Advanced Quantitative Methods: Regression diagnostics

Johan A. Elkink

University College Dublin

16 February 2017

1 Specification

Outliers, leverage, influence

Multicollinearity

2 Heteroscedasticity

Solutions

Diagnostics

3 Autocorrelation

Diagnostics

10 participants	March 2017 Mon 6		Tue 7		Fri 10	Tue 14	
	10:00 – 12:00	14:00 – 16:00	09:00 – 11:00	14:00 – 16:00	09:00 – 11:00	10:00 – 12:00	13:00 – 15:00
TIANYANG SONG	✓	✓			✓	✓	✓
Giulia Santomauro	✓		✓		✓		✓
Kevin Lacourse	✓	✓		✓			
Marco Schito	✓	✓	✓		✓	✓	✓
Tyler McDonough		✓	✓	✓	✓		✓
Yulu Guo			✓	✓	✓	✓	✓
Andrea Salvi			✓	✓	✓	(✓)	(✓)
Xiaodong Li	✓	✓		✓			
Samuel Almeida	✓	✓	✓	✓	✓	✓	✓
Eleonora La Spada			✓	✓	✓	✓	✓
Jos Elkink	Yes (Yes) No	Yes (Yes) No	Yes (Yes) No	Yes (Yes) No	Yes (Yes) No	Yes (Yes) No	Yes (Yes) No
	10:00 – 12:00	14:00 – 16:00	09:00 – 11:00	14:00 – 16:00	09:00 – 11:00	10:00 – 12:00	13:00 – 15:00

	Mon 6		Tue 7		Fri 10	Tue 14	
	March 2017						
Yes	6	6	7	7	**8**	5	7
Ifneedbe	0	0	0	0	**0**	1	1
No	4	4	3	3	**2**	4	2

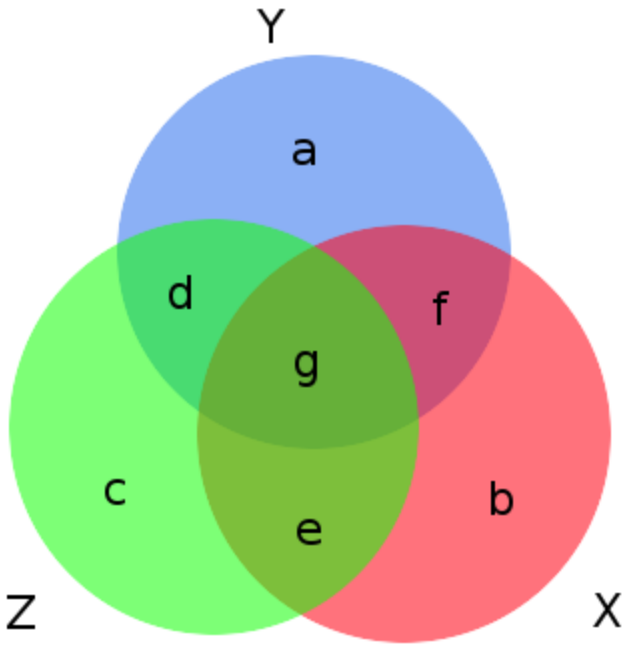
Outline

Omitting a relevant independent variable

For omitted variable $\mathbf{z}$:

- $\hat{\beta}^{OLS}$ will be biased iff $cor(\mathbf{z}, \mathbf{X}) \neq 0$.
- intercept will be biased iff $E(\mathbf{z}) \neq 0$.
- $\hat{\sigma}^2$ will be biased upward
- $\Longrightarrow V(\hat{\beta}^{OLS})$ will be biased upward

Omitting a relevant variable z: graphical intuition



If **z** omitted, areas g and f reflect information used to estimate $\hat{\beta}_{\mathbf{X}}^{OLS}$

If **z** included, only area f would be used to estimate $\hat{\beta}_{\mathbf{X}}^{OLS}$

Only area a is used to estimate $\hat{\sigma}^2$, except when **z** excluded, and then area d is also used

If **X** is orthogonal to **z**, then no area g and bias disappears

Omitted variables

If a variable is omitted which is correlated with one of the independent variables included, then the effect of this omitted variable will be absorbed by the error term, which will thus correlate with the independent variable.

Model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$$

Estimated model:

$$y_i = \beta_0^* + \beta_1^* x_i + u_i$$
$$u_i = \beta_2 z_i + \varepsilon_i$$

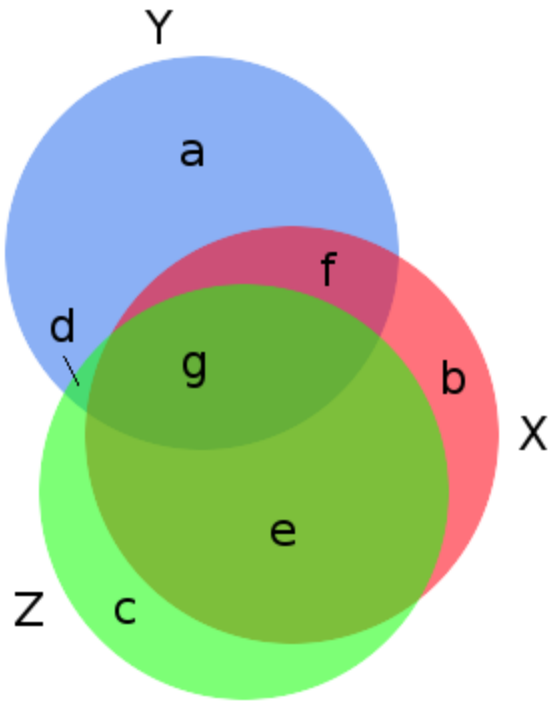
If $cor(x_i, z_i) \neq 0$, x_i and u_i will be correlated.

Including an irrelevant independent variable

For unnecessarily included variable **z**:

- $\hat{\beta}_{\mathbf{X}}^{OLS}$ and $V(\hat{\beta}_{\mathbf{X}}^{OLS})$ will remain unbiased
- $\hat{\beta}^{OLS}$ will be less efficient (increases MSE)

Adding an irrelevant variable: graphical intuition



Area f refects variation in **y** due entirely to **X**, so $\hat{\beta}_{\mathbf{X}}^{OLS}$ unbiased

Since area f $<$ area (f + g), $V(\hat{\beta}_{\mathbf{X}}^{OLS})$ increases

Area a used to estimate $\hat{\sigma}^2$ unbiased so $V(\hat{\beta}^{OLS})$ remains unbiased

If **z** is orthogonal to **X** then no area g and then no efficiency loss

Non-linearity

If there is non-linearity in the variables, but not in the parameters, there is no problem. E.g.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

can be estimated with OLS.

If there are other non-linearities, sometimes the equation can be transformed. E.g.

$$\begin{aligned} y_i &= \beta_0 x_i^{\beta_1} \varepsilon_i \\ \log(y_i) &= \log(\beta_0) + \beta_1 \log(x_i) + \log(\varepsilon_i) \\ y_i^* &= \beta_0^* + \beta_1 x_i^* + \varepsilon_i^* \end{aligned}$$

Functional forms for additional non-linear transformations

log-linear as with the previous example

semi-log has two forms:

$$y_i = \beta_0 + \beta_1 \log(x_i),$$

where β_1 is Δy due to $\%\Delta x$

$$\log(y_i) = \beta_0 + \beta_1 x_i,$$

where β_1 is $\%\Delta y$ due to Δx

inverse or reciprocal: $y_i = \beta_0 + \beta_1 \frac{1}{x_i}$

polynomial $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$

Testing for non-linearity

"As there are many possible forms of nonlinearity it is likely that no one test will be powerful against them all, so several tests may be needed."

(Teräsvirta, Tjøstheim and Granger, 1992)

R^2

Defined in terms of sums of squares:

$$\begin{aligned} R^2 &= \frac{SSE}{SST} \\ &= 1 - \frac{SSR}{SST} \\ &= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \end{aligned}$$

Interpretation: the proportion of the variation in **y** that is explained linearly by the independent variables.

A much over-used statistic: it may not be what we are interested in at all.

Adjusted R^2

One of the problems with looking at R^2 is that the more independent variables, the higher R^2, which discourages parsimony. One solution for this the **adjusted** R^2:

$$adjR^2 = 1 - \frac{n-1}{n-k}(1 - R^2)$$

So this R^2 has a penalty for having many parameters (high k).

Akaike Information Criterion (AIC)

Another approach to making a similar balance between parsimony and explained variance is **Akaike Information Criterion**:

$$AIC = \log\left(\frac{\mathbf{e}'\mathbf{e}}{n}\right) + \frac{2k}{n}$$

Thus the smaller AIC, the better.

AIC variations

$$AIC = \log\left(\frac{\mathbf{e}'\mathbf{e}}{n}\right) + \frac{2k}{n}$$

$$BIC = \log\left(\frac{\mathbf{e}'\mathbf{e}}{n}\right) + \frac{(\log n)k}{n}$$

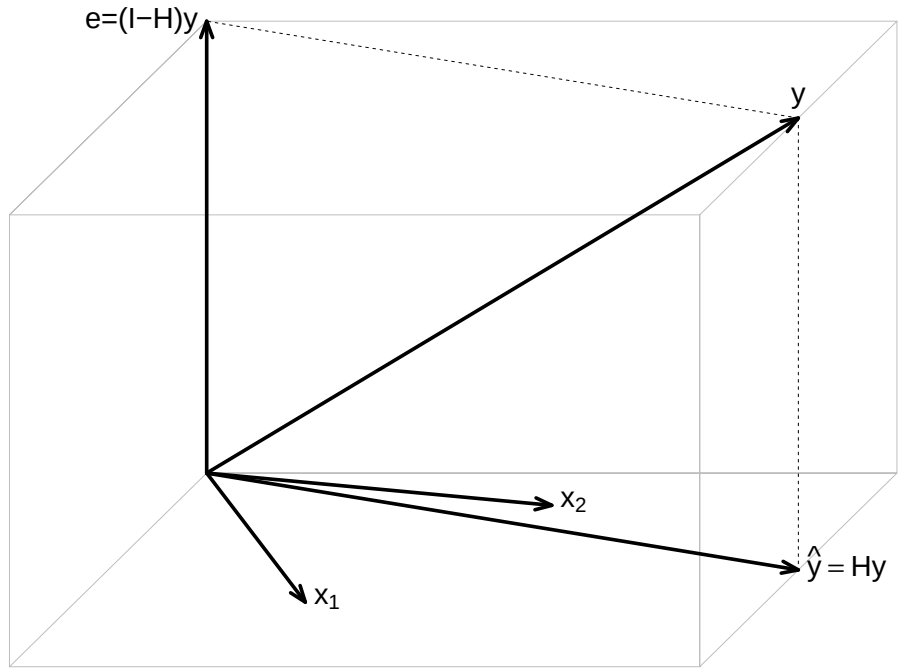
$$AICc = AIC + \frac{2k(k+1)}{n-k-1}$$

and there are many more similar variations.

Outline

Hat matrix

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{X}\hat{\beta}^{OLS} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \mathbf{H}\mathbf{y}\end{aligned}$$



$$var(\hat{\mathbf{y}}) = \sigma^2\mathbf{H}$$
$$var(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$$

H is called the **hat matrix** (it "puts a hat on **y**"), or sometimes **prediction matrix P**.

Leverage

The elements on the diagonal of **H** are called the **leverage** of each case — the higher the leverage, the more this particular case contributed to the predicted dependent variable.

For the remainder we will use:

$$h_i = \mathbf{H}_{ii} = \mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i',$$

thus h_i represents the leverage of observation i ($\mathbf{x}_i$ is a row vector of the independent variables for case i).

Note that $0 \leq h_i \leq 1$ and $\sum_{i=1}^{n} h_i = k$.

A high h_i means that $\mathbf{x}_i$ is far from the mean of $\mathbf{X}$.

Outliers

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

To account for the potential variables in the sampling variances of the residuals, we calculate **externally studentized residuals** (or studentized deleted residuals), where a large absolute value indicates an outlier. A test could be based on the fact that in a model without outliers, they should follow a $t(n-k)$ distribution.

(Kutner et al., 2005, 390–398)

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.

Studentized residuals

- The **internally studentized residual** is:

$$r_i = \frac{e_i}{\sqrt{s^2(1-h_i)}}$$

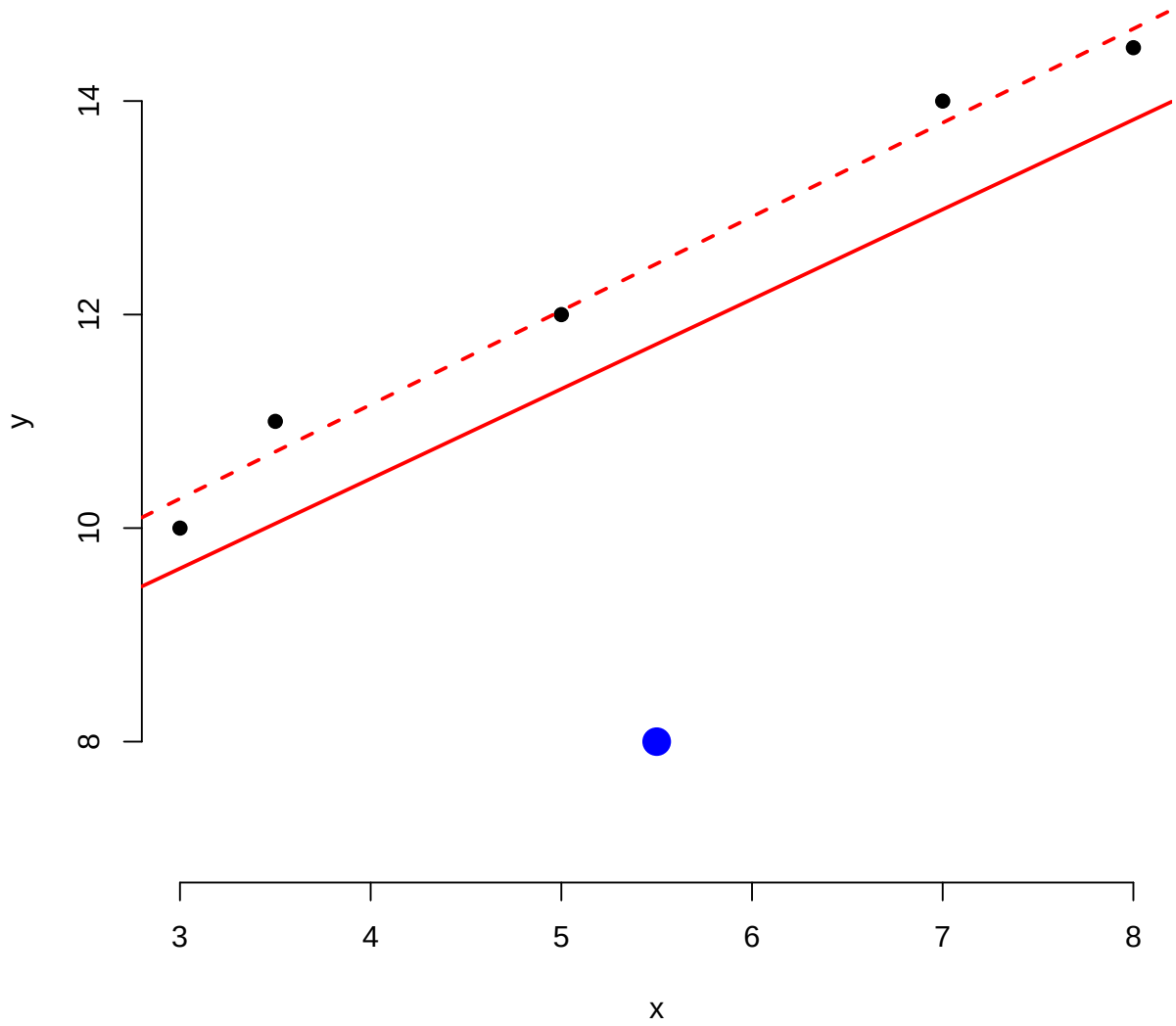
- The **deleted residual** is $d_i = y_i - \hat{y}_{i(-i)}$, whereby $\hat{y}_{i(-i)}$ is the predicted value of y_i based on a regression with the ith observation omitted.
- The **externally studentized residual** is:

$$t_i = \frac{d_i}{\sqrt{s_d^2(1-h_{i(-i)})}} = \frac{e_i}{\sqrt{s_{(-i)}^2(1-h_i)}}$$

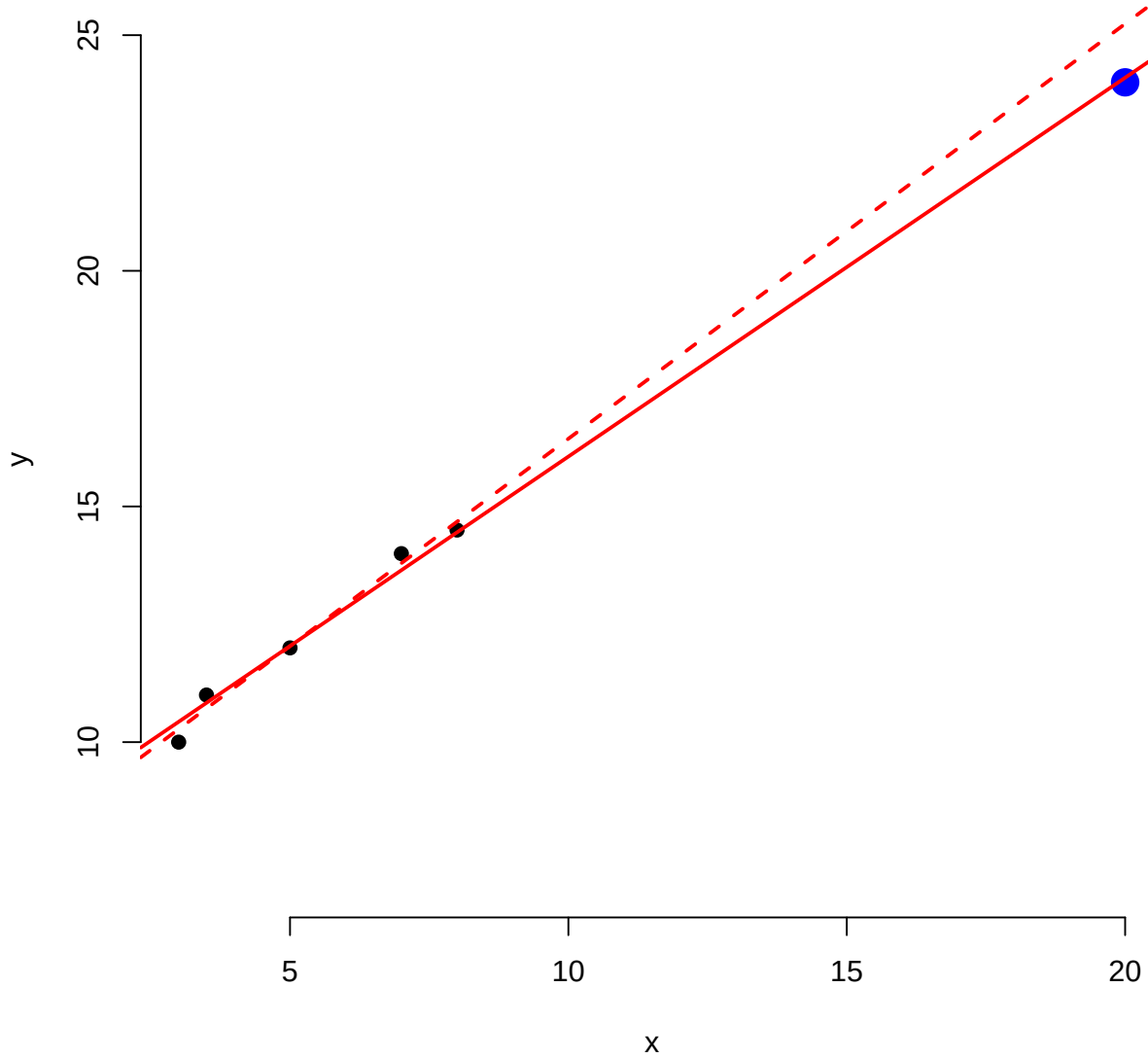
$$= e_i \sqrt{\frac{n-k-1}{s^2(1-h_i)-e_i^2}},$$

with $s_{(-i)}^2$ representing s^2 for the model without observation i.

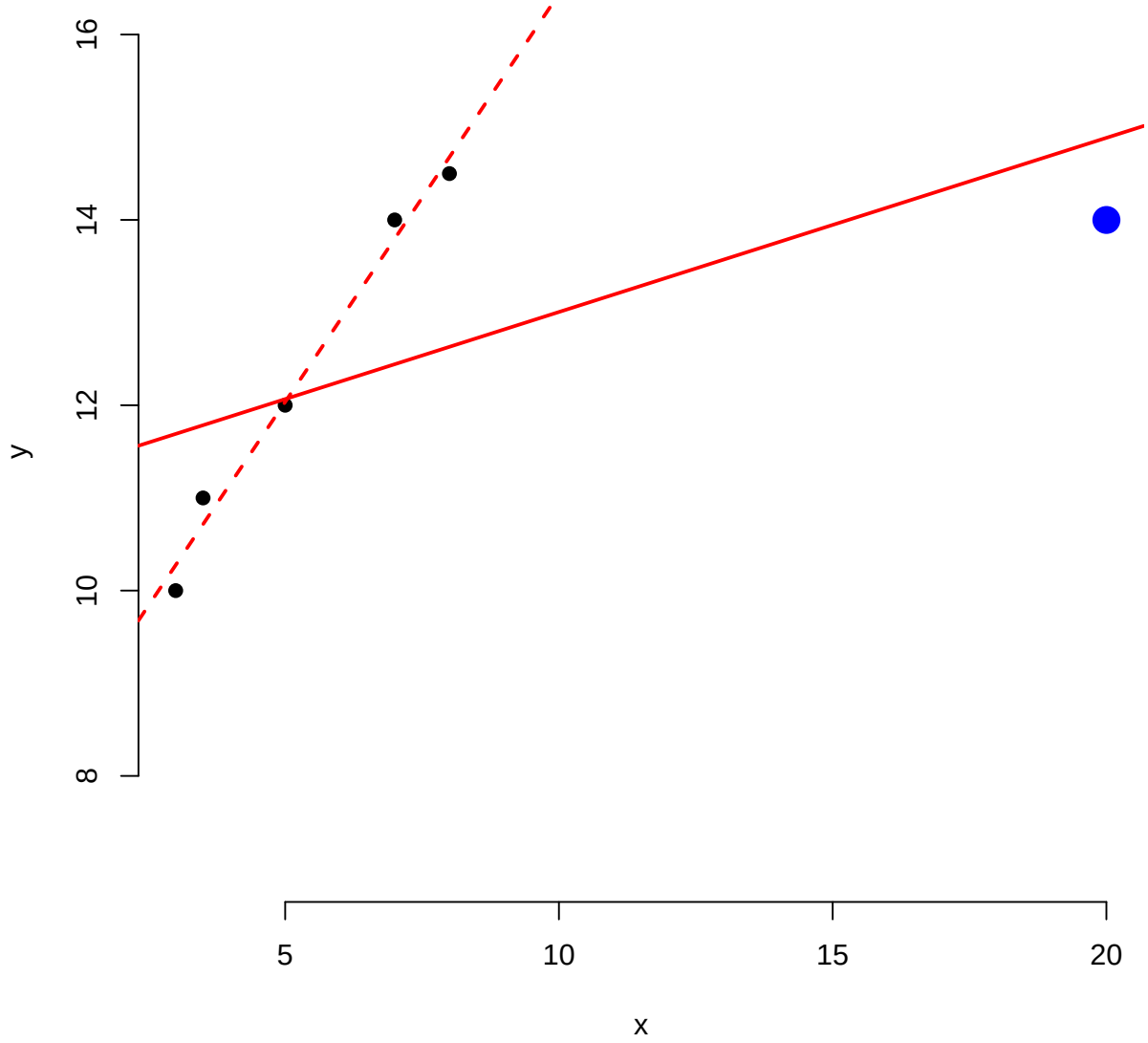
Outlier, low leverage, low influence



High leverage, low influence



High leverage, high influence



Cook's Distance

$$
\begin{aligned}
D_i &\equiv \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(-i)})^2}{ks^2} = \frac{(\hat{\beta}^{OLS}_{(-i)} - \hat{\beta}^{OLS})'\mathbf{X}'\mathbf{X}(\hat{\beta}^{OLS}_{(-i)} - \hat{\beta}^{OLS})}{ks^2} \\
&= \left(\frac{e_i}{s\sqrt{1-h_i}}\right)^2 \frac{h_i}{k(1-h_i)} \\
&= \frac{t_i^2}{k}\frac{var(\hat{y}_i)}{var(e_i)} \\
&\sim F(k, n-k)
\end{aligned}
$$

The F-test here refers to whether $\hat{\beta}^{OLS}$ would be significantly different if observation i were to be removed ($H_0 : \beta = \beta_{(-i)}$) (Cook, 1979, 169).

Cook's Distance

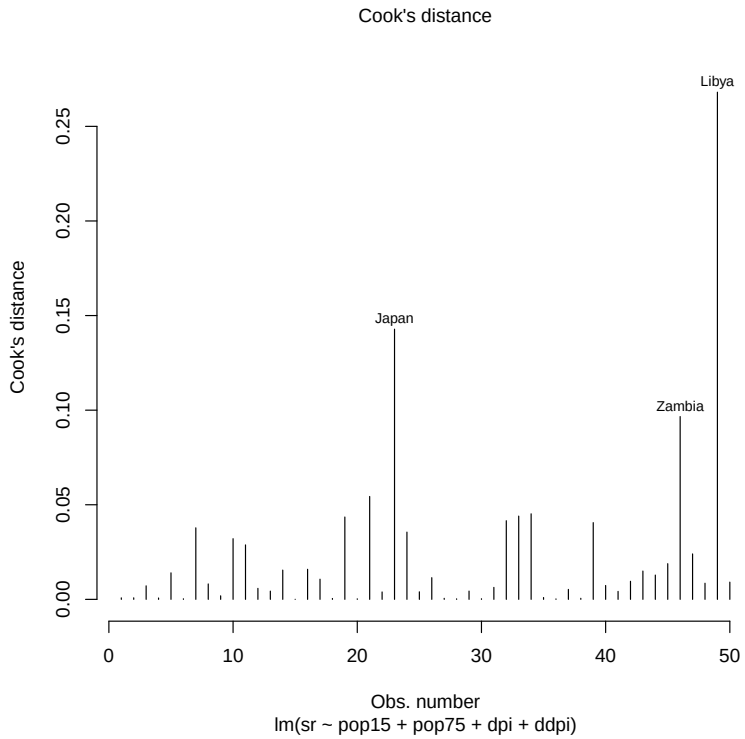
$$D_i = \frac{t_i^2}{k} \frac{var(\hat{y}_i)}{var(e_i)}$$

"t_i^2 is a measure of the degree to which the ith observation can be considered as an outlier from the assumed model."

"The ratios $\frac{var(\hat{y}_i)}{var(e_i)}$ measure the relative sensitivity of the estimate, $\hat{\beta}^{OLS}$, to potential outlying values at each data point."

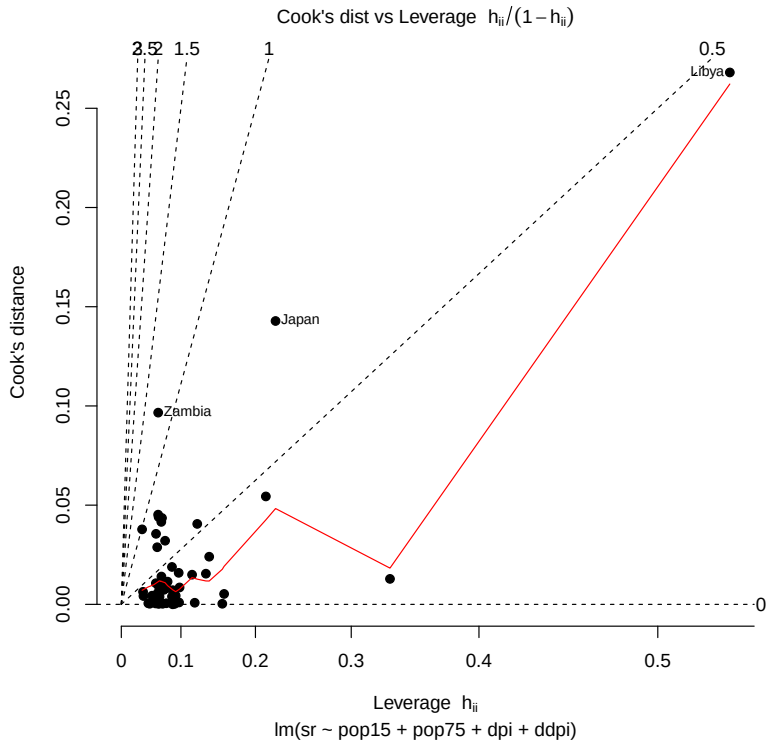
(Cook, 1977, 16)

Cook's Distance plot




```
plot(m, which = 4,
     pch = 19,
     bty = "n")
```

Cook's Distance vs leverage




```
plot(m, which = 6,
     pch = 19,
     bty = "n")
```

What to do with outliers?

Options:

1. Ignore the problem
2. Investigate *why* the data are outliers — what makes them unusual?
3. Consider respecifying the model, either by tranforming a variable or by including an additional variable (but beware of **overfitting**)
4. Consider a variant of "robust regression" that downweights outliers

Diagnosing problems in R

- A very easy set of diagnostic plots can be accessed by plotting a `lm` object, using `plot.lm()`
- This produces, in order:
  1. residuals against fitted values
  2. Normal Q-Q plot
  3. scale-location plot of $\sqrt{|e_i|}$ against fitted values
  4. Cook's distances versus row labels
  5. residuals against leverages
  6. Cook's distances against leverage/(1-leverage)
- Note that by default, `plot.lm()` only gives you 1,2,3,5

Outline

Collinearity

When some variables are linear combinations of others then we have exact (or perfect) collinearity, and there is no unique least squares estimate of $\boldsymbol{\beta}$. $(\mathbf{X}'\mathbf{X})^{-1}$ will not exist if $r(\mathbf{X}) < k$.

When $\mathbf{X}$ variables are highly correlated, we have **multicollinearity**.

Detecting multicollinearity:

- look at correlation matrix of predictors for *pairwise* correlations
- regress $\mathbf{x}_j$ on $\mathbf{X}_{(-j)}$ to produce R_j^2, and look for high values (close to 1.0)
- examine eigenvalues of $\mathbf{X}'\mathbf{X}$

Multicollinearity

The extent to which multicollinearity is a problem is debatable.

The issue is comparable to that of sample size: if n is too small, we have difficulty picking up effects even if they really exist; the same holds for variables that are highly multicollinear, making it difficult to separate their effects on $\mathbf{y}$.

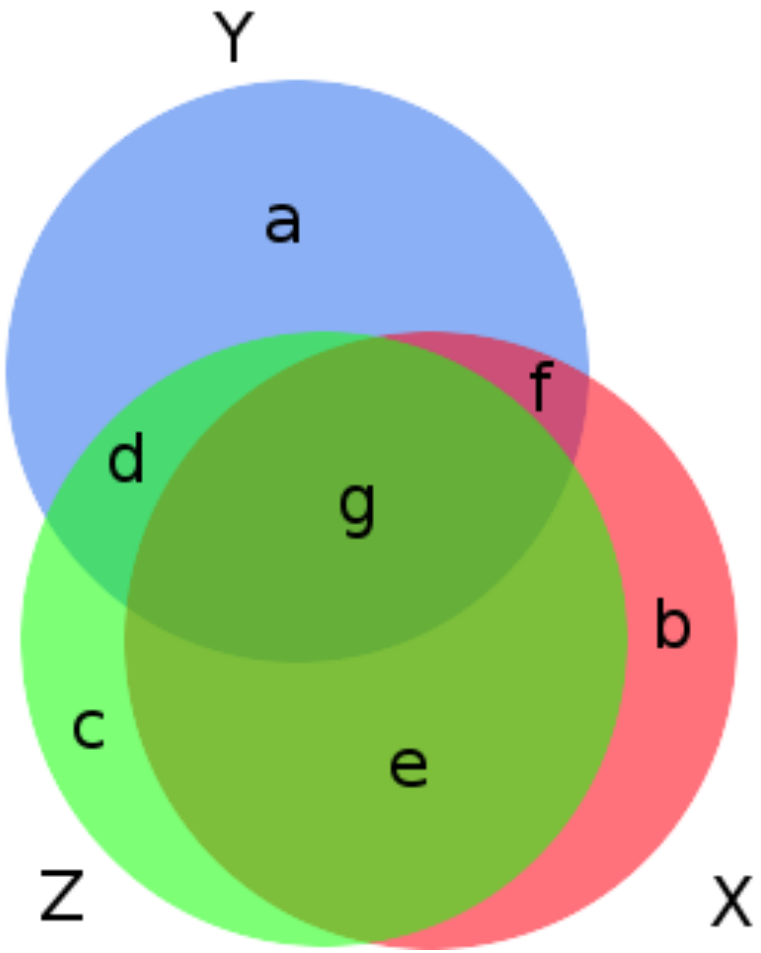
Multicollinearity

However, some problems with high multicollinearity:

- Small changes in data can lead to large changes in estimates
- High standard errors but joint significance
- Coefficients may have "wrong" sign or implausible magnitudes

(Greene, 2002, 57)

Multicollinearity



Variance of $\hat{\beta}^{OLS}$

$$var(\hat{\beta}_k^{OLS}) = \frac{\sigma^2}{(1 - R_k^2)\sum_i^n (x_{ik} - \bar{x}_k)^2}$$

- σ^2: all else equal, the better the fit, the lower the variance
- $(1 - R_k^2)$: all else equal, the lower the R^2 from regressing the kth independent variable on all other independent variables, the lower the variance
- $\sum_i^n (x_{ik} - \bar{x}_k)^2$: all else equal, the more variation in x, the lower the variance

(Greene, 2002, 57)

Variance Inflation Factor

$$var(\hat{\beta}_k^{OLS}) = \frac{\sigma^2}{(1 - R_k^2)\sum_i^n (x_{ik} - \bar{x}_k)^2}$$

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

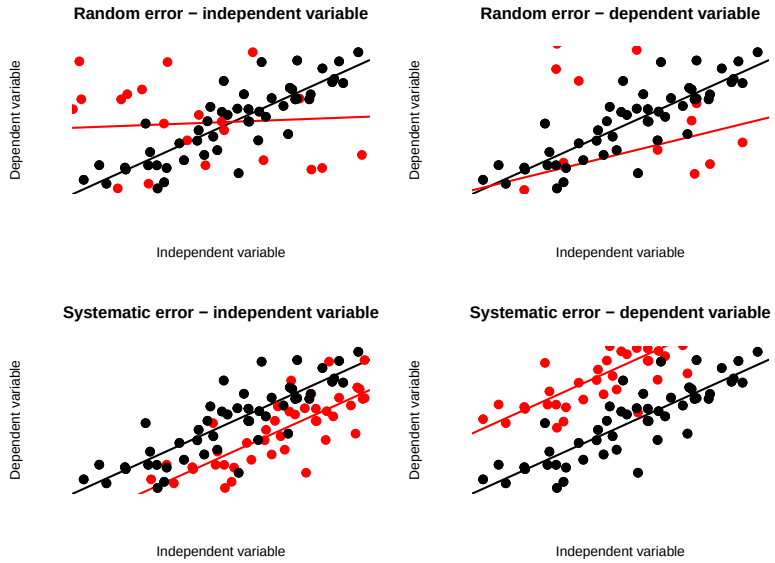
thus VIF_k shows the increase in the $var(\hat{\beta}_k^{OLS})$ due to the variable being collinear with other independent variables.

```
library(faraway)
vif(lm(...))
```

Multicollinearity: solutions

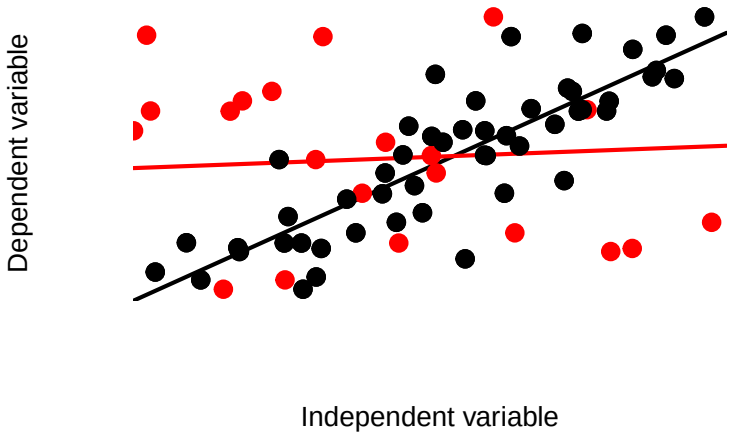
- Check for coding or logical mistakes (esp. in cases of perfect multicollinearity)
- Increase n
- Remove one of the collinear variables (apparently not adding much)
- Combine multiple variables in indices or underlying dimensions
- Formalise the relationship

Effects of measurement error

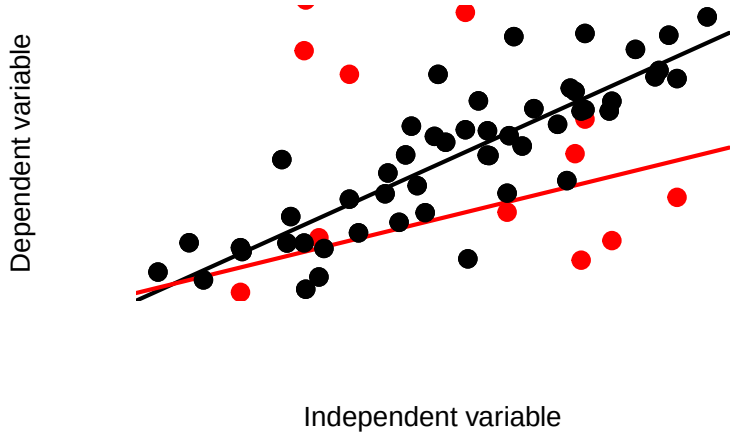


	Random	Systematic
Descriptive	Uncertainty	Bias
Independent	Underestimation	–
Dependent	Uncertainty	–

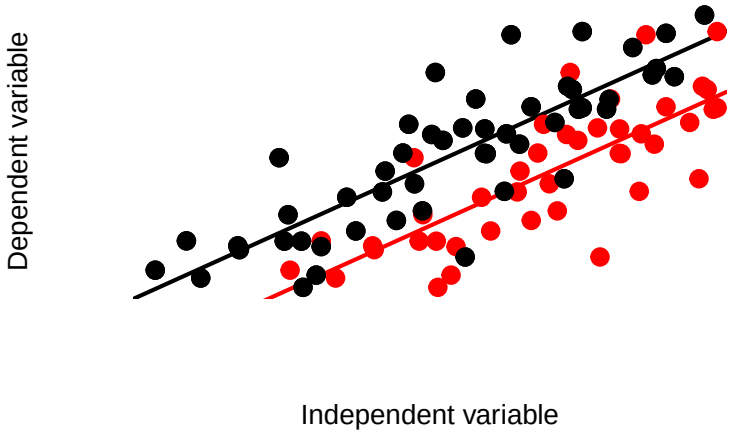
Random error − independent variable



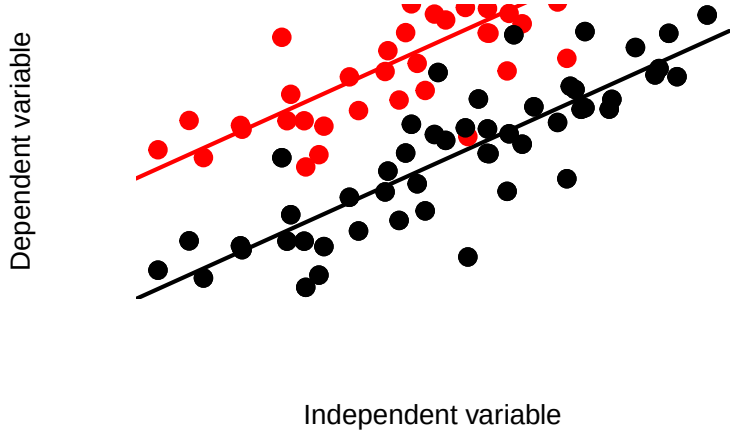
Random error − dependent variable



Systematic error − independent variable



Systematic error − dependent variable



Measurement error

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Assume there is measurement error in x, such that $x_i^* = x_i + v_i$, $v_i \sim N(0, \omega)$.

$$\begin{aligned} y_i &= \beta_0 + \beta_1 (x_i - v_i) + \varepsilon_i \\ &= \beta_0 + \beta_1 x_i + (\varepsilon_i - \beta_1 v_i) \\ &= \beta_0 + \beta_1 x_i + u_i, \end{aligned}$$

therefore, if $\beta_1 \neq 0$, x_i^* will be correlated with the error term u_i.

$$var(u_i) = \sigma^2 + \beta_1^2 v_i^2,$$

so variances will be inflated.

$$E(u_i|x_i) = E(u_i|v_i) = -\beta_1 v_i,$$

thus u_i and x_i will be correlated.

$\hat{\beta}^{OLS}$ will be **biased** and **inconsistent**. (Davidson and MacKinnon, 1999, 311)

Measurement error: bias

$$
\begin{aligned}
E(\varepsilon) &\neq 0 \\
E(\hat{\beta}^{OLS}|\mathbf{X}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \varepsilon) \\
&= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon
\end{aligned}
$$

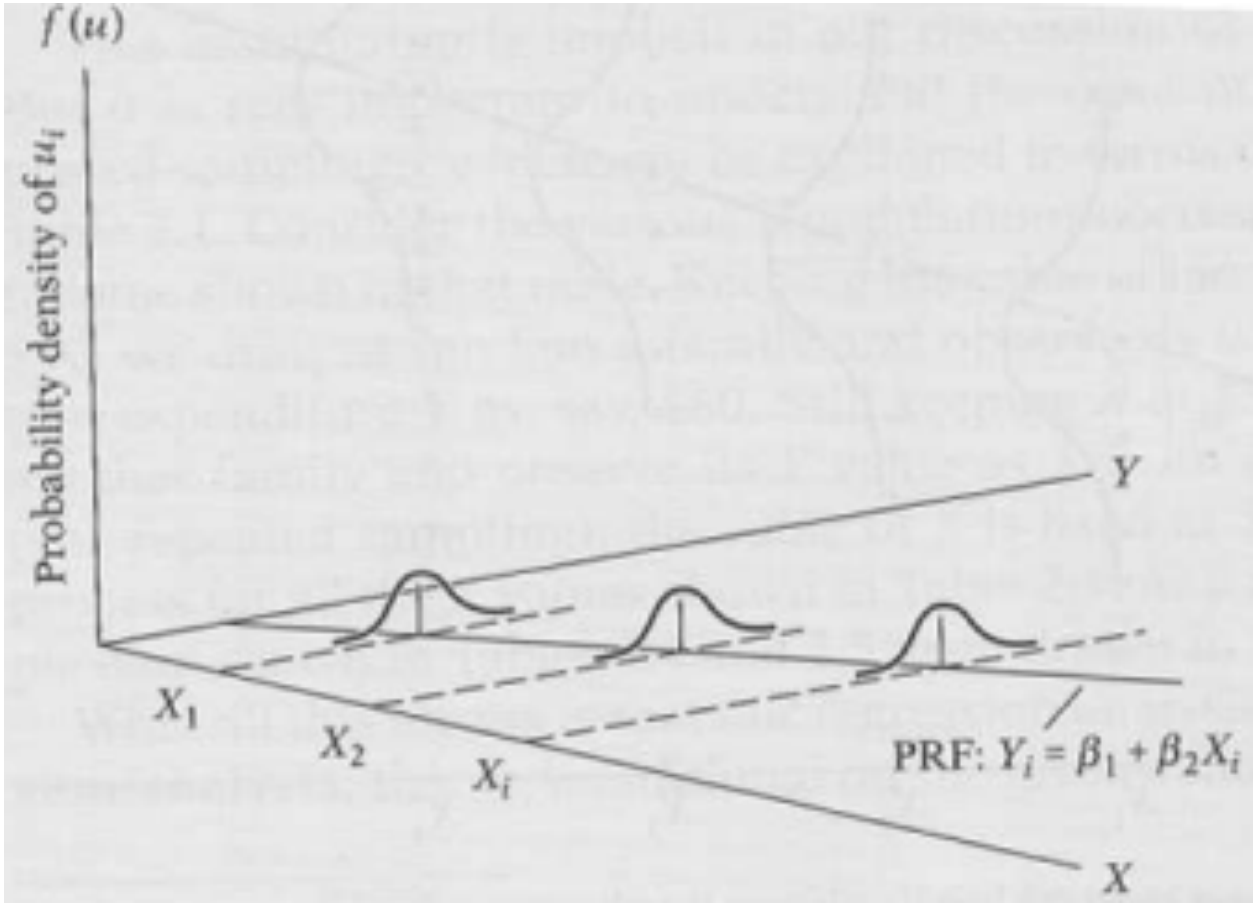
Since $E(u_i|x_i) = -\beta_1 v_i$, this means β_1 will be **underestimated** if $\beta_1 > 0$.

(Greene, 2002, 76)

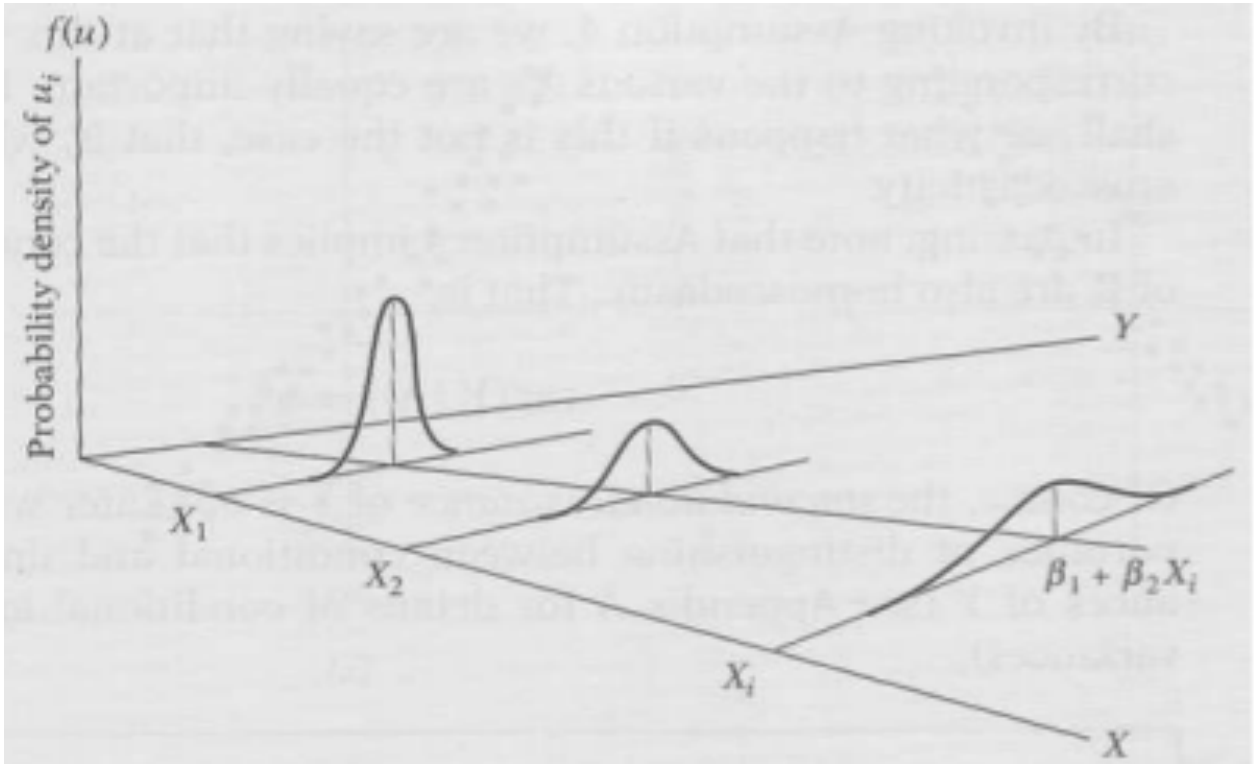
Instrumental variables is one method of dealing with measurement error in the independent variables (later in course).

Outline

Homoscedasticity



Heteroscedasticity



Heteroscedasticity

Regression **disturbances** whose variances are not constant across observations are **heteroscedastic**.

Under heteroscedasticity, the OLS estimators remain unbiased and consistent, but are no longer BLUE or asymptotically efficient.

(Thomas, 1985, 94)

Causes of heteroscedasicity

- More variation for larger sizes (e.g. profits of firms varies more for larger firms)
- More variation across different groups in the sample
- Learning effects in time-series
- Variation in data collection quality (e.g. historical data)
- Turbulence after shocks in time-series (e.g. financial markets)
- Omitted variable
- Wrong functional form
- Aggregation with varying sizes of populations
- etc.

Heteroscedasticity

Since OLS is no longer BLUE or asymptotically efficient,

- other linear unbiased estimators exist which have smaller sampling variances;
- other consistent estimators exist which collapse more quickly to the true values as n increases;
- we can no longer trust hypothesis tests, because $var(\hat{\beta}^{OLS})$ is biased.
  - $cov(\mathbf{X}_i^2, \sigma_i^2) > 0$, then $var(\hat{\beta}^{OLS})$ is underestimated
  - $cov(\mathbf{X}_i^2, \sigma_i^2) = 0$, then no bias in $var(\hat{\beta}^{OLS})$
  - $cov(\mathbf{X}_i^2, \sigma_i^2) < 0$, then $var(\hat{\beta}^{OLS})$ is overestimated (inefficient)

(Thomas 1985, 94–95; Judge et al. 1985, 422)

Heteroscedasticity

Normally we assume:

$$E(\varepsilon\varepsilon'|\mathbf{X}) = \sigma^2\mathbf{I} = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

For the heteroscedastic model we have:

$$E(\varepsilon\varepsilon'|\mathbf{X}) = \mathbf{\Omega} = \begin{bmatrix} \omega_1 & 0 & \dots & 0 \\ 0 & \omega_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \omega_n \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

Deriving $var(\hat{\beta}^{OLS})$

$$
\begin{aligned}
var(\hat{\beta}^{OLS}) &= E[(\hat{\beta}^{OLS} - \beta)(\hat{\beta}^{OLS} - \beta)'] \\
&= E[((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon)((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon)'] \\
&= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon\varepsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\varepsilon\varepsilon']\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}
$$

Deriving $var(\hat{\beta}^{OLS})$ under heteroscedasticity

$$
\begin{aligned}
var(\hat{\beta}^{OLS}) &= E[(\hat{\beta}^{OLS} - \beta)(\hat{\beta}^{OLS} - \beta)'] \\
&= E[((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon)((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon)'] \\
&= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon\varepsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\varepsilon\varepsilon']\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1},
\end{aligned}
$$

which cannot be simplified further and requires knowledge of $\mathbf{\Omega}$ to estimate.

Efficiency

Because observations with low variance will contain more information about the parameters than observations with high variance, an estimator which weighs all observations equally, like OLS, will not be the most efficient.

(Davidson and MacKinnon, 1999, 197)

Outline

Generalized Least Squares

More in general, if $var(\varepsilon_i) = \sigma^2\lambda_i$, with λ_i being some function of $\mathbf{X}_i$, then we can always transform our model by dividing all variables by $\sqrt{\lambda_i}$.

This is referred to as **generalized least squares** (GLS). (It is a generalization, because of $\lambda_i = 1$, we have OLS.)

With GLS, observations with lower σ^2 are weighted more heavily.

(Thomas 1985, 98; Judge et al. 1985, 421

Estimated Generalized Least Squares

To perform GLS estimation, σ_i^2 has to be known. In some cases we can estimate σ_i^2, in which case we talk of **estimated generalized least squares** (EGLS).

To estimate a model with minimal restrictions on σ_i^2, we are estimating a model with $n + k$ unknown parameters - i.e. the number of parameters to be estimated increases as n increases and the estimator is by definition inconsistent.

(Judge et al., 1985, 423)

Estimated Generalized Least Squares

Special cases where estimation might be possible:

- $\boldsymbol{\sigma}^2$ constant within subgroups
- $\boldsymbol{\sigma}^2 = (\mathbf{Z}\boldsymbol{\alpha})^2$, i.e. σ is linear function of exogenous variables
- $\boldsymbol{\sigma}^2 = \mathbf{Z}\boldsymbol{\alpha}$, i.e. σ^2 is linear function of exogenous variables
- $\boldsymbol{\sigma}^2 = \sigma^2(\mathbf{X}\boldsymbol{\beta})^p$, i.e. $var(\mathbf{y})$ is proportional to a power of its expectation
- $\boldsymbol{\sigma}^2 = e^{\mathbf{Z}\boldsymbol{\alpha}}$, "multiplicative heteroscedasticity"
- $e_t = v_t\sqrt{\alpha_0 + \alpha_1 e_{t-1}^2}$, "autoregressive conditional heteroscedasticity" (ARCH)

See Judge et al. (1985, 424ff) for an overview of estimators.

White's HCCM

When the form of the heteroscedasticity is unknown, we can get consistent estimates of $var(\hat{\beta}^{OLS})$ using a **heteroscedasticity consistent covariance matrix** (HCCM).

$$var(\hat{\beta}^{OLS}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

HCCM: estimate $\hat{\omega}_{ii} = (e_i - 0)^2 = e_i^2$, so that we have variance estimator

$$var(\hat{\beta}^{OLS}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'diag(e_i^2)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

Since there are several variations, this is called HC0 in the literature.

(Long and Ervin, 2000)

Residuals vs errors

Note that:

$$h_{ii} = \mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i'$$
$$var(e_i) = \sigma^2(1 - h_{ii}) \neq \sigma^2,$$

therefore $var(e_i)$ underestimates σ^2 and even when the errors (ε) are homoscedastic, the residuals ($\mathbf{e}$) are not.

So e_i^2, used in White's HC0, is, even though consistent, a biased estimator. The small sample properties turn out not to be very good.

(Long and Ervin, 2000)

HCCM variations

$$\text{HC0} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'diag(e_i^2)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

$$\text{HC1} = \frac{n}{n-k}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'diag(e_i^2)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \frac{n}{n-k}\text{HC0}$$

$$\text{HC2} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'diag\left(\frac{e_i^2}{1-h_{ii}}\right)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

$$\text{HC3} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'diag\left(\frac{e_i^2}{(1-h_{ii})^2}\right)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

Based on Monte Carlo analyses, HC3 is best in small samples.

(Long and Ervin, 2000)

White's HCCM in R

```
library(car)

m <- lm(...)
summary(m)

vcov <- hccm(m, type="hc3")
sqrt(diag(vcov))
```

(See notes for "manual" version.)

Bootstrap

Another solution for dealing with heteroscedasticity is to **bootstrap** to acquire standard errors.

```
se <- NULL
for (i in 1:1000) {
  sel <- sample(1:n, n, TRUE)
  mbs <- lm(y ~ x1 + x2, data=data[sel,])
  se <- rbind(se, sqrt(diag(vcov(mbs))))
}
colMeans(se)
```

Outline

Residual plots: heteroscedasticity

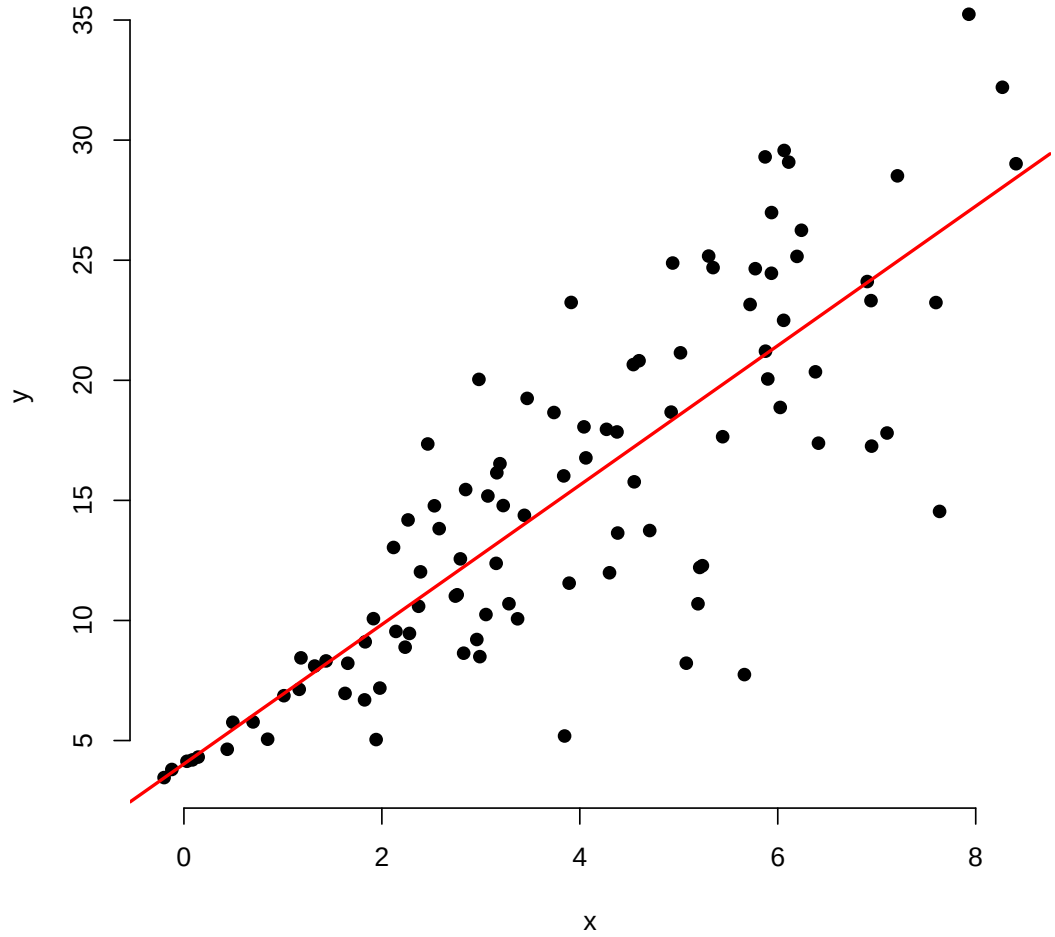
To detect heteroscedasticity (unequal variances), it is useful to plot:

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

Usually, the first one is sufficient to detect heteroscedasticity, and can simply be found by:

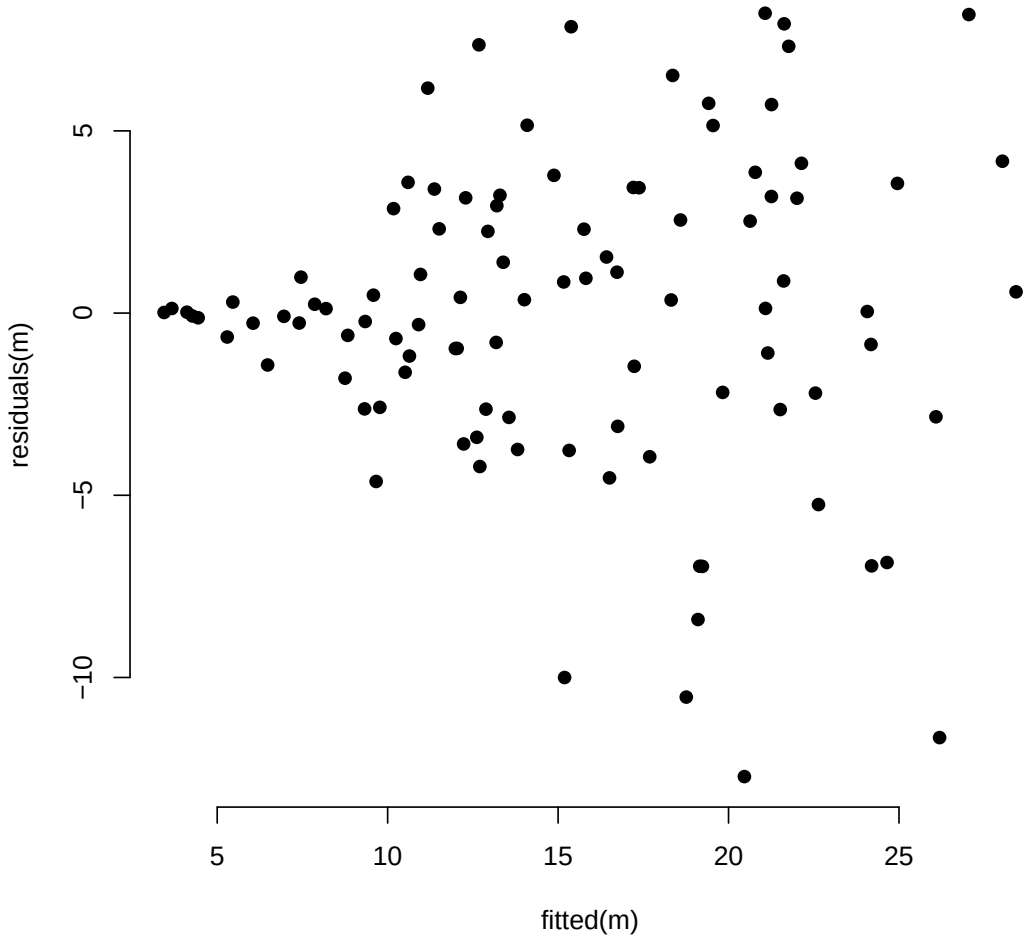
```
m <- lm(y ~ x)
plot(m)
```

Residual plots: heteroscedasticity



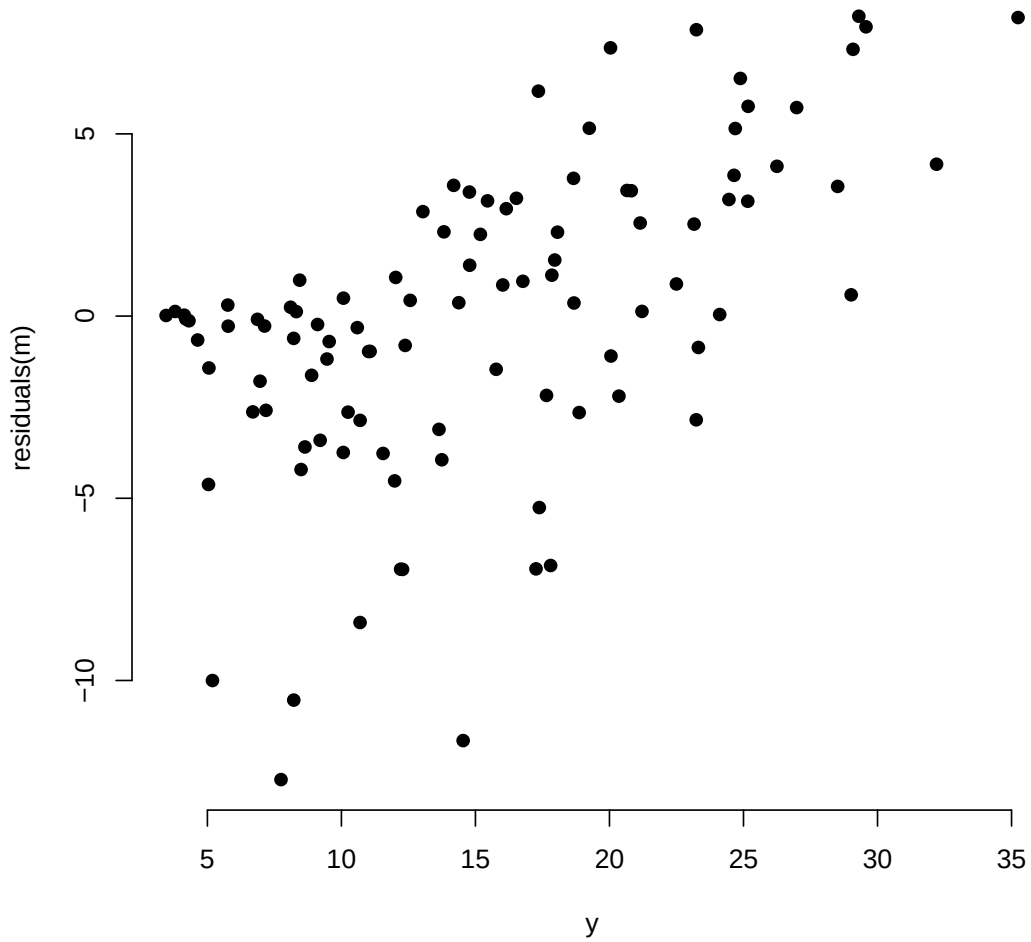

```
m <- lm(y ~ x)
```

Residual plots: heteroscedasticity



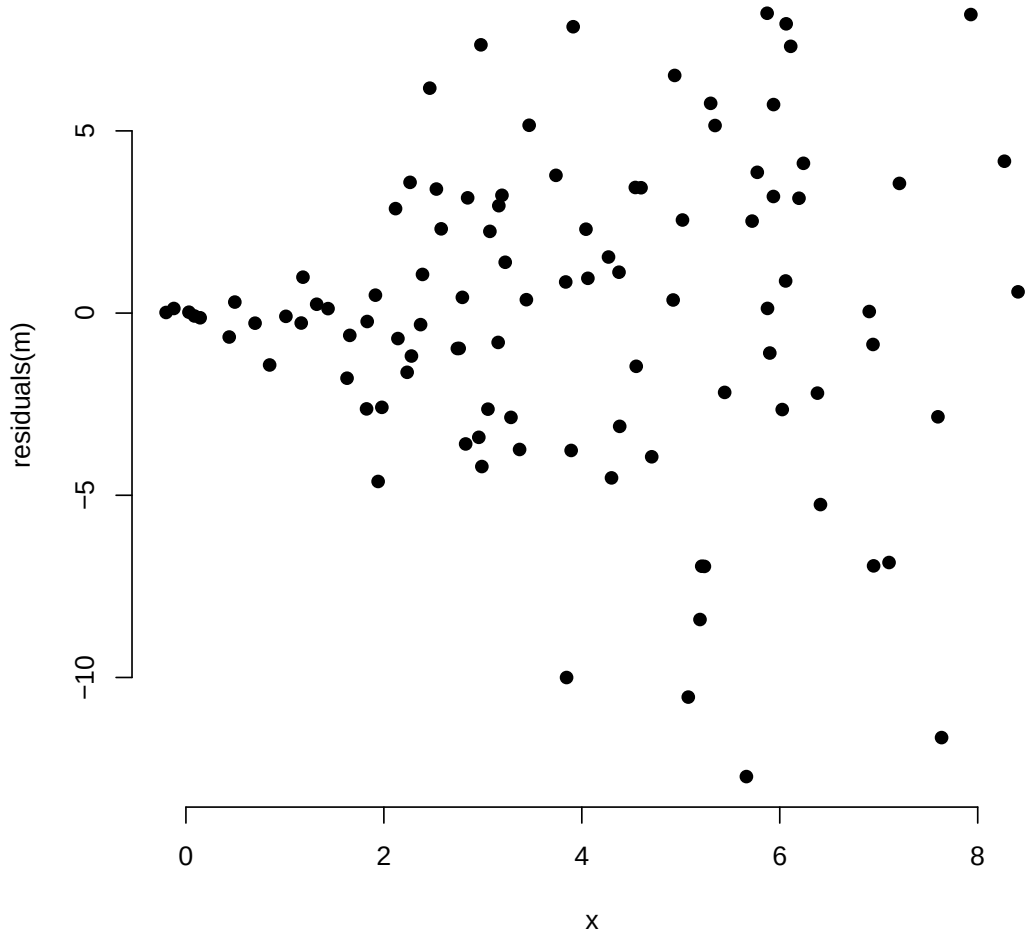

```
plot(residuals(m) ~ fitted(m))
```

Residual plots: heteroscedasticity



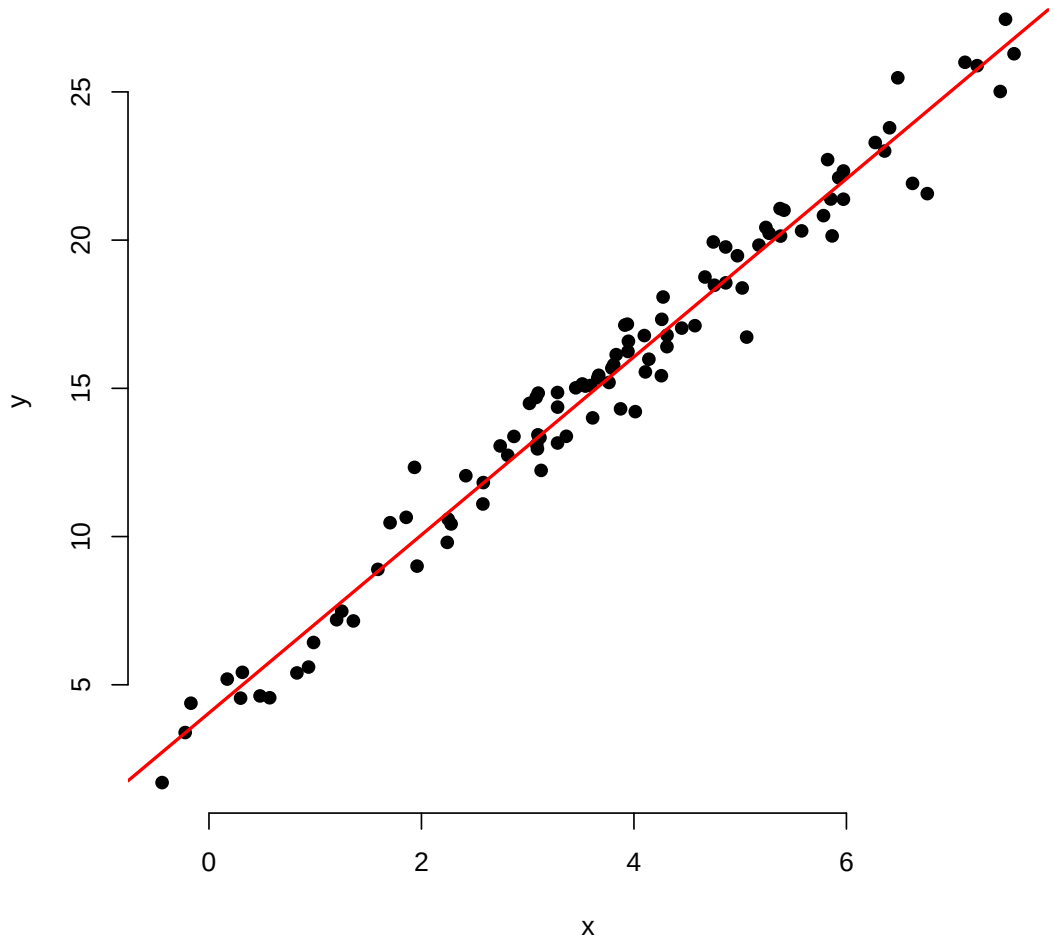

```
plot(residuals(m) ~ y)
```

Residual plots: heteroscedasticity



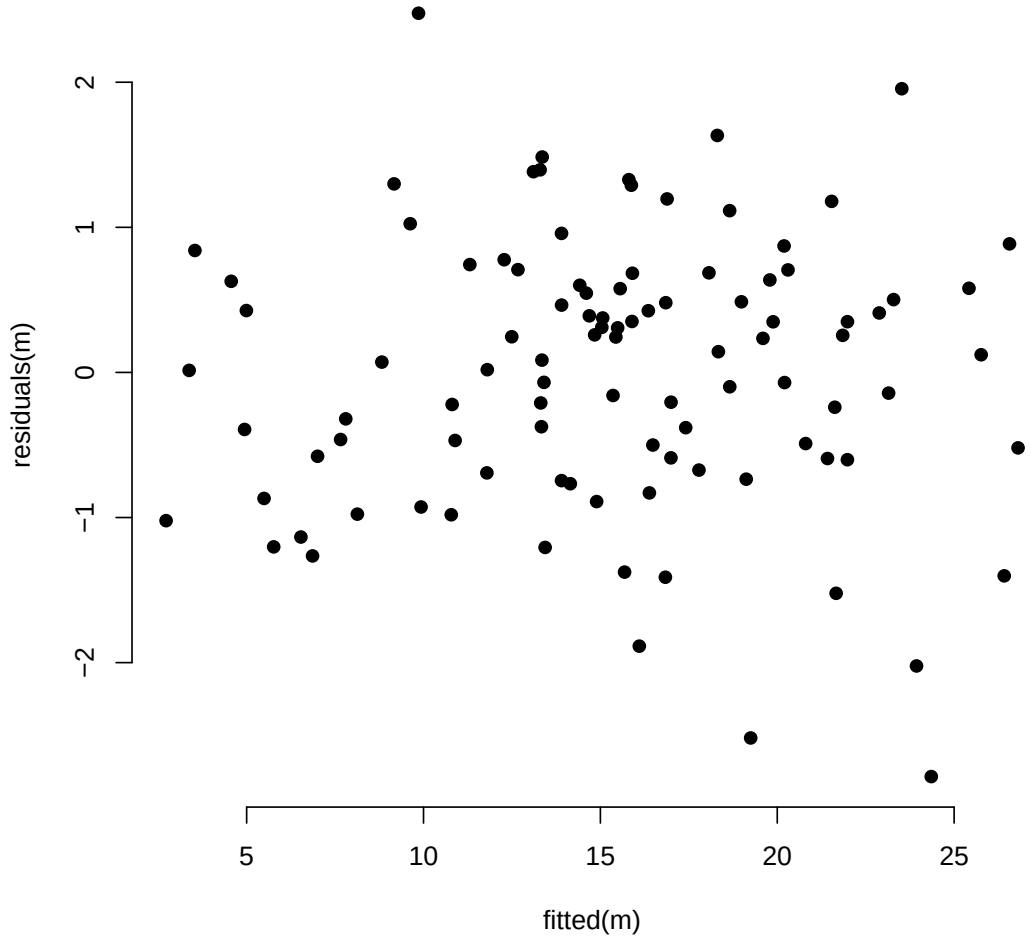

```
plot(residuals(m) ~ x)
```

Residual plots: homoscedasticity



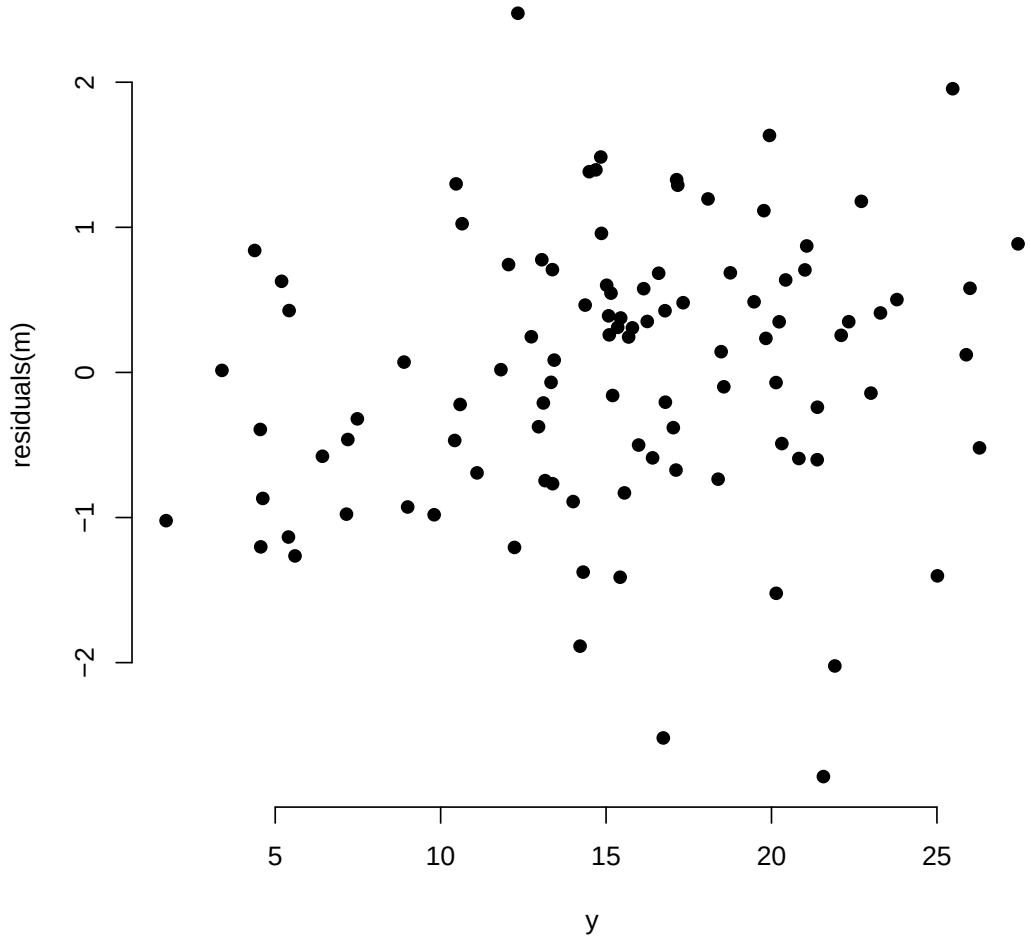

```
m <- lm(y ~ x)
```

Residual plots: homoscedasticity



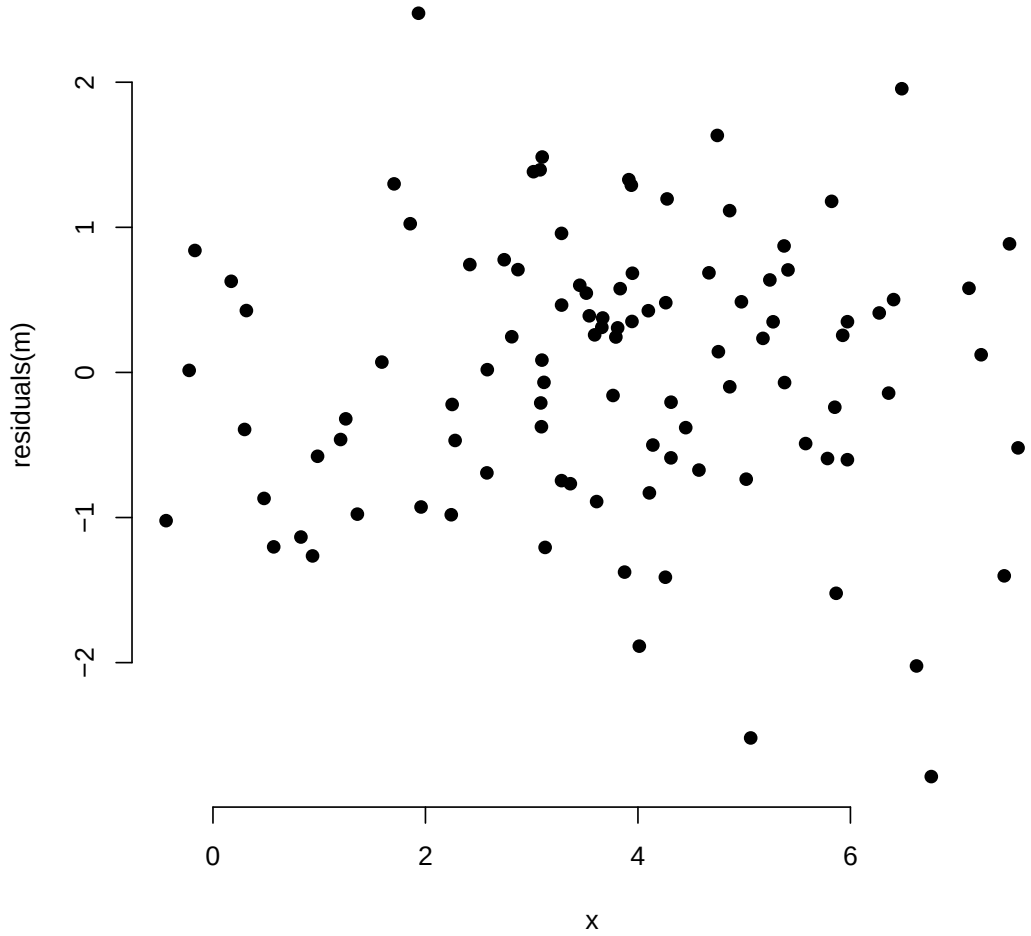

```
plot(residuals(m) ~ fitted(m))
```

Residual plots: homoscedasticity



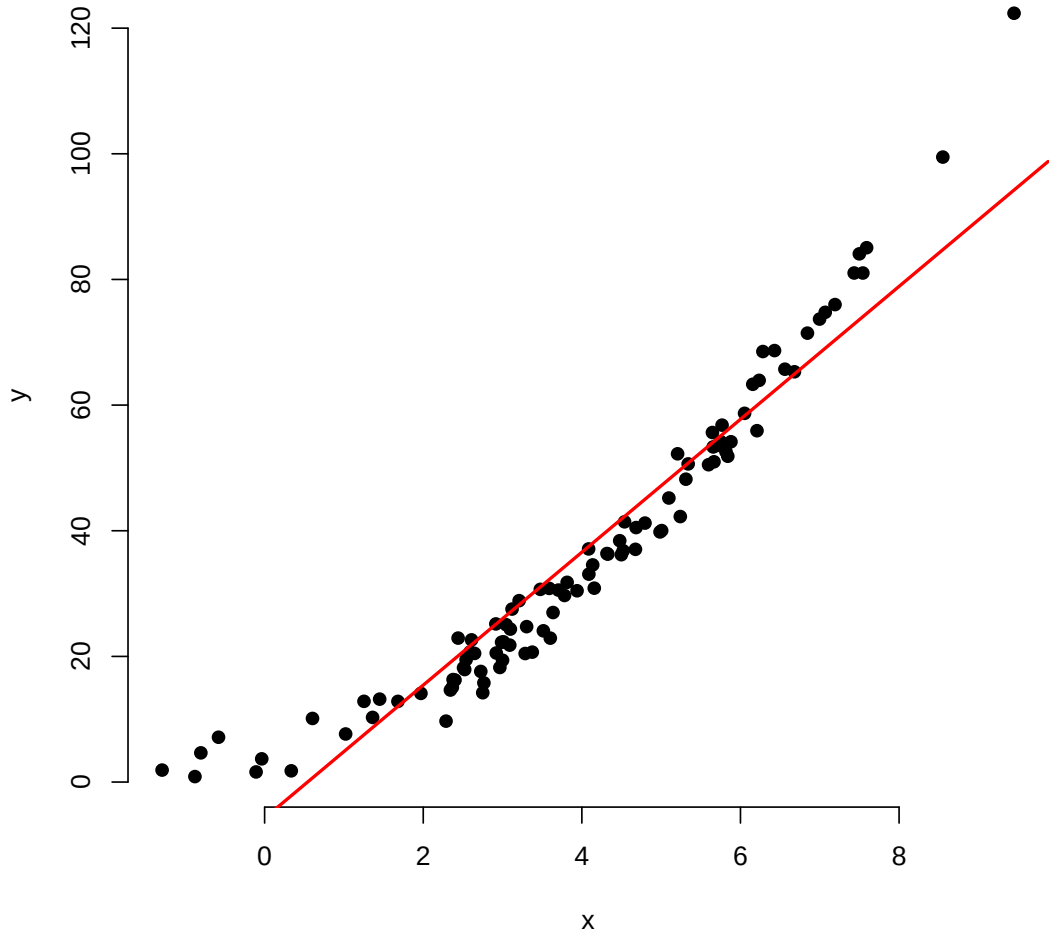

```
plot(residuals(m) ~ y)
```

Residual plots: homoscedasticity



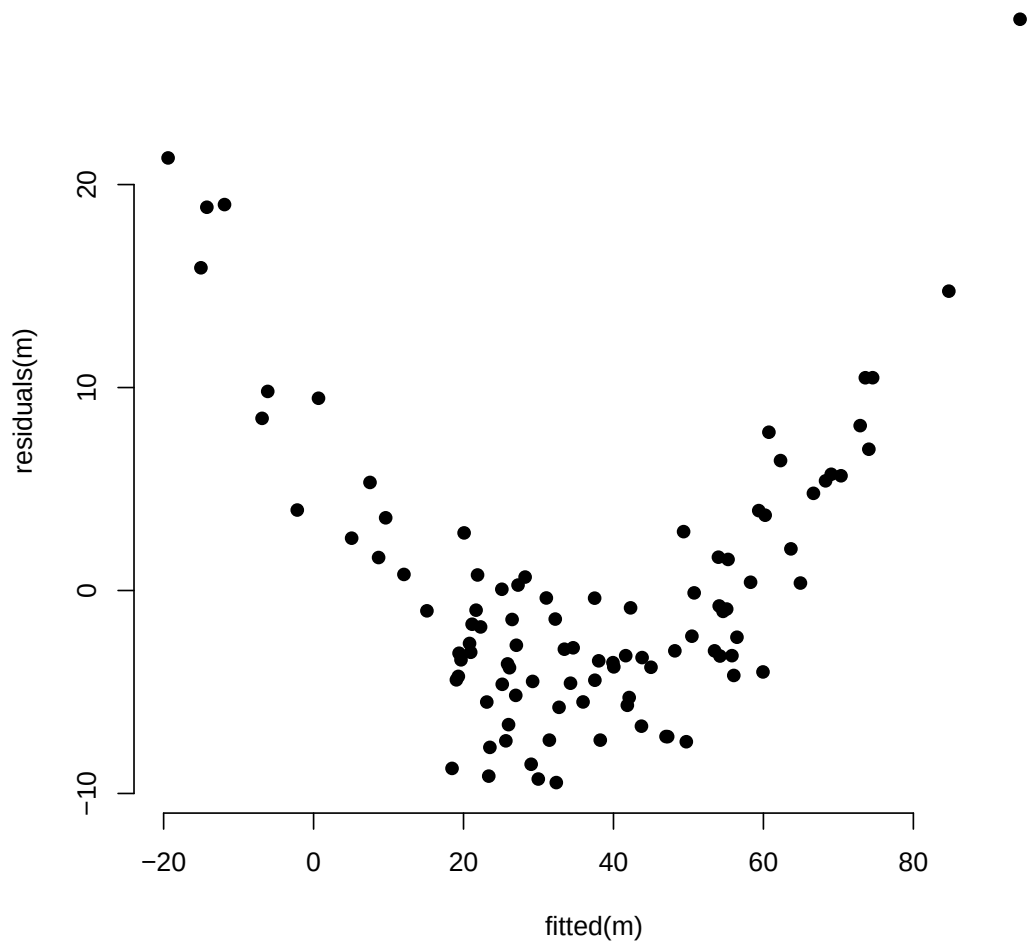

```
plot(residuals(m) ~ x)
```

Residual plots: heteroscedasticity



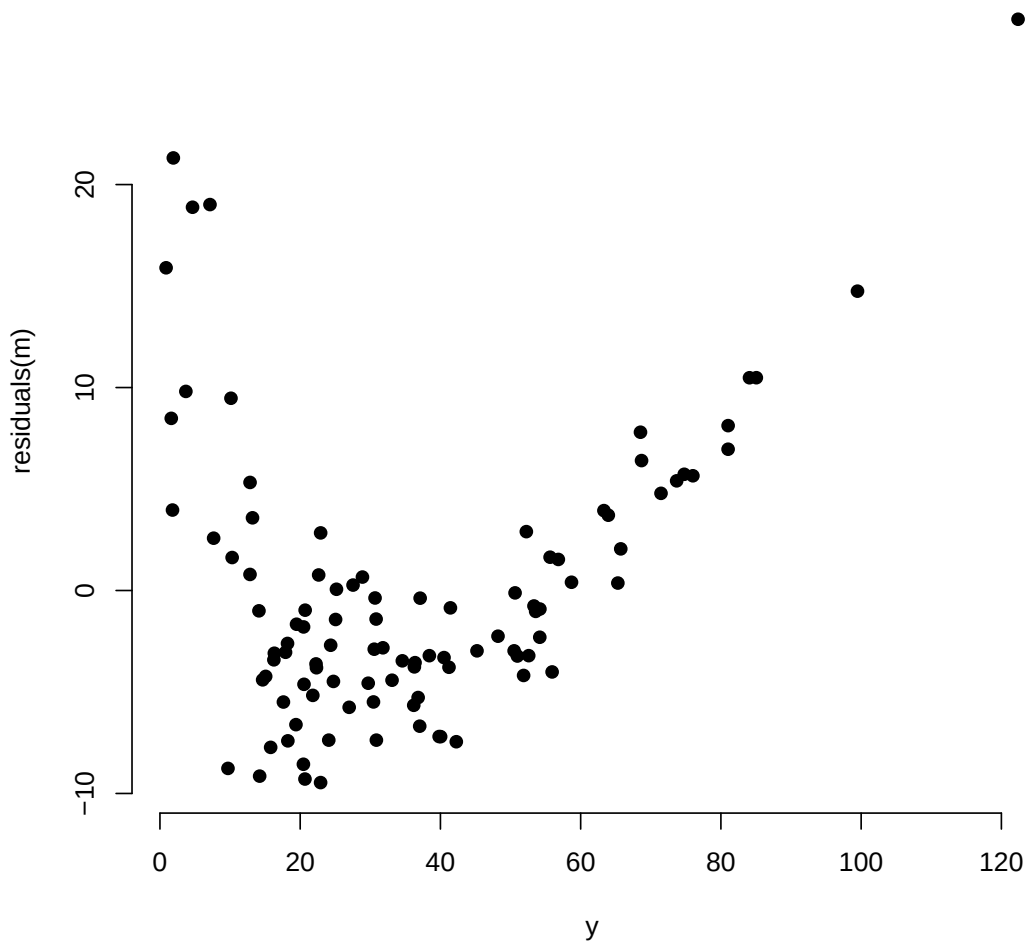

```
m <- lm(y ~ x)
```

Residual plots: heteroscedasticity



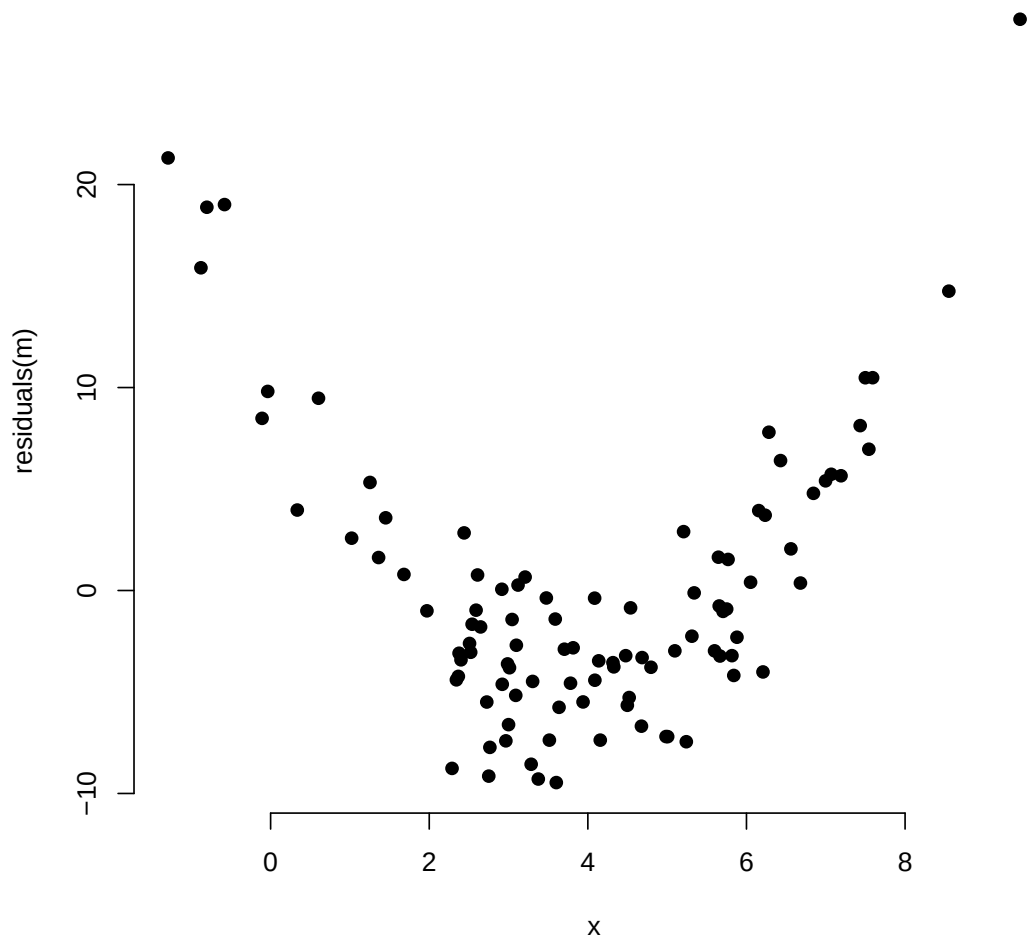

```
plot(residuals(m) ~ fitted(m))
```

Residual plots: heteroscedasticity




```
plot(residuals(m) ~ y)
```

Residual plots: heteroscedasticity




```
plot(residuals(m) ~ x)
```

Known groups

One way of testing for heteroscedasticity is if you expect that the variances might differ between two groups, is to run two separate regressions, for the two groups:

$$\frac{SSR_1/(n_1-k)}{SSR_2/(n_2-k)} \sim F(n_1-k, n_2-k)$$

$$H_0 : \sigma_1^2 = \sigma_2^2$$

(Wallace and Silver, 1988, 267)

Breusch-Pagan test

$$\sigma_i^2 = f(\mathbf{Z}\boldsymbol{\alpha})$$
$$\boldsymbol{\alpha} = \begin{bmatrix}\alpha_0 & \boldsymbol{\alpha}^*\end{bmatrix}'$$
$$H_0 : \boldsymbol{\alpha}^* = \mathbf{0}$$
$$H_1 : \boldsymbol{\alpha}^* \neq \mathbf{0}$$

with $f(\mathbf{Z}\boldsymbol{\alpha})$ being any function of $\mathbf{Z}\boldsymbol{\alpha}$ that does not depend on t. So this includes scenarios where $\sigma_i^2 = (\mathbf{Z}\boldsymbol{\alpha})^2$, or $\sigma_i = \mathbf{Z}\boldsymbol{\alpha}$, or $\sigma_i = e^{\mathbf{Z}\boldsymbol{\alpha}}$. If $\mathbf{Z}$ contains dummies for groups, it also includes heteroscedasticity due to different variances across subgroups.

Assumes $e_i^2 \sim N(0, \sigma_i^2)$

(Judge et al., 1985, 446)

Breusch-Pagan test

$$\eta = \frac{\mathbf{q}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{q}}{2\hat{\sigma}^4}$$
$$\sim \chi^2(s-1) \quad \text{asymptotically,}$$

where

$$q_i = e_i^2 - \hat{\sigma}^2$$
$$\hat{\sigma}^2 = \frac{1}{n}\mathbf{e}'\mathbf{e}$$

and $\mathbf{Z}_{n\times s}$ a matrix of exogenous variables.

```
bptest(lm(...), studentize=F)
```

Breusch-Pagan test

With more than one independent variable, an alternative approach is to look at an **auxiliary regression**:

$$e_i^2 = \gamma_0 + \gamma_1 \hat{y}_i^2 + v_i$$

If the model is homoscedastic and the variance is unrelated to $\hat{\mathbf{y}}$, then $H_0 : \gamma_1 = 0$. For this regression, $nR^2 \sim \chi^2(1)$.

```
summary(lm(residuals(m)^2 ~ fitted(m)))$r.sq * n
```

(Thomas 1985, 96-97)

Breusch-Pagan test

```
library(lmtest)

m <- lm(y ~ x)

bptest(m)
bptest(m, ~ z1 + z2)
```

By default, R assumes $\mathbf{Z} = \mathbf{X}$.

Goldfeld-Quandt test

To run a Goldfeld-Quandt test:

1. Omit r central observations from the data
2. Run two separate regressions, one for the first $(n-r)/2$ observations and one for the last
3. Calculate $R = SSR_1/SSR_2$
4. Perform test based on $R \sim F(\frac{1}{2}(n-r-2k), \frac{1}{2}(n-r-2k))$.

(Judge et al., 1985, 449)

Goldfeld-Quandt test

To run a Goldfeld-Quandt test:

1. Omit r central observations from the data
2. Run two separate regressions, one for the first $(n - r)/2$ observations and one for the last
3. Calculate $R = SSR_1/SSR_2$
4. Perform test based on $R \sim F(\frac{1}{2}(n - r - 2k), \frac{1}{2}(n - r - 2k))$.

(Judge et al., 1985, 449)

```
library(lmtest)
gqtest(m, n-40)
```

White's test

One solution for dealing with heteroscedasticity is calculating **White's heteroscedasticity-corrected standard errors**. The reasoning behind the **White test** is very straightforward: if there is homoscedasticity, the corrected standard errors should not be significantly different from the normal ones.

White's test

1. Regress e_i^2 on $\mathbf{x}_i$, all the variables in $\mathbf{x}_i$ squared, and all cross-products of $\mathbf{x}_i$; e.g. if

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

then run regression

$$e_i^2 = \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 x_{i1}^2 + \gamma_4 x_{i2}^2 + \gamma_5 x_{i1} x_{i2} + v_i$$

and calculate R^2;

2. Perform test on basis of $nR^2 \sim \chi^2(p-1)$, whereby p is the number of regressors in the auxiliary regression (6 in the example).

(Greene, 2003, 222)

White's test in R

```
m <- lm(y ~ x1 + x2)
bptest(m, ~ x1 * x2 + I(x1^2) + I(x2^2))
```

I.e. there does not appear to be an implementation of White's test in R, but it is equivalent to the Breusch-Pagan test with the independent variables as discussed.

Heteroscedasticity tests

In general,

- many of these tests require some idea about the shape of the heteroscedasticity;
- many of these tests have weak **power**, depending on the type of heteroscedasticity;
- if there is good reason to suspect heteroscedasticity, it is generally better to just use some robust estimation rather than test first — the tests are not reliable enough.

Outline

Notation: lagged variables

Instead of y_i to indicate each of n observations, we will use y_t to refer to each of T observations on a time-series.

y_{t-1} refers to the **lagged value**, i.e. the value of variable y at time $t-1$, the observation just one time period before time t.

A lag can have any length k ($k > 0$), y_{t-k}.

Notation: first differences

The difference between y_t and y_{t-1}, or the change in variable y at time t, is called the first difference, $\Delta y_t = y_t - y_{t-1}$.

Again, differences can have different lag lengths:
$\Delta y_{t-k} = y_{t-k} - y_{t-k-1}$.

Note that this means $\Delta y_{t-k} \neq y_t - y_{t-k}$, which some other authors might use instead.

The problem

Ignoring this autocorrelation leads to:

- $\hat{\beta}^{OLS}$ unbiased but inefficient (as long as $E(\varepsilon|\mathbf{X}) = 0$)
- $V(\hat{\beta}^{OLS})$ may be an under- or overestimate - the F- and t-tests cannot be trusted. If the autocorrelation is positive, $V(\hat{\beta}^{OLS})$ will be an underestimate.
- The residual variance is likely to be underestimated and R^2 overestimated.
- Risk of spurious regressions

Spurious regressions

When two variables are uncorrelated, but nonstationary, they often lead to highly significant estimates of their correlation in "naive" linear regression. Assume:

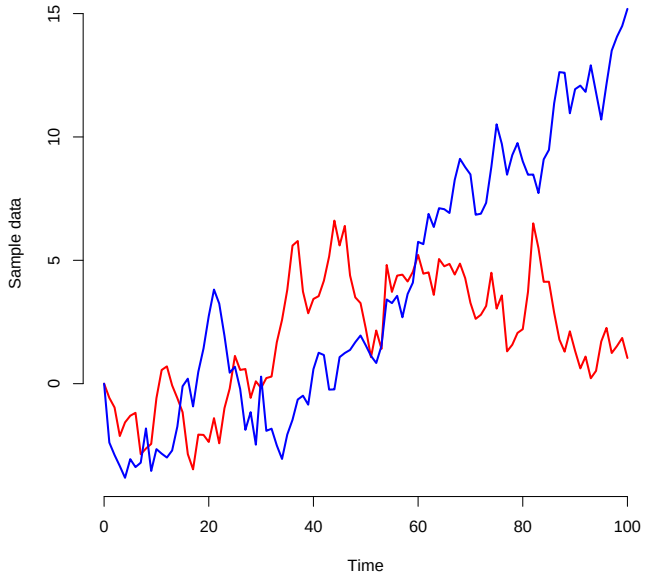
$$y_t = y_{t-1} + \varepsilon_{1,t}$$
$$x_t = x_{t-1} + \varepsilon_{2,t}.$$

Then OLS estimation of:

$$y_t = \alpha + \beta x_t + \varepsilon_t$$

will lead to a significant t-test on β.

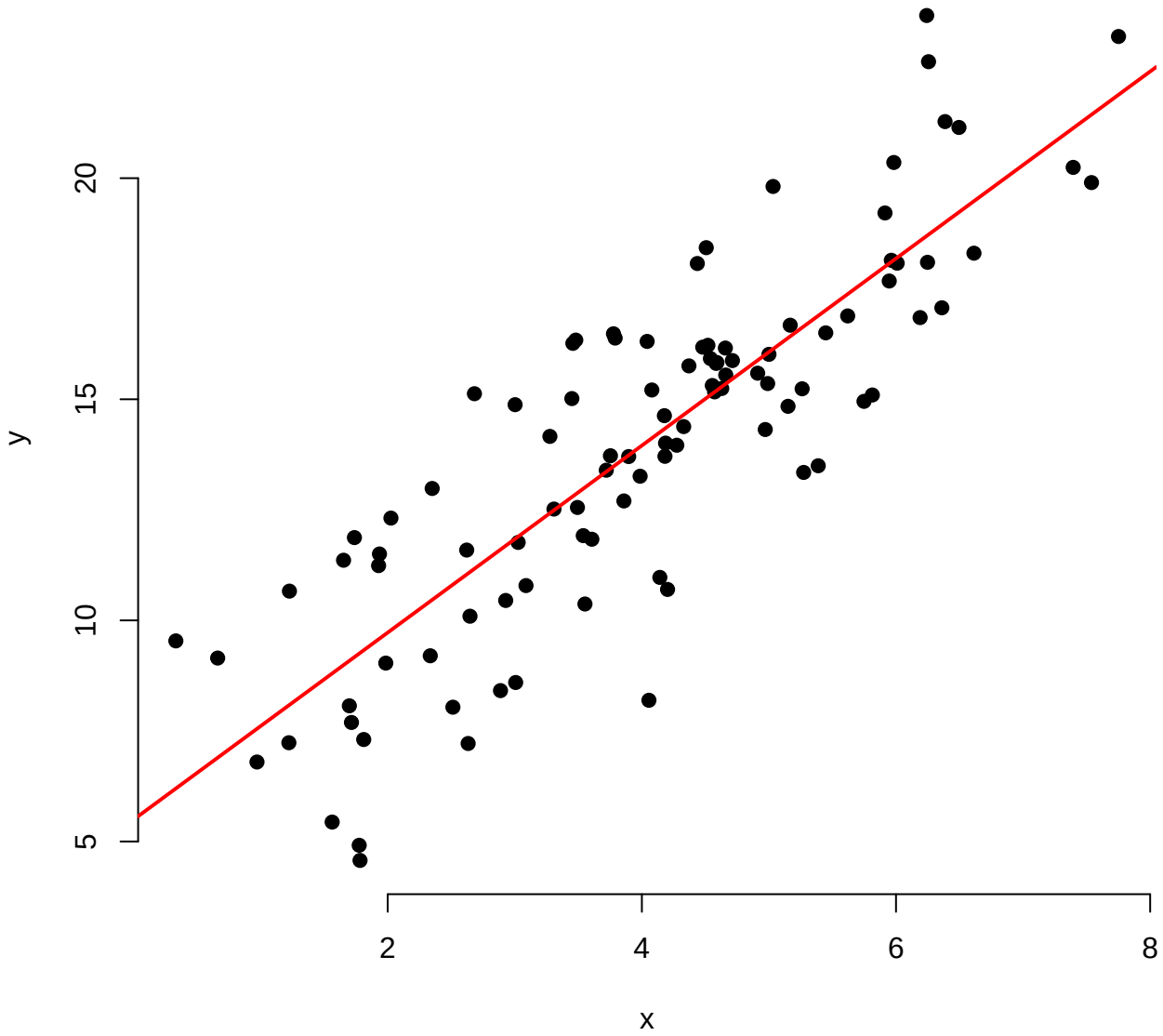
Spurious regression



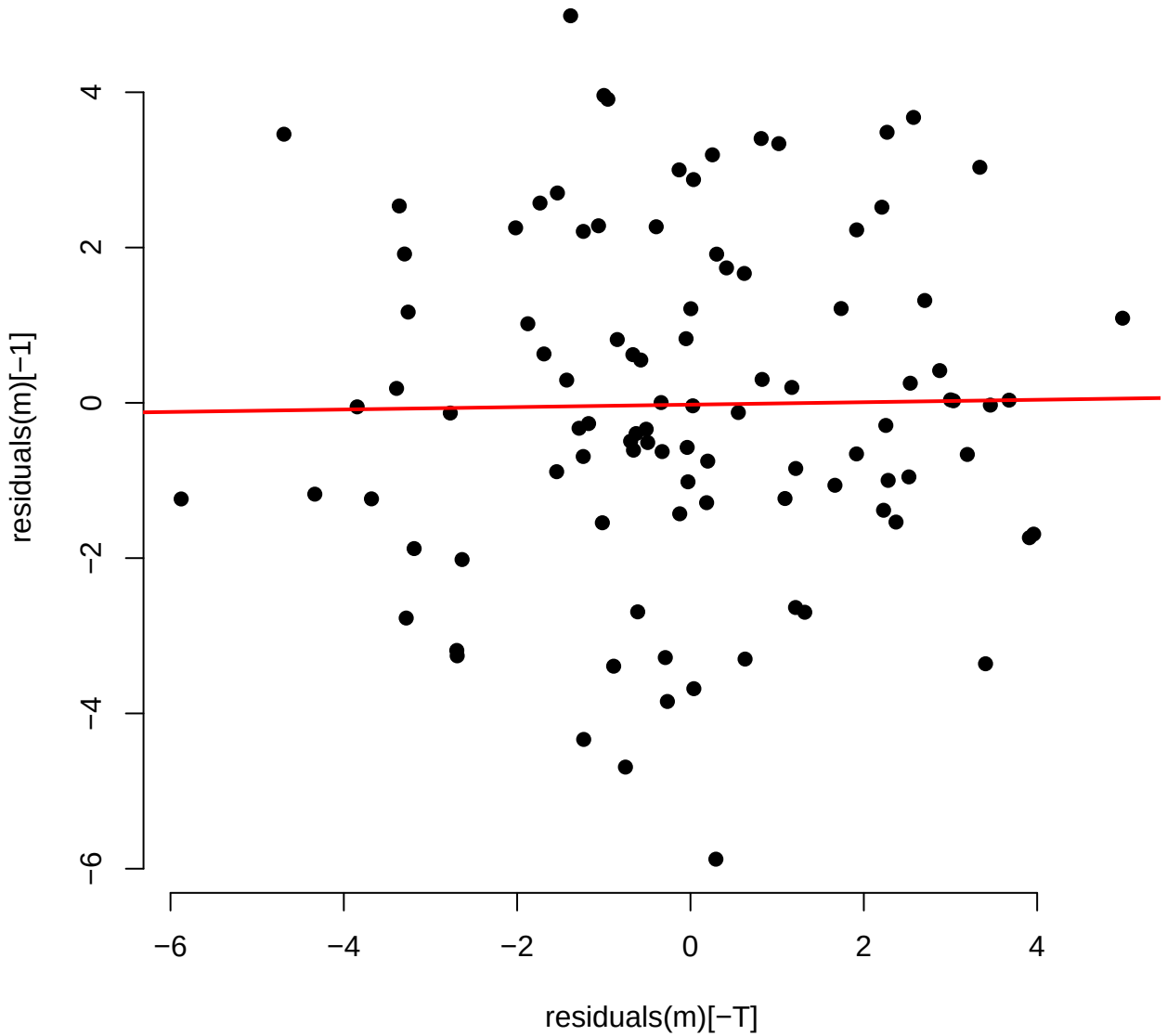

```
lm(formula = y ~ x)

             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.9646     0.3626  -2.660  0.00911 **
x            -0.9207     0.1002  -9.185 6.54e-15 ***

Residual standard error: 3.021 on 99 degrees of freedom
Multiple R-Squared: 0.4601, Adjusted R-squared: 0.4547
F-statistic: 84.37 on 1 and 99 DF,  p-value: 6.544e-15
```

Outline

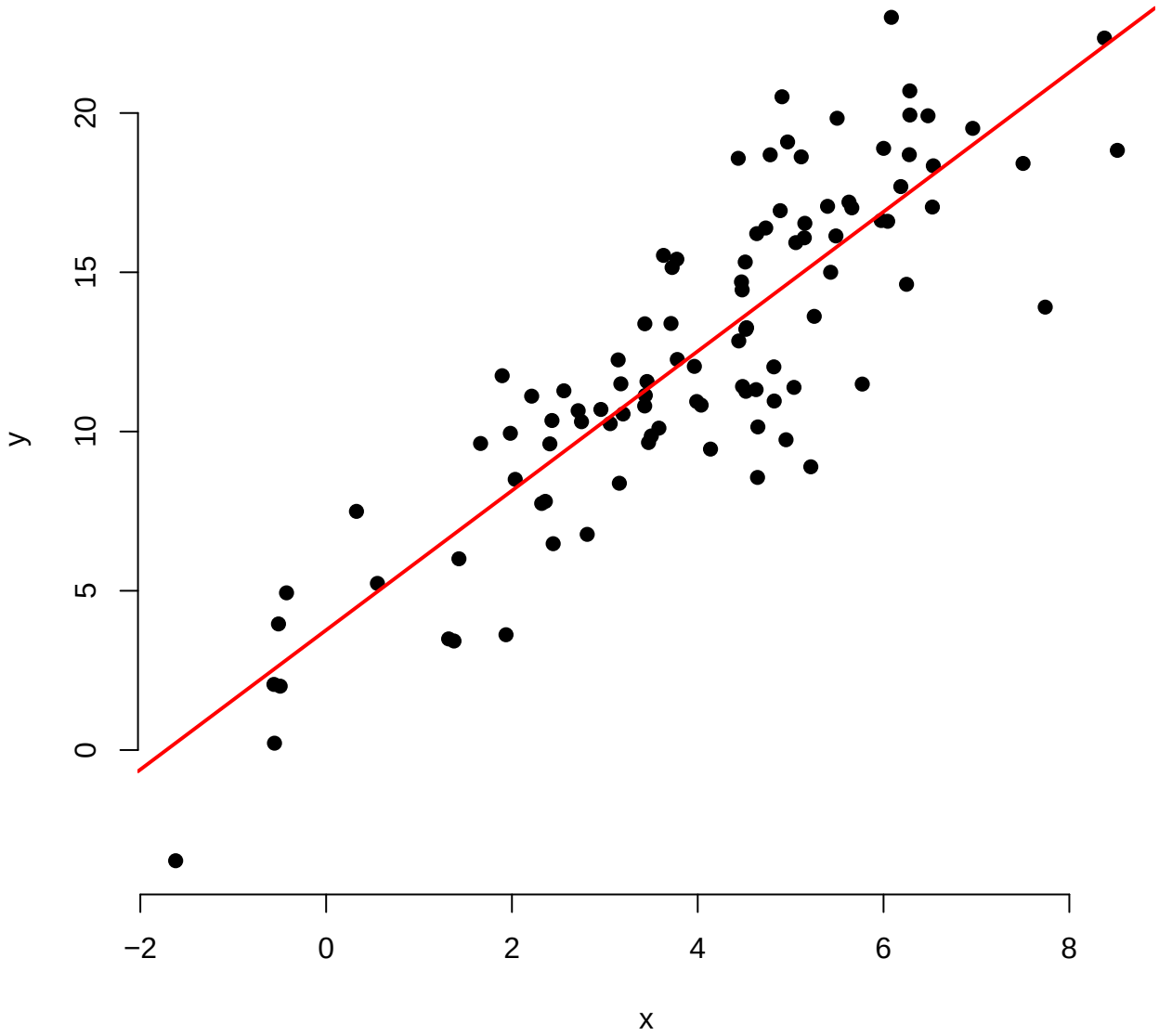
Residual plots: no autocorrelation



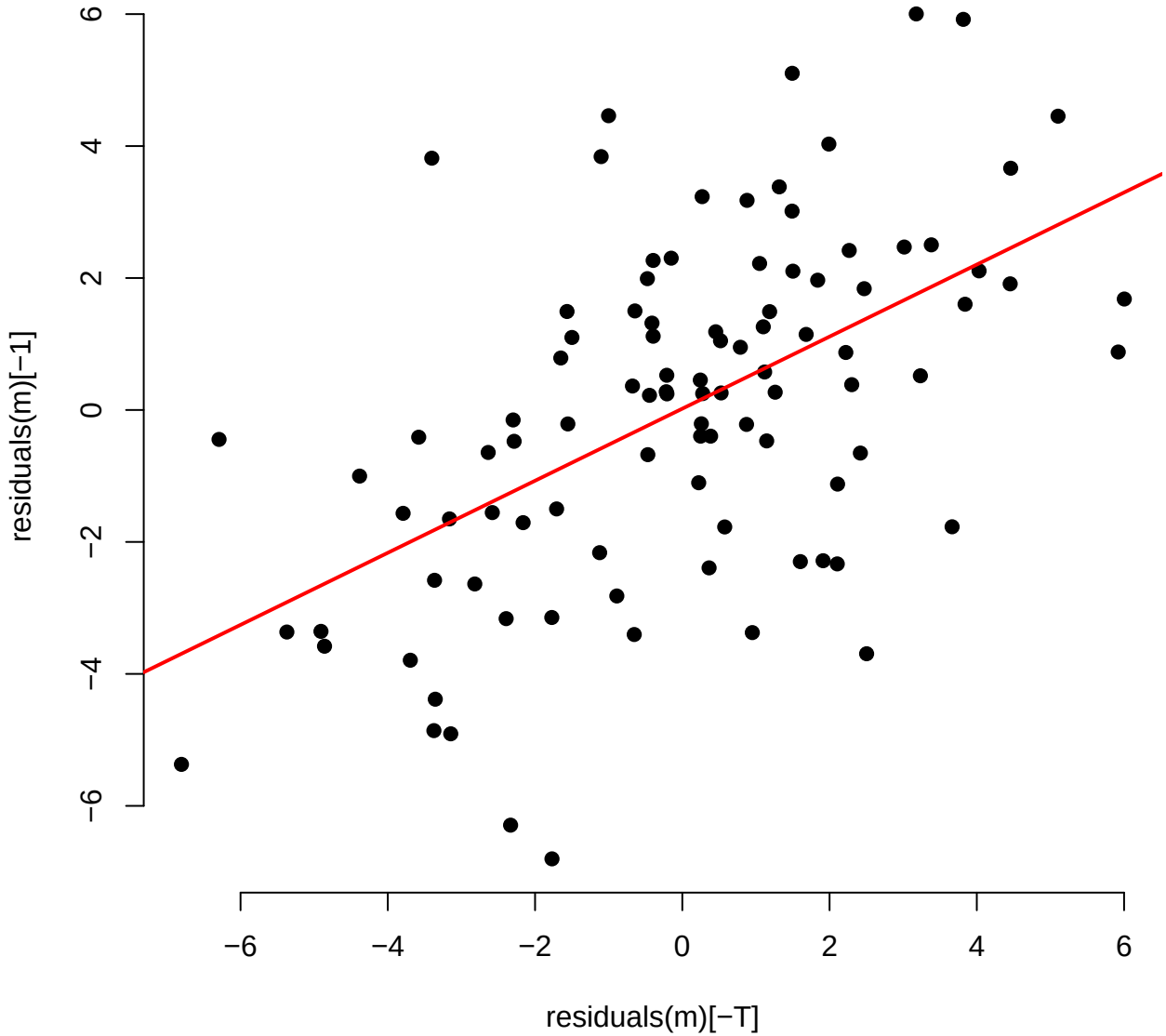
Residual plots: no autocorrelation



Residual plots: autocorrelation



Residual plots: autocorrelation



Durbin-Watson

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

If $\rho = cor(\varepsilon_t, \varepsilon_{t-1})$ and $\hat{\rho} = cor(e_t, e_{t-1})$, then $d \approx 2(1 - \hat{\rho})$. Thus, if d is close to 0 or 4, there is high first-order serial autocorrelation.

Note that $E[d] \approx 2 + \frac{2(k-1)}{n-k}$, thus biased.

Durbin-Watson

In matrix algebra, it could be written as:

$$d = \frac{\varepsilon' \mathbf{M}\mathbf{A}\mathbf{M}\varepsilon}{\varepsilon' \mathbf{M}\varepsilon} \qquad \mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}',$$

whereby

$$\mathbf{A} = \begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 2 & -1 \\ 0 & 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix}$$

The sampling distribution thus depends on $\mathbf{X}$.

Durbin-Watson

When the probability distribution of d is not exactly known, we can use threshold values. Given T and k, boundary values d_L and d_U have been tabulated.

E.g., if $T = 50$, $k = 6$, $\alpha = .05$ then $d_L = 1.335$ and $d_U = 1.771$, so we reject $H_0 : \rho > 0$ if $d < d_L$ and we do not reject if $d > d_U$, but in between we are undecided.

These threshold values are approximations and, depending on the speed at which regressors change, can be more or less appropriate.

Durbin-Watson

```
library(lmtest)
dwtest(model)
```

- Somewhat "old-fashioned" test, requiring special table.
- Assumes normally distributed errors.
- Model must include intercept.
- Requires **X** to be non-stochastic.
- Only tests for presence of AR(1) process.

Durbin's h test

The Durbin-Watson statistics cannot be used when there is a lagged dependent variable in the model. You should, with such variable, always test for remaining autocorrelation, however. One possible test is Durbin's h-test.

$$h = (1 - \frac{1}{2}d)\sqrt{\frac{T}{1 - T \cdot V(\hat{\beta}_{y_{t-1}})}} \overset{a}{\sim} N(0,1).$$

Breusch-Godfrey LM test

A more powerful test, which can handle higher order autoregressions, is the Breusch-Godfrey LM test.

1. Estimate OLS
2. Regress $\mathbf{e}$ on $\mathbf{X}$ and lagged values of $\mathbf{e}$ $(e_{t-1}, e_{t-2}, \cdots, e_{t-k})$
3. $(T-k)R^2 \overset{a}{\sim} \chi^2(k)$

```
library(lmtest)
bgtest(model, order = 3)
```

This assumes normally distributed errors. A slightly more general Gauss-Newton regression would not make this assumption.

Gauss-Newton regression

Assume an AR(1) process: $y_t = \mathbf{x}_t'\boldsymbol{\beta} + u_t,\ u_t = \rho\varepsilon_{t-1} + \varepsilon_t.$

In this case, we can simply first regress $\mathbf{y}$ on $\mathbf{X}$, and then use the residuals from this regression ($\hat{\mathbf{u}}$) to regress $\hat{\mathbf{u}}$ on $\mathbf{X}$ and $\tilde{\mathbf{u}}$, whereby $\tilde{u}_1 = 0$ and $\tilde{u}_t = \hat{u}_{t-1} \quad \forall \quad t > 1$:

$$\hat{\mathbf{u}} = \mathbf{X}\tilde{\boldsymbol{\beta}} + \tilde{\mathbf{u}}\tilde{\rho} + \tilde{\boldsymbol{\varepsilon}}$$

The test can easily be extended by including multiple lags and performing an F-test on all $\tilde{\rho}$'s.

The test is also valid for testing MA(q) or ARMA(p,q) processes.

(Davidson and MacKinnon, 1993, 357–360)

Gauss-Newton regression

```
m <- lm(y ~ x1 + x2)
T <- dim(m$model)[1]
u <- residuals(m)
u.tilde <- c(0, u[-T])
summary(lm(u ~ x1 + x2 + u.tilde))
```

and then check the t-test for the $\tilde{\mathbf{u}}$ variable.

Appendix

Simultaneity

In many scenarios, the causal relationship between y and x might be in both directions.

E.g. more economically developed countries are more likely to be democratic and democracies are likely to perform better economically.

Simultaneity

Assuming reverse causality:

$$y_i = \alpha_1 + \beta_1 x_i + u_i$$
$$x_i = \alpha_2 + \beta_2 y_i + v_i$$

$$x_i = \alpha_2 + \beta_2(\alpha_1 + \beta_1 x_i + u_i) + v_i$$
$$x_i - \beta_2\beta_1 x_i = (\alpha_2 + \beta_2\alpha_1) + \beta_2 u_i + v_i$$
$$x_i = \frac{\alpha_2 + \beta_2\alpha_1}{1 - \beta_1\beta_2} + \frac{\beta_2}{1 - \beta_1\beta_2} u_i + \frac{1}{1 - \beta_1\beta_2} v_i,$$

thus x_i and u_i will be correlated.

Simultaneity

You can model a set of simultaneous equations such as this, but if you want to estimate just one of the equations, **instrumental variable** estimation is an option.

Heteroscedasticity: aggregation example

Imagine we have the following model:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij},$$

whereby i indicates the individual, and j the region of this individual, with n_j individuals per region.

Say we only have regional level data, $\bar{y}_j = \frac{1}{n_j} \sum_i^{n_j} y_{ij}$ and $\bar{x}_j = \frac{1}{n_j} \sum_i^{n_j} x_{ij}$:

$$\bar{y}_j = \beta_0 + \beta_1 \bar{x}_j + \bar{\varepsilon}_j,$$

where $\bar{\varepsilon}_j = \frac{1}{n_j} \sum_i^{n_j} \varepsilon_{ij}$.

(Thomas, 1985, ∼98)

Heteroscedasticity: aggregation example

$$\bar{y}_j = \beta_0 + \beta_1 \bar{x}_j + \bar{\varepsilon}_j$$

$$E(\bar{\varepsilon}_j) = 0$$

$$E(\bar{\varepsilon}_j^2) = \frac{1}{n_j^2} E(\sum_i^{n_j} \varepsilon_{ij}) = \frac{n_j}{n_j^2}\sigma^2 = \frac{1}{n_j}\sigma^2$$

Therefore, $var(\bar{\varepsilon}_j)$ depends on n_j and thus varies across cases.

(Judge et al., 1985, 419–420)

Heteroscedasticity: aggregation example

$$\bar{y}_j = \beta_0 + \beta_1 \bar{x}_j + \bar{\varepsilon}_j$$

In this case the fix is actually easy: since $var(\varepsilon_j) = \sigma^2/n_j$, $var(\sqrt{n_j}\varepsilon_j) = \sigma^2$, so the heteroscedasticity can be avoided by transforming the variables:

$$\sqrt{n_j}\bar{y}_j = \beta_0\sqrt{n_j} + \beta_1\sqrt{n_j}\bar{x}_j + \varepsilon_j^*$$

(Thomas, 1985, ∼98)

Heteroscedasticity: solution

When the type of heteroscedasticity is known, we can often transform the data. An example is the multiplication with $\sqrt{n_j}$ of each term in the equation for the group means regression. Another example: if $var(\varepsilon_i) = \sigma^2 x_{i1}^2$, then $var(\varepsilon_i / x_{i1}) = \sigma^2$, so:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

$$\frac{y_i}{x_{i1}} = \beta_0 \frac{1}{x_{i1}} + \beta_1 \frac{x_{i1}}{x_{i1}} + \beta_2 \frac{x_{i2}}{x_{i1}} + \frac{\varepsilon_i}{x_{i1}}$$

$$y_i^* = \beta_1 + \beta_0 x_{i1}^* + \beta_1 x_{i2}^* + \varepsilon_i^*$$

(note the intercepts interpretation)

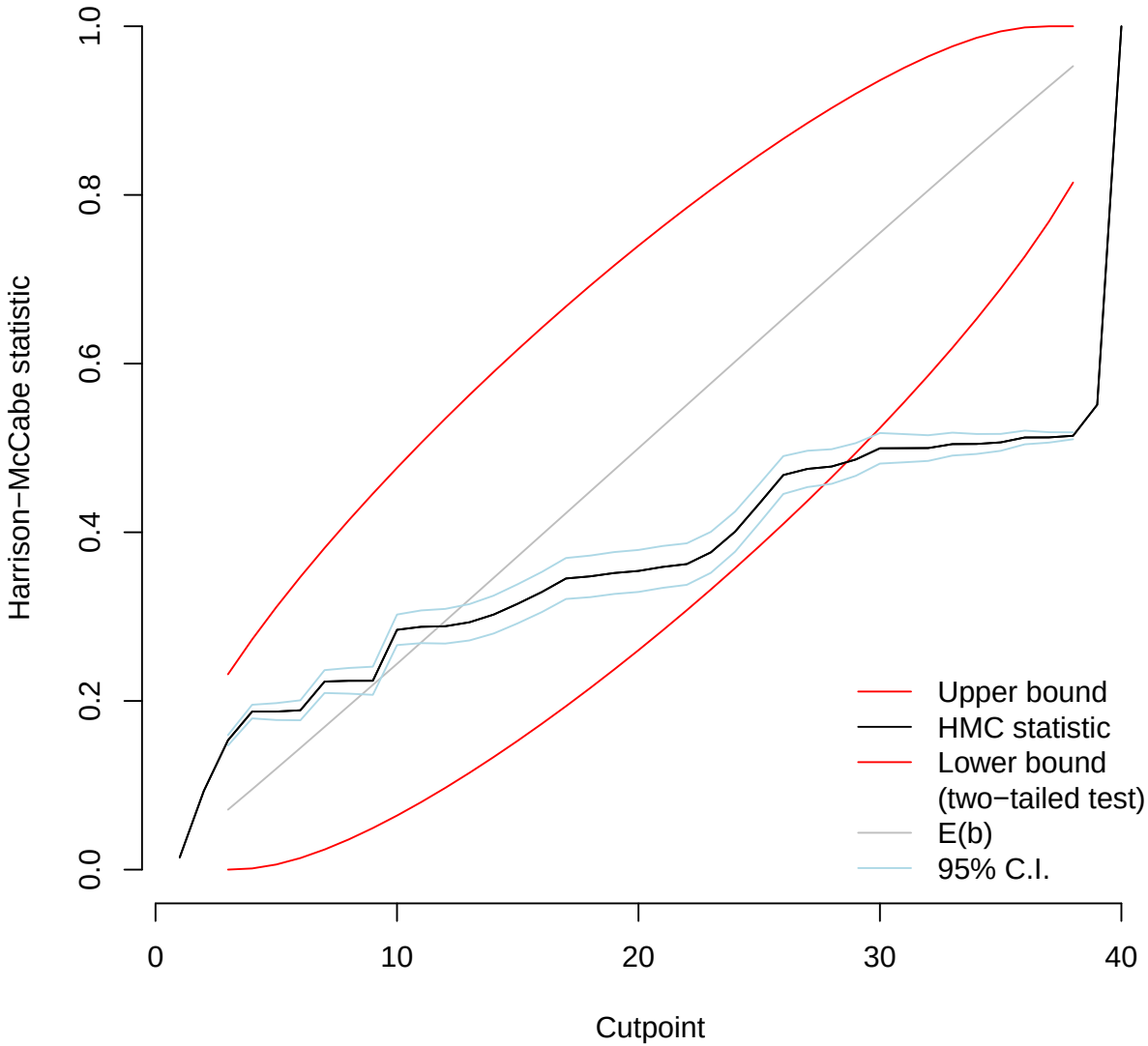
(Thomas, 1985, 98)

Harrison-McCabe test

For the above tests we always run several regressions, because even if errors are uncorrelated, residuals are not independent. If residuals are not independent, a ratio of subsets of these residuals do not have an F-distribution, while if we run separate regressions, the residuals will be independent (if the errors are) and such a ratio will have an F-distribution.

Harrison and McCabe (1979) suggest that such a ratio of subsets of the residuals do have a β-distribution, however.

Harrison-McCabe test



Cook, R. Dennis. 1977. "Detection of influential observation in linear regression." *Technometrics* pp. 15–18.

Cook, R. Dennis. 1979. "Influential observations in linear regression." *Journal of the American Statistical Association* 74(365):169–174.

Davidson, Russell and James G. MacKinnon. 1993. *Estimation and inference in econometrics*. Oxford: Oxford University Press.

Davidson, Russell and James G. MacKinnon. 1999. *Econometric theory and methods*. Oxford: Oxford University Press.

Greene, William H. 2002. *Econometric analysis*. London: Prentice Hall.

Greene, William H. 2003. *Econometric analysis*. 5th ed. Upper Saddle River: Prentice Hall.

Harrison, Michael J. and Brendan P.M. McCabe. 1979. "A test for heteroscedasticity based on ordinary least squares residuals." *Journal of the American Statistical Association* 74(366):494–499.

Judge, George G., William E. Griffiths, R. Carter Hill, Helmut Lutkepohl and Tsoung-Chao Lee. 1985. *The Theory and Practice of Econometrics*. New York: John Wiley and Sons.

Kutner, Michael H., Christopher J. Nachtsheim, John Neter and William Li. 2005. *Applied linear statistical models*. 5th ed. McGraw-Hill.

Long, J. Scott and Laurie H. Ervin. 2000. "Using heteroscedasticity consistent standard errors in the linear regression model." *The American Statistician* 54(3):217–224.

Teräsvirta, Timo, Dag Tjøstheim and Clive W.J. Granger. 1992. "Aspects of modelling nonlinear time series." *Research Report* 1.

Thomas, R. Leighton. 1985. *Introductory econometrics: Theory and applications*. Longman Harlow, Essex.

Wallace, T. Dudley and J. Lew Silver. 1988. *Econometrics: An introduction*. Addison-Wesley Longman.