Prediction and Residual Analysis

Suppose the model is

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u$$

Suppose y is the salary and x are the individual characteristics (experience, tenure, etc.) After the estimation of the model, suppose we want to predict the salary of an individual with certain experience, tenure, etc. $(x_1 = c_1, ..., x_k = c_k)$
The predicted salary is:

$$\hat{\theta}_0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + ... + \hat{\beta}_k c_k$$

In order to derive the confidence interval, we need to calculate the standard error $s.e.\left(\hat{\theta}_0\right)$

Easy way to do this:

Rewrite the linear model as follows:

$$y = \beta_0^* + \beta_1(x_1 - c_1) + \ldots + \beta_k(x_k - c_k) + u$$

The regression result is:

$$\hat{y} = \hat{\theta}_0 + \hat{\beta}_1(x_1 - c_1) + \ldots + \hat{\beta}_k(x_k - c_k)$$

You want to predict salary at $(x_1 = c_1, \ldots, x_k = c_k)$. Plug in those values:

$\hat{y} = \hat{\theta}_0$, which is the intercept.

So, in order to look at the standard error of the prediction $\hat{\theta}_0$, just take a look at the standard error of the intercept.

95% confidence interval:

$\left[\hat{\theta}_0 - t_{0.025} se(\hat{\theta}_0), \hat{\theta}_0 + t_{0.025} se(\hat{\theta}_0)\right]$, which is called the prediction interval.

But this is the confidence interval of $\hat{\theta}_0$, which is the predictor, which essentially is the prediction of the **average** salary at $(x_1 = c_1, ..., x_k = c_k)$. This tells you about the accuracy of the prediction of the average salary. But what would be more useful is the accuracy of the prediction with respect to the realized salary of an individual, which we want to predict.

Prediction error:

The true (realized) value:

$$y^0 = \beta_0 + \beta_1 c_1 + ... + \beta_k c_k + u^0$$

The predicted value:

$$\hat{\theta}_0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + ... + \hat{\beta}_k c_k$$

The difference, which is the prediction error:

$$\hat{e}^0 = y^0 - \hat{\theta}_0 = \left(\beta_0 - \hat{\beta}_0\right) + \left(\beta_1 - \hat{\beta}_1\right)c_1 + \ldots + \left(\beta_k - \hat{\beta}_k\right)c_k + u^0$$

Because u^0 is the new error term, not in the data, it is independent to all u_i's in the data, hence, orthogonal to $\hat{\beta}_i$'s.

$$E\left[\beta_l u^0\right] = E\left[\left(\beta_l + \frac{\sum r_{il} u_i}{\sum r_{il}^2}\right)u^0\right] = \beta_l E\left(u^0\right) + E\left[\frac{\sum r_{il} u_i}{\sum r_{il}^2}\right]E\left(u^0\right) = 0$$

Hence,

$$Cov\left(\hat{\theta}_0, u^0\right) = Cov\left(\hat{\beta}_0 + \hat{\beta}_1 c_1 + \ldots + \hat{\beta}_k c_k, u^0\right) = 0$$

So,

$$Var\left(\hat{e}^0\right) = Var\left(\hat{\theta}^0\right) + Var\left(y^0\right) + 2Cov\left(\hat{\theta}^0, y^0\right)$$

$$= Var\left(\hat{\theta}_0\right) + Var\left(u^0\right) + 0$$

The estimated standard error is:

$$se\left(\hat{e}^0\right) = \sqrt{Var\left(\hat{\theta}_0\right) + \hat{\sigma}^2}$$

95% **prediction interval** of the predictor is

$$\left[\hat{\theta}_0 - t_{0.025} s.e.\left(\hat{e}^0\right), \hat{\theta}_0 + t_{0.025} s.e.\left(\hat{e}^0\right)\right]$$

Example: housing price

Data:
Rooms: number of rooms
Baths : number of bathrooms.
Age : age of the house
Nbh : neighborhood rating (0 to 6)
Dist : distance to nearest incinerator (waste treatment)

Summary statistics.

	Obs.	Mean
price	321	96100.66
Annual price	321	7207.55
rooms	321	6.58567
baths	321	2.339564
age	321	18.00935
agesq	321	1381.567
nbh	321	2.208723
dist	321	20715.58

Annualize the price:

```
generate aprice = price*0.075
```

price regression

```
regress aprice rooms baths age agesq nbh dist
```

	coefficient	t-stat
const	1466.687	1.13
rooms	463.6997	2.31
baths	1792.686	6.61
age	-44.45106	-2.83
agesq	.1892436	1.92
nbh	-178.4794	-2.70
dist	-.0276951	-1.41

Now, do the procedure suggested by Wooldridge.
Predict at the mean:

c_1=6.58567, c_2 =2.339564, c_3 =18.00935
c_4 =1381.567, c_5 =2.208723, c_6 =20715.58

```
generate roomsp = rooms - 6.58567
generate bathsp  = baths - 2.339564
generate agep     = age   - 18.00935
generate agesqp = agesq - 1381.567
generate nbhp    = nbh   - 2.208723
generate distp     = dist  - 20715.58
```

regress aprice roomsp bathsp agep agesqp nbhp distp

	coefficient	Std. error
const	7207.55	136.1336
rooms	463.6997	200.5598
baths	1792.686	271.1221
age	-44.45105	15.732
agesq	.1892436	.0985865
nbh	-178.4794	66.21634
dist	-.0276951	.0195843

source	SS	df	MS
Model	1.4950e+09	6	249162257
Residual	1.8679e+09	314	5948884.71
Total	3.3629e+09	320	10509135.4

Now, the constant term is the predicted house price at the mean characteristics, which turns out to be exactly the mean housing price, 7207.55

Standard error of the prediction:

$$\sqrt{Variance(\theta_0)+\hat{\sigma}^2} = \sqrt{136.1+5948884} = 2442.8$$

Now, try to predict the effect of the unit deterioration in neighbhorhood quality.

$$c_5' = c_5 + 1$$

replace nbhp = nbhp-1

regress aprice roomsp bathsp agep agesqp nbhp distp

	coefficient	Std. error
const	7029.07	151.38
rooms	463.6997	200.5598
baths	1792.686	271.1221
age	-44.45105	15.732
agesq	.1892436	.0985865
nbh	-178.4794	66.21634
dist	-.0276951	.0195843

Change in predicted housing price: from 7207.55 to 7029.07.

source	SS	df	MS
Model	1.4950e+09	6	249162257
Residual	1.8679e+09	314	5948884.71
Total	3.3629e+09	320	10509135.4

Standard error of the prediction:

$$\sqrt{Variance(\theta_0)+\hat{\sigma}^2} = \sqrt{151.4^2 + 5948884} = 2443.7$$

Predicting y when $\log(y)$ is the Dependent Variable

Linear model:

$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$$

As we discussed before, the predictor of $\log(y)$ at x is

$$\hat{\log}(y) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k$$

Where $\hat{\beta}_j$'s are the OLS estimates.

Now, but suppose you want to predict y:

$$y = \exp[\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k] \times \exp(u)$$

Suppose that $u \sim N(0, \sigma_u^2)$, and is independent of x's. Then, it is well known that $\exp(u)$ is log normally distributed with mean $\exp\left[\frac{\sigma_u^2}{2}\right]$.

Therefore,

$$E[y \mid x] = \exp\left[\frac{\sigma_u^2}{2}\right]\exp[\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k]$$

Hence, if you want to predict y, you should use

$$\hat{y} = \exp\left[\frac{\hat{\sigma}_u^2}{2}\right]\exp\left[\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k\right]$$

But how about if we don't want to impose the assumption that the error term is normally distributed while still assuming that the error term is independent of the explanatory variables.

Then, we need to get the multiplicative constant $\hat{\alpha}_0$, i.e.

$$\hat{y} = \hat{\alpha}_0 \exp\left[\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k\right]$$

Step 1: Regress $\log(y_i)$ on $x_{il}, l = 1,\ldots,k$ to get $\hat{\beta}_j$, $j = 0,\ldots,k$.

Step 2: Regress y_i on $\exp\left[\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik}\right]$ without constant to get the coefficient $\hat{\alpha}_0$.

Multiple Regression Analysis with Qualitative Information: Binary (or Dummy) Variables.

Single Dummy Independent Variable:

$grant = 1$ if the firm received a training grant
$= 0$ if otherwise.

Wage equation:

$$hrstrain = \beta_0 + \delta_0 grant + \beta_1 \log(sales) + \beta_2 \log(employ) + u$$

$hrstrain$: hours of training per employee for a firm.

$sales$: annual sales

$employ$: number of employees.

Then, the coefficient δ_0 indicates the difference of hours of training between firms receiving a training grant and those that did not. That is, given sales and employment,

$$\delta_0 = E[grant \mid sales, employ] - E[no\ grant \mid sales, employ]$$

Intercept shift:
Intercept for firms without grant: β_0

Intercept for firms with grant: $\beta_0 + \delta_0$

Notice that we did not put dummy for no grant. Suppose we put a dummy no grant as follows

$ngrant = 1$ if the firm did not receive a training grant
$= 0$ if otherwise.

Then, the linear equation becomes

$$hrstrain = \beta_0 + \delta_0 grant + \delta_1 ngrant$$
$$+ \beta_1 \log(sales) + \beta_2 \log(employ) + u$$

Then, you cannot estimate the linear equation by OLS. This is because of the perfect collinearity of the independent variables. Notice that

$$grant + ngrant = 1$$

Which is the independent variable corresponding to the constant term.
This is called the **dummy variable trap**.

That is, if you have k categories and want to use dummy variables, you should drop one category and have only $k-1$ dummy variables.

In this case, intercept is the coefficient estimate of the excluded category, which is the base group.

If you absolutely want to use all the k dummy variables, then don't include the intercept.

$$hrstrain_i = \delta_0 grant_i + \delta_1 ngrant_i$$
$$+ \beta_1 \log(sales_i) + \beta_2 \log(employ_i) + u_i$$

Using Dummy Variables for Multiple Categories.

$$\hat{\log}(wage) = \underset{(0.110)}{0.321} + \underset{(0.055)}{0.213} marrmale - \underset{(0.058)}{0.198} marrfem$$
$$- \underset{(0.056)}{0.110} singfem + \underset{(0.007)}{0.079} educ - \underset{(0.005)}{0.027} exper - \underset{(0.00011)}{0.00054} exper^2$$
$$+ \underset{(0.007)}{0.029}\, tenure - \underset{(0.00023)}{0.00053}\, tenure^2$$

$n = 526$, $R^2 = 0.461$

Notice that we have 4 categories: female married, female single, male married and male single and we drop the male single from the dummy variables and make it the base group. For example:

$$marrymale = 1 \text{ if the individual is a married male}$$
$$= 0 \text{ otherwise.}$$

Holding education, experience and tenure constant, married males have higher wages than single ones, whereas the opposite holds for females.

Incorporating Ordinal Information by Using Dummy Variables:

Suppose the linear model is

$$\log(wage) = \beta_0 + \beta_1 educ + u$$

Where education in the data is:

$educ = 0$ if the individual did not graduate from highschool.

$educ = 1$ if the individual graduated from highschool.

$educ = 2$ if the individual graduated from college.

$educ = 3$ if the individual has graduate degree.

Then, the implicit assumption is that the marginal rate of return from highschool graduation is the same as that of college graduation and that of graduate school.

A way to estimate returns to education allowing different returns for every schools is to define following dummies.

$D1 = 1$ if the individual graduated from highschool
$= 0$ otherwise

$D2 = 1$ if the individual graduated from college
$= 0$ otherwise

$D3 = 1$ if the individual graduated from grad school
$= 0$ otherwise

$$\log(wage) = \beta_0 + \delta_1 D1 + \delta_2 D2 + \delta_3 D3 + u$$

δ_1: marginal returns to highschool degree.

δ_2: marginal returns to college degree.

δ_3: marginal returns to graduate degree.

$$\log(wage) = \beta_0 + \beta_1 educ + u$$

is a restriction of the above dummy model because

$$\delta_1 = \beta_1,\ \delta_2 = \beta_1,\ \delta_3 = \beta_1$$

The restriction is

$$\delta_1 = \delta_2 = \delta_3$$