

# Statistical Modeling

CH.7 - Correlated Errors

SS 2022 | | Prof. Dr. Buchwitz

Wir geben Impulse

- 1 Evaluation
- 2 Organizational Information
- 3 Autocorrelation
- 4 Handling Autocorrelation: Transformation
- 5 Autocorrelation and missing Variables

# Bitte evaluieren Sie den Kurs!

<http://evasys.fh-swf.de/evasys/online.php?pswd=K94FQ>

- 1 Evaluation
- 2 Organizational Information
- 3 Autocorrelation
- 4 Handling Autocorrelation: Transformation
- 5 Autocorrelation and missing Variables

# Course Contents

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 Evaluation
- 2 Organizational Information
- 3 Autocorrelation
- 4 Handling Autocorrelation: Transformation
- 5 Autocorrelation and missing Variables

- One of the **standard regression assumptions** is that the error terms  $\epsilon_i$  and  $\epsilon_j$  (of the  $i$ -th and  $j$ -th observation) are **uncorrelated**.
- **Correlation in the error terms suggests that there is additional information in the data that has not been exploited in the model.** When observations have a *natural sequential order*, the correlation is referred to as **autocorrelation**.
- Adjacent residuals tend to be similar (in temporal and spatial dimensions). Successive residuals in time series tend to be positively correlated.
- If the observations of an **omitted variable** are correlated, the errors from the estimated model will appear to be correlated.

## Consequences of Autocorrelation:

- 1) Least squares estimates of the regression coefficients are unbiased but not efficient in the sense that they no longer have minimum variance.
- 2) The estimate of  $\sigma^2$  and the standard errors of the regression coefficients may be seriously understated, giving a *spurious* impression of accuracy.
- 3) The confidence intervals and tests of significance would no longer strictly valid.



**We will cover two types of autocorrelation:**

- 1** Autocorrelation due to **omission of a variable**. Once the missing variable is uncovered, the autocorrelation problem is resolved.
- 2** **Pure autocorrelation**, that can be dealt with by applying transformations to the data.

# Example: Consumer Expenditure and Money Stock

P211

##	Year	Quarter	Expenditure	Stock
## 1	1952	1	214.6	159.3
## 2	1952	2	217.7	161.2
## 3	1952	3	219.6	162.8
## 4	1952	4	227.2	164.6
## 5	1953	1	230.9	165.9
## 6	1953	2	233.3	167.9
## 7	1953	3	234.1	168.3
## 8	1953	4	232.3	169.7
## 9	1954	1	233.7	170.5
## 10	1954	2	236.5	171.6
## 11	1954	3	238.7	173.9
## 12	1954	4	243.2	176.1
## 13	1955	1	249.4	178.0
## 14	1955	2	254.3	179.1
## 15	1955	3	260.9	180.2
## 16	1955	4	263.3	181.2
## 17	1956	1	265.6	181.6
## 18	1956	2	268.2	182.5
## 19	1956	3	270.4	183.3
## 20	1956	4	275.6	184.3

## Data Description

Expenditure Consumer  
expenditure (bn dollar)

Stock Stock of money (bn dollar)

Year Calendrical year of  
observation

Quarter Quarter of observation

$$y_t = \beta_0 + \beta_1 x_t + \epsilon_t$$

- The regression model above can be seen as a **simplified** model of the quantity theory of money.
- The coefficient  $\beta_1$  is called the *multiplier* and of interest for economists and is an important measure in fiscal and monetary policy.
- Since the observations are ordered in time, it is reasonable to expect that autocorrelation may be present.

## Example: Consumer Expenditure and Money Stock

```
mod <- lm(Expenditure ~ 1 + Stock, data=P211)
summary(mod)
```

```
##
## Call:
## lm(formula = Expenditure ~ 1 + Stock, data = P211)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -7.18  -3.40   1.40   2.93   6.36
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -154.719     19.850   -7.79  3.5e-07 ***
## Stock         2.300       0.115   20.08  9.0e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.98 on 18 degrees of freedom
## Multiple R-squared:  0.957, Adjusted R-squared:  0.955
## F-statistic: 403 on 1 and 18 DF, p-value: 8.99e-14
```

The analysis were complete if the basic regression assumptions were valid (which requires checking the residuals). If autocorrelation is present the model needs to be reestimated.

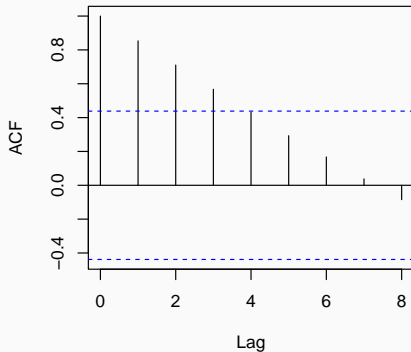
# Autocorrelation Function

```
par(mfrow=c(1,2))  
acf(P211$Expenditure, lag.max = 8)  
acf(P211$Stock, lag.max = 8)
```

**Series P211\$Expenditure**



**Series P211\$Stock**

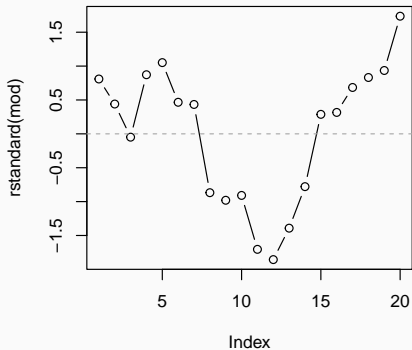


# Residuals

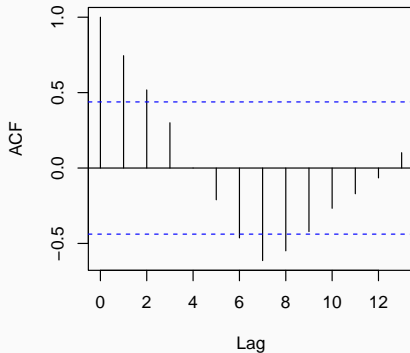
```
par(mfrow=c(1,2))  
plot(rstandard(mod), type="b", main="Standardized Residuals")  
abline(h=0, col="darkgrey", lty="dashed")  
acf(rstandard(mod))
```

The sequence run length of the sign of the residuals suggests departure from randomness.

Standardized Residuals



Series rstandard(mod)



- The Durbin-Watson statistic is the basis of a popular test of autocorrelation in regression analysis. It is based on the assumption that successive errors are correlated:

$$\epsilon_t = \rho\epsilon_{t-1} + \omega_t \quad \text{with} \quad |\rho| < 1$$

- Here  $\rho$  is the correlation coefficient between  $\epsilon_t$  and  $\epsilon_{t-1}$ , and  $\omega_t$  is normally independently distributed with zero mean and constant variance.
- Given that  $\rho$  is significant, the errors are said to have **first-order autoregressive structure** or first-order autocorrelation.
- Generally errors will have a more complex dependency structure and the simple first-order dependency is taken as a **simple approximation** of the actual error structure.

The Durbin-Watson statistic is defined as:

$$d = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2}$$

- $e_i$  is the  $i$ -th OLS residual.
- The tested hypotheses are  $H_0 : \rho = 0$  versus  $H_1 : \rho > 0$ . Where  $\rho = 0$  means that the  $\epsilon_i$ 's are uncorrelated.
- Determining the distribution of  $d$  is not trivial, and for determining the  $p$ -values multiple procedures exist (which we do not discuss here).



# Durbin-Watson Test

```
lmtest::dwtest(mod) # p-value based on linear combination of chi-square values
```

```
##  
## Durbin-Watson test  
##  
## data: mod  
## DW = 0.33, p-value = 2e-08  
## alternative hypothesis: true autocorrelation is greater than 0
```

```
car::durbinWatsonTest(mod) # p-value based on bootstrapping
```

```
## lag Autocorrelation D-W Statistic p-value  
## 1 0.7506 0.3282 0  
## Alternative hypothesis:  $\rho \neq 0$ 
```

- 1 Evaluation
- 2 Organizational Information
- 3 Autocorrelation
- 4 Handling Autocorrelation: Transformation
- 5 Autocorrelation and missing Variables

# Transformations for Handling Autocorrelation

$$\begin{aligned}\epsilon_t &= y_t - \beta_0 - \beta_1 x_t \\ \epsilon_{t-1} &= y_{t-1} - \beta_0 - \beta_1 x_{t-1}\end{aligned}$$

Substituting in  $\epsilon_t = \rho\epsilon_{t-1} + \omega_t$  yields:

$$y_t - \beta_0 - \beta_1 x_t = \rho (y_{t-1} - \beta_0 - \beta_1 x_{t-1}) + \omega_t$$

Rearranging yields:

$$\begin{aligned}y_t - \rho y_{t-1} &= \beta_0(1 - \rho) + \beta_1(x_t - \rho x_{t-1}) + \omega_t \\ y_t^* &= \beta_0^* + \beta_1^* x_t^* + \omega_t\end{aligned}$$

# Transformations for Handling Autocorrelation

- Since the  $\omega_t$ 's are uncorrelated, the transformed model represents a linear model with uncorrelated errors.
- This suggests to estimate OLS on the transformed variables  $y_t^*$  and  $x_t^*$ . The relation between the parameters in the transformed and original model are:

$$\hat{\beta}_0 = \frac{\hat{\beta}_0^*}{1 - \hat{\rho}} \quad \text{and} \quad \hat{\beta}_1 = \hat{\beta}_1^*$$

The strength of the autocorrelation is unknown, so that  $\rho$  needs to be estimated!

## Summary of the Procedure (Cochrane and Orcutt)

- 1 Compute the OLS estimates of  $\beta_0$  and  $\beta_1$  by fitting  $y_t = \beta_0 + \beta_1 x_t + \epsilon_t$  to the data.
- 2 Compute the residuals from the OLS model and estimate  $\rho$  using 
$$\hat{\rho} = \sum_{t=2}^n e_t e_{t-1} / \sum_{t=1}^n e_t^2.$$
- 3 Refit a linear model  $y_t^* = \beta_0^* + \beta_1^* x_t^* + \omega_t$  using the transformed variables  $y_t^* = y_t - \rho y_{t-1}$  and  $x_t^* = x_t - \rho x_{t-1}$ .
- 4 Examine the residuals of the newly fitted model. If the new residuals continue to show autocorrelation, repeat the entire procedure using the current model as starting point.

# Cochrane-Orcutt Estimation (Manually)

```
# Functions
d <- function(e){sum((head(e,length(e)-1) - tail(e,length(e)-1))^2) / sum(e)}
rho <- function(e){sum(head(e,length(e)-1) * tail(e,length(e)-1)) / sum(e)}

# Model 1 (OLS)
mod <- lm(Expenditure ~ 1 + Stock, data=P211)

# Model 2 (Cochrane Orcutt)
df <- P211
df$Expenditure_lag1 <- c(NA, head(df$Expenditure,nrow(df)-1))
df$Stock_lag1 <- c(NA, head(df$Stock,nrow(df)-1))
df$y_new <- df$Expenditure - rho(residuals(mod)) * df$Expenditure_lag1
df$x_new <- df$Stock - rho(residuals(mod)) * df$Stock_lag1
mod.co <- lm(y_new ~ 1 + x_new, data=df)

# Comparison: Both models in terms of the original Data
c(coef(mod), beta1_se=summary(mod)$coefficients[2,2])
```

```
## (Intercept)      Stock      beta1_se
##   -154.7192     2.3004      0.1146
```

```
c(coef(mod.co)[1] / (1 - rho(residuals(mod))), coef(mod.co)[2],
  beta1_se=summary(mod.co)$coefficients[2,2])
```

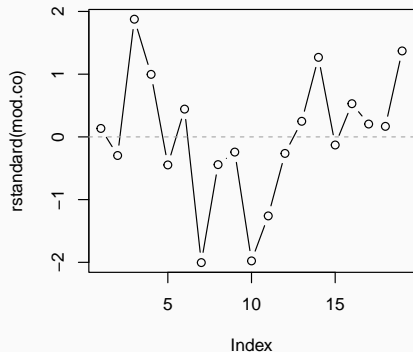
```
## (Intercept)      x_new      beta1_se
##   -215.3110     2.6434      0.3069
```

The  $\beta_1$  coefficient only changed slightly, however, the standard error increased by a factor of almost 3.

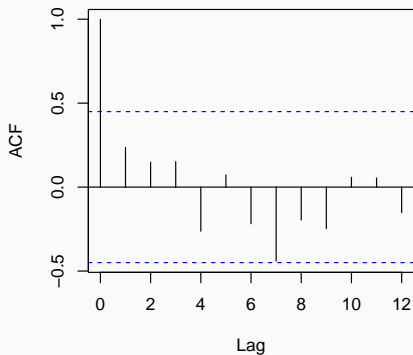
# Cochrane-Orcutt Estimation (Manually)

```
par(mfrow=c(1,2))  
plot(rstandard(mod.co), type="b", main="Standardized Residuals")  
abline(h=0, col="darkgrey", lty="dashed")  
acf(rstandard(mod.co))
```

**Standardized Residuals**



**Series rstandard(mod.co)**



- A more direct approach is estimating values of  $\rho$ ,  $\beta_0$  and  $\beta_1$  directly, instead of the classical two-step Cochrane-Orcutt procedure. This can be achieved by integrating  $\rho$  as parameter in the transformed model and simultaneously minimizing the sum of squares.

$$S(\beta_0, \beta_1, \rho) = \sum_{t=2}^n [y_t - \rho y_{t-1} - \beta_0(1 - \rho) - \beta_1(x_t - \rho x_{t-1})]^2$$

- The standard error of  $\beta_1$  can then be calculated using  $\hat{\sigma} = S(\hat{\beta}_0, \hat{\beta}_1, \hat{\rho}) / (n - 2)$  (treating  $\hat{\rho}$  as known) like

$$s.e(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum [x_t - \hat{\rho}x_{t-1} - \bar{x}(1 - \hat{\rho})]^2}}$$



# Iterative Cochrane-Orcutt-Style Estimation (R)

```
(mod.coit <- orcutt::cochrane.orcutt(mod))
```

```
## Cochrane-orcutt estimation for first order autocorrelation
##
## Call:
## lm(formula = Expenditure ~ 1 + Stock, data = P211)
##
## number of interaction: 13
## rho 0.8241
##
## Durbin-Watson statistic
## (original): 0.32821 , p-value: 2.303e-08
## (transformed): 1.60103 , p-value: 1.261e-01
##
## coefficients:
## (Intercept)      Stock
## -235.488      2.753
```

- 1 Evaluation
- 2 Organizational Information
- 3 Autocorrelation
- 4 Handling Autocorrelation: Transformation
- 5 Autocorrelation and missing Variables

- When an index plot of the residuals shows a pattern described previous (e.g. positive or negative clusters), it is reasonable to suspect that this may be due to the **omission of variables that change over time**.
- Exploring additional regressors is better than reverting to an autoregressive model, as it is less complex and potentially easier to understand. **The transformations that correct for pure autocorrelation may be viewed as an action of last resort.**
- In general a high value of the Durbin-Watson statistic should be seen as an indicator that a problem exists (missing variable and pure autocorrelation are possible).

# Example: Housing Starts

P219

##		H	P	D
## 1	0.09090	2.200	0.03635	
## 2	0.08942	2.222	0.03345	
## 3	0.09755	2.244	0.03870	
## 4	0.09550	2.267	0.03745	
## 5	0.09678	2.280	0.04063	
## 6	0.10327	2.289	0.04237	
## 7	0.10513	2.289	0.04715	
## 8	0.10840	2.290	0.04883	
## 9	0.10822	2.299	0.04836	
## 10	0.10741	2.300	0.05160	
## 11	0.10751	2.300	0.04879	
## 12	0.11429	2.340	0.05523	
## 13	0.11048	2.386	0.04770	
## 14	0.11604	2.433	0.05282	
## 15	0.11688	2.482	0.05473	
## 16	0.12044	2.532	0.05531	
## 17	0.12125	2.580	0.05898	
## 18	0.12080	2.605	0.06267	
## 19	0.12368	2.631	0.05462	
## 20	0.12679	2.658	0.05672	
## 21	0.12996	2.684	0.06674	
## 22	0.13445	2.711	0.06451	
## 23	0.13325	2.738	0.06313	
## 24	0.13863	2.766	0.06573	
## 25	0.13964	2.793	0.07229	

## Data Description

H Housing Starts

P Population Size (millions)

D Availabilit for Mortgage Money  
Index

## Example: Housing Starts

- The goal of the model is to better understand the relationship between housing starts (indicator for privately owned new houses on which construction has been started) and population growth.
- A **starting point** is the simple (and naive) model which relates housing starts and population

$$H_t = \beta_0 + \beta_1 P_t + \epsilon_t$$

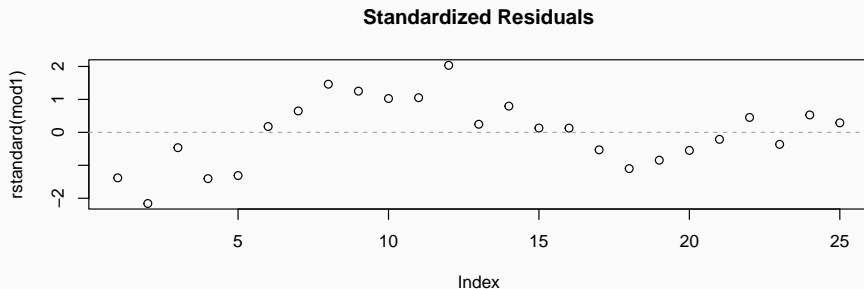
## Example: Housing Starts

```
mod1 <- lm(H ~ 1 + P, data=P219)
summary(mod1)
```

```
##
## Call:
## lm(formula = H ~ 1 + P, data = P219)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.008368 -0.002133  0.000525  0.002557  0.008075
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.06088     0.01042   -5.85  5.9e-06 ***
## P           0.07141     0.00423   16.87  1.9e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.00408 on 23 degrees of freedom
## Multiple R-squared:  0.925, Adjusted R-squared:  0.922
## F-statistic: 285 on 1 and 23 DF, p-value: 1.91e-14
```

## Example: Housing Starts

```
plot(rstandard(mod1), main="Standardized Residuals")  
abline(h=0, col="darkgrey", lty="dashed")
```



```
car::durbinWatsonTest(mod1)
```

```
## lag Autocorrelation D-W Statistic p-value  
## 1 0.6511 0.6208 0  
## Alternative hypothesis: rho != 0
```

## Example: Housing Starts

- The residual index plot and the Durbin-Watson-Test suggest autocorrelation.
- The importance of additional variables for the relationship like, *unemployment rate, social trends in marriage and family formation, goverment programs for housing and availability of construction and mortgage funds* cannot be neglected.

```
mod2 <- lm(H ~ 1 + P + D, data=P219)
car::durbinWatsonTest(mod2) # Adding Money Indicator removes autocorrelation!
```

```
## lag Autocorrelation D-W Statistic p-value
## 1 0.03957 1.852 0.436
## Alternative hypothesis: rho != 0
```



## Example: Housing Starts

```
mod3 <- lm(scale(H) ~ 1 + scale(P) + scale(D), data=
texreg::texreg(list(mod1, mod2, mod3))
```

	Model 1	Model 2	Model 3
(Intercept)	-0.06*** (0.01)	-0.06*** (0.01)	-0.06*** (0.01)
P	0.07*** (0.00)	0.03*** (0.00)	0.03*** (0.00)
D		0.76*** (0.12)	0.76*** (0.12)
scale(P)			0.47*** (0.09)
scale(D)			0.54*** (0.09)
R <sup>2</sup>	0.93	0.97	0.97
Adj. R <sup>2</sup>	0.92	0.97	0.97
Num. obs.	25	25	25

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

The standardized model shows that the mortgage index has a larger effect (and thus is more important for modeling the relationship). If **D** increases by one standard deviation **H** increases by 0.54 standard deviations.

**Table 2:** Statistical models

- If the pattern of time dependence is other than first order, the plot of residuals will still be informative.
- The Durbin-Watson statistic is, however, not designed to capture higher-order time dependence and may not yield much valuable information.

# Example: Ski Sales

P224

##	Quarter	Sales	PDI	Season
## 1	Q1/64	37.0	109	1
## 2	Q2/64	33.5	115	0
## 3	Q3/64	30.8	113	0
## 4	Q4/64	37.9	116	1
## 5	Q1/65	37.4	118	1
## 6	Q2/65	31.6	120	0
## 7	Q3/65	34.0	122	0
## 8	Q4/65	38.1	124	1
## 9	Q1/66	40.0	126	1
## 10	Q2/66	35.0	128	0
## 11	Q3/66	34.9	130	0
## 12	Q4/66	40.2	132	1
## 13	Q1/67	41.9	133	1
## 14	Q2/67	34.7	135	0
## 15	Q3/67	38.8	138	0
## 16	Q4/67	43.7	140	1
## 17	Q1/68	44.2	143	1
## 18	Q2/68	40.4	147	0
## 19	Q3/68	38.4	148	0
## 20	Q4/68	45.4	151	1
## 21	Q1/69	44.9	153	1
## 22	Q2/69	41.6	156	0
## 23	Q3/69	44.0	160	0
## 24	Q4/69	48.1	163	1
## 25	Q1/70	49.7	166	1
## 26	Q2/70	43.9	171	0
## 27	Q3/70	41.6	174	0
## 28	Q4/70	51.0	175	1
## 29	Q1/71	52.0	180	1
## 30	Q2/71	46.2	184	0
## 31	Q3/71	47.1	187	0

## Data Description

Quarter Quarter

Sales Sales

PDI Personal Disposable Income

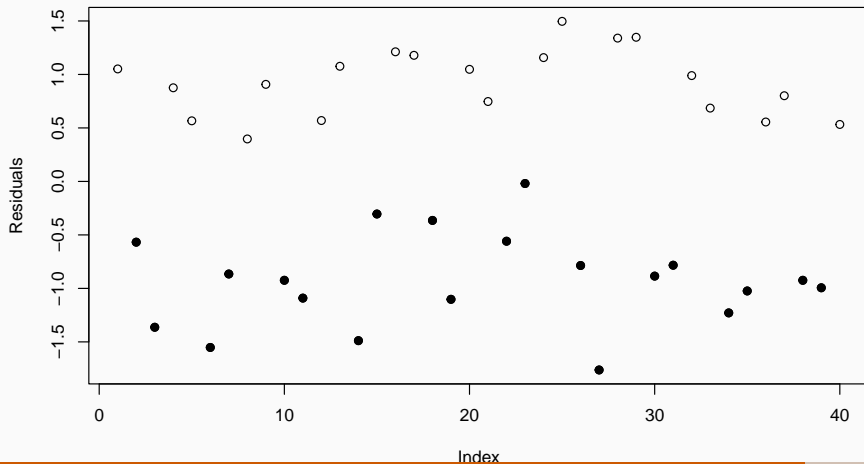
Season Indicator of Season (1 for Q1 and Q4, 0 otherwise)

# Example: Ski Sales

```
mod1 <- lm(Sales ~ 1 + PDI, data=P224)
d(residuals(mod1)) # Durbin-Watson Statistic (own Function defined above)
```

```
## [1] 1.968
```

**Standardized Residuals (values in Season are White)**



## Example: Ski Sales

```
mod2 <- lm(Sales ~ 1 + PDI + Season, data=P224)
texreg::texreg(list(mod1,mod2))
```

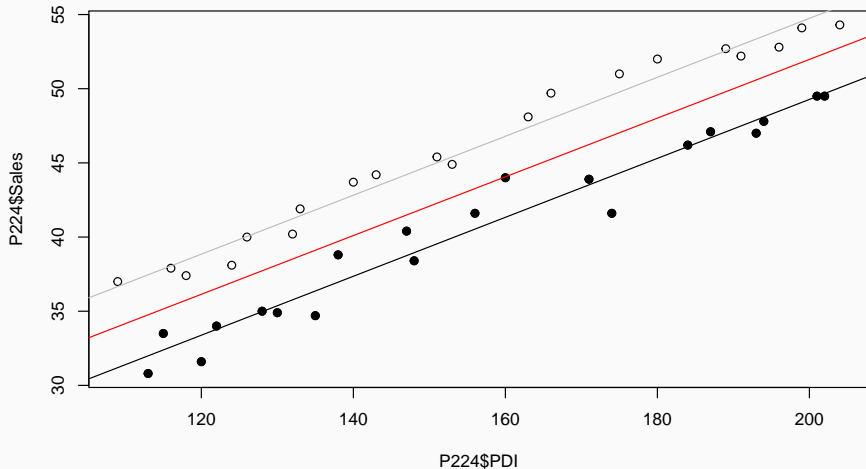
	Model 1	Model 2
(Intercept)	12.39*** (2.54)	9.54*** (0.97)
PDI	0.20*** (0.02)	0.20*** (0.01)
Season		5.46*** (0.36)
R <sup>2</sup>	0.80	0.97
Adj. R <sup>2</sup>	0.80	0.97
Num. obs.	40	40

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

**Table 3:** Statistical models

## Example: Ski Sales

Pooled vs. different Intercept based on Season-Dummy

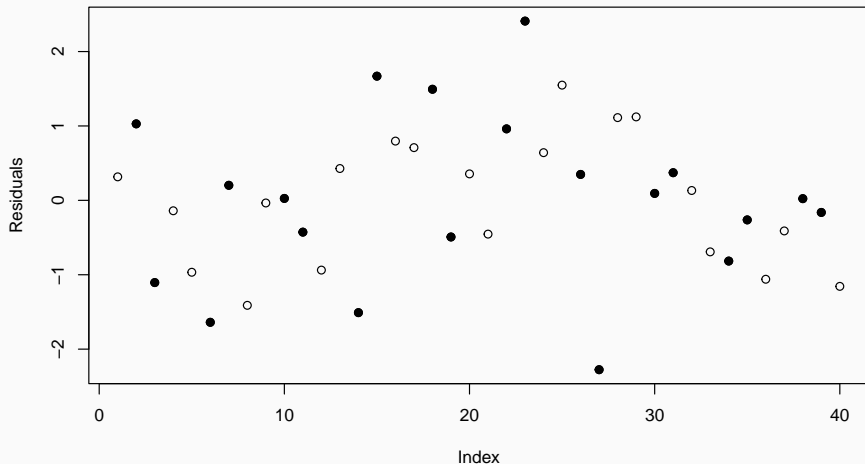


# Example: Ski Sales

```
d(residuals(mod2)) # Durbin-Watson Statistic (own Function defined above)
```

```
## [1] 1.772
```

**Standardized Residuals (values in Season are White)**



- The Durbin-Watson statistic is only sensitive to correlated errors, when the correlation occurs between adjacent observations (first-order autocorrelation).
- There are other tests that may be used for detection of higher-order autocorrelations (e.g. the Box-Pierce statistic), which we not cover here.
- The plot of the residuals is capable of revealing correlation structures of any order.
- If autocorrelation is identified, the model needs to be adapted.
- No autocorrelation is equivalent that the Durbin-Watson statistic is close to 2 (as  $d \propto 2 \cdot (1 - \rho)$ ).



- The data used here is mostly time series data instead of cross-sectional data (all observations captured at one point in time).
- The problem of autocorrelation is not relevant for cross-sectional data as the ordering of the observations is **often arbitrarily**. The correlation of adjacent observations is thus an effect of the organization of the data.
- Time series data often contains trends, which are direct functions of time a time variable  $t$ . So variables such as  $t$  or  $t^2$  could be included in the list of predictor variables.
- Additional variables such as lagged values of an regressor could be included in a model so that e.g.  $y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{1,t-1} + \beta_3 x_{2,t}$ .