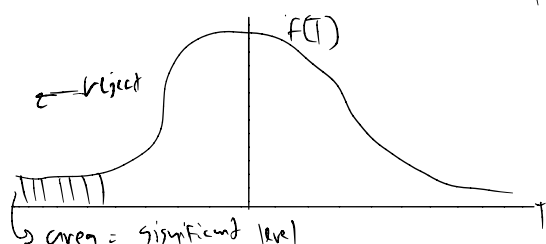


another possible hypothesis test (one-tailed alternative)

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

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

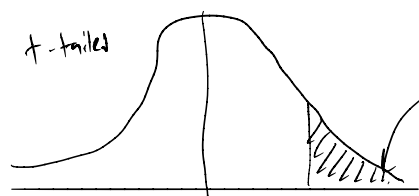
• It is assumed that β_1 would not be more than β_2 (returns to a 2-year college would never be more than return to university education)



$$t = \frac{(\hat{\beta}_1 - \hat{\beta}_2) - 0}{\text{S.E.}(\hat{\beta}_1 - \hat{\beta}_2)}$$

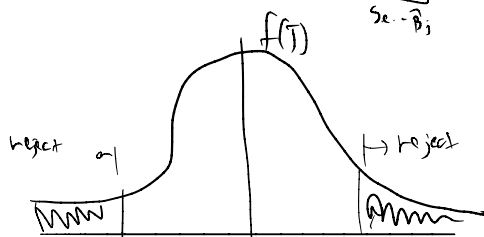
5 Computing p-Values for t-Tests

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



This shaded area in the rejection region is the p-value

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\text{S.E.}(\hat{\beta}_1 - \hat{\beta}_2)}$$



$$t = \text{computed } t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\text{S.E.}(\hat{\beta}_1 - \hat{\beta}_2)}$$

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

T = t -distributed random variable
with d.f. = $n - k - 1$

t = computed t -statistic

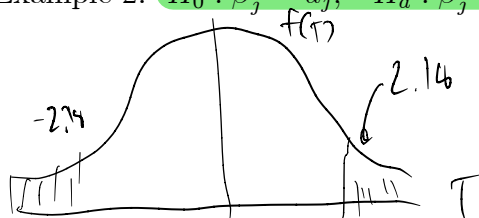
$\Rightarrow$ p-value = prob that a random t value will be
greater than H_0 test

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%?

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

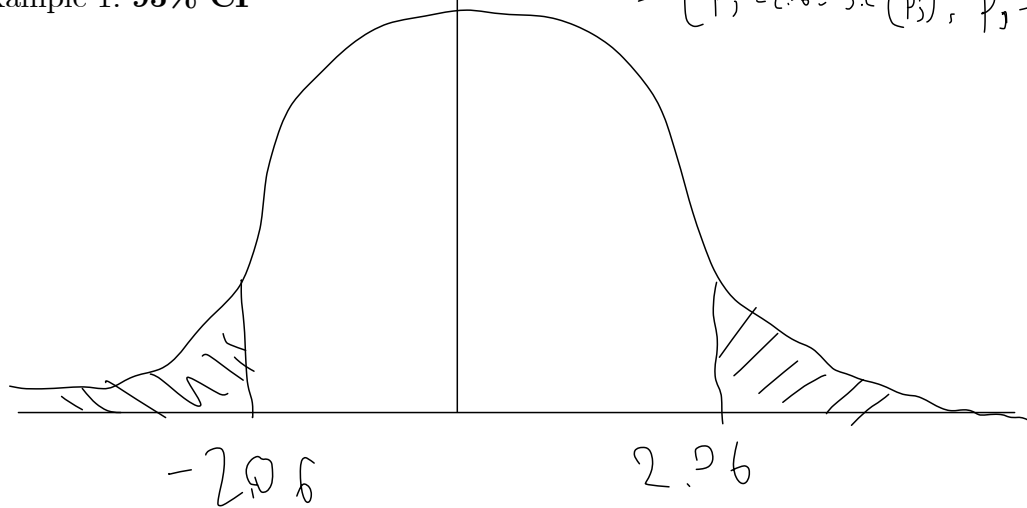


suppose the calculated $t_{\hat{\beta}_j} = -2.18$

- From the t-table, the value -2.18 corresponds to area = (0.02, 0.05)
- Thus, p-value = between 0.02 and 0.05
- Would we reject H_0 if we use the significance level = 5%?
 YES, reject H_0 because the area is less than 0.05 or
 p-value < 0.05

6 Confidence Intervals (CI)

- Confidence Intervals for the POPULATION PARAMETER (β_j)
- A 95% CI of β_j is given by

Example 1: **95% CI**

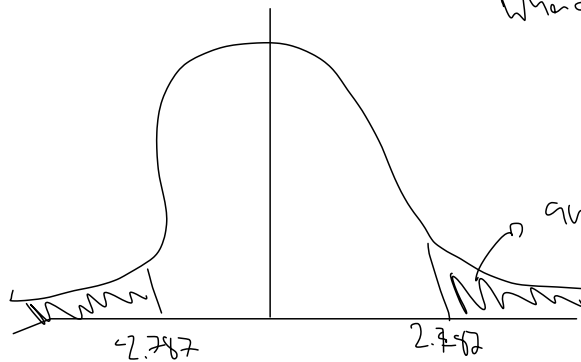
The 95% CI for $\hat{\beta}_j$

$$= [\hat{\beta}_j - 2.06 \cdot \text{s.e.}(\hat{\beta}_j), \hat{\beta}_j + 2.06 \cdot \text{s.e.}(\hat{\beta}_j)]$$

Example 2: **99% CI**

$$\text{d.f.} = 29$$

What percentile 99.5



given = $\frac{1\%}{2} = 0.005$

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

7 Testing Multiple Linear Restrictions: The F-test

Suppose the model is specified by

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

$$H_0 : \beta_1 = 0 \text{ and } \beta_2 = 0 \rightarrow \text{want to test if } x_1, x_2 \text{ both have no impact on } y,$$

$$H_1 : H_0 \text{ is not true}$$

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:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u \quad \text{is true}$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k + u \quad \rightarrow \text{reject } H_0$$

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

$$y = \beta_0 + \beta_1 x_1 + u \quad \text{is true} \rightarrow \text{do not reject } H_0$$

Suppose there are q number of β that we want to perform a joint-test of $= 0$

e.g., in the model $q=2$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{k-q} x_{k-q} + u$$

$$H_0 : \beta_{k-2+1} = \beta_{k-2+2} = \dots = \beta_{k-q}$$

$$H_a : H_0 \text{ is not true}$$

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

this term always positive and $SSR_r < SSR_{ur}$

if at the ur model

3. Some useful facts

① $R^2_{ur} > R^2_r$ because any additional x would $\uparrow R^2$

② By including more x , the model is certainly better explained

However, we would like to reject H_0 if the inclusion of extra variables doesn't improve the model enough.

4. Other ways to calculate the F-statistics:

from R^2 is $1 - \frac{SSR}{SST}$

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

of β that are set to "0"

obs # of slope

intercept

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

- the unrestricted model (ur) is defined by

UP model

. regress log_salary years gamesyr bavg hrunsyr rbisyr

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

- the restricted model (r) is defined by

. regress log_salary years gamesyr

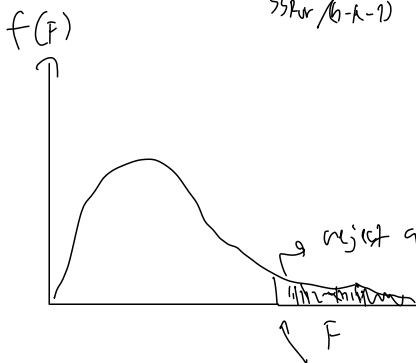
Source	SS	df	MS
Model	293.864058	2	146.932029
Residual	198.311477	350	.566604221
Total	492.175535	352	1.39822595

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

When considering each of the performance
X one by one none of them has significant
impact at 5%.

Now, our H_0 and H_a becomes

$$F = \frac{SSR_{ur} - SSR_{or} / k}{SSR_{ur} / (n - k - 1)} = \frac{194.311 - 198.186 / 3}{(198.186) / (353 - 5 - 1)} \approx 9.55$$



reject at 5%. (let's use 5% level of sig)

Since $F \approx 9.55 > 2.6$ we reject H_0

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)

⇒ These t – *statistics* are to test $H_0 : \beta_i = 0$

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

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

⇒ **When $t < 1.96$** we can say that β_i is **not statistically significant** at 5% level.

⇒ If $t < 1.96$ we can drop x_i from the model

⇒ 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

Multiple Regression Analysis : Further Issues

1 Data scaling on OLS statistics

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

$$\widehat{bwght} = \hat{\beta}_0 + \hat{\beta}_1 cigs + \hat{\beta}_2 faminc,$$

where

$bwght$ = child birth weight, in grams.

$cigs$ = number of cigarettes smoked by the mother while pregnant, per day.

$faminc$ = annual family income, in thousands of dollars.

What if we use $bwght$ in kilograms?

$$\widehat{bwght}_K = \frac{\widehat{bwght}_g}{1000} = \frac{\hat{\beta}_0}{1000} + \frac{\hat{\beta}_1}{1000} cigs + \frac{\hat{\beta}_2}{1000} faminc$$

$$= \hat{\alpha}_0 + \hat{\alpha}_1 cigs + \hat{\alpha}_2 faminc$$

$$\rightarrow \hat{\alpha}_0 = \frac{\hat{\beta}_0}{1000}, \hat{\alpha}_1 = \frac{\hat{\beta}_1}{1000}, \hat{\alpha}_2 = \frac{\hat{\beta}_2}{1000}$$

What if we use $faminc$ in USD (instead of thousands)

$$\widehat{bwght}_g = \hat{\beta}_0 + \hat{\beta}_1 cigs + \hat{\beta}_2 faminc_{USD}$$

$$\rightarrow \hat{\theta}_2 = \frac{\hat{\beta}_2}{1000}$$

2 More on functional forms

• Logarithmic Functional Form

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

$$\beta_1 = \frac{\frac{d \log(y)}{d \log(x)}}{\frac{d \log(x)}{d \log(x)}} = \frac{\frac{1}{y} dy}{\frac{1}{x_1} dx_1} = \frac{\frac{1}{y} dy}{\frac{1}{x_1} dx_1} = \frac{\log x \frac{1}{y} dy}{\log x \frac{1}{x_1} dx_1} = \frac{1}{x}$$

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

If we want the upper term to be 1% change, then

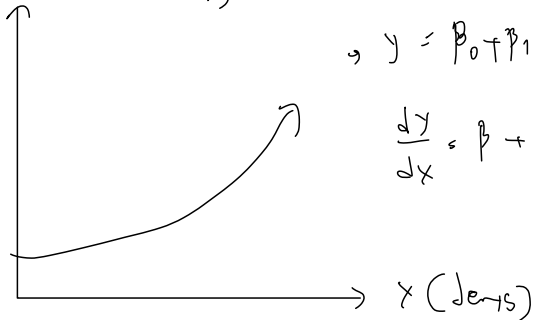
$$\log \beta_2 = \log \frac{1}{y} \frac{dy}{dx_2} = \frac{1\% \frac{dy}{dx_2}}{dx_2}$$

only 1%

Ex

• Models with Quadratics

y (Price of houses)



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

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

Relationship between

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

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

$$m = 10$$

$$\pi = (p - mc)q$$

$$\pi = (100 - 90)q$$

Example : Effects of Pollution on Housing Prices

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

$$F_{0.1} = 0.9022$$

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

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	
Total	84.582225	505	.167489554	Prob > F =	0.0000	
				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

Consider the effect of "room"

$$\frac{\Delta \ln(\text{price})}{\Delta (\text{rooms})} = \beta_3 + 2\beta_4 \text{rooms} = -0.553 + 2(0.062) \text{rooms}$$

$\ln(\text{price})$ $\Delta (\text{rooms})$

$$0 = -0.553 + 2(0.062) \text{rooms}$$

$$\text{rooms} = 4.4$$

implying at 4.4 rooms or more.

at ~ 5 rooms or more

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

$$\frac{\Delta \ln(\text{price})}{\Delta \text{rooms}} = -0.553 + 2(0.062) \cdot \text{rooms}$$

$$\frac{100 \cdot \frac{1}{\text{price}} \Delta \text{price}}{\Delta \text{rooms}} = 0.067 \times 100 = 6.7\% \text{ increase}$$

What about % in price when # rooms increases from 5 to 6?

$$\% \Delta \text{price} = 100(-0.553 + 2(0.062) \cdot 6) = 19.7\%$$

3 Models with Interaction Terms

Consider

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

where

price = housing price

sqft = house size (square feet)

bdrms = number of bedrooms

bthrms = number of bathrooms

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

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

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

- Adding more regressors ALWAYS improve fit

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

Consider Model 1

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

Consider Model 2

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

Multiple Regression Analysis with Qualitative Information:

1 Outline

- Describing qualitative information
- Using a single dummy independent variable
- Using dummy variables for multiple categories
- Interactions involving dummy variables
- A binary dependent variable (Y variable): The linear probability model

2 Describing Qualitative Information

- "Female" and "Married" are qualitative variable.
- We arbitrarily assign a dummy variable to describe them.

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

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

TABLE 7.1

A Partial Listing of the Data in WAGE1.RAW

<i>person</i>	<i>wage</i>	<i>educ</i>	<i>exper</i>	<i>female</i>	<i>married</i>
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

3 Models with a single dummy independent variable

Consider

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

where

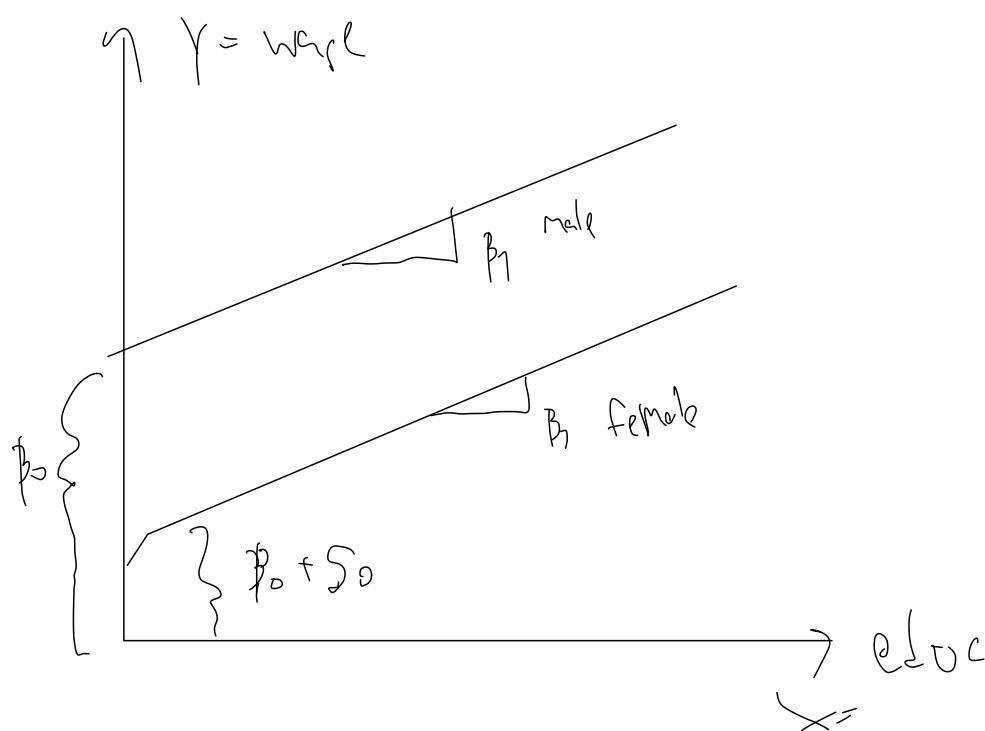
$$female = \begin{cases} 1 & \text{if female} \\ 0 & \text{otherwise (or if male)} \end{cases}$$

In this case, the δ_0 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.

$$\begin{aligned} \textcircled{1} E(wage | female, educ) &= E(\beta_0 + \delta_0 female + \beta_1 educ + u | female, educ) \\ &= \beta_0 + \delta_0 female + \beta_1 educ + E(u | female, educ) \\ &= \beta_0 + \delta_0 female + \beta_1 educ \end{aligned}$$

Thus

$$\textcircled{2} \quad \begin{aligned} \text{♀} : E(wage | female = 1, educ) &= \beta_0 + \delta_0(1) + \beta_1 educ = \beta_0 + \delta_0 + \beta_1 educ \\ \text{♂} : E(wage | female = 0, educ) &= \beta_0 + \delta_0(0) + \beta_1 educ = \beta_0 + \beta_1 educ \end{aligned}$$



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*.

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

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

Source	SS	df	MS	Number of obs = 526		
Model	65.6482326	7	9.37831895	F(7, 518) = 58.76		
Residual	82.6815188	518	.159616832	Prob > F = 0.0000		
Total	148.329751	525	.28253286	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

Comments:

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

female workers are expected
to earn less than males workers
by 29.02%, holding other factors
for same.

$$\frac{\Delta \log(\text{wage})}{\Delta \text{female}} = \log(-0.29)$$

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

Consider a model which includes dummy variables for each gender/marital status combination— *marrmale*, *marrfem* and *singfem*.

$$\log(\text{wage}) = \beta_0 + \delta_0 \text{marrmale} + \delta_1 \text{marrfem} + \delta_2 \text{singfem} + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{exper}^2 + \beta_4 \text{tenure} + \beta_5 \text{tenure}^2 + u. \quad (8.1)$$

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

Source	SS	df	MS	Number of obs = 526		
Model	68.3617623	8	8.54522029	F(8, 517) = 55.25		
Residual	79.9679891	517	.154676961	Prob > F = 0.0000		
Total	148.329751	525	.28253286	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

Comments: This regression is not the same as the previous one. It uses "single male" as the base group.

- So measure the expected diff in wage of married compared with single males, holding other factors constant
- So measure the expected diff in wage of married

• $\delta_2 \rightarrow$ same relation

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

$$\log(\text{salary}) = \beta_0 + \delta_0 \text{top10} + \delta_1 r11_25 + \delta_3 r26_40 + \delta_4 r41_60 + \beta_1 \text{LSAT} + \beta_2 \text{GPA} + \beta_3 \log(\text{libvol}) + \beta_4 \log(\text{cost}) + u.$$

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.

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

Source	SS	df	MS	Number of obs =	136
Model	9.16538532	8	1.14567316	F(8, 127) =	120.15
Residual	1.2109665	127	.009535169	Prob > F =	0.0000
				R-squared =	0.8833
				Adj R-squared =	0.8759
Total	10.3763518	135	.076861865	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

Comments:

Rank	top10	r11-25	r26-40	etc
1	1	0	0	
2	1	0	0	
3	1	0	0	
⋮	1	0	0	
10	1	0	0	
11	0	1	0	
12	0	1	0	
⋮	0	1	0	
25	0	1	0	
26	0	0	1	
⋮	0	0	1	
40	0	0	1	

1) δ_0 measures the difference in expected $\log(\text{salary})$ of a law-school graduate from a top 10 university compared to expected $\log(\text{salary})$ of those who graduated from the school.