Lecture 03: Classical Inferential Statistics I: Basics and Confidence Intervals

3.1. Ordinary Least Squares (OLS) Estimator: Expectation and Covariance

- Only stochasticity: residual errors $\boldsymbol{\epsilon}$ according to $\boldsymbol{y} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\epsilon}$
- The OLS estimator is linear in $\boldsymbol{y}$:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{y} \\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'(\mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ &= \underline{\underline{\boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\epsilon}}}\end{aligned}$$

Expectation value

$$E(\hat{\boldsymbol{\beta}}) = E(\boldsymbol{\beta}) + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'E(\boldsymbol{\epsilon}) = \boldsymbol{\beta}$$

> The OLS estimator of parameter-linear models is **unbiased** under the mild condition $E(\boldsymbol{\epsilon}) = \mathbf{0}$ for all the data points

OLS estimator: variances and covariances

- Gauß-Markow conditions $\rightarrow \epsilon \sim$ i.i.d$N(0, \sigma^2) \rightarrow \hat{\boldsymbol{\beta}}$ is normal distributed
- In this case, the complete error characteristics are specified by the expectation value and the **variance-covariance matrix** $\mathbf{V}_{\hat{\boldsymbol{\beta}}}$

$$
\begin{aligned}
\mathbf{V}_{\hat{\boldsymbol{\beta}}} & \stackrel{\text{def}}{=} E\left((\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\right) \\
\text{[insert } \hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} \rightarrow] & = E\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}\right)'\right) \\
\text{[transpose and inversion rules } \rightarrow] & = E\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right) \\
[E(.) \text{ acts only on } \boldsymbol{\epsilon} \rightarrow] & = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
\text{[Gauß-Markow } \rightarrow] & = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma_\epsilon^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
\text{[def inverse matrix } \rightarrow] & = \sigma_\epsilon^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}
$$

The variance-covariance matrix depends only on the values of the exogenous factors!

Results

- Ordinary least squares (OLS) estimator:

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

- **Variance-Covariance matrix** of the estimation errors (provided the errors are i.i.d.) can be written in terms of the **Hesse matrix H** of the objective function SSE:

$$\begin{aligned} \mathbf{V}_{\hat{\boldsymbol{\beta}}} &= E\left((\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\right) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = 2\sigma^2\mathbf{H}^{-1}, \\ H_{jk} &= \left.\frac{\partial^2 S}{\partial\beta_j\ \partial\beta_k}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = 2(\mathbf{X}'\mathbf{X})_{jk} \end{aligned}$$

- Variances of estimation errors: $V(\hat{\beta}_j) = V_{jj}$
- Correlation of estimation errors: $\text{Corr}(\hat{\beta}_j, \hat{\beta}_k) = \frac{V_{jk}}{\sqrt{V_{jj}V_{kk}}}$
- Distribution of the normalized estimation errors: $\frac{\hat{\beta}_j - \beta_j}{\sqrt{V_{jj}}} \sim N(0,1)$

Estimation of the residual variance

The above cannot be applied directly since the residual variance σ^2 is unknown and must be estimated by the minimum SSE $S(\hat{\beta})$:

$$\hat{\sigma}^2 = \frac{1}{n-J-1} \sum_i \left(y_i - \hat{y}(\boldsymbol{x}_i)\right)^2 = \frac{S(\hat{\beta})}{n-J-1}$$

Under the Gauß-Markow assumptions, this can be expressed as the sum of squared Gaussians as follows (derivation for the experts):

$$\begin{aligned}(n-J-1)\hat{\sigma}^2 &= (\hat{\boldsymbol{y}}-\boldsymbol{y})'(\hat{\boldsymbol{y}}-\boldsymbol{y}) \\ &= (\mathbf{X}\hat{\boldsymbol{\beta}}-\boldsymbol{y})'(\mathbf{X}\hat{\boldsymbol{\beta}}-\boldsymbol{y}) \\ &= (\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{X}\hat{\boldsymbol{\beta}}) - (\mathbf{X}\boldsymbol{\beta})'\boldsymbol{y} - \boldsymbol{y}'(\mathbf{X}\boldsymbol{\beta}) + \boldsymbol{y}'\boldsymbol{y}\end{aligned}$$

With following rule for scalar products: $\boldsymbol{a}'\boldsymbol{b} = \boldsymbol{b}'\boldsymbol{a}$ it follows that the two middle terms are equal. Replacing $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y}$ we see that, interestingly, the first term is the negative of each of the two middle terms resulting in

$$(n-J-1)\hat{\sigma}^2 = \boldsymbol{y}'\left(\mathbf{1} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\boldsymbol{y}$$

Estimation of the residual variance (ctned)

Finally, we replace the observed endogeneous data vector $\boldsymbol{y}$ by the model $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ Notice: the true and, according to the Gauß-Markow assumptions, immutable parameter vector $\boldsymbol{\beta}$ is used here!:

$$
\begin{aligned}
(n-J-1)\hat{\sigma}^2 &= (\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon})' \left(\mathbf{1} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}) \\
&= \boldsymbol{\epsilon}'(\mathbf{1} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\boldsymbol{\epsilon} \\
&\quad + 2(\mathbf{X}\boldsymbol{\beta})'(\mathbf{1} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\boldsymbol{\epsilon} + \boldsymbol{\beta}'\mathbf{X}'(\mathbf{1} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}
$$

After doing the simplification, we realize that the second and third term are each equal to zero, so we have the final result

$$
(n-J-1)\hat{\sigma}^2 = \boldsymbol{\epsilon}'(\mathbf{1} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\boldsymbol{\epsilon}
$$

With the Gauß-Markow-assumptions, this is proportional to a sum of $(n-J-1)$ squared Gaussians, i.e., a $\chi^2(n-J-1)$ distributed random variable

Results if the variance needs to be estimated

- Estimated variance-covariance matrix:

$$\hat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}} = 2\hat{\sigma}^2 \mathbf{H}^{-1} = \hat{\sigma}^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

- The normalized approximate estimation errors are student-t distributed (a Gaussian in the numerator and the square root of a χ^2 distributed random variable in the denominator):

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{V}_{jj}}} \sim T(n-1-J)$$

Multivariate distribution function of $\hat{\beta}$

The distribution of the errors $\Delta\hat{\boldsymbol\beta} = \hat{\boldsymbol\beta} - \boldsymbol\beta$ obeys a multivariate normal distribution:

$$f_{\hat{\boldsymbol\beta}}(\Delta\hat{\boldsymbol\beta}) \propto \exp\left[-\frac{1}{2}\Delta\hat{\boldsymbol\beta}' \, \mathbf{V}^{-1} \, \Delta\hat{\boldsymbol\beta}\right] = \exp\left[-\frac{\Delta\hat{\boldsymbol\beta}' \, \mathbf{X}'\mathbf{X} \, \Delta\hat{\boldsymbol\beta}}{2\sigma_\epsilon^2}\right].$$

Relation to the maximum-likelihood-method (→ Lecture 07:)

Expand the SSE $S(\boldsymbol\beta)$ around $\hat{\boldsymbol\beta}$ to second order:

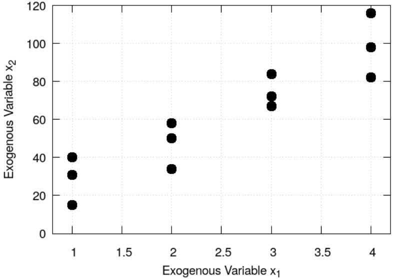
$$S(\boldsymbol\beta) - S(\hat{\boldsymbol\beta}) \approx \frac{1}{2}\Delta\hat{\boldsymbol\beta}' \, \mathbf{H} \, \Delta\hat{\boldsymbol\beta} = \Delta\hat{\boldsymbol\beta}' \, \mathbf{X}'\mathbf{X} \, \Delta\hat{\boldsymbol\beta}$$

$$\Rightarrow \qquad \boxed{f_{\hat{\boldsymbol\beta}}(\Delta\hat{\boldsymbol\beta}) \propto \exp\left[-\frac{S(\boldsymbol\beta) - S(\hat{\boldsymbol\beta})}{2\sigma_\epsilon^2}\right]}$$

and with the estimated residual variance $\hat\sigma_\epsilon^2 = S(\hat{\boldsymbol\beta})/(n-J-1)$

$$\hat{f}_{\hat{\boldsymbol\beta}}(\boldsymbol\beta) \propto \exp\left[-\frac{(n-J-1)}{2}\left(\frac{S(\boldsymbol\beta)}{S(\hat{\boldsymbol\beta})} - 1\right)\right]$$

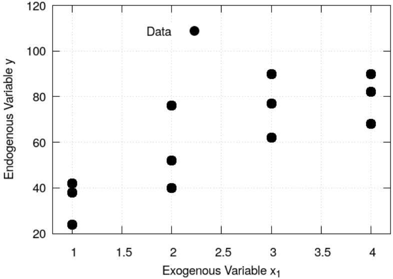
Example of correlated errors: modeling the demand for hotel rooms

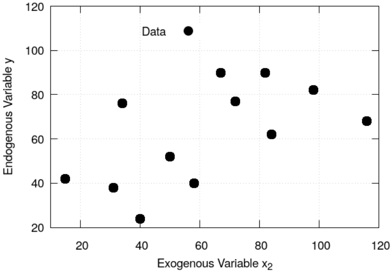


- The example of Lecture 02:

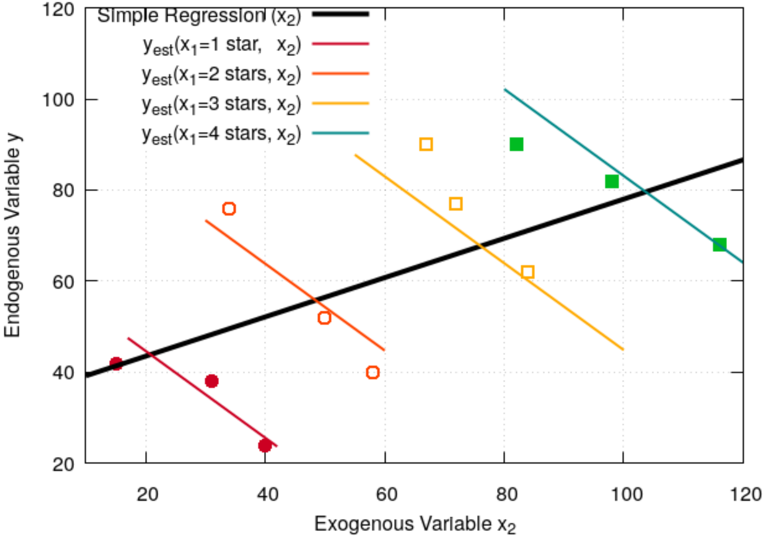
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

- Exogenous factors: $x_0 = 1$, x_1: proxy for quality [# stars]; x_2: price [€/night].
- Endogenous variable: booking rate [%]
- The demand is positively correlated with both the quality and the price

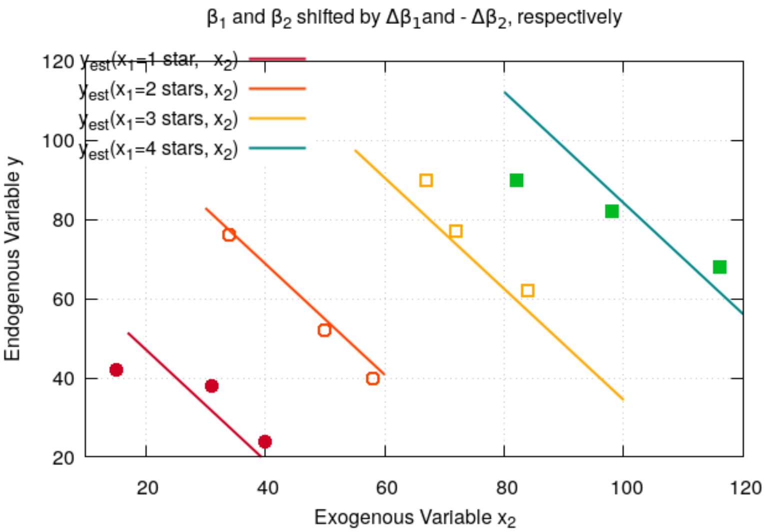




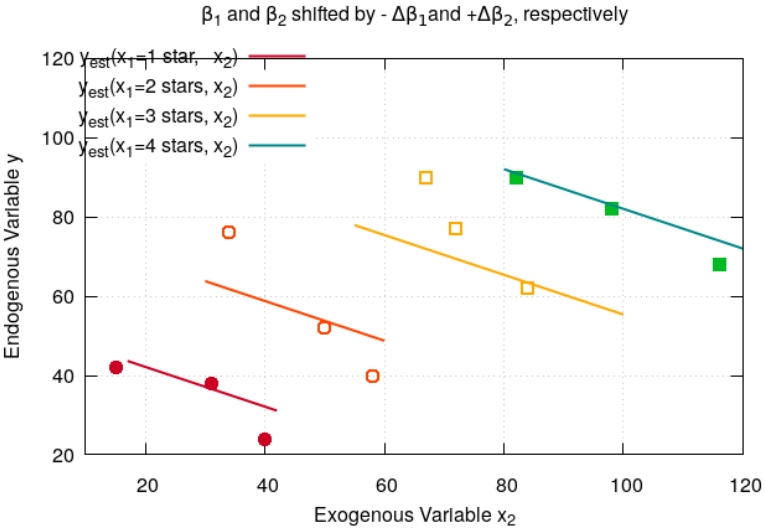
Residual errors for fitted parameters



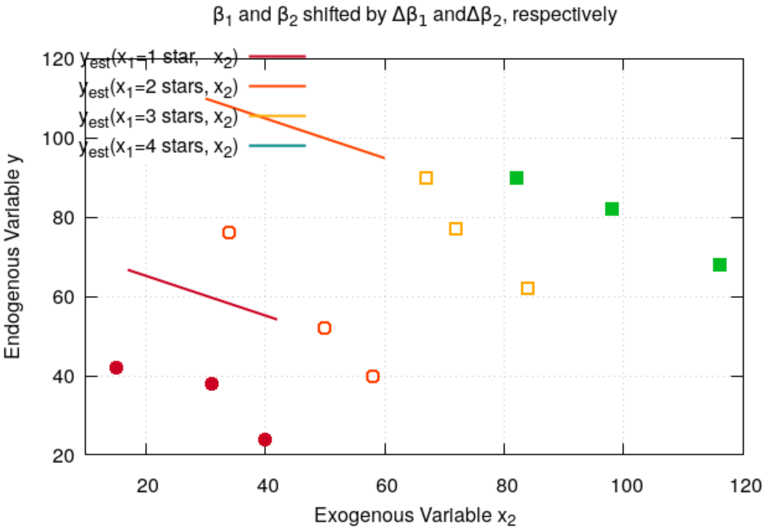
Effect of mis-fit parameters I: small effect if β_1 and β_2 have opposite misfits



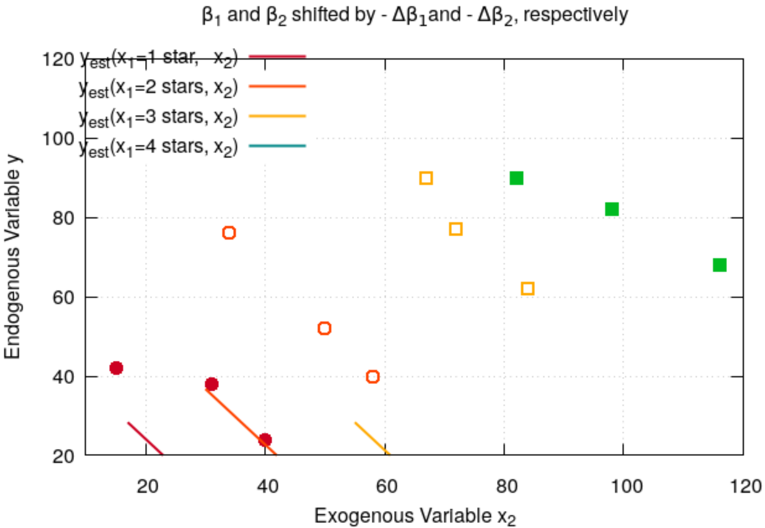
Effect of mis-fit parameters II: small effect if β_1 and β_2 have opposite misfits



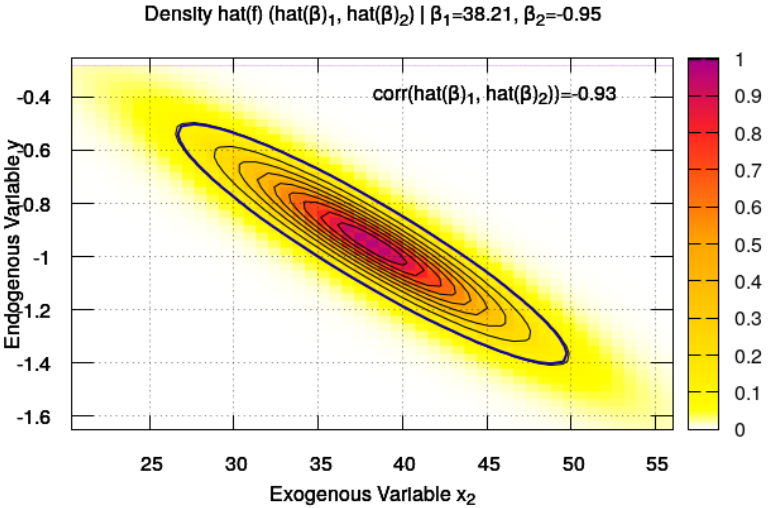
Effect of mis-fit parameters III: large effect if β_1 and β_2 have both positive misfits



Effect of mis-fit parameters IV: large effect if β_1 and β_2 have both negative misfits



All this results in a negative correlation between the estimation errors for β_1 and β_2



Special case 1: No exogenous variables

- Model: $y = \beta_0 + \epsilon := \mu + \epsilon$
- System matrix: $\mathbf{X} = (1, 1, ..., 1)'$
- OLS estimator:

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n}, \quad \mathbf{X}'\boldsymbol{y} = \sum_i y_i = n\bar{y},$$
$$\hat{\beta}_0 = \hat{\mu} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y} = \bar{y}$$

- Variance: $V_{00} = V(\hat{\mu}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} = \frac{\sigma^2}{n}, \quad \hat{V}_{00} = \frac{\hat{\sigma}^2}{n}$
- Distribution of the estimator (if $\epsilon \sim i.i.d N(\mu, \sigma^2)$)

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{V_{00}}} = \frac{\bar{y} - \mu}{\sigma}\sqrt{n} \sim N(0, 1),$$
$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\hat{V}_{00}}} = \frac{\bar{y} - \mu}{\hat{\sigma}}\sqrt{n} \sim T(n-1)$$

Special case 2: Simple linear regression

- Model (with $x_1 = x$): $y = \beta_0 + \beta_1 x + \epsilon$
- System matrix:

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \mathbf{X}'\mathbf{X} = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum x_i^2 \end{pmatrix}$$

- OLS estimator (with $s_x^2 = 1/n(\sum x_i^2 - n\bar{x})$):

$$\left(\mathbf{X}'\mathbf{X}\right)^{-1} = \frac{1}{ns_x^2} \begin{pmatrix} \frac{\sum x_i^2}{n} & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}, \quad \mathbf{X}'\boldsymbol{y} = \begin{pmatrix} n\bar{y} \\ \sum x_i y_i \end{pmatrix}$$

$$\begin{aligned} \hat{\beta}_1 &= \left(-\frac{\bar{x}}{ns_x^2}, \frac{1}{ns_x^2}\right) \begin{pmatrix} n\bar{y} \\ \sum x_i y_i \end{pmatrix} = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}} = \frac{s_{xy}}{s_x^2}, \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned}$$

Simple linear regression (ctnd)

- Variance-covariance matrix (assuming w/o loss of generality $\bar{x} = 0$):

$$\mathbf{V}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1} = \sigma^2 \begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{n s_x^2} \end{pmatrix}$$

- Variance of the estimator $\hat{y}(x)$ (x is deterministic):

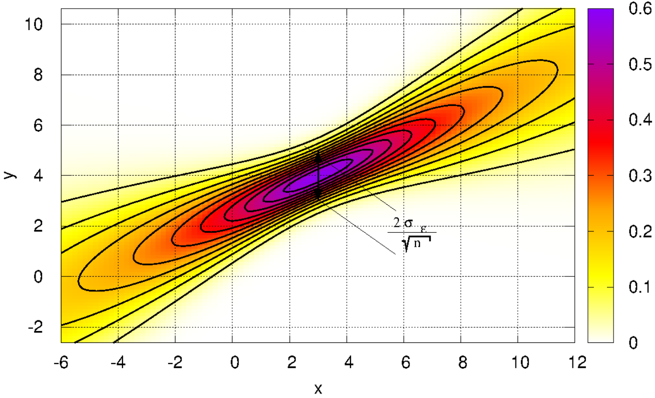
$$V(\hat{y}(x)) = V(\hat{\beta}_0 + \hat{\beta}_1 x) = V_{00} + x^2 V_{11} + 2x V_{01} = \frac{\sigma^2}{n}\left(1 + \frac{x^2}{s_x^2}\right)$$

- Distribution of the estimator for $y(x)$:

$$\hat{y}(x) \sim N\big(y(x), V(\hat{y}(x))\big)$$

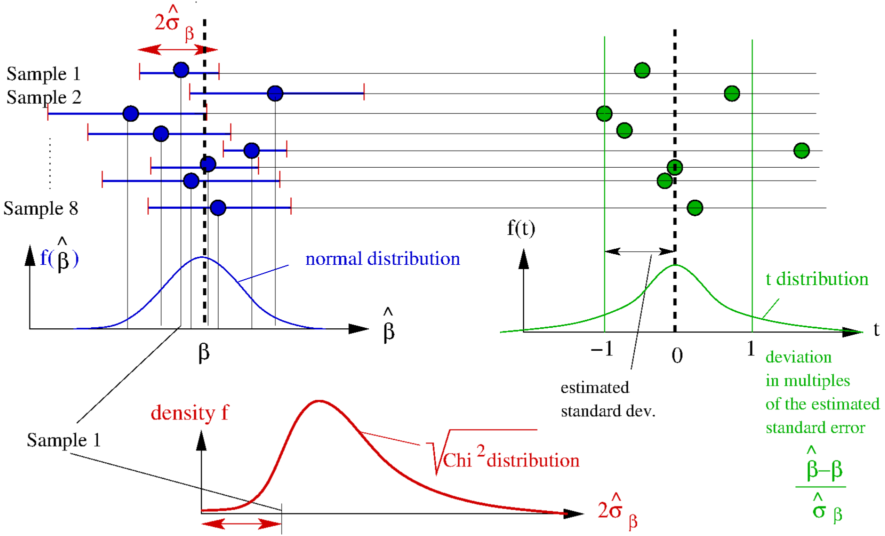
  If σ^2 has to be estimated by $\hat{\sigma}^2$, the normalized estimators for β_0, β_1 and $y(x)$ are $\sim T(n-2)$.

Probability density for $\hat{y}(x)$ for simple linear regression

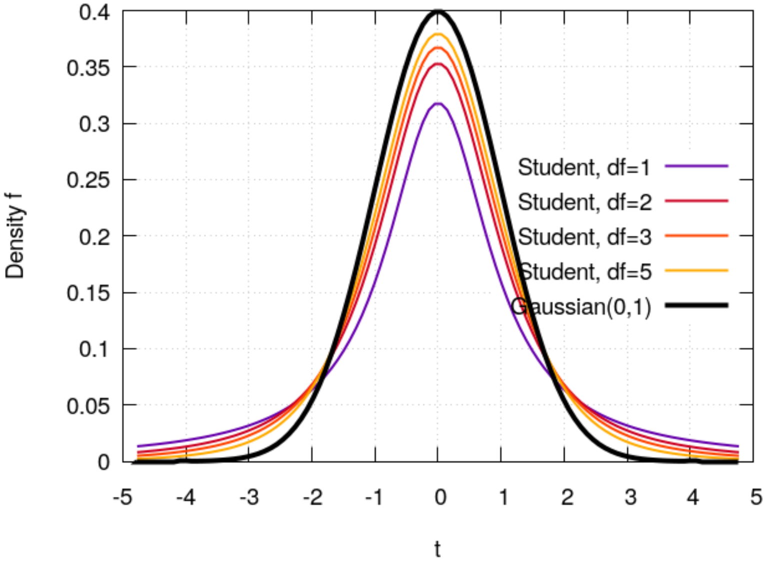


- If the Gauß-Markov assumptions apply, the model estimation errors $\hat{y}(x) - y(x)$ are Gaussian distributed
- The expectation and variance depends on x; the standard error is hyperbola-shaped.

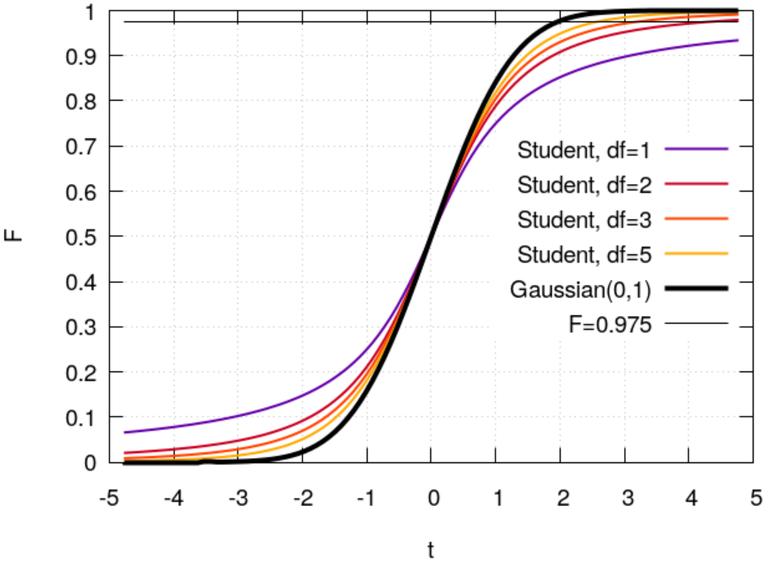
3.2. Confidence Intervals: where the Student-t distribution comes from



Densities of standard normal vs. Student-t distribution



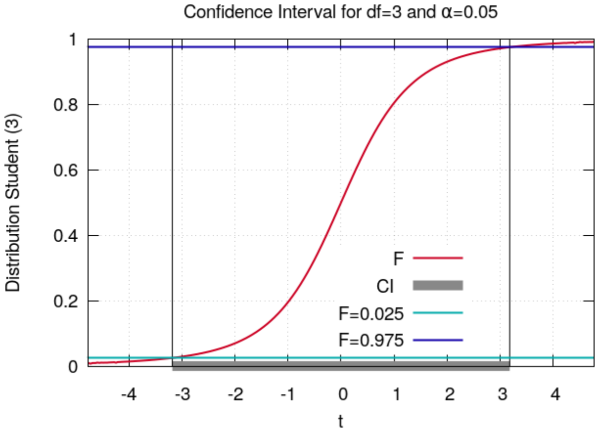
Distributions of standard normal vs. Student-t-distribution



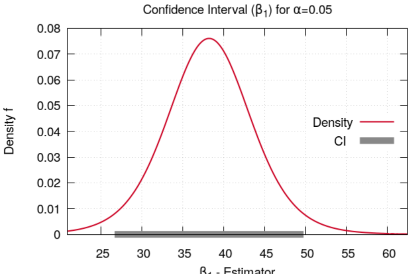
Calculation of the confidence intervals (CI)

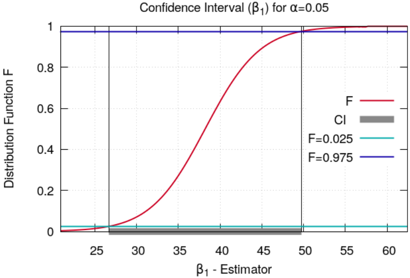
$$\mathrm{CI}_{\beta_j}^{(\alpha)} : \beta_j \in \left[\hat{\beta}_j - \Delta\hat{\beta}_j, \hat{\beta}_j + \Delta\hat{\beta}_j \right], \quad \Delta\hat{\beta}_j = t_{1-\alpha/2}^{(n-J-1)} \hat{\sigma}_{\hat{\beta}_j}.$$

- $t_{1-\alpha/2}$: Quantile (inverse of) the distribution function
- CI "uncertainty principle": Higher sensitivity implies higher α error.



Hotel example: CI for the appraisal for "stars" β_1 (full model)





Model: $y(\boldsymbol{x}) = \sum_j \beta_j x_j + \epsilon$

Factors:
$x_0 = 1$, x_1: #stars, x_2: price

Confidence interval (CI):

$$\beta_1 \in \left[\hat{\beta}_1 - \Delta\hat{\beta}_1^{(\alpha)}, \hat{\beta}_1 + \Delta\hat{\beta}_1^{(\alpha)}\right]$$

$$\Delta\hat{\beta}_1^{(\alpha)} = t_{1-\alpha/2}^{(n-3)} \sqrt{\hat{V}(\hat{\beta}_1)}$$

$$\hat{V}(\hat{\beta}_1) = \hat{\sigma}_\epsilon^2 \left[(\mathbf{X}'\mathbf{X})^{-1}\right]_{11}$$

$$\hat{\sigma}_\epsilon^2 = \frac{1}{n-3} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$