# Regression analysis

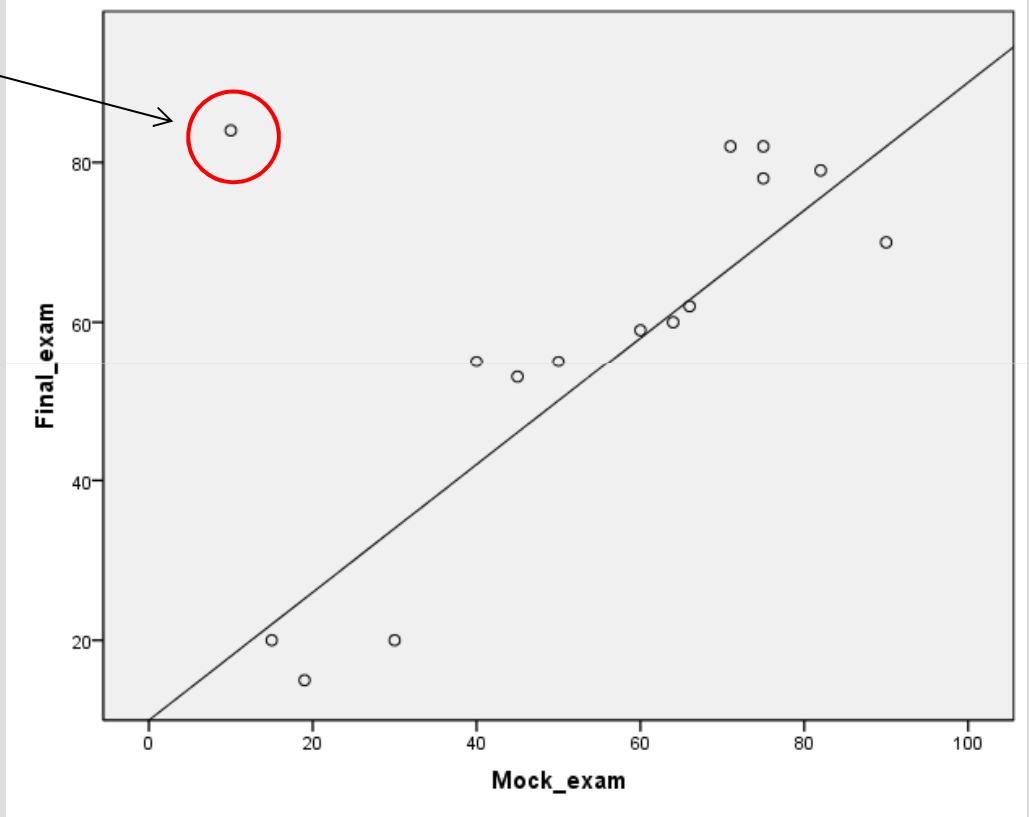
# Example 1

A teacher wants to find out if the results of a exam written earlier during the semester correlate with the results of the final exam. Perform a correlational analysis.

| Students | Mock exam | Final exam |
|---|---|---|
| Bill | 50 | 55 |
| Jane | 30 | 20 |
| Jack | 60 | 59 |
| Pat | 75 | 78 |
| John | 40 | 55 |
| Susan | 90 | 70 |
| Anna | 15 | 20 |
| Margret | 19 | 15 |
| Peggy | 64 | 60 |
| Joe | 80 | 84 |
| William | 25 | 82 |
| Ron | 82 | 79 |
| Bob | 66 | 62 |
| Sally | 71 | 82 |
| Marry | 45 | 53 |

# Example 1

Outlier

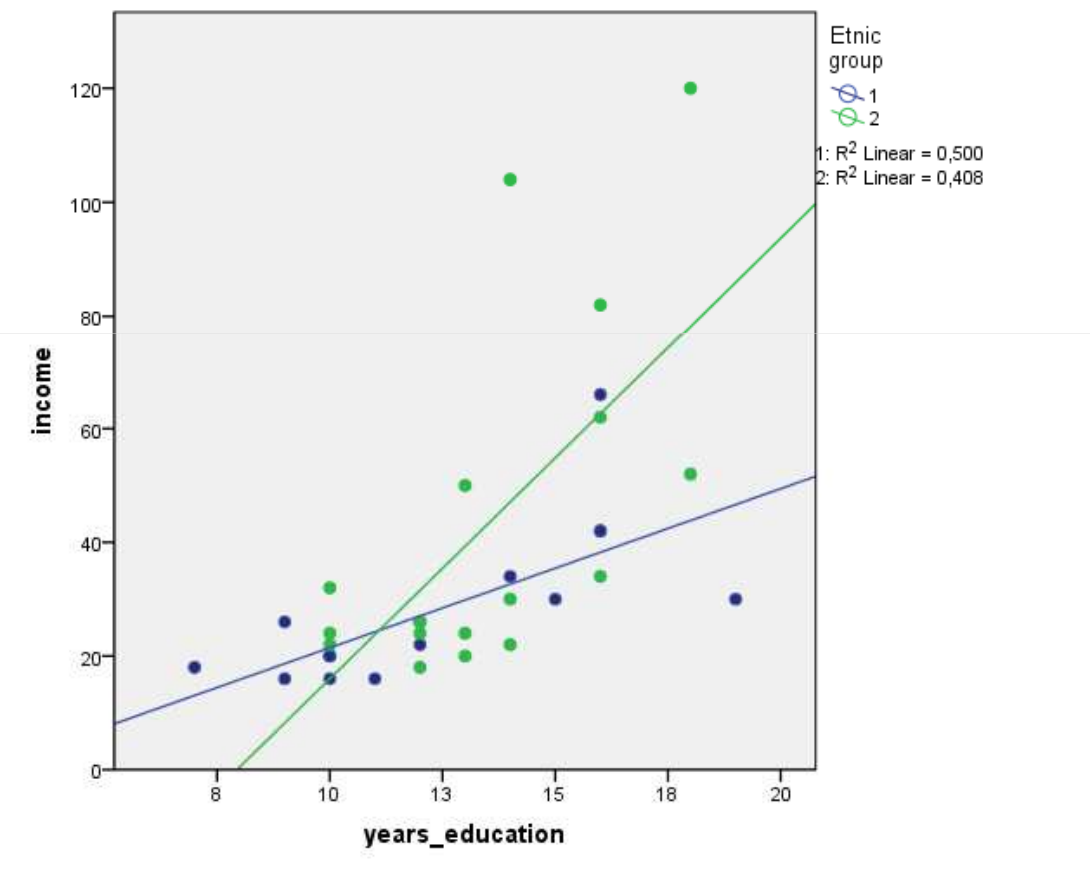


# Example 2

Is there a correlation between years of education, income and ethnic group?

| Black years of education | Blacks income | Black years of education | Blacks income |
|---|---|---|---|
| 10 | 16 | 16 | 62 |
| 7 | 18 | 10 | 24 |
| 9 | 26 | 13 | 50 |
| 11 | 16 | 10 | 32 |
| 14 | 34 | 16 | 34 |
| 12 | 22 | 18 | 52 |
| 16 | 42 | 12 | 24 |
| 16 | 42 | 14 | 22 |
| 9 | 16 | 13 | 20 |
| 10 | 20 | 14 | 30 |
| 16 | 66 | 13 | 24 |
| 12 | 26 | 18 | 120 |
| 10 | 20 | 10 | 22 |
| 15 | 30 | 16 | 82 |
| 10 | 20 | 12 | 18 |
| 19 | 30 | 12 | 26 |
| | | 14 | 104 |

# Example 2


Etnic group
1
2
1: R2 Linear = 0,500
2: R2 Linear = 0,408
income
years_education
0
20
40
60
80
100
120
8
10
13
15
18
20

# Example 2

**Korrelationen**

| | | income | years_ education |
|---|---|---|---|
| income | Korrelation nach Pearson | 1 | ,707** |
| | Signifikanz (2-seitig) | | ,002 |
| | N | 16 | 16 |
| years_education | Korrelation nach Pearson | ,707** | 1 |
| | Signifikanz (2-seitig) | ,002 | |
| | N | 16 | 16 |

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

**Korrelationen**

| | | income | years_ education |
|---|---|---|---|
| income | Korrelation nach Pearson | 1 | ,638** |
| | Signifikanz (2-seitig) | | ,006 |
| | N | 17 | 17 |
| years_education | Korrelation nach Pearson | ,638** | 1 |
| | Signifikanz (2-seitig) | ,006 | |
| | N | 17 | 17 |

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

# Bivariate linear regression

# Correlation - Regression

Correlational analysis gives us a measure that indicates how closely the data points are associated.

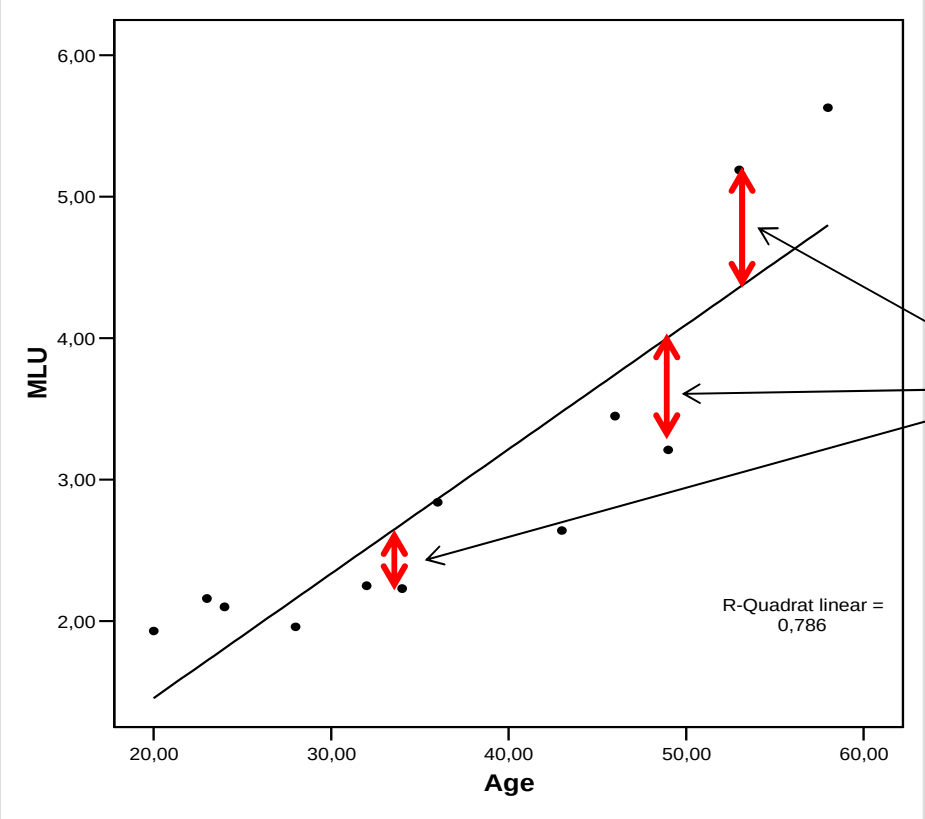
Regression analysis measures the effect of the predictor variable x on the dependent variable (= criterion) y. – How much does y change if you change x.

A correlational analysis is purely descriptive, whereas a regression analysis allows us to make predictions.

# Types of regression analysis

| | **Predictor variable** | **Criterion (target) variable** |
|---|---|---|
| Linear regression | 1 interval | 1 interval |
| Multiple regression | 2+ (some of the variables can be categorical) | 1 interval |
| Logistic regression<br>Discriminant analysis | 1+ | 1 categorical |

# Line of-best-fit




Residuals: difference between predicted and observed values

# Linear regression

$$y = bx + a$$

y = variable to be predicted
x = given value on the variable x
b = value of the slope of the line
a = the intercept (or constant), which is the place where the line-of-best-fit intercepts the y-axis

# Linear regression

Given a score of 20 on the x-axis, a slope of b = 2, and an interception point of a = 5, what is the predicted score?

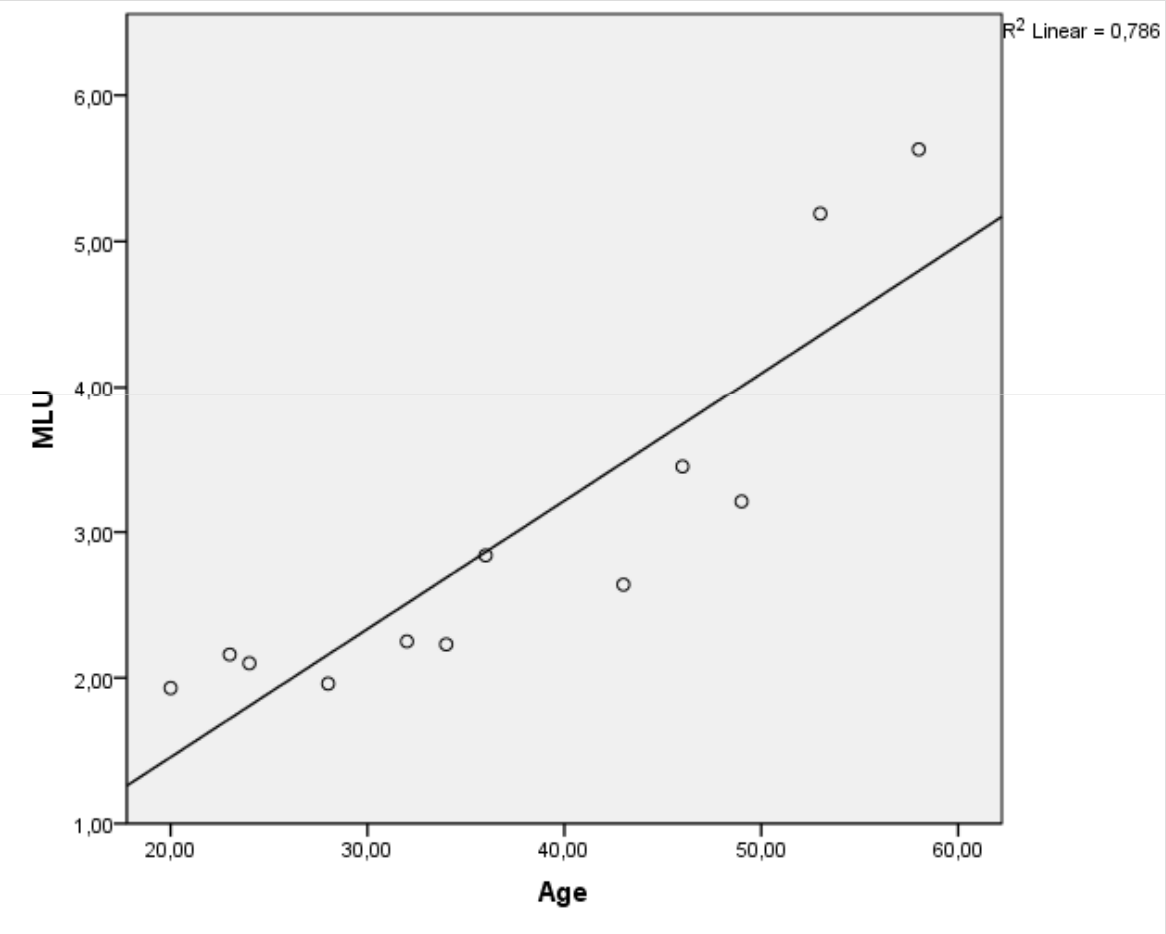
$$y = (2 \times 20) + 5$$
$$y = 45$$

# Example: Age and MLU

Ein wichtiges Maß für den Entwicklungsstand eines Kindes ist die Mittlere Länge von Äußerungen (MLU = Mean length of Utterances). Danach nimmt die Anzahl der Wörter und/oder Silben in einer Äußerung mit dem Alter zu. Überprüfen Sie diese Hypothese an den folgenden Daten.

| Cases | Children | Age in months | MLU |
|---|---|---|---|
| 1 | John | 24 | 2.10 |
| 2 | Bill | 23 | 2.16 |
| 3 | Sue | 32 | 2.25 |
| 4 | Jane | 20 | 1.93 |
| 5 | Ann | 43 | 2.64 |
| 6 | Susan | 58 | 5.63 |
| 7 | Jack | 28 | 1.96 |
| 8 | William | 34 | 2.23 |
| 9 | Mary | 53 | 5.19 |
| 10 | Peter | 46 | 3.45 |
| 11 | Pete | 49 | 3.21 |
| 12 | Allan | 36 | 2.84 |

# Example: Age and MLU



# Example: Age and MLU

**Modellzusammenfassung**

| Modell | R | R-Quadrat | Korrigiertes R-Quadrat | Standardfehler des Schätzers |
|---|---|---|---|---|
| 1 | ,887[a] | ,786 | ,765 | ,60242 |

a. Einflußvariablen : (Konstante), Age

Effect size

# Correlational coefficient and effect size

| Correlation coefficient | Shared variance (effect size) | |
|---|---|---|
| r = 0.0 | 0.00 | Kein Zusammenhang |
| r = 0.1 | 0.01 (1%) | Geringe Korrelation |
| r = 0.2 | 0.04 (4%) | |
| r = 0.3 | 0.09 (9%) | Mittlere Korrelation |
| r = 0.4 | 0.16 (16%) | |
| r = 0.5 | 0.25 (25%) | |
| r = 0.6 | 0.36 (36%) | Hohe Korrelation |
| r = 0.7 | 0.49 (49%) | |
| r = 0.8 | 0.64 (64%) | Sehr hohe Korrelation |
| r = 0.9 | 0.81 (81%) | |
| r = 1.0 | 1.00 (100%) | |

# Example: Age and MLU

**ANOVA[b]**

| Modell | | Quadratsumme | df | Mittel der Quadrate | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 13,369 | 1 | 13,369 | 36,838 | ,000[a] |
| | Nicht standardisierte Residuen | 3,629 | 10 | ,363 | | |
| | Gesamt | 16,998 | 11 | | | |

a. Einflußvariablen : (Konstante), Age
b. Abhängige Variable: MLU

# Example: Age and MLU

Intercept

**Koeffizienten[a]**

| Modell | | Nicht standardisierte Koeffizienten | | Standardisierte Koeffizienten | T | Sig. |
|---|---|---|---|---|---|---|
| | | RegressionskoeffizientB | Standardfehler | Beta | | |
| 1 | (Konstante) | -,304 | ,566 | | -,536 | ,603 |
| | Age | ,088 | ,014 | ,887 | 6,069 | ,000 |

a. Abhängige Variable: MLU

Regression coefficient

Sig of regression coefficient

# Example: Age and MLU

There is s strong association between age and MLU (R = 0.887). Specifically, it was found that the children's MLU increases by an average of .088 words each months (t = 6,069, p < 0.001), which amounts to about a word a year. Since the F-value (36,838, df = 1) is highly significant (p < 001), these results are unlikely to have arisen from sample error.

# Multivariate regression

# Multiple regression

Several predictor variables influence the criterion.

$$y = b_1x_1 + b_2x_2 + b_3x_3 \ldots + a$$

Plane-of-best-fit

1. Simultaneous multiple regression
2. Stepwise multiple regression

# Multiple regression

Eine Universität möchte wissen, welche Faktoren am besten dazu geeignet sind, den Lernerfolg ihrer Studenten vorherzusagen. Als Indikator für den Wissensstand der Studenten gilt die Punktzahl in einer zentralen Abschlussklausur. Als mögliche Faktoren werden in Betracht gezogen: (1) Punktzahl beim Eingangstest, (2) Alter, (3) IQ Test, (4) Punktzahl bei einem wissenschaftlichen Projekt.

# Multiple regression

**Predictor**:

- entrance exam
- age
- IQ
- Scientific project

**Criterion**:

- final exam

# Multiple regression

**Modellzusammenfassung**

| Modell | R | R-Quadrat | Korrigiertes R-Quadrat | Standardfehler des Schätzers |
|---|---|---|---|---|
| 1 | ,875[a] | ,765 | ,731 | 16,913 |

a. Einflußvariablen : (Konstante), Project, Age, IQ, Entrance

# Multiple regression

**ANOVA[b]**

| Modell | | Quadratsumme | df | Mittel der Quadrate | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 26067,689 | 4 | 6516,922 | 22,783 | ,000[a] |
| | Nicht standardisierte Residuen | 8009,220 | 28 | 286,044 | | |
| | Gesamt | 34076,909 | 32 | | | |

a. Einflußvariablen : (Konstante), Project, Age, IQ, Entrance
b. Abhängige Variable: Final

# Multiple regression

**Koeffizienten[a]**

| Modell | | Nicht standardisierte Koeffizienten | | Standardisierte Koeffizienten | T | Sig. |
|---|---|---|---|---|---|---|
| | | Regressionskoeffizient B | Standardfehler | Beta | | |
| 1 | (Konstante) | -267,843 | 48,263 | | -5,550 | ,000 |
| | Entrance | 2,492 | ,459 | ,576 | 5,431 | ,000 |
| | Age | 1,057 | 1,059 | ,099 | ,998 | ,327 |
| | IQ | 1,511 | ,328 | ,447 | 4,606 | ,000 |
| | Project | ,503 | ,355 | ,141 | 1,417 | ,168 |

a. Abhängige Variable: Final

Standardisierter Wert für Neigungswinkel (=b)

# Multiple regression

There is a strong association between the predictor variables and the result of the final exam (Multiple R = 0.875; F = 22,783, df = 4, p = .001). Together they account for 73% of the variation in the exam success. If we look at the four predictor variables individually we find that the result of the entrance exam (B = .576, t = 5.431, p = .001) and the IQ score (B = .447, t = 4.606, p = .001) make the strongest contributions (i.e. they are the best predictors). The predictive value of age (B = .099, t = 5.431, p = .327 and the score on the scientific project is not significant (B = 0.141, t = 1,417, p = 0.168.).

# Multivariate regression

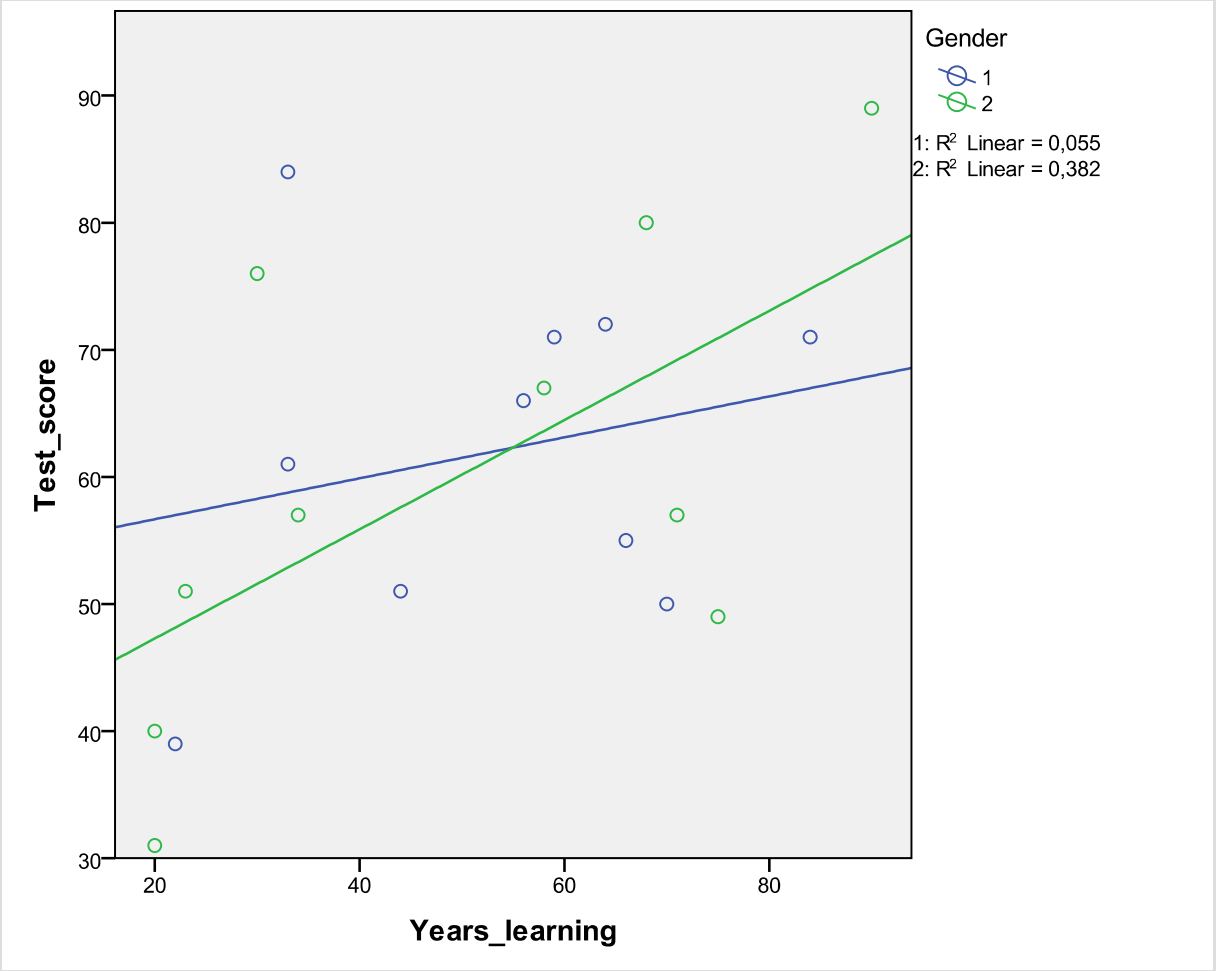
# Exercise

A researcher has carried out an experiment with 20 L2 learners of English testing their command of different types of relative clauses. Now he wants to find out if and to what extend the following factors influence the result of the experiment: years of foreign language education, time spent abroad (in month), and gender.

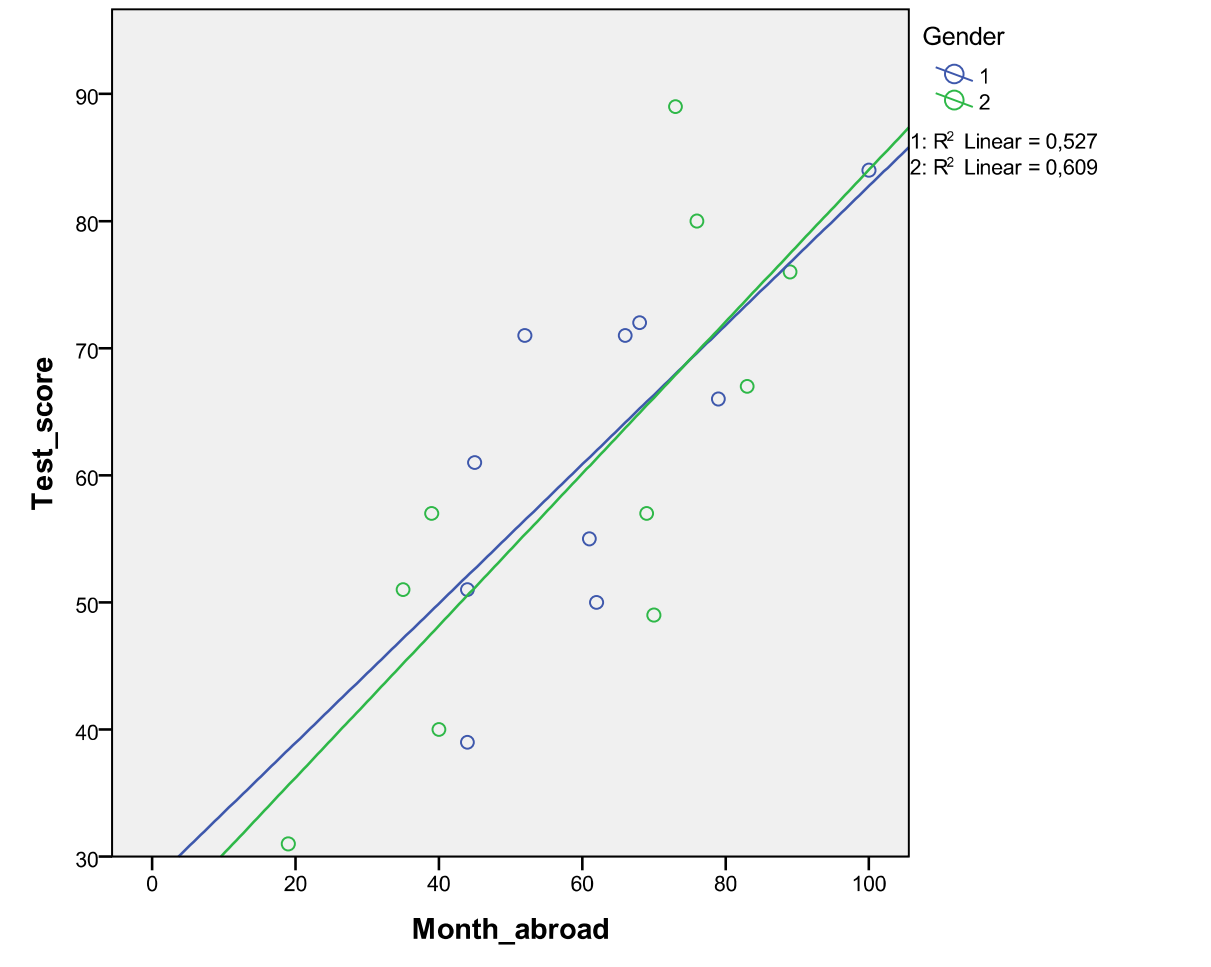
Inspect the data (scatterplots), perform a multiple regression analysis and interpret the output in form of a written report.

| Gender | Years_learning | Month_abroad | Test_score |
|---|---|---|---|
| 1 | 33 | 45 | 61 |
| 1 | 64 | 68 | 72 |
| 1 | 33 | 100 | 84 |
| 1 | 22 | 44 | 39 |
| 1 | 70 | 62 | 50 |
| 1 | 66 | 61 | 55 |
| 1 | 59 | 52 | 71 |
| 1 | 84 | 66 | 71 |
| 1 | 56 | 79 | 66 |
| 1 | 44 | 44 | 51 |
| 2 | 23 | 35 | 51 |
| 2 | 68 | 76 | 80 |
| 2 | 30 | 89 | 76 |
| 2 | 20 | 40 | 40 |
| 2 | 75 | 70 | 49 |
| 2 | 71 | 69 | 57 |
| 2 | 34 | 39 | 57 |
| 2 | 90 | 73 | 89 |
| 2 | 58 | 83 | 67 |
| 2 | 20 | 19 | 31 |

# Results


Gender
1
2
1: R² Linear = 0,055
2: R² Linear = 0,382
Test_score
90
80
70
60
50
40
30
20
40
60
80
Years_learning

# Results


Gender
1
2
1: R² Linear = 0,527
2: R² Linear = 0,609
Test_score
90
80
70
60
50
40
30
0
20
40
60
80
100
Month_abroad

# Results

**Modellzusammenfassung**

| Modell | R | R-Quadrat | Korrigiertes R-Quadrat | Standardfehler des Schätzers |
|---|---|---|---|---|
| 1 | ,775[a] | ,601 | ,526 | 10,755 |

a. Einflußvariablen : (Konstante), Month_abroad, Gender, Years_learning

# Results

**ANOVA[b]**

| Modell | | Quadratsumme | df | Mittel der Quadrate | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 2787,662 | 3 | 929,221 | 8,033 | ,002[a] |
| | Nicht standardisierte Residuen | 1850,888 | 16 | 115,680 | | |
| | Gesamt | 4638,550 | 19 | | | |

a. Einflußvariablen : (Konstante), Month_abroad, Gender, Years_learning
b. Abhängige Variable: Test_score

# Results

**Koeffizienten[a]**

| Modell | | Nicht standardisierte Koeffizienten | | Standardisierte Koeffizienten | T | Sig. |
|---|---|---|---|---|---|---|
| | | Regressionskoeffizient B | Standardfehler | Beta | | |
| 1 | (Konstante) | 24,099 | 11,298 | | 2,133 | ,049 |
| | Gender | -,376 | 4,834 | -,012 | -,078 | ,939 |
| | Years_learning | ,110 | ,124 | ,159 | ,881 | ,392 |
| | Month_abroad | ,523 | ,137 | ,686 | 3,813 | ,002 |

a. Abhängige Variable: Test_score

# Results

There is a strong association between the predictor variables and the result of the final exam (Multiple R = 0.775; F = 8,033, df = 3, p = .002). Together they account for 60% of the variation in the exam success. However, if we look at the three predictor variables individually we find that the months spent abroad (B = .686, t = 3.813, p = .002) is the only significant predictor. The predictive value of gender (B = …) and years of learning are not significant (B = …).

# Stepwise regression

# Stepwise multiple regression

**Koeffizienten[a]**

| Modell | | Nicht standardisierte Koeffizienten | | Standardisierte Koeffizienten | T | Signifikanz |
|---|---|---|---|---|---|---|
| | | B | Standardfehler | Beta | | |
| 1 | (Konstante) | -46,305 | 25,477 | | -1,817 | ,079 |
| | Entrance | 3,155 | ,532 | ,729 | 5,925 | ,000 |
| 2 | (Konstante) | -221,659 | 41,888 | | -5,292 | ,000 |
| | Entrance | 2,521 | ,431 | ,582 | 5,854 | ,000 |
| | IQ | 1,591 | ,336 | ,471 | 4,736 | ,000 |

a. Abhängige Variable: Final

# Assumptions of multiple regression

1. At least 15 cases/subjects
2. Interval data
3. Linear relationship between predictor variables and criterion.
4. No outliers (or delete them)
5. Predictor variables should be independent of each other

# Logistic regression

# Logistic regression

- Multiple predictor variables (continuous + categorical)
- A categorical dependent variable (with two or more levels)

# Logistic regression

What determines the order of object and particle in the English verb particle construction?

(1) He looked the number up.
(2) He looked up the number.

Previous research suggests that the following factors may be relevant: the length and complexity of the direct object, the meaning and definiteness of the object, the NP type of the object (pronoun vs. lexical NP), and the occurrence of a locational PP at the end of the sentence.

# Logistic regression