AREC422 Notes

Youpei Yan

February 7, 2019

1 Simple Linear Regression

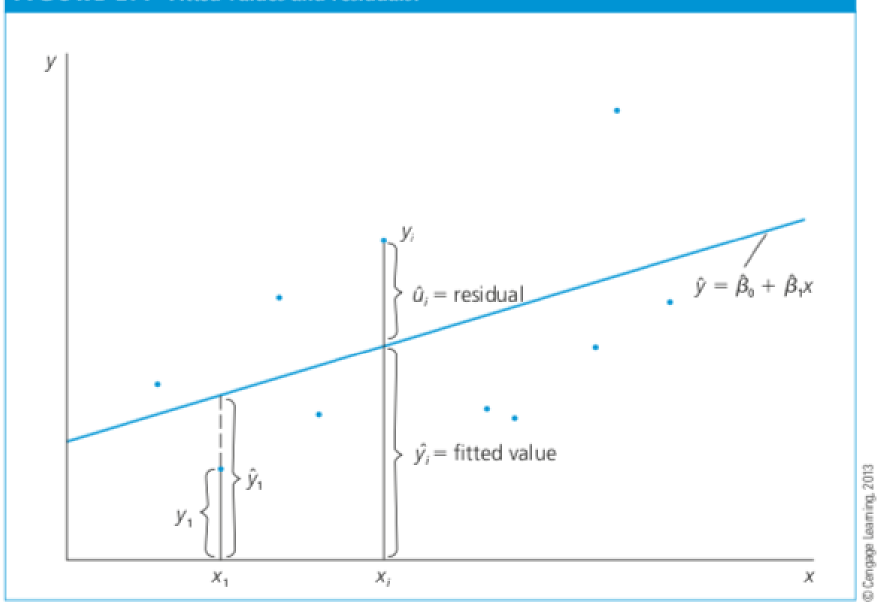
Continued from the last notes:

We have two equations with 2 unknowns: $\hat{\beta}_0$ and $\hat{\beta}_1$. We add hats on them because even the models are correct, they may not be the true β_0 and β_1, due to the sample collection.

Fitted value for y: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
Note that $\hat{y}_i$ is the values on the fitted line, it's not the same as the original y.

Residual: $\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$

FIGURE 2.4 Fitted values and residuals.



Sum of squared residuals (SSR)

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

We need it as small as possible to estimate the line (which means the estimation of $\hat{\beta}_0$ and $\hat{\beta}_1$.

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

also returns two equations (using the first order conditions by taking the derivatives). These two equations are the same as the equations we used above.

Anyway, we can derive $\hat{\beta}_0$ and $\hat{\beta}_1$ using either of the methods, and get the results:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{sample\ covariance}{sample\ variance\ of\ x_i}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

They are called the ordinary least square estimates (OLS estimates).

Use the price-demand example, where demand = y: (10,9,7,6,3,1), and price = x: (1,2,3,4,5,6). $\bar{y} = 6$, $\bar{x} = 3.5$.

Then plug all the numbers into the formulas:

$$\hat{\beta}_1 = \frac{(1-3.5)(10-6)+(2.3.5)(9-6)+(3-3.5)(7-6)+(4-3.5)(6-6)+(5-3.5)(3-6)+(6-3.5)(1-6)}{(1-3.5)^2+(2-3.5)^2+(3-3.5)^2+(4-3.5)^2+(5-3.5)^2+(6-3.5)^2}$$

And we get: $\hat{\beta}_1 = -1.83$, $\hat{\beta}_0 = 6 - (-1.83) * 3.5 = 12.40$

The fitted line is $\hat{y} = 12.40 - 1.83x$
When $x_1 = 1$, $\hat{y}_1 = 10.57$, which is close to 10 but not the same. The residual $\hat{u}_1 = 10 - 10.57 = -0.57$. The first point is below the line.
When $x_2 = 2$, $\hat{y}_2 = 12.40 - 1.83 * 2 = 8.74$, which is close to 9 but not the same. The residual $\hat{u}_2 = 9 - 8.74 = 0.26$. The second point is above the line.

Some relationships in OLS:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}\hat{u}_i^2$$

or SST = SSE + SSR.

The so called goodness of fit, R^2, is defined as

$$R^2 \equiv \frac{SSE}{SST} = 1 - \frac{SSR}{SST}, R^2 \in [0,1]$$

Remember that a low R^2 can also give a good estimate of the causal relationship. For instance, in the pollution example, even we control the number of cars and polluted rivers, there are still lots of things we don't observe that may influence pollution levels as well (say, weather info). All of them will be in the error term u. Residuals as the prediction of the errors can also be very large, which gives a low R^2 to us.

As long as all the weather info are not correlated with the number of firms, number of cars, or other things we controlled in the model, we are fine with the two assumptions we had. In this case, we have a valid/legit model.

2 Simple Linear Regression: Functional Forms

We use the "linear" functions in a broader way in this class, as long as that it has the following form: dep.var = constant + β_1 XXX + ...+ β_n XXX + error

So, the following models can all be estimated using the OLS technique, and they are all "linear functions" in our class.

$log(y) = \beta_0 + \beta_1 log(x) + u$
log-log model
$log(y) = \beta_0 + \beta_1 x + u$
log-level model
$y = \beta_0 + \beta_1 log(x) + u$
level-log model
$y = \beta_0 + \beta_1 x^2 + u$
quadratic model
$log(y) = \beta_0 + \beta_1 x^2 + \beta_2 log(x) + u$
$log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u$
model with an interaction term

Why do we need log functions?
Answer: because of the log function's derivative. Note that:

$$\frac{\partial log(x)}{\partial x} = \frac{dx}{x}$$

and we know that dx is just a small change in x (Δx). Then it is convenient for us to do:
(1) $\Delta x/x$ is approximately the percentage change in x
(2) elasticity analysis.

$$elasticity = \frac{\Delta y/y}{\Delta x/x}$$

For instance, the price elasticity of demand is

$$e_{(p)} = \frac{dQ/Q}{dP/P}$$

With a log-log function: $log(y) = \beta_0 + \beta_1 log(x) + u$, we have the following relationship:

$$\beta_1 = \frac{dy/y}{dx/x}$$

We estimate a $\hat{\beta}_1$ to represent this ratio.

Note that the underlying assumption of the log-log function is that y and x have a constant elasticity. The log-log function is also known as a constant-elasticity model. But we only use it when we have made the constant elasticity assumption, because elasticity in real life may be not constant.