# Predictive Modeling

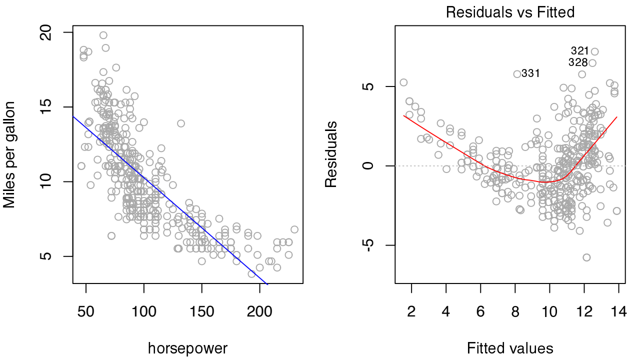
## Polynomial Regression, Residual Analysis and Model Selection

Mirko Birbaumer

HSLU T&A

# Non-linear Relationships and Polynomial Regression

**Example**: `Auto` data set: `mpg` (gas mileage in miles per gallon) versus `horsepower` is shown for a number of cars, see example 4.13 in the Multiple Linear Regression chapter

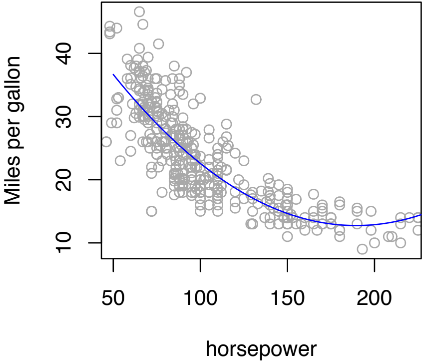



**Conclusion**: relationship between `mpg` and `horsepower` is **non-linear**

# Polynomial Regression - Example `Auto`

New Model:

$$\texttt{mpg} = \beta_0 + \beta_1 \cdot \texttt{horsepower} + \beta_2 \cdot \texttt{horsepower}^2 + \varepsilon$$

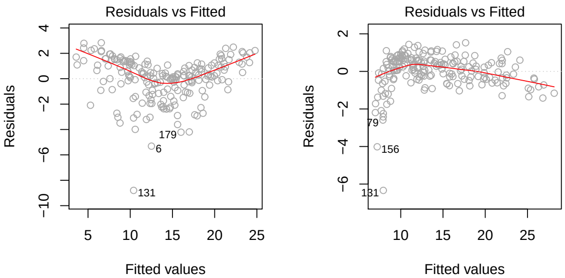



See example 4.13 in the `Multiple Linear Regression` chapter

# Residual Analysis: Example `Advertising`

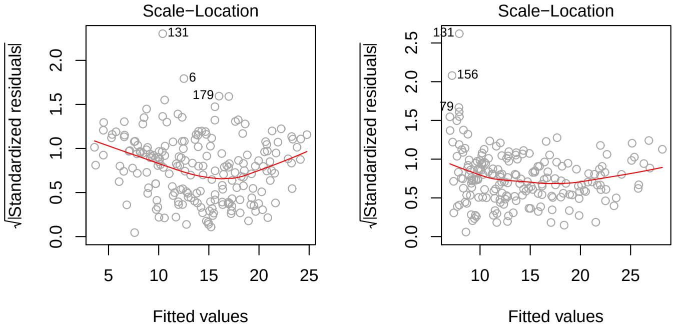
$$\texttt{sales} = \beta_0 + \beta_1 \cdot \texttt{TV} + \beta_2 \cdot \texttt{radio} + \varepsilon$$

$$\texttt{sales} = \beta_0 + \beta_1 \cdot \texttt{TV} + \beta_2 \cdot \texttt{radio} + \beta_3 \cdot \texttt{TV} \cdot \texttt{radio} + \varepsilon$$




**Tukey-Anscombe plots** for the two models; *left* with predictor variables `TV` and `radio`; *right* with predictor variables `TV`, `radio` and interaction term `radio` $\cdot$ `TV`. See example 4.14 in the Multiple Linear Regression chapter.

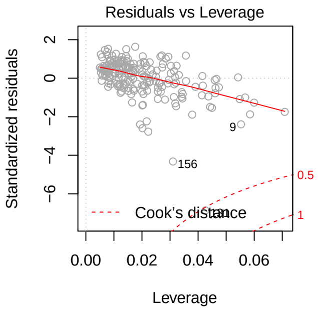
# Residual Analysis: Example `Advertising`




**Scale-location plot** for two models: *left* with predictor variables `TV` and `radio`; *right* with predictor variables `TV`, `radio` and interaction term `radio` · `TV`. See example `4.14` in the `Multiple Linear Regression` chapter

# Leverage Statistic and Cook's Distance: `Advertising`

Scatter plots for model including the interaction term with **leverage statistic** $h_i$ and **standardized residual** $\tilde{r}_i$. Contour lines of Cook's distance with $d_i = 0.5, 1$ are plotted as well.




*Left*: **with** outliers 131 and 156 ; *right* **without** observations 131 and 156 $\Rightarrow$ not dangerously influential! See example 4.14 in the `Multiple Linear Regression` chapter

# Example `Credit`

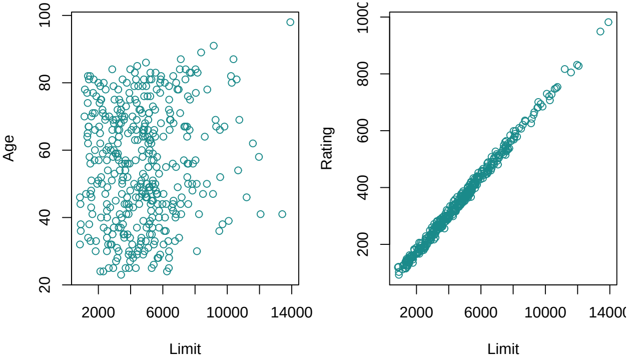
Data set `Credit` was recorded in the USA:

- **Response Variable** `balance` : average credit card debt for a number of individuals
- **Quantitative predictor variables**:
  - `age`
  - `cards` : number of credit cards
  - `education` : years of education
  - `income` : income in thousand of dollars
  - `limit` : credit card limit
  - `rating` : credit rating
- **Qualitative predictor variables (factors)**:
  - `gender`
  - `student` : student status
  - `ethnicity` : Caucasian, African Amercian or Asian
- Regression of `balance` (as response variable) on `age`, `rating` and `limit`

# Collinearity - Example `Credit`

**Collinearity** refers to the situation in which two or more predictor variables are closely related to one another

**Example**: Scatter plots of `Credit` data set: `age` versus `limit` and `rating` versus `limit`.



# Collinearity: Example `Credit`

| | | Coefficient | Std.error | t-statistic | p-value |
|---|---|---|---|---|---|
| Model 1 | `Intercept` | $-173.411$ | 43.828 | $-3.957$ | $< 0.0001$ |
| | `age` | $-2.292$ | 0.672 | $-3.407$ | 0.0007 |
| | `limit` | 0.173 | 0.005 | 34.496 | $< 0.0001$ |
| Model 2 | `Intercept` | $-377.537$ | 45.254 | $-8.343$ | $< 0.0001$ |
| | `rating` | 2.202 | 0.952 | 2.312 | 0.0213 |
| | `limit` | 0.025 | 0.064 | 0.384 | 0.7012 |

- **Model 1**: p-values of `age` and `limit` are highly significant
- **Model 2**: Collinearity between `limit` and `rating` causes the standard errors of coefficient estimate for `limit` to increase by a factor of 12 and the p-value to increase to 0.701
- Importance of the `limit` variable has been masked due to the presence of collinearity

# Identification of Collinearity in the Data

- **Correlation matrix**: very hight correlation between `limit` and `rating` : 0.997

- See example `4.17` in the `Multiple Linear Regression` chapter

- Not all collinearity problems can be detected by inspection of the correlation matrix: correlation matrix reveals only correlation between **two** variables

- Collinearity may occur between three or more variables even if no pair of variables has a particularly high correlation: **multicollinearity**

# Identification of Collinearity in the Data

- **Variance inflation factor** (VIF) to identify **multicollinearity**

$$\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1 - \mathrm{R}^2_{X_j|X_{-j}}}$$

$\mathrm{R}^2_{X_j|X_{-j}}$ represents the $\mathrm{R}^2$-value for a regression model of $X_j$ (response variable) onto all of the other predictors

  - **VIF**-value between 5 and 10 : indicates a problematic amount of collinearity
  - Smallest possible value of **VIF** is 1 : indicates complete absence of collinearity

# Identification of Collinearity in the Data

- **Example** `Credit`: Regression of `balance` (as response variable) on `age`, `rating`, and `limit` indicates that predictors have considerable collinearity.

- **VIF** values of
  - `age` : 1.01
  - `rating` : 160.67
  - `limit` : 160.59 considerable collinearity

- See examples 4.18 and 4.19 in the `Multiple Linear Regression` chapter

# The Marketing Plan

`Advertising` data set: `sales` of a particular product depending on the advertising budgets for `TV`, `radio` and `newspaper`

- *Is there a relationship between* `sales` *and advertising budget?*
- *How strong is the relationship between* `sales` *and budget?*
- *Which media contribute to* `sales`*?*
- *How large is the effect of each medium on* `sales`*?*
- *How accurately can we predict future* `sales`*?*
- *Is the relationship linear?*
- *Is there synergy among the advertising media?*

# Example: `Advertising`

See example 5.1 in the `Multiple Linear Regression` chapter

## 1. Is there a relationship between `sales` and advertising budget?

This question can be answered by fitting a multiple regression model of `sales` onto `TV`, `radio`, and `newspaper`

$$\texttt{sales} = \beta_0 + \beta_1 \cdot \texttt{TV} + \beta_2 \cdot \texttt{radio} + \beta_3 \cdot \texttt{newspaper} + \varepsilon$$

1. We test the null hypothesis

$$H_0 : \quad \beta_{\texttt{TV}} = \beta_{\texttt{radio}} = \beta_{\texttt{newspaper}} = 0$$

2. p-value associated with F-Statistic (`F-Statistic`) is approx. zero $\Rightarrow$ we **reject** null hypothesis

3. **Conclusion**: There is a relationship between advertising and sales

# 2. How strong is the relationship between `sales` and budget?

Two measures to assess **model accuracy**:

- RSE (`Residual Standard error`): average deviation of the response from the (true) population regression line
  - For the `Advertising` data, the RSE is 1.681 units
  - Mean value for the response is 14.022, indicating a percentage error of roughly 12 %
- $R^2$-value (`Multiple R-squared`): records the percentage of variability in the response that is explained by the predictors
  - The predictors explain almost 90 % of the variance in `sales`

# 3. Which media contribute to `sales`?

This question is answered by considering the p-values associated with each predictor's t-statistic (`t value`):

- In the multiple linear regression analysis, the p-values (`Pr(>|t|)`) for `TV` and `radio` are low, but the p-value for `newspaper` is not.

- This suggests that only `TV` and `radio` are related to `sales`

- Systematic discussion: see next chapter about **variable selection**

# 4. How large is the effect of each medium on **sales**?

This question is answered by confidence intervals for the regression coefficients $\beta_j$:

```
Advertising <- read.csv("../Daten/Advertising.csv")
round(confint(lm(sales ~ TV + radio + newspaper, data = Advertising)),
    digits = 3)

##                2.5 % 97.5 %
## (Intercept)  2.324  3.554
## TV           0.043  0.049
## radio        0.172  0.206
## newspaper   -0.013  0.011
```

- Confidence intervals for **TV** and **radio** for $\beta_j$ are narrow and far from zero, providing evidence that these media are ralated to **sales**
- Confidence interval for **newspaper** includes zero, indicating that the variable is not statistically significant given the values of **TV** and **radio**

# 5. How accurately can we predict future `sales`?

There are two possibilities to quantify the accuracy of a prediction:

- We wish to predict an individual response $Y = f(X_1, \ldots, X_p) + \varepsilon$
  $\Rightarrow$ **Prediction interval**
- We wish to predict the average response $Y$
  $\Rightarrow$ **Confidence interval**

# 6. Is the relationship linear?

- **Residual plots** (in particular Tukey-Anscombe plot) showed in the case of the `Advertising` data a pattern that reveals a **non-linear** relationship
- If the relationships are **linear**, then the residual plots should display **no** pattern
- **Solution**: Taking **interaction effects** into account

# 7. Is there synergy among the advertising media?

- Standard linear regression model assumes an **additive** relationship between the predictors and the response
- An additive model is easy to interpret because the effect of each predictor on the response is **unrelated** to the values of the other predictors
- Including an **interaction term** in the model results in a substantial increase in $R^2$, from around 90 % to almost 97 %

# Variable Selection: Example `Credit`

Data set `Credit` was recorded in the USA:

- **Response Variable** `balance` : average credit card debt for a number of individuals
- **Quantitative predictor variables**:
  - `age`
  - `cards` : number of credit cards
  - `education` : years of education
  - `income` : income in thousand of dollars
  - `limit` : credit card limit
  - `rating` : credit rating
- **Qualitative predictor variables (factors)**:
  - `gender`
  - `student` : student status
  - `ethnicity` : Caucasian, African Amercian or Asian

**Question**: From which subset consisting of $q$ predictor variables results the **best** model? Number of possible models: $2^p$

# Example `Credit` : Forward Stepwise Selection

1. We begin with the **null model** $\mathcal{M}_0$ which contains no predictors

$$\texttt{Balance} = \beta_0 + \varepsilon$$

2. We **add** a predictor variable to the null model: See example 2.1 in the `Linear Model Selection` chapter

3. We now choose the **best** variable in the sense that adding this variable leads to the regression model with the lowest RSS or the highest $R^2$ : `Rating`
New model: $\mathcal{M}_1$

$$\texttt{Balance} = \beta_0 + \beta_1 \cdot \texttt{Rating} + \varepsilon$$

# Example `Credit` : Forward Stepwise Selection

4. We now add a further predictor variable to the model $\mathcal{M}_1$ which leads, when added, to the lowest RSS, etc.

5. Repetition of this procedure until we have obtained 11 models $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_{10}$

6. We select the single **best** model on the basis of one of the following criteria: AIC, BIC, $C_p$ or adjusted $R^2$

See example `2.1` in the `Linear Model Selection` chapter

# Forward Stepwise Selection

**Algorithm: Forward stepwise selection**

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
    1. Consider all $p-k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
    2. Choose the *best* among these $p-k$ models, and call it $\mathcal{M}_{k+1}$. Here **best** is defined as having smallest RSS or highest $\mathrm{R}^2$.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using $C_p$, AIC, BIC or adjusted $\mathrm{R}^2$.

# Backward Stepwise Selection: Example `Credit`

1. We begin with the **full model**, that is $\mathcal{M}_{10}$, which contains **all** $p$ predictors of the `Credit` data set

$$\begin{aligned}\texttt{Balance} = \beta_0 &+ \beta_1 \cdot \texttt{Income} + \beta_2 \cdot \texttt{Limit} + \beta_3 \cdot \texttt{Rating} + \beta_4 \cdot \texttt{Cards} \\ &+ \beta_5 \cdot \texttt{Age} + \beta_6 \cdot \texttt{Education} + \beta_7 \cdot \texttt{Gender} + \beta_8 \cdot \texttt{Student} \\ &+ \beta_9 \cdot \texttt{Married} + \beta_{10} \cdot \texttt{Ethnicity} + \varepsilon\end{aligned}$$

2. We **remove** one predictor variable from the model: see example `2.2` in the `Linear Model Selection` chapter

3. We remove the **least useful** variable: the one yielding the reduced regression model with the lowest RSS or the highest $\mathrm{R}^2$. Its removal improves the model most significantly with respect to RSS. Most redundant variable here: `Education`

# Backward Stepwise Selection: Example `Credit`

3. **New Model**: $\mathcal{M}_9$

$$\begin{aligned}\texttt{Balance} = \beta_0 &+ \beta_1 \cdot \texttt{Income} + \beta_2 \cdot \texttt{Limit} + \beta_3 \cdot \texttt{Rating} + \beta_4 \cdot \texttt{Cards} \\ &+ \beta_5 \cdot \texttt{Age} + \beta_6 \cdot \texttt{Gender} + \beta_7 \cdot \texttt{Student} + \beta_8 \cdot \texttt{Married} \\ &+ \beta_9 \cdot \texttt{Ethnicity} + \varepsilon\end{aligned}$$

4. We iterate this procedure until **no** predictor is left in regression model

5. This iterative procedure yields 11 different models: $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_{10}$

6. We identify the **best** among these models on the basis of AIC, BIC, $C_p$ or adjusted $R^2$

See example `2.2` in the `Linear Model Selection` chapter

# Backward Stepwise Selection

**Algorithm: Backward stepwise selection**

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all $p$ predictors.
2. For $k = p, p-1, \ldots, 1$:
   1. Consider all $k$ models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors
   2. Choose the *best* among these $k$ models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest $\mathrm{R}^2$.
3. Select a single best model among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using $C_p$, AIC, BIC or adjusted $\mathrm{R}^2$.

# Best Subset Selection

- Model with $p$ predictor variables has $2^p$ possible submodels

- If $p = 20$ : $1\,048\,576$ models need to be fitted and evaluated
  $\Rightarrow$ **Best subset selection**: computationally infeasible for $p > 40$

- Comparison: Forward stepwise selection with $p = 20$ predictor variables leads to 211 models

- If computationally feasible: Go for it!

# Hybrid stepwise selection

- **Hybrid stepwise selection**:
  - We start with a model containing $k$ predictor variables
  - RSS of all models that result from adding to or removing each variable from the reference model is calculated
  - We iterate this procedure until the RSE stops decreasing
- Result is similar to best subset selection while retaining computational advantages of forward and backward stepwise selection

# Model Selection Criteria - Adjusted $R^2$

- Recall: $R^2$ is defined as follows

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum(y_i - \widehat{y}_i)^2}{\sum(y_i - \overline{y})^2}$$

- **Problem**: RSS always **decreases** as more predictors are added to the model $\Rightarrow$ $R^2$ always **increases** as more predictors are added

- **Solution**: add *penalty* to RSS which penalizes adding further predictor variables

# Model Selection Criteria - Adjusted $\mathrm{R}^2$

- **adjusted $\mathbf{R}^2$** is defined as

$$\text{adjusted } \mathrm{R}^2 = 1 - \frac{\mathrm{RSS}/(n-p-1)}{\mathrm{TSS}/(n-1)}$$

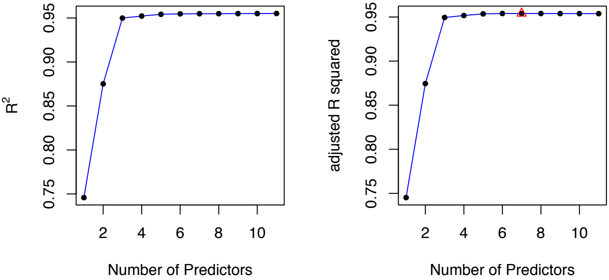
  $p$ : # predictor variables of least squares model
  $n$ : # data points

- To **maximize** adjusted $\mathrm{R}^2 \Rightarrow$ **minimize**

$$\frac{\mathrm{RSS}}{n-p-1}$$

- See example `2.3` in the `Linear Model Selection` chapter

# adjusted $R^2$ - Example `Credit`




$R^2$ : values are steadily increasing, whereas adjusted $R^2$ reaches maximum for **seven** predictors (see example 2.3) $\Rightarrow$ Best regression model among 11 models found by *forward stepwise selection*:

$$\texttt{Balance} = \beta_0 + \beta_1 \cdot \texttt{Income} + \beta_2 \cdot \texttt{Limit} + \beta_3 \cdot \texttt{Rating} + \beta_4 \cdot \texttt{Cards} \\ + \beta_5 \cdot \texttt{Age} + \beta_6 \cdot \texttt{Gender} + \beta_7 \cdot \texttt{Student} + \varepsilon$$

# AIC - Akaike information criterion

- **AIC** considers **goodness-of-fit** to the data and **penalizes** complexity of the model

$$\text{AIC} = -2\log(L) + 2q$$

where $L$ denotes the value of the likelihood function for a particular model and $q$ is the number of variables of this model.

- If errors $\varepsilon$ in linear regression model follow a normal distribution with expected value 0 and constant variance, then the **AIC** is

$$\text{AIC} = \frac{1}{n\hat{\sigma}^2}\left(\text{RSS} + 2p\hat{\sigma}^2\right)$$

  - $\hat{\sigma}$ : estimated standard deviation
  - $2p\hat{\sigma}^2$ is the **penalty term**: increases if more predictors are added to the model compensating the decrease in the RSS

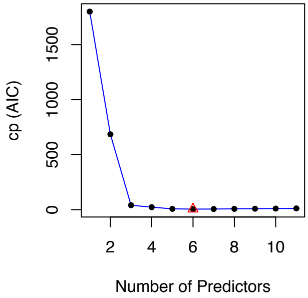
# Mallow's $C_p$-statistic

- For least squares models: AIC is proportional to **Mallow's** $C_p$-statistic

$$C_p = \frac{1}{n}\left(\mathrm{RSS} + 2p\hat{\sigma}^2\right)$$

- See example `2.4` in the `Linear Model Selection` chapter

# Model Selection with AIC: Example `Credit`




**Best Model** among 11 models found by *forward stepwise selection*: model with 6 predictor variables, see example 2.5

$$\texttt{Balance} = \beta_0 + \beta_1 \cdot \texttt{Income} + \beta_2 \cdot \texttt{Limit} + \beta_3 \cdot \texttt{Rating} + \beta_4 \cdot \texttt{Cards} + \beta_5 \cdot \texttt{Age} + \beta_6 \cdot \texttt{Student} + \varepsilon$$

# BIC - Bayesian information criterion

- The **BIC** is defined as

$$\text{BIC} = -2\log(L) + 2\log(n)q$$

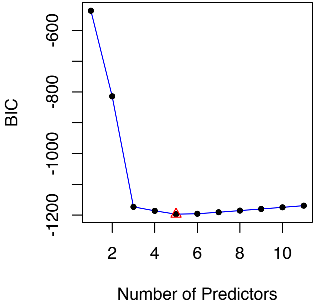
  where $L$ denotes the likelihood function for a particular model and $q$ is the number of estimated parameters of the model

- For the least squares model with $p$ predictors, the BIC is, up to irrelevant constants, given by

$$\text{BIC} = \frac{1}{n}\left(\text{RSS} + \log(n)p\hat{\sigma}^2\right)$$

  - $\hat{\sigma}$ : estimated standard deviation
  - $\log(n)p\hat{\sigma}^2$ penalty term: increases BIC when more predictors are added to the model

# Model Selection with BIC: Example Credit




**Best model** among 11 models found by forward stepwise selection: model with 5 predictor variables, see example 2.6

$$\begin{aligned}\texttt{Balance} = \beta_0 + \beta_1 \cdot \texttt{Income} + \beta_2 \cdot \texttt{Limit} + \beta_3 \cdot \texttt{Rating} + \beta_4 \cdot \texttt{Cards} \\ + \beta_5 \cdot \texttt{Student} + \varepsilon\end{aligned}$$