1) Box - transform

Box - Cox test

Zarembki transformation

① $F = \frac{y^{\lambda} - 1}{\lambda} \Rightarrow \begin{cases} y - 1 & \lambda = 1 \\ \to \log y & \lambda = 0 \end{cases}$

$y^* = \frac{y^{\lambda_1} - 1}{\lambda_1} \qquad x^* = \frac{x^{\lambda_2} - 1}{\lambda_2}$

lin vs log-linear

$y^* = \beta_0 + \beta_1 x^* + u$

1) estimate $\hat{\beta}_0$ $\hat{\beta}_1$ using OLS

2) $\hat{\lambda}_1$ and $\hat{\lambda}_2$ est. using NLS

t-test: $\lambda_1 = 1$, $\lambda_2 = 1$ linear

$\lambda_1 = 1$ $\lambda_2 = 0$ lin-log mod

. . . .

linear vs log-linear :

Zarembki Test.

1) Transformation using geom. means

$\bar{\bar{y}} = \sqrt[n]{\prod_{i=1} y_i} = \left(e^{\log \prod_{i=1} y_i}\right)^{1/n} =$

$e^{1/n \cdot \sum \log y_i} = e^{\overline{\log y_i}}$

$$Y_i^* = \frac{Y_i}{\bar{Y}}$$

H_0: no dif.

H_a: is a sig. dif

2) Est. $Y_i^* \mid X_i \to RSS_1$

$\log Y_i^* \mid X_i \to RSS_2$

$$\frac{n}{2} \cdot \left| \log \frac{RSS_1}{RSS_2} \right| \sim \chi_1^2$$

Problem 1. (UoL Exam). The rise in prices for public transport leads to lower corporate earnings, as people tend to choose cheaper alternatives. The student tries to find the best form of dependence of the volume of transportation T_i of some 50 transportation companies (in millions of dollars) from the prices of transportation P_i (in cents per one kilometer of transportation). She runs regressions (1-4) (linear, logarithmic and semi-logarithmic functions), she also runs two auxiliary regressions (5-6) performing Zarembka transformation (variable TZ_i is defined as $TZ_i = T_i / \sqrt[n]{T_1 \cdot T_2 \cdot \ldots \cdot T_n}$):

$P \uparrow 1\%$ $T \uparrow \beta/100$ $P \uparrow 1$ $T \uparrow \beta \cdot 100\%$

$Y = \beta_0 + \beta_1 \ln X_i$

$\frac{dY_i}{d \ln X_i} = \beta_1$

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable	T_i	T_i	$\log(T_i)$	$\log(T_i)$	TZ_i	TZ_i
Independent variable\Constant	8.74	12.26	2.175	2.635	1.171	1.641
P_i	-0.339	-	-0.0045		-0.0045	
$\log(P_i)$	-	-1.362	-	-0.179	-	-0.179
R^2	0.638	0.738	0.665	0.755	0.638	0.738
RSS	4.481	3.247	0.068	0.051	0.080	0.058

Y^* $\ln X$

$\ln Y$ $\ln X$

(a) Explain the differences in the values of a slope coefficient in regression (1) and (4) giving interpretation to both regressions. ✓

(b) Explain the differences in the values of a slope coefficient in regression (2) and (3) giving interpretation to both regressions. ✓

(c) Explain using some math why your interpretation of regression (4) is correct using different methods. Do the same for regressions 2-3.

(d) Which pairs of regression are comparable directly without Zarembka transformation). Which regressions becomes comparable after Zarembka transformation? Compare some regressions performing appropriate tests.

c) Algebraic approach

$$\log Y = a + b \cdot \log X \qquad \log T = 2.6 - 0.2 \cdot \log P$$

$$X \uparrow 1\% \qquad X\left(1 + \frac{1}{100}\right)$$

$$\log X + \log\left(1 + \frac{1}{100}\right)$$

$$\ln(1+\alpha) \approx \alpha$$

$$\approx$$

$$\log X + \frac{1}{100}$$

$$\log(Y + \Delta Y) = a + b\left(\log X + \frac{1}{100}\right)$$

$$\log(Y + \Delta Y) - \log Y = \frac{b}{100}$$

Problem 1. (UoL Exam). The rise in prices for public transport leads to lower corporate earnings, as people tend to choose cheaper alternatives. The student tries to find the best form of dependence of the volume of transportation T_i of some 50 transportation companies (in millions of dollars) from the prices of transportation P_i (in cents per one kilometer of transportation). She runs regressions (1-4) (linear, logarithmic and semi-logarithmic functions), she also runs two auxiliary regressions (5-6) performing Zarembka transformation (variable TZ_i is defined as $TZ_i = T_i / \sqrt[n]{T_1 \cdot T_2 \cdot \ldots \cdot T_n}$):

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable	T_i	T_i	$\log(T_i)$	$\log(T_i)$	TZ_i	TZ_i
Independent variable\Constant	8.74	12.26	2.175	2.635	1.171	1.641
P_i	-0.339	-	-0.0045		-0.0045	
$\log(P_i)$	-	-1.362	-	-0.179	-	-0.179
R^2	0.638	0.738	0.665	0.755	0.638	0.738
RSS	4.481	3.247	0.068	0.051	0.080	0.058

(a) Explain the differences in the values of a slope coefficient in regression (1) and (4) giving interpretation to both regressions.

(b) Explain the differences in the values of a slope coefficient in regression (2) and (3) giving interpretation to both regressions.

(c) Explain using some math why your interpretation of regression (4) is correct using different methods. Do the same for regressions 2-3.

(d) Which pairs of regression are comparable directly without Zarembka transformation). Which regressions becomes comparable after Zarembka transformation? Compare some regressions performing appropriate tests.

Model (1) & (3) log-linear vs linear mod

$$\chi^2 = \frac{50}{2} \cdot \left| \log \frac{0{,}08}{0{,}068} \right| = 4{,}063$$

$$\chi^2_{1,\,0{,}95} = 3{,}84$$

$\Rightarrow$ Ha is sig. dif $\Rightarrow$ model (3) log-lin

Model (2) & (4) lin-log vs log-log

$$\chi^2 = \frac{50}{2} \cdot \left| \log \frac{0{,}058}{0{,}051} \right| = 3{,}2$$

$\Rightarrow$ H₀ is not rej $\Rightarrow$ no sig. dif between $\Rightarrow$ choose both

Question 7. (ICEF Exam)

An employee of a real estate agency in a Russian city with a developed subway network is interested in estimating of the influence of the distance from the city center $CENTER_i$ (in kilometers) on the price of an two-room apartment in millions of rubles. Based on the data of 21 apartments sold during a period under consideration she runs a regression.

$$\underset{(0.88)}{\widehat{PRICE}_i = 12.39} - \underset{(0.10)}{0.20} \cdot CENTER_i \qquad R^2 = 0.17 \quad RSS = 103.4 \qquad \textbf{(1)}$$

t-test

(a) □ Is the regression coefficient significant (take into account that the realtor did not know exactly the sign of its coefficient before the regression calculation)?

$H_a: \beta \neq 0$

$H_0: \beta \geq 0$

$H_a: \beta < 0$

□ Are the results of the estimation compatible with the hypothesis that true regression coefficient is positive?

□ Are the results of the estimation compatible with the hypothesis that true regression coefficient is 0.1?

□ How the conclusion on significance of the slope would change if the manager could use the assumption that the influence of the $CENTER_i$ on the apartment price is not positive?

$t_{19}^{97.5\%} = 2{,}11$

□ Is intercept of the equation significant? Summarize all information on the test results and discuss economic meaning of the equation (1).

The realtor, not satisfied with the obtained result, decided to take into account the additional factor – the distance to the nearest subway station $METRO_i$ (also in kilometers).

$$\widehat{PRICE}_i = \underset{(0.97)}{13.71} - \underset{(0.09)}{0.22} \cdot CENTER_i - \underset{(0.25)}{0.58} \cdot METRO_i \qquad R^2 = 0.37 \quad RSS = 79.29 \qquad \textbf{(2)}$$

During the discussion at the workshop, the realtor received advice from a colleague to use Ramsey's test for this equation. Since the realtor was not experienced enough in econometrics, a colleague helped her calculate appropriate equation (using in the right side of (3) estimated values $\widehat{PRICE}_i^*$ from equation (2):

$$\widehat{PRICE}_i = \underset{(6.04)}{0.023} + \underset{(0.18)}{0.13} \cdot CENTER_i + \underset{(0.47)}{0.35} \cdot METRO_i + \underset{(0.033)}{0.07} \cdot (\widehat{PRICE}_i^*)^2 \qquad R^2 = 0.51 \quad RSS = 60.64 \qquad \textbf{(3)}$$

Then the colleague helped her to estimate a new equation

$$\widehat{\log PRICE}_i = \underset{(0.10)}{2.62} - \underset{(0.0095)}{0.019} \cdot CENTER_i - \underset{(0.026)}{0.059} \cdot METRO_i \qquad R^2 = 0.32 \quad RSS = 0.8448 \qquad \textbf{(4)}$$

and did Ramsey's test again (using in the right side of (5) estimated values $\widehat{\log PRICE}_i^{**}$ from equation (4):

$$\widehat{\log PRICE}_i = \underset{(1.53)}{0.62} + \underset{(0.039)}{0.030} \cdot CENTER_i + \underset{(0.11)}{0.084} \cdot METRO_i + \underset{(0.0088)}{0.012} \cdot (\widehat{\log PRICE}_i^{**})^2 \qquad R^2 = 0.39 \quad RSS = 0.7672 \qquad \textbf{(5)}$$

(b) □ Help the realtor to understand the logic of her colleague in estimating these equations.

□ Explain what the Ramsey test is, what is the null hypothesis and what statistics it uses; use them to perform the necessary calculations.

She estimated non-linear regression (4) using logarithm of dependent variable

$$\widehat{\log PRICE}_i = \underset{(0.10)}{2.62} - \underset{(0.0095)}{0.019} \cdot CENTER_i - \underset{(0.026)}{0.059} \cdot METRO_i \qquad R^2 = 0.32 \quad RSS = 0.8448 \qquad (4)$$

and evaluates Ramsey test again

$$\widehat{\log PRICE}_i = \underset{(1.53)}{0.62} + \underset{(0.039)}{0.030} \cdot CENTER_i + \underset{(0.11)}{0.084} \cdot METRO_i + \underset{(0.0088)}{0.012} \cdot (\widehat{\log PRICE}_i^{**})^2 \qquad R^2 = 0.39 \quad RSS = 0.7672 \qquad (5)$$

□ What conclusions can be drawn from the results in this part of the study?

Reset Test (Ramsey test)

↳ functional form test

$H_0: \; y_i = \beta_0 + \beta_1 X_{1i} + \dots + \beta_k \cdot X_{ki} + \varepsilon_i \qquad (*)$

$H_a:$ functional $(*)$ is wrong

(omitted var. or quadr. should be added)

1) $y_i \mid X_1 \dots X_k \Rightarrow \hat{y}_i$

2) $y_i = \beta_1 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} + \gamma_1 \hat{y}_i^2 + \dots + \gamma_p \hat{y}_i^{p+1} + \varepsilon_i$

$p = 1$ or $p = 2$

Ramsey test: F-stat. $\gamma_1 = \dots = \gamma_p = 0$

$$F \sim F_{p,\, n-k-p}$$

The realtor, not satisfied with the obtained result, decided to take into account the additional factor – the distance to the nearest subway station $METRO_i$ (also in kilometers).

$$\hat{PRICE}_i = \underset{(0.97)}{13.71} - \underset{(0.09)}{0.22}\cdot CENTER_i - \underset{(0.25)}{0.58}\cdot METRO_i \qquad R^2 = 0.37 \quad RSS = 79.29 \qquad \textbf{(2)}$$

During the discussion at the workshop, the realtor received advice from a colleague to use Ramsey's test for this equation. Since the realtor was not experienced enough in econometrics, a colleague helped her calculate appropriate equation (using in the right side of (3) estimated values $PRICE_i^*$ from equation (2):

$$\hat{PRICE}_i = \underset{(6.04)}{0.023} + \underset{(0.18)}{0.13}\cdot CENTER_i + \underset{(0.47)}{0.35}\cdot METRO_i + \underset{(0.033)}{0.07}\cdot(\hat{PRICE}_i^*)^2 \qquad R^2 = 0.51 \quad RSS = 60.64 \qquad \textbf{(3)}$$

Then the colleague helped her to estimate a new equation

$$\log \hat{PRICE}_i = \underset{(0.10)}{2.62} - \underset{(0.0095)}{0.019}\cdot CENTER_i - \underset{(0.026)}{0.059}\cdot METRO_i \qquad R^2 = 0.32 \quad RSS = 0.8448 \qquad \textbf{(4)}$$

and did Ramsey's test again (using in the right side of (5) estimated values $\log PRICE_i^{**}$ from equation (4):

$$\log PRICE_i = \underset{(1.53)}{0.62} + \underset{(0.039)}{0.030}\cdot CENTER_i + \underset{(0.11)}{0.084}\cdot METRO_i + \underset{(0.0088)}{0.012}\cdot(\log \hat{PRICE}_i^{*})^2 \qquad R^2 = 0.39 \quad RSS = 0.7672 \qquad \textbf{(5)}$$

(b) □ Help the realtor to understand the logic of her colleague in estimating these equations.

□ Explain what the Ramsey test is, what is the null hypothesis and what statistics it uses; use them to perform the necessary calculations.

She estimated non-linear regression (4) using logarithm of dependent variable

$$\log PRICE_i = \underset{(0.10)}{2.62} - \underset{(0.0095)}{0.019}\cdot CENTER_i - \underset{(0.026)}{0.059}\cdot METRO_i \qquad R^2 = 0.32 \quad RSS = 0.8448 \qquad (4)$$

and evaluates Ramsey test again

$$\log PRICE_i = \underset{(1.53)}{0.62} + \underset{(0.039)}{0.030}\cdot CENTER_i + \underset{(0.11)}{0.084}\cdot METRO_i + \underset{(0.0088)}{0.012}\cdot(\log PRICE_i^{**})^2 \qquad R^2 = 0.39 \quad RSS = 0.7672 \qquad (5)$$

□ What conclusions can be drawn from the results in this part of the study?

t-test $t = \frac{0,07}{0,03} = 2,12$

$t_{18}^{97.5} = 2,1$

$\Rightarrow$ some non-lin. is present

Question 1. (17 marks)

Two students developed different models for their course papers: student (a) was convinced that the coefficient of variable Z_i is equal to 1, while student (b) believed that this coefficient is the opposite of the coefficient for the variable X_i

$$Y_i = \beta_1 + \beta_2 X_i + Z_i + u_i \qquad \textbf{(a)}$$

$$Y_i = \beta_1 + \beta_2 X_i - \beta_2 Z_i + v_i \qquad \textbf{(b)}$$

They think that in fact both models will give the same estimates of β_2, but both faced unexpected difficulties when trying to evaluate these models (what difficulties?). So they decided instead to calculate sample variances and covariances on the base of the same data (n observations): Var(Y) =4, Var(X) =3, Var(Z) =5, Cov(Y,X) =6, Cov(Y,Z) =1, Cov(X,Z) =2.

(a) □ Help the students to find the least squares estimates of β_2 for their models, indicating all necessary steps.

□ Are these estimates really the same as students think?

$\beta_3 = 1 \qquad \beta_3 = -\beta_2$

1) (a) $\underbrace{Y_i - Z_i}_{Y^*} = \beta_1 + \beta_2 X_i + u_i \Rightarrow \hat{\beta}_2 = \frac{Cov(X_i, Y^*)}{Var(X_i)}$

RSS_R

(b) $Y_i = \beta_1 + \beta_2 \underbrace{(X_i - Z_i)}_{X^*} + v_i$

(a) $\hat{\beta}_2 = \frac{Cov(X, Y-Z)}{Var(X)} = \frac{cov(X,Y) - cov(X,Z)}{Var(X)} =$

$(6-2)/3 = 4/3$

(b) $\hat{\beta}_2 = \frac{Cov(X-Z, Y)}{Var(X-Z)} = \frac{6-1}{3+5-2\cdot 2} = \frac{5}{4}$

(b) The scientific advisor told the students that both their models are restricted versions of the more general model

UR: $Y_i = \beta_1 + \beta_2 X_i + \beta_3 Z_i + w_i$ **(c).** R: $Y_i = \beta_1 + \beta_2 X_i + Z_i + u_i$ **(a)** R: $Y_i = \beta_1 + \beta_2 X_i - \beta_2 Z_i + v_i$ **(b)**

RSS_{UR}

□ What are restrictions in each case?

□ How to test the restriction for the model (a)? Indicate necessary steps.

□ What model should be chosen if both restrictions are invalid? What model (a), (b) or (c) should be chosen if only one restiction is valid?

UR model

□ Let both restrictions be valid. Which model out of (a) and (b) is preferable and why? Use numerical data above to choose between (a) and (b).

- (a) $\beta_3 = 1$ (b) $\beta_3 = -\beta_2$
- t - test $t = \frac{\hat{\beta}_3 - 1}{se(\hat{\beta}_3)}$

F - test $\quad F = \dfrac{(RSS_R - RSS_{UR})/1}{RSS_{UR}/n-3}$

(a) $\sigma^2_{\hat{\beta}_2} = \dfrac{\sigma^2_u}{n \cdot Var(X)}$

(b) $\sigma^2_{\hat{\beta}_2} = \dfrac{\sigma^2_v}{n \cdot Var(X^*)} = \dfrac{\sigma^2_v}{n \cdot Var(X-Z)}$

$Var(X) = 3$

$Var(X-Z) = 3 + 5 - 2 \cdot 2 = 4$

$\Rightarrow$ given that $\sigma^2_u = \sigma^2_v$ we prefer (b)

(2)

- $y_i = \beta \cdot x_i + u_i$

$$\boxed{TSS \neq ESS + RSS}$$

$\Rightarrow$ $R^2_{uc} = 1 - \dfrac{\sum \hat{u}_i^2}{\sum y_i^2}$

without const

with const $R^2 = 1 - \dfrac{\sum \hat{u}_i^2}{\underbrace{\sum (y_i - \bar{y})^2}_{TSS}}$

$k \uparrow \Rightarrow R^2 \uparrow$

- $R^2_{adj} \downarrow$ if new var. is insignif.

- $R^2_{adj} \uparrow$ if t-stat for new var. is greater than 1

- $R^2_{adj} \in (-\infty; 1)$

Question 5. [25 marks] A lazy student found an interesting paper with econometric models describing how $COVID$ - aggregate anti COVID expendidtures, depend on GDP, aggregate gross national product, and POP, total population, for a sample of 70 countries in the second quarter of 2020. $COVID$ and GDP are both measured in US$ billion. POP is measured in million. RSS – Residual Sum of Squares. He decided to use this article in his course paper pretending that he got all equations himself using original data (what in fact was not true). He wrote the equations on a paper:

$$\log\widehat{\frac{COVID}{POP}} = -3.74 + 1.27\log\frac{GDP}{POP} \qquad R^2 = 0.90 \quad RSS = 15.45 \tag{1}$$

$$\log\widehat{COVID} = -3.60 + 1.27\log GDP - 0.33\log POP \qquad R^2 = 0.95 \quad RSS = 13.90 \tag{2}$$

$$\log\widehat{\frac{COVID}{POP}} = -3.60 + 1.27\log\frac{GDP}{POP} - 0.06\log POP \qquad R^2 = 0.91 \quad RSS = 13.90 \tag{3}$$

But when he wrote his coursework some details seemed a little strange to him and he began to doubt that he had correctly rewritten the equations on paper. He asked your advice.

$\beta_1 + \beta_2 - 1 = \tilde{\beta}_2$

(a) □ The student now believes that by mistake he repeated the same coefficient 1.27 in equations (1), (2) and (3), as well as he repeated the intercept -3.60 in equations (2) and (3), but he does not remember the correct values. Are these coincidences really happened by mistake?

(1) $\log \dfrac{Cov}{Pop} = \beta_0 + \beta_1 \cdot \log \dfrac{GDP}{pop}$

$\log Cov = \beta_0 + \beta_1 \log GDP + (1-\beta_1)\log Pop$

(2) $\log Cov = \beta_0 + \beta_1 \cdot \log GDP + \beta_2 \cdot \log PoP$

(3) $\log Cov = \beta_0 + \beta_1 \cdot \log GDP + (1-\beta_1+\beta_2)\log Pop$

□ It seems strange to him that both coefficietns of the variable $\log POP$ in equations (2) and (3) are negative but different in absolute value. Help the student to understand these.

(2) β_2

$\tilde{\beta}_2 = 1 - \beta_1 + \beta_2$

$\beta_2 = \tilde{\beta}_2 + \beta_1 - 1$