
Course: Basics of R programming language for statistical analysis

Instructor: Marina FERENT [marinaferent@gmail.com]

Meeting 8

You are provided with **a dataset <<Wages.csv[i]>>** which contains data on about 400 employees in Thailand.

The dataset contains **data on**:

- *Gender*= 0 for male and 1 for female
- *Educ* = number of years in education
- *Jobcat* = 1 for worker; 2 for admin; 3 for manager
- *Salary* measured in Thai baht
- *Salbegin* = beginning salary in Thai baht
- *Jobtime* = years on current job
- *Prevexp* = previous experience in years
- *Polytech* = 1 if a politech

!!! Make sure to set Format cells to Numeric for all the variables. Otherwise you will receive the following error when trying to upload the .csv into R: x must be numeric (or similar).

Compulsory tasks [20 min]:

1. Run the regression (linear model): log(salary)=f(log(salbegin), educ, jobcat, gender, jobtime, prevexp, polytech). Name it regression.
2. Check **one of the** assumptions of the linear model:
    - assumption 1: **linearity of the model** – do a specification test – Ramsey Reset Test
    - assumption 4: perfect or near **multicollinearity** should not exist – check correlation matrix
    - assumption 5:
   - **Homoskedasticity** of errors may be accepted
   - **Heteroskedasticity** has further implications, affecting the inference (standard errors, p-values, etc.) not the consistency

     Tests: BP test, or White test
    - assumption 6: check the **normality** of the residuals – *Tests: Jarque-Bera, Shapiro-Wilk, Kolmogorv-Smirnof, Cramer-von Mises, or Anderson-Darling*

Additional tasks – one of the [25 min]:

1. Plot ln(salary) vs. ln(salbegin). Add the regression line.
2. Using a for loop, plot log(salary) against all the quantitative independent variables and store all the scatter plots into a single .pdf file.
3. Store the correlation matrix into a .csv file on your computer.
4. Using a conditional statement of your choice interpret the result of your test in point 2 compulsory. [HINT: If you type in attributes(test_name(regression)) you will receive the labels under which are stored the p-values. If p-values are stored as p.value you can retrieve it as test_name(regression)$p.value]
5. Using paste0 function, interpret the R-squared.
6. Store the regression results into a .csv file on your computer.
7. Using the paste0 function interpret the coefficient of prevexp.
8. Using the paste0 function interpret the coefficient of ln(salbegin)
9. Using a conditional statement of your choice:
   - interpret the coefficient of prevexp, if statistically significant
   - write “not statistically significant”, if prevexp not statistically significant

Basics of econometrics cheat sheet:

1. Ramsey Reset Test interpretation:
   H0: The model is correctly specified, i.e. the linear model is adequate.
   H1: The model is mis-specified, i.e. there are neglected nonlinearities in the model.
2. Breusch-Pagan or White Test interpretation:
   H0: The residuals are homoscedastic.
   H1: The residuals are heteroskedastic.
3. Normality tests:
   Test SW, CM, AD, JB:
   H0: The residuals are normally distributed
   H1: The residuals are not normally distributed
   Test KS:
   H0: The residuals are not normally distributed
   H1: The residuals are normally distributed
4. The p-value rule:

p-value (Pr>|t|)	Significance level (stars in “Business and Economics system”)
<0.01	1% = ***
>0.01, but <0.05	5% = **
>0.05, but <0.1	10% =*

5. Coefficient interpretation:

Type of Relationship	Dependent	Independent	Interpretation
level — level	Y	X	**1 unit increase in X** results in $\boldsymbol{\beta}$ **units increase in Y** on average, *ceteris paribus.*
log — log	$\ln Y$	$\ln X$	**1% increase in X** results **in** $\boldsymbol{\beta}\%$ **increase in Y** on average, *ceteris paribus.*
log — level	$\ln X$	X	**1 unit increase in X** results in $\boldsymbol{\beta \cdot 100}\%$ **increase in Y** on average, *ceteris paribus.*
level — log	Y	$\ln X$	**1% increase in X** results **in** $\boldsymbol{\beta/100}$ **units increase in Y** on average, *ceteris paribus.*
Quadratic	Y	$X + X^2$	**1 unit increase in X** results in $\boldsymbol{\beta_1 + 2 \cdot \beta_2}$ **units increase in Y** on average, *ceteris paribus.*
Quadratic	$\ln Y$	$X + X^2$	**1 unit increase in X** results in $(\boldsymbol{\beta_1 + 2 \cdot \beta_2}) \cdot \boldsymbol{100}\%$ **increase in Y** on average, *ceteris paribus.*

Level-dummy	Y	We expect $\boldsymbol{\beta}$ **units increase in Y** for *dummy=1* with reference to *dummy=0* on average, *ceteris paribus.*
Log-dummy	$\ln X$	We expect $\boldsymbol{\beta \cdot 100}\%$ **increase in Y** for *dummy=1* with reference to *dummy=0* on average, *ceteris paribus.*

---

[i] The data set is a slightly altered version of *engin* data from *Wooldridge, Jeffrey M. (2013). Introductory econometrics: a modern approach. Mason, Ohio: South-Western Cengage Learning*. *Wooldridge Source: Thada Chaisawangwong, a former graduate student at MSU, obtained these data for a term project in applied econometrics. They come from the Material Requirement Planning Survey carried out in Thailand during 1998*.
The original data set is available for download at:
(1)https://www.cengage.com/cgi-wadsworth/course_products_wp.pl?fid=M20b&product_isbn_issn=9781111531041 Or
(2) https://cran.r-project.org/web/packages/wooldridge/wooldridge.pdf
Current data set changed the definition of the gender variable and created the variable job category for instructional purposes (Manager=those employees that have higher than Q3 total experience and salary; Admin = those employees that have higher than Q2 salaries or total experience; Worker=the rest). I also dropped some of the variables of the original data set.