FINM3123 Introduction to Econometrics

Chapter 6

Multiple Regression Analysis: Further Issues

Multiple Regression Analysis: Further Issues

- <u>**More on Functional Form**</u>
- <u>**More on using logarithmic functional forms**</u>
  - Convenient percentage/elasticity interpretation
  - Slope coefficients of logged variables are invariant to rescalings
  - Taking logs often eliminates/mitigates problems with outliers
  - Taking logs often helps to secure normality and homoscedasticity
  - Variables measured in units such as years should not be logged
  - Variables measured in percentage points should also not be logged
  - Logs must not be used if variables take on zero or negative values
  - It is hard to reverse the log-operation when constructing predictions

Multiple Regression Analysis: Further Issues

Using quadratic functional forms

Concave experience profile

- **Example: Wage equation**

$$\widehat{wage} = \underset{(.35)}{3.73} + \underset{(.041)}{.298}\ exper - \underset{(.0009)}{.0061}\ exper^2$$

$$n = 526, R^2 = .093$$

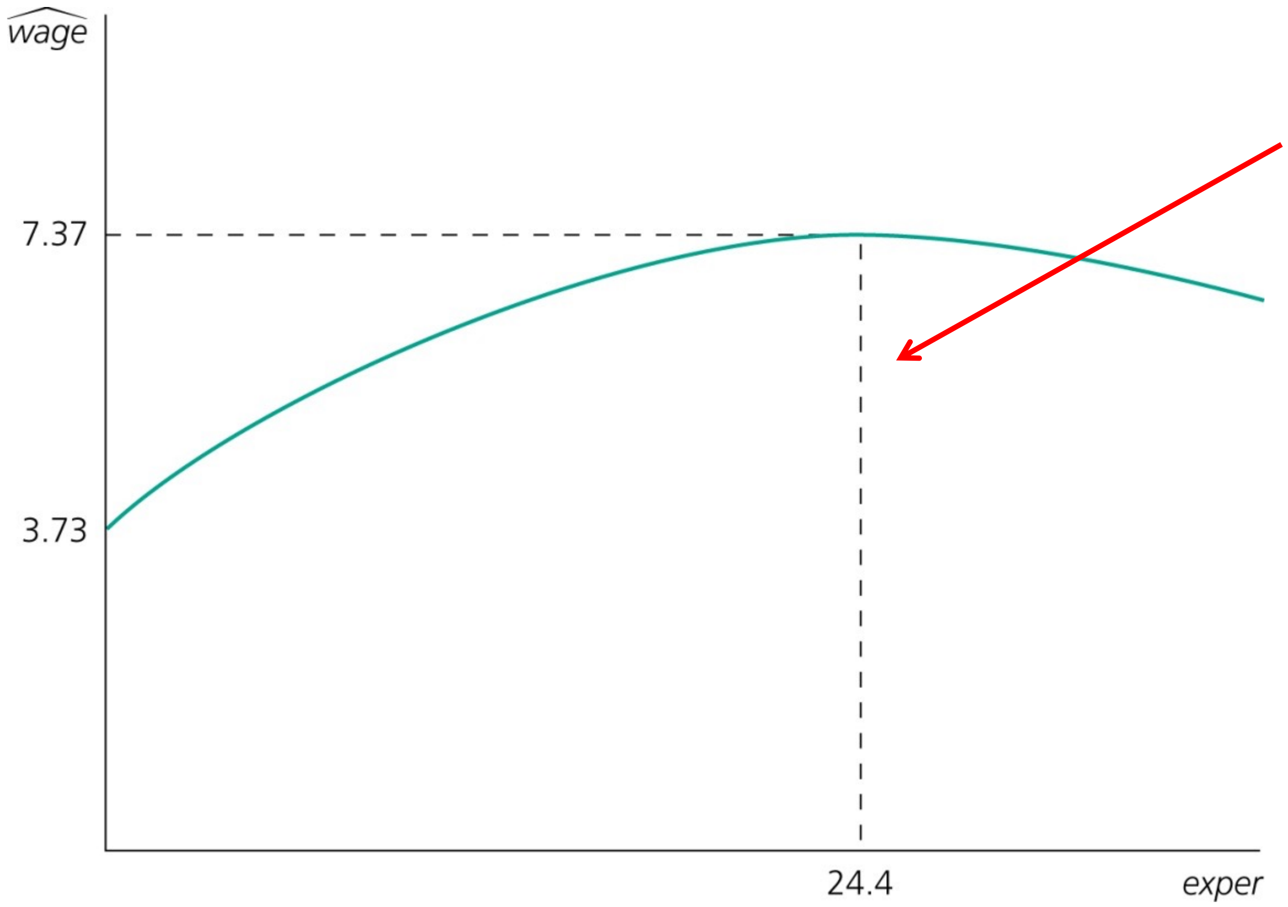
- **Marginal effect of experience**

$$\frac{\partial wage}{\partial exper} = .298 - 2(.0061)exper$$

The first year of experience increases the wage by some .30$, the second year by .298-2(.0061)(1) = .29$ etc.

Multiple Regression Analysis: Further Issues

Wage maximum with respect to work experience



$$x^* = \left|\frac{\hat{\beta}_1}{2\hat{\beta}_2}\right| = \left|\frac{.298}{2(.0061)}\right| \approx 24.4$$

Does this mean the return to experience becomes negative after 24.4 years?

Not necessarily. It depends on how many observations in the sample lie right of the turnaround point.

In the given example, these are about 28% of the observations. There may be a specification problem (e.g. omitted variables).

Multiple Regression Analysis: Further Issues

Example: Effects of pollution on housing prices

Nitrogen oxide in air, distance from employment centers, student/teacher ratio

$$\widehat{\log}(price) = \underset{(.57)}{13.39} - \underset{(.115)}{.902}\ \log(nox) - \underset{(.043)}{.087}\ \log(dist)$$

$$\underset{(.165)}{-\ .545}\ rooms + \underset{(.013)}{.062}\ rooms^2 - \underset{(.006)}{.048}\ stratio$$

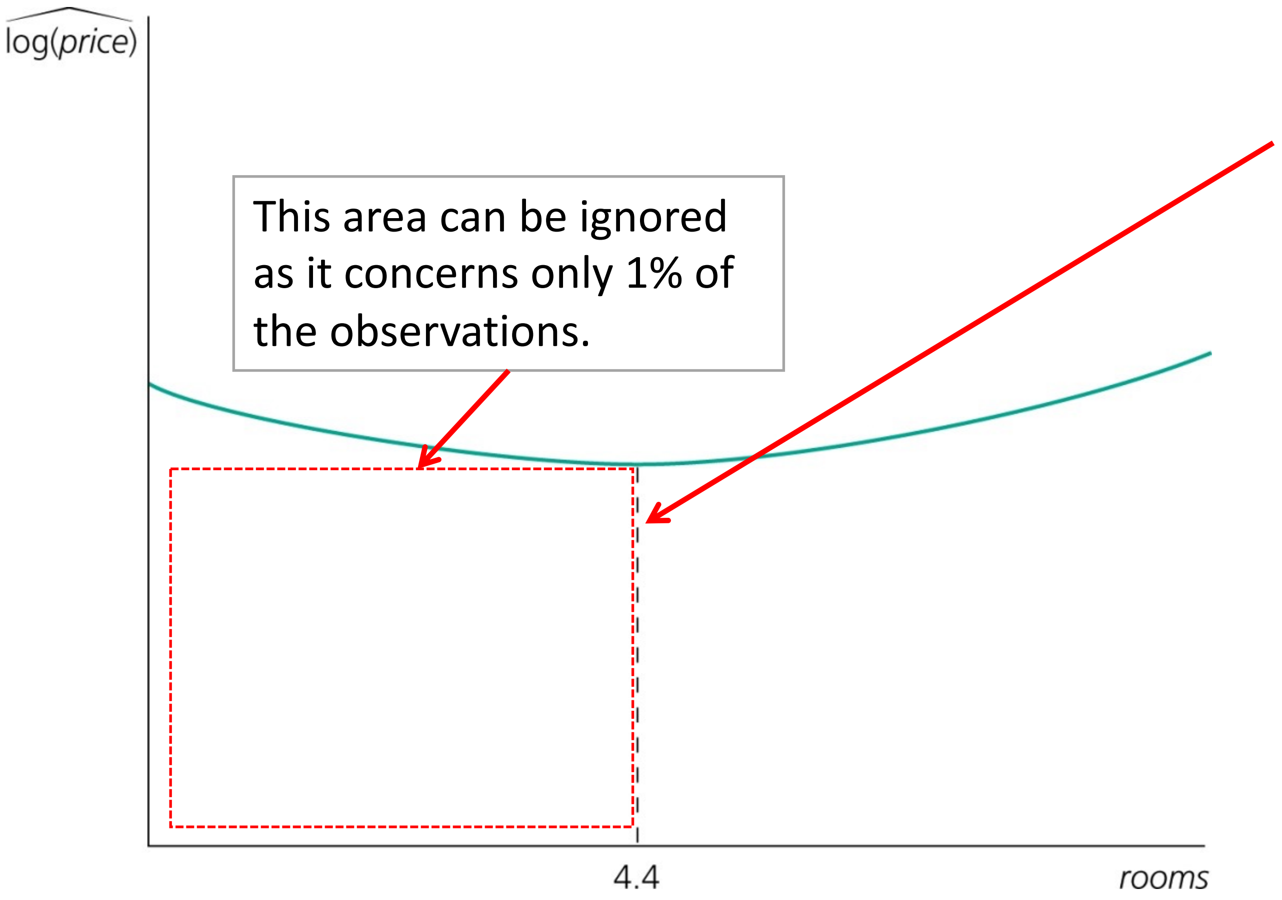
$$n = 506, R^2 = .603$$

Does this mean that, at a low number of rooms, more rooms are associated with lower prices?

$$\Rightarrow \quad \frac{\partial \log(price)}{\partial rooms} = \frac{\% \partial price}{\partial rooms} = -.545 + .124\, rooms$$

Multiple Regression Analysis: Further Issues

Calculation of the turnaround point



This area can be ignored as it concerns only 1% of the observations.

Turnaround point: $x^* = \left|\frac{-.545}{2(.062)}\right| \approx 4.4$

Increase rooms from 5 to 6:

$-.545 + .124(5) = +7.5\%\ price$

Increase rooms from 6 to 7:

$-.545 + .124(6) = +19.9\%\ price$

Multiple Regression Analysis: Further Issues

- **Other possibilities**

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2 [\log(nox)]^2$$

$$+\beta_3 crime + \beta_4 rooms + \beta_5 rooms^2 + \beta_6 stratio + u$$

$$\Rightarrow \quad \frac{\partial \log(price)}{\partial \log(nox)} = \frac{\% \partial price}{\% \partial nox} = \beta_1 + 2\beta_2[\log(nox)]$$

- **Higher polynomials**

$$cost = \beta_0 + \beta_1 quantity + \beta_2 quantity^2 + \beta_3 quantity^3 + u$$

Multiple Regression Analysis: Further Issues

- **Models with interaction terms**

$$\log(price) = \beta_0 + \beta_1 sqrft + \beta_2 bdrms$$

Interaction term

$$+\beta_3 \boxed{sqrft \cdot bdrms} + \beta_4 bthrms + u$$

$$\Rightarrow \quad \frac{\partial \log(price)}{\partial bdrms} = \beta_2 + \beta_3 sqrft$$

The effect of the number of bedrooms depends on the level of square footage

- **Interaction effects complicate the interpretation of parameters**

$\beta_2 =$ effect of number of bedrooms, but for a square footage of zero

Multiple Regression Analysis: Further Issues

- **Reparametrization of interaction effects**

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

Population means; may be replaced by sample means

$$y = \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + \beta_3 (x_1 - \mu_1)(x_2 - \mu_2) + u$$

Effect of x_2 if all variables take on their mean values

- **Advantages of reparametrization**
  - Easy interpretation of all parameters
  - Standard errors for partial effects at the mean values available
  - If necessary, interaction may be centered at other interesting values

Multiple Regression Analysis: Further Issues

<u>More on goodness-of-fit and selection of regressors</u>

- **General remarks on R-squared**
  - A high R-squared may result in misleading conclusions.
  - A low R-squared does not preclude precise estimation of partial effects
- **Adjusted R-squared**
  - What is the ordinary R-squared supposed to measure?

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{(SSR/n)}{(SST/n)} \quad \text{is an estimate for} \quad \boxed{1 - \frac{\sigma_u^2}{\sigma_y^2}}$$

Population R-squared

Multiple Regression Analysis: Further Issues

Correct degrees of freedom of numerator and denominator

- **Adjusted R-squared (cont.)**
  - A better estimate taking into account degrees of freedom would be

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

  - <u>The adjusted R-squared imposes a penalty for adding new regressors</u>
  - The adjusted R-squared increases if, and only if, the t-statistic of a newly added regressor is greater than one in absolute value

- **Relationship between R-squared and adjusted R-squared**

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

The adjusted R-squared may even get negative

Multiple Regression Analysis: Further Issues

Using adjusted R-squared to choose between nonnested models

- Models are nonnested if neither model is a special case of the other

$$rdintens = \beta_0 + \beta_1 \log(sales) + u$$ ← $R^2 = .061, \bar{R}^2 = .030$

$$rdintens = \beta_0 + \beta_1 sales + \beta_2 sales^2 + u$$ ← $R^2 = .148, \bar{R}^2 = .090$

- A comparison between the R-squared of both models would be unfair to the first model because the first model contains fewer parameters
- In the given example, even after adjusting for the difference in degrees of freedom, the quadratic model is preferred

Multiple Regression Analysis: Further Issues

Comparing models with different dependent variables

- R-squared or adjusted R-squared should **not** be used to compare models which differ in their definition of the dependent variable

Example: CEO compensation and firm performance

$$\boxed{\widehat{salary}} = \underset{(223.63)}{223.90} + \underset{(.0163)}{.0089}\ sales + \underset{(11.08)}{19.63}\ roe$$

$$n = 209, R^2 = .029, \bar{R}^2 = .020, \boxed{TSS = 391,732,982}$$

$$\boxed{\widehat{lsalary}} = \underset{(0.29)}{4.36} + \underset{(.033)}{.275}\ lsales + \underset{(.0040)}{.0179}\ roe$$

$$n = 209, R^2 = .282, \bar{R}^2 = .275, \boxed{TSS = 66.72}$$

There is much less variation in log(salary) that needs to be explained than in salary

Multiple Regression Analysis: Further Issues

Controlling for too many factors in regression analysis: <u>over controlling</u>

- **In some cases, certain variables should not be held fixed**
  - In a regression of traffic fatalities on state beer taxes (and other factors) one should not directly control for beer consumption
  - In a regression of family health expenditures on pesticide usage among farmers one should not control for doctor visits
- **Different regressions may serve different purposes**
  - In a regression of house prices on house characteristics, one would only include price assessments if the purpose of the regression is to study their validity; otherwise one would not include them

Multiple Regression Analysis: Further Issues

- **Adding regressors to reduce the error variance**
  - Adding regressors may exacerbate multicollinearity problems
  - On the other hand, adding regressors reduces the error variance
  - Variables that are uncorrelated with other regressors should be added because they reduce error variance without increasing multicollinearity
  - However, such uncorrelated variables may be hard to find
- **Example: Individual beer consumption and beer prices**
  - Including individual characteristics (such as age and education) in a regression of beer consumption on beer prices leads to more precise estimates of the price elasticity

Multiple Regression Analysis: Further Issues

Predicting y when log(y) is the dependent variable

$$\log y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

population model

$$\widehat{\log y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

prediction

The first idea to predict y could be to set $\hat{y} = \exp\left(\widehat{\log y}\right)$ but this *does not work*, in fact, this would always *underestimate* the expected value of y.

Take exponential on both sides of the population model:

$$y = \exp(u)\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)$$

Under Assumptions MLR.1-MLR.6, we know that $u \sim N(0, \sigma^2)$, therefore

$$\mathbb{E}(y|\boldsymbol{x}) \ = \exp(\sigma^2/2)\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)$$

Multiple Regression Analysis: Further Issues

Define $\widehat{m}_i \coloneqq \exp\left(\widehat{\log y}\right) = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k)$

A predictor for y is $\boldsymbol{\hat{y} = \exp\left(\hat{\sigma}^2/2\right)\widehat{m}_i}$

If we only assume MLR.1-MLR.5, then we don't know $\mathbb{E}(\exp u)$, but we can approximate it by $\hat{\alpha}_0 \coloneqq \frac{1}{n}\sum_{i=1}^{n} \exp(\hat{u}_i)$, where $\hat{u}_i = \log y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_k x_{ik}$

Then we can predict y by $\boldsymbol{\hat{y} = \hat{\alpha}_0 \widehat{m}_i}$

Another approximation of $\mathbb{E}(\exp u)$ can be obtained by a simple regression through the origin of y_i on $\widehat{m}_i$, since $y = \exp(u) m_i$.
The OLS slope estimate from the simple regression y_i on $\widehat{m}_i$ (no intercept) is $\check{\alpha}_0 = \frac{\sum_{i=1}^{n} \widehat{m}_i y_i}{\sum_{i=1}^{n} \widehat{m}_i^2}$

Then we can predict y by $\boldsymbol{\hat{y} = \check{\alpha}_0 \widehat{m}_i}$

These three predictors of y from $\widehat{\log y}$ are consistent but not unbiased.

Multiple Regression Analysis: Further Issues

Comparing R-squared of a logged and an unlogged specification

$$\boxed{\widehat{salary}} = \underset{(65.23)}{613.43} + \underset{(.0100)}{.0190}\ sales + \underset{(.0095)}{.0234}\ mktval + \underset{(5.61)}{12.70}\ ceoten$$

$$n = 177, R^2 = .201$$

R^2 is the square of the correlation between y_i and $\hat{y}_i$

$$\boxed{\widehat{lsalary}} = \underset{(.257)}{4.504} + \underset{(.039)}{.163}\ lsales + \underset{(.050)}{.109}\ mktval + \underset{(.0053)}{.0117}\ ceoten$$

$$n = 177, R^2 = .318$$

$$\tilde{R}^2 = .243$$

These are the R-squareds for the predictions of the unlogged salary variable (although the second regression is originally for logged salaries). Both R-squareds can now be directly compared.

$\tilde{R}^2$ can be defined as the square of the correlation between y_i and $\hat{y}_i = \hat{\alpha}_0 \widehat{m}_i$ (which is equal to the square of the correlation between y_i and $\widehat{m}_i$). Another possible definition is to define the residuals $\hat{r}_i = y_i - \hat{\alpha}_0 \widehat{m}_i$ and use them in the formula for R-squared from linear regression:

$$1 - \frac{\sum_{i=1}^{n} \hat{r}_i}{\sum_{i=1}^{n} (y_i - \bar{y})}$$

Summary

- Functional form
  - Logarithmic functional form
    - Interpretation: elasticity, semi-elasticity
    - When to use logarithmic form
  - Quadratic functional form
    - Turnaround point calculation
  - Interaction terms
    - Interpretation of parameters
    - Reparametrization

Summary

- Adjusted R-squared

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

  - Adjusted R-squared increases if and only if the t-statistic of a newly added regressor is greater than one in absolute value.
  - Relationship between R-squared and adjusted R-squared

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

  - Can be used to choose between nonnested models when dependent variable is the same

- Predicting y when log(y) is the dependent variable
  - Two approaches: $\hat{\alpha}_0, \check{\alpha}_0$.
  - Comparing R-squared of a logged and an unlogged specification.