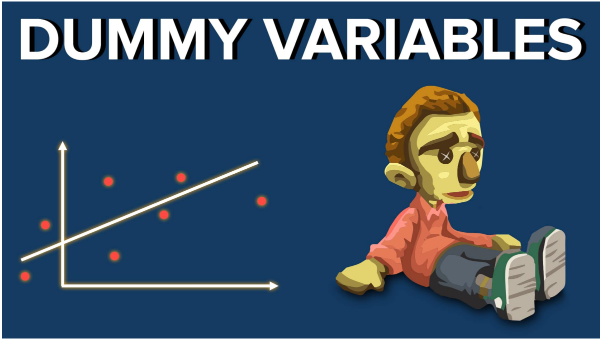
Fundamentals of Econometrics

Lecture 6: Multiple Regression with Qualitative Information



Section 1

Qualitative Information

Qualitative Information

- Some independent variables are qualitative in nature and are often difficult to include in a regression model.
- Example: gender, race, occupation, region, house type, rating grade, etc.
- A way to incorporate qualitative information is to use **dummy variables**.
- They may appear as the dependent variable or as independent variables.
- Dummy variables are binary variables that take on the value of 0 or 1.

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u$$

- δ_0 measures the difference in wages if the individual is a woman rather than a man, *ceteris paribus*.

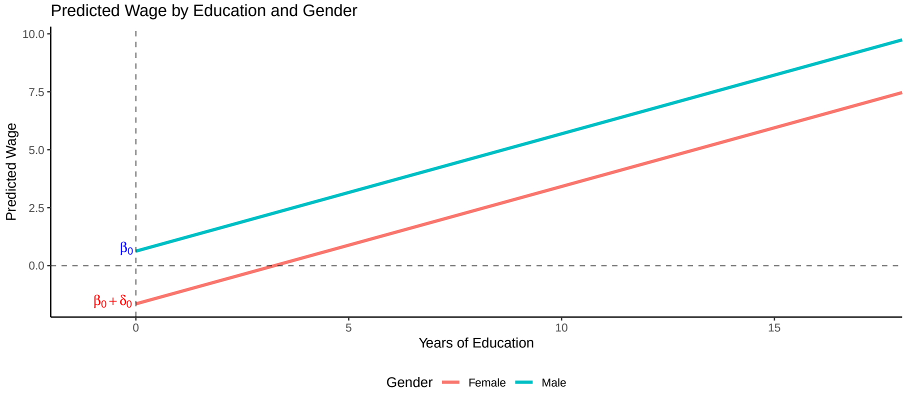
$$female = \begin{cases} 1, & \text{if the person is a woman} \\ 0, & \text{if the person is a man} \end{cases}$$

Section 2

Incorporating Dummies

Intercept Shifters

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.623     0.6725   0.926 3.55e-01
## educ           0.506     0.0504  10.051 7.56e-22
## female        -2.273     0.2790  -8.147 2.76e-15
```



Interpretating Dummies

An alternative interpretation of δ_0: This is the difference in the average wage between men and women, **with the same level of education**.

$$\delta_0 = E(\text{wage}|\text{female} = 1, \text{educ}) - E(\text{wage}|\text{female} = 0, \text{educ})$$

- **Dummy Variable Trap:** If we include a dummy variable for each category, we will have perfect multicollinearity.

$$wage = \beta_0 + \gamma_0 male + \delta_0 female + \beta_1 educ + u$$

- When using dummy variables, we must omit one category to avoid the dummy variable trap. - The omitted category is the **reference (or base) category**.

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u$$
$$wage = \beta_0 + \gamma_0 male + \beta_1 educ + u$$

Dummy Variables

- Alternatively, we could omit the intercept:

$$wage = \gamma_0 male + \delta_0 female + \beta_1 educ + u$$

Issues with this approach?

- Regression line is no longer forced to pass through the mean.
  - Not guaranteed: $\sum_{i=1}^{n} \hat{u}_i = 0$.
- More difficult to test for difference between parameters
- R^2 formula is invalid when the intercept is omitted.
- R^2 is inflated, unless $\bar{y}$ was truly zero.
  - Will have to run $\widehat{wage} \sim wage$ to get R^2.

```
(w1 <- lm(wage ~ female + educ + exper + tenure, wage1)) |> summary()
```

```
##
## Call:
## lm(formula = wage ~ female + educ + exper + tenure, data = wage1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.767 -1.808 -0.423  1.047 14.008
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.5679     0.7246   -2.16    0.031 *
## female       -1.8109     0.2648   -6.84  2.3e-11 ***
## educ          0.5715     0.0493   11.58  < 2e-16 ***
## exper         0.0254     0.0116    2.20    0.029 *
## tenure        0.1410     0.0212    6.66  6.8e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.96 on 521 degrees of freedom
## Multiple R-squared:  0.364,  Adjusted R-squared:  0.359
## F-statistic: 74.4 on 4 and 521 DF,  p-value: <2e-16
```

Does this mean women are discriminated against?

Difference in Means

```
(w.means <- lm(wage ~ female, wage1) )|> summary()
```

```
##
## Call:
## lm(formula = wage ~ female, data = wage1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.599 -1.849 -0.988  1.426 17.881
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.099      0.210   33.81   <2e-16 ***
## female        -2.512      0.303   -8.28    1e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.48 on 524 degrees of freedom
## Multiple R-squared:  0.116,  Adjusted R-squared:  0.114
## F-statistic: 68.5 on 1 and 524 DF,  p-value: 1.04e-15
```

```
w.means |> summary() |> coef()
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)     7.10      0.210   33.81 8.97e-134
## female         -2.51      0.303   -8.28  1.04e-15
```

Interpretation

- **Not holding other factors constant**, women earn \$2.512 less than men.
- This is the difference in the average wage for men and women.

Program Evaluation

Do training subsidies increase the hours of training per employee?

```
lm(hrsemp ~ grant + lsales + lemploy, jtrain, subset = (year==1988)) |> summary()
```

```
## 
## Call:
## lm(formula = hrsemp ~ grant + lsales + lemploy, data = jtrain, 
##     subset = (year == 1988))
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -36.87 -13.13  -3.64   4.77 119.62 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   46.665     43.412    1.07     0.28    
## grant         26.254      5.592    4.70  8.4e-06 ***
## lsales        -0.985      3.540   -0.28     0.78    
## lemploy       -6.070      3.883   -1.56     0.12    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 24.4 on 101 degrees of freedom
##   (52 observations deleted due to missingness)
## Multiple R-squared:  0.237,  Adjusted R-squared:  0.214 
## F-statistic: 10.4 on 3 and 101 DF,  p-value: 4.8e-06
```

Dummy Variables with log(y)

```
lm(lprice ~ llotsize + lsqrft + bdrms + colonial, hprice1) |> summary()
```

```
##
## Call:
## lm(formula = lprice ~ llotsize + lsqrft + bdrms + colonial, data = hprice1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6948 -0.0975 -0.0162  0.0915  0.7023
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.3496     0.6510   -2.07    0.041 *
## llotsize      0.1678     0.0382    4.40  3.2e-05 ***
## lsqrft        0.7072     0.0928    7.62  3.7e-11 ***
## bdrms         0.0268     0.0287    0.93    0.353
## colonial      0.0538     0.0448    1.20    0.233
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.184 on 83 degrees of freedom
## Multiple R-squared:  0.649,  Adjusted R-squared:  0.632
## F-statistic: 38.4 on 4 and 83 DF,  p-value: <2e-16
```

Dummies for multiple categories

- Define membership in each category with a dummy variable.
- Omit one category to avoid the dummy variable trap.

Example: Do single men earn more?

```
##
## Call:
## lm(formula = formula(lwage ~ marriedmale + marriedfemale + single
##     educ + exper + expersq + tenure + tenursq), data = full.wage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8970 -0.2406 -0.0269  0.2314  1.0920
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.321378   0.100009    3.21  0.00139 **
## marriedmale    0.212676   0.055357    3.84  0.00014 ***
## marriedfemale -0.198268   0.057835   -3.43  0.00066 ***
## singlefemale  -0.110350   0.055742   -1.98  0.04827 *
## educ           0.078910   0.006694   11.79  < 2e-16 ***
## exper          0.026801   0.005243    5.11  4.5e-07 ***
## expersq       -0.000535   0.000110   -4.85  1.7e-06 ***
## tenure         0.029088   0.006762    4.30  2.0e-05 ***
## tenursq       -0.000533   0.000231   -2.31  0.02153 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretations

- Holding other factors constant,
  - married women earn about 17.98% less than single men (our base category).
  - single women earn approximately 8.79% more than married women.
  - Not as straightforward to test if this difference is statistically significant, however.

```r
full.wage <- wage1 |> mutate(
  singlemale = ifelse(married==0 & female==0, 1, 0),
  singlefemale = ifelse(married==0 & female==1, 1, 0),
  marriedmale = ifelse(married==1 & female==0, 1, 0),
  marriedfemale = ifelse(married==1 & female==1, 1, 0))

lm(formula(lwage ~ marriedmale + marriedfemale + singlefemale +
             educ + exper + expersq + tenure + tenursq),
   data = full.wage) |> summary()
```

Incorporating Ordinal Information using Dummies

- Assume we are interested in how credit rating affects municipal bond interest rates.

  $$MBR = \beta_0 + \beta_1 CR + \text{other factors}$$

  where CR is an ordinal variable with values $\{0 = \text{worst}, 1, \ldots, 4 = \text{best}\}$.
- **Inappropriate model:** This specification assumes that the effect of a one-unit increase in $RATING$ is the same regardless of the starting point.
- **Instead:** We can use dummy variables to allow for different effects of each rating level.

  $$MBR = \beta_0 + \beta_1 CR_1 + \beta_2 CR_2 + \beta_3 CR_3 + \beta_4 CR_4 + \text{other factors}$$

What is the base category?

Interacting Dummy Variables: Slope Shifter

Example:

$$\widehat{log(wage)} = \beta_0 + \delta_0 female + \delta_1 married + \delta_2 \boxed{female \cdot married} + \dots$$

```
lm(lwage ~ female + married + female*married + educ + exper +
    expersq + tenure + tenursq, wage1) |> summary() |> coef()
```

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.321378   0.100009    3.21 1.39e-03
## female         -0.110350   0.055742   -1.98 4.83e-02
## married         0.212676   0.055357    3.84 1.37e-04
## educ            0.078910   0.006694   11.79 1.43e-28
## exper           0.026801   0.005243    5.11 4.50e-07
## expersq        -0.000535   0.000110   -4.85 1.66e-06
## tenure          0.029088   0.006762    4.30 2.03e-05
## tenursq        -0.000533   0.000231   -2.31 2.15e-02
## female:married -0.300593   0.071767   -4.19 3.30e-05
```

Interacting Dummy Variables: Slope Shifter

- Interacting a dummy with a **continuous variable** allows the slope to differ by category.

$$log(wage) = \beta_0 + \delta_0 female + \beta_1 educ + \delta_1 \boxed{female \cdot educ} + u$$

- β_0: intercept for men; β_1: slope on education for males
- $\beta_0 + \delta_0$: intercept for women; $\beta_1 + \delta_1$: slope on education

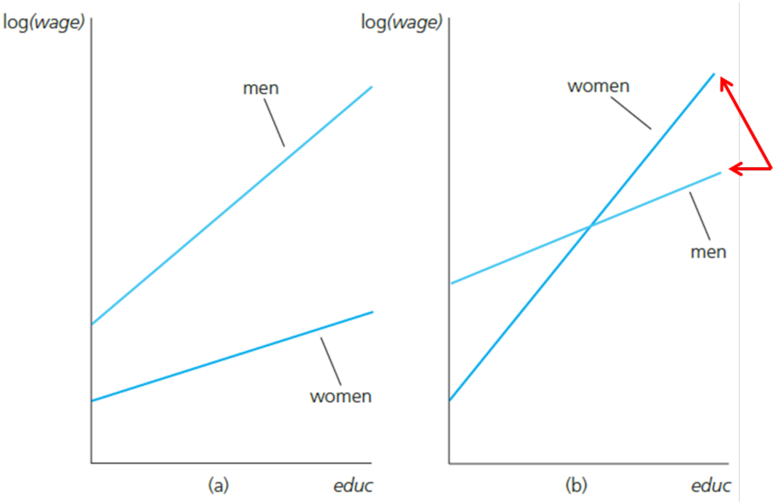
Interesting Hypotheses

- Returns to education is the same for men and women.

$$H_0 : \delta_1 = 0 \quad \text{vs.} \quad H_1 : \delta_1 \neq 0$$

- The whole wage equation is the same for men and women.

$$H_0 : \delta_0 = 0, \delta_1 = 0 \quad \text{vs.} \quad H_1 : \text{at least one} \neq 0$$


log(wage)
men
women
(a)
educ
log(wage)
women
men
(b)
educ

Testing for Differences between Groups

- Unrestricted Model:

$$\begin{aligned} cumgpa = \beta_0 + \delta_0 female + \beta_1 sat + \delta_1 female \cdot sat + \\ \beta_2 hsperc + \delta_2 female \cdot hsperc + \beta_3 tothrs + \\ \delta_3 female \cdot tothrs + u \end{aligned}$$

- Restricted Model:

$$cumgpa = \beta_0 + \beta_1 sat + \beta_2 hsperc + \beta_3 tothrs + u$$

- Null Hypotheses (all interaction effects are zero, the same regression applies to both men and women):

$$H_0 : \delta_0 = \delta_1 = \delta_2 = \delta_3 = 0$$

Table 1: Testing for Differences

	Dependent variable:	
	cumgpa	
	Unrestricted	Restricted
	(1)	(2)
female	−0.353	
	(0.411)	
sat	0.001^{***}	0.001^{***}
	(0.0002)	(0.0002)
hsperc	-0.008^{***}	-0.010^{***}
	(0.001)	(0.001)
tothrs	0.002^{***}	0.002^{***}
	(0.001)	(0.001)
female:sat	0.001^{*}	
	(0.0004)	
female:hsperc	−0.001	
	(0.003)	
female:tothrs	−0.0001	
	(0.002)	
Constant	1.480^{***}	1.490^{***}
	(0.207)	(0.184)
RSS	78.35	85.52
Observations	366	366
R^2	0.406	0.352
Adjusted R^2	0.394	0.346

```r
unrest <- lm(cumgpa~ female + sat + female*sat + hsperc +
               female*hsperc + tothrs + female*tothrs,
             data = gpa3, subset = (term==2))
rest <- lm(cumgpa~ sat + hsperc + tothrs, data = gpa3,
           subset = (term==2))

## Add RSS
RSS <- list(c("RSS",
              sprintf("%0.2f", sum(unrest$residuals^2)),
              sprintf("%0.2f", sum(rest$residuals^2))))
stargazer(unrest, rest, type = "latex",
          title = "Testing for Differences",
          column.labels = c("Unrestricted", "Restricted"),
          font.size = "footnotesize", add.lines = RSS)
```

- You can now manually compute your F test, as we did in Chapter 4 (Fstat = 8.18).
- Or use the `linearHypothesis` function from the `car` package.

```
car::linearHypothesis(unrest,
            c("female = 0", "female:sat = 0",
              "female:hsperc = 0","female:tothrs = 0"))
```

```
##
## Linear hypothesis test:
## female = 0
## female:sat = 0
## female:hsperc = 0
## female:tothrs = 0
##
## Model 1: restricted model
## Model 2: cumgpa ~ female + sat + female * sat + hsperc + fe
##     tothrs + female * tothrs
##
##   Res.Df  RSS Df Sum of Sq    F  Pr(>F)
## 1    362 85.5
## 2    358 78.4  4      7.16 8.18 2.5e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

```r
unrest <- lm(cumgpa~ female + sat + female*sat + hsperc +
               female*hsperc + tothrs + female*tothrs,
             data = gpa3, subset = (term==2))
rest <- lm(cumgpa~ sat + hsperc + tothrs, data = gpa3,
           subset = (term==2))

## Add RSS
RSS <- list(c("RSS",
              sprintf("%.2f", sum(unrest$residuals^2)),
              sprintf("%.2f", sum(rest$residuals^2))))
stargazer(unrest, rest, type = "text",
          title = "Testing for Differences",
          column.labels = c("Unrestricted", "Restricted"),
          add.lines = RSS)
```

Chow test

An alternative way to compute the F-statistic is to use the Chow test.

1. Run separate regressions for men and for women and collect respective SSRs (SSR_m, SSR_f).
    - Now $SSR_m + SSR_f = SSR_{ur}$.
2. Run restricted model and collect SSR (SSR_p). This is our pooled SSR.
3. Compute the Chow statistic:

$$F = \frac{(SSR_p - (SSR_m + SSR_f))/(k+1)}{(SSR_m + SSR_f)/(n - 2(k+1))}$$

   where k is the number of restrictions and n is the number of observations.

Important: The Chow test assumes that the error variances are equal across groups.

Chow test

```r
# Separate regressions
m <- lm(cumgpa ~ sat + hsperc + tothrs, data = gpa3, subset=(term==2 & female==0))
f <- lm(cumgpa ~ sat + hsperc + tothrs, data = gpa3, subset=(term==2 & female==1))
RSS_m <- sum(m$residuals^2)
RSS_f <- sum(f$residuals^2)

chow_num <- (sum(rest$residuals^2) - (RSS_m + RSS_f))/4
chow_denom <- (RSS_m + RSS_f)/(nrow(subset(gpa3,term==2)) - 2*(3 + 1))
Fstat_chow <- chow_num/chow_denom
cat("Chow F-statistic: ", Fstat_chow, "\n",
    "Critical value at 5%: ", qf(0.95, 4, 358))
```

```
## Chow F-statistic:  8.18
##  Critical value at 5%:  2.4
```

Chow test

- Again the Chow statistic is only valid under homoskedasticity.
- Another use of the Chow test is to determine whether parameter estimates vary over time. If we have time series data then the Chow test can be used to determine if there is a **structural break**.