Chapter 2

The Bivariate Regression Model

1

Linear model

- Sample of n observations, labeled as i=1,2,..n
  - $y_i = \beta_0 + x_i \beta_1 + \varepsilon_i$
- β_0 and β_1 are "population" values – represent the true relationship between x and y
- Unfortunately – these values are unknown
- The job of the researcher is to estimate these values

2

- Notice that if we differentiate y with respect to x, we obtain
- $\partial y / \partial x = \beta_1$
- β_1 represents how much y will change for a fixed change in x
  - Increase in income for more education
  - Change in crime or bankruptcy when casinos are opened
  - Increase in test score if you study more

3

Put some concreteness on problem

- Suppose a state is experiencing a significant budget shortfall
- Short-term solution – raise tax on cigarettes by 35 cents/pack
- Problem – a tax hike will reduce consumption (theory of demand)
- Question for state – as taxes are raised, how much will cigarette consumption fall

4

- Suppose y is a state's per capita consumption of cigarettes
- x represents taxes on cigarettes
- Question – how much will y fall if x is increased by 35 cents/pack?
- Note – there are many reasons why people smoke – cost is but one of them –

5

Benefits and Costs of Model

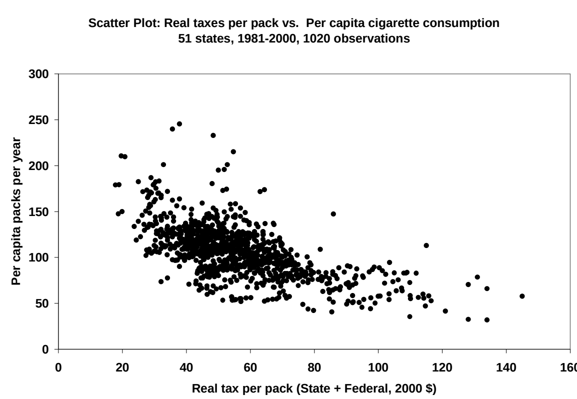
- Placed more structure on the model, therefore we can obtain precise statements about the relationship between x and y
- These statements will be true so long as the hypothesized relationship is true
- As you place more structure on any model, the chance that the assumptions of the model are correct declines.

6

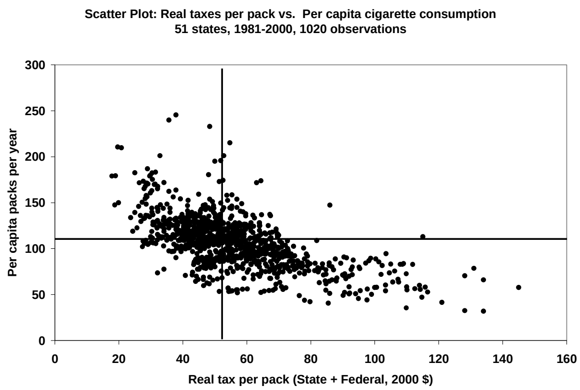
Data

- Data on state consumption/taxes, 1981-2000
- 51 states x 20 years = 1020 observations
- Y = per capita consumption
- X = tax (State + Federal) in real cents per pack
  - 2000 dollars

7



8



What is ε_i?

- There are many factors that determine a state's level of cigarette consumption
- Some of these factors we can measure, but for what ever reason, we do not have data
  - Education, age, income, etc.
- Some of these factors we cannot measure
  - Dislike of cigarettes, anti-smoking sentiment of your friends/neighbors/relatives
- ε_i identified what we cannot measure in our model

- Think of a difference way – draw a vertical line at any tax level (e.g., 40 cents).
- Notice that at this level, there are multiple values of Y that are present
- Therefore – on average, higher taxes will reduce consumption, but it cannot explain all of consumption across states

Current smoking rates By demographic group

- Adults
- Gender
  - Males
  - Females
- Age group
  - 18-44
  - 45-64
  - 75+
- Race
  - White
  - AA
  - AI/AN
  - Asian
- Hispanic origin
  - Hispanic
  - Non hispanic
- Education
  - < HS
  - HS
  - Some col.
  - College+
- Family Income
  - <$20K
  - $20-$35K
  - $35-$55K
  - $55-$75K
  - $>75K

- We can however estimate values of ε_i by estimating values of β_0 and β_1.
- Estimates have "hats": $\hat{\beta}_0$ *and* $\hat{\beta}_1$
- Our goal, is to choose values for $\hat{\beta}_0$ *and* $\hat{\beta}_1$ in an optimal way.
- Requires minimizing some function of the estimated errors associated with the model

13

Performance in the Olympics

- Medal count in the Olympics is a simple measure of output
- Countries vary by
  - Size
  - Resources
- How is performance once we control for these attributes?

14

Ranking by Total Medal Count

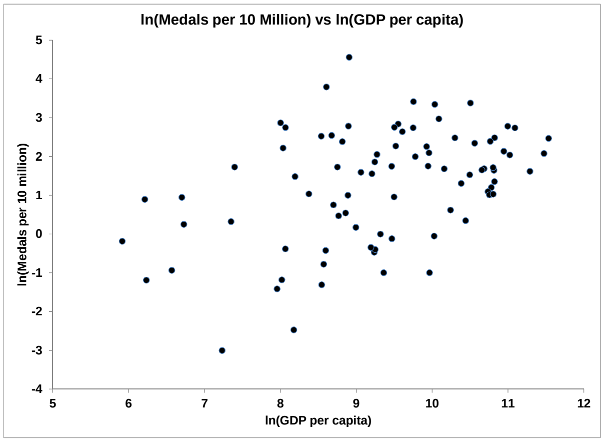
medals~k	name	country	medals
1	United States of America	USA	104
2	People's Republic of China	CHI	88
3	Russian Federation	RUS	82
4	Great Britain	GB	65
5	Germany	GER	44
6	Japan	JAP	38
7	Australia	AUS	35
8	France	FRE	34
9	Republic of Korea	SKOR	28
9	Italy	ITA	28

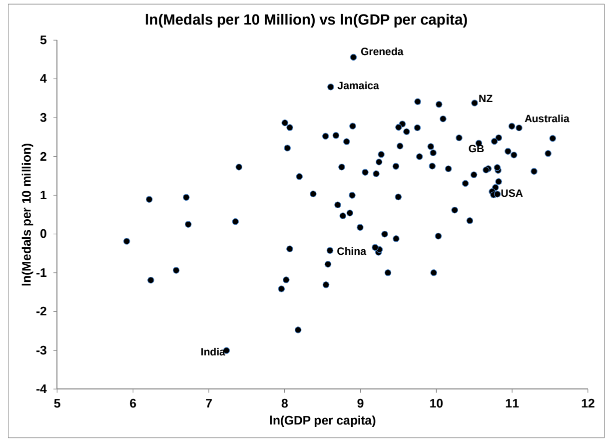
15

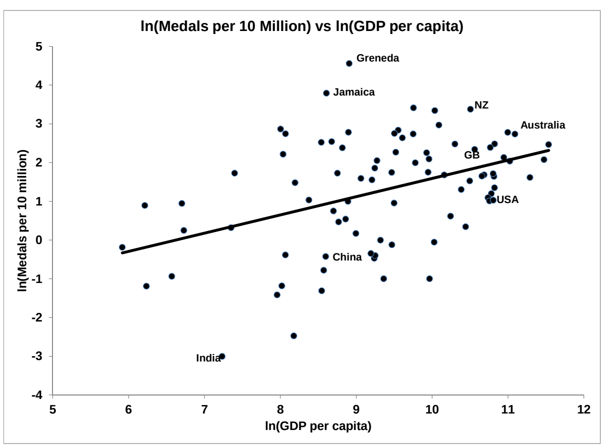
Ranking by Medals/10 million People

m~a_rank	name	country	medals	medals~a
1	Grenada	GREN	1	95.2381
2	Jamaica	JAM	12	44.34873
3	Trinidad and Tobago	T&T	4	30.3556
4	New Zealand	NZEL	13	29.31665
5	Bahamas	BAH	1	28.27591
6	Slovenia	SLOVE	4	19.43748
7	Mongolia	MONG	5	17.58087
8	Hungary	HUN	17	17.06485
9	Montenegro	MONT	1	16.12828
10	Denmark	DEN	9	16.11529

16





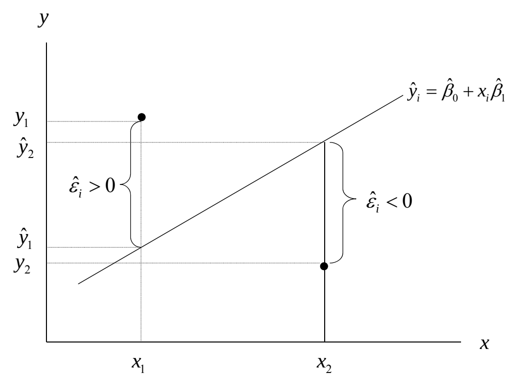


- Given linear model $y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$
- We can predict an level of consumption given parameter values
- $\hat{y}_i = \hat{\beta}_0 + x_i\hat{\beta}_1$
- The predicted value will not always be accurate
  - sometimes we will over or under predict the true value
- Because of the linear relationship between x and y, predictions will lie along a line

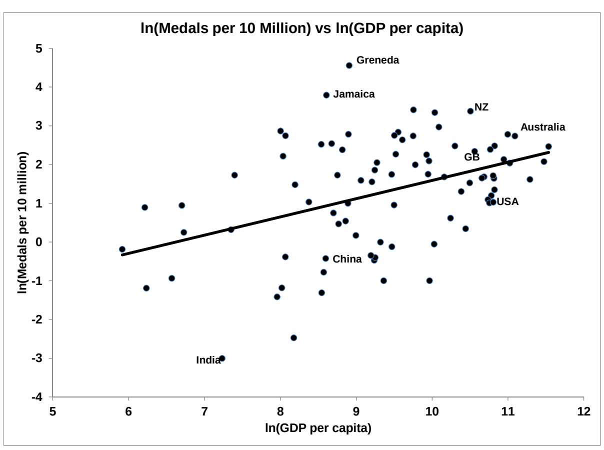
20

- Difference between the actual & predicted value
- $y_i - \hat{y}_i = y_i - \hat{\beta}_0 - x_i\hat{\beta}_1 = \hat{\varepsilon}_i$
- if $y_i - \hat{y}_i = \hat{\varepsilon}_i > 0$ you underpredict
  - (you did better than expected)
- If $y_i - \hat{y}_i = \hat{\varepsilon}_i < 0$ you overpredict
  - (you did worse than expected)

21



22



Estimation

- Estimated errors measure what we don't know
- Want to minimize these errors as much as possible
- There are N errors in each model
- Need to select a criteria to somehow minimize all these errors

24

Criteria: Least squares

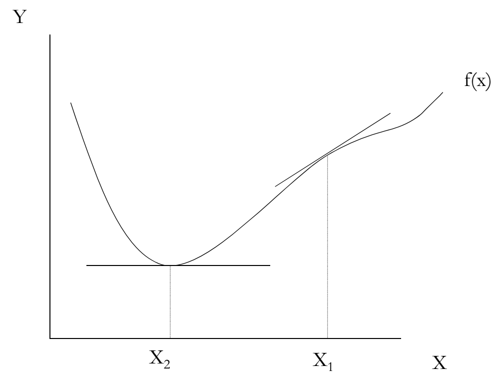
let $\hat{\beta}_0$ *and* $\hat{\beta}_1$ *be candidate values for the parameters. The estimated error is then*

$$\hat{\varepsilon}_i = y_i - \hat{\beta}_0 - x_i\hat{\beta}_1$$

Objective : min *the sum of squared errors*

$$SSE = \sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2} = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - x_i\hat{\beta}_1)^2$$

25



26

Cigarette example

- Data available on web page
  - state_cig_data.dta
  - Already in a Stata data file
  - To use,
    - Download to a folder
    - Change directory to the folder
    - type “use state_cig_data”

27

```
. * dscribe data
. desc

Contains data from state_cig_data.dta
  obs:         1,020
 vars:             7                          28 Aug 2008 15:39
 size:        29,580 (99.8% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
state           str2   %9s                    2-digit state code
year            int    %8.0g                  year
state_tax       float  %9.0g                  state tax in cents per pack
retail_price    float  %9.0g                  average retail price, nominal
federal_tax     byte   %8.0g                  federal tax in cents per pack
packs_pc        float  %9.0g                  packs of cigarettes per capita
cpi             float  %9.0g                  consumer price index, 2000=1.000
-------------------------------------------------------------------------------
```

28

```
. * generate real taxes
. gen real_tax=(state_tax+federal_tax)/cpi

.
. * get means of real tax and per capita consumption
. sum packs_pc real_tax

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    packs_pc |      1020    106.6021    28.29377       31.9      245.4
    real_tax |      1020    56.05339    18.45741   17.84456        145

. * get correlation coefficient
. corr packs_pc real_tax
(obs=1020)

             | packs_pc real_tax
-------------+------------------
    packs_pc |   1.0000
    real_tax |  -0.6115   1.0000
```

Generate real tax
By dividing by CPI

Std deviations

Means

Correlation coefficient

$$\hat{\beta}_1 = \frac{\hat{\rho}_{xy}\hat{\sigma}_y}{\hat{\sigma}_x} = \frac{-0.6115(28.29377)}{18.45741} = -0.937$$

$$\hat{\beta}_0 = \overline{y} - \overline{x}\hat{\beta}_1 = 106.60 - (56.05)(-0.937) = 159.1$$

Run regression in stata

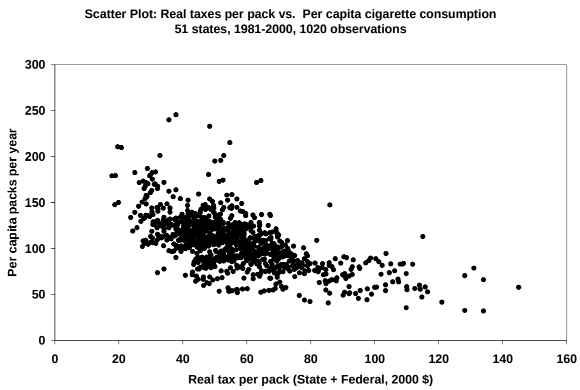
SSM, SSE, SST

R^2

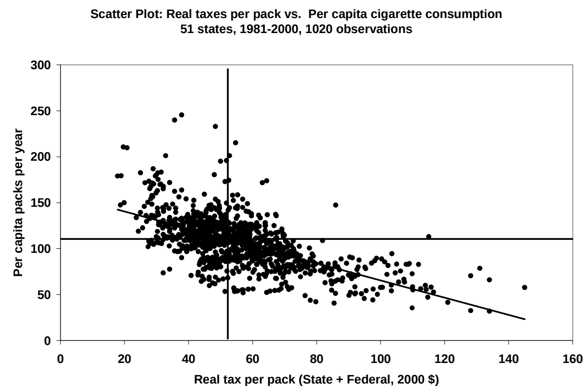
```
. * get regression estimate
. reg packs_pc real_tax

      Source |       SS       df       MS              Number of obs =    1020
-------------+------------------------------           F(  1,  1018) =  607.95
       Model |  305010.398     1  305010.398           Prob > F      =  0.0000
    Residual |  510736.995  1018  501.706282           R-squared     =  0.3739
-------------+------------------------------           Adj R-squared =  0.3733
       Total |  815747.392  1019  800.537186           Root MSE      =  22.399

------------------------------------------------------------------------------
    packs_pc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    real_tax |  -.9373448    .038016   -24.66   0.000    -1.011944    -.862746
       _cons |   159.1434   2.243373    70.94   0.000     154.7412    163.5456
------------------------------------------------------------------------------
```

β_0 and β_1

- Notice that SSE + SSM = SST
- R^2 = SSM/SST = 305010.4/815747.4 = 0.3739



33



34

Using the results

$$\hat{\beta}_1 = -0.9373$$

- For every penny increase in taxes, per capita consumption falls by 0.94 packs per year
- A 35 cent increase in taxes will reduce consumption by (35)(0.94) = 32 packs per person per year

35

Output residuals

```
. * output residuals
. predict error_packs_pc, residuals

.
. * show that the means of the errors are zero
. sum error_packs_pc

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
error_pack~c |      1020    9.24e-09    22.38781  -57.57952   121.7705

.
. * get correlation between x and error
. corr error_packs_pc real_tax
(obs=1020)

             | error_~c real_tax
-------------+------------------
error_pack~c |   1.0000
    real_tax |  -0.0000   1.0000
```

Predicted error Has zero mean

Predicted error not Correlated with x

36

Example 2

- Do better performing teams have higher attendance?
- Data on wins and average attendance/game for 2004 baseball season
- 30 observations
- attendance.dta

37

```
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
team            str13  %13s                   team city
attendance      long   %12.0g                 avg attendance per game
wins            int    %8.0g                  wins during year
-------------------------------------------------------------------------------
```

```
. * get means of wins and payroll
. sum wins attendance

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        wins |        30    80.83333    14.75334         55        103
  attendance |        30     28157.3      9317.7      10031      43712
```

38

```
. * get detailed data for attendance
. sum attendance, detail

                   avg attendance per game
-------------------------------------------------------------
      Percentiles      Smallest
 1%        10031          10031
 5%        10038          10038
10%      15169.5          13157       Obs                  30
25%        20703          17182       Sum of Wgt.          30

50%      28934.5                      Mean            28157.3
                        Largest       Std. Dev.        9317.7
75%        34527          39494
90%      39828.5          40163       Variance       8.68e+07
95%        43323          43323       Skewness      -.2430352
99%        43712          43712       Kurtosis       2.256339
```

39

```
. * run simple regression
. reg attendance wins

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  1,    28) =    8.89
       Model |   606784507     1   606784507           Prob > F      =  0.0059
    Residual |  1.9110e+09    28  68249360.1           R-squared     =  0.2410
-------------+------------------------------           Adj R-squared =  0.2139
       Total |  2.5178e+09    29  86819537.6           Root MSE      =  8261.3

------------------------------------------------------------------------------
  attendance |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        wins |   310.0473   103.9825     2.98   0.006     97.04894    523.0457
       _cons |    3095.14   8539.507     0.36   0.720    -14397.25    20587.53
------------------------------------------------------------------------------
```

$\hat{\beta}_1 = 310.05$ .... *for every addition win, attendance increase by* 310 / *game.*

An addition 10 *wins put* 3100.5 *more people in seats*

40

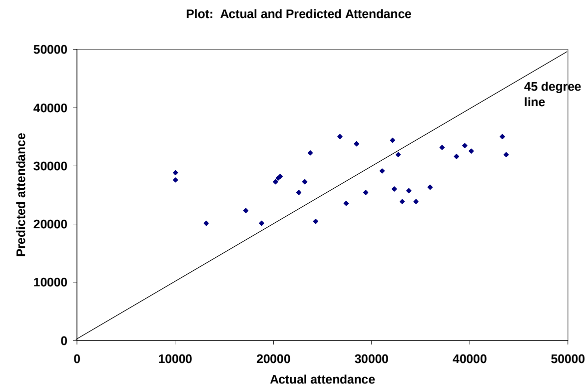
```
. * output predicted attendance
. predict pred_att, xb

.
. list team wins attendance pred_att

     +-------------------------------------------+
     |         team   wins   attend~e   pred_att |
     |-------------------------------------------|
  1. |      Seattle     93      43712   31929.54 |
  2. |  New York-AL    103      43323   35030.02 |
  3. |    Baltimore     67      33122   23868.31 |
  4. |       Boston     93      32717   31929.54 |
  5. |    Cleveland     74      32308   26038.64 |
     |-------------------------------------------|
  6. |        Texas     72      29405   25418.55 |
  7. |      Anaheim     99      28464   33789.83 |
  8. |      Oakland    103      26788   35030.02 |
  9. |    Minnesota     94      23759   32239.59 |
 10. |   Chicago--AL     81      20703   28208.97 |
     |-------------------------------------------|
```

Construct a new variable, The predicted value of Y

Consider Seattle

$Y^p = 3095.1 + (310.05)(93) = 31929$

41



42

Example 3: Education and Earnings

- Stylized fact: log wages or earnings is linear in education (above a certain range)
- Interpreted as a "return to education"
- Theoretical models why this would be the case
- Linear model:
  - y=ln(weekly wages) – endogenous variable
  - x=years of education – exogenous factor
  - $y_i = \beta_0 + x_i \beta_1 + \varepsilon_i$

43

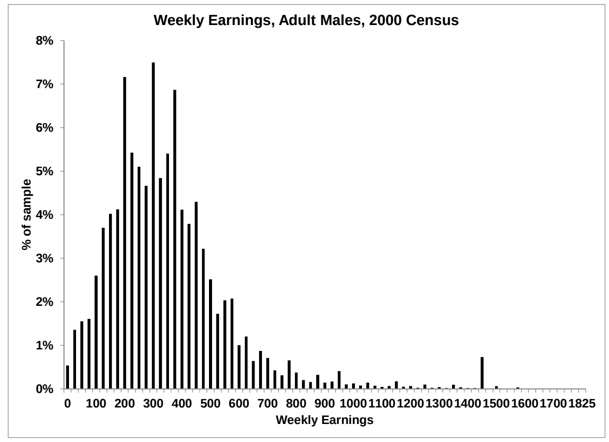
- Notice that β_1 has a different interpretation
- β_1 =dY/dX
- In this case, y=ln(Wages)
- dln(Wages)/dX = (1/wages)dWages/dX
- dWages/wages = % change in changes
  - (change in wages over base wages)
- when the endogenous variable is a natural log,
- β_1 =dY/dX is interpreted as "% change in y for a unit change in x'

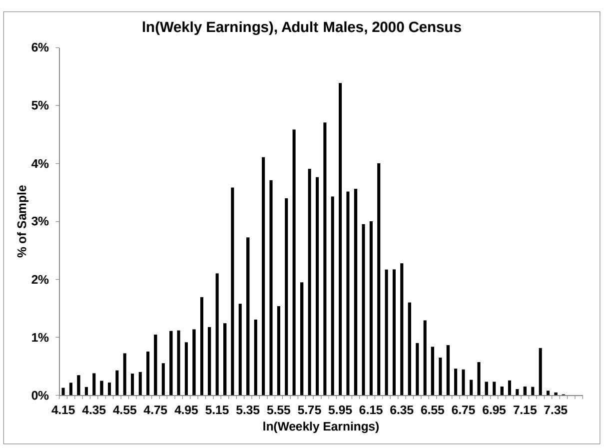
44

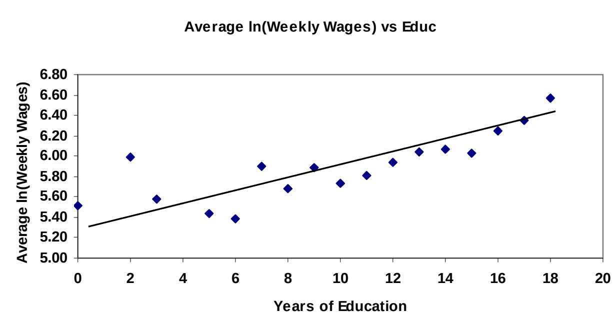
Data

- cps87.dta
- 19,906 observations from 1987 Current Population Survey on
  - Full time (>30 hours)
  - Males
  - Aged 21-64

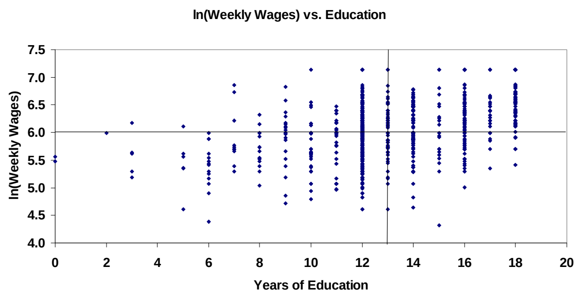
45







48



Generate ln(weekly earnings)

```
. * construct ln weekly earnings
. gen ln_weekly_earn=ln(weekly_earn)
.
. * get descriptive information
. sum weekly_earn ln_weekly_earn years_educ

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
 weekly_earn |     19906     488.264    236.4713         60        999
ln_weekly_~n |     19906    6.067307     .513047   4.094345   6.906755
  years_educ |     19906    13.16126    2.795234          0         18
```

Means of the variables


```
. * run simple regression
. reg ln_weekly_earn years_educ

      Source |       SS       df       MS              Number of obs =   19906
-------------+------------------------------           F(  1, 19904) = 3877.62
       Model |  854.28055     1  854.28055             Prob > F      =  0.0000
    Residual | 4385.05814 19904  .220310397            R-squared     =  0.1631
-------------+------------------------------           Adj R-squared =  0.1630
       Total | 5239.33869 19905  .263217216            Root MSE      =  .46937

------------------------------------------------------------------------------
ln_weekly_~n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  years_educ |   .0741141   .0011902    62.27   0.000     .0717813     .076447
       _cons |   5.091872   .0160138   317.97   0.000     5.060484    5.123261
------------------------------------------------------------------------------
```


Example 4:London Olympics

```
* generate measure of medals per person;
* divide by 10,000,000 people;
gen medals_capita=medals/(population/10000000);
label var medals_capita "medals per 10,000,000 people";

* take natural ln of medals_capita and gdp_capita;
gen ln_medals_capita=ln(medals_capita);
gen ln_gdp_capita=ln(gdp_capita);

* regress ln(medals/population) on ln(gdp/population);
reg ln_medals_capita ln_gdp_capita;
```

Example 4: London Olympics

```
. * regress ln(medals/population) on ln(gdp/population);
. reg ln_medals_capita ln_gdp_capita;

      Source |       SS       df       MS              Number of obs =      85
-------------+------------------------------           F(  1,    83) =   17.93
       Model |  31.8688335     1  31.8688335           Prob > F      =  0.0001
    Residual |  147.537169    83  1.77755625           R-squared     =  0.1776
-------------+------------------------------           Adj R-squared =  0.1677
       Total |  179.406002    84  2.13578574           Root MSE      =  1.3333

------------------------------------------------------------------------------
ln_medals_~a |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ln_gdp_cap~a |   .4702596   .1110622     4.23   0.000     .2493614    .6911579
       _cons |  -3.111996    1.04655    -2.97   0.004    -5.193543    -1.03045
------------------------------------------------------------------------------
```

53