Multiple Linear Regression: Model Checking and Diagnostics

Author: Nicholas G Reich, Jeff Goldsmith

This material is part of the ***statsTeachR*** *project*

Today's Lecture

- Model checking and diagnostics
- Variable transformations

Model selection vs. model checking

Assume $y|\mathbf{x} = f(\mathbf{x}) + \epsilon$

- model checking asks whether the ϵ match the assumed form, whether there are systematic and diagnosable (and fixable!) deviations from assumed model structure.
- model selection (coming soon!) focuses on how you construct $f(\cdot)$;

Model checking: possible challenges

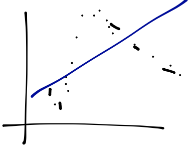
Two major areas of concern

- Global lack of fit, or general breakdown of model assumptions
  - Linearity
  - Unbiased, uncorrelated errors $E(\epsilon|x) = E(\epsilon) = 0$
  - Constant variance $Var(y|x) = Var(\epsilon|x) = \sigma^2$
  - Independent errors
  - Normality of errors
- Effect of influential points and outliers

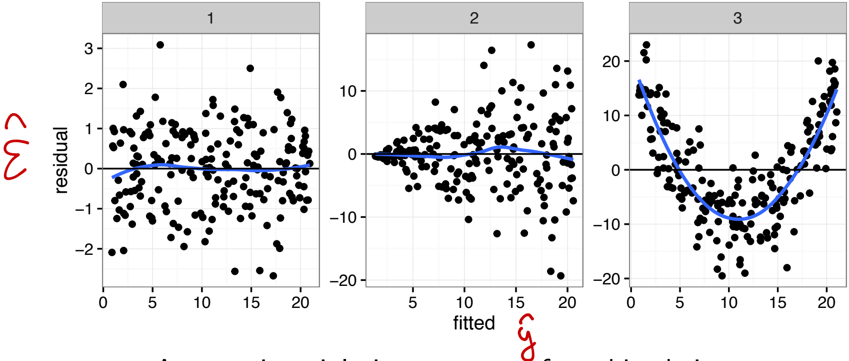
Model checking: possible solutions and strategies

- Global lack of fit, or general breakdown of model assumptions
    - Residual analysis – QQ plots, residual plots against fitted values and predictors
    - Adjusted variable plots
- Effect of influential points and outliers
    - Measure of leverage, influence, outlying-ness

Residual plots: verifying assumptions



Which assumptions (if any) do these plots show violations of?



Assumption violations are not often this obvious
(but sometimes they are!).

QQ-plots for checking Normality of residuals

q	d_1	d_2
.01	$q^1_{.01}$	$q^2_{.01}$
.02		
.03		
⋮		
.99		

QQ plot defined

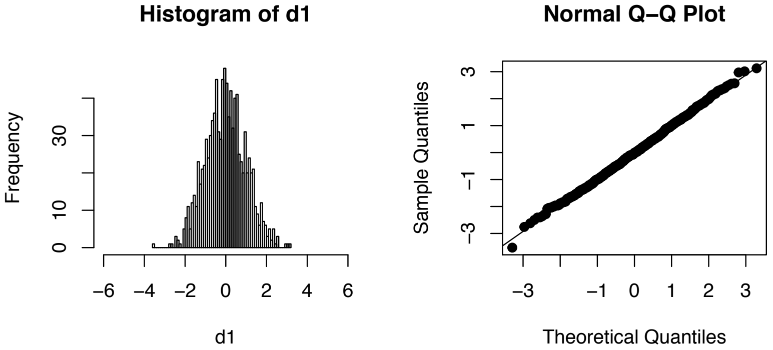
QQ-plot stands for quantile-quantile plot, and is used to compare two distributions. If the two distributions are the same, then each point (which represents a quantile from each distribution) should lie along a line.

For a single (x, y) point

- x = a specific quantile for the N(0,1) distribution
- y = the same quantile from the sample of data => sample residuals

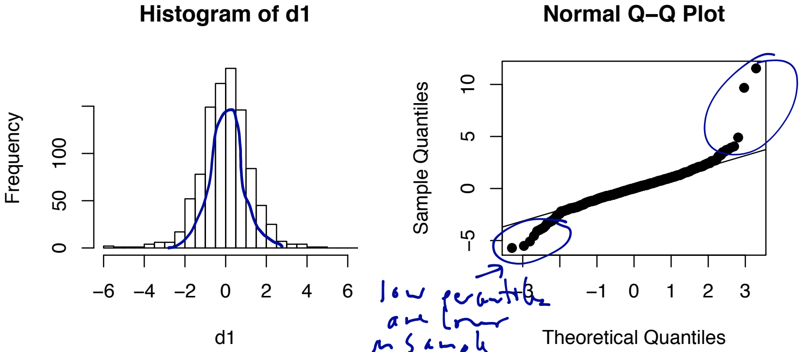
example: Gaussian or Normal(0,1) distribution

```r
d1 <- rnorm(1000)
layout(matrix(1:2, nrow=1))
hist(d1, breaks=50, xlim=c(-6, 6))
qqnorm(d1, pch = 19)
qqline(d1)
```



example: Student's T-distribution with 6 d.f.

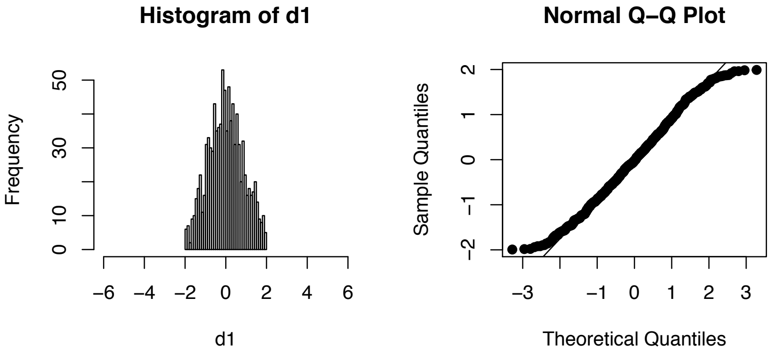
```
d1 <- rt(1000, df=5)
layout(matrix(1:2, nrow=1))
hist(d1, breaks=50, xlim=c(-6, 6))
qqnorm(d1, pch = 19)
qqline(d1)
```



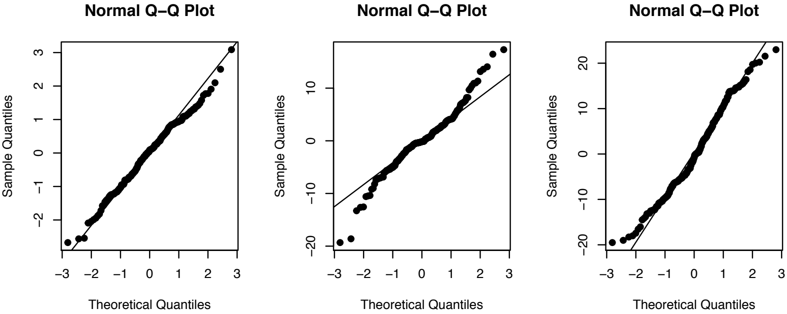
low percentiles are lower in sample

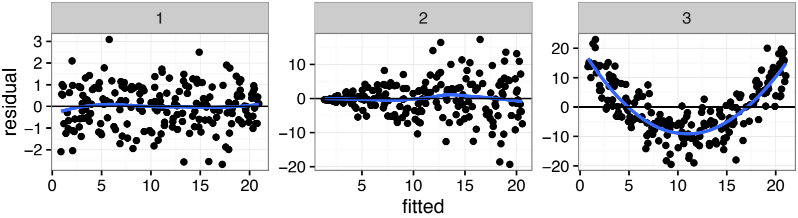
example: Truncated Gaussian

```r
d1 <- rnorm(1000)
d1 <- subset(d1, abs(d1)<2)
layout(matrix(1:2, nrow=1))
hist(d1, breaks=50, xlim=c(-6, 6))
qqnorm(d1, pch = 19)
qqline(d1)
```



QQ-plots for our three fits from earlier





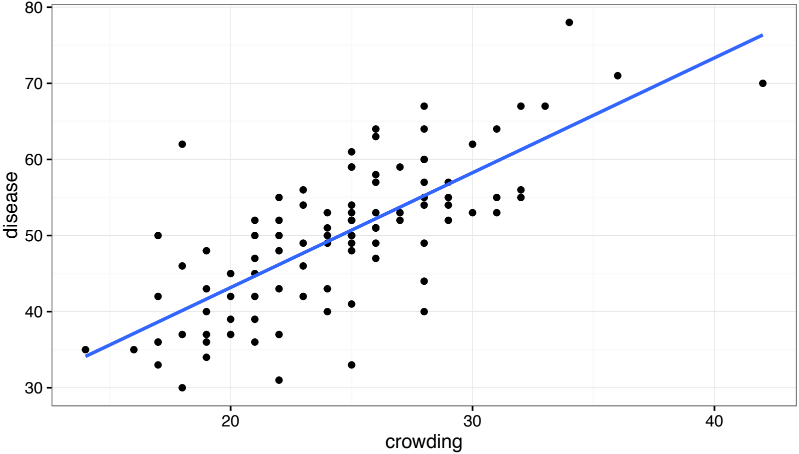
Recall: Lung Data Example

99 observations on patients who have sought treatment for the relief of respiratory disease symptoms.
The variables are:

- `disease` measure of disease severity (larger values indicates more serious condition).
- `education` highest grade completed
- `crowding` measure of crowding of living quarters (larger values indicate more crowding)
- `airqual` measure of air quality at place of residence (larger number indicates poorer quality)
- `nutrition` nutritional status (larger number indicates better nutrition)
- `smoking` smoking status (1 if smoker, 0 if non-smoker)

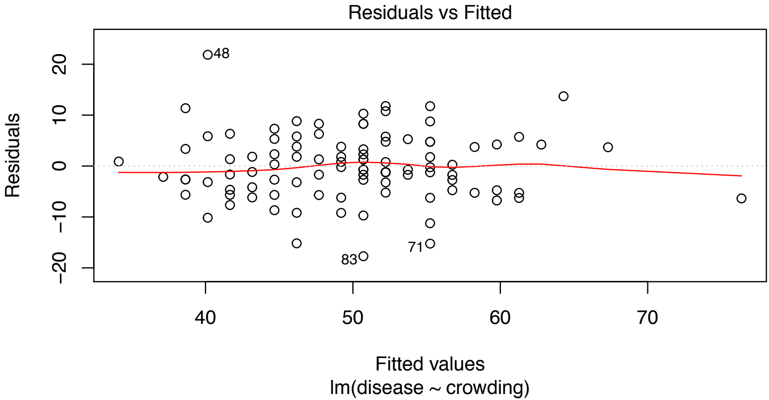
Typical regression plot: fitted line

```
ggplot(data, aes(crowding, disease)) +
    geom_point() + geom_smooth(method="lm", se=FALSE)
```



Typical residual plot: fitted vs. residuals

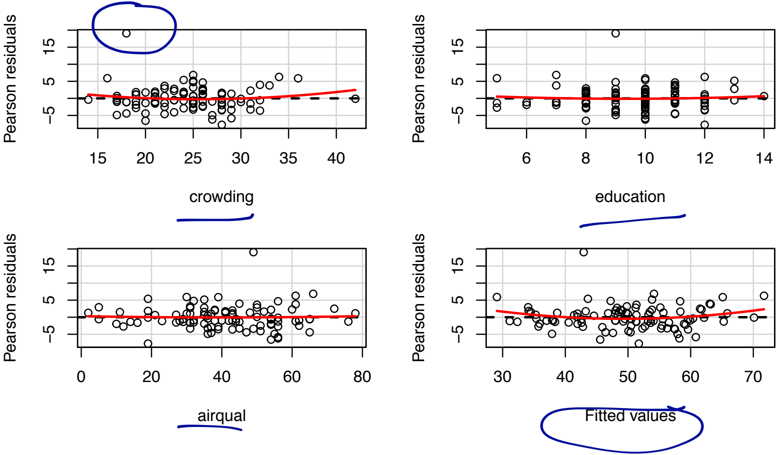
```
slr1 <- lm(disease ~ crowding, data=data)
plot(slr1, which=1)
```



But this is more complicated with MLR: how do we visualize adjusted multivariable relationships?

Predictor vs. residual plots

```r
library(car)
mlr1 <- lm(disease ~ crowding + education + airqual, data=data)
residualPlots(mlr1, tests=FALSE)
```



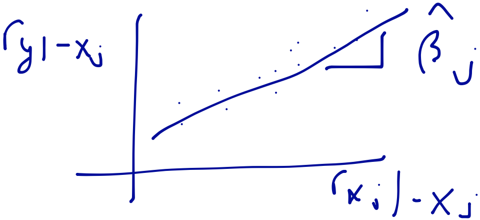
Checking model structure: adjusted variable plots!

- You can plot residuals against each of the predictors, or plot outcomes against predictors, BUT...
- Keep in mind the MLR uses adjusted relationships; scatterplots don't show that adjustment!

Adjusted variable plots (partial regression plots, added variable plots) can be useful.

Adjusted (or added) variable plots

- Regress y on everything but x_j; take residuals $r_{y|-x_j}$
- Regress x_j on everything but x_j; take residuals $r_{x_j|-x_j}$
- Regress $r_{y|-x_j}$ on $r_{x_j|-x_j}$; slope of this line will match β_j in the full MLR
- Plot of $r_{y|-x_j}$ against $r_{x_j|-x_j}$ shows the "adjusted" relationship
- This figure can be used to diagnose violations of linearity in MLR models.

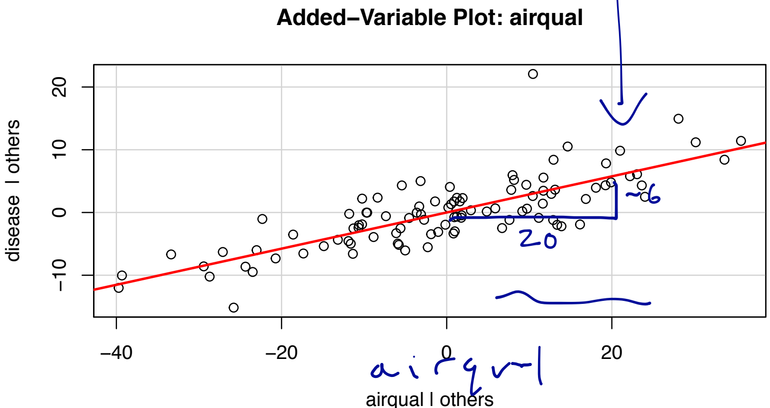


AV plots

```
coef(mlr1)

## (Intercept)    crowding   education     airqual
##  -7.7505215   1.3127837   1.4376563   0.2880687

avPlot(mlr1, variable="airqual")
```

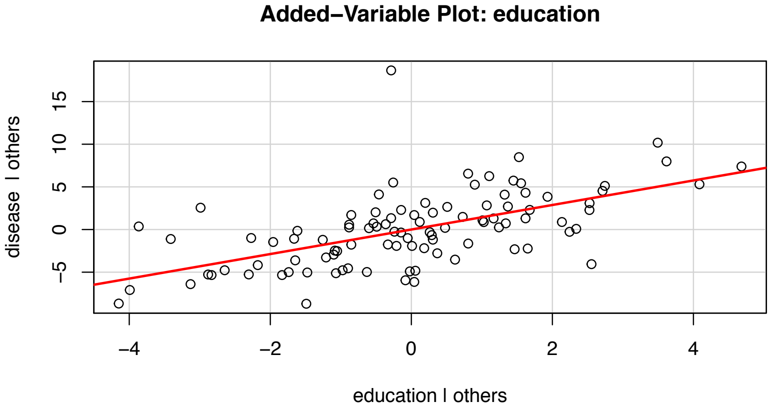


AV plots

```
coef(mlr1)

## (Intercept)    crowding   education     airqual
##  -7.7505215   1.3127837   1.4376563   0.2880687

avPlot(mlr1, variable="education")
```



Model checking: possible solutions

- Global lack of fit, or general breakdown of model assumptions
  - ✓ Residual analysis – QQ plots, residual plots against fitted values and predictors
  - ✓ Adjusted variable plots
- Effect of influential points and outliers
  - Measure of leverage, influence, outlying-ness

Isolated points

Points can be isolated in three ways

- Leverage point – outlier in x, measured by hat matrix
- Outlier – outlier in y, measured by residual
- Influential point – a point that largely affects $\boldsymbol{\beta}$
  - Deletion influence; $|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)}|$
  - Basically, a high-leverage outlier

Quantifying leverage

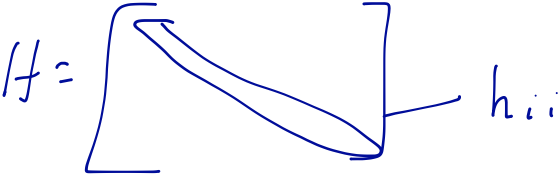
$$\frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}$$

We measure leverage (the "distance" of $\mathbf{x}_i$ from the distribution of $\mathbf{x}$) using

$$h_{ii} = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i$$

where h_{ii} is the $(i,i)^{th}$ entry of the hat matrix. Where, recall

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$



Quantifying Leverage via the Hat Matrix

Note that

$$\sum_i h_{ii} \overset{def}{=} tr(\mathbf{H}) = p$$

where p is the total number of independent predictors (i.e. βs) in your model (including a β_0 if you have one).

What counts as "big" leverage?

- Average leverage is p/n
- Typical rules of thumb are $2p/n$ or $3p/n$
- Leverage plots can be useful as well

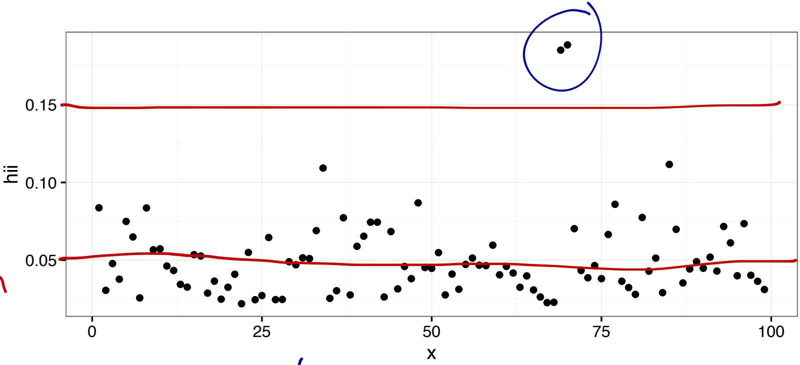
Example Leverage plot with lung data

$p = 5 \quad n = 99$

```
mlr <- lm(disease ~ nutrition+ airqual + crowding + smoking,
          data=data)
hii <- hatvalues(mlr)
x <- 1:length(hii)
qplot(x, hii, geom="point")
```

$3p/n$

p/n



$1 - 99$

Example Leverage plot with lung data

Can be useful to investigate specific points.

```
cols <- c("disease", "crowding", "education", "airqual")
summary(data[,cols])

##     disease         crowding        education         airqual     
##  Min.   :30.00   Min.   :14.00   Min.   : 5.000   Min.   : 2.00  
##  1st Qu.:42.50   1st Qu.:21.00   1st Qu.: 8.000   1st Qu.:31.00  
##  Median :51.00   Median :25.00   Median :10.000   Median :41.00  
##  Mean   :49.92   Mean   :24.47   Mean   : 9.566   Mean   :40.92  
##  3rd Qu.:55.00   3rd Qu.:28.00   3rd Qu.:11.000   3rd Qu.:54.00  
##  Max.   :78.00   Max.   :42.00   Max.   :14.000   Max.   :78.00  

(d <- data[which(hii>.15), cols])

##    disease crowding education airqual
## 69      39       20         8      54
## 70      70       42        12      19
```

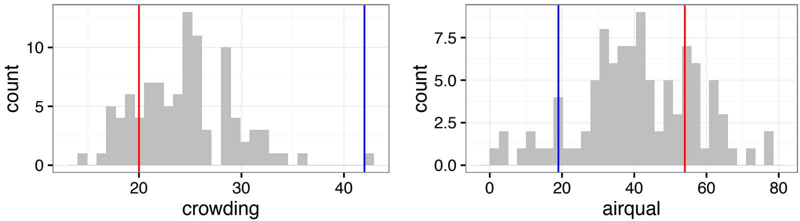
low

maximum

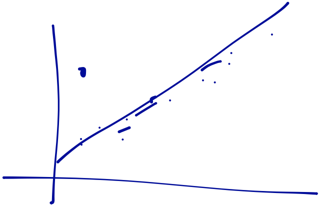
Example Leverage plot with lung data

Can be useful to investigate specific points.

```
library(gridExtra)
p1 <- ggplot(data) + geom_histogram(aes(x=crowding), fill="grey") +
    geom_vline(xintercept=d[1,"crowding"], color="red") +
    geom_vline(xintercept=d[2,"crowding"], color="blue")
p2 <- ggplot(data) + geom_histogram(aes(x=airqual), fill="grey") +
    geom_vline(xintercept=d[1,"airqual"], color="red") +
    geom_vline(xintercept=d[2,"airqual"], color="blue")
grid.arrange(p1, p2, ncol=2)
```



Outliers



- When we refer to "outliers" we typically mean "points that don't have the same mean structure as the rest of the data"
- Residuals give an idea of "outlying-ness", but we need to standardize somehow
- We can use the fact that $Var(\hat{\epsilon}_i) = \sigma^2(1 - h_{ii})$ ...

Outliers

The *standardized* residual is given by

$$\hat{\epsilon}_i^* = \frac{\hat{\epsilon}_i}{\sqrt{Var(\hat{\epsilon}_i)}} = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{(1 - h_{ii})}}$$

The *Studentized* residual is given by

$$t_i = \frac{\hat{\epsilon}_{(-i)}}{\hat{\sigma}_{(-i)}\sqrt{(1 - h_{ii})}} = \hat{\epsilon}_i^* \left(\frac{n - p}{n - p - \hat{\epsilon}_i^{*2}} \right)^{1/2}$$

Studentized residuals follow a t_{n-p-1} distribution.

Influence

Intuitively, “influence” is a combination of outlying-ness and leverage. More specifically, we can measure the “deletion influence” of each observation: quantify how much $\hat{\boldsymbol{\beta}}$ changes if an observation is left out.

- $|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)}|$
- Cook’s distance is

$$\begin{aligned} D_i &= \frac{\boxed{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})^T(\mathbf{X}^T\mathbf{X})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}}{p\hat{\sigma}^2} \\ &= \frac{(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(-i)})^T(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(-i)})}{p\hat{\sigma}^2} \\ &= \frac{1}{p}\hat{\epsilon}_i^2 \frac{h_{ii}}{1 - h_{ii}} = f(h_{ii}, \varepsilon_i) \end{aligned}$$

Handy `R` functions

Suppose you fit a linear model in `R`;

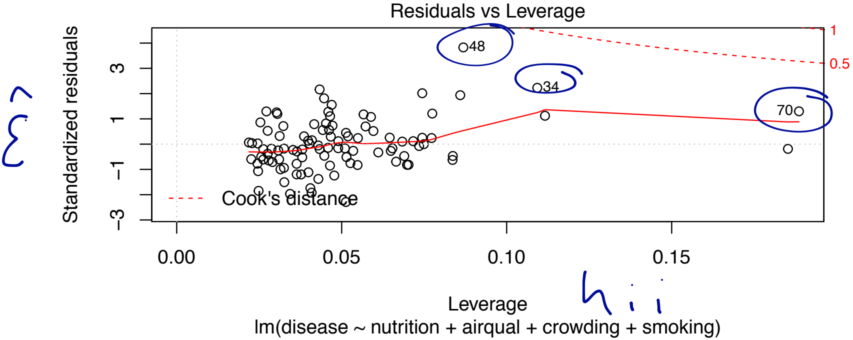
- `hatvalues` gives the diagonal elements of the hat matrix h_{ii} (leverages)
- `rstandard` gives the standardized residuals
- `rstudent` gives the studentized residuals
- `cooks.distance` gives the Cook's distances

Built-in R plots for `lm` objects

You can also use the `plot.lm()` function to look at leverage, outlying-ness, and influence all together. Recall that

$$D_i = \frac{1}{p}\hat{\epsilon^*}_i^2 \frac{h_{ii}}{1-h_{ii}}$$

```r
plot(mlr, which=5)
```

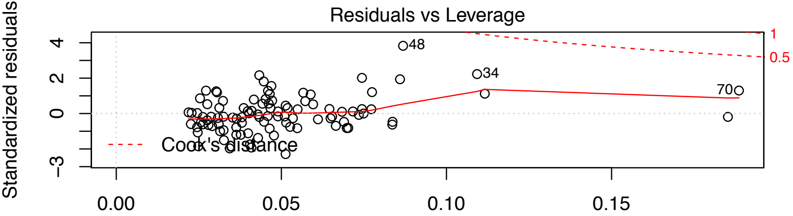


Model checking summary

You are looking for...

- Points that show worrisome level of influence $\Longrightarrow$ sensitivity analysis!
- Systematic departures from model assumptions $\Longrightarrow$ transformations, different model structure
- Unrealistic outliers $\Longrightarrow$ check your data!

No points show worrisome influence in this lung data analysis, although observation 70 showed up in both of our analyses.



Back to the outline

- Model checking and diagnostics
- Variable transformations

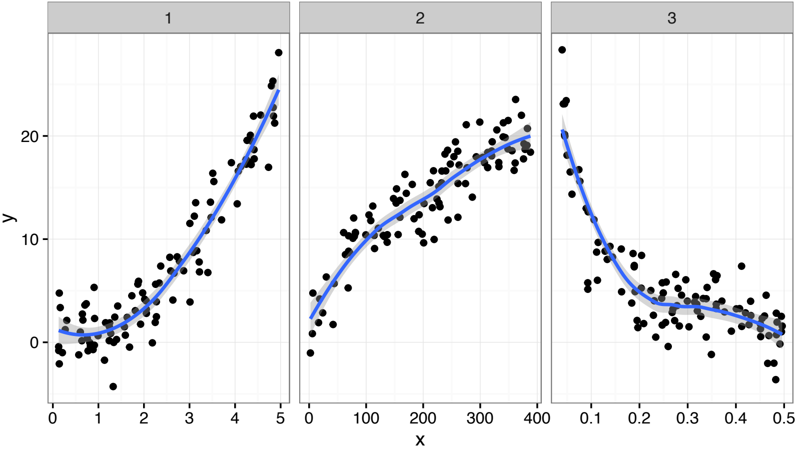
Overview of variable transformations

The problems

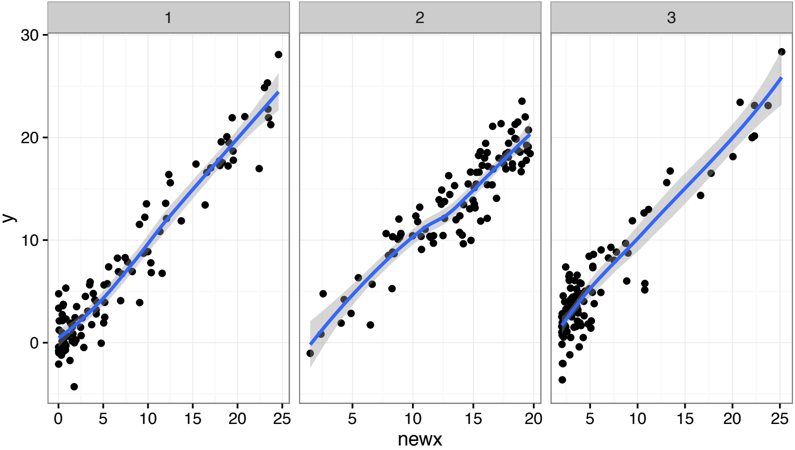
- Non-linearity between X and Y $\longrightarrow$ transform X
- Skewed distribution of Xs/points with high leverage $\longrightarrow$ transform X
- Non-constant variance $\longrightarrow$ transform Y

Transforming your X variables

Transforming predictor variables can help with constant-variance non-linear relationships.



Transforming your X variables



β interpretations with transformed Xs

Transforming predictor variables can help with non-linearities, but can make coefficient interpretations hard.

Possible solutions

- Interpret βs qualitatively across a region of interest: “We found strong evidence for an inverse association, where values of Y were inversely proportional to X across the observed range (a, b).
- Occasionally, a “one unit change in X” can be meaningful: e.g. $\log_a X$. A one unit change in $\log_a X$ indicates a a-fold increase in X.

$\log_2 X \Rightarrow$ “doubling”

β interpretations with transformed Xs

- Transforming predictor variables can help with non-linearities, but can make coefficient interpretations hard.
- Can also use polynomials, splines (more soon!).

Transforming Ys for non-constant variance

What to do ...

- Nothing; just use least squares and bootstrap
- Use weighted LS, GLS (Biostat Methods 3?)
- Use a variance stabilizing transformation
- Consider a generalized linear model (more soon)

Box-Cox Transformations

Outcome is raised to the λ power:

$$y_i^\lambda = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

- Estimate λ, a new parameter, by maximum likelihood.
- Some well-known choices of λ: 2, -1, 1/2
- By definition, when $\lambda = 0$, we specify $y_i^\lambda = \log_e y_i$

Today's big ideas

- Model checking
- Variable transformations
- Next up: inference about MLR parameters