

# Statistical Modeling

CH.6 - Weighted Least Squares

2023 || Prof. Dr. Buchwitz

1 Organizational Information

2 Weighted Least Squares

---

Session	Topic
1	Simple Linear Regression
2	Multiple Linear Regression
3	Regression Diagnostics
4	Qualitative Variables as Predictors
5	Transformation of Variables
6	Weighted Least Squares
7	Correlated Errors
8	Analysis of Collinear Data
9	Working with Collinear Data
10	Variable Selection Procedures
11	Logistic Regression
12	Further Topics

---

1 Organizational Information

2 Weighted Least Squares

- Estimation using *weighted least squares* is equivalent to performing OLS on the transformed variables.
- We discuss **WLS** as a method of dealing with heteroscedasticity of errors as well as an estimation method in its own (WLS performs better for e.g. fitting *logistic models* or *dose-response-curves*).
- WLS allows relaxing the assumption of equal error variance, so that the  $\epsilon_i$ 's are assumed to be **independently distributed with zero mean and**  $\text{Var}(\epsilon_i) = \sigma_i^2$  instead of  $\text{Var}(\epsilon_i) = \sigma^2$ .

- Obtaining the **WLS estimates** of  $\beta_0, \beta_1, \dots, \beta_p$  requires minimizing

$$\sum_{i=1}^n \omega_i (y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_p x_{ip})^2$$

- Usually the  $\omega_i$  are weights that are inversely proportional to the variance of the residuals, like  $\omega_i = 1/\sigma_i^2$ .
- Any observation with a small weight will be severely discounted by WLS in determining the estimates of  $\beta_0, \beta_1, \dots, \beta_p$ .
- In the extreme case where  $\omega_i = 0$  the  $i$ -th observation will be excluded from the estimation process.

- The approach we take here when estimating WLS is a **two step estimation** approach, which assumes that the weights  $\omega$  are unknown.
  - 1) We collect knowledge about the process that generates the data (DGP) and evidence from an **OLS** fit to detect the heteroscedastic problem.
  - 2) The **OLS** fit itself or the gathered evidence serves as basis for determining the weights  $\omega$ . Those weights are used in the **WLS** fit.

- 1) Variance proportional to a regressor
- 2) Heterogeneity of variance as consequence of data collection
- 3) Unknown source of heteroscedasticity and empirical identification of the structure



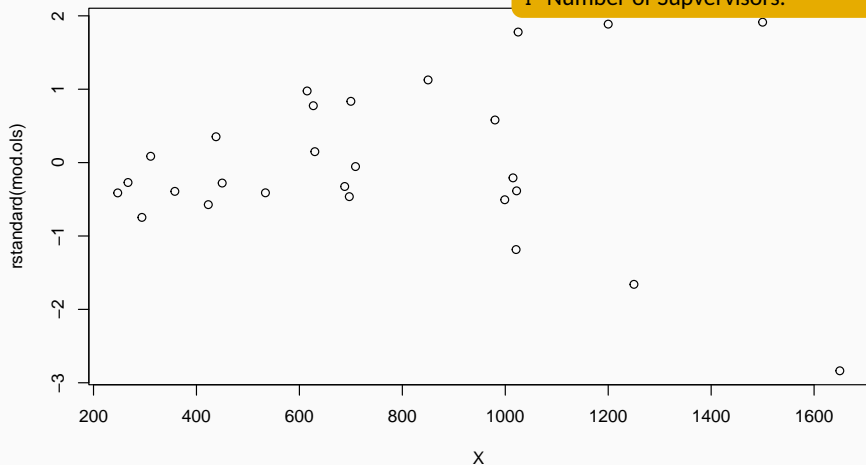
# Heteroscedastic Models

```
mod.ols <- lm(Y ~ 1 + X, data=P176)  
plot(P176$X,rstandard(mod.ols), xlab = "X")
```

## Data Description

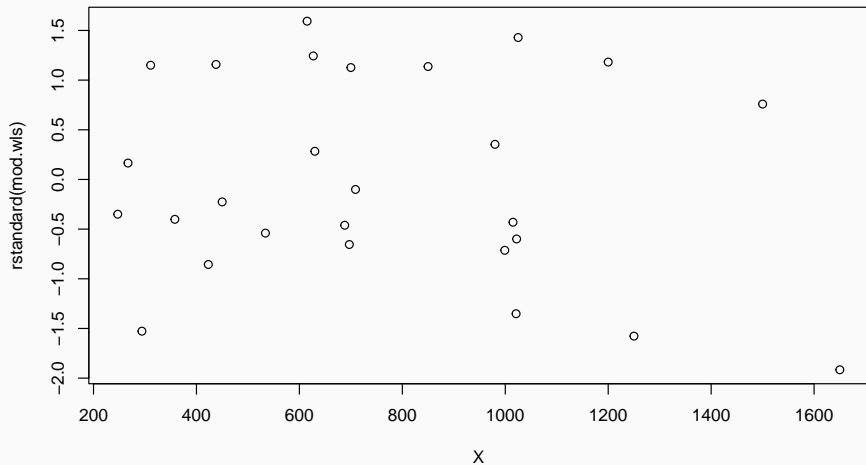
X Number of supervised workers.

Y Number of Supervisors.



# Type 1: Heteroscedastic Models

```
mod.wls <- lm(Y ~ 1 + X, weights = 1/X^2, data=P176)  
plot(P176$X,rstandard(mod.wls), xlab = "X")
```



## Type 1: Heteroscedastic Models

- In the given example we argued that the variance of  $\epsilon_i$  is proportional to the size of the establishment measured by  $x_i^2$ , so that  $\text{Var}(\epsilon_i) = \sigma_i^2 = k^2 x_i^2$ , with  $k > 0$ .
- The same approach also works in multiple regression, given that the variance of the residuals is only affected by one of the predictors (e.g.  $X_2$ ), the estimates of the parameters are determined by minimizing:

$$\sum_{i=1}^n \frac{1}{x_{i2}^2} (y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_p x_{ip})^2$$

All modern statistical packages provide WLS procedures. The transformation approach to WLS discussed in the previous chapter is to foster your understanding and (usually) not used when estimating models.

## Type 2: Heteroscedastic Models

- Another type of heteroscedasticity often occurs in surveys, where the observations are **averages of individual sampling units** taken over distinct groups or clusters.
- Due to the properties of the mean (which is a random variable) the variance is proportional to the square root of the sample size, on which the average is based, that is  $\sigma_{\bar{y}_i} = \sigma / \sqrt{n_i}$ . Here  $\sigma$  is the standard deviation of  $Y$  in the population.
- This leads to  $\omega_i = 1/\sigma_i^2$  as weights for the WLS approach.

## Type 2: Heteroscedastic Models

- The precision of measurement is the justification for weighting the observations in this fashion. Averages that are based on few observations (high variance) should play a smaller role in estimating the overall effect.

Estimation is then carried out by minimizing:

$$\begin{aligned} S &= \sum_{i=1}^n \omega_i (y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_p x_{ip})^2 \\ &= \sum_{i=1}^n \frac{1}{\sigma_i^2} (y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_p x_{ip})^2 \end{aligned}$$

As  $\sigma_i^2 = \sigma^2/n_i$ :

$$= \sum_{i=1}^n n_i (y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_p x_{ip})^2$$

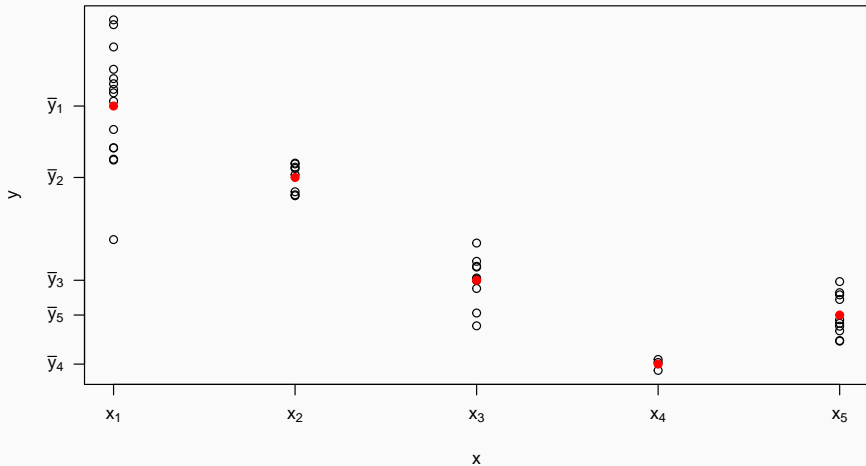
### Remember

The value  $x_{ij}$  is an average calculated based on  $n_j$  observations!

## Type 3: Heteroscedastic Models

- We deal with heteroscedasticity by transformation of variables, where the transformations are constructed from information in the raw data.
- In the following only the indication for heteroscedasticity is drawn from the raw data and the structure is determined empirically. Therefore the estimation requires **two stages**.

## Type 3: Heteroscedastic Models



## Type 3: Heteroscedastic Models

- The regression model for the shown sample data could be stated as follows, where  $\text{Var}(\epsilon_{ij}) = \sigma_j^2$ .

$$y_{ij} = \beta_0 + \beta_1 x_j + \epsilon_{ij} \quad \text{with } i = 1, 2, \dots, n_j \text{ and } j = 1, 2, 3, 4, 5$$

- The observed residual for the  $i$ -th observation in the  $j$ -th group is  $e_{ij} = y_{ij} - \hat{y}_{ij}$ .
- Adding and subtracting the mean of the response  $\bar{y}_j$  reveals that the residual has two parts which occur because of **pure error** and **lack of fit** respectively.

$$e_{ij} = \underbrace{(y_{ij} - \bar{y}_j)}_{\text{pure error}} + \underbrace{(\bar{y}_j - \hat{y}_{ij})}_{\text{lack of fit}}$$



## Type 3: Heteroscedastic Models

- The weights for fitting using **WLS** can be determined based on the pure error so that they are inversely proportional to the variance in the group

$$\omega_{ij} = 1/s_j^2.$$

$$s_j^2 = \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 / (n_j - 1)$$

- The question what constitutes a group can only be answered in a specific setting and a plausible way to explore heteroscedasticity is by clustering observations according to prior, natural and meaningful associations (that are often already available as variable or can also be constructed with modest effort).

# Example: Education Expenditure

P198

##	Y	X1	X2	X3	Region
## ME	235	3944	325	508	1
## NH	231	4578	323	564	1
## VT	270	4011	328	322	1
## MA	261	5233	305	846	1
## RI	300	4780	303	871	1
## CT	317	5889	307	774	1
## NY	387	5663	301	856	1
## NJ	285	5759	310	889	1
## PA	300	4894	300	715	1
## OH	221	5012	324	753	2
## IN	264	4908	329	649	2
## IL	308	5753	320	830	2
## MI	379	5439	337	738	2
## WI	342	4634	328	659	2
## MN	378	4921	330	664	2
## IA	232	4869	318	572	2
## MO	231	4672	309	701	2
## ND	246	4782	333	443	2
## SD	230	4296	330	446	2
## NB	268	4827	318	615	2
## KS	337	5057	304	661	2
## DE	344	5540	328	722	3
## MD	330	5331	323	766	3
## VA	261	4715	317	631	3
## WV	214	3828	310	390	3
## NC	245	4120	321	450	3
## SC	233	3817	342	476	3
## GA	250	4243	339	603	3
## FL	243	4647	287	805	3
## KY	216	3967	325	523	3
## TN	212	3946	315	588	3

## Data Description

Y Per capita expenditure on education projected for 1975.

X1 Per capita income in 1973.

X2 Number of residents per thousand under 18 years of age in 1974.

X3 Number of residents per thousand living in urban areas in 1970.

Region Geographic Region: (1) Northeast, (2) North Central, (3) South and (4) West.

## Example: Education Expenditure

- The objective is to get the best representation of the relationship between **expenditure on education** ( $Y$ ) and the other variables for all 50 states in the dataset.
- The data are grouped in a *natural way*, by geographic region.
- Our assumption is that, although the relationship is structurally the same for each region, the coefficients and residual variances may differ from region to region.
- The different variances constitute a case of heteroscedasticity that can be **treated directly in the analysis**.

## Example: Education Expenditure

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

- The model above is the model for the estimation. It should be noted that the data could be analyzed using indicator variables to look for effects associated with the regions.
- However, our objective here is to develop one relationship that can serve as the best representation for all regions and all states.
- The goal is accomplished by taking regional differences into account through the **WLS** estimation process.

## Example: Education Expenditure

- We assume that there is a unique residual variance associated with each of the four regions. The variances are denoted as  $(c_1\sigma)^2$ ,  $(c_2\sigma)^2$ ,  $(c_3\sigma)^2$  and  $(c_4\sigma)^2$ , where  $\sigma$  is the common part and the  $c_j$ 's are unique to the regions. The regression coefficients should be determined by minimizing

$$S_{\omega} = S_1 + S_2 + S_3 + S_4$$

- The individual sum of squares for each region  $j = 1, 2, 3, 4$  is given below:

$$S_j = \sum_{i=1}^{n_j} \frac{1}{c_j^2} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_3 x_{i3})^2$$

## Example: Education Expenditure

The factors  $1/c_j^2$  are the weights that determine how much influence each observation has in estimating the regression coefficients. The weighting scheme can be justified in two ways:

- 1) Arguing that observations that are most erratic (large error variance) should have little influence in determining the coefficients.
- 2) The **WLS** approach allows *transforming the data* so that the residual variance is constant. This can be achieved by dividing  $Y$ ,  $X_1$ ,  $X_2$  and  $X_3$  by the appropriate  $c_j$  and yields an error term, that is (in concept) also divided by  $c_j$ . The resulting residuals have common variance and the desired least squares properties.

## Example: Education Expenditure

- The values of the  $c_j$ 's are unknown and must be estimated in the same sense that  $\sigma^2$  and the  $\beta$ 's must be estimated.
- Basis for the estimation are the residuals from an **OLS** fit to the raw data, that are grouped together by region.

$$\hat{c}_j^2 = \frac{\hat{\sigma}_j^2}{\frac{1}{n} \sum_{i=1}^n e_i^2}$$

Before estimating the  $c_j$ 's, we check the residuals for obvious misspecification of the model.

## Example: Education Expenditure

```
mod <- lm(Y ~ 1 + X1 + X2 + X3, data=P198)
summary(mod)
```

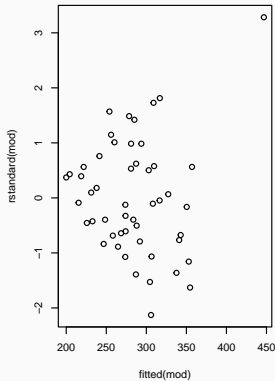
```
##
## Call:
## lm(formula = Y ~ 1 + X1 + X2 + X3, data = P198)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -84.88 -26.88  -3.83   22.25  99.24
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.57e+02   1.23e+02  -4.52  4.3e-05 ***
## X1           7.24e-02   1.16e-02   6.24  1.3e-07 ***
## X2           1.55e+00   3.15e-01   4.93  1.1e-05 ***
## X3          -4.27e-03   5.14e-02  -0.08   0.93
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.5 on 46 degrees of freedom
## Multiple R-squared:  0.591, Adjusted R-squared:  0.565
## F-statistic: 22.2 on 3 and 46 DF, p-value: 4.94e-09
```



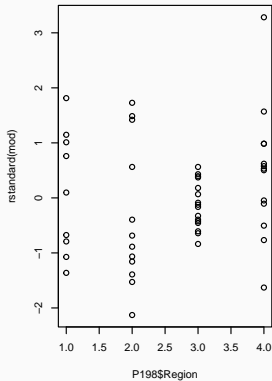
# Example: Education Expenditure

```
par(mfrow=c(1,3))  
plot(fitted(mod), rstandard(mod), main="Residuals vs. fitted Values")  
plot(P198$Region, rstandard(mod), main="Residuals grouped by Region")  
plot(hatvalues(mod), main="Leverage Values")
```

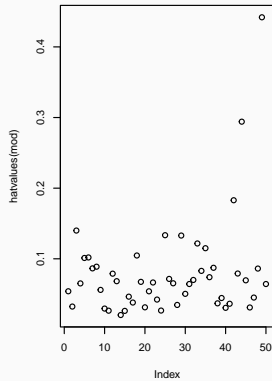
Residuals vs. fitted Values



Residuals grouped by Region



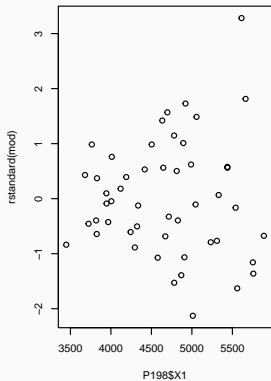
Leverage Values



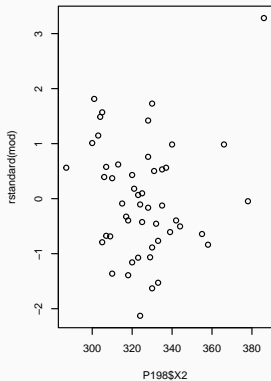
# Example: Education Expenditure

```
par(mfrow=c(1,3))  
plot(P198$X1, rstandard(mod), main="Residuals vs. Predictor X1")  
plot(P198$X2, rstandard(mod), main="Residuals vs. Predictor X2")  
plot(P198$X3, rstandard(mod), main="Residuals vs. Predictor X3")
```

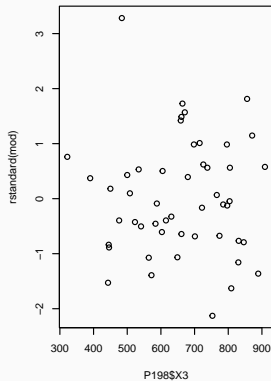
Residuals vs. Predictor X1



Residuals vs. Predictor X2

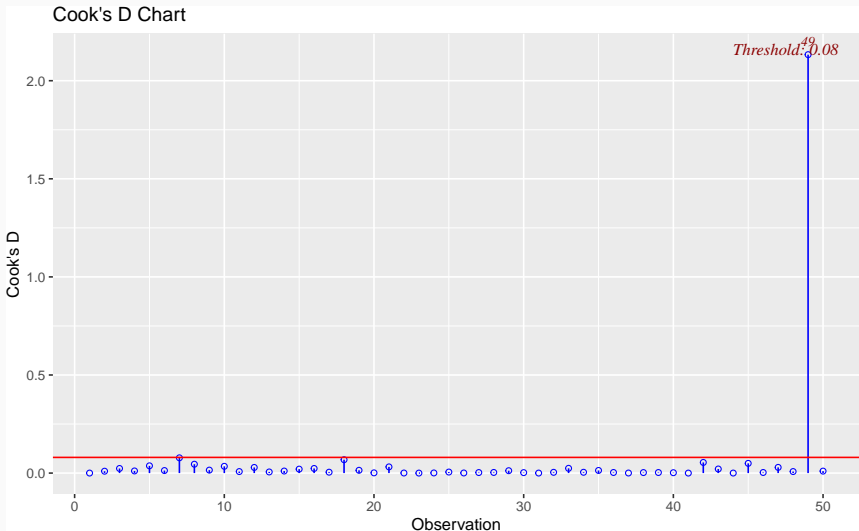


Residuals vs. Predictor X3



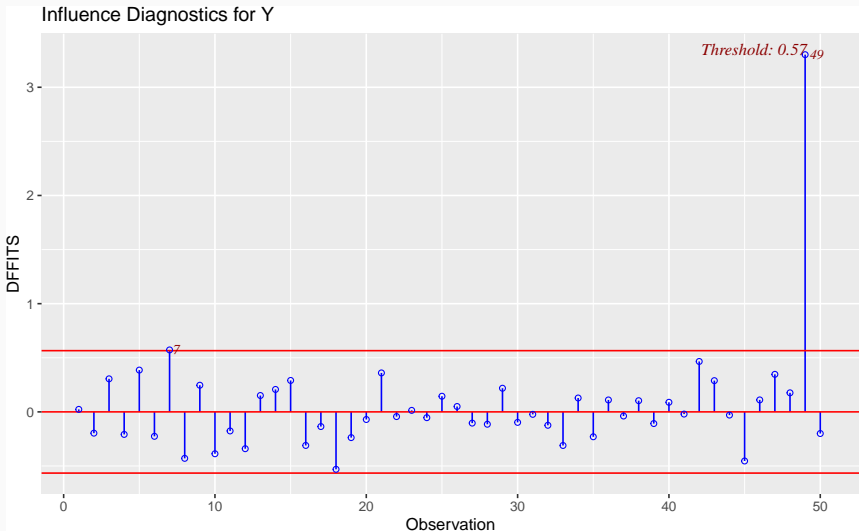
# Example: Education Expenditure

```
olsrr::ols_plot_cooksd_chart(mod)
```



# Example: Education Expenditure

```
olsrr::ols_plot_dffits(mod)
```



## Example: Education Expenditure

- Observation 44 (UT = Utah) and Observation 49 (AK = Alaska) are high-leverage points.

```
head(sort(hatvalues(mod), decreasing = T))
```

```
##      AK      UT      NM      VT      WV      FL  
## 0.4419 0.2941 0.1829 0.1400 0.1334 0.1329
```

- Only AK has high leverage and is influential (see DFITS and Cooks Distances plots). UT is only a high-leverage point without being influential.
- The data is from **1975**, Alaska has a very small population and an oil revenue boom. We judge that the respective education budget is not strictly comparable to the other states due to the unique situation. Therefore we exclude Alaska from the analysis.

```
d <- P198[-49,] # Remove Alaska (AK)
```

## Example: Education Expenditure

- After removing the outlier from the data and refitting an **OLS** model, the residuals can be used to estimate possible weights for the **WLS** estimation  $c_j$ .

```
d <- P198[-49,] # Remove Outlier
mod.ols <- lm(Y ~ 1 + X1 + X2 + X3, data=d)

res <- split(residuals(mod.ols), d$Region)
sigma_sq <- sapply(res, function(e){sum(e^2)/(length(e)-1)})
c_sq <- sigma_sq/(1/nrow(d)*sum(residuals(mod.ols)^2))
w <- 1/c_sq
knitr::kable(cbind(Region=1:4,n=sapply(res,length),sigma_sq, c=sqrt(c_sq), w),
              digits = 4, booktabs=T)
```

Region	n	sigma_sq	c	w
1	9	1632.5	1.1774	0.7213
2	12	2658.5	1.5026	0.4429
3	16	266.1	0.4753	4.4258
4	12	1036.8	0.9383	1.1357

## Example: Education Expenditure

- Below the complete code for the analysis is shown (somewhat redundant but hopefully useful for clarification).

```
mod <- lm(Y ~ 1 + X1 + X2 + X3, data=P198)

d <- P198[-49,] # Remove Outlier
mod.ols <- lm(Y ~ 1 + X1 + X2 + X3, data=d)

res <- split(residuals(mod.ols), d$Region)
sigma_sq <- sapply(res, function(e){sum(e^2)/(length(e)-1)})
d$sigma_sq <- rep(sigma_sq, times=sapply(res, length))

w <- 1/(sigma_sq/(1/nrow(d)*sum(residuals(mod.ols)^2)))
d$w <- rep(w, times=sapply(res, length))
mod.wls <- lm(Y ~ 1 + X1 + X2 + X3, weights=w, data=d)
```

## Example: Education Expenditure

```
texreg::texreg(list(mod, mod.ols, mod.wls), digits=3,  
  custom.model.names = c("OLS + Outlier", "OLS", "WLS"))
```

	OLS + Outlier	OLS	WLS
(Intercept)	-556.568*** (123.195)	-277.577* (132.423)	-316.024*** (77.419)
X1	0.072*** (0.012)	0.048*** (0.012)	0.062*** (0.008)
X2	1.552*** (0.315)	0.887* (0.331)	0.874*** (0.198)
X3	-0.004 (0.051)	0.067 (0.049)	0.029 (0.034)
R <sup>2</sup>	0.591	0.497	0.760
Adj. R <sup>2</sup>	0.565	0.463	0.744
Num. obs.	50	49	49

\*\*\*  $p < 0.001$ ; \*\*  $p < 0.01$ ; \*  $p < 0.05$

**Table 3:** Statistical models



## Example: Education Expenditure

- R computes the  $R^2$  in a different way for **WLS** models and employs a weighted mean instead of simply  $R^2 = [\text{Cor}(Y, \hat{Y})]^2$ .

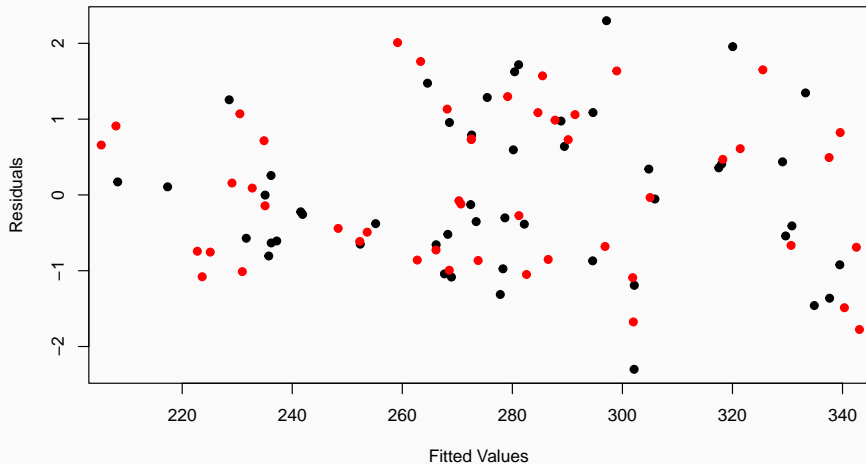
```
# R^2 for OLS and WLS models as we defined it  
c(R2.ols=cor(d$Y, fitted(mod.ols))^2, R2.wls = cor(d$Y, fitted(mod.wls))^2)
```

```
## R2.ols R2.wls  
## 0.4967 0.4861
```

- Only the  $R^2_{\text{OLS}}$  is the same as in the output table.
- For Details on how R calculates  $R^2$  in the case of **WLS** have a look at the links [here](#), [here](#) and [here](#).

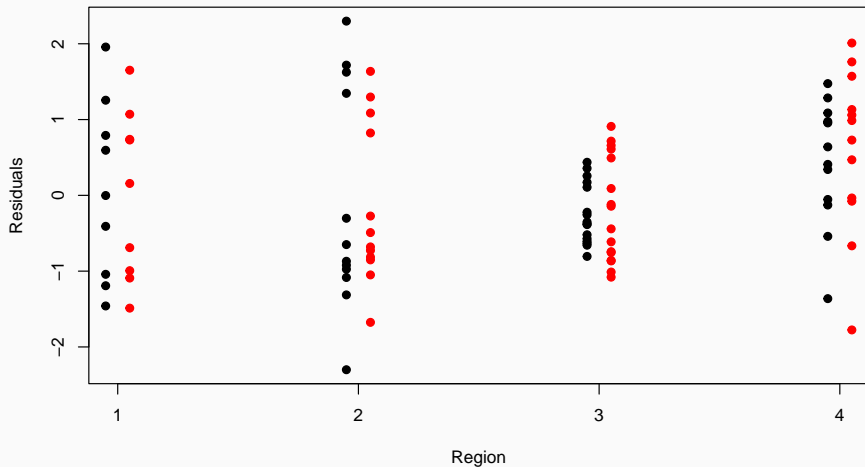
# Example: Education Expenditure

Residuals vs. fitted Values (WLS in red)



# Example: Education Expenditure

Residuals grouped by Region (WLS in red)



## Example: Education Expenditure

- The spread of the residuals has evened out for the **WLS** solution, which indicates that our treatment of heteroscedasticity was (at least partially) successful.
- The hypotheses for the **WLS** estimation are not exact since that estimation is carried out using **estimated** weights. The inherent uncertainty of the weights is not reflected in the hypotheses tests.
- The comparable  $R^2$  values show that  $R_{OLS}^2 > R_{WLS}^2$ , which must be the case as **OLS** provides a solution with minimum  $\hat{\sigma}$  (and thus maximum  $R^2$ ).
- The analysis is **far from perfect** as only roughly 50% of the variation in  $Y$  can be explained, so the search for additional variables and/or a better model should continue.

- Heteroscedasticity cannot only be treated by **WLS** estimation. In general heteroscedastic patterns in the residuals often disappear when sufficient indicator variables are introduced.
- In addition there are methods that are called **Sandwich Estimators** which adjust the standard errors of an OLS fit and produce **heteroscedasticity robust** standard errors. Those estimators are frequently used in econometrics, but not covered in this course.