

Econometrics



David Preinerstorfer

Institute for Statistics and Mathematics
Dept. of Finance, Accounting and Statistics

Spring 2026



Plan for today:

1. Consistency.
2. Omitted variables, variable selection.
3. Regression diagnostics.
4. Taking care of heteroskedasticity and autocorrelation.

Let $\hat{\beta}_N$ be (an arbitrary) estimator for β , based on sample size N . As illustrated above for the OLS estimator, for finite sample size N , $\hat{\beta}_N$ typically deviates from β .

To understand the statistical properties of an estimators, the sampling distribution of the random variable (r.v.) $\hat{\beta}_N$ is studied under the assumption that the data were generated by a given model with the true parameter being equal to β .

A minimum requirement for any estimator is asymptotic unbiasedness, also called consistency.

$\hat{\beta}_N$ is a **consistent** estimator for β , if the r.v. $\hat{\beta}_N$ converges to β in probability, i.e. for every $\epsilon > 0$, the following holds:

$$\lim_{N \rightarrow \infty} P\{|\hat{\beta}_N - \beta| \geq \epsilon\} = 0.$$

or, equivalently, $p \lim \hat{\beta}_N = \beta$.

- Consistency means that an estimator converges „in probability” to the true value with increasing number of observations N .
- A sufficient condition for convergence in probability is the following:

$$\lim_{N \rightarrow \infty} E(\hat{\beta}_N) = \beta, \quad \lim_{N \rightarrow \infty} V(\hat{\beta}_N) = 0.$$

Hence, also a biased estimators can be consistent, as long as the variance decreases sufficiently fast with increasing N .

Theorem: Let $\hat{\theta}_n$ be a sequence of estimators such that

$$E(\hat{\theta}_n) \rightarrow \theta \quad \text{and} \quad \text{Cov}(\hat{\theta}_n) \rightarrow 0.$$

Then $\hat{\theta}_n \rightarrow \theta$ in probability.

Proof: Let $\varepsilon > 0$ be fixed, and note that (triangle inequality)

$$P(|\hat{\theta}_n - \theta| \geq \varepsilon) \leq P\left(|\hat{\theta}_n - E(\hat{\theta}_n)| + |E(\hat{\theta}_n) - \theta| \geq \varepsilon\right),$$

which we can bound from above by

$$P\left(|\hat{\theta}_n - E(\hat{\theta}_n)| \geq \varepsilon/2\right) + P\left(|E(\hat{\theta}_n) - \theta| \geq \varepsilon/2\right).$$

The second probability converges to 0 (eventually, the number $|E(\hat{\theta}_n) - \theta|$ is smaller than $\varepsilon/2$).
The first probability is (by Markov's inequality) upper bounded by

$$4E(|\hat{\theta}_n - E(\hat{\theta}_n)|^2)/\varepsilon^2 = 4\text{trace}(\text{Cov}(\hat{\theta}_n))/\varepsilon^2 \rightarrow 0.$$

Consistency of the OLS estimator

For each $j = 1, \dots, K$, consistency follows under **AL**, **AR**, **AX** and **AH**:

- The OLS estimator is unbiased, i.e. $E(\hat{\beta}_j) = \beta_j$.
- The variance $V(\hat{\beta}_j)$ goes to 0 for $N \rightarrow \infty$:

$$\lim_{N \rightarrow \infty} V(\hat{\beta}_j) = \frac{\sigma^2}{N s_{x_j}^2 (1 - R_j^2)} = 0,$$

provided that $\lim_{N \rightarrow \infty} R_j^2 < 1$ (no perfect correlation with any other regressor) and $\lim_{N \rightarrow \infty} s_{x_j}^2 > 0$.

Assumption **AN** is not necessary to prove consistency.

The above argument requires the additional conditions that $\lim_{N \rightarrow \infty} R_j^2 < 1$ and $\lim_{N \rightarrow \infty} s_{x_j}^2 > 0$, and formally works only if X is non-random.

Note that we have already shown (under AL, AR, AX, and AH, and assuming that $E(X_i' X_i)$ is invertible) that in distribution

$$\sqrt{N}(\hat{\beta} - \beta) \rightarrow \text{Normal}_{K+1}(0, E(X_i' X_i)^{-1}).$$

Noting that

$$\hat{\beta} - \beta = \frac{1}{\sqrt{N}} \times \sqrt{N}(\hat{\beta} - \beta),$$

that $\frac{1}{\sqrt{N}} \rightarrow 0$ and Slutsky's lemma, we obtain $\hat{\beta} \rightarrow \beta$ in probability.

Alternatively: write $\hat{\beta}_n = \beta + (X'X)^{-1}X'\varepsilon$, recall that $(N^{-1}X'X)^{-1} \rightarrow E(X_i' X_i)^{-1}$ and observe that $N^{-1}X'\varepsilon = N^{-1} \sum_{i=1}^N X_i' \varepsilon_i \rightarrow 0$ (LLN).

Omitted and irrelevant predictors

Relevant variables may be omitted by mistake or by lack of data. Omitting variables has serious consequences. Suppose the following model is correct, and that assumption **AX** holds in that model:

$$y = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \varepsilon,$$

but the model was estimated without $\mathbf{X}_2$:

$$y = \mathbf{X}_1\beta_1 + \varepsilon_1, \quad \varepsilon_1 = \mathbf{X}_2\beta_2 + \varepsilon. \quad (51)$$

Hence, the error term in the misspecified model (51) captures the effects of the missing covariate.

The OLS estimator of β_1 from the misspecified model (51) is given by:

$$\hat{\beta}_1 = (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 \mathbf{y} = \beta_1 + \mathbf{H}_1 \varepsilon_1,$$

where $\mathbf{H}_1 = (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1$.

However, assumption **AX** is violated in model (51), unless the true value of β_2 is 0 or regressors $\mathbf{X}_1$ and $\mathbf{X}_2$ are orthogonal (i.e., uncorrelated), since:

$$E(\mathbf{X}'_1 \varepsilon_1 \mid \mathbf{X}_1) = E(\mathbf{X}'_1 \mathbf{X}_2 \beta_2 \mid \mathbf{X}_1) + E(\mathbf{X}'_1 \varepsilon \mid \mathbf{X}_1) = E(\mathbf{X}'_1 \mathbf{X}_2 \mid \mathbf{X}_1) \beta_2.$$

Thus $\hat{\beta}_1$ is **biased and inconsistent**, if omitted regressors are correlated with included ones. The omitted variable bias reads:

$$E(\hat{\beta}_1 | \mathbf{X}_1) = \beta_1 + E(\mathbf{H}_1 \varepsilon_1 | \mathbf{X}_1) = \beta_1 + E(\mathbf{B} | \mathbf{X}_1) \beta_2,$$

where $\mathbf{B}$ results from a regression of the omitted regressor $\mathbf{X}_2$ on the included regressors $\mathbf{X}_1$:

$$\mathbf{B} = (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbf{X}_2.$$

Since $\mathbf{AX}$ holds in the full model, this relation implies:

$$\hat{\beta}_1 = \hat{\beta}_1^f + \mathbf{B} \hat{\beta}_2^f,$$

where $\hat{\beta}_1^f$ and $\hat{\beta}_2^f$ are the estimators obtained from the full model.

Furthermore, the standard errors $\text{se}(\hat{\beta}_1)$ of $\hat{\beta}_1$ will be biased, since the error variance $\hat{\sigma}^2$ estimated from the misspecified model is biased.

Even if the omitted variable bias disappears, there are issues with the variance of the error term ε_1 in the misspecified model:

- if the regressors $\mathbf{X}_2$ are random, then the variance is heteroskedastic, since $V(\varepsilon_i) = \beta_2' \text{Cov}(\mathbf{x}_{2i}) \beta_2 + \sigma^2$, where $\mathbf{x}_{2i}$ is the i th row of $\mathbf{X}_2$.

Suppose that the correctly specified model is given by:

$$y = \mathbf{X}_1\beta_1 + \varepsilon,$$

but an **overfitting** model was estimated including the irrelevant variable $\mathbf{X}_2$, i.e. $\beta_2 = 0$:

$$y = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \varepsilon_2. \quad (52)$$

Including irrelevant variables leads to unbiased and consistent estimates, which tend to be inefficient.

Statistical hypotheses testing can be applied to test for the presence of irrelevant predictors.

Inefficiency of OLS under irrelevant predictors

Inefficiency is implied by the Gauss Markov theorem. From the overfitting model (52), we obtain the following OLS estimator of $\beta = (\beta_1, \beta_2)$:

$$\hat{\beta}^C = \mathbf{H}\mathbf{y}, \quad \mathbf{H} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \begin{pmatrix} \mathbf{C} \\ \mathbf{D} \end{pmatrix}$$

$\hat{\beta}_1^C$ is an unbiased estimator of β_1 given by:

$$\hat{\beta}_1^C = \mathbf{C}\mathbf{y},$$

which is usually different from the OLS estimator $\hat{\beta}_1$ of the correct model. Hence, $\hat{\beta}_C$ is less efficient than $\hat{\beta}_1$.

Irrelevant predictors III

The only situation where no loss of efficiency occurs, is if $\mathbf{X}_1$ and $\mathbf{X}_2$ are uncorrelated (orthogonal, i.e. $\mathbf{X}_1'\mathbf{X}_2 = 0$). In this case:

$$\begin{aligned}
 (\mathbf{X}'\mathbf{X})^{-1} &= \begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2'\mathbf{X}_2 \end{pmatrix}^{-1} \\
 &= \begin{pmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{X}_2'\mathbf{X}_2)^{-1} \end{pmatrix}.
 \end{aligned}$$

Hence, $\mathbf{C} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1$ and $\hat{\beta}_1^C$ is equal to the OLS estimator $\hat{\beta}_1$ of the correct model.

- The search for a correct specification of a regression model is usually difficult.
- Two ways possible: specific to general or general to specific. The second is preferable since the omission of relevant variables is a more serious issue.
- It is always recommended that variable selection on the basis of sound theory.
- If the p values of a coefficient is above the significance level the associated variable may be eliminated. If there are multiple, then start with the one with highest p value and re-estimate the model.
- Don't be misled by insignificant p -values (e.g. $p\text{-value} > 0.05$). If a variable is considered of key importance theoretically, keep it in the model. A failure to show significance occurs for small N .
- Significant variable should have the expected sign.

Variable selection by overfitting

There is a trade-off between including irrelevant variables (inefficiency) and omitting relevant variables (bias and inconsistency).

Due to unbiasedness, it is preferable to start with an overfitting model, rather than a simple model (which might suffer from omitted variable bias):

- Test the null hypothesis $\beta_j = 0$ for each coefficient.
- Use the Wald statistic or the F-test for testing $\beta_{j_i} = 0$ simultaneously for more than one coefficient.

Model Comparison Using R^2

Coefficient of determination R^2 (SST is the squared sum of residuals of the simple model without predictor):

$$R^2 = \frac{SST - SSR}{SST} = 1 - \frac{SSR}{SST}$$

Coefficient of determination R^2 useful for model evaluation (the closer to 1, the better).

However, variable selection using the coefficient of determination R^2 is not recommended

Problems with R^2

- Choosing the model with the smallest SSR (largest R^2) leads to overfitting: R^2 increases with increasing number of variables as SSR decreases.
- R^2 is 1 for $K = N - 1$, because $SSR = 0$, if we include as many predictors as observations, even if the predictors are useless.
- The increase of adding a useless predictor, however, is small $\Rightarrow$ penalize the ever decreasing SSR by including the number of parameters used for estimation which is an increasing function of the number of parameters.

One may use **adjusted** R^2 instead:

$$\bar{R}^2 = 1 - \frac{n-1}{n-K}(1 - R^2).$$

Example: Demand for Chicken (log-linear model)

Predictor included	SSR	R ²
pchick	0.273487	0.647001
income	0.041986	0.945807
income, pchick	0.015437	0.980074
income, pchick, ppork	0.014326	0.981509
income, pchick, ppork, pbeef	0.013703	0.982313

Model choice criteria

Definition of model choice criteria - model fit + penalty:

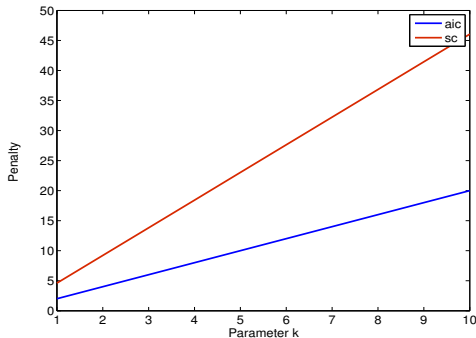
$$AIC = \log(\text{SSR}/N) + \frac{2 \cdot (K + 1)}{N},$$

$$SC = \log(\text{SSR}/N) + \frac{\log N \cdot (K + 1)}{N}.$$

Choose the model that minimizes a particular criterion.

Penalty for model fit is proportional to the number of parameters $K + 1$.

The various criteria may lead to different choices; SC has the larger and AIC has the smaller penalty for the same model size k , if the number of observations $N \geq 8 (> e^2)$. Comparing AIC and SC for $N = 100$; penalty plotted as a function of number k of parameters:



Example: Demand for Chicken (log-linear model) I

Predictor included	SSR	R ²	AIC	SC
pchick	0.273487	0.647001	-1.420206	-1.321468
income	0.041986	0.945807	-3.294124	-3.195386
income, pchick	0.015437	0.980074	-4.207711	-4.059603
income, pchick, ppork	0.014326	0.981509	-4.195488	-3.998011
income, pchick, ppork, pbeef	0.013703	0.982313	-4.152987	-3.906140

Example: Demand for Chicken (log-linear model) II

Non-nested models can be compared using AIC and SC

Predictor included	R^2	AIC	SC
pchick	0.705529	5.839137	5.765231
income, pchick	0.9108	4.633	4.781
income, pchick, ppork	0.936653	4.377760	4.575237
income, pchick, ppork, pbeef	0.942580	4.366473	4.613320
interaction term	0.972183	3.641719	3.888565
quadratic model	0.977124	3.533149	3.829365

Comparing linear and log-linear models

The residual sum of squares SSR depends on the scale of y_i , therefore AIC and SC are scale dependent

AIC and SC could not be used directly to compare a linear and a log-linear model.

AIC and SC of the log-linear model could be matched back to the original scale by adding 2 times the mean of the logarithmic values of y_i .

Correction formula:

$$AIC = AIC^* + \frac{2}{N} \sum_{i=1}^N \log(y_i)$$

$$SC = SC^* + \frac{2}{N} \sum_{i=1}^N \log(y_i)$$

AIC^* and SC^* are the model choice criteria for the log-linear model

Example: Demand for Chicken

Predictor included	SSR	R ²	AIC	SC
income, pchick (log-linear)	0.015437	0.980074	-4.207711	-4.059603
income, pchick (linear)	106.65	0.9108	4.633	4.781

Transform AIC and SC of the log-linear model: $AIC = -4.207711 + 2 \cdot 3.663887 = 3.1201$

$SC = -4.059603 + 2 \cdot 3.663887 = 3.2682$ log-linear model preferred

- Statistical Properties of OLS Estimation
- Specifications
- Regression diagnostics
- Time Series
- ARMA models

Hypothetical model:

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_K X_K + \varepsilon.$$

Estimated model:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \dots + \hat{\beta}_K x_{K,i} + \hat{\varepsilon}_i,$$

where $\hat{\varepsilon}_i$ is the OLS residual. Due to consistency, the OLS residual $\hat{\varepsilon}_i$ approaches the unobservable error ε_i , as N increases.

Use OLS residuals $\hat{\varepsilon}_i$ to test assumptions about ε_i .

Typical violations of assumption **AX** occur for the following reasons:

- ε_i is correlated with one of the regressors X_j (endogeneity)
- Reverse causality: X_j affects Y and Y simultaneously affects X_j .
- Misspecification: a relevant variables is omitted from the model or the functional form is misspecified. The true value of y_i at $X_j = x_{ji}$ will be
 - underrated, if $E(\varepsilon_i | X_j = x_{ji}) > 0$;
 - overrated, if $E(\varepsilon_i | X_j = x_{ji}) < 0$.

There is no simple way to test assumption **AX**.

Informal diagnostics:

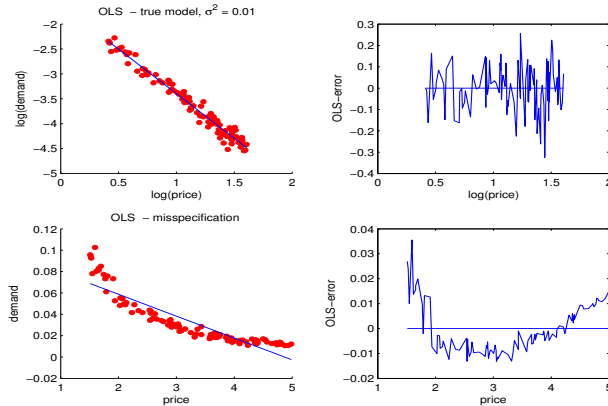
- Plot OLS-residuals versus the predictors X_j .

Example: simulate data from a simple log-linear regression model with $\tilde{\beta}_1 = 0.2$ and $\beta_2 = -1.8$:

$$y_i = 0.2 \cdot x_i^{-1.8} e^{\varepsilon_i},$$

- Residual plot over x_i for the linear regression model shows specification error.

Regression diagnostics III



Assumption AN implies: marginally, the error ε_i follows a normal distribution:

$$\varepsilon_i \sim \text{Normal}(0, \sigma^2).$$

- Roughly 95% of the OLS residuals lie between $[-2\hat{\sigma}, 2\hat{\sigma}]$; only 5% lie outside
- Assumption often violated, if outliers are present.
- Normality often improved through transformations.

Testing normality II

To test normality of ε_i , check normality of the OLS residuals $\hat{\varepsilon}_i$:

- Histogram
- Skewness coefficient m_3 close to 0?
- Kurtosis coefficient m_4 close to 3?

where

$$m_3 = \frac{1}{\hat{\sigma}^3} \left(\frac{1}{N} \sum_{i=1}^N \hat{\varepsilon}_i^3 \right), \quad m_4 = \frac{1}{\hat{\sigma}^4} \left(\frac{1}{N} \sum_{i=1}^N \hat{\varepsilon}_i^4 \right).$$

Jarque-Bera-Statistics:

$$J = \frac{N - K}{6} \left(m_3^2 + \frac{1}{4}(m_4 - 3)^2 \right). \quad (53)$$

- Null hypothesis H_0 : the errors follow a normal distribution
- Under H_0 , J follows asymptotically (i.e., N large) a χ_2^2 -distribution with 2 degrees of freedom (95%-quantile: $\chi_{2,0.95}^2 = 5.9915$).
- Reject H_0 , if p -value of J is close to 0.

Example: Firm Profits

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i, \quad (54)$$

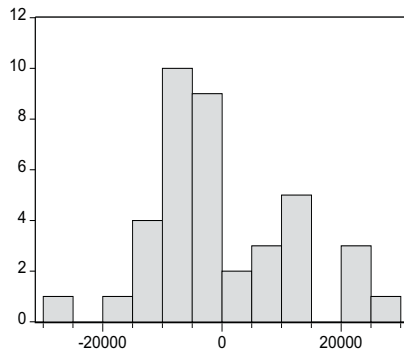
where

y_i ... profit 1994

$x_{1,i}$... profit 1993

$x_{2,i}$... turnover 1994

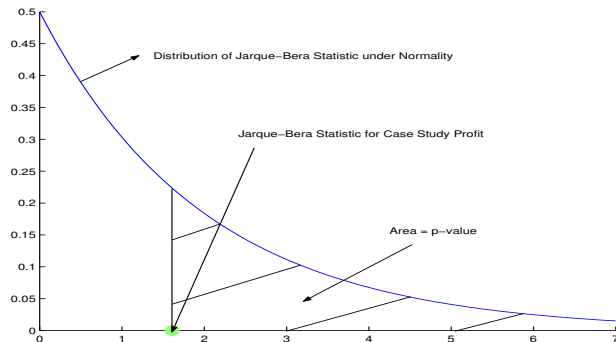
Consider only large firms ($i = 1, \dots, 39$)



Series: Residuals
Sample 1 39
Observations 39

Mean	-5.36E-13
Median	-2081.893
Maximum	29116.01
Minimum	-26654.90
Std. Dev.	11933.24
Skewness	0.497860
Kurtosis	3.010465

Jarque-Bera	1.611298
Probability	0.446798



Example: Yields

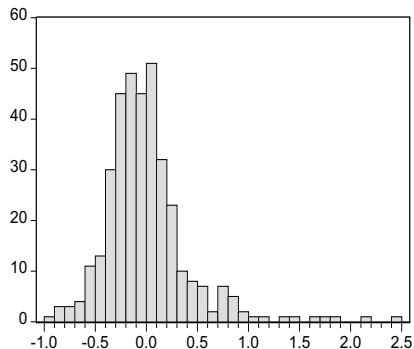
$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i, \quad (55)$$

where

y_i ... yield with maturity 3 months

$x_{2,i}$... yield with maturity 1 month

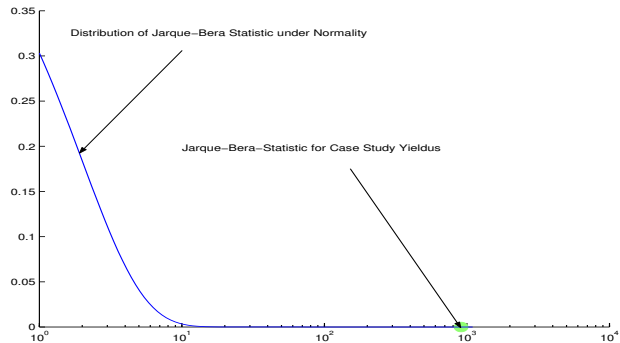
$x_{3,i}$... yield with maturity 60 months



Series: Residuals
Sample 1964:01 1993:12
Observations 360

Mean	-4.93E-18
Median	-0.047621
Maximum	2.454587
Minimum	-0.909939
Std. Dev.	0.419164
Skewness	1.872579
Kurtosis	9.829840

Jarque-Bera	910.0937
Probability	0.000000



Checking homoskedasticity

Assumption **AH** is often violated, because the variance $V(\varepsilon_i)$ is not constant over all ε_i . A typical deviation from assumption **AH** is that the errors are uncorrelated, but the variance is heteroskedastic:

$$\begin{aligned}
 V(\varepsilon_i | \mathbf{X}) &= \omega_i \sigma^2, \quad i = 1, \dots, N \\
 \text{Cov}(\varepsilon | \mathbf{X}) &= \sigma^2 \mathbf{\Omega}, \quad \mathbf{\Omega} = \text{Diag}(\omega_1 \dots \omega_N).
 \end{aligned}
 \tag{56}$$

Informal check: residual plot; formal tests are available to test, if $V(\varepsilon_i | \mathbf{X})$ depends on some of the regressors in $\mathbf{X}$.

The OLS estimator is unbiased (provided that **AX** holds), however, the covariance matrix of the sampling distribution is different from (33) and standard errors and p-values based on (33) are wrong, if assumption **AH** is violated.

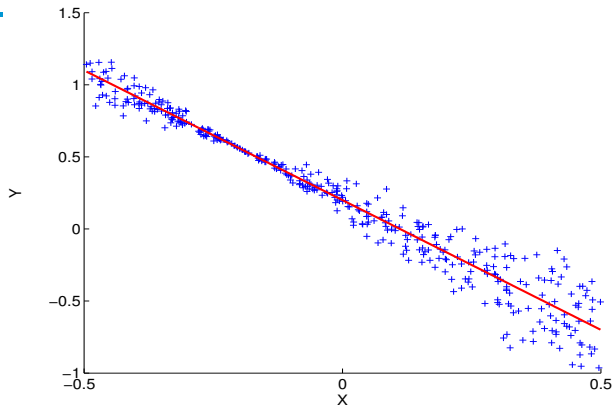
Using (34) (with $\mathbf{H} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$), we obtain

$$\text{Cov}(\hat{\beta} \mid \mathbf{X}) = \mathbf{H}\text{Cov}(\varepsilon \mid \mathbf{X})\mathbf{H}' = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{\Omega}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}. \quad (57)$$

Correct standard errors and p-values (if assumption **AN** holds) could be based on (57), however, the OLS estimator is no longer efficient, because it does not use all information. More efficient estimator are available, if we have knowledge about $\mathbf{\Omega}$.

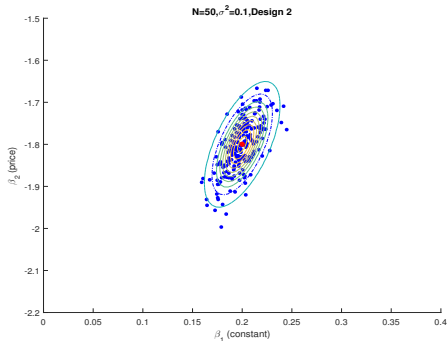
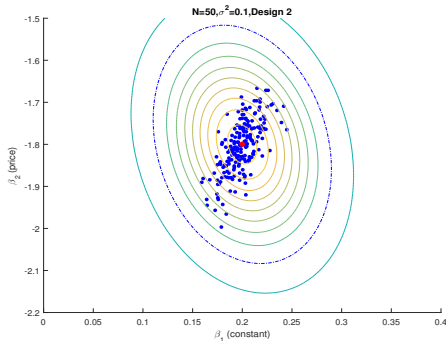
Simulate data from a regression model with $\beta_0 = 0.2$ and $\beta_1 = -1.8$ and heteroskedastic errors:

$$y_i = 0.2 - 1.8x_i + \varepsilon_i, \quad \varepsilon_i \sim \text{Normal}(0, \sigma_i^2),$$
$$\sigma_i^2 = \sigma^2 \cdot \omega_i, \quad \omega_i = (0.2 + x_i)^2.$$



Sampling distribution of the OLS estimator

Simulation study with 200 data sets (each $N = 50$ observations). True sampling distribution of the OLS estimator compared to sampling distribution derived under assumption **AH** (left hand side) and under the true covariance matrix Ω using (57) (right hand side)



Testing for Heteroskedasticity I

Tests for heteroskedasticity are based on the squared OLS-residuals $\hat{\varepsilon}_i^2$. The ***Breusch-Pagan test for heteroskedasticity*** is based on an auxiliary regression of $\hat{\varepsilon}_i^2$ on a constant and the regressors:

$$\hat{\varepsilon}_i^2 = \gamma_0 + \gamma_1 x_{1,i} + \dots + \gamma_K x_{K,i} + \xi_i. \quad (58)$$

Test the null hypothesis $\gamma_1 = \dots = \gamma_K = 0$ using the F-test for the entire regression model.

A related test statistic is $N \cdot R^2$, where R^2 is the coefficient of determination of the auxiliary regression model (58), which follows a χ_K^2 distribution, if the null hypothesis is true.

Testing for Heteroskedasticity II

The *White-test for heteroskedasticity* is based on an auxiliary regression of $\hat{\varepsilon}_i^2$ on not only the predictor variables, but also on the squared predictors.

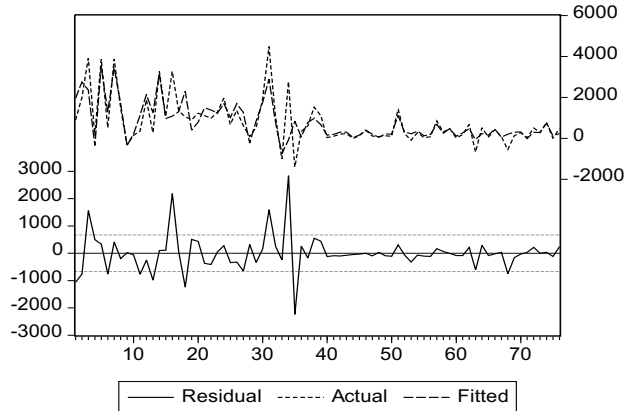
In a more general version of the White test, also cross-terms for all predictors are included.

Test statistics for White test are, again, the F-test for the entire regression model or $N \cdot R^2$, which follows a $\chi_{K\star}^2$ distribution under the null hypothesis, with $K\star$ being the number of parameters in auxiliary regression (excluding the constant).

As certain problem with these tests is, that the errors ξ_i of the auxiliary regression are typically not normal, hence p -value of the corresponding coefficients are only approximate.

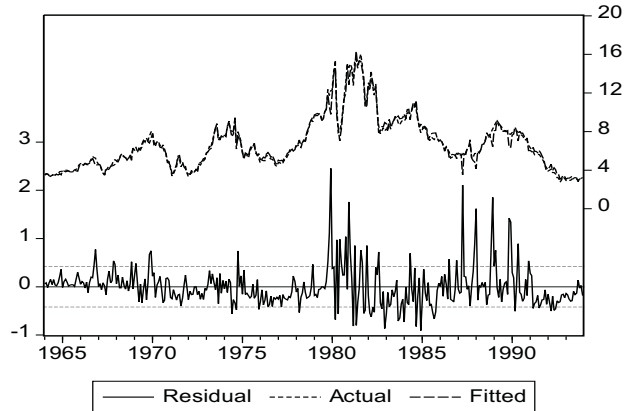
Testing for Heteroskedasticity III

In the full sample (76 firms), $\hat{\varepsilon}_i^2$ depends on the size of the firm



Testing for Heteroskedasticity IV

In the Example: Yields, $\hat{\varepsilon}_i^2$ exhibits volatility clusters:



White Heteroskedasticity consistent estimator

The *White heteroskedasticity consistent estimator* corrects the standard errors of the OLS estimator under heteroskedasticity, where Ω is unknown.

In (57), the term $\mathbf{X}'\Omega\mathbf{X}$ is estimated as:

$$\mathbf{X}'\Omega\mathbf{X} = \sum_{i=1}^N \sigma_i^2 \mathbf{x}_i' \mathbf{x}_i \approx \frac{N}{N - (K + 1)} \sum_{i=1}^N \hat{\varepsilon}_i^2 \mathbf{x}_i' \mathbf{x}_i.$$

where $\mathbf{x}_i$ is a row vector containing the covariates of y_i . Note that $\hat{\varepsilon}_i^2$ is a (one-point) estimator of σ_i^2 .

Use this estimator in t-statistics, F-statistics, confidence intervals, etc. Hugely influential!

A more efficient estimator than OLS is available, if Ω is known.

Assume the variance increases with an observed variable Z_i , i.e., $V(\varepsilon_i) = \sigma^2 \omega_i$, with $\omega_i = Z_i$.

A simple modification of regression model

$$y_i = \beta_0 + \beta_1 x_{1,i} + \dots + \beta_K x_{K,i} + \varepsilon_i,$$

leads to a model with homoskedastic variances:

$$\frac{y_i}{\sqrt{\omega_i}} = \beta_0 \frac{1}{\sqrt{\omega_i}} + \beta_1 \frac{x_{1,i}}{\sqrt{\omega_i}} + \dots + \beta_K \frac{x_{K,i}}{\sqrt{\omega_i}} + \varepsilon_i^*, \quad (59)$$

where $\varepsilon_i^* = \frac{\varepsilon_i}{\sqrt{\omega_i}}$ and $V(\varepsilon_i^*) = V(\varepsilon_i) / \omega_i = \sigma^2$.

Regression model (59) has identical parameters as the original model, but a modified response variable as well as modified predictors:

$$y_i^* = \beta_0 x_{0,i}^* + \beta_1 x_{1,i}^* + \dots + \beta_K x_{K,i}^* + \varepsilon_i^*, \quad (60)$$

where

$$y_i^* = \frac{y_i}{\sqrt{\omega_i}}, \quad x_{0,i}^* = \frac{1}{\sqrt{\omega_i}}, \quad x_{j,i}^* = \frac{x_{j,i}}{\sqrt{\omega_i}}, \quad \forall j = 1, \dots, K.$$

Note that model (60) *fulfills* assumption **AH**, i.e., it is a model with homoskedastic errors.

OLS estimation for the modified regression model (60) is the BLUE for β :

$$\hat{\beta} = ((\mathbf{X}^*)' \mathbf{X}^*)^{-1} (\mathbf{X}^*)' \mathbf{y}^*,$$

where

$$\mathbf{X}^* = \text{Diag}(\lambda_1 \cdots \lambda_N) \mathbf{X}, \quad \mathbf{y}^* = \text{Diag}(\lambda_1 \cdots \lambda_N) \mathbf{y}.$$

with $\lambda_i = 1/\sqrt{\omega_i}$, $i = 1, \dots, N$.

This is equal to following WLS estimator, which is expressed entirely in terms of the original variables:

$$\hat{\beta}_{\text{WLS}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}, \quad (61)$$

where $\mathbf{W} = \text{Diag}(\lambda_1^2 \cdots \lambda_N^2) = \text{Diag}(1/\omega_1 \cdots 1/\omega_N) = \mathbf{\Omega}^{-1}$.

The covariance matrix of the sampling distribution of $\hat{\beta}_{\text{WLS}}$ is given by:

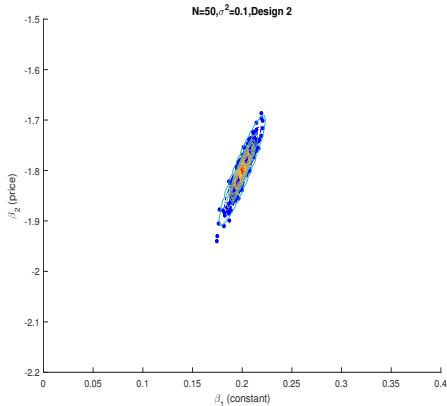
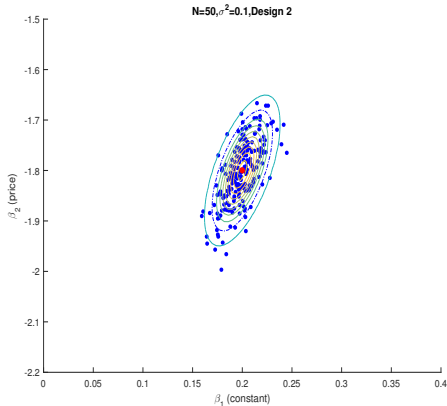
$$\text{Cov}(\hat{\beta}_{\text{WLS}}) = \sigma^{\hat{2},*}((\mathbf{X}^*)'\mathbf{X}^*)^{-1} = \sigma^{\hat{2},*}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1},$$

where $\sigma^{\hat{2},*}$ is estimated from the modified regression (60).

Sampling distribution of the WLS estimator

Simulation study with 200 data sets (each $N = 50$ observations), true covariance matrix Ω assumed to be known. Sampling distribution of the OLS estimator (left hand side) and the weighted least square estimator (right hand side)

Generalized least square under heteroskedasticity VI



The WLS estimator minimizes the sum of squared residuals in the modified model:

$$\text{SSR} = \sum_{i=1}^N (\varepsilon_i^*)^2 = \sum_{i=1}^N \varepsilon_i^2 \frac{1}{\omega_i}.$$

Hence, in the original regression model, residuals ε_i for observations with big variances $V(\varepsilon_i) = \sigma^2 \omega_i$ are down-weighted, while residuals for observations with small variances obtain a higher weight. Hence the name *weighted least square estimation*.

Autocorrelation

Assumption **AH** is often violated, because the errors ε_i are correlated, i.e.:
 $\text{Cov}(\varepsilon | \mathbf{X}) = \sigma^2 \mathbf{\Omega}$, where some (or all) off-diagonal elements of $\mathbf{\Omega}$ are different from 0. Mostly relevant for time series data.

The OLS estimator is unbiased (provided that **AX** holds), however, the covariance matrix of the sampling distribution is different from (33) and standard errors and p-values based on (33) are wrong, if the error are correlated.

Correct standard errors and p-values (if assumption **AN** holds) could be based on (57), if we have knowledge about $\mathbf{\Omega}$.

Typically, one tries to estimate $\mathbf{\Omega}$ and then uses this estimate as a plug-in for estimating the covariance matrix of the OLS estimator.

More details are provided in Part II, dealing with time series.