

MLR Model Checking

Author: Nicholas G Reich, Jeff Goldsmith

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

*Made available under the Creative Commons Attribution-ShareAlike 3.0 Unported
License: http://creativecommons.org/licenses/by-sa/3.0/deed.en_US*

Today's Lecture

- Model selection vs. model checking
- Continue with model checking (regression diagnostics)

Model selection vs. model checking

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

- model selection focuses on how you construct $f(\cdot)$;
- model checking asks whether the ϵ match the assumed form.

Model checking: possible challenges

Two major areas of concern

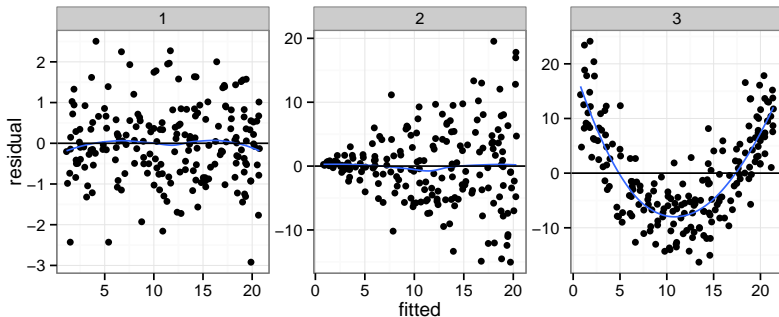
- Global lack of fit, or general breakdown of model assumptions
 - ▶ Linearity
 - ▶ Unbiased, uncorrelated errors $E(\epsilon|x) = E(\epsilon) = 0$
 - ▶ Constant variance $Var(y|x) = Var(\epsilon|x) = \sigma^2$
 - ▶ Independent errors
 - ▶ Normality of errors
- Effect of influential points and outliers

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 are these plots evaluating?



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

QQ-plots for checking Normality of residuals

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.

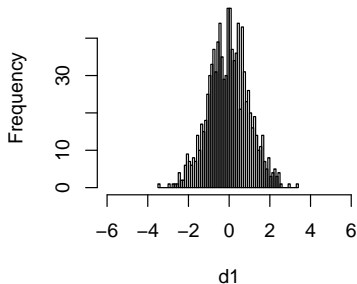
For a single (x, y) point

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

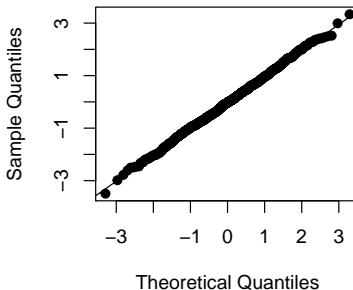
example: Gaussian or Normal(0,1) distribution

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

Histogram of d1



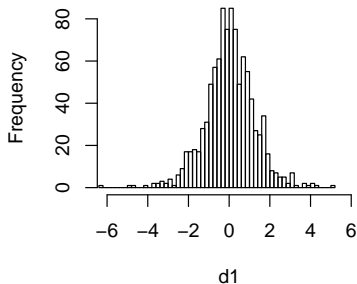
Normal Q-Q Plot



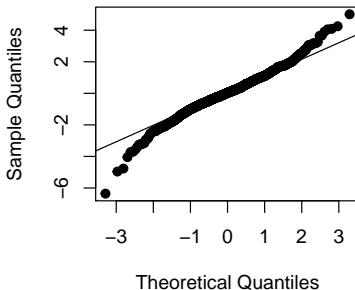
example: Student's T-distribution with 6 d.f.

```
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)
```

Histogram of d1



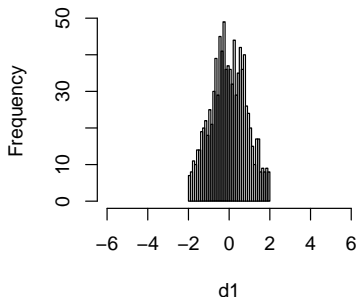
Normal Q-Q Plot



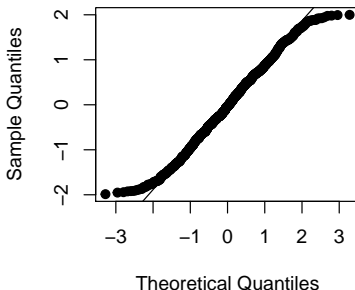
example: Truncated Gaussian

```
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)
```

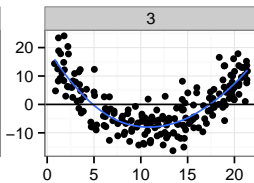
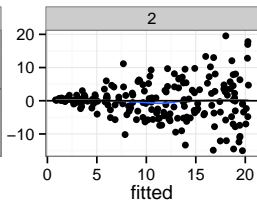
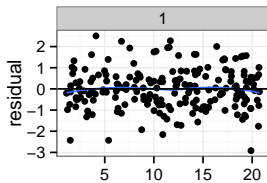
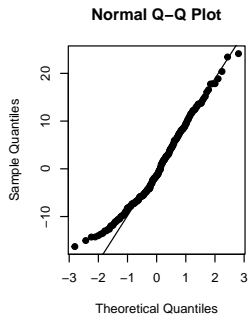
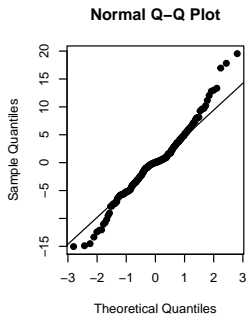
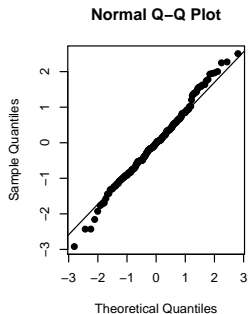
Histogram of d1



Normal Q-Q Plot



QQ-plots for our three fits from earlier



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 β
 - ▶ Deletion influence; $|\hat{\beta} - \hat{\beta}_{(-i)}|$
 - ▶ Basically, a high-leverage outlier

Quantifying leverage

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

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

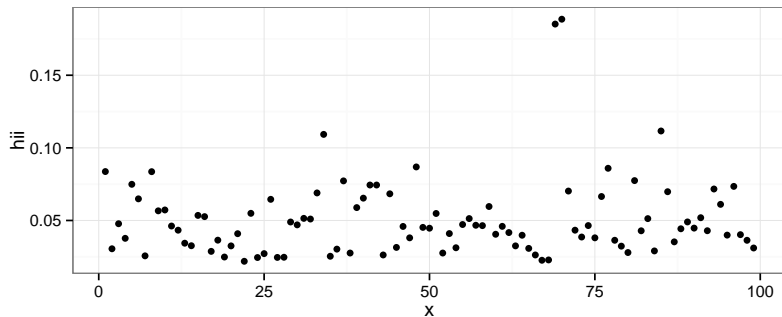
where p is the total number of independent predictors (i.e. β s) in your model (including a β_0 if you have one).

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

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



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 $\text{Var}(\hat{\epsilon}_i) = \sigma^2(1 - h_{ii}) \dots$

Outliers

The *standardized* residual is given by

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

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 $\hat{\beta}$ changes if an observation is left out.

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

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

Handy R functions

Suppose you fit a linear model in R;

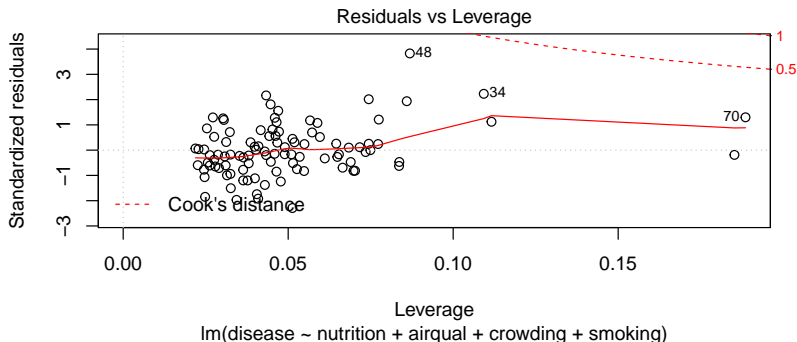
- `hatvalues` gives the diagonal elements of the hat matrix h_{ii} (leverages)
- `rstandard` gives the standardized residuals
- `rstudent` gives the studentized residuals
- `cooks.distance` gives the Cook's distances

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

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

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



Today's big ideas

- Model checking
- Up next: model **selection!**