Inference for Linear Regression

In the first module we learned using regression analysis we can answer 2 questions,

1. **The Prediction question:** How can we use the set of regressor X to predict Y well?

2. **The Inference question:** How does the predicted value(Y) change if we only change 1 component or 1 feature of X, keeping all other components constant.

In this section, we will provide an answer to the inference question.
To recall, we partition vectors of regressors X(x1, x2, ...xp)' into D and W, where D represents the "**target**" regressor of interest, and W represents other regressors, sometimes called the controls.
Therefore we write X as:

$$X = (D, W')' \quad \text{..... (1)}$$

and we re-write Y as equal to the predicted value,

$$Y = \underbrace{\beta_1 D + \beta_2' W}_{predicted\ \ value} + \underbrace{\epsilon}_{error} \quad \text{..... (2)}$$

And now we recall the inference question, which is: how does the predicted value of Y change if we increase D by a unit, holding W fixed?
Consider the wage example: **What is the difference in predicted wages between men and women with the same job-relevant characteristics?**
The answer to this question is the population regression coefficient **β1** corresponding to the target regressor D.
In the wage example, D is the female indicator(whether a person is a female or not) and β1 is the Gender Wage Gap.
Now, the next question is how to get this **β1?**
B1, can be found out using a method called *Partialling out.*
"Partialling-out" is an important tool that provides a conceptual understanding of the regression coefficient β1. The Steps for partialling out are:

1. We predict Y using W only and find its residuals. i.e. removing the dependence of W on Y.

2. We predict D using W and find its residuals, i.e. removing the dependence of W on D.
3. Then we model residuals from step 1 and step 2, and this will give us how Y is dependent on D only.

In the population, we define the partialling-out operation as a procedure that takes a random variable V and creates a "**residual**" V˜ by subtracting the part of V that is linearly predicted by W

$$\tilde{V} = V - \gamma_{VW}' W, \quad \gamma_{VW} = \arg\min_{\gamma} \mathrm{E}(V - \gamma' W)^2, \qquad \text{.....(3)}$$

When V is a vector, we apply the operation to each component. It can be shown that the partialling-out operation is linear:

$$Y = V + U \implies \tilde{Y} = \tilde{V} + \tilde{U}. \qquad \text{.....(4)}$$

We apply the partialling-out to both sides of our regression equation, to get

$$Y = \beta_1 D + \beta_2' W + \epsilon \qquad \text{.....(5)}$$

$$\tilde{Y} = \beta_1 \tilde{D} + \beta_2' \tilde{W} + \tilde{\epsilon}, \qquad \text{.....(6)}$$

which simplifies to the decomposition:

$$\tilde{Y} = \beta_1 \tilde{D} + \epsilon, \quad \mathrm{E}\epsilon\tilde{D} = 0. \qquad \text{.....(7)}$$

EϵD˜ is the same as the normal equation/first derivative, we saw in the second module.

This follows because partialling-out takes out $\beta'_2 W$, since W˜ = 0, and leaves ϵ untouched, ϵ˜ = ϵ, since ϵ is linearly unpredictable by X and therefore by W.

Moreover, EϵD˜ = 0 since D˜ is a linear function of X.

Frisch-Waugh Lovell, FWL Theorem

The decomposition (2) implies that EϵD˜ = 0 are the Normal Equations for the population regression of Y˜ on D˜. Thus:

Theorem (Frisch-Waugh-Lovell, FWL)

The population linear regression coefficient β_1 can be recovered from the population linear regression of $\tilde{Y}$ on $\tilde{D}$:

$$\beta_1 = \arg\min_{b_1} \mathrm{E}(\tilde{Y} - b_1\tilde{D})^2 = (\mathrm{E}\tilde{D}^2)^{-1}\mathrm{E}\tilde{D}\tilde{Y},$$

where β_1 is uniquely defined if D cannot perfectly predicted by W, i.e. $\mathrm{E}\tilde{D}^2 > 0$.

The FWL theorem is a remarkable fact. It asserts that β1 can be interpreted as a (univariate) regression coefficient of residualized Y on residualized D, where the residuals are defined by partialling-out the linear effect of W from Y and D.

How to do Estimation of beta1?

- In the sample, we will mimic the partialling-out in the population.
- When p/n is small, we can do this by sample linear regression.
- When p/n is not small, using sample linear regression for partialling-out is not a good idea. What we can do instead is penalized regression and variable selection, which we will discuss in other modules.

So let us assume that p/n is small, so it is appropriate to use the sample linear regression for partialling-out. Of course, by the FWL Theorem applied to the sample instead of the population, the sample linear regression of Y on D and W gives us an estimator $\hat{\beta}1$, which is numerically identical to the estimator obtained via sample partialling-out. It is still useful to give the formula for $\hat{\beta}1$ in terms of sample partialling-out, where we use **checked** quantities to denote the residuals that are left after predicting the variables in the sample with the controls. The population partialling-out is replaced here by the sample partialling out, where we replace the population expectation by the empirical expectation.

Multiple Linear Regression parameters can be estimated in two ways,

1. OLS
2. Partialling out Approach(FWL)

In both of the cases, we will get the same value for the parameter in which we are interested. This is explained in the FWL theorem. It follows the assumption that D(variable of interest) cannot perfectly predict the control variables(W).

So if we have OLS then why are we using Partialling out approach if they produce the same results?

1. This is useful in a variety of settings. For example, there may be cases where we would like to obtain the effect from a model that includes many regressors, and therefore a computationally infeasible matrix.
2. Its mostly used in the context of econometrics. Partialling out approach is used for panel data sets. The OLS approach would require you to run a regression with thousands of regressors, which is not a good idea numerically even nowadays with fast computers it would be very expensive(Heavy computation).

Mainly this is used in the context of high regressors and where computation is a bit heavy. There we can find the coefficient for the variable of interest by this approach.

Inference Result

Theorem (Inference)

If p/n is small, then the estimation error in $\check{D}_i$ and $\check{Y}_i$ has no first order effect on $\widehat{\beta}_1$, and

$$\widehat{\beta}_1 \overset{a}{\sim} N(\beta_1, V/n)$$

where

$$V = (\mathrm{E}\tilde{D}^2)^{-1}\mathrm{E}(\tilde{D}^2\epsilon^2)(\mathrm{E}\tilde{D}^2)^{-1}.$$

Using the formula, it can be shown that the following result holds. If p/n is small, then the estimation error in the estimated residualized quantities has a negligible effect on $\widehat{\beta}$ 1, and $\widehat{\beta}1$ is approximately distributed as a Normal variable with mean β1 and variance V/n, where the expression of the variance appears on the slide. we can say that the

estimator $\widehat{\beta}1$ concentrates in a $\sqrt{v/n}$ neighborhood of β1, with deviations controlled by the normal law.

We can now define the standard error of $\widehat{\beta}1$ as p $\sqrt{\widehat{v}/n}$, where $\widehat{v}$ is an estimator of V.

The result implies that the interval, given by the estimate plus/minus two standard errors, covers the true value β1 for most realizations of the data sample.

$$[\widehat{\beta}_1 - 2\sqrt{\widehat{V}/n}, \widehat{\beta}_1 + 2\sqrt{\widehat{V}/n}]$$

More precisely, approximately 95% of realizations of the data sample. If we replace 2 with other constants, we get other coverage probabilities. In other words, if our data sample is not extremely unusual, the interval covers the truth. For this reason, this interval is called the confidence interval.

For example, in the Wage Example, our estimate of the gender hourly wage gap is about −2$ and the 95% confidence interval is about −1$ to −3$.

We request you to check under Practice case studies Gender Wage Gap notebook, where this is applied.