Fachhochschule Südwestfalen
University of Applied Sciences

Statistical Modeling

CH.8 - Analysis of Collinear Data

SS 2021 || Prof. Dr. Buchwitz

Wir geben Impulse

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

- The interpretation of the coefficients in a multiple regression equation depend implicitly on the assumption that the predictors are **not strongly interrelated**.
- The common interpretation of regression coefficient is the change in the response when the corresponding predictor is increased by one unit and all other predictors are held constant.

This interpretation may not be valid if there are string linear relationships among the regressors.

Multicollinearity

- When there is complete absence of linear relationships among the predictor variables, they are said to be *orthogonal*.
- In most applications the regressors are not orthogonal. However, in some situations the predictor variables are so strongly interrelated that the regression resuls ts are ambigious.
- The condition of severe nonorthogonality is also referred to as the problem of **multicollineartiy**.
- This problem is *not a specification error* and thus cannot be detected in teh residuals.
- **Multicollinearity** is a condition of deficient data.

We cover the following topics:

1. How does collinearity affect statistical inference and forecasting?
2. How can collinearity be detected?
3. What can be done to resolve the difficulties associated with collinearity (**next Session**).

In an analysis these questions cannot be answered separately. When multicollinearity all therre issues must be treated simultaneously.

Outline

Example: Effects on Inference

```
P236
```

```
##         ACHV      FAM     PEER   SCHOOL
## 1  -0.43148  0.60814  0.03509  0.16607
## 2   0.79969  0.79369  0.47924  0.53356
## 3  -0.92467 -0.82630 -0.61951 -0.78635
## 4  -2.19081 -1.25310 -1.21675 -1.04076
## 5  -2.84818  0.17399 -0.18517  0.14229
## 6  -0.66233  0.20246  0.12764  0.27311
## 7   2.63674  0.24184 -0.09022  0.04967
## 8   2.35847  0.59421  0.21750  0.51876
## 9  -0.91305 -0.61561 -0.48971 -0.63219
## 10  0.59445  0.99391  0.62228  0.93368
## 11  1.21073  1.21721  1.00627  1.17381
## 12  1.87164  0.41436  0.71103  0.58978
## 13 -0.10178  0.83782  0.74281  0.72154
## 14 -2.87949 -0.75512 -0.64411 -0.56986
## 15  3.92590 -0.37407 -0.13787 -0.21770
## 16  4.35084  1.40353  1.14085  1.37147
## 17  1.57922  1.64194  1.29229  1.40269
## 18  3.95689 -0.31304 -0.07980 -0.21455
## 19  1.09275  1.28525  1.22441  1.20428
## 20 -0.62389 -1.51938 -1.27565 -1.36598
## 21 -0.63654 -0.38224 -0.05353 -0.35560
## 22 -2.02659 -0.19186 -0.42605 -0.53718
## 23 -1.46692  1.27649  0.81427  0.91967
## 24  3.15078  0.52310  0.30720  0.47231
## 25 -2.18938 -1.59810 -1.01572 -1.48315
## 26  1.91715  0.77914  0.87771  0.76496
```

Data Description

`ACHV` Student achievements.

`FAM` Faculty credentials

`PEER` Influence of peer group in school.

`SCHOOL` School facilities.

All variables are normalized indices. Goal is to evaluate the effect of school inputs on achievements.

Example: Effects on Inference

- The goal of the analysis is to measure the effect of the school inputs on achievements to asses *Equal Education Opportunity*. The variable `SCHOOL` is an index and we assume that it measures those aspects of the school environment that would affect achievement (physical plant, teaching materials, special programs, etc.).
- `ACHV` is an index constructed based on normalized test scores.
- Before we can assess the effect of the school we need to account for other variables that may influence `ACHV`, like the peer group and the personal environment. We assume that thos are captured in the indices for `PEER` and `FAM`.

$$ACHV = \beta_0 + \beta_1 FAM + \beta_2 PEER + \beta_3 SCHOOL + \epsilon$$

Example: Effects on Inference

- The contribution of the `SCHOOL` variable can be testes using the *t*-Test for β_3.
- The *t*-Test checks wheteher `SCHOOL` is necessary in the equation, given that `FAM` and `PEER` are already included.
- This can be interpreted as checking for an effect after the `ACHV` index has been adjusted for `FAM` and `PEER`.

$$ACHV - \beta_1 FAM - \beta_2 PEER = \beta_0 + \beta_3 SCHOOL + \epsilon$$

Note: This model is only for the sake of interpretation the model on the previous page is sufficient for the actual analysis.

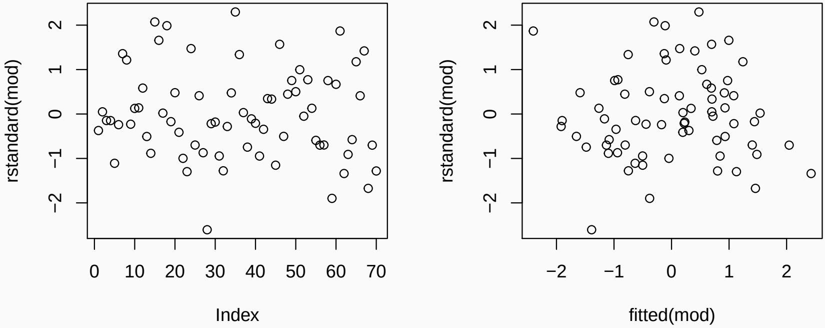
Example: Effects on Inference

```r
mod <- lm(ACHV ~ 1 + FAM + PEER + SCHOOL, data=P236)
summary(mod)
```

```
##
## Call:
## lm(formula = ACHV ~ 1 + FAM + PEER + SCHOOL, data = P236)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2096 -1.3934 -0.2947  1.1415  4.5881
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.06996    0.25064  -0.279    0.781
## FAM          1.10126    1.41056   0.781    0.438
## PEER         2.32206    1.48129   1.568    0.122
## SCHOOL      -2.28100    2.22045  -1.027    0.308
##
## Residual standard error: 2.07 on 66 degrees of freedom
## Multiple R-squared:  0.2063, Adjusted R-squared:  0.1702
## F-statistic: 5.717 on 3 and 66 DF,  p-value: 0.001535
```

Example: Effects on Inference

```r
par(mfrow=c(1,2))
plot(rstandard(mod))
plot(fitted(mod), rstandard(mod))
```



Example: Effects on Inference

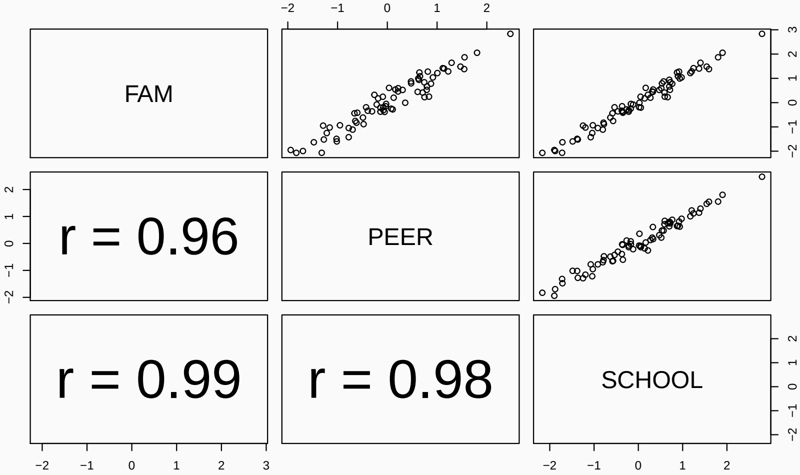
Observation:

- The regression model accounts for 20.63% of the data.
- The F-Statistic with a value of 5.7168 is significant and indicates a joint effect of the variables.
- All t-Statistics are small and indicate that none of the variables individually are significant.

Conclusion:

- The given situation is common for settings where **multicollinearity** occurs.
- The small t-values suggest that any of the variables can be dropped and the joint R^2 is affected by the realtionship among the predictors.

Example: Effects on Inference



Example: Effects on Inference

Combination	FAM	PEER	SCHOOL
1	+	+	+
2	+	+	-
3	+	-	+
4	-	+	+
1	+	-	-
2	-	+	-
3	-	-	+
4	-	-	-

A "+" indicates a value above average in the data. The dataset only contains combiantions 1 and 8 and is deficient so that not all partial effects can be estimated.

Example: Effects on Inference

- The dataset contains *missing combinations* which leads to the empy regions in the pairsplot. There may be two reasons for this:
  1) Incomplete data collection, so that collecting additional data leads do disappearing multicollinearity.
  2) The ground truth (population) only contains a specific set of combinations. Then it is not possible to separate effects and estimate the individual effects on achievement. A detailed investigation may lead to additinal variables thate are *more basic* determinants for the response.

Example: Effects on Forecasting

- We now examine the effect of multicollinearity on **forecasting**.
- The considered dataset (imports in the French economy) is index by time (variable `YEAR`).
- To generate **forecasts for the response**, future values of the predictor variables are plugged into the estimated regression equation.
- The future values of the predictor variables must be known or need to be forecasted themselfes (not discussed in this course).
- We assume that the future values of the predictor variables are **given**, which is highly **unrealistic and only for explanatory purposes**.

Example: Effects on Forecasting

```
P241
```

```
##    YEAR IMPORT DOPROD STOCK CONSUM
## 1    49   15.9  149.3   4.2  108.1
## 2    50   16.4  161.2   4.1  114.8
## 3    51   19.0  171.5   3.1  123.2
## 4    52   19.1  175.5   3.1  126.9
## 5    53   18.8  180.8   1.1  132.1
## 6    54   20.4  190.7   2.2  137.7
## 7    55   22.7  202.1   2.1  146.0
## 8    56   26.5  212.4   5.6  154.1
## 9    57   28.1  226.1   5.0  162.3
## 10   58   27.6  231.9   5.1  164.3
## 11   59   26.3  239.0   0.7  167.6
## 12   60   31.1  258.0   5.6  176.8
## 13   61   33.3  269.8   3.9  186.6
## 14   62   37.0  288.4   3.1  199.7
## 15   63   43.3  304.5   4.6  213.9
## 16   64   49.0  323.4   7.0  223.8
## 17   65   50.3  336.8   1.2  232.0
## 18   66   56.6  353.9   4.5  242.9
```

Data Description

`YEAR` Year of Observation.

`IMPORT` Import Volume.

`DOPROD` Domestic Production.

`STOCK` Stock Formation.

`CONSUM` Domestic Consumption.

Variables are measured in billion French francs.

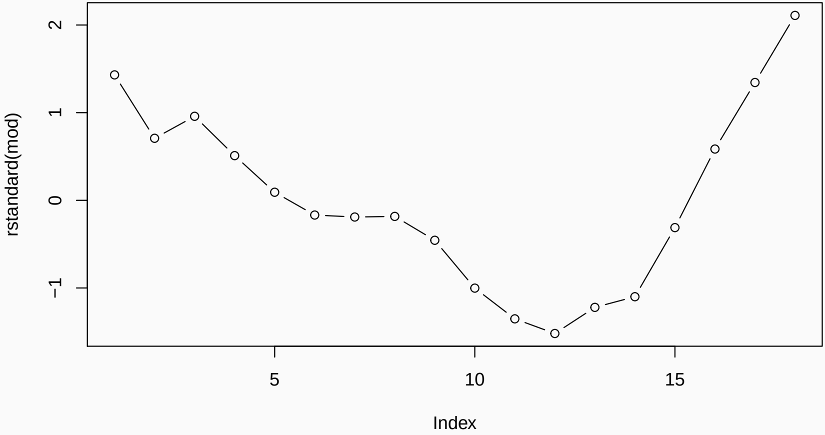
Example: Effects on Forecasting

```r
mod <- lm(IMPORT ~ 1 + DOPROD + STOCK + CONSUM, data=P241)
summary(mod)
```

```
##
## Call:
## lm(formula = IMPORT ~ 1 + DOPROD + STOCK + CONSUM, data = P241)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7208 -1.8354 -0.3479  1.2973  4.1008
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -19.7251     4.1253  -4.782 0.000293 ***
## DOPROD        0.0322     0.1869   0.172 0.865650
## STOCK         0.4142     0.3223   1.285 0.219545
## CONSUM        0.2427     0.2854   0.851 0.409268
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.258 on 14 degrees of freedom
## Multiple R-squared:  0.973,  Adjusted R-squared:  0.9673
## F-statistic: 168.4 on 3 and 14 DF,  p-value: 3.212e-11
```

Example: Effects on Forecasting

```
plot(rstandard(mod), type="b")
```



Example: Effects on Forecasting

$$\text{IMPORT} = \beta_0 + \beta_1\text{DOPROD} + \beta_2\text{STOCK} + \beta_3\text{CONSUM} + \epsilon$$

- The index plots of the residuals suggests that the model is not well specified, even though the R^2 is high.
- The problem reflected in the data is that the European Common Market began operations in 1960, causing changes in import-export relationships.
- Our objective is to study the effect of **multicollinearity**, we decide to ignore the dynamics after 1959 and only anlyze the first 11 years of data.

Example: Effects on Forecasting

```
mod <- lm(IMPORT ~ 1 + DOPROD + STOCK + CONSUM, data=head(P24
summary(mod)
```

```
##
## Call:
## lm(formula = IMPORT ~ 1 + DOPROD + STOCK + CONSUM, data =
##     11))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.52367 -0.38953  0.05424  0.22644  0.78313
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.12799    1.21216  -8.355  6.9e-05 ***
## DOPROD       -0.05140    0.07028  -0.731 0.488344
## STOCK         0.58695    0.09462   6.203 0.000444 ***
## CONSUM        0.28685    0.10221   2.807 0.026277 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4889 on 7 degrees of freedom
## Multiple R-squared:  0.9919, Adjusted R-squared:  0.9884
## F-statistic: 285.6 on 3 and 7 DF,  p-value: 1.112e-07
```

An increase in Domestic Production should cause an increase in the imports, when `STOCK` and `CONSUM` are held constant. Contrary to the prior model and to our believes, the coefficient for `DOPROD` is not statistically significant. The

residuals show no suspicious patterns.

Example: Effects on Forecasting

```
kable(round(cor(head(P241,11)),4))
```

	YEAR	IMPORT	DOPROD	STOCK	CONSUM
YEAR	1.0000	0.9476	0.9952	-0.0329	0.9952
IMPORT	0.9476	1.0000	0.9653	0.2507	0.9719
DOPROD	0.9952	0.9653	1.0000	0.0259	0.9973
STOCK	-0.0329	0.2507	0.0259	1.0000	0.0357
CONSUM	0.9952	0.9719	0.9973	0.0357	1.0000

- Investigation reveals that correlation between `CONSUM` and `DOPROD` is very high throughout the 11 year period.

Example: Effects on Forecasting

- The estimated relationship between `CONSUM` and `DOPROD` is given below.

$$\widehat{\text{CONSUM}} = 6.259 + 0.686(\text{DOPROD})$$

- Even in the presence of severe multicollinearity the regression equation *may* produce some good forecasts. The forecasting equation follows directly from the regression output.

$$\widehat{\text{IMPORT}} = -10.128 - 0.051(\text{DOPROD}) + 0.587(\text{STOCK}) + 0.287(\text{CONSUM})$$

For our purpose we must be confident that the character and strength of the overall relationship will hold into future periods (which is untrue in the given case, but ignored for convenience of explanation).

Example: Effects on Forecasting

- If we forecast the change in `IMPORT` next year corresponding to tan in crease in `DROPROD` of 10 units while holding `STOCK` and `CONSUM`at their current levels:

$$\text{IMPORT}_{1960} \approx \text{IMPORT}_{1959} - 0.051 \cdot 10$$

- This leads to an decrease in `IMPORT` by ≈ 0.51 units. However, if the relationship between `DOPROD`and `CONSUM` is kept intact, `CONSUM` will increase as well and the forecasted results changes and yields a forecased increase in `IMPORT`.

$$\text{IMPORT}_{1960} \approx \text{IMPORT}_{1959} - 0.051 \cdot 10 + 0.287 \cdot 0.686 \cdot 10$$

Outline

Simple Signs of Collinearity

- In the following we review the discussed ideas and introduce additional criteria that indicate multicollinearity.
- Besides simple indicators we are going to consider the two criteria **Variance Inflation Factors** (VIF) and **Condition Indices**.
- Simple indicators of multicollinearity are usually encountered during the process of adding, deleting or transforming variables or data points while searching for a good model.

Simple Signs of Collinearity

Indications of multicollinearity that appear as instability in the estimated coefficients are as follows:

- Large changes in the estimated coefficients when a variables is added or deleted.
- Large changes in the estimated coefficients when a data point is added or deleted.

Once the residual plots indicate that the model has been satisfactorily specified, collinearity may be present if:

- The algebraic signs of estimated coefficients do not conform to prior expectations.
- Coefficients of variables that are expected to be important have large standard errors (small t-values).

Simple Signs of Collinearity

- The table shows the effect of adding an removing a variable for the French economy data. We see that the presence or absence of certain variables has a large effect on the other coefficients.
- This problem is visible in the pairwise correlation coefficients.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
(Intercept)	-6.56^{*}	19.61^{***}	-8.01^{**}	-8.44^{***}	-8.88^{*}	-9.74^{***}	-10.13^{***}
	(2.59)	(3.25)	(2.44)	(1.44)	(2.85)	(1.06)	(1.21)
DOPROD	0.15^{***}			0.15^{***}	-0.11		-0.05
	(0.01)			(0.01)	(0.17)		(0.07)
STOCK		0.69		0.62^{**}		0.60^{***}	0.59^{***}
		(0.89)		(0.13)		(0.09)	(0.09)
CONSUM			0.21^{***}		0.37	0.21^{***}	0.29^{*}
			(0.02)		(0.24)	(0.01)	(0.10)
R^2	0.93	0.06	0.94	0.98	0.95	0.99	0.99
Adj. R^2	0.92	-0.04	0.94	0.98	0.93	0.99	0.99
Num. obs.	11	11	11	11	11	11	11

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

Table 4: Statistical models

Simple Signs of Collinearity

The source of multicollinearity may be more subtle than the simple relationship between two variables so that it **may not be possible to detect such a relationship with a simple coerrelation coefficient**.

```
kable(cor(P248)) # Advertising Data
```

	St	At	Pt	Et	At.1	Pt.1
St	1.0000000	-0.1704051	0.5401769	0.8108732	-0.3051611	-0.0520353
At	-0.1704051	1.0000000	-0.3569539	-0.1285210	-0.1397402	-0.4959912
Pt	0.5401769	-0.3569539	1.0000000	0.0625887	-0.3164655	-0.2963645
Et	0.8108732	-0.1285210	0.0625887	1.0000000	-0.1664279	0.2081124
At.1	-0.3051611	-0.1397402	-0.3164655	-0.1664279	1.0000000	-0.3577610
Pt.1	-0.0520353	-0.4959912	-0.2963645	0.2081124	-0.3577610	1.0000000

Simple Signs of Collinearity

```
P248
```

```
##          St      At  Pt   Et    At.1 Pt.1
## 1  20.11371 1.98786 1.0 0.30 2.01722  0.0
## 2  15.10439 1.94418 0.0 0.30 1.98786  1.0
## 3  18.68375 2.19954 0.8 0.35 1.94418  0.0
## 4  16.05173 2.00107 0.0 0.35 2.19954  0.8
## 5  21.30101 1.69292 1.3 0.30 2.00107  0.0
## 6  17.85004 1.74334 0.3 0.32 1.69292  1.3
## 7  18.87558 2.06907 1.0 0.31 1.74334  0.3
## 8  21.26599 1.01709 1.0 0.41 2.06907  1.0
## 9  20.48473 2.01906 0.9 0.45 1.01709  1.0
## 10 20.54032 1.06139 1.0 0.45 2.01906  0.9
## 11 26.18441 1.45999 1.5 0.50 1.06139  1.0
## 12 21.71606 1.87511 0.0 0.60 1.45999  1.5
## 13 28.69595 2.27109 0.8 0.65 1.87511  0.0
## 14 25.83720 1.11191 1.0 0.65 2.27109  0.8
## 15 29.31987 1.77407 1.2 0.65 1.11191  1.0
## 16 24.19041 0.95878 1.0 0.65 1.77407  1.2
## 17 26.58966 1.98930 1.0 0.62 0.95878  1.0
## 18 22.24466 1.97111 0.0 0.60 1.98930  1.0
## 19 24.79944 2.26603 0.7 0.60 1.97111  0.0
## 20 21.19105 1.98346 0.1 0.61 2.26603  0.7
## 21 26.03441 2.10054 1.0 0.60 1.98346  0.1
## 22 27.39304 1.06815 1.0 0.58 2.10054  1.0
```

Data Description

S_t Sales Volume.

A_t Advertising Expenditures.

P_t Promotion Expenditures.

E_t Sales Expense.

A_{t-1} and P_{t-1} are the lagged one-year variables.

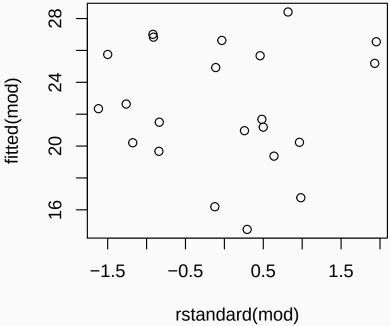
Simple Signs of Collinearity

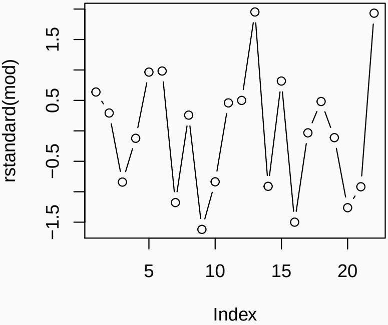
```r
mod <- lm(St ~ 1 + At + Pt + Et  + At.1 + Pt.1, data=P248)
summary(mod)
```

```
##
## Call:
## lm(formula = St ~ 1 + At + Pt + Et + At.1 + Pt.1, data = P248)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8601 -0.9848  0.1323  0.7017  2.2046
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -14.194     18.715  -0.758   0.4592
## At             5.361      4.028   1.331   0.2019
## Pt             8.372      3.586   2.334   0.0329 *
## Et            22.521      2.142  10.512 1.36e-08 ***
## At.1           3.855      3.578   1.077   0.2973
## Pt.1           4.125      3.895   1.059   0.3053
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.32 on 16 degrees of freedom
## Multiple R-squared:  0.9169, Adjusted R-squared:  0.8909
## F-statistic:  35.3 on 5 and 16 DF,  p-value: 4.289e-08
```

Simple Signs of Collinearity

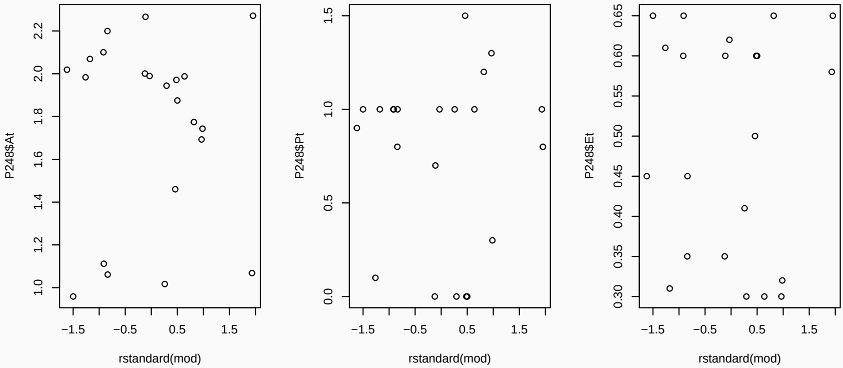
```
par(mfrow=c(1,2))
plot(rstandard(mod), fitted(mod))
plot(rstandard(mod), type="b")
```





Simple Signs of Collinearity

```
par(mfrow=c(1,3))
plot(rstandard(mod), P248$At)
plot(rstandard(mod), P248$Pt)
plot(rstandard(mod), P248$Et)
```



Simple Signs of Collinearity

- The residual plots do not exhibit clear signs of misspecification and the correlation between the predictors is moderate and does not indicate a problem.
- Experimentation shows that dripping the advertising variable A_t leads to severe changes in the coefficients (coefficient of P_t drops significantly, coefficients of lagged values change signs.)

```
mod.experiment <- lm(St ~ 1 + Pt + Et  + At.1 + Pt.1, data=P248)
coef(mod.experiment)
```

```
## (Intercept)          Pt          Et        At.1        Pt.1
##  10.5093791   3.7017522  22.7941828  -0.7691882  -0.9686601
```

Simple Signs of Collinearity

- The reason for the multicollinearity in the previous example is a budget constraint so that the sum of A_t, A_{t-1}, P_t and P_{t-1} was held approximately constant:

$$A_t + A_{t-1} + P_t + P_{t-1} \approx 5$$

- This can be empirically confirmed by regressing A_t on A_{t-1}, P_t and P_{t-1}.

```
mod.constraint <- lm(At ~ 1 + Pt + At.1 + Pt.1, data=P248)
equatiomatic::extract_eq(mod.constraint, use_coef=T)
```

$$\widehat{\text{At}} = 4.63 - 0.87(\text{Pt}) - 0.86(\text{At.}\ 1) - 0.95(\text{Pt.}\ 1)$$

Variance Inflation Factors (VIF)

- A thorough investigation of muticollinearity will involve examining the value of R^2 that restulst from regression **each of the predictors against all others**.
- The resulting effects can be judged by examining a quantitiy called *variance inflation index* (**VIF**).

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2} \quad \text{with} \quad j = 1, \ldots, p$$

- R_j^2 denotes the multiple correlation coefficient from regression the predictor X_j on all other $p - 1$ predictor variables.
- When X_j has a strong linear relationship with the other variables, R_j^2 will be close to 1 and VIF_j will be large.

A VIF > 10 is often taken as indicator that the data has multicollinearity problems.

Variance Inflation Factors (VIF)

- When R_j^2 is close to zero $VIF \approx 1$. The departure from 1 indicates departure from orthogonality and tendency toward collinearity.
- The naming is derived from the fact that VIF_j measures the amount by which the variance of the j-th regression coefficient is increased due to the linear association of X_j with other predictors **relative** to the value of the variance that would result in absence of a linear relation.
- As R_j^2 approaches 1, the VIF_j for $\hat{\beta}_j$ tends to infinity.

Variance Inflation Factors (VIF)

- The precision of the OLS estimates is measured by its variance, which is proportional to the variance of the error term in the regression model σ^2.
- The **constant of proportionality** is the VIF.
- The VIF's therefore can be used to obtain an expression for the expected squared distance of the OLS estimators from their true values. The smaller D^2 the more accurate are the estimates.

$$D^2 = \sigma^2 \sum_{j=1}^{p} \text{VIF}_j$$

Variance Inflation Factors (VIF)

- If the predictors were orthogonal, the VIF's would be equal to 1 and $D^2 = p\sigma^2$. It follows that the ratio $\overline{\text{VIF}}$ measures the squared error in the OLS estimators relative to the size of the error if the data were orthogonal.

$$\overline{\text{VIF}} = \frac{\sigma^2 \sum_{i=1}^{p} \text{VIF}_i}{p\sigma^2} = \frac{\sum_{i=1}^{p} \text{VIF}_i}{p}$$

- $\overline{\text{VIF}}$ can also be used as an **index for multicollinearity**.

Variance Inflation Factors (VIF)

```r
# Equal Education Opportunity Data
mod.eeo <- lm(ACHV ~ 1 + FAM + PEER + SCHOOL, data=P236)
vif.eeo <- car::vif(mod.eeo)
c(vif.eeo, averageVIF = mean(vif.eeo))
```

```
##        FAM       PEER     SCHOOL averageVIF
##   37.58064   30.21166   83.15544   50.31591
```

```r
# Import Data
mod.imp <- lm(IMPORT ~ 1 + DOPROD + STOCK + CONSUM, data=P241)
vif.imp <- car::vif(mod.imp)
c(vif.imp, averageVIF = mean(vif.imp))
```

```
##     DOPROD      STOCK     CONSUM averageVIF
## 469.742135   1.049877 469.371343 313.387785
```

```r
# Advertising Data
mod.adv <- lm(St ~ 1 + At + Pt + Et  + At.1 + Pt.1, data=P248)
vif.adv <- car::vif(mod.adv)
c(vif.adv, averageVIF = mean(vif.adv))
```

```
##         At         Pt         Et       At.1       Pt.1 averageVIF
##  36.941513  33.473514   1.075962  25.915651  43.520965  28.185521
```

Condition Indices

- Another way to detect collinearity in the data is to examine the condition indices fo the *correlation matrix* of the predictor variables.
- The condition indices are based on the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ of them correlation matrix. If any $\lambda = 0$, there is perfect linear relationship, which is an extreme case of collinearity. Strong heterogeneity in the eigenvalues (one value much smaller than the others) also indicates mulitcollinearity.
- An empirical criterion for the presence of collinearity is given by the sum of the reciprocals of the eigenvalues of the correlation matrix. If that sum is much larger (e.g. 5 times larger) than the number of predictor variables p, collinearity is present.

$$\sum_{j=1}^{p} \frac{1}{\lambda_i}$$

Condition Indices

- The condition indices measures the overall collinearity of the variables. The j-th condition index is given by

$$\kappa_j = \sqrt{\frac{\lambda_1}{\lambda_p}} \quad \text{for} \quad j = 1, 2, \ldots, p$$

- The largest condition index is called *condition number* of the matrix. If that condition number is small, then the predictor variables are not collinear. A large condition number indicates strong evidence of collinearity.
- Corrective actions should be taken, when the conditio number **exceeds 15** (which means that λ_1 is more than 225 times λ_p)

Condition Indices

```r
# Equal Education Opportunity Data
mod.eeo <- lm(ACHV ~ 1 + FAM + PEER + SCHOOL, data=P236)
round(olsrr::ols_eigen_cindex(mod.eeo), 4)
```

```
##   Eigenvalue Condition Index intercept    FAM   PEER SCHOOL
## 1     2.9547          1.0000    0.0005 0.0030 0.0037 0.0014
## 2     0.9974          1.7211    0.9756 0.0000 0.0000 0.0000
## 3     0.0400          8.5996    0.0004 0.3068 0.4428 0.0008
## 4     0.0079         19.2826    0.0235 0.6903 0.5535 0.9978
```

```r
# Import Data
mod.imp <- lm(IMPORT ~ 1 + DOPROD + STOCK + CONSUM, data=head(P241,11))
round(olsrr::ols_eigen_cindex(mod.imp), 4)
```

```
##   Eigenvalue Condition Index intercept DOPROD  STOCK CONSUM
## 1     3.8384          1.0000    0.0010 0.0000 0.0109 0.0000
## 2     0.1484          5.0863    0.0053 0.0001 0.9385 0.0001
## 3     0.0132         17.0732    0.7743 0.0015 0.0330 0.0011
## 4     0.0001        265.4613    0.2193 0.9984 0.0175 0.9989
```

```r
# Advertising Data
mod.adv <- lm(St ~ 1 + At + Pt + Et  + At.1 + Pt.1, data=P248)
round(olsrr::ols_eigen_cindex(mod.adv), 4)
```

```
##   Eigenvalue Condition Index intercept     At     Pt     Et   At.1   Pt.1
## 1     5.2810          1.0000    0.0000 0.0000 0.0002 0.0023 0.0001 0.0002
## 2     0.3798          3.7291    0.0000 0.0000 0.0075 0.0003 0.0000 0.0118
## 3     0.2272          4.8209    0.0000 0.0015 0.0160 0.0000 0.0011 0.0054
## 4     0.0601          9.3777    0.0000 0.0047 0.0004 0.2912 0.0160 0.0006
```

Conclusion

- Using the described techniques we can now detect multicollinearity.
- However, it is unclear how to deal with variables that cause collinearity issues. Removing those variables is often not a viable option.
- We will learn better ways of dealing with collinearity in the next chapter.