Regression & Correlation Analyzer

Mathematical Documentation

English

Wolfgang Boehme

wolfgang.boehme@economic-modeling.de

www.economic-modeling.de

6 November 2026

The documentation on this website was developed with the support of the AI Claude (Anthropic).

Table of Contents

1 Descriptive Statistics **3**
- 1.1 Measures of Location . . . 3
  - 1.1.1 Arithmetic Mean . . . 3
  - 1.1.2 Median . . . 3
  - 1.1.3 Quantiles . . . 3
- 1.2 Measures of Dispersion . . . 3
  - 1.2.1 Variance . . . 3
  - 1.2.2 Standard Deviation . . . 3
  - 1.2.3 Interquartile Range (IQR) . . . 4
- 1.3 Shape Parameters . . . 4
  - 1.3.1 Skewness . . . 4
  - 1.3.2 Kurtosis . . . 4

2 Correlation Analysis **4**
- 2.1 Pearson Correlation . . . 4
- 2.2 Spearman Rank Correlation . . . 5

3 Simple Linear Regression **5**
- 3.1 Model . . . 5
- 3.2 Parameter Estimation (OLS) . . . 5
- 3.3 Coefficient of Determination (R^2) . . . 5
- 3.4 Error Measures . . . 5
  - 3.4.1 Root Mean Square Error (RMSE) . . . 5
  - 3.4.2 Mean Absolute Error (MAE) . . . 5

4 Polynomial Regression **6**
- 4.1 Model of Degree p . . . 6
- 4.2 Matrix Notation . . . 6
- 4.3 Normal Equations . . . 6

5 Multiple Linear Regression **6**
- 5.1 Model . . . 6
- 5.2 Adjusted R^2 . . . 6
- 5.3 Multicollinearity: Variance Inflation Factor (VIF) . . . 6

6 Nonlinear Regression Models **7**
- 6.1 Exponential Model . . . 7
- 6.2 Power Model . . . 7
- 6.3 Logarithmic Model . . . 7
- 6.4 Logistic Function . . . 7
- 6.5 Gompertz Function . . . 7
- 6.6 Gaussian Function . . . 7

7 Residual Analysis **7**
- 7.1 Residuals . . . 7
- 7.2 Standardized Residuals . . . 8
- 7.3 Leverage . . . 8

7.4 Cook's Distance . . . 8
7.5 Q-Q Plot . . . 8

8 Model Selection **8**
8.1 Akaike Information Criterion (AIC) . . . 8

9 Example 1: Consumption Function USA 1960–2008 **9**
9.1 Data Set . . . 9
9.2 Descriptive Statistics . . . 9
9.3 Correlation Analysis . . . 10
9.4 Simple Linear Regression: Keynesian Consumption Function . . . 10
9.5 Residual Analysis . . . 10
9.6 Multiple Regression . . . 11
9.7 Model Comparison: Growth over Time . . . 11

10 Example 2: Engel Curve **11**
10.1 Power Model and Income Elasticity . . . 11
10.2 Multiple Regression with Household Size . . . 12

11 Example 3: Cobb-Douglas Production Function **12**
11.1 Linearization and Multiple Regression . . . 12
11.2 Intensive Form . . . 12

12 Summary: Model Comparison **13**

13 References **13**

14 Appendix: Numerical Methods **13**
14.1 Gaussian Elimination . . . 13
14.2 Approximation of the Inverse Error Function . . . 14

Abstract

This documentation describes the mathematical foundations and formulas of the regression and correlation analyzer. All implemented statistical procedures are formally defined, including descriptive statistics, correlation analysis, linear and nonlinear regression models, and diagnostic procedures. Three economic examples – consumption function (real data), Engel curve and Cobb-Douglas production function – illustrate the application of the methods; they are included in the tool as example datasets.

1 Descriptive Statistics

Let $X = \{x_1, x_2, \ldots, x_n\}$ be a sample with n observations.

1.1 Measures of Location

1.1.1 Arithmetic Mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{1}$$

1.1.2 Median

The median is defined as:

$$\text{Median}(X) = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & \text{if } n \text{ is even} \end{cases} \tag{2}$$

where $x_{(i)}$ denotes the i-th order statistic (sorted values).

1.1.3 Quantiles

The p-quantile ($0 \leq p \leq 1$) is defined by linear interpolation:

$$Q_p = x_{(k)} + (x_{(k+1)} - x_{(k)}) \cdot r \tag{3}$$

where $h = p(n-1)$, $k = \lfloor h \rfloor$, and $r = h - k$.

1.2 Measures of Dispersion

1.2.1 Variance

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{4}$$

1.2.2 Standard Deviation

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{5}$$

1.2.3 Interquartile Range (IQR)

$$\text{IQR} = Q_{0.75} - Q_{0.25} \tag{6}$$

1.3 Shape Parameters

1.3.1 Skewness

Skewness measures the asymmetry of the distribution:

$$\gamma_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3 \tag{7}$$

Interpretation:

- $\gamma_1 \approx 0$: symmetric distribution
- $\gamma_1 > 0$: right-skewed distribution (longer right tail)
- $\gamma_1 < 0$: left-skewed distribution (longer left tail)

1.3.2 Kurtosis

Kurtosis measures the "heaviness of the tails":

$$\gamma_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)} \tag{8}$$

Interpretation (excess kurtosis):

- $\gamma_2 \approx 0$: mesokurtic (like the normal distribution)
- $\gamma_2 > 0$: leptokurtic (heavy tails, sharp peak)
- $\gamma_2 < 0$: platykurtic (light tails, flat peak)

2 Correlation Analysis

Let $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_n\}$ be two variables.

2.1 Pearson Correlation

The Pearson correlation coefficient measures linear dependence:

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{9}$$

Equivalent formula:

$$r_{XY} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \cdot \sqrt{n \sum y_i^2 - (\sum y_i)^2}} \tag{10}$$

It holds that: $-1 \leq r_{XY} \leq 1$

2.2 Spearman Rank Correlation

For ranks $\text{rank}(x_i)$ and $\text{rank}(y_i)$:

$$\rho_{XY} = r_{\text{rank}(X),\text{rank}(Y)} \tag{11}$$

This is the Pearson correlation coefficient of the ranks, robust to outliers.

3 Simple Linear Regression

3.1 Model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n \tag{12}$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d.

3.2 Parameter Estimation (OLS)

Minimization of the sum of squared errors:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{13}$$

Estimators:

$$\hat{\beta}_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \tag{14}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{15}$$

3.3 Coefficient of Determination ($\mathbf{R^2}$)

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{16}$$

where $\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$ is the total sum of squares.
Interpretation: R^2 gives the proportion of explained variance ($0 \leq R^2 \leq 1$).

3.4 Error Measures

3.4.1 Root Mean Square Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{17}$$

3.4.2 Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{18}$$

4 Polynomial Regression

4.1 Model of Degree p

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \varepsilon_i \tag{19}$$

4.2 Matrix Notation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{20}$$

where

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & \cdots & x_2^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^p \end{pmatrix} \tag{21}$$

4.3 Normal Equations

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{22}$$

The solution is obtained via Gaussian elimination.

5 Multiple Linear Regression

5.1 Model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \tag{23}$$

5.2 Adjusted R^2

The adjusted R^2 takes the number of parameters into account:

$$R^2_{\text{adj}} = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1} \tag{24}$$

where p is the number of predictors.

5.3 Multicollinearity: Variance Inflation Factor (VIF)

For each predictor x_j:

$$\text{VIF}_j = \frac{1}{1 - R_j^2} \tag{25}$$

where R_j^2 is the coefficient of determination of the regression of x_j on all other predictors.

Interpretation:

- $\text{VIF}_j < 5$: no concerning multicollinearity
- $5 \leq \text{VIF}_j \leq 10$: moderate multicollinearity
- $\text{VIF}_j > 10$: high multicollinearity (problematic)

6 Nonlinear Regression Models

6.1 Exponential Model

$$y = a \cdot e^{bx} \tag{26}$$

Linearization: $\ln(y) = \ln(a) + bx$

6.2 Power Model

$$y = a \cdot x^b \tag{27}$$

Linearization: $\ln(y) = \ln(a) + b\ln(x)$

6.3 Logarithmic Model

$$y = a + b\ln(x) \tag{28}$$

6.4 Logistic Function

$$y = y_{\min} + \frac{L}{1 + e^{-k(x-x_0)}} \tag{29}$$

where:

- L: capacity (difference between maximum and minimum)
- k: growth rate
- x_0: inflection point

6.5 Gompertz Function

$$y = a \cdot e^{-b \cdot e^{-cx}} \tag{30}$$

Asymmetric growth curve, frequently used in biology.

6.6 Gaussian Function

$$y = a \cdot e^{-\frac{(x-b)^2}{2c^2}} \tag{31}$$

Symmetric bell curve with:

- a: amplitude (maximum)
- b: center (mean)
- c: width (standard deviation)

7 Residual Analysis

7.1 Residuals

$$e_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n \tag{32}$$

7.2 Standardized Residuals

$$r_i = \frac{e_i}{s_e} \tag{33}$$

where $s_e = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2}$ is the standard deviation of the residuals.

7.3 Leverage

For simple linear regression:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \tag{34}$$

It holds that: $0 \leq h_i \leq 1$ and $\sum_{i=1}^{n} h_i = 2$ (simple regression).
Threshold: $h_i > 2p/n$ indicates high leverage (with p = number of parameters).

7.4 Cook's Distance

$$D_i = \frac{r_i^2}{p} \cdot \frac{h_i}{1 - h_i} \tag{35}$$

Interpretation: $D_i > 1$ indicates an influential data point.

7.5 Q-Q Plot

Compares the quantiles of the standardized residuals with the theoretical quantiles of the standard normal distribution:

$$\Phi^{-1}\left(\frac{i - 0.5}{n}\right) \quad \text{vs.} \quad r_{(i)} \tag{36}$$

where $r_{(i)}$ are the sorted standardized residuals and Φ^{-1} is the quantile function of the standard normal distribution.

8 Model Selection

8.1 Akaike Information Criterion (AIC)

$$\text{AIC} = n \ln(\text{RSS}/n) + 2k \tag{37}$$

where:

- $\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$: Residual Sum of Squares
- k: number of estimated regression coefficients (linear, exponential, logarithmic, power: $k = 2$; polynomial of degree p: $k = p + 1$; multiple regression with p variables: $k = p + 1$)

Principle: Choose the model with the smallest AIC (balance between goodness of fit and model complexity). In the model comparison, the tool sorts the models by R^2 and also shows the AIC.

9 Example 1: Consumption Function USA 1960–2008

The tool contains three economic example datasets (selection list next to *Load example*). Example 1 is worked through in detail here, Examples 2 and 3 in the following sections. All figures correspond to the output of the tool.

9.1 Data Set

49 annual values 1960–2008: Consumption = real personal consumption expenditures, Income = real disposable income, GDP = real gross domestic product (each in billions of US dollars, 2005 prices), InterestRate = 3-month Treasury bill rate (%), Unemployment = unemployment rate (%), Inflation (%). Source: Federal Reserve Bank of St. Louis (FRED) and U.S. Bureau of Labor Statistics ("macrodata" dataset of the Python package statsmodels), annual averages of quarterly data.

Year	Consumption	Income	InterestRate	Unemployment
1960	1784.4	1964.0	2.71	5.58
1965	2241.8	2514.6	4.01	4.53
1970	2740.1	3108.5	6.11	5.00
1975	3214.1	3691.4	5.66	8.48
1980	3766.2	4295.0	11.68	7.18
1985	4540.4	5144.8	7.42	7.20
1990	5316.2	5896.3	7.38	5.60
1995	6079.0	6651.7	5.43	5.62
2000	7608.1	8161.6	5.80	3.95
2005	8819.0	9277.3	3.30	5.08
2008	9290.9	9911.1	1.15	5.80

Table 1: Extract from Example 1 (every fifth year; complete data in the tool)

9.2 Descriptive Statistics

	Consumption (Y)	Income (X)
Mean	4821.1	5307.9
Standard deviation	2241.1	2340.0
Median	4315.6	4989.0
$Q_{0.25}$ / $Q_{0.75}$	3019.5 / 6291.2	3406.5 / 6871.0
IQR	3271.7	3464.5
Skewness	0.565	0.464
Kurtosis	−0.739	−0.806

Both series are slightly right-skewed (mean > median): because of growth, the higher values of the later years are further apart.

9.3 Correlation Analysis

With $n = 49$, $\sum x_i = 260085$, $\sum y_i = 236234$, $\sum x_i y_i = 1\,510\,639\,978.65$ and $\sum x_i^2 = 1\,648\,793\,107.72$:

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 256\,743\,654.4, \qquad \sum (x_i - \bar{x})^2 = 268\,299\,082.7,$$
$$r_{XY} = 0.9992, \qquad \rho_{XY} \text{ (Spearman)} = 0.9998.$$

Interpretation: An almost perfect linear relationship. Caution: for trending time series, a high correlation is partly a consequence of common growth. Consumption correlates weakly negatively with the interest rate ($r = -0.29$) and unemployment ($r = -0.14$).

9.4 Simple Linear Regression: Keynesian Consumption Function

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{256\,743\,654.4}{268\,299\,082.7} = 0.9569,$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 4821.10 - 0.9569 \cdot 5307.86 = -258.15$$

Regression equation:

$$\boxed{\widehat{\text{Consumption}} = -258.15 + 0.9569 \cdot \text{Income}} \tag{38}$$

- The slope is the **marginal propensity to consume**: of each additional dollar of disposable income, about 96 cents are consumed and about 4 cents saved.
- The average propensity to consume $\bar{y}/\bar{x} = 0.908$ is lower; because $\hat{\beta}_0 < 0$ it rises slightly with income.

Goodness of fit:

$$\text{SST} = 246\,097\,042, \quad \text{SSE} = 411\,133, \quad R^2 = 1 - \frac{411\,133}{246\,097\,042} = 0.9983,$$
$$\text{RMSE} = \sqrt{411\,133/49} = 91.60, \quad \text{MAE} = 76.48 \quad \text{(billion US dollars)}.$$

Prediction for an income of 10 000: $\hat{y} = -258.15 + 0.9569 \cdot 10\,000 = 9311.2$.

9.5 Residual Analysis

Standard deviation of the residuals $s_e = 91.60$. Example year 2005 (largest Cook's distance):

$$\hat{y} = -258.15 + 0.9569 \cdot 9277.3 = 8619.58, \quad e = 8819.0 - 8619.58 = 199.42,$$
$$r = 199.42/91.60 = 2.18, \quad h = \tfrac{1}{49} + \tfrac{(9277.3 - 5307.86)^2}{268\,299\,082.7} = 0.079,$$
$$D = \tfrac{2.18^2}{2} \cdot \tfrac{0.079}{1 - 0.079} = 0.204.$$

Interpretation:

- Only two years have $|r| > 2$ (1984: -2.19, 2005: 2.18); all $D_i < 1$ – no influential outliers.
- In the residual plot, positive and negative residuals follow each other in longer phases: the residuals are **autocorrelated** (correlation of successive residuals 0.82, Durbin-Watson statistic 0.35; neither value is calculated by the tool). This is typical for time series – the standard assumption of independent errors is violated, so R^2 looks better than it is.

9.6 Multiple Regression

With income, interest rate and unemployment as regressors:

$$\widehat{\text{Consumption}} = 56.84 + 0.9494 \cdot \text{Income} - 18.21 \cdot \text{InterestRate} - 30.17 \cdot \text{Unemployment} \quad (39)$$

$R^2 = 0.99938$, $R^2_{\text{adj}} = 0.99933$, RMSE = 56.00 (instead of 91.60 for the simple regression). VIF: 1.08, 1.22, 1.14 – no multicollinearity.

Interpretation: For a given income, an interest rate one percentage point higher reduces consumption by about 18 billion US dollars, one percentage point more unemployment by about 30 billion US dollars (precautionary saving). The marginal propensity to consume remains almost unchanged at 0.95.

9.7 Model Comparison: Growth over Time

With Y = Consumption and X = Year, the model *Exponential*

$$\widehat{\text{Consumption}} = a \cdot e^{0.0340 \cdot \text{Year}}, \qquad R^2 = 0.9967, \quad (40)$$

shows an average real consumption growth of $e^{0.034} - 1 \approx 3.5\,\%$ per year (a is extremely small because calendar years are used as X and cannot be interpreted).

10 Example 2: Engel Curve

40 households (fictitious, realistic values): Income = net household income; Food, Housing, Leisure = monthly expenditure (euros); HouseholdSize (persons); FoodShare (% of income).

10.1 Power Model and Income Elasticity

In the power model $y = a \cdot x^b$, $b = \dfrac{\mathrm{d}\ln y}{\mathrm{d}\ln x}$ is the **income elasticity**. Results with X = Income (tab *Simple regression*):

Y	a	b	R^2	Classification
Food	14.84	0.412	0.718	necessity ($0 < b < 1$)
Housing	0.587	0.885	0.971	necessity, close to 1
Leisure	0.00903	1.296	0.960	luxury good ($b > 1$)

Table 2: Income elasticities (power model)

Engel's law: The food share falls with income ($r = -0.84$). Power model for the share: $\widehat{\text{FoodShare}} = 1478.9 \cdot \text{Income}^{-0.588}$, $R^2 = 0.848$ – it fits clearly better than the straight line ($R^2 = 0.706$), because the share falls convexly. As expected, $-0.588 \approx 0.412 - 1$.

10.2 Multiple Regression with Household Size

For food, income alone explains only 72 % of the variance. With household size:

$$\widehat{\text{Food}} = 91.32 + 0.0538 \cdot \text{Income} + 46.47 \cdot \text{HouseholdSize} \tag{41}$$

$R^2 = 0.916$, $R^2_{\text{adj}} = 0.911$, VIF = 1.05. Each additional person raises food expenditure by about 46 euros per month for a given income; of each additional euro of income, about 5 cents go to food.

11 Example 3: Cobb-Douglas Production Function

30 industrial firms (fictitious, realistic values): Output = value added, Capital = capital stock (each in thousand euros), Employees; in addition Productivity (output per employee), CapitalIntensity (capital per employee) and the natural logarithms.

11.1 Linearization and Multiple Regression

$$Y = A\,K^{\alpha}L^{\beta} \quad \Longleftrightarrow \quad \ln Y = \ln A + \alpha \ln K + \beta \ln L \tag{42}$$

Multiple regression of ln_Output on ln_Capital and ln_Employees:

$$\widehat{\ln Y} = 0.933 + 0.3113 \cdot \ln K + 0.6920 \cdot \ln L, \qquad R^2 = 0.9982,\ R^2_{\text{adj}} = 0.9981. \tag{43}$$

- **Output elasticities:** 1 % more capital raises output by 0.31 %, 1 % more employees by 0.69 %.
- **Returns to scale:** $\alpha + \beta = 1.003 \approx 1$ – constant returns to scale.
- $A = e^{0.933} = 2.54$ (total factor productivity).
- **Multicollinearity:** VIF = 11.1 (tool: "High"), because large firms have much capital and many employees ($r(\ln K, \ln L) = 0.95$). Here the estimate remains stable thanks to small disturbances; in real data, α and β often become imprecise as a result.

Omitted-variable bias: The simple regression of ln_Output on ln_Capital alone gives the slope 0.918 – far above $\alpha = 0.31$, because ln_Capital picks up the influence of the missing employees.

11.2 Intensive Form

With constant returns to scale, $Y/L = A\,(K/L)^{\alpha}$. Power model with Y = Productivity, X = CapitalIntensity:

$$\widehat{\text{Productivity}} = 2.568 \cdot \text{CapitalIntensity}^{0.312}, \qquad R^2 = 0.862. \tag{44}$$

The exponent 0.312 agrees with α from the multiple regression.

Model	R^2	RMSE	AIC
Linear	0.99833	91.60	446.71
Polynomial degree 2	0.99952	49.15	387.69
Polynomial degree 3	0.99964	42.74	376.00
Polynomial degree 6	0.99968	40.39	376.47
Exponential	0.94932	504.51	613.91
Logarithmic	0.93672	563.76	624.79
Power ($b = 1.035$)	0.99832	91.75	446.87
Multiple regression	0.99938	56.00	402.49

Table 3: Comparison of regression models for the consumption function

12 Summary: Model Comparison

Tab *Model comparison* for Example 1 with Y = Consumption, X = Income (AIC as in the tool, section Model Selection):

Recommendation: The tool ranks the models by R^2; the polynomial of the highest degree almost always wins. The AIC penalizes additional parameters; here the polynomials from degree 3 upwards are practically level. Economically, however, the linear consumption function (interpretable marginal propensity to consume) or the multiple regression with interest rate and unemployment, which achieves a large part of the improvement with only four parameters, are the sensible choices. High-degree polynomials adapt to the sample and are unsuitable for forecasts outside the data range.

13 References

- Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). *Regression: Models, Methods and Applications*. Springer.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.
- Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. Wiley.
- Seber, G. A. F., & Wild, C. J. (2003). *Nonlinear Regression*. Wiley.
- Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis*. Wiley.

14 Appendix: Numerical Methods

14.1 Gaussian Elimination

To solve the linear system of equations $\mathbf{Ax} = \mathbf{b}$:

Algorithm:

1. Forward elimination: bring $\mathbf{A}$ into upper triangular form
2. Back substitution: solve for $\mathbf{x}$

14.2 Approximation of the Inverse Error Function

For the Q-Q plot, the inverse error function erf^{-1} is required:

$$\mathrm{erf}^{-1}(x) \approx \mathrm{sgn}(x)\sqrt{\sqrt{\left(\frac{2}{\pi a}+\frac{\ln(1-x^2)}{2}\right)^2-\frac{\ln(1-x^2)}{a}}-\left(\frac{2}{\pi a}+\frac{\ln(1-x^2)}{2}\right)} \tag{45}$$

with $a = 0.147$ (Winitzki approximation).
The theoretical quantiles for the Q-Q plot are then:

$$q_i = \sqrt{2}\cdot\mathrm{erf}^{-1}\left(2\cdot\frac{i-0.5}{n}-1\right), \quad i = 1,\ldots,n \tag{46}$$