

Model validation

Bert van der Veen

Department of Mathematical Sciences, NTNU

Outline

- ▶ Normal distribution
- ▶ Linear models
- ▶ Assumption violation

Recall:

$$y_i = \alpha + \mathbf{x}_i \beta + \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- ▶ the systematic component: the regression line
- ▶ the random component: left-over stuff

Models make assumptions

Our models make assumptions. Sometimes, these are realistic.

All models are wrong, but some are useful

Is is important to:

- ▶ Know which assumptions your model makes
- ▶ Consider if they are realistic
- ▶ Determine if they are violated
- ▶ Assess the consequences

Tests make assumptions too, but more on that later

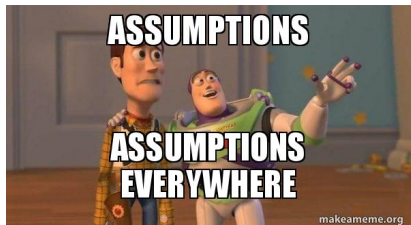
Linear regression assumptions

$$y_i = \alpha + \sum_{k=1}^p x_{ik}\beta_k + \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- ▶ Linearity
- ▶ Normality: normally distributed errors
- ▶ Homoscedasticity: same variance for all errors
- ▶ Independence of errors
- ▶ Lack of perfect multicollinearity ($\mathbf{x}_1 \neq \mathbf{x}_2$)
- ▶ No outliers
- ▶ No error in the explanatory variables

Violating assumptions

- ▶ Linearity: Eek
- ▶ Normality: cannot trust tests and confidence intervals
- ▶ Homoscedasticity: cannot trust tests and confidence intervals
- ▶ Independence of errors: Eek
- ▶ Lack of perfect multicollinearity
- ▶ No outliers: biased parameter estimates



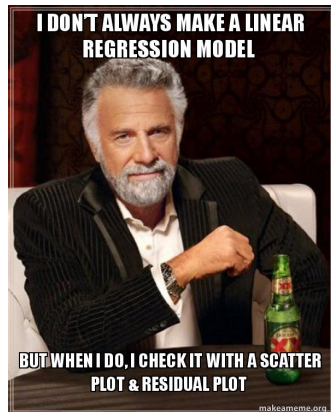
Assumption checking: methods

Tests:

- ▶ Score test for non-constant error variance
`car::ncvTest`
- ▶ Durbin-Watson test for Autocorrelated errors
`car::durbinWatsonTest`
- ▶ Bonferroni outlier test
`car::outlierTest`
- ▶ Shapiro-Wilk normality test `shapiro.est`

Rather: **Residual plots**

- ▶ `plot(model)`
- ▶ Various packages (e.g., `car` or `DHARMa`)



Residuals are the **observed** deviation from the regression line.

In \mathbb{R}

```
model <- lm(y~x)
residuals(model) # estimated random part
fitted(model) # estimated systematic part
```

Residual plots

Residual plots should have no discernible patterns.

- ▶ No curvature
- ▶ No shapes whatsoever
- ▶ No outliers
- ▶ No change in variance

This can be difficult with few observations.

Standardised residuals

$$\frac{\hat{\epsilon}_i}{s\sqrt{1-h_i}}, \quad \text{where } h_i \text{ is "influence"} \quad (1)$$

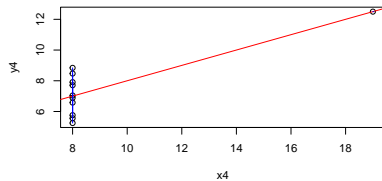
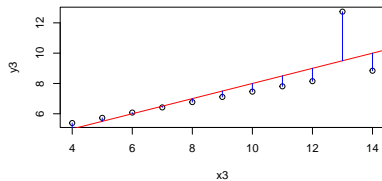
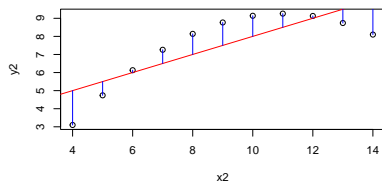
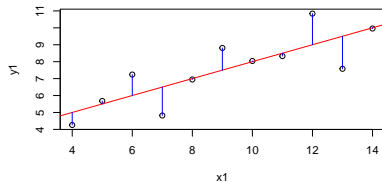
Studentized residuals

- ▶ omit observation i
- ▶ re-compute residual variance
- ▶ calculate standardised residual (with hat matrix)

`rstand` or `rstudent`

All of these regressions have the same parameter estimates

Anscombe's quartet: adding residuals



Our toolbox

- ▶ Residuals versus fitted: $\hat{\epsilon}_i$ vs. $\hat{\mu}_i$
- ▶ Observed versus theoretical quantiles:
 $Pr(X \leq y_i)$ vs. $Pr(X \leq x)$
- ▶ Cook's distance plot: $\hat{\epsilon}_i$ vs. $\hat{\mu}_i$
- ▶ Standardized residuals vs. fitted (Scale-location)
- ▶ Leverage and influence

Leverage: hat matrix

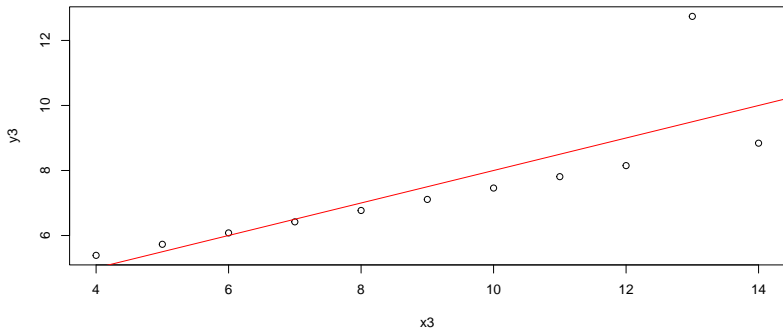
$$\hat{\mu}_i = \hat{\alpha} - \sum_{i=1}^n x_{ik} \hat{\beta}$$

but we can also write:

$$\hat{\mu} = \mathbf{H}\mathbf{y}$$

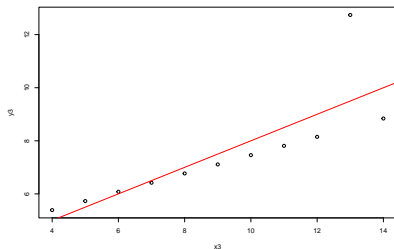
- ▶ **H** is the “hat” matrix
- ▶ $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$
- ▶ Its diagonal gives “leverage” and is between 0 and 1
- ▶ Leverage: observation that is far from others on the explanatory variable

Influence



An observation that has undue impact on the regression line.

Outliers



- ▶ Outliers are observations far away from the regression line
- ▶ Can be erroneous
- ▶ Might be real data
- ▶ Remove with caution
- ▶ Better: sensitivity analysis

Cook's distance

Cook's distance: how much the fitted values change if we remove an observation

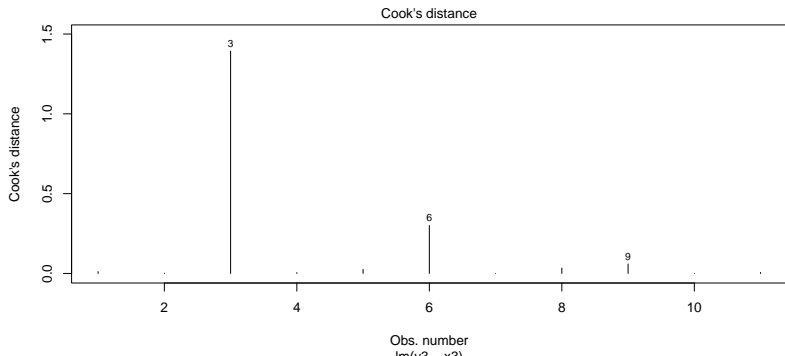
$$D_j = \frac{\sum_{i=1}^n (\hat{y}_i - \hat{y}_{i(-j)})^2}{\sigma^2}$$

- ▶ \hat{y}_i - prediction
- ▶ $\hat{y}_{i(-j)}$ - prediction for model with data point j removed
- ▶ σ^2 - residual variance

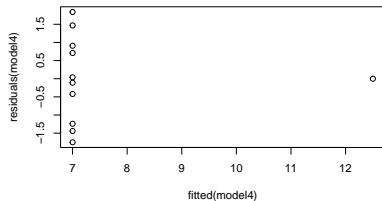
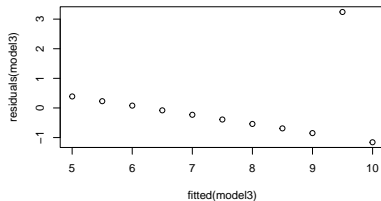
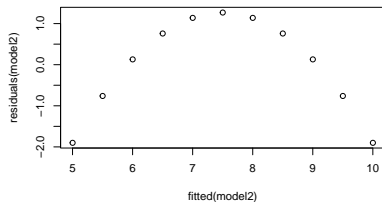
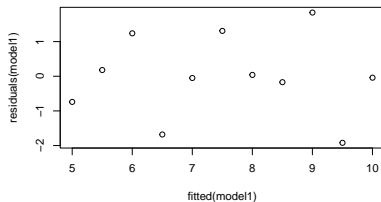
What is influential?

Large values of D_i mean a large influence

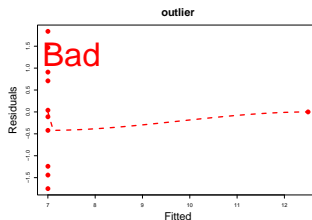
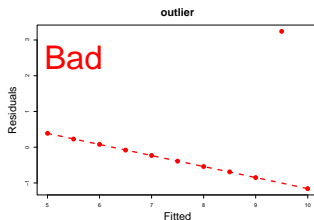
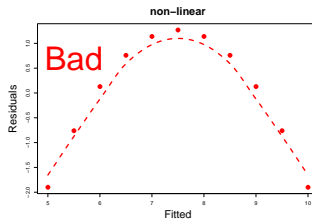
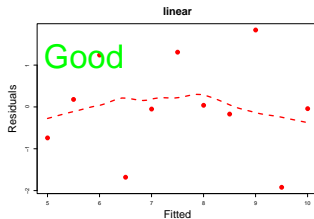
Usually $D_i > 1$, or $4/n$



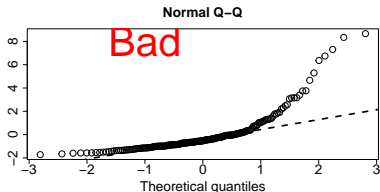
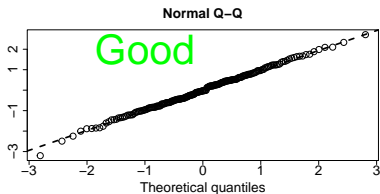
Anscombe's quartet: residuals vs. fitted



Anscombe's quartet: residuals vs. fitted

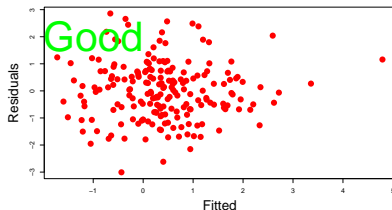


Residual diagnostics: normality with QQ-plot

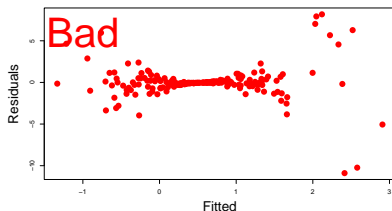


1. compare residuals to theoretical quantiles
2. points should follow the line

Residual diagnostics: constant variance



1. compare residuals to fitted
2. points should be randomly distributed



Addressing assumption violations

- ▶ Curvature? Add quadratic term
- ▶ Outlier? Sensitivity analysis (if not error)
- ▶ Normality? GLM
- ▶ Constant variance? GLM

Alternative: data transformation

Data transformation

$$\mathbb{E}\{h(y_i)|x_i\} = \alpha + \beta x_i \quad (2)$$

- ▶ $h(\cdot)$ data transformation (e.g. $\sqrt{(\cdot)}$ or $\log(\cdot)$)
- ▶ To stabilize variance
- ▶ To normalize residuals

Data transformation

$$\mathbb{E}\{h(y_i)|x_i\} = \alpha + \beta x_i \quad (2)$$

- ▶ $h(\cdot)$ data transformation (e.g. $\sqrt{(\cdot)}$ or $\log(\cdot)$)
- ▶ To stabilize variance
- ▶ To normalize residuals

Not the same as

$$h(\mathbb{E}\{y_i|x_i\}) = \alpha + \beta x_i \quad (3)$$

Use a Generalised linear model instead

Summary

Now we have a valid model

We still need to check **how** good it is (more on that tomorrow)