

Econometrics



David Preinerstorfer

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

Spring 2026



Recap:

1. Linear models
2. OLS estimator $(X'X)^{-1}X'Y$, predicted values $\hat{Y} = X\hat{\beta}$, residuals $\hat{\varepsilon} = Y - \hat{Y}$.
3. Assumptions AL, AR, AX, AH (below); can you recall them?
4. Properties: sum of squares decomposition $TSS = RSS + ESS$, $R^2 = ESS/TSS$, consequences of the *normal equation* $0 = X'Y - X'X\hat{\beta} = X'\hat{\varepsilon}$.

We recall briefly the main assumptions from slides 37, 44, 72, 75, and 77:

The multiple regression model describes the relation between the response variable Y and the predictor variables $X_1, \dots, X_K$ through a linear assumption:

Assumption AL:

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_K X_K + \varepsilon, \quad (12)$$

$\beta_0, \beta_1, \dots, \beta_K$ are unknown parameters.

A basic assumption

The matrix $\mathbf{X}'\mathbf{X}$ is a quadratic matrix with $(K + 1)$ rows and columns.

Assumption AR (rank assumption):

We assume that $\mathbf{X}$ has full rank equal to $(K + 1)$ (i.e., the columns of $\mathbf{X}$ are linearly independent).

If $\mathbf{X}$ has full rank, then $\mathbf{X}'\mathbf{X}$ is positive definite and the inverse matrix $(\mathbf{X}'\mathbf{X})^{-1}$ of $\mathbf{X}'\mathbf{X}$ exists.

Assumption AX: strict exogeneity:

the conditional expectation of ε conditional on all observations and variables in $\mathbf{X}$ is zero:

$$E(\varepsilon | \mathbf{X}) = \mathbf{0}, \quad E(\varepsilon_i | \mathbf{X}) = 0. \quad (25)$$

Knowing $\mathbf{X}$ does not give us any information about ε .

The term **endogeneity** (i.e., one or all explanatory variables are endogenous) is used, if **AX** does not hold (broadly speaking, if $\mathbf{X}$ and ε are correlated).

Assumption AH (homoskedasticity, assuming that AX holds):

$$V(\varepsilon_i | \mathbf{X}) = E(\varepsilon_i^2 | \mathbf{X}) - E(\varepsilon_i | \mathbf{X})^2 = E(\varepsilon_i^2 | \mathbf{X}) = \sigma^2. \quad (29)$$

If assumption (29) is violated, e.g., if $V(\varepsilon | \mathbf{X}) = \sigma^2 h(\mathbf{X})$, then we say that the error term is *heteroskedastic*.

Assumption AH :(Uncorrelatedness, assuming that AX holds):

$$\begin{aligned}\text{Cov}(\varepsilon_i, \varepsilon_j \mid \mathbf{X}) &= E(\varepsilon_i, \varepsilon_j \mid \mathbf{X}) - E(\varepsilon_i \mid \mathbf{X})E(\varepsilon_j \mid \mathbf{X}) \\ &= E(\varepsilon_i, \varepsilon_j \mid \mathbf{X}) = 0, \\ \text{Cov}(\boldsymbol{\varepsilon} \mid \mathbf{X}) &= E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}) = \sigma^2 \mathbf{I},\end{aligned}\tag{30}$$

where $\mathbf{I}$ is the identity matrix.

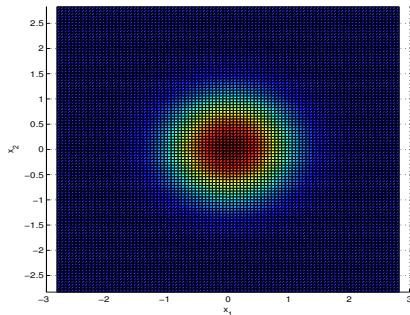
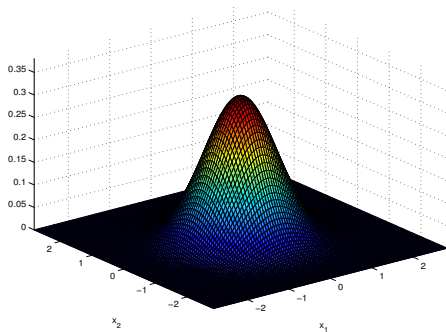
This means that the errors are uncorrelated, regardless of the predictor variables $X_1, \dots, X_K$.

Today:

1. Statistical properties of the OLS estimator.
2. Multicollinearity.
3. Estimating σ^2 .
4. Distribution of the OLS estimator.
5. Gauss-Markov-Theorem.
6. Testing, confidence intervals, prediction.

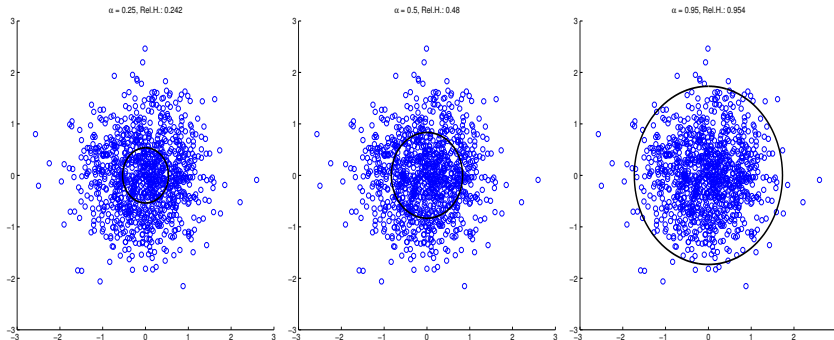
Multivariate normal distributions - independent components I

Density of the bivariate normal distribution $\text{Normal}_2(\mathbf{0}, \sigma^2 \mathbf{I})$ with $\sigma^2 = 0.5$.



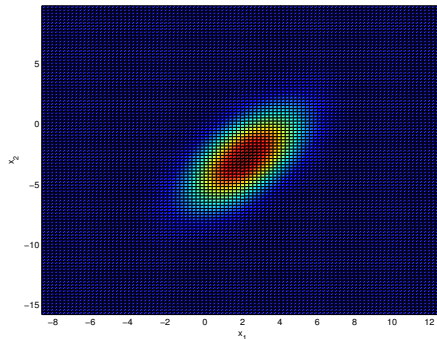
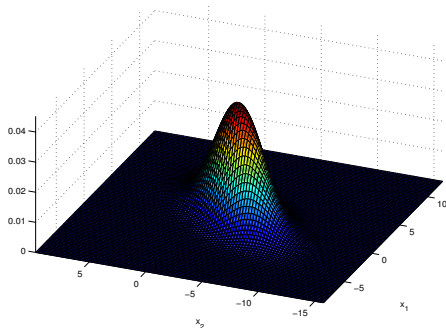
1000 observations from $\text{Normal}_2(\mathbf{0}, 0.5\mathbf{I})$ in comparison to 100%-confidence region
(from the left to the right: $\alpha = 0.25, \alpha = 0.5, \alpha = 0.95$)

Multivariate normal distributions - independent components II



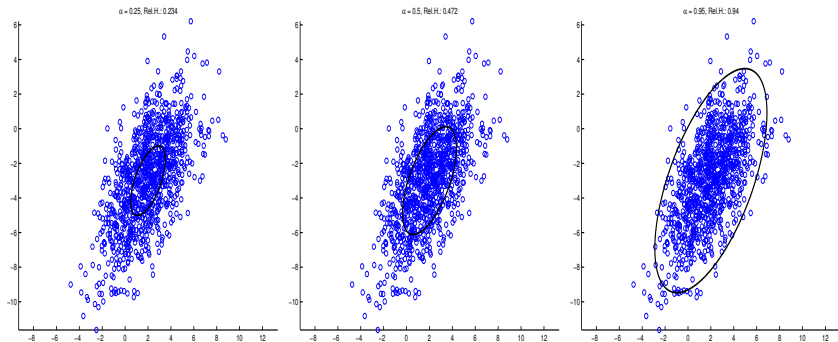
Multivariate normal distributions - dependent components I

Density of the bivariate normal distribution $\text{Normal}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = (2, -3)'$ and $\boldsymbol{\Sigma} = \begin{pmatrix} 4 & 3.2 \\ 3.2 & 7 \end{pmatrix}$



Multivariate normal distributions - dependent components II

1000 observations from $\text{Normal}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ in comparison to $100\alpha\%$ -confidence region
(from the left to the right: $\alpha = 0.25, \alpha = 0.5, \alpha = 0.95$)



Assumption AN (normality):

assumptions **AX** and **AH** imply that the mean and variance of ε are $\mathbf{0}$ and $\sigma^2 \mathbf{I}$. Adding the assumption of normality we have that the vector ε follows a multivariate normal distribution with independent components:

$$\varepsilon \mid \mathbf{X} \sim \text{Normal}_N(\mathbf{0}, \sigma^2 \mathbf{I}).$$

The error ε in the multiple regression model is independent of $X_1, \dots, X_K$ and follows a normal distribution:

$$\varepsilon \sim \text{Normal}(0, \sigma^2). \quad (31)$$

It follows that the conditional distribution of Y given $X_1, \dots, X_K$ follows a normal distribution:

$$Y \mid X_1, \dots, X_K \sim \text{Normal}(\beta_0 + \beta_1 X_1 + \dots + \beta_K X_K, \sigma^2).$$

If assumption **AL** holds, i.e., the data are generated by the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, then the OLS estimator $\hat{\boldsymbol{\beta}}$ may be expressed as:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \\ &= \boldsymbol{\beta} + \mathbf{H}\boldsymbol{\varepsilon}, \quad \mathbf{H} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}',\end{aligned}\tag{32}$$

provided that assumption **AR** holds. This representation allows to derive the sampling distribution of the OLS $\hat{\boldsymbol{\beta}}$.

Unbiasedness of the OLS estimator I

The *conditional expectation* $E(\hat{\beta} \mid \mathbf{X})$ for *fixed* $\mathbf{X}$ is given by:

$$E(\hat{\beta} \mid \mathbf{X}) = \beta + E(\mathbf{H}\epsilon \mid \mathbf{X}) = \beta + \mathbf{H}E(\epsilon \mid \mathbf{X}).$$

Under assumption **AX**, $E(\epsilon \mid \mathbf{X}) = 0$, hence the OLS estimator is unbiased, i.e.,

$$E(\hat{\beta} \mid \mathbf{X}) = \beta.$$

Unconditional expectation $E(\hat{\beta})$ for *random* $\mathbf{X}$:

$$E(\hat{\beta}) = E_{\mathbf{X}} \left(E(\hat{\beta} \mid \mathbf{X}) \right) = \beta + E_{\mathbf{X}} (\mathbf{H}E(\epsilon \mid \mathbf{X})).$$

Under assumption **AX**, $E(\epsilon \mid \mathbf{X}) = 0$, hence the OLS estimator is unbiased also for random $\mathbf{X}$.

Unbiasedness of the OLS estimator II

- Note that assumption **AH** and **AN** are not necessary.
- Unbiasedness implies:

$$\begin{aligned}
 E(\hat{\beta}_j) &= \beta_j, & j = 0, \dots, K, \\
 E(\hat{\beta} - \beta) &= 0.
 \end{aligned}$$

- For any linear function $\mathbf{z} = \mathbf{A}\beta + \mathbf{b}$ of β , the “plug-in” estimator $\hat{\mathbf{z}} = \mathbf{A}\hat{\beta} + \mathbf{b}$ is an unbiased estimator of $\mathbf{z}$:

$$E(\hat{\mathbf{z}}) = E(\mathbf{A}\hat{\beta} + \mathbf{b}) = \mathbf{A}E(\hat{\beta}) + \mathbf{b} = \mathbf{A}\beta + \mathbf{b} = \mathbf{z}.$$

Covariance of the OLS estimator I

Due to unbiasedness, the expected value $E(\hat{\beta}_j)$ of the OLS estimator is equal to β_j for $j = 0, \dots, K$. Hence, the variance $V(\hat{\beta}_j)$ measures the variation of the OLS estimator $\hat{\beta}_j$ around the true value β_j :

$$V(\hat{\beta}_j) = E\left((\hat{\beta}_j - E(\hat{\beta}_j))^2\right) = E\left((\hat{\beta}_j - \beta_j)^2\right).$$

The covariance $\text{Cov}(\hat{\beta}_j, \hat{\beta}_k)$ of different coefficients of the OLS estimators measures, if deviations between the estimator and the true value are correlated.

$$\text{Cov}(\hat{\beta}_j, \hat{\beta}_k) = E\left((\hat{\beta}_j - \beta_j)(\hat{\beta}_k - \beta_k)\right).$$

This information is summarized for all possible pairs of coefficients in the covariance matrix of the OLS estimator:

$$\text{Cov}(\hat{\beta}) = E((\hat{\beta} - \beta)(\hat{\beta} - \beta)').$$

Covariance of the OLS estimator II

Under the additional assumption **AH**, the covariance matrix of the OLS estimator $\hat{\beta}$ is given by:

$$\text{Cov}(\hat{\beta} \mid \mathbf{X}) = \sigma^2 (\mathbf{X}' \mathbf{X})^{-1}. \quad (33)$$

Proof. Using (32), we obtain:

$$\hat{\beta} - \beta = \mathbf{H}\varepsilon, \quad \mathbf{H} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'.$$

The following holds:

Covariance of the OLS estimator III

$$\text{Cov}(\hat{\beta} \mid \mathbf{X}) = \text{E}(\mathbf{H}\varepsilon\varepsilon'\mathbf{H}' \mid \mathbf{X}) = \mathbf{H}\text{E}(\varepsilon\varepsilon' \mid \mathbf{X})\mathbf{H}' = \mathbf{H}\text{Cov}(\varepsilon \mid \mathbf{X})\mathbf{H}'. \quad (34)$$

Under the additional assumption $\mathbf{A}\mathbf{H}$, $\text{Cov}(\varepsilon \mid \mathbf{X}) = \sigma^2\mathbf{I}$, therefore:

$$\text{Cov}(\hat{\beta} \mid \mathbf{X}) = \mathbf{H}(\sigma^2\mathbf{I})\mathbf{H}' = \sigma^2\mathbf{H}\mathbf{H}' = \sigma^2(\mathbf{X}'\mathbf{X})^{-1},$$

since

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

The unconditional variance depends on the population covariance of the regressors. Use

variance decomposition:

$$\text{Cov}(\hat{\beta}) = E_X \left(\text{Cov}(\hat{\beta} \mid \mathbf{X}) \right) + \text{Cov}_X \left(E(\hat{\beta} \mid \mathbf{X}) \right).$$

Since $E(\hat{\beta} \mid \mathbf{X}) = \beta$, the second term is zero and therefore:

$$\text{Cov}(\hat{\beta}) = E_X \left(\text{Cov}(\hat{\beta} \mid \mathbf{X}) \right) = E_X \left(\sigma^2 (\mathbf{X}' \mathbf{X})^{-1} \right) = \sigma^2 E_X \left((\mathbf{X}' \mathbf{X})^{-1} \right).$$

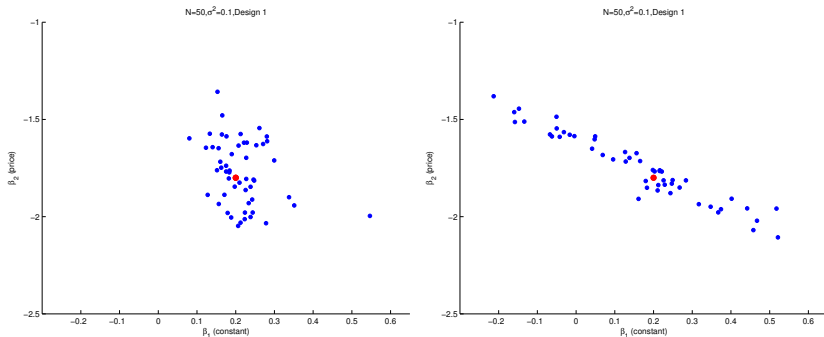
The diagonal elements of the matrix $\sigma^2 (\mathbf{X}' \mathbf{X})^{-1}$ define the variance $V \left(\hat{\beta}_j \right)$ of the

OLS estimator for each component. The standard deviation $SD_{\hat{\beta}_j}$ of each OLS estimator is defined as:

$$SD_{\hat{\beta}_j} = \sqrt{V(\hat{\beta}_j)} = \sigma \sqrt{(\mathbf{X}'\mathbf{X})_{j+1,j+1}^{-1}}. \quad (35)$$

It measures the estimation error on the same unit as β_j . Evidently, the standard deviation is the larger, the larger the variance of the error. What other factors influence the standard deviation?

Design 1: $x_i \sim -.5 + \text{Uniform}[0, 1]$ (left hand side) versus Design 2:
 $x_i \sim 1 + \text{Uniform}[0, 1]$ ($N = 50$, $\sigma^2 = 0.1$) (right hand side)



Variance of the OLS estimator in a simple regression I

In a simple regression model, we obtain:

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} N & \sum_{i=1}^N x_i \\ \sum_{i=1}^N x_i & \sum_{i=1}^N x_i^2 \end{pmatrix}$$

$$\begin{aligned} (\mathbf{X}'\mathbf{X})^{-1} &= \frac{1}{N \sum_{i=1}^N x_i^2 - (\sum_{i=1}^N x_i)^2} \begin{pmatrix} \sum_{i=1}^N x_i^2 & -\sum_{i=1}^N x_i \\ -\sum_{i=1}^N x_i & N \end{pmatrix} \\ &= \frac{1}{N s_x^2} \begin{pmatrix} s_x^2 + \bar{x}^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \end{aligned}$$

The variance of the slope estimator $\hat{\beta}_1$, given by:

Variance of the OLS estimator in a simple regression II

$$V\left(\hat{\beta}_1 \mid \mathbf{X}\right) = \frac{\sigma^2}{N s_x^2}, \quad (36)$$

is the larger,

- the smaller the number of observations N ;
- The variance of the slope estimator is the larger, the larger the error variance σ^2 ;

Variance of the OLS estimator in a simple regression III

The variance of the intercept $V(\hat{\beta}_0 | \mathbf{X})$, given by:

$$V(\hat{\beta}_0 | \mathbf{X}) = \frac{\sigma^2}{N} \frac{s_x^2 + \bar{x}^2}{s_x^2}. \quad (37)$$

depends on the ratio $(s_x^2 + \bar{x}^2)/s_x^2$. It can be minimized, if the covariates are centered, i.e., $\tilde{x}_i = x_i - \bar{x}$, i.e., $V(\hat{\beta}_0 | \tilde{\mathbf{X}}) = \sigma^2/N$. This does not affect $\hat{\beta}_1$, since $V(\hat{\beta}_1 | \tilde{\mathbf{X}}) = V(\hat{\beta}_1 | \mathbf{X})$. Centering the covariate X has also the effect that the

Variance of the OLS estimator in a simple regression **IV**

covariance of $\hat{\beta}_0$ and $\hat{\beta}_1$ is 0, i.e., the estimation errors are uncorrelated:

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1 \mid \tilde{\mathbf{X}}) = 0.$$

Explains differences in Design 1 and Design 2 ($N = 50$, $\sigma^2 = 0.1$) in the example, where the regressors are random:

- Design 1: $x_i \sim -0.5 + \text{Uniform}[0, 1]$, therefore: $E(x_i) = -0.5 + 0.5 = 0$, i.e., regressors are standardized.
- Design 2: $x_i \sim 1 + \text{Uniform}[0, 1]$, therefore: $E(x_i) = 1 + 0.5 = 1.5$, i.e., regressors are not standardized.

In practical regression analysis very often high (but not perfect) multicollinearity is present.

How well may X_j be explained by the other regressors?

Consider X_j as left-hand variable in the following regression model, whereas all the remaining predictors remain on the right hand side:

$$X_j = \tilde{\beta}_0 + \tilde{\beta}_1 X_1 + \dots + \tilde{\beta}_{j-1} X_{j-1} + \tilde{\beta}_{j+1} X_{j+1} + \dots + \tilde{\beta}_K X_K + \tilde{\varepsilon}.$$

Use OLS estimation to estimate the parameters and let $\hat{x}_{j,i}$ be the values predicted from this (OLS) regression.

- Define R_j as the correlation between the observed values $x_{j,i}$ and the predicted values $\hat{x}_{j,i}$ in this regression.
- If R_j^2 is close to 0, then X_j cannot be predicted from the other regressors. X_j contains additional, “independent” information.
- The closer R_j^2 is to 1, the better X_j is predicted from the other regressors and multicollinearity is present. X_j does not contain much „independent” information.

Using R_j , the variance $V(\hat{\beta}_j)$ of the OLS estimators of the coefficient β_j corresponding to X_j may be expressed in the following way for $j = 1, \dots, K$:

$$V(\hat{\beta}_j | \mathbf{X}) = \frac{\sigma^2}{N s_{x_j}^2 (1 - R_j^2)}.$$

Hence, the variance $V(\hat{\beta}_j | \mathbf{X})$ of the estimate $\hat{\beta}_j$ is large, if the regressors X_j is highly redundant, given the other regressors (R_j^2 close to 1, multicollinearity).

A naive estimator of $\sigma^2 = V(\varepsilon)$ would be the sample variance of the OLS residuals $\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_N$:

$$\tilde{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N \left(\hat{\varepsilon}_i - \frac{1}{N} \sum_{i=1}^N \hat{\varepsilon}_i \right)^2 = \frac{1}{N} \sum_{i=1}^N \hat{\varepsilon}_i^2 = \frac{\text{SSR}}{N},$$

where we used (19) and $\text{SSR} = \sum_{i=1}^N \hat{\varepsilon}_i^2$ is the sum of squared OLS residuals. However, due to the linear dependence between the OLS residuals, $\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_N$ is not an independent sample. Hence, $\tilde{\sigma}^2$ is a biased estimator of σ^2 . Due to the linear

dependence between the OLS residuals, only $\text{df} = (N - K - 1)$ residuals can be chosen independently. df is also called the degrees of freedom. An unbiased estimator of the error variance σ^2 in a homoskedastic multiple regression model obeying **assumption AH** is given by:

$$\hat{\sigma}^2 = \frac{\text{SSR}}{\text{df}}, \quad (38)$$

where $\text{df} = (N - K - 1)$, N is the number of observations, and K is the number of predictors $X_1, \dots, X_K$ (not including the constant).

The standard errors of the OLS estimator

The covariance matrix of the OLS estimator depends on σ^2 . To evaluate the estimation error for a given data set, σ^2 is substituted by the estimator (38). This yields the so-called *standard error* $\text{se}(\hat{\beta}_j)$ of the OLS estimator:

$$\text{se}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2} \sqrt{(\mathbf{X}'\mathbf{X})_{j+1,j+1}^{-1}}.$$

Using (32), we obtain under assumption **AN** the following sampling distribution:

$$\hat{\beta} - \beta \sim \text{Normal}_{K+1} \left(\mathbf{0}, \text{Cov}(\hat{\beta}) \right), \quad \text{Cov}(\hat{\beta}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}.$$

All marginal distributions are normal, hence:

$$\hat{\beta}_j - \beta_j \sim \text{Normal} \left(0, \text{SD}_{\hat{\beta}_j}^2 \right). \quad (39)$$

More about the distribution of the OLS estimator

- For any subset of coefficients $\tilde{\beta} = (\beta_{j_1}, \dots, \beta_{j_q})'$, the OLS estimator $\hat{\tilde{\beta}} = (\hat{\beta}_{j_1}, \dots, \hat{\beta}_{j_q})'$, follows a multivariate normal distribution:

$$\hat{\tilde{\beta}} - \tilde{\beta} \sim \text{Normal}_q \left(\mathbf{0}, \text{Cov}(\hat{\tilde{\beta}}) \right), \quad (40)$$

where $\text{Cov}(\hat{\tilde{\beta}})$ is obtained from the rows and columns $j_1, \dots, j_q$ of $\text{Cov}(\hat{\beta})$.

- This result may be used to construct 95%-confidence ellipsoids for all pairs of parameters $(\beta_{j_1}, \beta_{j_2})$.

In an i.i.d. setting, i.e., where (Y_i, X_i) are i.i.d. (and assumption AL, AR, AX, AH are satisfied) and $E(X_i'X_i)$ is invertible, OLS estimator is asymptotically normally distributed.

To see this, note, again, that

$$\sqrt{N}(\hat{\beta} - \beta) = \sqrt{N}(X'X)^{-1}X'\varepsilon = (N^{-1}X'X)^{-1}N^{-1/2}X'\varepsilon. \quad (1)$$

Note: $N^{-1}X'X = N^{-1} \sum_{i=1}^N X_i'X_i$ which converges to $E(X_i'X_i)$ by the LLN.

Note: $N^{-1/2}X'\varepsilon = N^{-1/2} \sum_{i=1}^N X_i'\varepsilon_i$ with $E(X_i'\varepsilon_i) = 0$ and $E(X_i'X_i\varepsilon_i^2) = \sigma^2 E(X_i'X_i)$.

Hence, the (multivariate) CLT implies $N^{-1/2}X'\varepsilon \rightarrow N(0, \sigma^2 E(X_i'X_i))$ in distribution.

By the continuous mapping theorem $(N^{-1}X'X)^{-1} \rightarrow E(X_i'X_i)^{-1}$ (in probability), so that Slutsky's theorem shows that (in distribution)

$$\sqrt{N}(\hat{\beta} - \beta) = (N^{-1}X'X)^{-1}N^{-1/2}X'\varepsilon \rightarrow N(0, \sigma^2 E(X_i'X_i)^{-1}). \quad (2)$$

The χ_q^2 distribution I

Let $Z_1, \dots, Z_q$ be a sequence of iid standard normal random variables. Define a random variable Y as the sum of Z_i^2 , i.e.:

$$Y = \sum_{i=1}^q Z_i^2.$$

Then the random variable Y follows a χ_q^2 -distribution with q degrees of freedom.

Moments of the χ_q^2 -distribution:

- If $Z \sim \text{Normal}(0, 1)$ follows a standard normal distribution, then $Y = Z^2$ follows a χ_1^2 -distribution; i.e., $E(Y) = E(Z^2) = 1$,
 $V(Y) = V(Z^2) = E(Z^4) - E(Z^2)^2 = 3 - 1 = 2$.

The χ_q^2 distribution II

- Since $Y \sim \chi_q^2$ is the sum of q independent squared standard normal random variables $Z_1, \dots, Z_q$, we obtain:

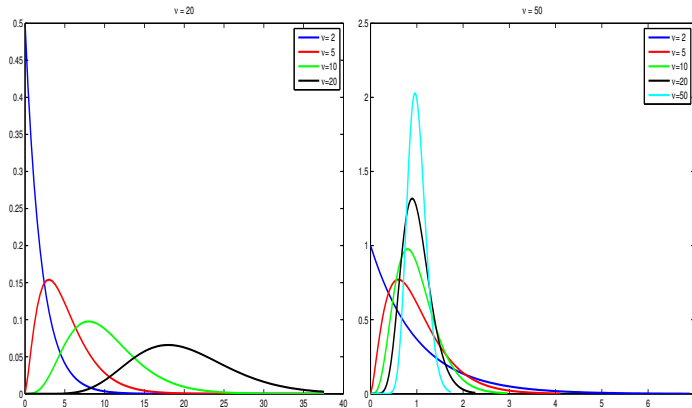
$$E(Y) = E\left(\sum_{i=1}^q Z_i^2\right) = \sum_{i=1}^q E(Z_i^2) = q,$$

$$V(Y) = V\left(\sum_{i=1}^q Z_i^2\right) = \sum_{i=1}^q V(Z_i^2) = 2q.$$

The χ_q^2 distribution III

Left hand side: density of the χ_q^2 -distribution; right hand side: density of the random variable X/q , where $X \sim \chi_q^2$ -distribution ($q = 2, 5, 10, 20$)

The χ^2_q distribution IV



The χ_q^2 distribution **V**

- If we standardize $Y \sim \chi_q^2$ by q , then $E(Y/q) = 1$ and $V(Y/q) = 2/q$. As q increases, Y/q converges in probability to 1.

It can be shown, that under the assumption **AN**,

$$\frac{\hat{\epsilon}'\hat{\epsilon}}{\sigma^2} \sim \chi_{df},$$

where $df = N - (K + 1)$. Hence, $\hat{\sigma}^2$ is an unbiased estimator of σ^2 :

$$E\left(\frac{\hat{\sigma}^2}{\sigma^2}\right) = E\left(\frac{\hat{\epsilon}'\hat{\epsilon}/df}{\sigma^2}\right) = 1.$$

Theorem

The estimator $\hat{\sigma}^2$ is unbiased for σ^2 under assumptions AL, AR, AX, AH (i.e., without the normality assumption).

The sum of the diagonal entries of a square matrix A , say, is called its *trace* and written as $\text{tr}(A)$. It has the property that $\text{tr}(AB) = \text{tr}(BA)$ (if the products make sense).

Exercise: Verify that for a random square matrix A it holds that $E(\text{tr}(A)) = \text{tr}(E(A))$.

Note that $\hat{\sigma}^2 = (N - K - 1)^{-1} \hat{\varepsilon}' \hat{\varepsilon}$ and that $(AL, AR) \hat{\varepsilon} = MY = M\varepsilon$ with $M = I - X(X'X)^{-1}X'$.

Exercise: Verify that $M = M' = M^2$ and that $\text{tr}(M) = N - K - 1$.

The unbiasedness claim above can now be established as follows:

First, note that

$$E(\hat{\sigma}^2) = E((N - K - 1)^{-1} \hat{\varepsilon}' \hat{\varepsilon}) = (N - K - 1)^{-1} E(\varepsilon' M' M \varepsilon) = (N - K - 1)^{-1} E(\varepsilon' M \varepsilon). \quad (3)$$

But

$$E(\varepsilon' M \varepsilon) = E(\text{tr}(\varepsilon' M \varepsilon)) = E(\text{tr}(M \varepsilon \varepsilon')) = \text{tr}(E(M \varepsilon \varepsilon')) = \text{tr}(M E(\varepsilon \varepsilon')) = \sigma^2 \text{tr}(M)$$

which equals $\sigma^2(N - K - 1)$; we used AX and AH in the 5th equality.

We next study the question in which sense the OLS estimator is “best”.

We only consider *linear* estimators, which are estimators of the form $\tilde{\beta} = C(X)Y$.

[Note that the OLS estimator is linear with $C(X) = (X'X)^{-1}X'$.]

We say that a linear estimator $\tilde{\beta}$ is *better* than another linear estimator $\check{\beta}$ if, for every vector γ , we have

$$V(\gamma\tilde{\beta} \mid X) \leq V(\gamma\check{\beta} \mid X);$$

equivalently

$$V(\tilde{\beta} \mid X) \leq V(\check{\beta} \mid X)$$

in the sense that the *matrix* $V(\check{\beta} \mid X) - V(\tilde{\beta} \mid X)$ is non-negative definite.

[Recall that a square matrix A is *non-negative definite* if $\gamma' A \gamma \geq 0$ for every γ .]

The Gauss Markov Theorem I

The Gauss Markov Theorem.

Under assumptions **AL**, **AR**, **AX**, **AH**, the OLS estimator is BLUE, i.e., the

- Best
- Linear
- Unbiased
- Estimator

Here “best” means that any other linear unbiased estimator has a larger variance than the OLS estimator. The OLS estimator is a linear combination of the outcomes $\mathbf{y}$, with

The Gauss Markov Theorem II

weights defined by the matrix $\mathbf{H}$. Suppose we use a different weight matrix $\mathbf{C}$, i.e., $\hat{\beta}_C = \mathbf{C}\mathbf{y}$, such that $E(\hat{\beta}_C | \mathbf{X}) = \beta$ is unbiased.

Under $\mathbf{A}\mathbf{X}$, this implies:

$$E(\mathbf{C}\mathbf{y} | \mathbf{X}) = E(\mathbf{C}\mathbf{X}\beta | \mathbf{X}) + E(\mathbf{C}\varepsilon | \mathbf{X}) = \mathbf{C}\mathbf{X}\beta = \beta, \quad (41)$$

hence $\mathbf{C}\mathbf{X} = \mathbf{I}$. Under $\mathbf{A}\mathbf{H}$, the covariance matrix of $\hat{\beta}_C$ is given by:

$$\text{Cov}(\hat{\beta}_C | \mathbf{X}) = \sigma^2 \mathbf{C}\mathbf{C}'.$$

It can be shown that:

The Gauss Markov Theorem III

$$\text{Cov}(\hat{\beta}_C | \mathbf{X}) = \text{Cov}(\hat{\beta}_{\text{OLS}} | \mathbf{X}) + \sigma^2 \mathbf{A} \mathbf{A}', \quad (42)$$

where $\mathbf{A}$ is defined as $\mathbf{A} = \mathbf{C} - \mathbf{H}$. Hence, for any vector of fixed values $\mathbf{w}$,

$$\begin{aligned}
 V(\mathbf{w}' \hat{\beta}_C | \mathbf{X}) &= \mathbf{w}' \text{Cov}(\hat{\beta}_C | \mathbf{X}) \mathbf{w} \\
 &= \mathbf{w}' \text{Cov}(\hat{\beta}_{\text{OLS}} | \mathbf{X}) \mathbf{w} + \sigma^2 \mathbf{w}' \mathbf{A} \mathbf{A}' \mathbf{w} \\
 &\geq V(\mathbf{w}' \hat{\beta}_{\text{OLS}} | \mathbf{X}),
 \end{aligned}$$

since the quadratic form $\mathbf{w}' \mathbf{A} \mathbf{A}' \mathbf{w} \geq 0$. Proof of (42). Since $\mathbf{C} = \mathbf{H} + \mathbf{A}$,

The Gauss Markov Theorem IV

$$\text{Cov}(\hat{\beta}_C | \mathbf{X}) = \sigma^2 \mathbf{C} \mathbf{C}' = \sigma^2 (\mathbf{H} + \mathbf{A})(\mathbf{H} + \mathbf{A})'.$$

Note that due to unbiasedness of $\hat{\beta}_C$,

$$\mathbf{I} = \mathbf{C} \mathbf{X} = (\mathbf{H} + \mathbf{A}) \mathbf{X} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{X} + \mathbf{A} \mathbf{X} = \mathbf{I} + \mathbf{A} \mathbf{X}.$$

Therefore, $\mathbf{A} \mathbf{X} = \mathbf{O}$ and consequently,

$$\begin{aligned} \mathbf{A} \mathbf{H}' &= \mathbf{A} \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} = \mathbf{O}, & \mathbf{H} \mathbf{A}' &= (\mathbf{A} \mathbf{H}')' = \mathbf{O}, \\ (\mathbf{H} + \mathbf{A})(\mathbf{H} + \mathbf{A})' &= \mathbf{H} \mathbf{H}' + \mathbf{A} \mathbf{H}' + \mathbf{H} \mathbf{A}' + \mathbf{A} \mathbf{A}' = \mathbf{H} \mathbf{H}' + \mathbf{A} \mathbf{A}'. \end{aligned}$$

Some comments:

The Gauss Markov Theorem **V**

- The Gauss Markov Theorem also holds, if the regressors are random.
- If in addition, assumption **AN** holds, then a stronger optimality result holds.

Efficiency of OLS estimation.

Under assumption **AN**, the OLS estimator $\hat{\beta}$ is the minimum variance unbiased estimator.

Any other unbiased estimator $\tilde{\beta}$ (which need not be a linear estimator) has larger standard deviations than the OLS estimator.

Example: CAPM I

The CAPM states a relation between the expected return $\mu_i = E(y^i)$ of an asset i , the risk-free return r_f , and the expected return $\mu_m = E(y^m)$ of a market portfolio:

$$\mu_i - r_f = \beta_i(\mu_m - r_f),$$

also known as security market line. The so-called beta factor β_i defined as

$$\beta_i = \frac{\text{Cov}(y^i, y^m)}{V(y^m)},$$

measure the systematic risk. Assets with $\beta_i > 1$ imply more risk than the market. Use

Example: CAPM II

the *market model* to estimate β_i :

$$y_t^i = \alpha_i + \beta_i y_t^m + \varepsilon_t^i,$$

where

- $y_t^i; t = 1, \dots, N$ are the observed returns of asset i ;
- $y_t^m; t = 1, \dots, N$ are the observed returns of a market portfolio;
- $\alpha_i = (1 - \beta_i)r_f$

Example: CAPM III

If we use excess returns $x_t^i = y_t^i - r_f$ and $x_t^m = y_t^m - r_f$, then

$$x_t^i = \beta_i x_t^m + \varepsilon_t^i.$$

The market model implies the following variance decomposition of $\sigma_i^2 = V(y^i)$:

$$\sigma_i^2 = \beta_i^2 \sigma_m^2 + \sigma_{\varepsilon^i}^2,$$

where σ_m^2 is the variance of the market portfolio and $\sigma_{\varepsilon^i}^2$ is a firm-specific (idiosyncratic) risk.

Example: CAPM IV

- The coefficient of determination R^2 measure the proportion of market-specific (systemic) risk in total variance:

$$R^2 = \frac{\beta_i^2 \sigma_m^2}{\sigma_i^2}.$$

- The testable implication of the CAPM is that the constant term in a simple regression model should be 0.

Multiple regression model:

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_j X_j + \dots + \beta_K X_K + \varepsilon, \quad (43)$$

Does the predictor variable X_j exercise an influence on the expected mean $E(Y)$ of the response variable Y , if we control for all other variables $X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_K$?
Formally,

$$\beta_j = 0?$$

Understanding the testing problem I

Repeat the following experiment many times:

- Simulate data from a multiple regression model with $\beta_0 = 0.2$, $\beta_1 = -1.8$, and $\beta_2 = 0$:

$$Y = 0.2 - 1.8X_1 + 0 \cdot X_2 + \varepsilon, \quad \varepsilon \sim \text{Normal}(0, \sigma^2).$$

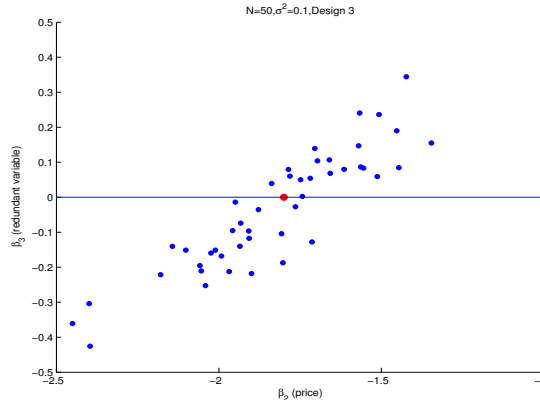
- Run OLS estimation for a model where β_2 is unknown:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \quad \varepsilon \sim \text{Normal}(0, \sigma^2),$$

to obtain $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$. Is $\hat{\beta}_2$ different from 0?

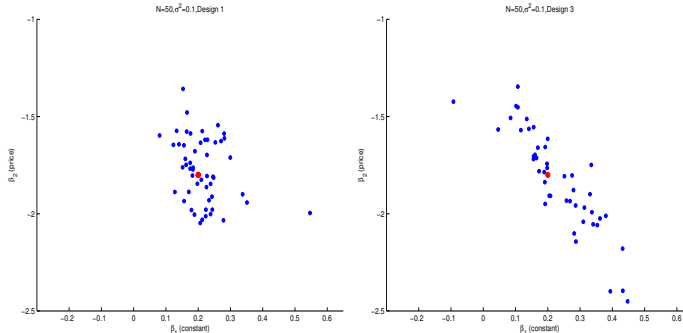
Understanding the testing problem II

The OLS estimator $\hat{\beta}_2$ of $\beta_2 = 0$ differs from 0 for a single data set, but is 0 on average.



Understanding the testing problem III

OLS estimation for the true model in comparison to estimating a model with a redundant predictor variable: including the redundant predictor X_2 increases the estimation error for the other parameters β_0 and β_1 .



- What may we learn from the data about hypothesis concerning the unknown parameters in the regression model, especially about the hypothesis that $\beta_j = 0$?
- May we reject the hypothesis $\beta_j = 0$ given data?
- Testing, if $\beta_j = 0$ is not only of importance for the substantive scientist, but also from an econometric point of view, to increase efficiency of estimation of non-zero parameters.

It is possible to answer these questions under assumption **AN**.

Testing a single coefficient - the t-test:

If σ^2 is known, then all marginal distributions are normal, hence:

$$\hat{\beta}_j - \beta_j \sim \text{Normal} \left(0, \text{SD}_{\hat{\beta}_j}^2 \right)$$

If the null hypothesis $\beta_j = 0$ is true, then possible differences between the OLS-estimator $\hat{\beta}_j$ and 0 may be quantified using the following t -statistic:

$$t_j = \frac{\hat{\beta}_j}{\text{SD}_{\hat{\beta}_j}} \sim \text{Normal} (0, 1) ,$$

which follows a standard normal distribution. If **AN** holds and σ^2 is known, then t_j

follows a standard normal distribution under the null hypothesis:

- Choose a significance level α .
- Determine the corresponding critical value c_α as the $(1 - \alpha/2)$ -quantile of the standard normal distribution.
- If $|t_j| > c_\alpha$: reject the null hypothesis (risk to reject the null hypothesis although it is true is equal to α).
- If $|t_j| \leq c_\alpha$: do not reject the null hypothesis (risk to “accept” a wrong null hypothesis may be arbitrarily large).

Choice of c_α , when σ^2 is unknown:

If σ^2 is unknown and estimated as described above, then $SD_{\hat{\beta}_j}$ is substituted by $se(\hat{\beta}_j)$, yielding the test statistic:

$$t_j = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}.$$

Choosing the quantiles of the normal distributions would lead to a test which rejects the null-hypothesis more often than desired, e.g., for $\alpha = 0.05$ and $K = 3$:

N	10	20	30	40	50	100
Prob(reject H_0)	0.09	0.07	0.06	0.05	0.05	0.05

The reason is that t_j no longer follows a normal distribution, but a t_{df} -distribution where $df = (N - K - 1)$. The critical values $t_{df,1-\alpha/2}$ depend on df and are equal to the quantiles of the t_{df} -distribution. For $\alpha = 0.05$ and for a regression model with 3 parameters, e.g., these values are:

$df = N - 3$	7	17	27	37	47	97
$t_{df,0.975}$	2.37	2.11	2.05	2.02	2.00	1.96

The p -value is derived from the distribution of the t -statistics under the null hypothesis and is easier to interpret than the t -statistics which has to be compared to the right quantiles:

- Choose a significance level α .
- If $p < \alpha$: reject the null hypothesis (risk to reject the null hypothesis although it is true is at most equal to α).
- If $p \geq \alpha$: do not reject the null hypothesis (risk to “accept” a wrong null hypothesis may be arbitrarily large).

If σ^2 is unknown, then (39) implies with $df = (N - K - 1)$:

$$\frac{\hat{\beta}_j - \beta_j}{\text{se}(\hat{\beta}_j)} \sim t_{df},$$

This yields, with $t_{df,p}$ being the p -quantiles of the t_{df} -distribution:

- β_j lies in $[\hat{\beta}_j - t_{df,1-\alpha/2}\text{se}(\hat{\beta}_j), \hat{\beta}_j + t_{df,1-\alpha/2}\text{se}(\hat{\beta}_j)]$
- $\hat{\beta}_j + t_{df,1-\alpha}\text{se}(\hat{\beta}_j)$ is an upper bound for β_j ;
- $\hat{\beta}_j - t_{df,1-\alpha}\text{se}(\hat{\beta}_j)$ is a lower bound for β_j .

Testing more than one coefficient I

- Testing the null hypothesis $\beta_j = 0$ based on t_j is only valid, if all other parameters remain in the model.
- Often, we want to test joint hypotheses about our parameters.
- E.g.; if the t_j -statistics is not significant for more than one parameter $j_1, \dots, j_q$, then one needs to test, if $\beta_{j_1} = 0, \dots, \beta_{j_q} = 0$ simultaneously.
- We cannot simply check each t_j -statistic separately. It is possible for jointly insignificant regressors to be individually significant (and vice versa).

Testing more than one coefficient II

Given the data, is it possible to reject the null hypothesis $\beta_{j_1} = 0, \dots, \beta_{j_q} = 0$?

Reject the null hypothesis, if the distance between the OLS estimator

$\tilde{\beta} = (\hat{\beta}_{j_1}, \dots, \hat{\beta}_{j_q})'$ and 0 is “large” (one-sided test).

The corresponding test statistic has to take into account that

- the standard deviations of the various OLS estimators are different.
- deviations of the OLS estimators from the true value are likely to be correlated.

Aggregate $t_{jl} = \hat{\beta}_{jl} / \text{SD}_{\hat{\beta}_{jl}}$ for $l = 1, \dots, q$, e.g., by taking the sum of squared t statistics? If the deviations of the OLS estimators $\hat{\beta}_{j1}, \dots, \hat{\beta}_{jq}$ from the true values are uncorrelated, then the aggregated test statistic

$$SSR_q = \sum_{l=1}^q \frac{\hat{\beta}_{jl}^2}{\text{SD}_{\hat{\beta}_{jl}}^2}$$

is the sum of q independent squared standard normal random variables, hence it follows a χ_q^2 distribution.

Obtaining uncorrelated coefficients

If $\beta \sim \text{Normal}_{K+1}(\mu, \Sigma)$, then the variable $\mathbf{H}\beta$, where $\mathbf{H}$ is a $q \times (K + 1)$ -matrix follows also a normal distribution: $\mathbf{H}\beta \sim \text{Normal}_q(\mathbf{H}\mu, \mathbf{H}\Sigma\mathbf{H}')$. Obtaining uncorrelated coefficients:

- decompose $\Sigma = \mathbf{L}\mathbf{L}'$; then $\tilde{\beta} = \mathbf{L}^{-1}\beta \sim \text{Normal}_q(\mathbf{0}, \mathbf{I})$:

$$\begin{aligned}
 \tilde{\beta} &= \mathbf{L}^{-1}\beta \sim \text{Normal}_{K+1}(\mathbf{0}, \mathbf{L}^{-1}\Sigma(\mathbf{L}')^{-1}) \\
 &\sim \text{Normal}_{K+1}(\mathbf{0}, \mathbf{L}^{-1}\mathbf{L}\mathbf{L}'(\mathbf{L}')^{-1}) \sim \text{Normal}_{K+1}(\mathbf{0}, \mathbf{I}).
 \end{aligned}$$

Therefore, $\tilde{\beta}_0, \dots, \tilde{\beta}_{K+1}$ are iid standard normal and $\tilde{\beta}'\tilde{\beta} = \beta'(\mathbf{L}')^{-1}\mathbf{L}^{-1}\beta = \beta'\Sigma^{-1}\beta$ follows a χ^2_{K+1} distribution.

Testing more than one coefficient - the Wald statistic

Under the nullhypothesis, the deviations of the OLS estimators $\hat{\beta}_{j_1}, \dots, \hat{\beta}_{j_q}$ from zero are usually correlated.

- The Wald-statistics transform the deviations to a coordinate system with independent standard normal random variables

$$W = \tilde{\beta}' \text{Cov}(\tilde{\beta})^{-1} \tilde{\beta} \sim \chi_q^2,$$

and follow a χ_q^2 -distribution with q degrees of freedom.

- Note: the χ_q^2 -distribution results only, if σ^2 is known.

Testing more than one coefficient - the F statistic I

The F-Test

- The F -statistic is obtained by substituting the unknown variance σ^2 by $\hat{\sigma}^2$ and dividing by q .
- The F -statistic is the ratio of two (independent) sum of squares, divided by the degrees of freedom, i.e., a χ_q^2/q and χ_{df}^2/df , where $df = N - K - 1$.
- If the null hypothesis $\beta_{j_1} = 0, \dots, \beta_{j_q} = 0$ is true, then the F -statistic follows a $F_{q,df}$ -distribution with parameters q (number of tested coefficients) and $df = N - K - 1$.
- Remark: for $q = 1$, $F = t_j^2$, where t_j is the t-statistic.

Testing more than one coefficient - the F statistic II

Reject the null-hypothesis, if

- the F -statistic is larger than the critical value from the corresponding $F_{q,df}$ -distribution (one-sided test).
- the corresponding p -value is smaller than the significance level. A p -value close to 0 shows that the value observed for the F -statistic is in conflict with the null hypothesis.

⇒ At least one of the coefficients $\beta_{j_1}, \dots, \beta_{j_q}$ is different from 0.

Testing more than one coefficient - the F statistic III

Do not reject the null-hypothesis, if

- the F -statistic is smaller than the critical value from the corresponding $F_{q,df}$ -distribution (one-sided test).
- The corresponding p -value is larger than the significance level. A p -value considerably larger than 0 shows that the observed value for the F -statistic is not in conflict with the null hypothesis.

⇒ There is no evidence in the data that we should reject the null hypothesis *that all coefficients $\beta_{j_1}, \dots, \beta_{j_q}$ are equal to 0.*