$S_e^* = \sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n}}$ reziduális szórás: a megfigyelt y értékek átlagosan mennyivel térnek el a becsült $\hat{y}$ értékektől

Korrigált ... $n-k-1$ ... Ha egy ... [illegible] paraméterekkel/változóval több, akkor választhatjuk, 5%/10% cen... tél magasabb az ördög c.p.

$$COV(X,Y) = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{N}$$

$\sum(y-\hat{y})^2 \to \min$

⇒ Normálegyenletek:

$\sum y_i = N\hat{\beta}_0 + \hat{\beta}_1 \sum x_i$

$\sum x_i y_i = \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2$

↳ $\hat{\beta}_1 = \frac{\sum d_x d_y}{\sum d_x^2}$ $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Kontraszt-kódolás: a dummy-k együtthatója ne a referencia-csoporthoz, hanem az átlaghoz képesti eltérést jelentse:

	C_A	C_B
A	1	0
B	0	1
C	-1	-1

Így: $\beta_0 + \beta_{C_A} + 0 = \bar{y}_A$

$\beta_0 + 0 + \beta_{C_B} = \bar{y}_B$

$\beta_0 - \beta_{C_A} - \beta_{C_B} = \bar{y}_C$

$+ \Rightarrow \beta_0 = \frac{\bar{y}_A + \bar{y}_B + \bar{y}_C}{3}$

β_0 a főátlag (ha azonosak a csoportok elemszámai, különben súlyozni kell...)

Referencia-kódolás: az együtthatók az eltéréseket jelentik a ref. csop.-tól (ami pedig a konstansban kerül)

	D_A	D_B
A	1	0
B	0	1
C	0	0

Marginális hatás: a magy. vált. kis (egységnyi) növelésének hatására mekkora az eredményváltozó egységnyi magy. vált.-növelésre jutó hatása.

$$\overset{SST}{\sum(y_i-\bar{y})^2} = \overset{SSR}{\sum(\hat{y}_i-\bar{y})^2} + \overset{SSE}{\sum(y_i-\hat{y}_i)^2}$$

Interakció: az egyik magy. változó hatásának nagyságát befolyásolja a másik szintje

IC-k: $AIC = \frac{SSE}{n} e^{2(k+1)/n}$, $BIC = \frac{SSE}{n} n^{(k+1)/n}$

$HQC = \frac{SSE}{n} (\ln n)^{2(k+1)/n}$ Minél kisebb, annál jobb

$k = q + m$ → plusz változók száma

Parciális korr.: y és x_i kapcsolata úgy, hogy mindkét változóból kiszűrjük az összes többi magy. vált. hatását

Általánosan: $r_{yj} = \frac{-q_{yj}}{\sqrt{q_{yy} q_{jj}}}$, ahol $q_{ij} = R_{ij}^{-1}$

Háromváltozóval: $r_{y1} = \frac{r_{y1} - r_{y2} r_{12}}{\sqrt{(1-r_{y2}^2)(1-r_{12}^2)}}$

$R^2 = \frac{SSR}{SST} = 1 - \frac{1}{q_{yy}} \overset{3\text{ vált}}{=} \frac{r_{y1}^2 + r_{y2}^2 - 2 r_{y1} r_{y2} r_{12}}{1 - r_{12}^2}$

Wald F-próba:

$$F = \frac{(SSE_R - SSE_U)/m}{SSE_U/(n-k-1)} = \frac{(R_U^2 - R_R^2)/m}{(1-R_U^2)/(n-k-1)} \overset{H_0}{\sim} F(m; n-k-1)$$

Ha $m = k \Leftrightarrow$ Globális F (ANOVA), ha $m = 1 \Leftrightarrow$ parc. t-próba

LM-próba: Segédregressziót készítünk, megnézzük, hogy a plusz változókkal az eredeti (szűkebb) modell reziduumainak varianciáját 0-tól szignifikánsan különböző mértékben tudjuk-e magyarázni az új változós modellel (bővebb).

$n \cdot R^2_{\hat{e}_R | x_1, \dots, x_{k+m}} \overset{H_0}{\sim} \chi^2(m)$ vagy $(n-k-1) \cdot \frac{R_U^2 - R_R^2}{1 - R_U^2} \overset{H_0}{\sim} \chi^2(m)$

ANOVA / Globális F-próba (j. oldali)

$$F = \frac{SSR/k}{SSE/(n-k-1)} = \frac{R^2/k}{(1-R^2)/(n-k-1)} \overset{H_0}{\sim} F(k; n-k-1)$$

OLS torzítatlan: $E(\hat{\beta}) = \beta$

Norm. eloszlású: $N(\beta_j, var(\hat{\beta}_j))$

$\sqrt{var(\hat{\beta}_j)} = S_e \sqrt{[(X^T X)^{-1}]_{jj}}$

$S_{\hat{\beta}_j} = \frac{S_e}{\sqrt{\sum d_x^2}}$ (két vált)

$F(1, v) = t^2(v)$

Az ANOVA egyoldalú F próbája ekvivalens a kétoldalú t-próbával, HA 1 két változós a regr.

t-próba:

$$\frac{\hat{\beta}_j - \beta_j}{S_{\hat{\beta}_j}} = \frac{\hat{\beta}_j}{S_{\hat{\beta}_j}} \overset{H_0}{\sim} t(n-k-1)$$

readxl → read.excel

moments → skewness, kurt.

psych → describe

Hmisc → smean.cl.normal

gmodels → CrossTable (változók vizsg. kapcs.)

ggplot2 → geom_smooth (method = loess)

ppcor → pcor

corrplot → corrplot

lmtest → coeftest (model)

→ waldtest (..., test = "F" vagy "Chi")

broom → glance

stargazer → stargazer

car → linearHypothesis (model, c("A = B")) H_0: A = B

margins → sei. hatás (dydx (beres, model3, "exper"))

by(Practice[, c(2,3)], Practice$Cond, summary)

prop.table(table(Town, Cond))

sapply(Practice[, c(3,4,5)], quantile)

qt, qnorm ... ⇒ inverz

var.test, t.test(var.equal = F...) → ha nem egyezik meg

shapiro.test

aov(Price ~ Cond, data = ...) ⇒ vegyes kapcs. (előfelt.: részsokaságok norm. + csoportok azonos szórásúak) → $F = \frac{SSK/(M-1)}{SSB/(n-M)}$

cov() ⇒ kov. mtrx

cor ⇒ korr. mtrx

pf, pt ... ⇒ p-érték (F-nél: lower.tail = F)

model.fitted.values ⇒ becsült y-k ($\hat{y}$)

anova(model)[1:2, 2] ⇒ SSR, SSE

read.csv(2) ⇒ ha a 2 kell

confint(model, level = 0.95)

OLS modellfeltevései:

0) Nagy a mintánk (~~és FAE~~) vagy $Y \sim N(\mu, \sigma^2)$

1) <u>Linearitás</u>

- A sokaságot valóban leíró modell a feltételezett, azaz $Y = X\beta + \varepsilon$

2) <u>Nincs egzakt multikollinearitás</u>

- $P(\text{rang } X = k+1) = 1$ (1 valószínűséggel teljes oszloprangú)

3) <u>Erős/Szigorú exogenitás</u> → a hibák függetlenek a magy. vált.-któl

- $E(\varepsilon_i | X_i) = 0 \Rightarrow E(E(\varepsilon_i | X_i)) = E(\varepsilon_i) = 0$ és $cov(x_{ij}, \varepsilon_j) = 0$

↳ Szigorúan erősebb, mint a korrelálatlanság

Sérülésének tipikus okai:

- Van olyan változó, ami lényeges magy. vált. lenne, de mégsem szerepel a modellben, miközben legalább egy magy. vált.-val korrelál (OVB) - <u>confounding</u>: egy változó csak azért korrelál egy másikkal, mert valójában egy vagy több másik változó hatását közvetíti, a sajátján túl → korreláció ≠ kauzalitás
- Mérési hiba magy. vált.-nál (van zaj)
- Szimultaneitás (<u>több változós modelleknél</u>)

4) <u>Homoszkedaszticitás</u>

- $\forall i = 1, \dots, n$: $D^2(\varepsilon_i | X) = \sigma^2$ (állandó) - hogy éppen hányadik elemet

5) <u>Autokorrelálatlanság</u>

- A különböző megfigyelésekhez tartozó hibák korrelálatlanok egymással
- Ha a minta FAE, akkor automatikusan igaz
- Nem FAE ⇒ $cov(\varepsilon_i, \varepsilon_j | X) = 0$

4) és 5) együtt: a kovarianciamtrx. $\sigma^2 I$ alakú

→ Ha mindegyik modellfeltevés teljesül, akkor az OLS becslések <u>BLUE</u>-k:

Best (minimális varianciájú) 4) 5)

Linear (lineáris a mintaelemekben) 1)

Unbiased (torzítatlan) 1) 2) 3)

Magy. vált. hatása = közvetlen + közvetett

<u>Parciális elaszticitás</u>: az x_j változó adott értékének 1%-os változása hány %-os változással jár együtt az eredményváltozóban c.p.

$El(\hat{y}, x_j) = \frac{\beta_j x_j}{\hat{y}}$

Két vált. esetben: $cor(y, \hat{y}) = R^2$ (okt. együ...)

Lin-Log: X 1%-os változása várhatóan $\beta/100$ egységgel változtatja Y-t

Log-Log: az X-ben elő idézett 1%-os változás Y-ban β%-os változást okoz

Log-Lin: ha X értékét 1 egységgel növeljük, akkor az Y várhatóan β ... [illegible]

... hiba nagysága független attól,

Log-Lin: Ha X_j értéke 1 egységgel nagyobb, akkor e^{β_j}-szoros nagyobb lesz y c.p. várhatóan

Lin-Log: X_j 1%-os változására várhatóan $\beta_j \cdot \ln(1{,}01) \approx \frac{\beta_j}{100}$ egységgel változtatja y-t.

Log-Log: Az X_j-ben előidézett 1%-os változás, y-ban β_j százalékos változást idéz elő

Ramsey-Reset:

$$\hat{Y}_R = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_k X_k + \gamma_1 \hat{Y}^2 + \gamma_2 \hat{Y}^3 + \varepsilon$$

$H_0: \gamma_1 = \gamma_2 = 0$

F-próba vagy LM-próba

Wald-F $k = q + (m) \to$ plusz ok

$$\frac{(R_u^2 - R_R^2)/m}{(1-R_u^2)/(n-k-1)} \overset{H_0}{\sim} F(m, n-k-1)$$

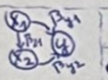
LM $n \cdot R^2_{e_R | x_1, \ldots, x_{k+m}} \sim \chi^2(m)$

t: $\hat{\beta}_j / s_{\hat{\beta}_j} \sim t(n-k-1)$

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

$AIC = \frac{SSE}{n} e^{2(k+1)/n}$ $BIC = \frac{SSE}{n} \cdot n^{(k+1)/n}$

$y = \beta_0 + \beta_1 X_1 + \varepsilon$



$y = \beta_{y0} + \beta_{y1} X_1 + \beta_{y2} X_2 + \varepsilon$

$X_2 = \beta_{20} + \beta_{21} X_1 + \varepsilon$

$\Rightarrow \beta_1 = \beta_{y1} + \beta_{21} \cdot \beta_{y2}$

telj. közvetlen közvetett

0.) Nagy minta
1.) Lineáris
2.) Nincs egz. multikoll.
3.) Erős exogenitás $(E(\varepsilon_i | X_i) = 0)$
4.) Homoszked.
5.) Autokorrelálatlanság

A cut val. változtatása a klassz.-ek hatásával a modell nem lesz jobb

Youden: specif. + szenzit. → max
Bal felső: $(1-Se)^2 + (1-Sp)^2 \to$ min

Multikoll (nem egzakt):

1.) Természetes jelenség, korrelálnak a változók
2.) Strukturális multikoll: mi visszük a modellbe a jobb függvényformáért (interakció, kvadratikus tag...)

→ A becslés torzítatlan marad $(E(\hat{\beta}_j) = \beta_j)$

→ Nem lesz hatásos (két torzítatlan becslőfv. közül az a hatásosabb, amelyiknek kisebb a varianciája, ha torzítottak, akkor $MSE = Var(\hat{\theta}) + Bs^2(\theta)$ $(Bs(\hat{\theta}) = \theta - E(\hat{\theta}))$)

→ A magy.-változók közti korrel. miatt megnő $Var(\beta_j)$, így nem hatásos ⇒ csökken a t-próba értéke ⇒ Megnő a p-érték ⇒ Nem szignifikáns olyan vált., ami a multikoll. nélkül az lenne

---

$$VIF(j) = \frac{1}{1 - R^2_{X_j | X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k}} = 33{,}46$$

⇒ A súly változó paraméterének becslési varianciája 33,46-szorosára nőtt a multikoll. miatt

$10 <$! A korrelációs mtrx. inverzének főátlójában a VIF-ek vannak!

$Tol(j) = 1 - R_j^2 = (VIF(j))^{-1} = 0{,}03$ ⇒ A súly változó összesen 3%-olyan egyedi információval rendelkezik, amelyet a többi magy. vált. nem tud magyarázni $! < \frac{1}{10}$

Főkomponenselemzés ⇒ 0.) Normáljuk az adatokat

⇒ $R = C = \frac{1}{n} X^T X = B \Lambda B^T$ (Spektrálfelbontás, B unitér C sajátvektorokkal, Λ megfelelő sajátértékekkel) ⇒ XB kov. mtrx-a az Λ ⇒ Ezek az oszlopok a főkomponensek (varianciák egyenlők a sajátértékekkel) ⇒ azt tartjuk meg, ahol $\lambda = Var \geq 1$

Nem mindig kell mindent szerepeltetni (érdemes azokat ahol a korrel. $|r| \geq 0{,}8$)

Korrel mtrx-el látjuk mi mennyit a főkompba. (sajátvektorokból(s))

Heteroszkedaszticitás ⇒ Amikor $D^2(\varepsilon | X)$ nem állandó

→ Lehetséges okok: 1.) Rossz függvényforma 2.) Csoportosított adatok (pl. háztartások) vizsgálata (méretkhatások miatt) 3.) Bővülő lehetőségek halmaza, fogyasztási szokások változása, stb.

→ köv: A becslés torzítatlan marad $(E(\hat{\beta}_j) = \beta_j)$

A paraméterbecslés varianciája és a standard hiba torzított lesz és inkonzisztens ⇒ t-próba nem követ t eloszlást, F nem követ F-eloszlást... ⇒ Nem megbízhatók a következtetések

---

White-teszt $n \cdot R^2_{\varepsilon^2 | \overbrace{X_1, \ldots, X_k}^{\text{kvadr}}, X_1^2, \ldots, X_k^2, X_{12}, \ldots, X_{ij}, \ldots, X_{k1}} \sim \chi^2(m)$

m a magyarázó változók száma (β_0 NEM!)

$H_0: R^2 = 0 \Leftrightarrow$ Homoszkedaszticitás

BP: $\frac{SSR}{} \sim \chi^2(m)$ ⇒ Feltevi, hogy a hibatag $\sim N(\mu, \sigma^2)$,

---

Heteroszked. ↓

1.) Standard hibák rosszak ⇒ újraszámolásra érdemes ⇒ White-féle, HCCME módon korrigált standard hibák ⇒ jó p-értékek!

2.) GLS ⇒ Olyan β-k, aminek stand. hibái konzisztensek $\beta = (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} y$, ahol Ω diag. alakú, a diag.-ban a White-teszt segédregressziójából becsült reziduum négyzetek vannak

---

Logisztikus regr. $y \in \{0,1\} \to P_x = P(y=1|X) \in [0,1] \to odds_x = \frac{P_x}{1-P_x} \in [0, \infty)$

$\left(P_x = \frac{odds_x}{1+odds_x}\right) \to logit_x = \ln(odds_x) \in (-\infty, \infty) = X\hat{\beta} + \varepsilon$

$$\Rightarrow P_x = \frac{e^{X\hat{\beta}}}{1+e^{X\beta}} \xrightarrow{ML} L = \prod_{i=1}^{n} P_{X_i}^{y_i} (1-P_{X_i})^{1-y_i} \Big| P_{X_i} = \frac{e^{X_i\beta}}{1+e^{X_i\beta}} \to \max_\beta$$

e^{β_i} ⇒ Ha a saját tőke aránya 1%-ponttal magasabb, akkor várhatóan 4,52%-kal alacsonyabb a csődbemenés esélye c.p. $(e^{\beta_i} = 0{,}9548)$

null deviance: az üres modell −2·loglikelihood értéke $= (n_1 \ln p_1 + n_0 \ln p_0)(-$

McFadden/Pseudo R^2: (null.deviance − deviance)/null.deviance ⇒ milyen messze van a modellünk a szaturált modelltől / nullmodelltől

	$\hat{y}=1$	$\hat{y}=0$
$y=1$	TP	FN
$y=0$	FP	TN

Helyes klasszifik. arány (accuracy): $\frac{TP+TN}{n}$

Szenzitivitás (TPR): $\left(\frac{TP+FN}{TP}\right)^{-1}$

Specificitás (TNR): $\frac{TN}{TN+FP}$

jó klasszifik. képesség: ROC görbe alatti T

auc > 0,8 (ajánlás)

Az 1-esek és 0-sok találati aránya csak egymás rovására javítható, ha javítani akarom TPR-t ⇒ alacsonyabb cut value

---

$$D^2[\hat{\beta}|X] = E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)^T|X] = E[((X^TX)^{-1}X^T(X\beta+\varepsilon)-\beta) \cdot ((X^TX)^{-1}X^T(X\beta+\varepsilon)-\beta)^T|X] = E[(\cancel{\beta}+(X^TX)^{-1}X^T\varepsilon-\cancel{\beta})(\cancel{\beta}+(X^TX)^{-1}X^T\varepsilon-\cancel{\beta})^T|X] = (X^TX)^{-1}X^TE[\varepsilon\varepsilon^T|X]X(X^TX)^{-1} \underset{\text{homo. autok.}}{=} (X^TX)^{-1}X^T\sigma^2 I X(X^TX)^{-1} = \sigma^2(X^TX)^{-1}$$

Spec (2 vált.): $D^2(\hat{\beta}_1) = \frac{\sigma^2}{SST_x}$

MLE → n elemű minta (FAE), θ paraméter

Diszkrét: $L(y_1, \ldots, y_n, \theta) = P(y_1 = k_1, \ldots, y_n = k_n | \theta)$

Folytonos: $L(y_1, \ldots, y_n, \theta) = f(y_1, \ldots, y_n | \theta)$

$f_\xi(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}}$ $\theta > 0$ $x \geq 0$ $E(\xi) = \theta$ $D^2(\xi) = \theta^2$

n elemű FAE: $L(x_1, \ldots, x_n, \theta) = \prod_{i=1}^{n} f(x_i|\theta) = \left(\frac{1}{\theta}\right)^n e^{-\frac{\sum x_i}{\theta}}$ / ln() ⇒

$\Rightarrow \ln L = -n \ln(\theta) - \sum x_i \cdot \frac{1}{\theta} \xrightarrow{\partial\theta} \partial_\theta \ln L = -\frac{n}{\theta} + \frac{\sum x_i}{\theta^2} = 0 \Rightarrow$

$\Rightarrow \hat{\theta}_{MLE} = \frac{\sum x_i}{n} = \bar{x}$

$x_i \in \{0,1\}$ $\forall i \in \{1, \ldots, n\}$-re

$f(x_i|p) = p^{x_i}(1-p)^{1-x_i} \Rightarrow L(x_1, \ldots, x_n, \theta) = p^{\sum x_i}(1-p)^{n-\sum x_i}$ / ln() ⇒

$\Rightarrow \ln L = \sum x_i \ln p + (n - \sum x_i) \ln(1-p) \xrightarrow{\partial\theta} \partial_\theta \ln L = \frac{\sum x_i}{p} - \frac{\sum(1-x_i)}{1-p} = 0 \Rightarrow$

$\Rightarrow (1-p)\sum x_i = p \sum(1-x_i) = 0 \Rightarrow \hat{p}_{MLE} = \frac{\sum x_i}{n}$

```
lmtest → resettest(model)
vcov(model) ~ β-k kov. mtrx-e
corrplot → corrplot(cor(data), method = "color")
car → vif(model) = diag(solve(cor(data)))
scale(data) ~ adatok normálása
eigen(cov/cor(scale(data)))$vectors/values
psych → describe(fokomponensek)[, c("mean", "sd")]
prcomp(data, center = T, scale = T) ~ főkomponenselemzés
summary(fokomp)        ↳ fokomponens$x[, 1:3] ~ megtartott látens
cbind(data, fokomp[, 1:2])
broom → glance(model)
skedastic → white(model, interactions = T/F)$statistic/p.value
paste0(colnames(data[, 2:15]), "-negyzet")
lmtest → bptest(model, studentize = T/F)
ks.test(data, "pnorm")
car & lmtest → coeftest(model, vcov = hccm(model)) ~ korrigált SH-k számolt p-értékek
nlme → gls(formula, data)
glm(formula, data, family = binomial(link = "logit")) ~ logit modell
summary(model)$null.deviance/deviance ~ McFadden-hez
pROC → roc(data$csod, model$fitted.values)
plot(ROC, col = "red", main = "Csőd modell ROC görbe", xlim = c(1,0),
ylim = c(0,1), type = "l")
pROC → auc(ROC)
predict(logitmodel, data) ~ logit értékek kiszámolása ⇒ type = "response"
paraméterrel P-k kiszámolása
csod$predicted > 0.5 = csod$becsles → table(csod$csod, csod$becsles)
pROC → coords(ROCcurve, "best", ret = "threshold", best.method = "closest.topleft"/"youden") ~ cut value választás
DescTools → PseudoR2(logitmodel, "McFadden")
```

MASS csomag

```
anova(model)[1/2, 2] ~
~ SSR/SSE
confint(model, level = 0.95)
margins → dydx
```

WN ≈ FAE

autokorr. ∅ (gyenge) stac

$\forall t = 0, \dots, T: E(y_t) = \mu, Var(y_t) = \sigma^2$

$\forall k > 0: ACF(k) = PACF(k) = 0$

Ha normális eloszlású a WN ⇒ korrelálatlanság = függetlenség

stacioner folyamat 0-ad rendben integrált

```
set.seed(10)
rnorm(db, μ, σ)
runif(db, a, b)
```

quantmod → getSymbols(Ticker, from, to, auto.assign = F)

```
par(mfrow = c(2,1))
acf(df)
pacf(df)
dev.off()
```

hiányzó adatok: na.approx(df)

lmtest → coeftest(model)
· Z-próba (u.i. ML becslés)

arima(df, order = c(p, i, q))

arima.sim(list(ar = _), n = 500)
(egyből ts)

forecast → Arima(df, order = c(...), x.reg = as.ma..., include.constant = ...)

tseries → jarque.bera.test(resid(model))
(H0: normális)

qqnorm(resid(model))
qqline(resid(model))

Ha Arima(.) → forecast(model, h = 10)
Ha át kell másolni (exp):
[...]$x = exp([...]$
$mean = ...
$lower = ...
$upper = ...

Próbák WN-re:

→ Durbin-Watson
· kizárólag elsőrendű autokorr. tesztelésére
· H_0: nincs elsőrendű autokorr.
· Ha a próbafv. 1,8-2,2 közé esik, akkor várhatóan ∅ elsőrendű autokorr.
· lmtest → dwtest(df~1, alternative = "two.sided")

plot(predict.forum, x.lim = c(T-10, T+10), y.lim = c(...))

In sample előrejelzés
plot(df) — statikus – valódi / dinamikus
lines(fitted(model))

Out-of-sample előrejelzés
predict(model, n.ahead = 15)
ts.plot(df, predikció, lty = c(1,3), col = c(5,2))
csak rövid távú előrejelzésre jó!
(gyorsan konvergál $E(y_t)$-hez)

→ Ljung-Box (jobban H_0 párti)
· H_0: $\rho_1 = \rho_2 = \dots = \rho_p = 0$ (p lagig autokorrelálatlan)
· H_1: $\exists j \in \{1, \dots, p\}: \rho_j \neq 0$
· Próbafv.: $$Q^* = T(T+2)\sum_{j=1}^{p} \frac{\hat\rho^2(j)}{T-j} \underset{H_0}{\sim} \chi^2(p-b)$$
(b a becsült paraméterek száma ⇒ ált. 0)
· Box.test(df, lag = 10, type = "Ljung-Box")

stabil idősor
1.) (gyengén) stacioner
2.) ergodikus
↳ sokkok hatása nem marad a rendszerben

Ha AR(3) rossz ⇒ AR(1) és AR(2) is
Ha MA(1) jó ⇒ nem romtja el a korrázható AR-ek

ACF, PACF standard hiba $1/\sqrt{T}$ és ~N

RWD pl.: $y_t = y_{t-1} + c + u_t \sim y_t = c \cdot t + u_t$ (det. trend és hasonló)

→ Breusch-Godfrey
· Segédregresszió ⇒ LM teszt (~ χ^2 eloszlás H_0 esetén)
· $y_t = \alpha + \sum_{j=1}^{p} \phi_j y_{t-j}$, H_0: $\phi_1 = \phi_2 = \dots = \phi_p = 0$
· lmtest → bgtest(df~1, order = p)

ADF: $\Delta y_t = \alpha + \delta_1 y_{t-1} + \sum_{i=1}^{p} \gamma_i y_{t-i} + u_t$
H_0: $\phi_1 = 1 \Leftrightarrow \delta_1 = 0 = \phi_1 - 1$ — lehet még trend + cikl.
RW folyamat differenciastac.
tseries → adf.test()
Másik tesztek: kpss.test H_0: stacioner
PP.test H_0: nem stac.
urca → ur.df

FinTS → ArchTest(df, lags = 12)

AR(p) folyamat → ACF lecseng, PACF p lag után letörik

$$y_t = \delta + \sum_{j=1}^{p} \phi_j y_{t-j} + u_t, \quad u_t \sim NWN(0, \sigma^2)$$

$$E(y_t) = E\left(\delta + \sum_{j=1}^{p} \phi_j y_{t-j} + u_t\right) \overset{\text{stacioner}}{=} \delta + E(y_t) \cdot \sum_{j=1}^{p} \phi_j$$

Var és ACF ⇐ Yule-Walker egyenletek:

$$\begin{cases} ACF(0) = 1 \\ ACF(k) = ACF(-k), \text{ ha } k < 0 \\ ACF(k) = \sum_{j=1}^{p} \phi_j \cdot ACF(k-j) \\ Var(y_t) = \sum_{j=1}^{p} \phi_j (ACF(j) \cdot Var(y_t)) + \sigma^2 \end{cases}$$

⇒ AR(1)-re: $ACF(k) = \phi_1^k$, $Var(y_t) = \frac{\sigma^2}{1-\phi_1^2}$

AR(2)-re: $ACF(1) = \frac{\phi_1}{1-\phi_2}$, $ACF(2) = \frac{\phi_1^2}{1-\phi_2} + \phi_2$, $ACF(k) = \phi_1 \cdot ACF(k-1) + \phi_2 \cdot ACF(k-2)$, $k \geq 2$, $Var(y_t) = \frac{1-\phi_2}{\dots} \cdot \frac{\sigma^2}{\dots}$

MA(q) folyamat → ACF q lagig tér el 0-tól, PACF lecseng

$$y_t = \alpha + \sum_{j=1}^{q} \theta_j u_{t-j} + u_t, \quad u_t \sim NWN(0, \sigma^2)$$

$E(y_t) = \alpha$, $Var(y_t) = Var\left(\alpha + \sum_j \theta_j u_{t-j} + u_t\right) = \sum \theta_j \sigma^2 + \sigma^2$

ARMA(p,q)

$$y_t = \delta + \sum_{j=1}^{p} \phi_j y_{t-j} + \sum_{j=1}^{q} \theta_j u_{t-j} + u_t$$

karakterisztikus egy. gyökei absz. értékben 1-nél < ⇔ stac.

$$E(y_t) = \frac{\delta}{1 - \sum_{j=1}^{p} \phi_j}$$

ARMA(1,1):

$$Var(y_t) = \sigma^2 \cdot \frac{1 + 2\theta_1\phi_1 + \theta_1^2}{1-\phi_1^2}$$

$$ACF(1) = \frac{(\phi_1 + \theta_1)(1 + \theta_1\phi_1)}{1 + 2\theta_1\phi_1 + \theta_1^2}$$

$ACF(k) = \phi_1 \cdot ACF(k-1)$, $k \geq 1$

stac ARMA → ha az AR rész stac. ⇒ karakterisztikus gyökei $|\lambda| < 1$
$y_t \sim \lambda^t \Rightarrow \lambda^t - \sum_{j=1}^{p} \phi_j \lambda^{t-j} = \dots$
$\Downarrow / :$
$\lambda^p - \sum_{j=1}^{p} \phi_j \dots$

weekdays(df$Date)
months(df$Date)
quarters(df$Date)
relevel(df$Quarter, ref="Q4")
cor(), ppcor → pcor(), acf(), pacf()

readxl → read_excel()

LOESS: Locally estimated scatterplot smoothing

A megfigyelés adott sávszélességű (span) környezetében minimalizálja a hibát, illeszti az egyenest, és ebből kapunk egy simított trendet a teljes idősorra.

loess(célvált. ~ t, data=df, span=0,4)

Exp trend → log-lin (értelmezésnél e^{β_1}-szoros vált.)
↳ hibatag $Y_t/\hat{Y}_t$ vagy exp(hibatag)   Trend · Szezon · Zaj

Rezid.-szórás: $s_e=\sqrt{\frac{SSE}{T}}$

STL: Seasonal and Trend Decomp. using Loess

Ha nem állandó a szezonalitás (a decompose-ból add. és mult. esetben is megmarad a szezonh. a maradékban), akkor ez még szóba jöhet, mert ez gördülő ablak technikával kezeli a szezonhatást.

stl(df(Time Series-ként), s.window=7)
↳ 7 azonos hónap alapján becsül, aztán gördül tovább
plot() DecompSTL$time.series[,"seas..."]

Additív modellben a szezonális eltérés megmutatja, hogy az adott időszak átlagosan mennyivel tér el a trendtől/főátlagtól ⇒ Kontraszt kódolás!

Kontraszt-kódolás

* contrasts(df$napok) (most)
contr.sum(7) (alapon)
kontrmtrx = contr.sum(7)
rownames(kontrmtrx) = rownames(contrasts(df$napok))
colnames(kontrmtrx) = c(felsorolás a * szerint)

	C_A	C_B
A	1	0
B	0	1
C	−1	−1

$$\left.\begin{array}{l}\beta_0+\beta_{C_A}+0=\bar{y}_A\\ \beta_0+0+\beta_{C_B}=\bar{y}_B\\ \beta_0-\beta_{C_A}-\beta_{C_B}=\bar{y}_C\end{array}\right\}\Rightarrow \beta_0=\frac{\bar{y}_A+\bar{y}_B+\bar{y}_C}{3}$$

A kimaradtat a többi összegének ($\beta_{első}+\dots$) ellentettjeként kapom.

Chow-teszt

$$\frac{SSE-(SSE^{(1)}+SSE^{(2)})}{(SSE^{(1)}+SSE^{(2)})}\cdot\frac{T-2k}{k}\overset{H_0}{\sim}F(k,T-2k)$$

H_0: nincs str. törés

Dummy-nál: $D_A=GM_A-GM_B=GM_A-\beta_0$ (B a ref.)   M = mean(dummy-k)
Kontr.-nál: $C_A=GM_A-M$, $C_B=GM_B-M$ (GM - Group Mean)

Sztochasztikusnál a véletlennek folyamatépítő szerepe van

Strukt. törésnél M = tengelymetszet (β_0)   sorszámot ad, ezt df[.]-be
a szakaszt interaktálhatva is! (az interakció azt vizsgálja, hogy eltér-e a két szakasz meredeksége)
strucchange → breakpoints(df$célváltozó ~ 1)$breakpoints[k]

Kontraszt ért.: A vizsgált időszakban a szombati napokon átlagosan 2,2%-ponttal magasabb a mobilitás, mint a szezonális átlag

A szezonindex megmutatja, hogy adott periódus értéke hány %-kal tér el az átlagtól/trendtől ⇒ Multipl. modell   trend (átlag)

Mozgóátlaggal kiszedni a trendet (− vagy /) → aggregate(Trendszűrt ~ Month, data, FUN=mean) →
→ korrigálni (osztani a mean(Trendszűrt)-el) – az átlaguk 1 (mult.-nál, add.-nál kivonni)
⇒ kiosztással/vonással kapunk szezonálisan kiigazított idősort (Trend · / + zaj)

Time Series: ts(df$célvált., start=c(2011,1), frequency=12) → havi
52 → heti
365,25 → napi
↳ plot()
decompose(TimeSeries, type="additive"/"multiplicative")
↳ plot() $trend/seasonal

$\alpha=\frac{2}{n+1}$, $n=\frac{2}{\alpha}-1$

centrált MA: forecast → ma(df$célv., order=12)
nem centrált MA: zoo → rollmean(df$célv., k=12, fill=NA)
exponenc. MA: pracma → movavg(df$célv., n=9, type="e")
↳ $\hat{Y}_t=\alpha Y_t+(1-\alpha)\hat{Y}_{t-1}$, $\hat{Y}_1=Y_1$ (α: mekkora súlyt adunk az utolsó megfigyelésnek)

Hodrick-Prescott szűrő

$$\min_{\{g_t\}}\sum_{t=1}^{T}(y_t-g_t)^2+\lambda\sum_{t=2}^{T-1}\left((g_{t+1}-g_t)-(g_t-g_{t-1})\right)^2$$

jobb illesztés   túlillesztés bünt.

mFilter → hpfilter(df$célv., freq=100, type="lambda")
y_t-g_t → ciklus

	λ
éves	100
negyedéves	1600
havi	14400

RW: $X_1=u_1$, $X_t=X_{t-1}+u_t=\sum_{i=1}^{t}u_i$, $E(X_t)=0$, $Var(X_t)=t\cdot\sigma^2$

RWD: $X_1=\delta+u_1$, $X_t=\delta+X_{t-1}+u_t$, $E(X_t)=t\cdot\delta$

ADF:
$y_t=\delta+\varphi y_{t-1}+u_t$
$\Delta y_t=\delta+(\varphi-1)y_{t-1}+u_t$
H_0: egységgyök
H_1: trendstac.

ARIMA(p,d,q)
sta ⇒ d=0 I(0)

Sarima SARIMA(p,d,q)(P,D,Q)S

- ACF kilóg S-sel => } szezonális diff => D=1
- PACF — " — => }
- Még mindig kilóg => P=1, Q=1 (PACF) (ACF)

{ P=1 => y_{t-s} ; Q=1 => u_{t-s} } ezeket veszem fel a modellbe

-> A D-vel diff.-elt idősorra ARIMA

D=1 => $y_t - y_{t-s}$ ↳ diff(df, 12)

Pl.: SARIMA(1,1,1)(0,1,0, 12)

```
• arima(df, order = c(1,1,1),
        seasonal = list(order = c(0,1,0),
        period = 12))
• lmtest → bgtest(resid(model) ~ 1, order = 13)
```

Hamis regresszió: csak látszólagos a kapcsolat, valójában mindkettő egységgyökös folyamat

=> A stacioner idősorok közötti kapcsolatot kell nézni

Hüvelykujj szabály: az eredeti regresszióban a DW teszt értéke kisebb, mint az R^2, akkor az dlt. hamis regressziót jelez

```
• lmtest → dwtest(rural%1 ~ rural%2)
```

=> Többváltozós idősormodell csak stacioner idősorokra

VAR(p) (szükségtelenül nagy p => zajos becslés)

- Több változó => amikor tartozik egyenlet az endogén; ami hatással van az end. vált.-ra, de maga inkább a modellen kívül határozódik meg, az exogén

- Egységkörön belül gyökök! ↳ stabilitás

```
• df_ts = ts(df, start = c(1969), frequency = 1)
• tseries → lapply(df_ts[, 2:3], adf.test)
  (stacioner idősorok kellenek!) kpss.test
• vars → VARselect(df_ts_stac) => milyen p-t mond?
• vars → VAR(df_ts_stac, p=2, type = "const")
```

- Hibatagok vizsgálata külön:

```
• hiba = as.data.frame(resid(varmodell))
• lapply(hiba, function(x){ bgtest(x~1, order=5)})
```

- Együttes teszt (Portmanteau) (ha H_0 igaz, akkor külön-külön is igaz)

```
• vars → serial.test(varmodell, lags.pt=5, type="PT.asymptotic")
```

- H_0: minden maradéktag (együtt is) fehérzaj

Def.: X idősor Granger okozza Y-t, ha X valamelyik lagje szignifikáns hatással van Y aktuális értékére.

Wald-F próba (ha VAR(1) van, akkor ez ekv. parc. t-próbával:

$$F(1,v) = \frac{\chi^2_1/1}{\chi^2_v/v} = \frac{z^2}{\chi^2_v/v} = \left(\frac{z}{\sqrt{\chi^2_v/v}}\right)^2 = t^2$$

```
• aod → wald.test(
    b = coef(varmodell$varresult$d_Fidesz2),
    Sigma = vcov(varmodell$varresult$d_Fidesz2),
    Terms = c(3,6))
```

def. szerint

H_0: β-k együtt 0-k <=> Fogyasztói NEM Granger-okozza a Fideszt

---

Impulzusválaszfv.: adott változóban bekövetkező szórásegységnyi sokk hogyan cseng le a többi változóban

```
• plot(irf(modell, impulse = "d_Fo",
       n.ahead = 10, ortho = FALSE))
```

(Cholesky miatt sorrendfüggő volna => GVAR)

d_Fo sokkja hogyan hat a többi változóra (alapban 95% konf. int.)

Becslési variancia dekomp (FEVD): az előrejelzendő célváltozó becslési varianciájának mekkora részét lehet megmagyarázni a modellben lévő többi változóval. Pl.: „a Fidesz becslési varianciájának kb. 7,5%-át tudjuk megmagyarázni a fogyasztói bizalmi index ismeretével az 1. előrejelzett időszakban"

```
• vars → fevd(modell, n.ahead=8)
```

↳ sorrendfüggő!

```
• frequencyConnectedness →
  → genFEVD(modell, n.ahead = 8)
```

↳ nem sorrendfüggő

Kointegráció & VECM ⇒ rövid és hosszútávú hatások is modellezhető vele

- Engle-Granger teszt (2 változóra): $x_t = x_{t-1} + u_t$; $y_t = \gamma_0 + \gamma_1 x_t + \vartheta_t$

$\Rightarrow x_t \sim I(1)$; $y_t - y_{t-1} = \cancel{\gamma_0} + \gamma_1 x_t + \vartheta_t - \cancel{\gamma_0} - \gamma_1 x_{t-1} - \vartheta_{t-1} = \gamma_1(x_t - x_{t-1}) + (\vartheta_t - \vartheta_{t-1}$

$\Rightarrow y_t \sim I(1) \Rightarrow x_t, y_t$ kointegrált, ha $\exists$ lin. komb., ami stac idősort ad

(teszt: kointegráló regr. maradéktagja stacioner-e?)

- $\Delta X_t = C + B X_{t-1} + \sum_{i=1}^{k-1} A_i \Delta X_{t-i} + u_t$ → Hiba korrekció = kointegráló regresszió maradéktagja
↳ Hamis regressziónál NEM

- Nagy késleltetést javasol a VARselect ⇒ Valszeg. klaszterezödik a variancia
- tsDyn → VECM(df, lag=1, r=1, include = "const", estim = "ML")
Nem diff.elt! ↳ VAR rész ↳ rank(B) ⇒ Johansen teszt
- Johansen teszt: $r := rank(B)$ ⇝ sajátértékekkel (nem közvetlenül B sajátértékei...)

$H_0: r=0, H_1: r>0$ → Jobboldali
$H_0: r\leq 1, H_1: r>0$ → k db idősor esetén $r \leq k-1$ (különben $\exists B^{-1}$...)
⋮ → $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{k-1} \geq 0$ sajátértékek χ^2 eloszlás
$H_0: r \leq k-1, H_1: r > k-1$

- urca → ca.jo(df, type = "eigen"/"trace", ecdet = "const", spec = "longrun")
↳ summary(.) ↳ több dimenziós ts ⇝ ts(df[,2:3], start = c(1990,01,01), frequency=
- ~~frequency~~Connectedness → genFEVD(vec2var(koiteszt, r=5))
↳ ca.jo(...) ↳ VECM-ben r (rank(B))
- irf ugyanígy

ARMA stac ⇔ kar polinom gyökei $|\cdot|<1$, késleltet. operátor $|\cdot|>1$

ARIMAX esetén
- forecast → Arima(BJ$Y, order = c(2,1,0), xreg = as.matrix(BJ[,4:6]), include.constant=T

Hiányzó adat lineáris interpolációval: • euro = na.approx(euro)
Előrejelzés forecast csomagból: • pr = forecast(model, h=10)
Átszámolás, ha logaritmáltuk: • pr$x = exp(pr$x), pr$mean/lower/upper = exp(pr$mean/lower/upper)
Ábra: • Time = lenght(ferran) → plot(pr, xlim = c(Time-100, Time+10), ylim = c(420,540))
statikus: csak valós adatokkal
dinamikus: becsült értékekkel (is) késl. operátor gyökei: abs(polyroot(c(1,-model$model$phi))

VECM 2 változóval értelmezés: ECT ⇝ azonnal 1 időszak alatt az egyensúlyi elmozdulásból hány egységet dolgoz le az ATS (ATS célváltozó)
⇝ Mennyire gyorsan igazodik a DEM-hez ⇒ -0,69 ⇝ 0,69 egységet azonnal ledolgoz
⇒ 1/0,69 ≈ 1,45 ~ 1,5 nap alatt ledolgozza az ATS az egyensúlyi elmozdulást
⇒ Csak 2 változóra értelmezzük így!