

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 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**.

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 with another.

•in general if the persons were sampled in clusters.

Predicted values and residuals

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

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
```

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:

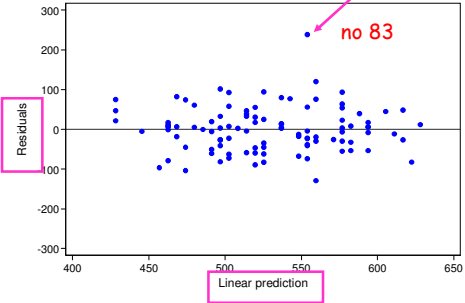
• **Residuals versus predicted**

• **Residuals versus x**

Checking the model:

Linearity and identical distributed errors

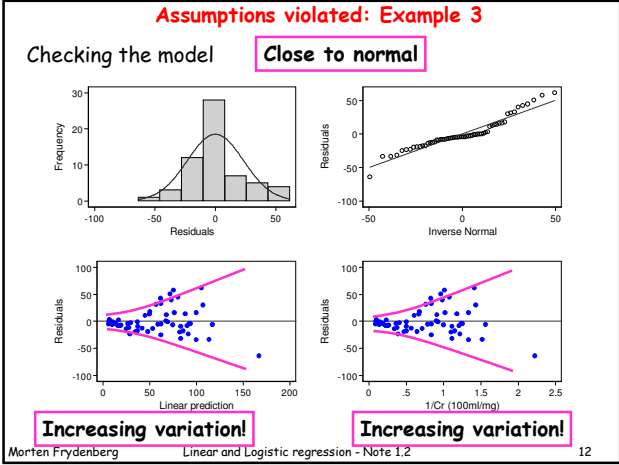
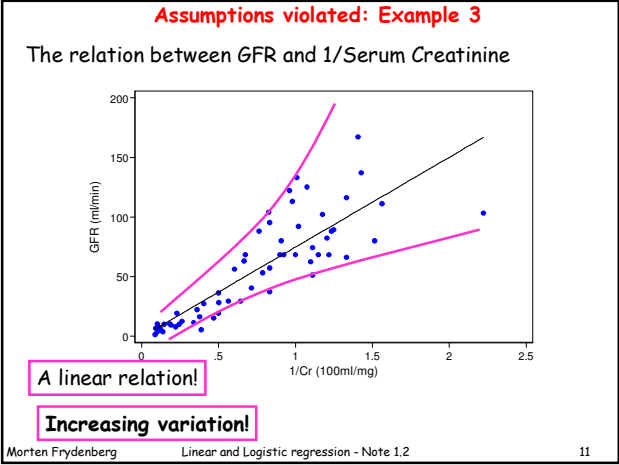
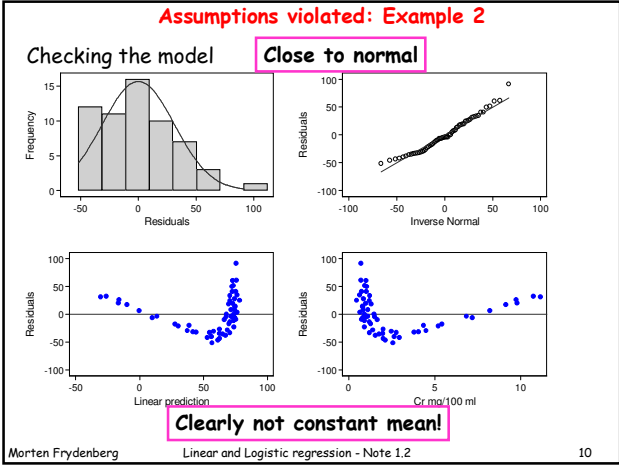
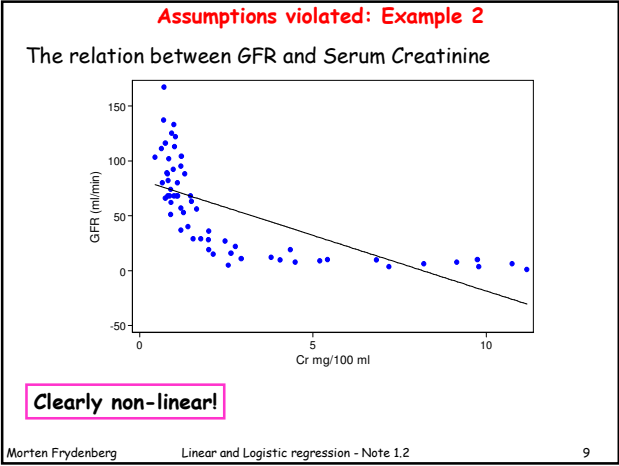
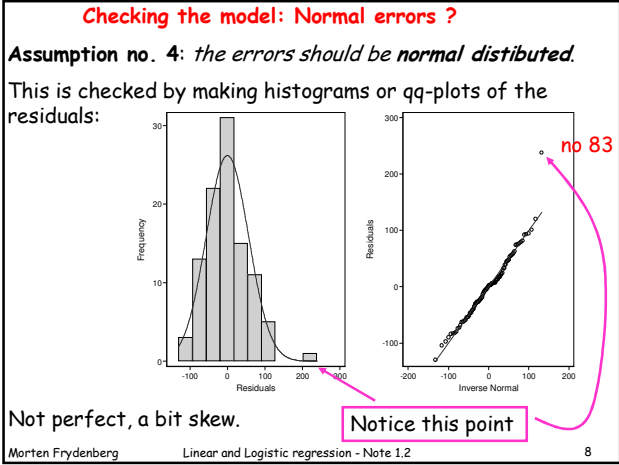
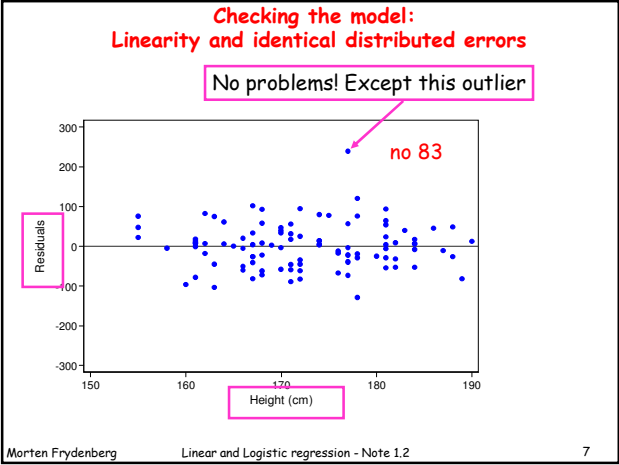
No problems! Except this outlier

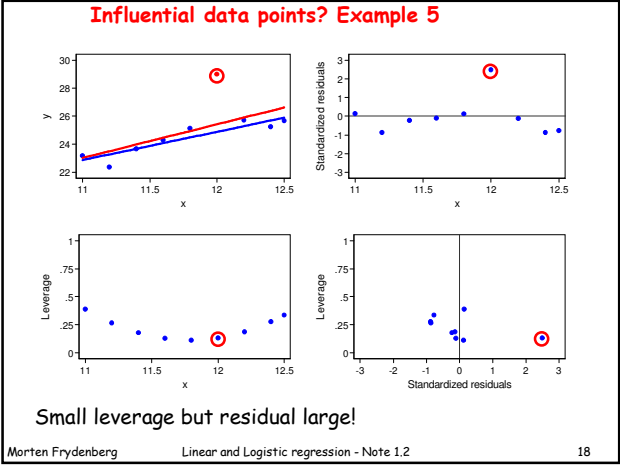
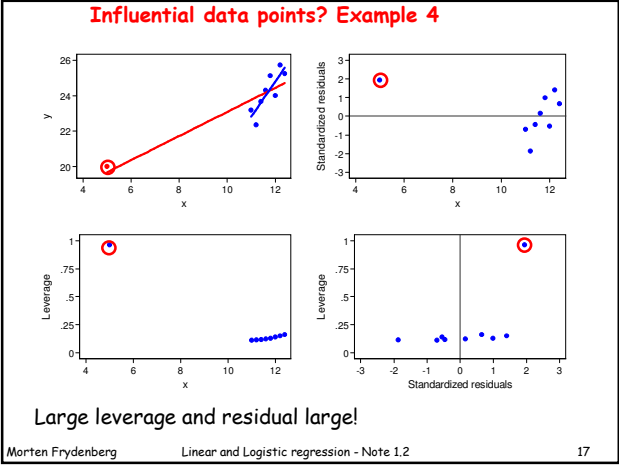
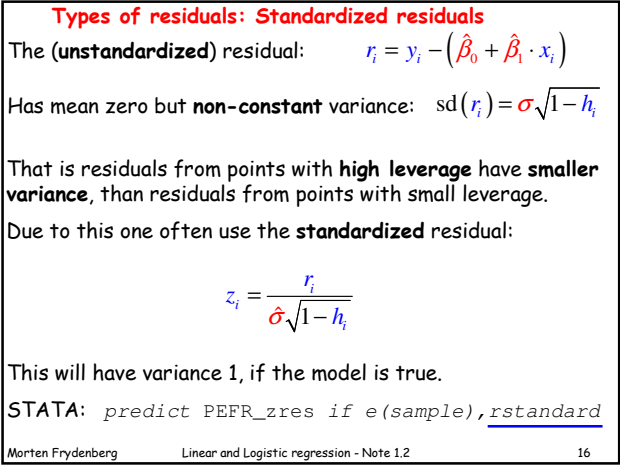
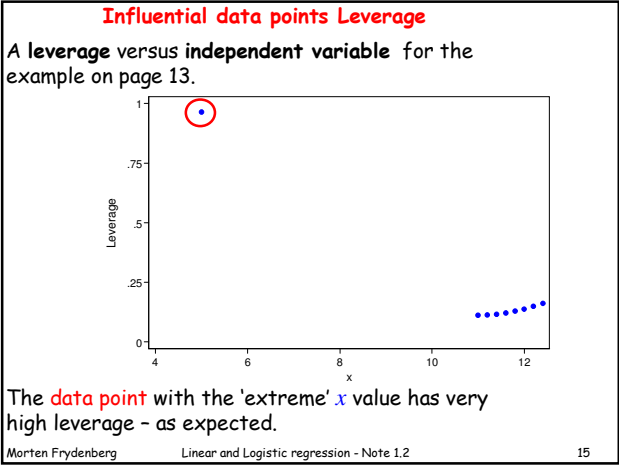
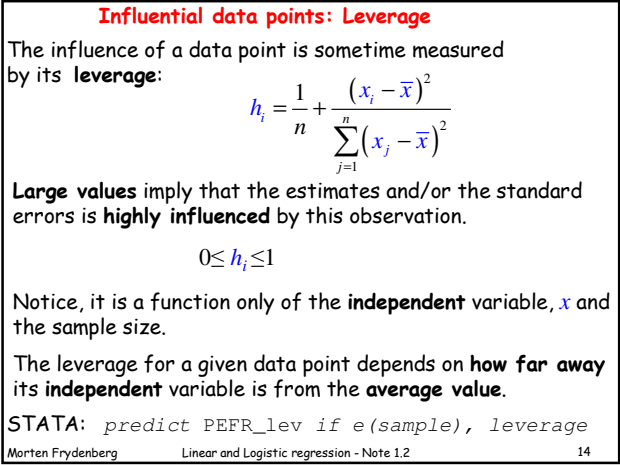
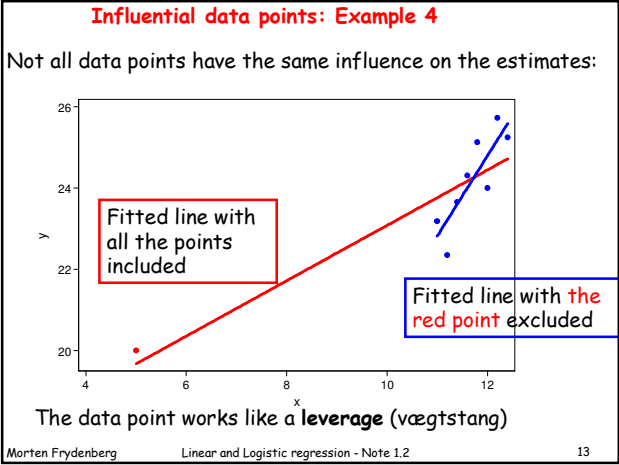


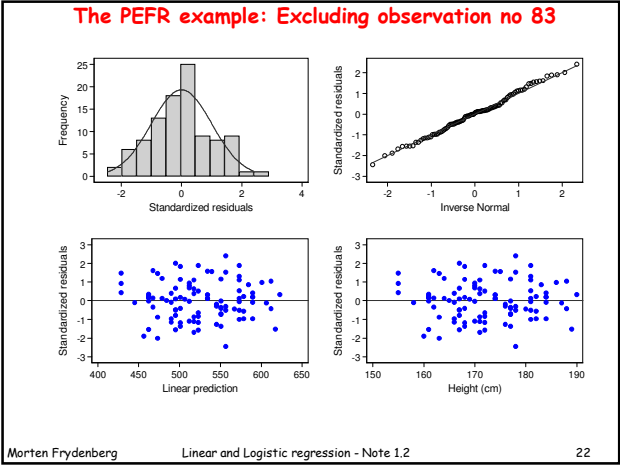
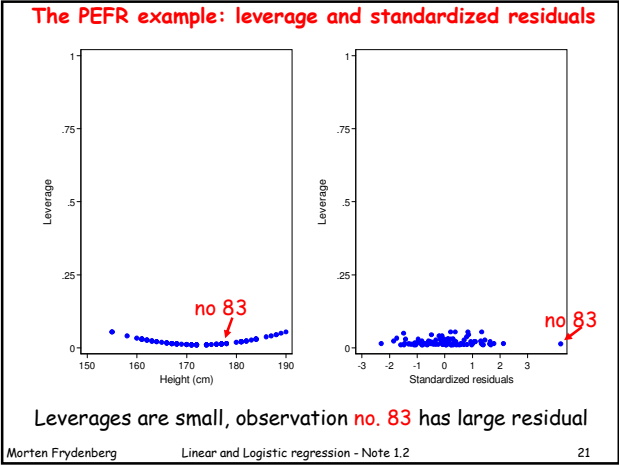
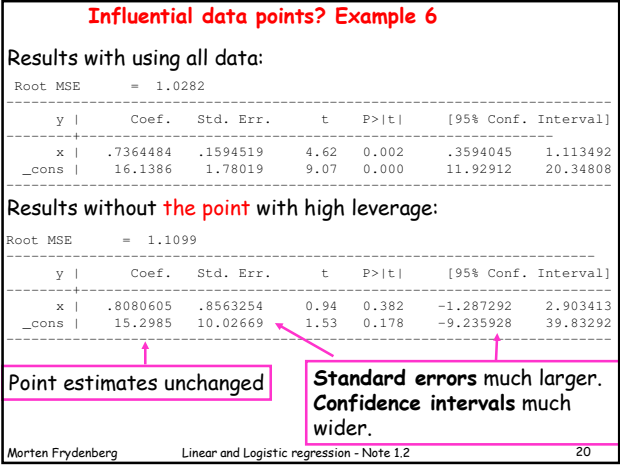
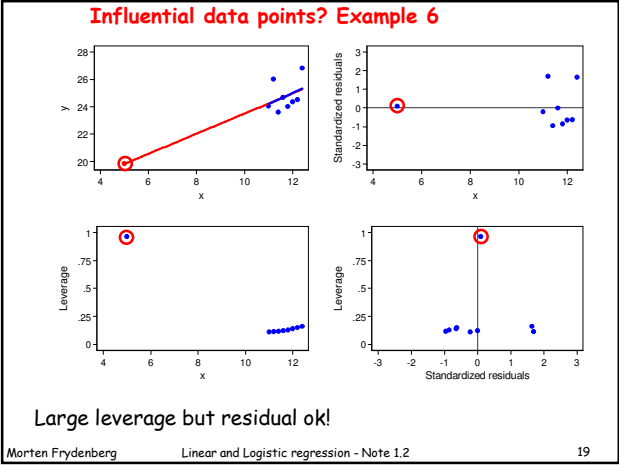
The figure is a scatter plot with 'Residuals' on the y-axis (ranging from -300 to 300) and 'Linear prediction' on the x-axis (ranging from 400 to 650). Most data points are clustered around the zero line on the y-axis. One point, labeled 'no 83', is a significant outlier located at approximately (550, 250).

Linear ang Logistic Regression: Note 1.2

1







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.

Morten Frydenberg Linear and Logistic regression - Note 1.2 23

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:

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

with the 95% (pointwise) prediction interval

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

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 24

