Simple Linear regression
Checking the model

Morten Frydenberg ©
Institut for Biostatistik

The assumptions.

Independent errors?

Predicted values and residuals.

Do the errors have the same distribution?

Normal errors?

Two examples, where the model is not valid.

Leverage: a measure of influence.

Standardized residuals.

Simple linear regression: The model

Let Y_i and x_i be the data for the ith person.

$$Y_i = \beta_0 + \beta_1 \cdot x_i + E_i \quad E_i \sim N\left(0, \sigma^2\right)$$

This model is based on the **assumptions**:

1. The **expected** value of Y is a **linear function** of x.
2. The **unexplained** random deviations are **independent**.
3. The unexplained random deviations have the **same distributions**.
4. This distribution is **normal**.

Checking the model: Independent errors ?

Assumption no. 2: *the errors should be* ***independent***, is mainly checked by considering **how the data was collected**.

The assumption is **violated** if

•some of the persons are **relatives** (and some are not) and the dependent variable have some **genetic** component.

•some of the persons were **measured** using one instrument and others using another.

•in general if the persons were sampled in **clusters**.

Predicted values and residuals

$$Y_i = \beta_0 + \beta_1 \cdot x_i + E_i \quad E_i \sim N\left(0, \sigma^2\right)$$

Based on the estimates we can calculate the **predicted** (fitted) values and the **residuals**:

$$\text{Predicted value:} \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i$$

$$\text{Residual:} \quad r_i = y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 \cdot x_i\right)$$

The **predicted value** is the best guess of y_i (based on the estimates) for the ith person.

The **residual** is a guess of E_i (based on the estimates) for the ith person.

STATA:
```
predict PEFR_hat if e(sample),xb
predict PEFR_res if e(sample),resid
```

Checking the model: Linearity and identical distributed errors

Assumption no. 1:
The **expected** value of Y is a **linear function** of x.
Assumption no. 3:
The unexplained random deviations have the **same distributions**.

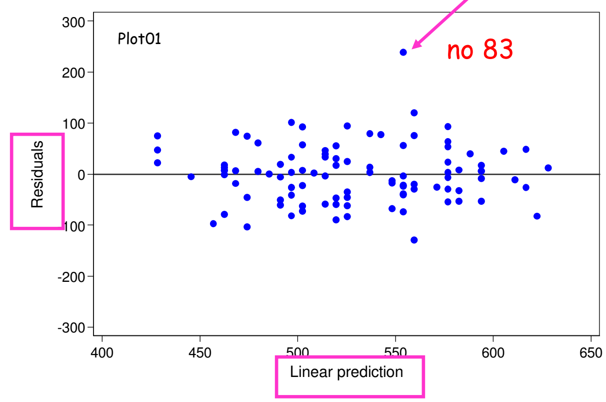
These are checked by inspecting the following plots of:

- **Residuals** versus **predicted**
- **Residuals** versus x

Morten Frydenberg Linear and Logistic regression - Note 1.2 5

Checking the model: Linearity and identical distributed errors

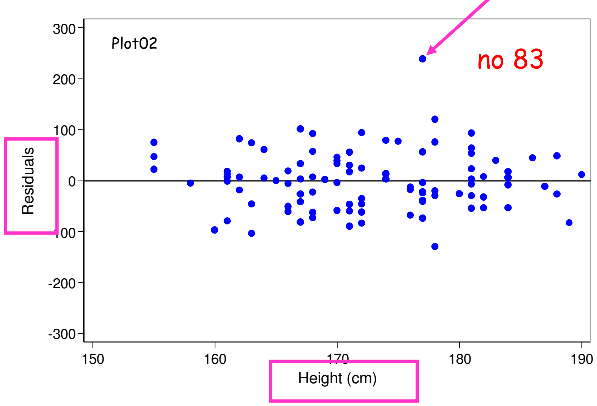
No problems! Except this outlier



Morten Frydenberg Linear and Logistic regression - Note 1.2 6

Checking the model: Linearity and identical distributed errors

No problems! Except this outlier

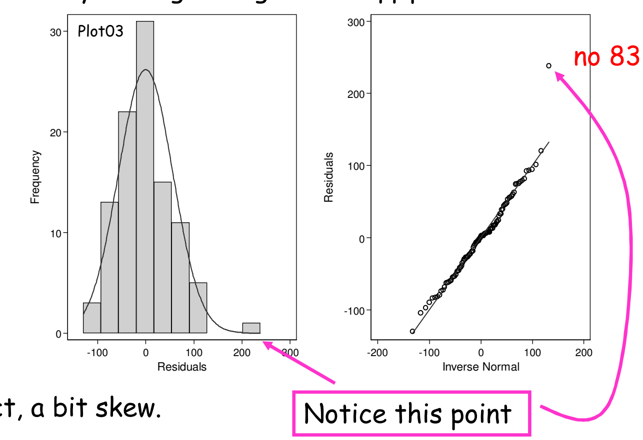


Morten Frydenberg Linear and Logistic regression - Note 1.2 7

Checking the model: Normal errors ?

Assumption no. 4: *the errors should be* ***normal distibuted***.

This is checked by making histograms or qq-plots of the residuals:

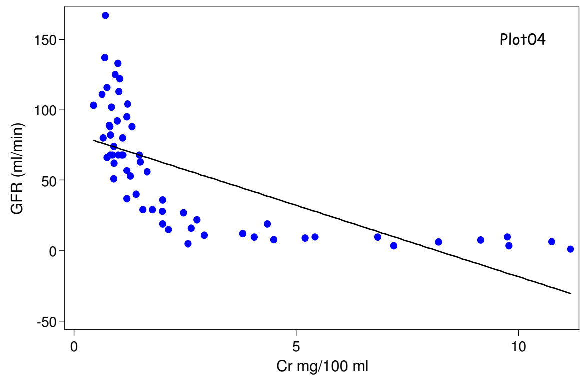


Not perfect, a bit skew.

Morten Frydenberg Linear and Logistic regression - Note 1.2 8

Assumptions violated: Example 2

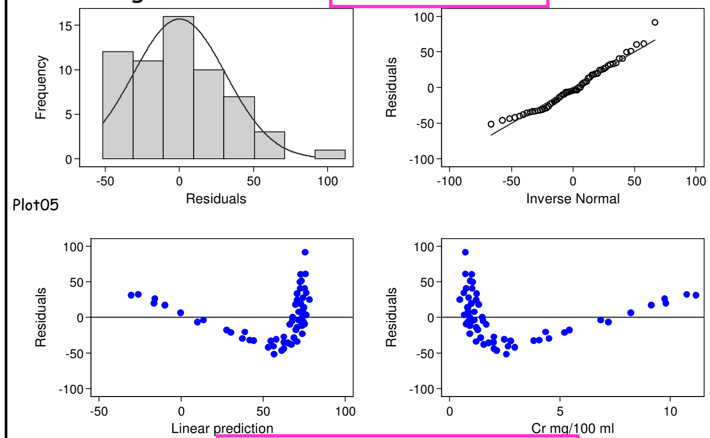
The relation between GFR and Serum Creatinine



Clearly non-linear!

Assumptions violated: Example 2

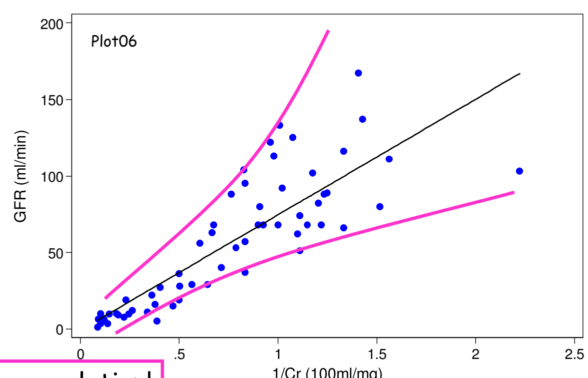
Checking the model **Close to normal**



Clearly not constant mean!

Assumptions violated: Example 3

The relation between GFR and 1/Serum Creatinine

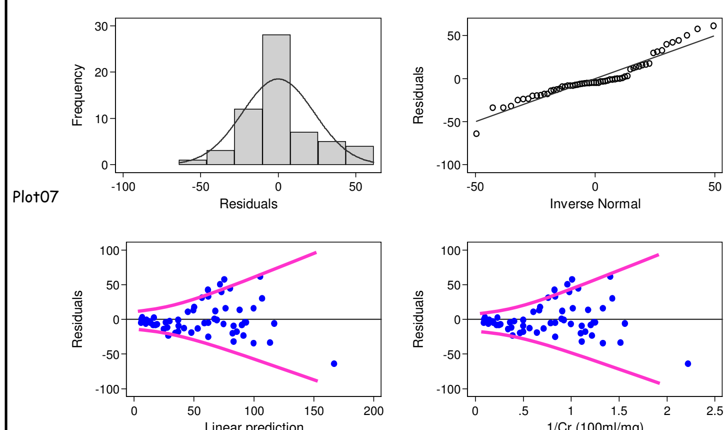


A linear relation!

Increasing variation!

Assumptions violated: Example 3

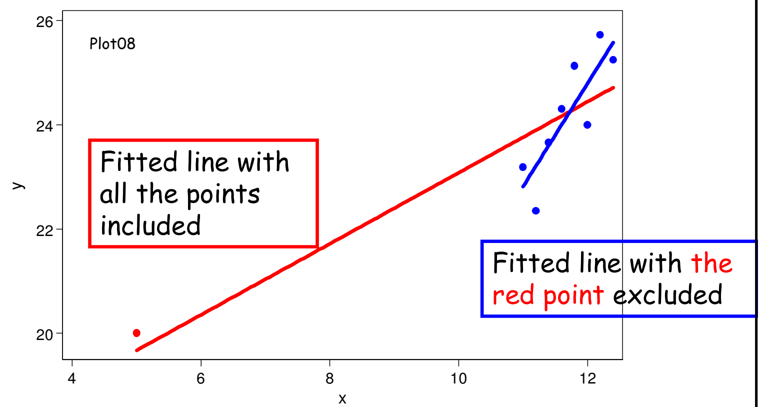
Checking the model **Close to normal**



Increasing variation! **Increasing variation!**

Influential data points: Example 4

Not all data points have the same influence on the estimates:



The data point works like a **leverage** (vægtstang).

Influential data points: Leverage

The influence of a data point is sometime measured by its **leverage**:

$$h_i = \frac{1}{n} + \frac{\left(x_i - \bar{x}\right)^2}{\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2}$$

Large values imply that the estimates and/or the standard errors is **highly influenced** by this observation.

$$0 \le h_i \le 1$$

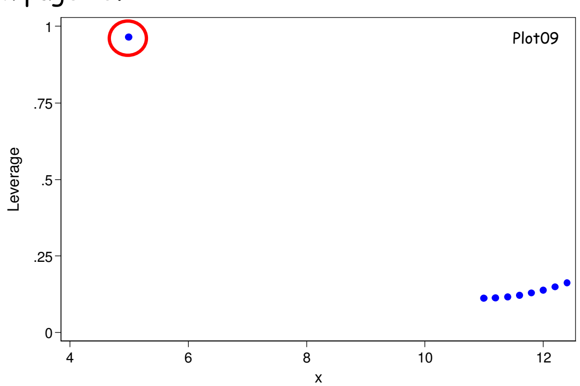
Notice, it is a function only of the **independent** variable, x and the sample size.

The leverage for a given data point depends on **how far away** its **independent** variable is from the **average value**.

STATA: `predict PEFR_lev if e(sample), leverage`

Influential data points Leverage

A **leverage** versus **independent variable** for the example on page 13.



The data point with the 'extreme' x value has very high leverage – as expected.

Types of residuals: Standardized residuals

The (**unstandardized**) residual: $r_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 \cdot x_i\right)$

Has mean zero but **non-constant** variance: $\text{sd}\left(r_i\right) = \sigma\sqrt{1-h_i}$

I. e. residuals from points with **high leverage** have **smaller variance**, than residuals from points with small leverage.

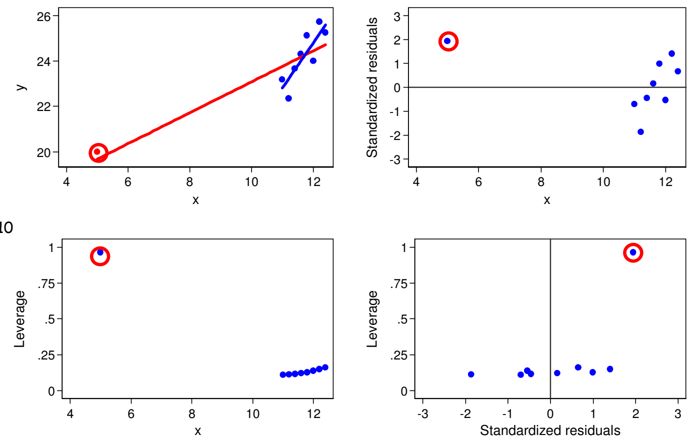
Due to this one often use the **standardized** residual:

$$z_i = \frac{r_i}{\hat{\sigma}\sqrt{1-h_i}}$$

This will have **variance 1**, if the **model is true**.

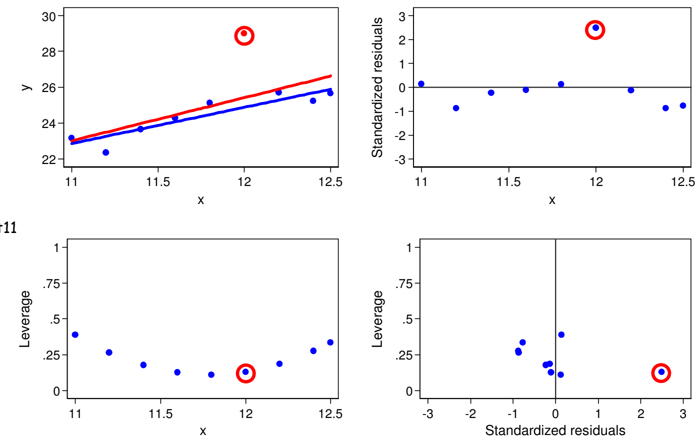
STATA: `predict PEFR_zres if e(sample),rstandard`

Influential data points? Example 4



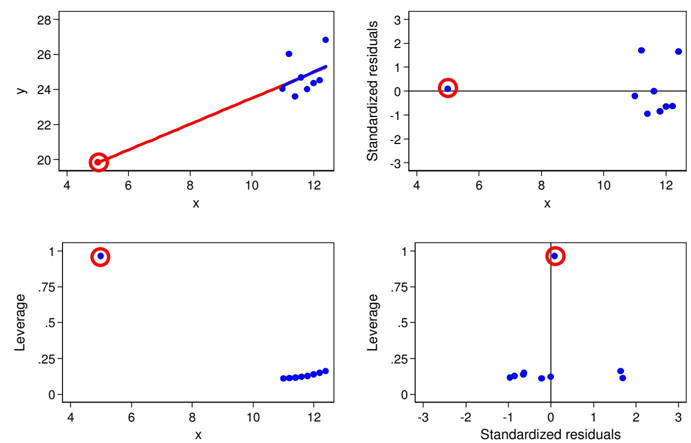
Large leverage and standardized residual large!

Influential data points? Example 5



Small leverage but standardized residual large!

Influential data points? Example 6



Large leverage but standardized residual ok!

Influential data points? Example 6

Results with using all data:

Root MSE = 1.0282

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	.7364484	.1594519	4.62	0.002	.3594045	1.113492
_cons	16.1386	1.78019	9.07	0.000	11.92912	20.34808

Results without the point with high leverage:

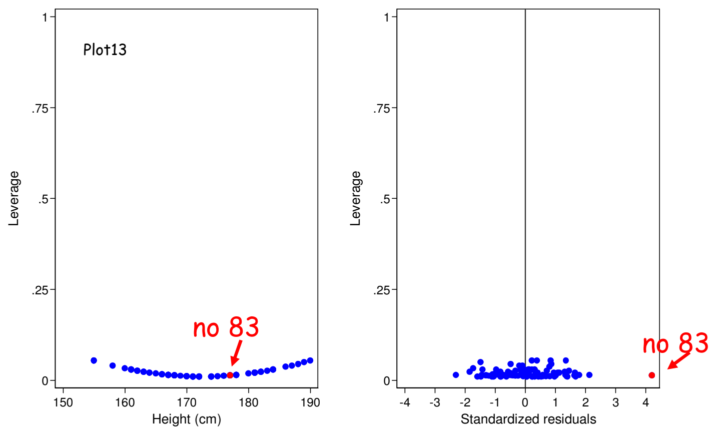
Root MSE = 1.1099

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	.8080605	.8563254	0.94	0.382	-1.287292	2.903413
_cons	15.2985	10.02669	1.53	0.178	-9.235928	39.83292

Point estimates unchanged

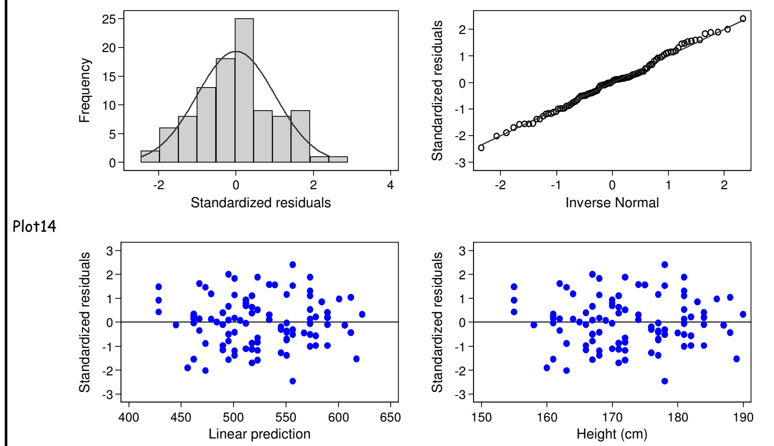
Standard errors much larger. **Confidence intervals** much wider.

The PEFR example: leverage and standardized residuals



Leverages are small, observation no. 83 has large residual

The PEFR example: Excluding observation no 83



Some comments on checking a (simple) linear regression

Always consider the design: **How was the data collected**?

This has implications for the validity of the **statistical model**.

And it has implications for the **interpretation** of the results.

Observations with **high leverages** have 'extreme' values of the **independent** variable.

These observation will have **high impact** on the results, but might not be 'representative'.

Sometimes it is best to **exclude** these from the analysis.

Observation with **large residuals**, that is observed y value far away from expected, should be **checked for errors**.

Prediction interval for future value

The **true line** is given as : $y=\beta_0+\beta_1\cdot x$

and **estimated** by plugging in the estimates $\hat{y}=\hat{\beta}_0+\hat{\beta}_1\cdot x$

The standard **deviation** for a **new observation** is given by:

$$\text{sd}\left(\hat{\beta}_0+\hat{\beta}_1\cdot x+E\right)=\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x-\overline{x})^2}{\sum(x_i-\overline{x})^2}}$$

with the 95% (pointwise) **prediction interval**

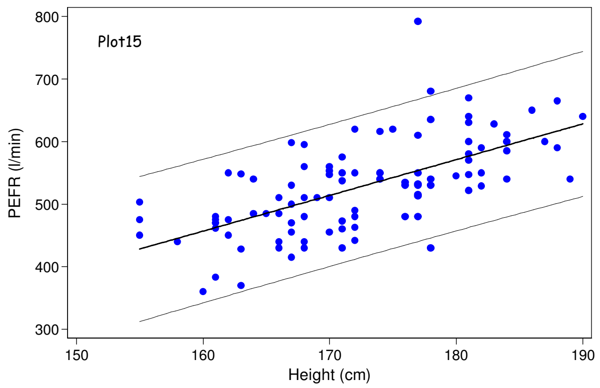
$$\hat{\beta}_0+\hat{\beta}_1\cdot x\pm t_{n-2}^{0.975}\cdot\text{sd}\left(\hat{\beta}_0+\hat{\beta}_1\cdot x+E\right)$$

Many programs can make a plot with the fitted line and its prediction limits.

In STATA its done by the `lfitci` and graph command, the option `stdf`

Prediction interval for future value

```
twoway ///
 (scatter PEFR height, mco(blue) msym(O))          ///
 (lfitci PEFR height, stdf clpat(1) cip(rline) )    ///
  ,legend(off) ytit("PEFR (l/min)")
```



Morten Frydenberg Linear and Logistic regression - Note 1.2 25