Intermediate Econometrics

Lecture 5

Vincenzo Gioia

04 November 2026

Econometric Example

wage1 dataset

- The dataset is extracted from Wooldridge's Introductory Econometrics (2016)
- **Goal**:

1. detect heteroscedasticity (non-constant error variance) in a simple wage regression
2. show how it can be mitigated by applying a logarithmic transformation to the dependent variable, or alternatively by estimating the model on the original scale using Generalized Least Squares (GLS).

- Estimate a linear regression:

$$wage_i = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 tenure_i + \varepsilon_i$$

- Test for heteroscedasticity, fit via OLS and GLS, then consider a transformation of the outcome

Econometric Example

wage1 dataset

- The dataset wage1 is included in the wooldridge R package
- It is a classic example in labor economics
- It contains cross-sectional data on 526 working individuals in the U.S.
- Variables

1. wage: Hourly wage (in USD)
2. educ: Years of education
3. exper: Years of labor market experience
4. tenure: Years with current employer
5. nonwhite, female, etc.: Demographic indicators

Econometric Example

```
library(wooldridge)
library(sandwich)
library(nlme)

data("wage1")
str(wage1)
```

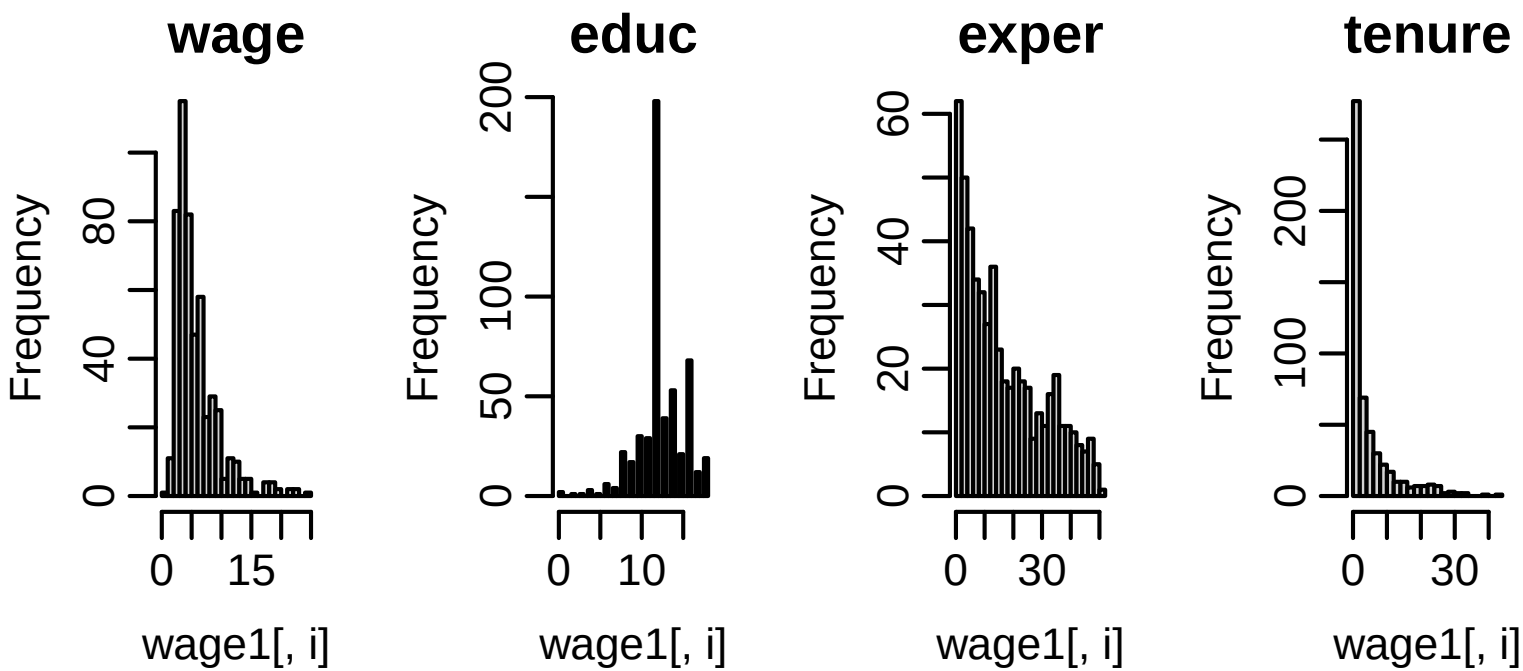
```
## 'data.frame':    526 obs. of  24 variables:
##  $ wage    : num  3.1 3.24 3 6 5.3 ...
##  $ educ    : int  11 12 11 8 12 16 18 12 12 17 ...
##  $ exper   : int  2 22 2 44 7 9 15 5 26 22 ...
##  $ tenure  : int  0 2 0 28 2 8 7 3 4 21 ...
##  $ nonwhite: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ female  : int  1 1 0 0 0 0 0 1 1 0 ...
##  $ married : int  0 1 0 1 1 1 0 0 0 1 ...
##  $ numdep  : int  2 3 2 0 1 0 0 0 2 0 ...
##  $ smsa    : int  1 1 0 1 0 1 1 1 1 1 ...
##  $ northcen: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ south   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ west    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ construc: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ndurman : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ trcommpu: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ trade   : int  0 0 1 0 0 0 1 0 1 0 ...
##  $ services: int  0 1 0 0 0 0 0 0 0 0 ...
##  $ profserv: int  0 0 0 0 0 1 0 0 0 0 ...
##  $ profocc : int  0 0 0 0 0 1 1 1 1 1 ...
##  $ clerocc : int  0 0 0 1 0 0 0 0 0 0 ...
##  $ servocc : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ lwage   : num  1.13 1.18 1.1 1.79 1.67 ...
##  $ expersq : int  4 484 4 1936 49 81 225 25 676 484 ...
##  $ tenursq : int  0 4 0 784 4 64 49 9 16 441 ...
##  - attr(*, "time.stamp")= chr "25 Jun 2011 23:03"
```

Econometric Example

```
summary(wage1[,1:4])
```

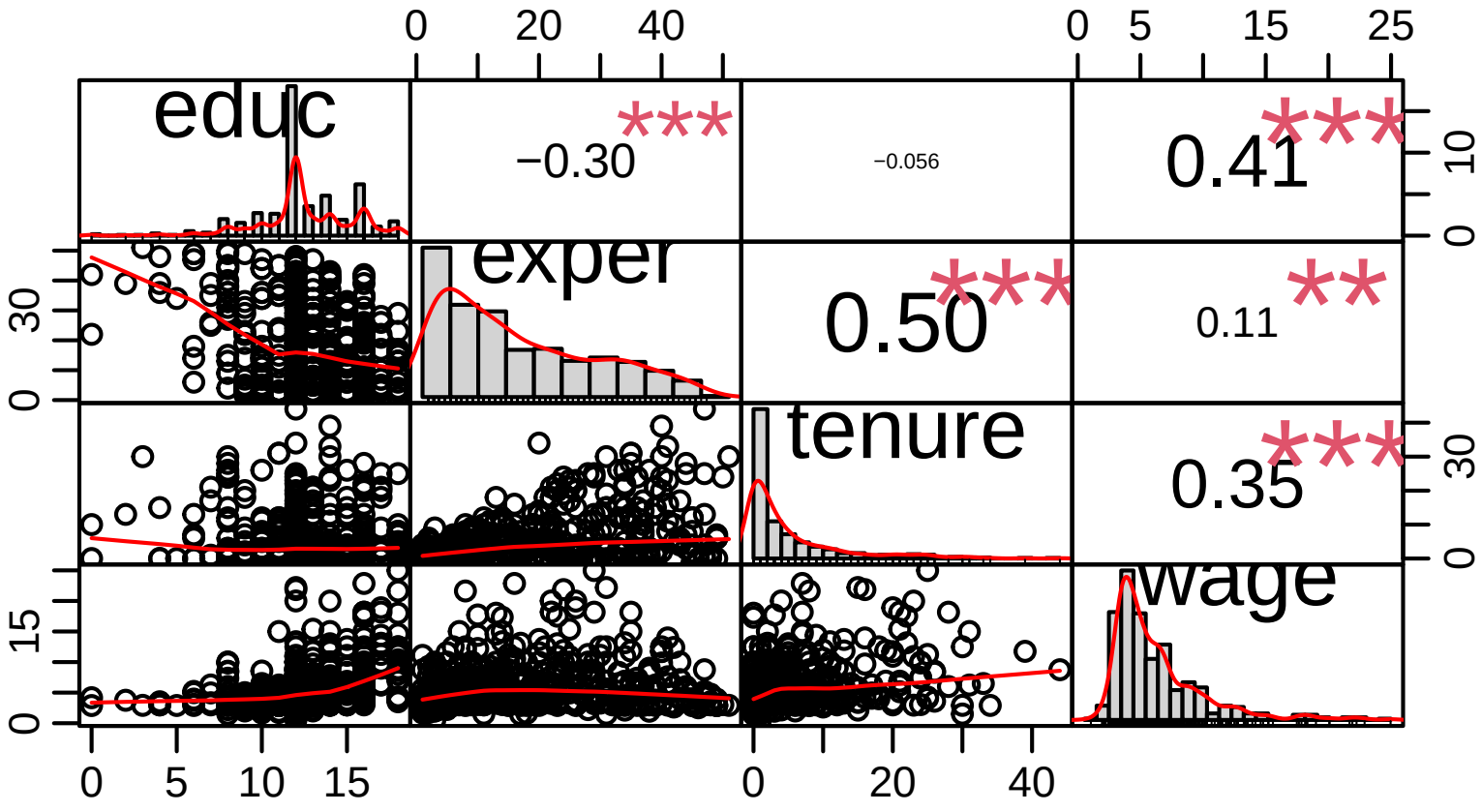
```
##       wage             educ           exper           tenure
##  Min.   : 0.530   Min.   : 0.00   Min.   : 1.00   Min.   : 0.000
##  1st Qu.: 3.330   1st Qu.:12.00   1st Qu.: 5.00   1st Qu.: 0.000
##  Median : 4.650   Median :12.00   Median :13.50   Median : 2.000
##  Mean   : 5.896   Mean   :12.56   Mean   :17.02   Mean   : 5.105
##  3rd Qu.: 6.880   3rd Qu.:14.00   3rd Qu.:26.00   3rd Qu.: 7.000
##  Max.   :24.980   Max.   :18.00   Max.   :51.00   Max.   :44.000
```

```
par(mfrow=c(1,4))
for(i in 1:4) hist(wage1[,i], main = names(wage1)[i], breaks = 30)
```



Econometric Example

```
library(PerformanceAnalytics)
chart.Correlation(wage1[,c(2,3,4,1)])
```



Econometric Example

OLS

- **educ**: Each additional year of education increases the hourly wage by about 0.60 dollars per hour, on average, holding fixed the other predictors

- **exper** and **tenure** typically have a positive but lower effect

```
ols <- lm(wage ~ educ + exper + tenure, data = wage1)
summary(ols)
```

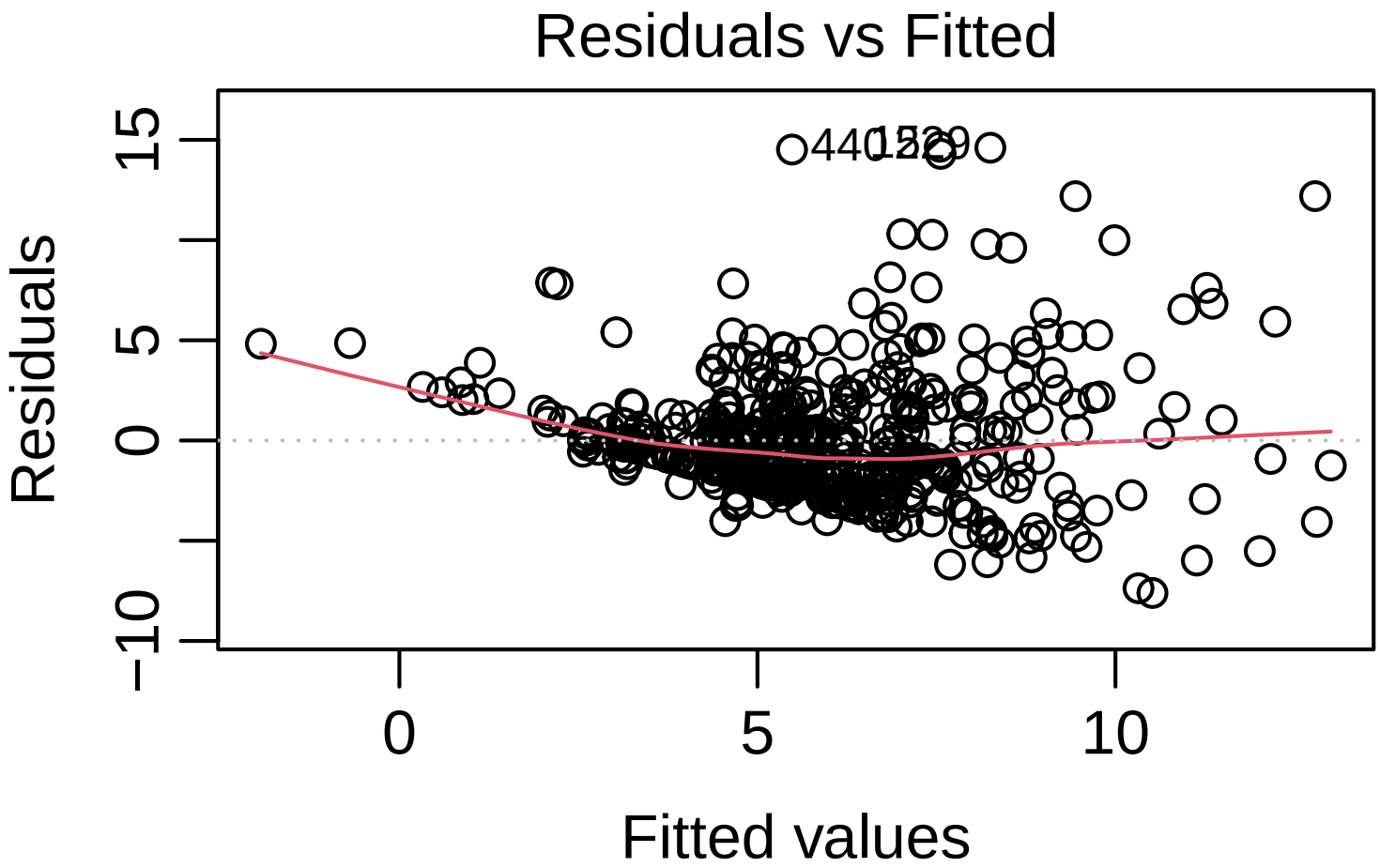
```
##
## Call:
## lm(formula = wage ~ educ + exper + tenure, data = wage1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6068 -1.7747 -0.6279  1.1969 14.6536
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.87273    0.72896  -3.941 9.22e-05 ***
## educ         0.59897    0.05128  11.679  < 2e-16 ***
## exper        0.02234    0.01206   1.853   0.0645 .
## tenure       0.16927    0.02164   7.820 2.93e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.084 on 522 degrees of freedom
## Multiple R-squared:  0.3064, Adjusted R-squared:  0.3024
## F-statistic: 76.87 on 3 and 522 DF,  p-value: < 2.2e-16
```

Econometric Example

wage1 dataset

- Residual variance increases for higher predicted wages — a clear sign of heteroscedasticity

```
plot(ols, 1)
```



Econometric Example

Testing for heteroscedasticity (Breusch-Pagan Test)

- Model $y_i = x_i^\top \beta + \varepsilon_i$, with $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$, $i = 1, \ldots, n$
- Test statistic:

$$LM = \frac{1}{2} f^\top W (W^\top W)^{-1} W^\top f$$

where

1. f is an n-dimensional vector composed of $(e_i^2/\hat{\sigma}^2 - 1)$, where e_i is the i-th residual of the OLS regression and $\hat{\sigma}^2$ is the estimate of the variance of the error term

2. W is a matrix of covariates. Indeed, we assume that σ_i^2 is a function of J covariates denoted by w_i, that is $\sigma_i^2 = h(\mathbf{w}_i^\top \delta)$, where the first element of the vector w_i is equal to 1

- The test statistic is, under $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2 = \sigma^2$ (homoscedasticity), distributed according to a Chi-square distribution with $J - 1$ degrees of freedom (as the first column of W is the intercept)

Econometric Example

wage1 dataset

- Breusch-Pagan Test: p-value close to 0 suggesting clear evidence for rejecting H_0 (so confirming the results of the graphical inspection)

```
library(lmtest)
bptest(ols)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  ols
## BP = 43.096, df = 3, p-value = 2.349e-09
```

- Let's explore the help page to understand the details of the test

Econometric Example

Testing for heteroscedasticity (Bartlett Test)

- Split the standardized residuals $\tilde{e}_i$ into G groups $\{\mathcal{G}_1, \ldots, \mathcal{G}_G\}$, according to the value on the x-axis of the scatterplot (an explanatory variable or the fitted values), and assume

$$\tilde{e}_i \sim \mathcal{N}(0, \sigma_g^2) \text{ if } \tilde{e}_i \in \mathcal{G}_g, \quad \text{independent}, \quad g = 1, \ldots, G$$

- The hypotheses are

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_G^2 \\ H_1 : \exists\, g, j \text{ such that } \sigma_g^2 \neq \sigma_j^2 \end{cases}$$

- **Note**: The standardized (or studentized) residuals are defined by

$$\tilde{e}_i = \frac{e_i}{S\sqrt{1 - p_{ii}}}$$

where p_{ii} is the i-th diagonal element of the projection matrix $P = X(X^\top X)^{-1}X^\top$

Econometric Example

Bartlett Test: test statistic

With n_g the size of group g and $n = \sum_{g=1}^{G} n_g$, the within-group variances and the pooled variance are

$$S_g^2 = \frac{\sum_{\tilde{e}_i \in \mathcal{G}_g} (\tilde{e}_i - \bar{\tilde{e}}_g)^2}{n_g - 1}, \qquad S_p^2 = \frac{1}{n - G} \sum_{g=1}^{G} (n_g - 1) S_g^2$$

$$T_B = \frac{(n - G) \ln S_p^2 - \sum_{g=1}^{G} (n_g - 1) \ln S_g^2}{1 + \frac{1}{3(G-1)} \left(\sum_{g=1}^{G} \frac{1}{n_g - 1} - \frac{1}{n - G} \right)}$$

- Under H_0, $T_B \dot{\sim} \chi^2_{G-1}$; reject if $T_B^{obs} > \chi^2_{G-1;1-\alpha}$
- T_B^{obs} is zero if all the group variances are equal
- The test is sensitive to departures from normality; the result depends on how the groups are chosen (there are some robust alternatives: Levene's test, etc.)

Econometric Example

wage1 dataset

- Standardized residuals of the OLS model, $G = 5$ groups of the fitted values

```
e_std <- rstandard(ols)
G <- 5
cuts <- quantile(fitted(ols), probs = seq(0, 1, length.out = G + 1))
grp <- cut(fitted(ols), breaks = cuts, include.lowest = TRUE)
table(grp)
```

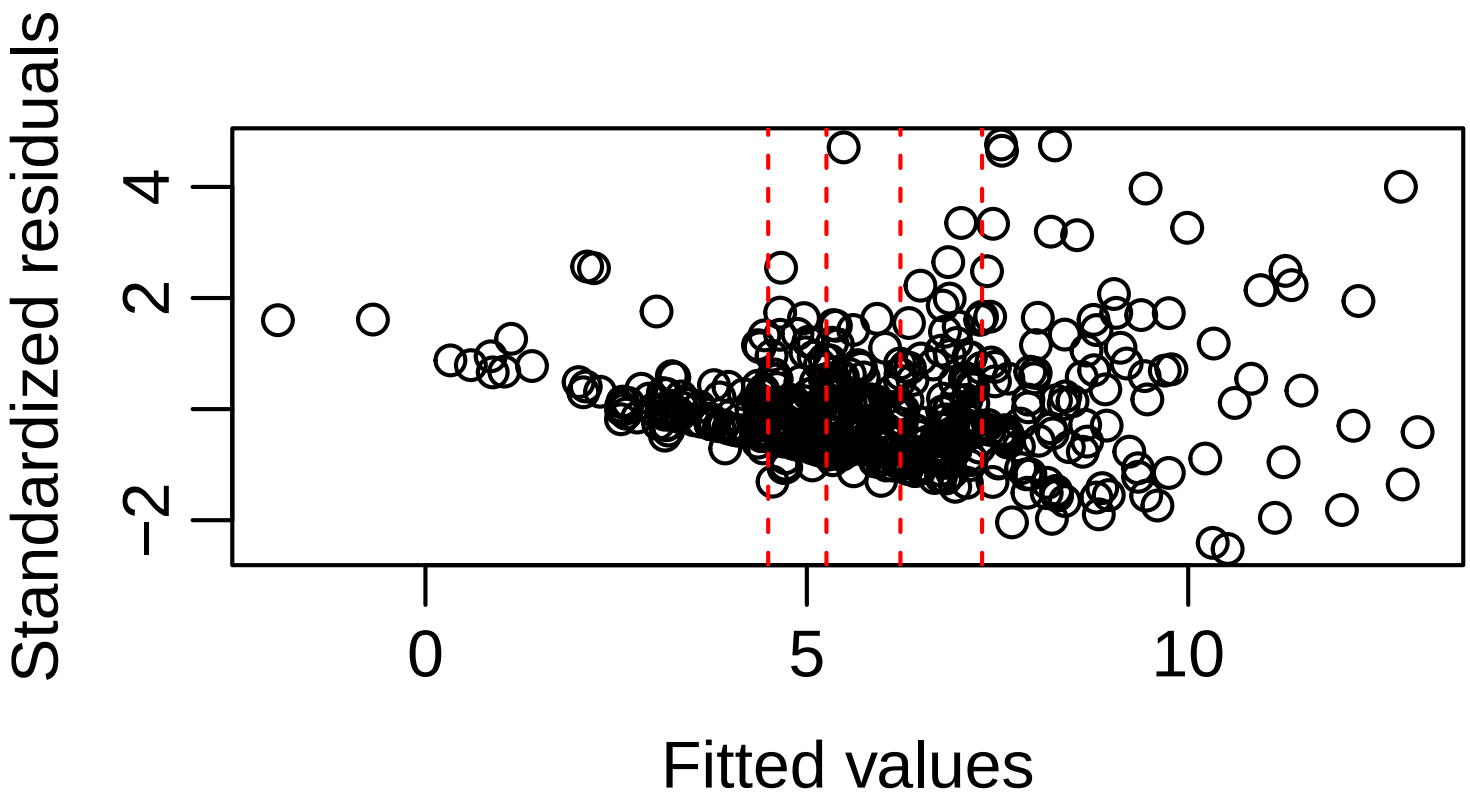
```
## grp
## [-1.93,4.49]  (4.49,5.26]  (5.26,6.23]   (6.23,7.3]     (7.3,13]
##          106           105           105          105          105
```

- The cut function allows you to split the data (here the fitted values) into groups according to the value of the breakpoints

Econometric Example

wage1 dataset

```
plot(fitted(ols), e_std, xlab = "Fitted values",
     ylab = "Standardized residuals")
abline(v = cuts[2:G], lty = 2, col = "red")
```



Econometric Example

wage1 dataset

- By using the function `bartlett.test` and by hand

```
bartlett.test(e_std ~ grp)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  e_std by grp
## Bartlett's K-squared = 151.22, df = 4, p-value < 2.2e-16
```

```
ng  <- tapply(e_std, grp, length)
S2g <- tapply(e_std, grp, var)
n   <- sum(ng)
S2p <- sum((ng - 1) * S2g) / (n - G)

TB <- ((n - G) * log(S2p) - sum((ng - 1) * log(S2g))) /
      (1 + (sum(1/(ng - 1)) - 1/(n - G)) / (3 * (G - 1)))
TB
```

```
## [1] 151.2237
```

```
pchisq(TB, df = G - 1, lower.tail = FALSE)
```

```
## [1] 1.113006e-31
```

Econometric Example
wage1 dataset

- Let's fit a GLS model. In this case we are saying that the variance of the errors is proportional to $|\hat{y}_i|^{2\delta}$ (i.e. the standard deviation is proportional to $|\hat{y}_i|^{\delta}$), with δ a parameter to be estimated

```
pred <- predict(ols)
glsFit <- gls(wage ~ educ + exper + tenure, data = wage1,
          weights = varPower(form = ~ pred), method = "ML")
summary(glsFit)
```

```
## Generalized least squares fit by maximum likelihood
##   Model: wage ~ educ + exper + tenure
##   Data: wage1
##        AIC      BIC    logLik
##   2590.165 2615.757 -1289.083
##
## Variance function:
##  Structure: Power of variance covariate
##  Formula: ~pred
##  Parameter estimates:
##     power
## 0.8036546
##
## Coefficients:
##                 Value Std.Error   t-value p-value
## (Intercept) 1.1086024 0.3790351  2.924801  0.0036
## educ        0.2844943 0.0278539 10.213815  0.0000
## exper       0.0155788 0.0080549  1.934075  0.0536
## tenure      0.1597151 0.0201473  7.927383  0.0000
##
##  Correlation:
##        (Intr) educ   exper
## educ   -0.897
## exper  -0.765  0.549
## tenure  0.073 -0.163 -0.262
##
```

Econometric Example

wage1 dataset

- The results are quite different (especially the estimated coefficients for **educ**)

- If there is no heteroscedasticity we expect similar estimated regression coefficients and standard errors

```
cbind(coef(ols), coef(glsFit))
```

```
##                     [,1]      [,2]
## (Intercept) -2.87273482 1.1086024
## educ         0.59896507 0.2844943
## exper        0.02233952 0.0155788
## tenure       0.16926865 0.1597151
```

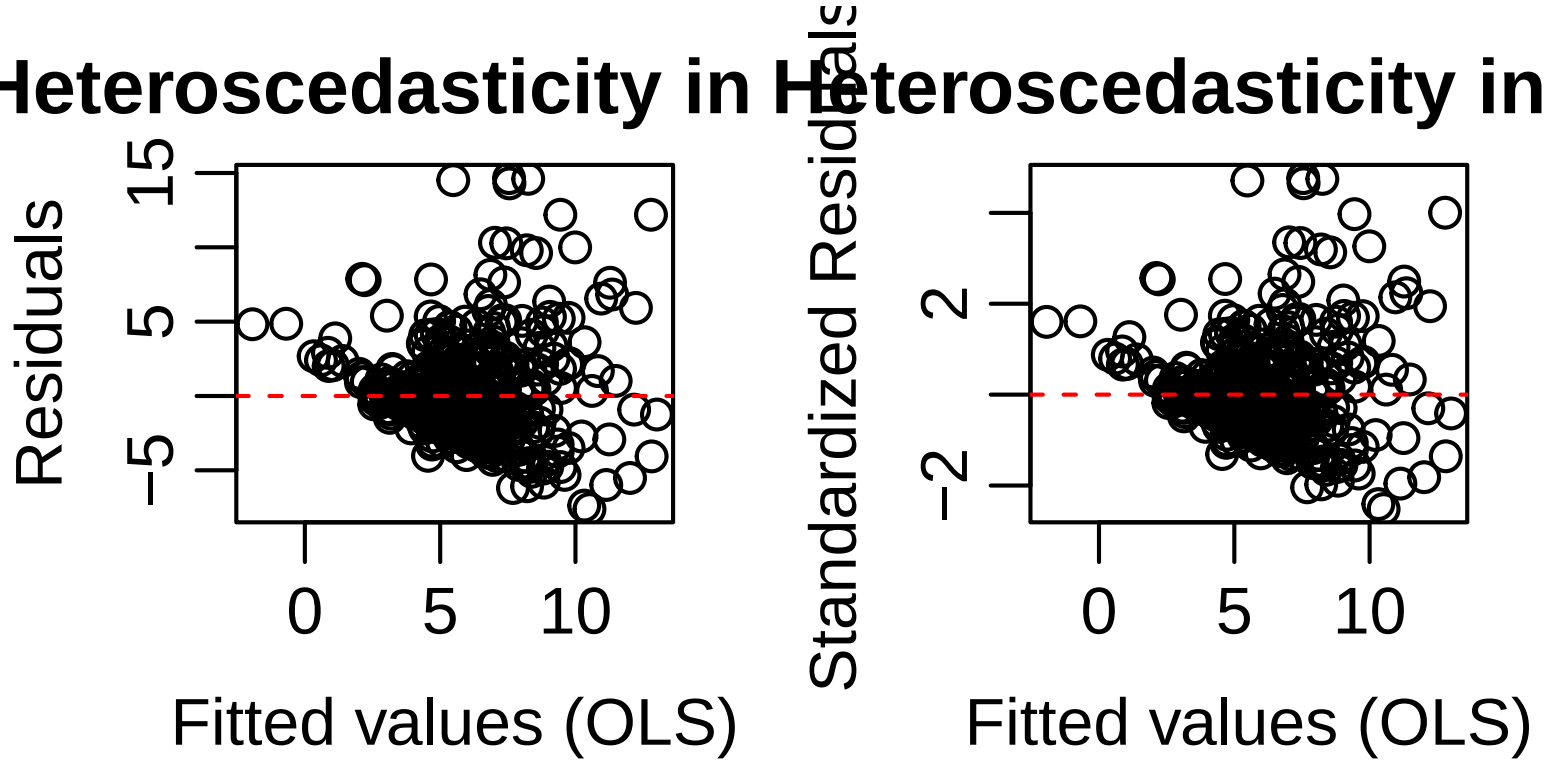
```
cbind(summary(ols)$coefficients[,2], sqrt(diag(glsFit$varBeta)))
```

```
##                    [,1]        [,2]
## (Intercept) 0.72896429 0.379035123
## educ        0.05128355 0.027853877
## exper       0.01205685 0.008054912
## tenure      0.02164461 0.020147272
```

Econometric Example

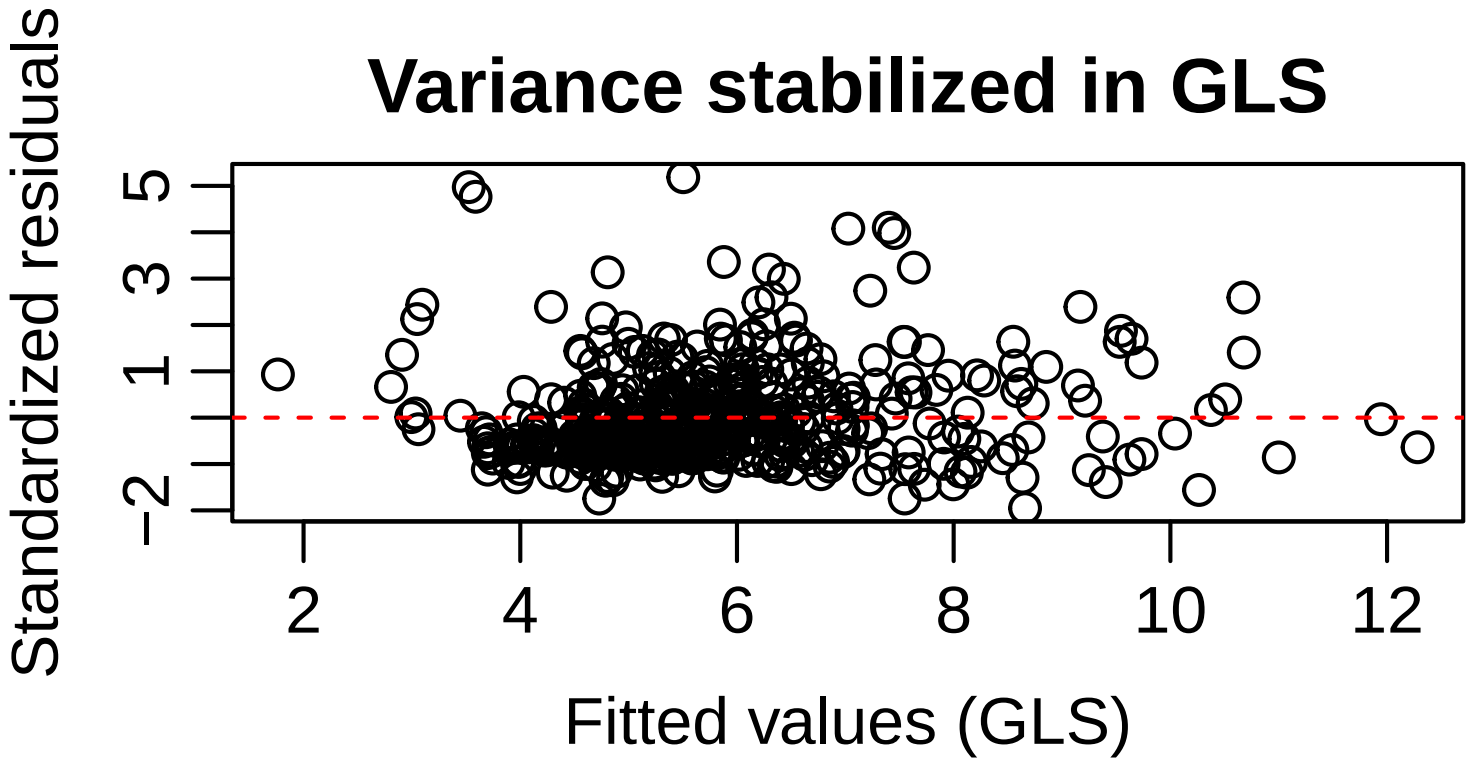
```
par(mfrow=c(1,2))
plot(fitted(ols), residuals(ols), xlab = "Fitted values (OLS)",
     ylab = "Residuals", main = "Heteroscedasticity in OLS")
abline(h = 0, col = "red", lty = 2)

plot(fitted(ols), rstandard(ols), xlab = "Fitted values (OLS)",
     ylab = "Standardized Residuals", main = "Heteroscedasticity in OLS")
abline(h = 0, col = "red", lty = 2)
```



Econometric Example

```
plot(fitted(glsFit), resid(glsFit, type="normalized"),     xlab = "Fitted values (GLS)",
     ylab = "Standardized residuals", main = "Variance stabilized in GLS")
abline(h = 0, col = "red", lty = 2)
```



Econometric Example

- What if we transform the outcome using the logarithm?

```
ols_log <- lm(log(wage) ~ educ + exper + tenure, data = wage1)
bptest(ols_log)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  ols_log
## BP = 10.761, df = 3, p-value = 0.01309
```

```
e_std <- rstandard(ols_log)
G <- 5
cuts <- quantile(fitted(ols_log), probs = seq(0, 1, length.out = G + 1))
grp <- cut(fitted(ols_log), breaks = cuts, include.lowest = TRUE)
bartlett.test(e_std ~ grp)
```

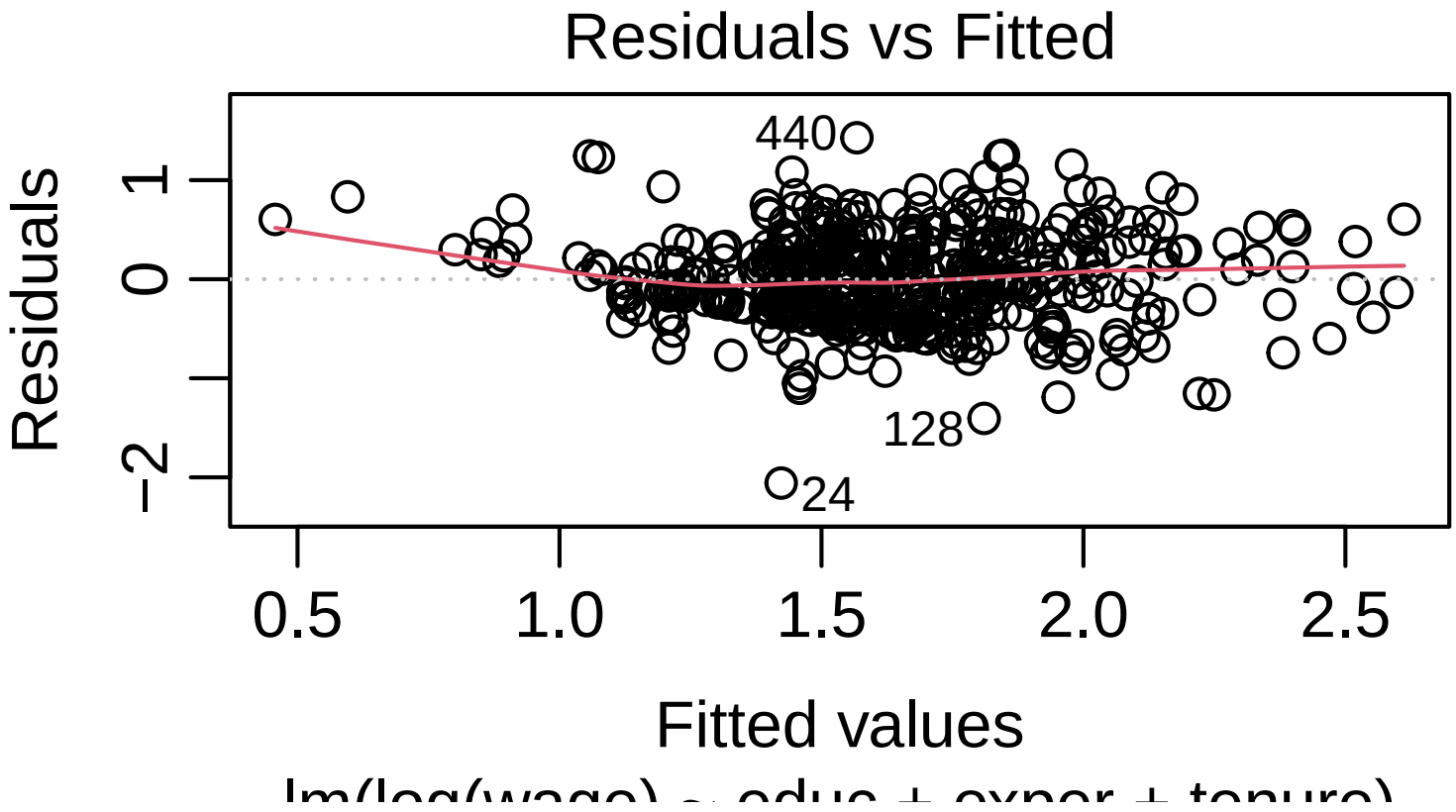
```
##
##  Bartlett test of homogeneity of variances
##
## data:  e_std by grp
## Bartlett's K-squared = 18.474, df = 4, p-value = 0.0009966
```

- The p-value is no longer very close to zero (under both tests), but there is a suspicion that handling heteroscedasticity in the models is still important (but remember, we just tried the logarithmic transformation and considered all the effects as linear: Statistical modeling is an art and a craft.)

Econometric Example

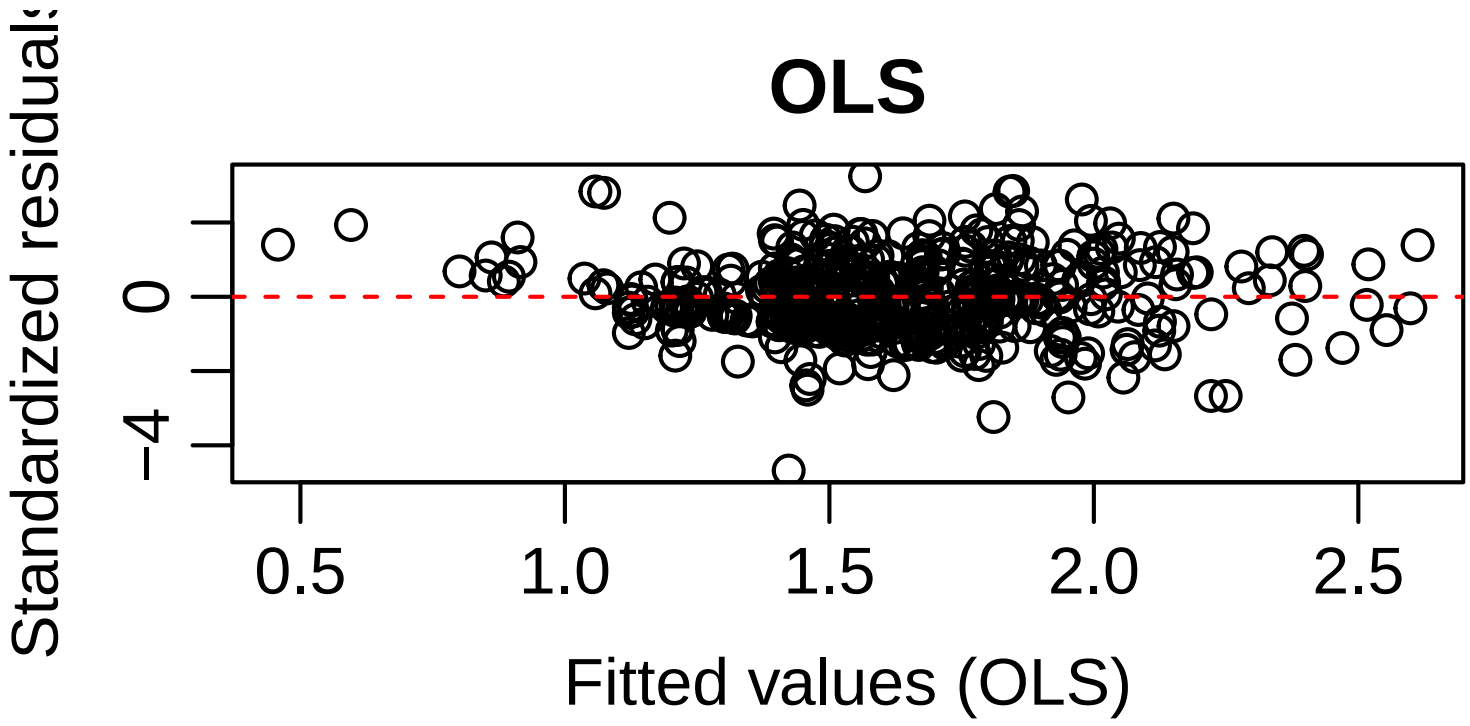
- Residual analysis (fitted vs residuals) does not clearly show a systematic increase of the variance w.r.t. the fitted values

```
plot(ols_log,1)
```



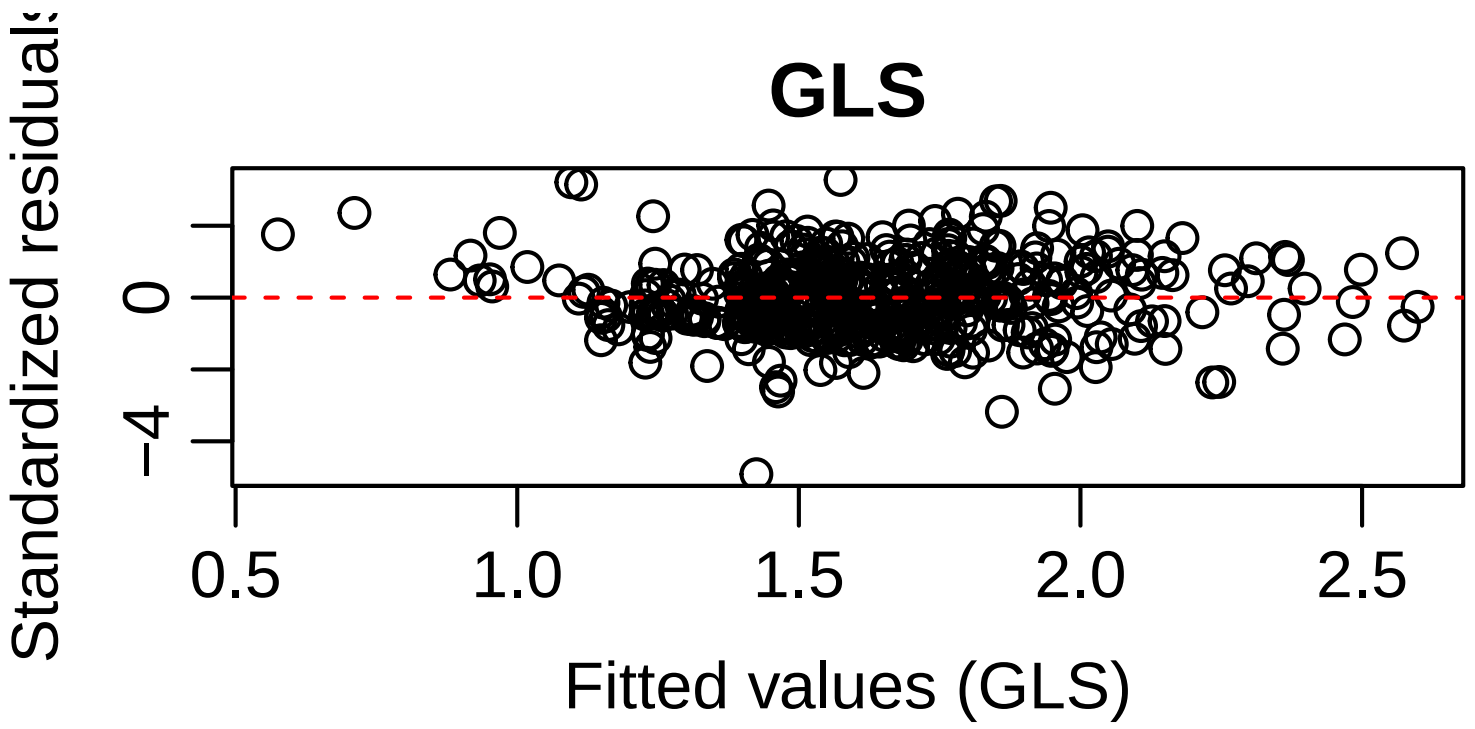
Econometric Example

```
pred <- predict(ols_log)
gls_log <- gls(log(wage) ~ educ + exper + tenure, data = wage1,
          weights = varPower(form = ~ pred), method = "ML")
plot(fitted(ols_log), rstandard(ols_log), xlab = "Fitted values (OLS)",
     ylab = "Standardized residuals", main = "OLS")
abline(h = 0, col = "red", lty = 2)
```



Econometric Example

```
plot(fitted(gls_log), resid(gls_log, type="normalized"),     xlab = "Fitted values (GLS)",
     ylab = "Standardized residuals", main = "GLS")
abline(h = 0, col = "red", lty = 2)
```



Econometric Example

wage1 dataset

- The results are no longer too different (especially the estimated coefficients for **educ**)

```
cbind(coef(ols_log), coef(gls_log))
```

```
##                      [,1]        [,2]
## (Intercept) 0.284359555 0.392450753
## educ        0.092028987 0.083069425
## exper       0.004121109 0.004347204
## tenure      0.022067217 0.022291483
```

```
cbind(summary(ols_log)$coefficients[,2], sqrt(diag(gls_log$varBeta)))
```

```
##                      [,1]        [,2]
## (Intercept) 0.104190378 0.096854820
## educ        0.007329923 0.006896749
## exper       0.001723277 0.001683689
## tenure      0.003093649 0.003227772
```

Linear Model

Comparing the models

- We discussed that we can compare models in terms of model accuracy by analyzing the RSE or the R^2 **but only if we are working in the same scale**

- However, for the GLS models the R^2 is no longer available because the variance decomposition is no longer unique (there are some pseudo R^2)

- We can compare them using the RSE (but only between models having the same scale of the response)

```
cbind(summary(ols)$sigma, summary(glsFit)$sigma)
```

```
##           [,1]      [,2]
## [1,] 3.084476 0.7118385
```

```
cbind(summary(ols_log)$sigma, summary(gls_log)$sigma)
```

```
##           [,1]      [,2]
## [1,] 0.440862 0.3681296
```

Linear Model

Akaike Information Criterion

- It is based on the log-likelihood of the fitted model and a penalization term

$$AIC = 2(p+1) - 2\ell(\hat{\beta}; \hat{\sigma}^2)$$

- In a linear model it is simply (up to an additive constant)

$$AIC = 2(p+1) + n \log \hat{\sigma}^2$$

- To compare models fitted in different scales, you must consider an additive term in the log-likelihood
- **The lower, the better**

Linear Model

Akaike Information Criterion

- The comparison leads to the conclusion that the GLS model is preferred to the OLS (although on the logarithmic scale there is a marginal benefit of adopting GLS)

```
cbind(AIC(ols), AIC(glsFit))
```

```
##          [,1]     [,2]
## [1,] 2683.662 2590.165
```

```
cbind(AIC(ols_log), AIC(gls_log))
```

```
##          [,1]     [,2]
## [1,] 637.0955 633.8292
```

```
AIC(ols_log) + sum(2*log(wage1$wage))
```

```
## [1] 2344.774
```

```
AIC(gls_log) + sum(2*log(wage1$wage))
```

```
## [1] 2341.508
```

Linear Model

Multicollinearity Problem

- It refers to situations in which two or more predictors are closely related to each other.

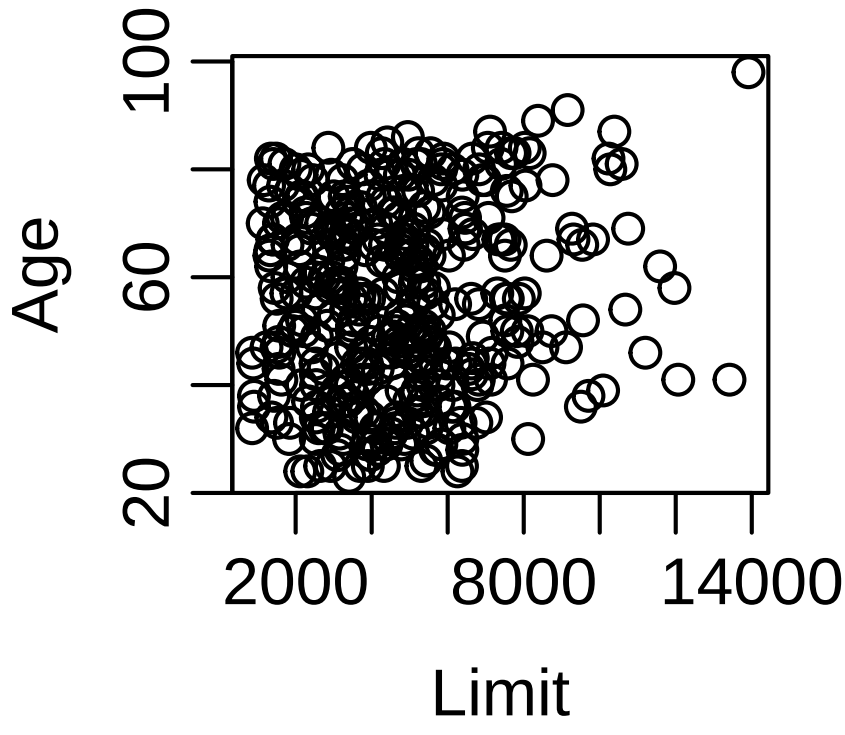
- See the example below

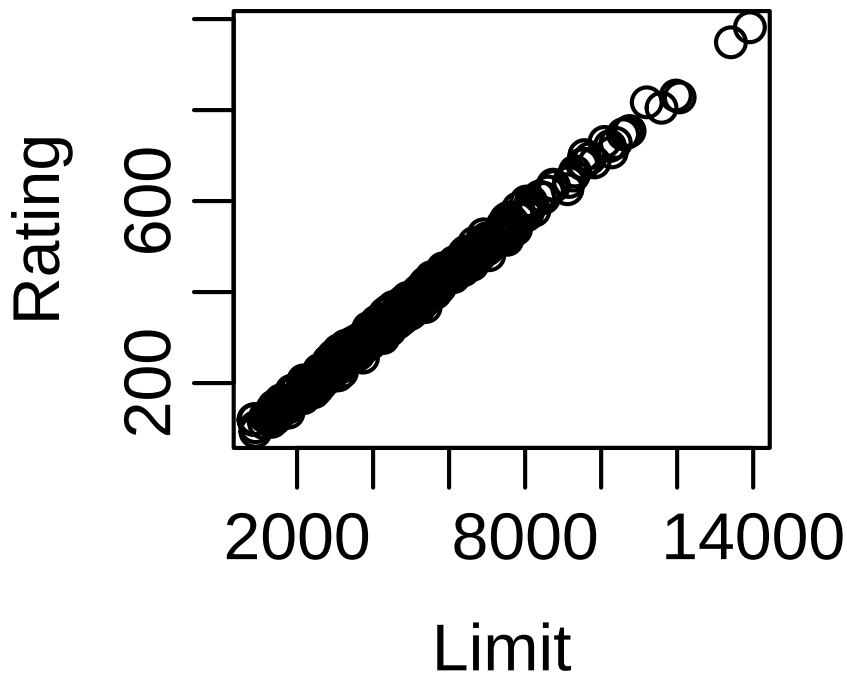
```
library(ISLR2)
data(Credit)
par(mfrow=c(1,2))
with(Credit, plot(Limit, Age))
with(Credit, plot(Limit, Rating))
```

Linear Model

Multicollinearity Problem

- While Limit and Age do not have an obvious relationship, limit and rating are highly correlated (we say that they are collinear)





Linear Model

Multicollinearity Problem

- Collinearity can pose problems in the context of regressions, since it can be difficult to separate individual effects of collinear variables on the response
- In other words, since `limit` and `rating` tend to increase or decrease together, it can be difficult to determine how each one separately is associated with the response (balance)
- Collinearity reduces the accuracy of the estimates: so it causes the standard error of $\hat{\beta}_j$ to grow
- Remember that the t-statistic for testing the nullity of the single coefficients is obtained by dividing the estimate $\hat{\beta}_j$ by its standard error
- Consequently, collinearity results in a decline of the t-statistic and we may fail to reject $H_0 : \beta_j = 0$

Linear Model

Multicollinearity Problem

- We fit the simple linear regression models

```
lmAge <- lm(Balance ~ Age, data = Credit)
summary(lmAge)$coefficients
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 517.2922247  77.851531 6.64459921 1.002280e-10
## Age           0.0489114   1.335991 0.03661057 9.708139e-01
```

```
lmLimit <- lm(Balance ~ Limit, data = Credit)
summary(lmLimit)$coefficients
```

```
##                 Estimate   Std. Error   t value      Pr(>|t|)
## (Intercept) -292.7904955 26.683414516 -10.97275  1.184152e-24
## Limit          0.1716373  0.005066234  33.87867 2.530581e-119
```

```
lmRating <- lm(Balance ~ Rating, data = Credit)
summary(lmRating)$coefficients
```

```
##                Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) -390.84634 29.0685146 -13.44569  3.073181e-34
## Rating         2.56624  0.0750891  34.17594 1.898899e-120
```

Linear Model

Multicollinearity Problem

- The estimated coefficient for `limit` is almost the same when including age in the model, as well as its standard error

```
lmAge <- lm(Balance ~ Age, data = Credit)
summary(lmAge)$coefficients
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 517.2922247  77.851531 6.64459921 1.002280e-10
## Age           0.0489114   1.335991 0.03661057 9.708139e-01
```

```
lmLimit <- lm(Balance ~ Limit, data = Credit)
summary(lmLimit)$coefficients
```

```
##                 Estimate   Std. Error   t value      Pr(>|t|)
## (Intercept) -292.7904955 26.683414516 -10.97275  1.184152e-24
## Limit          0.1716373  0.005066234  33.87867 2.530581e-119
```

```
lmAgeLimit <- lm(Balance ~ Age + Limit, data = Credit)
summary(lmAgeLimit)$coefficients
```

```
##                 Estimate   Std. Error   t value      Pr(>|t|)
## (Intercept) -173.410901 43.828387048 -3.956589  9.005366e-05
## Age           -2.291486  0.672484540 -3.407492  7.226468e-04
## Limit          0.173365  0.005025662 34.495944 1.627198e-121
```

Linear Model
Multicollinearity Problem

- Instead, there is a drastic change in both the estimated coefficients and the associated standard errors (increase of 12 times $\implies$ p-value of 0.7)
- The importance of `limit` is masked due to the presence of collinearity

```
lmRating <- lm(Balance ~ Rating, data = Credit)
summary(lmRating)$coefficients
```

```
##               Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) -390.84634 29.0685146 -13.44569  3.073181e-34
## Rating         2.56624  0.0750891  34.17594 1.898899e-120
```

```
lmLimit <- lm(Balance ~ Limit, data = Credit)
summary(lmLimit)$coefficients
```

```
##                Estimate   Std. Error   t value      Pr(>|t|)
## (Intercept) -292.7904955 26.683414516 -10.97275  1.184152e-24
## Limit          0.1716373  0.005066234  33.87867 2.530581e-119
```

```
lmLimitRating <- lm(Balance ~ Limit + Rating, data = Credit)
summary(lmLimitRating)$coefficients
```

```
##                 Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept) -377.53679536 45.25417619 -8.3425846 1.213565e-15
## Limit          0.02451438  0.06383456  0.3840298 7.011619e-01
## Rating         2.20167217  0.95229387  2.3119672 2.129053e-02
```

Linear Model

Multicollinearity Problem

- A simple way to detect a potential problem of collinearity is to look at the correlation matrix

- Unfortunately, not all the problems related to collinearity can be detected by analyzing the correlation matrix: it is possible for collinearity to exist between three or more variables, even if no pair of variables has a particularly high correlation (we call this situation **multicollinearity**)

```
cor(Credit[, c("Age", "Limit", "Rating")])
```

```
##                Age     Limit    Rating
## Age    1.0000000 0.1008879 0.1031650
## Limit  0.1008879 1.0000000 0.9968797
## Rating 0.1031650 0.9968797 1.0000000
```

Linear Model

Multicollinearity Problem

- An alternative is to explore the **Variance Inflation Factor (VIF)**: the ratio between the variance of $\hat{\beta}_j$ when fitting the full model and the variance of $\hat{\beta}_j$ when fitting the model on its own

- The lowest possible value is 1 and values of 5/10 indicate a problematic amount of multicollinearity

- In this case, the presence of multicollinearity is clear (values of 160)

```
lmFull <- lm(Balance ~ Age + Limit + Rating, data = Credit)
library(car)
vif(lmFull)
```

```
##        Age      Limit     Rating
##   1.011385 160.592880 160.668301
```

Linear Model

Multicollinearity Problem

- Possible solutions to deal with this problem

1. Removing one of the variables (for instance Rating)
2. Combining the predictors into a single one
3. Using regularization techniques (we do not see them)

```
vif(lmAgeLimit)
```

```
##       Age     Limit
## 1.010283 1.010283
```

Linear Model

Endogeneity

- The unbiasedness and the consistency of the OLS estimator rest on the hypothesis that the conditional expectation of the error is constant (and can be set to zero if the model contains an intercept)

- Consider the simple linear model:

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

with

$$\begin{aligned}\mathbb{E}(Y_i|x_i) &= \mathbb{E}[\beta_1 + \beta_2 x_i + \varepsilon_i|x_i] \\ &= \beta_1 + \beta_2 x_i + \mathbb{E}[\varepsilon_i|x_i] = \beta_1 + \beta_2 x_i\end{aligned}$$

if $\mathbb{E}(\varepsilon_i|x_i) = 0$

- When x is correlated with ε, we face **endogeneity**

- **Endogeneity** leads to **biased** and **inconsistent** OLS estimates

- It is common in observational data

Linear Model

Why endogeneity matters

- The same property can also be described using the covariance between the covariate and the error that can be written, using the rule of repeated expectation (tower property):

$$\begin{aligned}\text{cov}(x, \varepsilon) &= \mathbb{E}[(x - \mu_x)\varepsilon] = \mathbb{E}_x[\mathbb{E}_\varepsilon[(x - \mu_x)\varepsilon|x]] \\ &= \mathbb{E}_x[(x - \mu_x)\mathbb{E}_\varepsilon[\varepsilon|x]]\end{aligned}$$

- **Key consequence:**

$$\text{cov}(x, \varepsilon) \neq 0 \Rightarrow \hat{\beta}_{2,\text{OLS}} \xrightarrow{p} \beta_2 + \frac{\text{cov}(x, \varepsilon)}{\text{var}(x)}$$

where $\xrightarrow{p}$ stands for convergence in probability

- Indeed, considering a simple linear regression model

$$\begin{aligned}\hat{\beta}_{2,\text{OLS}} &= \frac{cov(x, y)}{var(x)} = \frac{cov(x, \beta_1 + \beta_2 x + \varepsilon)}{var(x)} \\ &= \beta_2 + \frac{cov(x, \varepsilon)}{var(x)}\end{aligned}$$

Linear Model

Why endogeneity matters

- Cases that we could encounter in applications:

1. if $cov(x, \varepsilon) > 0 \implies$ OLS overestimates β_2
2. if $cov(x, \varepsilon) < 0 \implies$ OLS underestimates β_2
3. if $cov(x, \varepsilon) = 0 \implies$ OLS is consistent (converges to β_2)

- OLS incorrectly attributes variation in y to x

- Standard errors are no longer meaningful

- Inference becomes unreliable