MLDS 401/IEMS 404-1 (Fall 2023): Lab 2 – 10/03/2023

3.5 (Omitted variables) Suppose that the true linear model is

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon},$$

where $\boldsymbol{X}_1 : n \times p_1, \boldsymbol{\beta}_1 : p_1 \times 1, \boldsymbol{X}_2 : n \times p_2, \boldsymbol{\beta}_2 : p_2 \times 1$. However, we mistakenly fit the model $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$ and estimate $\boldsymbol{\beta}_1$ by $\widehat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y}$.

a) Show that, in general, $\widehat{\boldsymbol{\beta}}_1$ is a biased estimator with

$$\text{Bias}(\widehat{\boldsymbol{\beta}}_1) = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\boldsymbol{\beta}_2.$$

Note that this formula generalizes that given in Exercise 2.4.

b) Under what condition on $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ is $\widehat{\boldsymbol{\beta}}_1$ unbiased? How is this condition related to that in Exercise 2.4 for $\widehat{\beta}_1$ to be unbiased?

3.10 (Cobb-Douglas production function) Data on 569 European companies on their capital (K) measured as the total fixed assets (in millions of euros) at the end of 1995, labor (L) measured as the number of workers and output (O) measured as the value added (in millions of euros) are available in file `cobbdouglas.csv`. The companies in this data set are from different industry sectors in which different Cobb-Douglas production functions may apply since their capital and labor requirements are different, but we ignore this problem.

a) Fit the Cobb-Douglas production function $y = \gamma K^{\alpha} L^{\beta} \varepsilon$ by making the log-transformation , so that the model becomes $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ln \varepsilon$, where $y = \ln O, x_1 = \ln K, x_2 = \ln L, \beta_0 = \ln \gamma, \beta_1 = \alpha$ and $\beta_2 = \beta$.

b) If $\alpha + \beta = \beta_1 + \beta_2 = 1$ then it is easy to check that if the capital and labor are changed by a common scaling factor then the output is changed by the same factor. In economics this is known as the **constant returns to scale**. Test the null hypothesis of the constant returns to scale for these data by testing $H_0 : \beta_1 + \beta_2 = 1$ using the t-statistic

$$t = \frac{\widehat{\beta}_1 + \widehat{\beta}_2 - 1}{\sqrt{(\widehat{\text{Var}}(\widehat{\beta}_1) + \widehat{\text{Var}}(\widehat{\beta}_2) + 2\widehat{\text{Cov}}(\widehat{\beta}_1, \widehat{\beta}_2))}},$$

where the estimates of $\text{Var}(\widehat{\beta}_1), \text{Var}(\widehat{\beta}_2)$ and $\text{Cov}(\widehat{\beta}_1, \widehat{\beta}_2)$ are available by using the `vcov` function in R.

c) An alternative way to do the above test is to use $y - x_2$ as the response variable and fit the model $y - x_2 = \beta_0 + \beta_1(x_1 - x_2) + (\beta_1 + \beta_2 - 1)x_2 + \ln \varepsilon$, where $x_1 - x_2$ and x_2 are the new predictor variables. Test the significance of the coefficient of x_2 in this regression, i.e., test $H_0 : \beta_1 + \beta_2 - 1 = 0$ using a t-test. Is the result the same as that obtained in Part (b)?

3.14 **(Salary data)** File `salaries.csv` contains data on annual salaries of 46 employees of a company and possible predictors. The variable definitions are given in Table 3.10. Use $\log_{10}$(Salary) as the response variable.

a) Fit a prediction model using the given data. Using Female and Advertising as reference categories for Gender and Dept categorical variables, check that the fitted equation is

$$\begin{aligned}\widehat{\log_{10}(\text{Salary})} &= 4.429 + 0.0075\ \text{YrsEm} + 0.0017\ \text{PriorYr} + 0.0170\ \text{Educ} + 0.0004\ \text{Super} \\ &\quad +0.0231\ \text{Female} - 0.0388\ \text{Adver} - 00057\ \text{Engg} - 0.0938\ \text{Sales}.\end{aligned}$$

b) If we use Female and Purchase as reference categories, what will be the new coefficients for Male and for the other three departments?

c) Predict a person's salary along with the associated prediction interval for the following values of predictor variables: YrsEm = 8, PriorYr = 10, Education= 12, Gender = Male, Dept = Sales, Super = 5.

d) The coefficient of Engg is highly nonsignificant with a P-value = 0.774 in the above regression. But if Sales is used as the reference category, the coefficient of Engg is highly significant with a P-value < 0.001. Interpret this result.

e) In the above model, the coefficient of Female is nonsignificant with $P = 0.115$, so there is not a significant difference between the salaries of Males and Females, controlling for other variables. Should we drop Gender as a predictor variable? For Dept the question is more complicated since the coefficients of Advert and Engg are nonsignificant but the coefficient of Sales in highly significant with $P = 0.0002$. Should the nonsignificant categories, Advert and Engg, be pooled with the reference category, Purchase, or should the categories be left unpooled or should the Dept be dropped as a predictor variable?

Table 3.10 Salary Data Variables

Variable	Explanation
Salary	Annual salary in $
YrsEm	No. of years employed with the company
PriorYr	No. of years of prior experience
Educ	No. of years of education after high school
Super	No. of people supervised
Gender	M = Male, F = Female
Dept	Advertising, Engineering, Purchase, Sales

Source: McKenzie and Goldman (1999, Temco Data Set)

4.4 (College GPA and entrance test scores: Checking normality and homoscedasticity) Refer to Example 3.15 in which we fitted the model

$$\text{GPA} = \beta_0 + \beta_1\text{Verbal} + \beta_2\text{Math} + \beta_3\text{Verbal}^2 + \beta_4\text{Math}^2 + \beta_5\text{Verbal} \times \text{Math} + \varepsilon.$$

a) Make the normal and fitted values plots of residuals. Comment on why the normality and especially the homoscedasticity assumption seem to be violated. Does the fitted values plot suggest the log transformation of GPA?

b) Fit the same model using log(GPA) as the response variable. Make the normal and fitted values plots of residuals. Are the normality and homoscedasticity assumptions satisfied now?