Chapter 4

Linear Regression

Regression: variety

Regression analysis is one of the most popular and flexible methods of statistical analysis. Most of other techniques can be linked to the regression.

Aim

- Analysis of the relationships between one dependent variable (regressand, output) and one or many independent variables (regressors, inputs);
- the relationships should be quantified and explained;
- data-driven problems with the model should be addressed in details.

- Reality is complex and it is almost never possible to find a functional form which precisely describes real-world relationships.
- A model is a simplification of the reality: only the most important characteristics are reflected.

one	one or many
dependent variable	independent variable(s)
metric	any scale
y	$X_1, X_2, \ldots, X_j, \ldots, X_K$

$$y \approx f(X_1, X_2, \ldots, X_K)$$

Example: relationship between the sales of a particular product and its price (X_1), costs of merchandizing (X_2), etc.

Notation: We write X_j for variables and x_{ij} for observations with $i = 1, ..., N$ and $j = 1, ..., K$.

Note: cross-sectional data, i.e. i is not a time index here!

- Causal dependence: the changes in y are responses to changes in in $X_1, \ldots, X_K$
  - sales vs. famous
  - sales of ice-cream vs. weather

- The choice of regressors should be economically motivated (else "spurious regression")

How to choose the function $f(\cdot)$?

- f is a linear function in the coefficients β_j and in the variables X_j $\leadsto$ linear regression

$$y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K$$

- Note: a very restrictive assumption; the impacts cannot be taken in the same linear form for all the variables and all the observations
- But: the linear function is often a good approximation to non-linear relationships
- Later: we extend the model to a non-linear function or transformed X-variables

Regression: Linear regression

Deterministic model: we model the relationship between wage W and experience E by

$$W = \beta_0 + \beta_1 \cdot E.$$

- the model is clearly an approximation;
- further factors characterizing the individuals;
- further factors characterizing the company;
- factors can be unobservable or difficult to obtain.

Problem: the model is purely deterministic, which makes it incompatible with the data.

Stochastic model: we model the relationship between wage and experience by

$$W = \beta_0 + \beta_1 \cdot E + u,$$

where u is a zero-mean random variable called disturbance/error term.

The disturbance may be due to:

- the fact that the relation between W and E is not linear;
- the fact that the coefficient may vary form one person to another;
- the omission of secondary variables;
- measurement errors on the variables W and E ...

$\rightsquigarrow$ the disturbance is interpreted as a summary of various kinds of ignorance. It can be used to construct measures of error associated with the model.

Data

```
salary  = 1999 salary + bonuses
totcomp = 1999 CEO total compensation
tenure  = # of years as CEO (=0 if less than 6 months)
age     = age of CEO
sales   = total 1998 sales revenue of firm i
profits = 1998 profits for firm i
assets  = total assets of firm i in 1998
```

1999 salary, total compensation, tenure, and age were collected from Forbes' 1999 list of Corporate America's Most Powerful People

1998 sales, profits, and assets were collected from Fortune Magazine's 1999 Fortune 500 list

Source: Wooldridge, J.M., Introductory Econometrics

```
The WAGES1 file contains 3294 USA working individuals with the
following information for 1987. The data are taken from the
National Longitudinal Survey.

exper   = experience in years
male    = 1 if male, 0 otherwise
school  = years of schooling
wage    = wage (in 1980 $) per hour
```

Source: Verbeek, M., A Guide to Modern Econometrics

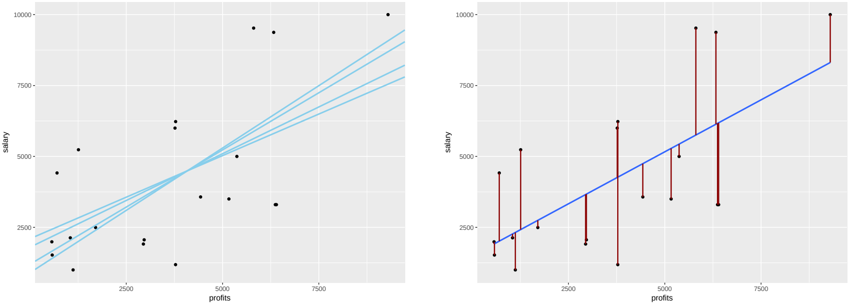
Setup of the simple LR

We postulate a linear relationship between y and a single explanatory variable X (or a linear impact of x on y).

$$y_i = \beta_0 + \beta_1 x_i + u_i, \qquad i = 1, ..., N$$

where u_i's are error terms and realizations of scalar RV's.

Note: The parameters β_0 (intercept) and β_1 (slope) are unknown and must be estimated



Ordinary Least Squares (OLS)

The unknown parameters should minimize the

$$\sum_{i=1}^{N} u_i^2 = \sum_{i=1}^{N} [y_i - (\beta_0 + \beta_1 x_i)]^2 \longrightarrow \min \quad \text{w.r.t.} \quad \beta_0, \beta_1!$$

Partial differentiation leads to:

$$\begin{aligned}\hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2} = \frac{\sum_{i=1}^N x_i y_i - N\bar{x}\bar{y}}{\sum_{i=1}^N x_i^2 - N\bar{x}^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}\end{aligned}$$

Example: CEO salary vs. profits

$$\hat{y} = 1497.9441 + 0.7329 \cdot X$$

Alternative approaches: least absolute deviation, total least squares, etc.

Setup of the multiple LR ($K > 1$)

We postulate a linear relationship between y and x's (or a linear impact of x's on y).

$$\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \cdots + \beta_K x_{1K} + u_1 \\
y_2 &= \beta_0 + \beta_1 x_{21} + \cdots + \beta_K x_{2K} + u_2 \\
&\vdots \\
y_N &= \beta_0 + \beta_1 x_{N1} + \cdots + \beta_K x_{NK} + u_N
\end{aligned}$$

where u_i are scalar RV's.

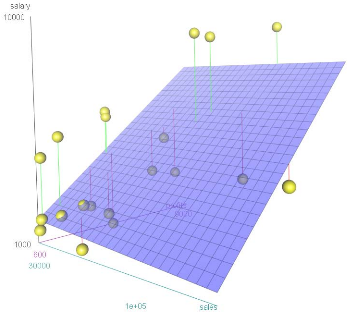
In matrix form:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u},$$

where $\boldsymbol{y} = (y_1, \ldots, y_N)'$, $\boldsymbol{u} = (u_1, \ldots, u_N)'$ and

$$\boldsymbol{X} = (x_{ik})_{i=1,\ldots,N;k=1,\ldots,K} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1K} \\ 1 & x_{21} & \ldots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \ldots & x_{NK} \end{pmatrix}.$$

- $\boldsymbol{X}$ known $(N, K+1)$-matrix with $N \geq K+1$;
- $\boldsymbol{y}$ known N-dimensional vector of observations of y
- $\boldsymbol{\beta}$ unknown $K+1$-dimensional vector of parameters/coefficients
- $\boldsymbol{u}$ a N-dimensional vector of error terms



Ordinary Least Squares (OLS)

The unknown parameters should minimize the

$$SSR(\boldsymbol{\beta}) = \boldsymbol{u}'\boldsymbol{u} = \sum_{i=1}^{N} u_i^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \longrightarrow \min \quad \text{w.r.t.} \quad \boldsymbol{\beta}!$$

We minimize the objective function by taking the derivative w.r.t. $\boldsymbol{\beta}$

$$\begin{aligned}\frac{\partial SSR(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} &= \frac{\partial}{\partial\boldsymbol{\beta}}\left(\boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}\right)\\ &= -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \overset{!}{=} \mathbf{0}\end{aligned}$$

If $rg(\boldsymbol{X}) = K+1$ then $\boldsymbol{X}'\boldsymbol{X}$ is positive definite and therefore invertible.

OLS estimator

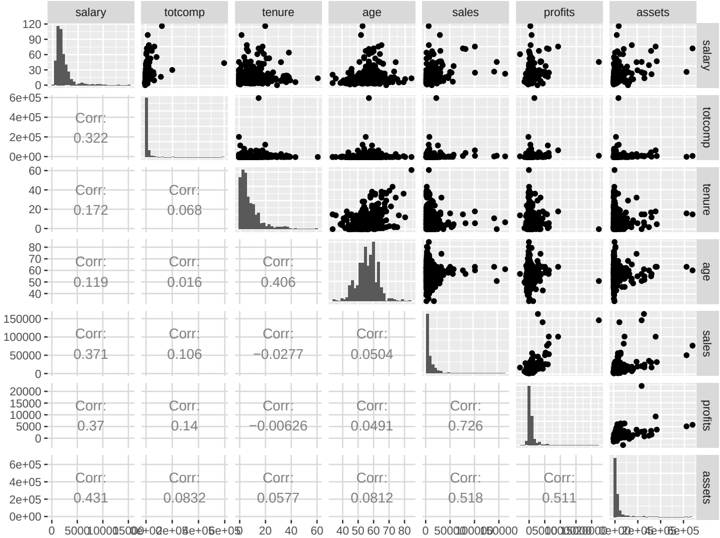
$$\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \qquad \left(\ \hat{\sigma}^2 = \frac{1}{N-K-1}\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}\ \right).$$

The vector of fitted values $\hat{\boldsymbol{y}}$ and the residuals $\hat{\boldsymbol{u}}$ are given by

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{P}\boldsymbol{y} \quad \text{and} \quad \hat{\boldsymbol{u}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = (\boldsymbol{I}-\boldsymbol{P})\boldsymbol{y},$$

where $\boldsymbol{P}$ is the projection matrix.

Example: full sample (447 observations)



```
> lm(salary ~ ., data = ceo)
Coefficients:
 (Intercept)       totcomp        tenure           age         sales       profits        assets
857.53755940    0.01430199   27.40055298    7.03434851    0.01397754    0.10356587    0.00756664
```

Note: the size of the coefficients cannot be used to decide about the importance of the variables

- Standardize all regressors (not binary/cathegoric) and the regressand:
$$x_{ik}^* = \frac{x_{ik} - \bar{x}_k}{s_k}.$$
- Standardized coefficients:
$$\hat{\beta}_{k,stand} = \hat{\beta}_k \cdot \frac{\text{standard deviation of } X_k}{\text{standard deviation of } y}$$

```
Standardized Coefficients:
(Intercept)     totcomp      tenure         age       sales     profits      assets
 0.00000000  0.26213182  0.13117910  0.02779590  0.13119609  0.09274202  0.28402494
```

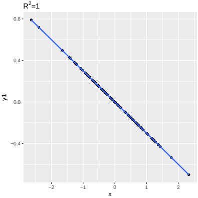
Goodness-of-fit

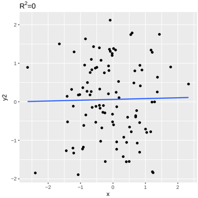
$$\begin{aligned} TSS &= RSS + ESS, \\ RSS &= \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \quad \text{residual sum of squares} \\ ESS &= \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2 \quad \text{explained sum of squares} \\ TSS &= \sum_{i=1}^{N} (y_i - \bar{y})^2 \quad \text{total sum of squares} \end{aligned}$$

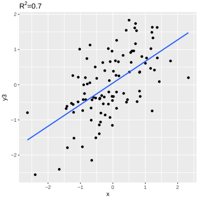
Def: The coefficient of determination R^2 is defined by

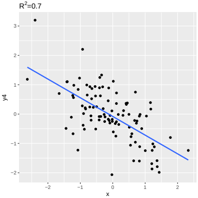
$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}$$

1. $0 \leq R^2 \leq 1$
2. $R^2 = 1 \Leftrightarrow RSS = 0 \Leftrightarrow y_i = \hat{y}_i$ for all i perfect fit
3. $R^2 = 0 \Leftrightarrow ESS = 0 \Leftrightarrow \hat{y}_i = \bar{y}$ for all i model explains nothing
4. For $K = 1$: we have $R^2 = r_{xy}^2$
5. For $K > 1$: R^2 equals the squared correlation between the observed values and the fitted values, i.e. $R^2 = r_{y,\hat{y}}^2$









Disadvantage 1: R^2 increases if K increases.

Thus R^2 cannot be used to select one model from a sequence of "nested" models.

Therefore: adjusted R^2

$$R^2_{adj} = 1 - \frac{RSS/(N-K-1)}{TSS/(N-1)}$$

Thus $-\frac{K}{N-K-1} \leq R^2_{adj} \leq 1$.

Disadvantage 2: If the model contains no constant, then R^2 can take arbitrary values and can also be negative.

Statistical properties of OLS

Note: OLS estimation is only an algebraic descriptive technique. To make inferences we need statistical assumptions: $\boldsymbol{u}$ is a random vector !

4. Mean of the residuals: $E(\boldsymbol{u}) = \boldsymbol{0}$
5. Homoscedasticity: the variance is the same for all residuals, i.e. $Var(u_i) = E(u_i^2) = \sigma_u^2$ with $0 < \sigma^2 < \infty$ for all $i = 1, \ldots, N$ $(Var(u_i|\boldsymbol{X}) = \sigma^2$ for stochastic regressors)
6. No autocorrelation: $Corr(u_i, u_j) = E(u_i u_j) = 0$ for $i \neq j$ $(E(u_i u_j|\boldsymbol{X}) = 0$ for stochastic regressors)
7. Normal distribution: $\boldsymbol{u} \sim \mathcal{N}_N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$

Since

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{u}) \\ &= \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}\end{aligned}$$

it follows that

$$\begin{aligned}E(\hat{\boldsymbol{\beta}}) &= \boldsymbol{\beta} + E((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}) \\ &= \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{u}) = \boldsymbol{\beta}\end{aligned}$$

Note: if $\boldsymbol{X}$ is stochastic then $E(\boldsymbol{u}|\boldsymbol{X}) = \mathbf{0}$ is usually appropriate for cross-sectional data, but not always for time-series data

$\leadsto$ OLS can be biased !!!

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}) &= E[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'] = E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}\boldsymbol{u}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}] \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{u}\boldsymbol{u}')\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \end{aligned}$$

```
> Zceo3= lm(salary  ~ profits +  sales, data=ceo)
> Zceo3$coefficients
  (Intercept)       profits         sales
 1.646503e+03  7.723880e-01 -5.239456e-03
> vcov(Zceo)
              (Intercept)      profits         sales
(Intercept) 1254680.79953 -51.75438711 -1.448409e+01
profits         -51.75439   0.07280971 -3.847250e-03
sales           -14.48409  -0.00384725  5.108337e-04
```

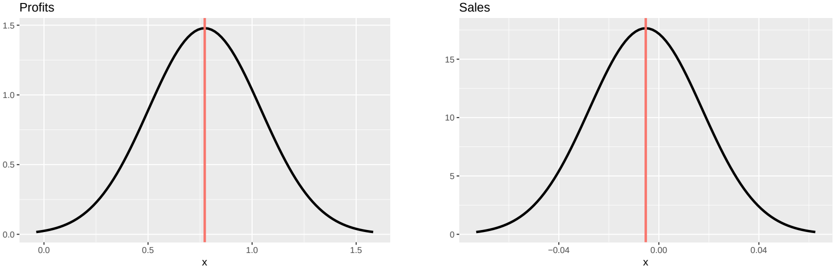
Theorem (Gauss-Markov)

The OLS estimator $\hat{\boldsymbol{\beta}}$ is more efficient than any other linear unbiased estimator $\tilde{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{y}$, i.e. it has the smallest possible variance among all unbiased estimators, which are linear in $\boldsymbol{y}$.

Assuming normally distributed residuals and deterministic regressors we obtain

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}_{K+1}\big(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\big)$$

Note: every parameter follows univariate normal distribution!



Testing the coefficients

Here: we test if the coefficients deviate from a given value. $\rightsquigarrow$ t-test

$$H_0 \ : \ \beta_j = \beta_j^* \quad \text{vs} \quad H_1 \ : \ \beta_j \neq \beta_j^*$$

Special case: $\beta_j^* = 0$

$$H_0 \ : \ \beta_j = 0 \quad \text{vs} \quad H_1 \ : \ \beta_j \neq 0$$

Using the t-test with $\beta_j = 0$ we test if X_j has a significant impact on y.

Statistics II:

$$H_0 \ : \ \mu = \mu_0 \quad \text{vs} \quad H_1 \ : \ \mu \neq \mu_0$$

- $\hat{\mu} = \bar{X}$ is an estimator of μ
- $\hat{\mu} \sim N(\mu, \sigma^2)$ for Gaussian samples
- Test statistics $t = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}/\sqrt{N}} \sim t_{N-1}$
- Use either the rejection area or the p-values

The test for β's follows the same path!

t-statistics

$$t_{emp} = \frac{\hat{\beta}_j - \beta_j^*}{\sqrt{\widehat{Var(\hat{\beta}_j)}}},$$

where $\widehat{Var(\hat{\beta}_j)}$ is the j-th element on the main diagonal of $\hat{\sigma}^2(\boldsymbol{X'X})^{-1}$.

Under H_0 and under the assumption of Gaussian residuals

$$t_{emp} \sim t_{N-K-1},$$

where t_n is the t distribution with n degrees of freedom.

Note: usually we use the p-value approach to run the test.

Note: if a coefficient ist insignificant, you may drop, but the remaining coefficients should be reestimated!

Example:

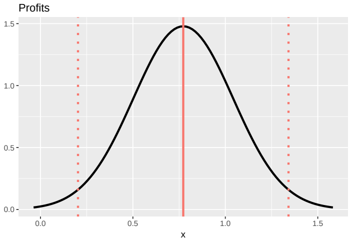
```
> summary(Zceo3)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.647e+03  1.120e+03   1.470   0.1598
profits      7.724e-01  2.698e-01   2.862   0.0108 *
sales       -5.239e-03  2.260e-02  -0.232   0.8194
---
Residual standard error: 2253 on 17 degrees of freedom
Multiple R-squared:   0.42,Adjusted R-squared:  0.3517
F-statistic: 6.154 on 2 and 17 DF,  p-value: 0.009758
```

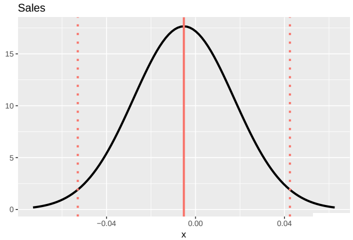
Confidence intervals for the regression coefficients

The unknown parameter β_j belongs with the probability of $(1-\alpha)$ to the interval

$$\left[\hat{\beta}_j - t_{N-K-1;1-\alpha/2} \cdot \sqrt{\widehat{Var(\hat{\beta}_j)}};\;\; \hat{\beta}_j + t_{N-K-1;1-\alpha/2} \cdot \sqrt{\widehat{Var(\hat{\beta}_j)}}\right]$$

```
> confint(Zceo)
                    2.5 %       97.5 %
(Intercept) -716.75514306 4.009761e+03
profits        0.20309069 1.341685e+00
sales         -0.05292473 4.244582e-02
```





- ***F*-test**: tests the joint explanatory power of all regressors

$$\begin{aligned} H_0 &: \quad \beta_1 = \cdots = \beta_K = 0 \\ H_1 &: \quad \beta_j \neq 0 \quad \text{for at least one parameter} \end{aligned}$$

Under H_0 all regressors have no impact on the dependent variable.

F-statistics

$$\begin{aligned} F_{emp} &= \frac{\text{explained variation}/K}{\text{unexplained variation}/(N-K-1)} \\ &= \frac{R^2/K}{(1-R^2)/(N-K-1)} \end{aligned}$$

Under H_0 and with Gaussian residuals it holds $F_{emp} \sim F_{K;N-K-1}$.

```
F-statistic: 6.154 on 2 and 17 DF,  p-value: 0.009758
```

Forecasting/prediction

Let $\hat{\boldsymbol{\beta}}$ be the OLS-estimator and $\boldsymbol{x}_0$ be a new vector of observations. We are interested in the forecast for y_0 and in the variance of the forecast error.

The forecast and the forecast error are given by

$$\hat{y}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}} \text{ and } e_0 = y_0 - \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}.$$

Moreover

$$\begin{aligned} E(e_0) &= E(\boldsymbol{x}_0'\boldsymbol{\beta} + u_0 - \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}) = 0 \\ Var(e_0) &= E(e_0^2) - [E(e_0)]^2 = E(\boldsymbol{x}_0'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + u_0)^2 \\ &= E(u_0^2) + E(\boldsymbol{x}_0'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}))^2 \\ &= \sigma^2 + \boldsymbol{x}_0' E[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})']\boldsymbol{x}_0 \\ &= \sigma^2(1 + \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0) \end{aligned}$$

Note: forecast errors here are not the same as $(\hat{u}_i)$!

Assuming Gaussian residuals the prediction interval for y_0 is given by:

$$\left[\boldsymbol{x}_0'\hat{\boldsymbol{\beta}} - t_{N-K-1;1-\alpha/2}\hat{\sigma}\sqrt{1+\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0};\; \boldsymbol{x}_0'\hat{\boldsymbol{\beta}} + t_{N-K-1;1-\alpha/2}\hat{\sigma}\sqrt{1+\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}\right]$$

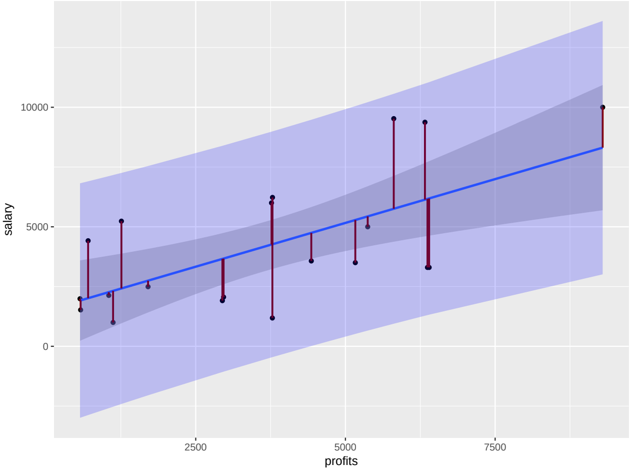
Note: Frequently the objective is to forecast $E(y_0|\boldsymbol{x}_0)$ and not y_0

$$Var(e_0^*) = E(\boldsymbol{x}_0'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}))^2 = \sigma^2\,\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0$$

These leads to confidence interval for $E(y_0|\boldsymbol{x}_0)$ is given by:

$$\left[\boldsymbol{x}_0'\hat{\boldsymbol{\beta}} - t_{N-K-1;1-\alpha/2}\hat{\sigma}\sqrt{\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0};\; \boldsymbol{x}_0'\hat{\boldsymbol{\beta}} + t_{N-K-1;1-\alpha/2}\hat{\sigma}\sqrt{\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}\right]$$

```
> predict.lm(Zceo2, data.frame(profits=5000), interval = "prediction")
       fit     lwr      upr
1 5162.584 406.782 9918.387
> predict.lm(Zceo2, data.frame(profits=5000), interval = "confidence")
       fit      lwr      upr
1 5162.584 3985.205 6339.963
```



Asymptotic properties I

Note: if the residuals are not normal, then the finite sample distribution of $\hat{\boldsymbol{\beta}}$ is unknown $\rightsquigarrow$ asymptotic properties

$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2\left(\sum_{i=1}^{N}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}.$$

- As N increases the matrix $\boldsymbol{X}'\boldsymbol{X}$ becomes larger and the inverse becomes smaller $(\boldsymbol{X}'\boldsymbol{X})^{-1} \to \mathbf{0}$.
- Assumption (A): $\lim_{N\to\infty}\frac{1}{N}\boldsymbol{X}'\boldsymbol{X} = \mathbf{Q}$ and $\mathbf{Q}$ is positive definite

Asymptotic properties II

- From Chebysheff inequality under (A)

$$P(|\hat{\beta}_j - \beta_j| > c) \leq \frac{Var(\hat{\beta}_j)}{c^2} \longrightarrow 0, \text{ as } N \to \infty$$

  thus the OLS estimator is a consistent estimator of $\boldsymbol{\beta}$.

- From the CLT the OLS-estimator is asymptotically normal:

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{a}{\sim} N(\mathbf{0}, \sigma^2 \mathbf{Q}^{-1}).$$

Thus:

- all the above tests and CIs are approx. valid in large samples (?).
- The normality of the residuals can be verified using goodness-of-fit tests.

Asymptotic properties III

- One can replace the t-quantiles with the normal quantiles.

```
  > resid = Zceo3$residuals
  > ks.test(resid,"pnorm", mean=mean(resid), sd = sqrt(var(resid)))
One-sample Kolmogorov-Smirnov test
data:  resid
D = 0.1348, p-value = 0.8141
alternative hypothesis: two-sided
```

Chapter 5

Linear Regression: further issues

Interpretation

$$E(y_i|\boldsymbol{x}_i) \quad = \quad E(\boldsymbol{x}_i'\boldsymbol{\beta} + u_i) = \boldsymbol{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK}$$

From the linearity of the model:

$$\frac{\partial E(y_i|\boldsymbol{x}_i)}{\partial x_{ij}} = \beta_j$$

Interpretation: β_j measures the expected change in y_i if x_{ij} changes by one unit and other x's DO NOT change (*ceteris paribus*)

Alternatively:

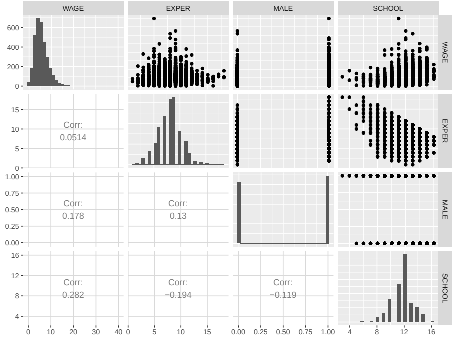
$$\begin{aligned} & E(y_i^*|x_{ij}+1) - E(y_i|x_{ij}) \\ = \quad & \beta_0 + \beta_1 x_{i1} + \beta_{j-1} x_{i,j-1} + \beta_j(x_{ij}+1) + \beta_{j+1} x_{i,j+1} + \cdots + \beta_K x_{iK} \\ - \quad & [\beta_0 + \beta_1 x_{i1} + \beta_{j-1} x_{i,j-1} + \beta_j x_{ij} + \beta_{j+1} x_{i,j+1} + \cdots + \beta_K x_{iK}] = \beta_j \end{aligned}$$

```
lm(formula = WAGE ~ ., data = wage)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.38002    0.46498  -7.269 4.50e-13 ***
EXPER        0.12483    0.02376   5.253 1.59e-07 ***
MALE         1.34437    0.10768  12.485  < 2e-16 ***
SCHOOL       0.63880    0.03280  19.478  < 2e-16 ***
Residual standard error: 3.046 on 3290 degrees of freedom
Multiple R-squared:  0.1326,Adjusted R-squared:  0.1318
F-statistic: 167.6 on 3 and 3290 DF,  p-value: < 2.2e-16
```

$$\text{WAGE}_i = -3.38 + 0.12 \cdot \text{EXPER}_i + 1.34 \cdot \text{MALE}_i + 0.64 \cdot \text{SCHOOL}_i$$

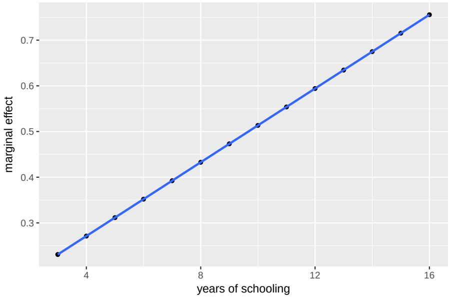
$\rightsquigarrow$ an additional year of schooling increases the expected hourly wage by 0.63880 (same gender, same experience!)

Note 1: frequently we use a quadratic (polynomial) regression



$$
\begin{aligned}
\mathrm{WAGE}_i &= \beta_0 + \beta_1 \mathrm{SCHOOL}_i + \beta_3 \mathrm{SCHOOL}_i^2 + \ldots \\
\frac{\partial E(y_i|\boldsymbol{x}_i)}{\partial \mathrm{SCHOOL}_i} &= \beta_1 + 2\beta_3 \mathrm{SCHOOL}_i
\end{aligned}
$$

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.83892    1.51609   1.213 0.225242
EXPER        0.10973    0.02408   4.556  5.4e-06 ***
MALE         1.33616    0.10750  12.429  < 2e-16 ***
SCHOOL      -0.27814    0.25568  -1.088 0.276751
SCHOOL2      0.04037    0.01116   3.616 0.000304 ***
Residual standard error: 3.041 on 3289 degrees of freedom
Multiple R-squared:  0.136,Adjusted R-squared:  0.135
F-statistic: 129.5 on 4 and 3289 DF,  p-value: < 2.2e-16
```



Note 2: dummy variables

$$\text{MALE}_i = \begin{cases} 1, & \text{if the } i\text{-th person is a male} \\ 0, & \text{if the } i\text{-th person is a female} \end{cases}$$

$$\text{WAGE}_i = \beta_0 \mathbb{1} + \beta_1 \text{EXPER}_i + \beta_2 \text{MALE}_i + \ldots$$

$\rightsquigarrow$ β_2 is wage difference between males and females (*CP*!)

Note 3: Interactions

$$\begin{aligned} \text{WAGE}_i &= \beta_0 + \beta_1 \text{SCHOOL}_i + \beta_2 \text{EXPER}_i + \beta_3 \text{SCHOOL}_i \cdot \text{EXPER}_i + \ldots \\ \frac{\partial E(y_i|\boldsymbol{x}_i)}{\partial \text{EXPER}_i} &= \beta_2 + \beta_3 \text{SCHOOL}_i \end{aligned}$$

$\rightsquigarrow$ the change in the wage due to an additional year of experience depends on the years of schooling

Note 4: regression in log's

$$\begin{aligned} \log y_i &= \beta_0 + \beta_1 \log x_{i1} + u_i \\ \frac{\partial E(y_i|x_{i1})}{\partial x_{i1}} \cdot \frac{x_{i1}}{E(y_i|x_{i1})} &\approx \frac{\partial E(\log y_i|\log x_{i1})}{\partial \log x_{i1}} = \beta_1 \end{aligned}$$

$\rightsquigarrow$ β_1 is the percentage change in y_i induces by a 1% change in x_{i1}

Note 5: regression in log's with dummies

$$\begin{aligned} \log y_i &= \beta_0 + \beta_1 x_{i1} + u_i \\ \frac{\partial E(y_i|x_{i1})}{\partial x_{i1}} \cdot \frac{1}{E(y_i|x_{i1})} &\approx \frac{\partial E(\log y_i|\log x_{i1})}{\partial x_{i1}} = \beta_1 \end{aligned}$$

$\rightsquigarrow$ β_1 is the relative change in y_i induces by a switch in x_{i1} from 0 to 1

Dummy variables

Def: A dummy variable takes the value one for some observations to indicate the presence of an effect or membership in a group and zero for the remaining observations.

But:

$$\text{WAGE}_i = \beta_0 + \beta_1 \text{EXPER}_i + \beta_2 \text{MALE}_i + \beta_3 \text{FEMALE}_i + u_i$$

where

$$\text{MALE}_i = \begin{cases} 1 & \text{person } i \text{ is a male} \\ 0 & \text{else} \end{cases}$$

$$\text{FEMALE}_i = \begin{cases} 1 & \text{person } i \text{ is a female} \\ 0 & \text{else} \end{cases}$$

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 & 1 & 0 \\ 1 & x_2 & 1 & 0 \\ 1 & x_3 & 0 & 1 \\ 1 & x_4 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

Note that

$$\boldsymbol{x}_1 = \boldsymbol{x}_3 + \boldsymbol{x}_4.$$

Thus the columns of $\boldsymbol{X}$ are perfectly colinear and $rg(\boldsymbol{X}) < K$.

- Drop one gender dummy (for example, the 2nd)

$$\begin{aligned} E(\text{WAGE}_i|\text{ male }) &= \beta_0 + \beta_1 \text{EXPER}_i + \beta_2 \\ E(\text{WAGE}_i|\text{ female}) &= \beta_0 + \beta_1 \text{EXPER}_i \end{aligned}$$

Example

$$\begin{aligned} E(\text{WAGE}_i|\text{ male }) &= 1.83892 + 0.10973 \cdot \text{EXPER}_i + 1.33616 + \dots \\ E(\text{WAGE}_i|\text{ female }) &= 1.83892 + 0.10973 \cdot \text{EXPER}_i + \dots \end{aligned}$$

- Drop the constant

$$\begin{aligned} E(\text{WAGE}_i|\text{ male}) &= \beta_1 \text{EXPER}_i + \beta_2 \\ E(\text{WAGE}_i|\text{ female}) &= \beta_1 \text{EXPER}_i + \beta_3 \end{aligned}$$

Now: Dummies for the slope

Q: is the impact of EXPER is different for males and females?

$$\begin{aligned} \mathrm{WAGE}_i &= \beta_0 + \beta_1 \mathrm{EXPER}_i + \beta_2 \mathrm{MALE}_i \cdot \mathrm{EXPER}_i + \beta_3 \mathrm{MALE}_i + \ldots + u_i \\ E(\mathrm{WAGE}_i |\text{ male }) &= \beta_0 + (\beta_1 + \beta_2)\mathrm{EXPER}_i + \beta_3 \\ E(\mathrm{WAGE}_i |\text{ female }) &= \beta_0 + \beta_1 \mathrm{EXPER}_i \end{aligned}$$

Example

```
> Z.wage = lm(WAGE ~ EXPER+MALE+SCHOOL+EXPER*MALE,  data=wage);
> summary(Z.wage)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.50145    0.49164  -7.122 1.30e-12 ***
EXPER        0.14484    0.03545   4.085 4.51e-05 ***
MALE         1.63186    0.39296   4.153 3.37e-05 ***
SCHOOL       0.63598    0.03301  19.269  < 2e-16 ***
EXPER:MALE  -0.03609    0.04744  -0.761    0.447
---
Residual standard error: 3.046 on 3289 degrees of freedom
Multiple R-squared:  0.1327,Adjusted R-squared:  0.1317
F-statistic: 125.9 on 4 and 3289 DF,  p-value: < 2.2e-16
```

Multicollinearity

Multicollinearity arises if there is strong correlation or dependence among the regressors.

Consequences of multicollinearity

- difficult to separate effects; problematic interpretation (CP):

$$\text{WAGE}_i = \beta_0 + \beta_1 \text{EXPER}_i + \beta_2 \text{AGE}_i + \ldots$$

- large variance of the OLS estimators induce insignificant coefficients, BUT R^2 is still high:

  ! high correlation between the columns of $\boldsymbol{X}$

  $\rightsquigarrow$ $\boldsymbol{X}'\boldsymbol{X}$ is close to being singular

  $\rightsquigarrow$ $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ consists of high values

$$\begin{pmatrix} 1 & 0.99 \\ 0.99 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 50.25126 & -49.74874 \\ -49.74874 & 50.25126 \end{pmatrix}$$

  $\rightsquigarrow$ $Var(\hat{\boldsymbol{\beta}}) = \sigma(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is a large matrix

Methods of detecting multicollinearity

- Simple correlation among regressors.
  If correlation is greater than 0.8 or greater than R^2, then MC is serious. Disadvantage: no insight into the complex interrelationships between the regressors.
- Variance inflation factor.

$$VIF_j = \frac{1}{1-R_j^2},$$

where R_j^2 is the coefficient of determination of the auxiliary regression of X_j on the remaining regressors, i.e.:

$$X_j = a_0 + a_1 X_1 + \cdots + a_{j-1}X_{j-1} + a_{j+1}X_{j+1} + \cdots + a_K X_K + v.$$

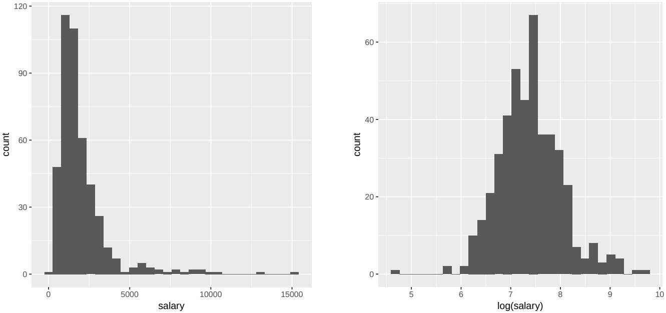
If $VIF_j \geq 5$ ($R_j^2 \geq 0.8$), then the contribution of the regressor to the multicollinearity is substantial.

```
Z = lm(salary ~ ., data=ceo)
> vif(Z)
 totcomp   tenure      age    sales  profits   assets
1.025532 1.211724 1.204613 2.262765 2.248375 1.454040
```

Specifying the Functional Form of the Regression Function

Note 1: y in logs

- If $y > 0$, then $\hat{y}$ might be negative
- Is y heavily skewed, so is $u \rightsquigarrow$ non-normal



$$ln(y_i) = z_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK} + u_i.$$

Using LS approach we estimate the parameters and obtain for $(x_{01}, \ldots, x_{0K})$ the forecasts $\hat{z}_0$.

But: it is in general wrong to forecast y_0 by $\hat{y}_0 = e^{\hat{z}_0}$! It holds

$$E(\hat{Z}_0 | x_{01}, \ldots, x_{0K}) = z_0$$

but

$$E(e^{\hat{z}_0} | x_{01}, \ldots, x_{0K}) \neq e^{z_0} = y_0$$

Thus the forecasts are biased.

If $Z \sim N(\mu, \sigma^2)$, then

$$E(e^Z) = e^{\mu + \frac{1}{2}\sigma^2}.$$

Thus if the residuals are Gaussian, then the following forecasts are optimal:

$$\hat{y}_0^{(opt)} = e^{\hat{z}_0 + \frac{1}{2}\widehat{Var(\hat{z}_0)}} = e^{\hat{z}_0 + \frac{1}{2}\widehat{Var(\hat{\varepsilon}_0)}}$$

Note:

- Compute both forecasts and choose the method with a better fit.
- The optimal forecasts depend on the type of transformation.

Note 2: transforming X's

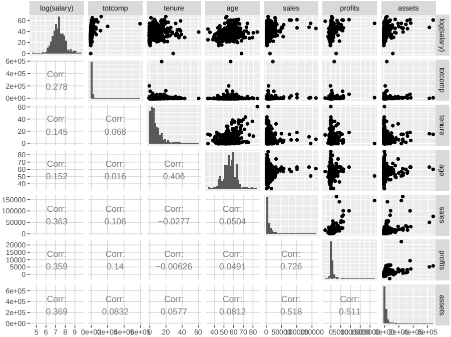
- a linear regression is linear in β_j's, but not necessarily in X_j's

$$y_i = \beta_0 + \beta_1 g_1(x_{i1}) + \beta_2 g_2(x_{i2}) + \cdots + \beta_K g_K(x_{iK}) + u_i$$

- Plot the scatter-plots and pick the most appropriate transformation

$$g(x) = \log(x), \quad 1/x, \quad e^x, \quad \sqrt{x}, \quad x + x^2$$

- The interpretation above is not valid any more ....



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.575e+02  5.963e+02   1.438  0.15111
totcomp     1.430e-02  2.179e-03   6.564 1.47e-10 ***
tenure      2.740e+01  9.067e+00   3.022  0.00266 **
age         7.034e+00  1.095e+01   0.642  0.52105
sales       1.398e-02  6.320e-03   2.212  0.02749 *
profits     1.036e-01  6.603e-02   1.569  0.11748
assets      7.567e-03  1.267e-03   5.973 4.80e-09 ***
Residual standard error: 1434 on 440 degrees of freedom
Multiple R-squared:  0.3158,Adjusted R-squared:  0.3065
F-statistic: 33.85 on 6 and 440 DF,  p-value: < 2.2e-16
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.592e+00  1.046e+00   3.435 0.000650 ***
log.totcomp   3.431e-01  2.074e-02  16.538  < 2e-16 ***
rec.tenure   -2.221e-01  9.067e-02  -2.450 0.014686 *
age           4.421e-03  2.840e-02   0.156 0.876377
age2         -3.779e-06  2.492e-04  -0.015 0.987906
log.sales     1.213e-01  3.510e-02   3.454 0.000606 ***
log.profit.p -4.542e-02  9.391e-02  -0.484 0.628857
sqrt.assets   8.783e-04  2.465e-04   3.562 0.000408 ***
Residual standard error: 0.4264 on 439 degrees of freedo
Multiple R-squared:  0.5594,Adjusted R-squared:  0.5524
F-statistic: 79.62 on 7 and 439 DF,  p-value: < 2.2e-16
```

Note 3: testing non-linearity using RESET tests

If the true model is truly nonlinear or contains other transformations of regressors, this should be captured by the auxiliary regression:

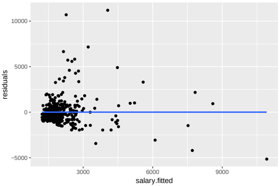
$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \alpha_2\hat{y}_i^2 + \alpha_3\hat{y}_i^3 + ... + \alpha_q\hat{y}_i^q + v_i$$

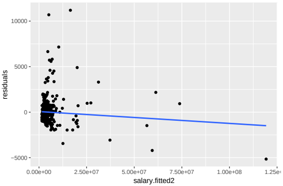
If

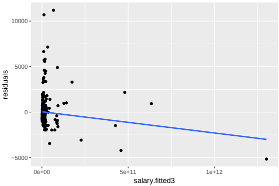
$$H_0 : \alpha_2 = \cdots = \alpha_q = 0$$

is rejected, then there is a potential misspecification of the functional form.

```
> library("lmtest")
> resettest(salary ~ .,  data=ceo, power = 2:3, type = c("fitted"))
RESET test
data:  salary ~ .
RESET = 23.088, df1 = 2, df2 = 438, p-value = 2.933e-10
```







Selecting Regressors

Selecting regressors is usually based on comparing different models. We have to distinguish between two

- ... nested models: same y and an extended set of regressors
- ... non-nested models: potentially different y's and different regressors

Here: nested models

Building the right nested model can follow one of two strategies:

- simple-to-general: start with a small model and add sequentially new significant variables
- general-to-simple: start with a large model and drop sequentially insignificant variables

The discussion on the next slides on omission of relevant and inclusion of irrelevant variables implies that the 2nd strategy should be preferred!

Omission of relevant regressors

Suppose that the correct model is

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}.$$

The model we estimate is misspecified and given by

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{u}.$$

OLS estimation leads to

$$\begin{aligned}\hat{\boldsymbol{\beta}}_1 &= (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y} \\ &= (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^{-1}\left(\boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}\right) \\ &= \boldsymbol{\beta}_1 + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\boldsymbol{\beta}_2 + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{u}\end{aligned}$$

Unless $\boldsymbol{X}_1'\boldsymbol{X}_2 = \boldsymbol{0}$ or $\boldsymbol{\beta}_2 = \boldsymbol{0}$ the estimator $\hat{\boldsymbol{\beta}}_1$ is biased.

$$E(\hat{\boldsymbol{\beta}}_1) = \boldsymbol{\beta}_1 + \boldsymbol{P}_{1.2}\boldsymbol{\beta}_2 \quad \text{with} \quad \boldsymbol{P}_{1.2} = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2.$$

Let $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ consist each of a single column. Then

$$Var(\hat{\boldsymbol{\beta}}_1|\boldsymbol{X}) = \frac{\sigma^2}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2}$$

$$Var(\hat{\boldsymbol{\beta}}_{1.2}|\boldsymbol{X}) = \frac{\sigma^2}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2} \cdot \frac{1}{1 - r_{12}^2},$$

where r_{12} is the correlation coefficient between the columns of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$.

$\rightsquigarrow$ dilemma: if r_{12} is high, should we use a larger model, with potentially insignificant but unbiased parameters, or a smaller model with biased but significant parameters?

Inclusion of irrelevant regressors

Assume that the true regression is given by

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{u},$$

but we estimate

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}.$$

The second model still leads to correct results. It simply fails to incorporate the restriction $\boldsymbol{\beta}_2 = \boldsymbol{0}$. Both $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are unbiased.

$$E(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{0} \end{pmatrix}$$

$$E\left(\frac{\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}}{N - k_1 - k_2}|\boldsymbol{X}\right) = \sigma^2.$$

⇝ model building - general to simple (!!!)

Stepwise model selection

Note:

- Due to multicollinearity it is nor optimal simply to drop insignificant variables from the model.
- There are $2^J - 1$ different models for given J regressors $\rightsquigarrow$ estimate unrealistically many model and pick up the model with the highest R^2_{korr}.
  Example: $2^{15} - 1 = 32767$
- We need a procedure to assess the significance of the difference between models.

- FORWARD selection simple-to-general:
  - Choose a single variable, which the best fit.
  - The next variables are added sequentially; in each step take the variable with the highest and significant contribution to the model fit.
  - The optimal model is achieved, when non of the remaining variables contributes significantly to the goodness of the model.

  Note: A variable, which was added at a previous step, cannot be dropped at further stages.
- BACKWARD selection - general-to-simple
  - Estimate the model with all available variables.
  - The variables are dropped sequentially; at each step drop the variable with the smallest and insignificant contribution to the model
  - The optimal model is achieved, when all remaining variables have significant contribution.

  Note: A variable, which was dropped, cannot be added to the model at later steps.
- STEPWISE selection:
  - The variables can be added or dropped from the model sequentially.

M	a model with K_1 variables
M_{in}	a model with an additional variable
M_{out}	a model with one variable less

Question I: Does the additional variable lead to a significant improvement of the model?

Question II: Does the elimination of one variable lead to an insignificant loss in the goodness of the model?

M vs. M_{in}: FORWARD

H_0 : both models are equivalent vs. H_1 : model M_{in} is better
$H_0 : R^2_M = R^2_{M_{in}}$ vs. $H_1 : R^2_{M_{in}} > R^2_M$

Formally we test the significance of the difference between the R^2's or between F-test statistics.

$$F - to - enter = \frac{R^2_{M_{in}} - R^2_M}{1 - R^2_{M_{in}}} \cdot (N - K_1 - 1) \sim F_{1;N-K_1-1}$$

M vs. M_{out}: BACKWARD

H_0 : both models are equivalent vs. H_1 : model M_{out} is worse
$H_0 : R^2_M = R^2_{M_{out}}$ vs. $H_1 : R^2_{M_{out}} < R^2_M$

$$F - to - remove = \frac{R^2_M - R^2_{M_{out}}}{1 - R^2_M} \cdot (N - K_1) \sim F_{1;N-K_1}$$

Further goodness-of-fit measures

Additionally to R^2_{adj} one frequently uses the following information criteria (IC)

- Akaike Information Criterion

$$AIC = \ln(\hat{\sigma}^2) + \frac{2K}{N}$$

- Bayes Information Criterion

$$BIC = \ln(\hat{\sigma}^2) + \frac{K}{N}\ln(N)$$

Note:

- The ICs find the trade-off between the number of regressors and the variance of residuals
- An IC is one possibility to choose the “best” model among nested models $\rightsquigarrow$ choose the specification with the lowest (!) IC

```
> Z.ceo = lm(salary~., data=ceo)
>   step(Z.ceo, direction ="backward")
Start:  AIC=6505.04
salary ~ totcomp + tenure + age + sales + profits + assets
          Df Sum of Sq       RSS    AIC
- age      1    848791 906269147 6503.5
<none>                 905420356 6505.0
- profits  1   5062573 910482928 6505.5
- sales    1  10066746 915487102 6508.0
- tenure   1  18793698 924214054 6512.2
- assets   1  73421542 978841897 6537.9
- totcomp  1  88670076 994090431 6544.8
Step:  AIC=6503.46
salary ~ totcomp + tenure + sales + profits + assets
          Df Sum of Sq       RSS    AIC
<none>                 906269147 6503.5
- profits  1   5085050 911354196 6504.0
- sales    1  10248755 916517902 6506.5
- tenure   1  26525466 932794612 6514.4
- assets   1  73994765 980263912 6536.5
- totcomp  1  88338982 994608129 6543.0
```

General Linear Hypothesis

Q: we add K^* more regressors to the model. Is the improvement in the goodness-of-fit statistically significant?

$$M_{small}: \; y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK} + u_i$$

$$M_{large}: \; y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK} + \ldots + \beta_{K+K^*} x_{i,K+K^*} + u_i$$

We are interested in the hypotheses:

$$H_0: \quad R^2_{small} = R^2_{large} \qquad \text{vs} \qquad H_1: \quad R^2_{small} < R^2_{large}$$

To test use:

$$F = \frac{R^2_{large} - R^2_{small}}{(1 - R^2_{large})/(N - K - K^* - 1)} \sim F_{1,N-K-K^*-1}$$

The above test is a special case of the General Linear Hypothesis
The general linear hypothesis is given by

$$H_0: \boldsymbol{R\beta} = \boldsymbol{r} \qquad \text{vs} \qquad H_1: \boldsymbol{R\beta} \neq \boldsymbol{r},$$

where $\boldsymbol{R}$ is a $q \times K+1$ matrix with $rg\,(\boldsymbol{R}) = q$. $\boldsymbol{R}$ and $\boldsymbol{r}$ are assumed to be known.

Examples:

a) $H_0 : \beta_i = 0$ then $\boldsymbol{R} = (0, .., 1, ..0)$ with 1 at the $i+1$-th position, $\boldsymbol{r} = 0$

b) $H_0 : \beta_i = \beta^*$ then $\boldsymbol{r} = \beta^*$ ($\boldsymbol{R}$ as in a)

c) $H_0 : \beta_i = \beta_j$ then $\boldsymbol{R} = (0, .., 1, .., -1, ..0)$ with 1 at the i–th position and -1 at the j–th position.

d) $H_0 : \beta_1 = ... = \beta_K = 0$ then $\boldsymbol{R} = (\boldsymbol{0}_K, \boldsymbol{I}_K)$, $\boldsymbol{r} = \boldsymbol{0}_K$

Suppose that the error terms are normally distributed, i.e. that $\boldsymbol{u} \sim N(0, \sigma^2 \boldsymbol{I})$. This implies

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}).$$

$$\boldsymbol{R}\hat{\boldsymbol{\beta}} \sim \mathcal{N}_q(\boldsymbol{R}\boldsymbol{\beta}, \sigma^2 \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')$$

$$\boldsymbol{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \mathcal{N}_q(0, \sigma^2 \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')$$

It follows under $H_0 : \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r}$

$$\frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})'(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})}{\sigma^2} \sim \chi^2_q. \qquad (*)$$

It can be shown that

$$(N - K - 1)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{N-K-1} \qquad (**).$$

an the random variables in (*) and (**) are independent.

Then under $H_0 : \boldsymbol{R\beta} = \boldsymbol{r}$ we get

$$T = \frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})')(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})}{q \cdot \hat{\sigma}^2} \sim F_{q,N-K-1}$$

Example: we test if the regressors `age, sales` and `profits` simultaneously have no impact on the salary.

$$H_0 : \ \beta_3 = \beta_4 = \beta_5 = 0 \qquad \text{vs} \qquad H_1 : \ \beta_j \neq 0 \text{ for at least one } j$$

$$\boldsymbol{R} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \qquad \boldsymbol{r} = (0,0,0)'$$

```
> library("gmodels");
> R = rbind(c(0,0,0,1,0,0,0), c(0,0,0,0,1,0,0), c(0,0,0,0,0,1,0));
> r = c(0,0,0);
> glh.test(Z.ceo, cm=R, d=r);
Test of General Linear Hypothesis
glh.test(reg = Z.ceo, cm = R, d = r)
F = 6.5869, df1 =   3, df2 = 440, p-value = 0.0002312
```

Comparing non-nested models

Q: how to compare two non-nested models with the same y_i(!)

$$\text{Model } A: \; y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i$$
$$\text{Model } B: \; y_i = \boldsymbol{z}_i'\boldsymbol{\gamma} + v_i$$

where $\boldsymbol{x}$ includes variables which are not in $\boldsymbol{z}$ and *vice versa.*

I Use R^2_{adj}, AIC, BIC ⇝ less formal approach

II Encompassing F-test: Let $\boldsymbol{x}_i = (\boldsymbol{x}_{1i}', \boldsymbol{x}_{2i}')'$, where $\boldsymbol{x}_{1i}$ are included in $\boldsymbol{z}_i$, but $\boldsymbol{x}_{2i}$ are not.

$$y_i = \boldsymbol{z}_i'\boldsymbol{\gamma} + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + v_i$$

If $H_0: \; \boldsymbol{\beta}_2 = 0$ is not rejected, then Model B encompasses Model A ⇝ simple GLH test

III *J*-test: is there additional information in the predictions from A?

$$y_i = \boldsymbol{z}_i'\boldsymbol{\gamma} + \delta \hat{y}_i^{(A)} + v_i$$

If $H_0: \ \delta = 0$ is not rejected, then Model B is superior $\rightsquigarrow$ simple t-test

Example: is the model based on the company or on the CEO characteristics better?

```
> jtest(salary ~ age+tenure+totcomp, salary ~ profits+sales+assets, data=ceo)
Model 1: salary ~ age + tenure + totcomp
Model 2: salary ~ profits + sales + assets
                Estimate Std. Error t value  Pr(>|t|)
M1 + fitted(M2)  0.92205   0.084195 10.9513 < 2.2e-16 ***
M2 + fitted(M1)  0.85055   0.110291  7.7119 8.291e-14 ***
```

Q: how to compare a model for y_i with a model for $\ln y_i$?

- $\hat{y}_i$ are the predictions from the model for y_i (linear model)
- $\widehat{\ln y_i}$ are the predictions from the model for $\ln y_i$ (log-linear model)

Is linear better then the log-linear?

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \delta_{lin}(\ln \hat{y}_i - \widehat{\ln y_i}) + u_i$$

If $H_0: \ \delta = 0$ is not rejected, then

- there is no additional valuable information in the log's useful to forecast y_i
- the linear model should be preferred

Is log-linear better then the linear?

$$\ln y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \delta_{log}(\hat{y}_i - exp\{\widehat{\ln y_i}\}) + u_i$$

Missing data

Here: just for explanatory variables

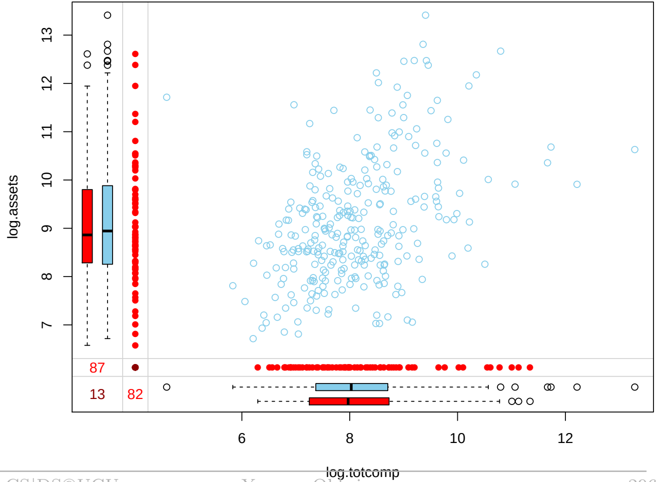
- *Missing completely at random* (MCAR) - the data is missing independently of the observed and missing values
  Example: each participant tosses a coin to decide if he/she wants to answer specific questions on depressions
- *Missing at random* (MAR) - the data is missing independently of the missing values
  Example: men tend to give no answers independently of the level of depression
- *Missing not at random* (MNAR) - if the data is missing strongly depends on the values
  Example: participants with high depression mainly do not respond

Solutions by MCAR and MAR

- Drop the rows with missing data
- *hot imputation*: the missing value is replaced with some random but typical
- *mean imputation*: the missing value is replaced with the mean of the observed values
- *regression imputation*: the missing value is forecasted using some model (regression)

For MNAR we have *informative missingness*. A more complex modelling is required.

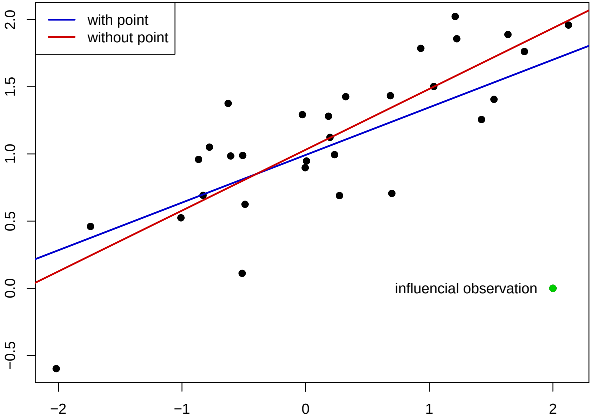
```
> library("mice")
> X = log(as.matrix(ceo[,c(2,7)]));
> colnames(X) = c("log.totcomp", "log.assets")
> X.missing = rbinom(dim(X)[1]*dim(X)[2],1, 0.20)
> dim(X.missing)=dim(X); X.missed = X;
> X.missed[which(X.missing==1)] = NA; X.missed= as.data.frame(X.missed);
> marginplot(X.missed[c(1,2)])
```

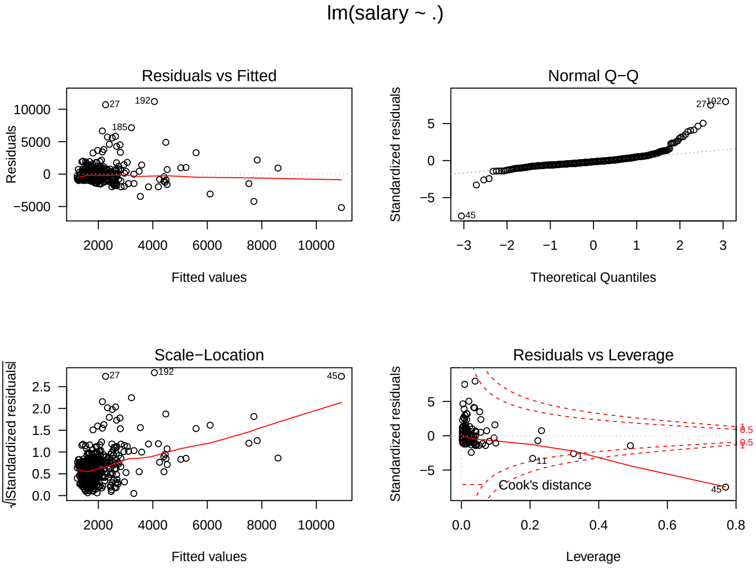


Outliers and influential observations

Note: the OLS estimators are a weighted average of the elements of $\boldsymbol{y}$

$\rightsquigarrow$ some observations may have larger impact than other




lm(salary ~ .)
Residuals vs Fitted
Residuals
10000
5000
0
−5000
2000
4000
6000
8000
10000
Fitted values
27
192
185
Normal Q−Q
Standardized residuals
5
0
−5
−3
−2
−1
0
1
2
3
Theoretical Quantiles
27
192
45
Scale−Location
√|Standardized residuals|
2.5
2.0
1.5
1.0
0.5
0.0
2000
4000
6000
8000
10000
Fitted values
27
192
45
Residuals vs Leverage
Standardized residuals
5
0
−5
0.0
0.2
0.4
0.6
0.8
Leverage
Cook's distance
11
1
45
1
0.5
0.5
1

An important tool in determining the influential points is leverage h_i

- h_i measures how far is the given observation from the rest of the values
- The leverage is defined as self-sensitivity
$$h_i = \frac{\partial \hat{y}_i}{\partial y_i}$$
- The leverage is also the diagonal element of the projection matrix $\boldsymbol{P}$: $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ (see above)
$$h_i = \{\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\}_{ii}$$
- The variance of the prediction error is defined (cf. above) as
$$Var(\hat{u}_i) = \sigma^2(1 - h_i^2)$$
- The standardized and studendized residuals are defined as
$$\hat{u}^2_{i,stan} = \frac{\hat{u}_i}{\hat{\sigma}\sqrt{1 - h_i}} \quad \text{snd} \quad \hat{u}^*_{i,stud} = \frac{\hat{u}_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_i}}$$
- It holds: $0 \leq h_i \leq 1$

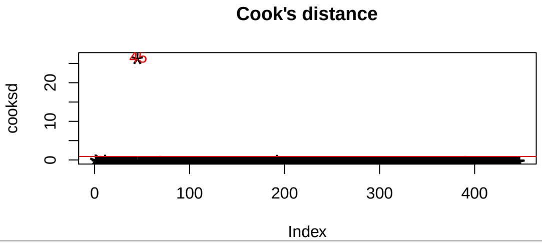
To identify influential observations we can use Cook's Distance

$$D_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})' \; \boldsymbol{X}'\boldsymbol{X} \; (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{(K+1)\hat{\sigma}^2},$$

where $\hat{\boldsymbol{b}}_{(i)}$ is the vector of estimated parameters with dropped i-th observation.

If $D_i > F_{K;N-K-1;0.5} \in [0.5; 1]$, then the observation i is influential.

```
> cooks.distance(Z.ceo)
```



Generalized Least Squares

Note: to guarantee the good properties of $\hat{\boldsymbol{\beta}}_{OLS}$ we have to assume

- $E(\boldsymbol{u}|\boldsymbol{X}) = \mathbf{0} \rightsquigarrow$ for unbiasedness, i.e. $E(\hat{\boldsymbol{\beta}}_{OLS}) = \boldsymbol{\beta}$
- $Var(\boldsymbol{u}|\boldsymbol{X}) = \sigma^2\boldsymbol{I} \rightsquigarrow$ for the smallest possible variance $Var(\hat{\boldsymbol{\beta}}_{OLS}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$

But: what happens if the 2nd assumption is not fulfilled?

$$Var(\boldsymbol{u}|\boldsymbol{X}) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1N} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{N1} & \sigma_{N2} & \dots & \sigma_N^2 \end{pmatrix} = \sigma^2\boldsymbol{\Psi},$$

where $\boldsymbol{\Psi}$ is a known (at least here) positive definite, symmetric square matrix and σ^2 is a scaling parameter.

Then

- $E(\hat{\boldsymbol{\beta}}_{OLS}) = \boldsymbol{\beta}$ is still fulfilled,
- ... but the variance changes

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}_{OLS}) &= Var(\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}) \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'Var(\boldsymbol{u})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Psi}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \end{aligned}$$

For the LR this implies:

- The classical formula $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is incorrect.
- The t-tests, F-test, the test of the GLH become incorrect.
- The Gauss-Markov theorem is not fulfilled and $\hat{\boldsymbol{\beta}}_{OLS}$ is not efficient!

Idea: can we transform the data, so that the original assumption becomes satisfied?

Consider the Cholesky decomposition of the matrix $\boldsymbol{\Psi}$:

$$\boldsymbol{\Psi}^{-1} \quad = \quad \boldsymbol{P}'\boldsymbol{P},$$

where $\boldsymbol{P}$ is a square (triangular) non-singular matrix. It holds

$$\begin{aligned} \boldsymbol{\Psi} &= (\boldsymbol{P}'\boldsymbol{P})^{-1} = \boldsymbol{P}^{-1}(\boldsymbol{P}')^{-1} \\ \boldsymbol{P}\boldsymbol{\Psi}\boldsymbol{P}' &= \boldsymbol{P}\boldsymbol{P}^{-1}(\boldsymbol{P}')^{-1}\boldsymbol{P}' = \boldsymbol{I} \end{aligned}$$

$$\begin{aligned}
\boldsymbol{y} &= \boldsymbol{X\beta} + \boldsymbol{u} \\
\boldsymbol{P}\cdot\boldsymbol{y} &= \boldsymbol{P}\cdot\boldsymbol{X\beta} + \boldsymbol{P}\cdot\boldsymbol{u} \\
\boldsymbol{y}^* &= \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{u}^*
\end{aligned}$$

But for the new error term $\boldsymbol{u}^*$ it holds

$$Var(\boldsymbol{u}^*) = Var(\boldsymbol{Pu}) = \boldsymbol{P}Var(\boldsymbol{u})\boldsymbol{P}' = \sigma^2\boldsymbol{P\Psi P}' = \sigma^2\boldsymbol{I}$$

Thus the new error term satisfies the classical assumptions!

The generalized least squares (GLS) estimator is then defined as:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{GLS} &= (\boldsymbol{X}^{*\prime}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*\prime}\boldsymbol{y}^* \\
&= (\boldsymbol{X}'\boldsymbol{P}'\boldsymbol{PX})^{-1}\boldsymbol{X}'\boldsymbol{P}'\boldsymbol{Py} \\
&= (\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{y}
\end{aligned}$$

with the variance

$$Var(\hat{\boldsymbol{\beta}}_{GLS}) = \sigma^2(\boldsymbol{X}^{*\prime}\boldsymbol{X}^*)^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}$$

Note 1:

- $\hat{\boldsymbol{\beta}}_{GLS}$ is BLUE and thus has a variance smaller than $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$
- $\hat{\boldsymbol{\beta}}_{GLS}$ is a consistent estimator of $\boldsymbol{\beta}$ if $p\lim \boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X}/N = \tilde{\mathbf{Q}}$.
- $\hat{\boldsymbol{\beta}}_{GLS}$ is normally distributed, if $\boldsymbol{u}$ is normal, i.e.

$$\hat{\boldsymbol{\beta}}_{GLS} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}).$$

- Note: R^2 cannot be used to compare the GLS models. R^2 is not bounded and should not necessarily increase/decrease if we add/drop a variable.

Note 2: $\boldsymbol{\Psi}$ is a known matrix up to now! If we have an estimator $\hat{\boldsymbol{\Psi}}$, then this results in the feasible generalized least squares (FGLS) estimator $\hat{\hat{\boldsymbol{\beta}}}_{FGLS}$

Note 3: we have three possibilities

I $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ with the “wrong” variance
$Var(\hat{\boldsymbol{\beta}}_{OLS}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$

II $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ with the “corrected” variance
$Var(\hat{\boldsymbol{\beta}}_{OLS}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$

III $\hat{\boldsymbol{\beta}}_{GLS} = (\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{y}$ with the “correct” variance
$Var(\hat{\boldsymbol{\beta}}_{GLS}) = \sigma^2(\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}$

The structure of $\boldsymbol{\Psi}$ allow for two special cases:

heteroscedasticity $$\sigma^2\boldsymbol{\Psi} = \sigma^2 \begin{pmatrix} \omega_1^2 & 0 & \dots & 0 \\ 0 & \omega_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \omega_N^2 \end{pmatrix}$$

autocorrelation $$\sigma^2\boldsymbol{\Psi} = \sigma^2 \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1N} \\ \rho_{21} & 1 & \dots & \rho_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{N1} & \rho_{N2} & \dots & 1 \end{pmatrix}$$

Heteroscedasticity

Heteroscedasticity implies that

$$Var(u_i) = \sigma^2 \omega_i^2 \qquad \text{with} \qquad \sum_{i=1}^{N} \omega_i = 1$$

and frequently arises in cross-sectional data sets.

Using the idea of GLS we obtain $P = diag\{\omega_i^{-1}\}$ and it holds

$$y_i^* = \boldsymbol{x}_i^* \boldsymbol{\beta} + \boldsymbol{u}_i^* \qquad \frac{y_i}{\omega_i} = \left(\frac{\boldsymbol{x}_i}{\omega_i}\right)' \boldsymbol{\beta} + \frac{\boldsymbol{u}_i}{\omega_i}$$

The GLS estimator is then:

$$\hat{\boldsymbol{\beta}}_{GLS} = \left(\sum_{i=1}^{N} \omega_i^{-2} \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} \sum_{i=1}^{N} \omega_i^{-2} \boldsymbol{x}_i y_i$$

- Each observation is weighted by a factor proportional to the variance $\rightsquigarrow$ weighted least squares.
- The parameters have to be interpreted in the context of the original model.
- All tests can be performed similarly as for the classical OLS.

How strong is the inefficiency of OLS?

$$
\begin{aligned}
& Var(\hat{\boldsymbol{\beta}}_{OLS}) - Var(\hat{\boldsymbol{\beta}}_{hetero}) \\
= \; & \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} - \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Psi}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\
= \; & \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left[\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}'\boldsymbol{\Psi}\boldsymbol{X}\right]\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
= \; & \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left[\sum_{i=1}^{N}(1-\omega_i)\boldsymbol{x}_i\boldsymbol{x}_i'\right]\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}
\end{aligned}
$$

If ω_i is not correlated with the data (this would be the case if, for example, w_i is high for high x's), then $\sum_{i=1}^{N}\boldsymbol{x}_i\boldsymbol{x}_i'$ would be close to $\sum_{i=1}^{N}\omega_i\boldsymbol{x}_i\boldsymbol{x}_i'$ (since $\sum\omega_i = 1$) and pure OLS variance $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ would not be too misleading.

Example (Davidson and MacKinnon (1993))

$$Y_t = 1 + x_t + u_t, \quad u_t \sim N(0, x_t^{\alpha}), \quad t = 1, \dots, 100,$$
$$x_t \sim U[0,1]$$

a) Generation of 20000 samples of 100 observations

b) Calculation of

i) $Var(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\,\boldsymbol{X}' \begin{pmatrix} x_1^{\alpha} & & 0 \\ & \ddots & \\ 0 & & x_{100}^{\alpha} \end{pmatrix} \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$ variance of OLS if the data are heteroscedastic

ii) $(\boldsymbol{X}'\boldsymbol{\Psi}\boldsymbol{X})^{-1}$ variance of the GLS estimator

iii) $\hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ variance of OLS if the data is falsely assumed to be homoscedastic

Example (cont.)

Correct and incorrect standard errors

	OLS intercept			**OLS slope**		
α	**Incorrect**	**Correct**	**GLS intercept**	**Incorrect**	**Correct**	**GLS slope**
0.5	0.164	0.134	0.110	0.285	0.277	0.243
1.0	0.142	0.101	0.048	0.246	0.247	0.173
2.0	0.116	0.074	0.0073	0.200	0.220	0.109
3.0	0.100	0.064	0.0013	0.173	0.206	0.056

Problem: the variances ω_i^2 are unknown

Solutions:

- Improve the classical OLS estimator without specifying the form of the heteroscedasticity
- Specify the form of the heteroscedasticity and use FGLS.

Heteroskedasticity-consistent variance estimator of White

Idea:

- The heteroskedasticity has no impact on the unbiasedness of $\hat{\beta}_{OLS}$, only on the variance.
- Is it possible to "correctly " estimate the variance?

Recall: under heteroscedasticity the variance of the OLS estimator is

$$Var(\hat{\boldsymbol{\beta}}_{OLS}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Psi}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1},$$

where $\boldsymbol{\Psi} = diag\{\omega_i\}$ with $i = 1, \ldots, N$.

$Var(u_i) = \sigma^2\omega_i$'s are unknown, but can be estimated using residuals:

$$\widehat{Var}(u_i) = \hat{u}_i^2.$$

This leads to the Heteroskedasticity-consistent (or White) variance estimator (FGLS)

$$\begin{aligned}
\widehat{Var}(\hat{\boldsymbol{\beta}}_{OLS}) &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' diag\{\hat{u}_i^2\}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= \left(\sum_{i=1}^{N} \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1} \sum_{i=1}^{N} \hat{u}_i^2 \boldsymbol{x}_i\boldsymbol{x}_i' \left(\sum_{i=1}^{N} \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1} .
\end{aligned}$$

It holds that

$$p\lim \widehat{Var}(\hat{\boldsymbol{\beta}}_{OLS}) = Var(\hat{\boldsymbol{\beta}}_{OLS}).$$

Example (Salary with White estimator)

We have no additional knowledge about the structure of the heteroscedasticity and use the White estimator.

```
  > Zceo= lm(salary  ~ ., data=ceo);
  > summary(Zceo)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.575e+02  5.963e+02   1.438  0.15111
totcomp     1.430e-02  2.179e-03   6.564 1.47e-10 ***
tenure      2.740e+01  9.067e+00   3.022  0.00266 **
age         7.034e+00  1.095e+01   0.642  0.52105
sales       1.398e-02  6.320e-03   2.212  0.02749 *
profits     1.036e-01  6.603e-02   1.569  0.11748
assets      7.567e-03  1.267e-03   5.973 4.80e-09 ***
---
Residual standard error: 1434 on 440 degrees of freedom
Multiple R-squared:  0.3158,Adjusted R-squared:  0.3065
F-statistic: 33.85 on 6 and 440 DF,  p-value: < 2.2e-16
```

```
  > library("sandwich")
  > library("lmtest")
  > Zceo= lm(salary  ~ ., data=ceo)
  > covWhite = vcovHC(Zceo, type="HC")
  > coeftest(Zceo, vcov=covWhite)

  t test of coefficients:
              Estimate Std. Error t value  Pr(>|t|)
(Intercept) 8.5754e+02 5.1864e+02  1.6534 0.0989539 .
totcomp     1.4302e-02 7.5349e-03  1.8981 0.0583378 .
tenure      2.7401e+01 1.1708e+01  2.3404 0.0197099 *
age         7.0343e+00 9.8739e+00  0.7124 0.4765850
sales       1.3978e-02 1.0141e-02  1.3784 0.1687908
profits     1.0357e-01 9.4396e-02  1.0971 0.2731771
assets      7.5666e-03 2.1462e-03  3.5257 0.0004667 ***
```

White test for heteroscedasticity

Aim: can we test heteroscedasticity without specifying its form?
Idea: use $\hat{u}_i^2$ as a proxy for σ^2 as in the White estimator

$$\begin{aligned}\hat{u}_i^2 &= \alpha_0 + \alpha_1 x_{1i} + \cdots + \alpha_J x_{Ji} \\ &\quad + \alpha_{J+1} x_{1i}^2 + \cdots + \alpha_{2J} x_{Ji}^2 + \text{ cross products} + v_i\end{aligned}$$

$$H_0: \ \boldsymbol{\alpha} = \mathbf{0} \qquad \text{vs} \qquad H_1: \ \boldsymbol{\alpha} \neq \mathbf{0}$$

$$V = N \cdot R^2 \sim \chi^2_{J^*},$$

where R^2 is the corresponding goodness-of-fit measure.

Example (White test for CEO data)

```
> ceo.sq = cbind(ceo, ceo[,-1]^2)
> # renaming the variables
> Zceo.white = lm(Zceo$residuals^2 ~ ., data=ceo.sq)
> summary(Zceo.white)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.543e+06  1.208e+07  -0.128   0.8984
salary       4.764e+03  2.214e+02  21.516   <2e-16 ***
totcomp     -7.771e+00  2.764e+01  -0.281   0.7787
tenure      -1.195e+05  8.733e+04  -1.368   0.1720
age         -4.710e+04  4.349e+05  -0.108   0.9138
sales       -8.776e+01  5.517e+01  -1.591   0.1124
profits     -1.436e+03  4.524e+02  -3.175   0.0016 **
assets      -1.535e+01  1.145e+01  -1.341   0.1808
totcomp2     4.439e-05  5.008e-05   0.887   0.3758
tenure2      3.513e+03  2.412e+03   1.457   0.1459
age2        -2.464e+02  3.879e+03  -0.064   0.9494
sales2       6.572e-04  3.916e-04   1.678   0.0940 .
profits2     2.200e-02  2.404e-02   0.915   0.3606
assets2     -4.195e-06  2.132e-05  -0.197   0.8441
---
Residual standard error: 6002000 on 433 degrees of freedom
Multiple R-squared:  0.6004,Adjusted R-squared:  0.5885
F-statistic: 50.05 on 13 and 433 DF,  p-value: < 2.2e-16
> 447*0.6004
[1] 268.3788
> qchisq(0.96, 12)
[1] 21.78511
```

A model with two unknown variances

Assume the data can be split into two groups: $i \in A$ (all males) and $i \in B$ (all females). Then

$$Var(u_i) \quad = \quad \begin{cases} \sigma_A^2, & \text{if } i \in A \\ \sigma_B^2, & \text{if } i \in B \end{cases}$$

Note that σ_A^2 can be estimated by

$$\hat{\sigma}_A^2 = \frac{1}{N_A - K} \sum_{i \in A} (y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_A)^2$$

and similarly for σ_B^2.

Then the FGLS estimator of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}}_{FGLS} = \left(\hat{\sigma}_A^{-2} \sum_{i \in A} \boldsymbol{x}_i \boldsymbol{x}_i' + \hat{\sigma}_B^{-2} \sum_{i \in B} \boldsymbol{x}_i \boldsymbol{x}_i' \right)^{-1} \left(\hat{\sigma}_A^{-2} \sum_{i \in A} \boldsymbol{x}_i y_i + \hat{\sigma}_B^{-2} \sum_{i \in B} \boldsymbol{x}_i y_i \right).$$

Example (Wage with two groups)

We believe that the variance of the residuals is different for males and for females.

```
> Zwage = lm(WAGE˜ ., data=wage)
> summary(Zwage)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.38002    0.46498  -7.269 4.50e-13 ***
EXPER        0.12483    0.02376   5.253 1.59e-07 ***
MALE         1.34437    0.10768  12.485  < 2e-16 ***
SCHOOL       0.63880    0.03280  19.478  < 2e-16 ***
---
Residual standard error: 3.046 on 3290 degrees of freedom
Multiple R-squared:  0.1326,Adjusted R-squared:  0.1318
F-statistic: 167.6 on 3 and 3290 DF,  p-value: < 2.2e-16
```

```
> sigma2a = var(Zwage$residuals[wage$MALE==1])
[1] 10.90651
> sigma2b = var(Zwage$residuals[wage$MALE==0])
[1] 7.477704
> weights.v = vector(,length=dim(wage)[1]);
> weights.v[wage$MALE==1] = 1/sqrt(sigma2a);
> weights.v[wage$MALE==0] = 1/sqrt(sigma2b);
> wage.star = t(t(wage)%*%diag(weights.v));
> one.v = rep(1,length=dim(wage)[1])*weights.v;
> wage.star = data.frame(one.v, wage.star);
> Zwage2.star = lm(WAGE˜ . -1, data=wage.star);
> summary(Zwage2.star)
Coefficients:
       Estimate Std. Error t value Pr(>|t|)
one.v  -3.25018    0.45583  -7.130 1.23e-12 ***
EXPER   0.12676    0.02344   5.407 6.88e-08 ***
MALE    1.33839    0.10675  12.537  < 2e-16 ***
SCHOOL  0.62657    0.03247  19.295  < 2e-16 ***
---
Residual standard error: 1 on 3290 degrees of freedom
Multiple R-squared:  0.7881,Adjusted R-squared:  0.7878
F-statistic:  3058 on 4 and 3290 DF,  p-value: < 2.2e-1
```

How to detect heteroscedasticity for groups?

$$H_0:\ \sigma_A^2 = \sigma_B^2 \qquad \text{vs} \qquad H_1:\ \sigma_A^2 \neq \sigma_B^2$$

Statistics II: two-sample test for variances:

$$V = \frac{\hat{\sigma}_A^2}{\hat{\sigma}_B^2} \sim F_{N_A-K,N_B-K}$$

Note: the test can be generalized to any subgroups of data ⇝ Bartlett's test

$$H_0:\ \sigma_1^2 = \sigma_2^2 = \ldots \sigma_J^2 \qquad \text{vs} \qquad H_1:\ \sigma_i^2 \neq \sigma_j^2$$

$$V = \frac{1}{const} \sum_{i=1}^{J} (N_i - K) \log \frac{\hat{\sigma}_i^2}{\hat{\sigma}^2} \overset{asymp}{\sim} \chi^2_{J-1},$$

where $\hat{\sigma}^2$ is a "pooled" estimator of the variance.

Example (Wage with two groups)

Is the variance in fact different for males and females?

```
> library("olsrr")
> ols_bartlett_test(Zwage$residuals,
                    group_var = factor(wage$MALE))

    Bartlett’s Test of Homogenity of Variances
------------------------------------------------
Ho: Variances are equal across groups
Ha: Variances are unequal for atleast two groups
         Test Summary
 ------------------------------
 DF             =    1
 Chi2           =    57.78661
 Prob > Chi2    =    2.92152e-14
```

Multiplicative heteroscedasticity

The most common form of heteroscedasticity assumes that the variance of error terms depends exponentially on J exogenous variables:

$$Var(u_i) = \sigma_i^2 = \sigma^2 e^{\boldsymbol{z}_i \boldsymbol{\gamma}},$$

where $\boldsymbol{z}_i$ is a subset of x's, their transformations or additional variables and $\boldsymbol{\gamma}$ is J-dim. vector of parameters.

Note:

- if $J = 1$ and z is a dummy then the model reduces to the model with two variances.
- The equation above can be written as:

  $$\log \sigma_i^2 = \log \sigma^2 + \boldsymbol{z}_i \boldsymbol{\gamma}$$

  leading to

  $$\log \hat{u}_i^2 = \log \sigma^2 + \boldsymbol{z}_i \boldsymbol{\gamma} + v_i \qquad (*)$$

To estimate the model follow the steps:

1. Using OLS obtain $\hat{\boldsymbol{\beta}}_{OLS}$
2. Compute $\log \hat{u}_i^2$
3. Run the regression (*) to obtain $\hat{\boldsymbol{\gamma}}$
4. Compute $\hat{\omega}_i^2 = e^{\boldsymbol{z}_i \hat{\boldsymbol{\gamma}}}$
5. Estimate the weighted regression to obtain $\hat{\boldsymbol{\beta}}_{FGLS}$

$$\frac{y_i}{\hat{\omega}_i} = \frac{\boldsymbol{x}_i'}{\hat{\omega}_i}\boldsymbol{\beta} + \frac{u_i}{\hat{\omega}_i}$$

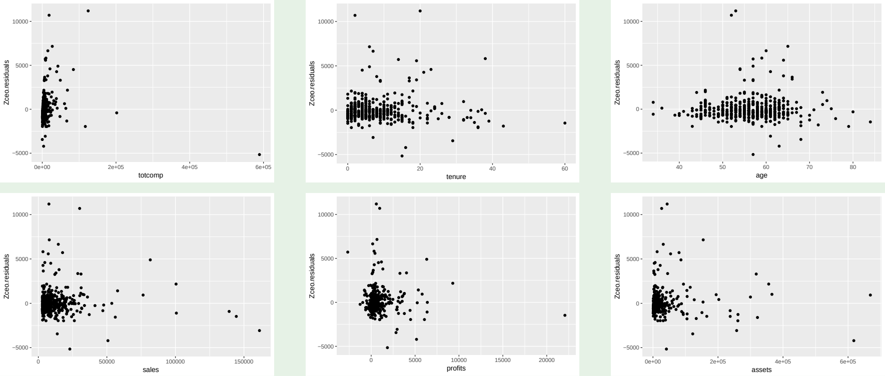
6. The constant σ^2 is estimated by

$$\hat{\sigma}^2 = \frac{1}{N-K} \sum_{i=1}^{N} \frac{(y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_{FGLS})^2}{\hat{\omega}_i^2}$$

7. The variance $\hat{\boldsymbol{\beta}}_{FGLS}$ is estimated by

$$\widehat{Var}(\hat{\boldsymbol{\beta}}_{FGLS}) = \hat{\sigma}^2 \left(\sum_{i=1}^{N} \frac{\boldsymbol{x}_i \boldsymbol{x}_i'}{\hat{\omega}_i^2} \right)^{-1}$$

Example (CEO Salary)



```
> library("nlme")
> Zceo.gls = gls(salary  ~ ., data=ceo, weights=varExp(form=~tenure))
Generalized least squares fit by REML
  Model: salary ~ .
  Data: ceo
      AIC      BIC    logLik
  7726.99 7763.771 -3854.495

Variance function:
 Structure: Exponential of variance covariate
 Formula: ~tenure
 Parameter estimates:
     expon
0.03311352

Coefficients:
               Value Std.Error  t-value p-value
(Intercept) 512.6636  529.7081 0.967823  0.3337
totcomp       0.0147    0.0023 6.295461  0.0000
tenure       37.7716   12.4337 3.037826  0.0025
age          12.0895    9.7766 1.236575  0.2169
sales         0.0141    0.0058 2.438940  0.0151
profits       0.0979    0.0546 1.792885  0.0737
assets        0.0074    0.0013 5.783951  0.0000

Residual standard error: 1031.879
```

Breush-Pagan test for multiplicative heteroscedasticity using (*)

$$H_0 : \boldsymbol{\gamma} = \mathbf{0} \qquad \text{vs} \qquad H_1 : \gamma_j \neq \mathbf{0}\text{for at least one } j$$

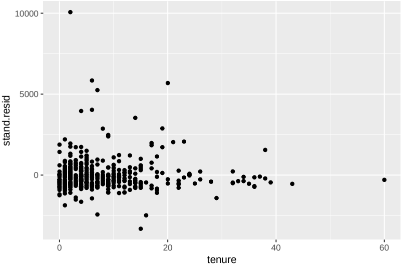
Note: the residuals of (*) are not normal, so the F-test is not reliable

$$V = N \cdot R^2 \sim \chi^2_J,$$

where R^2 is the goodness-of-fit measure of (*)

Example:

```
> summary(lm(log(Zceo$residuals^2)~ ceo$tenure))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.81048    0.15115  78.140  < 2e-16 ***
ceo$tenure   0.05602    0.01330   4.213 3.05e-05 ***
---
Residual standard error: 2.316 on 445 degrees of freedom
Multiple R-squared:  0.03836,Adjusted R-squared:  0.0362
F-statistic: 17.75 on 1 and 445 DF,  p-value: 3.05e-05
> 447*0.03836
[1] 17.14692
> qchisq(0.95,1)
[1] 3.841459
```



Autocorrelation

The problem of autocorrelation or serial correlation frequently arises for time series data.

Def: The time ordered sequence $\{y_t\}_{t=1,\dots,T}$ with systematic correlation between observations is called a time-series process.

A typical time series model links y_t with factors x_t and own past values $y_{t-1}, y_{t-2}, \dots$.

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 y_{t-1} + u_t.$$

Def: A time series $\{z_t\}$ is stationary, if

- $E(z_t)$ =const for all t
- $Var(z_t)$ =const for all t
- $Corr(z_t, z_{t-h}) = \rho_h$, i.e. the autocorrelation depends only on lag h, but not on t

If y_t and $\boldsymbol{x}_t$ are time series data, then frequently the error terms inherit the time dependence

$$\begin{aligned} y_t &= \boldsymbol{x}_t'\boldsymbol{\beta} + u_t \\ E(u_t) &= 0 \\ Var(\boldsymbol{u}) &= \sigma^2\boldsymbol{\Psi}, \end{aligned}$$

where

$$\sigma^2\boldsymbol{\Psi} = \begin{pmatrix} 1 & Corr(u_1,u_2) & \dots & Corr(u_1,u_T) \\ Corr(u_2,u_1) & 1 & \dots & Corr(u_2,u_T) \\ \vdots & \vdots & \ddots & \vdots \\ Corr(u_T,u_1) & Corr(u_2,u_T) & \dots & 1 \end{pmatrix},$$

where $\rho_h = Corr(u_t, u_{t-h})$ is the autocorrelation function.

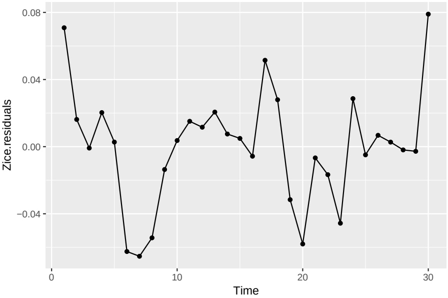
Note: ρ_h depends on h, but not on t (see stationarity below)!

Example (Ice cream consumption)

`cons:`	`consumption of ice cream per head (in pints);`
`income:`	`average family income per week (in US Dollars);`
`price:`	`price of ice cream (per pint);`
`temp:`	`average temperature (in Fahrenheit);`
`time:`	`index from 1 to 30`

Source: Veerbek, M. A guide to modern econometrics

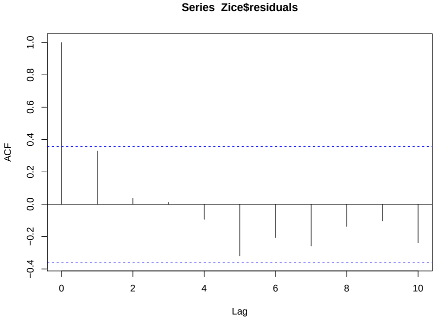
```
> library("Ecdat")
> data(Icecream)
> Zice = lm(cons~., data=Icecream)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1973151  0.2702162   0.730  0.47179
income       0.0033078  0.0011714   2.824  0.00899 **
price       -1.0444140  0.8343573  -1.252  0.22180
temp         0.0034584  0.0004455   7.762  3.1e-08 ***
---
Residual standard error: 0.03683 on 26 degrees of freedom
Multiple R-squared:  0.719,Adjusted R-squared:  0.6866
F-statistic: 22.17 on 3 and 26 DF,  p-value: 2.451e-07
```



The autocorrelation $\rho_h = Corr(u_t, u_{t-h})$ can be estimated in a similar way as the Pearson correlation:

$$\hat{\rho}_h = \frac{\frac{1}{T-h-1}\sum_{t=h+1}^{T}(\hat{u}_t - \bar{\hat{u}})(\hat{u}_{t-h} - \bar{\hat{u}})}{\frac{1}{T-1}\sum_{t=1}^{T}(\hat{u}_t - \bar{\hat{u}})^2}$$

```
> acf(Zice$residuals, lag.max=10)
```



Note: If we assume that the residuals are autocorrelated we may assume that they follow a time series model.

The simples time series model is an AR(1) process (autoregressive process of order 1)

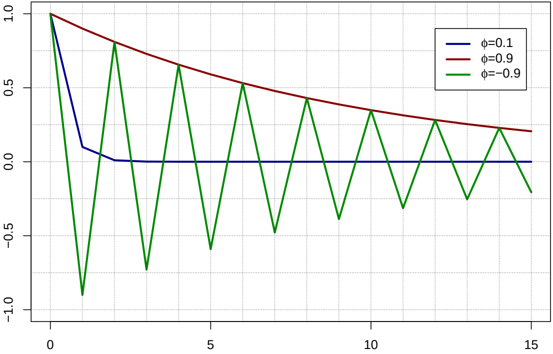
$$\begin{aligned}
u_t &= \phi u_{t-1} + v_t,\\
E(u_t) &= \phi E(u_{t-1}) + E(v_t) = 0,\\
Var(u_t) &= \phi^2 Var(u_{t-1}) + Var(v_t) = \sigma^2/(1-\phi^2) = \gamma_0,\\
Cov(u_t, u_{t-1}) &= \phi E(u_{t-1}u_{t-1}) + E(u_{t-1}v_t) = \phi\gamma_0\\
Cov(u_t, u_{t-h}) &= \phi^h \gamma_0
\end{aligned}$$

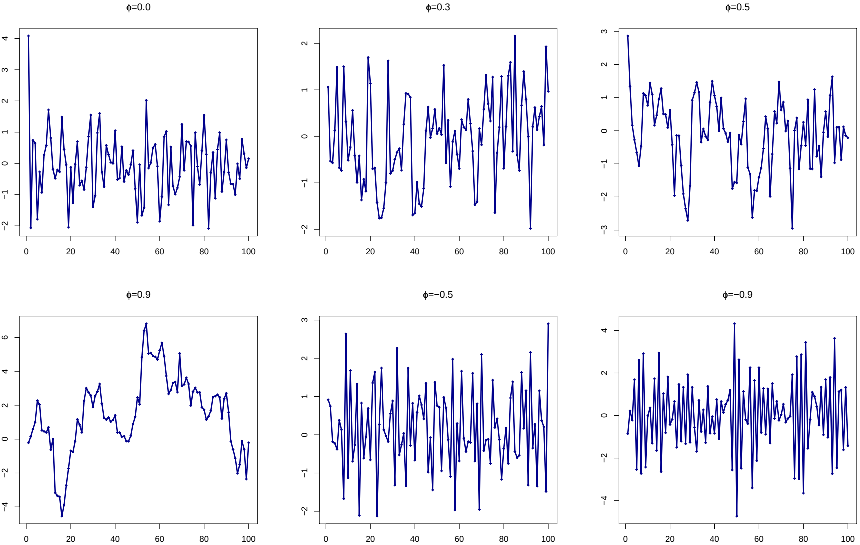
Note: it is necessary to assume $|\phi| < 1$.

$$Var(\boldsymbol{u}) = \frac{\sigma^2}{1-\phi^2}\begin{pmatrix} 1 & \phi & \dots & \phi^{T-1}\\ \phi & 1 & \dots & \phi^{T-2}\\ \vdots & \vdots & \ddots & \vdots\\ \phi^{T-1} & \phi^{T-2} & \dots & 1 \end{pmatrix}$$

Autocorrelations of an AR(1) process

$$Corr(u_t, u_{t-h}) = \frac{Cov(u_t, u_{t-h})}{\sqrt{Var(u_t)Var(u_{t-h})}} = \frac{\phi^h \gamma_0}{\sqrt{\gamma_0 \gamma_0}} = \phi^h$$




ϕ=0.0
ϕ=0.3
ϕ=0.5
ϕ=0.9
ϕ=−0.5
ϕ=−0.9
0
20
40
60
80
100
−2
−1
0
1
2
3
4
−4
6

Is autocorrelation damaging for OLS?

If there are no lagged dependent variables among the regressors, then the OLS estimator of $\boldsymbol{\beta}$ is still unbiased, consistent, asymptotically normal, but inefficient.

The regressors $\boldsymbol{x}_t$ should be stationary (and ergodic).

Def: A time series $\{z_t\}$ is stationary, if

- $E(z_t)$ =const for all t
- $Var(z_t)$ =const for all t
- $Corr(z_t, z_{t-h}) = \rho_h$, i.e. the autocorrelation depends only on lag h, but not on t

Example of an inconsistent OLS estimation

$$y_t = \beta y_{t-1} + u_t \qquad\qquad u_t = \phi u_{t-1} + v_t$$

$$\begin{aligned} Cov(y_{t-1}, u_t) &= Cov(y_{t-1}, \phi u_{t-1} + \varepsilon_t) = \phi Cov(y_{t-1}, u_{t-1}) \\ &= \phi Cov(\beta y_{t-2} + u_{t-1}, u_{t-1}) \\ &= \phi\beta Cov(y_{t-2}, u_{t-1}) + \phi Cov(u_{t-1}, u_{t-1}) \\ &= \phi\beta Cov(y_{t-1}, u_t) + \phi Cov(u_{t-1}, u_{t-1}) \end{aligned}$$

$$Cov(y_{t-1}, u_t) = \frac{\phi Var(u_t)}{1-\beta\phi} = \frac{\phi\sigma_v^2}{(1-\beta\phi)(1-\phi^2)}.$$

Thus the regressor is correlated with the residuals.

$$p\lim\hat{\beta} = \beta + \frac{Cov(y_{t-1}, u_t)}{Var(y_{t-1})}$$

$\leadsto$ the OLS estimator is neither consistent nor unbiased!!!

How to detect autocorrelation? - Durbin-Watson test

$$
\begin{aligned}
y_t &= \boldsymbol{x}_t'\boldsymbol{\beta} + u_t \\
u_t &= \phi\, u_{t-1} + v_t
\end{aligned}
$$

testing problem:

$H_0 : \phi = 0$		(zero autocorrelation)	against
$H_1 : \phi > 0$	(< 0)	(positive (negative) autocorr.)	

Test statistic:

$$
d = \frac{\sum\limits_{t=2}^{T} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum\limits_{t=1}^{T} \hat{u}_t^2}
$$

Note that

$$
\begin{aligned}
d &= \frac{\sum\limits_{t=2}^{T} \hat{u}_t^2 + \sum\limits_{t=2}^{T} \hat{u}_{t-1}^2 - 2\sum\limits_{t=2}^{T} \hat{u}_t\,\hat{u}_{t-1}}{\sum\limits_{t=1}^{T} \hat{u}_t^2} \\
&\approx 2(1-\hat{\rho}_1)
\end{aligned}
$$

where $\hat{\rho}_1 = \hat{\phi}$ being the first order autocorrelation.

heuristic:

$d \approx 2$ ⇝ zero autocorrelation of the $\hat{u}$'s

$d > 2$ ⇝ negative autocorrelation of the $\hat{u}$'s

$d < 2$ ⇝ positive autocorrelation of the $\hat{u}$'s

Example (Ice cream with Durbin Watson)

```
> dwtest(Zice)

Durbin-Watson test

data:  Zice
DW = 1.0212, p-value = 0.0003024
alternative hypothesis: true autocorrelation is greater than 0
```

Disadvantages of the test:

i) Low power for small T

ii) It cannot be applied to lagged y values

iii) The distribution of the test statistics follows a non-standard distribution

How to estimate $Var(\hat{\boldsymbol{\beta}})$?

Note: if $\boldsymbol{\Psi}$ is known, then the correct variance of $\hat{\boldsymbol{\beta}}$ is

$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Psi}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

Q: how to estimate it without specifying the structure of autocorrelation?
Note: as for the heteroscedasticity we may use the White estimator

$$\hat{\sigma}^2\boldsymbol{X}'\hat{\boldsymbol{\Psi}}\boldsymbol{X} = \sum_{i=1}^{T}\sum_{j=1}^{T}\hat{u}_i\hat{u}_j\boldsymbol{x}_i\boldsymbol{x}_j'.$$

Problem: the estimator is not necessarily positive definite!

The heteroscedasticity and autocorrelation (HAC) robust Newey-West estimator

$$\hat{\sigma}^2 \boldsymbol{X}'\hat{\boldsymbol{\Psi}}\boldsymbol{X} = \sum_{t=1}^{T} \hat{u}_t^2 \boldsymbol{x}_t \boldsymbol{x}_t' + \sum_{i=1}^{L} \sum_{t=i+1}^{T} \omega_i \hat{u}_t \hat{u}_{t-i} (\boldsymbol{x}_t \boldsymbol{x}_{t-i}' + \boldsymbol{x}_{t-i} \boldsymbol{x}_t'),$$

with $\omega_i = 1 - i/(L+1)$ and L must be determined in advance.

Example (Ice cream with Newey-West)

```
> summary(Zice)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1973151  0.2702162   0.730  0.47179
income       0.0033078  0.0011714   2.824  0.00899 **
price       -1.0444140  0.8343573  -1.252  0.22180
temp         0.0034584  0.0004455   7.762  3.1e-08 ***
---
Residual standard error: 0.03683 on 26 degrees of freedom
Multiple R-squared:  0.719,Adjusted R-squared:  0.6866
F-statistic: 22.17 on 3 and 26 DF,  p-value: 2.451e-07
```

```
> library("sandwich"
> coeftest(Zice, vcov=NeweyWest(Zice, 2))
t test of coefficients:

                Estimate   Std. Error t value   Pr(>|t|)
(Intercept)  0.19731507   0.32105598   0.6146   0.544172
income       0.00330776   0.00093645   3.5322   0.001563 *
price       -1.04441399   0.96076185  -1.0871   0.286982
temp         0.00345843   0.00054147   6.3871  9.139e-07 *
```

Estimation if $\boldsymbol{\Psi}$ is known/unknown

If the Durbin-Watson test confirms the autocorrelation of the residuals this has to be taken into account.

$$\begin{aligned} y_t &= \boldsymbol{x}_t'\boldsymbol{\beta} + u_t \\ u_t &= \phi\, u_{t-1} + v_t, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}), \quad |\phi| < 1. \end{aligned}$$

now: GLS estimation

First suppose that ϕ is known. Then

$$Var(\boldsymbol{u}) = \boldsymbol{\Psi} = \frac{\sigma^2}{1-\phi^2} \begin{pmatrix} 1 & \phi & \cdots & \phi^{n-1} \\ \phi & 1 & \cdots & \phi^{n-2} \\ \vdots & \vdots & & \vdots \\ \phi^{n-1} & \varphi^{n-2} & \cdots & 1 \end{pmatrix} = \frac{1}{1-\phi^2}(\phi^{|i-j|})$$

Then

$$\hat{\boldsymbol{\beta}}_{GLS} = (\boldsymbol{X}'\,\boldsymbol{\Psi}^{-1}\,\boldsymbol{X})^{-1}\,\boldsymbol{X}'\,\boldsymbol{\Psi}^{-1}\boldsymbol{y}.$$

If ϕ is unknown then one uses the Cochrane-Orcutt recursion suggested an iterative estimation procedure:

1. Starting estimator of $\phi : \hat{\phi}^{(1)}$ by regression $\hat{u}_t$ on $\hat{u}_{t-1}$ without a constant
2. Insert $\hat{\phi}^{(1)}$ in $\mathbf{\Psi}$ to obtain an estimator $\hat{\mathbf{\Psi}}$
3. Using $\hat{\mathbf{\Psi}}$ compute the FGLS estimator $\hat{\boldsymbol{\beta}}^{(1)}_{FGLS}$
4. Using $\hat{\boldsymbol{\beta}}^{(1)}_{FGLS}$ recompute residuals and reestimate $\hat{\varphi}^{(2)}$
5. proceed with the above steps....

Example (Ice cream with autocorrelation)

```
> summary(Zice)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1973151  0.2702162   0.730  0.47179
income       0.0033078  0.0011714   2.824  0.00899 **
price       -1.0444140  0.8343573  -1.252  0.22180
temp         0.0034584  0.0004455   7.762  3.1e-08 ***
---
Residual standard error: 0.03683 on 26 degrees of freedom
Multiple R-squared:  0.719,Adjusted R-squared:  0.6866
F-statistic: 22.17 on 3 and 26 DF,  p-value: 2.451e-07
```

```
> Zice.ar1 = gls(cons ~ ., data=Icecream, correlati
> summary(Zice.ar1)
Generalized least squares fit by REML
  Model: cons ~ .
  Data: Icecream
        AIC       BIC logLik
  -84.73199 -77.18341 48.366

Correlation Structure: AR(1)
 Formula: ~1
 Parameter estimate(s):
      Phi
0.9112057

Coefficients:
                 Value Std.Error   t-value p-value
(Intercept)  0.6583509 0.2948486  2.232844  0.0344
income      -0.0016118 0.0021112 -0.763442  0.4521
price       -0.9795943 0.7320736 -1.338109  0.1924
temp         0.0028192 0.0007224  3.902384  0.0006
Residual standard error: 0.07878502
Degrees of freedom: 30 total; 26 residual
```

Possible alternative solution: add a lagged x-variable to the models, e.g. lagged temperature