Chapter 5

Dummy Variable

Problem with Chow Test

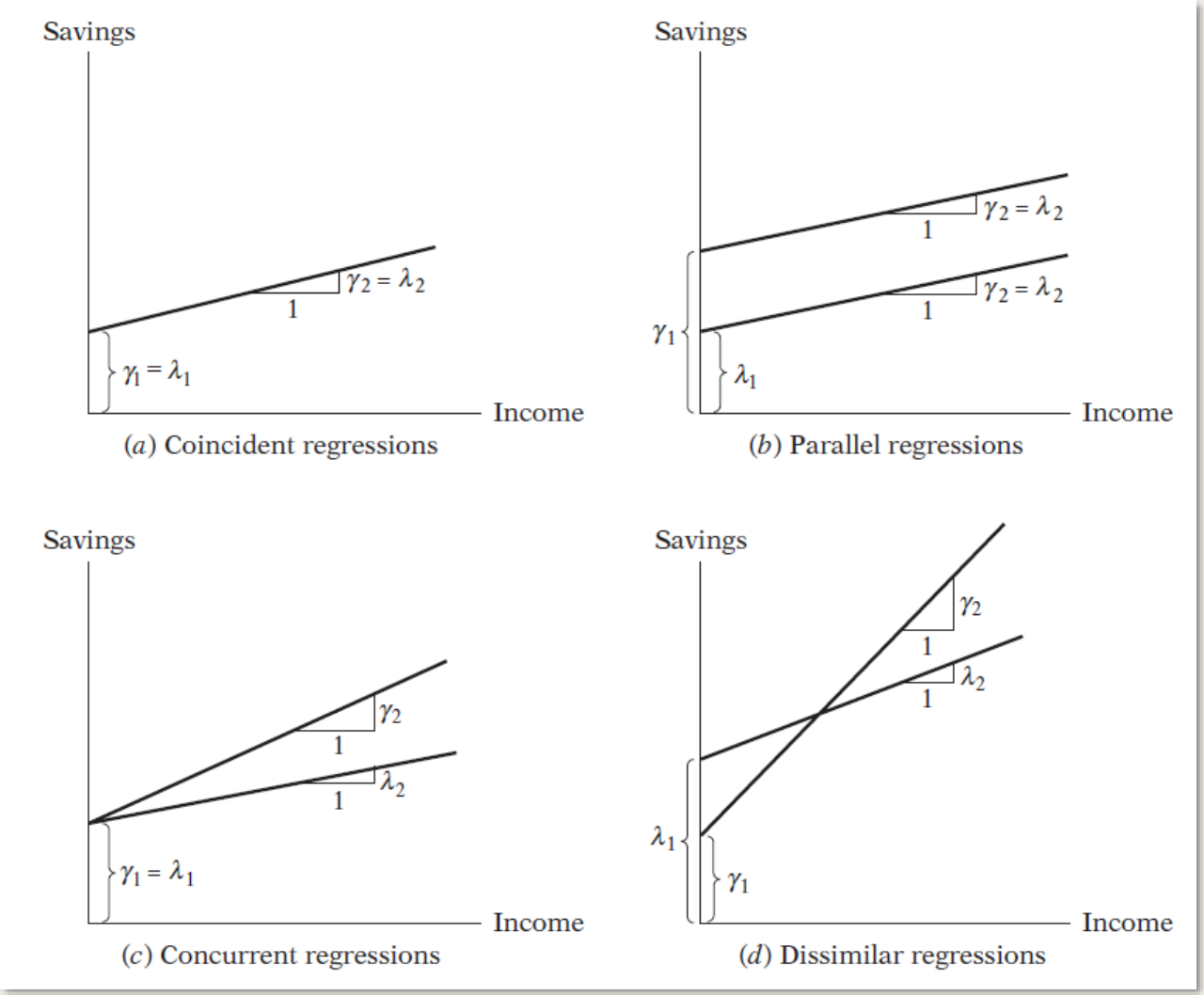
Recalling the Chow Test, a test for structural change, let's see all the possibilities from ex-ante and post crisis.

- $Y_t = \lambda_1 + \lambda_2 X_t + u_{1t}$ $\quad n_1 = 12$
- $Y_t = \gamma_1 + \gamma_2 X_t + u_{2t}$ $\quad n_2 = 14$
- $Y_t = \beta_1 + \beta_2 X_t + u_t$ $\quad n = (n_1 + n_2) = 26$

As discussed earlier, the major for overall test is that F-test is usually very general, when a null hypothesis is rejected.

Though a null hypothesis is rejected, we still do not know what and how, in this case, λ_1, γ_1 and λ_2, γ_2 are different. We would know if we keep nesting the F-test, which is way too much work.

Problem with Chow Test



Problem with Chow Test

If we can include a variable that separates ex-ante and post crisis period in a single equation, that would be ideal because we can see a difference, or no difference between pre and post crisis. We can see its significance with a single t-test.

Not only that, if we can include another variable that can capture the slope for pre and post crisis, that is also very helpful since we can test its significance with a t-test as well.

We are gradually going to implement this concept step-by-step.

(1) ANOVA *model*

So far, we have only dealt with continuous variables (weight, height, income, price, quantity, temperature, etc.) for both dependent and independent variables.

A natural problem arises since we know that there are so many real-world variables that is '**qualitative**'. A basic and most upfront example is gender. Consider our expenditure model here.

- $pexp_i = \beta_1 + \beta_2 sex_i + u_i$

where $pexp_i$ is monthly personal expenditure and sex_i is a binary variable, therefore there are only two possible encodings either

- sex_i = otherwise
- sex_i = female

This model is called ANOVA model, or a model containing only quantitative variables or **dummy variable**.

(1) ANOVA *model*

Given that the result of this regression model is

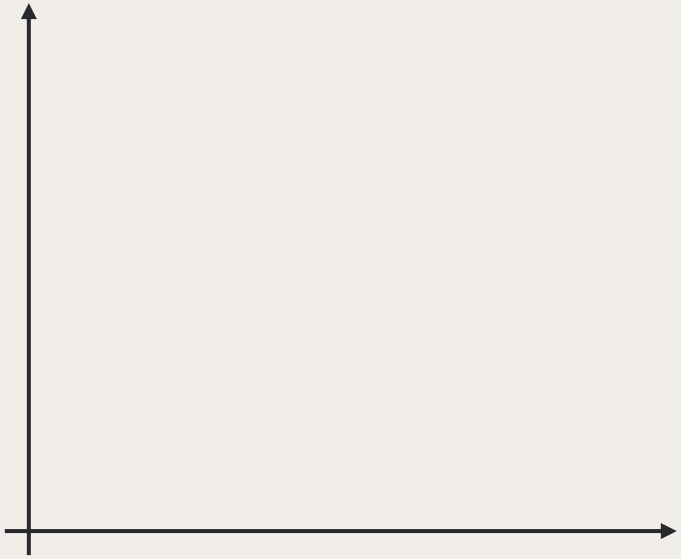
- $\widehat{pexp}_i = 11{,}871.43 - 6{,}890.347 sex_i$

First, we look at how this result should be interpreted by considering the expected value. Note that we encode $sex_i = 0$ for otherwise and $sex_i = 1$ for female

- $E(\widehat{pexp}_i | sex_i = 0) =$

- $E(\widehat{pexp}_i | sex_i = 1) =$

(1) ANOVA *model*



If we plot each expected value, we see a difference between each group on this graph.

Data for the estimation are in this table. To distinguish gender, we only need one dummy variable to accommodate this difference. To be precise, we need only $n - 1$ dummy variable(s) to incorporate n groups of the sample.

(2) *More than two categories*

Now let's assume that we will use another dummy variable, region, so we have 5 groups which are Bangkok and vicinity, center, north, northeast, south. Only one option can be chosen among these. Given that

- $D_{2i} = 0$ for otherwise ; $D_{2i} = 1$ for center
- $D_{3i} = 0$ for otherwise ; $D_{3i} = 1$ for north
- $D_{4i} = 0$ for otherwise ; $D_{4i} = 1$ for northeast
- $D_{5i} = 0$ for otherwise ; $D_{5i} = 1$ for south

The estimated model becomes

- $\widehat{pexp}_i = \hat{\beta}_1 + \hat{\beta}_2 D_{2i} + \hat{\beta}_3 D_{3i} + \hat{\beta}_4 D_{4i} + \hat{\beta}_5 D_{5i}$

(2) More than two categories

Find the expected value and interpretation for

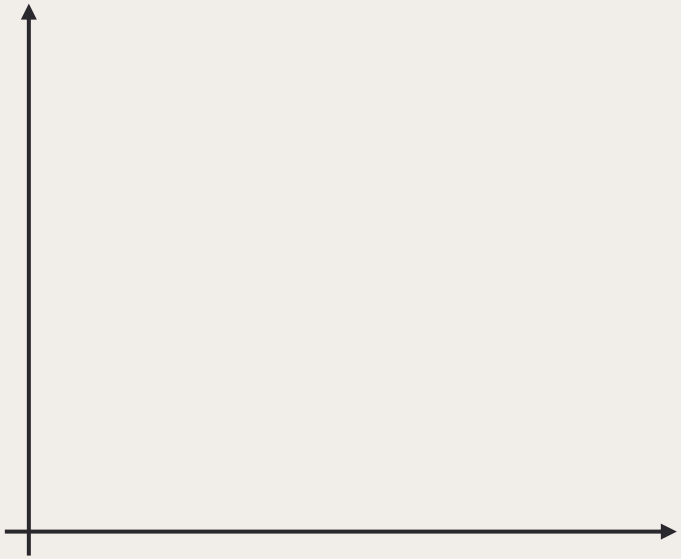
- $E(\widehat{pexp}_i | D_{2i} = 0; D_{3i} = 0; D_{4i} = 0; D_{5i} = 0) =$
- $E(\widehat{pexp}_i | D_{2i} = 1; D_{3i} = 0; D_{4i} = 0; D_{5i} = 0) =$
- $E(\widehat{pexp}_i | D_{2i} = 0; D_{3i} = 1; D_{4i} = 0; D_{5i} = 0) =$
- $E(\widehat{pexp}_i | D_{2i} = 0; D_{3i} = 0; D_{4i} = 1; D_{5i} = 0) =$
- $E(\widehat{pexp}_i | D_{2i} = 0; D_{3i} = 0; D_{4i} = 0; D_{5i} = 1) =$

(2) *More than two categories*

Given that the result of this regression model is

- $\widehat{pexp}_i = 6{,}948.889 - 5{,}298.889D_{2i} + 4{,}551.111D_{3i} - 1{,}198.889D_{5i}$

Plot each group onto this graph. Note that we don't have any student living in the northeast.



(3) *Two or more dummy variables*

Let's say we now have 2 quantitative variables, sex and self-declared personality (extrovert/introvert)

- $D_{2i} = 0$ for otherwise ; $D_{2i} = 1$ for female
- $D_{3i} = 0$ for otherwise ; $D_{3i} = 1$ for introvert

The estimated model becomes

- $\widehat{pexp}_i = \hat{\beta}_1 + \hat{\beta}_2 D_{2i} + \hat{\beta}_3 D_{3i}$

(3) *Two or more dummy variables*

Find the expected value and interpretation for

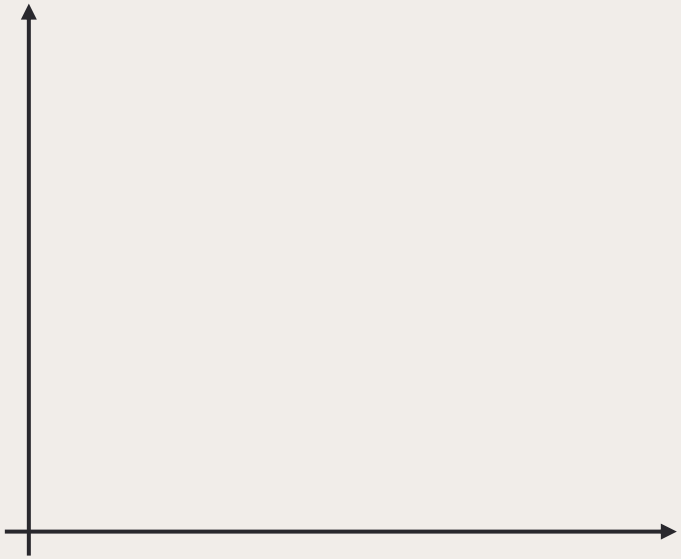
- $E(\widehat{pexp}_i|D_{2i}=0;D_{3i}=0)=$
- $E(\widehat{pexp}_i|D_{2i}=1;D_{3i}=0)=$
- $E(\widehat{pexp}_i|D_{2i}=0;D_{3i}=1)=$
- $E(\widehat{pexp}_i|D_{2i}=1;D_{3i}=1)=$

(3) *Two or more dummy variables*

Given that the result of this regression model is

- $\widehat{pexp}_i = 14{,}527.16 - 7{,}170.928D_{2i} - 3{,}380.019D_{3i}$

Plot each group onto this graph.



(4) ANCOVA *model*

Regression models containing a mixture of both types of variables are called **ANCOVA models**. (Analysis of covariance)

Let's go back to our real sample of the expenditure model, now we include height (a quantitative continuous variable) and gender (a qualitative discrete variable) in this model as follows.

- $pexp_i = \beta_1 + \beta_2 pinc_i + \beta_3 sex_i + u_i$

where $pexp_i$ is monthly personal expenditure, $pinc_i$ is monthly personal income and sex_i is a binary variable as usual.

For this ANCOVA model, we have both quantitative and qualitative variables included. Each of them determines shoe size in a different way. Therefore, we need to interpret both height and gender that coexist in the same model.

(4) ANCOVA *model*

Given that the result of this regression model is

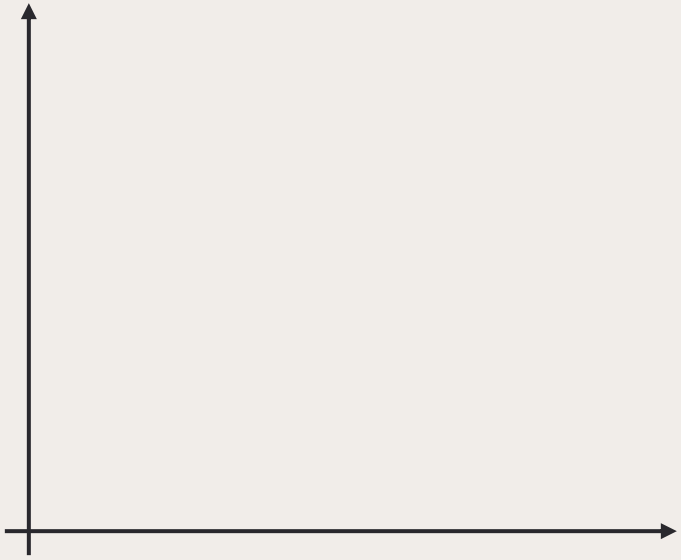
- $\widehat{pexp}_i = 6{,}604.296 + 0.2178 pinc_i - 3{,}195.217 sex_i$

Now consider each case when

- $E(\widehat{pexp}_i | sex_i = 0) =$

- $E(\widehat{pexp}_i | sex_i = 1) =$

(4) ANCOVA *model*



If we plot each expected value, we see a difference between each group on this graph.

(5) *Additional example*

Let's consider this example of two or more dummy variables. Given that we are considering two models with two dummies

- **1st model**: › $pexp_i = \beta_1 + \beta_2 sex_i + \beta_3 ei_i + u_i$
- **2nd model**: › $pexp_i = \beta_1 + \beta_2 sei_{2i} + \beta_3 sei_{3i} + \beta_4 sei_{4i} + u_i$

where sex_i is binary gender (0=male; 1=otherwise), ei_i is binary self-declared personality of extrovert or introvert (0=extrovert; 1=otherwise), and sei_i represent a cross variable

when sei_i = 0 : male, extrovert
= 1 : female, extrovert
= 2 : male, introvert
= 3 : female, introvert

(5) *Additional example*

Here are the results of tabulation.

```
. tab sex ei

Biological |  Extrovert/Introvert
    gender | Extrovert  Introvert |     Total
-----------+----------------------+----------
      Male |         3         11 |        14
    Female |        11         26 |        37
-----------+----------------------+----------
     Total |        14         37 |        51
```

```
. tab sei

      Interacted |
        variable |      Freq.     Percent        Cum.
-----------------+-----------------------------------
  Male/Extrovert |          3        5.88        5.88
Female/Extrovert |         11       21.57       27.45
  Male/Introvert |         11       21.57       49.02
Female/Introvert |         26       50.98      100.00
-----------------+-----------------------------------
           Total |         51      100.00
```

(5) *Additional example*

```
. reg pexp i.sex i.ei

      Source |       SS           df       MS      Number of obs   =        51
-------------+----------------------------------   F(2, 48)        =      8.56
       Model |   597454013         2   298727006   Prob > F        =    0.0007
    Residual |  1.6755e+09        48  34907240.7   R-squared       =    0.2628
-------------+----------------------------------   Adj R-squared   =    0.2321
       Total |  2.2730e+09        50  45460031.4   Root MSE        =    5908.2

------------------------------------------------------------------------------
        pexp | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         sex |
      Female |  -7170.928   1860.285    -3.85   0.000    -10911.28   -3430.574
             |
          ei |
   Introvert |  -3380.019   1860.285    -1.82   0.075    -7120.373    360.3346
       _cons |   14527.16   2151.698     6.75   0.000     10200.88    18853.44
------------------------------------------------------------------------------
```

```
. reg pexp i.sei

      Source |       SS           df       MS      Number of obs   =        51
-------------+----------------------------------   F(3, 47)        =      8.61
       Model |   806209960         3   268736653   Prob > F        =    0.0001
    Residual |  1.4668e+09        47  31208332.1   R-squared       =    0.3547
-------------+----------------------------------   Adj R-squared   =    0.3135
       Total |  2.2730e+09        50  45460031.4   Root MSE        =    5586.4

------------------------------------------------------------------------------
             pexp | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
------------------+----------------------------------------------------------------
              sei |
Female/Extrovert  |  -15409.09   3638.667    -4.23   0.000    -22729.14   -8089.037
  Male/Introvert  |  -11618.18   3638.667    -3.19   0.003    -18938.24   -4298.128
Female/Introvert  |  -16276.92   3406.332    -4.78   0.000    -23129.58   -9424.267
                  |
            _cons |      21000   3225.334     6.51   0.000     14511.47    27488.53
------------------------------------------------------------------------------
```

(5) *Additional example*

Let's sum up the results of expenditure to see that these two models are different.

Group	1st model	2nd model
1. Male / extrovert		
2. Female / extrovert		
3. Male / introvert		
4. Female / introvert		

Though we can list groups from both models, the results from the first one only shows **partial difference, not an interaction** between two dummies.

(1) *Dummy and dummy*

Dummy variables can be crossed to study differential effect from two or more dummies stacked together. Consider the following example of the same equation, instead we add a cross-product term here.

- $pexp_i = \beta_1 + \beta_2 sex_i + \beta_3 ei_i + \beta_4 (sex_i \cdot ei_i) + u_i$

Where β_4 represents **additional effect**. The term is called **interaction dummy**, effect of the two attributes considered individually.

Therefore, when both dummies or either sex_i or ei_i is zero, or both, β_4 will be zero.

Though it seems that this coefficient can then add the case of **Female/introvert** (when sex_i and ei_i are both 1), once we add this interaction term into this equation, both variables are considered interacted. See the comparison of results in the upcoming slide.

(1) *Dummy and dummy*

```
. reg pexp i.sex i.ei sex#ei
      Source |       SS           df       MS      Number of obs   =        51
-------------+----------------------------------   F(3, 47)        =      8.61
       Model |   806209960         3   268736653   Prob > F        =    0.0001
    Residual |  1.4668e+09        47  31208332.1   R-squared       =    0.3547
-------------+----------------------------------   Adj R-squared   =    0.3135
       Total |  2.2730e+09        50  45460031.4   Root MSE        =    5586.4
------------------------------------------------------------------------------
             pexp | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
------------------+---------------------------------------------------------------
              sex |
           Female |  -15409.09   3638.667    -4.23   0.000    -22729.14   -8089.037
                  |
               ei |
         Introvert |  -11618.18   3638.667    -3.19   0.003    -18938.24   -4298.128
                  |
           sex#ei |
Female#Introvert |   10750.35   4156.602     2.59   0.013     2388.345    19112.35
                  |
            _cons |      21000   3225.334     6.51   0.000     14511.47    27488.53
------------------------------------------------------------------------------
```

```
. reg pexp i.sei
      Source |       SS           df       MS      Number of obs   =        51
-------------+----------------------------------   F(3, 47)        =      8.61
       Model |   806209960         3   268736653   Prob > F        =    0.0001
    Residual |  1.4668e+09        47  31208332.1   R-squared       =    0.3547
-------------+----------------------------------   Adj R-squared   =    0.3135
       Total |  2.2730e+09        50  45460031.4   Root MSE        =    5586.4
------------------------------------------------------------------------------
             pexp | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
------------------+---------------------------------------------------------------
              sei |
Female/Extrovert |  -15409.09   3638.667    -4.23   0.000    -22729.14   -8089.037
  Male/Introvert |  -11618.18   3638.667    -3.19   0.003    -18938.24   -4298.128
Female/Introvert |  -16276.92   3406.332    -4.78   0.000    -23129.58   -9424.267
                  |
            _cons |      21000   3225.334     6.51   0.000     14511.47    27488.53
------------------------------------------------------------------------------
```

(2) *Dummy and continuous variable*

A dummy variable can be crossed with another continuous variable as well. Given that

- $pexp_i = \beta_1 + \beta_2 pinc_i + \beta_3 sex_i + \beta_4 (pinc_i \cdot sex_i) + u_i$

Let's find the expected value from estimated model.

- $E(\widehat{pexp}_i | sex_i = 0) =$

- $E(\widehat{pexp}_i | sex_i = 1) =$

(2) *Dummy and continuous variable*

```
. reg pexp pinc i.sex c.pinc#sex

      Source |       SS           df       MS      Number of obs   =        51
-------------+----------------------------------   F(3, 47)        =     16.52
       Model |  1.1668e+09         3   388924972   Prob > F        =    0.0000
    Residual |  1.1062e+09        47  23536737.3   R-squared       =    0.5133
-------------+----------------------------------   Adj R-squared   =    0.4823
       Total |  2.2730e+09        50  45460031.4   Root MSE        =    4851.5

------------------------------------------------------------------------------
        pexp | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        pinc |   .2048834   .0428632     4.78   0.000     .1186537    .2911131
             |
         sex |
      Female |  -5511.089   2332.389    -2.36   0.022    -10203.25    -818.927
             |
  sex#c.pinc |
      Female |   .2904635    .202925     1.43   0.159     -.117769     .698696
             |
       _cons |    6917.64   1659.898     4.17   0.000     3578.355    10256.92
------------------------------------------------------------------------------
```

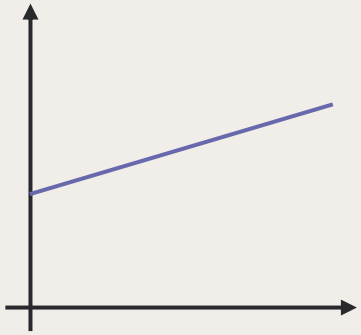
(2) *Dummy and continuous variable*

The model above is actually another type of Chow Test.

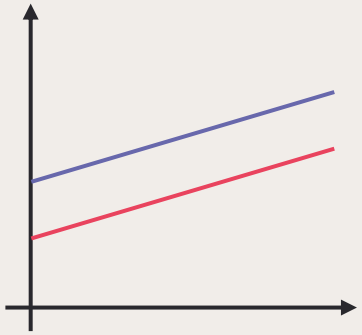
- $\hat{\beta}_1$ is the base intercept while $\hat{\beta}_3$ is the additional intercept for female.

- $\hat{\beta}_2$ is the base slope while $\hat{\beta}_4$ is the additional slope for female.

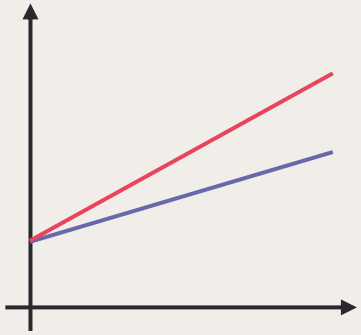
(a) Coincident regression



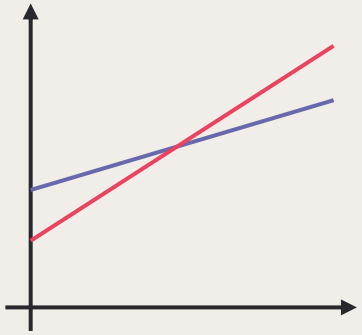
(b) Parallel regression



(c) Concurrent regression



(d) Dissimilar regression



(2) *Dummy and continuous variable*

Let's drop the interaction term and see the result below.

```
. reg pexp pinc i.sex

      Source |       SS           df       MS      Number of obs   =        51
-------------+----------------------------------   F(2, 48)        =     23.25
       Model |  1.1186e+09         2   559275739   Prob > F        =    0.0000
    Residual |  1.1545e+09        48  24051043.6   R-squared       =    0.4921
-------------+----------------------------------   Adj R-squared   =    0.4709
       Total |  2.2730e+09        50  45460031.4   Root MSE        =    4904.2

------------------------------------------------------------------------------
        pexp | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        pinc |    .217843   .0423514     5.14   0.000     .1326898    .3029962
             |
         sex |
      Female |  -3195.217   1698.243    -1.88   0.066    -6609.763    219.3287
       _cons |   6604.296    1663.28     3.97   0.000     3260.048    9948.544
------------------------------------------------------------------------------
```

Log-linear interpretation of dummy variable

We are interested in the change of Y when a dummy variable D switch from 0 to 1. Given a regression function of

- $\ln Y_i = \beta_1 + \beta_2 D_i + u_i$ when $D_i = 0,1$ or $Y_i = e^{\beta_1 + \beta_2 D_i + u_i}$

Now let's look at the change when $D_i = 1$ and $D_i = 0$

When $D_i = 1$: $E(Y_i|D_i = 1) = e^{\beta_1 + \beta_2 + u_i}$

When $D_i = 0$: $E(Y_i|D_i = 0) = e^{\beta_1 + u_i}$

$$E(Y_i|D_i = 1) - E(Y_i|D_i = 0) = e^{\beta_1 + \beta_2 + u_i} - e^{\beta_1 + u_i}$$
$$= e^{\beta_2} \cdot e^{\beta_1 + u_i} - e^{\beta_1 + u_i}$$
$$= e^{\beta_2} \cdot e^{\beta_1 + u_i} - e^{\beta_1 + u_i} = e^{\beta_1 + u_i}(e^{\beta_2} - 1)$$

Log-linear interpretation of dummy variable

Now consider ***exact*** percentage change

$$= \frac{e^{\beta_1+u_i}(e^{\beta_2}-1)}{e^{\beta_1+u_i}} \cdot 100 = \left(\boldsymbol{e^{\beta_2}-1}\right) \cdot \mathbf{100}$$

On the other hand, we can ***approximate*** the percentage change when switching from $D_i = 0$ to $D_i = 1$.

We can say that when $D_i = 0$ switch to $D_i = 1$, approximate percentage change in $\ln \hat{Y}_i$ is **100 x $\boldsymbol{\hat{\beta}_2}$**. This approximation works if the coefficient is small. See the example next page.

Log-linear interpretation of dummy variable

Coefficient	Coefficient x 100	$e^{\beta_2} - 1$	Diff
0.005	0.5	0.501252086	0.001252
0.01	1	1.005016708	0.005017
0.015	1.5	1.511306462	0.011306
0.02	2	2.020134003	0.020134
0.025	2.5	2.531512052	0.031512
0.03	3	3.045453395	0.045453
0.035	3.5	3.56197088	0.061971
0.04	4	4.081077419	0.081077
0.045	4.5	4.602785991	0.102786
0.05	5	5.127109638	0.12711
0.055	5.5	5.654061468	0.154061
0.06	6	6.183654655	0.183655
0.065	6.5	6.715902438	0.215902
0.07	7	7.250818125	0.250818
0.075	7.5	7.788415088	0.288415
0.08	8	8.328706767	0.328707
0.085	8.5	8.87170667	0.371707
0.09	9	9.417428371	0.417428
0.095	9.5	9.965885513	0.465886
0.1	10	10.51709181	0.517092
0.105	10.5	11.07106104	0.571061
0.11	11	11.62780705	0.627807
0.115	11.5	12.18734376	0.687344
0.12	12	12.74968516	0.749685
0.125	12.5	13.31484531	0.814845
0.13	13	13.88283833	0.882838

Log-linear interpretation of dummy variable

```
. reg lnexp i.sex

      Source |       SS           df       MS      Number of obs   =        51
-------------+----------------------------------   F(1, 49)        =      6.60
       Model |  4.61990689         1  4.61990689   Prob > F        =    0.0133
    Residual |  34.3084635        49  .700172724   R-squared       =    0.1187
-------------+----------------------------------   Adj R-squared   =    0.1007
       Total |  38.9283704        50  .778567408   Root MSE        =    .83676

------------------------------------------------------------------------------
       lnexp | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         sex |
     Female  |  -.6744299   .2625565    -2.57   0.013    -1.202057   -.1468028
       _cons |   8.969325   .2236344    40.11   0.000     8.519915    9.418735
------------------------------------------------------------------------------
```