

Simple Linear regression
Checking the model
Morten Frydenberg ©
Department of Biostatistics, Aarhus Univ, Denmark

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.

Morten Frydenberg Linear and Logistic regression - Note 1.2 1

Simple linear regression: The model

Let Y_i and x_i be the data for the i th person.

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

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**.

Morten Frydenberg Linear and Logistic regression - Note 1.2 2

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**.

Morten Frydenberg Linear and Logistic regression - Note 1.2 3

Predicted values and residuals

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

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

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

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

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

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

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

Morten Frydenberg Linear and Logistic regression - Note 1.2 4

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:

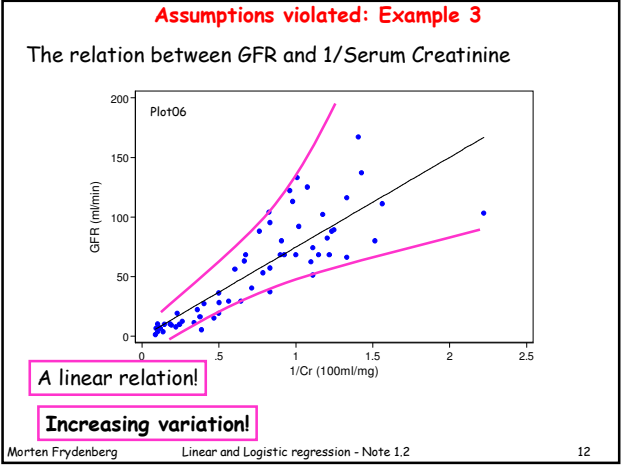
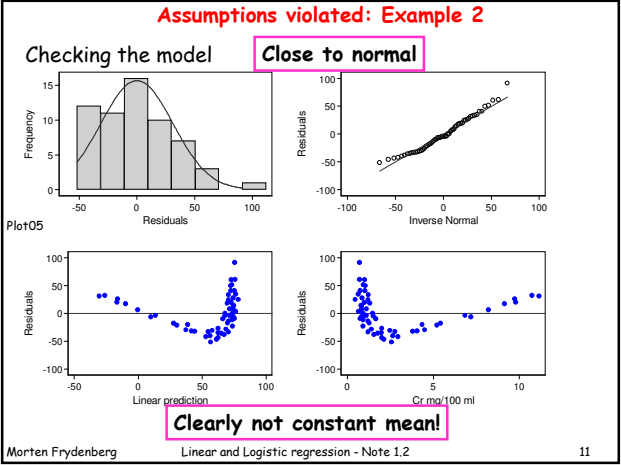
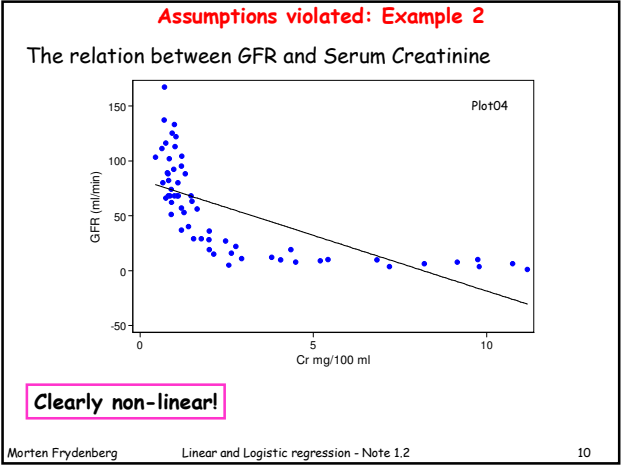
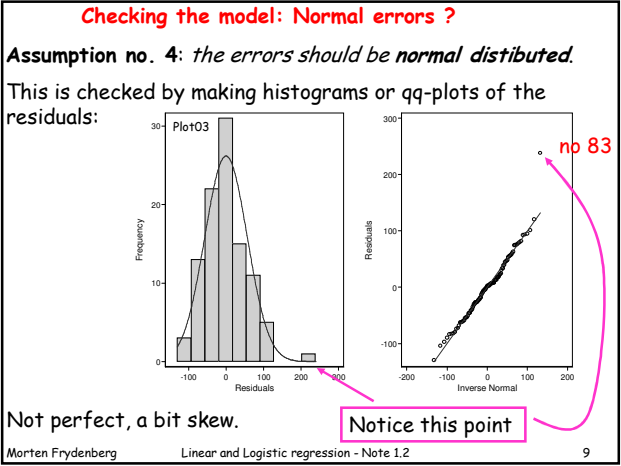
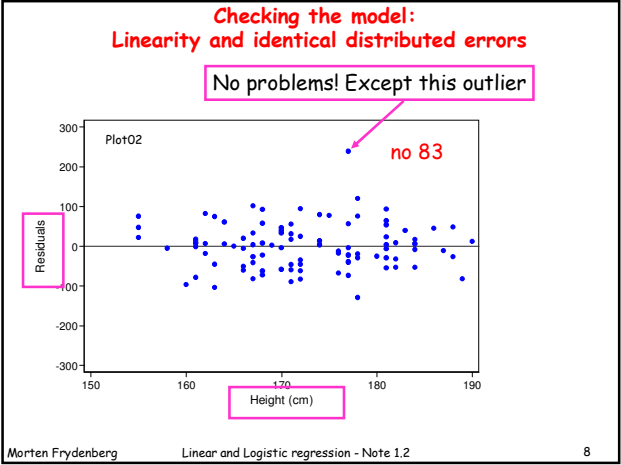
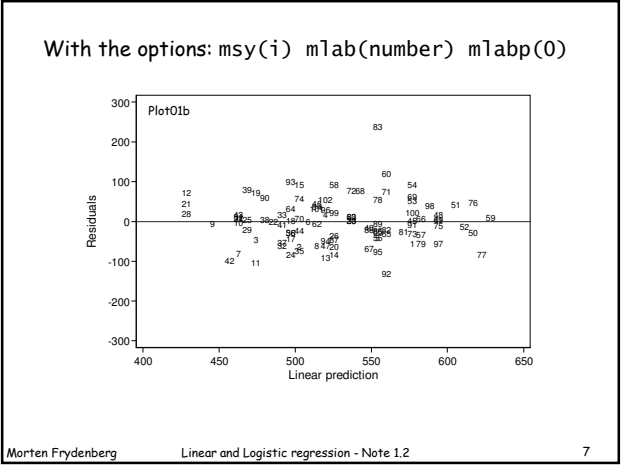
- **Residuals versus predicted**
- **Residuals versus x**

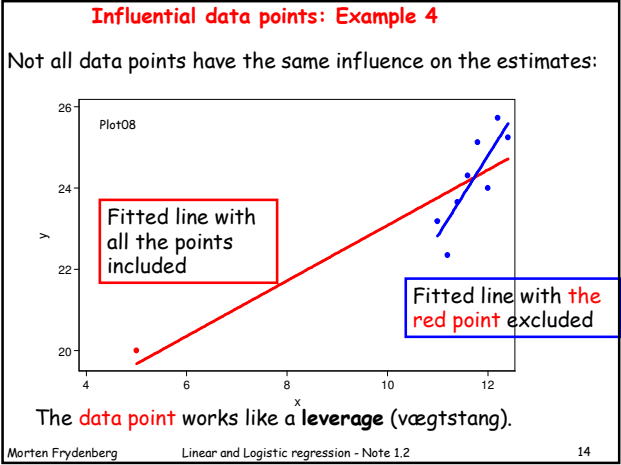
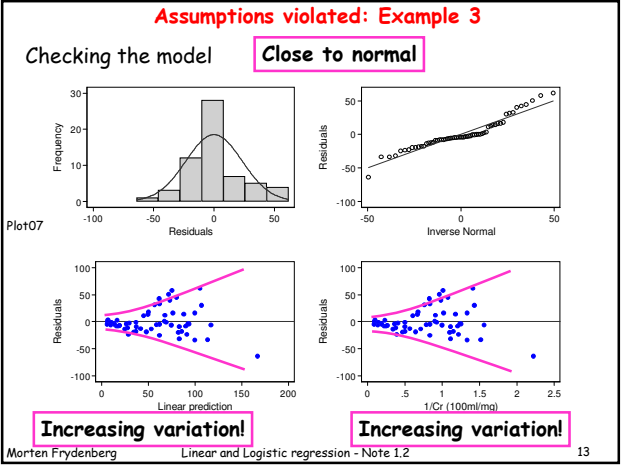
Morten Frydenberg Linear and Logistic regression - Note 1.2 5

Checking the model:
Linearity and identical distributed errors

No problems! Except this outlier

Morten Frydenberg Linear and Logistic regression - Note 1.2 6





Influential data points: Leverage

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

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

A **large value** implies that the estimates and/or the standard errors are **highly influenced** by this observation.

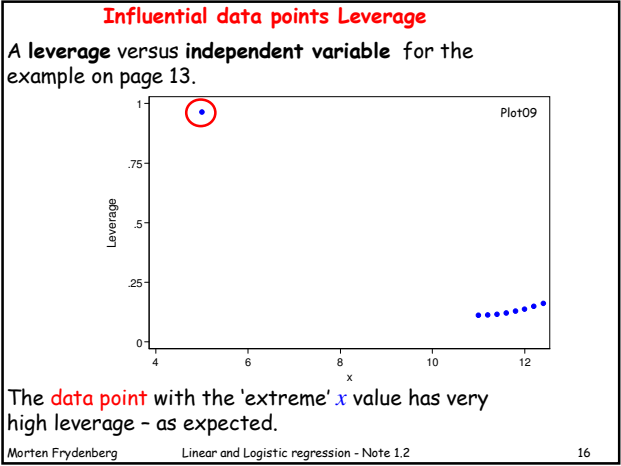
Note that $0 \leq h_i \leq 1$

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`

Morten Frydenberg Linear and Logistic regression - Note 1.2 15



Types of residuals: Standardized residuals

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

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

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

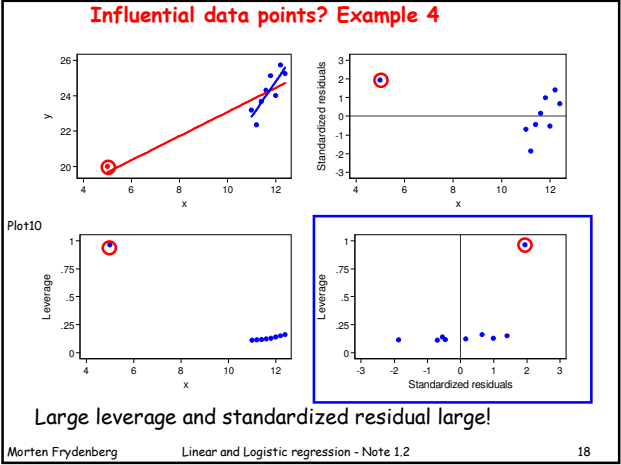
Due to this one often use the **standardized** residual:

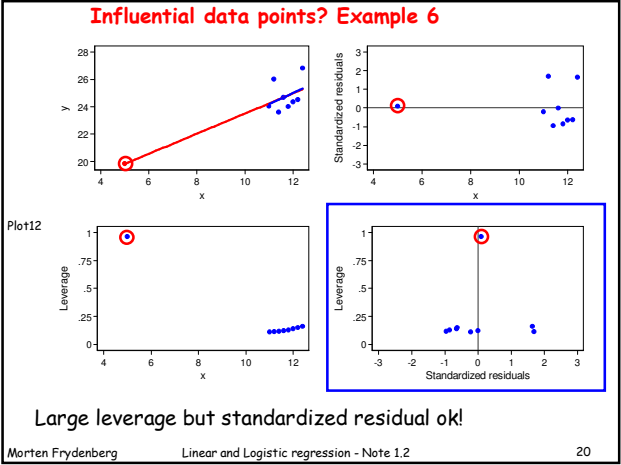
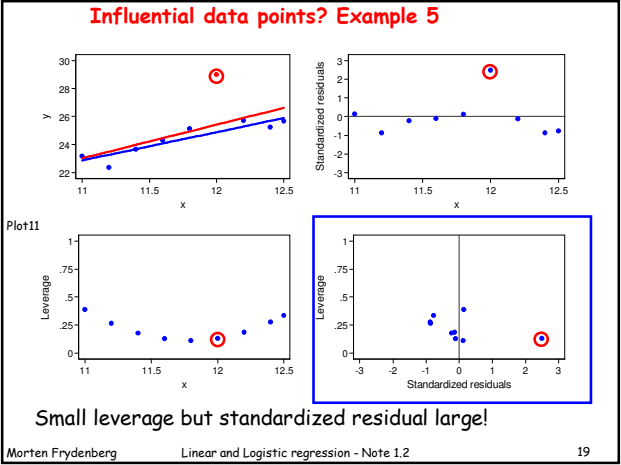
$$z_i = \frac{r_i}{\hat{\sigma} \sqrt{1 - h_i}}$$

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

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

Morten Frydenberg Linear and Logistic regression - Note 1.2 17





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

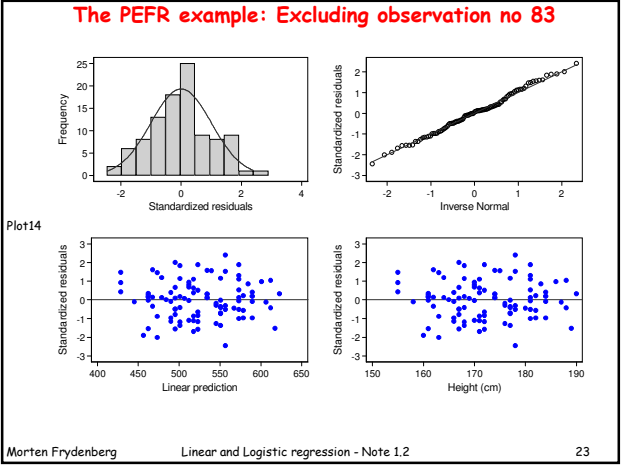
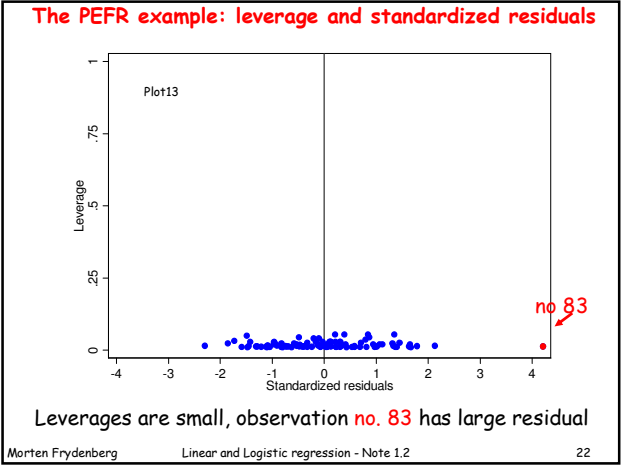
Results without the point with high leverage:
Root MSE = 1.1099

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.

Morten Frydenberg Linear and Logistic regression - Note 1.2 21



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 observations will have **high impact** on the results, but might not be 'representative'.
Sometimes it is best to **exclude** these from the analysis.

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

Morten Frydenberg Linear and Logistic regression - Note 1.2 24

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:

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

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

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

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***

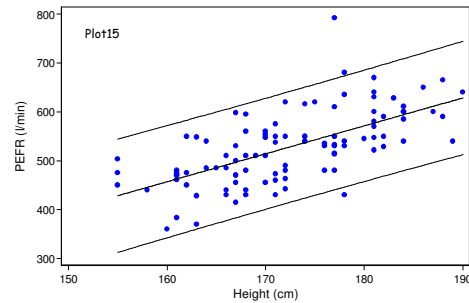
Morten Frydenberg

Linear and Logistic regression - Note 1.2

25

Prediction interval for future value

```
twoway ///
(scatter PEFR height, mco(blue) msym(0)) ///
(lfitci PEFR height, stdf c1pat(1) cip(r1line)) ///
. llegend(off) vtitle("PEFR (l/min)")
```



Morten Frydenberg

Linear and Logistic regression - Note 1.2

26