Fachhochschule Südwestfalen
University of Applied Sciences

Statistical Modeling

CH.7 - Correlated Errors

SS 2021 | | Prof. Dr. Buchwitz

Outline

Bitte evaluieren Sie den Kurs!

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

Outline

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

Outline

Introduction

- One of the **standard regression assumptions** is that the error terms ϵ_i and ϵ_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 σ^2 and the standard errors rof 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 his 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

Example: Consumer Expenditure and Money Stock

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

- The regression model above can be seen as a **simplified** model of the quantity theory of money.
- The coefficient β_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

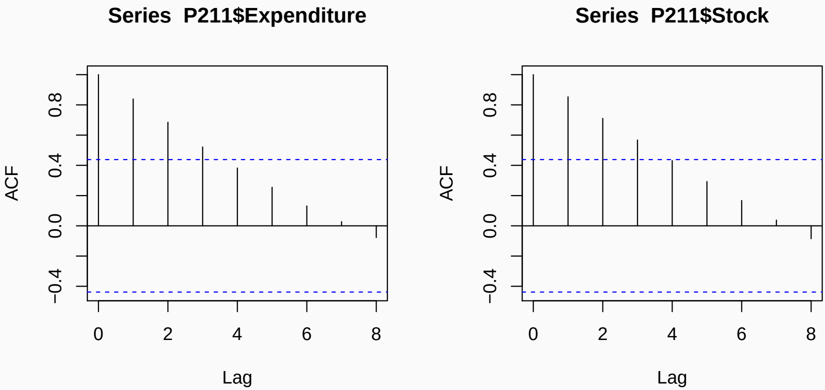
```r
mod <- lm(Expenditure ~ 1 + Stock, data=P211)
summary(mod)
```

```
##
## Call:
## lm(formula = Expenditure ~ 1 + Stock, data
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.176 -3.396   1.396   2.928   6.361
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -154.7192    19.8500  -7.794 3.54e-07 ***
## Stock          2.3004     0.1146  20.080 8.99e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.983 on 18 degrees of freedom
## Multiple R-squared:  0.9573, Adjusted R-squared:  0.9549
## F-statistic: 403.2 on 1 and 18 DF,  p-value: 8.988e-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

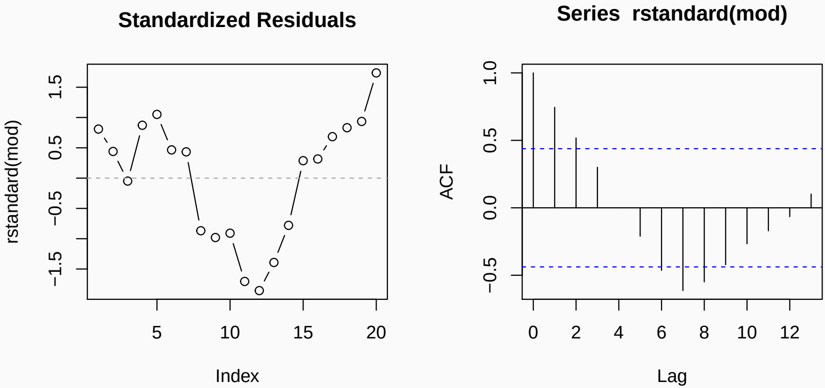
```r
par(mfrow=c(1,2))
acf(P211$Expenditure, lag.max = 8)
acf(P211$Stock, lag.max = 8)
```



Residuals

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

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



Durbin-Watson Test

- 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 ρ is the correlation coefficient between ϵ_t and ϵ_{t-1}, and ω_t is normally independenlty distribution with zero mean and constant variance.
- Given that ρ is significant, the errors are said to have **first-order autoregressive strucutre** 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.

Durbin-Watson Test

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 ϵ_i's are uncorrelated.
- Determining the distribution of d is not trivial, and for determinig the p-values multiple procedures exist (which we do not discuss here).

Durbin-Watson Test

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

```
##
##  Durbin-Watson test
##
## data:  mod
## DW = 0.32821, p-value = 2.303e-08
## alternative hypothesis: true autocorrelation is greater than 0
```

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

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.7506122     0.3282113       0
##  Alternative hypothesis: rho != 0
```

Outline

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:

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

Rearranging yields:

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

Transformations for Handling Autocorrelation

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

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

The strength of the autocorrelation is unknown, so that ρ needs to be estimated!

Transformations for Handling Autocorrelation

Summary of the Procedure (Cochrane and Orcutt)

1. Compute the OLS estimates of β_0 and β_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 ρ 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)

The β_1 coefficient only changed slightly, however, the standard error increased by a factor of almost 3.

```
# Functions
d <- function(e){sum((head(e,length(e)-1) - tail(e,length(e)-
rho <- function(e){sum(head(e,length(e)-1) * tail(e,length(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])
```

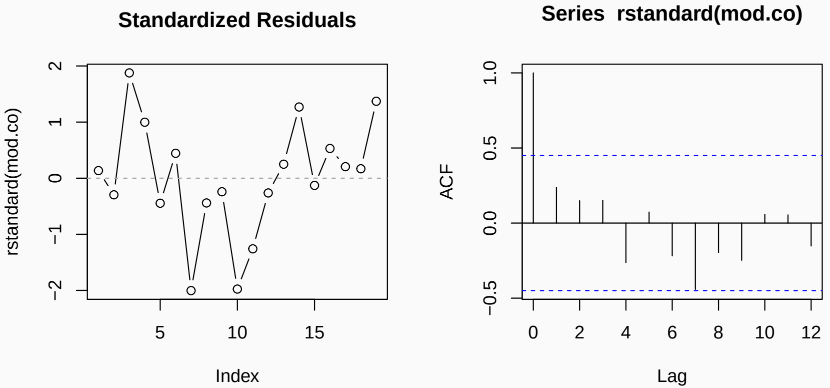
```
##  (Intercept)        Stock     beta1_se
## -154.7191620    2.3003707    0.1145583
```

```
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
## -215.310969    2.643443    0.306882
```

Cochrane-Orcutt Estimation (Manually)

```r
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))
```



Iterative Cochrane-Orcutt-Style Estimation

- A more direct approach is estimating values of ρ, β_0 and β_1 directly, instead of the classical two-step Cochrane-Orcutt prodecure. This can be achived by integrating ρ 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 β_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)

```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 0.824054
##
## Durbin-Watson statistic
## (original):    0.32821 , p-value: 2.303e-08
## (transformed): 1.60103 , p-value: 1.261e-01
##
##  coefficients:
## (Intercept)       Stock
## -235.488248    2.753057
```

Outline

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 an 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 (missign 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 owened ney 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

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

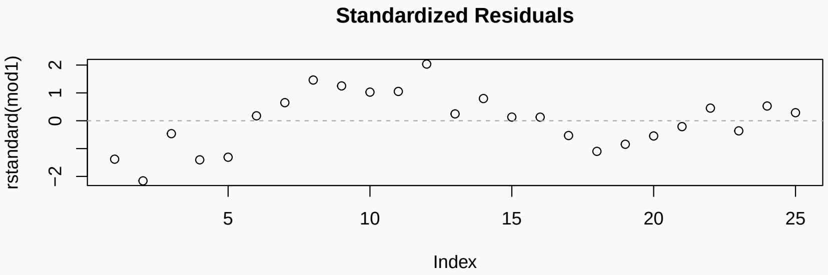
Example: Housing Starts

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

```
##
## Call:
## lm(formula = H ~ 1 + P, data = P219)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0083683 -0.0021329  0.0005252  0.0025572  0.0080754
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.060884   0.010416  -5.845 5.89e-06 ***
## P            0.071410   0.004234  16.867 1.91e-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:  0.9252, Adjusted R-squared:  0.922
## F-statistic: 284.5 on 1 and 23 DF,  p-value: 1.911e-14
```

Example: Housing Starts

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




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

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.6511468     0.6208403       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.

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

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.03957229      1.852409   0.432
##  Alternative hypothesis: rho != 0
```

Example: Housing Starts

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

	Model 1	Mod	
(Intercept)	-0.06^{***}	$-0.$	
	(0.01)	(0.(	
P	0.07^{***}	0.03	
	(0.00)	(0.(	
D		0.76	
		(0.1	
scale(P)			0.47^{***}
			(0.09)
scale(D)			0.54^{***}
			(0.09)
R^2	0.93	0.97	0.97
Adj. R^2	0.92	0.97	0.97
Num. obs.	25	25	25

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 2: Statistical models

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.

Limits of the Durbin-Watson Test

- If the pattern of time dependence is other than first order, teh 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
```

Data Description

`Quarter` Quarter

`Sales` Sales

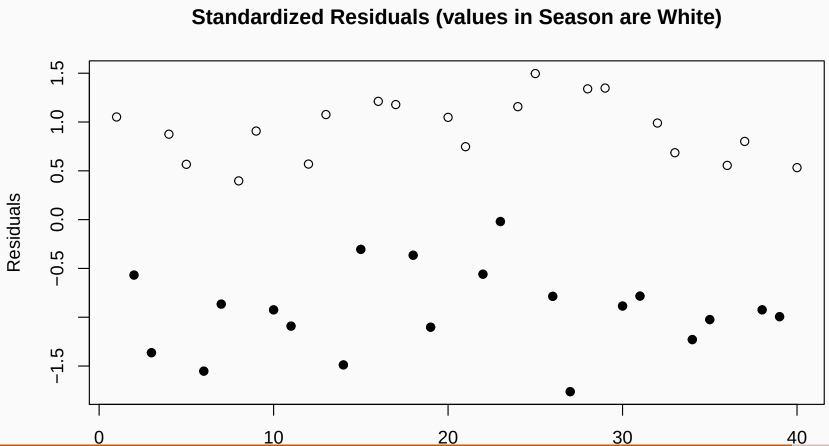
`PDI` Personal Disposable Income

`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] 1.968394
```



Example: Ski Sales

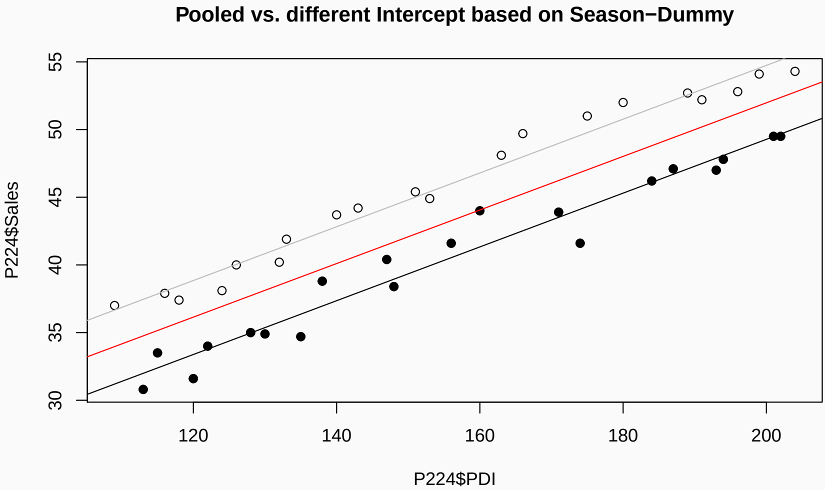
```r
mod2 <- lm(Sales ~ 1 + PDI + Season, data=P224)
texreg(list(mod1,mod2))
```

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

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 3: Statistical models

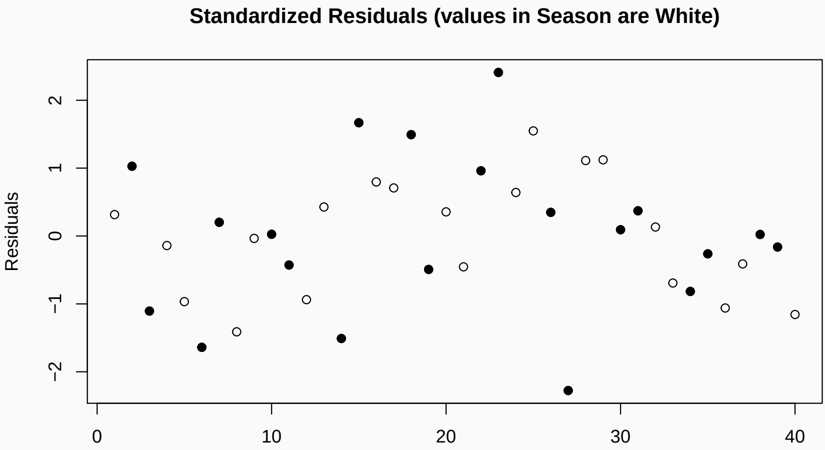
Example: Ski Sales



Example: Ski Sales

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

```
## [1] 1.771843
```



Conclusion 1/2

- 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 strucutres 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 $d \propto 2 \cdot (1 - \rho)$).

Conclusion 2/2

- The data used here is mostly time series data instead of cross-sectional data (all observations caputred 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 trens, which are 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. $y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{1,t-1} \beta_3 x_{2,t}$.