# OLS and Model Inference

Isabel Casas
Office: V5-206a-2
icasas@sam.sdu.dk

# Using R in the Terminalrums

- Go to Start
- Choose Run and type cmd
- In the MS-DOS page S:\R\R-2.12.1\bin\i386\Rgui.exe
- If some needed libraries are not there, let me know. They need to install them.

# Outline

- OLS estimator
  - Reminder
  - Bias
  - Variance
  - Asymptotic properties: consistency and asymptotic normality
- Model diagnostic
  - T-test on the parameter estimators
  - F-test on a group of parameters
- Problems with the OLS assumptions
  - Endogeneity
  - Identificability
  - Heteroskedasticity

# Fitted model

The multivariate linear model may be written in **matrix form**:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{k1} \\ 1 & x_{12} & \dots & x_{k2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \dots & x_{kn} \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \qquad \text{fitted values}$$

$$\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \hat{\mathbf{y}} \qquad \text{residuals}$$

# Estimation Methodology Assumptions

OLS.1 $E(\mathbf{X}'\boldsymbol{\epsilon}) = 0$. This is weaker than MLR.4 $(E(\boldsymbol{\epsilon}|\mathbf{x}_1, \ldots, \mathbf{x}_k) = 0)$

OLS.2 No perfect collinearity: none of the $x$'s are constant and one cannot be written as a linear relationship of the others. The rank of $E(\mathbf{X}'\mathbf{X})$ is $k + 1$

OLS.3 $E(\boldsymbol{\epsilon}^2\mathbf{X}'\mathbf{X}) = \sigma^2 E(\mathbf{X}'\mathbf{X})$. Homokedasticity, no need to assume normal errors.

# Estimation of MLR by OLS

The real value of the parameter:

$$\boldsymbol{\beta} = (E(\mathbf{X}'\mathbf{X}))^{-1}E(\mathbf{X}'\mathbf{y})$$

The OLS estimator:

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y})$$

OLS.1 and OLS.2 must be satisfied to obtain this estimator

# Properties about $\hat{\boldsymbol{\beta}}_{OLS}$

- Is it biased?
- What is its variance?
- Is it asymptotically normal?
- Is it efficient?

# Unbiased

- We say that $\hat{\boldsymbol{\beta}}$ is unbiased if $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.
- This condition is satisfied if $E(\boldsymbol{\epsilon}|\mathbf{X}) = 0$ and if $rank E(\mathbf{X}'\mathbf{X}) = k + 1$.
- We are assuming less $E(X'\epsilon) = 0$ so we cannot prove unbiaseness.
- The OLS estimator is not necessarily unbiased if OLS.1 and OLS.2 are satisfied. However, it is if we impose MLR.4 and OLS.2.
- Some estimators might be biased but **asymptotically unbiased**. This means that their bias tends to zero as $n \to \infty$.

Q: Find $E(\hat{\boldsymbol{\beta}})$

# Question:

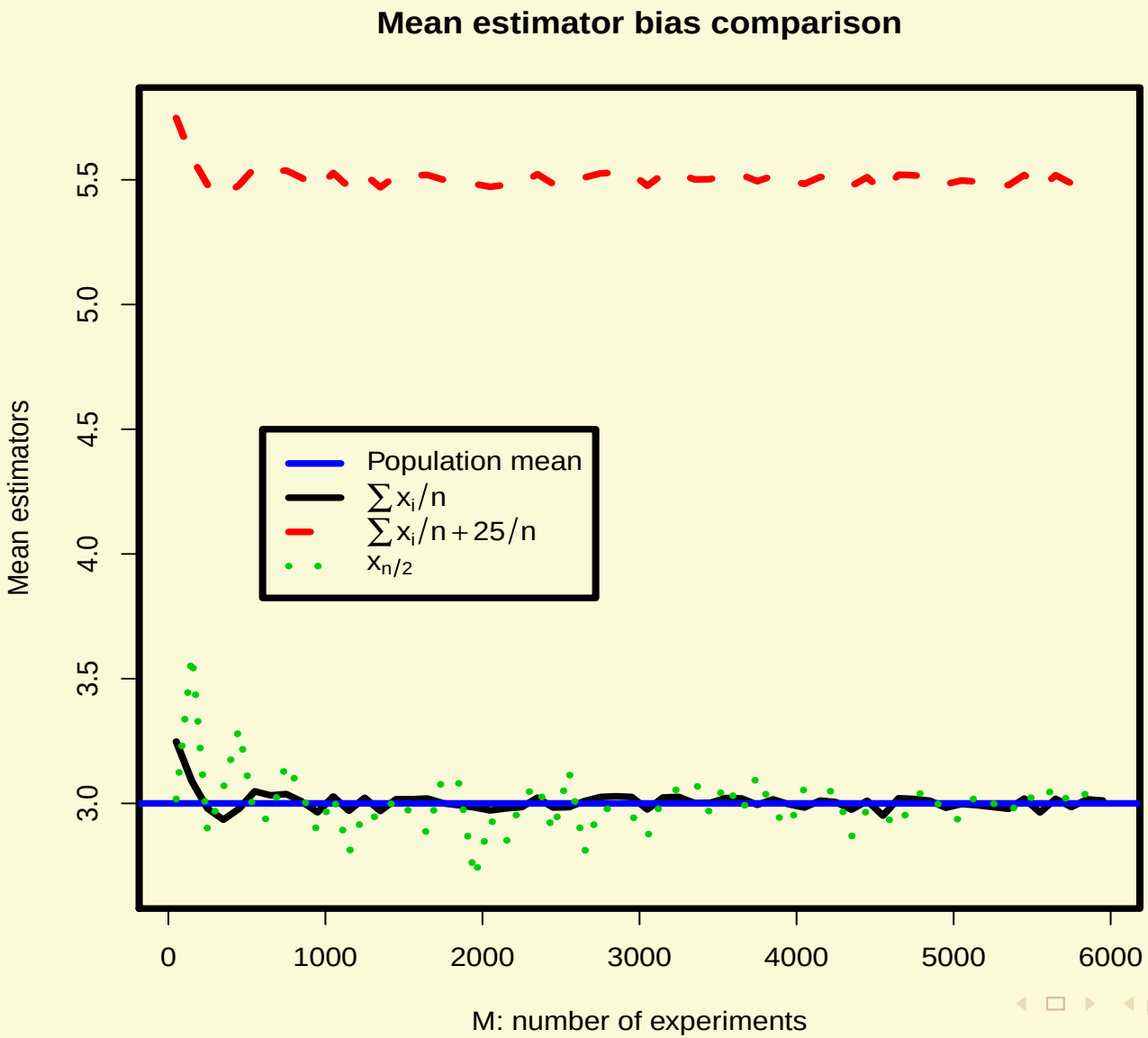
Let $\{x_1, x_2, \ldots, x_n\}$ a random sample of the variable $X \sim N(3, 25)$ and $n = 10$.

Which of the following three estimators of $\mu$ is unbiased:

1. $\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$
2. $\hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{25}{n}$
3. $\hat{\mu}_3 = x_{[n/2]}$: the middle point of the sample

And asymptotically unbiased?

# Solution:


Mean estimator bias comparison
Mean estimators
3.0
3.5
4.0
4.5
5.0
5.5
0
1000
2000
3000
4000
5000
6000
M: number of experiments
Population mean
$\sum x_i/n$
$\sum x_i/n + 25/n$
$x_{n/2}$

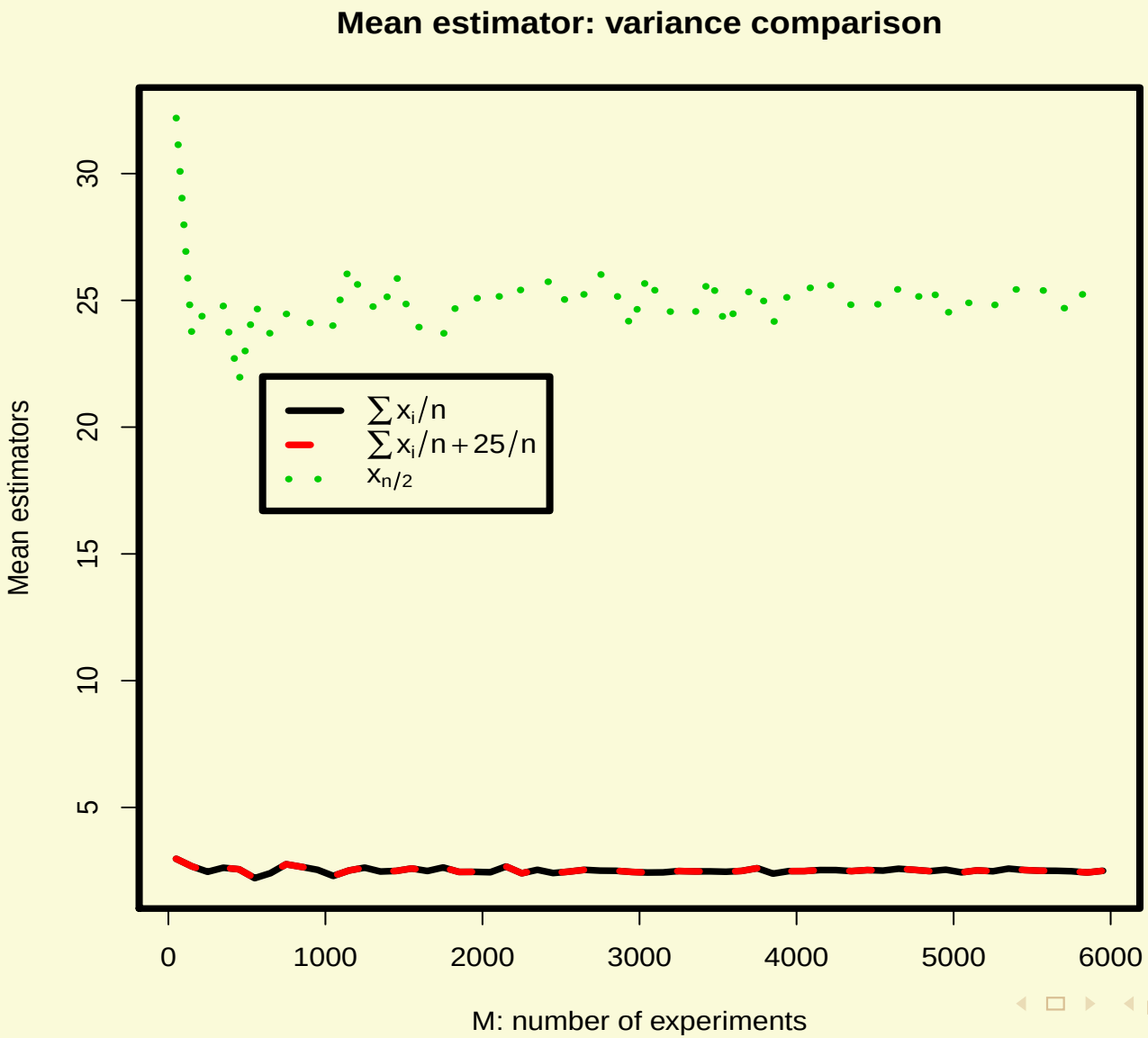
# Question on variance

Let $\{x_1, x_2, \ldots, x_n\}$ a random sample of the variable $X \sim N(3, 25)$ and $n = 10$.

Which of the following three estimators have greater variance:

1. $\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$
2. $\hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{25}{n}$
3. $\hat{\mu}_3 = x_{[n/2]}$: the middle point of the sample

# Solution:


Mean estimator: variance comparison
Mean estimators
M: number of experiments
0
1000
2000
3000
4000
5000
6000
5
10
15
20
25
30
$\sum x_i/n$
$\sum x_i/n + 25/n$
$x_{n/2}$

# Consistency

- The property of consistency assures that the estimator is, with a very high probability, very close to the real value of the parameter if the sample size if sufficiently large.
- An estimator $\hat{\theta}_n$ is consistent if and only if

$$\forall \epsilon > 0 \quad \lim_{n \to \infty} P\{|\hat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \boldsymbol{\theta}| \leq \epsilon\} = 1,$$

that is, the succession of $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ (which depends on $n$) converges in probability to the real parameter $\theta$. It is denoted by plim $_{n \to \infty} \hat{\boldsymbol{\theta}}_{\boldsymbol{n}} = \boldsymbol{\theta}$ or simply plim$\hat{\boldsymbol{\theta}}_{\boldsymbol{n}} = \boldsymbol{\theta}$.

# Consistency

- Intuitively, if an estimator is consistent, as the size of the sample grows, the probability that the estimation is close to the real value also grows.
- The OLS parameter $\hat{\boldsymbol{\beta}}$ is consistent, so the probability that is close to $\boldsymbol{\beta}$ is large when $n \to \infty$.
- This condition is satisfied if OLS.1 and OLS.2 are satisfied.

# Consistency

How do we proof consistency?

- Using the definition of consistency, this tends to be difficult.
- Try to show that the estimator is asymptotically unbiased and its variance tends to zero as the sample size goes to $\infty$. This consdition is sufficient but not absolutely necessary.

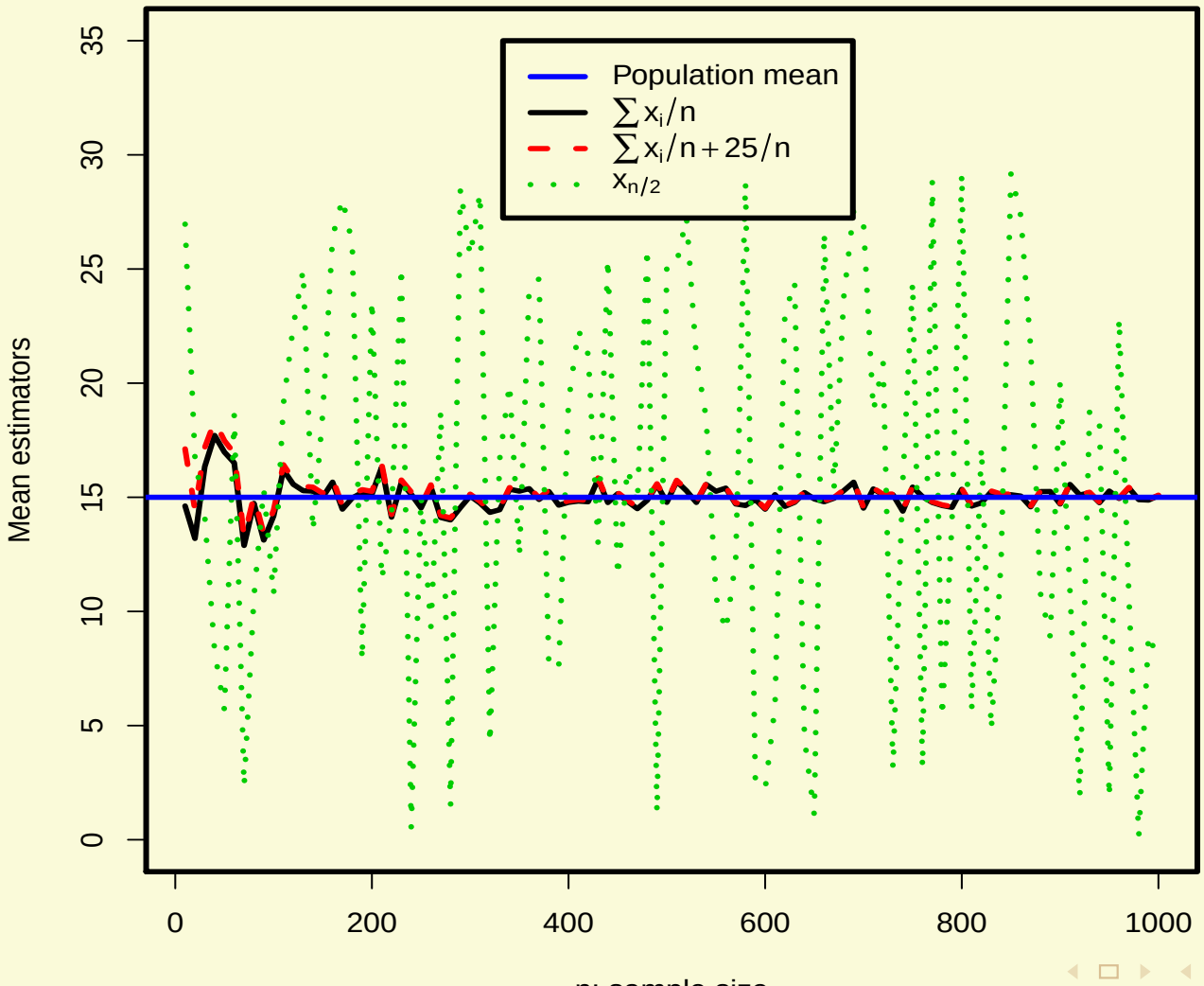
# Question:

Let $\{x_1, x_2, \ldots, x_n\}$ a random sample of the variable $X \sim U(0, 30)$ with mean $\mu = 15$ and variance $\sigma^2 = 75$.

Which of the following three estimators of $\mu$ are consistent:

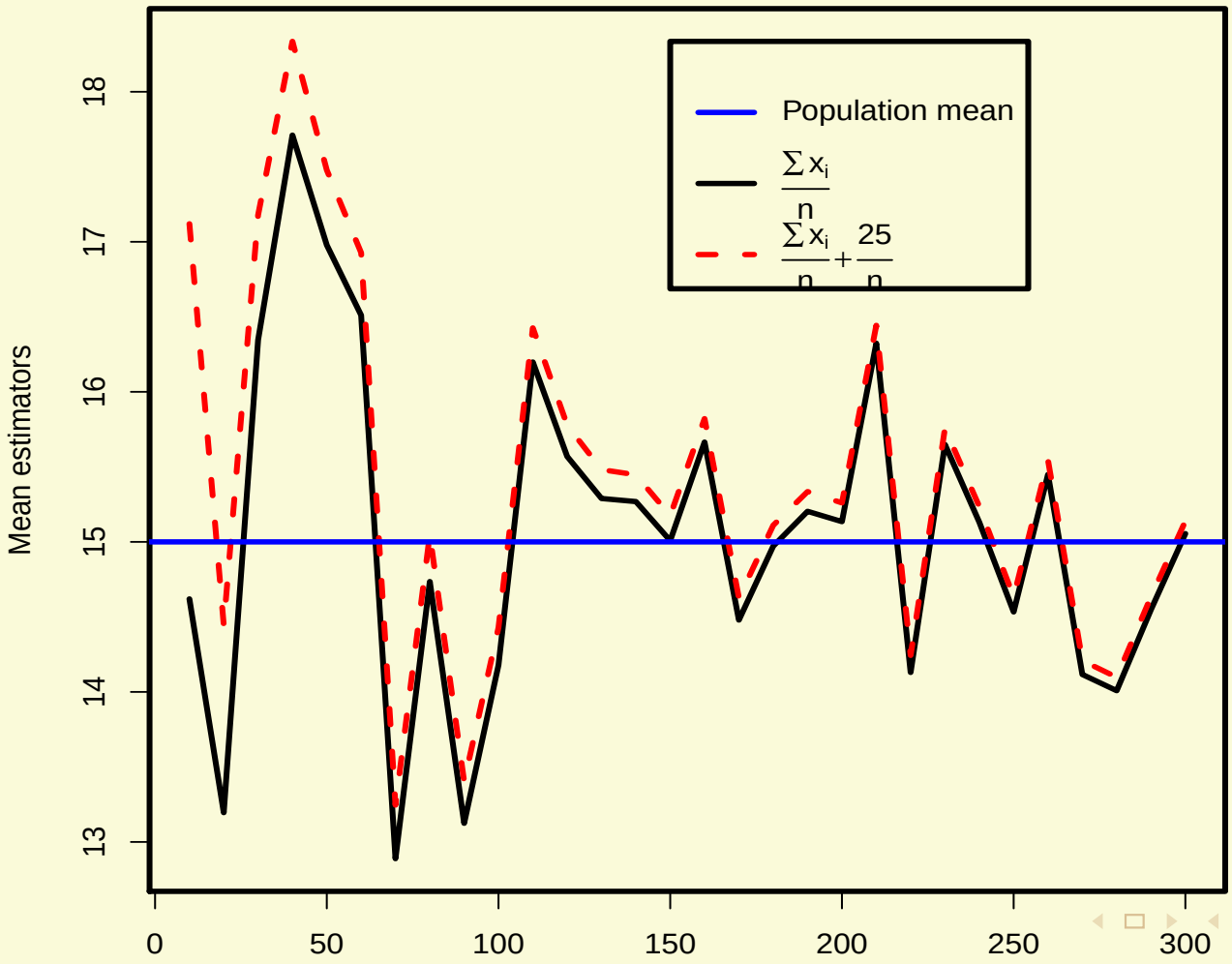
1. $\hat{\mu}_1 = \frac{1}{n} \sum_{i=1}^{n} x_i$
2. $\hat{\mu}_2 = \frac{1}{n} \sum_{i=1}^{n} x_i + \frac{25}{n}$
3. $\hat{\mu}_3 = x_{[n/2]}$: the middle point of the sample

# Solution:


Population mean
$\sum x_i/n$
$\sum x_i/n + 25/n$
$x_{n/2}$
Mean estimators
0
5
10
15
20
25
30
35
0
200
400
600
800
1000
n: sample size

# Solution:

## A closer look



# Exercise: $\hat{\boldsymbol{\beta}}$ in terms of $\boldsymbol{\beta}$

Mnemotechnic rule:
The OLS estimator $\hat{\boldsymbol{\beta}}^{\text{OLS}}$ is BLUE= Best Linear Unbiased Estimator of $\boldsymbol{\beta}$.

# OLS efficiency

Assuming MLR.5, the estimators $\hat{\beta}_i$ are efficient (in the sense that have the minimum variance amongst all the linear unbiased linear estimators – Gauss–Markov Theorem).

In fact their variance is:

$$Var(\hat{\boldsymbol{\beta}}) = \mathbf{V} = \frac{\sigma^2}{n} E(\mathbf{X}'\mathbf{X})^{-1} \tag{1}$$

$\sigma^2$ and $E(\mathbf{X}'\mathbf{X})$ are unknown, substituing by their estimators:

$$\hat{\mathbf{V}} = \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{n-k-1} = \frac{SSR}{n-k-1} = \frac{s^2}{n}\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} = s^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

# Variance estimation by OLS

- The standard error of $\hat{\beta}_i$ is given by:

$$se(\hat{\beta}_i) = \sqrt{diag(\hat{\mathbf{V}})[i+1]}$$

- Note the difference between the standard deviation and the standard error of the OLS estimator.

# Variance estimation ($\hat{\sigma}$) by OLS

Notation:

- $SST = \sum(y_i - \bar{\mathbf{y}})^2$, total sum of squares
- $SSE = \sum(\hat{y}_i - \bar{\mathbf{y}})^2$, the explained sum of squares
- $SSR = \sum(y_i - \hat{y}_i)^2 = \sum \hat{\epsilon}_i^2$, the residual sum of squares
- $SST = SSE + SSR$: the total variation of $y_i$ is equal to the total variation of the fitted values $\{\hat{y}_i\}$ and the total variation of the residuals $\{\hat{\epsilon}_i\}$.

Q: Is $\hat{\sigma}^2_{OLS}$ biased?

# Asymptotically normal

- Commonly it is difficult to derive analytically the exact distribution of an estimator.
- A solution to this problem is to look at the distribution probability in the infinity.
- If the distribution of our parameter gets close to one of the known distributions as the sample size increases, then we can use this known distribution as an approximation of the distribution of our estimator for large samples.
- If the error term is normally distributed, then the OLS estimator is normally distributed
- If the error term is not normally distributed, then the OLS estimator is *asymptotically* normally distributed
- Under conditions OLS.1–OLS.3, the OLS estimator is asymptotically normal:

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim^{a} N(0, \sigma^2 \Sigma^{-}1)$$

# Asymptotic properties

- Proof of Consistency
- Proof of Asymptotic Normality

# Asymptotically normal

Therefore the asymptotic distribution of the OLS estimator is

$$\hat{\boldsymbol{\beta}}^{\text{OLS}} \sim^{a} N\left(\boldsymbol{\beta}, \frac{\sigma^2}{n} E(\mathbf{X}'\mathbf{X})^{-1}\right)$$

- Now that we know the asymptotic distribution of our estimator, we can do inference and test economical hypothesis.
- The $(1-\alpha)\%$ asymptotic confident interval of $\beta_j$ is:

$$[\hat{\beta}_j \pm t_{(n-k-1);\alpha} se(\hat{\beta}_j)]$$

- The CI includes the real value of the parameter $(1-\alpha)\%$ of the time.
- So we can do inference if the sample is large enough

# T-test of a population parameter

We would like to study whether $\beta_j$ is *statistically* non-zero for a given $j = 0, \ldots, k$.

| Null hypothesis | Alternative hypothesis | Test statistic | Critical Region |
|---|---|---|---|
| $H_0 : \beta_j = 0$ | $H_1 : \beta_j > 0$<br>$H_1 : \beta_j < 0$<br>$H_1 : \beta_j \neq 0$ | $t = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$ | $t > t_{(n-k-1),\alpha}$<br>$t < t_{(n-k-1),\alpha}$<br>$\lvert t \rvert > t_{(n-k-1),\alpha}$ |

where $\alpha$ is the significance level

# Example: School performance and school size, baby Wooldridge

It is of great interest to know the effect of the school size in the performance of the students. Some say that a small number of students is better.
Download the file school_performance.txt from Blackboard. The data have been recorded in 408 high schools in Michigan during 1993. We can use these data to test the hypothesis
$H_0$: "The size of the high school does not affect the marks in the evaluation tests" versus
$H_1$ : "The size has a negative effect on performance".

# Example: School performance and school size, baby Wooldridge

- food_perc: percentage of students in the food program.
- enrolled: number of students enrolled.
- staff: staff size (teachers, admin...) for each 1000 students.
- expense: Student expenditure ($)
- salary: mean teacher salary ($).
- benefits: mean teacher benefits ($).
- quit: percentage of students who quit that high school
- graduation: percentage of graduates
- maths: percentage of students who pass maths.
- science: percentage of students who pass science.
- compensation: salary+ benefits of teachers
- bensal: benefits/salary

# Example: School performance and school size, baby Wooldridge

The performance is measured by the percentage of students that pass the maths exam in the 10th course. The size of the school is measured by the variable enrolled. Therefore the test of interest is $H_0 : \beta_{enrolled} = 0$ vs $H_1 : \beta_{enrolled} < 0$

We would like to include the compensation and staff variables to take into account the quality of the school. The linear model is

$$maths = \beta_0 + \beta_1\ compensation + \beta_2\ staff + \beta_3\ enrolled + \epsilon$$

# Example: School performance and school size, baby Wooldridge

Do the linear model in R:

- How is the fitted model?
- What are the standard errors of the model parameters?
- What is the determination coefficient?
- Is it a good model?

# Example: School performance and school size (UG)

```
> rm(list=ls())
> #Read data
> mydata<-read.table("school_performance.txt", h=T)
> model.lm<-lm(maths~1+compensation+staff+enrolled, data=mydata)
> summary(model.lm)

Call:
lm(formula = maths ~ 1 + compensation + staff + enrolled, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-22.235  -7.008  -0.807   6.097  40.689

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.2740214  6.1137938   0.372    0.710
compensation 0.0004586  0.0001004   4.570 6.49e-06 ***
staff        0.0479199  0.0398140   1.204    0.229
enrolled    -0.0001976  0.0002152  -0.918    0.359
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 10.24 on 404 degrees of freedom
Multiple R-squared:  0.05406,        Adjusted R-squared:  0.04704
F-statistic: 7.697 on 3 and 404 DF,  p-value: 5.179e-05
```

# Example: School performance and school size (UG)

- The fitted model is

$$\begin{aligned}\widehat{maths} =& 2.274 + 0.00046\ compensation \\ &+ 0.048\ staff - 0.0002\ enrolled\end{aligned}$$

- The standard errors of the estimators $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ are 6.114, 0.0001, 0.040 and 0.0002 respectively.
- The coefficient of determination is of 5%.
- The coefficient of the variable *enrolled* is negative which makes sense with the alternative hypothesis. As expected, the coefficients of *compensation* and *staff* are positive

# Example: School performance and school size (UG)

- To convince ourselves of the negative effect of the variable *enrolled*, we do a $t$-test on its parameter.
- We have $k = 3$ then the degrees of freedom of the test are 404.
- We take $\alpha = 0.05$, therefore the critical value $-t_{404,0.05} = -1.65$ (in R $qt(0.05, 404)$).
- The test statistics of *enrolled* is $-0.0002/0.000215 \approx -0.92$.
- This value is not in the critical region and therefore we cannot reject the null hypothesis.
- What value of $\alpha$ is needed to reject $H_0$?

# Example: School performance and school size (UG)

- The variable *compensation* is statistically significant even for $\alpha = 1\%$.
- Therefore, we reject the null hypothesis for the test $H_0 : \beta_{compensation} = 0$ versus $H_1 : \beta_{compensation} > 0$, at 1% level.
- What about the test for the variable *staff*?.

# Example: School performance and school size (UG)

Let see that a change in the structural model will affect the conclusions. Let us the logarithmic form for all the regressors:

$$\begin{aligned} maths =& \beta_0 + \beta_1 \log(compensation) + \beta_2 \log(staff) \\ &+ \beta_3 \log(enrolled) + \epsilon \end{aligned}$$

# Example: School performance and school size (UG)

```
> model.log<-lm(maths~1+log(compensation)+log(staff)+log(enrolled), data=mydata)
> summary(model.log)

Call:
lm(formula = maths ~ 1 + log(compensation) + log(staff) + log(enrolled),
    data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-22.735  -6.838  -0.835   6.139  39.718

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -207.6649    48.7031  -4.264 2.50e-05 ***
log(compensation)   21.1550     4.0555   5.216 2.92e-07 ***
log(staff)           3.9800     4.1897   0.950   0.3427
log(enrolled)       -1.2680     0.6932  -1.829   0.0681 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 10.18 on 404 degrees of freedom
Multiple R-squared:  0.06538,        Adjusted R-squared:  0.05844
F-statistic:  9.42 on 3 and 404 DF,  p-value: 4.974e-06
```

# Example: School performance and school size (UG)

The fitted model

$$\widehat{maths} = -207.66 + 21.16 \log(compensation) + 3.98 \log(staff) - 1.27 \log(enrolled)$$

We can reject the null hypothesis $H_0 : \beta_{log(enrolled)} = 0$ because its statistics is -1.829 which is under -1.65.

Q: Which model do we choose?

# Example: School performance and school size (UG)

The fitted model

$$\widehat{maths} = -207.66 + 21.16 \log(compensation) + 3.98 \log(staff) - 1.27 \log(enrolled)$$

We can reject the null hypothesis $H_0 : \beta_{log(enrolled)} = 0$ because its statistics is -1.829 which is under -1.65.

Q: Which model do we choose?

We check the $R^2$ and we see that this is bigger for the model with the log variables.

# Example: School performance and school size (UG)

Observations

- Large standard errors might be due to multicollinearity or a large correlation between some regressors.
- When the sample size is large, parameters are estimated precisely and the standard errors are commonly very small in relation to the value of the coefficients. Therefore the value of $\alpha$ tend to be small 1% or 5%.
- When the sample size is small, we might consider greater significance levels (greater $\alpha$).

# F-test

- The t-test is useful to do inference about individual model parameters.
- The F–test is used to test hypothesis over a set of model parameters.
- It tests the effect of a set of variables in the dependent variable.

# F-test

We have two models:

- The non–restricted model which contains all the regressors

$$\mathbf{y} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \mathbf{X}_1 + \ldots + \boldsymbol{\beta}_k \mathbf{X}_k + \boldsymbol{\epsilon}$$

- The restricted model which contains the regressors we consider necessary

$$\mathbf{y} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \mathbf{X}_1 + \ldots + \boldsymbol{\beta}_{k-q} \mathbf{X}_{k-q} + \boldsymbol{\epsilon}$$

We use an F-test to test the hypothesis of type

$$H_0 : \ \beta_{k-q+1}, \ldots, \beta_k = 0$$

$H_1$ : at least one of those parameters is non-zero

# F-test

The test statistics:

$$F = \frac{(SSR_{res} - SSR_{nores})/q}{SRC_{nores}/(n-k-1)} \sim F_{q,n-k-1}$$

- $k$ is the initial number of regressors
- $q = df_{res} - df_{nores}$ is the number of variables of the restricted model + 1.
- This statistics is always positive,
- Why? because $SSR_{res} \geq SSR_{nores}$ then the $CR = \{F \geq F_{q,n-k-1;\alpha}\}$
- Another way to calculate the test statistics is using the coefficient of determination:

$$F = \frac{(R^2_{nores} - R^2_{res})/q}{(1 - R^2_{nores})/(n-k-1)} \sim F_{q,n-k-1}$$

# F-test

We want to determine if $\beta_1, \ldots, \beta_k$ are statistically significant.
The test:
$H_0 : \beta_1 = \ldots = \beta_k = 0$ vs $H_1$ : at least one $\beta_j \neq 0$
The test statistics:

$$F = \frac{SSE/k}{SSR/(n-k-1)} = \frac{\text{variance explained by the model}}{\text{variance not explained by the model}} \sim F_{k,n-k-1}$$

Like $R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$ then

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)} \sim F_{k,n-k-1}$$

# F-test

The ANOVA table of a multiple regression is:

| Source | SS | df | variance | F |
|---|---|---|---|---|
| Regressors | $SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ | $k$ | $\hat{s}^2_{exp} = \frac{SSE}{k}$ | $F = \frac{s^2_{exp}}{s^2_{resid}}$ |
| Residuals | $SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \hat{\epsilon}_i^2$ | $n - k - 1$ | $\hat{s}^2_{resid} = \frac{SSR}{n-k-1}$ | |
| Total | $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$ | n-1 | $\hat{s}^2_y = \frac{SSR}{n-1}$ | |

# Example (UG)

Let us consider a model to explain the salary of baseball players in 1993 (UGB). File: baseball.txt (tab delimited, . decimal)

- salary: 1993 season salary
- years: years in major leagues
- games: career games played
- rbis: career runs batted in
- hruns: career home runs
- runs: career runs scored

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 games + \beta_3 rbis + \beta_4 hruns + \beta_5 runs + \epsilon$$

# Example (UG)

```
salary.lm <-lm(log(salary)~1 + years+ games+ rbis+hruns+runs)
> summary(salary.lm)
Call:
lm(formula = log(salary) ~ 1 + years + games + rbis + hruns +
    runs)
Residuals:
     Min       1Q   Median       3Q      Max
-3.02508 -0.45034 -0.04013  0.47014  2.68924

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.119e+01  2.888e-01  38.752  < 2e-16 ***
years       6.886e-02  1.211e-02   5.684 2.79e-08 ***
games       1.255e-02  2.647e-03   4.742 3.09e-06 ***
rbis        9.786e-04  1.104e-03   0.887    0.376
hruns       1.443e-02  1.606e-02   0.899    0.369
runs        1.077e-02  7.175e-03   1.500    0.134
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7266 on 347 degrees of freedom
Multiple R-Squared: 0.6278,     Adjusted R-squared: 0.6224
F-statistic: 117.1 on 5 and 347 DF,  p-value: < 2.2e-16
```

# Example (UG)

The fitted unrestricted model

$$\widehat{\log(salary)} = 11.19 + 0.0689 years + 0.0126 games + 0.00098 rbis + 0.014 hruns + 0.0108 runs$$

- $R^2 \approx 63\%$.
- Both *years* and *games* are significant
- Variables *rbis*, *hruns* and *runs* are not significant
- $SSR_{nores} = sum(residuals(salary.lm)^2) =$ 183.186.
- $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ vs $H_1 :$ at least one $\beta_j \neq 0$ for $j = 3, 4, 5$.

# Example (UG)

Restricted model (under $H_0$)

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 games + \epsilon$$

Fitted model:

$$\widehat{\log(salary)} = 11.22 + 0.0713 years + 0.0202 games$$

- $R^2$= 60%
- $SSR_{res} = 198.311.$

# Example (UG)

```
> summary(lm(log(salary)~1 + years+ games))

Call:
lm(formula = log(salary) ~ 1 + years + games)

Residuals:
     Min       1Q   Median       3Q      Max
-2.66858 -0.46412 -0.01177  0.49219  2.68829

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.223804   0.108312 103.625  < 2e-16 ***
years        0.071318   0.012505   5.703 2.50e-08 ***
games       0.020174   0.001343  15.023  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Example (UG)

In our example,

- k = 5, q= 3, n-k-1= 347,
- $F = 9.55$, $RR = \{2.60, \infty\}$ at $\alpha = 0.05$.
- There is enough evidence to reject the null hypothesis with signification level 5%.
- Therefore, we include all the variables in the model.

# Possible violations of the OLS assumptions

1. OLS.1 $E(X'\epsilon) \neq 0$ fails. The errors are correlated with the regressors, that is, $X$ is endogenous
   - Consequences: The estimates $\hat{\beta}$ will be biased and inconsistent
   - Solution: Proxy variables, instrumental variables (IV) and Panels
   - Start with IV next week

# Possible violations of the OLS assumptions

2. OLS.2 Rank $E(X'X) < k+1$ fails: the number of normal equations is less than the number of parameters $\beta$
   - Consequence: the parameters cannot be identified, we cannot find an unique solution
   - Solution: identify the dependent variables in $X$ and remove them from the model

# Possible violations of the OLS assumptions

3. OLS. 3 $E(\epsilon^2 X'X) = \sigma^2 h(X) \neq \sigma^2$ fails. Heteroskedasticity: the variance of the errors depends on $X$.
   - The value of the asymptotic variance of $\hat{\beta}$ is changed.
   - Luckily, the consistency and the unbiasedness of the estimator are preserved, as well as the asymptotic normality.
   - Solution: WLS, or use robust standard errors for the tests.

# Test of heteroskedasticity: Graphics

- Plot the residuals. We can get an intuitive idea if there is a continuous increase/decrease of the errors
- Plot the abosolute value of the residuals vs the explanatory variable we suspect is responsible of the heteroskedasticity of the model. Is there a pattern in the variabilidad?

# Test of heteroskedasticity: Breush–Pagan

If $E(\boldsymbol{\epsilon}|\mathbf{X}) = 0$ then

- The OLS estimator of a MLR is unbiased and consistent
- $Var(\boldsymbol{\epsilon}|\mathbf{X}) = \sigma^2 = E(\boldsymbol{\epsilon}^2|\mathbf{X}) = E(\boldsymbol{\epsilon}^2)$.

The null hypothesis of the BP test is:

$H_0 : Var(\boldsymbol{\epsilon}|\mathbf{X}) = \sigma^2$ which is equivalent to $H_0 : E(\boldsymbol{\epsilon}^2|\mathbf{X}) = E(\boldsymbol{\epsilon}^2)$

.

# Test of heteroskedasticity: Breush–Pagan

- We assume that $\boldsymbol{\epsilon}^2$ is not correlated with any explanatory variable
- The null hypothesis is false if $E(\boldsymbol{\epsilon}^2) = h(\mathbf{X}_i)$ for some $i = 1, \ldots, k$ and any function $h(\cdot)$ cualquier function.
- We could assume that this function is linear

$$\boldsymbol{\epsilon}^2 = \delta_0 + \delta_1 \mathbf{X}_1 + \ldots + \delta_k \mathbf{X}_k + \mathbf{v}$$

where $\mathbf{v}$ is an error such that $E(\mathbf{v}|\mathbf{X}) = 0$

# Test of heteroskedasticity: Breush–Pagan

Then, the null hypothesis of the BP test can be written as:

$$H_0 : \delta_1 = \ldots = \delta_k = 0$$

- Because $\boldsymbol{\epsilon}$ is unknown, then we use the residuals.

$$\hat{\boldsymbol{\epsilon}}^2 = \delta_0 + \delta_1 \mathbf{X}_1 + \ldots + \delta_k \mathbf{X}_k + error \tag{2}$$

- We find the $F$ statistics using the determinations coefficient for (2)

$$F = \frac{R^2_{\hat{\epsilon}^2}/k}{(1 - R^2_{\hat{\epsilon}^2})/(n-k-1)} \sim F_{k,n-k-1}$$

where $k$ is the number of independent variables.

# Test of heteroskedasticity: LM

Another useful statistic is the Lagrange multiplier.

$$LM = nR^2_{\hat{\epsilon}^2} \sim \chi^2_k \quad \text{Lagrange multiplier}$$

1. Estimate the MLR model by OLS and obtain $\hat{\epsilon}^2$.
2. Regress the residuals vs. the explanatory variables and get $R^2_{\hat{\epsilon}^2}$.
3. Calculate the F or LM statistics and find the correspondent p-value

# Example: housing price (Wooldridge)

What variables does the average of housing price depend upon? (houseprice.txt – semicolon delimited, . decimal)

- price: selling price
- rooms: # rooms in house
- area: square footage of house
- land: square footage lot

# Example: housing price (Wooldridge)

```
> mydata<-read.table("houseprice.txt", h=T, sep=";", dec=",")
> house.lm<- lm(price~land+area+rooms, data=mydata)
> summary(house.lm)

Call:
lm(formula = price ~ land + area + rooms, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-120.026  -38.530   -6.555   32.323  209.376

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.177e+01  2.948e+01  -0.739  0.46221
land         2.068e-03  6.421e-04   3.220  0.00182 **
area         1.228e-01  1.324e-02   9.275 1.66e-14 ***
rooms        1.385e+01  9.010e+00   1.537  0.12795
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 59.83 on 84 degrees of freedom
Multiple R-squared:  0.6724,        Adjusted R-squared:  0.6607
F-statistic: 57.46 on 3 and 84 DF,  p-value: < 2.2e-16
```

# Example: housing price (Wooldridge)

The fitted model:

$$\widehat{price} = -21.7703(29.48) + 0.0021(0.00064)\ land + 0.12(0.013)\ area + 0.0021(9.01)\ rooms$$

with $n = 88$ y $R^2 = 67\%$.

Take a look at the model diagnostic plots with $plot(house.lm)$

# Example: housing price (Wooldridge)

```
> residuals=residuals(house.lm)
> epsilon.sq<-residuals^2
> house2.lm<-lm(epsilon.sq~land+area+rooms, data=mydata)
> summary(house2.lm)

Call:
lm(formula = epsilon.sq ~ land + area + rooms, data = mydata)

Residuals:
   Min     1Q Median     3Q    Max
 -9044  -2212  -1256    -97  42582

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.523e+03  3.259e+03  -1.694  0.09390 .
land         2.015e-01  7.101e-02   2.838  0.00569 **
area         1.691e+00  1.464e+00   1.155  0.25128
rooms        1.042e+03  9.964e+02   1.046  0.29877
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 6617 on 84 degrees of freedom
Multiple R-squared:  0.1601,        Adjusted R-squared:  0.1301
F-statistic: 5.339 on 3 and 84 DF,  p-value: 0.002048
```

# Example: housing price (Wooldridge)

```
> k<-house2.lm$rank
> n<-length(mydata$price)
> R_2<-summary(house2.lm)$r.squared
> LM= n*R_2
> p.value<-pchisq(LM, df=k-1, lower.tail=F)
> p.value
```

```
[1] 0.00278206
```

We reject the null hypothesis and conclude that there is heteroskedasticity in the model (the variance of the estimator is not independent of the individuals).

So the s.e are not trustable.

# Example: housing price (Wooldridge)

```
> library(lmtest)
> bptest(price~land+area+rooms, data=mydata)

        studentized Breusch-Pagan test

data:  price ~ land + area + rooms
BP = 14.0924, df = 3, p-value = 0.002782
```

Q: Run the test for the log–log model.

# Test of heteroskedasticity: White

- The asumption $Var(\boldsymbol{\epsilon}|\mathbf{X}) = \sigma^2$ can be replace by the weaker assumption that $\boldsymbol{\epsilon}^2$ are uncorrelated with the rest of the independent variables ($\mathbf{X}_i$), their squared values ($\mathbf{X}_i\mathbf{X}_i'$) and all their cross products ($\mathbf{X}_i\mathbf{X}_h', \ i \neq h$).
- White (1980) uses this observation to create the following heteroskedasticity test
- Let assume we have $k = 3$ independent variables, the White test is based on the estimation of:

$$\begin{aligned}\hat{\boldsymbol{\epsilon}}^2 = \quad & \delta_0 + \delta_1\mathbf{X}_1 + \delta_2\mathbf{X}_2 + \delta_3\mathbf{X}_3 + \delta_4\mathbf{X}_1^2 + \delta_5\mathbf{X}_2^2 + \delta_6\mathbf{X}_3^2 \\ & + \delta_7\mathbf{X}_1\mathbf{X}_2 + \delta_8\mathbf{X}_1\mathbf{X}_3 + \delta_9\mathbf{X}_2\mathbf{X}_3 + error\end{aligned} \tag{3}$$

# Test of heteroskedasticity: White

- The White test regression has 6 regressors more than the BP regression for $k = 3$
- $H_0 : \delta_1 = \ldots = \delta_9 = 0$
- In this case 9 restrictions are tested
- The lagrange multiplier test statistics $LM = nR^2 \sim \chi^2_9$
- We could also use the $F$ test statistics

# Test of heteroskedasticity: White

Steps of the White test

1. Estimate the MLR model by OLS and obtain $\hat{\epsilon}^2$.
2. Regress the residuals vs. the explanatory variables, their squares and cross-products and get $R^2_{\hat{\epsilon}^2}$.
3. Calculate the F or LM statistics and find the correspondent p-value

# Test of heteroskedasticity: White

- The main problem of this test is the number of parameters to test
- For $k = 3$ we have 9 parameters
- For $k = 6$ we have 27 parameters
- This test needs a large data set if the model includes a moderate number of independent variables

# Test of heteroskedasticity: White

A simpler test runs the regression:

$$\hat{\boldsymbol{\epsilon}}^2 = \delta_0 + \delta_1 \hat{\mathbf{y}} + \delta_2 \hat{\mathbf{y}}^2 + error \tag{4}$$

- The test: $H_0 : \delta_1 = \delta_2 = 0$.
- Note that we always have only two variables for (4).
- This is a special case of the White test. It is useful when we have reasons to believe that the variance in the errors changes with $E(\mathbf{y}|\mathbf{X})$.

# Example: housing price (Wooldridge)

White test:

```
> x1<-mydata$land
> x2<-mydata$area
> x3<-mydata$rooms
> x1.2<-x1^2
> x2.2<-x2^2
> x3.2<-x3^2
> x1.x2<-x1*x2
> x1.x3<-x1*x3
> x2.x3<-x2*x3
> house3.lm<-lm(epsilon.sq~1+x1+x2+x3+x1.2+x2.2+x3.2+x1.x2+x1.x3+x2.x3)
> k<-house3.lm$rank
> R_2<-summary(house3.lm)$r.squared
> LM= n*R_2
> p.value<-pchisq(LM, df=k-1,lower.tail=F)
> p.value

[1] 9.952941e-05
```

# Example: housing price (Wooldridge)

## Special case of White test

```
> y.hat<-fitted(house.lm)
> y.hat.2<-y.hat^2
> house4.lm<-lm(epsilon.sq~1+ y.hat+y.hat.2)
> k<-house4.lm$rank
> R_2<-summary(house4.lm)$r.squared
> LM= n*R_2
> p.value<-pchisq(LM, df=k-1, lower.tail=F)
> p.value

[1] 0.0002933311
```

# Solving heteroskedasticity: solution I, robust se

- We will see robust se next week.
- Below WLS to solve heteroskedasticity. This does not enter in the exam.

# Solving heteroskedasticity: solution II, WLS

The Weighted Least Squares (WLS)

- We know that $Var(\boldsymbol{\epsilon}|\mathbf{X}) = E(\boldsymbol{\epsilon}^2|\mathbf{X}) = \sigma^2 h(\mathbf{X})$, and

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- What do you suggest?

# Solving heteroskedasticity: solution II, WLS

The Weighted Least Squares (WLS)

- We know that $Var(\boldsymbol{\epsilon}|\mathbf{X}) = E(\boldsymbol{\epsilon}^2|\mathbf{X}) = \sigma^2 h(\mathbf{X})$, and

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- What do you suggest?
- We assume that $h(\mathbf{X})$ is known and $\sigma^2$ is unknown but estimable
- For a random individual $j$,
  $\sigma_j^2 = Var(\epsilon_j|X_{1j}, \ldots, X_{kj}) = \sigma^2 h(X_{1j}, \ldots, X_{kj}) = \sigma^2 h_j$
- All variables $y_j, x_{ij}$ of individial $j$ are divided by $\sqrt{h_j}$

# Solving heteroskedasticity: solution II, WLS

The Weighted Least Squares (WLS)

- We know that $Var(\boldsymbol{\epsilon}|\mathbf{X}) = E(\boldsymbol{\epsilon}^2|\mathbf{X}) = \sigma^2 h(\mathbf{X})$, and

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- What do you suggest?
- We assume that $h(\mathbf{X})$ is known and $\sigma^2$ is unknown but estimable
- For a random individual $j$,
  $\sigma_j^2 = Var(\epsilon_j|X_{1j}, \ldots, X_{kj}) = \sigma^2 h(X_{1j}, \ldots, X_{kj}) = \sigma^2 h_j$
- All variables $y_j, x_{ij}$ of individial $j$ are divided by $\sqrt{h_j}$
- So, $Var(\epsilon_j/\sqrt{h_j}|X_{ij}) = \sigma^2$ and we can apply OLS to these transformed variables

# Solving heteroskedasticity: solution II, WLS

The new model is

$$\mathbf{y}^* = \mathbf{X}^* \boldsymbol{\beta} + \boldsymbol{\epsilon}^*$$

Note that $X_{0j} = 1/h_0$.

- This equation is linear in the parameters $\beta_i$
- The errors $\boldsymbol{\epsilon}^*$ have mean zero and conditional variance $\sigma^2$
- If $\boldsymbol{\epsilon}$ follows a normal distribution then $\boldsymbol{\epsilon}^*$ also does.

Therefore the transformed equation satisfies the assumptions of the MLR and can be estimated consistently and efficiently by OLS.

# Solving heteroskedasticity: solution II, WLS

In addition:

$$\hat{\sigma}^2 = \frac{1}{n-k} \sum (\hat{\epsilon}_j^*)^2$$

is unbiased.

- The interpretation of the estimations are done related to the original equation
- The $(R^*)^2$ of the transformed model can be used for the $F$ test but it is not meaningful as a measure of model fitness.

# Solving heteroskedasticity: solution I.2

We have assumed that $h(\cdot)$ is known. What if it is not?

# Solving heteroskedasticity: solution I.2

We have assumed that $h(\cdot)$ is known. What if it is not?

- Often we can specified certain form of $h_j$, so we use $\hat{h}_j$ in our transformation
- We have the Feasible Weighted Least Squares estimator.
- For example, let us assume that

$$Var(\boldsymbol{\epsilon}|\mathbf{X}) = \sigma^2 \exp(\delta_0 + \delta_1 \mathbf{X}_1 + \ldots + \delta_k \mathbf{X}_k),$$

- $h(X) = \exp(\delta_0 + \delta_1 X_1 + \ldots + \delta_k X_k)$.
- We use an exponential function to ensure that our variance is positive
- Of course, the parameters $\delta_i$ must be estimated first. How?

# Solving heteroskedasticity: solution I.2

We assume that

$$\boldsymbol{\epsilon}^2 = \sigma^2 \exp(\delta_0 + \delta_1 \mathbf{X}_1 + \ldots + \delta_k \mathbf{X}_k) + \mathbf{v}$$

where $E(\mathbf{v}|\mathbf{X}) = 1$.
If $\mathbf{v}$ is independent of $\mathbf{X}_i$, then we can write:

$$\log(\boldsymbol{\epsilon}^2) = \alpha_0 + \delta_1 \mathbf{X}_1 + \ldots + \delta_k \mathbf{X}_k + \mathbf{e}$$

with $\mathbf{e}$ an error of mean zero and independent of $\mathbf{X}_j$.

Because $\boldsymbol{\epsilon}$ is unknown but we can use $\hat{\boldsymbol{\epsilon}}$ to obtain $\hat{g}_j = \widehat{\log(\hat{\epsilon}_j^2)}$ and $\hat{h}_j = exp(\hat{g}_j)$.
Now we transform the original equation dividing by $1/\hat{h}_j$.

# Solving heteroskedasticity: solution I.2

Steps:

1. Regress $\mathbf{y}$ over $\mathbf{X}_1, \ldots, \mathbf{X}_k$ and obtain the residuals $\hat{\boldsymbol{\epsilon}}$.
2. Create the variable $\log(\hat{\boldsymbol{\epsilon}}^2)$ and regress it over $\mathbf{X}_1, \ldots, \mathbf{X}_k$ to obtain the fitted values $\hat{\mathbf{g}}$.
3. Calculate $\hat{\mathbf{h}} = \exp(\hat{\mathbf{g}})$
4. Estimar the equation $\mathbf{y}^* = \mathbf{X}^* \boldsymbol{\beta}$ where $\mathbf{X}_j^* = \mathbf{X}_j / \sqrt{\hat{h}_j}$

The feasible weighted least squares estimator of $\sigma^2$ is biased. Although it is consistent and asymptotically efficient.

# Solving heteroskedasticity: solution I.2

Another option is using the regression

$$\log(\hat{\boldsymbol{\epsilon}}^2) = \delta_0 + \delta_1 \hat{\mathbf{y}} + \delta_2 \hat{\mathbf{y}}^2 + error$$

to find $\hat{h}(X_j)$.
Step 3 is then no necessary.