Multivariate Regression

Jerome Dumortier

17 August 2023

Lecture Overview

Extension of the bivariate model to multivariate regression:

- One dependent variable but multiple independent variables

Topics associated with multivariate regression models covered in this lecture:

- Dummy variables
- Natural logarithm
- Functional forms
- Interaction terms

Introduction

Bivariate regression model (one independent and one dependent variable)

$$y = \beta_0 + \beta_1 \cdot x_1 + \epsilon$$

Multivariate linear regression model (multiple independent variables)

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k + \epsilon$$

Whether we consider a bivariate or multivariate model, the objective is to minimize the sum of squared errors, hence ordinary least square (OLS) model. The equation of a line can be determined using slope and intercept, i.e.:

$$E(y|x) = \beta_0 + \beta_1 \cdot x$$

A model with two independent variables (predictors) describes a plane.

Assumptions

Key assumptions of the model are identical to the bivariate model:

- Zero mean of the error terms, i.e., $E(\epsilon|x_1, \ldots, x_k) = 0$.
- No auto-regression and no serial correlation, i.e., $Cov(\epsilon_i, \epsilon_j) = 0 \quad \vee \quad i \neq j$
- Homoscedasticity, $Var(\epsilon_i) = \sigma^2$
- Linearity, i.e., the model can be expressed as a linear relationship between y and $x_1, \ldots, x_k$.

With more than one independent variable, absence of perfect multicollinearity becomes important.

- For example, including wealth and income in a regression model may result in multicollinearity.

Multivariate Regression Models

Purpose:

- Measuring the effect of one independent variable on the dependent variable
- Crucial issue: We have to control for everything else that could influence the dependent variable!

Why we have to control for everything:

- Weekly grocery bill as a function of years of education

Example: Crime rate depends on unemployment, population density, and high school dropout rate:

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3$$

where x_1 = population density, x_2 = unemployment, and x_3 = high school dropout rate

North Carolina County Data

```
##
## Call:
## lm(formula = violentcrimerate ~ popdense + unemployment + publicschoolenrollment,
##     data = crime)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31585 -0.06984 -0.00229  0.06183  0.87121
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -1.927e-01  1.462e-01  -1.318    0.191
## popdense                5.609e-04  8.278e-05   6.776 1.03e-09 ***
## unemployment            3.660e-02  8.629e-03   4.242 5.15e-05 ***
## publicschoolenrollment  1.497e-02  9.193e-03   1.628    0.107
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1701 on 95 degrees of freedom
## Multiple R-squared:  0.3777, Adjusted R-squared:  0.358
## F-statistic: 19.22 on 3 and 95 DF,  p-value: 8.036e-10
```

Child Mortality: Setup

- We think that Child Mortality (CM), measured as the number of deaths per 1000 births, is related to the per-capita Gross National Product (PGNP) and the Female Literacy rate (FLFP).
- We might hypothesize that wealthier countries have lower child mortality rates, and countries with higher female literacy rates also have lower child mortality rates.
- Why? Maybe higher PGNP means more money for healthcare; Maybe better literacy rates mean more educated mothers that can take advantage of health related information.
- The data set also contains the total fertility rate (TFR) between 1980 and 1985, i.e., the average number of children born to a woman.

Child Mortality: R Analysis

```
bhat=lm(cm~pgnp+flfp,data=childmortality)
summary(bhat)
```

```
##
## Call:
## lm(formula = cm ~ pgnp + flfp, data = childmortality)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -84.267 -24.363   0.709  19.455  96.803
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) 263.641586  11.593179  22.741  < 2e-16 ***
## pgnp         -0.005647   0.002003  -2.819  0.00649 **
## flfp         -2.231586   0.209947 -10.629 1.64e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.75 on 61 degrees of freedom
## Multiple R-squared:  0.7077, Adjusted R-squared:  0.6981
```

Child Mortality: Interpretation

- It tells us the parameter estimates for PGNP and FLR. It also tells us if our estimate is statistically different from zero.
- As expected, both variables lower child mortality.
- As per capita GNP increases by one dollar, child mortality decreases by 0.565 units, holding the effects of female literacy constant.
- Since the units of CM are number of deaths per 1000 births, it is easier to say that for a $1000 dollar increase in per capita GNP, the number of childhood deaths decreases by 5.6 per 1000. (since $0.00565 * 1000 = 5.6$ units)
- Similarly, for a one unit (one percent) increase in the female literacy rate, the childhood mortality decreases by 2.23 deaths per 1000, holding the effects of PGNP constant.

Child Mortality Data: Purpose of OLS I

Forecasting

- If a country as a per capita GNP of $3,000 and a female literacy rate of 50, then the expected number of childhood deaths per 1000 births is 135.26.
- β_1 and β_2 as partial regression or partial slope coefficients:
  - β_1 quantifies the change in the expected mean of Y, i.e., $E(Y)$, per one unit increase/decrease in x_1 holding the value of x_2 constant.
  - Important: No need to know the value of x_2.

Child Mortality Data: Purpose of OLS I

Hypothesis testing

- We reject the null hypothesis that the effect of Per Capita GNP on childhood mortality is zero at the 1% critical level, with a t-value of -2.82 and p-value of 0.0065.
- Similarly, we reject the null hypothesis of no relationship between female literacy rate and childhood mortality at the 1% critical level. FLR has a t-value of -10.63 and a p-value of 0.0001.

F-Test: Accident Data

```r
bhat = lm(accidents~temperature+precipitation,data=accidents)
summary(bhat)
```

```
##
## Call:
## lm(formula = accidents ~ temperature + precipitation, data = accidents)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6766 -1.6334 -0.0497  1.0756  4.4419
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   21.19866    0.81411  26.039  < 2e-16 ***
## temperature   -0.18839    0.01582 -11.907 2.96e-12 ***
## precipitation  0.02608    0.42042   0.062    0.951
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.82 on 27 degrees of freedom
```

F-Test: Accident Data

Hypothesis: H_0: $\beta_1 = \beta_2 = 0$

$$\frac{R^2/(k-1)}{(1-R^2)/(n-k)} \sim F_{k-1,n-k}$$

Given the accident data: $n = 30, k = 3$

$$\frac{0.8553107/2}{(1-0.8553107)/(30-3)} = 79.8034063$$

This is different and an extension from the previous versions of hypothesis testing. We are not conducting hypothesis tests on the individual parameters but on all slope coefficients simultaneously.

Comparing two R^2 values only if identical sample size and identical dependent variable.

Dummy Variables

Overview

Representation of a single qualitative characteristics coded as 0 or 1

- Examples: Religion, gender, nationality, all-wheel drive (AWD), hardwood floors

Used car examples where the *price* depends on *miles* and *AWD* (i.e., a dummy variable):

$$price_i = \beta_0 + \beta_1 \cdot miles_i + \beta_2 \cdot AWD_i + \epsilon_i$$

with $AWD_i = 1$ for an all-wheel drive car and $AWD_i = 0$ for a car with no all-wheel drive. This regression can theoretically be separated into two single equations:

- RWD: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
- AWD: $Y_i = (\beta_0 + \beta_2) + \beta_1 X_i + \epsilon_i$

Interpretation:

- Knowledge on how the dummy-variable was coded.
- If the coefficient of the dummy-variable "adds" (or "subtracts" if sign is negative) compared to the 0-group.

Dummay Variables: Interpretation

```r
bhat=lm(price ~ miles + allwheeldrive, data = bmw)
summary(bhat)
```

```
##
## Call:
## lm(formula = price ~ miles + allwheeldrive, data = bmw)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3874.1 -1724.0  -176.5  1604.5  5355.0
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.047e+04  1.711e+03  23.660  < 2e-16 ***
## miles         -2.728e-01  4.044e-02  -6.745 3.05e-07 ***
## allwheeldrive  3.429e+03  1.063e+03   3.227  0.00327 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2449 on 27 degrees of freedom
```

Regressions Involving Natural Logarithms I

Consider the log-linear model:

$$y_i = \beta_0 \cdot x_i^{\beta_1} \cdot \epsilon_i$$

Taking the natural logarithm on both sides

$$\ln(y_i) = \ln(\beta_0) + \beta_1 \cdot \ln(x_i) + \epsilon_i$$

You can choose which variables you want to transform using the natural log. You can transform just the dependent variable and/or all (or just some) of the independent variables. However, the interpretation of the β coefficients will change depending on your approach.

Regressions Involving Natural Logarithms II

Dep. Var.	Indep. Var.	Interpretation
y	x	$\Delta y = \beta \cdot \Delta x$
y	$\ln(x)$	$\Delta y = (\beta/100)\% \cdot \Delta x$
$\ln(y)$	x	$\%\Delta y = (100 \cdot \beta) \cdot \Delta x$
$\ln(y)$	$\ln(x)$	$\%\Delta y = \beta\% \cdot \Delta x$

For example, consider the following regression:

$$\ln(consumption) = \beta_0 + \beta_1 \cdot \ln(income)$$

Assume $\beta_1 = 0.8$: A 1 percent increase in income results in a $0.8 \cdot 1\% = 0.8\%$ increase in consumption.

Dummy Variables and Natural Logarithm I

Consider the following model:

$$\ln(y) = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot D + \epsilon$$

In this case, X is the continuous independent variable and D is the dummy variable. β_2 is interpreted as follows:

- If D switches from 0 to 1, the percent impact of D on Y is $100 \cdot (e^{\beta_2} - 1)$.
- If D switches from 1 to 0, the percent impact of D on Y is $100 \cdot (e^{\beta_2} - 1)$.

Dummy Variables and Natural Logarithm I

Interpretation when the Dependent Variable is $\ln(\cdot)\}$ Consider the following model:

$$\ln(y) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot D + \epsilon$$

In this case, the interpretation of β_1 is $e^{\beta} - 1$. So in the regression on the next slide, we have the coefficient for colonial which is 0.0538. Thus the feature “colonial” adds 5.53 percent to the value of the house.

Dummy Variables and Natural Logarithm III

```
bhat=lm(log(price)~log(lotsize)+log(sqrft)+bdrms+
    colonial,data=housing1)
summary(bhat)
```

```
## 
## Call:
## lm(formula = log(price) ~ log(lotsize) + log(sqrft) + bdrms + 
##     colonial, data = housing1)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.69479 -0.09750 -0.01619  0.09151  0.70228 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  -1.34959    0.65104  -2.073   0.0413 *  
## log(lotsize)  0.16782    0.03818   4.395 3.25e-05 ***
## log(sqrft)    0.70719    0.09280   7.620 3.69e-11 ***
## bdrms         0.02683    0.02872   0.934   0.3530    
## colonial      0.05380    0.04477   1.202   0.2330    
## 
```

Functional Form

Examples of Functional Forms

- Relation between consumption and income: Change in consumption due to extra income may decrease with income.
- Relationship between income and education: Change in income due to more education may decrease with more education

Consider the following relationships between y and x:

- $y = \beta_0 + \beta_1 x + \beta_2 x^2$
- $y = \beta_0 + \beta_1 x^{\beta_2}$

If a nonlinear relation can be expressed as a linear relation by redefining variables we can estimate that relation using ordinary least square.

Functional Form

Relationship 1:

- Linear in the regression coefficients, i.e. it can be expressed as a linear relation between y and independent variables x_1 and x_2: $x_1 = x$ and $x_2 = x^2$

Relationship 2:

- Taking the log of the dependent/independent variable can help making the model linear.

Squared/Quadratic Terms

Consider a model with x_2 included as a squared term:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2$$

Change in y due to a change in x:

$$\Delta \hat{y} \approx (\hat{\beta}_2 + 2 \cdot \hat{\beta}_2) \Delta x$$

Squared/Quadratic Terms in R: wage

```r
bhat=lm(income~educ+exper+I(exper^2),data=wage)
summary(bhat)
```

```
##
## Call:
## lm(formula = income ~ educ + exper + I(exper^2), data = wage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1134 -2.1056 -0.5476  1.2517 15.0251
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.0079730  0.7552203  -5.307 1.65e-07 ***
## educ         0.5992640  0.0532414  11.256  < 2e-16 ***
## exper        0.2686777  0.0370474   7.252 1.49e-12 ***
## I(exper^2)  -0.0046121  0.0008253  -5.588 3.70e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.179 on 522 degrees of freedom
## Multiple R-squared:  0.2696, Adjusted R-squared:  0.2654
## F-statistic: 64.23 on 3 and 522 DF,  p-value: < 2.2e-16
```

Squared/Quadratic Terms in R: hprice2

```
##
## Call:
## lm(formula = log(price) ~ log(nox) + log(dist) + rooms + I(rooms^2) +
##     stratio, data = hprice2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.04285 -0.12774  0.02038  0.12650  1.25272
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.385477   0.566473  23.630  < 2e-16 ***
## log(nox)    -0.901682   0.114687  -7.862 2.34e-14 ***
## log(dist)   -0.086781   0.043281  -2.005  0.04549 *
## rooms       -0.545113   0.165454  -3.295  0.00106 **
## I(rooms^2)   0.062261   0.012805   4.862 1.56e-06 ***
## stratio     -0.047590   0.005854  -8.129 3.42e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2592 on 500 degrees of freedom
## Multiple R-squared:  0.6028, Adjusted R-squared:  0.5988
## F-statistic: 151.8 on 5 and 500 DF,  p-value: < 2.2e-16
```

Interaction Effects

Interaction Effects: Overview

Assumptions so far:

- Change in an independent variable translates into variations of the dependent variable irrespective of the level of some other independent variable.

Interaction term: The impact of one independent variable depends on the level of another independent variable.

```
wage2$pareduc= wage2$meduc+wage2$feduc
bhat = lm(log(wage)~educ+educ:pareduc+exper+tenure,
          data=wage2)
```

Interaction Effects in R

```
##
## Call:
## lm(formula = log(wage) ~ educ + educ:pareduc + exper + tenure,
##     data = wage2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.85839 -0.23760  0.01424  0.25882  1.28750
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.6465188  0.1295593  43.582  < 2e-16 ***
## educ         0.0467522  0.0104767   4.462 9.41e-06 ***
## exper        0.0188710  0.0039429   4.786 2.07e-06 ***
## tenure       0.0102166  0.0029938   3.413 0.000679 ***
## educ:pareduc 0.0007750  0.0002107   3.677 0.000253 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3834 on 717 degrees of freedom
##   (213 observations deleted due to missingness)
## Multiple R-squared:  0.169,  Adjusted R-squared:  0.1643
## F-statistic: 36.44 on 4 and 717 DF,  p-value: < 2.2e-16
```