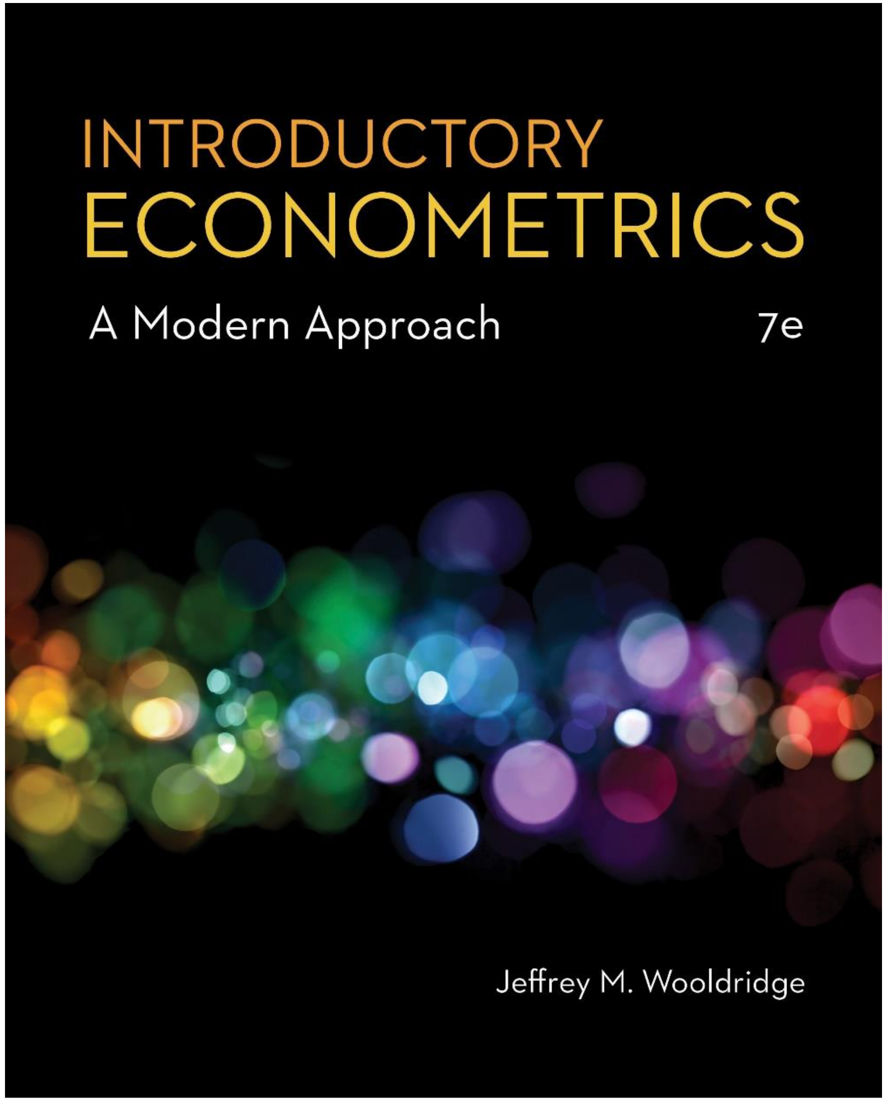
Chapter 6

Multiple Regression Analysis: Further Issues



Multiple Regression Analysis: Further Issues (1 of 16)

- **More on Functional Form**

- More on using logarithmic functional forms:
  - 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 homoskedasticity
  - 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 (2 of 16)

- **Using quadratic functional forms**
- Example: Wage equation

WAGE1.DTA

Concave experience profile

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

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

- Marginal effect of experience

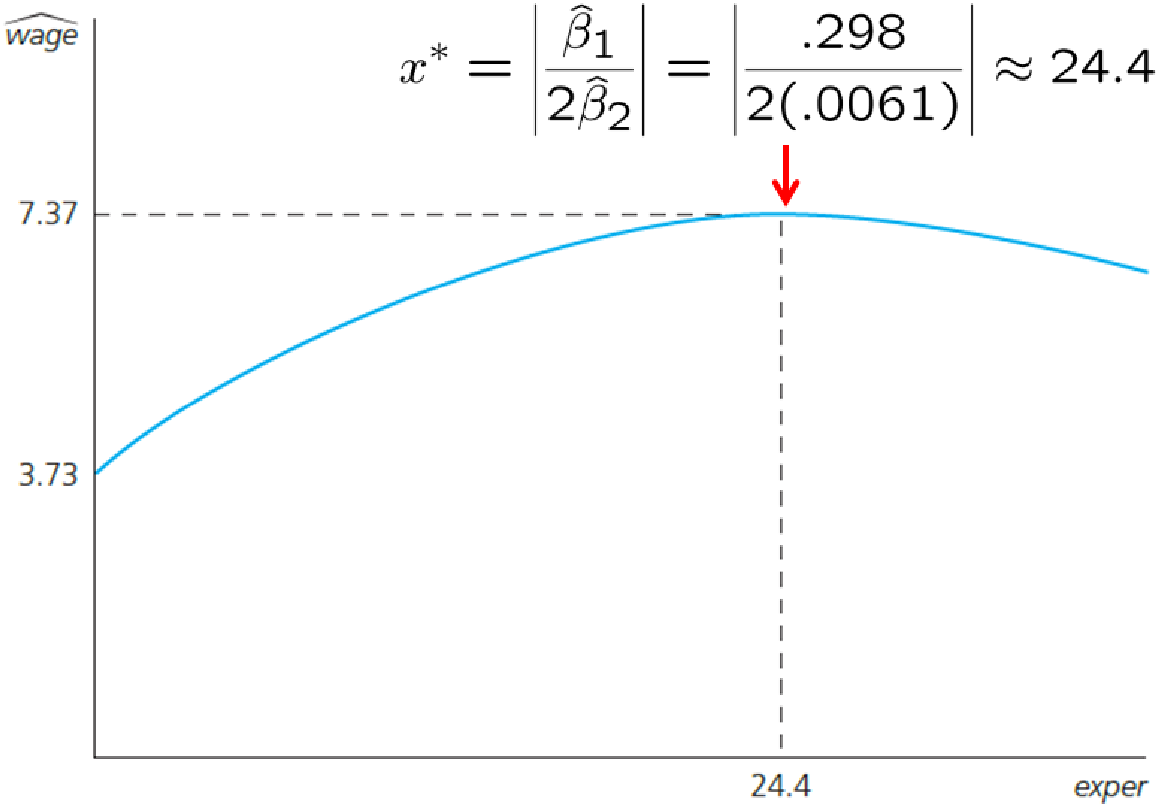
$$\frac{\Delta wage}{\Delta exper} = .298 - 2(.0061)exper$$

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 (3 of 16)

- **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 to the 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 (4 of 16)

- **Example: Effects of pollution on housing prices**

HPRICE2.DTA

$$\widehat{\log(price)} = \underset{(.57)}{13.39} - \underset{(.115)}{.902}\log(nox) - \underset{(.043)}{.087}\log(dist)$$
$$- \underset{(.165)}{.545}\, rooms + \underset{(.013)}{.062}\, rooms^2 - \underset{(.006)}{048}\, stratio$$
$$n = 506,\ R^2 = .603$$

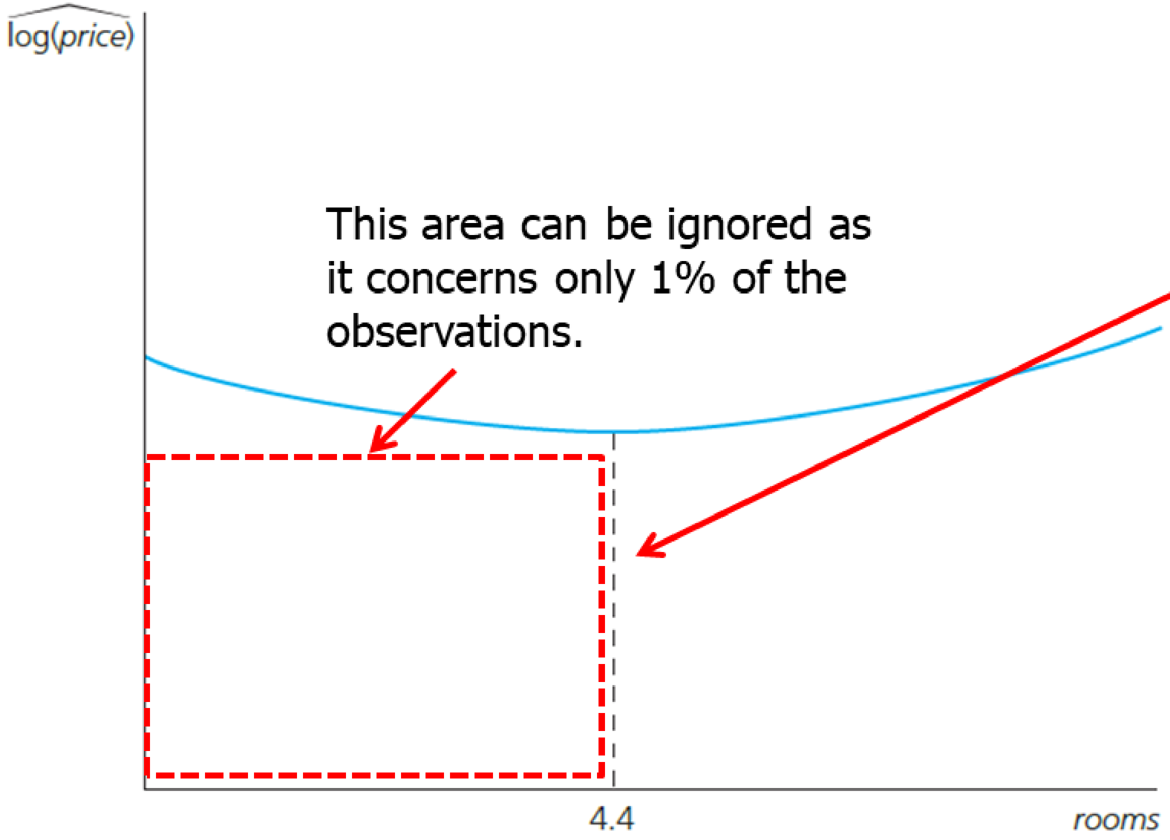
nox: nitrogen oxide in the air
$dist$: distance from employment centers
$rooms$: number of rooms
$stratio$: average student/teacher ratio

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

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

Multiple Regression Analysis: Further Issues (5 of 16)

- **Calculation of the turnaround point**



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 (6 of 16)

- Other possibilities

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2[\log(nox)]^2$$
$$+\beta_3 crime + \beta_4 rooms + \beta_5 rooms^2 + \beta_6 stratio + u$$

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

- Higher polynomials

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

Multiple Regression Analysis: Further Issues (7 of 16)

- **Models with interaction terms**

hprice1.dta

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms$$

$$+\beta_3 sqrft \cdot bdrms + \beta_4 bthrms + u$$

Interaction term

$$\Rightarrow \quad \frac{\Delta price}{\Delta bdrms} = \beta_2 + \beta_3 sqrft$$

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

- Interaction effects complicate interpretation of parameters

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

Multiple Regression Analysis: Further Issues (8 of 16)

- **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 (9 of 16)

- **Average partial effects**
- In models with quadratics, interactions, and other nonlinear functional forms, the partial effect depend on the values of one or more explanatory variables
- Average partial effect (APE) is a summary measure to describe the relationship between dependent variable and each explanatory variable
- After computing the partial effect and plugging in the estimated parameters, average the partial effects for each unit across the sample

Multiple Regression Analysis: Further Issues (10 of 16)

- **More on goodness-of-fit and selection of regressors**

- General remarks on R-squared
  - A high R-squared does not imply that there is a causal interpretation
  - A low R-squared does not preclude precise estimation of partial effects

- Adjusted R-squared
  - What is the ordinary R-squared supposed to measure?

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

Population R-squared

Multiple Regression Analysis: Further Issues (11 of 16)

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

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

  - The adjusted R-squared imposes a penalty for adding new regressors
  - 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 (12 of 16)

- **Using adjusted R-squared to choose between nonnested models**
  - Models are nonnested if neither model is a special case of the other

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

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

- 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 (13 of 16)

- **Comparing models with different dependent variables**
  - R-squared or adjusted R-squared must not be used to compare models which differ in their definition of the dependent variable

- Example: CEO compensation and firm performance

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

$$\widehat{salary} = \underset{(223.90)}{830.63} + \underset{(.0089)}{.0163}\, sales + \underset{(11.08)}{19.03}\, roe$$

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

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

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

Multiple Regression Analysis: Further Issues (14 of 16)

- **Controlling for too many factors in regression analysis**

- 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 (15 of 16)

- **Adding regressors to reduce the error variance**
  - Adding regressors may excarcerbate 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 in a regression of beer consumption on beer prices leads to more precise estimates of the price elasticity

Multiple Regression Analysis: Further Issues (16 of 16)

- **Predicting y when log(y) is the dependent variable**

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

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

- Under the additional assumption that u is independent of $x_1,...,x_k$:

$$\Rightarrow \quad E(y|\mathbf{x}) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k) E(\exp(u))$$

$$\Rightarrow \quad \hat{y} = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k)(\frac{1}{n}\sum_{i=1}^{n} \exp(\hat{u}_i))$$