The Classical Regression Model

The error u in the multiple regression model (50) is independent of $X_1, \ldots, X_K$ and follows a normal distribution:

$$u \sim \text{Normal}\left(0, \sigma^2\right). \tag{51}$$

This assumption implies the more general assumptions (28) and (39):

$$\begin{aligned}
\mathrm{E}(u|X_1, \ldots, X_K) &= \mathrm{E}(u) = 0, \\
\mathrm{Var}\left(u|X_1, \ldots, X_K\right) &= \mathrm{Var}\left(u\right) = \sigma^2.
\end{aligned}$$

The Classical Regression Model

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

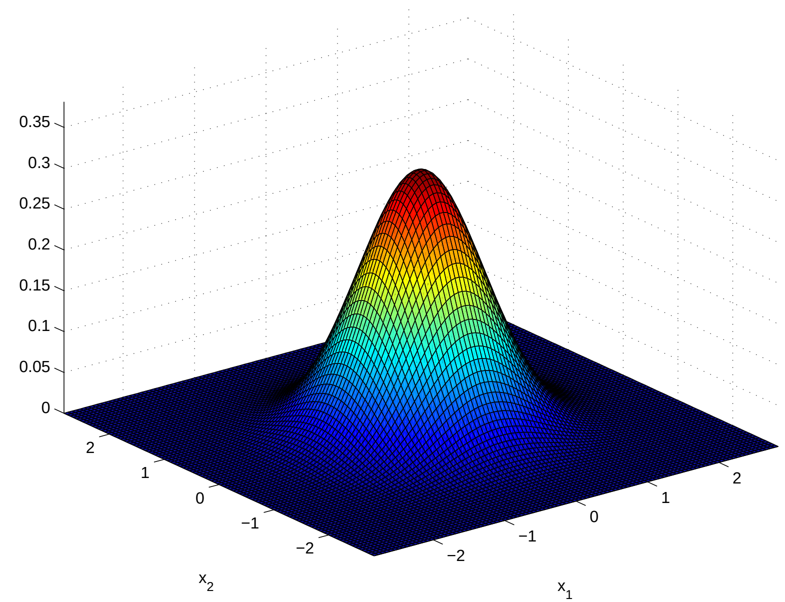
$$Y|X_1, \ldots, X_K \sim \text{Normal}\left(\beta_0 + \beta_1 X_1 + \ldots + \beta_j X_j + \ldots + \beta_K X_K, \sigma^2\right).$$

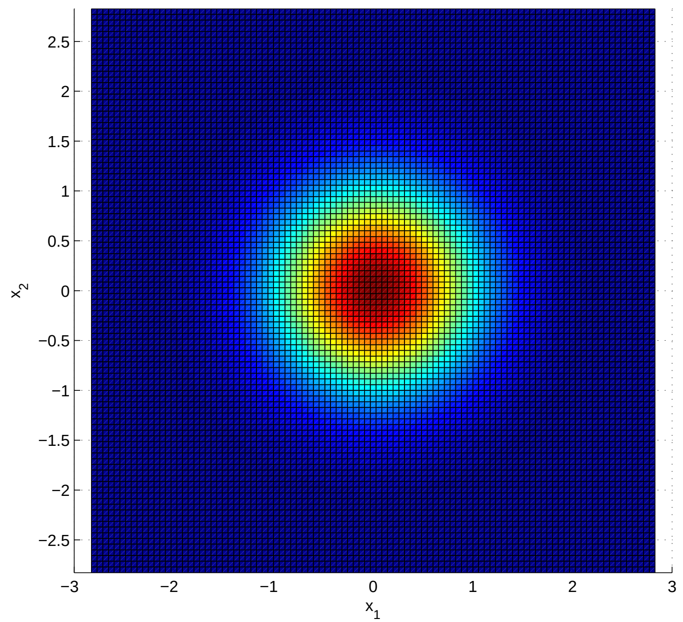
Furthermore, because the observations are a random sample, the vector $\boldsymbol{u}$ has a multivariate normal distribution with independent components:

$$\boldsymbol{u} \sim \text{Normal}_N\left(\mathbf{0}, \sigma^2 \mathbf{I}\right).$$

Multivariate normal distributions - independent componen

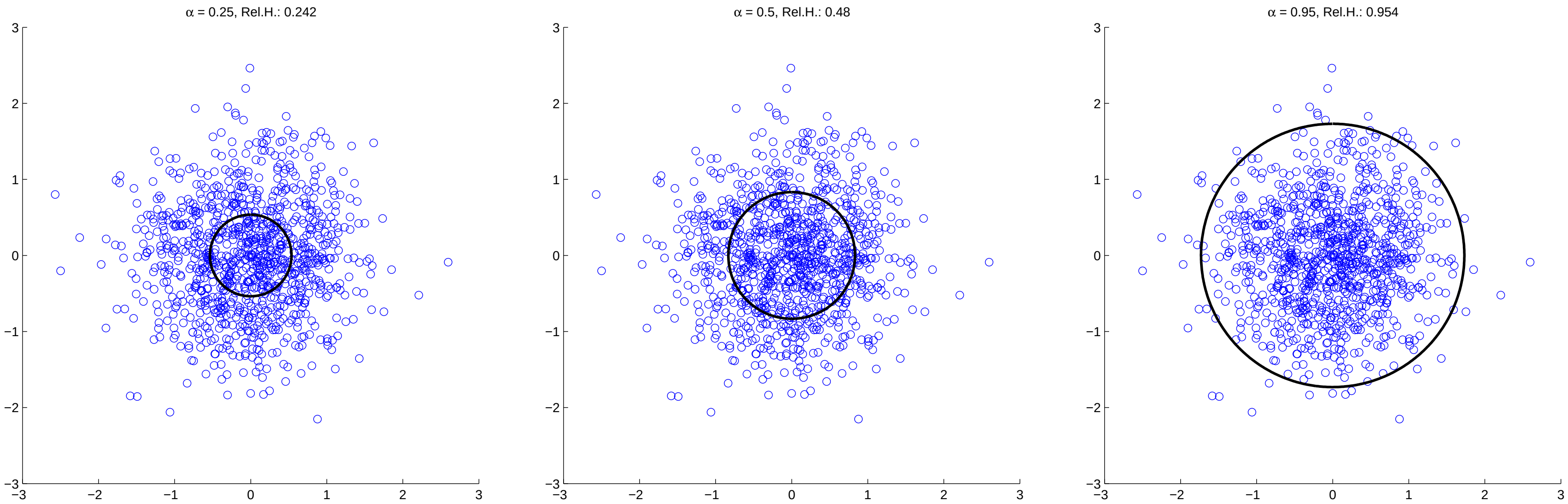
Density of the bivariate normal distribution $\mathrm{Normal}_2\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$ with $\sigma^2 = 0.5$.





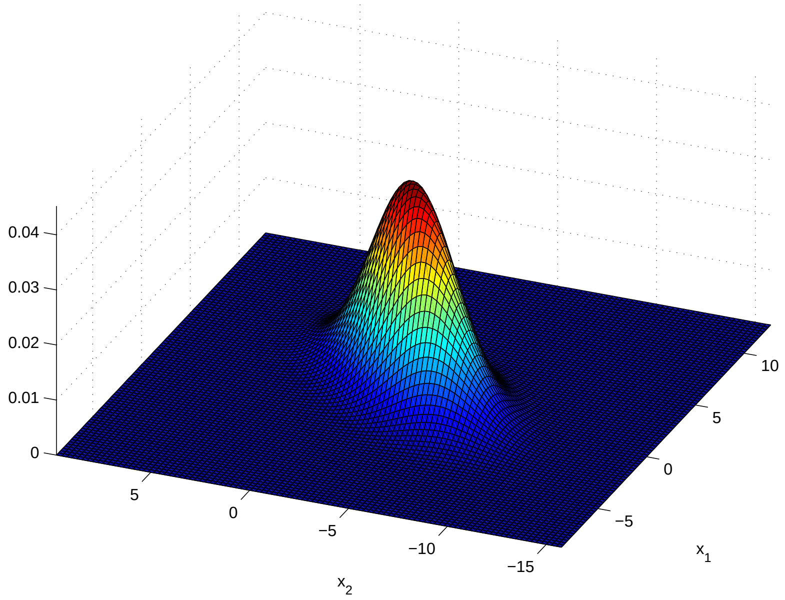
Multivariate normal distributions - independent componen

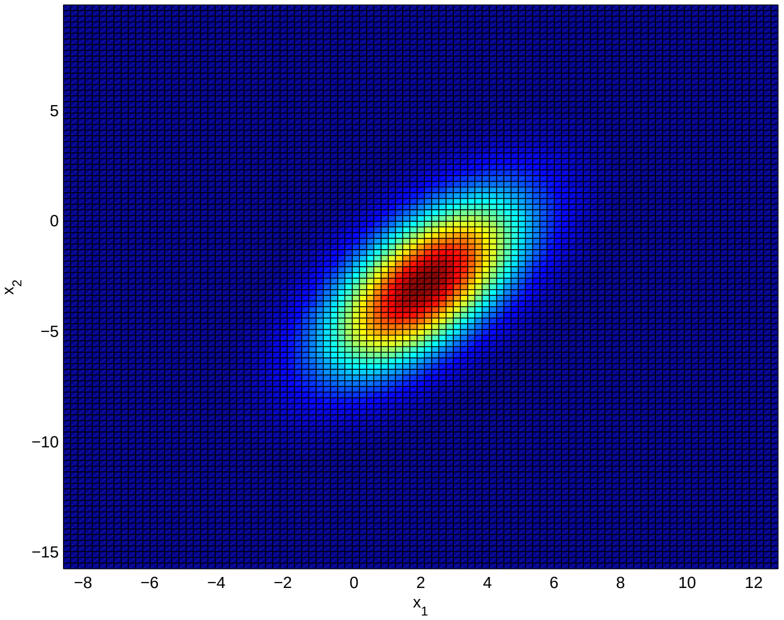
1000 observations from $\mathrm{Normal}_2\left(\mathbf{0}, 0.5\mathbf{I}\right)$ in comparison to 100α%-confidence region (from the left to the right: $\alpha = 0.25, \alpha = 0.5, \alpha = 0.95$)



Multivariate normal distributions - dependent components

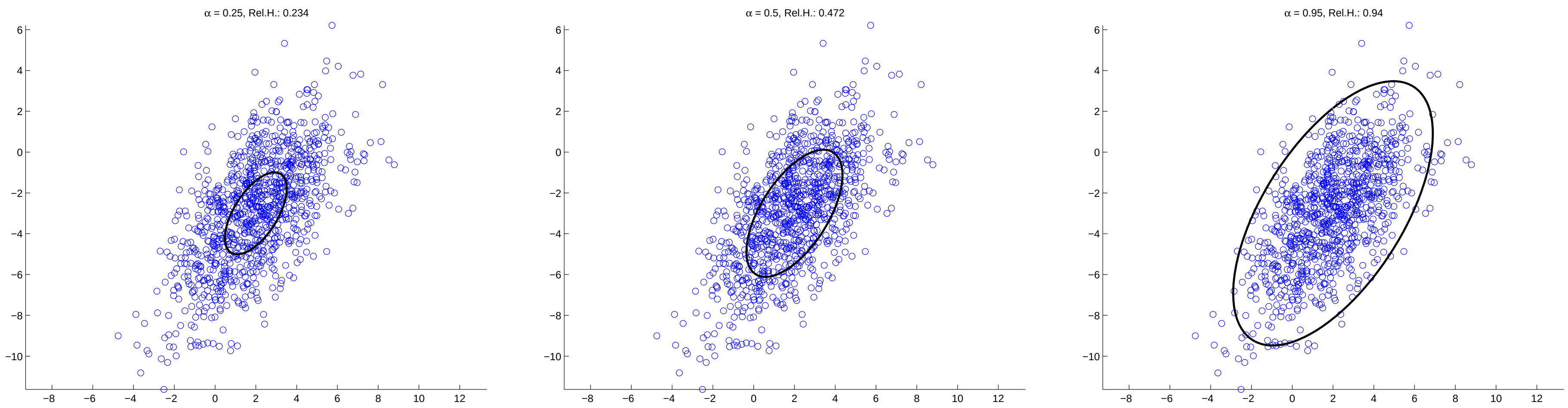
Density of the bivariate normal distribution $\mathrm{Normal}_2\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ with $\boldsymbol{\mu} = (2, -3)'$ and $\boldsymbol{\Sigma} = \begin{pmatrix} 4 & 3.2 \\ 3.2 & 7 \end{pmatrix}$.





Multivariate normal distributions - dependent components

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



Distribution of the OLS estimator

Using (38), we obtain:

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

Deviations between the true value and the OLS estimator are usually correlated.

All marginal distributions are normal, hence:

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

$$\frac{\hat{\beta}_j - \beta_j}{\text{sd}(\hat{\beta}_j)} \sim \text{Normal}\left(0, 1\right). \tag{52}$$

Testing a single coefficient - the t-test

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

$$\frac{|\hat{\beta}_j|}{\mathrm{sd}(\hat{\beta}_j)} \leq c_\alpha, \tag{53}$$

where c_α is equal to the $(1-\alpha/2)$-quantile of the standard normal distribution. (53) is used to construct a test statistic:

$$t_j = \frac{\hat{\beta}_j}{\mathrm{sd}(\hat{\beta}_j)}. \tag{54}$$

Testing a single coefficient - the t-test

If (51) 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.
- 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 $\mathrm{sd}(\hat{\beta}_j)$ is substituted by $\mathrm{se}(\hat{\beta}_j)$, yielding the test statistic:

$$t_j = \frac{\hat{\beta}_j}{\mathrm{se}(\hat{\beta}_j)}. \tag{55}$$

Choosing the quantiles of the normal distributions would lead to a test which rejects the null-hypothesis more often than desired, e.g. for $\alpha = 0.95$ 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

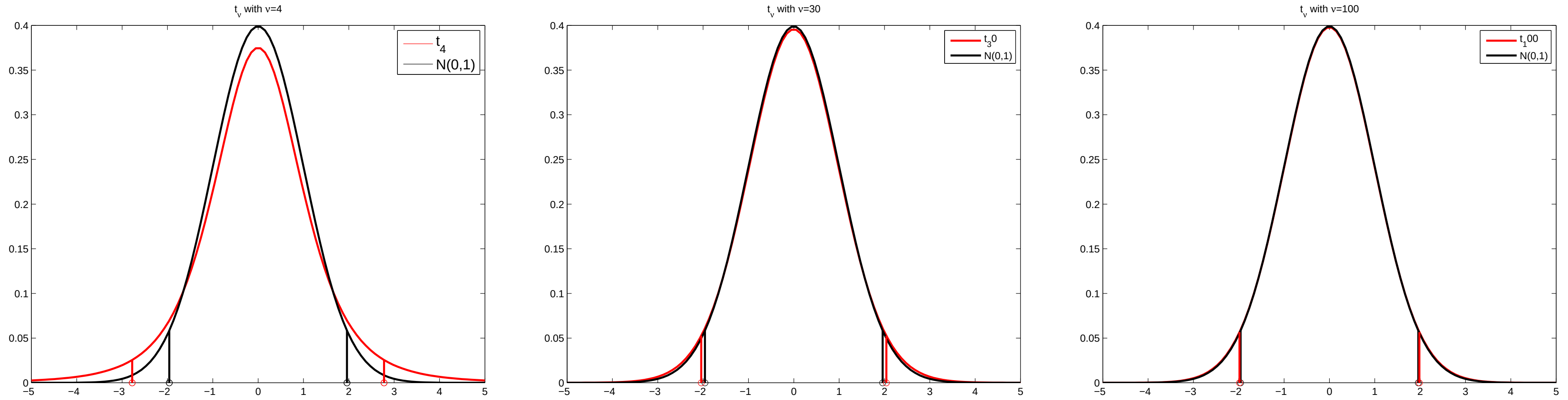
Choice of c_α, when σ^2 is unknown

The reason is that t_j no longer follows a normal distribution, but a t_{df}-distribution where $\mathrm{df} = (N-K-1)$. The critical values $t_{\mathrm{df},1-\alpha/2}$ depend on df and are equal to the quantiles of the t_{df}-distribution. For $\alpha = 0.95$ and for a regression model with 3 parameters, e.g. these values are:

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

The student t-distribution

The following figures show the density of the student t_{df} distribution for various degrees of freedom (from the left to the right: $\mathrm{df} = 4, \mathrm{df} = 30, \mathrm{df} = 100$)



The t_{df}-distribution converges to the standard normal distribution, as df goes to infinity.

The p-value

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).

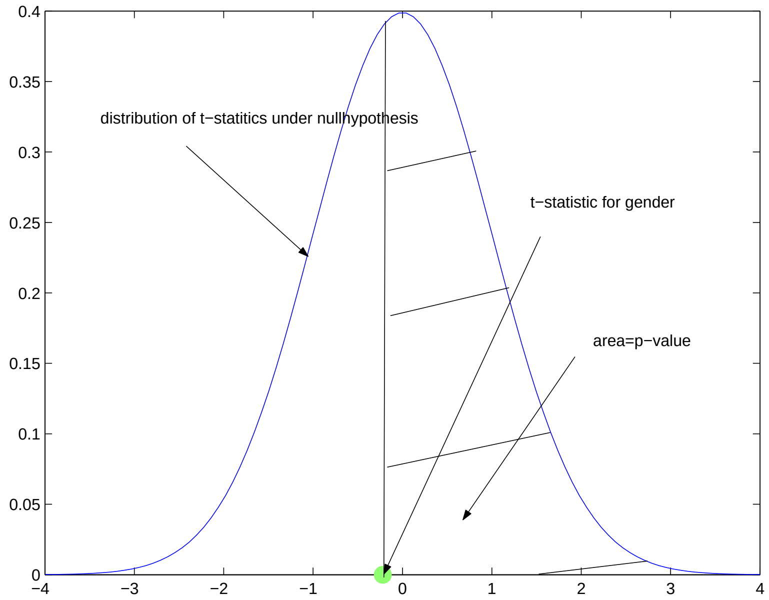
EVIEWS Exercise II.6.1

Discuss in EVIEWS how to formulate sensible null hypotheses and how to test them using the t-statistic and the p-value.

- Case Study profit, workfile `profit`;
- Case Study Chicken, workfile `chicken`;
- Case Study Marketing, workfile `marketing`;

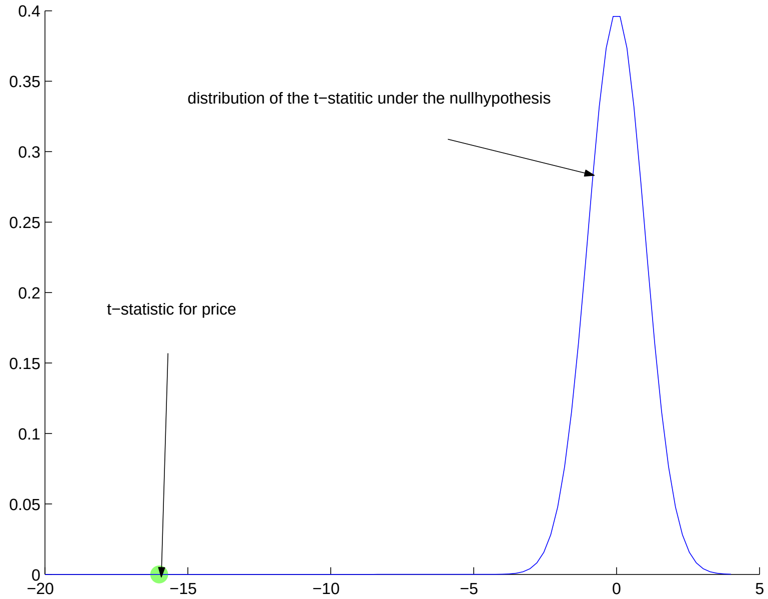
Case Study Marketing

The t-statistic for the variable `gender` is equal to -0.38, p-value:0.704



Case Study Marketing

The t-statistic for the variable `price` is equal to -16.1, p-value:0



Understanding p-values

- A small p-value shows that the value observed for the t-statistic is unlikely under the null hypothesis, thus we reject the null hypothesis for small p-values.
  $\Rightarrow$ There is high evidence in the data that $\beta_j \neq 0$.

- A p-value considerable larger than 0 shows that the observed value for the t-statistic is likely under the null hypothesis. Do not reject the null hypothesis that $\beta_j = 0$.
  $\Rightarrow$ There is no evidence in the data that we should reject the null hypothesis. Note, however, this does not necessarily mean that $\beta_j = 0$ (risk to accept a wrong null hypothesis is not controlled).

Confidence intervals for the unknown coefficients

The marginal distribution (52) is also useful for obtaining $100(1-\alpha)$ confidence regions for the unknown regression coefficients (e.g. $\alpha = 0.05$ leads to a 95% confidence region).

Two-sided confidence regions:

$$\Pr\{-c_{1-\alpha/2} \leq \frac{\hat{\beta}_j - \beta_j}{\mathrm{sd}(\hat{\beta}_j)} \leq c_{1-\alpha/2}\} = 1 - \alpha, \qquad (56)$$

where c_p is the p-quantile of the standard normal distribution, i.e. the confidence interval reads:

$$[\hat{\beta}_j - c_{1-\alpha/2}\mathrm{sd}(\hat{\beta}_j), \hat{\beta}_j + c_{1-\alpha/2}\mathrm{sd}(\hat{\beta}_j)]$$

Confidence intervals for the unknown coefficients

One-sided confidence regions:

$$\Pr\{\frac{\hat{\beta}_j - \beta_j}{\mathrm{sd}(\hat{\beta}_j)} \leq c_{1-\alpha}\} = 1 - \alpha,$$

$$\Pr\{-c_{1-\alpha} \leq \frac{\hat{\beta}_j - \beta_j}{\mathrm{sd}(\hat{\beta}_j)}\} = 1 - \alpha.$$

This yields (with probability $1 - \alpha$):

- $\hat{\beta}_j - c_{1-\alpha}\mathrm{sd}(\hat{\beta}_j)$ is a lower bound for β_j,
- $\hat{\beta}_j + c_{1-\alpha}\mathrm{sd}(\hat{\beta}_j)$ is an upper bound for β_j.

Confidence intervals for the unknown coefficients

If σ^2 is unknown, then $\mathrm{sd}(\hat{\beta}_j)$ is substituted by $\mathrm{se}(\hat{\beta}_j)$. Instead of (52), we obtain with $\mathrm{df} = (N - K - 1)$:

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

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

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

More about the distribution of the OLS estimator

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

$$\tilde{\hat{\boldsymbol{\beta}}} - \tilde{\boldsymbol{\beta}} \sim \mathrm{Normal}_q \left(\mathbf{0}, \mathrm{Cov}(\tilde{\hat{\boldsymbol{\beta}}}) \right), \tag{57}$$

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

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

Testing more than one coefficient

- 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 $j_1, \ldots, j_q$, then one needs to test, if $\beta_{j_1} = 0, \ldots, \beta_{j_q} = 0$ 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

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

Reject the null hypothesis, if the distance between the OLS estimator $\tilde{\hat{\boldsymbol{\beta}}} = (\hat{\beta}_{j_1}, \ldots, \hat{\beta}_{j_q})'$ 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.

Testing more than one coefficient

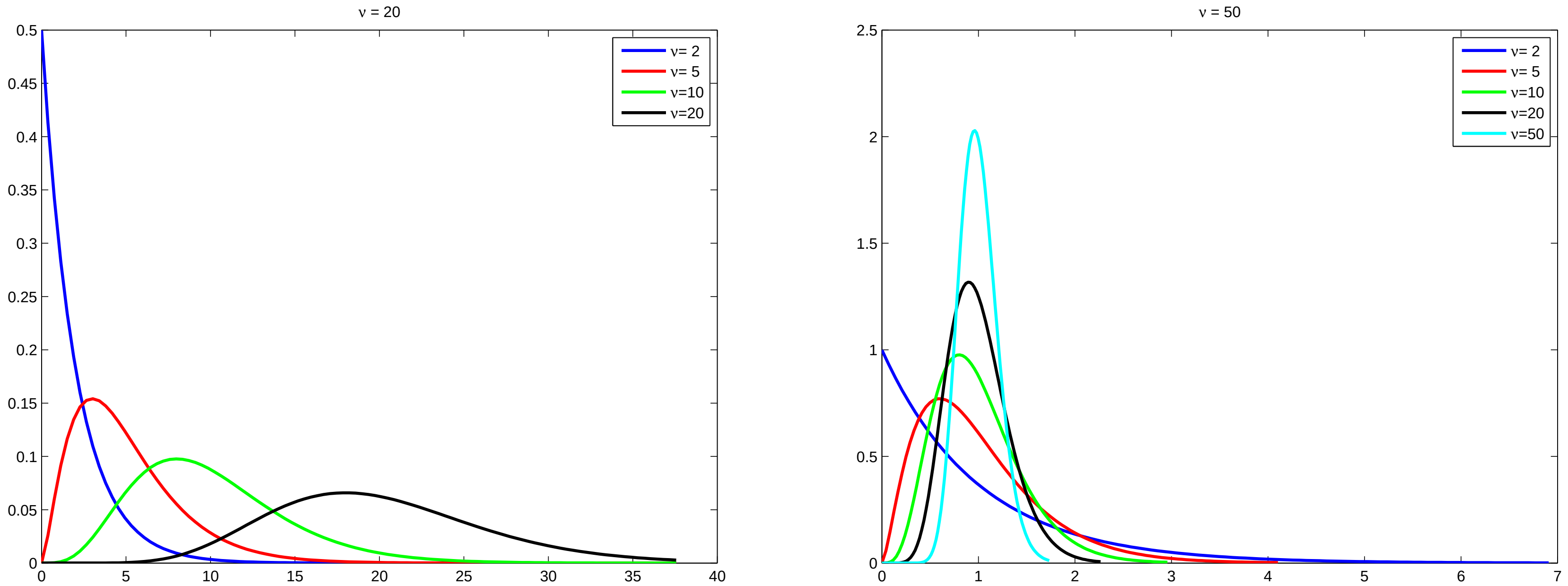
Aggregate $t_{j_l} = \hat{\beta}_{j_l}/\mathrm{sd}(\hat{\beta}_{j_l})$ for $l = 1, \ldots, q$, e.g. by taking the sum of squared t statistics?

If the deviations of the OLS estimators $\hat{\beta}_{j_1}, \ldots, \hat{\beta}_{j_q}$ from the true values are uncorrelated, then the aggregated test statistic

$$\sum_{l=1}^{q} \frac{\hat{\beta}_{j_l}^{\,2}}{\mathrm{sd}(\hat{\beta}_{j_l})^2}$$

is the sum of q independent squared standard normal random variables. Such a random variable follows a χ^2_q-distribution with q degrees of freedom.

The χ^2_q distribution



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

Testing more than one coefficient

Usually, the deviations of the OLS estimators $\hat{\beta}_{j_1}, \ldots, \hat{\beta}_{j_q}$ from the true values are correlated:

- Transform the deviations to a coordinate system with independent standard normal random variables. In this new coordinate system, the sum of squared deviations follow a χ^2_q-distribution with q degrees of freedom. The appropriate transformation reads:

$$\tilde{\hat{\boldsymbol{\beta}}}' \mathrm{Cov}(\tilde{\hat{\boldsymbol{\beta}}})^{-1} \tilde{\hat{\boldsymbol{\beta}}} \sim \chi^2_q$$

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

The F-Test

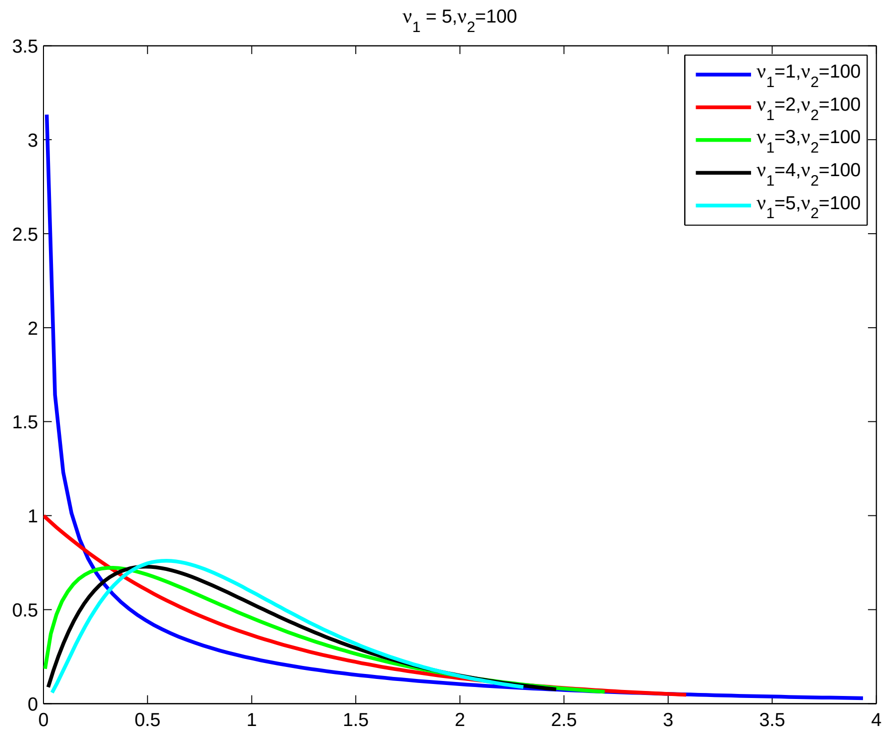
- The F-statistic is obtained by substituting the unknown variance σ^2 by $\hat{\sigma}^2$ and dividing by q.

- The F-statistic is the ratio of two (independent) sum of squares, divided by the degrees of freedom, i.e. a χ^2_q/q and $\chi^2_{\mathrm{df}}/\mathrm{df}$, where $\mathrm{df} = N - K - 1$.

- If the null hypothesis $\beta_{j_1} = 0, \ldots, \beta_{j_q} = 0$ is true, then the F-statistic follows a $F_{q,\mathrm{df}}$-distribution with parameters q (number of tested coefficients) and $\mathrm{df} = N - K - 1$.

- Remark: for $q = 1$, $F = t_j^2$, where t_j is the t-statistic.

The F-distribution



Density of the $F_{q,\mathrm{df}}$-distribution with parameters $\mathrm{df} = 100$ and $q = 1, \ldots, 5$.

The F-Test

Reject the null-hypothesis, if

- the F-statistic is larger than the critical value from the corresponding $F_{q,\mathrm{df}}$-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 unlikely under the null hypothesis.

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

The F-Test

Do not reject the null-hypothesis, if

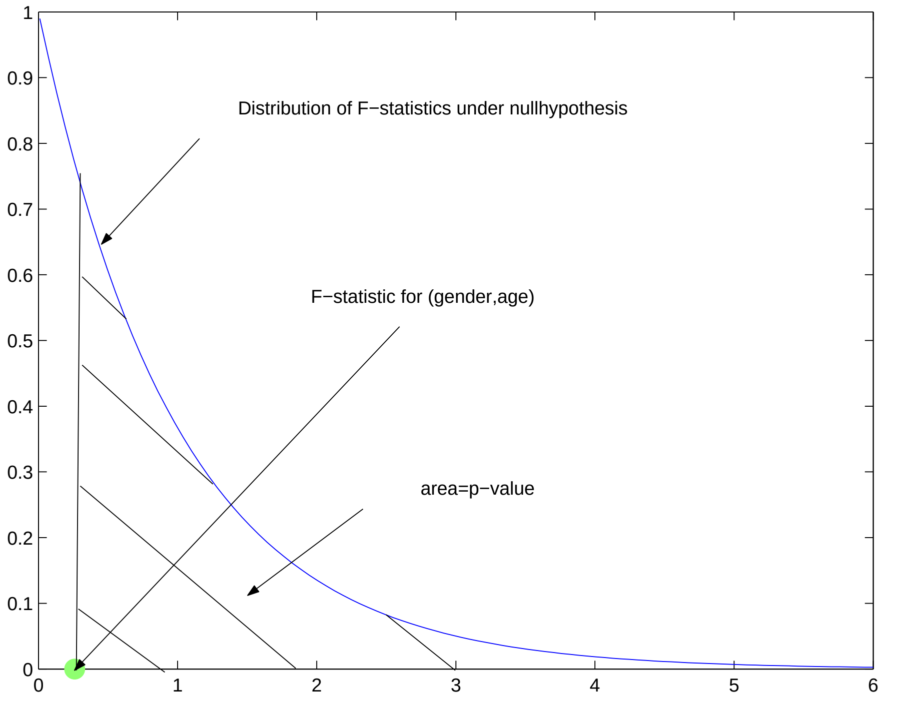
- the F-statistic is smaller than the critical value from the corresponding $F_{q,\mathrm{df}}$-distribution (one-sided test).

- 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 likely under the null hypothesis.

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

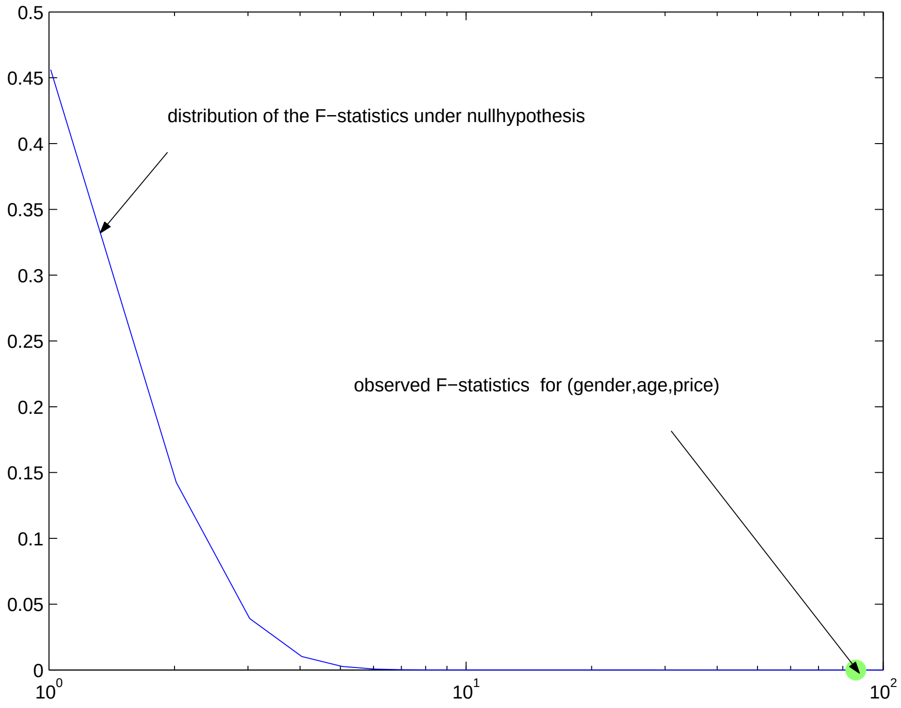
The F-Test

The F-statistic for testing the variables `gender,age` is equal to 0.263, p-value:0.769



The F-Test

The F-statistic for testing the variables `gender,age,price` is equal to 86.24, p-value:0.



An alternative form of the F-statistic

Equivalent forms of the F-statistic show that the F-statistic measures the loss of fit from imposing the q restrictions on the model:

$$F = \frac{(\mathrm{SSR}_r - \mathrm{SSR})/q}{\mathrm{SSR}/\mathrm{df}}, \qquad F = \frac{(\mathrm{R}^2 - \mathrm{R}_r^2)/q}{(1-\mathrm{R}^2)/\mathrm{df}},$$

- SSR is the minimum sum of squared residuals and R^2 is the coefficient of determination for the unrestricted regression model.
- SSR_r is the minimum sum of squared residuals and R_r^2 is the coefficient of determination for the restricted regression model.
- Note that $\mathrm{SSR}_r > \mathrm{SSR}$ and $\mathrm{R}_r^2 < \mathrm{R}^2$.

Testing the whole regression model

In the standard regression output of EViews, a F-statistic is available by default. This F-statistics test the hypothesis that none of the predictor variables influences the response variable:

$$\beta_1 = 0, \dots, \beta_K = 0$$

In this case, $\mathrm{R}_r^2 = 0$, and the F-statistic reads:

$$F = \frac{\mathrm{R}^2/K}{(1-\mathrm{R}^2)/\mathrm{df}}.$$

Under the null hypothesis, $F \sim F_{K,\mathrm{df}}$-distribution. Hopefully, the corresponding p-value is close to 0. Otherwise, the usefulness of the whole regression model is somewhat doubtful.