Linear Regression Properties and Goodness of fit

Babak Rezaee Daryakenari
The Institute of Political Science
Leiden University
s.rezaeedaryakenari@fsw.leidenuniv.nl
For more research methods handouts:
https://babakrezaee.github.io

Spring 2020

Review of the last session

The Gauss-Markov Assumptions

1. **Linearity assumption:** $y = \beta_0 + \beta_1 x + \epsilon$
2. X **is a full rank matrix.**
3. $E(\epsilon|X) = 0$
4. $E(\epsilon\epsilon'|X) = \sigma^2 I$
5. X **and** y **are orthogonal** $X \perp \epsilon$
6. $\epsilon|X \sim N(0, \sigma^2 I)$

Th assumption 6 is not actually required for the Gauss-Markov Theorem. However, we often assume it to make hypothesis testing easier.

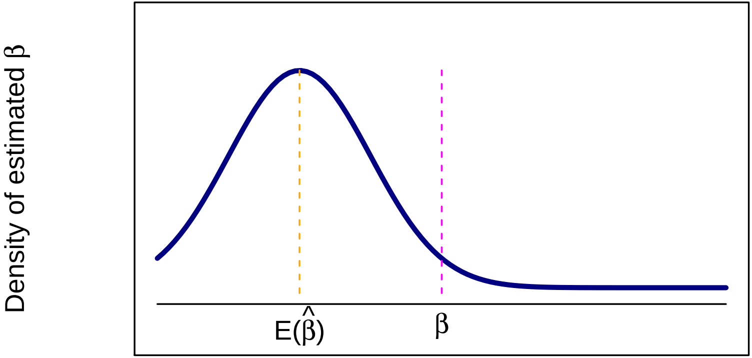
Gauss-Markov Theorem

The **Gauss-Markov Theorem** states that, conditional on assumptions 1-5, OLS estimator is the Best Linear, Unbiased, and Efficient estimator **(BLUE)**:

1. $\hat{\beta}$ is an unbiased estimator of β.
2. $\hat{\beta}$ is a linear estimator of β.
3. $\hat{\beta}$ has minimal variance among all linear and unbiased estimators.

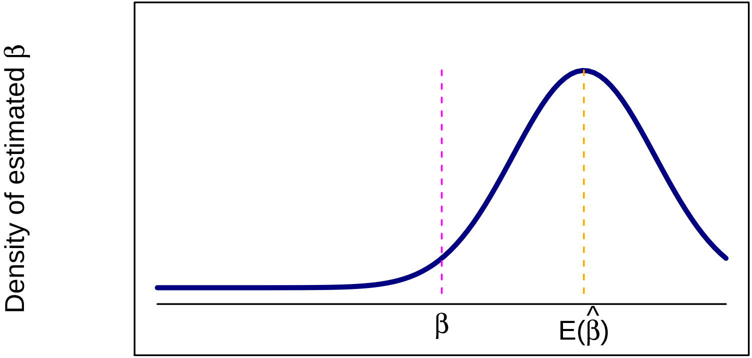
Biased vs. unbiased estimation

- $E(\hat{\beta}) < \beta$



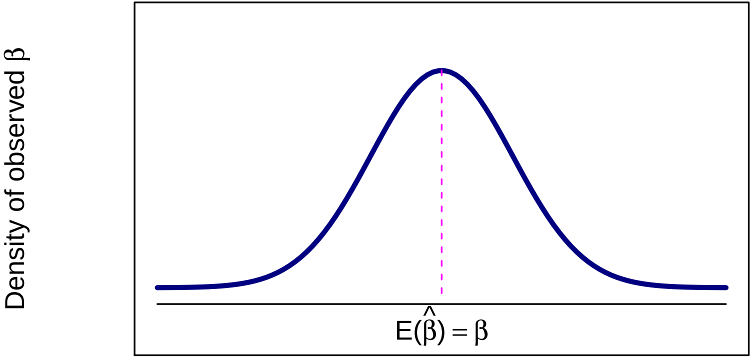
Biased vs. unbiased estimation

- $E(\hat{\beta}) > \beta$



Biased vs. unbiased estimation

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



Properties of OLS

OLS regression: an example

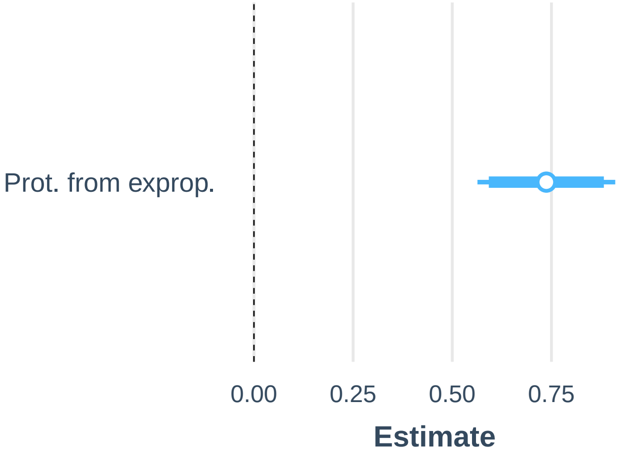
A regression model of $y = \beta_0 + \beta_1 x + \epsilon$ estimated using OLS method has some properties that can help us to analyze the estimated model and whether some of the **Gauss-Markov Theorem** assumptions are satisfied.

```
# try jtools library for cleaner regression report and of course some other interesting tools
library(jtools)
# OLS model
lin.mod <- lm(logpgp95~avexpr, data=rawData)
summ(lin.mod)
```

```
## MODEL INFO:
## Observations: 57
## Dependent Variable: logpgp95
## Type: OLS linear regression
##
## MODEL FIT:
## F(1,55) = 72.32, p = 0.00
## R² = 0.57
## Adj. R² = 0.56
##
## Standard errors: OLS
## ------------------------------------------------
##                      Est.   S.E.   t val.      p
## ----------------- ------ ------ -------- ------
## (Intercept)         4.89   0.38    12.79   0.00
## avexpr              0.49   0.06     8.50   0.00
## ------------------------------------------------
```

Plotting the regression results

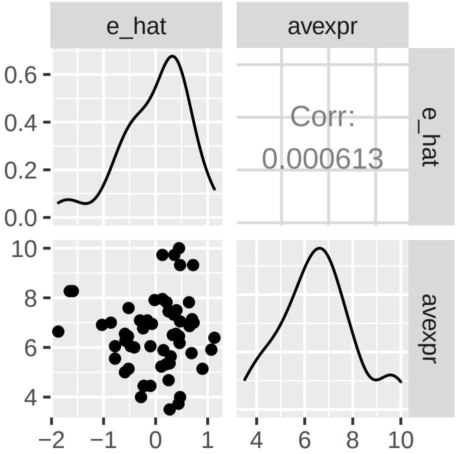
```
## You need "ggstance" package installed for plot_summs
plot_summs(lin.mod,  scale = TRUE,
           coefs = c("Prot. from exprop."="avexpr" ),
           inner_ci_level = .9)
```



Properties of OLS: 1

The observed values of X are uncorrelated with the residuals(regression errors).

```
#Predictions
rawData$logpgp95_hat<- predict(lin.mod)
#Errors (residuals)
rawData$e_hat<- rawData$logpgp95-rawData$logpgp95_hat
```



Properties of OLS:2

The sum of the residuals is zero.

If there is a constant, then the first column in X will be a column of ones. This means that for the first element in the $X'e$ vector (i.e. $X_{11}e_1 + x_{21}e_2 + \cdots + x_{N1}e_N$) to be zero, it must be the case that $\sum_1^N e_i = 0$.

Properties of OLS:2

The sum of the residuals is zero.

If there is a constant, then the first column in X will be a column of ones. This means that for the first element in the $X'e$ vector (i.e. $X_{11}e_1 + x_{21}e_2 + \cdots + x_{N1}e_N$) to be zero, it must be the case that $\sum_1^N e_i = 0$.

```
sum(rawData$e_hat)
```

```
## [1] 0.00000000000001865175
```

```
round(sum(rawData$e_hat), 3)
```

```
## [1] 0
```

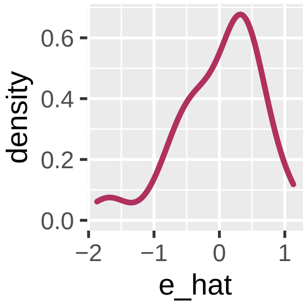
Properties of OLS: 3

The sample mean of the residuals is zero. This follows straightforwardly from the previous property, $\bar{e} = \frac{\sum_1^N e_i}{N} = 0$.

```
mean_ehat=mean(rawData$e_hat)
round(mean_ehat,2)
```

```
## [1] 0
```

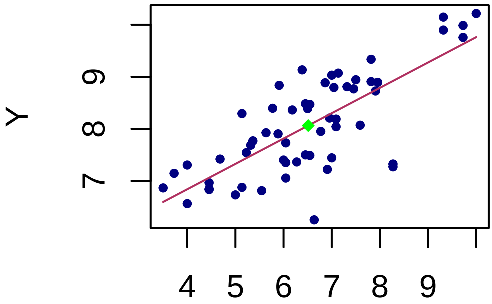
```
library(ggplot2)

ggplot(rawData, aes(x = e_hat)) +
  geom_density(col='maroon', lwd=1)
```



Properties of OLS: 4

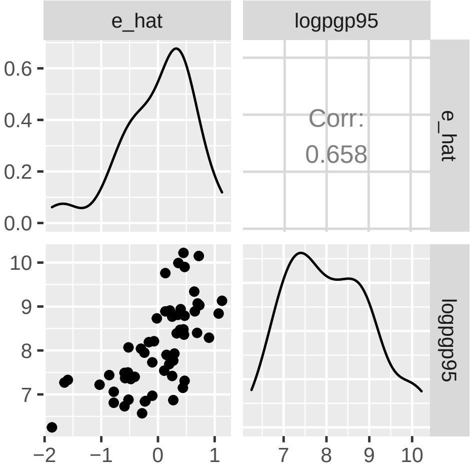
The regression hyperplane, line in a bivariate model, passes through the means of the observed values ($\bar{x}$ and $\bar{y}$).

```
plot(rawData$avexpr, rawData$logpgp95 , col="navy", pch=19, cex=.5,
     xlab="X", ylab="Y")
lines(logpgp95_hat~avexpr,data=rawData , col="maroon", lwd=1)
points(mean(rawData$avexpr), mean(rawData$logpgp95), pch = 18, col = "green", cex = .9)
```



Properties of OLS: 5

The predicted values of y are uncorrelated with the residuals (errors).



Goodness of fit: R^2 and R^2-adjusted

How good is an estimated regression?

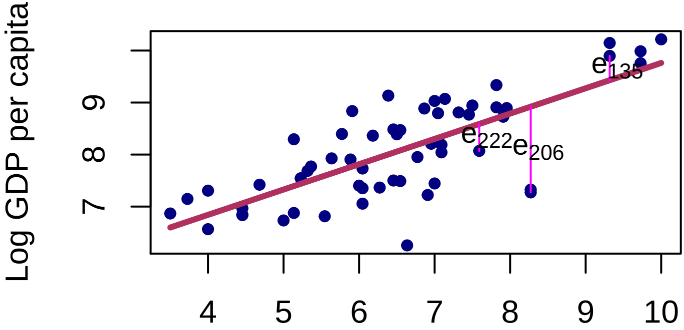
- We learned about the concept of (un)biased estimation. An unbiased estimation shows that how well we quantified the association between variable x, or variable xs in multivariate regression, and variable y.
- Another measure of a good estimation is how much of variations in the outcome variable y is explained using the independent variable(s).
- There is a subtle difference between these two measures. The first part is the focus of *causal inference* models while the second one is the focus of *prediction/forecasting* models.
- This part of the course focuses on the goodness of fit/prediction power/forecasting power/explanatory power of a regression model.

Root of Mean Square Error

- The sum of squared errors SSE_{OLS}:

$$SSE_{OLS} = \sum_{i=1}^{N} e^2 = \sum_{i=1}^{N} (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_N - \hat{y}_N)^2 \quad (1)$$

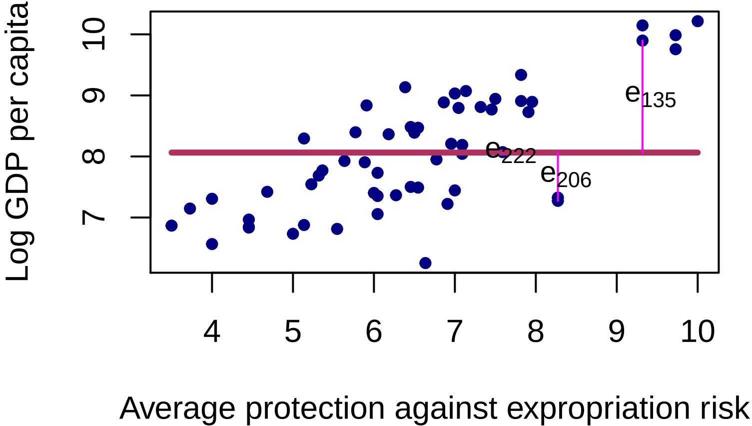
$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} e^2}{N}} \quad (2)$$



R-squared (coefficient of determination)

- What if we want to compare RMSE with a benchmark? We can use $SSE_{\bar{y}}$:

$$SSE_{\bar{y}} = \sum_{i=1}^{N} (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_N - \bar{y})^2 \qquad (3)$$



R-squared

- How much of the variation in y, measured by $SSE_{\bar{y}}$, is not explained by the OLS model, measured SSE_{OLS}:

$$\frac{SSE_{OLS}}{SSE_{\bar{y}}} \tag{4}$$

- Therefore, the amount variation in outcome y that is explained by OLS is:

$$R^2 = 1 - \frac{SSE_{OLS}}{SSE_{\bar{y}}} \tag{5}$$

R^2: An example

```
## MODEL INFO:
## Observations: 57
## Dependent Variable: logpgp95
## Type: OLS linear regression
##
## MODEL FIT:
## F(1,55) = 28.37, p = 0.00
## R² = 0.34
## Adj. R² = 0.33
##
## Standard errors: OLS
## ------------------------------------------------
##                      Est.   S.E.   t val.      p
## ----------------- ------ ------ -------- ------
## (Intercept)          7.76   0.12    64.31   0.00
## democ00a             0.19   0.04     5.33   0.00
## ------------------------------------------------
```

R^2: An example

```
## MODEL INFO:
## Observations: 57
## Dependent Variable: logpgp95
## Type: OLS linear regression
##
## MODEL FIT:
## F(1,55) = 28.37, p = 0.00
## R² = 0.34
## Adj. R² = 0.33
##
## Standard errors: OLS
## ------------------------------------------------
##                      Est.   S.E.   t val.      p
## ----------------- ------ ------ -------- ------
## (Intercept)          7.76   0.12    64.31   0.00
## democ00a             0.19   0.04     5.33   0.00
## ------------------------------------------------
```

```
## MODEL INFO:
## Observations: 57
## Dependent Variable: logpgp95
## Type: OLS linear regression
##
## MODEL FIT:
## F(2,54) = 13.94, p = 0.00
## R² = 0.34
## Adj. R² = 0.32
##
## Standard errors: OLS
## ------------------------------------------------
##                      Est.   S.E.   t val.      p
## ----------------- ------ ------ -------- ------
## (Intercept)          7.74   0.19    40.15   0.00
## democ00a             0.19   0.04     5.20   0.00
## cons1                0.01   0.05     0.14   0.89
## ------------------------------------------------
```

R^2-adjusted

How can we penalize unnecessary complexity of a regression model?

$$R^2_{adjusted} = 1 - (1 - R^2)[\frac{n-1}{n-k-1}] \tag{6}$$

where k is the number of regressors.