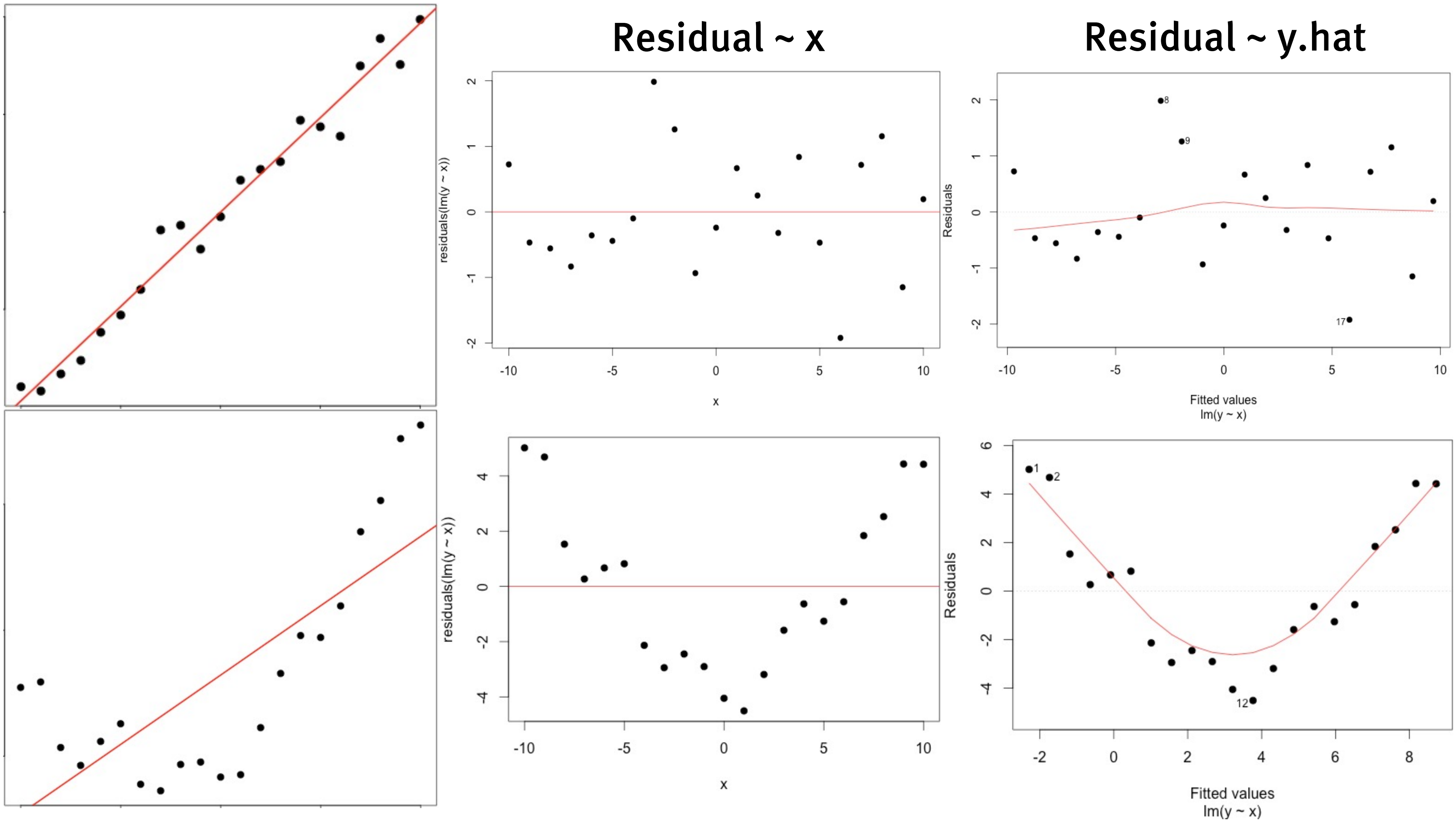
Useful to check for error aurocorrelation

Residuals as a function of leverage plot(lm, 5)



Useful for detecting extreme influence.

Checking for non-linearity



Residual ~ x

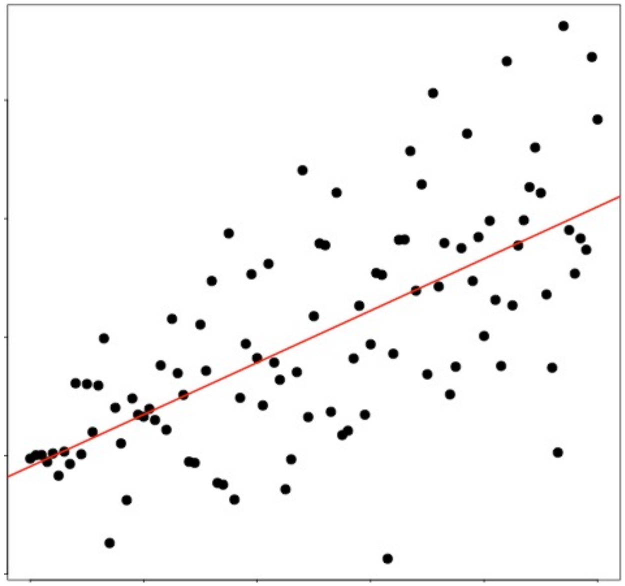
Residual ~ y.hat

Residual plots highlight the non-linearity

For high dimensional data, only Residual ~ y.hat is really possible to look at.

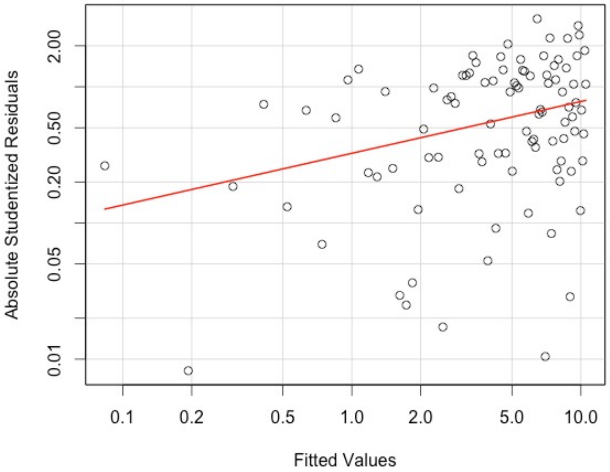
Checking for homoscedasticity

Homoscedasticity: variance of residuals is constant



|residual| ~ y.hat

```
spreadLevelPlot(lm(y~x))
plot(lm, 3)
```



Test for non-constant variance (heteroscedasticity) based on regression of error^2 as a function of fitted y values (for regression): "Breusch-Pagan test"

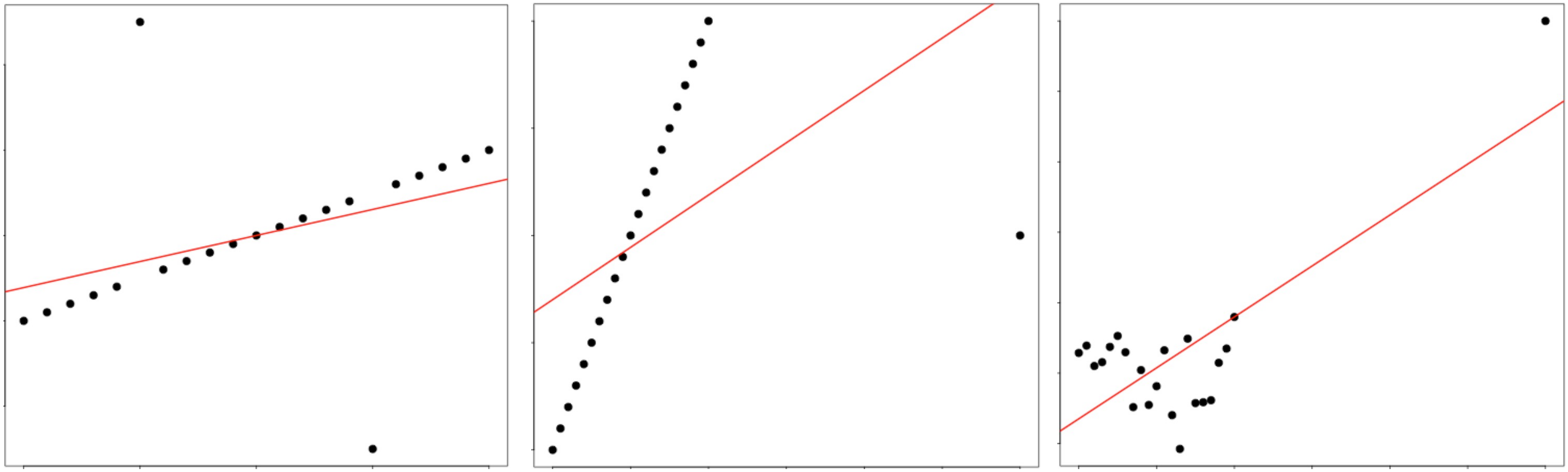
(different, and somewhat more powerful procedure for categorical predictors)

```
ncvTest(lm(y~x))
```

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 10.68375    Df = 1     p = 0.00108081
```

Outliers and extreme influence

Data points with large residuals, and/or high *leverage*



How do we measure this apparent extreme influence?

Outlier detection

```
    qqPlot
    outlierTest
```

Look at residuals as a function of leverage

```
    plot(lm(y~x), which=5)
```

Compute Cook’s distance

```
    plot(lm(y~x), which=4)
```

Studentized / Standardized residuals

Residuals (estimated error)

Deviation of real y value from line

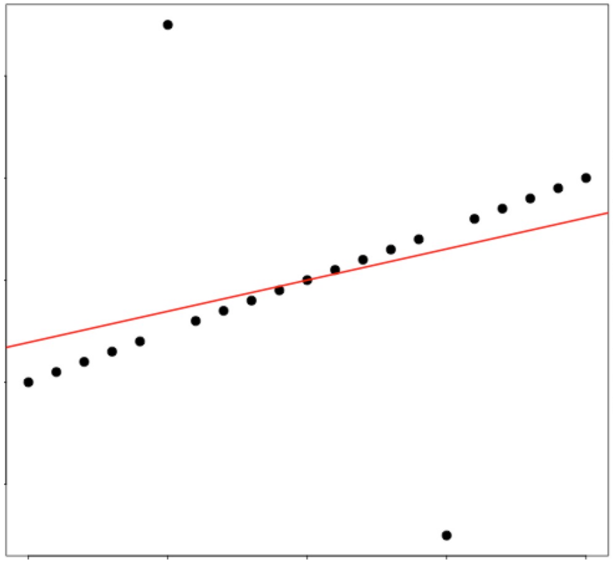
$$\hat{\varepsilon}_i = \left(y_i - \hat{y}_i\right)$$

Standardized residuals

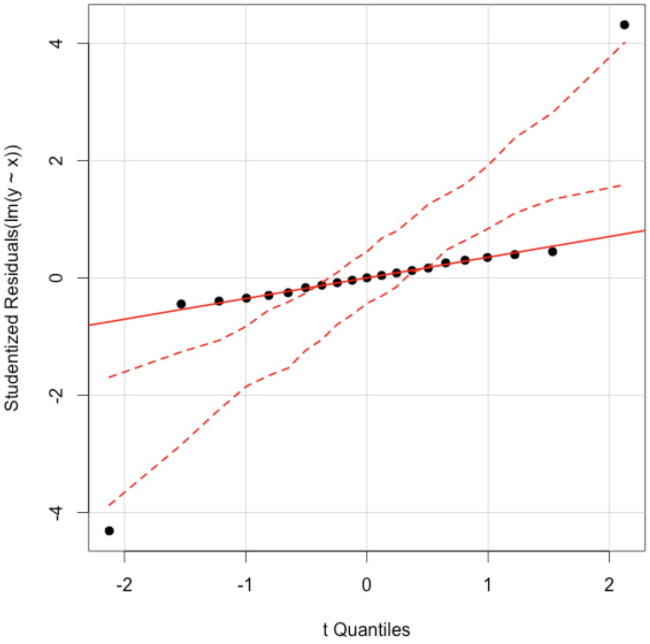
Residual divided by sd of residuals

$$\hat{\varepsilon}_i^{(S)} = \hat{\varepsilon}_i / s_r$$

These should be t distributed, so we can compare to t distribution to look for abnormalities / outliers.



```
qqPlot(lm(y~x))
```

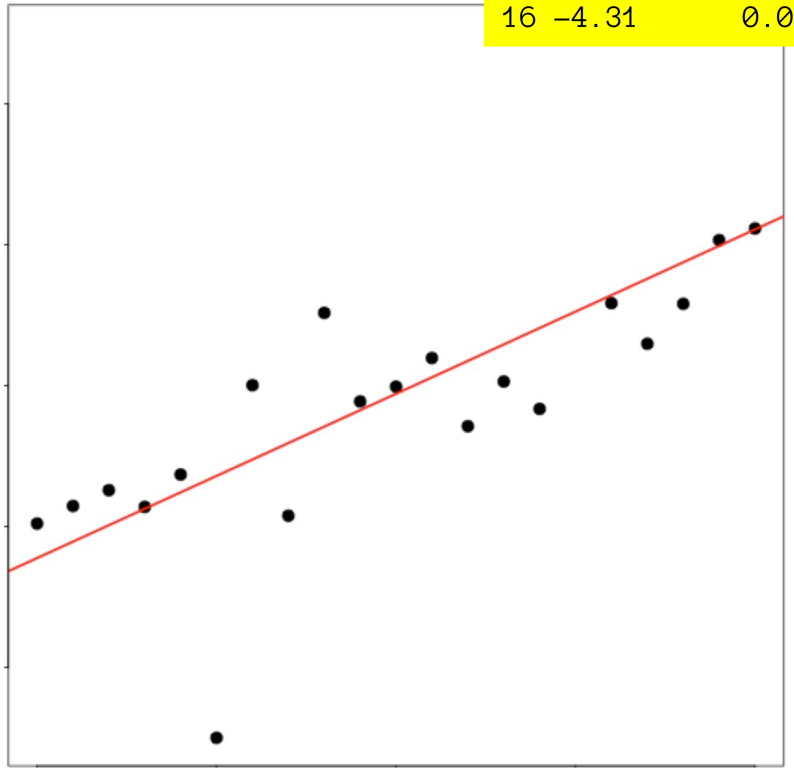


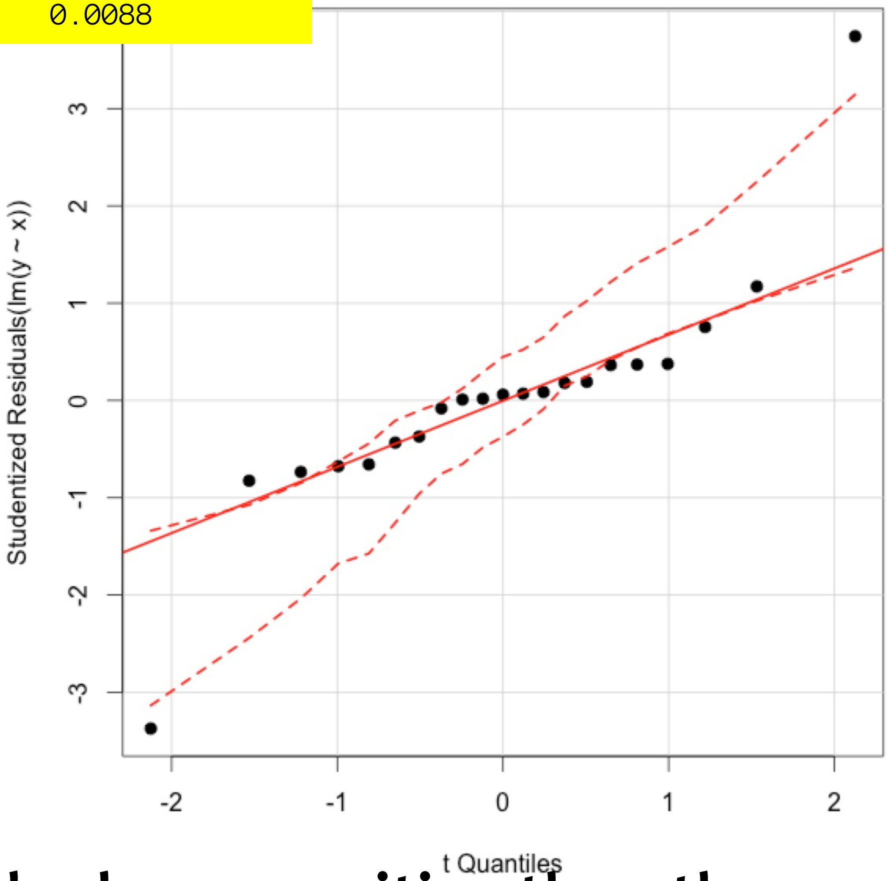
Large deviations from theoretical t distribution can be tested for (via t-test!) and extreme outliers will be evident this way.

Testing for outliers

```
outlierTest(lm(y~x))
```

```
  student    uncorrectedBonferonni
#  error       p-value      p-value
6    4.31        0.0004       0.0088
16 -4.31        0.0004       0.0088
```

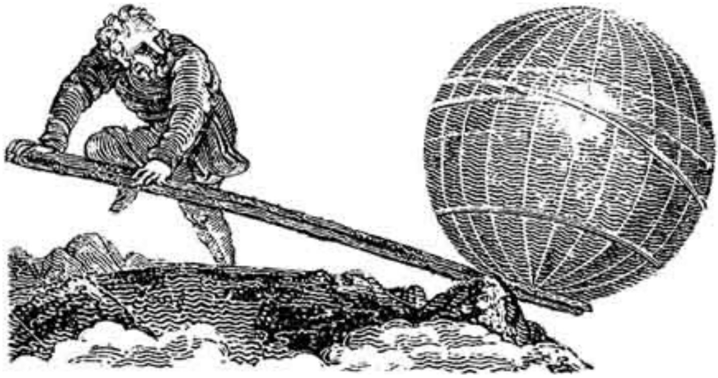




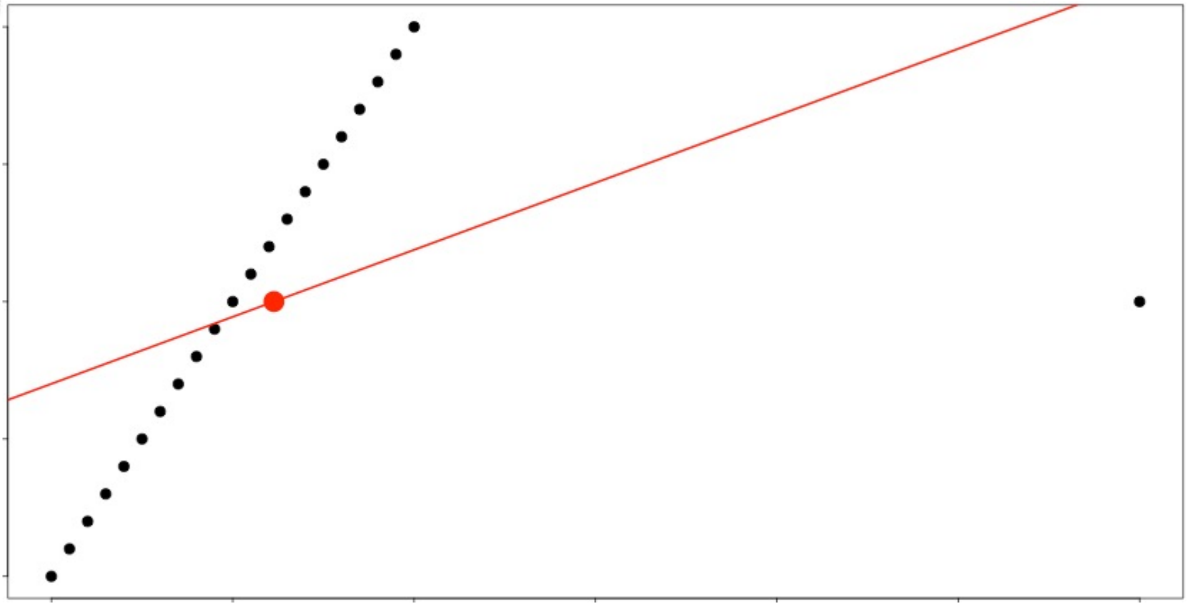
These tests for outliers tend to be less sensitive than the eye:

if there is a significant outlier, we will be able to see it,
but if we can see it, it may still not be significant.

Leverage



Leverage in statistics is like leverage in physics: with a long enough lever (a predictor far enough away from the mean) you can make a regression line do whatever you want.



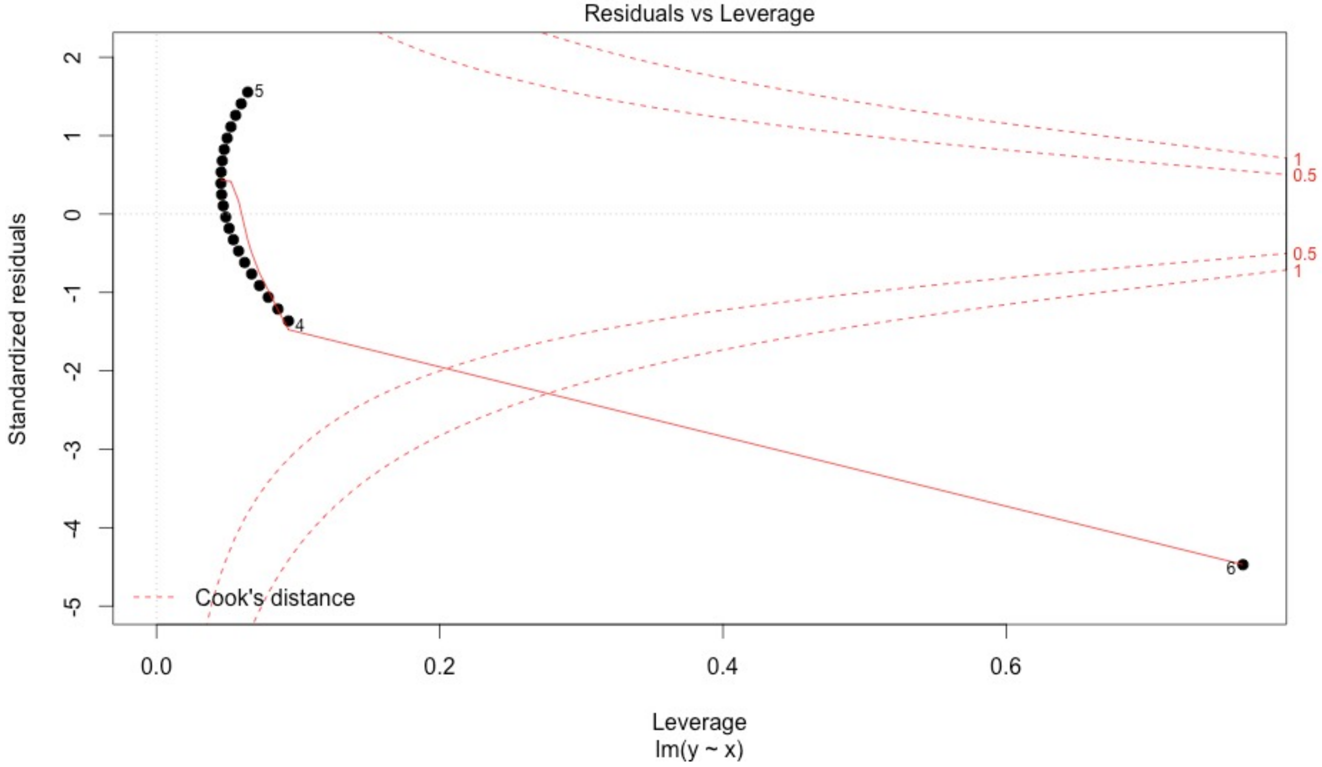
Leverage is potential influence.

With many predictors what matters is ~Mahalanobis distance: distance from the center of mass scaled by the covariance matrix.

This is hard to visualize, so it's useful to just look at the leverage numbers, and particularly, whether there are large residuals at large leverage – that is bad.

Cook’s distance

```
plot(lm(y~x), which=5)
```



A data point with a lot of leverage and large residuals is exerting undue influence on the regression.

Cook’s distance measures this.

Several, equally correct, ways to think about Cook’s distance:

(1) How much will my regression coefficients change without this data point?
(2) How much will the predicted Y values change without this data point?
(3) A combination of leverage and residual to ascertain point’s influence.

Cook's distance

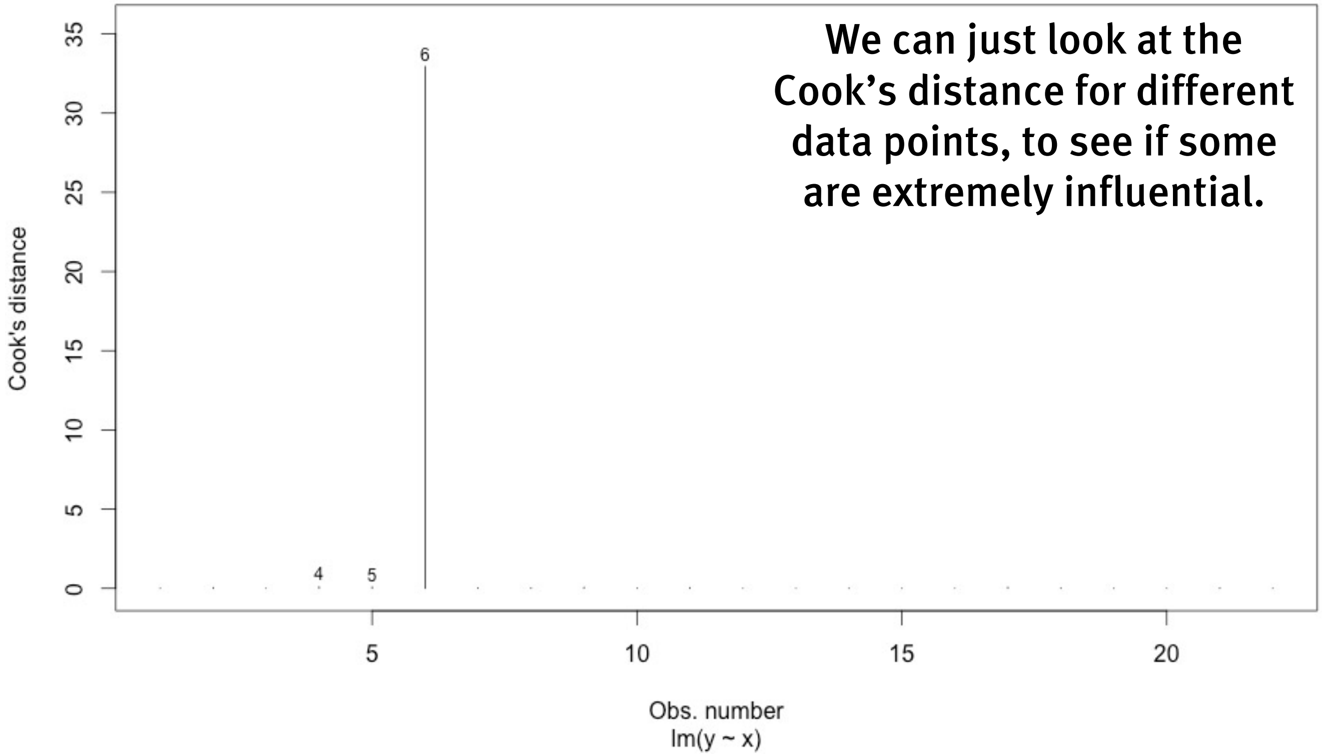
Several, equally correct, ways to think about Cook's distance:

(1) How much will my regression coefficients change without this data point?

(2) How much will the predicted Y values change without this data point?

(3) A combination of leverage and residual to ascertain point's influence.

```
plot(lm(y~x), which=4)
```



We can just look at the Cook's distance for different data points, to see if some are extremely influential.

How much influence is too much?

(a) D > 1 ?

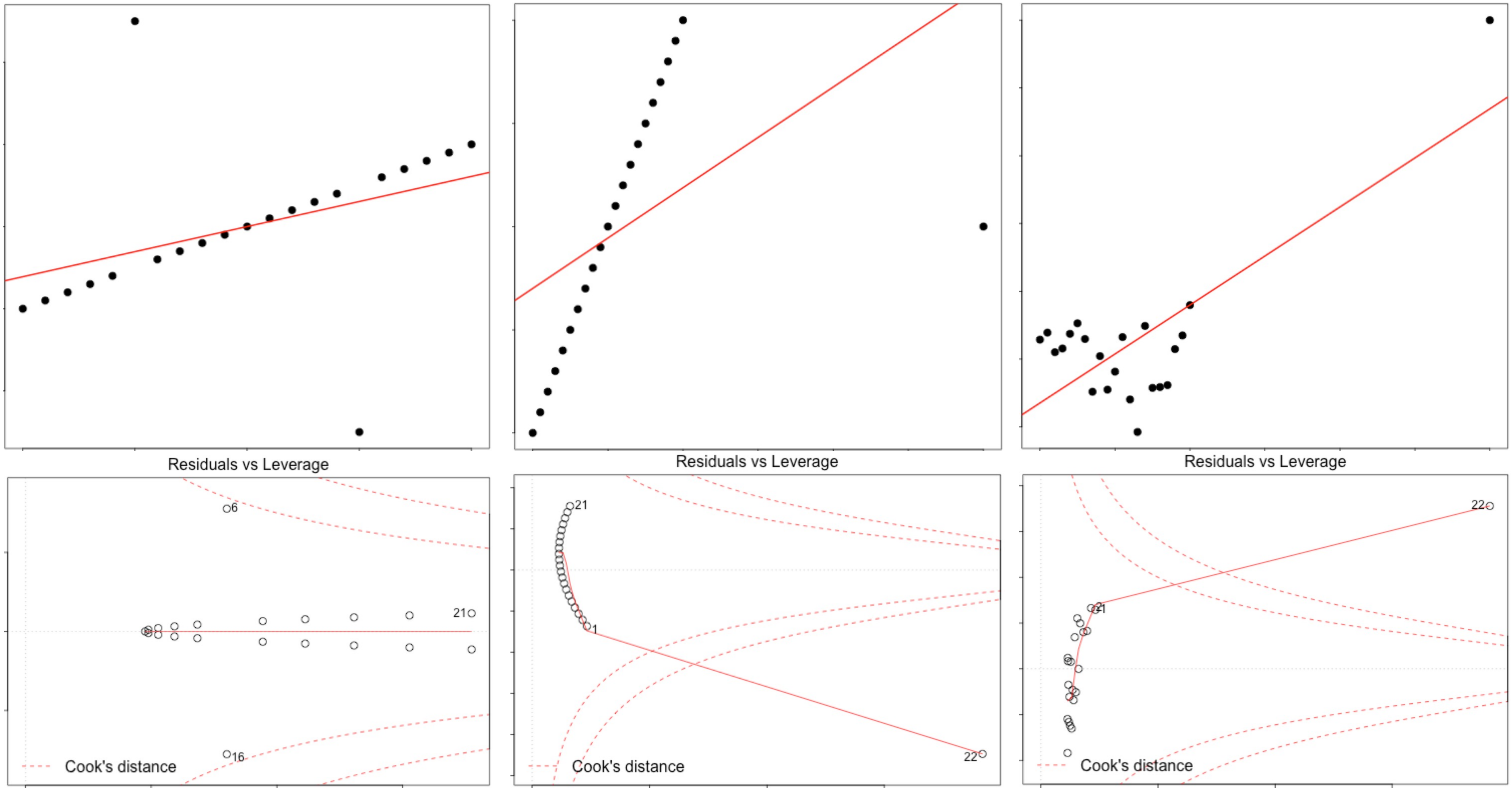
(b) D > (4/n) ?

(c) D > (4/(n-k-1)) ?

Different folks have different standards…

Outliers and extreme influence

Data points with large residuals, and/or high *leverage*



Checking for autocorrelated errors

Sometimes errors might be autocorrelated
(when there is a particular dependence in sample acquisition)

This is rarely considered unless we are dealing with clearly time-based data. (although our subjects vary over the quarter!)

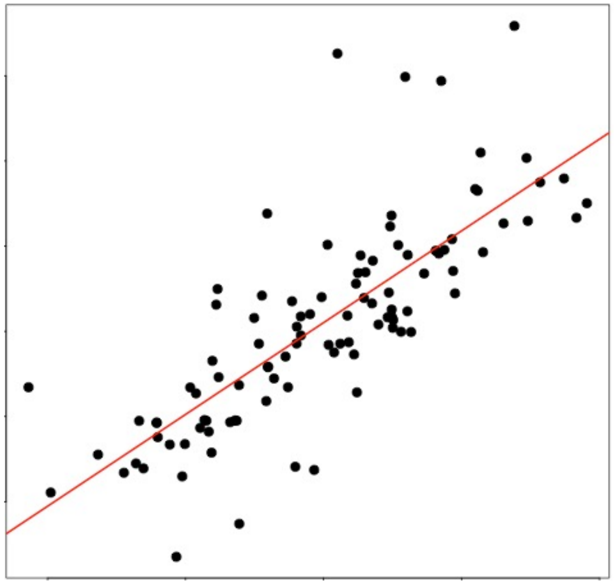
Check for this by looking at residuals ~ observation_number

Test for this via durbinWatson test
(default: tests for lag-1 autocorrelation, can consider higher lags)

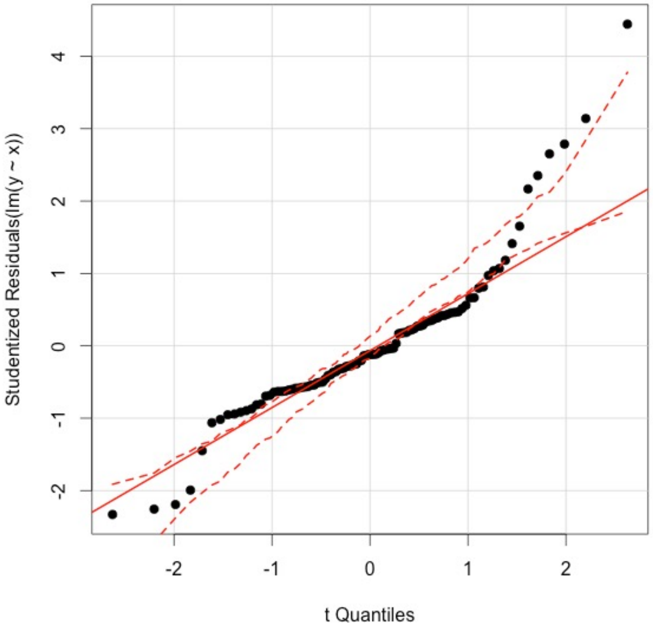
If very concerned: look at autocorrelation plots of residuals…

Checking for normal residuals

Look at qq plot, test with Kolmogorov-Smirnov test



```
qqPlot(lm(y~x))
```



```
ks.test(rstudent(lm(y~x)), "pt", length(y)-2)
```

```
        One-sample Kolmogorov-Smirnov test
data:  rstudent(lm(y ~ x))
D = 0.1398, p-value = 0.04002
alternative hypothesis: two-sided
```

Generally though, it's fine to ignore slight but significant deviations

Checking for linear model assumptions

Linearity, Homoscedasticity, Uncorrelated residuals, Normal residuals

If you are really paranoid about making sure all assumptions are valid, you can even consider the “Global validation test for linear model assumptions”

```
library(gvlma)
gvlma(lm(y~x))
```

```
ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = lm(y ~ x))

                     Value   p-value                   Decision
Global Stat        65.6446 1.882e-13 Assumptions NOT satisfied!
Skewness           21.3914 3.745e-06 Assumptions NOT satisfied!
Kurtosis           43.8742 3.502e-11 Assumptions NOT satisfied!
Link Function       0.2748 6.002e-01    Assumptions acceptable.
Heteroscedasticity  0.1043 7.467e-01    Assumptions acceptable.
```

Global statistic here combines statistics measuring skewness, kurtosis of residuals (for non-normality, outliers), **link function linearity** (based on residuals being consistent across y.hat values), **and constant variance, uncorrelated variance** (based on squared residuals as function of observation order).

With enough real data, it will ~always tell you assumptions are violated.