F-test

we want to test the significance of a group of hypotheses (multiple hypotheses)

$$Grade_{325} = \beta_0 + \beta_1 \#times\_front + \beta_2 \#times\_back + \beta_3 hr\_study + \beta_4 past\_GPA + \beta_5 gender + u$$

H_0 : seat position doesn't have impact on GPA

$\beta_1 = 0$ and $\beta_2 = 0$ $\rightarrow$ $\beta_1 = \beta_2 = 0$

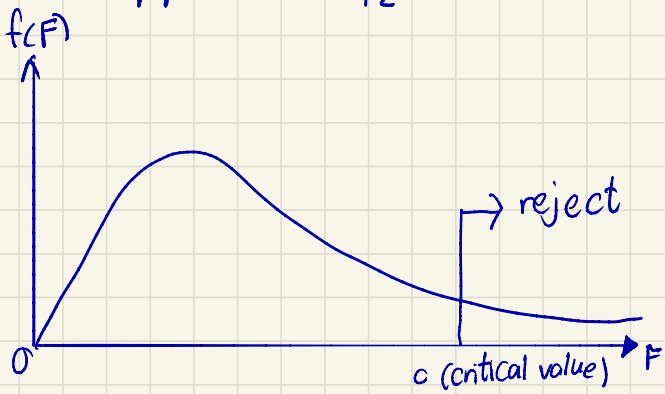
H_a : seat position matters

$\beta_1 \neq 0$ and $\beta_2 \neq 0$

or $\beta_1 \neq 0$ and $\beta_2 = 0$

or $\beta_1 = 0$ and $\beta_2 \neq 0$

} at least one of the $\beta_1, \beta_2 \neq 0$



$H_0 : \beta_2 = \beta_3 = 0$

H_a : H_0 not true

$F \sim F_{q, n-k-1}$

q : # of hypotheses being tested

$n-k-1$: d.f. of ur model

7 Testing Multiple Linear Restrictions: The F-test

Suppose the model is specified by

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

→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 (contain all variables)

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$ is true → reject H_0

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

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" numbers of β that we would like to perform a joint-test of $= 0$

→ e.g. in this model, $q = 2$

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

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

(the last q βs = 0

H_a : H_0 is not true

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

SSR_{ur} smaller than SSR_r everytime you add 1 more x, the model will be better explained

q ← number of explainatory variables we want to test

$n-k-1$ ← d.f. of "ur" model

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

→ everytime we add 1 more x, $var(\hat{\beta}_s)$ ↑ making the prediction of β less precise.

→ so we want to keep the additional X_3 if it can improve the model enough, SSR ↓ R^2 ↑ (significant)

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 H_0 if the inclusion of variables does not explain the model enough

4. Other ways to calculate the F-statistics:

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

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

q: # of β that are set to 0; n: # of observ.; k: # of slope β; $n-k-1$: intercept

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

$H_0: \beta_1 = \beta_2 = \beta_3 \ldots = \beta_k = 0$

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

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
- *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

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

H_a: otherwise is true

- the unrestricted model (ur) is defined by

y x

```
. regress log_salary years gamesyr bavg hrunsyr rbisyr
```

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

SSE SSR SST

% of the variables that the model can explain

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

1 2 3 4 5

<1.96 >0.05

intercept

When considering each of the performance x one by one, none of them has a significant impact at 5%. (0.05)

- the restricted model (r) is defined by

```
. regress log_salary years gamesyr
```

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

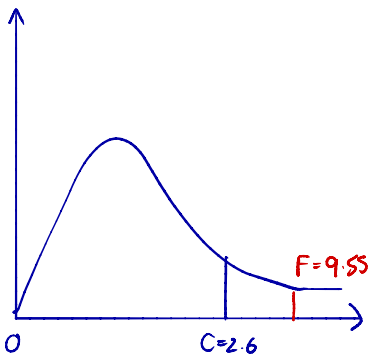
SSE SSR SST

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 F test, performances have joint impact

Now, our H_0 and H_a becomes

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



since 9.55 > 2.6, we reject H_0 at 5% 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 with 5% sig level

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

   $\Rightarrow$ **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 →

Company performance

CEO characteristics

like a simple regression with 1X omitted variable bias

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 bweight in kilograms, 1 kg = 1000

$$\widehat{bweight}_{kg} = \frac{bweight_g}{1000} = \frac{\hat{\beta}_0}{1000} + \frac{\hat{\beta}_1}{1000} cigs + \frac{\hat{\beta}_2}{1000} faminc$$

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

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

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

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

$$= \hat{\beta}_0 + \hat{\beta}_1 cigs + \hat{\theta}_2 faminc_{USD}$$ ← 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
$\hat{\beta}_2$ = 1 ——" 1000 USD ↑ increase in income

What if we use bweght in kg & income in THB

$$bweight_{kg} = \frac{\hat{\beta}_0}{1000} + \frac{\hat{\beta}_1}{1000} + \frac{\hat{\beta}_2}{1000} faminc_{THB}$$

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

2 More on functional forms

$\beta_0, \beta_1, \beta_2$ have to be linear
x_1, x_2 – anything

- Logarithmic Functional Form

usually means natural log

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

$\Delta y = y_1 - y_2$
$\Delta x = x_{11} - x_{12}$

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

With the log(y) & log(x) format, the coefficient is going to be the elasticity (x_1 elasticity of y)

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

if we want the upper term to be % change

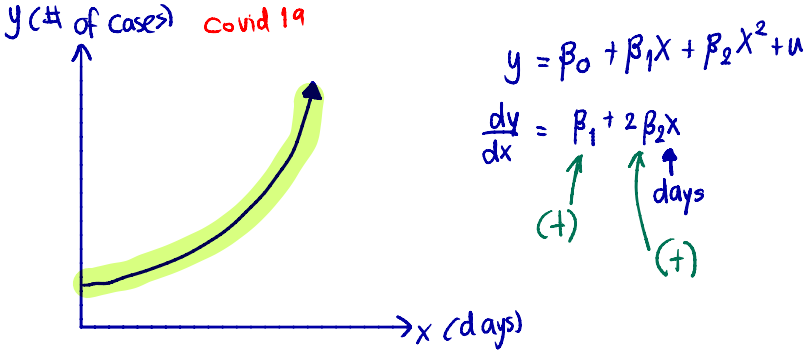
$$100\beta_2 = \frac{100\frac{1}{y}\Delta y}{\Delta x_2}$$

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

$100\beta_2$ = % Δ in y given that x_2 increases by 1 unit

- Models with Quadratics (squares)

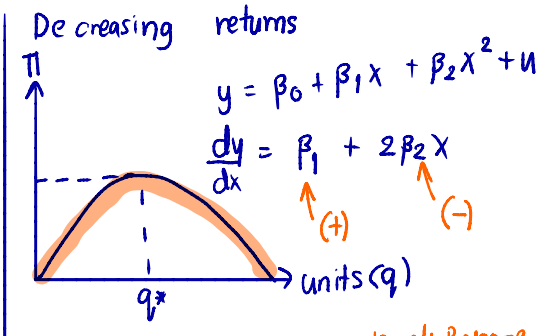
→ capture increasing / decreasing marginal effects (slope of the relationship between x & y is not constant



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

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

β_1 (+), β_2 (+), days



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

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

β_1 (+), β_2 (−)

$\Pi = (p - mc)q$; $mc = 10$, demand: $p = 100 - q$

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

FOC $\frac{\partial \Pi}{\partial q} = 0 = 90 - 2q$ (β_2 is (−))

Example : Effects of Pullution on Housing Prices

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