Intermediate Econometrics

12-14th 2025 - Vincenzo Gioia

Linear Model

Limits and extension

- Although a powerful tool to explore the relationship between the covariates and the outcome, the linear model is based on the assumptions:

1. **Linearity**: The expected value of Y is a linear function of the explanatory variables $\mathrm{E}(Y_i) = \mu_i = \mathbf{x}_i^\top \boldsymbol{\beta}, \quad i = 1, \ldots, n$
2. **Normality**: The variables Y_i have distribution $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$ (for the errors $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$)
3. **HOmoschedasticity**: The variance of Y_i (or the error term) is not dependent on i
4. **Independence**: Y_i is independent on Y_j for each couple i, j with $i \neq j$ (equivalently in terms of errors)
5. **Linear independence between explanatory variables**: the model matrix X is not stochastic and of full rank (p)

- If the assumptions are not aligning with the data, model-based inference can be misleading

Linear Model

Linearity

- If we have evidence of non-linearity we can explore the following options
- Transforming the predictors $Y_i = \beta_1 + \beta_2 g(x_{i2}) + \ldots + \beta_p g(x_{ip}) + \varepsilon_i$

1. Several transformations (log, square root, inverse, …)
2. Polynomial regression

- Transforming the outcome: $f(Y_i) = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i^*$

1. For instance $f(Y_i) = \log(Y_i)$; in such a case we are assuming that $\log(Y_i)$ (and ε_i^*) are normally distributed, so Y_i and $\varepsilon_i = \exp(\varepsilon_i^*)$ are distributed accordingly to he lognormal distribution

- Transform either the predictors and the outcome

Interpretation

- We are loosing the simplicity of the interpretation (the simple interpretation of the linear model coefficients must be done in the transoformed scale)

Linear Model

Normality

- Also without assuming the normality of errors, the OLS estimator has good properties
- However, without the normality assumption, the inferential results (confidence intervals and hypothesis test) cannot be obtained easily
- Sometimes, we try to recover the normality assumption via transformations, for instance $\log()$

1. A very popular transformation is the Box-Cox transformation, which includes the logarithmic transformation as a special case (we do not see)
2. Considering a different class of models: Generalized Linear Models (GLMs), that also includes the normal linear model, where we assume that the response variable is distributed according to a certain distribution (we will see some of them)

Linear Model

Homoschedasticity

- Let's suppose that we are in a case where $V(\varepsilon_i) = V(Y_i) = \sigma_i^2$, for $i = 1, \ldots, n$: the variance is not constant $\implies$ **heteroschedasticity**

1. The OLS estimator $\hat{\beta}$ has mean β (unbiased)
2. The OLS estimator $\hat{\beta}$ is no more efficient; indeed its variance covariance matrix is

$$V(\hat{\beta}) = V((X^\top X)^{-1} X^\top Y) = (X^\top X)^{-1} X^\top V(Y) X (X^\top X)^{-1}$$

3. The distribution is still normal $\hat{\beta} \sim \mathcal{N}(\beta, V(\hat{\beta}))$

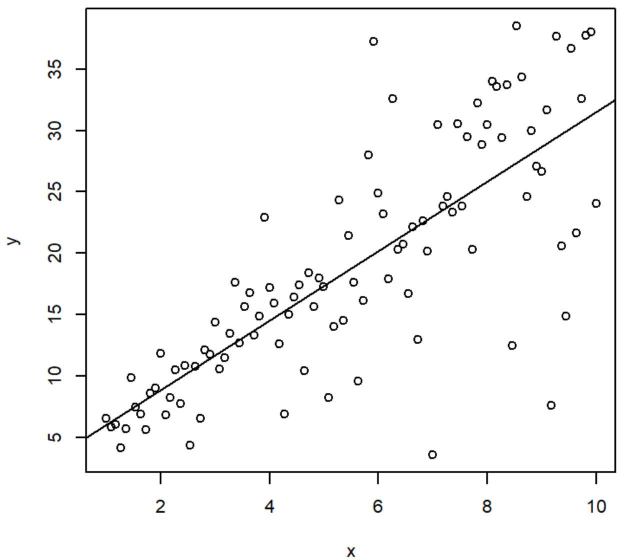
Problems

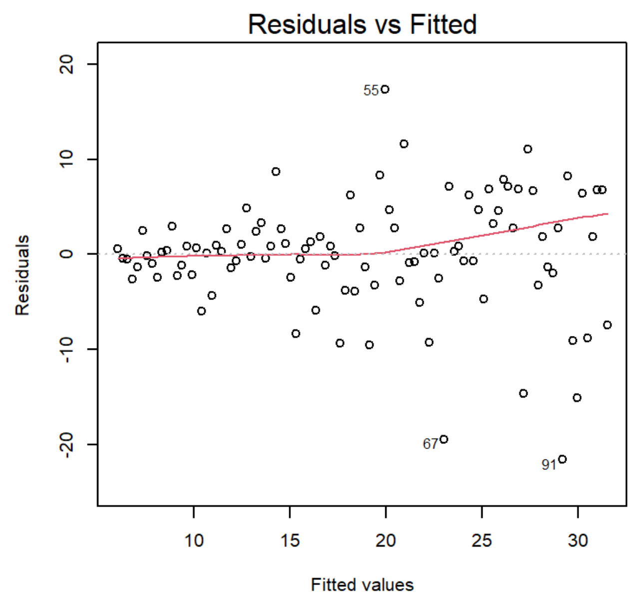
- Despite the normality, the heteroschedasticity implies that the usual procedures for obtaining hypothesis test and confidence intervals are no more valid

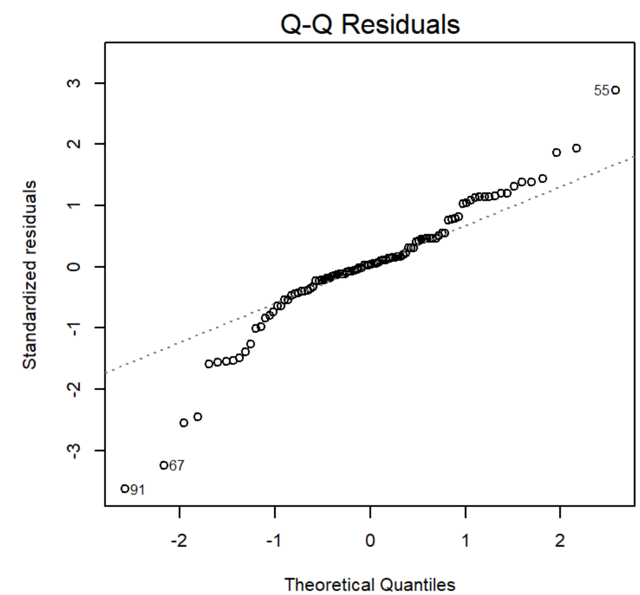
Linear Model

How to deal with eteroschedasticity

1. Trasforming the outcome (there are some trasformations that are able to stabilize the variance, e.g. square root, logarithm, inverse square root, arcsin of the square root). Consider the following data



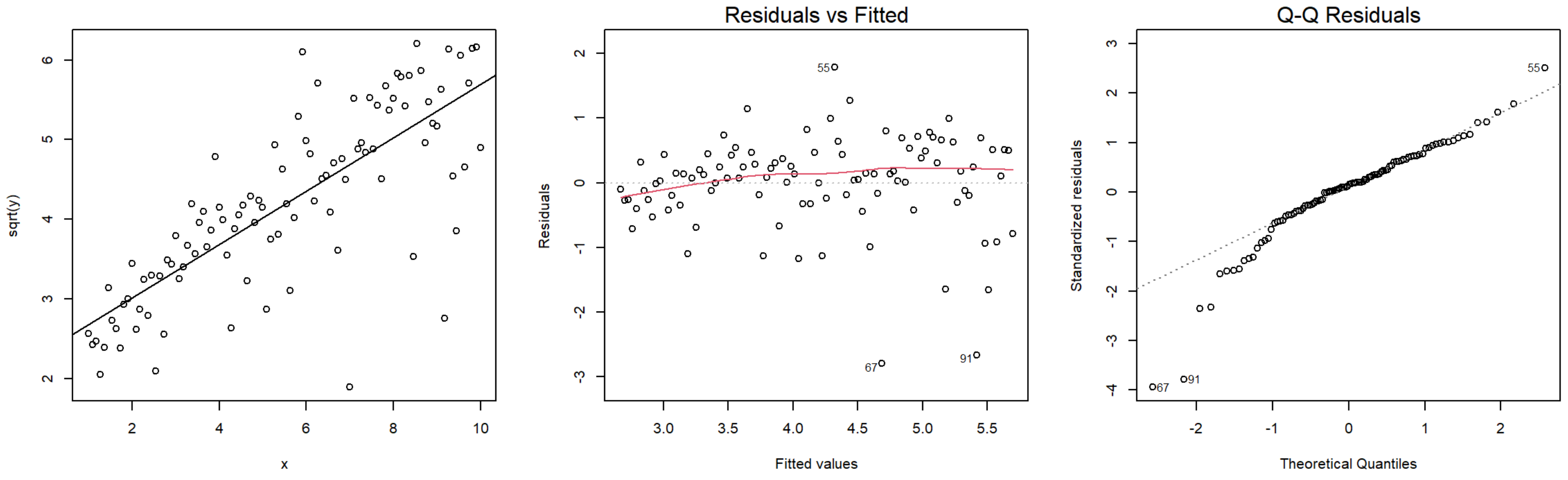




Linear Model

How to deal with eteroschedasticity

1. Trasforming the outcome: Consider the square root



Linear Model

Heteroschedasticity

2. Change the estimation approach: GLS (generalized least square)

- The model structure remains the same:

$$Y = X\beta + \varepsilon$$

but we are substituting the homoschedasticity ($V(\varepsilon) = \sigma^2 I$) assumption with

$$V(\varepsilon) = \sigma^2 \Omega$$

where Ω is a known diagonal matrix (obviously not the identity)

Linear Model

Heteroschedasticity

2. Change the estimation approach: GLS (generalized least square)

- The log-likelihood for (β, σ^2) is

$$\ell(\beta, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y - X\beta)^\top\Omega^{-1}(y - X\beta)$$

- The maximum likelihood estimate of β is

$$\hat{\beta} = \arg\min_{\beta}\,(y - X\beta)^\top\Omega^{-1}(y - X\beta) = (X^\top\Omega^{-1}X)^{-1}X^\top\Omega^{-1}y$$

- So, it is easy to derive $V(\hat{\beta})$ ($\hat{\beta}$ is still normally distributed)
- Since Ω is diagonal, by using the generalized least squares we are minimizing the function

$$RSS_g(\beta) = \sum_{i=1}^{n}\frac{1}{w_{ii}}(y_i - x_i^\top\beta)^2$$

where ω_{ii} is the i-th diagonal element of Ω

Linear Model

Heteroschedasticity

2. Change the estimation approach: GLS (generalized least square)

- In summary, the contribution to the sum of squares are weighted by $1/\omega_{ii}$: greater is ω_{ii}, that is greater is the variance of the error ε_i for the i-th observations, and lower is the weight associated to the contribution of the sum
- Note: If Ω is not diagonal, the generalized least squares can be used to deal with error's dependence, in addition to the heteroschedasticity
- The case illustrated above belongs to the class of **non-spherical disturbance errors**

Econometric Example

wage1 dataset

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

1. detect heteroschedasticity (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 using Generalized Least Squares (GLS).

- Estimate a linear regression:

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

- Testing for heteroschedasticity, fit via OLS and GLS, then considering 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

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

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

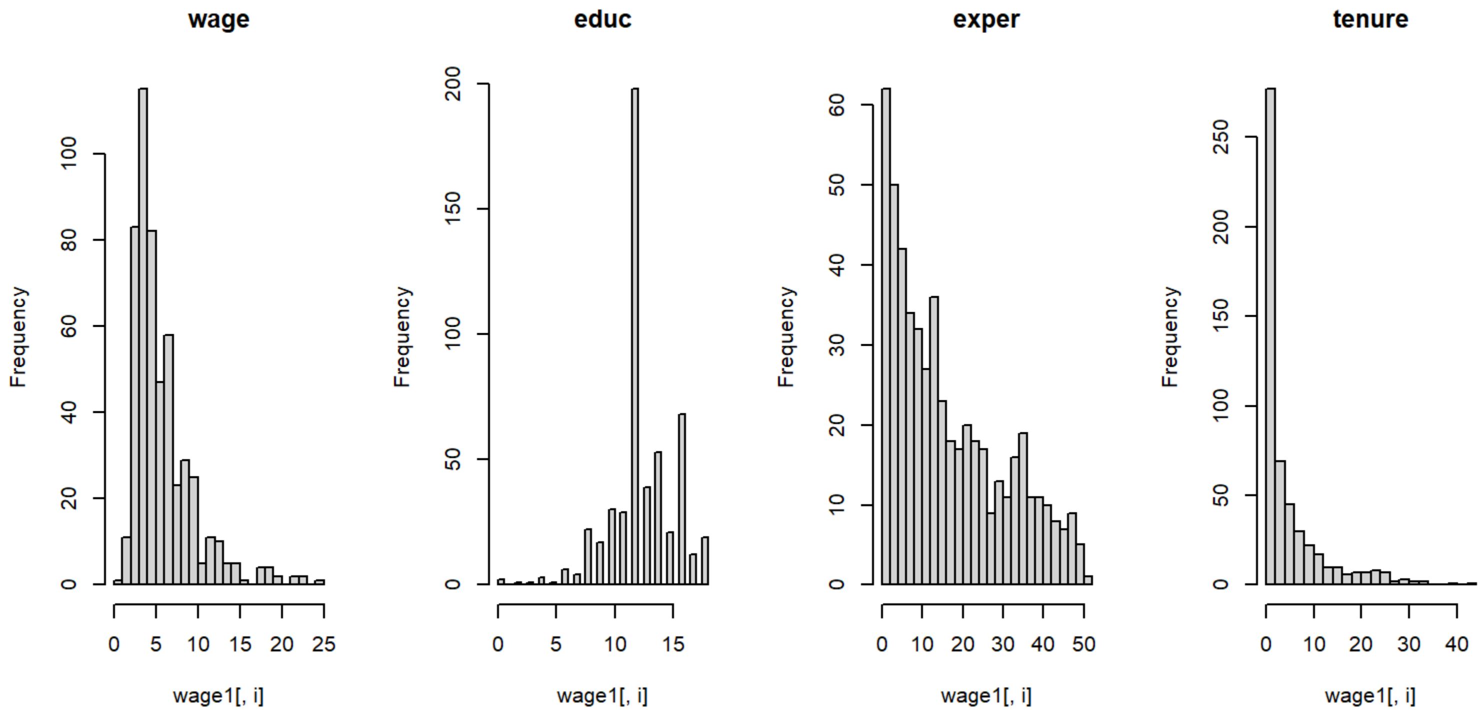
```
'data.frame':	526 obs. of  24 variables:
 $ wage    : num  3.1 3.24 3 6 5.3 ...
 $ educ    : int  11 12 11 8 12 16 18 12 12 17 ...
 $ exper   : int  2 22 2 44 7 9 15 5 26 22 ...
 $ tenure  : int  0 2 0 28 2 8 7 3 4 21 ...
 $ nonwhite: int  0 0 0 0 0 0 0 0 0 0 ...
 $ female  : int  1 1 0 0 0 0 0 1 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 ...
```

Econometric Example

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

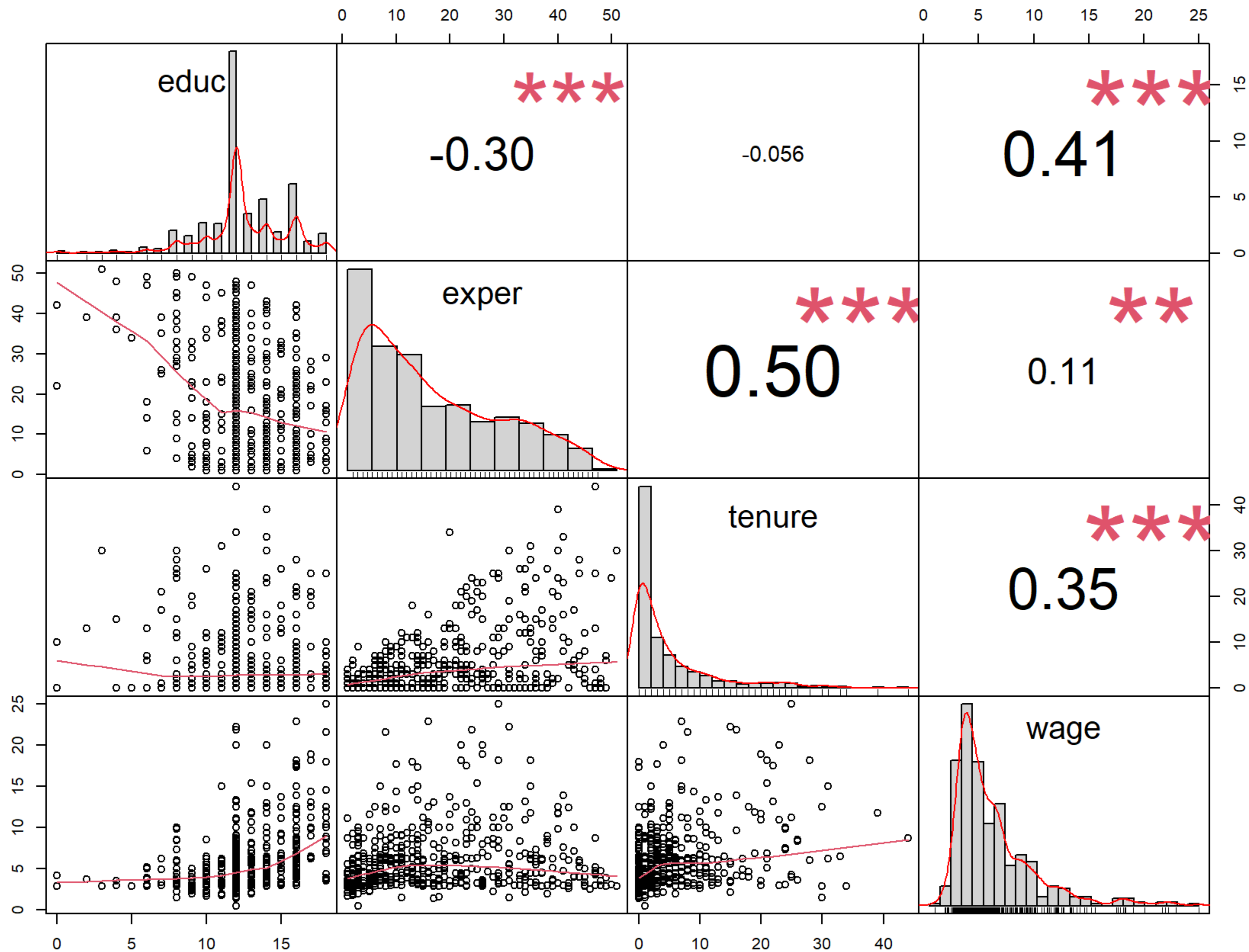
```
      wage             educ            exper           tenure      
 Min.   : 0.530   Min.   : 0.00   Min.   : 1.00   Min.   : 0.000  
 1st Qu.: 3.330   1st Qu.:12.00   1st Qu.: 5.00   1st Qu.: 0.000  
 Median : 4.650   Median :12.00   Median :13.50   Median : 2.000  
 Mean   : 5.896   Mean   :12.56   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  
```

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


wage
educ
exper
tenure
Frequency
wage1[, i]

Econometric Example

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



educ
exper
tenure
wage
-0.30
***
-0.056
0.41
***
0.50
***
0.11
**
0.35
***

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 positive but a lower effect

```r
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 ***
```

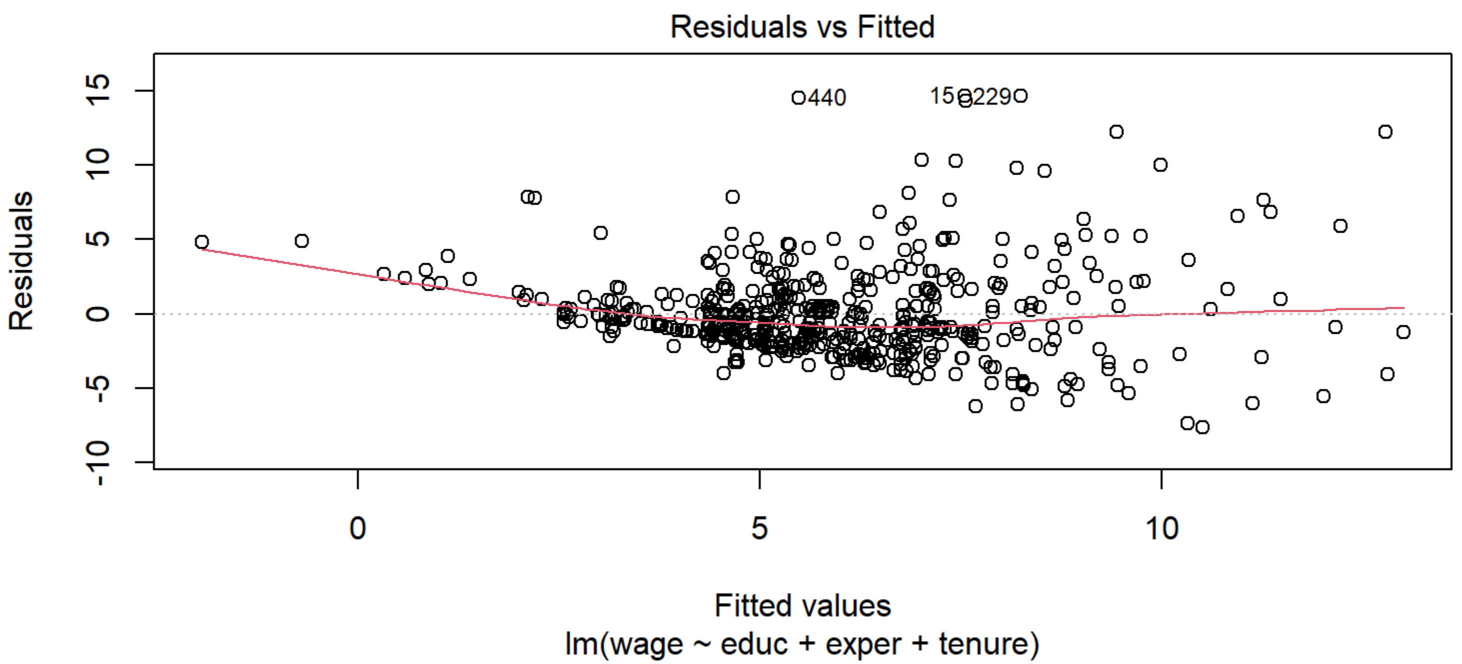
```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Econometric Example

wage1 dataset

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

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



Econometric example

Testing for heteroschedasticity (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:

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

where

1. f is a n-dimensional vector composed by $(e_i^2 / \hat{\sigma}^2 - 1)$, 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 $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2 = \sigma^2$ (homoschedasticity), distributed according to a Chi-square distribution with J degrees of freedom

Econometric example

wage1 dataset

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

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

```
	studentized Breusch-Pagan test

data:  ols
BP = 43.096, df = 3, p-value = 2.349e-09
```

Econometric example

wage1 dataset

- Let's fit a GLS model, the R function is gls (in addition to the model and the data formula you must provide the weights)
- In this case we are saying that the variance of the errors is proportional to $\hat{y}_i^{\delta}$, with δ a parameter to be estimated (you have several options)

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

```
Generalized least squares fit by REML
  Model: wage ~ educ + exper + tenure
  Data: wage1
       AIC      BIC    logLik
  2612.449 2637.995 -1300.225

Variance function:
 Structure: Power of variance covariate
 Formula: ~pred
 Parameter estimates:
    power
```

```
0.7879627

Coefficients:
                Value Std.Error   t-value p-value
(Intercept) 1.0550929 0.3850044  2.740470  0.0063
educ        0.2887735 0.0282548 10.220328  0.0000
```

Econometric Example

wage1 dataset

- The results are quite different (especially the estimated coefficients for **educ**)
- If there is no heteroschedasticity we expect similar estimated regression coefficients and standard errors
- In the following slide, the residuals are plotted

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

```
                   [,1]       [,2]
(Intercept) -2.87273482 1.05509285
educ         0.59896507 0.28877352
exper        0.02233952 0.01607207
tenure       0.16926865 0.15947728
```

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

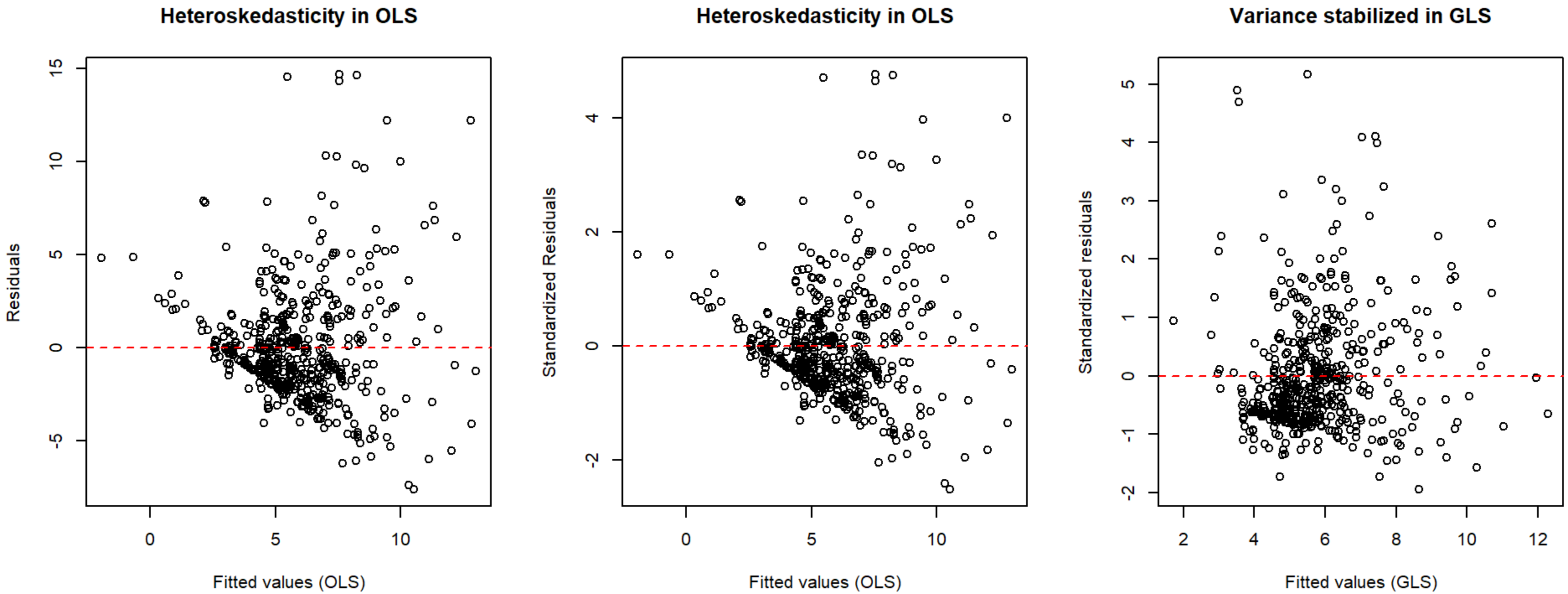
```
                  [,1]        [,2]
(Intercept) 0.72896429 0.385004357
educ        0.05128355 0.028254819
exper       0.01205685 0.008164635
tenure      0.02164461 0.020268395
```

Econometric Example

```r
par(mfrow=c(1,3))
plot(fitted(ols), residuals(ols), xlab = "Fitted values (OLS)",
     ylab = "Residuals", main = "Heteroskedasticity in OLS")
abline(h = 0, col = "red", lty = 2)

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

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


Heteroskedasticity in OLS
Residuals
Fitted values (OLS)
Heteroskedasticity in OLS
Standardized Residuals
Fitted values (OLS)
Variance stabilized in GLS
Standardized residuals
Fitted values (GLS)

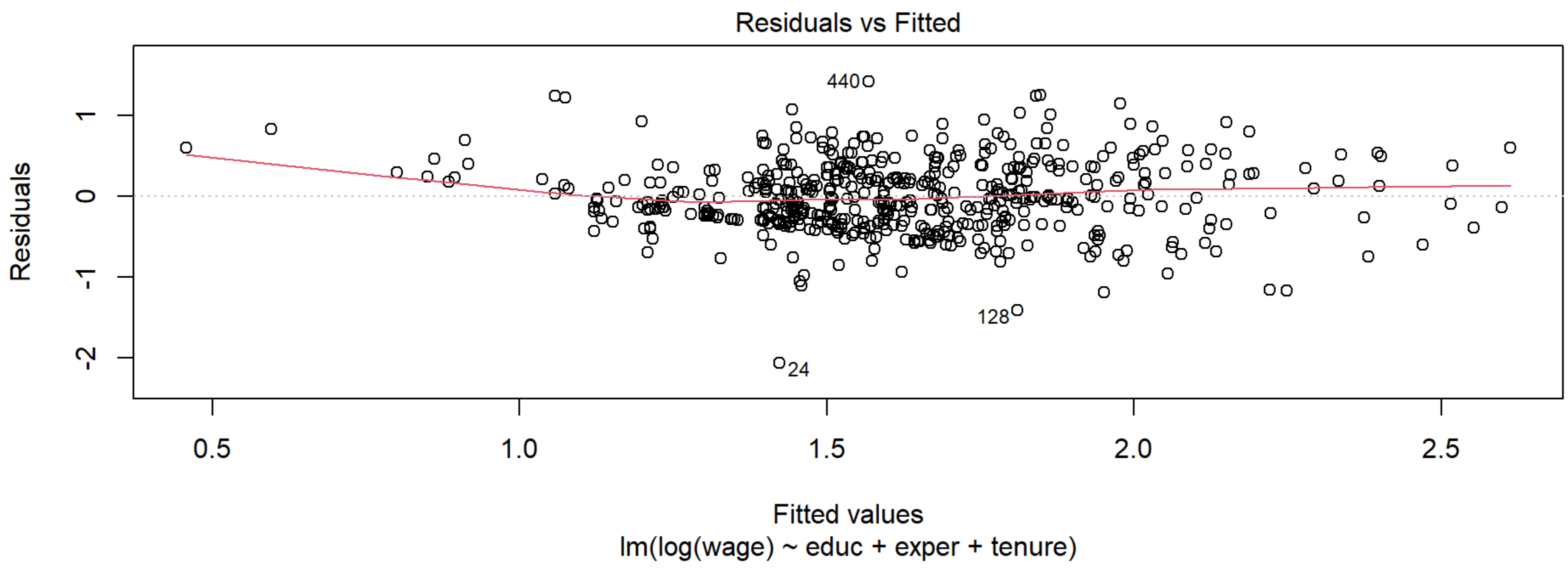
Econometric Example

```r
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
```

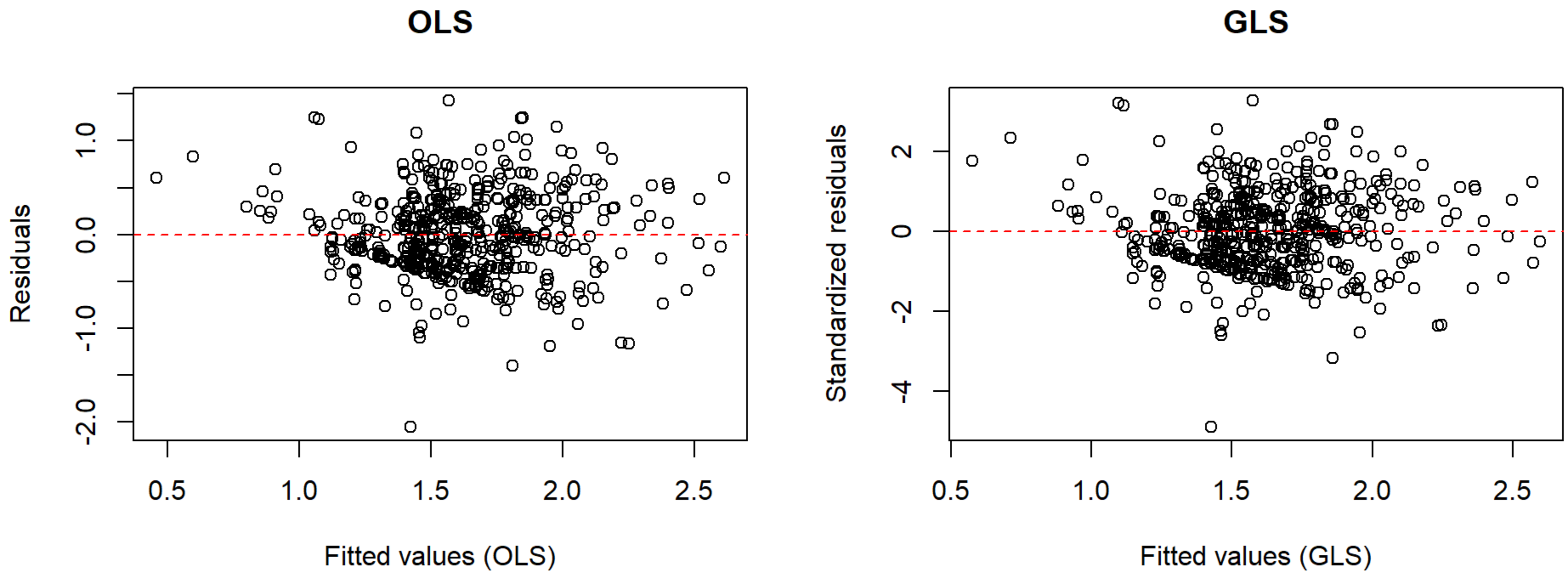
```r
plot(ols_log,1)
```



Econometric Example

```r
pred <- predict(ols_log)
gls_log <- gls(log(wage) ~ educ + exper + tenure, data = wage1,
          weights = varPower(form = ~ pred))
par(mfrow=c(1,2))
plot(fitted(ols_log), residuals(ols_log), xlab = "Fitted values (OLS)",
     ylab = "Residuals", main = "OLS")
abline(h = 0, col = "red", lty = 2)

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


OLS
Residuals
-2.0
-1.0
0.0
1.0
0.5
1.0
1.5
2.0
2.5
Fitted values (OLS)
GLS
Standardized residuals
-4
-2
0
2
0.5
1.0
1.5
2.0
2.5
Fitted values (GLS)

Econometric Example

wage1 dataset

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

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

```
                  [,1]        [,2]
(Intercept) 0.284359555 0.391014486
educ        0.092028987 0.083188829
exper       0.004121109 0.004344837
tenure      0.022067217 0.022288778
```

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

```
                  [,1]        [,2]
(Intercept) 0.104190378 0.096951203
educ        0.007329923 0.006902781
exper       0.001723277 0.001684173
tenure      0.003093649 0.003226156
```

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 more unique (there exists some pseudo R^2)
- We can compare them using the RSE (but only within the model having the same scale of the answer)

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

```
         [,1]      [,2]
[1,] 3.084476 0.7339895
```

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

```
         [,1]      [,2]
[1,] 0.440862 0.3702997
```

Linear Model

Akaike Information Criteria

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

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

- In a linear model it is simply

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

- To compare models fitted in different scales, you must consider an additive term in the log-likelihood
- Lower is the best

Linear Model

Akaike Information Criteria

- The comparison

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

```
         [,1]     [,2]
[1,] 2683.662 2612.449
```

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

```
         [,1]     [,2]
[1,] 637.0955 669.0064
```

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

```
[1] 2344.774
```

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

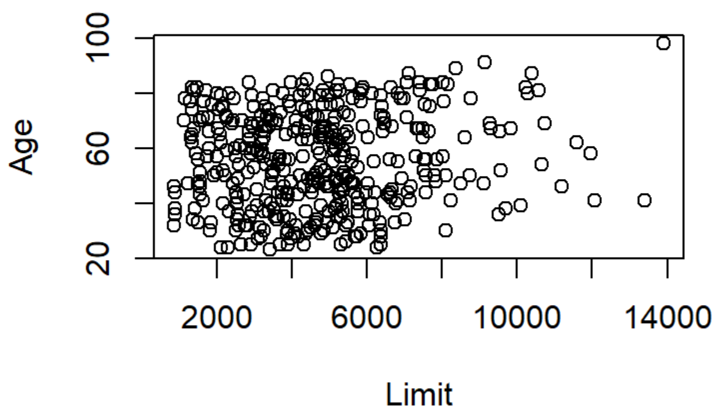
```
[1] 2376.685
```

Linear Model

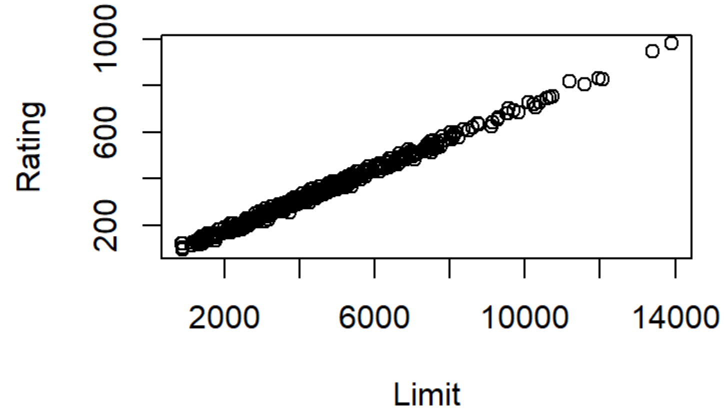
Multicollinearity Problem

- It refers to situations in which two or more predictors are closely related to one other.
- See the example below: while Limit and Age have not an obvious relationship, limit and rating are highly correlated (we say that they are collinear)

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



Age
20
60
100
2000
6000
10000
14000
Limit



Rating
200
600
1000
2000
6000
10000
14000
Limit

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 togheter , 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 of $\hat{\beta}_j$ for 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

```r
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
```

```r
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
```

```r
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 similar when including age in the model, as well as its standard error

```r
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
```

```r
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
```

```r
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

- Here we can see that there is a drastic change on both the estimated coefficients and they associated standard errors (increase of 12 times implying a p-value of 0.7)
- The importance of the limit variable 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 see the correlation matrix
- Unfortunately, not all the problems related with collinearity can be detected by analyzing the correlation matrix: it is possible for collinearity to exist between three or more variable, even if no pair of variables has a particularly high correlation (we call this situation **multicollinearity**)

```r
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 Inflaction 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 lower value is 1 and values of 5/10 indicates a problematic amount of multicollinearity
- In this case, it is clear the presence of multicollinearity (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. Use regularization techniques (we do not see)

```r
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

$$\mathrm{E}(Y_i | x_i) = \mathrm{E}[\beta_1 + \beta_2 x_i + \varepsilon_i | x_i] = \beta_1 + \beta_2 x_i + \mathrm{E}[\varepsilon_i | x_i] = \beta_1 + \beta_2 x_i$$

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

- When x is correlated with ε, we face **endogeneity**
- Leads to **biased** and **inconsistent** OLS estimates
- 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's property):

$$\mathrm{cov}(x, \varepsilon) = \mathrm{E}[(x - \mu_x)\varepsilon] = \mathrm{E}_x[\mathrm{E}_\varepsilon[(x - \mu_x)\varepsilon \,|\, x]] = \mathrm{E}_x[(x - \mu_x)\mathrm{E}_\varepsilon[\varepsilon \,|\, x]]$$

- **Key consequence:**

$$\mathrm{cov}(x, \varepsilon) \neq 0 \Rightarrow \overset{\wedge}{\beta}_{2_{OLS}} \overset{p}{\rightarrow} \beta_2 + \frac{\mathrm{cov}(x, \varepsilon)}{\mathrm{var}(x)}$$

  where $\overset{p}{\rightarrow}$ stay for convergence in probability

- Indeed, considering a simple linear regression model

$$\overset{\wedge}{\beta}_{2_{OLS}} = \frac{cov(x, y)}{var(x)} = \frac{cov(x, \beta_1 + \beta_2 x + \varepsilon)}{var(x)} = \beta_2 + \frac{cov(x, \varepsilon)}{var(x)}$$

Linear Model

Why endogeneity matters

- Cases that we could encounter in application:
  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 no longer meaningful
- Inference becomes unreliable

Linear Model

Endogeneity

- If the conditional expectation of ε is a constant, $\mathrm{E}_{\varepsilon}[\varepsilon \,|\, x] = \mu_{\varepsilon}$ (not necessarily 0), the covariance is

$$cov(x, \varepsilon) = \mu_{\varepsilon}\mathrm{E}_{x}[x - \mu_{x}] = 0$$

- Stated in a different way, x is supposed to be exogenous, or x is assumed to be uncorrelated with ε
- Endogeneity when $cov(x, \varepsilon) \neq 0$
- Sources of endogeneity:

1. There are errors in the variables
2. There are omitted variables
3. Simultaneity bias

Linear Model

1. Errors in the variables (outcome)

- Data used in economics, especially micro-data, are prone to errors of measurement (either outcome and predictors)
- Suppose that the model that we seek to estimate is

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

  where the covariates is exogenous ($cov(x^*, \varepsilon^*) = 0$)
- Suppose that the response is observed with error, namely that the observed value of the response is

$$y_i^* = y_i - v_i$$

  where v_i is the measurement error of the respnse. Then

$$y_i = \beta_1 + \beta_2 x_i^* + (\varepsilon_i^* + v_i)$$

- The error of the model is $\varepsilon_i = \varepsilon_i^* + v_i$, which is still uncorrelated with x if v is uncorrelated with x, which means that the error of measurement of the response is uncorrelated with the covariate
- The measurement error only increases the size of the error, which implies that the coefficients are estimated less precisely and that the R^2 is lower compared to a model with a correctly measured response

Linear Model

1. Errors in the variables (predictor)

- Let's suppose now that the covariate is observed with error, namely that the observed value of the covariate is

$$x_i^* = x_i - v_i$$

  where v_i is the measurement error of the covariate.

- If the measurement error is uncorrelated with the value of the covariate, the variance of the observed covariate is

$$\sigma_x^2 = \sigma_x^{*2} + \sigma_v^2$$

  and the covariance between the observed covariate and the measurement error is equal to the variance of the measurement error, that is $cov(x, v) = \mathrm{E}(x^* + v - \mu_x)v) = \sigma_v^2$ because the measurement error is uncorrelated with the covariate

- So, rewriting the model in terms of x, we get

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

  with $\varepsilon_i = \varepsilon_i^* - \beta_2 v_i$

- The error of the model is correlated with x, as $cov(x, \varepsilon) = cov(x^* + v, \varepsilon - \beta v) = -\beta\sigma_v^2$

Linear Model

1. Errors in the variables (predictor)

- The OLS estimator can be written as usual as

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \beta_2 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Taking the expectation, we have $\mathrm{E}[(x - \bar{x})\varepsilon]) = -\beta_2\sigma_v^2$ and the expected value of the estimator is

$$\mathrm{E}(\hat{\beta}_2) = \beta_2\left(1 - \frac{\sigma_v^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2/n}\right) = \beta_2\left(1 - \frac{\sigma_v^2}{\hat{\sigma}_x^2}\right)$$

- The OLS estimator is biased and the term in brackets is the minus the share of the variance of that is due to measurement errors

Linear Model

1. **Errors in the variables (predictor)**

$$E(\hat{\beta}_2) = \beta_2 \left(1 - \frac{\sigma_v^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2 / n} \right) = \beta_2 \left(1 - \frac{\sigma_v^2}{\hat{\sigma}_x^2} \right)$$

- Then, $|\hat{\beta}_2| < \beta_2$
- This is called **attenuation bias**: it can be either a lower or an upper bias depending on the sign of β
- This bias clearly doesn't attenuate in large samples. As n grows, the empirical variances/covariances converge to the population ones, and the estimator therefore converges to $\beta_2(1 - \sigma_v^2/\sigma_x^2)$

Linear Model

2. Omitted variables bias

- Suppose that the true model is:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$$

  where $E[\varepsilon \,|\, x, z] = 0$ and the model can be estimated consistently using OLS

- **Consider that z is unobserved**
- The model to be estimated is

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

  where $\varepsilon_i^* = \varepsilon_i + \beta_3 z_i$

Linear Model

2. Omitted variables bias

- The omission of relevant covaritate $\beta_3 \neq 0$ has two consequences

1. The variance of the error is

$$\sigma^2_{\varepsilon^*} = \beta_3^2 \sigma_z^2 + \sigma_\varepsilon^2$$

   and it is greater than the one of the initial model for which z is observed and used as a covariate

2. the covariance between the error and x is

$$\text{cov}(x, \varepsilon^*) = \beta_3 \text{cov}(x, z)$$

   if the covariate is correlated with the omitted variable, the covariate and the error of the model are correlated.

Linear Model

2. Omitted variables bias

- As the variance of the OLS estimator is proportional to the variance of the errors, omission of a relevant covariate will always induce a less precise estimation of the slopes and a lower R^2
- Moreover, if the omitted covariate is correlated with the covariate used in the regression, the estimation will be biased and inconsistent
- This omitted variable bias can be computed as follows:

$$\hat{\beta}_2 = \beta_2 + \frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_3 z_i + \varepsilon_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_2 + \beta_3 \frac{\sum_{i=1}^n (x_i - \bar{x})(z_i + \bar{z})}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- Taking the conditional expectation, the last term disappear, so that

$$\mathrm{E}(\hat{\beta}_2 \mid x, z) = \beta_2 + \beta_3 \widehat{cov}(x, z)/\hat{\sigma}_x^2$$

1. upper bias if the signs of the covariance between x and z and β_3 are the same
2. a lower bias if they have opposite signs

Linear Model

2. Omitted variables bias

- There is an upper bias if the signs of the covariance between x and z and β_3 are the same, and a lower bias if they have opposite signs.
- As $n \to \infty$ the OLS estimator converges to

$$\hat{\beta}_2 \xrightarrow{p} \beta_2 + \beta_3 cov(x, z)/var(x) = \beta_2 + \beta_3 \times \beta^{*2}$$

  where

- β^{*2} is the true value of the slope of the regression of z on x
- This formula makes clear what $\hat{\beta}_2$ really estimates in a linear regression:

1. The direct effect of x on y
2. The indirect effect of x on y which is the product of x in z (β^{2*}) times the effect of z on ε^*

Econometric Example

2. **Returns from education**

- A classic example of omitted variable bias occurs in the estimation of a **Mincer earning function**, which relates **wage** (w), **education** (e) and **experience** (s)

$$\log(w_i) = \beta_1 + \beta_2 e_i + \beta_3 s_i + \beta_4 s_i^2 + \varepsilon_i$$

where β_2 is the percentage increase of the wage for one more year of education, holding fixed the other predictors. Indeed

$$\beta_2 = \frac{d\log(w)}{de} = \frac{dw/w}{de}$$

- Numerous studies of the Mincer function deal with this problem of endogeneity of the education level

Econometric Example

2. Returns from education

- Sample of 303 white males taken from the National Longitudinal Survey of Youth in 1992
- Variables:

1. **wage**: Log hourly wage
2. **educ**: Education (years of schooling)
3. **AFQT**: Ability based on standardized AFQT test score
4. **educSibl**: Education of oldest sibling (years of schooling)
5. **exper**: Experience (total weeks of labor market experience)
6. **tenure**: Tenure (weeks on current job)
7. **educM**: Mother's education (years of schooling)
8. **educF**: Father's education (years of schooling)
9. **urban**: Dummy variable for urban residence
10. **home**: Broken home dummy (dummy for living with both parents at age 14)

```
data <- read.table("kpt.dat")
```

```r
colnames(data) <- c("wage", "educ", "AFQT", "educSibl", "exper",
                    "tenure", "educM", "educF", "urban", "home")
```

Econometric Example

2. Returns from education

- Let’s fit an OLS regression: one more year of education increases the average by 10%, holding fixed the experience

```r
data$exper <- data$exper/52

fit1 <- lm(wage ~ educ + poly(exper, 2), data = data)
summary(fit1)$coefficients
```

```
                  Estimate Std. Error  t value     Pr(>|t|)
(Intercept)      1.1114866 0.15949528 6.968774 2.045533e-11
educ             0.1000376 0.01179666 8.480160 1.049101e-15
poly(exper, 2)1 2.3912857 0.45538678 5.251109 2.874745e-07
poly(exper, 2)2 0.4821923 0.45418215 1.061672 2.892415e-01
```

Econometric Example

2. Returns from education

- Concern: the individuals have different abilities (a), and that more abilities have a positive effect on wage, but may also have a positive effect on education
- If so, adding ability in the model

$$\log(w_i) = \beta_1 + \beta_2 e_i + \beta_3 s_i + \beta_4 s_i^2 + \beta_5 a_i + \varepsilon_i$$

will provide $\beta_5 > 0$ and regressing ability on the education, that is

$$e_i = \beta_1^* + \beta_2^* a_i + \varepsilon_i^*$$

will provide $\beta_2^* > 0$

- Thus

$$\hat{\beta}_2 \xrightarrow{p} \beta_2 + \beta_5 \times \beta_2^* > \beta_2$$

and the OLS estimator is upwarded biased

Econometric Example

2. Returns from education

- This is the case, because more education

1. increases, for a given level of ability, the expected wage is β_2
2. on average, the level of ability is higher, this effect being β_2^*, so the wage will also be higher

- In our data set we have the ability (**AFQT**), which is the standardized AFQT test score.
- If we introduce ability in the regression, education is no more endogenous and least squares will give a consistent estimation of the effect of education on wage.
- We first check that education and ability are positively correlated:

```
with(data, cor(educ, AFQT))
```

```
[1] 0.6055756
```

Econometric Example

2. Returns from education

- Adding ability as a covariate should decrease the coefficient on education
- The effect of one more year of education is now an increase of 8.7% of the wage (compared to the 10%)
- The relatively small decrease suggests that the correlation between education and ability is limited, so the omitted variable bias in this example is not very large

```r
fit2 <- lm(wage ~ educ + poly(exper, 2) + AFQT, data = data)
summary(fit2)$coefficients
```

```
                  Estimate Std. Error  t value     Pr(>|t|)
(Intercept)     1.23953056 0.19077860 6.497220 3.428526e-10
educ            0.08874818 0.01498128 5.923938 8.660934e-09
poly(exper, 2)1 2.30936000 0.45993514 5.021056 8.861583e-07
poly(exper, 2)2 0.44963379 0.45459292 0.989091 3.234210e-01
AFQT            0.04693974 0.03844733 1.220884 2.230949e-01
```

Linear Model

3. Simultaneity bias

- Often in economics, the phenomenon of interest is not described by a single equation, but by a system of equations
- Consider for example a market equilibrium. The two equations relate the quantity demanded / supplied (q^d and q^s) to the unit price and to some specific covariates to the demand and to the supply side of the market
- The equilibrium on the loan market is then defined by a system of three equations:

$$\begin{cases} q^d = \beta_1^d + \beta_2^d p + \beta_3^d d + \varepsilon^d \\ q^s = \beta_1^s + \beta_2^s p + \beta_3^o s + \varepsilon^s \\ q^d = q^s \end{cases}$$

where

1. q is the quantity (in logarithm)
2. p is the price
3. d and s two rates

Linear Model

3. Simultaneity bias

$$\begin{cases} q^d = \beta_1^d + \beta_2^d p + \beta_3^d d + \varepsilon^d \\ q^s = \beta_1^s + \beta_2^s p + \beta_3^o s + \varepsilon^s \\ q^d = q^s \end{cases}$$

- The demand curve should be decreasing: $\beta_2^d < 0$
- The supply curve should be increasing: $\beta_2^s > 0$
- By fitting the OLS regression we can identify the correct signs…
- However, the fit of the supply equation is very bad

Linear Model

3. Simultaneity bias

- What is actually observed for each observation in the sample is a price-quantity combination at an equilibrium.
- A positive shock on the demand equation will move upward the demand curve and will lead to a new equilibrium with a higher equilibrium quantity $q^{'}$ and also a higher equilibrium price $p^{'}$ (except in the special case where the supply curve is vertical, which means that the price elasticity of supply is infinite).
- This means that p is correlated with ε^d, which leads to a bias in the estimation of β_2^d via OLS
- The same reasoning applies of course to the supply curve.

Linear Model

3. Instrumental variable estimator

- Instrumental variable regression can eliminate the bias when $\mathrm{E}(\varepsilon \,|\, x) \neq 0$ using an instrumental variable (let's say z)
- We will explore very quickly the theoretical part (see the iv2 slides), in particular we will see an application in R