4 Testing Hypotheses about a Single Linear Combination of the Parameter

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 $\beta_1 = \beta_2$.

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

against

$H_a: \beta_1 \neq \beta_2 \gg H_0: \beta_1 - \beta_2 \neq 0$

2tailed-test

$t = \dfrac{(\hat{\beta}_1 - \hat{\beta}_2) - 0}{s.e.(\hat{\beta}_1 - \hat{\beta}_2)}$ } we compute this t-statistics and compare with the critical vale

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

not very straight forward to calculate

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

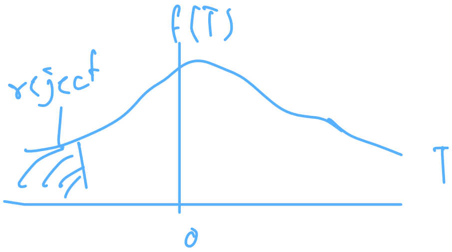
>> we use a variable transformation trick → see notes

another possible hypothesis test (one-tailed alternative)

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

$H_a: \beta_1 < \beta_2 \Rightarrow H_0: \beta_1 - \beta_2 < 0$

- It is assumed that β_1 would be not more than β_2 (returns to a 2yrs colledge would never be more than returns to university education)



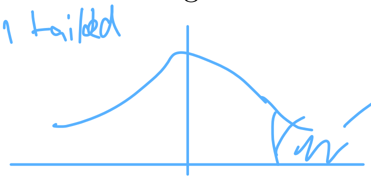
$$t = \frac{(\hat{\beta}_1 - \hat{\beta}_2) - 0}{s.e(\hat{\beta}_1 - \hat{\beta}_2)}$$

Then, go to the extra note

5 Computing p-Values for t-Tests

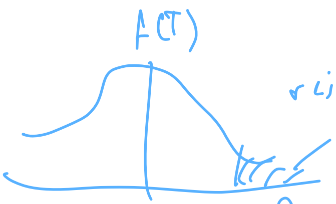
- What is the significance level given the computed t-statistics?

1 tailed



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

$$t = \frac{\hat{\beta}_j - \beta_j}{s.e.\hat{\beta}_j}$$



reject

total area from 2 sides

$$t = \frac{\hat{\beta}_j - \beta_j}{s.e.\hat{\beta}_j}$$

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

T = t - distributed random variables with df = n-k-1

t = computed t stats.

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

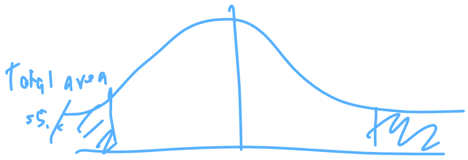
- p-value = what should be the sig level, given the critical value of -2.75 ? Find the shaded area

suppose the calculated $t_{\hat{\beta}_j} = -2.75$ >> $t_{\hat{\beta}_j} = \frac{(\hat{\beta}_j - \beta_j)}{s.e.(\hat{\beta}_j)}$

- 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

RULE = we reject H_0 if p-value < sig lvl.

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



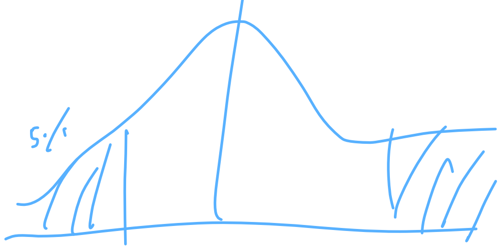
suppose the calculated $t_{\hat{\beta}_j} = -2.18$

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

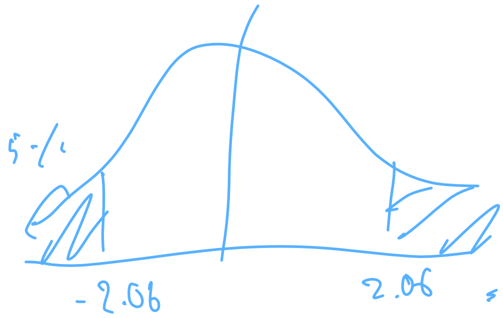
6 Confidence Intervals (CI)

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

CI = $\hat{\beta}_j \pm c \cdot s.e.(\hat{\beta}_j)$
c is the 97.5 percentage in the t-distribution with n-k-1 d.f.



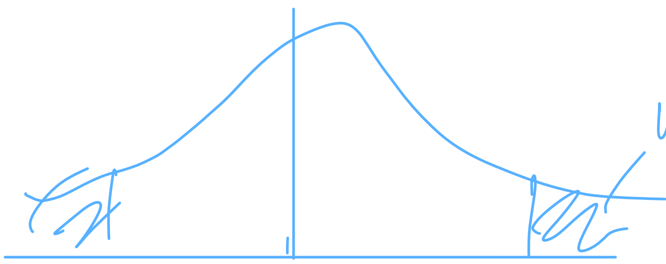
Example 1: **95% CI**



= value of 97.5th percentage in the t_{25} distribution

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

Example 2: **99% CI** df = 25



What pctile? 99.5

or area = 1 - 0.005 = 0.995

→ area = $\frac{1\%}{2} = \frac{0.01}{2} = 0.005$

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

7 Testing Multiple Linear Restrictions: The F-test

Suppose the model is specified by

$$
\begin{aligned}
y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u \\
H_0 &: \beta_1 = 0 \text{ and } \beta_2 = 0 \\
\text{Ha}, H_1 &: H_0 \text{ is not true}
\end{aligned}
$$

>> want to test if x_1 and x_2 both have no impact on y

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: ↗ big model

   $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$ is true >> Reject Ho

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

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 Ho

· suppose there are "q" number of β that we would like to perform a joint-test of = 0

e.g. in this model q=2

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

$H_0 : \beta_{k-q+1} = \beta_{k-q+2} = \dots = \beta_k = 0$

(the last q βs = 0)

Ha = Ho is not true

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{k-q} x_{k-q} + \beta_{k-q+1} x_{k-q+1} + \beta_{k-q+2} + \dots \beta_k x_k + u$

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

← This is always (+) b.c. $SSR_{ur} < SSR_r$
Everytime you add 1 more x, the model will be better explained

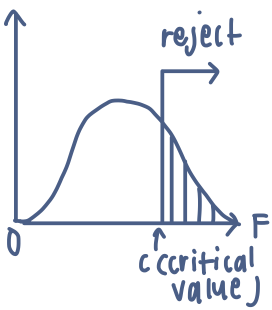
↳ d.f. of the "ur" model

· So, if everytime you add 1 more x variable, the SSR↓ and R^2↑, why don't we just keep the additional x in the model

>> because everytime we add 1 more x, $Var(\hat{\beta}s)$ will increase, making the prediction of β less precise

So we only keep the addition xs if it/they can improve the model enough

(and SSR↓ (R^2↑) enough
can significantly ↓ SSR and R^2↑



$H_0: \beta_2 = \beta_3 = \ldots = 0$

H_a: H_0 is not true

$F \sim F_{q, n-K-1}$ ← d.f. of the ur model

# of joint hypothesis's being tested

3. Some useful facts

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

>> $SSR_{ur} < SSR_r$

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

4. Other ways to calculate the F-statistics:

From $R^2 = 1 - \frac{SSR \rightarrow RSS}{SST \rightarrow TSS}$

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

# of β that are set to "0" → q

# of obs → n - k - 1 → intercept

↓ # of slope β

>> If we want to test the overall significance of the model

$H_0 : \beta_1 = \beta_2 = \beta_3 = .. = \beta_K = 0$, H_a = otherwise

$F \equiv \frac{R^2/K}{(1 - R^2)/(n-k-1)}$

R^2 of the model ≃ uR

the "r" model has no x at all

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

Y	*salary*	= season salary
r, ur	*years*	= years in major leagues
r, ur	*gamesyr*	= games per year in the league
ur	*bavg*	= career batting average
ur	*hrunsyr*	= homeruns per year
ur	*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 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
------------------------------------------------------------------------------
```

y: log_salary; x: years gamesyr bavg hrunsyr rbisyr (1, 2, 3, 4, 5); $q = 3$ (bavg, hrunsyr, rbisyr); _cons: intc.

when considering each of the performance x one-by-one, none of them has a significant impact at 5%

- 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
------------------------------------------------------------------------------
```

y: log_salary; x: years gamesyr; SSE: Model; SSR: Residual; SST: Total

• But when performing an F-test, performances have joint impact.

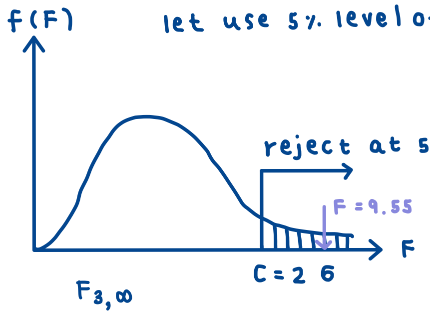
Now, our H_0 and H_a becomes

$$F \equiv \frac{\frac{SSR_r - SSR_{ur}}{q}}{\frac{SSR_{ur}}{n-k-1}} \equiv \frac{\frac{198.311 - 183.186}{3}}{\frac{183.186}{353-5-1}} \approx 9.55$$

HW.

$$F \equiv \frac{\frac{R^2}{q}}{\frac{1-R^2}{n-k-1}} \equiv ?$$

let use 5% level of sig.



Since F = 9.55 > 2.6, we reject H_0 at 5% sig level and conclude that performances have joint effects on salary.

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

   $\Rightarrow$ If the d.f. > 30, then when $t \ > 1.96$, we can reject H_0 with 5% sig. level

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

   $\Rightarrow$ **When** $t \ <$ **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

sales →

other company performance

CEO characteristics

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 *R*-squared	177 .281	177 .304	177 .353

like a simple regression w/1 X

Multiple Regression Analysis : Further Issues

1 Data scaling on OLS statistics

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

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

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 bweght in kilograms ??

  1 kg = 1000 g

  $$\widehat{bweght}_{kg} = \frac{\widehat{bweght}_g}{1000} = \frac{\hat\beta_0}{1000} + \frac{\hat\beta_1}{1000} cigs + \frac{\hat\beta_2}{1000} faminc$$

  $$= \hat\alpha_0 + \hat\alpha_1 cigs + \hat\alpha_2 faminc$$

  $$>> \hat\alpha_0 = \frac{\hat\beta_0}{1000}, \quad \hat\alpha_1 = \frac{\hat\beta_1}{1000}, \quad \hat\alpha_2 = \frac{\hat\beta_2}{1000}$$

- What if we use faminc in USD (instead of 1000 USD)

  $$bweght_g = \hat\beta_0 + \hat\beta_1 cigs + \hat\beta_2 faminc_{USD}$$

  The value of this variable is going to be 1000 times larger than faminc

  $$>> \hat\theta_2 = \frac{\hat\beta_2}{1000}$$

  in other words $\hat\theta_2$ = impact of 1 USD ↑ in income

  $\hat\beta_2$ = impact of 1000 USD ↑ in income

- What if we use bweght in kg & income in THB

  $$bweght_{kg} = \frac{\hat\beta_0}{1000} + \frac{\hat\beta_1}{1000} cigs + \frac{\left(\frac{\hat\beta_2}{1000}\right)}{1000} faminc_{THB}$$

  This value is going to be 30,000 times more than faminc.

2 More on functional forms

- Logarithmic Functional Form

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

$\Delta y = y_1 - y_2$

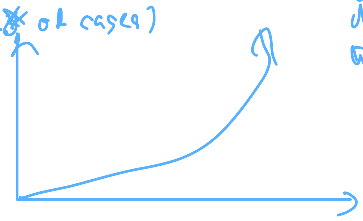
$\Delta x = x_{11} - x_{12}$

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

- Models with Quadratics

Covid19 example

y (# of cases)

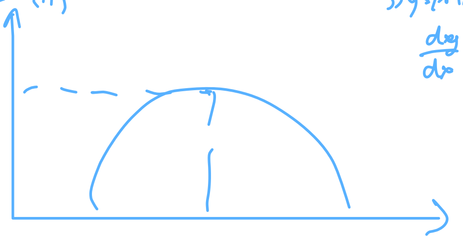


$\gg y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$

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

Decreasing returns

prof (π)



$\gg y = \beta_0 + \beta_1 x_1 + \beta_2 x^2 + u$

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

$\pi = (P - MC)q$

$\{900 - q - 90$

Example : Effects of Pullution on Housing Prices

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

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 school in the area w/o having to take any test. So, the lower stratio, the better school

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

Source	SS	df	MS
Model	51.4933152	5	10.298663
Residual	33.0889098	500	.06617782
Total	84.582225	505	.167489554

Number of obs =	506
F(5, 500) =	155.62
Prob > F =	0.0000
R-squared =	0.6088
Adj R-squared =	0.6049
Root MSE =	.25725

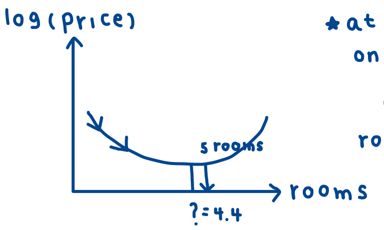
log(price) → lprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
log(nox) → lnox	β_1 -.9767545	.0995938	-9.81	0.000	-1.172429	-.7810806
dist	β_2 -.0321972	.0094013	-3.42	0.001	-.050668	-.0137264
rooms	β_3 -.5528032	.1612965	-3.43	0.001	-.8697056	-.2359007
rooms_sq	β_4 .0624697	.0124867	5.00	0.000	.0379368	.0870025
stratio	β_5 -.0486667	.0058131	-8.37	0.000	-.0600879	-.0372455
_cons	intc 13.59154	.5650901	24.05	0.000	12.4813	14.70178

$|t| > 1.96$ ↑ ↑ all < 0.05

>> all variables are significant

Consider the effect of "room"

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



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

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

$rooms = 4.4$

ANS at 4.4 rooms or more

at 5 rooms or more

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

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

$$100 \times \frac{\frac{1}{price}\,d\,price}{d\,rooms} = 100(-0.553 + 2(0.062)\cdot 5)$$

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

• total % Δ in price when # rooms ↑ from 5 to 7 is $6.7 + 19.1 = 25.8\%$

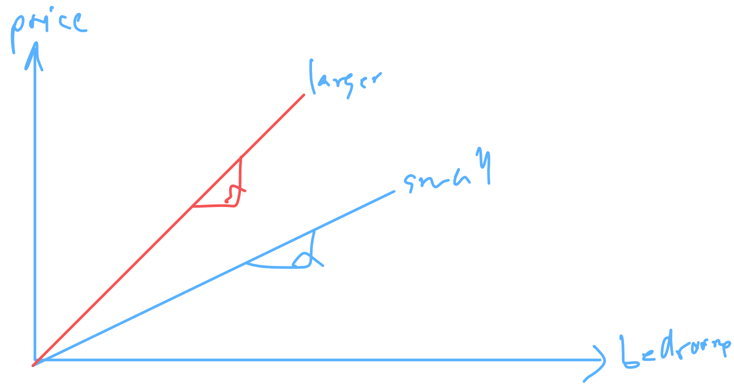
>> what about % in price when # rooms increases from 5 to 7 ??

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

3 Models with Interaction Terms

Consider

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



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

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

> 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

But we lose the "dof"

(d.f. = free data point used to estimate the parameters)

>> 1 data point is sacrificed every time we estimate a parameter

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

Consider Model 1

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

Consider Model 2

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

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 arbitarily assign a dummy variable to decribe them.

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

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

3 Models with a single dummy independent variable

Consider

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

where

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

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.

1) $E(wage \mid female, edu) = E(\beta_0 + \delta_0 female + \beta_1 edu + u \mid female, edu)$

$= \beta_0 + \delta_0 female + \beta_1 edu + E(u \mid female, edu)$

$= \beta_0 + \delta_0 female + \beta_1 edu$

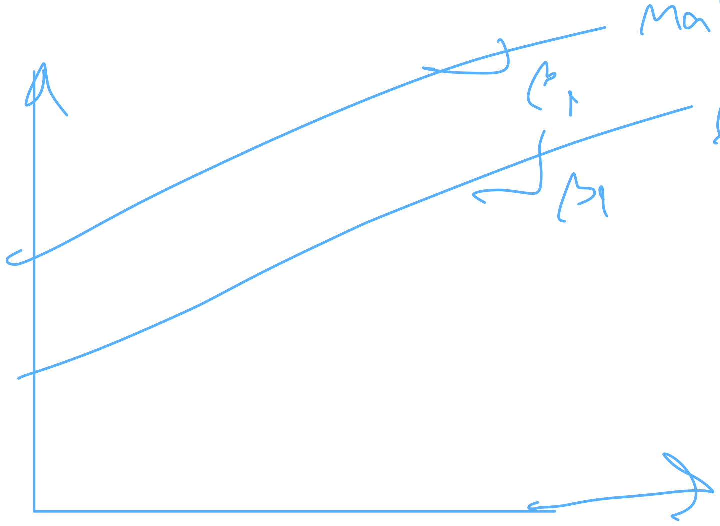
2) Two

f: $E(wage \mid female = 1, edu) = \beta_0 + \delta_0 (1) + \beta_1 edu = \beta_0 + \delta_0 + \beta_1 edu$

♂: $E(wage \mid female = 0, edu) = \beta_0 + \delta_0 (0) + \beta_1 edu$

male: $\beta_0 + \beta_1 edu = \beta_0 + \beta_1 edu$

female: $\underbrace{\beta_0 + \delta_0}_{\text{intercept}} + \beta_1 edu$



4 It is not possible to include all of the dummy alternatives in the same model

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

For example:

$$\begin{aligned} 1 &= female + male \\ female &= male + 1 \end{aligned}$$

or

$$\begin{aligned} 1 &= winter + spring + summer + fall \\ winter &= 1 - spring - summer - fall \end{aligned}$$

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

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

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  4,   521) =   75.27
       Model |  54.3265253     4  13.5816313           Prob > F      =  0.0000
    Residual |  94.0032262   521  .180428457           R-squared     =  0.3663
-------------+------------------------------           Adj R-squared =  0.3614
       Total |  148.329751   525   .28253286           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
------------------------------------------------------------------------------
```

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 $married$.

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

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

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  7,   518) =   58.76
       Model |  65.6482326     7  9.37831895           Prob > F      =  0.0000
    Residual |  82.6815188   518  .159616832           R-squared     =  0.4426
-------------+------------------------------           Adj R-squared =  0.4351
       Total |  148.329751   525  .28253286            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:

1. δ_0 measures the expected value bet female/male workers given the same marital status and other factors.

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

$$= \frac{100 \cdot \frac{1}{wage}\, dwage}{\partial female} = 100(-0.29)$$

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

2) δ_1 measures the impact of the married

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

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

```
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

Comments: The regression is not the same as the previous, therefore single male as the base group

· δ_0 measure the expected diff in wage of married as compared with single males

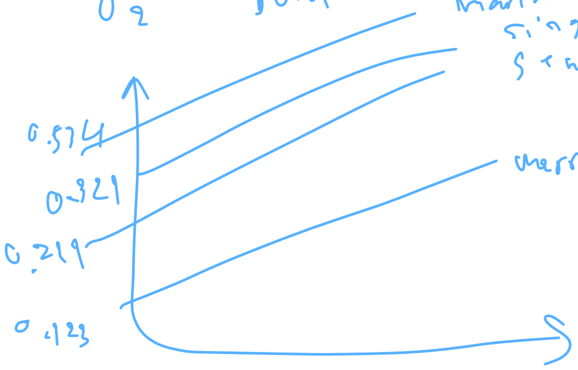
· δ_1 measures the expected diff in diff. in wage married females as compared.

· δ_2 → same rationale

married = $\beta_0 + 0.2127$ = 0.321 + 0.2127

single = m. β_0

female = f. $\beta_0 + \delta_2$



married female = $\beta_0 + \delta_1$ = 0.321 − 0.1986

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

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

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
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

Comments: 1) δ_0 measures the difference 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 just the same rationale