Tutorial 5

keywords: estimators, unbiasedness, expected value, variance, multiple linear regression, interpretation, ceteris paribus, prediction, interpretation, R squared

estimated reading time: 30 minutes

January 16, 2023

Question 1

(*Discussed in class*)

Question 2

Multiple linear regression model and interpreting coefficients

EViews workfile: $hprice.wf1$

i. Estimate the model of $price$ on a constant, $sqrft$ and $bdrms$ and write out the results in equation form.

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u$$

- $price$ - house price ($'000)
- $sqrft$ - area of the house (square foot)
- $bdrms$ - no. of bedrooms

$$Quick \rightarrow Estimate\ Equation$$

$$Equation\ Estimation : price\ c\ sqrft\ bdrms$$

Dependent Variable: PRICE
Method: Least Squares
Sample: 1 88
Included observations: 88

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	−19.31500	31.04662	−0.622129	0.5355
SQRFT	0.128436	0.013824	9.290506	0.0000
BDRMS	15.19819	9.483517	1.602590	0.1127

R-squared	0.631918	Mean dependent var	293.5460
Adjusted R-squared	0.623258	S.D. dependent var	102.7134
S.E. of regression	63.04484	Akaike info criterion	11.15907
Sum squared resid	337845.4	Schwarz criterion	11.24352
Log likelihood	−487.9989	Hannan-Quinn criter.	11.19309
F-statistic	72.96353	Durbin-Watson stat	1.858074
Prob(F-statistic)	0.000000		

Table 1: Regression output of $price$ on a constant, $sqrft$ and $bdrms$

When reporting the estimated model, we must not forget to include a 'hat' above the dependent variable and $se(\hat{\beta}_j)$ underneath $\hat{\beta}_j$ in parenthesis,

$$\widehat{price} = \underset{(se(\hat{\beta}_0))}{\hat{\beta}_0} + \underset{(se(\hat{\beta}_1))}{\hat{\beta}_1} sqrft + \underset{(se(\hat{\beta}_2))}{\hat{\beta}_2} bdrms$$

$$\widehat{price} = \underset{(31.0466)}{-19.3150} + \underset{(0.0138)}{0.1284} sqrft + \underset{(9.4835)}{15.1982} bdrms$$

ii. What is the estimated increase in price for a house with one more bedroom, holding square footage constant?

Background

Interpretation of estimated coefficients for multiple linear regression models

Suppose we estimate a model of y on a constant, x_1, and x_2,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

if x_1 and x_2 changes by Δx_1 and Δx_2 respectively then,

$$x_1 \textit{ becomes } x_1 + \Delta x_1$$

$$x_2 \textit{ becomes } x_2 + \Delta x_2$$

which will change $\hat{y}$,

$$\hat{y} \textit{ becomes } \hat{y} + \Delta\hat{y}$$

This then gives us the following equation,

$$\begin{aligned}
\hat{y} + \Delta\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1(x_1 + \Delta x_1) + \hat{\beta}_2(x_2 + \Delta x_2) \\
&= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2 \\
&= \hat{y} + \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2
\end{aligned}$$

Since $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$, it must follow that,

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2$$

$\therefore$ the change in $\hat{y}$ for a 1-unit change in x_1, holding x_2 constant, is $\hat{\beta}_1$,

$$\Delta x_2 = 0$$

$$\Delta x_1 = 1$$

$$\begin{aligned}
\Delta\hat{y} &= \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2 \\
&= \hat{\beta}_1 \times 1 + \hat{\beta}_2 \times 0 \\
&= \hat{\beta}_1
\end{aligned}$$

and the change in $\hat{y}$ for a 1-unit change in x_2, holding x_1 constant, is $\hat{\beta}_2$,

$$\Delta x_2 = 1$$

$$\Delta x_1 = 0$$

$$\begin{aligned}
\Delta\hat{y} &= \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2 \\
&= \hat{\beta}_1 \times 0 + \hat{\beta}_2 \times 1 \\
&= \hat{\beta}_2
\end{aligned}$$

As we can see, $\hat{\beta}_1$ and $\hat{\beta}_2$ have a partial effect (ceteris paribus) interpretation!

From our estimated model, the changed in estimated house price depends on the change square footage and no. of bedrooms,

$$\Delta\widehat{price} = \hat{\beta}_1 \Delta sqrft + \hat{\beta}_2 \Delta bdrms$$

Note: The estimated intercept coefficient does not change the estimated house price.

If square footage is held constant,

$$\Delta sqrft = 0$$

then the change in the estimated house price depends only on the change in no. of bedrooms,

$$\begin{aligned} \Delta\widehat{price} &= \hat{\beta}_1 \Delta \times 0 + \hat{\beta}_2 \Delta bdrms \\ &= \hat{\beta}_2 \Delta bdrms \end{aligned}$$

Therefore, the estimated increase in house price for for an additional bedroom, holding square footage constant,

$$\begin{aligned} \Delta\widehat{price} &= \hat{\beta}_2 \times 1 \\ &= 15.1982 \times 1 \\ &= 15.1982 \end{aligned}$$

$$\$15,198.20$$

iii. What is the estimated increase in price for a house with an additional bedroom that is 140 square feet in size? Compare this to your answer in part (ii).

$$\Delta bdrms = 1$$

$$\Delta sqrft = 140$$

$$\begin{aligned} \Delta\widehat{price} &= \hat{\beta}_1 \Delta sqrft + \hat{\beta}_2 \Delta bdrms \\ &= 0.1284 \times 140 + 15.1982 \times 1 \\ &= 33.12 \end{aligned}$$

$$\$33,120$$

The change in estimated house price is greater here than in ii) because we are also increasing the size of the house. In ii), we estimated the change in house price for an additional bedroom but kept the size of the house the same.

iv. What percentage of the variation in price is explained by square footage and number of bedrooms?

$$R^2 = 63.2\%$$

63.2% of the variation in house price is explained by square footage and number of bedrooms.

v. The first house in the sample has $sqrft = 2438$ and $bdrms = 4$. Find the predicted selling price for this house from the OLS regression line.

$$\begin{aligned}\widehat{price} &= -19.3150 + 0.1284 sqrft + 15.1982 bdrms \\ \widehat{price}_1 &= -19.3150 + 0.1284 sqrft_1 + 15.1982 bdrms_1 \\ &= -19.3150 + 0.1284 \times 2438 + 15.1982 \times 4 \\ &= 354.6052\end{aligned}$$

$$\$354,605$$

To perform this calculation in EViews,

$$Command\ Window: scalar\ prediction\ = c(1) + c(2)^*2438 + c(3)^*4$$

$$(press\ Enter\ to\ execute\ code)$$

vi. The actual selling price of the first house in the sample was \$300,000 (so $price_1$=300). Find the residual for this house. Does it suggest that the buyer underpaid or overpaid for the house?

$$\begin{aligned}\hat{u}_i &= price_i - \widehat{price}_i \\ \hat{u}_1 &= price_1 - \widehat{price}_1 \\ &= 300 - 354.605 \\ &= -54.605\end{aligned}$$

$$-\$54,605$$

Based on our estimated model, the buyer underpaid, however, we have not considered other features that impact house price e.g. number of baths, age of house, whether it has been renovated etc.

Question 3

We would like to make an "app" where users input their easy to measure body characteristics and the app predicts their body fat percentage. We start with making an app for men. We have data on body fat percentage ($BODY\_FAT$), weight in kg (WKG) and abdomen circumference in cm ($ABDOMEN$) for 251 adult men. The matrix of scatter plots of each pair of these three variables in our sample is given below.

The plots in the first row are: the scatter plot of *body fat* against *body fat* (which is the 45 degree line) at the left corner, the scatter plot of body fat against abdomen circumference in the middle, and the scatter plot of body fat against weight in the top right corner. You can create these matrices in EViews by graphing more than 2 variables and them choosing scatter plot

$$Quick \rightarrow Graph \rightarrow Scatter \rightarrow Multiple\ Series : Scatterplot\ Matrix$$

Without estimating any regressions, explain what these plots can tell us about each of the following (the correct answer for one of these is "nothing"):

Background

OLS estimator for a simple linear regression model

For the following simple linear regression model,

$$y = \beta_0 + \beta_1 x_1 + u$$

the OLS estimator of β_0 and β_1 can be expressed by the following formulas,

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1$$
$$\hat{\beta}_1 = \frac{\widehat{Cov(y, x_1)}}{\widehat{Var(x_1)}}$$

or in matrix notation,

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \bar{y} - \hat{\beta}_1 \bar{x}_1 \\ \dfrac{\widehat{Cov(y, x_1)}}{\widehat{Var(x_1)}} \end{bmatrix}$$

since $\widehat{Var(x_1)} > 0$, the sign of $\hat{\beta}_1$ depends directly on the sign of $\widehat{Cov(y, x_1)}$.

For the multiple linear regression model,

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

the OLS estimator of β_1 is not equal to $\dfrac{\widehat{Cov(y, x_1)}}{\widehat{Var(x_1)}}$,

$$\hat{\beta}_1 \neq \frac{\widehat{Cov(y, x_1)}}{\widehat{Var(x_1)}}$$

$\therefore$ the sign of $\hat{\beta}_1$, in the estimated multiple linear regression model, does not depend directly on the sign of $\widehat{Cov(y, x_1)}$.

(a) the sign of the coefficient of $ABDOMEN$ in a regression of $BODY\_FAT$ on a constant and $ABDOMEN$

$$BODY\_FAT = \beta_0 + \beta_1 ABDOMEN + u$$
$$\widehat{BODY\_FAT} = \hat{\beta}_0 + \hat{\beta}_1 ABDOMEN$$

For the simple regression model of $BODY\_FAT$ on a constant and $ABDOMEN$ the OLS estimator of β_0 and β_1 are given by the following formulas,

$$\hat{\beta}_0 = \overline{BODY\_FAT} - \hat{\beta}_1 \overline{ABDOMEN}$$
$$\hat{\beta}_1 = \frac{\widehat{Cov(BODY\_FAT, ABDOMEN)}}{\widehat{Var(ABDOMEN)}}$$

From the scatter plot, we can see that $BODY\_FAT$ and $ABDOMEN$ have a positive linear relationship,

$$\therefore \widehat{Cov(BODY\_FAT, ABDOMEN)} > 0$$
$$\implies \hat{\beta}_1 > 0$$

(b) the sign of the coefficient of WKG in a regression of $BODY\_FAT$ on a constant and WKG

$$(different\ model\ so\ I'm\ using\ a\ different\ Greek\ letter)$$

$$BODY\_FAT = \alpha_0 + \alpha_1 WKG + u$$
$$\widehat{BODY\_FAT} = \hat{\alpha}_0 + \hat{\alpha}_1 WKG$$

For the simple regression model of $BODY\_FAT$ on a constant and WKG the OLS estimator of α_0 and α_1 are given by the following formulas,

$$\hat{\alpha}_0 = \overline{BODY\_FAT} - \hat{\alpha}_1 \overline{WKG}$$
$$\hat{\alpha}_1 = \frac{Cov(\widehat{BODY\_FAT}, WKG)}{\widehat{Var(WKG)}}$$

From the scatter plot, we can see that $BODY\_FAT$ and WKG have a positive linear relationship,

$$\therefore Cov(\widehat{BODY\_FAT}, WKG) > 0$$
$$\implies \hat{\alpha}_1 > 0$$

(c) which of the two regressions explained in parts (a) and (b) is likely to have a better fit?

$$\widehat{BODY\_FAT} = \hat{\beta}_0 + \hat{\beta}_1 ABDOMEN \quad (1)$$
$$\widehat{BODY\_FAT} = \hat{\alpha}_0 + \hat{\alpha}_1 WKG \quad (2)$$

The first estimated model is likely to fit the data better than the second. Why?

R^2, which is a measure of the models goodness of fit with the data is given by the formula,

$$R^2 = 1 - \frac{SSR}{SST}$$

where

- SSR - sum of squared residual

$$SSR = \sum_{i=1}^{n} (BODY\_FAT_i - \widehat{BODY\_FAT}) = \sum_{i=1}^{n} \hat{u}_i^2$$

- SST - sum of squared total

$$SST = \sum_{i=1}^{n} (BODY\_FAT_i - \overline{BODY\_FAT})$$

From the scatter plot of $BODY\_FAT$ against WKG, an OLS regression line of (1) through the scatter plot of $BODY\_FAT$ against $ABDOMEN$ would have a smaller SSR than the OLS regression line of (2),

$$SSR_{(1)} < SSR_{(2)}$$

Since SST depends only on *body fat* and sample mean of *body fat* it is the same for both estimated models,

$$SST_{(1)} = SST_{(2)} = SST$$

then the R^2 of (1) will be higher than the R^2 of (2),

$$R^2_{(1)} = 1 - \frac{SSR_{(1)}}{SST} > R^2_{(2)} = 1 - \frac{SSR_{(2)}}{SST}$$

In the scatter plot of $BODY\_FAT$ against $ABDOMEN$, $BODY\_FAT$ is less dispersed around $\overline{BODY\_FAT}$ for each value of $ABDOMEN$ than it is for WKG. (Think about R^2.)

(d) the sign of the coefficient of WKG in a regression of $BODY\_FAT$ on a constant, $ABDOMEN$ and WKG.

$$BODY\_FAT = \beta_0 + \beta_1 ABDOMEN + \beta_2 WKG + u$$

Scatter plots cannot tell us anything about the correlation of body fat and weight after controlling for the influence of abdomen circumference. So even though the scatter plot tells us that body fat and weight have a positive linear relationship,

$$\widehat{Cov}(BODY\_FAT, WKG) > 0$$

we do not know whether weight still has positive impact body fat after controlling for abdomen circumference.

(In fact, it actually has a negative impact after controlling for abdomen circumference! Think about 2 people with the same abdomen circumference i.e. controlling for $ABDOMEN$, but one weights more than the other. Since both have the same abdomen circumference, the one that is heavier will have weight distributed elsewhere in his body e.g. broader shoulders, thicker quads, fuller chest etc. If both males have the same abdomen circumference, the one with the bigger shoulders, quads, chest, etc. is likely to have a better physique and quite likely less body fat.)

Question 4

EViews workfile: *bodyfat.wf1*

With the same data as above, we have estimated three regressions:

$$\widehat{BODY\_FAT} = -12.63 + 0.39WKG, \qquad R^2 = 0.385, \bar{R}^2 = 0.382$$
$$\widehat{BODY\_FAT} = -38.60 + 0.62ABDOMEN, \qquad R^2 = 0.681, \bar{R}^2 = 0.679$$
$$\widehat{BODY\_FAT} = -42.94 + 0.91ABDOMEN - 0.27WKG, \qquad R^2 = 0.724, \bar{R}^2 = 0.722$$

(a) The signs of the R^2s of the first two regressions must agree with your answers to parts (a), (b), and (c) of the previous question. If they don't then discuss these in the tutorial or during consultation

They do!

(b)

(*Discussed in Question* 3(*d*))

(c) If weight was measured in pounds regard than kilograms (each kilogram is 2.2 pounds), how would the above regression results change? Check your answers by running the regressions using $bodyfat.wf1$ file.

$$WEIGHTLB = 2.2 \times WKG$$
$$WKG = \frac{WEIGHTLB}{2.2}$$

$$\begin{aligned}\widehat{BODY\_FAT} &= -12.63 + 0.39WKG \\ &= -12.63 + \frac{0.39}{2.2}WEIGHTLB \\ &= -12.63 + 0.177WEIGHTLB\end{aligned}$$

So we should find that when regressing body fat on a constant and weight in pounds, the estimated coefficient of weight in pounds equals to the estimated coefficient of weight in kg divided by 2.2

$$\frac{0.39}{2.2} = 0.177$$

The R^2 and other estimated coefficients stay the same.

Since weight in pounds equals to weight in kg times 2.2, a regression of body fat on a constant and weight in pounds is given by,

$$BODY\_FAT = \beta_0 + \beta_1 WKG^{*}2.2 + u$$

which gives the followed estimated regression model,

$$\widehat{BODY\_FAT} = \underset{(2.5550)}{-12.63} + \underset{(0.0142)}{0.1768} WKG$$

(d) If body fat was regressed on a constant only, what would the OLS estimate of the constant be? Answer it first and then check your answer using *bodyfat.wf1* file.

$$BODY\_FAT = \beta_0 + u$$

Since the regression model has only one parameter to estimate,

$$\beta_0$$

the OLS formula gives an estimate of β_0,

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \hat{\beta}_0$$

$$\mathbf{X}'\mathbf{X} = \underset{1 \times n}{\begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}} \underset{n \times 1}{\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}} = n$$

$$\therefore (\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n}$$

$$\mathbf{X}'\mathbf{y} = \underset{1 \times n}{\begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}} \underset{n \times 1}{\begin{bmatrix} BODY\_FAT_1 \\ BODY\_FAT_2 \\ \vdots \\ BODY\_FAT_n \end{bmatrix}} = \sum_{i=1}^{n} BODY\_FAT_i$$

$$\begin{aligned} \therefore \widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \frac{1}{n}\sum_{i=1}^{n} BODY\_FAT_i \\ &= \overline{BODY\_FAT} \\ &= \hat{\beta}_0 \end{aligned}$$