Regression diagnostics & model fit

Johan A. Elkink
School of Politics & International Relations
University College Dublin

18 November 2019

Outline

Unbiasedness

An unbiased estimator of a coefficient β is an estimator where the **mean of the sampling distribution** is identical to the true β. I.e. $E(\hat{\beta}) = \beta$.

The bias of an estimator is thus $E(\hat{\beta}) - \beta$.

Best unbiased

Often, many estimators could be defined whereby $E(\hat{\beta}) - \beta = 0$, i.e. that are unbiased. The **best unbiased** estimator is the estimator that leads to an unbiased estimated with the **smallest variance**.

Another term for this is **efficiency** - a smaller standard error means a more efficient estimator.

If all assumptions underlying OLS hold, OLS is BLUE, i.e. the **best linear unbiased estimator** of β.

Mean Squared Error

Although unbiasedness is often considered more important than high efficiency, there is a certain trade-off between the two. It is better to have a slightly biased but highly efficient estimator than an unbiased but very inefficient estimator.

The **Mean Squared Error** (MSE) refers to the **weighted square error** of an estimator, with equal weights for variance and bias.

Outline

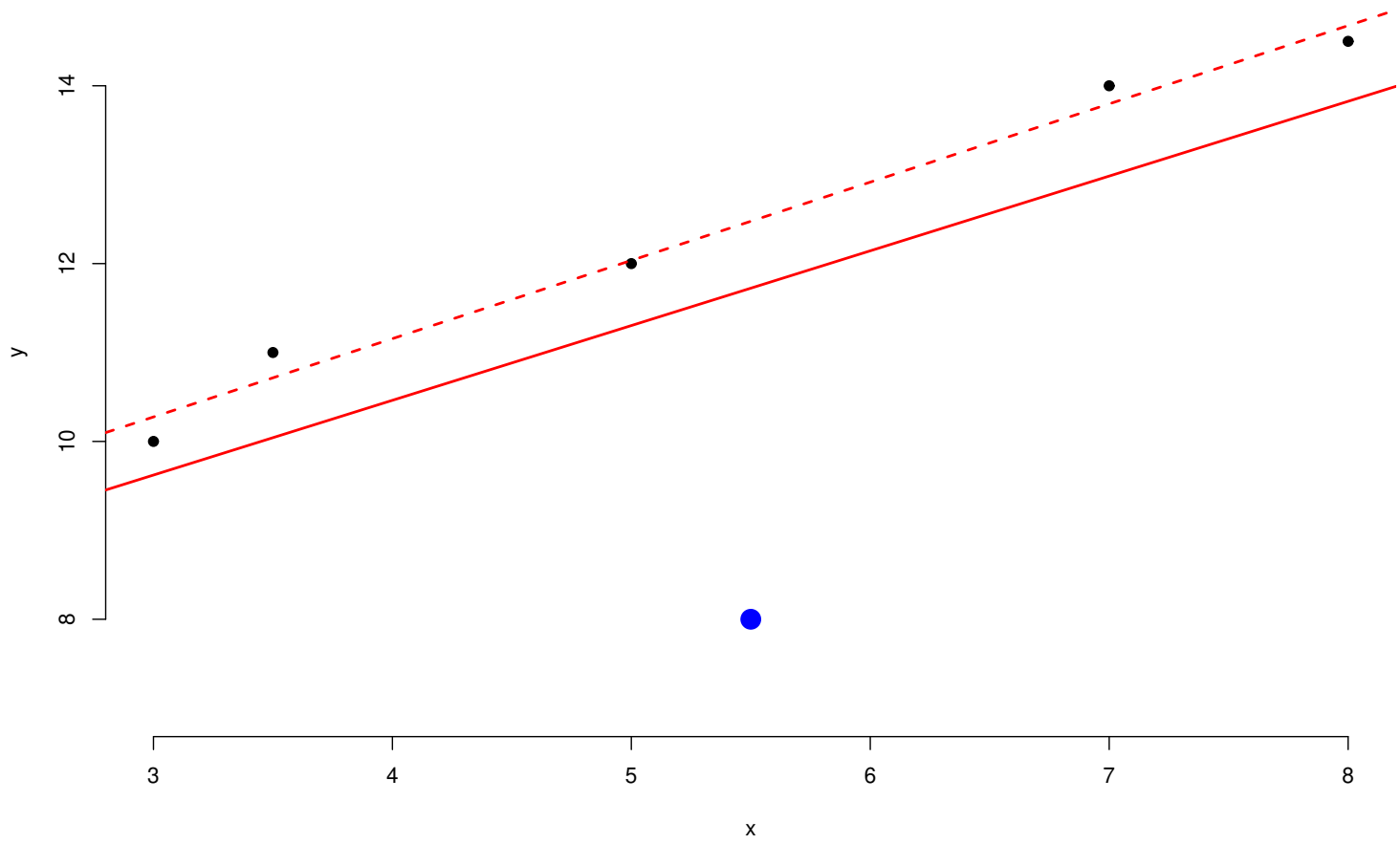
Outliers

An **outlier** is a point on the regression line where the error is large.

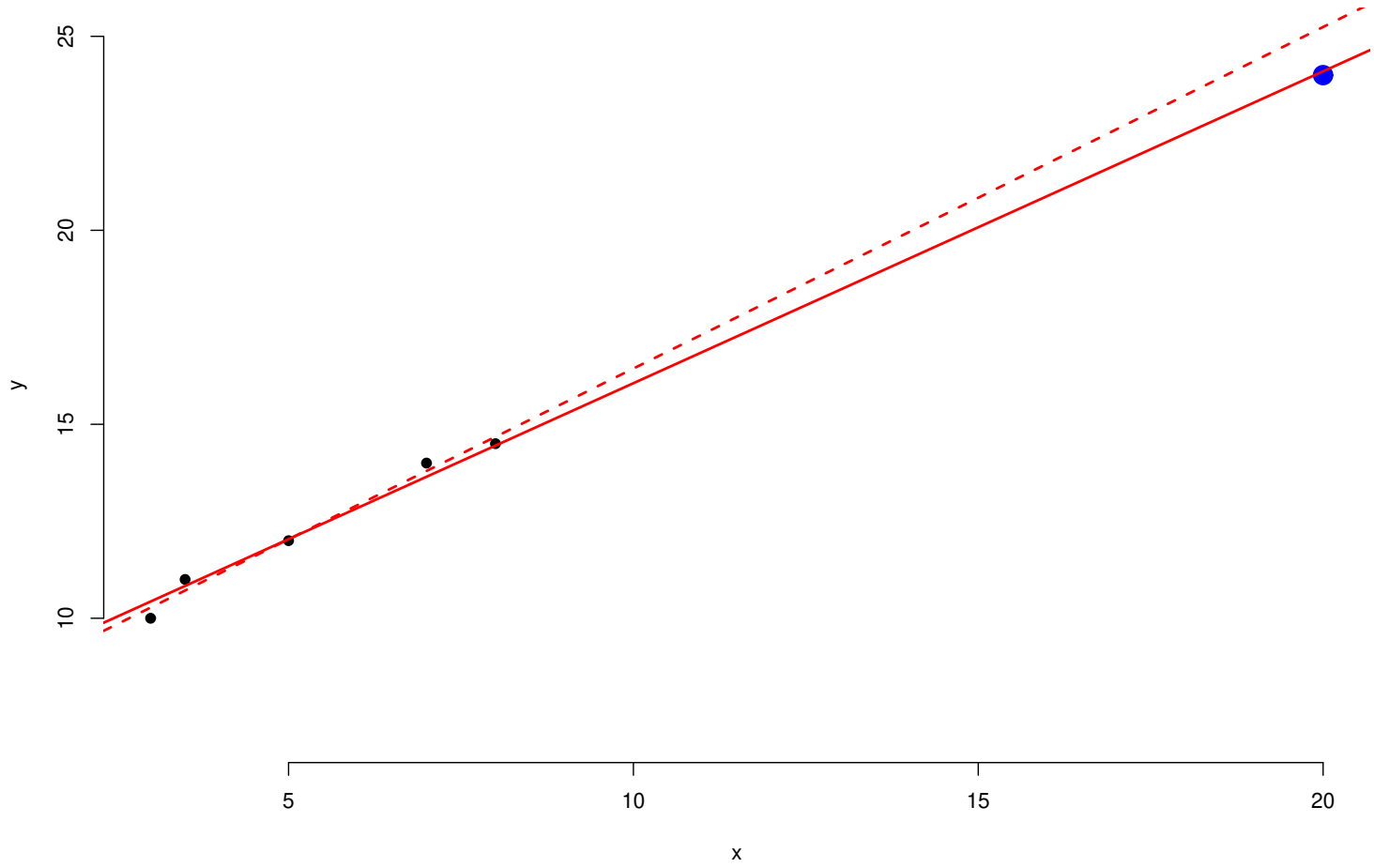
A point with high **leverage** is located far from the other points.

A high leverage point that strongly influences the regression line is called an **influential point**.

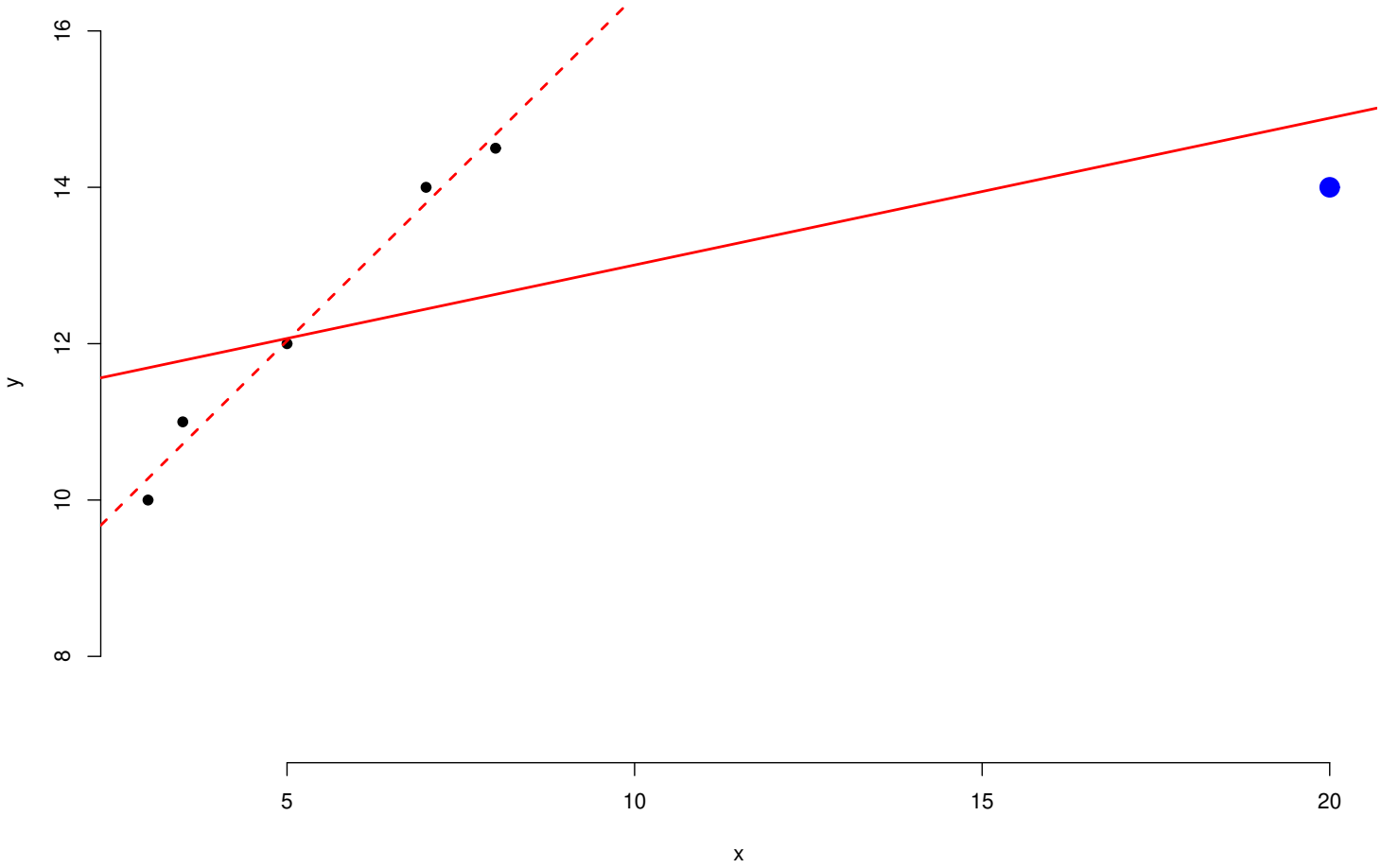
Outlier, low leverage, low influence



High leverage, low influence



High leverage, high influence



Outliers: solution

"Automatic rejection of outliers is not always a wise procedure. Sometimes the outlier is providing information that other data points cannot due to the fact that it arises from an unusual combination of circumstances which may be of vital interest and requires further investigation rather than rejection. As a general rule, outliers should be rejected out of hand only if they can be traced to causes such as errors of recording the observations or setting up the apparatus [in a physical experiment]. Otherwise, careful investigation is in order." (Draper & Smith (1998), as cited in Gujarati (2003))

OLS assumptions: specification

- Linear in parameters.

Note that this does not imply that you cannot include non-linearly transformed variables, e.g. $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$ can be estimated with OLS.

Nonlinearity

The estimator assumes a linear relation, since the model is always of the form $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$.

A nonlinear relation can often be made linear, however, by transforming one of the variables, e.g. by taking a square or a log. Other forms of nonlinearity require specialised models.

OLS assumptions: specification

- Linear in parameters.
- No extraneous variables in X.
- No omitted independent variables.

Omitted relevant variables

A relevant variable is a variable that is correlated with the dependent variable.

If the omitted variable (Z) is correlated with an independent variable (X), the estimate of $\hat{\beta}_X$ will be biased. If $\bar{Z} \neq 0$, $\hat{\beta}_{intercept}$ will be biased.

If Z is uncorrelated with X, the estimated standard error for $\hat{\beta}_X$ is biased upwards.

Specification tests

F-tests can be used to test a full model against a restricted model.

This is material for a more advanced course.

OLS assumptions: specification

- Linear in parameters.
- No extraneous variables in X.
- No omitted independent variables.
- Parameters to be estimated are constant.

OLS assumptions: specification

- Linear in parameters.
- No extraneous variables in X.
- No omitted independent variables.
- Parameters to be estimated are constant.
- Number of parameters is less than the number of cases, $k < n$.

OLS assumptions: errors

- Errors have an expected value of zero given X.

$E(\varepsilon) = 0$

The mean of the errors is assumed to be zero.

This is violated when

- there is systematic measurement error in the dependent variable
- a relevant variable with a non-zero mean is excluded
- the dependent variable is not continuous or is truncated or censored
- a constant is not included and should have been

If $E(\varepsilon) \neq 0$, the estimate of the intercept is biased.

Outline

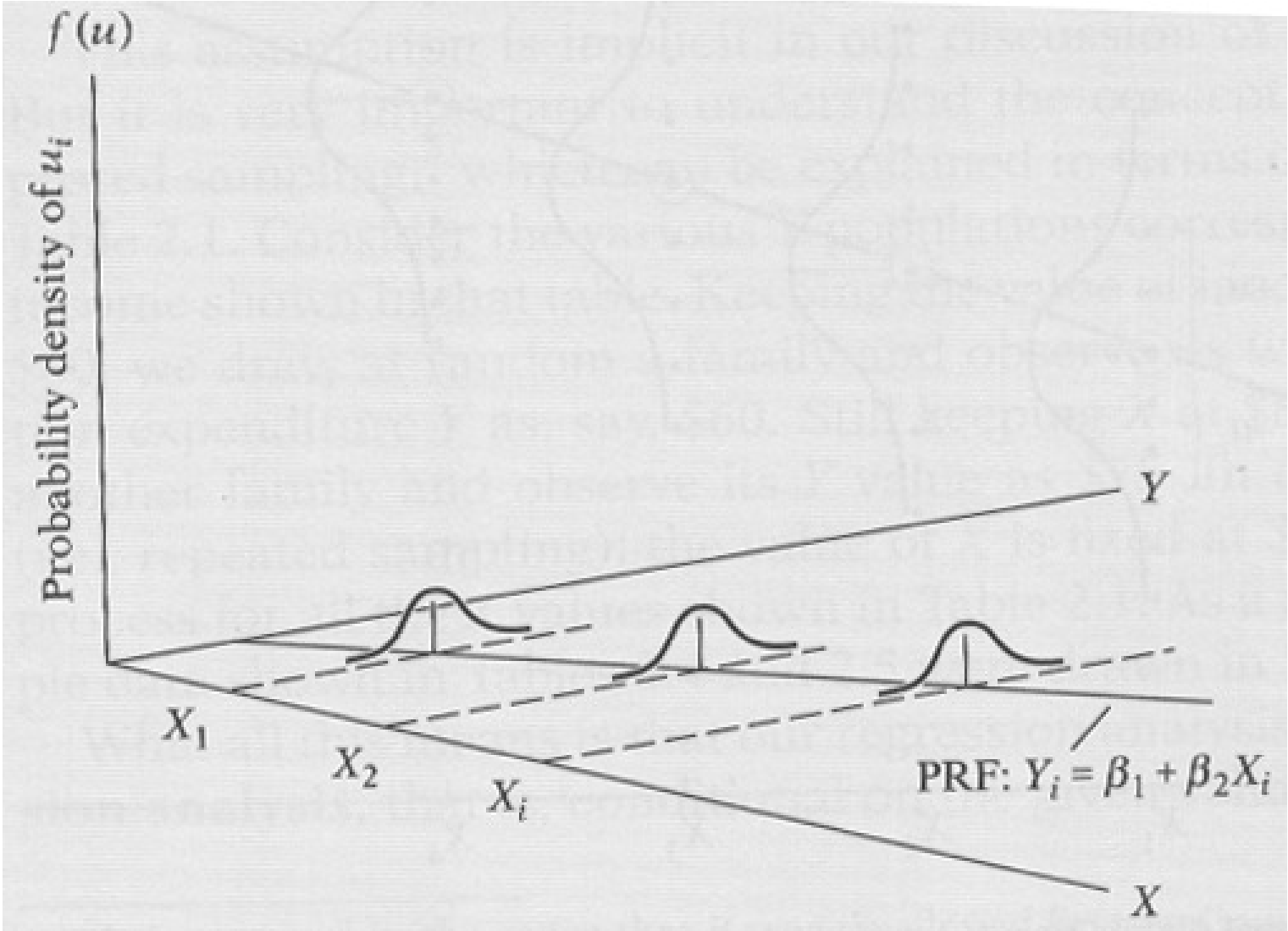
OLS assumptions: errors

- Errors have an expected value of zero given X.
- Errors are normally distributed.

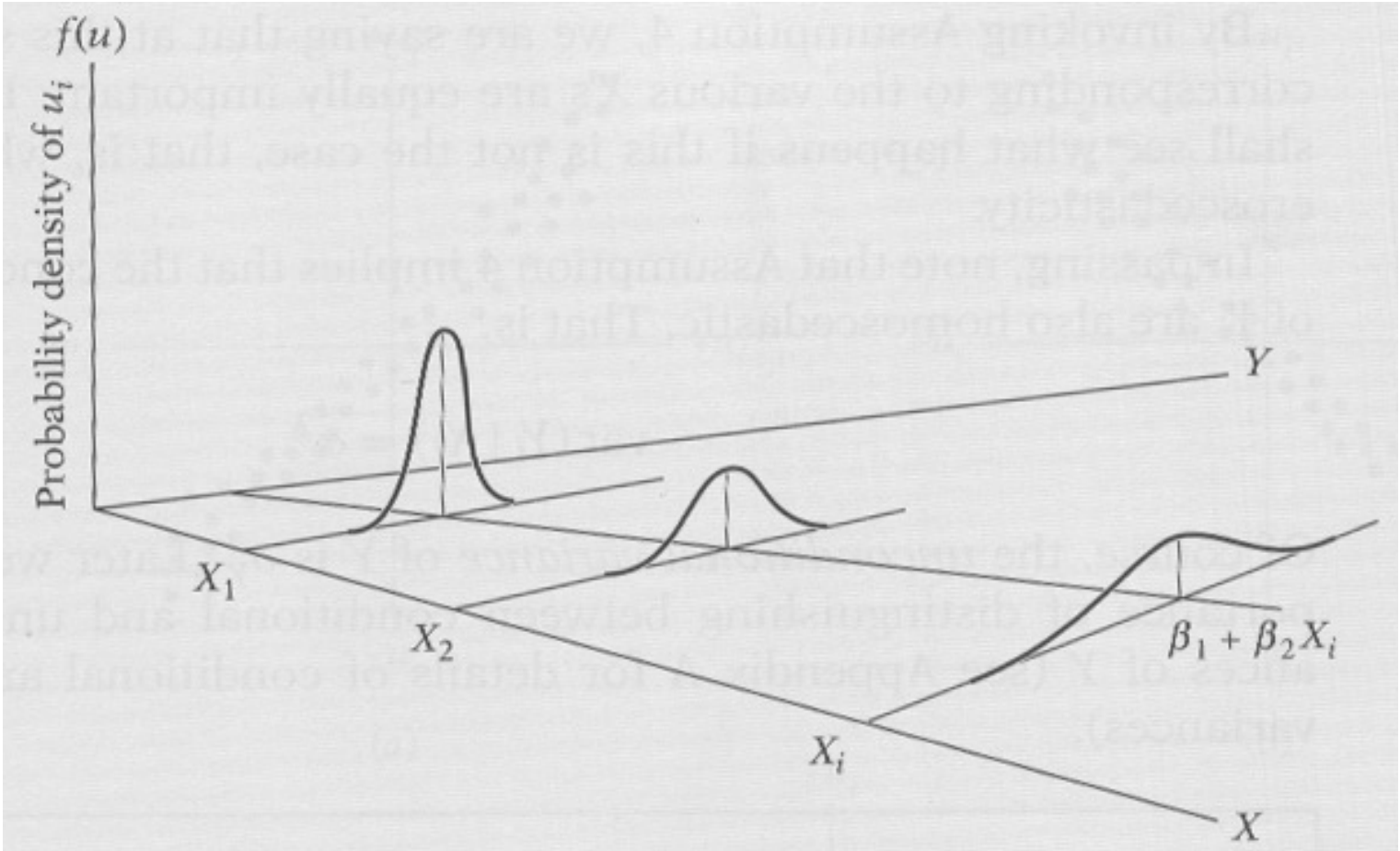
OLS assumptions: errors

- Errors have an expected value of zero given X.
- Errors are normally distributed.
- Errors have a constant variance.

Homoscedasticity



Heteroscedasticity

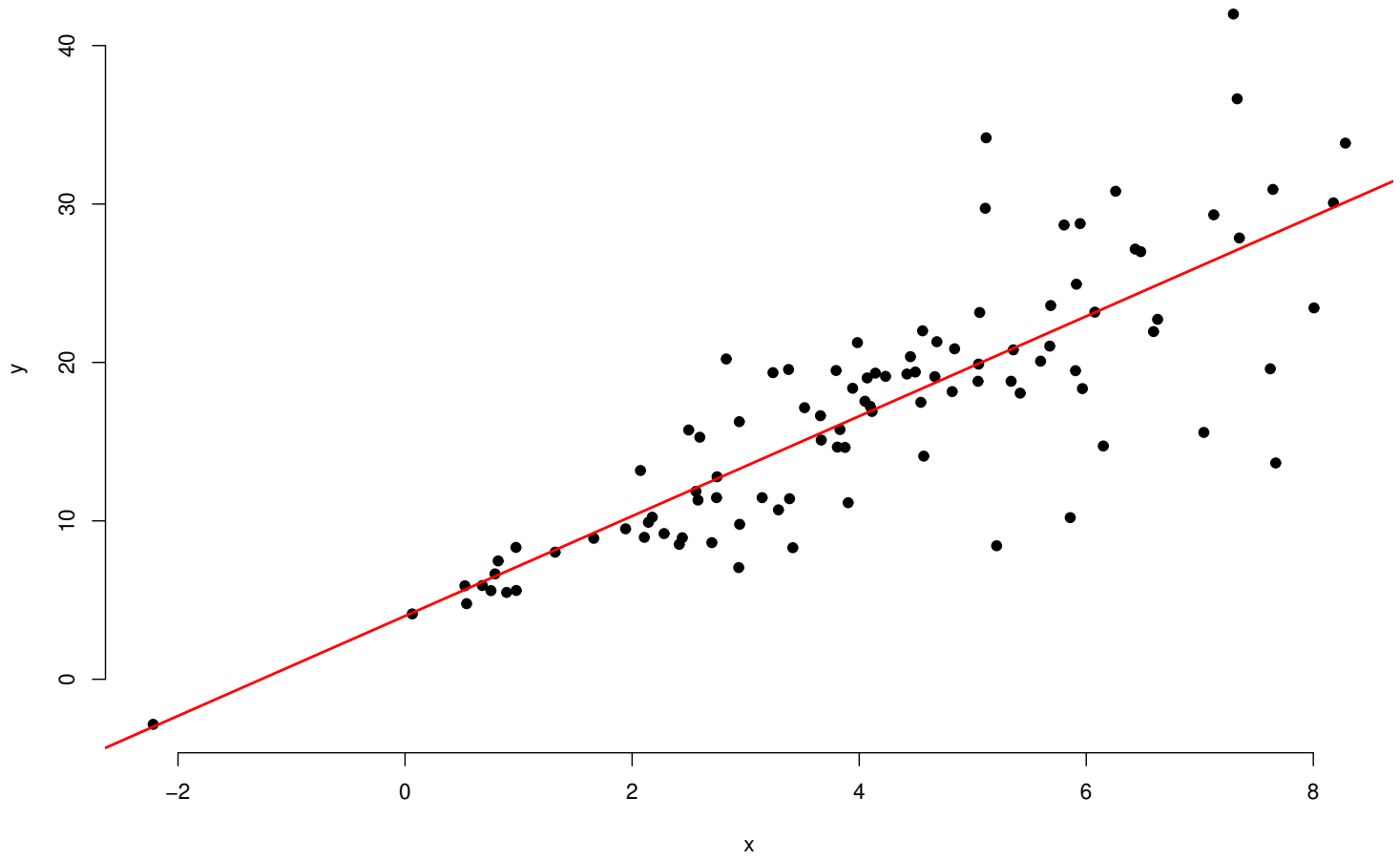


Residual plots: heteroscedasticity

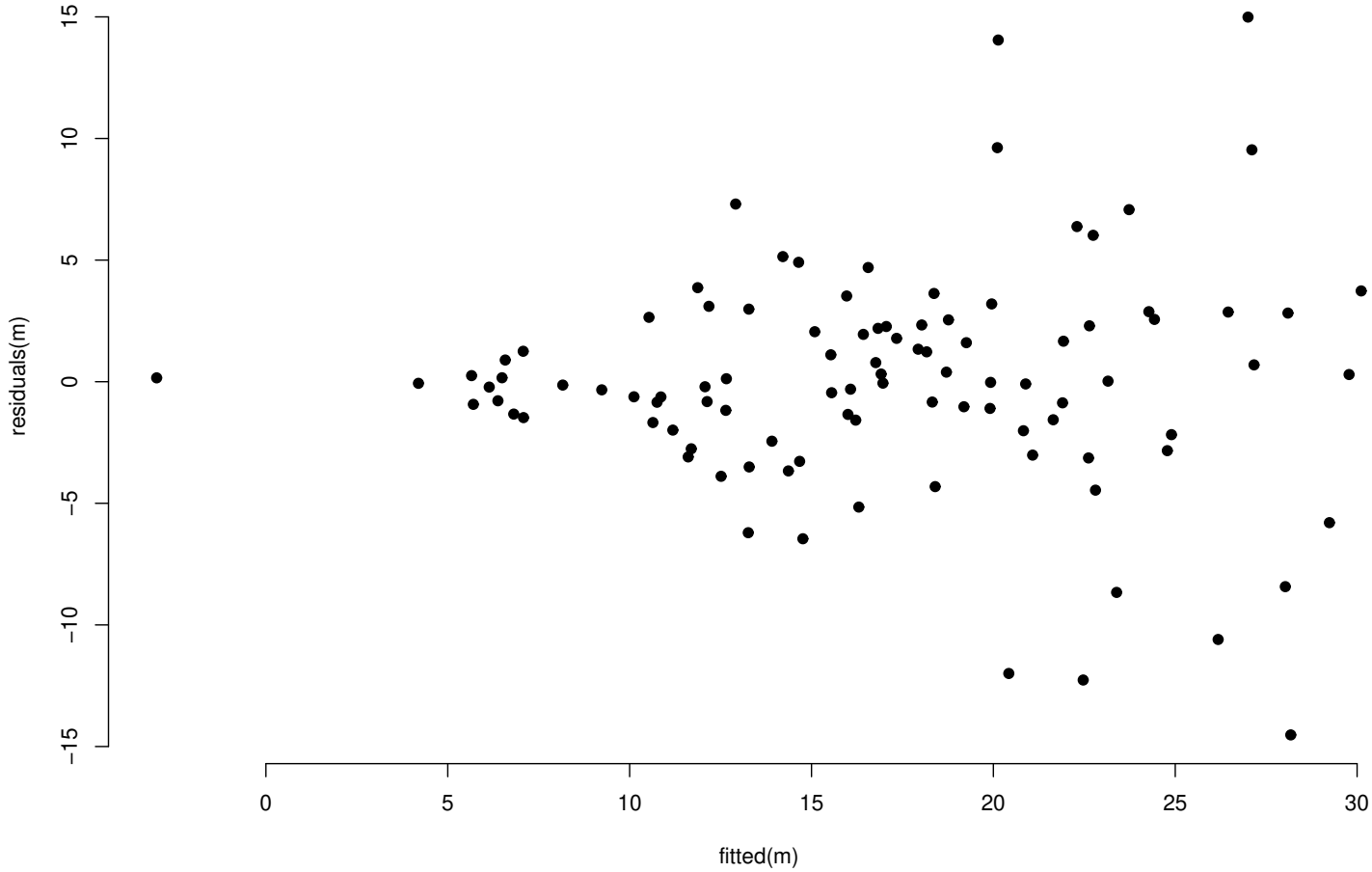
To detect heteroscedasticity (unequal variances), it is useful to plot:

- Residuals against fitted values
- Residuals against dependent variable
- Residuals against independent variable(s)

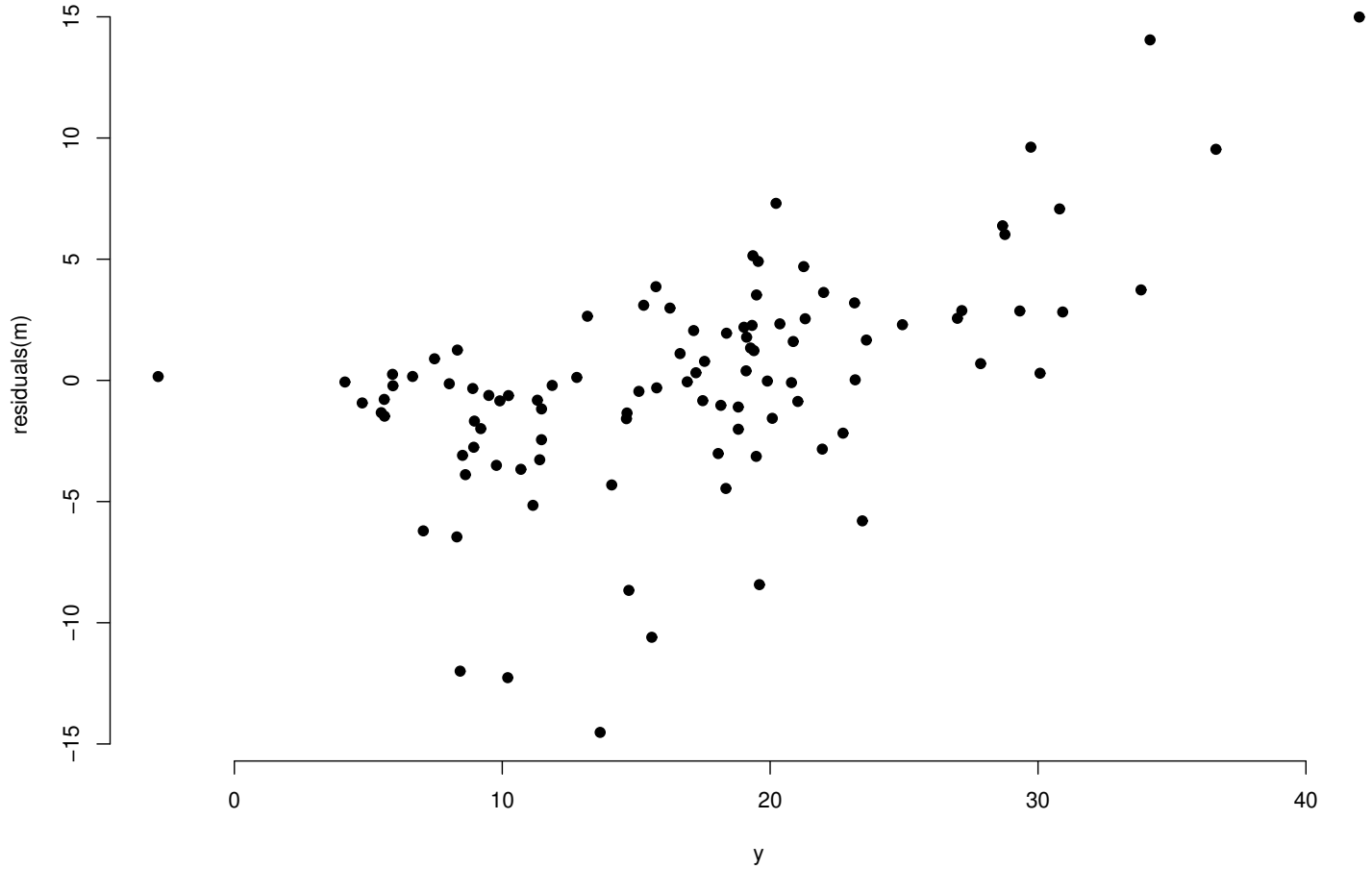
Residual plots: y by x



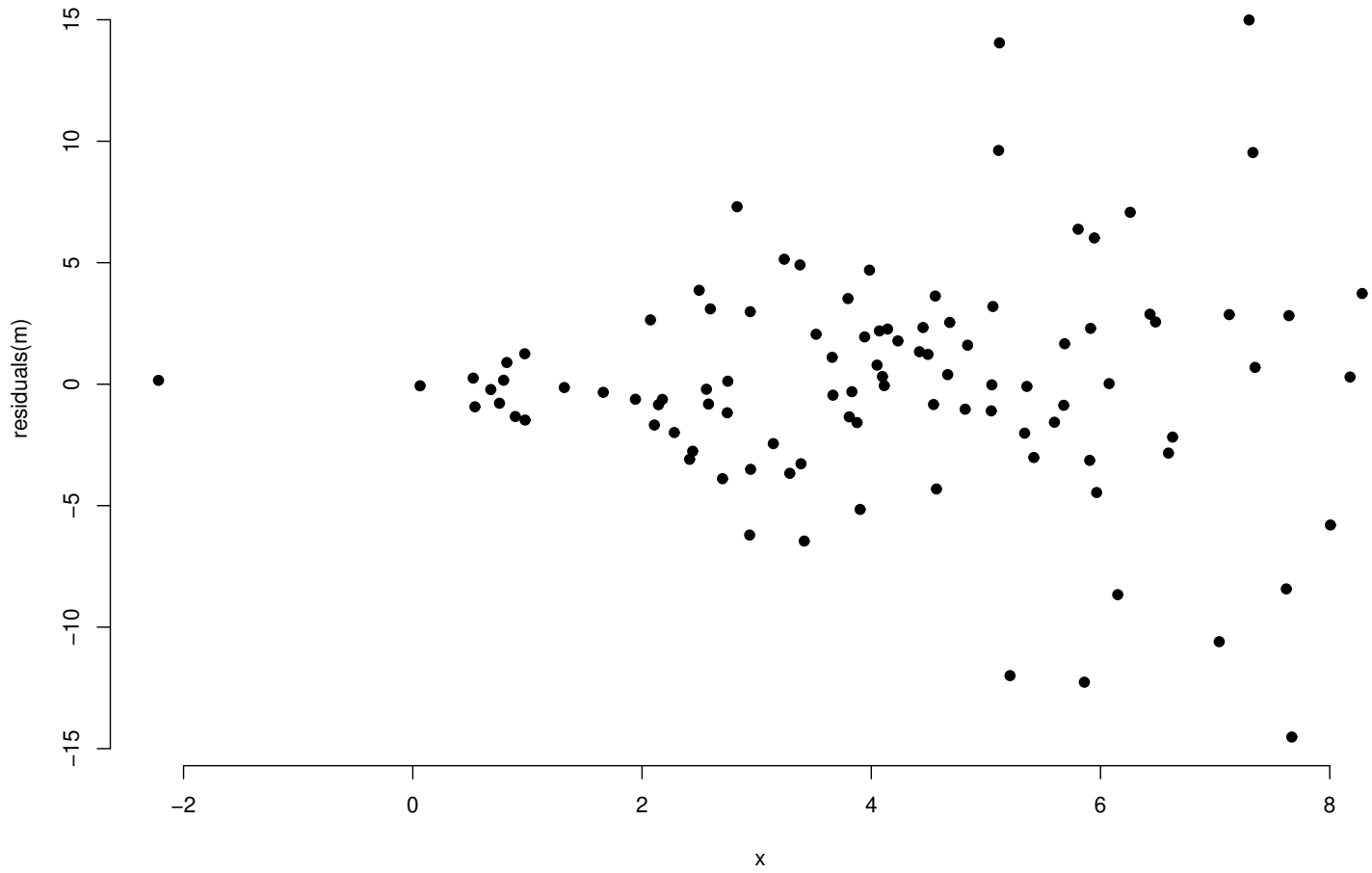
Residual plots: ε by $\hat{y}$



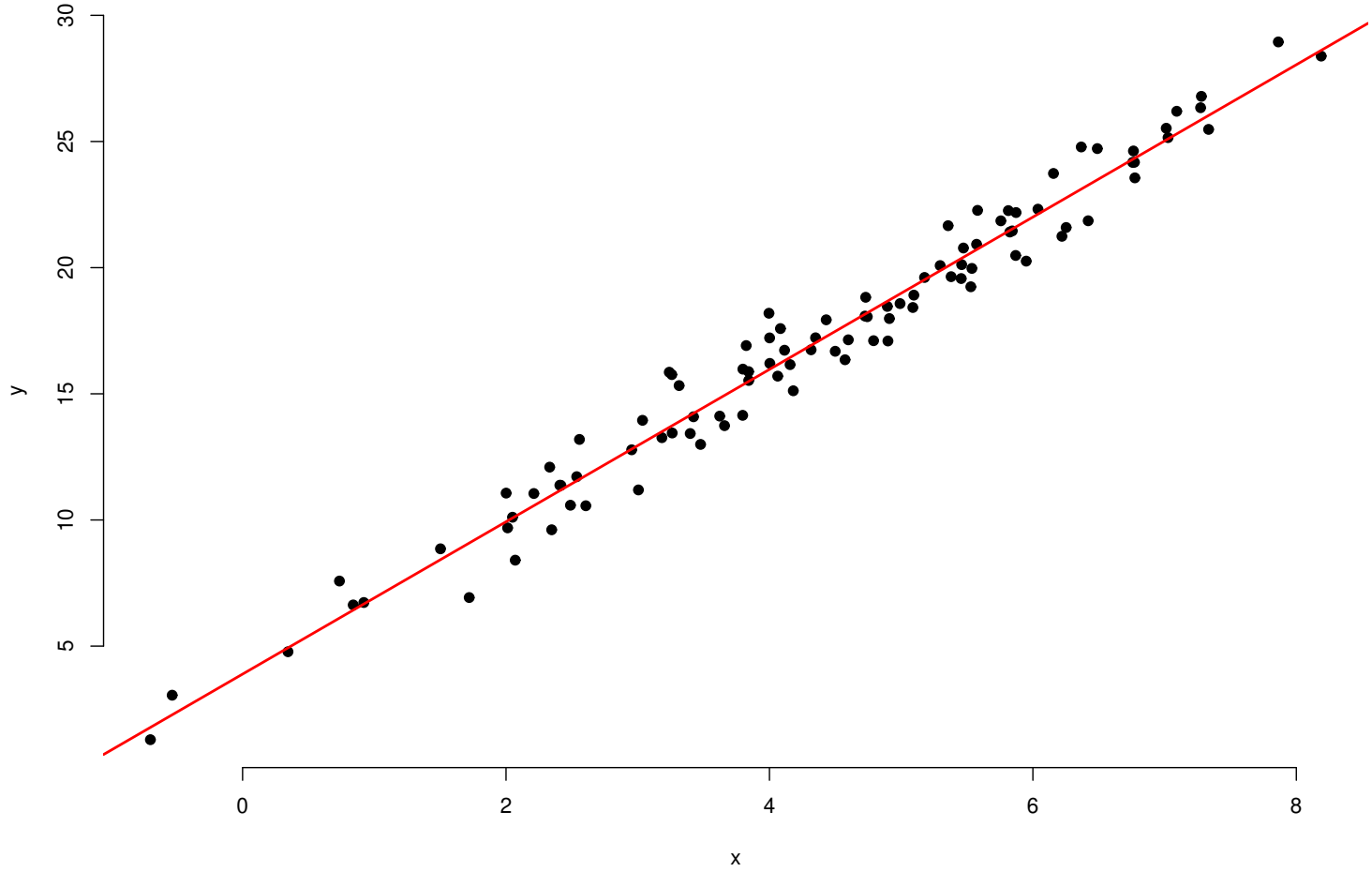
Residual plots: ε by y



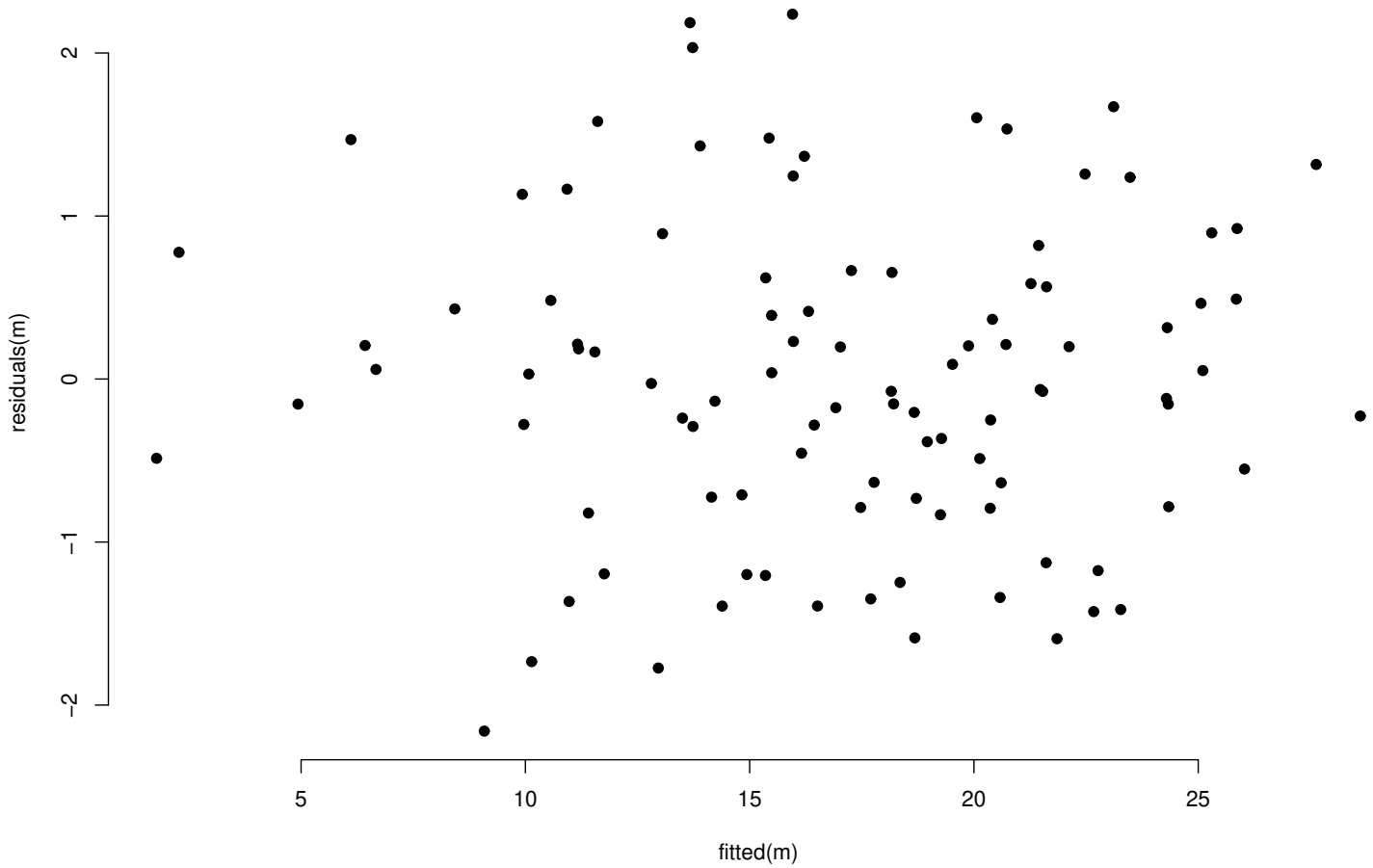
Residual plots: ε by x



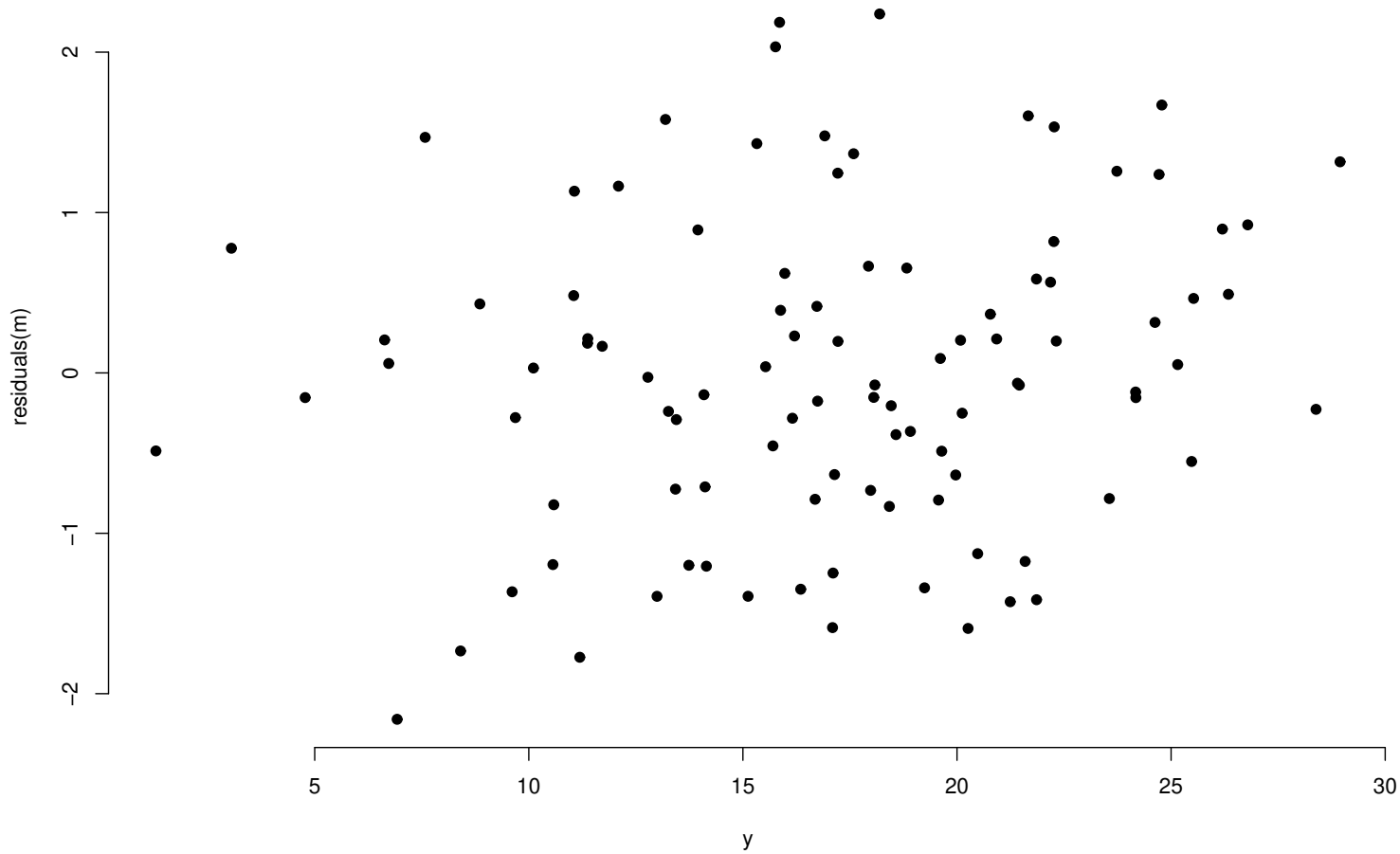
Residual plots: y by x



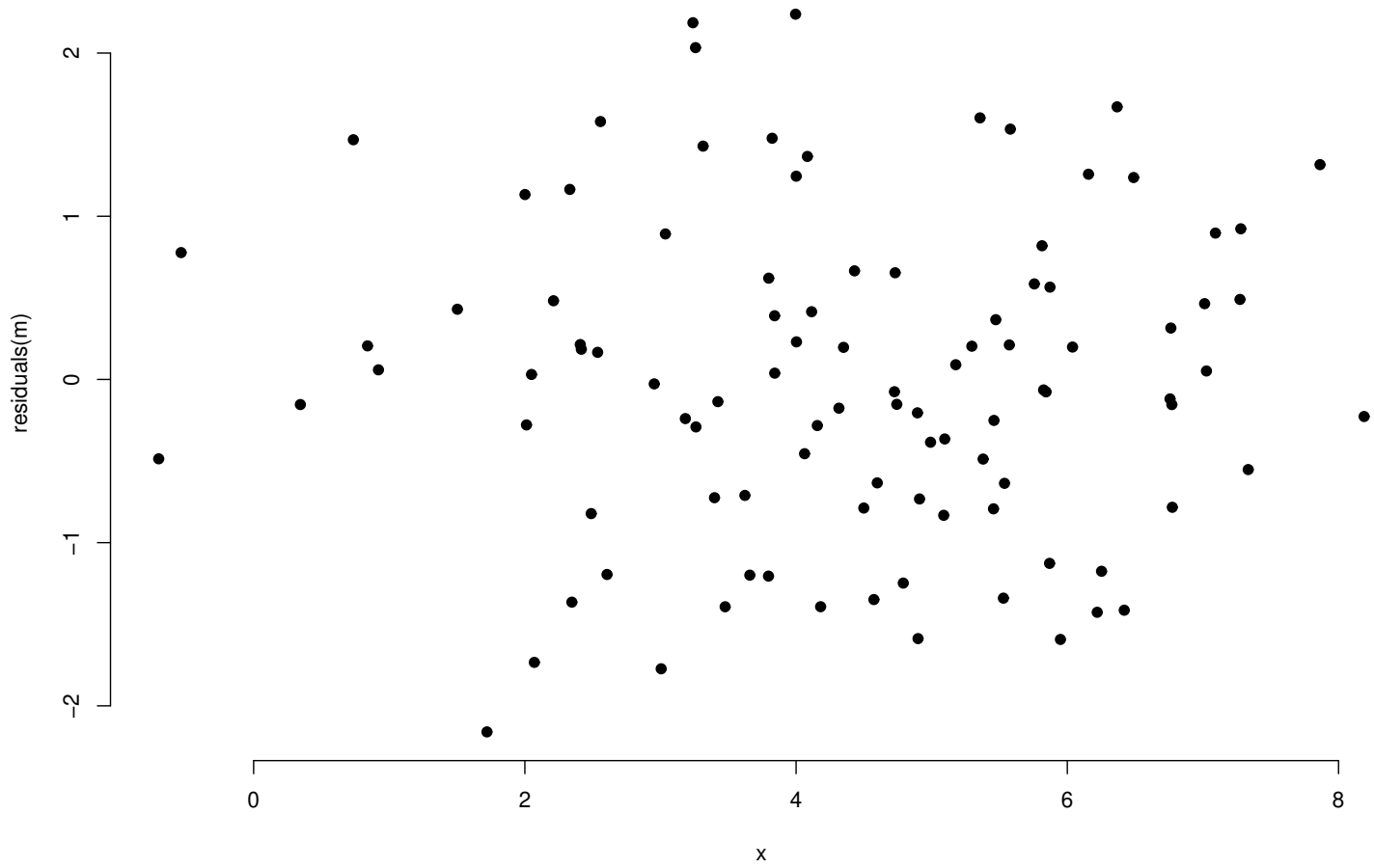
Residual plots: ε by $\hat{y}$



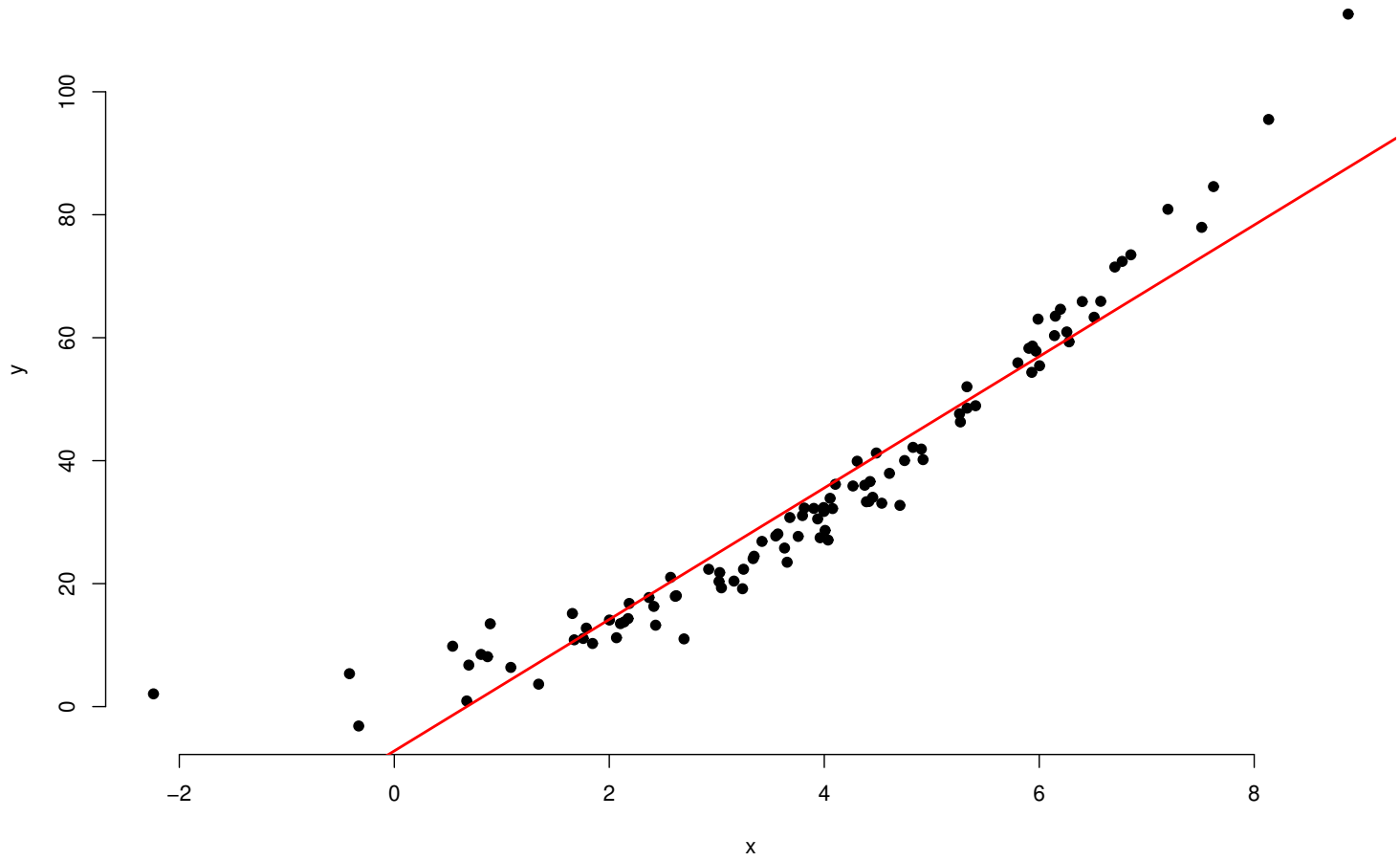
Residual plots: ε by y



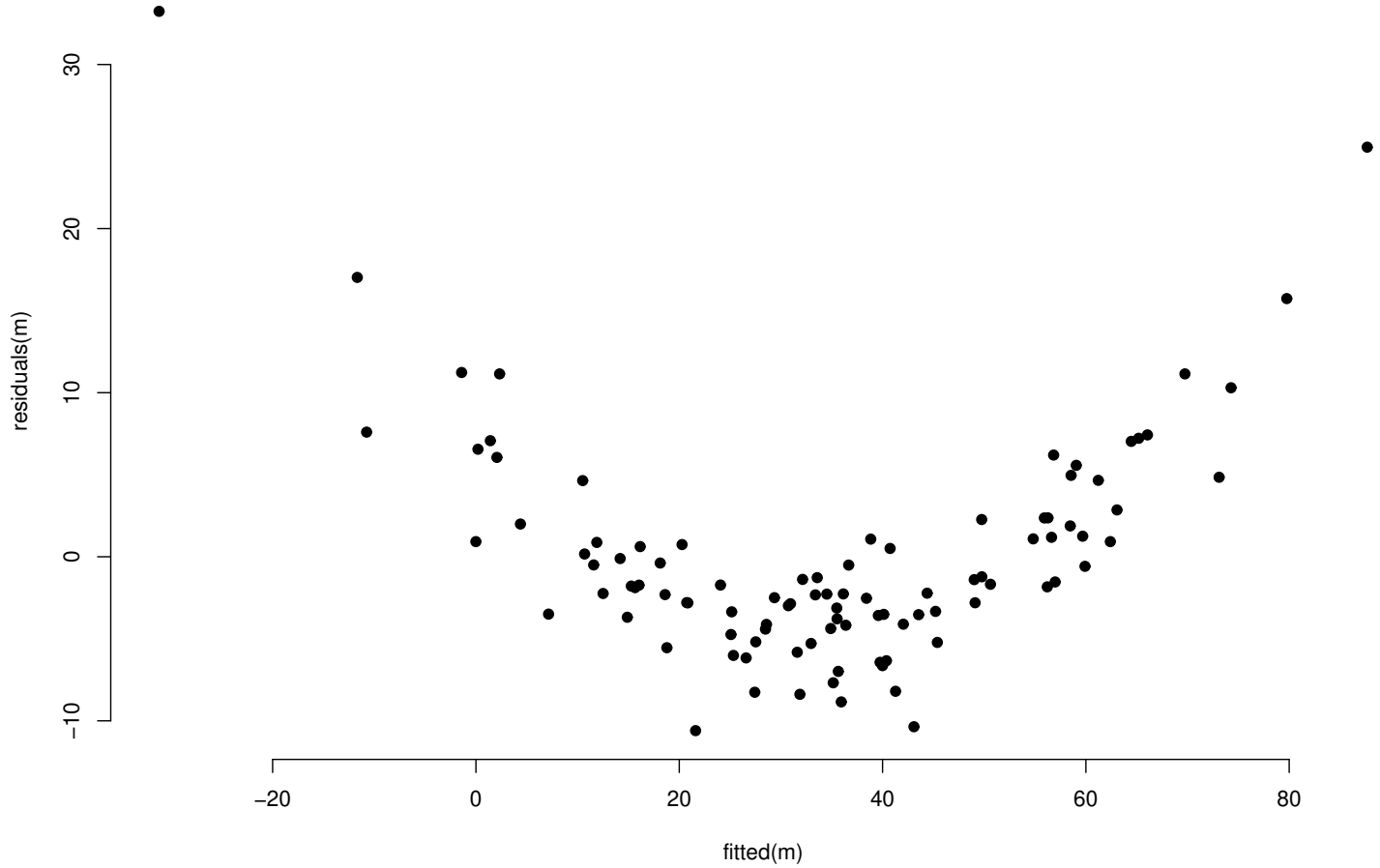
Residual plots: ε by x



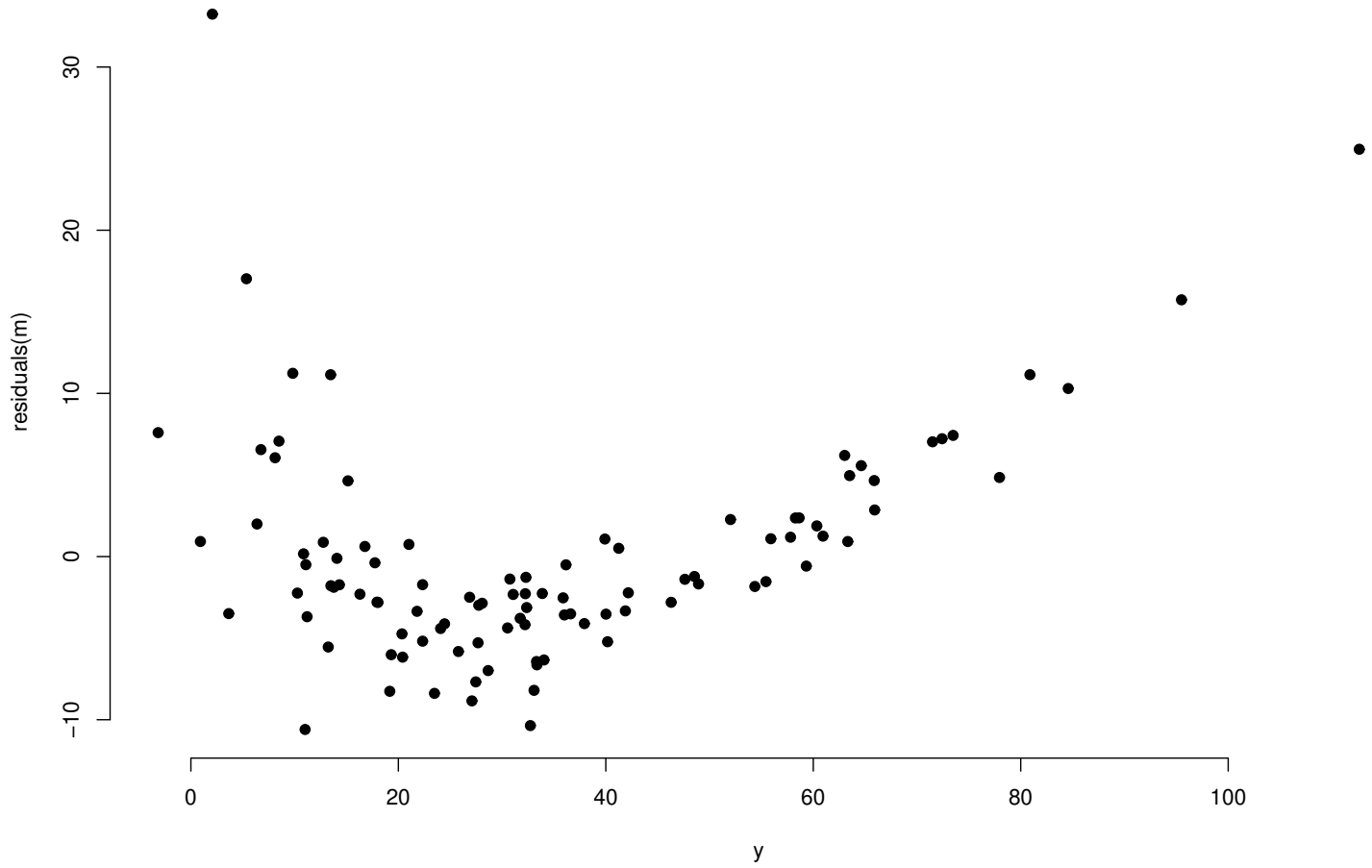
Residual plots: y by x



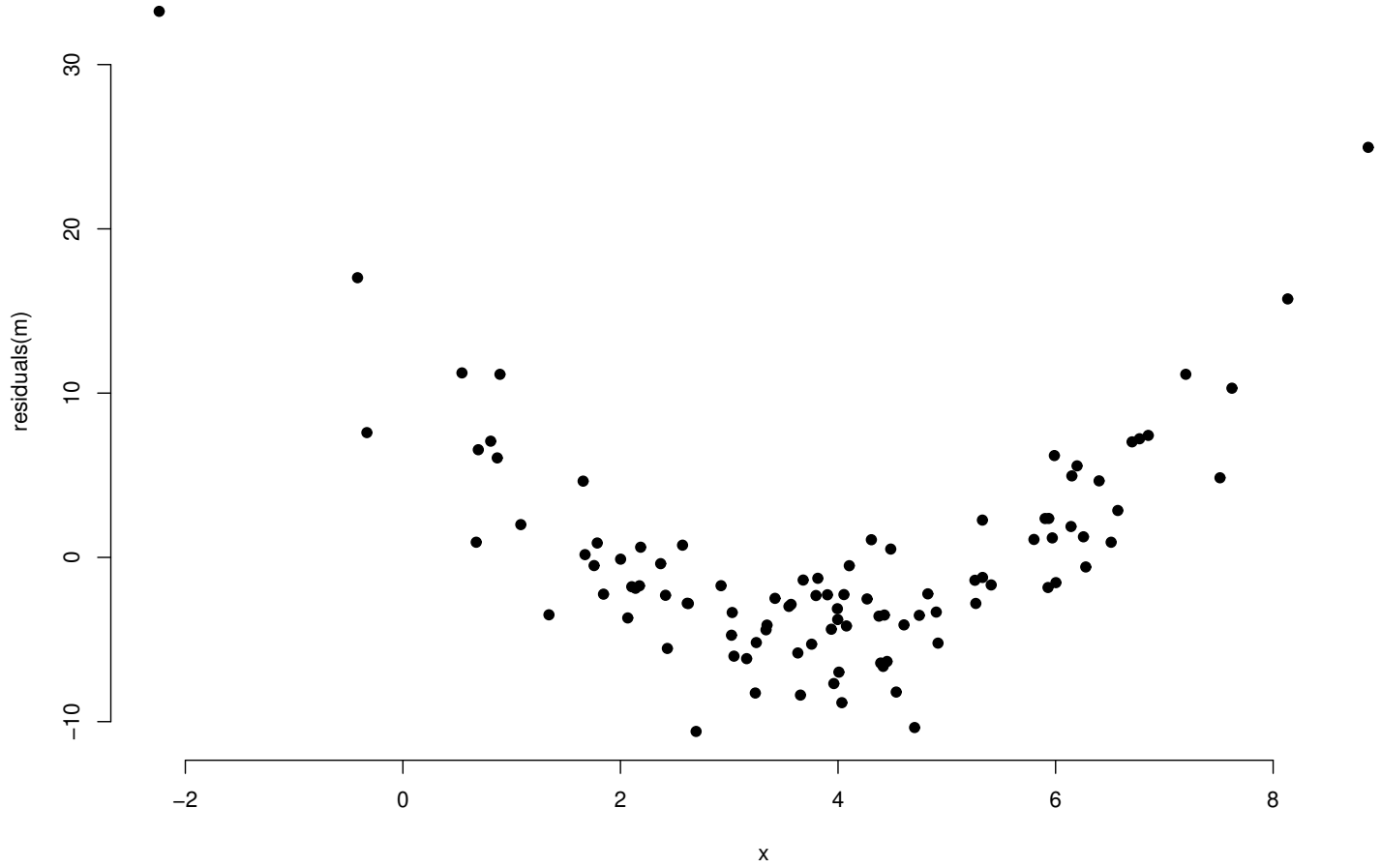
Residual plots: ε by $\hat{y}$



Residual plots: ε by y



Residual plots: ε by x



Heteroscedasticity: effect

With heteroscedastic disturbances, $\hat{\beta}$ will be unbiased but inefficient.

Heteroscedasticity: solution

First, check whether you can see a misspecification (e.g. the relationship is quadratic).

Otherwise, some specialised model, like a regression with **robust standard errors** or a **weighted least squares** model can be used.

Outline

OLS assumptions: errors

- Errors have an expected value of zero given X.
- Errors are normally distributed.
- Errors have a constant variance.
- Errors are not autocorrelated.

Autocorrelation

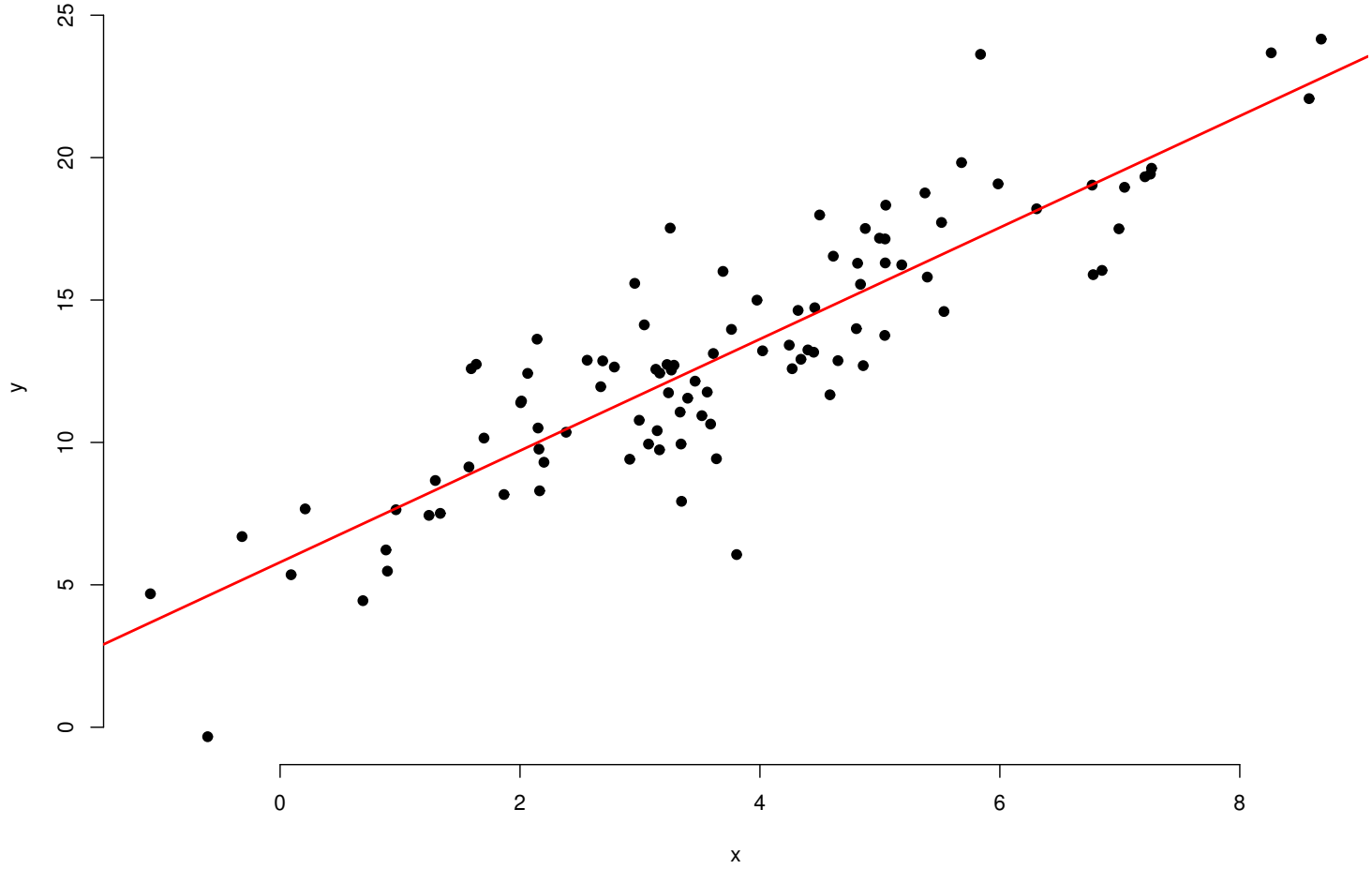
Autocorrelation refers to the fact that residuals might be correlated with eachother. This can occur for various reasons:

- Spatial autocorrelation
- Temporal autocorrelation
- Persistent shocks
- Inertia / psychological conditioning
- Partial adjustments over time

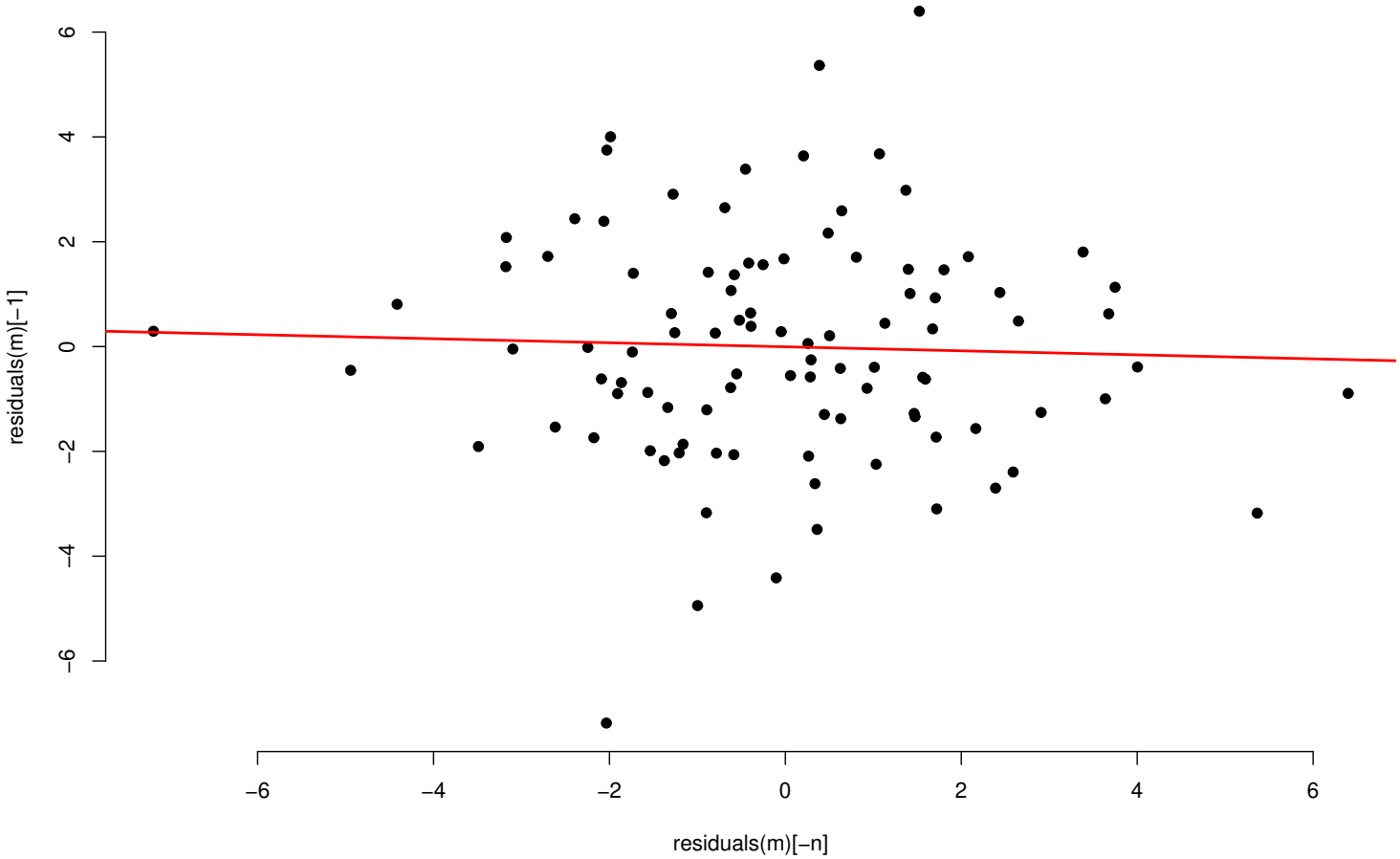
Autocorrelation: consequences

Autocorrelated residuals leads to an inflated R^2 and the estimates for $\hat{\beta}$ will be unbiased but inefficient.

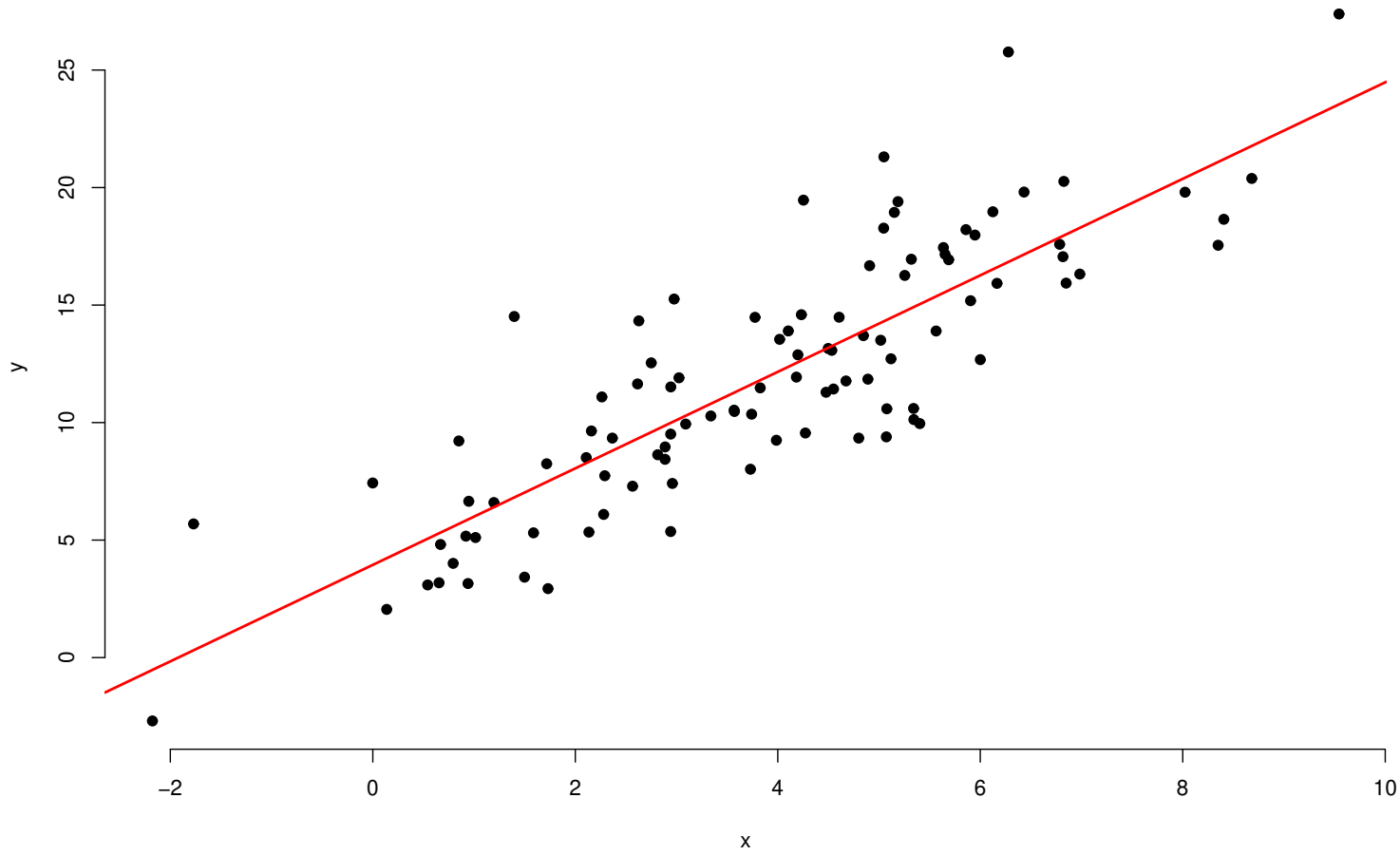
Residual plots: y by x



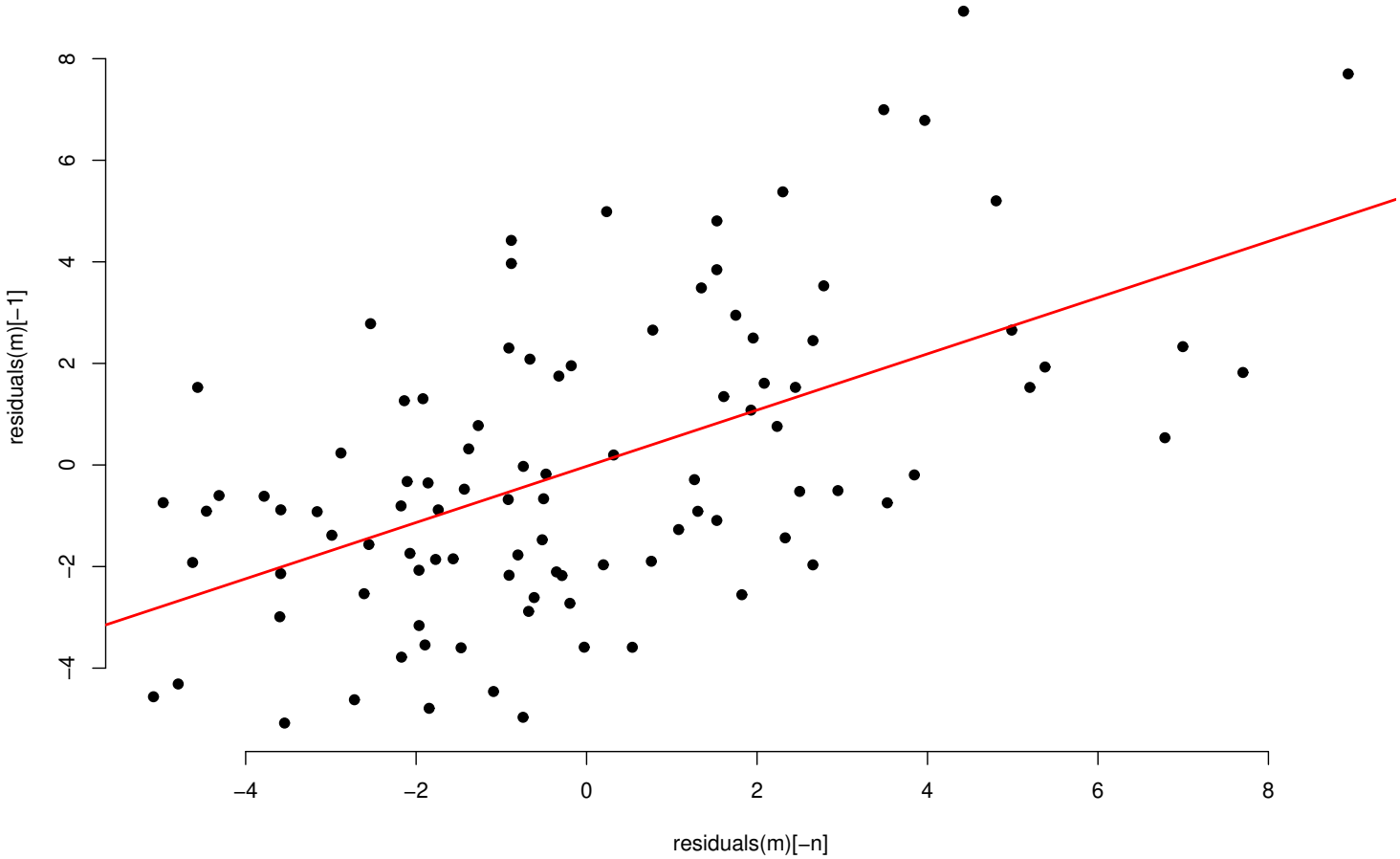
Residual plots: ε_t by ε_{t-1}



Residual plots: y by x



Residual plots: ε_t by ε_{t-1}



Autocorrelation: solution

Estimating a regression model with autocorrelation is possible, but not straightforward. Many solutions exist. The fields that are of primary concern when dealing with such data are **time-series analysis** and **spatial econometrics**.

Both are well beyond this course.

Outline

OLS assumptions: errors

- Errors have an expected value of zero given X.
- Errors are normally distributed.
- Errors have a constant variance.
- Errors are not autocorrelated.
- Errors are not correlated with X.

OLS assumptions: regressors

- X varies

OLS assumptions: regressors

- X varies
- X is of full column rank (note: requires $k < n$)

Detecting multicollinearity

To detect multicollinearity problems, you can simply regress all independent variables on all other independent variables or look at the variance inflation factors (VIFs), which are based on the R^2 of these auxilliary regressions:

$$VIF_k = \frac{1}{1 - R_k^2},$$

with R_k^2 the R^2 for the kth auxilliary regression.

Multicollinearity: solution

If you have very high multicollinearity, you might simply decide to drop one of the offending variables. It is apparently not providing a lot of additional information. If it is still providing *some* information, the results estimates will be biased.

Alternatives are to leave the model as is, or develop an index combining multiple variables.

Outline

OLS assumptions: regressors

- X varies
- X is of full column rank (note: requires $k < n$)
- No measurement error in X

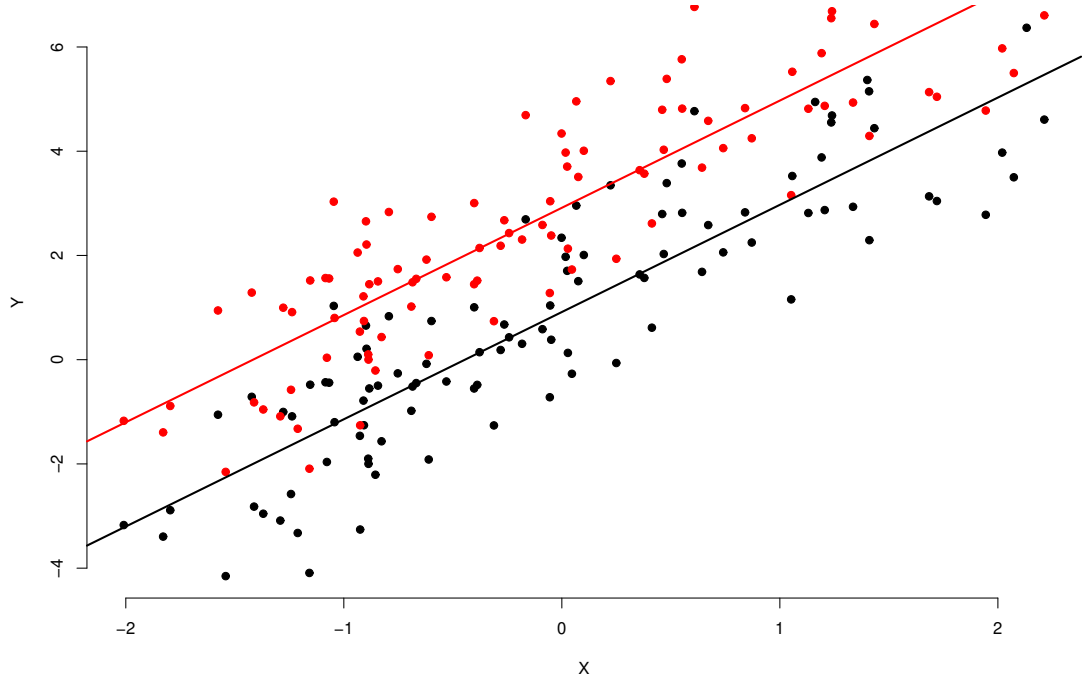
Measurement error

For measurement error, important distinctions are between **systematic** and **random** measurement error and between measurement errors in dependent and in independent variables.

The consequences for the estimation are different.

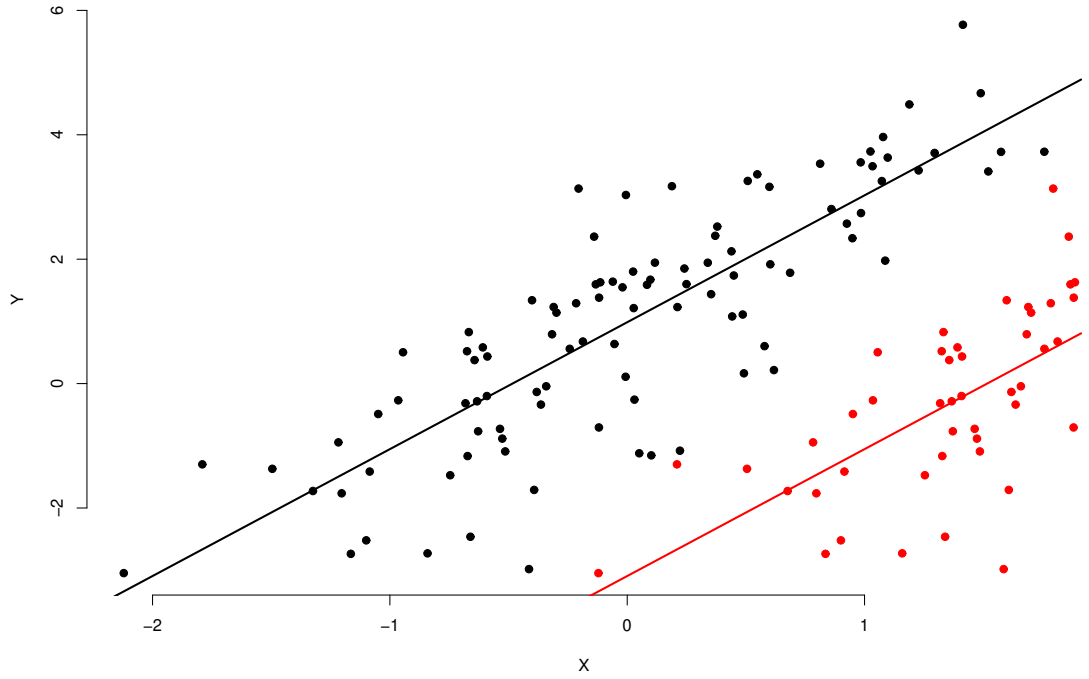
Measurement error

A systematic error in the dependent variable leads to a biased estimate of $\hat{\beta}_{intercept}$:



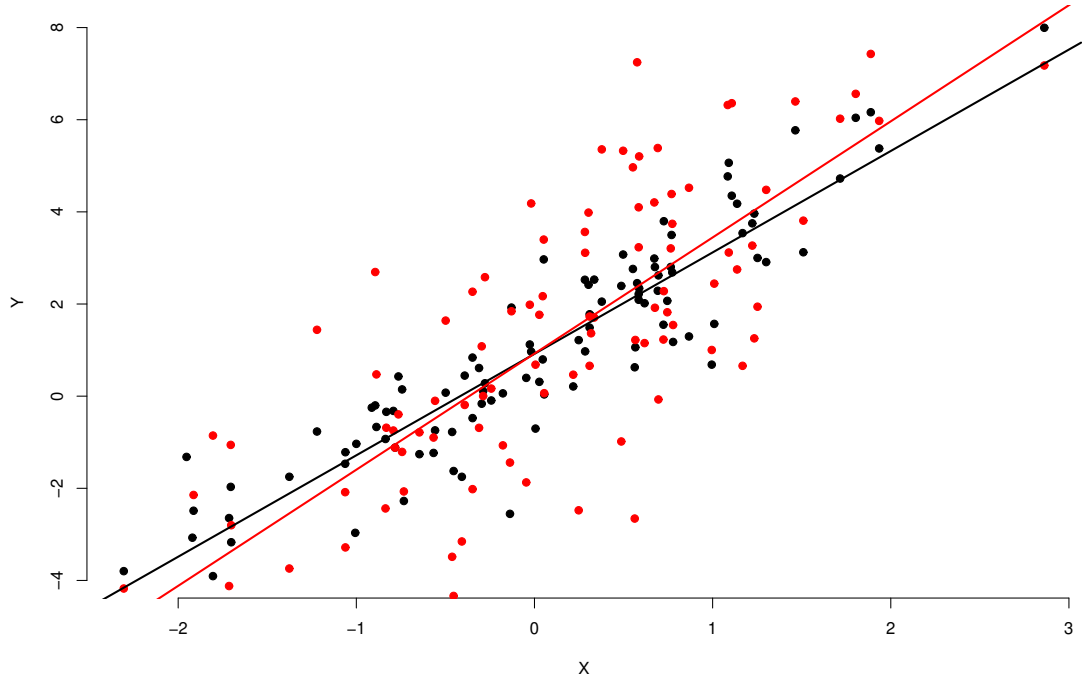
Measurement error

A systematic error in the independent variable leads to a biased estimate of $\hat{\beta}_{intercept}$:



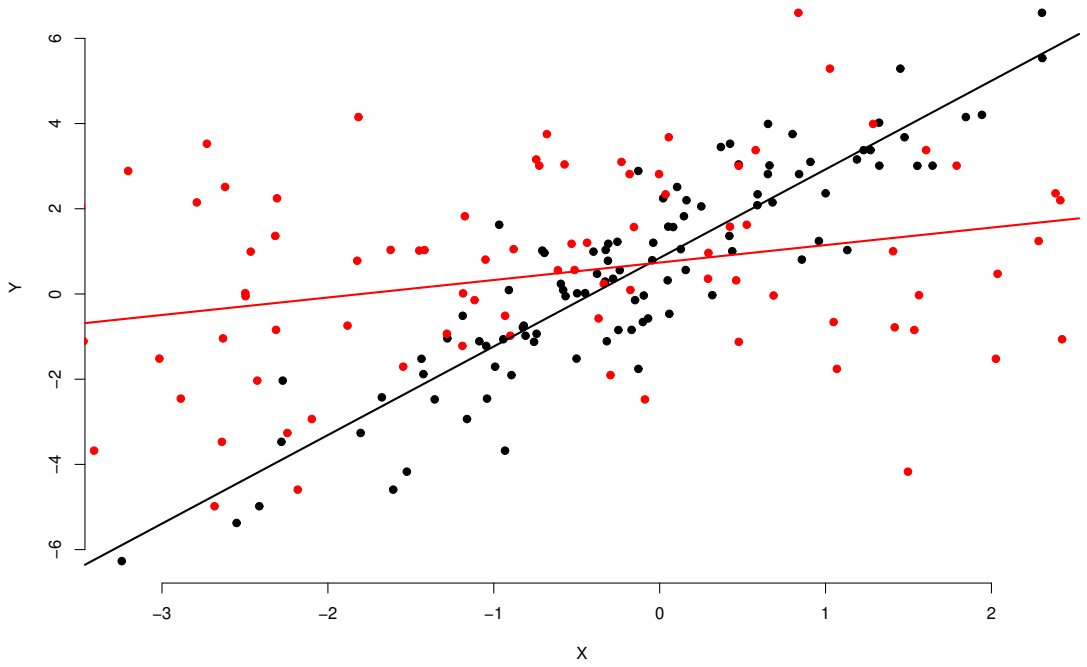
Measurement error

A random error in the dependent variable leads to a less efficient estimate of $\hat{\beta}$:



Measurement error

A random error in the independent variable leads to a biased estimate of $\hat{\beta}$:



OLS assumptions: regressors

- X varies
- X is of full column rank (note: requires $k < n$)
- No measurement error in X
- No endogenous variables in X