Tackling important statistical assumptions.

1) The issue of normality (last lecture):

2) The issue of homogeneity of variances (today):

- Standard (e.g., ANOVAs, regressions) assume homoscedasticity.
- Robust approaches (Welch's ANOVA, Weighted least squares) are good to deal with heteroscedasticity.

1

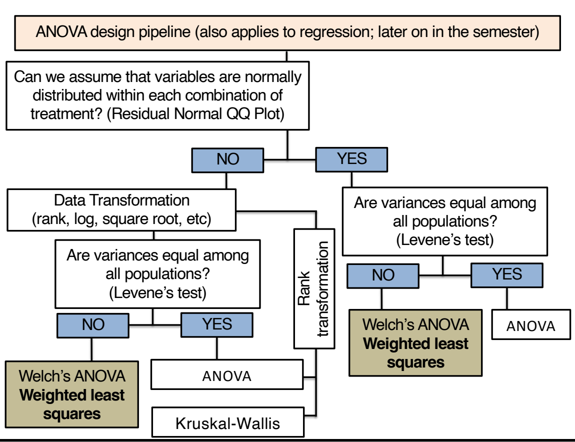
REMINDER: Classic non-parametric tests (ranked data, permutation tests) are often considered those tests that can handle non-normal data.

There is a common misunderstanding (however) in the statistical literature, including many biostatistics books, that non-parametric tests can also handle differences in variances among samples (because the term "non-parametric", it is often assumed that they are completely assumption free.

THIS IS NOT TRUE! They are also affected by variance differences among groups (e.g., the Kruskal-Wallis, ANOVAs on ranks).

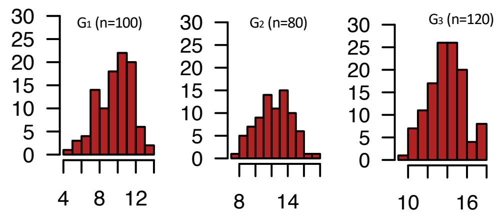
Example: test variance differences in ranks (rarely done in the literature but necessary)!

2



3

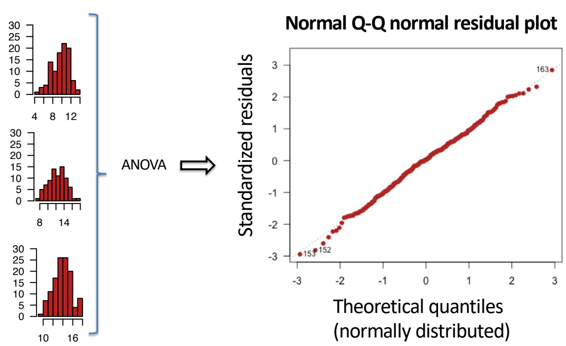
ANOVA design pipeline – let's use some normally distributed ***homoscedastic*** simulated data to understand **the Weighted Least Squares approach (WLS)**



One-factorial design - 3 groups, normally distributed homoscedastic data ($\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 4$), varying in means ($\mu_1^2 = 10, \mu_2^2 = 12, \mu_3^2 = 14$)

4

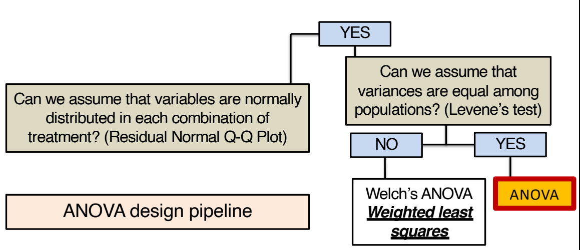
[1] – Can we assume that variables are normally distributed within each combination of treatment? (Residual Normal Q-Q Plot)



$Y = Factor(G1, G2) + residuals$

5

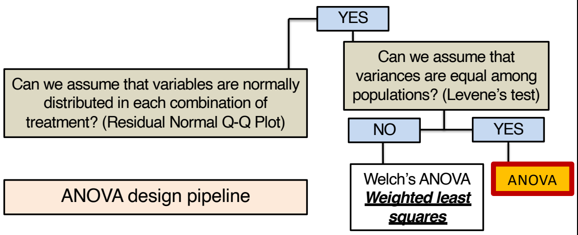
[2] – Can we assume that variances are equal among populations? (Levene's test)



6

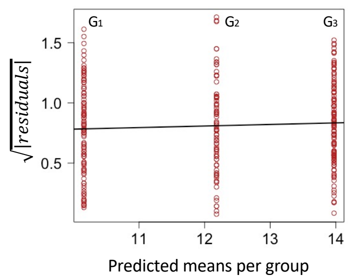
[2] – Can we assume that variances are equal among populations? (Levene's test); well, we simulated data, so no big surprises

```
> leveneTest(values ~ as.factor(groups))
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   2   0.223 0.8003
      297
```



7

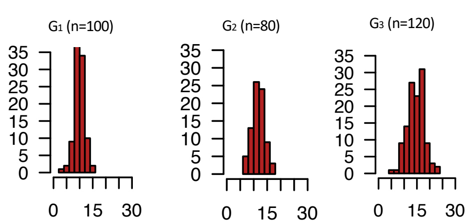
Can we assume that variances are equal among all populations? (alternative to the Levene's test and they way to understand WLS)



The plot between the square root of the absolute ANOVA residuals (i.e., deviations from the predicted mean group) against predicted mean group (you will see this one in the tutorial) should look like a straight line (constant variance). The Breusch-Pagan test can be employed to determine whether a deviation from a straight line is significant (we will use this test to assess homoscedasticiy in regressions).

8

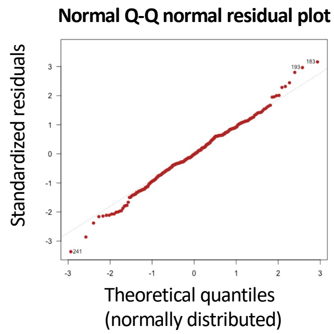
ANOVA design pipeline – let's use some normally distributed ***heteroscedastic*** simulated data to understand Weighted Least Squares



One-factorial design - 3 groups, normally distributed heteroscedastic data ($\sigma_1^2 = 4, \sigma_2^2 = 6.25, \sigma_3^2 = 9$), varying in means ($\mu_1^2 = 10, \mu_2^2 = 12, \mu_3^2 = 14$)

9

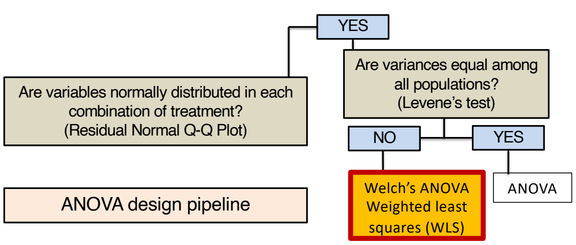
[1] – Can we assume that variables are normally distributed within each combination of treatment? (Residual Normal Q-Q Plot)



10

Can we assume that variances are equal among populations? (Levene's test)

```
> leveneTest(values ~ as.factor(groups))
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)
group   2  12.295 7.414e-06 ***
      297
```



11

Can we assume that variances are equal among populations? (Levene's test)

We can use the square of the residuals to assess that;
Note that the average of residuals is always zero.

First, we estimate the residuals of the ANOVA:

$$Y = Factor(G1, G2) + residuals$$

Then, for each group, square their respective residuals

$residuals$	$\sqrt{\lvert residuals \rvert}$		
-0.9723056	0.9860556	Group 1	$var(residuals^2)$=0.005018537
-0.8426648	0.9179678		
-0.7130241	0.8444075		
0.1944611	0.4409774	Group 2	$var(residuals^2)$=0.142741
0.9723056	0.9860556		
1.3612278	1.1667167		

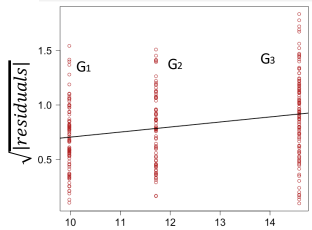
$\sqrt{\lvert residuals \rvert}$=square root of absolute values

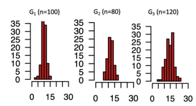
Using here a "tiny" small number of observations for demonstration purposes

12

Can we assume that variances are equal among populations?
(alternative to the Levene's test and they way to understand WLS)

$(\sigma_1^2 = 4, \sigma_2^2 = 6.25, \sigma_3^2 = 9)\ (\mu_1^2 = 10, \mu_2^2 = 12, \mu_3^2 = 14)$





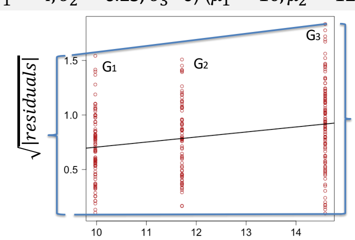
Predicted means per group (order of groups from small to large variance)

The plot between the square root of the absolute ANOVA residuals (i.e., deviations from the predicted mean group) against predicted mean group (you will see this one in the tutorial) should look like a straight line (constant variance). It doesn't here, so clearly indicating heteroscedasticity in the data.

13

Can we assume that variances are equal among populations?
(alternative to the Levene's test and they way to understand WLS)

$(\sigma_1^2 = 4, \sigma_2^2 = 6.25, \sigma_3^2 = 9)\ (\mu_1^2 = 10, \mu_2^2 = 12, \mu_3^2 = 14)$



Predicted means per group

The plot between the square root of the absolute ANOVA residuals (i.e., deviations from the predicted mean group) against predicted mean group (you will see this one in the tutorial) should look like a straight line (constant variance). It doesn't here, so clearly indicating heteroscedasticity in the data.

14

ANOVA is a regression model!

They differ in "design" but not in calculations!

15

The weighted least square (WLS) approach for dealing with heteroscedasticity

Welch's ANOVA covered in Intro Stats and can only deal with single factorial ANOVA designs

Today:

1) How does heteroscedasticity affect residual variation in ANOVAs?

And

2) How can we use the weighted least squares (WLS) approach to deal with heteroscedasticity in ANOVAs (original data or ranked-based ANOVA)

16

The weighted least square (WLS) approach for dealing with heteroscedasticity

Welch's ANOVA covered in Intro Stats and can only deal with single factorial ANOVA designs

Today:

1) How does heteroscedasticity affect residual variation in ANOVAs?

And

2) How can we use the weighted least squares (WLS) approach to deal with heteroscedasticity in ANOVAs (original data or ranked-based ANOVA)

But first we need to understand that:

ANOVA is a regression model

17

ANOVA is a regression model where the response variable is continuous and the predictors are categorical; the categorical predictors are coded in such a way that an ANOVA becomes a regression problem

Let's use a tiny fictional example with 2 groups (control, Group_1)

Response	Factor (predictor)
1.2	control
2.7	control
3.1	control
4.1	Group_1
5.3	Group_1
6.1	Group_1

18

ANOVA is a regression model where the response variable is continuous and the predictors are categorical.

Response	Factor (predictor)	Contrast
1.2	control	0
2.7	control	0
3.1	control	0
4.1	Group_1	1
5.3	Group_1	1
6.1	Group_1	1

Contrasts are numerical values that can be used directly into a regression model so that ANOVA becomes estimating a regression model; The ANOVA of the regression model has then exactly the same results as the standard ANOVA.

19

ANOVA is a regression model where the response variable is continuous and the predictors are categorical.

A tiny example:

```
groups <- c("control","control","control","Group_1","Group_1","Group_1")
values <- c(1.2,2.7,3.1,4.1,5.3,6.1)
```

Running ANOVA using the R function *aov*:

```
> summary(aov(values~groups))
            Df Sum Sq Mean Sq F value Pr(>F)
groups       1 12.042  12.042   11.94 0.0259 *
Residuals    4  4.033   1.008
---
```

20

ANOVA is a regression model where the response variable is continuous and the predictors are categorical.

A tiny example:

```
groups <- c("control","control","control","Group_1","Group_1","Group_1")
values <- c(1.2,2.7,3.1,4.1,5.3,6.1)
```

Running ANOVA using the R function *aov*:

```
> summary(aov(values~groups))
            Df Sum Sq Mean Sq F value Pr(>F)
groups       1 12.042  12.042   11.94 0.0259 *
Residuals    4  4.033   1.008
---
```

Running ANOVA using the R function *lm* (linear model = regression) setting group as a ***factor***:

```
> anova(lm(values~factor(groups)))
Analysis of Variance Table

Response: values
               Df  Sum Sq Mean Sq F value  Pr(>F)
factor(groups)  1 12.0417 12.0417  11.942 0.02592 *
Residuals       4  4.0333  1.0083
```

21

Let's (quickly) revisit a simple regression model (as seen in Intro Stats). More on regressions later in our Multiple Regression module

$Y = \underbrace{\beta_0 + \beta_1 X} + e$ - e represents the vector of residual values.

$\beta = (X^T X)^{-1} X^T Y$ - Slope and intercept estimated by one single operation via Ordinary Least Squares (OLS).

22

Simple regression model

$Y = \underbrace{\beta_0 + \beta_1 X} + e$ - e represents the vector of residual values.

$\beta = (X^T X)^{-1} X^T Y$ - Slope and intercept estimated by one single operation via Ordinary Least Squares (OLS).

$\hat{Y} = \underbrace{\beta_0 + \beta_1 X}$ - $\hat{Y}$ is called Y-hat and is a vector containing predicted values.

23

Simple regression model

$Y = \beta_0 + \beta_1 X + e$ - e represents the vector of residual values.

$\beta = (X^T X)^{-1} X^T Y$ - Slope and intercept estimated by one single operation via Ordinary Least Squares (OLS).

$\hat{Y} = \beta_0 + \beta_1 X$ - $\hat{Y}$ is called Y-hat and is a vector containing predicted values.

$e = Y - \hat{Y}$ - e represents the vector of residual values.

24

ANOVA as a regression model

$Y = \beta_0 + \beta_1 X + e$

$\hat{Y} = \beta_0 + \beta_1 X$

$\beta = (X^T X)^{-1} X^T Y$

$\beta_0 = 2.333 \therefore \beta_1 = 2.833$

back to our tiny example

X

Response (Y)	Constant (β_0)	Predictor (β_1)
1.2	1	0
2.7	1	0
3.1	1	0
4.1	1	1
5.3	1	1
6.1	1	1

25

ANOVA as a regression model

$Y = \beta_0 + \beta_1 X + e$

$\hat{Y} = \beta_0 + \beta_1 X$

$\beta = (X^T X)^{-1} X^T Y$

$\beta_0 = 2.333 \therefore \beta_1 = 2.833$

$\hat{Y} = 2.333 + 2.833 X_1$

$e = Y - \hat{Y}$

- $\hat{Y}$ is called Y-hat and represents the vector of predicted values.
- e represents the vector of residual values.

Response (Y)	Constant (β_0)	Predictor X_1 (β_1)	$\hat{Y}$	e
1.2	1	0	2.33	-1.13
2.7	1	0	2.33	0.37
3.1	1	0	2.33	0.77
4.1	1	1	5.17	-1.07
5.3	1	1	5.17	0.13
6.1	1	1	5.17	0.93

26

ANOVA as a regression model

Response (Y)	Constant (β_0)	Predictor (β_1)	$\hat{Y}$	e
1.2	1	0	2.33	-1.13
2.7 $\bar{X}$	1	0	2.33	0.37
3.1	1	0	2.33	0.77
4.1	1	1	5.17	-1.07
5.3 $\bar{X}$	1	1	5.17	0.13
6.1	1	1	5.17	0.93

In ANOVAs, predicted values are the predicted mean values per group

27

ANOVA as a regression model

Response (Y)	Constant (β_0)	Predictor (β_1)	$\hat{Y}$	e
1.2	1	0	2.33	-1.13
2.7 $\bar{X}$	1	0	2.33	0.37
3.1	1	0	2.33	0.77
4.1	1	1	5.17	-1.07
5.3 $\bar{X}$	1	1	5.17	0.13
6.1	1	1	**5.17**	**0.93**

$e_6 = 6.10 - 5.17 = 0.93$

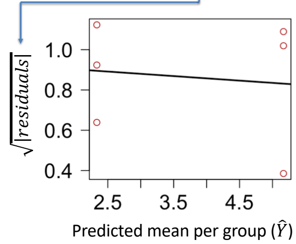
In ANOVAs, predicted values are the predicted mean values per group, and **residuals (e) represent variation around the observed group mean not explained by the regression model (or ANOVA).**

28

Plot between the square root of the absolute ANOVA residuals (i.e., deviations from the predicted mean group) against predicted mean per group

Response (Y)	Constant (β_0)	Predictor (β_1)	$\hat{Y}$	e
1.2	1	0	2.33	-1.13
2.7	1	0	2.33	0.37
3.1	1	0	2.33	0.77
4.1	1	1	5.17	-1.07
5.3	1	1	5.17	0.13
6.1	1	1	5.17	0.93

Variance of residuals looks ok, particularly given the small number of replicates per group.



29

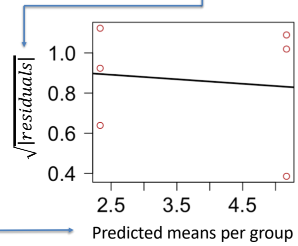
Plot of residuals on predicted values (ANOVA as a regression model) **versus** standard Levene's test for testing for homoscedasticity among groups

Response (Y)	Constant (β_0)	Predictor (β_1)	$\hat{Y}$	e
1.2	1	0	2.33	-1.13
2.7	1	0	2.33	0.37
3.1	1	0	2.33	0.77
4.1	1	1	5.17	-1.07
5.3	1	1	5.17	0.13
6.1	1	1	5.17	0.93

Levene's test

```
      Df F value Pr(>F)
group  1  0.0034 0.9562
       4
```

Variance of residuals are ok!



30

Coding for predictors with 3 groups (more groups and more factors, more predictors)

Response	Factor	Constant (β_0)	Predictor (β_1)	Predictor (β_2)
1.2	control	1	0	0
2.7	control	1	0	0
3.1	control	1	0	0
4.1	Group_1	1	1	0
5.3	Group_1	1	1	0
6.1	Group_1	1	1	0
8.1	Group_2	1	0	1
9.4	Group_2	1	0	1
10.1	Group_2	1	0	1

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$$

Multifactorial ANOVAs become then multiple regression models

31

How does heteroscedasticity affect variance of residual variation in ANOVAs?

32

Here we will understand:

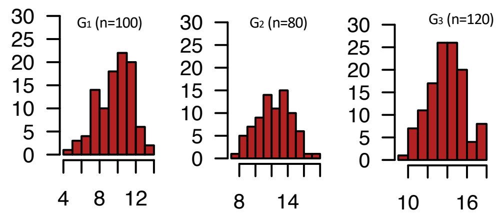
1) How heteroscedasticity affects variance of residual variation in ANOVAs

And

2) How weighted least squares (WLS) approach can be used to deal with heteroscedasticity in ANOVAs

33

GOING BACK TO the simulated normally distributed *homoscedastic* simulated data

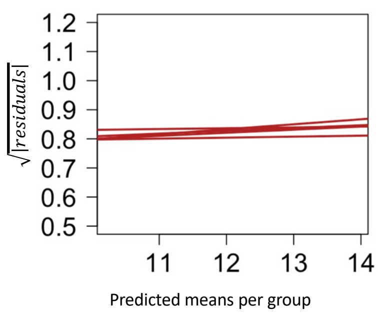


One-factorial design - 3 groups, normally distributed homoscedastic data ($\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 4$), varying in means ($\mu_1^2 = 10, \mu_2^2 = 12, \mu_3^2 = 14$)

34

One-factorial design - 3 groups, normally distributed homoscedastic data ($\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 4$), varying in means ($\mu_1^2 = 10, \mu_2^2 = 12, \mu_3^2 = 14$)

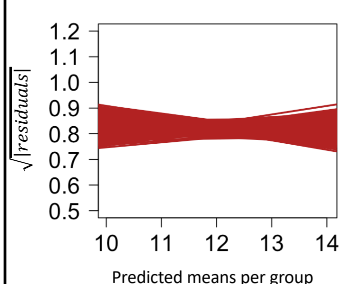
5 simulated data sets



35

One-factorial design - 3 groups, normally distributed homoscedastic data ($\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 4$), varying in means ($\mu_1^2 = 10, \mu_2^2 = 12, \mu_3^2 = 14$)

1000 simulated data sets



Sample variation from homoscedastic populations -

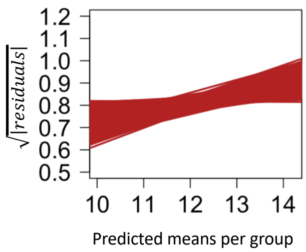
H_0 for population mean (μ) differences is set to **FALSE**.

H_0 for population variance differences (σ) is set to **TRUE**.

36

One-factorial design - 3 groups, normally distributed
heteroscedastic data ($\sigma_1^2 = 4, \sigma_2^2 = 6.25, \sigma_3^2 = 9$),
varying in means ($\mu_1^2 = 10, \mu_2^2 = 12, \mu_3^2 = 14$)

1000 simulated data sets



Sample variation from heteroscedastic populations -

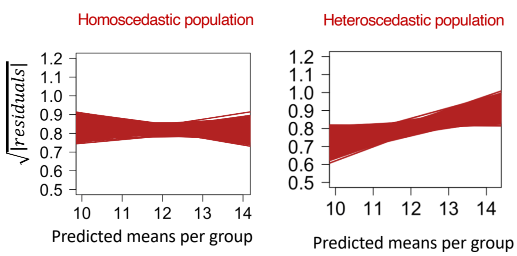
H_0 for population mean (μ) differences is set to **FALSE**.

H_0 for population variance differences (σ) is set to **FALSE**.

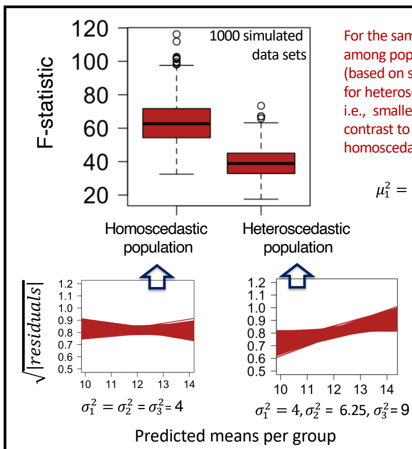
37

One-factorial design - 3 groups, normally distributed
homoscedastic data ($\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 4$) and heteroscedastic data ($\sigma_1^2 = 4, \sigma_2^2 = 6.25, \sigma_3^2 = 9$),
varying in means ($\mu_1^2 = 10, \mu_2^2 = 12, \mu_3^2 = 14$)

1000 simulated data sets



38



For the same mean differences among populations, the F-statistic (based on samples) is much smaller for heteroscedastic populations, i.e., smaller statistical power in contrast to the F-statistic for homoscedastic populations.

$\mu_1^2 = 10, \mu_2^2 = 12, \mu_3^2 = 14$

39

On the other hand, when samples are taken from populations with the same means, but their variances vary (heteroscedastic) then Type I error can increase!

(this will be demonstrated in TUTORIAL 5)

Sample variation from heteroscedastic populations:

H_0 for population mean (μ) differences is set to **TRUE**.

H_0 for population variance differences (σ) is set to **FALSE**.

AUSTRIAN JOURNAL OF STATISTICS
Volume 36 (2007), Number 3, 179–188

How to keep the Type I Error Rate in ANOVA if Variances are Heteroscedastic

Karl Moder
Institute of Applied Statistics and Computing,
University of Natural Resources and Applied Life Sciences, Vienna

Abstract: One essential prerequisite to ANOVA is homogeneity of variances in underlying populations. Violating this assumption may lead to an increased type I error rate. The reason for this undesirable effect is due to the calculation of the corresponding F-value. A slightly different test statistic keeps the level α. The underlying distribution of this alternative method is Hotelling's T^2. As Hotelling's T^2 can be approximated by a Fisher's F-distribution, this alternative test is very similar to an ordinary analysis of variance.

40

Here we will understand:

1) How heteroscedasticity affects variance of residual variation in ANOVAs ✓

And

⇨ 2) **How weighted least squares (WLS) approach can be used to deal with heteroscedasticity in ANOVAs**

41

The weighted least squares (WLS) approach

$$\beta = (X^{\mathrm{T}}X)^{-1}\, X^{\mathrm{T}}Y\ (OLS)$$

$$\beta = (X^{\mathrm{T}}WX)^{-1}\, X^{\mathrm{T}}WY\ (WLS)$$

$W =$

1	0	0	0	0	0
0	1	0	0	0	0
0	0	1	0	0	0
0	0	0	1	0	0
0	0	0	0	1	0
0	0	0	0	0	1

OLS and WLS are equal when W is an identity matrix in which all (main) diagonal elements equal to 1, i.e., all observations have the same weight in the regression estimates.

42

The weighted least squares (WLS) approach
Let's understand how weights change statistical estimates
(the case of the weighted mean)

$$\frac{1\times2+2\times3+3\times4+4\times5}{14}=2.86$$
Weighted mean
Weights = **2,3,4,5**

$$\frac{1+2+3+4}{4}=2.5$$
regular mean

$$\frac{1\times5+2\times4+3\times3+4\times2}{14}=2.14$$
Weighted mean
Weights = **5,4,3,2**

43

The weighted least squares (WLS) approach
Let's understand how weights change statistical estimates
(the case of the weighted mean)

$$\frac{1\times2+2\times3+3\times4+4\times5}{14}=2.86$$
Weighted mean
Weights = **2,3,4,5**

$$\frac{1+1+2+2+2+3+3+3+3+4+4+4+4+4}{14}=\frac{40}{14}=2.86$$

44

The weighted least squares (WLS) approach

$$\beta=(X^{T}X)^{-1}\,X^{T}Y\ (OLS)$$

$$\beta=(X^{T}WX)^{-1}\,X^{T}WY\ (WLS)$$

Response (Y)	Constant (β_0)	Predictor (β_1)	$\hat{Y}$	e
1.2	1	0	2.33	-1.13
2.7	1	0	2.33	0.37
3.1	1	0	2.33	0.77
4.1	1	1	5.17	-1.07
5.3	1	1	5.17	0.13
6.1	1	1	5.17	0.93

Variance of residuals per group
1.003333
1.013333

45

The weighted least squares (WLS) approach –
more variance, less influence in the regression estimation

$$\beta = (X^{T}X)^{-1}\, X^{T}Y\ (OLS)$$

$$\beta = (X^{T}WX)^{-1}\, X^{T}WY\ (WLS)$$

$$W = 1/s^{2}_{group}$$

In **OLS**, each observation has the same weight (inform the model in the same way. In **WLS**, we treat each observation as more (smaller group residual variance) or less (larger groups residual variance) **informative** about the underlying relationship between X and Y.

46

The weighted least squares (WLS) approach –
more variance, less influence in the regression estimation

$$\beta = (X^{T}X)^{-1}\, X^{T}Y\ (OLS)$$

$$\beta = (X^{T}WX)^{-1}\, X^{T}WY\ (WLS)$$

1 / Variance of residuals per group

$W = 1/$

0.997	0	0	0	0	0
0	0.997	0	0	0	0
0	0	0.997	0	0	0
0	0	0	0.990	0	0
0	0	0	0	0.990	0
0	0	0	0	0	0.990

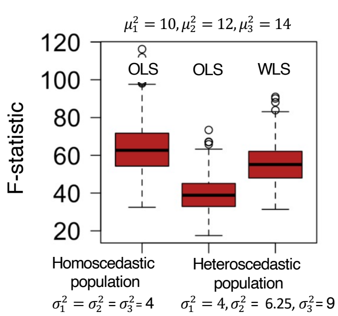
1 / 1.003333

1 / 1.013333

The influence of each observation is the inverse of its group residuals variance (i.e., reciprocal, 1/variance)

47

For the same mean differences among populations, the F-statistic (based on samples) is much smaller for heteroscedastic populations, i.e., smaller statistical power in contrast to the F-statistic for homoscedastic populations. The WLS makes it more powerful (larger F-values) and much closer to what is expected for homoscedastic populations.



48