Elements of Econometrics. 2022-2023.
Class 6. Variables Transformations

Box-Cox (Zarembka)

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 $TZ_i = T_i / \sqrt[n]{T_1 \cdot T_2 \cdot ... \cdot T_n}$):

	(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.

Question 2. (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.

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

(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)?

□ 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?

□ 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 $\cdot \hat{PRICE}_i^*$ from equation (2):

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

Then the colleague helped her to estimate a new equation

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

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

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

(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

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

and evaluates Ramsey test again

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

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