LECTURE #7

Econometrics I

DATA SCALING & FUNCTIONAL FORMS SELECTION OF EXPLANATORY VARIABLES

Jiri Kukacka, Ph.D.
Institute of Economic Studies, Faculty of Social Sciences, Charles University

Summer semester 2024, April 9

Midterm summary

- Aggregate results
- Midterm summary statistics
- Quick midterm review: TOP mistakes

In the previous lecture #6

- We introduced testing multiple linear restrictions using F test:

$$F = \frac{(SSR_R - SSR_U)/q}{SSR_U/(n-k-1)} = \frac{(R_U^2 - R_R^2)/q}{(1-R_U^2)/(n-k-1)} \sim F_{q,n-k-1}.$$

- Under MLR.1 through MLR.4, the **OLS estimator** $\hat{\beta}_j$ **is consistent** for all $j = 0, \ldots, k$.
- Even **MLR.4′ Zero mean and zero correlation:** $\mathbb{E}(u) = 0$ and $\text{Cov}(x_j, u) = 0$ for $j = 1, 2, \ldots, k$, is sufficient.
- Under MLR.1 through MLR.5, $\hat{\beta}_j$ is **asymptotically normally distributed** and **asymptotically efficient**.
- We discussed Lagrange multiplier tests: $LM = nR_u^2 \sim \chi_q^2$.
- Readings for lecture #7:
  - Chapter 6: 6.1–6.3

Outline

Effects of data scaling on OLS statistics

More on functional form

- More on logarithmic forms
- Models with quadratics
- Models with interaction terms

More on goodness-of-fit

Selection of explanatory variables

Outline

Effects of data scaling on OLS statistics

More on functional form
- More on logarithmic forms
- Models with quadratics
- Models with interaction terms

More on goodness-of-fit

Selection of explanatory variables

Effects of data scaling on OLS statistics

Why do we rescale the data?

Mainly for reporting purposes:

- Too many decimals
- Different units of measurement

Effects of scaling (independent variable)

Instead of a standard population model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u,$$

consider a model ($c \neq 0$ being a constant)

$$y = \gamma_0 + \gamma_1 (c x_1) + \gamma_2 x_2 + u.$$

Assuming MLR.1-MLR.6 hold, what is the relationship between:

- $\hat{\beta}_1$ and $\hat{\gamma}_1$, $\hat{\beta}_2$ and $\hat{\gamma}_2$?
- $se(\hat{\beta}_1)$ and $se(\hat{\gamma}_1)$?
- Relevant t statistics?
- Relevant p-values?
- Hint: use the definition $se(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sqrt{SST_j(1-R_j^2)}}$.

Effects of scaling (independent variable)

- Estimate:

$$\hat{\beta}_1 = \hat{\gamma}_1 c \quad \longrightarrow \quad \boxed{\hat{\gamma}_1 = \frac{\hat{\beta}_1}{c}} \quad \longrightarrow \quad \hat{y} = \hat{\beta}_0 + \boxed{\frac{\hat{\beta}_1}{c}} c x_1 + \hat{\beta}_2 x_2$$

- Standard error:
  - $SST_{x_1} = \sum (x_{1i} - \bar{x}_1)^2$
  - $SST_{cx_1} = \sum (cx_{1i} - c\bar{x}_1)^2 = c^2 \sum (x_{1i} - \bar{x}_1)^2 = c^2 SST_{x_1}$
  - $\boxed{se(\hat{\gamma}_1)} = \frac{\hat{\sigma}}{\sqrt{SST_{cx_1}(1-R^2_{x1})}} = \frac{\hat{\sigma}}{\sqrt{c^2 SST_{x_1}(1-R^2_{x1})}} =$
    $\frac{1}{c}\frac{\hat{\sigma}}{\sqrt{SST_{x_1}(1-R^2_{x1})}} = \boxed{\frac{se(\hat{\beta}_1)}{c}}$
- t statistic: $\boxed{\hat{t}_{\hat{\gamma}_1}} = \frac{\hat{\gamma}_1}{se(\hat{\gamma}_1)} = \frac{\hat{\beta}_1/c}{se(\hat{\beta}_1)/c} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \boxed{\hat{t}_{\hat{\beta}_1}}$
- p-value is based on the t statistic, i.e., it does not change either.

Effects of scaling (dependent variable and logs)

What happens when:

- Dependent variable is rescaled?
- Dependent variable in the logarithmic form is rescaled?
- Independent variable in the logarithmic form is rescaled?

Effects of scaling (dependent variable)

Dependent variable rescaled:

$$
\begin{aligned}
cy &= \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + cu \quad \longrightarrow \quad c\hat{y} = \boxed{c\hat{\beta}_0} + \boxed{c\hat{\beta}_1} x_1 + \boxed{c\hat{\beta}_2} x_2 \\
y &= \frac{\gamma_0}{c} + \gamma_1 \frac{x_1}{c} + \gamma_2 \frac{x_2}{c} + u
\end{aligned}
$$

- Effect on **all parameters and estimates** is parallel to the independent variable rescaling, i.e., $\gamma_k = c\beta_k$ and $\hat{\gamma}_k = c\hat{\beta}_k$, $i = 0, \ldots, k$.
- $SST_{x_k/c}$ changes again:
  - $SST_{x_k/c} = \sum (x_{ki}/c - \bar{x}_k/c)^2 = \frac{1}{c^2} SST_{x_k}$
  - $se(\hat{\gamma}_k) = \frac{\hat{\sigma}}{\sqrt{\frac{SST_{x_k}}{c^2}(1-R^2_{x_k})}} = c \cdot se(\hat{\beta})$
- In the end, the estimate is rescaled by the factor of c, *se* also by the factor of c, i.e., t stats and p-values remain unchanged.

Effects of scaling (logarithms)

- Dependent variable in logs rescaled:

$$\begin{aligned}
\log(y) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \\
\log(cy) &= \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + v \\
\log(c) + \log(y) &= \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + v \\
\log(y) &= \gamma_0 - \log(c) + \beta_1 x_1 + \beta_2 x_2 + v \\
&\Rightarrow \boxed{\gamma_0 = \beta_0 + \log(c)} \\
\Rightarrow \widehat{\log(cy)} &= \boxed{\hat{\beta}_0 + \log(c)} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + v
\end{aligned}$$

  $\Rightarrow$ no effect on the slope parameters, only the intercept is adjusted.

- Independent variable in logs rescaled: parallel to the case above, only the intercept is adjusted.
- Rescaling within the logarithmic transforms thus translates only into the intercept estimates.

Illustrative example

TABLE 6.1 Effects of Data Scaling

Dependent Variable	(1) *bwght*	(2) *bwghtlbs*	(3) *bwght*
Independent Variables			
cigs	−.4634 (.0916)	−.0289 (.0057)	—
packs	—	—	−9.268 (1.832)
faminc	.0927 (.0292)	.0058 (.0018)	.0927 (.0292)
intercept	116.974 (1.049)	7.3109 (.0656)	116.974 (1.049)
Observations	1,388	1,388	1,388
R-Squared	.0298	.0298	.0298
SSR	557,485.51	2,177.6778	557,485.51
SER	20.063	1.2539	20.063

Source: Wooldridge (2012)

Another type of 'scaling' (not mandatory): Standardized 'beta coefficients'

Outline

Outline

More on logarithmic forms

- Consider the following model:

$$\log(y) = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + u.$$

- **Semi-elasticity** $100\hat{\beta}_2$ is only an **approximate percentage change** in y, but it always lies between the absolute values for an increase and a decrease of x_2.
- Exact change is: $\%\Delta\hat{y} = 100\big(\exp(\hat{\beta}_2 \Delta x_2) - 1\big).$
- **Log-level form** is often used for relationships with an increasing rate of change (wages, sales, market value).
- **Level-log form** for a diminishing rate of change.
- Models with $\log(y)$ often more closely satisfy the CLM assumptions (linearity, homoskedasticity, normality).
- Logarithmic transformation also generally reduces the variation of variables and narrows its range (large salaries, big populations).

Outline

Models with quadratics

- Polynomial forms are good approximations of various nonlinear relationships (Taylor expansions) and are useful for their interpretation.
- Consider the following model: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + u.$
- Partial derivative then implies:

$$\Delta \hat{y} = (\hat{\beta}_1 + 2\hat{\beta}_2 x)\Delta x.$$

- Effect of x on $\hat{y}$ is thus a function of x dependent on values of $\hat{\beta}_1$ and $\hat{\beta}_2$.
- x^2 used for modelling a **diminishing effect** (of a cumulative variable): $\hat{\beta}_1 > 0$, $\hat{\beta}_2 < 0$, $\hat{\beta}_1 >> |\hat{\beta}_2|$, $TP = -\hat{\beta}_1/(2\hat{\beta}_2)$.
- Higher polynomials can be utilized using the same logic (but the interpretation, of course, becomes a bit trickier).

Models with quadratics: Logarithms

- We may combine the quadratic relationship with the logarithmic one.
- A square of the logarithm must be used, not a square of x in the logarithm!
- Interpretation becomes a bit more complicated.
- Consider the following model:

$$\log y = \beta_0 + \beta_1 \log x + \beta_2 (\log x)^2 + u.$$

- To get the relationship between x and y, we again use the partial derivate but now with respect to $\log x$:

$$\%\Delta \hat{y} = (\hat{\beta}_1 + 2\hat{\beta}_2 \log x)\%\Delta x.$$

Outline

Models with interaction terms

- Sometimes, the effect of one explanatory variable on the dependent variable can itself be dependent on the magnitude of yet another explanatory variable.
- To make this less confusing, consider the following model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u.$$

- To see the effect of x_1 on Δy, we can again use the partial derivative to get

$$\Delta \hat{y} = (\hat{\beta}_1 + \hat{\beta}_3 x_2)\Delta x_1.$$

- This is called an **interaction effect**.
- Significance of β_3, i.e., the existence of the interaction effect, can be easily tested using the standard procedures (t test, possibly F test).

Note on reporting

- As in these specific cases, the respective βs do not give us the effect straight away; it is useful to report the effects for some interesting values (average, median, quantiles) of the independent variable of interest.
- Connected to the previous point, reporting a range (min to max) of the effect is sometimes useful.
- Remember that when an effect depends on an independent variable, so is the estimator's variance.
- The transformations you use should be reasonable and possible to interpret.

Outline

Goodness-of-fit measures

- So far, we have considered the coefficient of determination R^2.
- Adding an independent variable to a model never decreases R^2, which makes it useless for variable selection/model construction.
- As an alternative goodness-of-fit measure, the adjusted R^2 labelled as $\bar{R}^2$ controls for the number of explanatory variables:

$$R^2 = 1 - \frac{SSR}{SST} \quad = \quad 1 - \frac{\frac{SSR}{n}}{\frac{SST}{n}},$$

$$\boxed{\bar{R}^2} = 1 - (1 - R^2)\frac{n-1}{n-k-1} \quad = \quad \boxed{1 - \frac{\frac{SSR}{n-k-1}}{\frac{SST}{n-1}}}.$$

- Final formula provides the logic behind $\bar{R}^2$, i.e., the interplay between SSR and df.
- Asymptotically, $\bar{R}^2 = R^2$.
- Interestingly, $\bar{R}^2$ can be even negative.

Adjusted $\bar{R}^2$

Adjusted coefficient of determination $\bar{R}^2$ has two interesting and potentially useful properties:

- When we add a new independent variable, $\bar{R}^2$ increases if and only if the t statistic on the new variable is greater than one in absolute value.
- When a group of new independent variables is added, $\bar{R}^2$ increases if and only if the F statistic on the group of new variables is greater than one.

Outline

Choosing between nonnested models

- Consider two models:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u,$$
$$y = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_3 + v.$$

- These are called **nonnested models** as neither one is a special case of the other.
- We cannot use the F statistics for nonnested models (creating a broader composite model with all independent variables can lead to ambiguous results of hypotheses testing).
- R^2 can be used only for the same number of independent variables.
- $\bar{R}^2$ can also be used for comparing various specifications (e.g., the logarithmic vs. quadratic form of the explanatory variable).
- However, the **dependent variable** needs to have **the same functional form** across all competing models.

Controlling for too many explanatory variables

- 'Quest for high R^2' can lead to an **over-specified/ over-fitted** model, i.e., with too many explanatory variables.
- As the adjusted $\bar{R}^2$ penalizes for additional variables, it can also serve as a signal to stop adding explanatory variables.
- $\bar{R}^2$ construction tells us whether the reduction in error variance is sufficient.
- As an alternative to R^2 and $\bar{R}^2$, various information criteria (AIC, SBIC, HQC) provide higher flexibility in penalizing additional explanatory variables. These can be used for cross-sectional data but are usually used for time series analysis and probabilistic models.

Four important variable selection criteria

Does an explanatory variable belong to the model?

1. **Theory:** Is including a variable in the equation unambiguous and theoretically sound? Does intuition suggest that it should be included? Also, the modeling purpose is crucial:
   - prediction/explanation
   - vs. testing a specific theoretical/empirical relationship
2. **Omitted variable bias reduction:** Do estimated coefficients of other variables change considerably when the variable is added to the model? It is essential to avoid serious OVB.
3. **Adjusted $\bar{R}^2$:** Does the overall fit of the equation improve (enough) when the variable is added to the model?
4. t **test and** F **test:** Is its coefficient statistically significant in the expected direction? F *test* can help us when considering excluding multiple variables or for step-wise elimination.

Seminars and the next lecture

- Seminars:
  - data scaling
  - practicing logarithmic and quadratic functional forms
  - practicing interaction terms
  - model comparison
- Next lecture #8:
  - prediction and residual analysis
  - multiple regression with qualitative information
    - single binary/dummy independent variable
    - using dummy variables for multiple categories
- Readings for lecture #8:
  - Chapter 6: 6.4, Chapter 7: 7.1–3

Appendix: Standardized 'beta coefficients' (not mand.)

- Sometimes, we are not interested in interpreting the effects in measurement units (usually due to non-standard measurement units of the variable of interest) but in **standard deviations**.
- We thus **standardize** the variables of interest to obtain the standardized **'beta coefficients'**.
- All variables are put on 'equal grounds'.
- Effects are interpreted as changes in standard deviations, often used in sociology and for survey analyses.
- Statistical significance **unaffected** (t stats, p-values).
- In a simple regression model, the beta coefficient b equals the correlation coefficient between the dependent and the independent variable.

Standardized 'beta coefficients': Derivation

- Starting from the estimated model

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{u}_i,$$

  we can subtract the corresponding averages from each variable in the equation to get

$$\hat{y}_i - \bar{y} = 0 + \hat{\beta}_1 (x_{1i} - \bar{x_1}) + \ldots + (\hat{u}_i - 0).$$

- Having $\hat{\sigma}_y$ as a sample standard deviation of y and $\hat{\sigma}_j$, $i = 1, \ldots, k$, as sample standard deviations of the independent variables, we can rewrite the equation as

$$\frac{y_i - \bar{y}}{\hat{\sigma}_y} = \hat{\beta}_1 \frac{\hat{\sigma}_1}{\hat{\sigma}_y} \frac{x_{1i} - \bar{x_1}}{\hat{\sigma}_1} + \ldots + \frac{\hat{u}_i}{\hat{\sigma}_y}.$$

- Finally, using the z-score notation, we can write the equation as

$$z_y = \hat{b}_1 z_1 + \ldots + \hat{e}_i,$$

  giving us the **standardized beta coefficients** as $\boxed{\hat{b}_j = \frac{\hat{\sigma}_j}{\hat{\sigma}_y} \hat{\beta}_j}$.