14th Oct 2020

Advanced Topics in Biostatistics 2020/21:

Multiple Linear Regression

Dr. Diana Tichy, d.tichy@dkfz.de

GERMAN CANCER RESEARCH CENTER IN THE HELMHOLTZ ASSOCIATION

- Revision: ANOVA
- Introduction: Simple linear regression
- Multiple linear regression
    - Estimation of the coefficients
    - Model diagnostics: Assumptions Checking
    - Model diagnostics: Multicollinearity
    - Variable selection
- Conclusions and Outlook

Revision

Prostate volume is typically related to age ...

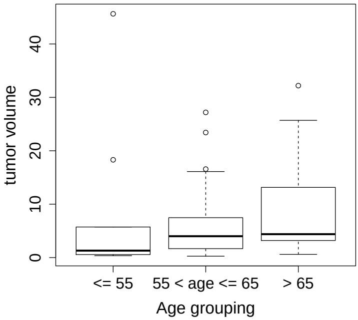
Subject	tumor volume	Age grouping
1	0.6	≤ 55
2	0.4	≤ 55
$\vdots$	$\vdots$	$\vdots$
11	2.1	$55 < \ age \ \leq 65$
12	1.3	$55 < \ age \ \leq 65$
$\vdots$	$\vdots$	$\vdots$
55	4.4	> 65
56	4.7	> 65

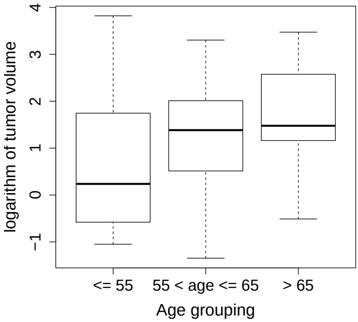
Recall: Differences between two or more groups:

Is there a difference in mean tumor volume with respect to age?

$\hookrightarrow$ Apply One-way ANOVA

One-way ANOVA





H_0 : Age grouping has no effect on mean tumor volume.

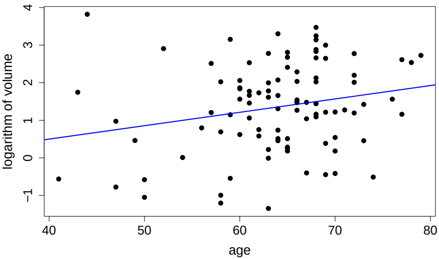
$$F_{treat} = \frac{\textit{variation between groups}}{\textit{variation within groups}} = \frac{3.9}{1.3} = 2.9 \qquad \hookrightarrow \qquad p = 0.057 \qquad \text{(F-test)}$$

$\hookrightarrow$ Test result does not contradict H_0 ...

Simple linear regression

Does age effect the tumor volume?

$$log.vol \;=\; \beta_0 + \beta_1 \cdot \; age + \varepsilon$$



$\hookrightarrow$ Linear regression models!

$$H_0 \,:\; \beta_1 = 0 \,:\;\; \text{Age has no effect on tumor volume.}$$

Can the effect of age be adjusted for other potential factors?

Dataset: Prostate tumor volume

Data from $n = 97$ men who are supposed to undergo prostatectomy:

	vol	lcavol	lweight	age	lbph	lcp	lpsa	age.group
17	0.66	-0.4155154	3.5160	70	1.244155	-0.59784	1.47018	> 65
18	9.86	2.2884862	3.6494	66	-1.386294	0.37156	1.49290	> 65
19	0.57	-0.5621189	3.2677	41	-1.386294	-1.38629	1.55814	<= 55
20	1.20	0.1823216	3.8254	70	1.658228	-1.38629	1.59939	> 65
21	3.15	1.1474025	3.4194	59	-1.386294	-1.38629	1.63900	55 < age <= 65
22	7.84	2.0592388	3.5010	60	1.474763	1.34807	1.65823	55 < age <= 65

- `vol` tumor volume
- `lcavol`: tumor volume on log scale
- `lweight`: prostate weight on log scale
- `age`: age
- `lbph`: benign prostatic hyperplasia amount on log scale
- `lcp`: capsular penetration on log scale
- `lpsa`: prostate specific antigen on log scale
- `age.group`: age grouping

Motivation

Goal: model the relationship between the *response variable* tumor volume and the *predictors* prostate weight,age, benign prostatic hyperplasia, capsular penetration, prostate specific antigen

Model	Reponse Variable	Predictors
Linear Regression	continuous	continuous, categorical
Logistic Regression	binary	continuous, categorical
Cox PH regression	survival times	continuous, categorical

If predictors are categorical with more than two levels, use *dummy variables*:

- `Sigma Plot`: `Analysis -> Statistical -> Dummy Variables`
- `R`: `as.factor()`

Reminder: Simple linear regression

Model for the relationship between the response variable and *one* predictor variable (e.g. `age`).

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

$$\text{Variables} = \begin{cases} Y_i & = \text{observations of the response variable} \\ X_i & = \text{observations of the predictor variable} \\ \varepsilon_i & = \text{residuals} \end{cases}$$

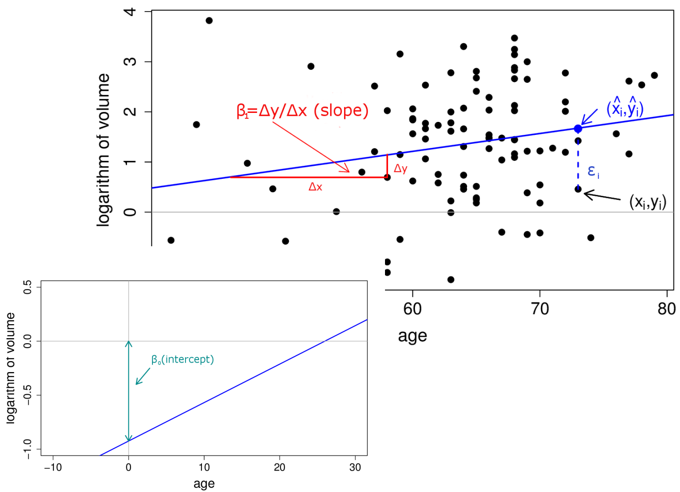
$$\text{Parameters} = \begin{cases} \beta_0 & = \text{intercept} \\ \beta_1 & = \text{slope} \end{cases}$$

Questions:

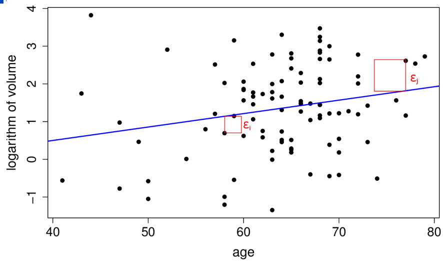
$\hookrightarrow$ Estimate coefficients β_0, β_1.
$\hookrightarrow$ Is there an effect of the predictor on the response?, e.g.: Test for

$$H_0 : \beta_1 = 0$$

Scatterplot and the regression line



Ordinary least squ

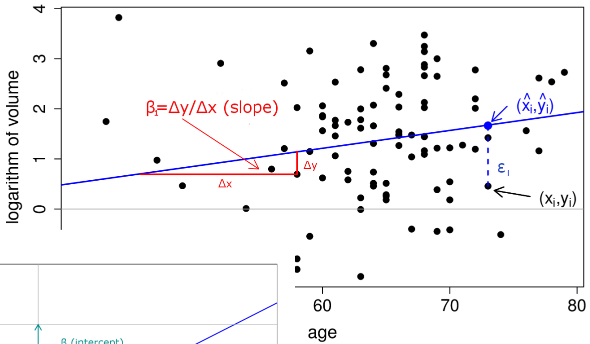


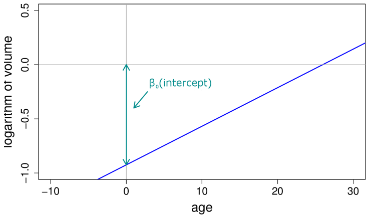
Estimate β_0, β_1, such that the sum of the squared residuals (residual $= \varepsilon_i = \hat{Y}_i - Y_i$) is minimized:

$$\sum_{i=1}^{n} (Y_i - \underbrace{(\beta_0 + \beta_1 \cdot X_i)}_{=\hat{Y}_i})^2 \stackrel{!}{=} min$$

$\hookrightarrow$ The *least squares estimators* $\widehat{\beta}_0, \widehat{\beta}_1$ of β_0, β_1 are

$$\widehat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \quad \text{and} \quad \widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \cdot \bar{X}$$





$$\widehat{log.vol} = \hat{\beta}_0 + \hat{\beta}_1 \cdot age$$
$$\hat{y}_i = -0.92 + 0.04 \cdot x_i$$

$$H_0 : \beta_1 = 0 : \hookrightarrow \; t = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{0.036}{0.016} = 2.25 \quad \hookrightarrow p = 0.027 \quad (Wald - test)$$

$\hookrightarrow$ Univariable model finds a significant effect of age on tumor volume.

Multiple linear regression

To model the joint influence of two or more predictor variables, repeated simple linear regressions are not appropriate:

- The combination of two or more variables usually gives a better prediction for the response than considering these variables separately (age and sex of a child predicts height better than just age or sex)
- Correlation Structure is neglected $\Rightarrow$ Spurious Correlations (Social Status may be significant for lung cancer in a simple linear regression, but not significant when adjusting for smoking)

Multiple Linear regression enables us to

- Find the best linear prediction for the response using a set of predictor variables
- Test if a predictor variable shows a significant effect on the response variable *adjusted* for the other predictor variables

Multiple linear regression

Model equation:

$$Y_i = \beta_0 + \beta_1 \cdot X_{1,i} + \beta_2 \cdot X_{2,i} + \ldots + \beta_K \cdot X_{K,i} + \varepsilon_i, \quad i = 1, \ldots, n$$

$$\text{Variables} = \begin{cases} Y_i & = \text{ observations of the response variable} \\ X_{1,i}, \ldots, X_{K,i} & = \text{ observations of the K predictor variables} \\ \varepsilon_i & = \text{ residuals} \end{cases}$$

$$\text{Parameters} = \begin{cases} \beta_0 & = \text{ regression coefficient for intercept} \\ \beta_1, \ldots, \beta_K & = \text{ regression coef. for predictor } X_1 \text{ to } X_K \end{cases}$$

Estimation of the regression coefficients

Least squares estimation (analogously to simple linear regression)

$$\sum_{i=1}^{n}(Y_i - \underbrace{(\beta_0 + \beta_1 \cdot X_{1,i} + \ldots + \beta_K \cdot X_{K,i}}_{=\hat{Y}_i})^2 \overset{!}{=} \mathit{min}$$

$\hookrightarrow \ \hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_K)$: *Least squares estimator (LS-estimator)*

Assumptions of multiple linear regression

1. Linear relationship between the response and the predictor variables
2. The errors $\varepsilon_1, \ldots, \varepsilon_n$
   2.1 are uncorrelated
   2.2 are normally distributed
   2.3 have the same variance (*homoscedasticity*)

These assumptions are required to ensure, that

- the choice of a *linear* regression model is correct
- the test statistics (see later on) are appropriate to test for the significance of the regression coefficients

What might happen, if assumptions are violated:

Wald-test on volume:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.38065    6.98126   0.627    0.532
age          0.04103    0.10858   0.378    0.706
```

Wald-test on log-transformed volume:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.92486    1.01749  -0.909   0.3657
age          0.03562    0.01583   2.251   0.0267
```

Dataset: Prostate tumor volume

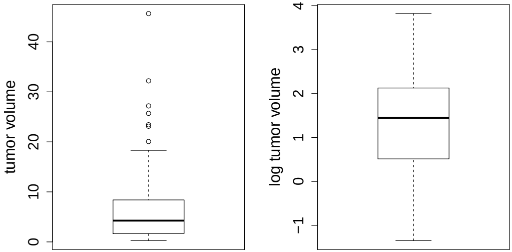
Data from $n = 97$ men who are supposed to undergo prostatectomy:

	vol	lcavol	lweight	age	lbph	lcp	lpsa	age.group
17	0.66	-0.4155154	3.5160	70	1.244155	-0.59784	1.47018	> 65
18	9.86	2.2884862	3.6494	66	-1.386294	0.37156	1.49290	> 65
19	0.57	-0.5621189	3.2677	41	-1.386294	-1.38629	1.55814	<= 55
20	1.20	0.1823216	3.8254	70	1.658228	-1.38629	1.59939	> 65
21	3.15	1.1474025	3.4194	59	-1.386294	-1.38629	1.63900	55 < age <= 65
22	7.84	2.0592388	3.5010	60	1.474763	1.34807	1.65823	55 < age <= 65

- `vol` tumor volume
- `lcavol`: tumor volume on log scale
- `lweight`: prostate weight on log scale
- `age`: age
- `lbph`: benign prostatic hyperplasia amount on log scale
- `lcp`: capsular penetration on log scale
- `lpsa`: prostate specific antigen on log scale
- `age.group`: age grouping

Data Transformation

- It is often advisable to transform the responses Y_i (and continuous predictors..) to eliminate skewness.
- If the distributon of the $Y_i' s$ is skewed, the errors ε_i are unlikely to be normally distributed (Assumption 2.2).

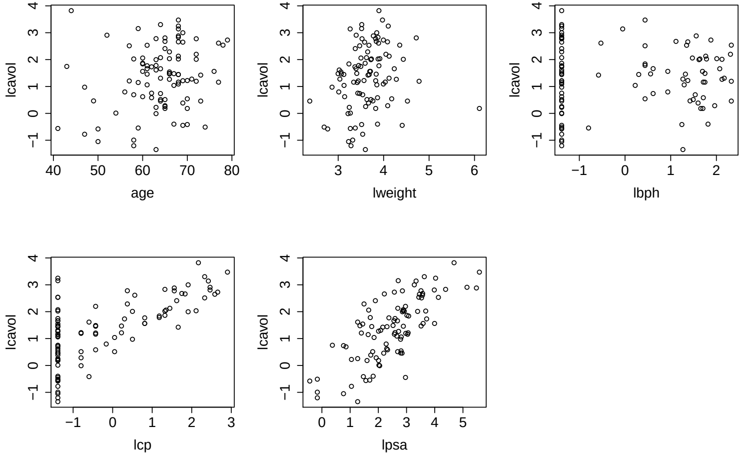


$\hookrightarrow$ Log-transformation of tumor volume eliminates skewness!

Assumptions of multiple linear regression

1. Linear relationship between the response and the predictor variables
2. The errors $\varepsilon_1, \ldots, \varepsilon_n$
   2.1 are uncorrelated
   2.2 are normally distributed
   2.3 have the same variance (*homoscedasticity*)

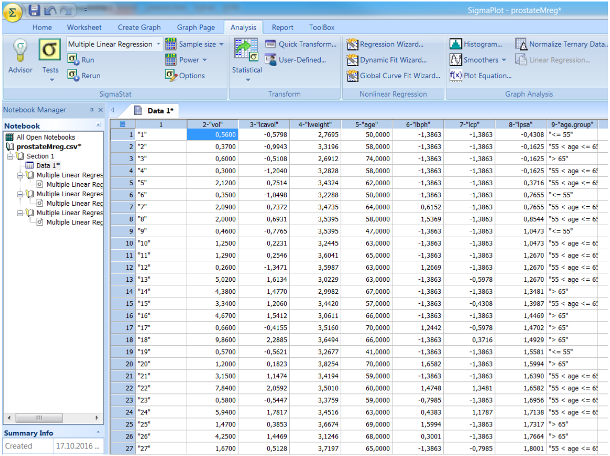
Check the linear relationship



Run Multiple Linear Regression

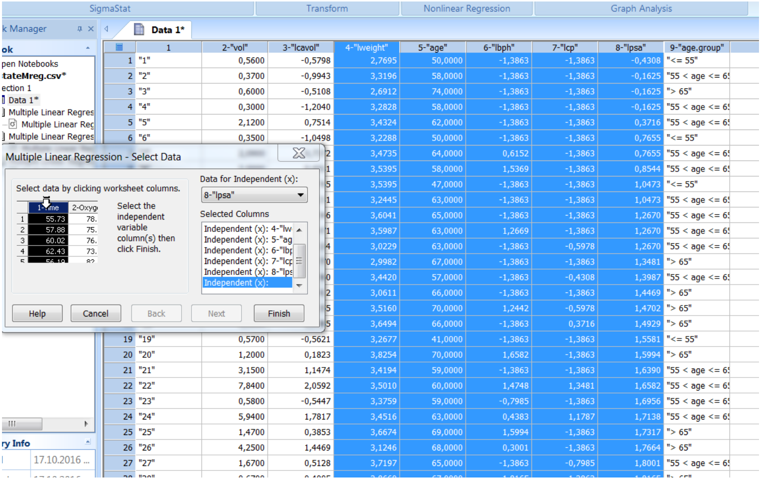
R:lmfit <- lm(lcavol∼ lweight+age+lbph+lcp+lpsa,data=dat)

Sigmaplot:



	1	2-"vol"	3-"lcavol"	4-"lweight"	5-"age"	6-"lbph"	7-"lcp"	8-"lpsa"	9-"age.group"
1	"1"	0,5600	-0,5798	2,7695	50,0000	-1,3863	-1,3863	-0,4308	"<= 55"
2	"2"	0,3700	-0,9943	3,3196	58,0000	-1,3863	-1,3863	-0,1625	"55 < age <= 6
3	"3"	0,6000	-0,5108	2,6912	74,0000	-1,3863	-1,3863	-0,1625	"> 65"
4	"4"	0,3000	-1,2040	3,2828	58,0000	-1,3863	-1,3863	-0,1625	"55 < age <= 6
5	"5"	2,1200	0,7514	3,4324	62,0000	-1,3863	-1,3863	0,3716	"55 < age <= 6
6	"6"	0,3500	-1,0498	3,2288	50,0000	-1,3863	-1,3863	0,7655	"<= 55"
7	"7"	2,0900	0,7372	3,4735	64,0000	0,6152	-1,3863	0,7655	"55 < age <= 6
8	"8"	2,0000	0,6931	3,5395	58,0000	1,5369	-1,3863	0,8544	"55 < age <= 6
9	"9"	0,4600	-0,7765	3,5395	47,0000	-1,3863	-1,3863	1,0473	"<= 55"
10	"10"	1,2500	0,2231	3,2445	63,0000	-1,3863	-1,3863	1,0473	"55 < age <= 6
11	"11"	1,2900	0,2546	3,6041	65,0000	-1,3863	-1,3863	1,2670	"55 < age <= 6
12	"12"	0,2600	-1,3471	3,5987	63,0000	1,2669	-1,3863	1,2670	"55 < age <= 6
13	"13"	5,0200	1,6134	3,0229	63,0000	-1,3863	-0,5978	1,2670	"55 < age <= 6
14	"14"	4,3800	1,4770	2,9982	67,0000	-1,3863	-1,3863	1,3481	"> 65"
15	"15"	3,3400	1,2060	3,4420	57,0000	-1,3863	-0,4308	1,3987	"55 < age <= 6
16	"16"	4,6700	1,5412	3,0611	66,0000	-1,3863	-1,3863	1,4469	"> 65"
17	"17"	0,6600	-0,4155	3,5160	70,0000	1,2442	-0,5978	1,4702	"> 65"
18	"18"	9,8600	2,2885	3,6494	66,0000	-1,3863	0,3716	1,4929	"> 65"
19	"19"	0,5700	-0,5621	3,2677	41,0000	-1,3863	-1,3863	1,5581	"<= 55"
20	"20"	1,2000	0,1823	3,8254	70,0000	1,6582	-1,3863	1,5994	"> 65"
21	"21"	3,1500	1,1474	3,4194	59,0000	-1,3863	-1,3863	1,6390	"55 < age <= 6
22	"22"	7,8400	2,0592	3,5010	60,0000	1,4748	1,3481	1,6582	"55 < age <= 6
23	"23"	0,5800	-0,5447	3,3759	59,0000	-0,7985	-1,3863	1,6956	"55 < age <= 6
24	"24"	5,9400	1,7817	3,4516	63,0000	0,4383	1,1787	1,7138	"55 < age <= 6
25	"25"	1,4700	0,3853	3,6674	69,0000	1,5994	-1,3863	1,7317	"> 65"
26	"26"	4,2500	1,4469	3,1246	68,0000	0,3001	-1,3863	1,7664	"> 65"
27	"27"	1,6700	0,5128	3,7197	65,0000	-1,3863	-0,7985	1,8001	"55 < age <= 6

Run Multiple Linear Regression



Results of Multiple Linear Regression

Multiple Linear Regression 4 Report *

Multiple Linear Regression Donnerstag, Oktober 16, 2014, 11:50:39

Data source: Excel1 in prostateMreg.xls

log vol = -1,016 - (0,0735 * lweight) + (0,0208 * age) - (0,0812 * lbph) + (0,307 * lcp) + (0,553 * lpsa)

N = 97 Missing Observations = 1

R = 0,815 Rsqr = 0,664 Adj Rsqr = 0,646

Standard Error of Estimate = 0,702

	Coefficient	**Std. Error**	**t**	**P**	**VIF**
Constant	-1,016	0,822	-1,236	0,220	
lweight	-0,0735	0,171	-0,429	0,669	1,412
age	0,0208	0,0105	1,975	0,051	1,197
lbph	-0,0812	0,0570	-1,425	0,158	1,334
lcp	0,307	0,0623	4,923	<0,001	1,479
lpsa	0,553	0,0797	6,933	<0,001	1,653

Analysis of Variance:

	DF	**SS**	**MS**	**F**	**P**
Regression	5	88,571	17,714	35,992	<0,001
Residual	91	44,788	0,492		
Total	96	133,359	1,389		

Interpretation of the results

Model:

$$lcavol = -1,016 - (0,0735 * lweight) + (0,0208 * age) - (0,0812 * lbph) + (0,307 * lcp) + (0,553 * lpsa)$$

Regression coefficients:

- Positive effects of `age`, `lcp` and `lpsa` (higher psa $\Rightarrow$ higher tumor volume)
- Negative effects of `lweight` and `lbph` (higher prostate weight $\Rightarrow$ lower tumor volume)

Tests:

- The effects of `lweight, age` and `lbph` are not significant
- `lcp` and `lpsa` are significant
- `age` is close to significance and may be worth further investigations

$R^2_{adjust} = 0.664$: two thirds of the variance in the data is explained by the above regression model.

Model diagnostics

Variance decomposition in the linear regression model:

$$\underbrace{\sum(Y_i - \bar{Y})^2}_{SS_{total}} = \underbrace{\sum(\hat{Y}_i - \bar{Y})^2}_{SS_{regression}} + \underbrace{\sum(Y_i - \hat{Y}_i)^2}_{SS_{error}}$$

$SS =$ Sum of Squares

The *multiple coefficient of determination* R^2

$$R^2 = \frac{SS_{regression}}{SS_{total}}$$

- is a measure for the model fit.
- returns the proportion of variance in the data, which is explained by the regression model.
- For $K = 1$, R^2 equals the squared Pearson correlation ρ^2.

Model diagnostics

Problem: R^2 always increases when adding predictor variables to the model, even when these predictors are independent of the response.
$\hookrightarrow$ A model with 10 predictors tends to have a larger R^2 than a model with 4 predictors, even if the 6 additional variables have nothing to do with the reponse.

The *adjusted coefficient of determination*

$$R^2_{adjust} : 1 - (1 - R^2)\left(\frac{n-1}{n-p}\right)$$

adjusts for this effect.

Model diagnostics: Assumption Checking

1. Linear relationship between the response and the predictor variables
2. The errors $\varepsilon_1, \ldots, \varepsilon_n$
   2.1 are uncorrelated
   2.2 are normally distributed
   2.3 have the same variance (*homoscedasticity*)

Model diagnostics: Assumption Checking

Correlation of the residuals $\varepsilon_1, \ldots, \varepsilon_n$ is rarely a problem.

Possible sources of Correlation:

- Different observations correspond to measurements of the same patient
- Unobserved groups (patients from the same family...)

Tools/Tests:

- Durbin-Watson Test

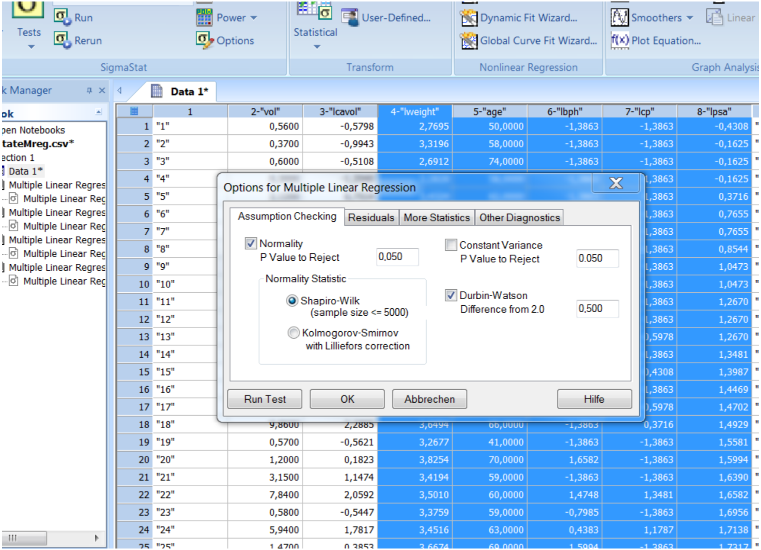
Model diagnostics: Assumption Checking

1. Linear relationship between the response and the predictor variables
2. The errors $\varepsilon_1, \ldots, \varepsilon_n$
   - **2.1** are uncorrelated
   - **2.2** are normally distributed
   - **2.3** have the same variance (*homoscedasticity*)

Tests:

- Shapiro-Wilk / Kolmogorov-Smirnov Test (Normality)
- Tests for Constant Variance

Model diagnostics: Assumption Checking



Model diagnostics: Assumption Checking

Column	SSIncr	SSMarg
"lweight"	5,026	0,0907
"age"	4,025	1,919
"lbph"	1,639	0,999
"lcp"	54,221	11,930
"lpsa"	23,660	23,660

The dependent variable "lcavol" can be predicted from a linear combination of the independent variables:

	P
"lweight"	0,669
"age"	0,051
"lbph"	0,158
"lcp"	<0,001
"lpsa"	<0,001

Not all of the independent variables appear necessary (or the multiple linear model may be underspecified).
The following appear to account for the ability to predict "lcavol" (P < 0.05): "lcp" , "lpsa"

Durbin-Watson Statistic = 2,354 Passed

Normality Test (Shapiro-Wilk) Passed (P = 0,349)

Constant Variance Test: Failed (P = 0,013)

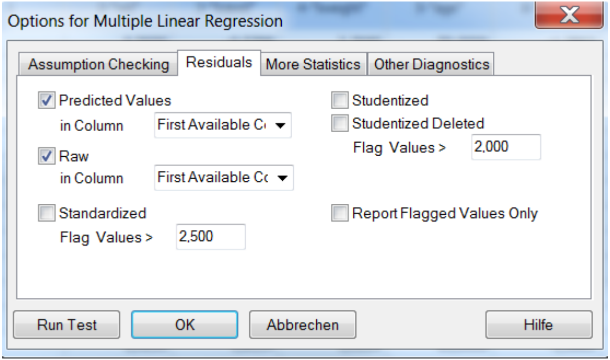
Power of performed test with alpha = 0,050: 1,000

Model diagnostics: Assumption Checking

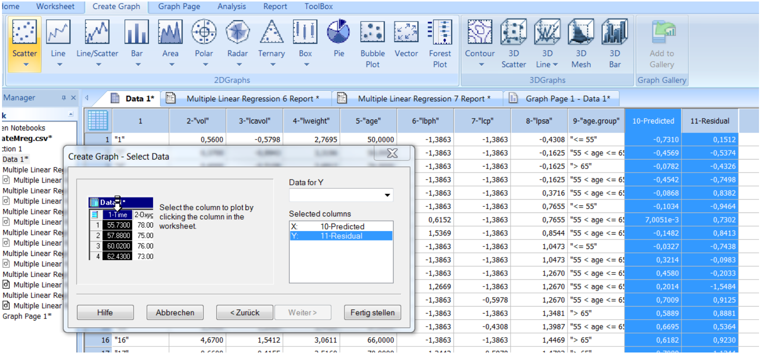
Tests are often overly sensitive for the requirements of linear regression.

$\hookrightarrow$ Additionally use graphical graphical tools

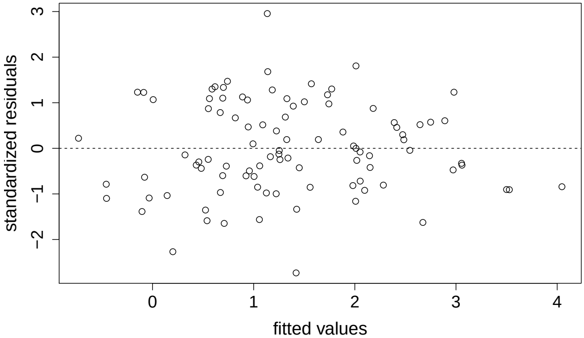
$\hookrightarrow$ Plot residuals ε_i vs. fitted values $\hat{Y}_i$



Model diagnostics: Assumption Checking



Residual vs. fitted plot for the full model



Can be used to ...

- Check variance homogeneity: If standardized residuals are homoscedastic and the linear regression model is correct, then the residuals spread equally around 0.
- Check linear relationship between predictors and response variable

Residual vs. fitted plot

Faraway, J.J. (2005). Linear models with R

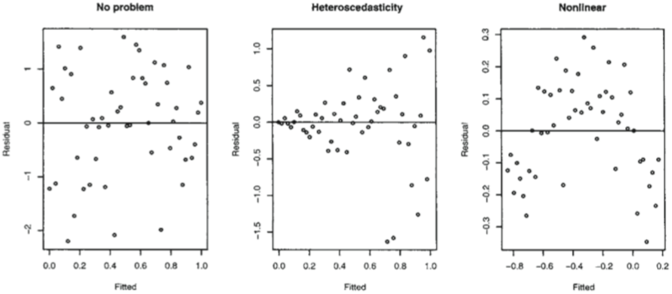


Figure 4.1 *Residuals vs. fitted plots—the first suggests no change to the current model while the second shows nonconstant variance and the third indicates some nonlinearity, which should prompt some change in the structural form of the model.*

Model diagnostics: Futher reading

Further Plots for Assumption Checking:

- Q-Q-Plot (Normality)
- Scale-Location (homoscedasticity)

Tools for detecting influential observations/outliers:

- Residuals vs. Leverage - Plot
- Cook's Distance
- dfbetas

Model Diagnostic Plots (Res vs. Fitted, Q-Q etc.) in

```
R:lmfit <- lm(lcavol∼ lweight+age+lbph+lcp+lpsa,data=dat)
  plot(lmfit)
```

Model diagnostics: Multicollinearity

- Multicollinearity occurs when two or more of the predictors are highly correlated
- This phenomenon leads to a high variance of the estimators $\hat{\beta}_i$

The *variance inflation factor*

$$VIF_i$$

is the factor by which the variance of $\hat{\beta}_i$ is increased compared to the case, where X_i is uncorrelated with the other predictors.

Rules of Thumb:

- $VIF_i > 4 \rightarrow$ There may be a problem with multicollinearity
- $VIF_i > 10 \rightarrow$ There is a severe problem with multicollinearity

Model diagnostics: Multicollinearity

Multiple Linear Regression 4 Report *

Multiple Linear Regression Donnerstag, Oktober 16, 2014, 11:50:39

Data source: Excel1 in prostateMreg.xls

log vol = -1,016 - (0,0735 * lweight) + (0,0208 * age) - (0,0812 * lbph) + (0,307 * lcp) + (0,553 * lpsa)

N = 97 Missing Observations = 1

R = 0,815 Rsqr = 0,664 Adj Rsqr = 0,646

Standard Error of Estimate = 0,702

	Coefficient	**Std. Error**	**t**	**P**	**VIF**
Constant	-1,016	0,822	-1,236	0,220	
lweight	-0,0735	0,171	-0,429	0,669	1,412
age	0,0208	0,0105	1,975	0,051	1,197
lbph	-0,0812	0,0570	-1,425	0,158	1,334
lcp	0,307	0,0623	4,923	<0,001	1,479
lpsa	0,553	0,0797	6,933	<0,001	1,653

Analysis of Variance:

	DF	**SS**	**MS**	**F**	**P**
Regression	5	88,571	17,714	35,992	<0,001
Residual	91	44,788	0,492		
Total	96	133,359	1,389		

Model diagnostics: Multicollinearity

Possible Reasons for multicollinearity:

- Highly correlated predictors (`bmi` and `weight`)
- Two variables represent transformations of the same quantity (`bph` and `log(bph)`)

	Coefficient	Std. Error	t	P	VIF
Constant	-0,857	0,829	-1,034	0,304	
lweight	-0,160	0,184	-0,867	0,388	1,642
age	0,0198	0,0105	1,879	0,064	1,204
lbph	-0,227	0,130	-1,750	0,084	6,972
lcp	0,321	0,0631	5,082	<0,001	1,527
lpsa	0,557	0,0796	7,002	<0,001	1,656
bph	0,0860	0,0688	1,251	0,214	8,010

If possible, avoid multicollinearity by selecting only one of two highly correlated predictors before running the multiple linear regression.

Variable selection

Reasons for variable selection:

- Number of observations low compared to number of variables ($<$ 10 observations per variable)
- High multicollinearity between the predictors
- Desire to obtain a simple model (for scores etc.)

If possible, select the variables before running the multiple linear regression using expert knowledge/literature.

Drawbacks of variable selection:

- Overfitting (model may depend heavily on the specific dataset)
- Test statistics of the reduced model do not account for variable selection and can not be used

Variable selection

DO NOT:

- just select variables which are significant in a simple (univariable) regression model
- report only the estimators and *p*-values of the significant predictors!

There is nothing wrong with non-significant predictors!

SINCE:

- If the true model is multivariable, a univariable model does not correctly estimate the effect, the *p*-values and parameters estimates are wrong!
- If the true model is univariable, a multivariable correctly estimates the effect (but with higher variance).

Variable Selection

Forward Selection:

- Start with an empty model
- In every step, add the predictor which yields the largest improvement of the fit
- Stop, when the fit can no longer be considerably improved

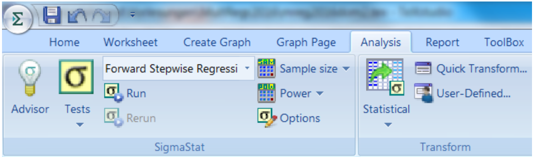
Backward Selection:

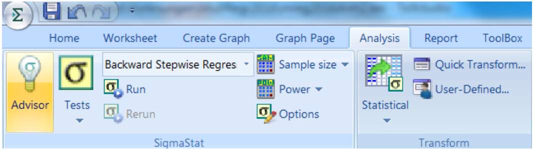
- Start with the full model
- In every step, remove the predictor which yields the smallest decline of the fit
- Stop, when the fit is considerably worsened

Variable Selection

Recommendation:

- Run forward selection and backward selection
- If both methods result in the same model, this choice can be considered as sufficiently robust





Results: Forward Selection

Step 2: "lcp" Entered
R = 0,803 Rsqr = 0,645 Adj Rsqr = 0,638
Standard Error of Estimate = 0,709

Analysis of Variance:

Group	DF	SS	MS	F	P
Regression	2	86,081	43,040	85,575	<0,001
Residual	94	47,278	0,503		

Variables in Model

Group	Coef.	Std. Coeff.	Std. Error	F-to-Remove	P
Constant	0,0913		0,205		
"lcp"	0,328	0,390	0,0619	28,119	<0,001
"lpsa"	0,532	0,521	0,0750	50,229	<0,001

Variables not in Model

Group	F-to-Enter	P
"lweight"	0,262	0,610
"age"	2,092	0,151
"lbph"	1,126	0,291

Summary Table

Step #	Vars. Entered	Vars. Removed	R	RSqr	Delta RSqr	Vars in Model
1	"lpsa"		0,734	0,539	0,539	1
2	"lcp"		0,803	0,645	0,106	2

The dependent variable "lcavol" can be predicted from a linear combination of the independent variables:

	P
"lcp"	<0,001
"lpsa"	<0,001

Results: Backward Selection

Step 3: Column E Removed
R = 0,803 Rsqr = 0,645 Adj Rsqr = 0,638
Standard Error of Estimate = 0,709

Analysis of Variance:

Group	DF	SS	MS	F	P
Regression	2	86,081	43,040	85,575	<0,001
Residual	94	47,278	0,503		

Variables in Model

Group	Coef.	Std. Coeff.	Std. Error	F-to-Remove	P
Constant	0,0913		0,205		
lcp	0,328	0,390	0,0619	28,119	<0,001
lpsa	0,532	0,521	0,0750	50,229	<0,001

Variables not in Model

Group	F-to-Enter	P
lweight	0,262	0,610
age	2,092	0,151
lbph	1,126	0,291

Summary Table

Step #	Vars. Entered	Vars. Removed	R	RSqr	Delta RSqr	Vars in Model
1		lweight	0,815	0,663	0,663	4
2		age	0,808	0,653	-0,0102	3
3		lbph	0,803	0,645	-0,00780	2

The dependent variable can be predicted from a linear combination of the independent variables:

	P
lcp	<0,001
lpsa	<0,001

The following variables did not significantly add to the ability of the equation to predict Column C and were not included in the final equation: lweight age lbph

Subset of two predictors:

Variables in Model Rsqr=0.64

Group	Coef.	Std. Coeff.	Std. Error	F-to-Remove	P
Constant	0,0913		0,205		
lcp	0,328	0,390	0,0619	28,119	<0,001
lpsa	0,532	0,521	0,0750	50,229	<0,001

The above subset shows a similar fit as the full model:

R = 0,815 Rsqr = 0,664 Adj Rsqr = 0,646

Standard Error of Estimate = 0,702

	Coefficient	Std. Error	t	P	VIF
Constant	-1,016	0,822	-1,236	0,220	
lweight	-0,0735	0,171	-0,429	0,669	1,412
age	0,0208	0,0105	1,975	0,051	1,197
lbph	-0,0812	0,0570	-1,425	0,158	1,334
lcp	0,307	0,0623	4,923	<0,001	1,479
lpsa	0,553	0,0797	6,933	<0,001	1,653

- Always additionally report the results for the full model.
- The *p*-values of the Wald Tests in the reduced model are wrong.

Conclusions/Guideline

1. Always visualise your data first!
2. Eventually transform response variables (and continuous predictors) to eliminate skewness
3. Check if your model contains highly correlated predictors (e.g `bmi` and `weight`, `psa` and `log(psa)`). Only keep one of the respective variables
4. Run the multiple linear regression.
5. Check assumptions using tests (Shapiro-Wilk...) and graphical tools (Residuals vs. Fitted Plot)
6. Check for multicollinearity (and outliers/influential observations)
7. Interpret the regression coefficients and the corresponding test results (Wald-test).

Your are very welcome to contact us ...

Get statistical support:

$\hookrightarrow$ contact us : biostatistics-consulting@dkfz.de

Department of Biostatistics (C060)

Outlook:

21 Oct 2020: Logistic regression

Model	Reponse Variable	Predictors
Linear Regression	continuous	continuous, categorical
Logistic Regression	binary	continuous, categorical
Cox PH regression	survival times	continuous, categorical