Fachhochschule Südwestfalen
University of Applied Sciences

Statistical Modeling

CH.3 - Regression Diagnostics

SS 2021 || Prof. Dr. Buchwitz

Wir geben Impulse

Outline

1 Organizational Information

2 Regression Diagnostics

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

1 Organizational Information

2 Regression Diagnostics

Introduction

- In this chapter we talk about the **standard regressions assumptions**, the consequences when violating them and how to detect violations so that we can focus on the remainder of the course on methods of how to correct or compensate for violations.
- When those assumptions are violated, the discussed and derived results for making inferences about the regression coefficients do not hold, which essentially means that **conclusions drawn on the corresponding models are wrong**.
- The majority of the discussed methods are **graphical methods** which means that they may be somewhat subjective here or there, which needs to be kept in mind when interpreting diagnostic plots.

Overview

1. Assumptions about the form of the model.
2. Assumptions about the errors.
3. Assumptions about the predictors.
4. Assumptions about the observations.

The properties of the least squares estimators (BLUE) are based on the discussed assumptions!

Assumption 1: Model

- The model that relates Y and $X_1, X_2, \ldots, X_p$ is assumed to be **linear in the regression parameters** $\beta_0, \beta_1, \ldots, \beta_p$ so that

$$\begin{aligned} \text{Model:} \quad & Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon \\ \text{Observation:} \quad & y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i, \ \ i = 1, 2, \ldots, n \end{aligned}$$

- This assumption is called the **linearity assumption**.
- In simple linear regression checking can be done using a scatterplot of Y versus X. For multiple linear regression there are other plotting techniques which we will discuss.
- When the linearity assumptin does not hold, transforming the data may lead to linearity (Transformations are discussed at a later point).

Assumption 2: Errors

- The errors $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are assumed to be **independently and identically distributed** (iid) normal random variables each with mean zero and commonon variance σ^2. This implies:
  - **Normality Assumption:** The error $\epsilon_i, i = 1, 2, \ldots, n$ has a normal distribution.
  - The errors $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ have mean zero.
  - **Constant Variance Assumption:** The errors have the same (but unknown) variance σ^2. When this assumtion does not hold we have the *heteroscedasticity problem*.
  - **Independent errors Assumption:** $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are independent of each other (pairwise covariances are zero). Violations lead to the *autocorrelation problem*.

Assumption 3: Predictors

- Thre are three assumptions for the predictor variables.
  - The predictor variables $X_1, X_2, \ldots, X_p$ are **nonrandom**. This means the values $x_{1j}, x_{2,j}, \ldots, x_{nj}$ with $j = 1, 2, \ldots, p$ are fixed (which is usually only fully satisfied under experimental conditions). In practice the results presented hold, but results are conditional on the data.
  - The values $x_{1j}, x_{2,j}, \ldots, x_{nj}$ are measured without error (which is hardly ever satisfied). In practice it is sufficient, when the measurement error is small compared to the random error ϵ_i.
  - The predictor variables $X_1, X_2, \ldots, X_p$ arre assumend to be linaerly independent of each other. This assumtion guarantees the uniqueness of the lest squares solution. If this assumption is violated this is to refeered as the *collinearity problem*.

The first two assumptions cannot be checked and do not play a role in our analysis. They have to be kept in mind when collecting data.

Assumption 4: Observations

- All observations are equally reliable and have an approximately equal role in determining the regression results. This means that they are equally relied on when drawing conclusions.

Assumption 4: Observations

- All observations are equally reliable and have an approximately equal role in determining the regression results. This means that they are equally relied on when drawing conclusions.

Conclusion

Small or minor violations of the underlying assumptions do not invalidate the inferences or conclusions drawn from the analysis. Gross validations, however, can seriously distort conlusions. **It is essential to investigate all signs of assumption validations by *always* checking the structure of the residuals and the data patterns at least using graphs!**

Residuals

- Analysing residuals is a simple and effective method for detecting model deficiencies in regression analysis. In most analysis it is probably the **most important** part of an analysis.
- Residual plots may lead to suggestiosn fo structure or point to informaiton in the data tha tmight be missed or overlooked. Those clues can lead to a better understanding (and possibly a better model) of the underlying process.
- Starting point for the analysis are the **ordinary** least squares residuals that can be calculated after obtained the fitted values:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip}$$
$$e_i = y_i - \hat{y}_i \quad \text{for} \quad i = 1, 2, \ldots, n$$

Residuals

- The fitted values can also be written as function of the predictor variables, where p_{ij} only depends on the predictor variables (essentially values from the hat matrix **P**).

$$\hat{y}_i = p_{i1}y_1 + p_{i2}y_2 + \ldots + p_{in}y_n$$

- When $i = j$ the value p_{ii} represents the weight (leverage) given to y_i in determining the i-th fitted value $\hat{y}_i$. The n **leverage values** $p_{11}, p_{22}, \ldots, p_{nn}$ can be obtained by

$$p_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$$

A high laverage value indicates some "extremeness" in X.

Residuals

- The *ordinary least squares residuals* $e_1, e_2, \ldots, e_p$ do not have **unequal variances** $Var(e_i) = \sigma^2(1 - p_{ii})$. Analyzing requires **standardized residuals** by calculating

$$z_i = \frac{e_i}{\sigma\sqrt{1 - p_{ii}}}$$

- This requires an unbiased estimate for the unknown standard deviation σ of ϵ for which we have two unbiased estimates to choose from

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n - p - 1} = \frac{SSE}{n - p - 1} \quad \text{with} \quad \hat{\sigma}^2_{(i)} = \frac{SSE_{(i)}}{(n-1) - p - 1} = \frac{SSE_{(i)}}{n - p - 2}$$

- $SSE_{(i)}$ is the sum of squared residuals when the i-th observation is left out so that the model is fitted using $n - 1$ observations.

Residuals

The choice of variance estimates results in two different types of residuals, although both are unbiased estimates.

Internally studentized residuals (using $\hat{\sigma}^2$)

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - p_{ii}}} \quad \text{with} \quad \hat{\sigma}^2 = \frac{SSE}{n - p - 1}$$

Externally studentized residuals (using $\hat{\sigma}^2_{(i)}$)

$$r_i^* = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1 - p_{ii}}} \quad \text{with} \quad \hat{\sigma}^2_{(i)} = \frac{SSE_{(i)}}{n - p - 2}$$

Called *externally studentized* because e_i is not involved in (external to) $\hat{\sigma}^2_{(i)}$.

In practice the difference between r_i and r_i^* is small and both could be used, so the difference is ignored in the following notation.

Graphical Methods

Dimensionality:

- One-dimensional graphs, inidcate the distribution of a particular variable (e.g. symmetry, skewedness) and allow identification of outliers.
- Two-dimensional graphs allow exploration of relationships (by pairing variables) and general patterns.

Step in Model Selection Process:

- Graphs **before** fitting a model, to e.g. correct data errors, seelect variables and preparation for model selection.
- Graphs **after** fitting a model to check assumptions and assessing the goodness of fit.

Example: Hamiltons Data

```
P103
```

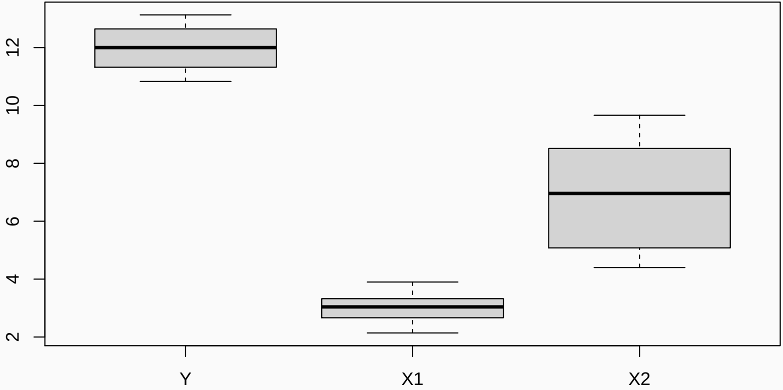
```
##        Y   X1   X2
## 1  12.37 2.23 9.66
## 2  12.66 2.57 8.94
## 3  12.00 3.87 4.40
## 4  11.93 3.10 6.64
## 5  11.06 3.39 4.91
## 6  13.03 2.83 8.52
## 7  13.13 3.02 8.04
## 8  11.44 2.14 9.05
## 9  12.86 3.04 7.71
## 10 10.84 3.26 5.11
## 11 11.20 3.39 5.05
## 12 11.56 2.35 8.51
## 13 10.83 2.76 6.59
## 14 12.63 3.90 4.90
## 15 12.46 3.16 6.96
```

Graphs before fitting a Model

1. Boxplot
2. Histogram
3. Pairsplot

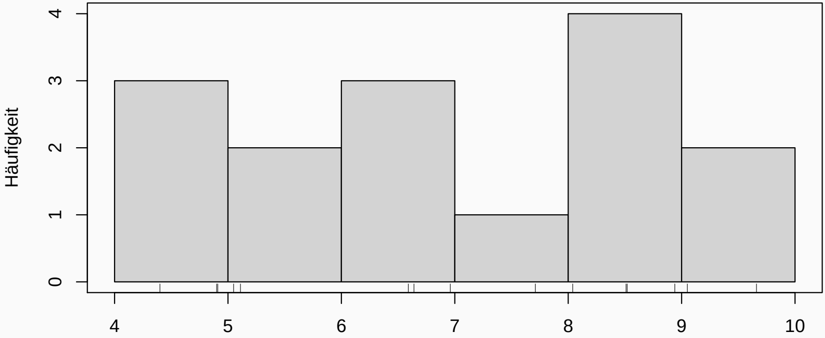
Boxplot

```r
boxplot(P103)
```



Histogram

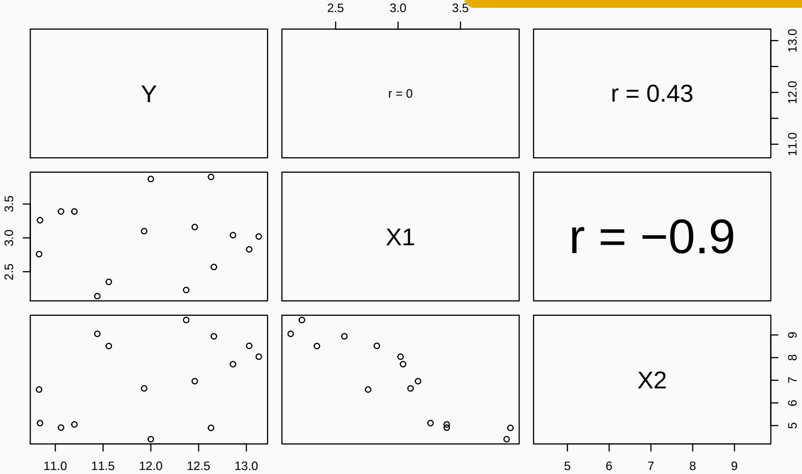
```r
hist(P103$X2, main = "", ylab = "Häufigkeit", xlab = "")
rug(P103$X2)
box()
```



Pairplot

Your turn

Test



Model Fitting

```r
mod <- ols_regress(Y ~ 1 + X1 + X2, data=P103)
mod
```

```
##                          Model Summary
## --------------------------------------------------------------
## R                       1.000       RMSE               0.011
## R-Squared               1.000       Coef. Var          0.089
## Adj. R-Squared          1.000       MSE                0.000
## Pred R-Squared          1.000       MAE                0.009
## --------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                  ANOVA
## ----------------------------------------------------------------------
##                 Sum of
##                Squares        DF    Mean Square        F          Sig.
## ----------------------------------------------------------------------
## Regression      9.007          2          4.504    39222.343    0.0000
## Residual        0.001         12          0.000
## Total           9.009         14
## ----------------------------------------------------------------------
##
##                                    Parameter Estimates
## -----------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta       t        Sig      lower     upper
## -----------------------------------------------------------------------------------------
## (Intercept)    -4.515         0.061                 -73.851    0.000    -4.649    -4.382
```

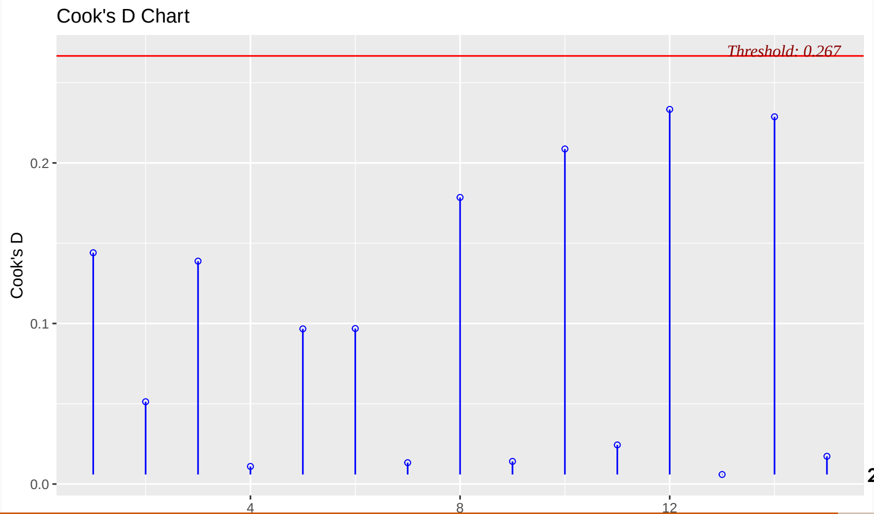
Graphs after fitting a Model

1. Graphs for cheking the linearity and normality assumptions
2. Graphs for the detection of ouliers and influential observations
3. Diagnostic plots for the effect of variables

Checking Linearity and Normality Assumptions

Influence

```
model <- lm(Y ~ 1 + X1 + X2, data = P103)
ols_plot_cooksd_chart(model)
```



BACKUP

$$Y = f(X_1, X_2, \ldots, X_p) + \epsilon$$

- We now generalize the simple linear regression model so that the relation between the response Y and p predictor variables $X_1, X_2, \ldots, X_p$ can be studied.
- We still assume that **within the range** of the data the true relation between Y and the predictors can be approximated using a linear function.
- The previously discussed simple linear regression model can be seen as a special case of the general linear regression model where $p = 1$.

Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

- Each regressor needs its own constant β so that the regression coefficients are now $\beta_0, \beta_1, \ldots, \beta_p$.
- The random disturbance is noted using ϵ. This term measures the discrepancy in the approximation and ϵ contains **no systematic information for determing** Y that is not already captured by the X's.