6 Interactions involving dummy variables

Case 1 Interactions among dummies

** We can use interactions among dummies to account for the effect of each combination of dummies as well:

A different way to estimate eq.(8.1) is

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

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

$$\log(wage) = \beta_0 + \delta_0 female + \delta_1 married + \delta_3 female \cdot married + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 tenure + \beta_5 tenure^2 + u.$$

where $female{\cdot}married = female \times married.$

id	female	married	female_married
1	1	1	1
2	0	1	0
3	0	0	0
4	1	0	0

```
. gen female_married = female*married

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

Y: lwage, X_1: female, X_2: married, X_2: female_married, X_4: educ, X_5: exper, X_6: expersq, ...

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]	
female	$\hat\delta_0$ -.1103502	.0557421	-1.98	0.048	-.219859	-.0008414
married	$\hat\delta_1$ .2126757	.0553572	3.84	0.000	.103923	.3214284
female_married	$\hat\delta_3$ -.3005931	.071767	-4.19	0.000	-.4415838	-.1596024
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	$\hat\beta_0$ .3213781	.100009	3.21	0.001	.1249041	.5178521

Comments: This type of regression can estimate specific impact of being a married female

→ Sometimes the impact of one variable depends on the value of a dummy variable.
In this case, we suspect that the impact of "being married" may not be the same for male & female.

So, instead of $lwage = \beta_0 + \delta_0 female + \delta_1 married + \text{other vari.}$ assume that the impact of married is the same for ♂, ♀

we can add an interaction

$lwage = \beta_0 + \delta_0 female + \delta_1 married + \delta_2 female \cdot married + \ldots$

δ_1: impact on lwage if married for ♂, ♀

δ_2: additional impact of married on female ♀

$$\log(wage) = \beta_0 + \delta_0 female + \delta_1 married + \delta_3 female \cdot marr.$$

intercept for

single male (singmale) $= \beta_0 = 0.321$

singfemale $= \beta_0 + \delta_0 = 0.321 - 0.11$

marrmale $= \beta_0 + \delta_1 = 0.321 + 0.213$

marrfemale $= \beta_0 + \delta_0 + \delta_1 + \delta_3 = 0.321 - 0.11 + 0.213 - 0.301$

* Calculate the intercepts. They should be the same as the ones on page 85

But for this type of model, we can see the impact of being married for the female group more explicitly.

Case 2 Interaction between a dummy and a continuous variable

The slope of Y & the interacted continuous variable is going to be different for different dummy groups.

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

```
. gen female_educ = female*educ

        Y      X1      X2          X3       . . .
. regress lwage female married female_educ educ exper expersq tenure tenursq
```

Source	SS	df	MS		
Model	65.677852	8	8.2097315	Number of obs =	526
Residual	82.6518994	517	.159868277	F(8, 517) =	51.35
				Prob > F =	0.0000
				R-squared =	0.4428
				Adj R-squared =	0.4342
Total	148.329751	525	.28253286	Root MSE =	.39984

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
female	δ_0 -.2197774	.1675154	-1.31	0.190	-.5488721	.1093172
~~married~~	~~.0529779~~	~~.0407884~~	~~1.30~~	~~0.195~~	~~-.0271535~~	~~.1331092~~
female_educ	δ_1 -.0056186	.0130532	-0.43	0.667	-.0312625	.0200254
educ	β_1 .0813472	.0085008	9.57	0.000	.0646469	.0980476
exper	.0268542	.005335	5.03	0.000	.0163733	.0373351
expersq	-.0005375	.0001124	-4.78	0.000	-.0007583	-.0003167
tenure	.0314803	.0068669	4.58	0.000	.0179898	.0449709
tenursq	-.0005792	.0002351	-2.46	0.014	-.0010412	-.0001173
_cons	.389629	.1186101	3.28	0.001	.1566119	.6226461

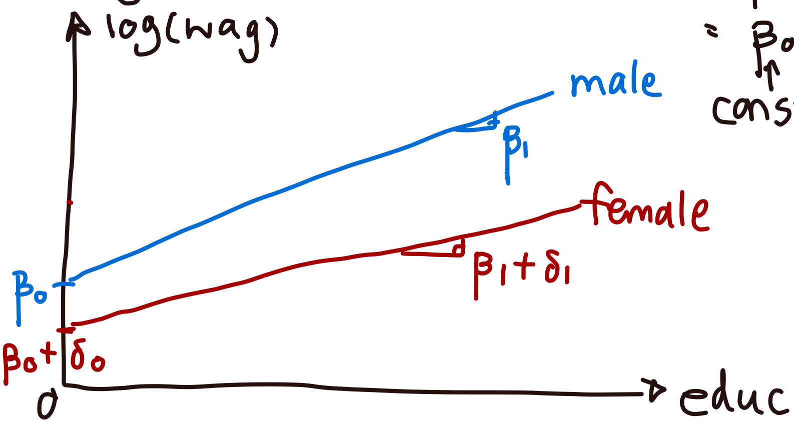
Comments: * Sometimes the impact of one variable depends on the value of a dummy variable. e.g. impact of educ on wage may be different for male & female.

$$E[\log(wage) \mid female=1, x_1, \ldots x_k] = \beta_0 + \delta_0 \cdot 1 + \beta_1 educ + \delta_1 educ + \ldots + u$$

$$= \underbrace{\beta_0 + \overset{(-)}{\delta_0}}_{\text{constant}} + \underbrace{(\beta_1 + \delta_1)}_{\text{slope}} educ + \ldots + u$$

$$E[\log(wage) \mid female=0, x_1, \ldots x_k] = \beta_0 + \delta_0 \cdot 0 + \beta_1 educ + \delta_1\, 0 \cdot educ + \ldots + u$$

$$= \underset{\text{constant}}{\beta_0} + \underset{\text{slope}}{\beta_1} educ + \ldots + u$$



* If the coeff. are significantly different from zero (assume), then, ♀ workers would have lower expected wage & lower returns to education.

7 Testing for Differences in Regression Functions across Groups

- Is it reasonable to believe that the population regression function that explains the dependent variable is the same across subsamples of populations?
- For example, is it reasonable to believe that the function that explains "GPA of college athlete" is the same for male and female students?

Consider

$$cumgpa = \beta_0 + \beta_1 sat + \beta_2 hsperc + \beta_3 tothrs + u,$$

In this case the interaction between female & SAT will allow us to test whether being a female contributes to any different impact of SAT on GPA as compared with male students.

where

$cumgpa$ = cummulative GPA
sat = SAT score
$hsperc$ = high school rank percentile
$tothrs$ = total hours of college courses

If $\delta_1 \neq 0$ → impact of SAT on CumGPA would be different between ♂, ♀

- If we want to test whether "male" students and "female" students have the same values of $\beta_0, \beta_1, \beta_2, \beta_3$ we can estimate the following model

interaction → impact of one variable may depend on the value of another.

$$\begin{aligned} cumgpa \;=\; & \beta_0 + \boxed{\delta_0 female} + \beta_1 sat + \boxed{\delta_1 female \cdot sat} + \beta_2 hsperc + \boxed{\delta_2 female \cdot hsperc} \\ & + \beta_3 tothrs + \boxed{\delta_3 female \cdot tothrs} + u, \end{aligned}$$

If at least one of these δ is $\neq 0$ ⇒ we need to use separate models (diffrent values of the slopes & intercept) for ♂ & ♀

and the null hypothesis would be

We can use same model for all group. →

$$\begin{aligned} H_0 &: \delta_0 = \delta_1 = \delta_2 = \delta_3 = 0 \\ H_a &: otherwise \text{ (at least one } \delta_j \neq 0) \end{aligned} \tag{8.2}$$

We can use the F-test to test this type of null hypothesis:

1. The restricted model (r)

```
. regress cumgpa sat hsperc tothrs
```

Source	SS	df	MS
Model	168.533658	3	56.1778861
Residual	547.364897	728	.751874858
Total	715.898555	731	.979341389

Number of obs =	732
F(3, 728) =	74.72
Prob > F =	0.0000
R-squared =	0.2354
Adj R-squared =	0.2323
Root MSE =	.86711

cumgpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sat	.0009028	.0002079	4.34	0.000	.0004947	.0013109
hsperc	-.0063791	.0015678	-4.07	0.000	-.0094572	-.0033011
tothrs	.0119779	.0009314	12.86	0.000	.0101494	.0138064
_cons	.9291105	.2285515	4.07	0.000	.4804118	1.377809

2. The unrestricted model (ur)

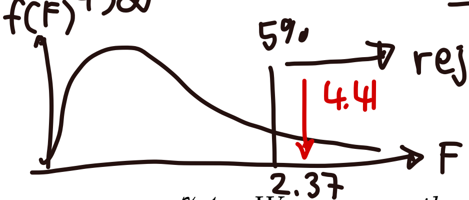
```
. gen female_sat = female*sat

. gen female_hsperc = female*hsperc

. gen female_tothrs = female*tothrs

. regress cumgpa female sat female_sat hsperc female_hsperc tothrs female_tothrs

      Source |       SS           df       MS      Number of obs   =       732
-------------+----------------------------------   F(7, 724)       =     35.15
       Model |  181.589407         7  25.9413439   Prob > F        =    0.0000
    Residual |  534.309148       724   .73799606   R-squared       =    0.2537
-------------+----------------------------------   Adj R-squared   =    0.2464
       Total |  715.898555       731  .979341389   Root MSE        =    .85907

------------------------------------------------------------------------------
       cumgpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+----------------------------------------------------------------
       female |  -1.113638    .528539    -2.11   0.035    -2.15129    -.0759859
          sat |   .0006113    .000235     2.60   0.009    .0001499     .0010727
   female_sat |   .0011167      .0005     2.23   0.026    .0001351     .0020984
       hsperc |  -.0059675   .0017765    -3.36   0.001   -.0094551    -.0024798
female_hsperc |   .0000508   .0041025     0.01   0.990   -.0080035      .008105
       tothrs |   .0103004   .0010928     9.43   0.000    .0081549     .0124459
female_tothrs |   .0055599   .0020696     2.69   0.007    .0014968      .009623
        _cons |   1.213984   .2648281     4.58   0.000    .6940617     1.733907
------------------------------------------------------------------------------
```

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

use 5% sig. level

$F_{4,\infty} = 2.37$

$= \dfrac{(547.36 - 534.31)/4}{534.31/(732-7-1)} = \dfrac{3.26}{0.74} = 4.41$



We reject H_0 of same model for ♂, ♀

So, we will need different models for ♂, ♀ college athletes.

7.1 We can use the "Chow statistics" to test this type of hypothesis as well

- When there are many variables in the model, adding an interaction for every explanatory variable would make the regression analysis messy.

- In which case, we can use the "Chow test" or "Chow statistic" to test the hypothesis expressed in (8.2).

- Chow statistic is a type of F-statistic.

↳ Test for different models. We don't need to create interactions.

$$F = \frac{SSR_p - (SSR_1 + SSR_2)}{SSR_1 + SSR_2} \cdot \frac{[n - 2(k+1)]}{k+1},$$

where n is the total number of observations.

$SSR_p = SSR$ from the pooled model (include observations from both subsamples)

But we will need to estimate 3 models.

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

e.g male athletes in the previous example.
female

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

include only ♂ samples

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

732

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

include only ♀ samples

```
      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: $F \equiv \dfrac{SSR_p - (SSR_1 + SSR_2)}{SSR_1 + SSR_2} \cdot \dfrac{(n - 2(k+1))}{k+1}$

$$\equiv \frac{547.36 - (390.62 + 143.69)}{(390.62 + 143.69)} \cdot \frac{(732 - 2(3+1))}{3+1}$$

$$\equiv 4.42$$

$\rightarrow$ Since $F_{4,\infty}$ at 5% $= 2.37$, we reject H_0 because $4.42 > 2.37$

$\Rightarrow$ We need different models for diff sub groups.

Example from a past seminar paper.

ii. Regression Result

Table 2: Regression Result

VARIABLES	(1) gpax	(2) gpax	(3) gpax
female	0.129**	0.112*	0.0389
	(0.0549)	(0.0580)	(0.162)
hs_gpax	0.415***	0.409***	0.416***
	(0.0622)	(0.0637)	(0.0623)
edu_dad	-0.000535	-0.00168	-0.000472
	(0.00858)	(0.00878)	(0.00855)
edu_mom	-0.0105	-0.0107	-0.0102
	(0.00808)	(0.00841)	(0.00824)
time_sm	-0.0295***	-0.0346***	-0.0510***
	(0.0109)	(0.0117)	(0.0161)
time_smedu	0.0436**	0.0456**	0.0674**
	(0.0189)	(0.0193)	(0.0338)
time_edu	0.0127	0.00992	0.0160
	(0.0243)	(0.0246)	(0.0350)
time_eduexam	0.0107	0.00953	0.00942
	(0.0118)	(0.0120)	(0.0217)
communicate		-0.00638	
		(0.0955)	
update		0.00807	
		(0.0540)	
news		-0.0782	
		(0.0631)	
edu		0.0363	
		(0.0666)	
star		0.0293	
		(0.0627)	
shopping		0.0538	
		(0.0584)	
femalextime_sm			0.0327
			(0.0215)
femalextime_smedu			0.0344
			(0.0403)
femalextime_edu			-0.00574
			(0.0475)
femalextime_eduexam			-0.000853
			(0.0260)
Constant	1.733***	1.821***	1.788***
	(0.244)	(0.263)	(0.261)
Observations	269	269	269
R-squared	0.217	0.226	0.222

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

coefficients ($\hat{\beta}$, $\hat{\delta}$)

std. errs.

Time spent on social media

social media usage purposes

Time spent on social media "interact" with female (gender)

δ_1, δ_2, δ_3, δ_4

H_0: gender does not have any effect on the impact of social media

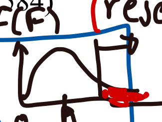
$\delta_1 = \delta_2 = \delta_3 = \delta_4 = 0$

H_a = otherwise.

So, we don't reject H_0

$$F = \frac{(R^2_{UR} - R^2_R)/q}{(1-R^2_{UR})/(n-k-1)} = \frac{(0.222-0.217)/4}{(1-0.222)/(269-12-1)} = 0.411$$

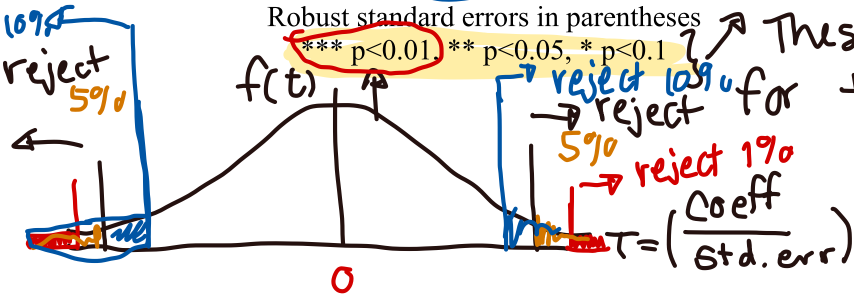
$F_{4,\infty}(5\%) = 2.37$



By only looking at the coefficients, do you think the effects of time spent on social media is different for for different genders??

These p-values are for testing $H_0: \beta = 0$ or the coeff $= 0$

$H_a: \beta \neq 0$



$T = \left(\frac{\text{coeff}}{\text{std. err}}\right)$

*** $p<0.01$ (strongly significant effect)

** $p<0.05$

* $p<0.10$ (10% significant level) (weakly sig)

8 A Binary Dependent Variable (y variable): The Linear Probability Model

Y variable is

- So far, our Y variables are continuous.
- What if we are interested in explaining a qualitative Y variable (that is, Y is a dummy variable)?

Consider

could be continuous or 0,1

$$
\begin{aligned}
y &= \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u \\
E(y|\mathbf{x}) &= \beta_0 + \beta_1 x_1 + ... + \beta_k x_k,
\end{aligned}
$$

where $\mathbf{x}$ denotes all of the explanatory variables $(x_1, ..., x_k)$.

$Y = \begin{cases} 1 & \text{if rain} \\ 0 & \text{if no rain} \end{cases}$, $Y = \begin{cases} 1 & \text{covid+} \\ 0 & \text{if not} \end{cases}$, $Y = \begin{cases} 1 & \text{if using iPad for note-taking} \\ 0 & \text{if not} \end{cases}$

vector of $X_1, X_2, \ldots X_k$

$E(Y|X) = \text{Prob}(Y=1|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$

If this is done in a normal Linear regression

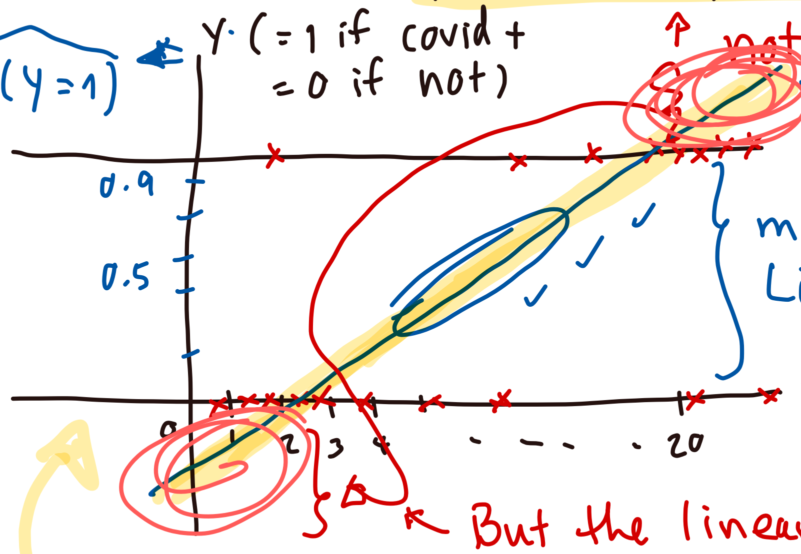
$\Rightarrow$ we can have $\text{Prob}(Y=1|X) > 1$ or < 0

$\hat{Y} = \text{Prob}(Y=1)$

Y: (= 1 if covid+, = 0 if not)

not realistic for prob < 1, < 0

$\hat{Y} \rightarrow \text{Prob}(Y=1)$



make sense in the Linear Probability model

X = (# of times out of the house/week)

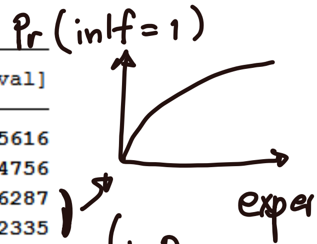
But the linear probability can also give us some unrealistic prediction.

- It is possible for Y to be binary (0,1). Here are some options
  1) Linear Probability Model $\Rightarrow$ same estimation as OLS but treat $\hat{Y} = \text{Prob}(Y=1)$
     However, this can give $\hat{Y} > 1, < 0 \Rightarrow$ unrealistic
  2) Logit, Probit models ! will force $\hat{Y}$ to be $\hat{Y} \in [0,1] \Rightarrow$ don't get unrealistic results.

doing a linear regression (OLS)

y X_1

```
. regress inlf nwifeinc educ exper expersq age kidslt6 kidsge6

      Source |       SS       df       MS              Number of obs =     753
-------------+------------------------------           F(  7,   745) =   38.22
       Model |  48.8080578     7  6.97257969           Prob > F      =  0.0000
    Residual |  135.919698   745  .182442547           R-squared     =  0.2642
-------------+------------------------------           Adj R-squared =  0.2573
       Total |  184.727756   752  .245648611           Root MSE      =  .42713

------------------------------------------------------------------------------
        inlf |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nwifeinc |  -.0034052   .0014485    -2.35   0.019    -.0062488   -.0005616
        educ |   .0379953    .007376     5.15   0.000      .023515    .0524756
       exper |   .0394924   .0056727     6.96   0.000     .0283561    .0506287
     expersq |  -.0005963   .0001848    -3.23   0.001    -.0009591   -.0002335
         age |  -.0160908   .0024847    -6.48   0.000    -.0209686    -.011213
     kidslt6 |  -.2618105   .0335058    -7.81   0.000    -.3275875   -.1960335
     kidsge6 |   .0130122    .013196     0.99   0.324    -.0128935    .0389179
       _cons |   .5855192    .154178     3.80   0.000     .2828442    .8881943
------------------------------------------------------------------------------
```



↳ Because β_{exper} (+)

β_{exper^2} (−)

in the labor force (working)

where

inlf = 1 if the woman reports working for a wage outside the home at some point during the year, zero otherwise.

X_1 *nwifeinc* = husband's earnings (in thousands of dollars)

educ = years of education

exper = past years of labor market experience

age = age

kidslt6 = number of children less than 6 years old

kidsage6 = number of kids between 6 - 18 years old

$$\widehat{\text{Prob}(inlf = 1 \mid X_1, X_2, \ldots X_K)} = \hat{\beta}_0 + \hat{\beta}_1 nwifeinc + \hat{\beta}_2 educ + \hat{\beta}_3 exper \ldots$$

$$= 0.586 - 0.003\, nwifeinc + 0.038\, educ + \ldots$$

- $$\frac{d\,\widehat{\text{Prob}(inlf=1 \mid X_1, \ldots X_K)}}{d\, nwifeinc} = \hat{\beta}_1 = -0.003$$

  → every USD 1,000 increase in the husband's income would decrease the prob that the wife works by 0.003.

- $$\frac{d\,\widehat{\text{Prob}(inlf=1 \mid X_1, \ldots X_K)}}{d\, kidslt6} = \hat{\beta}_{kidslt6} = -0.262$$

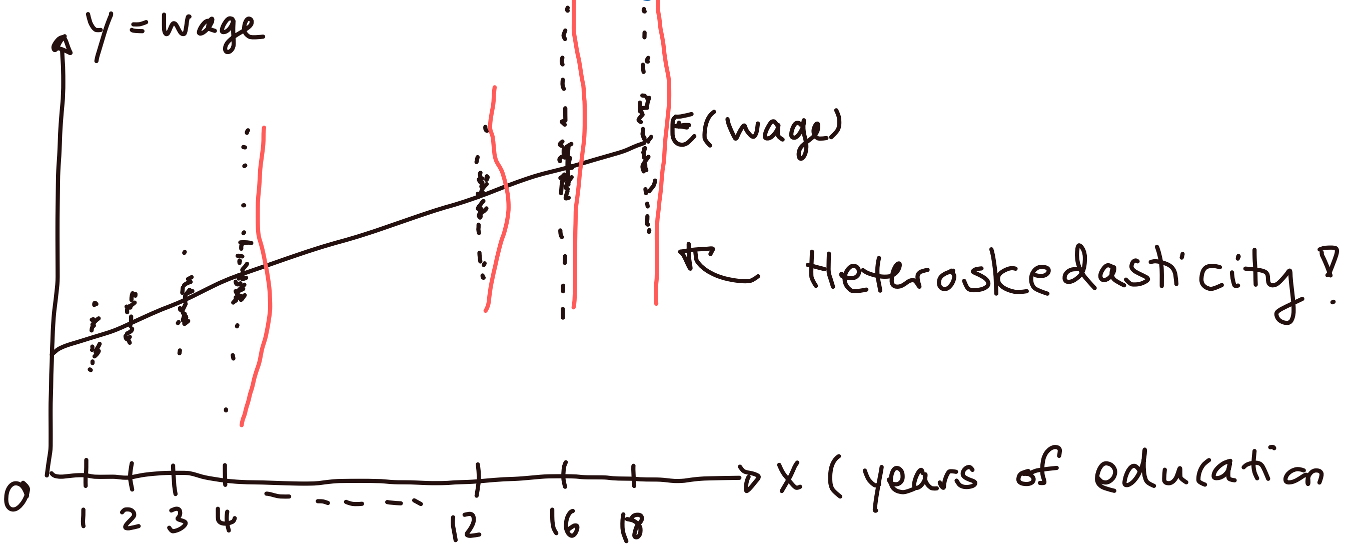
  → Each additional kid under 6 years old would decrease the prob that the wife works by 0.262 (26%)

Heteroscedasticity Problem

1 Nature and Consequences of heteroscedasticity for OLS

- Heteroskedasticity (broadly) - different sub-population / sub-samples have different variability from others.
- Heteroskedasticity (in econometrics) - variance of the error term (u) (unobserved factors) changes across different sub-populations (or changes according to different values of X variables

1.1 *Nature of Heteroskedasticity*



$$Var(u|X_{edu}=1) \neq Var(u|X_{edu}=2) \neq \dots Var(u|X_{edu}=12) \neq \dots$$

$$\boxed{\text{Heteroskedasticity}: Var(u_i|X_i) = \sigma_i^2}$$

variance depends on the value of X

1.2 *Consequences of Heteroskedasticity*

1. Does not affect the biasedness of the OLS estimators

- We only need assumption MLR 1 – 4 to be satisfied. Heteroskedasticity problem violates ass. MLR 5. So, as long as MLR 1 – 4 are satisfied, $\hat{\beta}_{OLS}$ will be unbiased.

Borrowed from Aj. Wasin's note
Please insert between p.95-96

Figure 9.1: Homoscedasticity

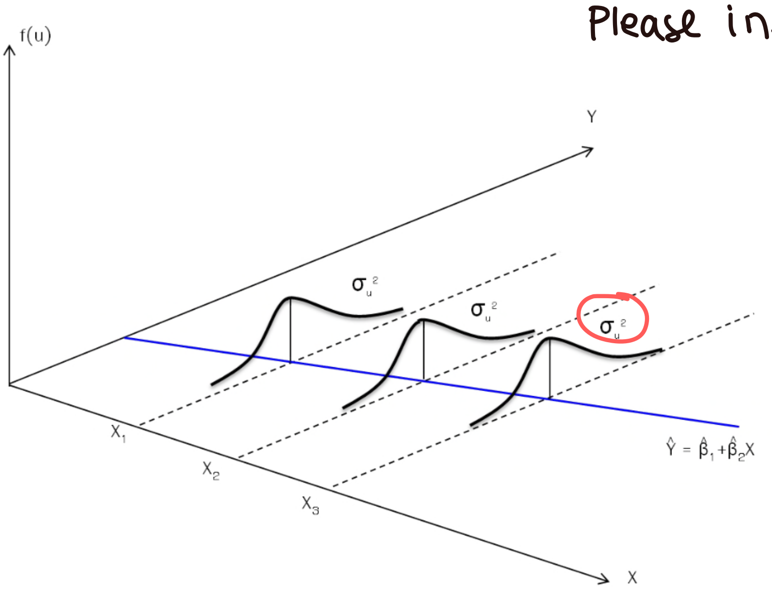
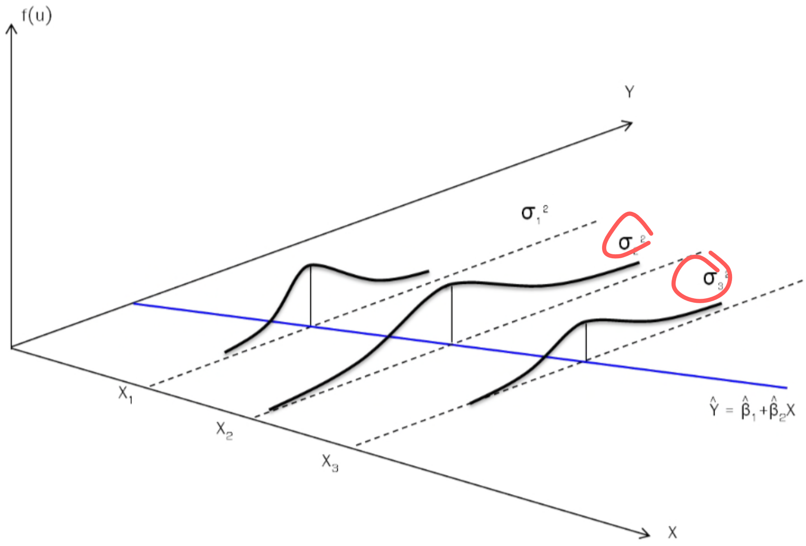


Figure 9.2: Heteroscedasticity



Generally, there is a variety of causes for the existence heteroscedasticity problem in the regression model studied by most economists. Yet, only 6 main sources are discussed here.

1. Normally, error learning is the nature of human. At the initial stage of their work, people probably commit a large number of mistakes. As they carry on working and become more specialized, the amount of errors would be reduced. In this case, it seems that the variance at the initial stage will be high but would decrease as people learn from their errors.

2. Considering the relationship between independent variable X and dependent variable Y, there seems to be possible that as the value of independent variable increase, the variance of the value of dependent one will increases. The feasible reason is the characteristics of those variables such as the relationship between the profit X of the company (independent variable) and dividend Y (dependent variable). As the profit rises, the board of director may have a variety of dividend policies. Some companies may pay a small amount of dividend in order to keep the profit for further development. Some may pay a large amount of dividend to satisfy the shareholders. On the contrary, if the profit is low, the dividend policy will not diverge across the companies since the companies seem to be at the growth stage and decide to keep their profit as retained earnings.

3. The collection of data is another source. As the collection technique employed by the researcher is improved, the collection error would be lower. Contrarily, with the poor collection technique, the data obtained to construct the regression model would probably incur more and more errors, causing the condition variance of disturbance term to vary.

some values that are far-off from expected values

4. The existence of outliers in the independent and dependent variables may make the conditional variance of disturbance term on independent variables volatile. Mostly, if the researchers collect too small amount of data, that set of data tends to include the outliers and undermine the regression analysis. As the amount of data increases, those outliers may become normal relative to other observations.

5. The misspecification of the model could result in the heteroscedasticity problem as well. In some cases, econometricians drop some important and necessary independent variable from the regression model. The disturbance term, then, will incorporate the characteristics of the missing variables, resulting in heteroscedasticity problem. For example, suppose the researchers want to establish the model to explain the relationship of price and quantity of good X, as suggested by the theory of demand. However, if it turns out that good Y is the substitute for good X. This mistake of failing to include price of good Y, which has the explanatory power over the demand for good X, will result in misspecification error. The disturbance term will have the characteristics of good Y, which, in turn, leads to heteroscedasticity.

6. Heteroscedasticity problem usually happens in the model that applies the **cross-sectional data** which tends to be highly diverse because the data is collected in the same time period. To illustrate, the census acquired from numerous provincial areas may cover the wide range of value and results in the stated problem. On the other hand, for the **time series data**, it is prone to be the collection of the same sample for different period of time. With the same sample, the range of the value covered seems to be narrow. Hence, the stated problem, generally, does not occur with this kind of data.

2. Does not affect the value of R^2 and $adj.R^2$

We can still refer to R^2, Adj. R^2

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \approx 1 - \frac{\sigma_u^2}{\sigma_y^2}$$

These two are the variance of the whole dataset (all samples) so, are not affected by the difference of $\sigma^2_{u_i}$ across groups.

3. Make the estimated value of $Var(\hat{\beta}_{OLS})$ wrong

→ And because $\sqrt{Var(\hat{\beta}_{OLS})}$ is wrong, our estimate for $\sqrt{Var(\hat{\beta}_{OLS})}$ = std. err. will be wrong

F-test uses std.err. too. So, it will be invalid as well.

4. Affect the correctness of our inference

$H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$

$$t = \frac{\hat{\beta}_1 - \beta_1}{\text{std.err}}$$

$\sqrt{Var(\hat{\beta}_{OLS})}$

So, if std.err is incorrect, t-statistic will be ~. We cannot rely on our conclusion about the hypothesis testing.

$H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$

1.3 How can the estimated value of $Var(\hat{\beta}_{OLS})$ be wrong?

Suppose

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

Given that assumption 1 to 4 are true, but assumption 5 (homoskedasticity) is violated. Thus,

$$Var(u_i|x_i) = \sigma_i^2$$

indicates heterosk. because it says that variance depends on the value of x_i

And from the OLS estimation steps, we can write

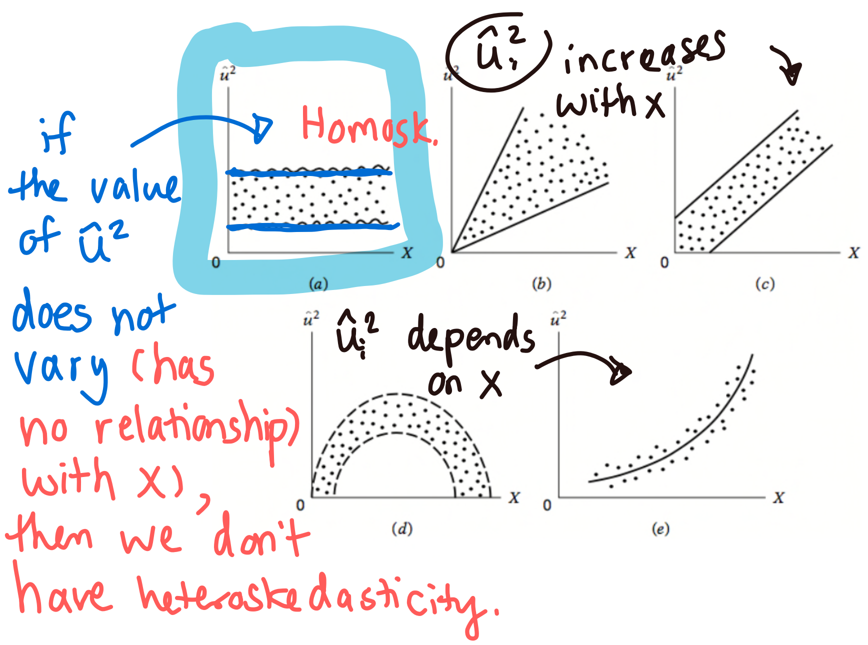
look up $\boxed{\text{P.36-7}}$

$$\boxed{\hat{\beta}_1 = \beta_1 + \frac{\sum_i (x_i - \bar{x}) u_i}{\sum_i (x_i - \bar{x})^2}}$$

$$Var(\hat{\beta}_1) = 0 + Var\left(\frac{\sum_i (x_i - \bar{x}) u_i}{\sum_i (x_i - \bar{x})^2}\right)$$

Figure 9.4: The relationship between u_i^2 and regressor X of the model suffering heteroscedasticity problem

proxy of σ^2



using assumption 4 (MLR, SLR), x is conditional upon (is given) → constant for each observation

We get

$$Var(\hat{\beta}_1) = \left[\frac{1}{(\sum_i (x_i-\bar{x})^2)}\right]^2 Var\left[\sum_i (x_i-\bar{x})u_i\right]$$

$$= \left[\frac{1}{\sum_i (x_i-\bar{x})^2}\right]^2 Var\left[(x_1-\bar{x})u_1 + (x_2-\bar{x})u_2 + \dots + (x_k-\bar{x})u_k\right]$$

$$= \frac{1}{SST_x^2}\cdot\left[(x_1-\bar{x})^2 \overbrace{Var(u_1)}^{\sigma_1^2} + (x_2-\bar{x})^2 \overbrace{Var(u_2)}^{\sigma_2^2} + \dots\right]$$

Heterosk.

$$\boxed{Var(\hat{\beta}_1) = \frac{1}{SST_x^2}\sum (x_i-\bar{x})^2\sigma_i^2 \quad \text{under heteroskedasticity}}$$

$$Var(\hat{\beta}_1) = \frac{1}{SST_x^2}\underbrace{\sum (x_i-\bar{x})^2}_{SST_x}\sigma^2 = \frac{\sigma^2}{SST_x} \quad \text{under homoskedasticity}$$

* If we have heterosk., we cannot use the usual OLS calculation for $Var(\hat{\beta})$, $\widehat{Var(\hat{\beta})}$ ↗ std. err.

1.4 Two types of remedies

1. Passive – We do not fix or alter the data used for estimation. We only employ the correct calculation for the std. err.

i → j

id	x_1 (educ)	x_j (inc)
1		
2		
3		
4		
5		
100		

$$Var(\hat{\beta}_j) = \frac{\sum_i (x_{ij}-\bar{x})^2\sigma_i^2}{SST_x^2(1-R_j^2)}$$

The R^2 from regressing x_j on all other xs.

EX: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$ (age)

$$Var(\hat{\beta}_3) = \frac{\sum_i (x_{i3}-\bar{x})\sigma_i^2}{SST_x^2(1-R_3^2)}$$

from $x_3 = \theta_0 + \theta_1 x_1 + \theta_2 x_2 +$

estimator

$$\boxed{\widehat{Var(\hat{\beta}_j)} = \frac{\sum_i (\hat{r}_{ij}^2\,\hat{u}_i^2)}{SSR_j^2} \quad \text{(std. err.)}}$$

called Heteroskedasticity-robust std. err.

where $\hat{r}_{ij}$ = i^{th} residual from regressing x_j on all other xs.

$\hat{u}_i$ = error ———————— y on all xs. (not y) (from the main regression)

SSR_j = SSR from regressing x_j on all other xs.

STATA command: regress y x1 x2 ... xk, robust

2. Active - manipulate (change / alter / transform) the data so that we will have the homoske. pattern in the dataset. ⇒ use the usual OLS std. err.

⇝ use methods such as weighted-least-squares.

2 Testing for heteroskedasticity

- The main point - Heteroskedasticity happens when $Var(u_i|x_i) = \sigma_i^2$ ⇝ variance depends on X

However, for the heterosk. test, we don't know the true value of σ_i^2. So, we are going to use $\hat{u}_i^2$ as a proxy for σ_i^2 !

Suppose

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

want to test if $Var(u|x_1,..x_k)$ is constant.

Assume that assumption 1 to 4 are true. Our hypotheses to test for heteroskeroskedasticity would be

$\hat{\beta}_s$ are unbiased.

$H_0: Var(u|x_1, x_2, ..., x_k) = \sigma^2$

H_a: H_0 is not true ⇝ Heterosk.

Tip: $=, \leq, \geq$ should be with H_0

⇒ if we reject H_0, we have heterosk.

We know that $Var(u|\mathbf{x}) = E(u^2|\mathbf{x}) - [E(u|\mathbf{x})]^2$. But $E(u|X) = 0$ (MLR 4) according to assumption 4. Thus, H_0 and H_a can be written as

$H_0: E(u^2|x_1, x_2 ... x_k) = \sigma^2$

H_a: H_0 is not true

2.1 Breusch-Pagan test (BP test)

To test the above hypothesis, we follow these steps.

1) We need to find out whether u^2 is correlated with X. This can be done by regressing

$$u^2 = \delta_0 + \delta_1 X_1 + \dots \delta_k X_k + \text{error}$$

population parameter b/c u = real error

Then perform a joint-hypothesis testing for

$$H_0: \delta_1 = \delta_2 = \delta_3 = \dots = \delta_k = 0 \quad \text{(homosk.)}$$

against H_a: H_0 is not true. So, if at least one of the $\delta \neq 0$, we already have heterosk.

2) We don't know the true value of u^2. So, we use $\hat{u}^2$ instead.

$R^2_{\hat{u}^2}$ is the R^2 from this regression

$$\hat{u}^2 = \delta_0 + \delta_1 X_1 + \dots + \delta_k X_k + \text{error} \quad (1)$$

Then, use F-test

$$F_{k,\, n-k-1} = \left[R^2_{\hat{u}^2}/k\right] \Big/ (1-R^2_{\hat{u}^2})/(n-k-1)$$

# of observation

# of regressors in (1)

To perform the Breusch-Pagan Test in STATA

STATA commands (in case $k = 4$):

```
regress y x1 x2 x3 x4
predict u_hat, residual
generate u_hat_sq = u_hat^2
regress u_hat_sq x1 x2 x3 x4
```

** Then, check the F-statistic on the top right-hand corner of the result table.

Example: Finding the determinants of GPA.

Or use the Lagrange-multiplier test (LM)

```
. regress  termgpa attend priGPA final frosh soph
```

Source	SS	df	MS
Model	226.077541	5	45.2155081
Residual	142.319996	674	.211157264
Total	368.397537	679	.542558964

Number of obs =	680
F(5, 674) =	214.13
Prob > F =	0.0000
R-squared =	0.6137
Adj R-squared =	0.6108
Root MSE =	.45952

termgpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
attend	.046594	.0036082	12.91	0.000	.0395093	.0536787
priGPA	.5329307	.0403281	13.21	0.000	.4537468	.6121146
final	.0503197	.0040339	12.47	0.000	.0423992	.0582403
frosh	.0974307	.0560211	1.74	0.082	-.0125662	.2074276
soph	.0689273	.0467006	1.48	0.140	-.0227689	.1606236
_cons	-1.361077	.1316861	-10.34	0.000	-1.619642	-1.102513

$LM \sim \chi^2_k$

$LM = n \cdot R^2_{\hat{u}^2}$

if $LM >$ critical value in the χ^2-distribution, we reject H_0