Multivariate forecasting: Multiple regression

PD Dr. Ralf Stecking and **Abigail Opokua Asare**

Department of Business Adminstration,
Economics and Law

Institute of Economics
Carl von Ossietzky University Oldenburg

18 November, 2025

Overview

1. The linear regression model
2. The (adjusted) determination coefficient
3. The quality of the regression function
4. Population and Sample
5. The t- and F-statistic
6. Betas
7. Multicollinearity
8. The autocorrelation in the residuals

Multiple linear regression analysis

- In general: **Most important** and **most frequently** used method of multivariate data analysis.
- Idea: Our time series can be predicted by a **linear combination** of one or more different time series.
- Also: **description** and **explanation** of relationships.
- Assumption: level of measurement of all time series variables is **metric** or **nominal binary**.

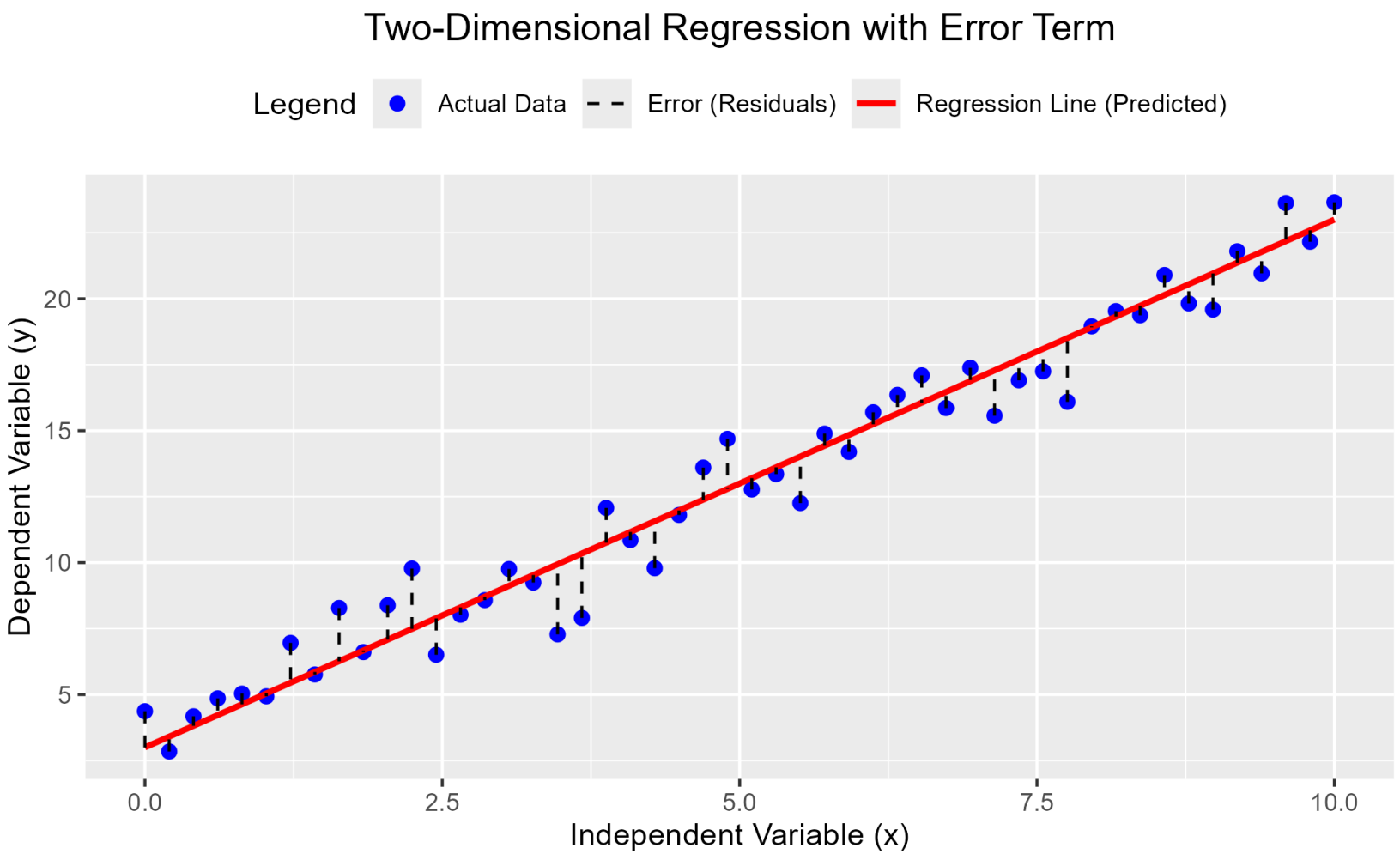
Linear regression model

$$y_{t+1} = b_0 + b_1 x_{1_t} + ... + b_J x_{J_t} + u_t$$

with y as depending time series to be forecasted, $x_1, x_2, ..., x_J$ as explaining time series with time lags and $u_t = y_{t+1} - \hat{y}_{t+1}$ as **random fluctuations** between observed and predicted time series value.
The linear coefficients $b_0, b_1, ..., b_J$ are determined by

$$\sum_{t=1}^{T} u_t^2 = \sum_{t=1}^{T} (y_{t+1} - \hat{y}_{t+1})^2 = \sum_{t=1}^{T} (y_{t+1} - b_0 - b_1 x_{1_t} - ... - b_J x_{J_t})^2 \rightarrow$$

$Min!$



Carl von Ossietzky Universität Oldenburg

(A) Decomposition of total deviation

$$Total\ deviation = \quad explained\ deviation + \quad residual$$

$$y_t - \bar{y} \qquad = \qquad \hat{y}_t - \bar{y} \qquad + \quad y_t - \hat{y}$$

(B) Decomposition of total variation

$$Total\ variation = \quad explained\ variation \ + \ not\ explained\ variation$$

$$\sum_{t=1}^{T}(y_t - \bar{y})^2 \quad = \quad \sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2 \quad + \quad \sum_{t=1}^{T}(y_t - \hat{y})^2$$

This non-trivial relationship applies only for **linear regression** with coefficients estimated by method of least squares!

Carl von Ossietzky Universität Oldenburg

The determination coefficient r^2

$$R^2 = \frac{explained\ variation}{Total\ variation} = \frac{\sum_{t=1}^{T}(\hat{x}_t - \bar{x})^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2}$$

The **determination coefficient** R^2 measures the goodness-of-fit of the trend line towards the time series. R^2 is normalized between zero and one. The higher the determination coefficient, the better the time series is described by the trend line.

The adjusted determination coefficient

The $\bar{R}^2$ is designed to provide a more accurate measure of goodness-of-fit in regression models, particularly when comparing models with different numbers of predictors or sample sizes.

$$\bar{R}^2 = R^2 - \frac{J(1-R^2)}{T-J-1}$$

- R^2: Ordinary determination coefficient (explains the proportion of variance in the dependent variable accounted for by the model).
- J: Number of explanatory variables (predictors).
- T: Total number of observations (sample size).
- $\bar{R}^2$ can be negative if the model performs poorly or if irrelevant variables are added.
- It is generally lower than R^2, reflecting the adjusted measure's stricter criteria.

Standard error of the estimation

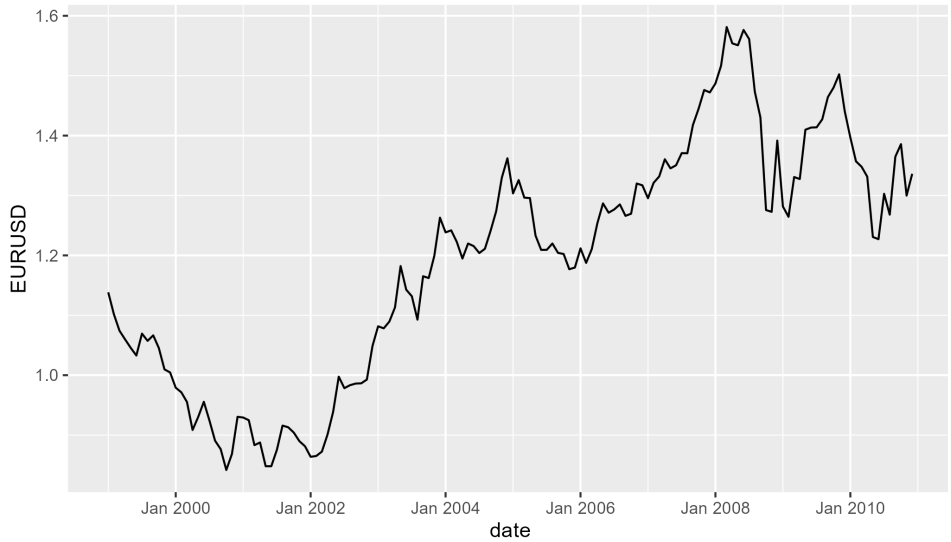
$$s = \sqrt{\frac{\sum_{t=1}^{T} u_t^2}{T - J - 1}} = \sqrt{\frac{not\ explained\ variation}{No.\ of\ observ. - No.\ of\ expl.\ - 1}}$$

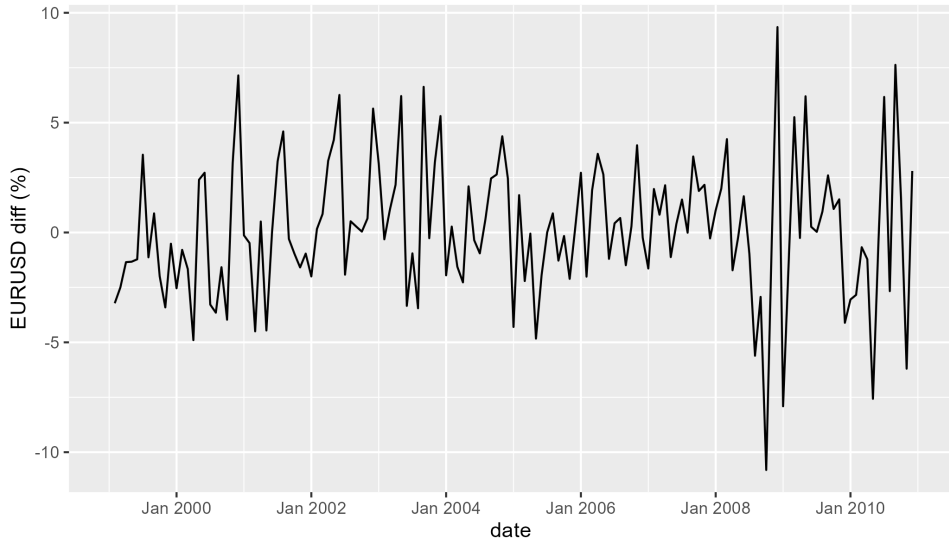
indicates the **mean error** of applying the regression function in order to estimate the values of the forecasting time series y.

Forecast of monthly percentage rate of change of EURO/USD exchange rate

DATE	EURUSD	EURUSD t+1 diff (%)	EURUSD diff (%)	Eurostoxx diff (%)	DJ diff (%)	Retail sales diff12 (%)	Unemployment diff (%)
JAN 2010	1.40	-2.84	-3.05	-6.35	-3.46	-3.61	10.46
FEB 2010	1.36	-0.67	-2.84	-1.74	2.56	1.04	0.70
MAR 2010	1.35	-1.22	-0.67	7.43	5.15	8.71	-2.06
APR 2010	1.33	-7.57	-1.22	-3.90	1.40	-2.86	-4.50
MAY 2010	1.23	-0.29	-7.57	-7.33	-7.92	1.07	-4.82
JUN 2010	1.23	6.17	-0.29	-1.42	-3.58	5.88	-2.71
JUL 2010	1.30	-2.67	6.17	6.56	7.08	4.47	1.22
AUG 2010	1.27	7.63	-2.67	-4.35	-4.31	3.63	-0.12
SEP 2010	1.36	1.53	7.63	4.76	7.72	3.06	-4.91
OCT 2010	1.39	-6.20	1.53	3.53	3.06	1.05	-2.83
NOV 2010	1.30	2.80	-6.20	-6.82	-1.01	5.53	-0.48
DEC 2010	1.34	NA	2.80	5.35	5.19	2.26	2.89

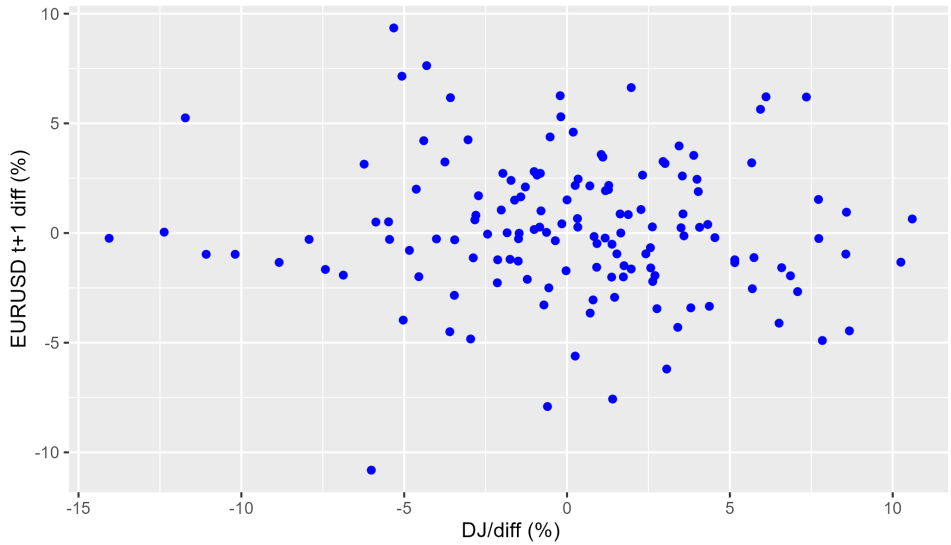
Time series plots / original vs. differences

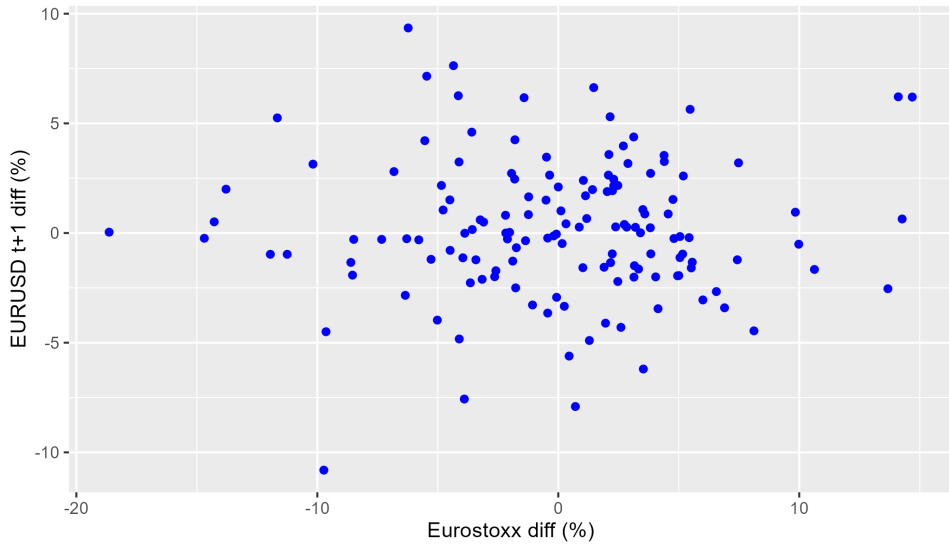




Carl von Ossietzky Universität Oldenburg

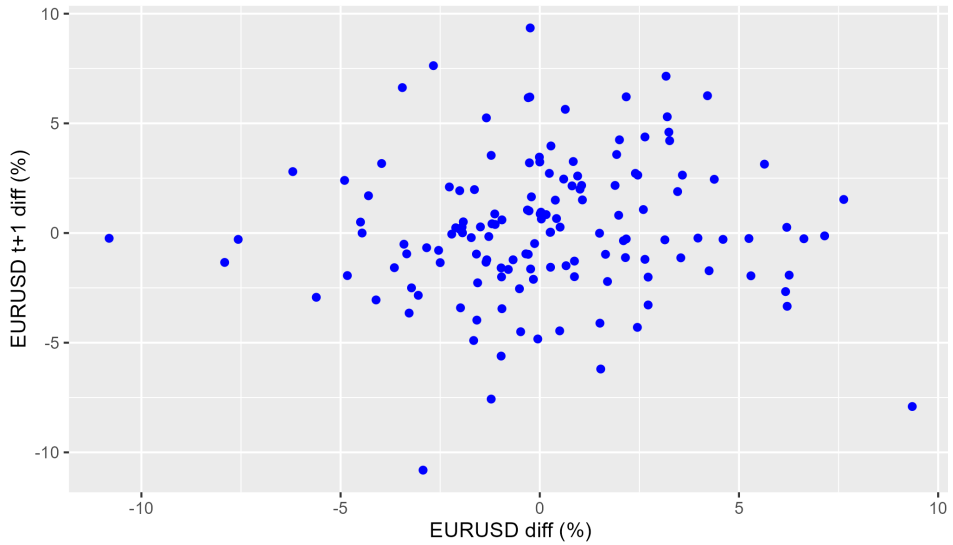
Scatter-plots (I)

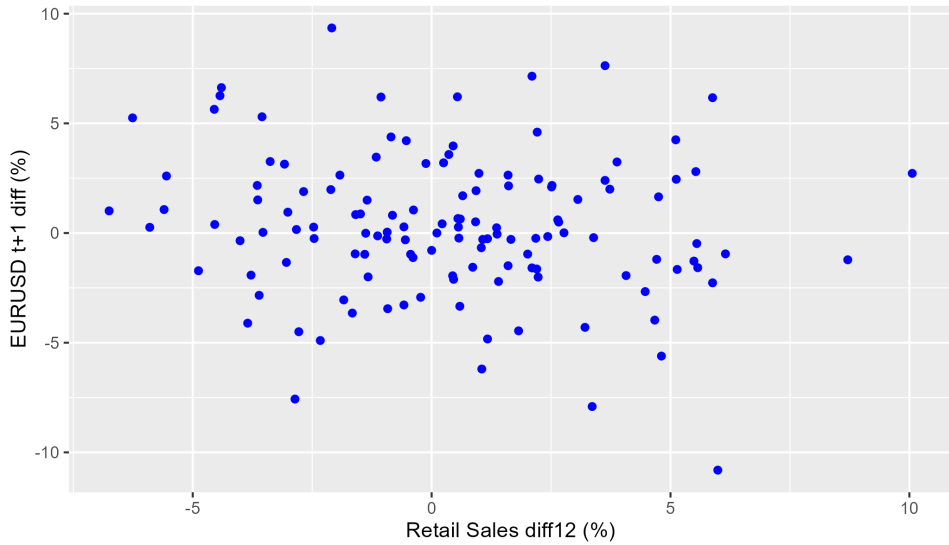




Carl von Ossietzky Universität Oldenburg

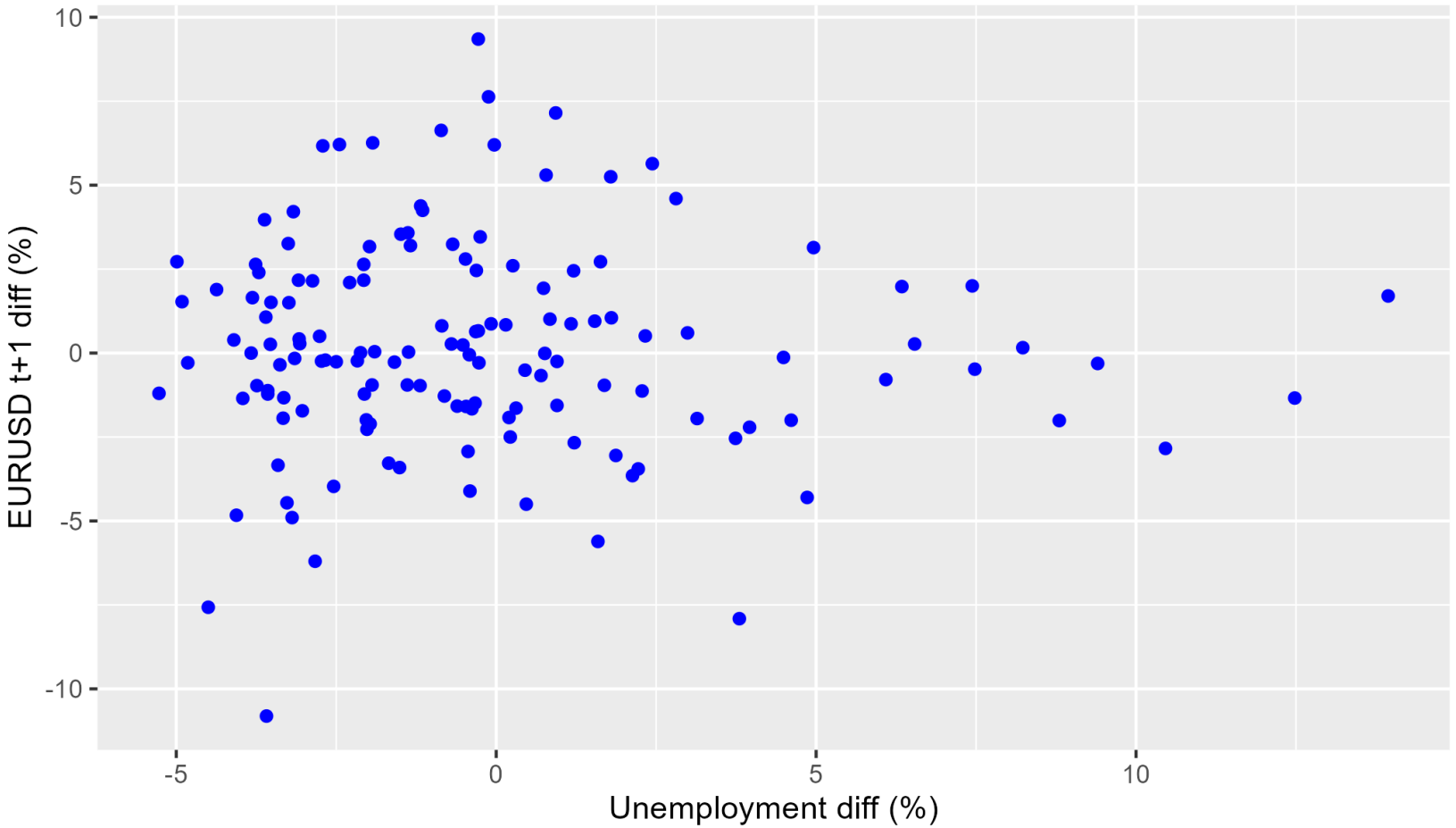
Scatter-plots (II)





Carl von Ossietzky Universität Oldenburg

Scatter-plots (III)



Correlation table

	EURUSD t+1 diff (%)	EURUSD diff (%)	Eurostoxx diff (%)	DJ diff (%)	Retail Sales diff12 (%)	Unemployment diff (%)
EURUSD t+1 diff (%)	1.00	0.05	0.04	-0.06	-0.14	-0.04
EURUSD diff (%)	0.05	1.00	0.11	0.24	-0.11	-0.07
Eurostoxx diff (%)	0.04	0.11	1.00	0.81	0.05	-0.08
DJ diff (%)	-0.06	0.24	0.81	1.00	-0.05	-0.10
Retail Sales diff12 (%)	-0.14	-0.11	0.05	-0.05	1.00	-0.09
Unemployment diff (%)	-0.04	-0.07	-0.08	-0.10	-0.09	1.00

Table: Pearson Correlation Table

Multiple regression analysis

	(1)
Constant	0.401
	(0.275)
EURUSD diff (%)	−0.253*
	(0.138)
Eurostoxx diff (%)	0.182
	(0.114)
DJ diff (%)	0.073
	(0.107)
Retail Sales diff12 (%)	−0.167*
	(0.098)
Unemployment diff (%)	−0.048
	(0.067)
Num.Obs.	131
R2	0.064
R2 Adj.	0.026
AIC	682.5
BIC	702.6
RMSE	3.10

The Quality of the Regression Function

- The determination coefficient R^2 is a measure of the **goodness of fit** of the regression function towards the empirical data.
- However, R^2 increases if the number of explanatory variables J is high or if the number of observation points T is low.
- The adjusted determination coefficient R^2_{adj} accounts for the influence of J and T, providing a more accurate measure.
- Question: What is the critical value of R^2_{adj}, i.e., at what value can we consider the **quality of the regression function** to be sufficient?

Population and Sample

Assumption: The regression coefficients $b_0, b_1, \ldots, b_J$ are random samples from a distribution, serving as **estimators** for the unknown population coefficients $\beta_0, \beta_1, \ldots, \beta_J$.

- **Null Hypothesis (H_0):**

$$\beta_1 = \beta_2 = \cdots = \beta_J = 0$$

This implies no relationship/dependency between forecasting and explaining time series.

- If H_0 is true:
  - The "true" determination coefficient ρ^2 equals zero.
  - However, the empirical coefficients $b_1, b_2, \ldots, b_J$ and R^2 may not be zero due to:
    - Sampling variability.
    - Random noise in the data.

Population and Sample

Key Question: How large must $b_0, b_1, \ldots, b_J$ or R^2 be to reject H_0 and conclude that $\beta_0, \beta_1, \ldots, \beta_J$ or $\rho^2 > 0$? This depends on:

- Sample size T.
- Number of predictors J.
- Chosen significance level α.

F Statistic

- The F statistic tests whether there is a linear relationship between the dependent variable y and the independent variables $x_1, x_2, \ldots, x_J$.
- **Key Question:** Does at least one of the independent variables contribute to explaining the variation in y?
- Equivalent: Is at least one regression coefficient β_j significantly different from zero?

Hypotheses in the F Test

- **Null Hypothesis (H_0):**

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_J = 0$$

This implies:

  - There is no relationship between y and any of the independent variables.
  - The regression model does not improve upon a simple prediction.

- **Alternative Hypothesis (H_1):** At least one regression coefficient β_j is not zero, suggesting a linear relationship exists.

Step 1: Computing the Empirical F statistic

- The F statistic is calculated as a ratio:

$$F_{emp} = \frac{\frac{\sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2}{J}}{\frac{\sum_{t=1}^{T}(y_t - \hat{y})^2}{T-J-1}} = \frac{explained\ variation/J}{not\ explained\ variation/T-J-1}$$

- Components:
  - **Explained variation:** Variation due to the regression model, adjusted for the number of predictors J.
  - **Unexplained variation:** Residual variation, adjusted for degrees of freedom $(T - J - 1)$.

Step 2: Specification of the Error Probability

- The error probability α defines the likelihood of rejecting the null hypothesis H_0 when it is actually true (**Type I error**).
- **Common Values of α:**
  - $\alpha = 0.05$ (5%) is the most commonly used level.
  - Other common values: $\alpha = 0.01$ (1%) or $\alpha = 0.10$ (10%).
- **Confidence Level:** Equal to $1 - \alpha$. For example:

$$\text{If } \alpha = 0.05, \text{ then confidence level } = 95\%.$$

Step 3: Finding the Theoretical F statistic

- The theoretical F value F_{theor} is obtained from the F-distribution table.
- **Required Inputs:**
  - **Numerator Degrees of Freedom:** $df_1 = J$, the number of predictors.
  - **Denominator Degrees of Freedom:** $df_2 = T - J - 1$, based on the number of observations T and predictors J.
- The choice of significance level α (e.g., 5%) determines the threshold for rejecting H_0.
- Use the F-distribution table to find F_{theor} for the given df_1, df_2, and α.

Step 4: Comparing Empirical and Theoretical F statistic

- **Decision Rule:**
  - If $F_{emp} > F_{theor}$:
    - Reject the null hypothesis H_0.
    - Conclude that the regression model is **statistically significant**.
  - If $F_{emp} \leq F_{theor}$:
    - Fail to reject the null hypothesis H_0.
    - Conclude that there is **no significant relationship** between the variables.
- This comparison determines whether the relationship described by the regression function is meaningful or likely due to chance.

The Connection Between F and R^2

- The coefficient of determination R^2 is defined as:

$$R^2 = \frac{\text{explained variation (EV)}}{\text{total variation (TV)}} = 1 - \frac{\text{not explained variation (NEV)}}{\text{total variation (TV)}}$$

- The F statistic can be expressed in terms of R^2:

$$F = \frac{\frac{\text{EV}}{\text{TV}}/J}{\frac{\text{NEV}}{\text{TV}}/(T-J-1)} = \frac{R^2/J}{(1-R^2)/(T-J-1)}$$

- This equation links F and R^2, allowing determination of model significance using R^2.

Beta Values

- **Definition:** Beta values (β_j) are standardized regression coefficients:

$$\beta_j = b_j \cdot \frac{s_{x_j}}{s_y}$$

Where:

- b_j: Non-standardized regression coefficient for predictor x_j.
- s_{x_j}: Standard deviation of predictor x_j.
- s_y: Standard deviation of the dependent variable y.

- **Reasons:**
  - Standardizes the influence of predictors, making them comparable across variables with different units or scales.
  - Helps identify which predictor has the largest influence on the dependent variable y.
- **Insight:**
  - The predictor x_j with the largest absolute Beta value ($|\beta_j|$) has the strongest effect on y.
  - The sign of β_j indicates the direction of the relationship.

Example

Scenario: Study the impact of Income (x_1), Education (x_2), and Age (x_3) on SWB (y).

- Non-standardized regression coefficients:

$$b_1 = 0.03, \quad b_2 = 0.8, \quad b_3 = -0.1$$

- Standard deviations:

$$s_{x_1} = 15,000, \quad s_{x_2} = 2, \quad s_{x_3} = 10, \quad s_y = 5$$

Empirical example–EURO/USD

- Forecast of monthly percentage rate of change of EURO/USD exchange rate
- Observation period JAN/1999 – DEC/2019

Correlation table

	EURUSD t+1 diff (%)	EURUSD diff (%)	Eurostoxx diff (%)	DJ diff (%)	Retail Sales diff12 (%)	Unemployment diff (%)
EURUSD t+1 diff (%)	1.00	0.01	0.01	-0.10	-0.10	-0.02
EURUSD diff (%)	0.01	1.00	0.12	0.26	-0.06	-0.01
Eurostoxx diff (%)	0.01	0.12	1.00	0.78	0.08	-0.01
DJ diff (%)	-0.10	0.26	0.78	1.00	0.01	-0.04
Retail Sales diff12 (%)	-0.10	-0.06	0.08	0.01	1.00	-0.04
Unemployment diff (%)	-0.02	-0.01	-0.01	-0.04	-0.04	1.00

Table: Pearson Correlation Table

Regression with Standardized Coefficients

	Coefficients	Standardized coefficients
Constant	0.346*	0.000*
	(0.208)	(0.000)
EURUSD diff (%)	0.061	0.061
	(0.079)	(0.079)
Eurostoxx diff (%)	0.152**	0.275**
	(0.071)	(0.128)
DJ diff (%)	−0.234***	−0.334***
	(0.085)	(0.121)
Retail Sales diff12 (%)	−0.100	−0.114
	(0.063)	(0.071)
Unemployment diff (%)	−0.026	−0.032
	(0.042)	(0.053)
Num.Obs.	239	239
R2	0.051	0.051
R2 Adj.	0.031	0.031
F-statistic	2.327	2.327

Carl von Ossietzky Universität Oldenburg

t-Test for the Regression Coefficient

- After using the F-test to check if at least one regression coefficient (β_j) is not zero, we test each coefficient individually to identify which ones are significant.
- The empirical t-value formula:

$$t_{\text{emp}} = \frac{b_j - \beta_j}{s_{b_j}}$$

Where:

- b_j: Estimated coefficient for the j-th variable (from the data).
- β_j: True (unknown) coefficient.
- s_{b_j}: Standard error of b_j, measuring the variability of b_j.

- Typically, the null hypothesis $H_0 : \beta_j = 0$ is tested:

$$t_{\text{emp}} = \frac{b_j}{s_{b_j}}$$

Four Steps of the t-Test

1. Compute empirical t-value:

$$t_{\text{emp}} = \frac{b_j}{s_{b_j}}$$

2. Set significance level (α), e.g., 5%.
3. Find theoretical t-value using degrees of freedom ($df = T - J - 1$).
   - The $t-test$ is a **two-sided test**. The theoretical t-value therefore must be found in the column $\frac{\alpha}{2}$ of the $t-table$!
4. Decision Rule:
   - Reject H_0 if $|t_{\text{emp}}| > t_{\text{theor}}$. It suggests that x_j significantly affects y.
   - Fail to reject H_0 if $|t_{\text{emp}}| \leq t_{\text{theor}}$.

Multicollinearity

- Multicollinearity occurs when predictors are highly correlated, leading to unreliable coefficient estimates.
- Why is it a problem?
  - Coefficients become unstable (small data changes result in large differences in estimates).
  - Inflated standard errors reduce the accuracy of t-tests.
- Detection Methods:
  1. **Correlation Matrix**: Check for correlations > 0.9 or < -0.9.
  2. **Tolerance (TOL)**: $TOL = 1 - R_j^2$, where R_j^2 is the determination coefficient of x_j against other predictors. Low $TOL (< 0.1)$ indicates multicollinearity.
  3. **Variance Inflation Factor (VIF)**: $VIF = \frac{1}{TOL}$. High $VIF (> 10)$ signals multicollinearity.

Strategies Against Multicollinearity

- **Extend Observation Period (T)**: Increasing the sample size can reduce correlations between predictors.
- **Combine or Transform Variables**: Highly correlated predictors can be merged into a composite index or transformed to reduce dependence.
- **Eliminate Predictors**: Remove variables that contribute most to multicollinearity without compromising the theoretical basis of the model.

Autocorrelation in Residuals

- Residuals ($u_t = y_t - \hat{y}_t$) should be **uncorrelated**.
- Problem: Autocorrelation inflates standard errors and produces unreliable t-values.
- **Durbin-Watson Statistic**:

$$d = \frac{\sum_{t=2}^{T}(u_t - u_{t-1})^2}{\sum_{t=1}^{T} u_t^2}$$

Measures the correlation of consecutive residuals.

- Interpretation:
  1. For positive autocorrelation ($u_t \approx u_{t-1}$) d tends to zero.
  2. For negative autocorrelation ($u_t \approx u_{t-1}$) d tends to $+4$.
  3. No autocorrelation ($u_t \approx u_{t-1}$) d tends to $+2$ for large values of T.
  - For large datasets (T is large), the expected value $E(d) = 2 + \frac{2 \cdot J}{T-J-1}$, d approaches 2, which is the ideal scenario of no autocorrelation.

Durbin-Watson Statistic in Practice

- Steps:
  1. Set significance level (α), e.g., 5%.
  2. Find critical values (d_{lo} and d_{up}) based on T and J.
  3. There is an **uncertainty area** for the Durbin-Watson statistic:
     a. $d_{lo} < d > d_{up}$.
     b. $(4 - d_{up}) < d > (4 - d_{lo})$.
  4. The **null hypothesis** H_0 reads $\rho = 0$, i.e.. the "true" **autocorrelation** coefficient between u_t and u_{t-1} is equal to zero.
  5. Compare d_{emp}:
     - $d_{\text{emp}} < d_{lo}$ or $d_{\text{emp}} > 4 - d_{lo}$: Reject H_0 (autocorrelation detected).
     - $d_{up} < d_{\text{emp}} < 4 - d_{up}$: Fail to reject H_0 (no autocorrelation detected).
- Example: For $T = 142$, $J = 5$, and $\alpha = 5\% : d_{lo} = 1.57, d_{up} = 1.78$:
  - Reject H_0 if $d_{\text{emp}} < 1.57$ or $d_{\text{emp}} > 2.43$.
  - Fail to reject H_0 if $1.78 < d_{\text{emp}} < 2.22$.