PS1

June 28, 2024

1. (a) *Proof.*

$$\begin{aligned}
\sum_{i=1}^{n} y_i &= \sum_{i=1}^{n} (x_{i1} + 1 - x_{i1}) \times y_i \\
&= \sum_{i=1}^{n} x_{i1} \times y_i + \sum_{i=1}^{n} (1 - x_{i1}) \times y_i
\end{aligned}$$

$\square$

(b)

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_{i1} - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_{i1} - \overline{x})^2} \\
&= \frac{\sum_{i=1}^{n} x_{i1} y_i - \overline{x}\sum_{i=1}^{n} y_i - \overline{y}\sum_{i=1}^{n} x_{1i} + \sum_{i=1}^{n} \overline{xy}}{\sum_{i=1}^{n} x_{1i} - 2\overline{x}\sum_{i=1}^{n} x_{1i} + \overline{x}\sum_{i=1}^{n} x_{1i}} \\
&= \frac{\sum_{i=1}^{n} x_{i1} y_i - \overline{x}\sum_{i=1}^{n} y_i - \overline{y}\sum_{i=1}^{n} x_{1i} + \overline{x}\sum_{i=1}^{n} y_i}{(1 - \overline{x})\sum_{i=1}^{n} x_{1i}} \\
&= \frac{\sum_{i=1}^{n} x_{i1} y_i - \overline{y}\sum_{i=1}^{n} x_{1i}}{(1 - \overline{x})\sum_{i=1}^{n} x_{1i}} \\
&= \frac{\sum_{i=1}^{n} x_{i1} y_i}{(1 - \overline{x})\sum_{i=1}^{n} x_{1i}} - \frac{\overline{y}}{1 - \overline{x}} \\
&= \frac{\sum_{i=1}^{n} x_{i1} y_i}{(1 - \overline{x})\sum_{i=1}^{n} x_{1i}} - \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n}(1 - x_{i1})} \\
&= \frac{\sum_{i=1}^{n} x_{i1} y_i}{(1 - \overline{x})\sum_{i=1}^{n} x_{1i}} - \frac{\sum_{i=1}^{n} x_{i1} y_i}{\sum_{i=1}^{n}(1 - x_{i1})} - \frac{\sum_{i=1}^{n}(1 - x_{i1}) y_i}{\sum_{i=1}^{n}(1 - x_{i1})} \\
&= \left(\sum_{i=1}^{n} x_{i1} y_i\right)\left(\frac{1}{(1 - \overline{x})\sum_{i=1}^{n} x_{1i}} - \frac{1}{\sum_{i=1}^{n}(1 - x_{i1})}\right) - \frac{\sum_{i=1}^{n}(1 - x_{i1}) y_i}{n_0} \\
&= \left(\sum_{i=1}^{n} x_{i1} y_i\right)\frac{n - \sum_{i=1}^{n} x_{1i}}{\sum_{i=1}^{n}(1 - x_{1i})\sum_{i=1}^{n} x_{1i}} - \overline{y}_0 \\
&= \frac{\sum_{i=1}^{n} x_{i1} y_i}{\sum_{i=1}^{n} x_{1i}} - \overline{y}_0 \\
&= \overline{y}_1 - \overline{y}_0 \\
\widehat{\beta}_0 &= \overline{y} - \widehat{\beta}_1 \overline{x} \\
&= \frac{\sum_{i=1}^{n} y_i}{n} - (\overline{y}_1 - \overline{y}_0)\frac{\sum_{i=1}^{n} x_{1i}}{n} \\
&= \overline{y}_0\frac{\sum_{i=1}^{n} x_{1i}}{n} + \frac{\sum_{i=1}^{n} y_i}{n} - \frac{\overline{y}_1\sum_{i=1}^{n} x_{1i}}{n} \\
&= \overline{y}_0\frac{\sum_{i=1}^{n} x_{1i}}{n} + \frac{\sum_{i=1}^{n} y_i}{n} - \frac{\overline{y}_1 \times n_1}{n} \\
&= \overline{y}_0\frac{\sum_{i=1}^{n} x_{1i}}{n} + \frac{\sum_{i=1}^{n}(1 - x_{i1}) y_i}{n} \\
&= \overline{y}_0\frac{\sum_{i=1}^{n} x_{1i}}{n} + \overline{y}_0\frac{\sum_{i=1}^{n}(1 - x_{1i})}{n} \\
&= \overline{y}_0
\end{aligned}$$

(c) $\widehat{\beta}_1$ represents the difference between y on $x = 1$ and y on $x = 0$, and $\widehat{\beta}_0$ represents the average of y on $x = 0$.

(d) *Proof.*

$$\begin{aligned}
\mathbb{E}[\widehat{\beta}_1] &= \beta_1 \\
&= \frac{\mathbb{E}[y] - \beta_0}{\mathbb{E}[x_1]} \\
&= \frac{\mathbb{E}[y|x_1 = 1]P(x_1 = 1) + \mathbb{E}[y|x_1 = 0]P(x_1 = 0) - \mathbb{E}[y|x_1 = 0]}{P(x_1 = 0) \times 0 + P(x_1 = 1) \times 1} \\
&= \frac{\mathbb{E}[y|x_1 = 1]P(x_1 = 1) - \mathbb{E}[y|x_1 = 0]P(x_1 = 1)}{P(x_1 = 1)} \\
&= \mathbb{E}[y|x_1 = 1] - \mathbb{E}[y|x_1 = 0]
\end{aligned}$$

□

2. (a) There's 3 possible explanations for the correlation between smoking and baby's weight.
   i. The more the mother smoked, the less the baby's weight.
   ii. The less the baby's weight, the more the mother smoked.
   iii. There's a third variable that affects both the mother's smoking and the baby's weight in the contrary way.

   (b) Because the more the mother smoked, the less the baby's weight, and the effect of smoking on baby's weight is quite big.

   (c) Yes, it's possible according to the common sense.

3. (a)

```
. use "EduIncome_24.dta"

. summarize birthyear wage schooling_yr

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
   birthyear |      2,429    1974.794      10.584       1955       1990
        wage |      2,429    58122.36    41021.76   2113.569   608707.9
schooling_yr |      2,429    7.656237    2.927878          0         15
```

(b)

```
. gen female = (gender == 2)

. summarize female

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      female |      2,429    .2984767     .457684          0          1
```

Female in the sample accounts for 29.85%.

(c)

```
. summarize wage if female == 1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        wage |        725    47644.76    32694.44   2242.335   355079.6

. summarize wage if female == 0

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        wage |      1,704    62580.26    43337.33   2113.569   608707.9

. regress wage female

      Source |       SS           df       MS      Number of obs   =     2,429
-------------+----------------------------------   F(1, 2427)      =     69.32
       Model |  1.1345e+11         1  1.1345e+11   Prob > F        =    0.0000
    Residual |  3.9723e+12     2,427  1.6367e+09   R-squared       =    0.0278
```

```
-------------+----------------------------------   Adj R-squared   =    0.0274
       Total |  4.0858e+12     2,428  1.6828e+09   Root MSE        =     40457

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   -14935.5   1793.902    -8.33   0.000    -18453.24   -11417.77
       _cons |   62580.26    980.063    63.85   0.000     60658.41    64502.11
------------------------------------------------------------------------------
```

So, we have wage $= 62580.26 - 14935.5 \times$ female .

Using my conclusion in Q1, this regression result means that the wage difference between female and male is -14935.5 yuan, and the average wage for male is 62580.26 yuan.

(d)

```
. gen age = 2023 - birthyear

. summarize age

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         age |      2,429    48.20585      10.584         33         68

. gen ln_wage = ln(wage)

. regress ln_wage age schooling_yr

      Source |       SS           df       MS      Number of obs   =     2,429
-------------+----------------------------------   F(2, 2426)      =     33.05
       Model |  37.4492937         2  18.7246468   Prob > F        =    0.0000
    Residual |  1374.39397     2,426  .566526782   R-squared       =    0.0265
-------------+----------------------------------   Adj R-squared   =    0.0257
       Total |  1411.84327     2,428  .581484047   Root MSE        =    .75268

------------------------------------------------------------------------------
     ln_wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0027053    .001456     1.86   0.063    -.0001498    .0055605
schooling_yr |   .0425877   .0052633     8.09   0.000     .0322666    .0529088
       _cons |   10.26689   .0867829   118.31   0.000     10.09671    10.43707
------------------------------------------------------------------------------
```

If `schooling_yr` increases by 1, the **log(`wage`)** will increase by 0.043.

(e)

```
. predict ln_wage_hat, xb

. gen residual = ln_wage - ln_wage_hat

. summarize ln_wage_hat residual

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
 ln_wage_hat |      2,429    10.72336    .1241931   10.35617   11.07885
    residual |      2,429    3.34e-09    .7523697  -3.072877   2.525533
```

The residual is very small, which suggests that the model fits the data well.

(f) In (d) , we have $R^2 = 0.0265$, indicating that there may be other factors not included in the model.