Basic Multivariate Regression

Jerome Dumortier

Lecture Overview

Extension of the bivariate model to a multivariate model

- One dependent and multiple independent variables

Topics covered besides the basic concepts

- Basic concepts of multivariate regression
- F-test
- Dummy variables
- Natural logarithm
- Functional forms (e.g., quadratic terms)
- Interaction terms

Introduction

Bivariate regression model (one independent and one dependent variable)

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

With the regression line determined by an intercept (β_0) and a slope (β_1), i.e.:

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

Multivariate linear regression model includes multiple independent variables

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

Ordinary least square (OLS) for bivariate and multivariate model

- Coefficients β_i such that sum of squared errors is minimal

Multivariate Regression Models

Purpose

- Measuring the effect of an independent variable on the dependent variable while including other independent variables to control for factors that may influence the dependent variable
- More technical language: Estimating the partial effect of an independent variable on the dependent variable while controlling for other relevant covariates

Example: Weekly grocery bill as a function of years of education

- Correlation between education and income, which affects grocery expenditures
- Household size affecting grocery spending and being correlated with age, marital status, and education
- Location: Food prices and education levels vary geographically
- Preferences (e.g., eating out vs. at home) vary with education

Estimated education effect captures those factors if excluded from OLS model

Life Expectancy: Setup of `wdi`

Using data for 2019 in `wdi` (World Development Indicators from the World Bank) to analyze life expectancy and how it is impacted by per capita gross domestic product (GDP) and literacy rate

- Dependent variable: *lifeexp* in years
- Independent variables: Income and education (i.e., *gdp* per capita, *litrate* in percent)

Hypothesis

- Increasing income and education levels lead to higher life expectancy

Sample size of 55 out of approximately 200 countries in total

- Likely systematic exclusion of low-income countries due to unavailable data

Life Expectancy: R Analysis

```r
df   = subset(wdi,year==2019,select=c("gdp","litrate","lifeexp"))
df   = na.omit(df)
bhat = lm(lifeexp~gdp+litrate,data=df)
summary(bhat)
```

```
##
## Call:
## lm(formula = lifeexp ~ gdp + litrate, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.9870  -1.5986   0.7002   3.2881   8.7156
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.603e+01  3.503e+00  13.141  < 2e-16 ***
## gdp         2.646e-04  8.033e-05   3.293  0.00178 **
## litrate     2.681e-01  4.280e-02   6.263 7.39e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.796 on 52 degrees of freedom
## Multiple R-squared:  0.5913, Adjusted R-squared:  0.5756
## F-statistic: 37.62 on 2 and 52 DF,  p-value: 7.884e-11
```

Life Expectancy: Interpretation

Overall observations

- Life expectancy positively associated with GDP per capita and literacy rate, which are both statistically significant (i.e., *p*-value below 10%)
- GDP per capita and literacy explain about 59% (R-squared) of the cross-country variation in life expectancy

GDP per capita

- Holding literacy constant, an increase of $10,000 in GDP per capita leads to $10,000 \cdot 0.0002646 = 2.646$ additional years of life expectancy

Literacy rate

- Holding GDP per capita constant, a 10 percentage point increase in literacy is associated with $10 \cdot 0.2681 = 2.681$ additional years of life expectancy

Overview

Overall explanatory power of the regression

- Joint significance of all slope coefficients with null hypothesis of all coefficients being equal to zero simultaneously
- Alternative hypothesis: At least one slope coefficient non-zero

Formula

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

with k and n being the number of all coefficients (including intercept) and number of observations, respectively

F-Test: Life Expectancy

Given the `wdi` data

- $n = 55$ and $k = 3$

Calculation

$$\frac{0.5912923/2}{(1-0.5912923)/(55-3)} = 37.6151521$$

Different and extending previous versions of hypothesis tests on individual coefficients

- F-test as a hypothesis test on all slope coefficients simultaneously

Overview

Representation of a single qualitative characteristics of an independent variable coded as 0 or 1. Examples:

- Male or female
- Presence or absence of hardwood floors in a house or all-wheel drive (AWD) in a car
- Home ownership
- Voting: Participation or candidate preference in a two-party system

One dummy variable less than categories

- One dummy variable for $\textit{hardwoodfloor} = 1$ with no hardwood floor being coded as 0
- Five religions (i.e., Christianity, Islam, Hinduism, Buddhism, and Judaism) requiring four dummy variables

Dummy Variable Example: All-Wheel Drive and `bmw`

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

```
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
## Multiple R-squared:  0.6287, Adjusted R-squared:  0.6012
## F-statistic: 22.86 on 2 and 27 DF,  p-value: 1.553e-06
```

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

```r
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
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1841 on 83 degrees of freedom
## Multiple R-squared:  0.6491, Adjusted R-squared:  0.6322
## F-statistic: 38.38 on 4 and 83 DF,  p-value: < 2.2e-16
```

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

```
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: 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
```