Unit root, differencing the time series, unit root test (ADF test)

Beáta Stehlíková

Time series analysis

Outline

- What is a unit root and what is its consequence
- If we have unit root - how to transform the data, so that we can use the ARMA methodology
- How to find from the data that there is a unit root $\rightarrow$ unit root tests

Examples

- Consider the process $y_t = y_{t-1} + u_t$:
  - ◇ it is a nonstationary AR(1) process with a unit root
  - ◇ for its differences $\Delta y_t = y_t - y_{t-1}$ we have $\Delta y_t = u_t$
  - ◇ so Δy_t is a stationary process
- Consider a nonstationary process with a unit root

$$(1 - \frac{1}{2}L)(1 - L)x_t = 1 + (1 - \frac{1}{3}L)u_t$$

Then for the differences

$$\Delta y_t = y_t - y_{t-1} = (1 - L)y_t$$

we have

$$(1 - \frac{1}{2}L)\Delta y_t = 1 + (1 - \frac{1}{3}L)u_t,$$

so Δy_t is a stationary process

Examples

- Consider nonstationary process with a double unit root
$$(1-\frac{1}{2}L)(1-L)^2x_t = 1 + (1-\frac{1}{3}L)u_t$$
The for the second differences
$$\Delta^2 y_t = \Delta(\Delta y_t) = (1-L)(1-L)y_t = (1-L)^2 y_t$$
we have
$$(1-\frac{1}{2}L)\Delta^2 y_t = 1 + (1-\frac{1}{3}L)u_t,$$
so $\Delta^2 y_t$ is a stationary process
- In general:
If the multiplicity of the unit root is k (and the others are outsider the unit circle), then its k-th differences are stationary.

ARIMA models

- We need need to differentiate the process k times, in order to get a stationary process, it is called intergrated process of order k, denoted by $I(k)$
- If these k-th differences follow ARMA(p,q), then we say that the original time series is ARIMA(p,k,q).
- For example x_t, if

$$(1-\frac{1}{2}L)(1-L)^2x_t = 1 + (1-\frac{1}{3}L)u_t,$$

is ARIMA(1,2,1) process.

Our aim

- Consider firstly AR(1) model: $x_t = \delta + \rho x_{t-1} + u_t$
- We want to:
  - test a hypothesis about a unit root (then, the process is nonstationary), so $H_0 : \rho = 1$
  - find out it can be rejected in favour of the stationarity - $H_1 : \rho < 1$

Unit root and the t-statistics

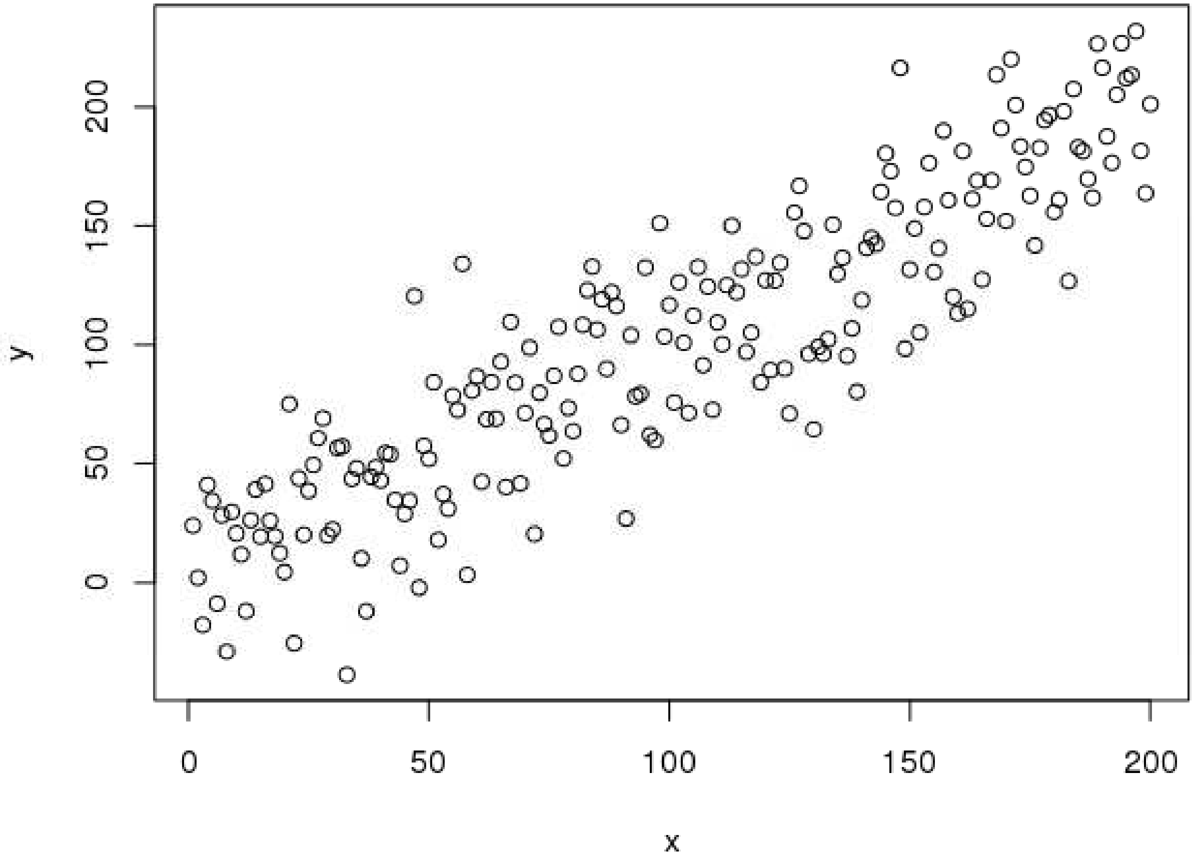
We try to use the hypothesis testing about coefficients of a regression model known from econometrics.

For example:

- Set x=1:200
- Simulate y=x+rnorm(200)*sigma
- Estimate the model $y = c + \rho x + \varepsilon$
- We note:
    - ◇ estimate of the parameter ρ
    - ◇ value of the t-statistics corresponding to hypothesis $H_0 : \rho = 1$ (which holds)
- Repeat 10^5 times and plot the histogram

Unit root and the t-statistics

- Example of the simulated data:



Unit root and the t-statistics

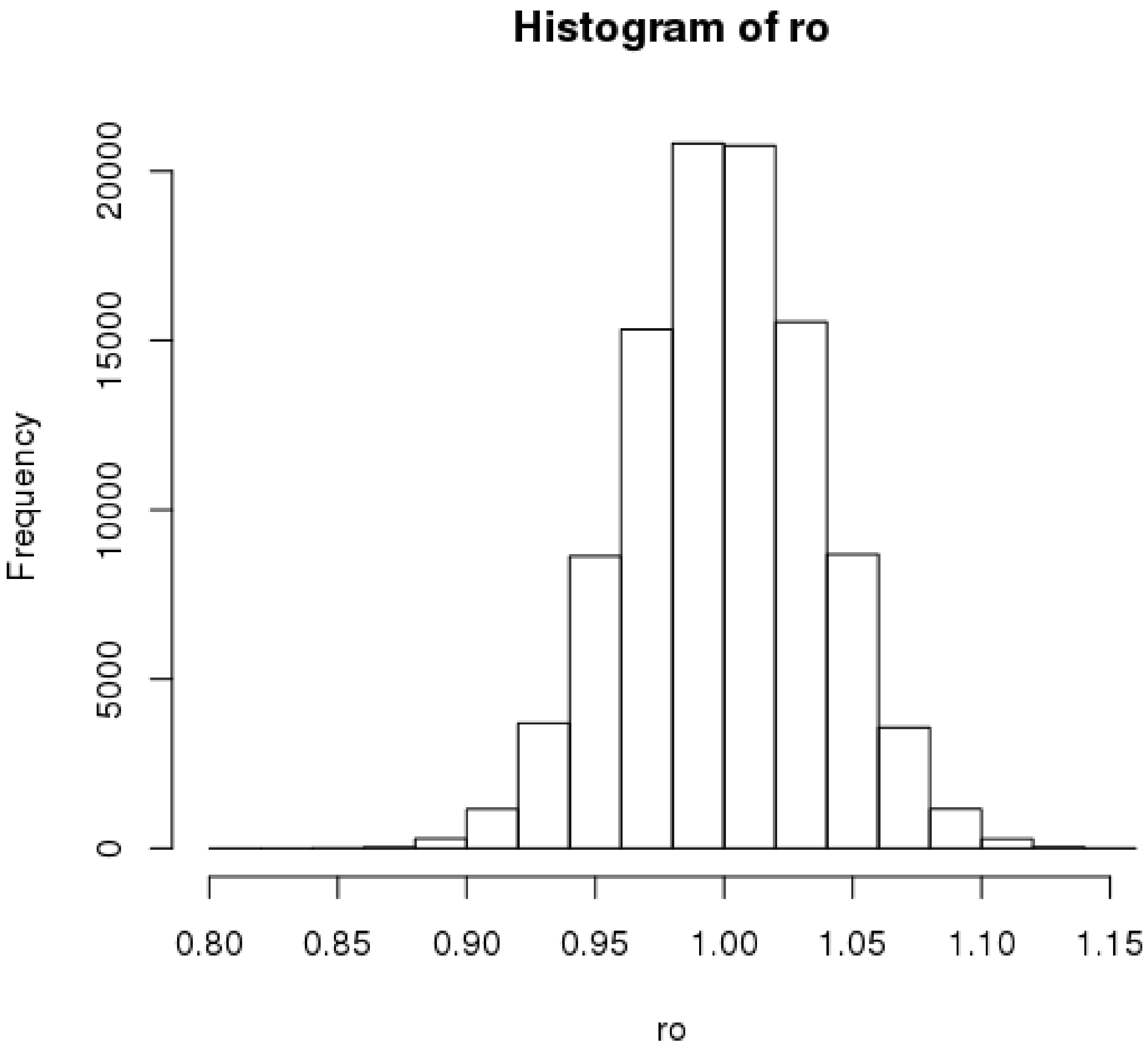
- Estimated regression:

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-75.067 -19.239   3.818  20.520  74.772

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.72079    4.00941   1.177     0.24
x            0.95689    0.03459  27.662   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28.24 on 198 degrees of freedom
Multiple R-squared: 0.7944,     Adjusted R-squared: 0.7934
F-statistic: 765.2 on 1 and 198 DF,  p-value: < 2.2e-16
```

- Estimated coefficient ρ is 0.95689
- T-statistics to test $\rho = 1$ is $\frac{0.95689-1}{0.03459}$

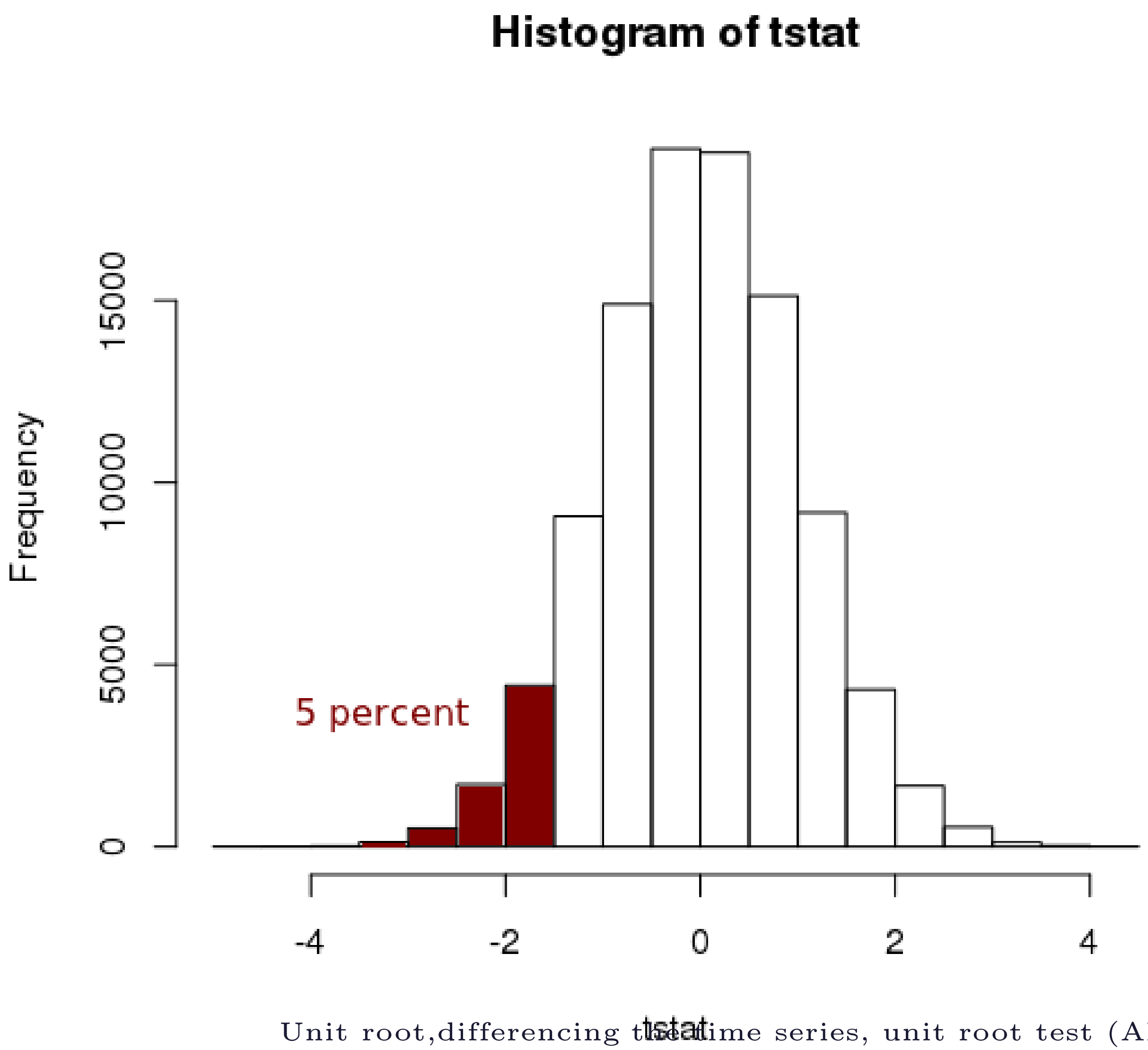
Unit root and the t-statistics

- Estimated of ρ: normal distribution



Unit root and the t-statistics

- t-statistics to test $H_0 : \rho = 1$: Student t-distribution



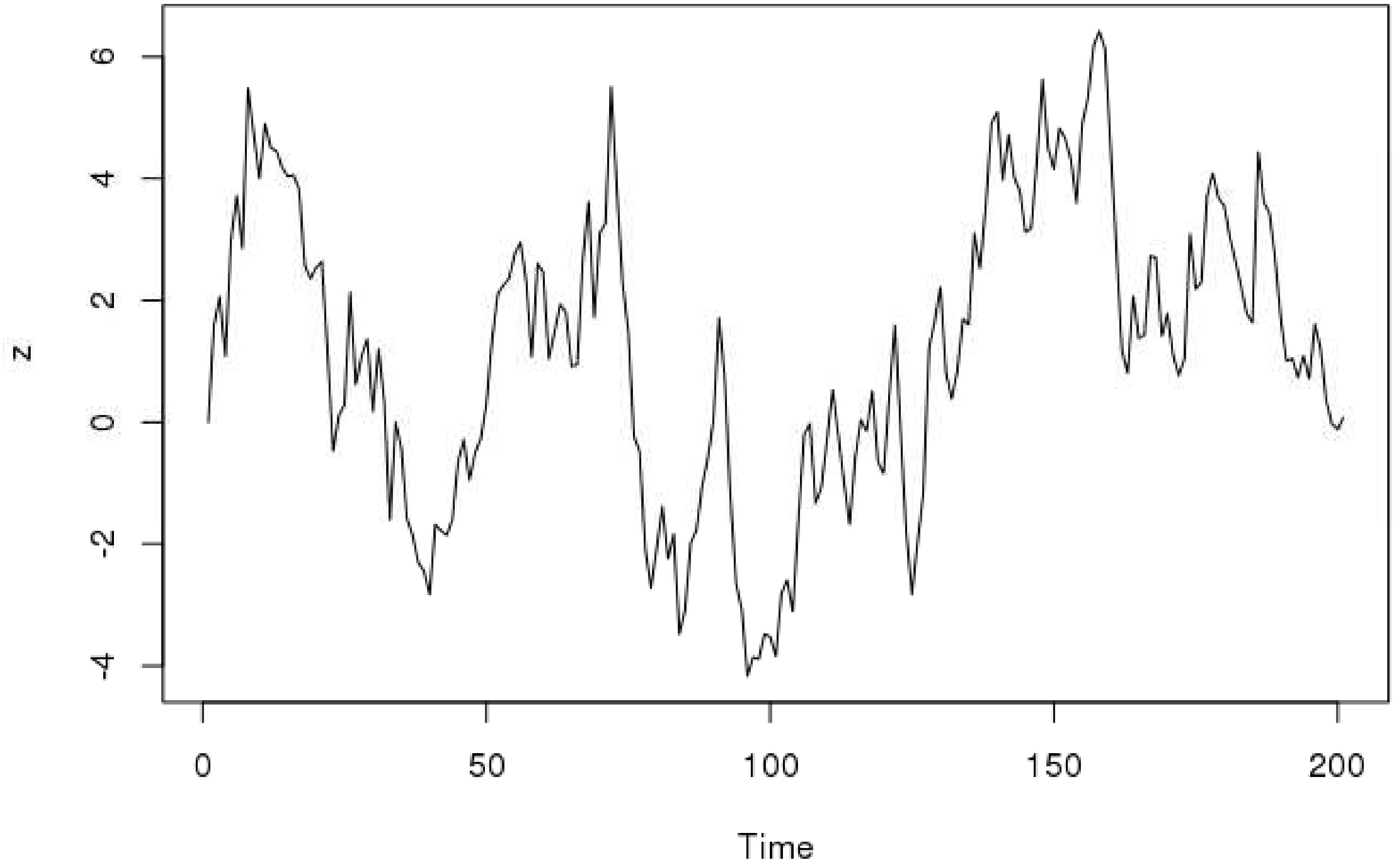
Unit root and the t-statistics

Another simulation:

- Consider vector z generated as $z_t = z_{t-1} + \varepsilon_t$
- Take x=z[1:200], y=z[2:201], so $y_t = z_t$, $x_t = z_{t-1}$
- Estimate the model $y = c + \rho x + \varepsilon$
- Note again:
  - ◇ estimate of parameter ρ
  - ◇ value of t-statistics to test $H_0 : \rho = 1$ (which holds)
- Repeat 10^5 times and plot the histogram

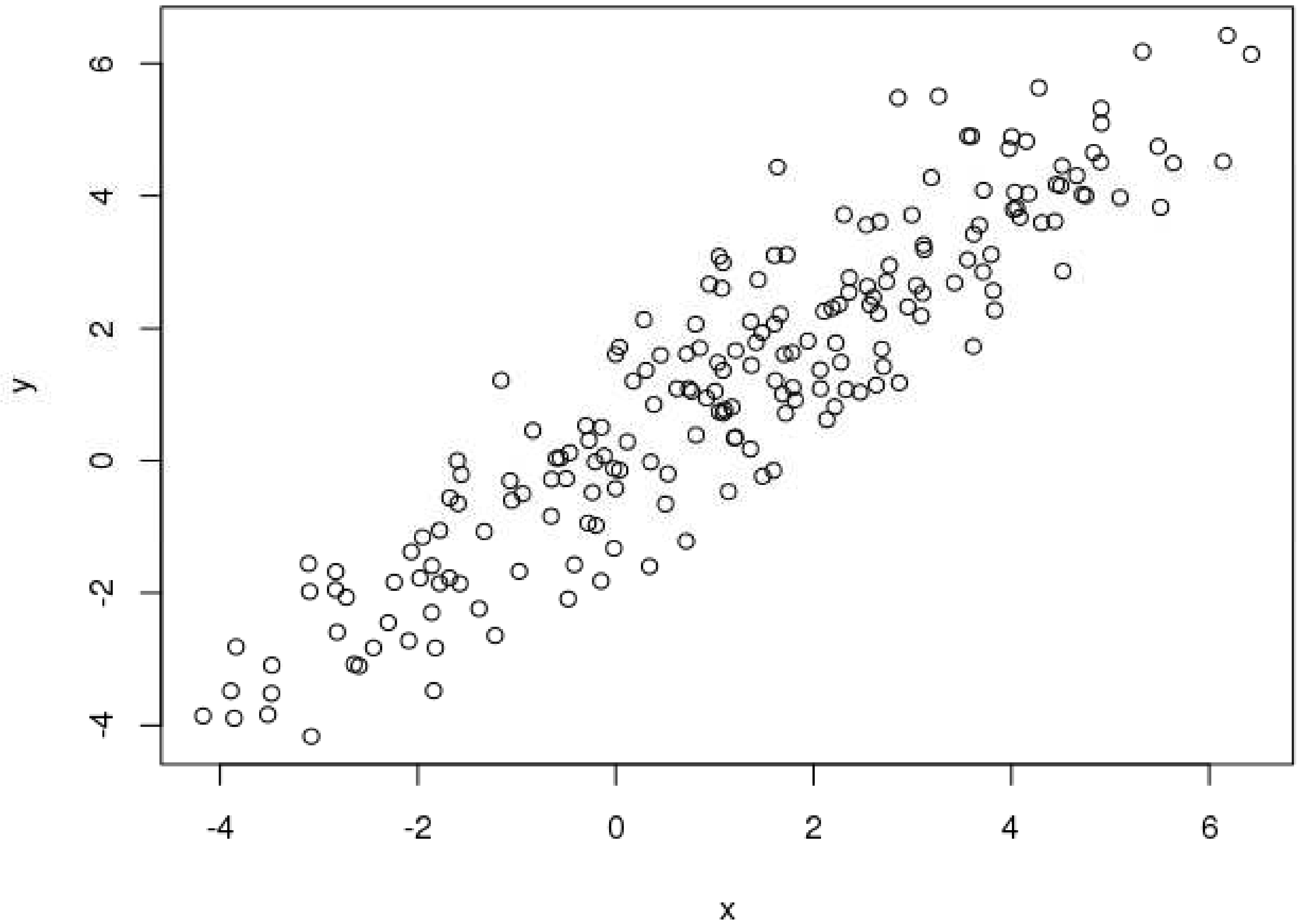
Unit root and the t-statistics

- Example of simulated data - time series s z:



Unit root and the t-statistics

- Example of simulated data - data for regression



Unit root and the t-statistics

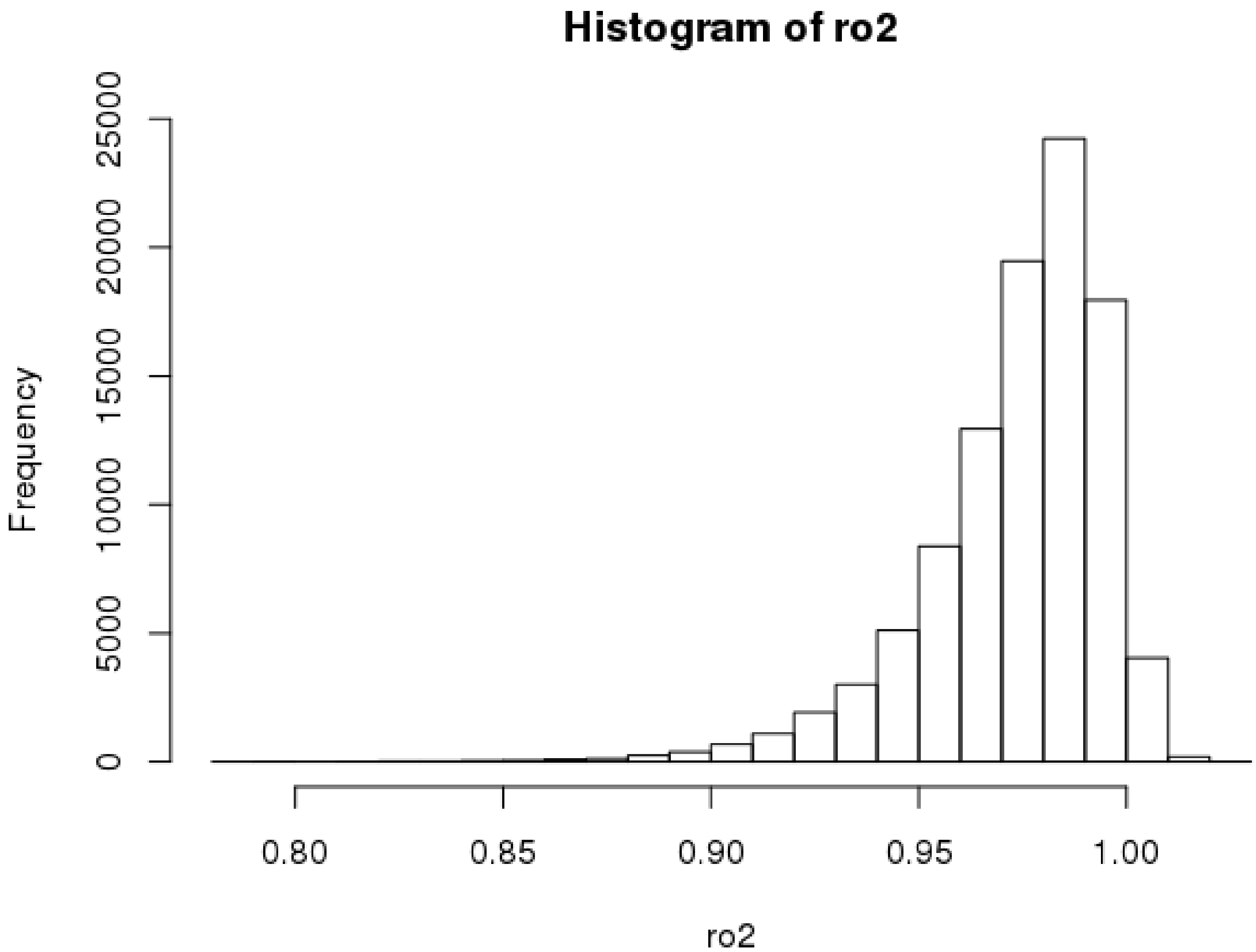
- Estimated regression:

```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-2.01047 -0.64723  0.00159  0.53995  2.84017

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.09589    0.07489    1.28    0.202
x            0.91962    0.02790   32.96   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9496 on 198 degrees of freedom
Multiple R-squared: 0.8458,     Adjusted R-squared: 0.845
F-statistic:  1086 on 1 and 198 DF,  p-value: < 2.2e-16
```

- Estimated coefficient ρ is 0.91962
- t-statistics to test $\rho = 1$ is $\frac{0.91962-1}{0.02790} = -2.88$

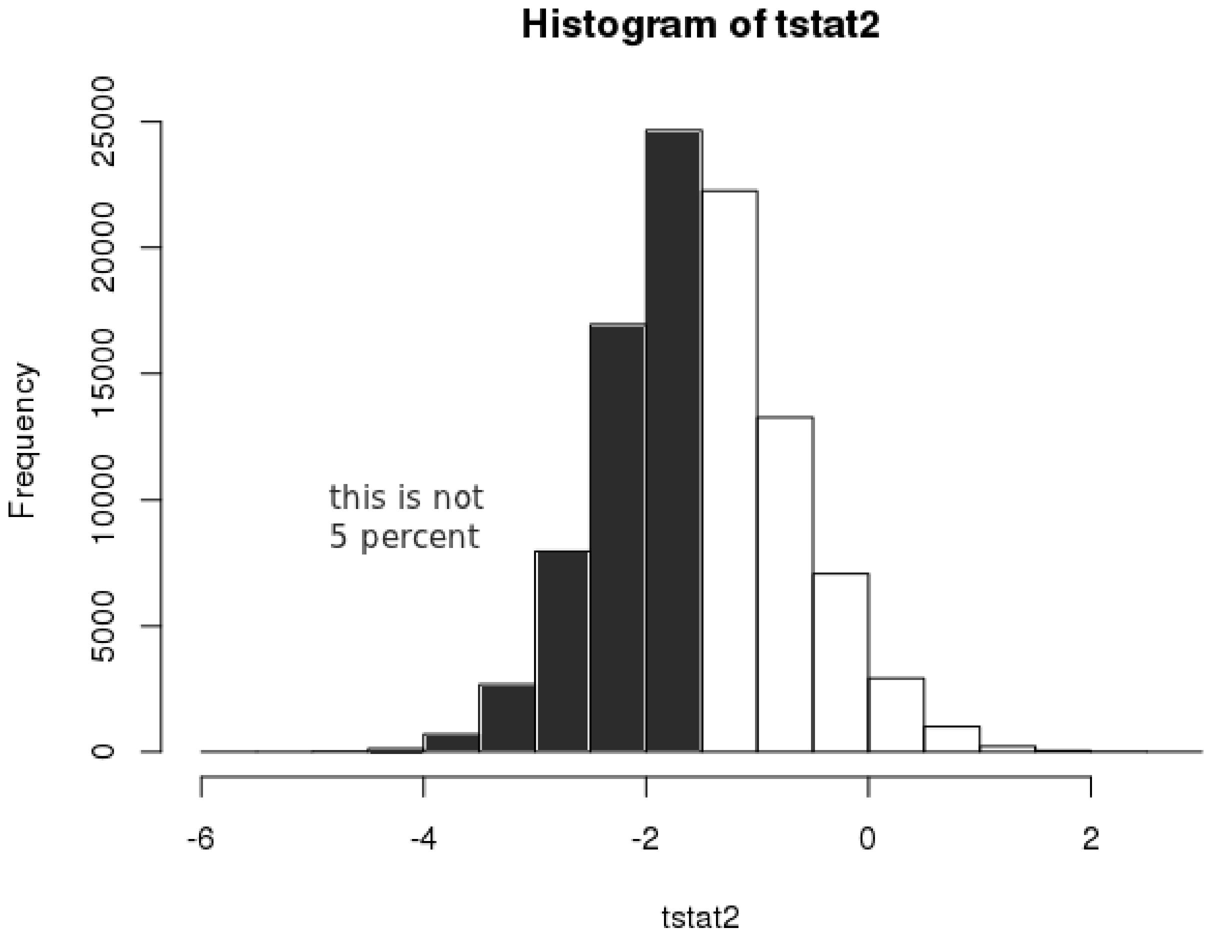
Unit root and the t-statistics

- Estimates of parameter ρ: not a normal distribution



Unit root and the t-statistics

- "t-statistics" to test $H_0 : \rho = 1$: does not have a t-distribution



- We cannot use critical values of the t-distribution

Unit root and the t-statistics

- Solution - the mail idea:
  - we keep the test statistics
  - but we use different critical values
- Quantiles from our simulations:

```
> quantile(tstat2,0.05)
      5%
-2.866481
> quantile(tstat2,0.01)
      1%
-3.462751
```

- critical values should be similar

- Question:
  - What if we have a process of a higher order instead of AR(1)?

Unit root tests

- AR(1) process:

  (1) $$y_t = \rho y_{t-1} + u_t$$

  unit root means that $\rho = 1$.
- Equivalently:

  $$\Delta y_t = (\rho - 1) y_{t-1} + u_t$$

  and we are interested in t-statistics from the significance of coefficient at y_{t-1} - but with another critical value
- This critical value
  - depends on number of data
  - changes, if equation (1) has a constant term or a linear drift
- In general: $\Delta y_t = \alpha + \beta t + (\rho - 1) y_{t-1} + u_t$

Unit root tests

- AR(p) process: $y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \dots \alpha_p y_{t-p} + u_t$
  unit root $\to \alpha_1 + \dots \alpha_p = 1.$
- Write in the form:

  $$y_t = \rho y_{t-1} + \theta_1 \Delta y_{t-1} + \theta_2 \Delta y_{t-2} + \dots + \theta_{p-1} \Delta y_{t-p+1} + u_t,$$

  where $\rho = \sum_{j=1}^{p} \alpha_j$, $\theta_i = -\sum_{j=i+1}^{p} \alpha_j$ $(i = 1, \dots, p-1)$
- Equivalently:

  $$\Delta y_t = (\rho - 1) y_{t-1} + \theta_1 \Delta y_{t-1} + \theta_2 \Delta y_{t-2} + \dots + \theta_{p-1} \Delta y_{t-p+1} + u_t$$

  and we are interested in t-statistics for coefficient at y_{t-1}
- In general: y can have a trend and/or intercept $\Rightarrow$
  $\Delta y_t = \alpha + \beta t + (\rho - 1) y_{t-1} + \theta_1 \Delta y_{t-1} + \theta_2 \Delta y_{t-2} + \dots$
  $+ \theta_{p-1} \Delta y_{t-p+1} + u_t$

Augmented Dickey-Fuller test (ADF)

Wayne A. Fuller (1976)

David A. Dickey, Wayne A. Fuller (1979, 1981)

- We estimate
  $\Delta y_t = \alpha + \beta t + (\rho - 1) y_{t-1} + \theta_1 \Delta y_{t-1} + \ldots + \theta_k \Delta y_{t-k} + u_t$
  where we have to
  - decide whether to include a constant α and/or linear trend β (depending on whether they are present in process y)
  - choose k
- Then, we are interested in t-statistics from significance test about coefficient in front of y_{t-1}, but with correct critical values

ADF test - critical values

- James G. MacKinnon (1991) - available as a part of an expanded version from 2010:

  James G. MacKinnon: **Critical Values for Cointegration Tests**. Queen's Economics Department Working Paper No. 1227, 2010..

  Dostupné online: http://ideas.repec.org/p/qed/wpaper/1227.html

- Values obtaines by simulations:

Table 1. Response Surface Estimates of Critical Values

N	Variant	Level	Obs.	β_∞	(s.e.)	β_1	β_2
1	no constant	1%	600	−2.5658	(0.0023)	−1.960	−10.04
		5%	600	−1.9393	(0.0008)	−0.398	
		10%	560	−1.6156	(0.0007)	−0.181	
1	no trend	1%	600	−3.4336	(0.0024)	−5.999	−29.25
		5%	600	−2.8621	(0.0011)	−2.738	−8.36
		10%	600	−2.5671	(0.0009)	−1.438	−4.48
1	with trend	1%	600	−3.9638	(0.0019)	−8.353	−47.44
		5%	600	−3.4126	(0.0012)	−4.039	−17.83
		10%	600	−3.1279	(0.0009)	−2.418	−7.58

ADF test - critical values

- If we use T data points in the regression, the critical value is $\beta_\infty + \beta_1/T + \beta_2/T^2$
- In our example from the simulations: constant without trend, $T = 200$:

N	Variant	Level	Obs.	β_∞	(s.e.)	β_1	β_2
1	no trend	1%	600	−3.4336	(0.0024)	−5.999	−29.25
		5%	600	−2.8621	(0.0011)	−2.738	−8.36
		10%	600	−2.5671	(0.0009)	−1.438	−4.48

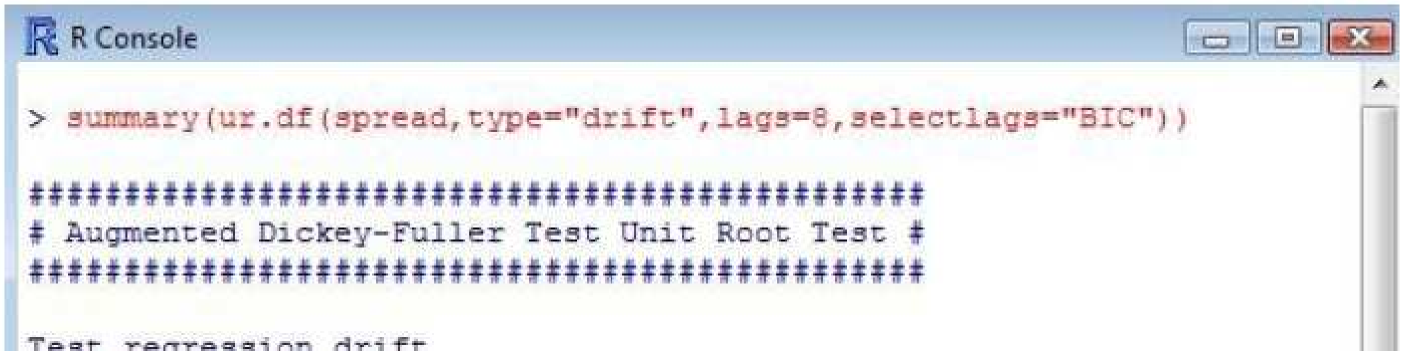
  - ◇ for 1 percent:
    $-3.4336 - 5.999/200 - 29.25/200^2 = -3.451$
  - ◇ for 5 percent:
    $-2.8621 - 2.738/200 - 8.36/200^2 = -2.879$
- Compare with t-distribution (different) and quantiles from simulations (ok)

ADF test in R

- Library urca
- Function ur.df (ur - unit root, df - Dickey-Fuller) with parameters:
  - type: possible values are drift (constant without linear trend), trend (constant and linear trend), none (nothing)
  - lags: maximal number of lags
  - selectlags: criterion for the choice of lags (information criteri: AIC, BIC)

ADF test in R

- Example: **spread** from the earlier lectures difference between long-term and short-term rates)
- In R:
  summary(ur.df(spread,type="drift",lags=8,selectlags="BIC"))
- summary in order to obtain also critical values, not only the test statistics



ADF test in R

- Výstup: estimated regression and the test statistics:

```
Call:
lm(formula = z.diff ~ z.lag.1 + 1 + z.diff.lag)

Residuals:
     Min       1Q   Median       3Q      Max
-2.62148 -0.37475 -0.01138  0.35785  2.57280

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.10035    0.05592   1.794 0.074240 .
z.lag.1     -0.10720    0.02741  -3.911 0.000125 ***
z.diff.lag   0.29007    0.06706   4.326 2.38e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7006 on 204 degrees of freedom
Multiple R-squared: 0.1216,     Adjusted R-squared: 0.113
F-statistic: 14.12 on 2 and 204 DF,  p-value: 1.806e-06


Value of test-statistic is: -3.9112 7.6595
```

ADF test in R

- Output: test statistics and critical values:

```
Value of test-statistic is: -3.9112 7.6595

Critical values for test statistics:
      1pct  5pct 10pct
tau2 -3.46 -2.88 -2.57
```

- Criterion: Hypothesis about the unit root is rejected, if the statistics is less than the critical value