1)

6. Multiple Regression Analysis (Inference) 67

4 Testing Hypotheses about a Single Linear Combination of the Parameters

Consider

$$\log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 \exp er + u$$

where jc = number of years attending a two-year college
$univ$ = number of years at a four-year college
$\exp er$ = months in the workforce.
We want to test whether $\beta_1 = \beta_2$.

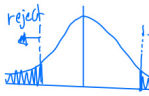
(we want to test) if the returns from 1 more year of education at a junior college is the same as that of the university.

$H_0 : \beta_1 = \beta_2 \Rightarrow H_0 : \beta_1 - \beta_2 = 0$

against

$H_a : \beta_1 \neq \beta_2 \Rightarrow H_a : \beta_1 - \beta_2 \neq 0$

2-tailed test



area = $\frac{\text{significant level}}{2}$

$$t = \frac{(\hat{\beta}_1 - \hat{\beta}_2) = 0}{s.e.(\hat{\beta}_1 - \hat{\beta}_2)}$$

We compute this t-statistic and compare with the critical value

where $s.e.(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{Var(\hat{\beta}_1 - \hat{\beta}_2)}$

$$= \sqrt{Var(\hat{\beta}_1) + Var(\hat{\beta}_2) - 2Cov(\hat{\beta}_1, \hat{\beta}_2)}$$

not very straightforward to calculate

⇒ We use a variable transformation trick ⇒ see notes!!!

2)

Inference ⇝ Hypothesis testing about β the true parameter.

$wage = \beta_0 + \beta_1 educ + \beta_2 experience + \ldots + u$

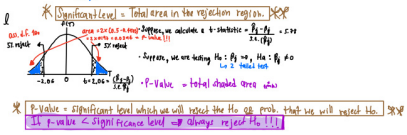
We want to test hypothesis about the true impact (β) of each X variables (educ, experience) on the independent variable(Y)

BUT, We don't know what the true β are. So, we use $\hat{\beta}$ (estimator) and s.e.($\hat{\beta}$) to test the hypothesis

1) Test if β = some number
eg: $\beta_j = 0 \rightarrow X_j$ has no impact on Y.
$\beta_j = 1 \rightarrow$ 1 unit ↑ in X_j correspond to 1 unit ↑ in Y.

⇒ t-test ⇒ How ??

$\frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)} \sim t_{d.f.}$



* [Significant level] = Total area in the rejection region. *

* p-value = significant level which we will reject the H_0 at prob. that we will reject H_0.
If p-value < Significance level ⇒ always reject H_0 !!!

3)

68 6. Multiple Regression Analysis (Inference)

another possible hypothesis test (one-tailed alternative)

$H_0 : \beta_1 = \beta_2 \Rightarrow H_0 : \beta_1 - \beta_2 = 0$

$H_a : \beta_1 < \beta_1 \Rightarrow H_a : \beta_1 - \beta_2 < 0$

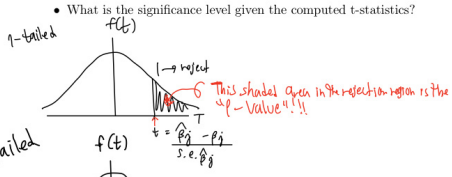
- It is assumed that β_1 would not be more than β_2 (returns to a 2-year college would never be more than returns to University education)

$$t = \frac{(\hat{\beta}_1 - \hat{\beta}_2) - 0}{s.e.(\hat{\beta}_1 - \hat{\beta}_2)}$$

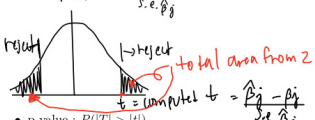
* Then, goto the extra note

5 Computing p-Values for t-Tests

- What is the significance level given the computed t-statistics?



2-tailed



- p-value : $P(|T| > |t|)$

T = t-distributed random variable with d.f. = n−k−1
t = computed t-statistic.
⇝ p-value = probability that a random T value will be greater (in the | | term) than our T in the H. test)

4)

In-class exercise

Consider the multiple regression model, assume MLR1-6 are satisfied.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

You would like to test $H_0 : \beta_1 - 3\beta_2 = 1$
H_a : otherwise is true

1st) write the t-statistic for testing H_0

$$t = \frac{(\hat{\beta}_1 - 3\hat{\beta}_2) - 1}{s.e.(\hat{\beta}_1 - 3\hat{\beta}_2)}$$

Trick 2nd) Define $\theta_1 = \beta_1 - 3\beta_2 \Rightarrow H_0 : \theta_1 = 1$, $H_a : \theta_1 \neq 1$

$$t = \frac{\hat{\theta}_1 - 1}{s.e.(\hat{\theta}_1)}$$

⇒ We need our regression to have θ_1 in it. So, STATA or OLS estimation will automatically give $\hat{\theta}_1$ & s.e. $\hat{\theta}_1$

Now, $\beta_1 = \hat{\theta}_1 + 3\hat{\beta}_2$

$\beta_1 = \theta_1 + 3\beta_2$

Substitute in the main regression & get

$$Y = \beta_0 + (\theta_1 + 3\beta_2) X_1 + \beta_2 X_2 + \beta_3 X_3 + u$$
$$= \beta_0 + \theta_1 X_1 + 3\beta_2 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$$
$$= \beta_0 + \theta_1 X_1 + \beta_2 (X_2 + 3X_1) + \beta_3 X_3 + u$$

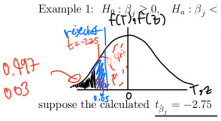
* Now, the explanatory variables are going to be X_1, $X_2 + 3X_1$ & X_3

- We can calculate $t = \frac{\hat{\theta}_1 - 1}{s.e.(\hat{\theta}_1)}$

5)

6. Multiple Regression Analysis (Inference) 69

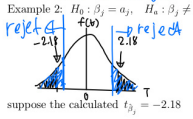
Example 1: $H_0 : \beta_j \geq 0$, $H_a : \beta_j < 0$, d.f.= 140.



suppose the calculated $t_{\hat{\beta}_j} = -2.75$

- From the z-table, the value -2.75 corresponds to area = 0.003
- Thus, p-value = 0.003
- Would we reject H_0 if we use the significance level = 5%? Yes.

Example 2: $H_0 : \beta_j = a_j$, $H_a : \beta_j \neq a_j$, d.f.= 18.



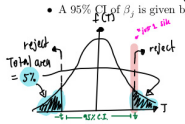
suppose the calculated $t_{\hat{\beta}_j} = -2.18$

- From the t-table, the value -2.18 corresponds to area = 0.02 to 0.05
- Thus, p-value = is between 0.02 - 0.05
- Would we reject H_0 if we use the significance level = 5%?

Yes, reject Ho bcz the area is less than 0.05 or p-value<0.05

6 Confidence Intervals (CI)

- Confidence Intervales for the POPULATION PARAMETER (β_j) → The range of values that would capture the true β_j at a 95% chance
- A 95% CI of β_j is given by



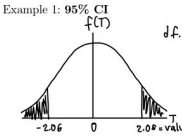
$CI = \hat{\beta}_j \pm c \times s.e.(\hat{\beta}_j)$

c is the 97.5 percentile in the t-distribution with n-k-1 d.f.

6)

70 6. Multiple Regression Analysis (Inference)

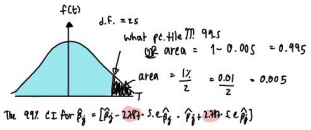
Example 1: 95% CI



2.06 = value of 97.5% percentile in the t_{25} distribution

The 95% C.I. for $\hat{\beta}_j$ = $[\hat{\beta}_j - 2.06 \cdot s.e.(\hat{\beta}_j + 2.06 \cdot s.e.(\hat{\beta}_j)]$

Example 2: 99% CI



The 99% CI for $\hat{\beta}_j$ = $[\hat{\beta}_j - 2.787 \cdot s.e.\hat{\beta}_j \cdot \hat{\beta}_j + 2.787 \cdot s.e.\hat{\beta}_j]$

1 Mar

F-test motivation

→ We want to test the significance of the group of hypotheses (Multiple Hypotheses)

Grade 325 = $\beta_0 + \beta_1$ # times_front + β_2 # times_back + β_3 hr_study + β_4 past_GPA + β_5 gender + u

Ho : seat position doesn't have impact on GPA
$\beta_1 = 0$ & $\beta_2 = 0$ ⇒ $\beta_1 = \beta_2 = 0$

Ha: seat position matters
$\beta_1 \neq 0$ & $\beta_2 \neq 0$
OR $\beta_1 \neq 0$ & $\beta_2 = 0$
OR $\beta_1 = 0$ & $\beta_2 \neq 0$
} at least one of the $\beta_1, \beta_2 \neq 0$

7)

6. Multiple Regression Analysis (Inference) 71

7 Testing Multiple Linear Restrictions: The F-test

Suppose the model is specified by

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

H_0 : $\beta_3 = 0$ and $\beta_2 = 0$ → want to test if x_1 & x_2 both have no impact on Y

Ha_1 H_1 : H_0 is not true

We can use the F-test to test this type of "multiple hypotheses".

1. Our full model is called the "unrestricted" model (ur). Suppose it can be expressed as: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$ is true ⇒ reject H_0

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + u$$

2. The model which takes out x (which we think its associated $\beta = 0$) is called the restricted model (r). small model

$Y = \beta_0 + \beta_1 x_1 + u$ is true ⇒ do not reject H_0

* Suppose there are "q" no. of β that we would like to perform a joint-test of =0
eg. in this model $q = 2$

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{k-q}$

$H_0 : \beta_{k-q+1} = \beta_{k-q+2} = \ldots = \beta_k = 0$ (2 restrictions)
(the last q $\beta_s = 0$)
$Ha : H_0$ is not true.

$$y = \underbrace{\beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots + \beta_{k-q} X_{k-q}}_{(r)} + \beta_{k-q+1} X_{k-q+1} + \beta_{k-q+2} X_{k-q+2} \ldots + \beta_k X_k + u$$

ur (unrestricted model)

$$F \equiv \frac{(SSR_r - SSR_{ur})}{q} \Big/ \frac{SSR_{ur}}{(n-k-1)}$$

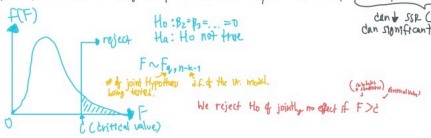
This is always (+) bcz $SSR_{ur} < SSR_r$. Every time you add a 1 more X_q the model will be better explained

d.f. of the "ur" model.

8)

- So, if every time you add 1 more x variable, the SSR ↓ and R^2 ↑, why don't we just keep the additional X in the model ??

Ans: Because everytime we add 1 more X_j, Var($\hat{\beta}_3$) will increase, making the prediction of β less precise. So, we only keep the additional X_j if it/they can improve the model enough.



can ↓ SSR ($\uparrow R^2$) enough.
can significantly ↓ SSR & ↑ R^2

9)

3. Some useful facts

① $R^2_{ur} > R^2_r$ because any additional X would increase R^2 (improve fit) $\Rightarrow SSR_{ur} < SSR_r$

② By including more Xs the model is certainly better explained. However, we would like to reject H_0 if the inclusion of extra variables does not improve the model enough.

4. Other ways to calculate the F-statistics:

$\Rightarrow$ From $R^2 = 1 - \frac{SSR}{SST}$

(Now) we have $F = \frac{(R^2_{ur} - R^2_r)/q}{(1 - R^2_{ur})/(n-k-1)}$

Example: Suppose we are interested in understanding the determinant of a baseball player's salary.

salary	= season salary
years	= years in major leagues
gamesyr	= games per year in the league
bavg	= career batting average
hrunsyr	= homeruns per year
rbisyr	= runs batted in per year

If we want to test whether performance has any impact on salary.
$H_0: \beta_{bavg} = \beta_{hrunsyr} = \beta_{rbisyr} = 0$
H_a: Otherwise is true

- the unrestricted model (ur) is defined by

10)

(unrestricted) ur model

```
. regress log_salary years gamesyr bavg hrunsyr rbisyr

      Source |       SS       df       MS              Number of obs =     353
-------------+------------------------------           F(  5,   347) =  117.06
       Model |  308.989208     5  61.7978416           Prob > F      =  0.0000
    Residual |  183.186327   347  .527914487           R-squared     =  0.6278
-------------+------------------------------           Adj R-squared =  0.6224
       Total |  492.175535   352  1.39822595           Root MSE      =  .72658

------------------------------------------------------------------------------
  log_salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       years |   .0688626   .0121145     5.68   0.000     .0450355    .0926898
     gamesyr |   .0125521   .0026468     4.74   0.000     .0073464    .0177578
        bavg |   .0009786   .0011035     0.89   0.376    -.0011918     .003149
     hrunsyr |   .0144295    .016057     0.90   0.369    -.0171518    .0460107
      rbisyr |   .0107657    .007175     1.50   0.134    -.0033462    .0248776
       _cons |   11.19242   .2888229    38.75   0.000     10.62435    11.76048
------------------------------------------------------------------------------
```

intercept ($\therefore k = 5$) (+1 intercept)

- the restricted model (r) is defined by

```
. regress log_salary years gamesyr

      Source |       SS       df       MS              Number of obs =     353
-------------+------------------------------           F(  2,   350) =  259.32
       Model |  293.864058     2  146.932029           Prob > F      =  0.0000
    Residual |  198.311477   350  .566604221           R-squared     =  0.5971
-------------+------------------------------           Adj R-squared =  0.5948
       Total |  492.175535   352  1.39822595           Root MSE      =  .75273

------------------------------------------------------------------------------
  log_salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       years |    .071318    .012505     5.70   0.000     .0467236    .0959124
     gamesyr |   .0201745   .0013429    15.02   0.000     .0175334    .0228156
       _cons |    11.2238    .108312   103.62   0.000     11.01078    11.43683
------------------------------------------------------------------------------
```

SSE / SSR / SST

bigger model → less R^2 → lower S.D of the 2 common explanatory variables

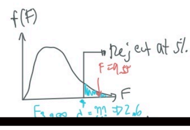
* But, when performing an F-test, performances have joint impact !!!

this: $F = \frac{(R^2/q)}{(1-R^2)/(n-k-1)} = ??$

Now, our H_0 and H_a becomes

$$F \equiv \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$$

$$= \frac{(198.311 - 183.186)/3}{(183.86)/(353-5-1)} \approx 9.55$$



Let's use 5% level of sig.
Since $F = 9.55 > 2.6$, we reject H_0 at 5% level & conclude that performances have joint effects on salary.

11)

8 How the Hypothesis Testing is done in Practice

1. Check the values of $t-statistic$ reported by the statistical software (i.e. STATA, SPSS, SAS)

   $\Rightarrow$ These $t-statistics$ are to test $H_0: \beta_i = 0$ — z-table with 5% sig. level

   $\Rightarrow$ If the d.f. > 30, then when $|t| > 1.96$, we can reject H_0

   $\Rightarrow$ **When** $|t| > \mathbf{1.96}$, we can say that β_i is **statistically significant** at 5% level. (value of $\beta_i \neq 0$)

   $\Rightarrow$ **When** $|t| < \mathbf{1.96}$ we can say that β_i is **not statistically significant** at 5% level.

   $\Rightarrow$ If $|t| < 1.96$ we can drop x_i from the model

   $\Rightarrow$ After we drop x_i, we estimate the new regression function and obtain a new set of $\hat{\beta}$.

2. We can also perform other hypothesis testings of interest.

   e.g. $H_0: \beta_i = \beta_j$

   or $H_0: \beta_i = 5$ etc.

   or perform an F-test for testing multiple linear restrictions

3. Usually, in economics, the estimation results are reported using this form

Dependent Variable: log(*salary*)			
Independent Variables	**(1)**	**(2)**	**(3)**
log(*sales*)	.224 (.027)	.158 (.040)	.188 (.040)
log(*mktval*)	——	.112 (.050)	.100 (.049)
profmarg	——	−.0023 (.0022)	−.0022 (.0021)
ceoten	——	——	.0171 (.0055)
comten	——	——	−.0092 (.0033)
intercept	4.94 (0.20)	4.62 0.25)	4.57 (0.25)
Observations	177	177	177
R-squared	.281	.304	.353

sales; other company performance; CEO characteristics

like a simple regression with 1X

12)

This is page
Printer: Opa

Multiple Regression Analysis : Further Issues

1 Data scaling on OLS statistics

e.g. currency, measurement & how the estimated parameter change?

When we change the unit of measurment of a variable, the value of estimators would chance accordingly. For example

$$\widehat{bweght} = \hat{\beta}_0 + \hat{\beta}_1 cigs + \hat{\beta}_2 fa\min c,$$

where
$bwght$ = child birth weight, in grams.
$cigs$ = number of cigarettes smoked by the mother while pregnant, per day.
$fa\min c$ = annual family income, in thousands of dollars.

- What if we use bweight in kilograms???
  1 kg = 1,000 g
  $\widehat{bweight}_{kg} = \frac{\widehat{bweight}_g}{1{,}000} = \frac{\hat{\beta}_0}{1000} + \frac{\hat{\beta}_1}{1000} cigs + \frac{\hat{\beta}_2}{1000} faminc$
  $= \hat{\alpha}_0 + \hat{\alpha}_1 cigs + \hat{\alpha}_2 faminc$
  $\Rightarrow \hat{\alpha}_0 = \frac{\hat{\beta}_0}{1000}, \hat{\alpha}_1 = \frac{\hat{\beta}_1}{1000}, \hat{\alpha}_2 = \frac{\hat{\beta}_2}{1000}$

- What if we use faminc in USD (instead of 1000 USD)
  $bweight_g = \hat{\beta}_0 + \hat{\beta}_1 cigs + \hat{\beta}_2 faminc_{USD}$
  $= \hat{\beta}_0 + \hat{\beta}_1 cigs + \hat{\theta}_2 faminc_{USD}$ (the value of this variable is going to be 1,000 times larger than faminc)
  $\Rightarrow \hat{\theta}_2 = \frac{\hat{\beta}_2}{1000}$
  in other words $\hat{\theta}_2$ = impact of 1 USD in income.
  $\hat{\beta}_2$ = 1,000 USD

- What if we use bweight in kg & income in THB
  $bweight_{kg} = \frac{\hat{\beta}_0}{1000} + \frac{\hat{\beta}_1}{1000} cigs + \left(\frac{\hat{\beta}_2}{1000}\right) faminc_{THB}$
  This value is going to be 3,000 times more than faminc.

13)

76 7. Multiple Regression Analysis : Further Issues

2 More on functional forms

- Logarithmic Functional Form

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

- $\beta_1 = \frac{d\log(y)}{d\log(x)} = \frac{\frac{1}{y}dy}{\frac{1}{x_1}dx_1} = \frac{\frac{1}{y}\Delta y}{\frac{1}{x_1}\Delta x_1} = \frac{100 \times \frac{1}{y}\Delta y}{100 \times \frac{1}{x_1}\Delta x_1} = \frac{\%\Delta y}{\%\Delta x}$

- $\beta_2 = \frac{d\log(y)}{dx_2} = \frac{\frac{1}{y}dy}{dx_2} = \frac{\frac{1}{y}\Delta y}{\Delta x_2}$

$\Rightarrow$ if we want the upper term to be % change, then

$100\beta_2 = \frac{100\frac{1}{y}\Delta y}{\Delta x_2}$

$100\beta_2 = \frac{\%\Delta y}{\Delta x_2}$

$\therefore 100\beta_2 = \%\Delta$ in y given that x_2 increases by 1 units.

- Models with Quadratics (squares)

$\Rightarrow$ capture increasing/decreasing marginal effects (slope of the relationship btw x & y is not constant.

$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$

$\frac{dy}{dx} = \beta_1 + 2\beta_2 x$

Example : Effects of Pullution on Housing Prices

$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2 \log(dist) + \beta_3 rooms + \beta_4 room^2 + \beta_5 stratio + u$

14)

7. Multiple Regression Analysis : Further Issues 77

where
price = housing price
nox = level of pollution
dist = distance from downtown
rooms = number of rooms
stratio = average student per teacher ratio
The estimation result is given by

In the US or many other countries, students can apply to schools in the area without having to take any test. So, the lower st-ratio, the better the school

```
regress lprice lnox dist rooms rooms_sq stratio
```

Source	SS	df	MS		
Model	51.4933152	5	10.298663	Number of obs	= 506
Residual	33.0889098	500	.06617782	F(5, 500)	= 155.62
				Prob > F	= 0.0000
Total	84.582225	505	.167489554	R-squared	= 0.6088
				Adj R-squared	= 0.6049
				Root MSE	= .25725

lprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnox	-.9767545	.0995938	-9.81	0.000	-1.172429	-.7810806
dist	-.0321972	.0094013	-3.42	0.001	-.050668	-.0137264
rooms	-.5528032	.1612965	-3.43	0.001	-.8697056	-.2359007
rooms_sq	.0624697	.0124867	5.00	0.000	.0379368	.0870025
stratio	-.0486667	.0058131	-8.37	0.000	-.0600879	-.0372455
_cons	13.59154	.5650901	24.05	0.000	12.4813	14.70178

Log(price) ← lprice
Log(nox) ← lnox

$H_0: \beta_1 = 0$ $H_1: \beta_1 \neq 0$ all < 0.05

$\Rightarrow$ all variable are significant

Consider the effect of "room"

$\frac{d\log(price)}{drooms} = \beta_3 + 2\beta_4 rooms = -0.553 + 2(0.062) \cdot rooms$

log(price) vs rooms

★ at how many rooms does 1 additional room has a positive impact on log(price)??

$0 = -0.553 + 2(0.062) \cdot rooms$

$rooms = 4.4$

Answer $\Rightarrow$ at 4.4 rooms or more

at $\rightarrow$ 5 rooms or more

What would be the % change in price when the number of room increases from 5 to 6?

- $\frac{d\log(price)}{drooms} = -0.553 + 2(0.062) \cdot rooms$

$100 \cdot \frac{1}{price} dprice = 100(-0.553 + 2(0.062) \cdot 5)$

$= 100 \times 0.067 = 6.7\%$ increase.

what about % in price when # rooms increases from 5 to 7 ???

$\%\Delta price = 100(-0.553 + 2(0.062) \cdot 6) = 19.1\%$

15)

78 7. Multiple Regression Analysis : Further Issues

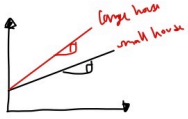
3 Models with Interaction Terms

Consider

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + \beta_3 sqrft \times bdrms + \beta_4 bthrms + u$$

(x_1, x_2, $x_3 = x_1 \cdot x_2$, x_4)

where
price = housing price
sqrft = house size (square feet)
bdrms = number of bedrooms
bthrms = number of bathrooms



- $\frac{\partial price}{\partial bdrms} = \beta_2 + \beta_3 sqrft$

$\Rightarrow$ If $\beta_2 > 0$ then, an additional bedroom would increase price more for a larger house!

16)

7. Multiple Regression Analysis : Further Issues 79

4 More on the Goodness-of-Fit and Selection of Regressors

- Adding more regressors ALWAYS improve fit $\rightarrow R^2$ always ↑
- But we lose the "degree of freedom" (d.f. = free data point used to estimate the parameter).
  ↝ 1 data point is sacrificed every time we estimate a parameter.
- Using R^2 would not punish "having too many regressors".
- We use adjusted-R^2 or R^{-2} when we want to punish adding too many regressors.

$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{SSR/n}{SST/n}$

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

If we have more k, d.f. $= n-k-1$ ↓, $SSR/(n-k-1)$ ↑,
adj.-R^2 ↓
(∴ adj-R^2 ↓ only we increase additional no. of k) ≈ depend on

Using adjusted R-squared to choose between non-nested models (one model is not a subset of another).

Consider Model 1

$$\widehat{salary} = 830.63 + 0.0163 sales + 19.63 roe$$
$$(223.90) \quad (0.0089) \quad (11.08)$$
$$n = 209, \ R^2 = 0.029, \ \bar{R}^2 = 0.020$$

Consider Model 2

$$\widehat{\log(salary)} = 4.36 + 0.2751 \log(sales) + 0.0179 roe$$
$$(0.29) \quad (0.033) \quad (0.004)$$
$$n = 209, \ R^2 = 0.282, \ \bar{R}^2 = 0.275$$

27.5% of variation in Y is explained
So, this model is better!!

(17)

This is page 1
Printer: Opaq

Multiple Regression Analysis with Qualitative Information:

info that is not quantitative e.g. gender, season, months

1 Outline

- Describing qualitative information
- Using a single dummy independent variable
- Using dummy variables for multiple categories (qual. info.)
- Interactions involving dummy variables
- A binary dependent variable (Y variable): The linear probability model
  ↳ what will happen if Y variable is also qualitative.

2 Describing Qualitative Information

- "Female" and "Married" are qualitative variable.
- We arbitarily assign a dummy variable to decribe them.

dummy variables

$$famale = \begin{cases} 1 & \text{if female} \\ 0 & \text{otherwise (or if male)} \end{cases}$$

$$married = \begin{cases} 1 & \text{if married} \\ 0 & \text{otherwise (of if single)} \end{cases}$$

TABLE 7.1

A Partial Listing of the Data in WAGE1.RAW

person	wage	educ	exper	female	married
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

(not married; otherwise; otherwise)

(18)

3 Models with a single dummy independent variable

(There are MLR with the pop. reg. fn. – 2 variables explaining wage)

Consider

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u \quad (1)$$

where

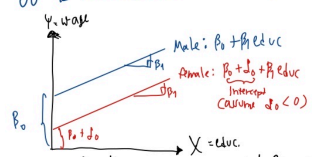
$$famale = \begin{cases} 1 & \text{if female} \\ 0 & \text{otherwise (or if male)} \end{cases}$$

In this case, the δ_0 (delta) notation is used to highlight the interpretation of the parameters multiplying dummy variables. In other cases, we can use any notation that is the most convenient.

① $E(wage|female, educ) = E(\beta_0 + \delta_0 female + \beta_1 educ + u | female, educ)$
$= \beta_0 + \delta_0 female + \beta_1 educ + E(u|female, educ)$
$= \beta_0 + \delta_0 female + \beta_1 educ$ ($= 0$, assume 1–4 hold)

② Thus
♀ : $E(wage|female=1, educ) = \beta_0 + \delta_0(1) + \beta_1 educ = \beta_0 + \delta_0 + \beta_1 educ$
♂ : $E(wage|female=0, educ) = \beta_0 + \delta_0(0) + \beta_1 educ = \beta_0 + \beta_1 educ$
$\delta_0 = E(wage|female=1, educ) - E(wage|female=0, educ)$
or $\delta_0 = E(wage|female, educ) - E(wage|male, educ)$

* given the same value of educ (same education level), δ_0 is the difference in the expected wage of females & males.



→ By the way, we model the regression fn., "female" is going to give a constant impact on wage, regardless of the level of educ.

(19)

4 It is not possible to include all of the dummy alternatives in the same model (as long as there is an intercept in the model)

If we include all alternatives of a dummy variable in the same model, we will face the "perfect collinearity" problem.

$wage = \beta_0 X_0 + \delta_0 female + \beta_1 educ + \delta_1 male + u$
(X_0 intercept $= 1$; X_1 female, X_2 educ, X_3 male)

For example:

$X_0 = X_1 + X_3$

$$1 = female + male$$
$$female = male + 1$$

ID	female	male	X_0
1	1	0	1
2	1	0	1
3	0	1	1
4	0	1	1
⋮	0	1	1
⋮	1	0	1
⋮	1	0	1

or If there are "n" categories, we omit "1" category to avoid multicollinearity.

$$1 = winter + spring + summer + fall$$
$$winter = 1 - spring - summer - fall$$

(Dummy trap)

winter $= \begin{cases} 1 & \text{if winter} \\ 0 & \text{otherwise} \end{cases}$

spring $= \begin{cases} 1 & \text{if spring} \\ 0 & \text{otherwise} \end{cases}$

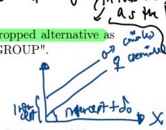
etc.

ID	winter	spring	summer	fall	X_0
1	1	0	0	0	1
2	1	0	0	0	1
3	0	0	1	0	1
4	0	0	1	0	1
1	0	0	1	0	1
1	0	1	0	0	1
1	0	1	0	0	1

In this case, we'll use it as the based (benchmark)

- At least one alternative has to be dropped. We treat the dropped alternative as the "BASE GROUP" or "BASELINE" or "BENCHMARK GROUP".

female $= \begin{cases} 1 & \text{if female} \\ 0 & \text{if male} \end{cases}$ male $= \begin{cases} 1 & \text{if male} \\ 0 & \text{if female} \end{cases}$




```
. regress lwage female male married educ exper
note: male omitted because of collinearity
```

Source	SS	df	MS
Model	54.3265253	4	13.5816313
Residual	94.0032262	521	.180428457
Total	148.329751	525	.28253286

Number of obs =	526
F(4, 521) =	75.27
Prob > F =	0.0000
R-squared =	0.3663
Adj R-squared =	0.3614
Root MSE =	.42477

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-.3251146	.0377061	-8.62	0.000	-.3991892	-.25104
male	0	(omitted)				
married	.1380145	.0411197	3.36	0.001	.0572338	.2187953
educ	.0872644	.0071554	12.20	0.000	.0732075	.1013213
exper	.0076213	.0015314	4.98	0.000	.0046129	.0106297
_cons	.4690918	.1040575	4.51	0.000	.264668	.6735156

Being female workers are expected to have less wage compared to male workers.

(20)

5 Using dummy variables for multiple categories

Case 1 We can use many dummy variables in the same model

Consider a model which includes 2 dummy variables– *female* and *married*.

($\begin{cases} 1 & \text{if female} \\ 0 & \text{if otherwise} \end{cases}$; $\begin{cases} 1 & \text{if married} \\ 0 & \text{otherwise} \end{cases}$)

$$\log(wage) = \beta_0 + \delta_0 female + \delta_1 married + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 tenure + \beta_5 tenure^2 + u.$$

```
regress lwage female married educ exper expersq tenure tenursq
```

Source	SS	df	MS
Model	65.6482326	7	9.37831895
Residual	82.6815188	518	.159616832
Total	148.329751	525	.28253286

Number of obs =	526
F(7, 518) =	58.76
Prob > F =	0.0000
R-squared =	0.4426
Adj R-squared =	0.4351
Root MSE =	.39952

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-.2901838	.0361121	-8.04	0.000	-.3611279	-.2192396
married	.0529219	.0407561	1.30	0.195	-.0271456	.1329894
educ	.0791547	.0068003	11.64	0.000	.0657952	.0925143
exper	.0269535	.0053258	5.06	0.000	.0164907	.0374163
expersq	-.0005399	.0001122	-4.81	0.000	-.0007603	-.0003196
tenure	.0312962	.0068482	4.57	0.000	.0178426	.0447499
tenursq	-.0005744	.0002347	-2.45	0.015	-.0010355	-.0001134
_cons	.4177837	.0988662	4.23	0.000	.2235557	.6120116

1) female; 2) δ_1 married

(Cannot conclude that there is any impact)

2) δ_1 measures the impact of be married. (marriage premium)
But, since $|t| < 1.96$ or $p > 0.05$, we don't reject H_0 of no impact

Comments:

1) δ_0 measures the expected difference b/w female & male workers given the same marital status & other factors.

$$\frac{\partial \log(wage)}{\partial female} = \frac{1}{wage}\frac{\partial wage}{\partial female} = -0.29$$

• female workers are expected to earn less than male workers by 29.02%, holding other factors the same.

$$= 100 \cdot \frac{1}{wage}\frac{\partial wage}{\partial female} = 100 \cdot -0.29$$

$$= \frac{\% \Delta wage}{\% \Delta female} = 29.02\%$$

	♀	♂
mar	mar fem	mar male
sing	sing fem	sing male

Base case (Baseline)

2i)

Consider a model which includes dummy variables for each gender/marital status combination– *marrmale*, *marrfem* and *singfem*. (or sing male are used as the base case)

$$\log(wage) = \beta_0 + \delta_0 marrmale + \delta_1 marrfem + \delta_3 singfem + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 tenure + \beta_5 tenure^2 + u. \quad (8.1)$$

```
regress lwage marrmale marrfem singfem educ exper expersq tenure tenursq
```

Source	SS	df	MS
Model	68.3617623	8	8.54522029
Residual	79.9679891	517	.154676961
Total	148.329751	525	.28253286

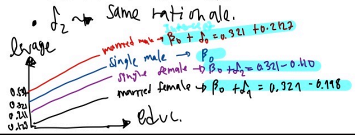
Number of obs =	526
F(8, 517) =	55.25
Prob > F =	0.0000
R-squared =	0.4609
Adj R-squared =	0.4525
Root MSE =	.39329

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
marrmale	.2126757	.0553572	3.84	0.000	.103923	.3214284
marrfem	-.1982676	.0578355	-3.43	0.001	-.311889	-.0846462
singfem	-.1103502	.0557421	-1.98	0.048	-.219859	-.0008414
educ	.0789103	.0066945	11.79	0.000	.0657585	.092062
exper	.0268006	.0052428	5.11	0.000	.0165007	.0371005
expersq	-.0005352	.0001104	-4.85	0.000	-.0007522	-.0003183
tenure	.0290875	.006762	4.30	0.000	.0158031	.0423719
tenursq	-.0005331	.0002312	-2.31	0.022	-.0009874	-.0000789
_cons	.3213781	.100009	3.21	0.001	.1249041	.5178521

$\hat{\beta}$

Comments: This regression is not the same as the previouse one. It uses "single male" as the base group. (the previous one use male & single as 2 base groups).

- δ_0 measures the expected diff. in wage of married male as compared with single males, holding other factors constant.
- δ_1 measures the expected diff. in wage of married female as compared with single males, holding other factors constant.
- δ_2 → same rationale.



22)

Case 2 We can use dummy variables to represent multiple categories of a variable Consider the relationship between law school rankings and starting salaries

$$\log(salary) = \beta_0 + \delta_0 top10 + \delta_1 r11\_25 + \delta_3 r26\_40 + \delta_4 r41\_60 + \beta_1 LSAT + \beta_2 GPA + \beta_3 \log(libvol) + \beta_4 \log(cost) + u.$$

where $top10$, $r11\_25$, $r26\_40$, $r41\_60$ would be equal to 1 when the variable $rank$ falls into the appropriate range.

** Rank below 60 would be the base case.

* In many cases, the "range of values" serve as a better explanatory variable than the "value" itself. e.g. age may explain the model better if split into generations. Young 0–15, gen z 16–29 etc.

```
. regress lsalary top10 r11_25 r26_40 r41_60 LSAT GPA llibvol lcost
```

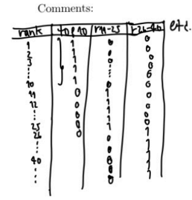
Source	SS	df	MS
Model	9.16538532	8	1.14567316
Residual	1.2109665	127	.009535169
Total	10.3763518	135	.076861865

Number of obs =	136
F(8, 127) =	120.15
Prob > F =	0.0000
R-squared =	0.8833
Adj R-squared =	0.8759
Root MSE =	.09765

lsalary	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
top10	✓ .5393428	.053542	10.07	0.000	.4333927	.6452928
r11_25	✓ .4716199	.0390921	12.06	0.000	.3942637	.548976
r26_40	✓ .2790977	.0346972	8.04	0.000	.2104383	.3477571
r41_60	✓ .182382	.0283098	6.44	0.000	.126362	.238402
LSAT	.0060482	.0034919	1.73	0.086	-.0008616	.012958
GPA	.1305893	.0818678	1.60	0.113	-.0314122	.2925908
llibvol	.0725522	.0289213	2.51	0.013	.0153221	.1297824
lcost	.0249169	.0283224	0.88	0.381	-.031128	.0809619
_cons	8.363103	.4457314	18.76	0.000	7.481081	9.245125

the baseline is ranking 61th and worse

Comments:



1) δ_0 measures the diff. in expected log(salary) of a law school graduate from a top 10 uni. compared to expected log(salary) of those who graduated from the school ranked 61th and worse.

2) δ_1 → use the same rationale.