Predictive Modeling

Testing Model Assumptions: Residual Analysis

Mirko Birbaumer

HSLU T&A

1. Repetition: Simple Linear Regression
2. Linear Regression: Model Assumptions
3. Diagnostics Tools
4. Therapeutical Treatments

Repetition: Simple Linear Regression

- **Simple Linear Regression Model**:

$$Y = \beta_0 + \beta_1 X + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

  - Y: **response variable**
  - X: **predictor variable**
  - ε: **error term**

- Example: `Advertising` data set where we want to predict the response variable `sales` by means of the predictor variable advertising budget for `TV`

Repetition: Simple Linear Regression

- 95 % confidence interval for β_1 takes approximately the form

$$\left[\hat{\beta}_1 - 2 \cdot \mathrm{se}(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot \mathrm{se}(\hat{\beta}_1)\right]$$

  where

$$\mathrm{se}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

  denotes the **standard error** which corresponds to the average deviation of $\hat{\beta}_1$ from the true β_1

- $\sigma^2 = \mathrm{Var}[\varepsilon]$ cannot be observed

Repetition: Simple Linear Regression

- $\varepsilon = Y - (\beta_0 + \beta_1 X)$ cannot be measured since β_0 and β_1 are unknown

- *Approximation* for ε: **residuals** $r_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$

- Residual Standard Error (RSE)

$$\text{RSE} = \sqrt{\frac{RSS}{n-2}} = \sqrt{\frac{r_1^2 + r_2^2 + \ldots + r_n^2}{n-2}}$$

- $\hat{\sigma} = \text{RSE}$

Repetition: Simple Linear Regression

- 95% **confidence interval** for **expected value** of $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$ for a given value x_0, that is for $\mathrm{E}[\hat{y}|x_0]$

$$[\hat{y}_0 - 2 \cdot \mathrm{se}(\hat{y}_0), \hat{y}_0 + 2 \cdot \mathrm{se}(\hat{y}_0)]$$

where

$$\mathrm{se}(\hat{y}_0)^2 = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \right)$$

- **Interpretation**: with a probability of 95%, the true regression line (*population regression line*) passes through this interval for given x_0

Repetition: Simple Linear Regression

- 95% **prediction interval** for future observation y_0 at a given value x_0

$$[\hat{y}_0 - 2 \cdot \mathrm{se}(y_0), \hat{y}_0 + 2 \cdot \mathrm{se}(y_0)]$$

where

$$\mathrm{se}(y_0)^2 = \hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right)$$

- **Interpretation**: a future observation y_0 falls with a probability of 95% into this interval

- All these (theoretical) confidence and prediction intervals rely on the assumption $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

- How do we know whether these assumptions are fulfilled?

Model Assumptions for the Error Terms ε_i

Model Assumptions for the Error Terms ε_i

All test and estimation methods rely on **model assumptions**: The error terms ε_i are independent and normally distributed random variables with a constant variance:

$$\varepsilon_i \quad \text{iid} \quad \mathcal{N}(0, \sigma^2)$$

1. For the *expected value* of all ε_i we have
$$\mathrm{E}[\varepsilon_i] = 0$$
2. The error terms ε_i all have the same constant *variance*
$$\mathrm{Var}[\varepsilon_i] = \sigma^2$$
3. The error terms ε_i are *normally distributed*
4. The error terms ε_i are *independent*

Residual Analysis

- **Residual Analysis**: we will verify every assumption underlying the linear regression model by means of summary statistics and graphical methods

- Error term $\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$ is unknown, since β_0 and β_1 are unknown

- We however can determine the **residuals**: $r_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x)$ which are relevant to estimate the standard deviation of the error terms

Residual Analysis

Aim of Residual Analysis

If one or several model assumptions are violated, we should see this as a chance or starting point to adapt and/or extend our regression model to find a better and more adapted model (**explorative data analysis**)

Residual Analysis

The **RSE** (residual standard error) is an estimate of the standard deviation of ε. Roughly speaking, it is the average amount that the response will deviate from the true regression line.

$$\text{RSE} = \sqrt{\frac{r_1^2 + \ldots + r_n^2}{n-2}} = \sqrt{\frac{(y_1 - \hat{y}_1)^2 + \ldots + (y_n - \hat{y}_n)^2}{n-2}}$$

Residual Standard Error - RSE

- See the `Advertising` example `2.4` in the chapter `Simple Linear Regression`

- RSE $= 3.26$: actual `sales` in each among the 200 markets deviate from the true regression line by approximately 3260 units, on average.

- Mean value of `sales` over all markets is approximately 14 000 units, and so the percentage error is

$$\frac{3.260}{14.000} \approx 0.23 = 23\,\%$$

- RSE is considered a measure of the **lack of fit** of the regression model to the data. What constitutes a good RSE?

R^2 Statistic

The R^2 statistic provides an alternative measure of fit

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} = 1 - \frac{\text{variance left after regression fit}}{\text{total variance}}$$

- R^2 takes the form of a **proportion** - the proportion of variance explained: R^2 always takes on a value between 0 and 1, and is independent of the scale of Y
- If model fits perfectly the data, then $\hat{y}_i = y_i$ for all $i \Rightarrow \quad R^2 = 1$

R^2 Statistic

- Interpretation of R^2: proportion of the variance in the data that is **explained** by the regression model
  - R^2-value of approximately 1 means that a **large** part of the variance in the data is *explained* by the model (evt. in physics)
  - R^2-value near 0 indicates that **little** of the variance in the data is explained by the model (sometimes in social sciences)
- `Advertising`: `Multiple R-squared` yields $R^2 = 0.61$, approx. $2/3$ of variability in `sales` is explained by linear regression on `TV`
- See example 2.4 in the Simple Linear Regression chapter

R^2 Statistic

Correlation Coefficient

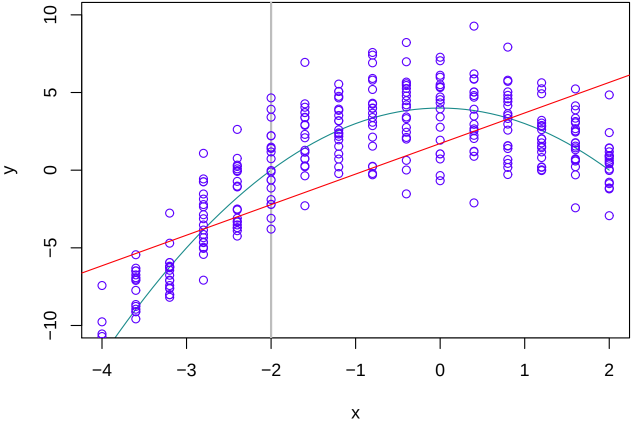
$$r = \mathrm{Cor}[X, Y] = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

is also a measure of the linear relationship between X and Y

- in simple linear regression setting: $r^2 = \mathrm{R}^2$
- R^2 statistic is a measure of the linear relationship between X and Y
- *Question*: Why not use $r = \mathrm{Cor}[X, Y]$ instead of R^2 in order to assess the fit of the linear model? *Answer*: Multiple Linear Regression
- See exercises on `Anscombe` data set (very high values of R^2 despite strong nonlinear relationship)

Diagnostics Tool for Testing Model Assumption $\mathrm{E}[\varepsilon] = 0$

The linear model assumes that there is a straight-line relationship between the predictor and the response. If f is **non-linear**, then model assumption $\mathrm{E}[\varepsilon_i] = 0$ is violated.



See example **2.3** in the **Testing Model Assumptions** chapter

Diagnostics Tool for Testing Model Assumption $\mathrm{E}[\varepsilon] = 0$

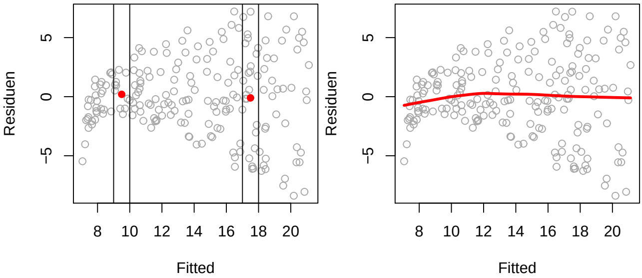
Goal: we want to *identify non-linearity* of the regression function f, that is, we want to verify the model assumption $\mathrm{E}[\varepsilon] = 0$; by means of the so-called **Tukey-Anscombe-Plot**.

Tukey-Anscombe-Plot:

- We plot on the vertical axis the **residuals** $r_i = y_i - \hat{y}_i$
- We plot on the horizontal axis the fitted or **predicted** values $\hat{y}_i$
- We thus plot the points $(\hat{y}_i, r_i)$ for $i = 1, \ldots, n$
- See example `2.4` in the `Testing Model Assumptions` chapter

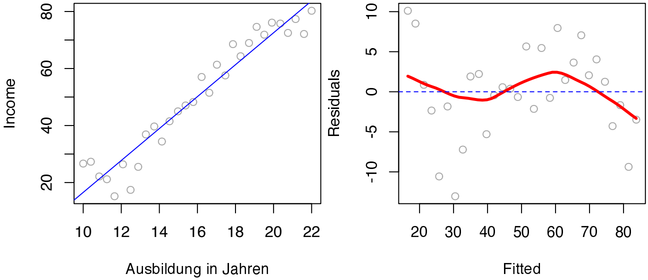
Tukey-Anscombe Plot: Smoothing Approach

The linear model fits the data well if the points in the Tukey-Anscombe plot scatter **evenly** around the $r = 0$ line. To visualize the relation between the residuals r_i and the predicted response values $\hat{y}_i$, we use the *smoothing approach*, in particular the LOESS smoother.



See Advertising example 2.5 in the Testing Model Assumptions chapter

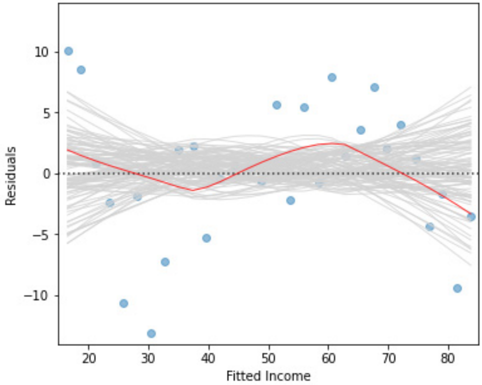
Example: `Income`



Question: How can we decide whether this wiggly smoothing curve systematically deviates from the $r = 0$ line and hence violates the assumption $\mathrm{E}[\varepsilon_i] = 0$ or when this is just due to a random variation?

Simulation of Plausible Smoothing Curves

Principle idea of resampling approach: simulating data points on the basis of the existing data set. For simulated data points we fit a smoothing curve and add it to Tukey-Anscombe plot.

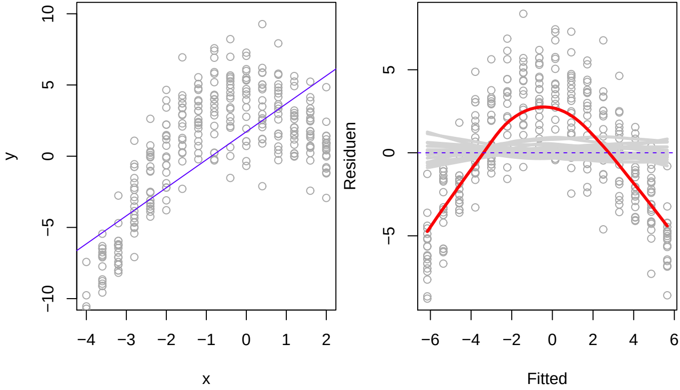


Simulation of Plausible Smoothing Curves

1. **Step 1** We keep the predicted values $\hat{y}_i$ as they are. Then, we assign to each $\hat{y}_i$ a *new* residual r_i^* which we obtained from sampling with replacement among the existing set of r_i
2. **Step 2** On the basis of the new data pairs $(\hat{y}_i, r_i^*)$, a smoothing curve is fitted, and is added to the Tukey-Anscombe plot as a grey line (the resampled data points are not shown)
3. **Step 3** The entire process is repeated for a number of times, e.g. one-hundred times.

See the `Income` example 2.6 in the `Testing Model Assumptions` chapter

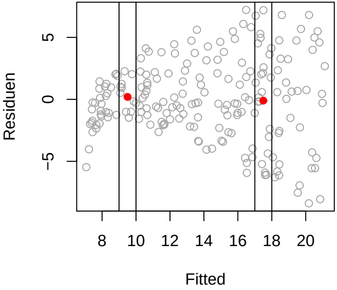
Tukey-Anscombe Plot for Non-Linear Regression Function

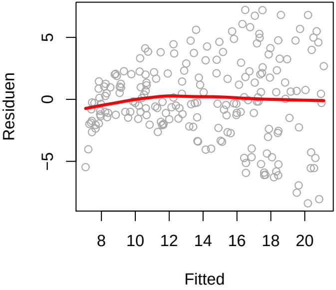


We compare red curve with 100 simulated smoothing curves to check whether deviation from $r = 0$ line is due to random variation or systematic.

Diagnostics Tool for Testing the Model Assumption $\mathrm{Var}[\varepsilon_i] =$ constant

- Non-constant variances in the errors ε_i: **heteroscedasticity**
- Example: `Advertising`





Testing the Model Assumption $\mathrm{Var}[\varepsilon_i] =$ constant

- Measure of scattering amplitude of errors: square root of the absolute value of the **standardized residuals**, that is

$$\sqrt{|\widetilde{r}_i|}$$

- **Standardized residuals** $\widetilde{r}_i$ are defined as follows

$$\widetilde{r}_i = \frac{r_i}{\hat{\sigma}\sqrt{1 - \left(\frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_i^n (x_i - \overline{x})^2}\right)}}$$

- $\hat{\sigma}$: estimate of standard deviation of error terms (estimated by RSE)

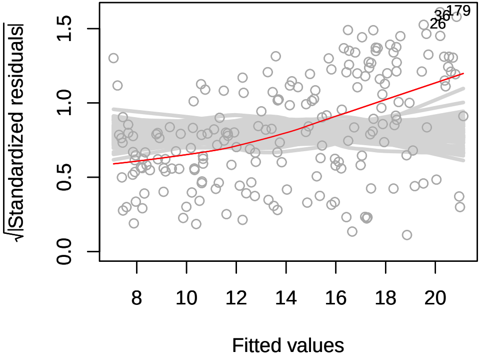
- If error terms ε_i are normally distributed, then

$$\widetilde{r}_i \sim \mathcal{N}(0, 1)$$

Scale-Location Plot

If we plot the square root of the absolute values of the standardized residuals versus the predicted values $\hat{y}_i$: **Scale-Location Plot**

See `Advertising` example `2.9` in `Testing Model Assumptions` chapter

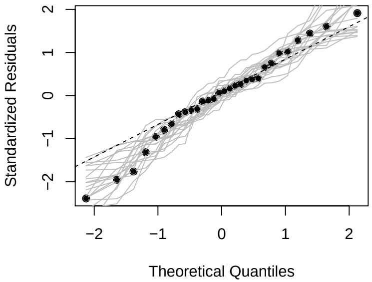


Red curve not within grey band of simulated curves: **heteroscedasticity**

Diagnostics Tool for the Normal Distribution Assumption of the Errors ε_i

We are not able to determine the error terms ε_i directly, we use the **standardized residuals** instead: $\tilde{r}_i$

We check the Normal Distribution Assumption of the errors by means of a normal plot.



See `Advertising` example `2.12` in `Testing Model Assumptions` chapter

Diagnostics Tool for Independence Assumption of Errors ε_i

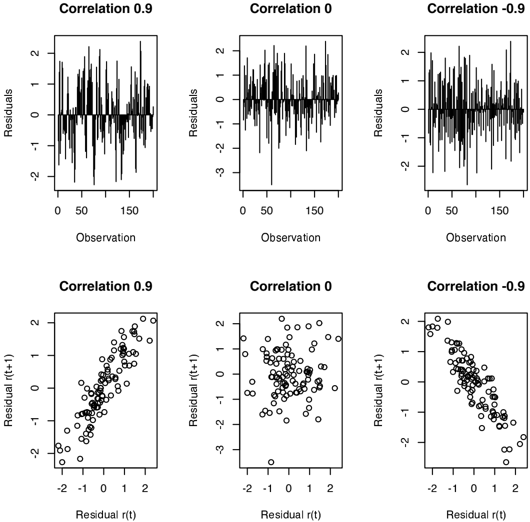
Example: the fact that ε_i is positive provides little or no information about the sign of ε_{i+1}

Consequences for case of correlated error terms

- The standard errors that are computed for the estimated regression coefficients or the fitted values are based on the assumption of **independent** error terms ε_i
- If there is correlation among the error terms, then the estimated standard errors will tend to **underestimate** the true standard errors. As a result, confidence and prediction intervals will be narrower than they should be

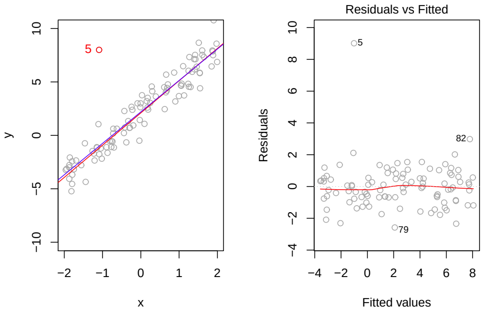
Diagnostics-Tool: if observations follow a time order

- Plot the residuals r_i from model as a function of time
- Generate scatter plot of the residuals r_{t+1} versus the residuals r_t



Outlier

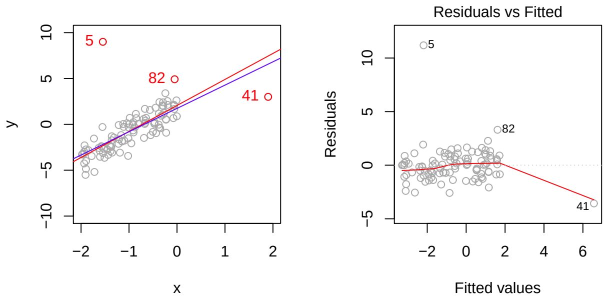
An **Outlier** is point for which y_i is far from value $\hat{y}_i$ predicted by model.



- Red regression line: **without** outlier; blue regression line: **with** outlier
- Removing outlier: **little** effect on β_0 and β_1
- BUT: **important** effect on RSE and R^2

High Leverage Points

Leverage Points: have an unusual value for x_i



- Blue regression line: **with** observation 41; red regression line: **without** observation 41
- Removing a high leverage observation has a much **more substantial** impact on the least squares line than removing an outlier

Leverage Points and Leverage Statistic h_i

Leverage Statistic:

$$h_i = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{i'=1}^{n} (x_{i'} - \overline{x})^2}$$

Properties h_i:

- h_i increases with the distance of x_i from $\overline{x}$
- High leverage point is a point having a large distance from the center of gravity $\overline{x}$ - *high momentum to turn the regression line around*
- h_i is always between $1/n$ and 1
- Average leverage for all the observations is always equal to $2/n$ (simple linear regression)

Leverage Points and Leverage Statistic h_i

For which values of h_i do we consider an observation as **high leverage point**?

- if a given observation has a leverage statistic that greatly **exceeds** $2/n$, then we may suspect that the corresponding point has **high leverage**

Cook's Distance

Cook's distance: measures the influence of an observation i

$$d_i = \frac{1}{\hat{\sigma}^2} \cdot \left(\underline{\hat{y}}_{(-i)} - \underline{\hat{y}}\right)^T \left(\underline{\hat{y}}_{(-i)} - \underline{\hat{y}}\right)$$

- $\underline{\hat{y}}_{(-i)}$ denotes the vector of predicted values if the ith observation is *removed*

Properties of Cook's Distance

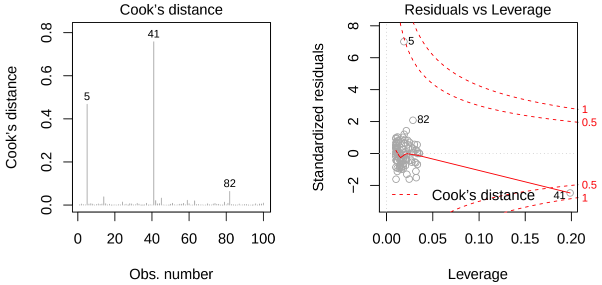
- Cook's distance d_i may be expressed as a function of the leverage statistic h_i and the standardized residual $\widetilde{r}_i$:

$$d_i = \widetilde{r}_i^2 \frac{h_i}{2(1 - h_i)}$$

- The larger the value of Cook's distance d_i is, the **higher** is the **influence** of observation i on the estimation of the predicted value $\hat{y}_i$

- An observation with a value of Cook's distance larger than 1 is considered as **dangerously influential**

- Cook's distances are shown either as bar plots or as **contour lines** in a scatter plot with standardized residuals versus leverage statistic

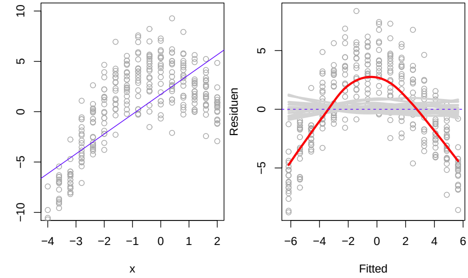


- Observation 41 is a high leverage point, but has a relatively small standardized residual value: **potentially dangereous**
- Observation 5 has a large residual value, but its leverage statistic is rather small: **not dangerous**

Example: Advertising

See examples 2.14 and 2.15 in the Testing Model Assumptions chapter

Therapeutical Treatment in the case of $\mathrm{E}[\varepsilon_i] \neq 0$

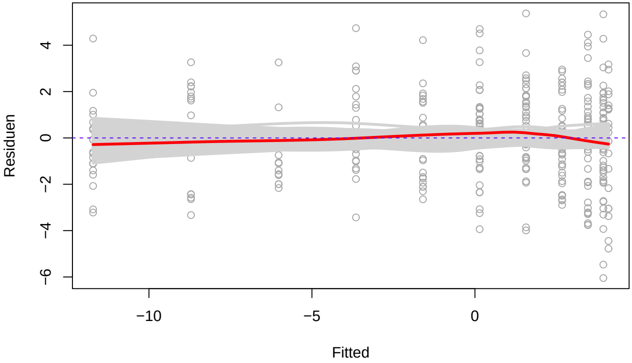


- If Tukey-Anscombe plot indicates **non-linear** structure in the data (f is non-linear), then a non-linear transformation of predictor such as
  - $\widetilde{X} = \log(X)$
  - $\widetilde{X} = \sqrt{X}$
  - $\widetilde{X} = X^2$

  may help to establish a **linear** relationship between transformed variable $\widetilde{X}$ and response variable Y

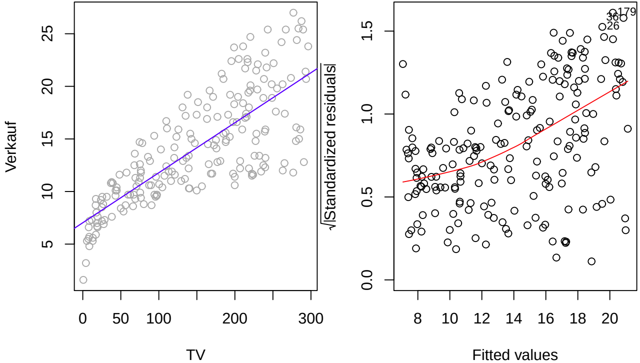
Therapeutical Treatment in the case of $\mathrm{E}[\varepsilon_i] \neq 0$

Solution for previous problem: variable transformation $\widetilde{X} = X^2$



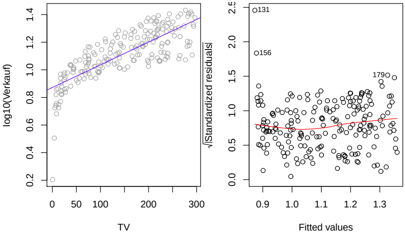
Therapeutical Treatment for $\mathrm{Var}[\varepsilon_i] \neq$ constant

The scattering magnitude of the residuals increases with the predicted values $\hat{y}_i$



Therapeutical Treatment: log-transformation of the response variable Y may lead to a constant variance

Therapeutical Treatment for $\mathrm{Var}[\varepsilon_i] \neq$ constant



Tukey's first aid principles

- log-transformation for concentration data and absolute values
- square root transformation for count data (discrete random variables)
- arcsine-transformation $\widetilde{Y} = \arcsin(\sqrt{Y})$ or the logit-transformation $\widetilde{Y} = \log\left(\frac{Y+0.005}{1.01-Y}\right)$ for percentage data

Therapeutic Treatment in the Case of Outliers and High Leverage Points

- **Fundamental Consideration for Outliers**: an observation is considered as an outlier with respect to a given model that is not fitting this observation

- **Variable transformations** may change the model so that the new model suddenly fits the observation that previously was considered an outlier: don't forget your ambitions for a Nobel Prize!

Therapeutic Treatment in the Case of Outliers and High Leverage Points

Procedure:

1. Check whether outlier has occured due to an error in data collection or recording
   - If an error may have occured: omit the data point
   - If an error can be excluded: go to 2
2. Attempt to transform predictor or response variables in order to make *disappear* the outlier. If no improvement, go to 3
3. Outlier occured due to an unusual random variation: If such outliers affect parameter estimations too much, then the observation may be removed (needs to be mentioned in the reports!)