3. Some useful facts

① $R^2_{ur} > R^2_r$ because any additional x would increase R^2 (compare AA)
$\Rightarrow SSR_{ur} < SSR_r$

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

4. Other ways to calculate the F-statistics:

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

we have $F = \dfrac{(R^2_{ur} - R^2_r)}{\#\text{ of } \beta \text{ that are set to "0"}} \Big/ \dfrac{(1 - R^2_{ur})}{n - k - 1}$ ← intercept

k ↑ # of slope β, # of obs.

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

$H_0: \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$, else = otherwise

$F = \dfrac{R^2/k}{(1-R^2)/(n-k-1)}$ ← 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.

If we want to test whether performance has any impact on salary

$$\begin{cases} ur \begin{cases} salary & = \text{season salary} \\ years & = \text{years in major leagues} \\ gamesyr & = \text{games per year in the league} \\ bavg & = \text{career batting average} \\ hrunsyr & = \text{homeruns per year} \\ rbisyr & = \text{runs batted in per year} \end{cases} \\ r \end{cases}$$

$H_0: \beta_{bavg} = \beta_{hrunsyr} = \beta_{rbisyr} = 0$

H_a: Otherwise is true

- the unrestricted model (ur) is defined by

```
. 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
				R-squared = 0.6278
Total	492.175535	352	1.39822595	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

intercept

- 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
Model	293.864058	2	146.932029	F(2, 350) = 259.32
Residual	198.311477	350	.566604221	Prob > F = 0.0000
				R-squared = 0.5971
Total	492.175535	352	1.39822595	Adj R-squared = 0.5948
				Root MSE = .75273

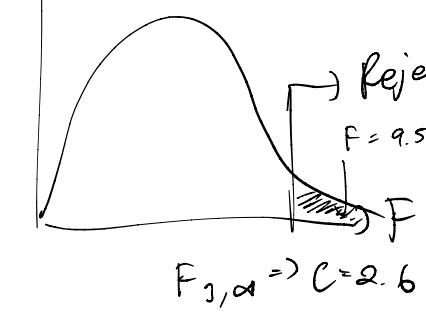
log_salary	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
years	.071318	.012505	5.70	0.000	.0467236	.0959124
gamesyr	.0201745	.0013429	15.02	0.000	.0175334	.0228156
_cons	11.2238	.108312	103.62	0.000	11.01078	11.43683

- But when performing an F-test, performance drop:

$F = \frac{(R^2/q)}{(1-R^2)/(n-k-1)} = \frac{0.6278/3}{(1-0.5971)/353-5-1} =$

Now, our H_0 and H_a becomes

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} = \frac{(198.311 - 183.186)/3}{(183.86)/(353-5-1)} \approx 9.55$$



let's use 5% level of sig since $F = 9.55 > 2.6$, we reject H_0 at 5% level and conclude these performing have joint effect 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 $H_0 : \beta_i = 0$

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

   $\Rightarrow$ **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

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

sales

other company performance

CEO characteristics

like a simple regression with 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

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

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 b weight in kilograms??

$1\,kg = 1000\,g$

$$\widehat{bweight_{kg}} = \frac{\widehat{bweight_g}}{1000} = \frac{\hat{\beta}_0}{1000} + \frac{\hat{\beta}_1}{1000} cigs + \frac{\hat{\beta}_2}{1000} famine$$

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

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

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

$$bweight_g = \hat{\beta}_0 + \hat{\beta}_1 cigs + \frac{\hat{\beta}_2}{1000} famine_{USD}$$

$$= \hat{\beta}_0 + \hat{\beta}_1 cigs + \hat{\theta}_2 famine_{USD}$$

This value of this variable is going to be 1000 times larger than famine

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

In other word $\hat{\theta}_2$ = impact of 1 USD ↑ in income

$\hat{\beta}_2$ = ⟵ 1000 USD ↑ in income

- what if we use b weight in kg & income in THB

$$bweight_{kg} = \frac{\hat{\beta}_0}{1000} + \frac{\hat{\beta}_1}{1000} cigs + \frac{\left(\frac{\hat{\beta}_2}{1000}\right)}{30{,}000} famine_{THB}$$

↑ this value is going to be 30,000 times more than famines

2 More on functional forms

- Logarithmic Functional Form

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

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

with the log y & log x format, the coefficient is going to be the elasticity!

(x_1: elasticity of y (demand) / (price))

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

⇒ if we want the upper term to be % change, then

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

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

$100\beta_2 = \%\Delta$ in y given that x_2 increase by 1 unit

- Models with Quadratics (Squares)

→ capture increasing / decreasing marginal effects (slope of the relationship between x & y is not constant)

COVID-19 example

y (# of cases) ⇒ $y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$

$\frac{dy}{dx} = \beta_1 + 2\beta_2 x$ (β_1: (+), β_2: (+) days)

x (days)

Decreasing returns

β_0 (# of C[T]) ⇒ $y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$

$\frac{dy}{dx} = \beta_1 + 2\beta_2 x$ (β_1: (+), β_2: (−))

Unit (q)

q^*

Assume MC = 10

$\Pi = (P - MC)q$; Demand: $P = 100 - q$

$\Pi = (100 - q - 10)q$

β_1 is positive

F.O.C $\frac{d\Pi}{dq} = 0 = 90 - 2Q^*$ ← β_2 is (−)

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 many other countries, students can apply to school in the area to school in the area without having to take any test. So, the lower stratio the better the 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]
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 (nox)

$|t| > 1.96$ ↑ all < 0.05 ↑

⇒ all variable are significant

Consider the effect of "room"

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

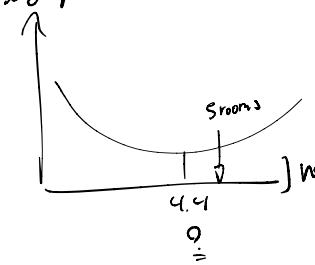
↳ at how many rooms does 1 additional room has a positive on log(price) ??

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

$rooms = 4.4$

Answer → 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?

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

total % Δ in price

when # rooms ↑ from 5 to 6 is + 19.1% = 25.8%

$$\frac{100 \cdot \frac{1}{price} d price}{d\, rooms} = 100 \times 0.062 = 6.2\% \text{ increase}$$

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

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

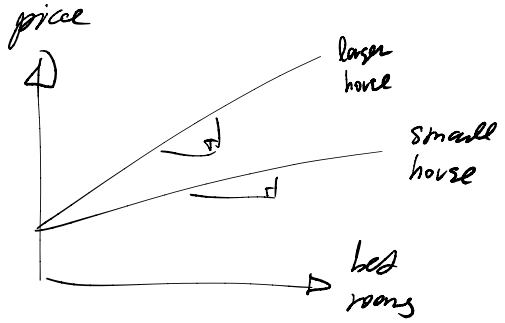
3 Models with Interaction Terms

⇒ used when the impact of one variable depends on the value (level) of another variable.

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 $\Rightarrow R^2$ 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 parater
- using R^2 would not punish "having too many regressors"
- we use adjusted-R^2 or $\bar{R}^2$ when we want to punish adding too many regressors

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

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

⇝ If we have more k, d.f. = n-k-1 ↓,
$\frac{SSR}{(n-k-1)}$ ↑, adj $-R^2$ ↓

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

Consider Model 1

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

Consider Model 2

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

← 27.5% of variation in Y is explained so, this model is better.

$SSR_1 = SSR$ from subsample 1
$SSR_2 = SSR$ from subsample 2

```
regress cumgpa sat hsperc tothrs if female == 0
```

Source	SS	df	MS	Number of obs =	552
				F(3, 548) =	41.94
Model	89.6937042	3	29.8979014	Prob > F =	0.0000
Residual	390.619421	548	.712809162	R-squared =	0.1867
				Adj R-squared =	0.1823
Total	480.313125	551	.871711661	Root MSE =	.84428

cumgpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sat	.0006113	.000231	2.65	0.008	.0001576	.001065
hsperc	-.0059675	.0017459	-3.42	0.001	-.0093969	-.002538
tothrs	.0103004	.001074	9.59	0.000	.0081907	.0124101
_cons	1.213984	.2602697	4.66	0.000	.7027359	1.725233

```
regress cumgpa sat hsperc tothrs if female == 1
```

Source	SS	df	MS	Number of obs =	180
				F(3, 176) =	34.08
Model	83.4816253	3	27.8272084	Prob > F =	0.0000
Residual	143.689727	176	.816418902	R-squared =	0.3675
				Adj R-squared =	0.3567
Total	227.171352	179	1.2691137	Root MSE =	.90356

cumgpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sat	.0017281	.0004642	3.72	0.000	.0008119	.0026442
hsperc	-.0059167	.0038895	-1.52	0.130	-.0135927	.0017594
tothrs	.0158603	.0018485	8.58	0.000	.0122122	.0195085
_cons	.1003465	.4810947	0.21	0.835	-.8491105	1.049803

Comments: