Econ 272

Homework 1

Name: Gianluca Crescenzo

Question 1. A key question in economics is whether credit access affects firm profits. This can be studied using the linear regression:

$$\text{Profits}_i = \beta_0 + \beta_1 \text{Credit}_i + \epsilon_i.$$

The unit of observation is the firm, denoted i. *Profits* and *Credit* are measured in dollars. *Credit* is the amount of bank loans a firm has. We test this relationship in a sample of 2,000 U.S. firms.

(a) A priori, do you think the population relationship (estimated by β_1) would be positive or negative?

(b) List 3 other variable which can affect firm profits. For each variable considered, hypothesize the following: (i) would that variable positively or negatively affect profits in the population? (ii) Would that variable be positively or negative correlated with whether a firm has bank loans? A one sentence intuitive explanation would be sufficient.

(c) Upon estimation, assume that we obtain the following coefficients: $\hat{\beta}_0 = 41,354$; $\hat{\beta}_1 = 0.1745$. Based on these values, what would be the predicted profits for a firm which has received USD 20,000 in bank loans?

Answer. (a) I believe the population relationship between profits and credit would be positive.

(b) Three other variables which might affect firm profits could be firm size, operating costs, and firm age. Firm size would have a positive effect on profits, as larger firms tend to earn higher profits. Operating costs would have a negative effect, as higher operating costs will reduce profits. Firm age would have a positive effect on profits, as an older firm might be more stable due to experience and/or establishment in markets.

(c) Plugging in the values for $\hat{\beta_0}$ and $\hat{\beta}_1$ gives:

$$\begin{aligned}\text{Profits}_i &= \hat{\beta}_0 + \hat{\beta}_1 \text{Credit}_i \\ &= 41354 + (0.1745)(20000) \\ &= 44844.\end{aligned}$$

Question 2. Consider some population relationship denoted by $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$. A sample is drawn to estimate this relationship using linear regressions, yielding coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$.

(a) Show that the sample mean will always lie on the regression line (hint: denote the sample mean of the independent variable as $\overline{x}$ and plug it into the estimated regression line. You want to show that the predicted value is the sample mean of the outcome variable, $\overline{y}$).

(b) Using the formulas of the OLS estimators of β_0 and β_1 derived in class, show that the OLS estimators are unaffected by units of measurement (hint: assume that there is some constant value c, and we multiply all observations in the independent variable x by c. Show that the OLS estimator for β_1 will now equal $\frac{1}{c}\hat{\beta}_1$).

Answer. (a) Observe that:

$$\begin{aligned}\hat{\beta}_0 + \hat{\beta}_1 \overline{x} &= \overline{y} - \hat{\beta}_1 \overline{x} + \hat{\beta}_1 \overline{x} \\ &= \overline{y}.\end{aligned}$$

(b) Let $\hat{\beta}_1'$ be the OLS estimator whose observations of x are all multiplied by c, Minimizing the squares of our residuals gives:

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \hat{\beta}_1'} \sum_{i=1}^{n} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1' c x_i) \right]^2 \\
&= -2c \sum_{i=1}^{n} (y_i \hat{\beta}_0 - \hat{\beta}_1' c x_i) x_i \\
&= -2c \left(\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - c\hat{\beta}_1' \sum_{i=1}^{n} x_i^2 \right).
\end{aligned}
$$

Dividing by $-2c$ on both sides and moving terms around gives:

$$
\begin{aligned}
\sum_{i=1}^{n} x_i y_i &= \hat{\beta}_0 \sum_{i=1}^{n} x_i + c\hat{\beta}_1' \sum_{i=1}^{n} x_i^2 \\
&= (\overline{y} - \hat{\beta}_1' \overline{x}) \sum_{i=1}^{n} x_i + c\hat{\beta}_1' \sum_{i=1}^{n} x_i^2 \\
&= \left(\frac{1}{n} \sum_{i=1}^{n} y_i - \hat{\beta}_1' \frac{1}{n} \sum_{i=1}^{n} c x_i \right) \sum_{i=1}^{n} x_i + c\hat{\beta}_1' \sum_{i=1}^{n} x_i^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i - c\hat{\beta}_1' \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2 + c\hat{\beta}_1' \sum_{i=1}^{n} x_i^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} x_1 \sum_{i=1}^{n} y_i + c\hat{\beta}_1' \left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2 \right).
\end{aligned}
$$

Isolating $c\hat{\beta}_1'$ now gives us:

$$
\begin{aligned}
c\hat{\beta}_1' &= \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2} \\
&= \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \\
&= \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)}.
\end{aligned}
$$

Whence $\hat{\beta}_1' = \frac{1}{c} \hat{\beta}_1$.

Question 3. We are now going to empirically test whether urbanization affects unemployment using data from India. Use the dataset hw1.dta for this exercise. The unit of observation is "districts" in India —somewhat similar to U.S. counties. *rural* is the share of rural population in a district; *unemp* is the share of unemployed individuals in a district; *seced* is the share of adults who have completed a high school or college education.

Answer. (a) I would expect $\hat{\beta}_1$ to be negative. One possible reason for this is that there are likely fewer people, fewer business, and people with less education, resulting in more people to be unemployed.

(b) The share of the population living in rural areas is 76%, and the rate of unemployment is 19%.

(e) From my .do file, $\hat{\beta}_1 = -0.1108817$.

(f) From my .do file, $\hat{\beta}_0 = .02779372$.

(g) From my .do file, $\sum_{i=1}^{n} e_i^2 = .21393375$.

(h) The correlation of the residuals and the share of rural population in a district was zero. This makes sense, since OLS residuals are uncorrelated with independent variables.

(i) Stata's regression gave $\hat{\beta}_0 = 0.277937$ and $\hat{\beta}_1 = -0.110882$. These are very similar to the ones I computed above.

(j) After multiplying *unemp* by 100, running a regression gave $\hat{\beta}_0 = 2.779371$ and $\hat{\beta}_1 = -1.108816$. The coefficients grew by a factor of 10.

(k) Including the share of educated adults in the regression, $\hat{\beta}_1 = .009741$, which is a significant change from the previous estimator. Previously, the coefficient only captured the effect of *rural* on *unemp*, without accounting for the potential effect of the share of educated adults. The coefficient $\hat{\beta}_2$ measures the effect of the share of educated adults on district unemployment.