San José State University

Math 261A: Regression Theory & Methods

Model Adequacy Checking

Dr. Guangliang Chen

This lecture is based on the following textbook sections:

- **Chapter 4: 4.1 – 4.3, 4.5**

Outline of this presentation:

- Introduction
- Residual analysis
- Residual plots
- Lack of fit tests
- Summary

Introduction

The major assumptions we have made in linear regression models

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

are

- The relationship between the response and regressors is linear.
- The error term ϵ has zero mean (no need to check).
- The error term ϵ has constant variance σ^2.
- The errors are uncorrelated.
- The errors are normally distributed.

We present several methods useful for diagnosing violations of the regression assumptions, by examining the model residuals:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$$

That is,

$$e_i = y_i - \hat{y}_i, \ i = 1, \ldots, n$$

Methods for dealing with model inadequacies, as well as additional, more sophisticated diagnostics, are discussed in Chapters 5 and 6.

Residual analysis

The residuals $\mathbf{e} = (e_1, \ldots, e_n)'$ can be shown to satisfy

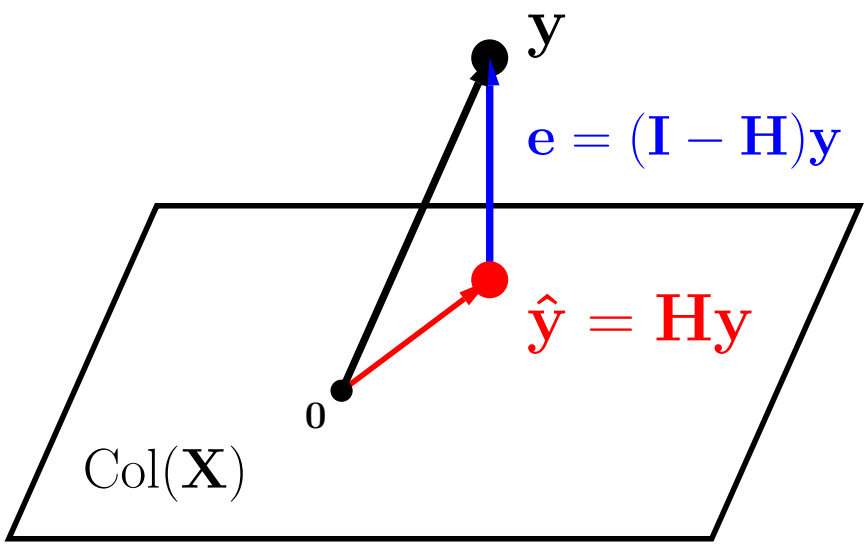
$$\mathbf{X}'\mathbf{e} = \mathbf{0}$$

That is, $\mathbf{e}$ is orthogonal to the columns of $\mathbf{X}$ (predictors).

Proof.

$$\begin{aligned}\mathbf{X}'\mathbf{e} &= \mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{0}.\end{aligned}$$

Geometric intuition:



More insights about $e_1, \ldots, e_n$:

- They can be viewed as observations of the model errors $\epsilon_1, \ldots, \epsilon_n$
- They have zero mean because $\sum e_i = 0$.
- They are not independent with only $n-p$ degrees of freedom because there are p equations constraining the residuals e_i:

$$\mathbf{X}'\mathbf{e} = \mathbf{0}$$

- They can be used to form an unbiased estimator for σ^2 as follows:

$$MSE = \frac{SS_{Res}}{n-p} = \frac{\sum e_i^2}{n-p}$$

Methods of scaling residuals

We introduce a few methods for scaling residuals, as they need to be looked at "relative to" the standard deviation of the model error σ:

- **Standardized residuals**, or **(internally) Studentized residuals**
- **Externally Studentized residuals (or R-Student)**

Remark. Since we do not know the true value of σ^2, we will compare the residuals against its point estimate MS_{Res}.

Assume a collection of residuals e_i from fitting a linear regression model to a data set.

A simple way to scale them is as follows:

$$d_i = \frac{e_i}{\sqrt{MS_{Res}}}, \ i = 1, \ldots, n$$

The normalized residuals d_i also have mean zero and are expected to be between -3 and 3 for most observations.

However, this is not a correct way to standardize the residuals (note that the book incorrectly calls d_i the standardized residuals).

To correctly standardize the residuals, we need to obtain the exact standard deviation of e_i.

Since

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

we have

$$\text{Var}(\mathbf{e}) = (\mathbf{I} - \mathbf{H}) \underbrace{\text{Var}(\mathbf{y})}_{=\sigma^2\mathbf{I}} (\mathbf{I} - \mathbf{H})' = \sigma^2(\mathbf{I} - \mathbf{H}).$$

That is,

$$\text{Var}(e_i) = \sigma^2(1 - h_{ii}), \quad \text{Cov}(e_i, e_j) = -\sigma^2 h_{ij}, \ i \neq j$$

Remark. Recall that h_{ii} is a measure of the remoteness (leverage) of the ith point relative to the full data in x space: In the setting of only predictor ($k = 1$), it has been shown that

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$$

Clearly, h_{ii} is smallest at the center and increases as we move away from the center.

This implies that the variance of e_i depends on where x_i lies. Generally, points near the center of the x space have larger variance (poorer least-squares fit) than residuals at more remote locations.

R demonstration

(Internally) Studentized residuals, or standardized residuals

Def 0.1. The standardized residuals, also called (internally) Studentized residuals, are defined as

$$r_i = \frac{e_i}{\sqrt{MS_{Res}(1 - h_{ii})}}, \ i = 1, \ldots, n$$

Remark. The standardized residuals have zero mean and constant variance regardless of the location of x_i when the form of the model is correct.

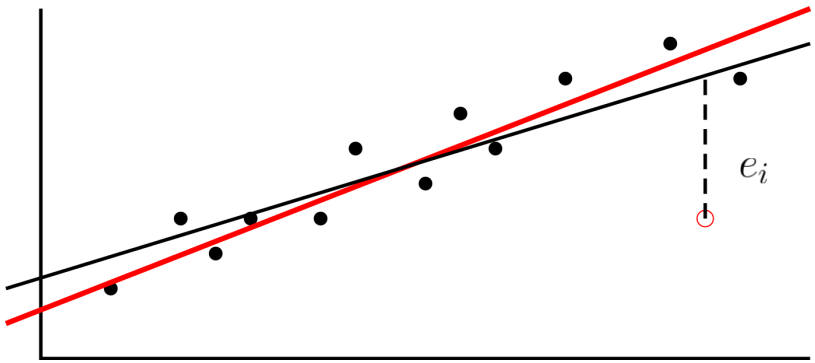
For large data sets, $h_{ii} \approx 0$, so d_i and r_i have little difference in those cases.

Externally Studentized residuals (or R-Student)

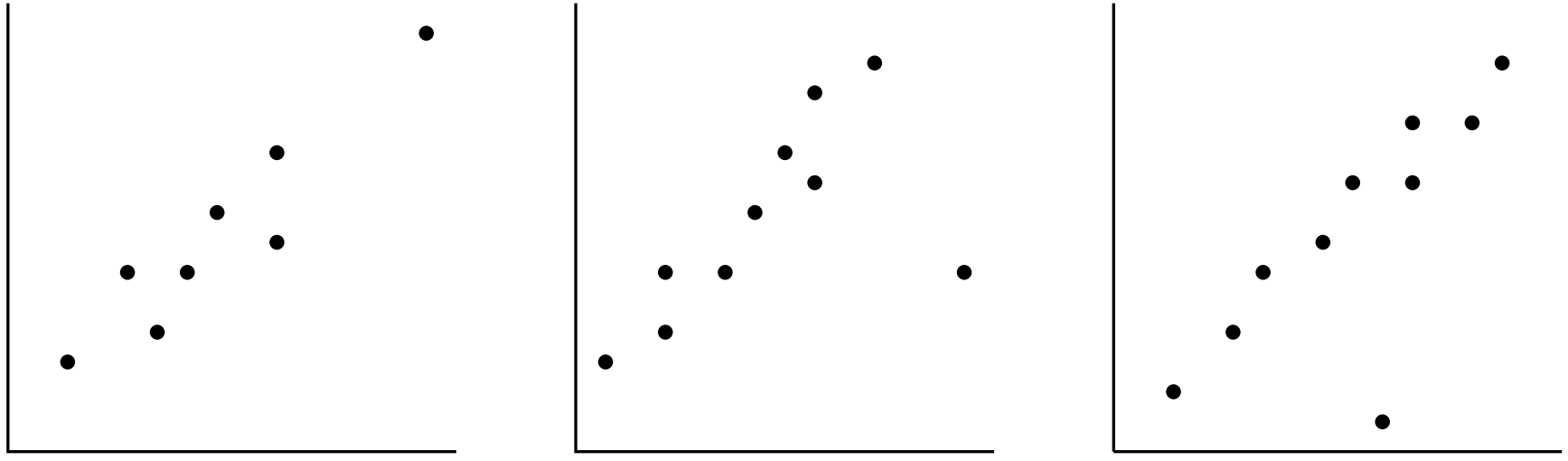
Def 0.2. The externally Studentized residuals are defined as

$$t_i = \frac{e_i}{\sqrt{S^2_{(i)}(1 - h_{ii})}}, \quad i = 1, \ldots, n$$

where $S^2_{(i)}$ is an estimate of σ^2 obtained by fitting a linear regression model to all data but the ith observation (note that the e_i are still computed from the model on the full data set).



Remark. The externally Studentized residuals t_i do not differ much from the internally Studentized residuals r_i, except for influential points.



Left: a leverage point (not influential); Middle: a leverage and influence point; Right: a point with little leverage or influence

Comparing with the d_i, the r_i **is more sensitive to leverage points** while the t_i is more sensitive to influential points.

Remark. It can be shown that for all i,

$$S^2_{(i)} = \frac{(n-p)MS_{Res} - e_i^2/(1-h_{ii})}{n-1-p}$$

This indicates that the $S^2_{(i)}$ can be computed from the model on the full data set.

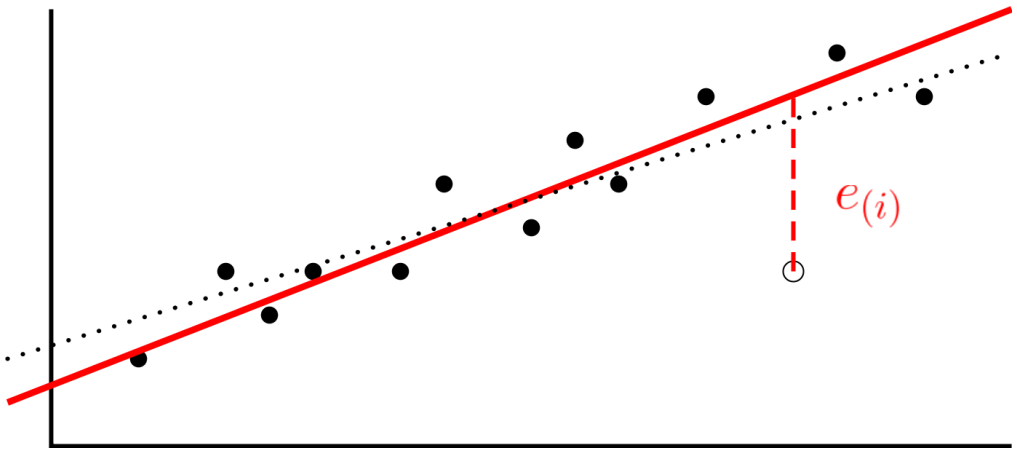
Thus, in practice, only one model based on the full data needs to be fit in order to compute all t_i simultaneously.

Deleted residuals

Def 0.3. The deleted residuals, also called PRESS residuals are defined as

$$e_{(i)} = y_i - \hat{y}_{(i)}, \ \ i = 1, \ldots, n$$

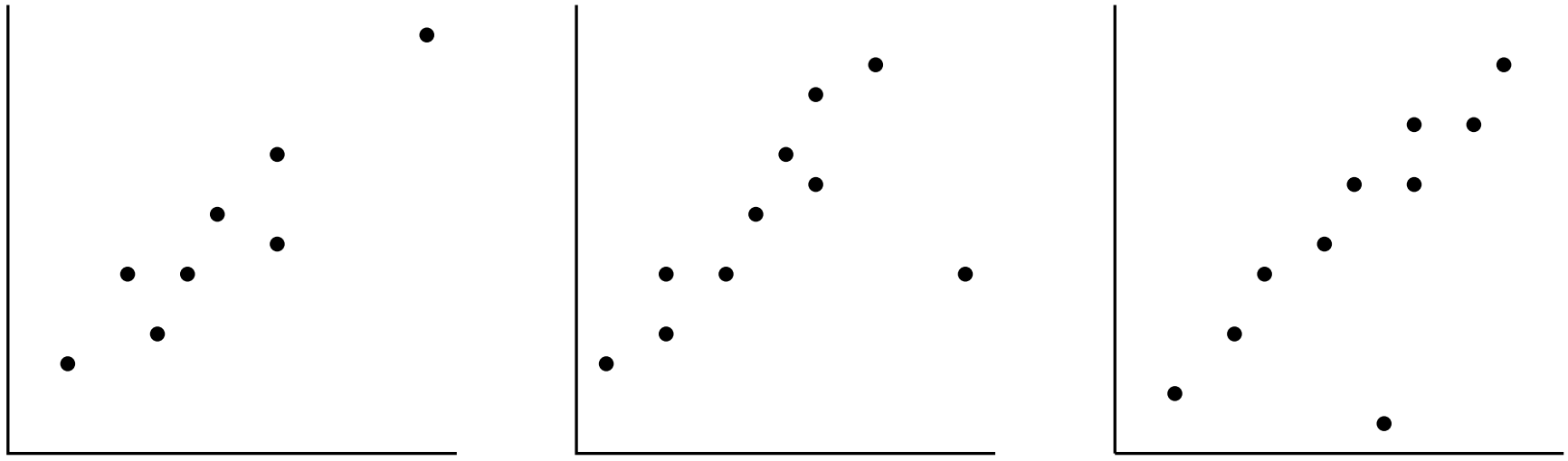
where $\hat{y}_{(i)}$ is the prediction of y_i based on the model fit over all observations except the ith one.



Remark. It can be shown that

$$e_{(i)} = \frac{e_i}{1 - h_{ii}}, \quad i = 1, \ldots, n$$

Thus, residuals associated with points for which both e_i and h_{ii} are large will have large PRESS residuals.



Remark. It turns out that the standardized PRESS residuals are identical to the Studentized residuals.

First,

$$\text{Var}(e_{(i)}) = \frac{1}{(1-h_{ii})^2}\text{Var}(e_i) = \frac{1}{(1-h_{ii})^2}\,\sigma^2(1-h_{ii}) = \frac{\sigma^2}{1-h_{ii}}.$$

It follows that

$$\frac{e_{(i)}}{\sqrt{\text{Var}(e_{(i)})}} = \frac{e_i/(1-h_{ii})}{\sqrt{\sigma^2/(1-h_{ii})}} = \frac{e_i}{\sqrt{\sigma^2(1-h_{ii})}}$$

Assessing predictive power of a model

Another way to use the PRESS residuals is to define the PRESS statistic for measuring how well a regression model will perform in predicting new data:

$$\text{PRESS} = \sum \underbrace{(y_i - \hat{y}_{(i)})}_{e_{(i)}}^2 = \sum \left(\frac{e_i}{1 - h_{ii}} \right)^2$$

Clearly, small values of the PRESS statistic are desired, and it should be looked at relative to SS_T:

$$R^2_{\text{prediction}} = 1 - \frac{\text{PRESS}}{SS_T}$$

Remark. PRESS $> SS_{Res}$ and thus $R^2_{\text{prediction}} < R^2$.

Using PRESS to Compare Models:

One very important use of the PRESS statistic is in comparing regression models of different sizes (in terms of predictive power).

Generally, a model with a small value of PRESS is preferable to one where PRESS is large.

Which other criterion can be used to compare different models of different sizes?

We will discuss the topic of model selection and comparison in detail in Chapter 10.

R commands for computing scaled residuals

- Raw residuals: *residuals(mymodel)* or *mymodel$residuals*
- Normalized residuals: *residuals(mymodel)/summary(mymodel)$sigma*
- Standardized residuals, also called (internally) Studentized residuals: *rstandard(mymodel)*
- Externally Studentized residuals (or R-Student): *rstudent(mymodel)*
- PRESS/deleted residuals: *mymodel$residuals/(1-hatvalues(mymodel))*

What's next

Residuals and their various scaled versions are useful in identifying outliers and diagnosing for leverage and influence. We will cover this topic in depth in Chapter 5.

In this lecture we focus on using residuals to check the model assumptions.

Residual plots

Graphical analysis is much more effective in trying to detect patterns in the residuals than looking at the raw numbers. There are different types of plots that can be employed to check the different model assumptions.

- Normal quantile plots (qq-plots) $\longleftarrow$ checking normality
- Residuals against fitted values $\longleftarrow$ checking constant variance, or nonlinearity
- Residuals against a regressor $\longleftarrow$ checking constant variance, or nonlinearity
- Residuals against time (if time known) $\longleftarrow$ checking autocorrelation

Normal Quantile Plots (qq-plots)

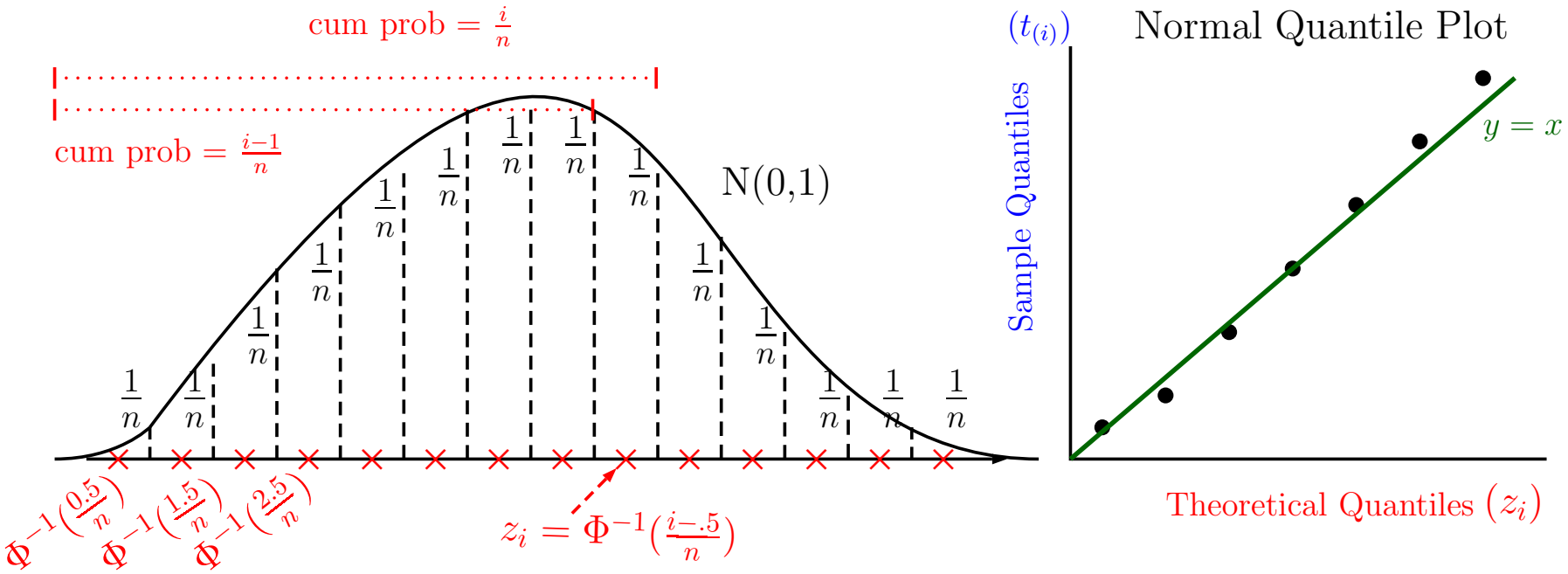
Assumption: $\epsilon_1, \ldots, \epsilon_n \sim N(0, \sigma^2) \longrightarrow e_1, \ldots, e_n \longrightarrow t_1, \ldots, t_n$

What: A graphical method for comparing a sample ($\{t_i\}$) with a target distribution (standard normal) to see if there is any obvious violation of the assumption that the sample is from the distribution.

How: Plot sample quantiles (sorted sample values $t_{(i)}$) against theoretical quantiles ($z_i = \Phi^{-1}(\frac{i-0.5}{n})$), which are expected samples from the distribution (standard normal).

Desired pattern: If the sample truly comes from the distribution, then the points in the qq-plot should closely follow the line $y = x$.

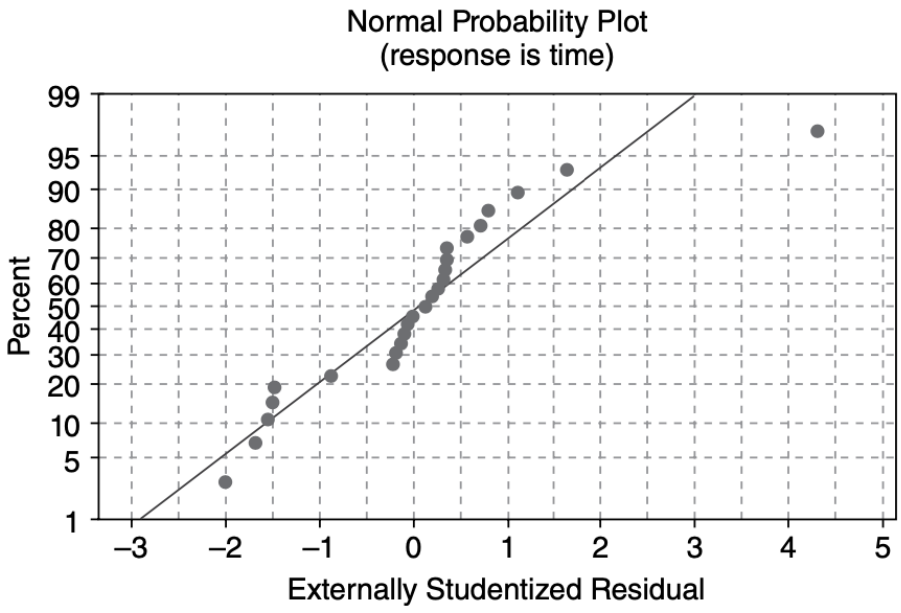
Graphical demonstration:



Remark. Note that the textbook uses normal probability plots:

- The externally Studentized quantiles ($t_{(i)}$) are shown on the horizontal axis
- The probabilities are shown on the vertical axis
- The vertical axis does not have linear scale!

Overall, the two plots are equivalent (we are looking for linear patterns in both of them).



(Figure 4.4, page 140 of textbook)

Residuals against fitted values

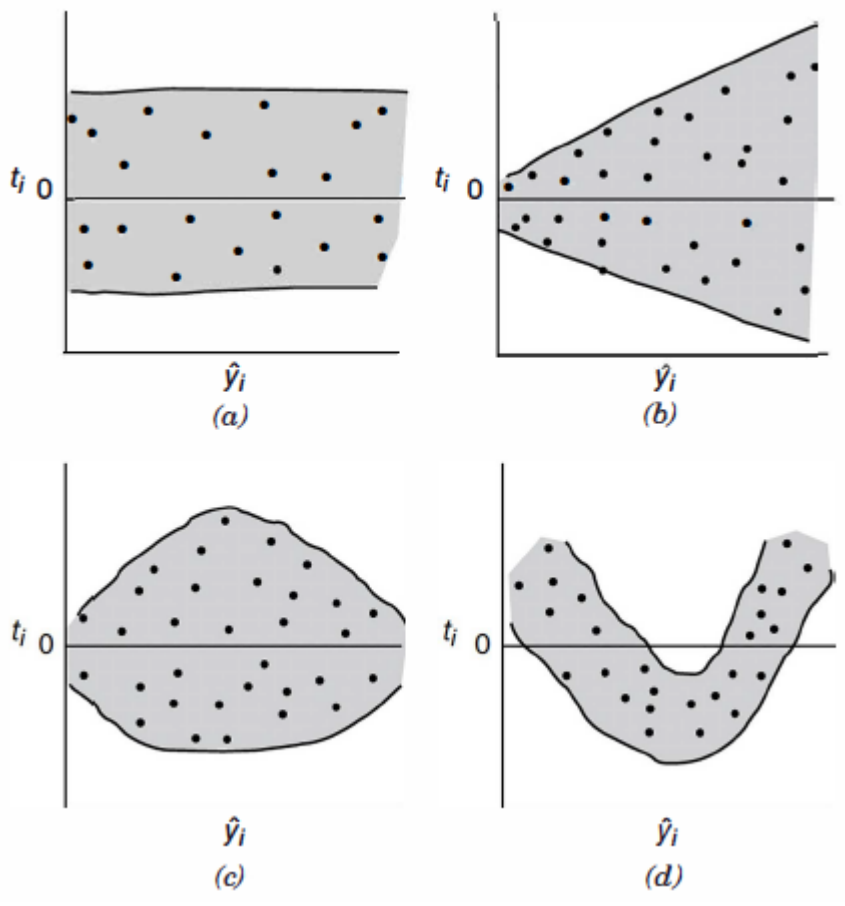
This plot may be used to check the constant-variance assumption of the model error (and also nonlinearity):

(a) Confetti in a box ✓

(b) Funnel

(c) Double bow

(d) Curvature (indication of a non-linear relationship between the response and the predictors)



Residuals against values of a regressor

Consider the following two cases:

- **If the regressor is already in the model (x_j)**: Such plots are equivalent to the plot of residuals against fitted values, useful for checking the constant-variance assumption (and if there is a nonlinear relationship between the response and the regressor)
- **If the regressor is a new one**: Such a plot is useful for determining whether the new regressor should be added to the model (based on the strength of the association) and if yes, in which way (based on the form of the association).

Experiments

(See in-class R demonstrations: simulation + bodydata example)

Remark. Interpreting the residual plots is not an easy task:

- A lot of randomness for small data sets (must set the bar high)
- Easier for moderate or large data sets
- Important to learn from simulations!

It is an art and requires experience.

Residuals against time

The time sequence plot of residuals may indicate that the errors at one time period are correlated with those at other time periods. The correlation between model errors at different time periods is called autocorrelation.

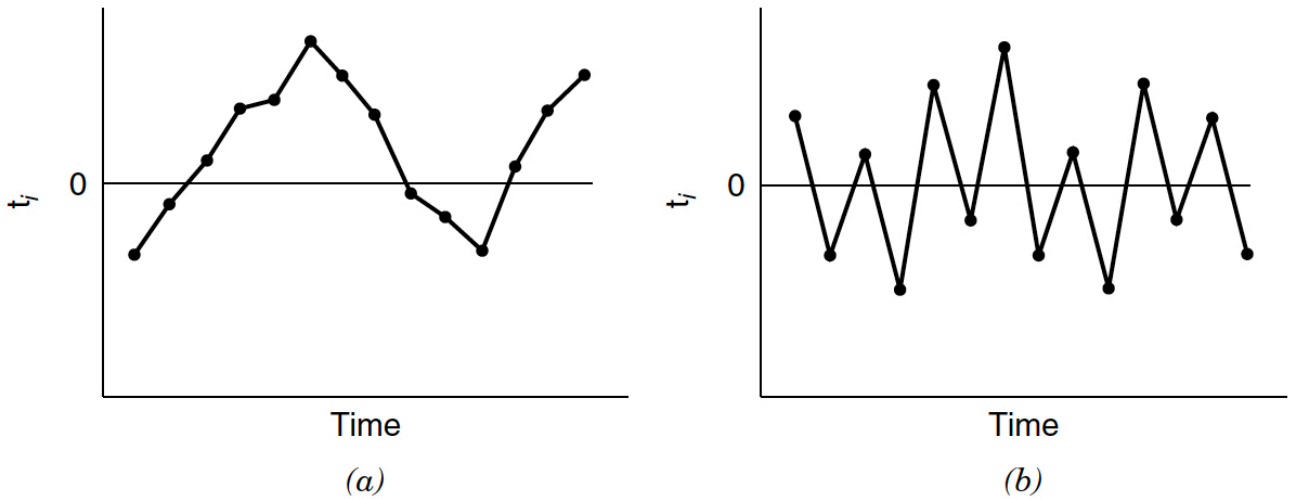


Figure 4.8 Prototype residual plots against time displaying autocorrelation in the errors: (*a*) positive autocorrelation; (*b*) negative autocorrelation.

Lack of fit test

The formal statistical test for the lack of fit of a linear regression model assumes that the following three requirements

- normality, independence, and constant-variance

are all met and that only the linear relationship is in doubt.

It is formulated as follows:

H_0 : There is no lack of fit (i.e., a linear model is valid)

H_1 : There is a lack of fit (i.e., a linear model is insufficient)

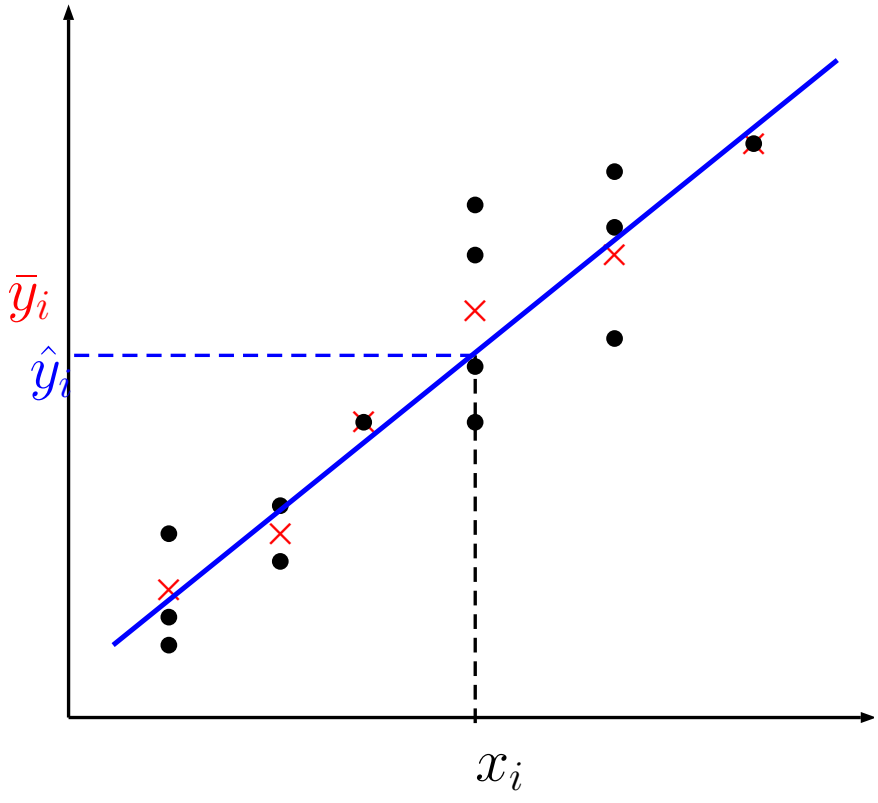
Assumption: We have **replicate observations** on the response y for at least one value of the predictor x.

Suppose that x has m distinct values (called levels) and there are n_i observations at each level $x_i, 1 \le i \le m$:

$$\{(x_i, y_{ij}) \mid j = 1, \ldots, n_i\}$$

such that $n = \sum n_i$.

We fit a regression line to all n points $\{(x_i, y_{ij}) \mid 1 \le j \le n_i, 1 \le i \le m\}$.



The fitted value at x_i is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, \ldots, m$$

To conduct the lack-of-fit test, we need to compute

$$SS_{Res} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_i)^2$$

$$SS_{PE} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 \qquad \text{where} \quad \bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$$

$$SS_{LOF} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{y}_i - \hat{y}_i)^2 = \sum_{i=1}^{m} n_i(\bar{y}_i - \hat{y}_i)^2$$

It can be shown that

$$SS_{Res} = SS_{PE} + SS_{LOF},$$

with degrees of freedom $n - 2 = (n - m) + (m - 2)$.

The **test statistic** for lack of fit is

$$F_0 = \frac{SS_{LOF}/(m-2)}{SS_{PE}/(n-m)} = \frac{MS_{LOF}}{MS_{PE}} \overset{H_0 \text{ true}}{\sim} F_{m-2,n-m}$$

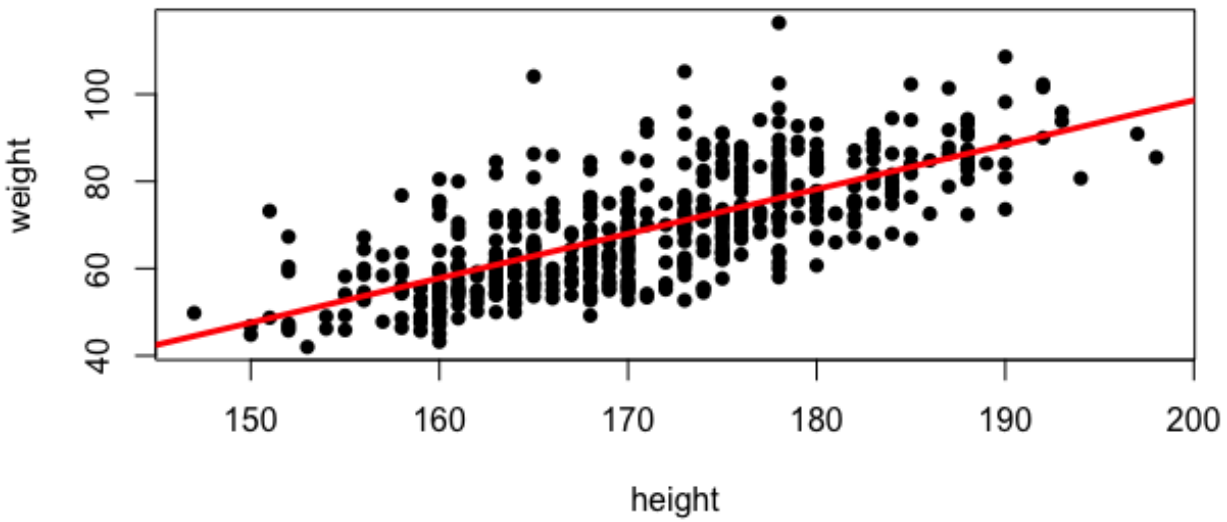
and large values of F_0 are evidence against H_0.

Therefore, to test for lack of fit, we would compute the test statistic F_0 and conclude that the regression function is not linear if

$$F_0 > F_{\alpha,m-2,n-m} \quad (\text{or } p\text{-value } < \alpha).$$

Example: weight∼height

This dataset has $n = 507$ observations and the predictor has $m = 47$ levels (after rounding off the values of height to nearest integers):



R code for conducting the lack of fit test:

```
mydata$height←round(mydata$height)

mymodel←lm(weight∼height, data=mydata)

mylofmodel←lm(weight∼as.factor(height), data=mydata)

anova(mymodel,mylofmodel)
```

Output

Model 1: weight $\sim$ height
Model 2: weight $\sim$ as.factor(height)

	Res Df	RSS	Df	Sum of Sq	F	Pr($>F$)
1	505	43725				
2	460	38956	45	4768.8	1.2514	0.1345

How to read the R output:

- $SS_{Res} = 43725$ with $df = n - 2 = 505$
- $SS_{PE} = 38956$ with $df = n - m = 460$
- $SS_{LOF} = SS_{Res} - SS_{PE} = 43725 - 38956 = 4768.8$ with $df = m - 2 = 45$
- $F_0 = \frac{SS_{LOF}/(m-2)}{SS_{PE}/(n-m)} = \frac{4768.8/45}{38956/460} = 1.2514$
- pval $= P(F_{45,460} > 1.2514) = 0.1345$, meaning that we fail to reject H_0 at level 5% (or less). Thus, it is reasonable to assume a linear model for this data set.

Summary

We talked about the following methods to check each assumption:

- The response and the regressors have a linear relationship. $\longleftarrow$ lack of fit test
- The error term ϵ has zero mean. $\longleftarrow$ no need to check
- The error term ϵ has a constant variance σ^2. $\longleftarrow$ residual plots
- The errors are uncorrelated. $\longleftarrow$ time plot (only reveals timewise dependence)
- The errors are normally distributed. $\longleftarrow$ normal quantile plot

Further learning

Section 4.2.4: Partial Regression Plots