AREC422 Notes

Youpei Yan

February 19, 2019

One thing forgot to mention: Confidence Intervals (CI) last time.
CI=$\hat{\beta}_j \pm c \cdot se(\hat{\beta}_j)$
where c is the critical values in the t-table.

For example: last time's demand $\sim$ price regression: read the table and find the critical values at the 5% level (2.776) and at the 1% level (4.604). Therefore, the 95% CI for β_j is $[-1.8286 - 2.776 * 0.1457, -1.8286 + 2.776 * 0.1457] = [-2.233, -1.4241]$

The other way to do the t-test with $H_0 : \beta_j = a_j$ vs. $H_1 : \beta_j \neq a_j$ is to see if a_j is in the CI. For instance, 0 is not in the range we calculated above, so we can reject H_0 at the 5% level.

1 Multiple Regression Analysis

Motivation:
SLR shows the simplest possible relationship between x and y, but it is very likely to violate GMAs ($E(u|X) = 0$). With more Xs, we are controlling more factors in the equation and moves more unobservables in the error term.
For instance, $housing\_price \sim dist\_incinerator + house\_char. + neighbor\_char.$ is more likely to give us unbiased estimators than $housing\_price \sim dist\_incinerator$.

We can also have richer functional forms to broaden the "linearity": $yield \sim temp + temp^2 + log(radia.$

OLS Estimates:

Equation: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$
With k independent variables $x_1, x_2, ...x_k$.

If we have a dataset for MLR, each of the variables is a vector (containing n observations):

# obs	y (pollution)	x_1 (# firms)	x_2 (# cars)	...	x_k (weather)
1	y_1	x_{11}	x_{21}		x_{k1}
2	y_2	x_{12}	x_{22}		x_{k2}
...					
n	y_n	x_{1n}	x_{2n}		x_{kn}

We should have a $(k+1) \cdot n$ matrix (and we must have $n \geq k + 1 to fit a line$)

Speaking of fitting a line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k$ is the fitted line. The residuals are still $\hat{u}_i = y_i - \hat{y}_i$.

The way to solve the model is still using OLS method to minimize SSR.

$$\min_{\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2 - ... - \hat{\beta}_k x_k)^2$$

First order condition will provide $k+1$ linear equations with $k+1$ unknowns simultaneously.

Recall that in the SLR, $\hat{\beta}_1$ is called the marginal effect. In the MLR,$\hat{\beta}_i$ is called the partial effect of x_i.

Holding $x_2, x_3, ...x_k$ fixed, $\hat{\beta}_1$ gives a ceteris paribus interpretation between x_1 and y. It is the same to say that after controlling for the effects of $x_2, x_3, ...x_k$, we estimate that 1 unit (or 1%, depending on the models we use) of x_1 will change y by $\hat{\beta}_1$ (or $100\hat{\beta}_1$%).

For instance, if we have a equation:

$$log(\widehat{housing\_price}) = 0.284 + 0.091 \cdot dist\_incinerator + 0.14 \cdot \#schools + 0.07 \cdot forestation\_area + ...$$

Holding $dist\_incinerator$ & $\#schools$ fixed, another acre of afforestation is predicted to increase the housing price by 7%.

With both increase of an acres of afforestation & 2km away from a garbage incinerator, the predicted housing price will increase by 25.2%. (0.091*2+0.07=0.252)

We can select one variable (with everyone else fixed) and examine its own impact to the dependent variable, or we can select a group of variables in the MLR.
(If the model is a level-log model, $y = \beta_0 + \beta_1 log(x_1) + u$, increasing x_1 by 1%, y increases by $\beta_1/100$ unit(s).)

In the MLR, GMAs also change/upgrade to the following version:
1. "linearity".
2. Random sampling.
3. No perfect collinearity. (None of the Xs is constant, no exact linear relationship among Xs.)
4. Zero conditional mean. $E(u|x_1, x_2, ...x_k) = 0$
5.Homoscedasticity. $Var(u|x_1, x_2, ...x_k) = \sigma^2$

From GMAs, we can derive 3 Theorems:
1. Unbiasedness of OLS: $E(\hat{\beta}_j) = \beta_j$, $j = 1, 2, ..., k$.
2. Unbiased estimation of σ^2. $E(\hat{\sigma}^2) = \sigma^2$

$$\hat{\sigma}^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} \hat{u_i}^2$$

3. (**Gauss-Markov Theorem**): The OLS estimator $\hat{\beta}_j$ for β_j is the redbest redlinear redunbiased redestimator (BLUE).

- "best": have the smallest variance for any estimator $\tilde{\beta}_j$ that is linear and unbiased. $Var(\hat{\beta}_j) < Var(\tilde{\beta}_j)$
- "linear": the estimator $\tilde{\beta}_j$ can be expressed as a linear function of the dependent variable: $\tilde{\beta}_j = \sum_{i=1}^{n} w_{ij} y_i$

Normality Assumption $u \sim N(0, \sigma^2)$+ GMA1-5: Classical linear model assumptions.