BLUE Properties and Multicollinearity

Pravat Uprety

BLUE Properties
(Best linear unbiased estimates)

- Unbiased ness
- Linearity
- Best ness (minimum variance)
- Consistency

BLUE properties

1. The estimators are linear, that is, they are linear functions of the dependent variable Y. Linear estimators are easy to understand and deal with compared to nonlinear estimators.
2. The estimators are unbiased, that is, in repeated applications of the method, on average, the estimators are equal to their true values.
3. In the class of linear unbiased estimators, OLS estimators have minimum variance. As a result, the true parameter values can be estimated with least possible uncertainty; an unbiased estimator with the least variance is called an efficient estimator.

Unbiased ness

Unbiased ness

Given the general linear model

$\underline{Y} = X\underline{\beta} + \underline{\epsilon}$

And we know that

$\underline{\hat{\beta}} = (X^t X)^{-1} (X^t Y)$

Now substituting $\underline{Y} = X\underline{\beta} + \underline{\epsilon}$ in the equation

we get,

$\underline{\hat{\beta}} = (X^t X)^{-1} X^t (X\underline{\beta} + \underline{\epsilon})$

$= (X^t X)^{-1} (X^t X)\, \underline{\beta} + (X^t X)^{-1} X^t \underline{\epsilon}$

$= \underline{\beta} + (X^t X)^{-1} X^t \underline{\epsilon}$

Taking expectation on both sides

$$E(\underline{\hat{\beta}}) = E[\underline{\beta} + (X^t X)^{-1} X^t \underline{\epsilon}]$$

$$= E(\underline{\beta}) + (X^t X)^{-1} X^t E(\underline{\epsilon})$$

$$= \underline{\beta} + (X^t X)^{-1} X^t (0)$$

$$E(\underline{\hat{\beta}}) = \underline{\beta}$$

This proves that $\underline{\hat{\beta}}$ is an unbiased estimator of $\underline{\beta}$.

Linearity

We know that

$$\underline{\hat{\beta}} = (X^t X)^{-1} (X^t Y)$$

And it is clear that $\underline{\hat{\beta}}$'s have a linear relation with $\underline{Y}$ with the weights being functions of X data, which are non stochastic.

Bestness

We know that

$\underline{\hat{\beta}} = \underline{\beta} + (\underline{X}^t X)^{-1} \underline{X}^t \underline{\epsilon}$ ----------(1)

Or, $\underline{\hat{\beta}} - \underline{\beta} = (\underline{X}^t X)^{-1} \underline{X}^t \underline{\epsilon}$

The variance of $\underline{\hat{\beta}}$ is

$\text{Var}(\underline{\hat{\beta}}) = E[(\underline{\hat{\beta}} - \underline{\beta})(\underline{\hat{\beta}} - \underline{\beta})^t]$

$= E[(\underline{X}^t X)^{-1} \underline{X}^t \underline{\epsilon} \{(\underline{X}^t X)^{-1} \underline{X}^t \underline{\epsilon}\}^t]$

$= E[(\underline{X}^t X)^{-1} \underline{X}^t \underline{\epsilon}\, \underline{\epsilon}^t X (\underline{X}^t X)^{-1}]$

$= (\underline{X}^t X)^{-1} \underline{X}^t E(\underline{\epsilon}\, \underline{\epsilon}^t) X (\underline{X}^t X)^{-1}$

$= (\underline{X}^t X)^{-1} \underline{X}^t \sigma^2 X (\underline{X}^t X)^{-1}$

$= \sigma^2 (\underline{X}^t X)^{-1} (\underline{X}^t \underline{X}) (\underline{X}^t X)^{-1}$

$= \sigma^2 (\underline{X}^t X)^{-1}$ ----------(2)

Example: ceo salary and return on equity

$$\hat{y} = 963.191 + 18.501\, x$$

n= 209 $R^2 = 0.0132$

The firm's return on equity explains only about 1.3% of the variation in salaries.

Log transformation

- Generally, log transformation is used to obtain a constant elasticity model.
- A constant elasticity model is
- log (salary) = β_0 + β_1 log(sales) + ε

Where sales is annual firm sales, measured in million of dollars.

β_1 is the elasticity of salary with respect to sales.

##Command in R

- lm (y~x, data = name of data frame)

##Open CEOSAL1 data

```
attach(CEOSAL1)
dim(CEOSAL1)
head(CEOSAL1)
```

##Using dplyr

```
library(dplyr)
CEOSAL1 %>%  summarise(avg_sal=mean(salary), avg_roe=mean(roe))
```

```
##Summary statistics using R
mean(salary)
mean(roe)
cov(salary, roe)
var(salary)
var(roe)
##Manual Calculation in R
(slope= cov(roe,salary)/var(roe))
(yintercept = mean(salary)-slope*mean(roe))
```

Regression analysis

- ##Direct regression
- lm(salary~roe)
- reg_ceo<-lm(salary~roe)
- summary(reg_ceo)
- salhat<-fitted(reg_ceo)
- uhat<-resid(reg_ceo)

- cbind(CEOSAL1, salhat,uhat)
- BIC(reg_ceo)
- detach(CEOSAL1)

Log linear model

- For the estimation of logarithmic or semi logarithmic models, the lm formula can be directly used
- lm(log(salary)~log(sales), data =CEOSAL1)

Common used technique

```
reg _ceo<-lm(salary~roe, data=CEOSAL1)
summary(reg_ceo)
max(CEOSAL1$salary)
plot(density(CEOSAL1$salary))
##Scatter diagram
plot(CEOSAL1$roe, CEOSAL1$salary)
plot(CEOSAL1$roe, CEOSAL1$salary, ylim=c(0,4000))
abline(reg_ceo)    [to draw the regression line in plot]
##Log linear model
ln_sal<-lm(log(salary)~log(sales), data =CEOSAL1)
summary(ln_sal)
```

Matrix approach

```
##determine sample size and no. of regressors:
n<-nrow(CEOSAL1); k<-1
n
#Extract y
y<- CEOSAL1$salary
x <- (cbind(1, CEOSAL1$roe))
dim(x)
head(x)
##Parameter estimate (matrix approach)
(bhat<-solve(t(x)%*%x) %*% t(x) %*%y)
#Residual
uhat<-y-x%*%bhat
Uhat
mean(uhat)
var(uhat)
```

Packages

- ##to get better presentation we can install jtools package
- install.packages("jtools")
- library(jtools)

- install.packages("huxtable")
- library(huxtable)

- install.packages("car")
- library(car)

- summary(reg_ceo)
- summ(reg_ceo)

Example: Wage data and dummy in independent variable

Wage: Hourly wage in dollars, which is the dependent variable.

The explanatory variables, or regressors, are as follows:

Female: Gender, coded 1 for female, 0 for male

Nonwhite: Race, coded 1 for nonwhite workers, 0 for white workers

Union: Union status, coded 1 if in a union job, 0 otherwise

Education: Education (in years)

##Opening file

- wage <- read_excel("wage.xls")
- library(dplyr)
- head(wage)

#Regression

- reg_wage<- lm(wage~female+nonwhite+union+education+exper, data=wage)
- reg_wage
- summary (reg_wage)

##to obtain confidence interval estimate

- confint(reg_wage)

Running regression

- reg_wage<- lm(wage~female+nonwhite+union+education+exper, data=wage)
- reg_wage1<- lm(wage~female+nonwhite+union+education, data=wage)
- summ(reg_wage1)

##To compare two or more models

- export_summs(reg_wage, reg_wage1)
- summ(reg_wage1, scale = TRUE, vifs = TRUE, part.corr = TRUE, confint = TRUE, pvals = FALSE)

Result using summary command

```
Residuals:
    Min      1Q  Median      3Q     Max
-20.781  -3.760  -1.044   2.418  50.414

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.18334    1.01579  -7.072 2.51e-12 ***
female      -3.07488    0.36462  -8.433  < 2e-16 ***
nonwhite    -1.56531    0.50919  -3.074  0.00216 **
union        1.09598    0.50608   2.166  0.03052 *
education    1.37030    0.06590  20.792  < 2e-16 ***
exper        0.16661    0.01605  10.382  < 2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Result using summ command after installing jtools (summ(reg_wage))

F(5,1283) = 122.61, *p* = 0.00
R^2 = 0.32
Adj. R^2 = 0.32

Standard errors:OLS

	Est.	S.E.	t val.	p
(Intercept)	-7.18	1.02	-7.07	0.00
female	-3.07	0.36	-8.43	0.00
nonwhite	-1.57	0.51	-3.07	0.00
union	1.10	0.51	2.17	0.03
education	1.37	0.07	20.79	0.00
exper	0.17	0.02	10.38	0.00

Measuring goodness of fit in multiple regression analysis

- The goodness of fit of the estimated model is understood in terms of value of R^2 (coefficient of multiple determination) and R^2 statistic provides a measure of proportion of total variation in the dependent variable that is explained by the independent/explanatory variables in the model.

$$R^2 = \frac{Explained\ Variation}{Total\ Variation} = \frac{ESS}{TSS}$$

Adj R^2 penalizes R^2 for adding additional explanatory variables in the model so that Adj R^2 values of models having different number of explanatory variables become comparable.

However, Adj R^2 is not only statistic (criteria) used by the researchers to select the best fit model from a set of alternative models.

other model selection criteria that

- Akaike Information Criteria (AIC)
- Schwarz Bayesian Criteria (SBC) (BIC)
- Hannan-Quinn Criteria (HQC)

$$AIC = -2 * log(L) + 2 * k$$

$$BIC = -2 * log(L) + k * ln(n)$$

$$HQ = -2 * log(L) + 2 * k * ln(ln(n))$$

While selecting a model based on these criteria, we select the one that reports minimum values for all these criteria (statistics) compared to an alternative model.

Model Selection

- ###Model Selection criteria
- anova(reg_wage1, reg_wage)
- ###By using AIC
- AIC(reg_wage1, reg_wage)
- BIC(reg_wage1, reg_wage)

- library(car)
- outlierTest(reg_wage)

Usefull Function/Command

Function	Action
`summary()`	Displays detailed results for the fitted model
`coefficients()`	Lists the model parameters (intercept and slopes) for the fitted model
`confint()`	Provides confidence intervals for the model parameters (95% by default)
`fitted()`	Lists the predicted values in a fitted model
`residuals()`	Lists the residual values in a fitted model
`anova()`	Generates an ANOVA table for a fitted model, or an ANOVA table comparing two or more fitted models
`vcov()`	Lists the covariance matrix for model parameters
`AIC()`	Prints Akaike's Information Criterion
`plot()`	Generates diagnostic plots for evaluating the fit of a model
`predict()`	Uses a fitted model to predict response values for a new dataset

Useful functions for regression diagnostics (car package)

Function	Purpose
`qqPlot()`	Quantile comparisons plot
`durbinWatsonTest()`	Durbin–Watson test for autocorrelated errors
`crPlots()`	Component plus residual plots
`ncvTest()`	Score test for nonconstant error variance
`spreadLevelPlot()`	Spread-level plots
`outlierTest()`	Bonferroni outlier test
`avPlots()`	Added variable plots
`influencePlot()`	Regression influence plots
`scatterplot()`	Enhanced scatter plots
`scatterplotMatrix()`	Enhanced scatter plot matrixes
`vif()`	Variance inflation factors

Multicollinearity

- Multicollinearity refers to a situation where there are high inter correlations among the explanatory variables of a multiple regression model. Multicollinearity problem arises only in the context of multiple regressions, it is considered as a problem because when the explanatory variables are highly correlated, most of their variation is common so that there is little variation unique to each variable.
- In empirical econometrics, while estimating multiple regression model, quite often we obtain unsatisfactory results in the sense that a good number of the estimated coefficients and found to be statistically insignificant. This happens when variances and hence standard errors of the estimated coefficients are large. This is possible when there is little variation in explanatory variables or high inter correlations among the explanatory variables or both.

Overall test (F test) ----------Reject

Individual test (t test) ------Do not reject

(problem of multicollinearity)

Perfect correlation between two explanatory variables ------- (-1 or +1)

We can not run the regression

Y	X1	X2	X3	X4	X5
		Highly correlated			

Sources of Multicollinearity

- Multicollinearity may arise for several reasons

1) Multicollinearity may arise because of faulty data collection method. For example, sampling over a limited range of values of explanatory variables creates this problem.

2) Multicollinearity problem arises when the explanatory variables share a common time trend. This is the case in time series regressions. For instance, in regression of GDP on money supply and prices, the two explanatory variables (money supply and prices) are likely to be highly correlated because when money supply rises in an economy, price level also rises.

3) When lagged values of the same variable are included as explanatory variables, we are likely to have multicollinearity problem. For example, in a time series regression of area under cultivation for a crop (say wheat) on its current and past prices, we may have multicollinerity problem because prices of the crop at different time points are generally correlated.

4) There are many cross section regressions where we face high inter correlations among the explanatory variables.

   - For example, in a regression of consumption expenditure of persons on their income and education levels, we typically have strong correlation between income and education variables as the persons reporting higher incomes are usually found to be more educated.

Sources

5. Multicollinearity arises for faulty specification of the model. For example, adding polynomial terms (X and X^2) when X ranges are small will create this problem.

6. Multicollinearity arises in the over determined model where number of explanatory variables (k) is greater than number of observations (n).

7. In a situation of dummy variable trap we have multicollinearity problem. In this situation, the model includes an intercept term and the number of dummies is equal to number of categories.

Consequences of Multicollinearity

- No multicollinearity

- Perfect multicollinearity
- under perfect multicollinearity it is not possible to compute the values of OLS estimates
- Imperfect multicollinearity

Detection of multicollinearity

1) Correlation matrix
2) F statistic and t statistic
3) Klen's rule of thump
4) Variance inflation factor (VIF)
5) Tolerance method (TOL)

- VIF = 1 (no multicollinearity)
- 1 < VIF < 5 (less multicollinearity)
- 5 ≤ VIF ≤ 10 (Moderate multicollinearity)
- VIF > 10 (High multicollinearity)

Correlation matrix

- Since multicollinearity is caused by inter correlation among the explanatory variables, some idea about this problem may be obtained by computing simple or zero order correlations between the explanatory variables. To understand these correlations, the researchers usually obtain the correlation matrix by using different software.
- However, it is to be remembered that using simple correlation coefficient to understand presence of multicollinearity is a valid procedure when the model has two explanatory variables. When the model includes more than two explanatory variables instead of simple correlations, we should consider the partial correlations which examine the influence of one variable upon another after eliminating the effects of all other variables. The statistical significance of partial correlation coefficient may also be tested by applying the t test procedure, but formula for computation of t is different here.
-

F statistic and t statistic

- F-statistic and t-statistic
- In the data having multicollinearity problem we observe that its presence generates the high variances for the OLS estimates, thereby providing low t-ratios and hence statistically insignificant regression results that is there seems contradictory result in F-test and t-test.

Klein's Rule of thumb:

Klein (1962) suggested a rule of thumb according to which multicollinearity would be regarded as a problem if $R_Y^2 < R_k^2$, where R_Y^2 is the squared multiple correlation coefficient between the dependent variable Y_i and explanatory variables X_{1i}, X_{2i},…….X_{ki}, and R_k^2 is squared multiple correlation coefficient between K^{th} explanatory variable and other explanatory variables.

Variance Inflation Factor

When the multicollinearity is present, the variance of the estimated coefficient of K^{th} explanatory variable is measured by

$$\text{Var}(\hat{\beta}_k) = \frac{\sigma_\varepsilon^2}{\Sigma x_{ki}^2 (1 - R_k^2)}$$

Under the ideal situation when there is no multicollinearity, $R_k^2 = 0$, so that,

$$\text{Var}(\hat{\beta}_k) = \frac{\sigma_\varepsilon^2}{\Sigma x_{ki}^2}$$

The VIF compares these two situations by taking a ratio of the two variances.

Decision criteria by using VIF

Thus,

$$\text{VIF}(\hat{\beta}_k) = \frac{\sigma_\varepsilon^2}{\Sigma x_{ki}^2 (1-R_k^2)} \Big/ \frac{\sigma_\varepsilon^2}{\Sigma x_{ki}^2} = \frac{1}{(1-R_k^2)}$$

VIF = 1 (no multicollinearity)

1 < VIF < 5 (less multicollinearity)

5 ≤ VIF ≤ 10 (Moderate multicollinearity)

VIF > 10 (High multicollinearity)

Tolerance method (TOL) = 1/VIF $= 1-R_k^2$

Remedial Technique

1) By increasing sample size
2) Transformation of variables
3) Using extraneous estimates
4) Dropping the variables
    - We have to drop the variable having highest VIF (>10) at a time. Again run the regression of remaining variables check whether there is VIF >10 or not and repeat the same process until we get VIF < 10.

QQ test for normality

- qqPlot(reg_wage)
- qqPlot(reg_wage, id.method="identify", simulate=TRUE, main="Q-Q Plot")

- ##To check the multicollinearity
- vif(reg_wage)
- residualPlot(reg_wage)

- ##To check the autocorrelation
- durbinWatsonTest(reg_wage)

Usefull Function/Command

Function	Action
`summary()`	Displays detailed results for the fitted model
`coefficients()`	Lists the model parameters (intercept and slopes) for the fitted model
`confint()`	Provides confidence intervals for the model parameters (95% by default)
`fitted()`	Lists the predicted values in a fitted model
`residuals()`	Lists the residual values in a fitted model
`anova()`	Generates an ANOVA table for a fitted model, or an ANOVA table comparing two or more fitted models
`vcov()`	Lists the covariance matrix for model parameters
`AIC()`	Prints Akaike's Information Criterion
`plot()`	Generates diagnostic plots for evaluating the fit of a model
`predict()`	Uses a fitted model to predict response values for a new dataset

Useful functions for regression diagnostics (car package)

Function	Purpose
`qqPlot()`	Quantile comparisons plot
`durbinWatsonTest()`	Durbin–Watson test for autocorrelated errors
`crPlots()`	Component plus residual plots
`ncvTest()`	Score test for nonconstant error variance
`spreadLevelPlot()`	Spread-level plots
`outlierTest()`	Bonferroni outlier test
`avPlots()`	Added variable plots
`influencePlot()`	Regression influence plots
`scatterplot()`	Enhanced scatter plots
`scatterplotMatrix()`	Enhanced scatter plot matrixes
`vif()`	Variance inflation factors