Wirtschaftsstatistik

Übungsblatt Modul 6

Korrelation und Regression

# Aufgabe 1

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Filiale** | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **I** | **J** |
| **Verkaufsfläche (1.000 qm)** | 7 | 5 | 6 | 3 | 8 | 2 | 4 | 6 | 4 | 7 |
| **Filialumsatz (Mio €)** | 35 | 22 | 41 | 15 | 38 | 12 | 34 | 28 | 25 | 52 |

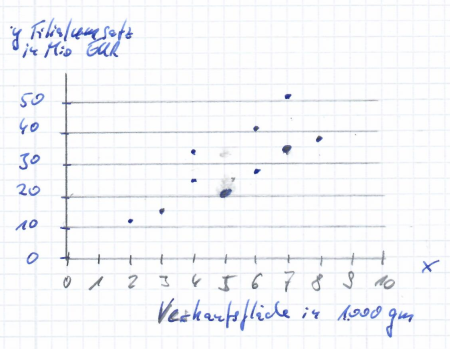
a)   
Visualisieren Sie für das obige Datenmaterial den Zusammenhang zwischen den beiden Merkmalen Verkaufsfläche und Filialumsatz in einem geeigneten Diagramm. Wie wird diese grafische Darstellung genannt?

b)

Welche Erkenntnisse liefert Ihnen eine erste qualitative Zusammenhangs- bzw. Abhängigkeitsanalyse auf der Basis des in a) erstellten Diagramms?

Lösung zu a)

Die grafische Lösung wird Streudiagramm genannt



Lösung zu b)

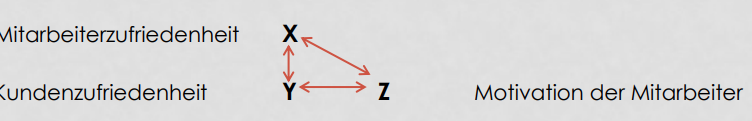
**Zusammenhangsanalyse (Interpendenzanalyse)**

Prüft,

* ob Zusammenhang zwischen Merkmalen (Variablen) besteht und, falls dem so ist,
* wie stark der Zusammenhang zwischen den Merkmalen (Variablen) ausgeprägt ist.
* Wie lässt sich die Stärke (der Grad, die Intensität) des Zusammenhangs messen
* Wie lässt sich der Zusammenhang (grafisch) darstellen?
* Lassen sich die Beobachtungswerte einer Variablen X durch mehrere andere Variablen (Y (Y1, Y2, …) näherungsweise bestimmen?
* Prüft die Wechselwirkung der Variablen untereinander
* Das Zusammenhangsmaß (Assoziationsmaß) bestimmt die Stärke und ggf. die Richtung eines Zusammenhangs

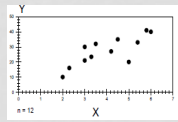
**KEIN gerichteter Zusammenhang**

* **zwischen den unabhängigen Variablen X, Y untereinander und zu einer unabhängiger Variablen Z besteht kein gerichteter Zusammenhang**
* der **Zusammenhang zwischen allen Variablen** (den unabhängigen Variablen X, Y untereinander und zwischen den unabhängigen Variablen X oder Y und der abhängigen Variablen Z) **ist bidirektional**  
  z. B. Zusammenhang zwischen den unabhängigen Variablen „Mitarbeiterzufriedenheit X“ und „Kundenzufriedenheit Y“ und der abhängigen Variable „Motivation der Mitarbeiter Z“

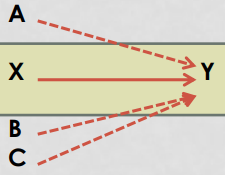


**Zusammenhangsanalyse zwischen metrischen Merkmalen X und Y heißt Korrelationsanalyse**

* Die **Korrelationsanalyse** prüft, ob ein **linearer Zusammenhang** zwischen den Variablen **besteht** und **wie stark der lineare Zusammenhang** ist.
* Grafische Darstellung **als Streudiagramm** (qualitative Zusammenhangsanalyse)



* **einfache Regressionsanalyse: Analyse des Zusammenhangs zwischen zwei metrischen Größen** 
  + **Eine Einflussgröße X (Regressor, unabhängiges Merkmal / unabhängige Variable)**
  + **Eine Zielgröße Y (Regressand, abhängiges Merkmal / abhängige Variable)**
  + **anhand von 2 Parametern wird eine Gerade durch die Punkwolke gezeichnet** und dadurch der lineare Zusammenhang gut veranschaulicht
* Es gibt **meist mehr als einen Einflussfaktor X (Regressor**, unabhängiges Merkmal / unabhängige Variable)
* **Probleme sind selten monokausal**
* **multiple Regressionsanalyse: Verallgemeinerung der einfachen linearen Regression** 
  + **mit k Regressoren (Einflussgrößen bzw. unabhängigen Merkmalen / Variablen)**, **welche den Regressand (abhängiges Merkmal / Variable) erklären sollen**
  + **mehrere Einflussgrößen** **X1, X2, …Xk** (Regressoren, unabhängige Merkmale / Variablen) und **eine Zielgröße Y** (Regressand / abhängiges Merkmal / Variable)  
    **X1, X2, …Xk → Y**



* **Zusammenhangsmaß** heißt **Korrelationskoeffizient r**
* **Korrelationskoeffizienten r** ist die **Quantifizierung der Korrelation**
* **Korrelationskoeffizienten r** informiert über,
  + **Stärke (Grad) des linearen Zusammenhangs** (der Korrelation )
  + **Richtung des linearen Zusammenhangs** (der Korrelation )
* **Korrelationskoeffizient r kann Werte zwischen -1 und +1 haben**  
  **-1 ≤ rxy ≤ +1**
* **Korrelationskoeffizient ist dimensionsloses Maß für den linearen Zusammenhang**
* **Korrelationskoeffizient r > 0: Positiver Zusammenhang**  
  **hohe Werte in einem Merkmal (Variablen) X gemeinsam mit hohen Werten in einem anderen Merkmal (Variablen) Y**
* **Korrelationskoeffizient r < 0: Negativer Zusammenhang**  
  **hohe Werte in einem Merkmal (Variablen) X mit niedrigen Werten in einem anderen Merkmal (Variablen) Y**
* **Korrelationskoeffizient r = +1: extrem starker positiver Zusammenhang** (in der Praxis sehr selten)  
  Punktewolke liegt auf einer Geraden mit positiver Steigung (von links unten nach rechts oben)
* **Korrelationskoeffizient r = -1: extrem starker negativer Zusammenhang** (in der Praxis sehr selten)  
  Punktewolke liegt auf einer Geraden mit negativer Steigung (von links oben nach rechts unten)
* **Korrelationskoeffizient r = 0: kein Zusammenhang**

**Korrelation**

* Ist **linearer (gerichteter) Zusammenhang**
* Ist **zahlenmäßiger statistischer Zusammenhang zwischen zwei metrischen Merkmalen** X und Y
* **Aufgrund des zahlenmäßiger Zusammenhangs nur bei** **metrischen Merkmalen**
* **Korrelationsanalyse: Zusammenhangsanalyse zwischen metrischen Merkmalen X und Y**
* **Positive Korrelation: beide Merkmale X und Y entwickeln sich gleichförmig:**bei höheren Werten von X hat auch Y höhere Werte  
  wenn Werte von X steigen, steigen auch die Werte von Y
* **Negative Korrelation: beide Merkmale X und Y entwickeln sich gegenläufig**bei höheren Werten von X hat Y niedrigere Wertewenn Werte von X steigen, fallen die Werte von Y
* **Kausaler Zusammenhang**: **zwischen zwei Merkmalen X und Y** existiert ein **Ursache-Wirkungs-Beziehung**  
  die **Veränderung eines abhängigen Merkmals Y ist eindeutig auf Veränderung des unabhängigen Merkmals X zurückzuführen**  
  **Korrelation** sagt **nichts über den kausalen Zusammenhang** und **nichts über die Kausalitätsrichtung** aus.  
  **Korrelation ≠ Kausalität**
* **Korrelation ist für Kausalität notwendig**, **aber nicht ausreichende Voraussetzung**nicht jede Korrelation ist ein kausaler Zusammenhang, aber jeder kausale Zusammenhang basiert auf einer Korrelation

**Korrelation**:

statistischer **Zusammenhang**

**zwischen metrischen Merkmalen**

**Scheinkorrelation**:

**Kein kausaler Zusammenhang**

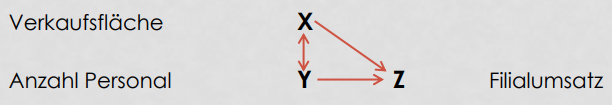
**Kausalität:**

**kausaler Zusammenhang**

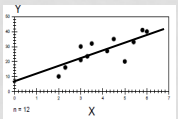
(Ursache-Wirkungs-Beziehung)

**Abhängigkeitsanalyse (Dependenzanalyse)**

* Wird **als Regressionsanalyse bezeichnet**
* **Unterscheidung zwischen abhängigen und unabhängigen Merkmalen**
* Unabhängige Merkmale beeinflussen abhängige Merkmale 🡆 **abhängige Variablen werden durch unabhängige Variablen beeinflusst**
* Es geht um **gerichteten Zusammenhang**
  + es besteht **vorab eine sachlogisch begründete Vorstellung bzw. Vermutung über den Zusammenhang zwischen den Merkmalen**
  + es i**st bekannt bzw. wird vermutet welche der Merkmale auf andere Merkmale einwirken**
  + **zwischen unabhängigen Variablen X, Y besteht ein gerichteter Zusammenhang zu einer unabhängiger Variablen Z** (Zusammenhang zwischen der unabhängigen Variablen X und der unabhängigen Variablen Z bzw. der unabhängigen Variablen Y und der unabhängigen Variablen Z ist unidirektional)  
    **zwischen den unabhängigen Variablen X, Y besteht kein gerichteter Zusammenhang** 🡆 der **Zusammenhang zwischen den unabhängigen Variablen X, Y ist bidirektional**  
    z. B. Zusammenhang zwischen den unabhängigen Variablen „Verkaufsfläche X“ und „Anzahl Personal Y“ und der abhängigen Variable „Umsatz Z“

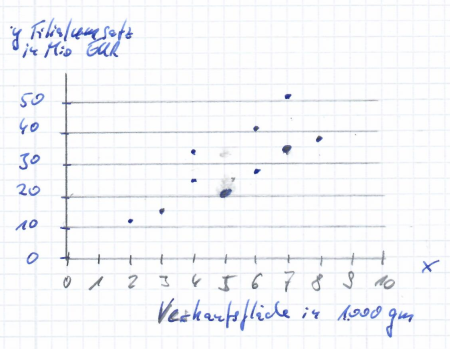


* **Abhängigkeitsanalyse muss sinnvoll** sein
* **Abhängigkeitsmaß ŷ**
* **Regressionsfunktion ŷ = a + b\*x**
* Grafische Darstellung **als Streudiagramm mit Regressionsgeraden** (qualitative Abhängigkeitsanalyse)



Lösung zu b)

Welche Erkenntnisse liefert Ihnen eine erste qualitative Zusammenhangs- bzw. Abhängigkeitsanalyse auf der Basis des in a) erstellten Diagramms?



**qualitative Zusammenhangsanalyse:**

Es besteht eine positive Korrelation (ein positiver Zusammenhang) zwischen der Verkaufsfläche und dem Filialumsatz. Eine größere Verkaufsfläche (höhere Werte zum Merkmal X „Verkaufsfläche in 1000 qm“) hat auch einen höheren Filialumsatz zur Folge (höhere Werte zum Merkmal Y „Filialumsatz in Mio EUR“).

Der lineare Zusammenhang ist „relativ“ stark.

Der Korrelationskoeffizient r muss > 0 sein.

**Qualitative Abhängigkeitsanalyse**

Es existiert ein kausaler Zusammenhang zwischen der Verkaufsfläche und dem Filialumsatz

Der Zusammenhang ist gerichtet und annähernd linear.

Die Größe der Verkaufsfläche hat Einfluss auf den Filialumsatz.

Das unabhängige Merkmal (die unabhängige Variable) beeinflusst das abhängige Merkmal (die abhängige Variable) Filialumsatz

# Aufgabe 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Monate** | **Januar** | **Februar** | **März** | **April** |
| **Produzierte Menge (in Mio. Stück)** | 5 | 3 | 2 | 4 |
| **Produktionskosten (Mio. €)** | 8 | 4 | 6 | 5 |

a)  
Berechnen Sie eine lineare Regressionsfunktion Ŷ = Ŷ(x) = a + b\*x, die den Zusammenhang zwischen produzierter Menge und Kosten möglichst gut charakterisiert   
(x produzierte Menge in Mio. Stück und Y Kosten in Mio. €)

Tipp: transponiere und erweitere obere Tabelle, um Zwischenberechnungen einzufügen

(s. z. B. Folien 20 und 39 Modul 6)

**Abhängigkeitsanalyse (lineare Regressionsfunktion Ŷ = a + b\*x): berechnet das Abhängigkeitsmaß**

**Voraussetzungen**:

* X und Y sind quantitative Merkmale
* gerichteter Zusammenhang zwischen X und Y (X -> Y)

**vorbereitende Arbeiten**

* prüfen, ob Abhängigkeitsanalyse sinnvoll ist
* Erhebung der Daten für X und Y

1. Visualisierung im Streudiagramm (qualitative Abhängigkeitsanalyse)
2. Auswahl des Funktionstyps (in diesem Seminar nur lineare Funktionen)
3. Berechnung der Regressionsfunktion (nach der Methode der kleinsten Quadrate)

**Regressionsfunktion Ŷ = a + b\*x nach der Methode der kleinsten Quadrate**

* **a und b** sind die **Regressionskoeffizienten (Kurvenparameter)**
* **Regressionskoeffizienten a und b werden so bestimmt, dass die Summe der quadratischen Abweichungen (Differenzen) von der Kurve der beobachteten Punkte minimal ist**

**∑(yi – Ŷ)²**

**∑(= (yi – (a + b\*x) )²**

**Vorgehen**

1. **REGRESSIONSRECHNUNG**
   1. Berechnungstabelle mit Hilfsgrößen **Xi \* Yi** und **Xi²** und Yi² erstellen

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Lfd. Nr | Monate | **Xi**  Produzierte Menge  (in Mio. Stück) | **Yi**  Produktionskosten  (in Mio. €) | **Xi \* Yi** | **Xi²** | Yi² |
| 1 | Januar | 5 | 8 | 40 | 25 | 64 |
| 2 | Februar | 3 | 4 | 12 | 9 | 16 |
| 3 | März | 2 | 6 | 12 | 4 | 36 |
| **4** | April | 4 | 5 | 20 | 16 | 25 |
| **Summe** |  | **14** | **23** | **84** | **54** | **141** |

* 1. **Berechnung der Regressionskoeffizienten a und b** mit Hilfe der Formeln

|  |  |
| --- | --- |
| **a =** ∑**Xi²** \* ∑**Yi** - ∑**Xi** \* ∑**Xi \* Yi**  **n** \* ∑**Xi²** - (∑**Xi**)² | **b = n** \* ∑**Xi \* Yi** - ∑**Xi** \* ∑**Yi**  **n** \* ∑**Xi²** - (∑**Xi**)² |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **a =** | ∑**Xi²** \* ∑**Yi** - ∑**Xi** \* ∑**Xi \* Yi**  **n** \* ∑**Xi²** - (∑**Xi**)² |  |  |  |
| **a =** | **54 \* 23 – 14 \* 84**  **4 \* 54 – 14²** | **1242 - 1176**  **216 - 196** | **66**  **20** | **3,3** |
|  | Hinweis:  Der Nenner muss nur 1 x berechnet werden, da er bei der Berechnung der Variablen b erneut verdendet wird. | | | |

**a = 3,3**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **b =** | **n** \* ∑**Xi \* Yi** - ∑**Xi** \* ∑**Yi**  **n** \* ∑**Xi²** - (∑**Xi**)² |  |  |  |
| **b =** | **4 \* 84 – 14 \* 23**  **4 \* 54 – 14²** | **336 - 322**  **20** | **14**  **20** | **0,7** |
|  | Hinweis:  Der Nenner muss nur 1 x berechnet werden, da er bei der Berechnung der Variablen b erneut verdendet wird. | | | |

**b = 0,7**

**Regressionsfunktion nach der Methode der kleinsten Quadrate**

**Ŷ = Ŷ(x) = a + b\*x**

**Ŷ = Ŷ(x) = 3,3 + 0,7\*x (Mio €)**

b)

Zeichnen Sie das Streudiagramm und die Regressionsfunktion in das Koordinatensystem ein. Beschriftung der Achsen nicht vergessen!

* 1. **Zeichnen des Streudiagramms anhand der Wertepaare (Korodinaten) für Xi und Yi aus der in Schritt 1.1 erstellten Tabelle**
  2. **Ermitteln der Abweichungsquadrate anhand der Formel ∑(yi – Ŷ)²**Für **Ŷ** die Werte der Regressionskoeffizienten a und b und für Xi den gegebenen Wert aus der Tabelle einsetzen  
     **ui = ∑(yi – Ŷ)² = ∑(yi – (a + b \* Xi)²**

**u ist das Residuum (unerklärte Abweichung), welches für die Regressionsgeraden** (Gerade zur Regressionsfunktion) **benötigt wird.**

**a, b sind die Regressionskoeffizienten**.   
  
Die Werte zu a, b wurden zuvor in der Nebenrechnung ermittelt.

Regressionskoeffizient **a = 3,3**

Regressionskoeffizient **b = 0,7**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Lfd. Nr | **Xi**  Menge  (in Mio. Stück) | **Yi**  Kosten  (in Mio. €) | **a** | **b** | b \* Xi | Ŷ =  a + (b \* xi) | Yi – Ŷ |
| 1 | **5** | 8 | **3,3** | **0,7** | **3,5** | 6,8 | 8 – 6,8 = **1,2** |
| 2 | **3** | 4 | **3,3** | **0,7** | **2,1** | 5,4 | 4 – 5,4 = **-1,4** |
| 3 | **2** | 6 | **3,3** | **0,7** | **1,4** | 4,7 | 6 – 4,7 = **1,3** |
| **4** | **4** | 5 | **3,3** | **0,7** | **2,8** | 6,1 | 5 – 6,1 = **-1,1** |
| **Summe** | **14** | **23** | **-** | **-** | - | - |  |

* 1. **Regressionsgerade für die Regressionsfunktion in das Streudiagramm einzeichnen**

Im Streudiagramm

* + **Koordinate für y = Ŷ = a + (b \* xi)**
* und die Punkte für die Gerade einzeichnen

1. Koordinatenpaar (x1 = 5, y1 = 8)

u1(y) = b \* x1 +a | **a = 3,3, b = 0,7**

u1(y) =0,7 \* 5 + 3,3 = 3,5 + 3,3 = 6,8

**Koordinaten (x,y) für Regressionsgerade: (5; 6,8)**

2. Koordinatenpaar (x2 = 3, y2 = 4)

u2(y) = b \* x2 +a | **a = 3,3, b = 0,7**

u2(y) =0,7 \* 3 + 3,3 = 2,1 + 3,3 = 5,4

**Koordinaten (x,y) für Regressionsgerade: (3; 5,4)**

3. Koordinatenpaar (x3 = 2, y3 = 6)

u3(y) = b \* x3 +a | **a = 3,3, b = 0,7**

u3(y) =0,7 \* 2 + 3,3 = 1,4 + 3,3 = 4,7

**Koordinaten (x,y) für Regressionsgerade: (2; 4,7)**

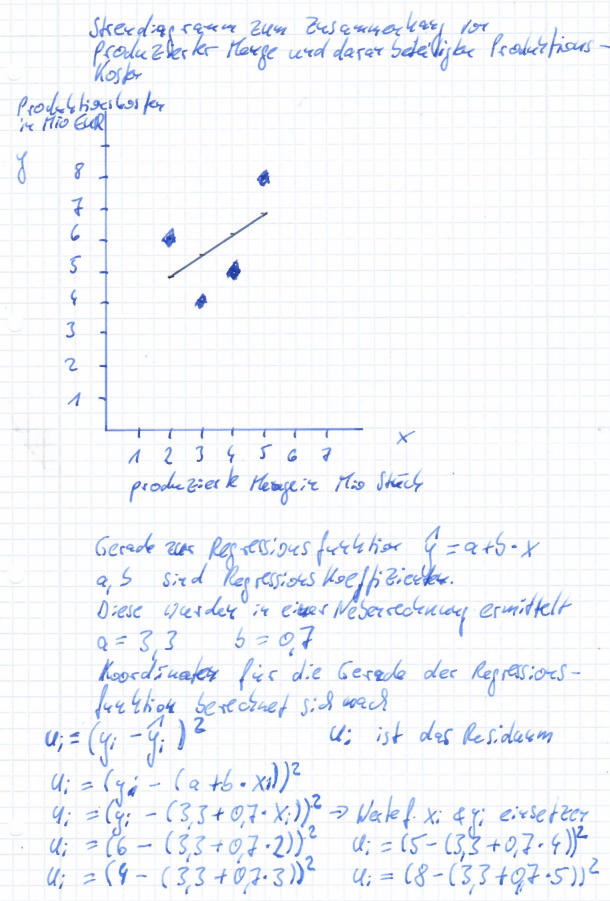
4. Koordinatenpaar (x4 = 4, y4 = 5)

u4(y) = b \* x4 +a | **a = 3,3, b = 0,7**

u4(y) =0,7 \* 4 + 3,3 = 2,8 + 3,3 = 6,1

**Koordinaten (x,y) für Regressionsgerade: (4; 6,1)**

Streudiagramm zum Zusammenhang der produzierten Menge und der daran beteiligten Produktionskosten



c)

Interpretieren Sie die beiden Regressionskoeffizienten der berechneten Regressionsfunktion betriebswirtschaftlich.

Wie nennt man eine solche Funktion in der Betriebswirtschaftslehre?

Regressionskoeffizient **a = 3,3**

Regressionskoeffizient **b = 0,7**

**Produzierte Menge in Mio Stück: 2**

**Produktionskosten in Mio EUR**

3,3 + 0,7 \* **2** = 3,3 + 1,4 = **4,7** 4,7 : 2 = 2,35 Mio EUR / 1 Mio Stück produzierte Menge

**Produzierte Menge in Mio Stück: 3**

**Produktionskosten in Mio EUR**

3,3 + 0,7 \* **3** = 3,3 + 2,1 = **5,4** 5,4 : 3 = 1,8 Mio EUR / 1 Mio Stück produzierte Menge

**Produzierte Menge in Mio Stück: 4**

**Produktionskosten in Mio EUR**

3,3 + 0,7 \* **4** = 3,3 + 2,8 = **6,1** 6,1 : 4 = 1,525 Mio EUR / 1 Mio Stück produzierte Menge

**Produzierte Menge in Mio Stück: 5**

**Produktionskosten in Mio EUR**

3,3 + 0,7 \* **5** = 3,3 + 3,5 = 6,8 6,8 : 5 = 1,36 Mio EUR / 1 Mio Stück produzierte Menge

Interpretation der Regressionskoeffizienten:

**a = 3,3 Mio sind Fixkosten,**

**b = 0,7 sind variable (Stück)-Kosten**

**Die Stückkosten fallen bei zunehmender produzierte Menge.**

**In der Kostenrechnung/BWL wird diese Funktion lineare Kostenfunktion genannt.**

d)

Erstellen Sie auf der Basis der in a) ermittelten Regressionsfunktion eine

Kostenprognose für den Monat Mai, in dem eine Produktionsmenge von 6 Mio. Stück geplant ist

Kostenprognose für den Monat Mai mit 6 Mio Stück produzierte Menge

Ŷ(6) = 3,3 + 0,7 \* 6 = 3,3 + 4,2 = 7,5 Mio EUR

e)

Berechnen Sie den Korrelationskoeffizienten und das Bestimmtheitsmaß zur obigen

Regressionsrechnung

**Zusammenhangsanalyse zwischen zwei metrischen Merkmalen X und Y**

**Korrelationskoeffizient (Zusammenhangsmaß)** r

-1 ≤ r ≤ +1

* **ist dimensionsloses Maß für den linearen Zusammenhang**
* **Korrelationskoeffizient r > 0: Positiver Zusammenhang**  
  **hohe Werte in einem Merkmal (Variablen) X gemeinsam mit hohen Werten in einem anderen Merkmal (Variablen) Y**
* **Korrelationskoeffizient r < 0: Negativer Zusammenhang**  
  **hohe Werte in einem Merkmal (Variablen) X mit niedrigen Werten in einem anderen Merkmal (Variablen) Y**
* **Korrelationskoeffizient r = +1: extrem starker positiver Zusammenhang**   
  Punktewolke liegt auf einer Geraden mit positiver Steigung (von links unten nach rechts oben)
* **Korrelationskoeffizient r = -1: extrem starker negativer Zusammenhang**   
  Punktewolke liegt auf einer Geraden mit negativer Steigung (von links oben nach rechts unten)
* **Korrelationskoeffizient r = 0: kein Zusammenhang**

**Korrelationsrechnung ist die „Standardisierung“ der Kovarianz**

* **Zwei metrische Merkmale X, Y müssen** 
  + **mindestens intervallskaliert** sein (intervallskalierte, verhältnisskalierte oder absolut-skalierte Merkmale)
  + **jeweils eine positive Standardabweichung** haben und
  + **eine Kovarianz COV(X, Y)** haben

Für die Korrelationsrechnung (nach Pearson) werden für jedes Merkmal X, Y

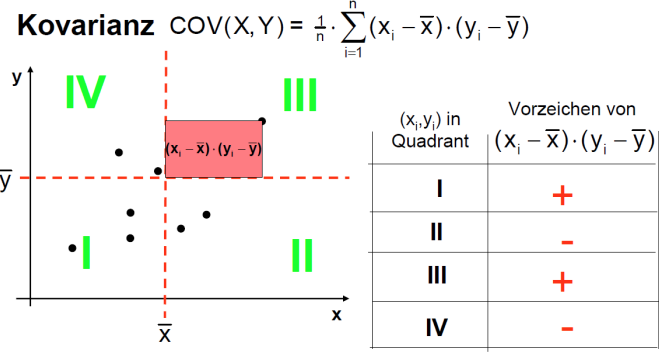
* die Kovarianz COV(X,Y) und
* die Standardabweichung Sx, Sy   
  (berechnet sich aus der Wurzel zur Summe der Quadrate aus der Differenz aus Merkmalswert und Mittelwert geteilt durch die Anzahl der Merkmale)  
  Sx = √∑((xi – x̅ )²) : n

benötigt

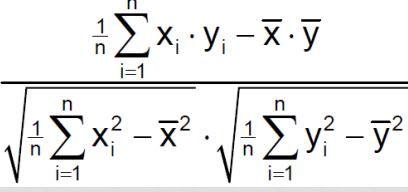
Korrelationskoeffizient 

**Kovarianz COV(X,Y)**

* informiert über gemeinsame Variabilität der Merkmale X und Y.
* Ist der **Zusammenhang (Korrelation) positiv, ist auch die Kovarianz positiv**
* ist der **Zusammenhang (Korrelation) negativ, ist auch die Kovarianz negativ**.
* **kein (linearer) Zusammenhang zwischen X und Y 🡆 Kovarianz liegt in der Nähe von 0**.
* **Bei positiver Kovarianz (= positivem Zusammenhang (Korrelation)) 🡆 steigende Gerade im Streudiagramm** (von links unten nach rechts oben)
* **Bei negativer Kovarianz (= positivem Zusammenhang (Korrelation)) 🡆 fallende Gerade im Streudiagramm** (von links oben nach rechts unten)



**Kovarianz COV(X,Y) = (xi – x̅) \* (yi – y̅) : n**



|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Lfd. Nr | Monate | **Xi**  Produzierte Menge  (in Mio. Stück) | **Yi**  Produktionskosten  (in Mio. €) | **xi \* yi** | **x²** | **y²** |  |
| 1 | Januar | 5 | 8 | 40 | 25 | 64 |  |
| 2 | Februar | 3 | 4 | 12 | 9 | 16 |  |
| 3 | März | 2 | 6 | 12 | 4 | 36 |  |
| **4** | April | 4 | 5 | 20 | 16 | 25 |  |
| **Summe** |  | **14** | **23** | **84** | **54** | **141** |  |

**x̅ = 14 : 4 = 3,50 x̅² = 12,25**

**y̅ = 23 : 4 = 5,75 y̅² = 33,06**

**x̅ \* y̅ = 3,5 \* 5,75 = 20,12**

rx,y =

¼ \* 84 – 3,5 \* 5,75 21 – 20,125 0,875

0,875 0,875 = 0,52914

1,118 \* 1,479 1,6536

Korrelationskoeffizient rx,y = 0,529

**Bestimmtheitsmaß 𝑹²**

Bei **einfacher linearen Regression (nur eine unabhängige Variable) entspricht das Bestimmtheitsmaß dem Quadrat des Korrelationskoeffizienten nach Pearson rXY**

𝑹² = (𝑟𝑋𝑌)²

**𝑹²** = 0,529² **= 0,28**

f)

Interpretieren Sie das Bestimmtheitsmaß

**Lehre zum Bestimmtheitsmaß**

*WS\_Modul\_6\_Korrelation\_und\_Regression.pdf, S. 31*

Anwendungen des Regressionsverfahren lassen sich in die Kategorien

* **Erstellen von Vorhersagemodellen (Prognosen)**
* **Qualifizierung der Stärke von Zusammenhängen (Korrelation)** 
  + dient der **Ermittlung der Merkmalsausprägungen xj, die keinen Zusammenhang mit dem Merkmal y haben** oder
  + der Ermittlung der Teilmengen xi, … , xj die redundante Informationen über y enthalten.

einteilen

**Vorhersagemodelle**

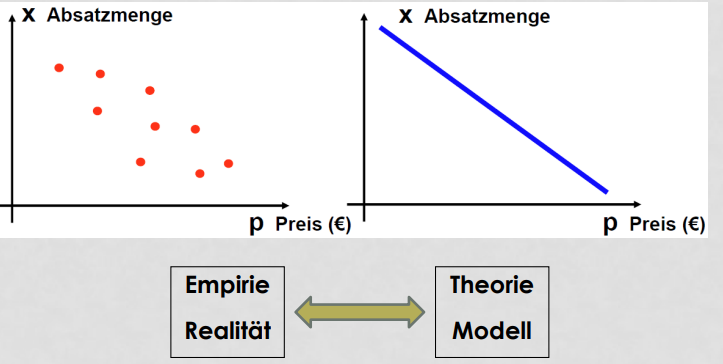
*WS\_Modul\_6\_Korrelation\_und\_Regression.pdf, S. 32 – S. 34*

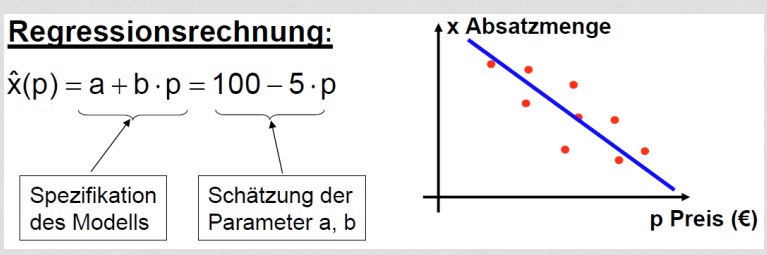
**MODELL VS. REALITÄT**

**Fragestellung/Problem:**

* **Wie gut beschreibt das Modell die Realität?** bzw.   
  **Wie gut wird die Realität durch das Modell repräsentiert**?

**Modell ist das vereinfachte Abbild der Realität**





* **Wie gut ist die Anpassungsgüte** („goodness of fit“)?
* **Wie gut beschreibt die Regressionsfunktion Ŷ = a + b\*x die Abhängigkeit?**
* Es wird ein **Gütemaß für die Schätzung der** in der Regressionsfunktion verwendeten **Parameter benötigt**

Bsp:

Unabhängiges Merkmal (Variable) abhängiges Merkmal (Variable)

**Verkaufsflächen** **Filialumsatz**

**unterschiedlich groß** **unterschiedlich hoch**

**WARUM?**

* Wie gut erklären die Unterschiede der Verkaufsflächen (unabhängiges Merkmal) die unterschiedlichen Filialumsätze (abhängiges Merkmal)
* **Wieviel Varianz wird durch das Modell nicht erklärt?**(**Varianz** ist der **arithmetische Mittel der Abweichungsquadrate 🡆   
   *s² =*  oder**

* Wie **gut erklärt die Regressionsfunktion die Abhängigkeit zwischen dem unabhängigen Merkmal** (unabhängigen Variable) Verkaufsfläche **und dem abhängigen Merkmal** (abhängigen Variable) Filialumsatz?
* **Wie groß ist die Erklärungskraft des Modells?**

**PROGNOSEWERTE UND RESIDUEN**

**Abweichungszerlegung:**

*WS\_Modul\_6\_Korrelation\_und\_Regression.pdf, S. 35 – S. 36*

**Residuum (unerklärte Abweichung) wird für die Regressionsgeraden (Gerade zur Regressionsfunktion) benötigt**

**Residuum ui** = **(𝒚𝒊 –** **ŷ)**

**ŷ = a + b \* x**

a, b sind die Regressionskoeffizienten

**Abweichungszerlegung:**

**(yi – y̅ = (yi – ŷi) + (ŷi – y̅i)**

**Bsp. für Filiale Nr. 3**

Koordinaten (xi; **yi**) = (**6**; **40**) xi = Verkaufsfläche, yi = Filialumsatz

**Mittelwert y̅ = 25**

Regressionskoeffizient a = 5

Regressionskoeffizient b = 5

**Regressionsfunktion: ŷ = a + b \* x 🡺 ŷ = 5 + 5 \* x 🡺 ŷ = 5 + 5 \* 6 🡺**  **ŷ = 5 +30 🡺**  **ŷ = 35**

**Abweichungszerlegung:**

**yi = abhängige Variable (abhängiges Merkmal)**

**y̅ = Mittelwert der abhängigen Variable y (abhängiges Merkmal)**

**ŷ = Ergebnis der Regressionsfunktion zur abhängigen y Variable (abhängiges Merkmal)**

|  |  |  |
| --- | --- | --- |
| **Gesamtabweichung** | **Unerklärte Abweichung**  **/Residuum** | **erklärte Abweichung** |
| **(yi – y̅ )** | **(yi – ŷi)** | **(ŷ – y̅)** |
| Hinweis:  Der **2. Parameter** zur Gesamtabweichung ist der **Mittelwert y̅** zur abhängigen Variable y  Gesamtabweichung der Beobachtung yi zum Mittelwert y̅  Vorhersage mit der Gesamtbeobachtung ist falsch, wenn diese mit dem Mittelwert y̅ erfolgt. | Hinweis:  Der **2. Parameter** zur unerklärten Abweichung ist das Ergebnis der **Regressionsfunktion ŷi**  unerklärte Abweichung der Beobachtung yi zur Regressionsgeraden (Residuum ui)  Dieser Fehler (die unerklärte Abweichung kann auch durch die Einbeziehung der unabhängigen Variable xi nicht vermieden werden | Hinweis:  Der **1. Parameter** zur erklärten Abweichung ist das **Ergebnis der Regressionsfunktion**  ŷi .  Der **2. Parameter ist der Mittelwert y̅** zur abhängigen Variable (Beobachtung).  erklärte Abweichung der Regressionsgeraden ŷi zum Mittelwert y̅.  Dieser Fehler ist durch die Einbeziehung der unabhängigen Variablen x vermeidbar. |
| **40 – 25 = 15** | **40 – 35 = 5** | **35 – 25 = 10** |

**yi – y̅ = (yi - ŷi) + (ŷi – y̅)**

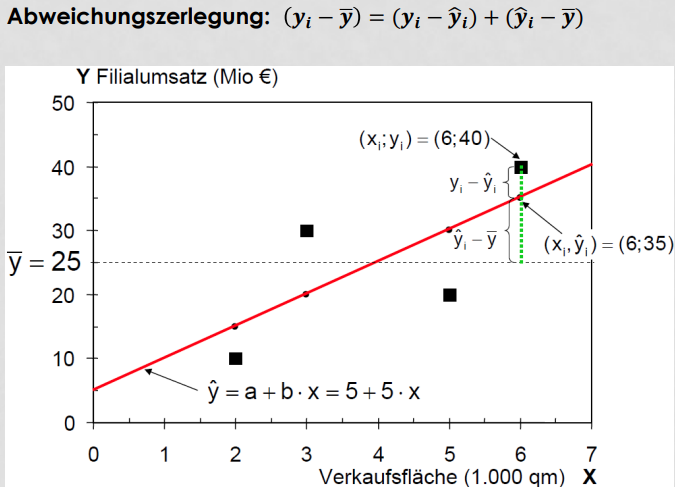
**Filiale Nr. 3:**

**xi = 6 Verkaufsfläche in 1.000 m²**

**yi = 40 Filialumsatz in Mio. Euro**

**40 – 25 = (40 – 35) + (35 – 25)**

**15 = 5 + 10**



**PROGNOSEWERTE UND RESIDUEN**

**Varianzzerlegung:**

*WS\_Modul\_6\_Korrelation\_und\_Regression.pdf, S. 37 – S. 39*

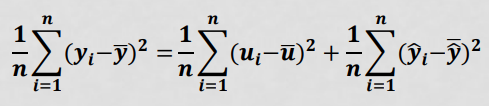
**WICHTIG:**

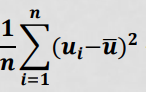
**Unterschied zwischen Abweichungszerlegung und Varianzzerlegung**

**Abweichungszerlegung ohne Potenzierung (Quadrierung)**

**yi – y̅ = (yi - ŷi) + ( ŷi – y̅ )**

**Varianzzerlegung mit Potenzierung (Quadrierung)**

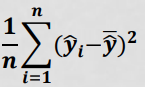


**su2 =  bzw. 🡆 Varianz des Residuum u (su2)**

Regressionsfunktion ŷ = yi - (a + b \* x)

Funktion zum Residuum = u = yi – ŷi

Varianz der Residuen su² = ( ∑ (yi - (a + b \* x) )² : n – ( ∑ui )² : n

**sŷ² =  bzw. 🡆Varianz der Regressionsfunktion ŷ (Sŷ²)**

Regressionsfunktion ŷ = yi - (a + b \* x)

Varianz der Regressionsfunktion **sŷ² = ( (∑ (a + b \* x)² : n ) – y̅²**

**Varianz des abhängigen Merkmals y (sy²) ist gleich der Summe aus der Varianz zum Residuum (su2 ) und der Varianz der Regressionsfunktion (sŷ²)**

**Sy² = su² + sŷ²**

* Die Varianz der Regressionswerte wird auch durch die Varianz des unabhängigen Merkmal x bestimmt.

Varianz = Regressions- \* Varianz

d. Regressionswerte koeffizient b² d. abhängigen Merkmals x

**sŷ² = b² \* sx²**

**Varianz zum Residuen**

Hinweis:

Zur Berechnung der Varianz der Residuen die Regressionsfunktion mit den Regressionsparametern berechnen **ŷ = (a + b \* x)**

Dann die Differenz aus Variable yi und Regressionsfunktion berechnen **yi - ŷ = yi - (a + b \* x)**

Jede Differenz quadrieren und mit den Quadraten der anderen Differenzen summieren.

Die Summe durch die Anzahl der Ausprägungen n dividieren.

Bei der Berechnung zur Varianz der Regressionswerte wurde bereits für jede Variable yi die Regressionsfunktion ŷ = (a + b \* x) berechnet.

Auf diesen Werten basieren die Residuen ui .

Zu diesen muss für jedes yi die Differenz gebildet werden

yi - ŷ

Die Differenz muss quadriert und die Quadrate summiert und durch die Anzahl n dividiert werden

ui² = ∑(yi – ŷ)² : n

Zu ui den Mittelwert u̅ bilden und den Mittelwert quadrieren u̅²

Das Quadrat zum Mittelwert von der Quadratsumme der Residuen subtrahieren ui² - u̅²

Regressionskoeffizient a = 5

Regressionskoeffizient b = 5

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Filiale-Nr i | **xi**  **Verkaufsfläche** | yi  Filialumsatz | a | **b** | Regressionswert  **ŷ =** a **+ b \* x** | **ui = yi - ŷ** | **ui²** |
| 1 | **3** | 30 | 5 | **5** | 20 | 10 | 100 |
| 2 | **2** | 10 | 5 | **5** | 15 | -5 | 25 |
| 3 | **6** | 40 | 5 | **5** | 35 | 5 | 25 |
| 4 | **5** | 20 | 5 | **5** | 30 | -10 | 100 |
| Summe | **16** | **100** | - | - | **100** | **0** |  |

su² = ( ( ∑ui )² : n ) - ( ( ∑ui )² : n ) : n – ( ∑ui )² : n

su² = ( ( 10² + (-5)² + 5² + (-10)² ) : 4 ) - 0² : 4 Hinweis:

Das Quadrat zu einem negativem Residuum

ist positiv

Hier (-5)² = 25, (-10)² = 100

Der Mittelwert berechnet sich jedoch aus der   
 Summe der Residuen zum Quadrat   
 geteilt durch die Anzahl der Beobachtungswerte  
 (∑ui )² : n  
 Summe der Residuen ist hier 0  
 10 + (-5) + 5 + (-10) = 0

su² = ( (100 + 25 + 25 + 100) : 4 )

su² = 250 : 4 🡆 su² = 62,5

**Varianz der unabhängigen Variablen x sx²**

**Varianz der abhängigen Variablen y sy²**

**Varianz der Regressionskoeffizienten ŷ sŷ²**

Regressionskoeffizient a = 5

Regressionskoeffizient b = 5

**x̅ = 16 : 4 = 4 x̅² = 4² = 16**

**y̅ = 100 : 4 = 25 y̅² = 25² = 625**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Filiale-Nr i | **xi**  **Verkaufsfläche** | yi  Filialumsatz | xi² | yi² | a | **b** | Regressionswert  **ŷ =** a **+ b \* x** | **ŷ²** |  |
| 1 | **3** | 30 | 9 | 900 | 5 | **5** | 20 | 400 |  |
| 2 | **2** | 10 | 4 | 100 | 5 | **5** | 15 | 225 |  |
| 3 | **6** | 40 | 36 | 1600 | 5 | **5** | 35 | 1225 |  |
| **4** | **5** | 20 | 25 | 400 | 5 | **5** | 30 | 900 |  |
| Summe | **16** | **100** | **74** | **3000** | - | - | **100** | **2750** |  |
|  |  |  |  |  | ŷ̅ = 100 : 4  ŷ̅ = 25 | | | ∑**ŷ² : n**  **2750 : 4** |  |
|  |  |  |  |  |  |  | **ŷ̅² = 625** | **687,5** |  |

sx² = ( ∑**xi²** ) : n – x̅² **74** : 4 – **16** = **2,5**

sy² = ( ∑yi² ) : n – y̅² **3000** : 4 – **625** = **125**

sŷ² = ( ∑ŷ² ) : n – ŷ̅² (20² + 15² + 35² + 30²) : 4 – 25² = ( **2750** : 4 ) – **625** = **687,5** – **625** = **62,5**

* Die Varianz der Regressionswerte wird auch durch die Varianz des unabhängigen Merkmal x bestimmt.

Varianz = Regressions- \* Varianz

d. Regressionswerte koeffizient b² d. abhängigen Merkmals x

**sŷ² = b² \* sx²**

62,5 = = 5² = 25 \* 2,5

62,5 = 62,5

**Bestimmtheitsmaß R² (Erklärungskraft d. Modells)**

*WS\_Modul\_6\_Korrelation\_und\_Regression.pdf, S. 37 – S. 39*

* Ist **Gütemaß des Modells**
* **Wie gut kann unabhängige Variable (abhängiges Merkmal)**, die **Varianz der abhängigen Variable (abhängiges Merkmal) erklären**

**R² = 0% 🡆 unbrauchbares Modell**

**R² = 100 % perfekte Modellanapassung**

**(unbrauchbares Modell) 0% ≤ R² ≤ 100% (perfekte Modellanpassung)**

* Bestimmtheitsmaß **R² zerlegt die Varianz der abhängigen Variablen** (abhängigen Merkmals) **in erklärte Varianz und nicht erklärte Varianz**

**ŷ = Ergebnis der Regressionsfunktion bzw. Regressionswert (a + b \* x)**

**Bestimmtheitsmaß R² ist das Verhältnis (der Quotient) aus der Streuung der Prognosewerte (erklärte Varianz) und unerklärte Varianz und der Gesamtstreuung der y-Werte**

* **Bei einfacher linearer Regression** (**nur eine unabhängige Variable**) entspricht des **Bestimmtheitsmaß R² dem Quadrat des Korrelationskoeffizienten** nach Pearson rxy

**R² = (rxy)²**

**Lösung zu f)**

Interpretieren Sie das Bestimmtheitsmaß

Korrelationskoeffizient rx,y = 0,529

**Bestimmtheitsmaß 𝑹²** = 0,529² **= 0,28**

28% der Varianz zu den Produktionskosten (abhängige Variable y) resultieren aus der Varianz der produzierten Menge (unabhängige Variable x). Das bedeutet, dass nur 28% der unterschiedlichen Produktionskosten im Monat durch die unterschiedlichen Produktionsmengen erklärt werden können.

72% sind nur durch andere Einflussfaktoren erklärbar.

g)

Berechnen Sie die Varianz der **Produktionsmengen, der Produktionskosten und der Regressionswerte** (Summen aus der Berechnungstabelle verwenden).

Zeigen Sie an diesem Beispiel, dass das Bestimmtheitsmaß angibt, welcher Anteil der Varianz der Produktionskosten erklärt wird durch die Regressionsfunktion bzw. die Varianz der Produktionsmengen

Tipp: Varianzzerlegung s. Modul 6

*WS\_Modul\_6\_Korrelation\_und\_Regression.pdf, S. 37 – S. 38*

Regressionskoeffizient a = 3,3

Regressionskoeffizient b = 0,7

**x̅** = 14 : 4 = **3,50** **x̅² = 12,25**

**y̅** = 23 : 4 = **5,75** **y̅² = 33,0625**

= 23 : 4 = **5,75 = 33,0625**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Lfd. Nr | **xi**  Menge  (in Mio. Stück) | **yi**  Kosten  (in Mio. €) | **xi \* yi** | **xi²** | x̅ | x̅² | **yi²** | y̅ | y̅² | Ŷ =  a + (b \* xi) | Ŷ² |
| 1 | **5** | 8 | 40 | 25 | **3,5** | **12,25** | 64 | **5,75** | **33,0625** | 6,8 | 46,24 |
| 2 | **3** | 4 | 12 | 9 | **3,5** | **12,25** | 16 | **5,75** | **33,0625** | 5,4 | 29,16 |
| 3 | **2** | 6 | 12 | 4 | **3,5** | **12,25** | 36 | **5,75** | **33,0625** | 4,7 | 22,09 |
| **4** | **4** | 5 | 20 | 16 | **3,5** | **12,25** | 25 | **5,75** | **33,0625** | 6,1 | 37,21 |
| **Summe** | **14** | **23** | **84** | **54** | **-** | **-** | **141** | **-** | **-** | **23,0** | **134,7** |
|  |  |  |  |  |  |  | = **23 : 4 = 5,75** | | | |  |
|  |  |  |  |  |  |  | **= 33,0625** | | | |  |

**Varianz der Produktionsmengen**

**sx² = ( ( ∑xi)²: n ) – x̅²**

**sx² = ( 54 : 4 ) – 3,5² = 13,5 - 12,25 = 1,25**

**sx² = 1,25**

**Varianz der Produktionskosten**

**sy² = ( ( ∑yi)²: n ) – x̅²**

**sy² = ( 141 : 4 ) – 5,75² = 35,25 – 33,0625 = 2,1875**

**sy² = 2,1875**

**Varianz der Regressionswerte**

Hinweis:

Der **Mittelwert für ŷ̅ ist der Mittelwert von y̅ (5,75)** (d. h. **NICHT** den Mittelwert von ŷ berechnen).

sŶ² = ( ( **∑**Ŷ²) : n ) – Ŷ̅²

sŶ² = ( ( 6,8² + 5,4² + 4,7² + 6,1²) : 4 ) – ( 23 : 4 ) ² = ( ( 46,24 + 29,16 + 22,09 + 37,21) : 4 ) – ( **5,75** )²

sŶ² = ( **134,7** : **4** ) – **( 5,75 )²**

sŶ² = **33,675 – 33,0625** = **0,6125**

**sŶ² = 0,6125**

Bestimmtheitsmaß R²

# Aufgabe 3

Die Marktforschungsabteilung eines Unternehmens will für ein neues Produkt die Abhängigkeit der Absatzmenge x vom Preis Y empirisch untersuchen. Dazu wird in 4

vergleichbar großen Testmärkten unter sonst annähernd gleichen Rahmenbedingungen 8 Wochen lang jeweils ein unterschiedlicher Preis gefordert. Die Ergebnisse dieses Tests stehen in der folgenden Tabelle:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Testmärkte** | **A** | **B** | **C** | **D** |
| **x: Preis pro Stück (€)** | 1,90 | 2,10 | 1,50 | 2,50 |
| **y: Absatzmenge in 8 Wochen (Stück)** | 5.000 | 4.000 | 8.000 | 3.000 |

a)

Berechnen Sie eine lineare Regressionsfunktion Ŷ = Ŷ(x) = a + b\*x, die die Abhängigkeit zwischen Preis und Absatzmenge möglichst gut charakterisiert. (x Preis in € und Y Absatzmenge in Stück)

Formel zur Berechnung der Regressionskoeffizienten a und b

Regressionskoeffizient

Regressionskoeffizient

**WICHTIGER HINWEIS:**

**Der Nenner ( ) muss nur einmal berechnet werden und wird in beiden Regressionskoeffizienten verwendet.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Testmärkte | x  Preis  pro Stück (€) | y  Absatzmenge  in 8 Wochen (Stück) | x² | y² | x \* y | Ŷ = a+b\*x |  |
| A | 1,90 | 5000 | 3,61 | 25.000.000 | 9.500 | 5.500 |  |
| B | 2,10 | 4000 | 4,41 | 16.000.000 | 8.400 | 4.500 |  |
| C | 1,50 | 8000 | 2,25 | 64.000.000 | 12.000 | 7.500 |  |
| D | 2,50 | 3000 | 6,25 | 9.000.000 | 7.500 | 2.500 |  |
| **SUMMEN** | **8,00** | **20.000** | **16,52** | **114.000.000** | **37.400** | **20.000** |  |
|  |  |  |  |  |  |  |  |

**Regressionsfunktion: Ŷ = Ŷ(x) = a + b\*x = Ŷ = 15.000 + (- 5.000) \* x (Mio €)**

Regressionswerte (Koordinaten für die Regressionsgeraden): **y = a + b\*x**

Ŷ = 15.000 + (- 5.000) \* 1,90 = **5.500**

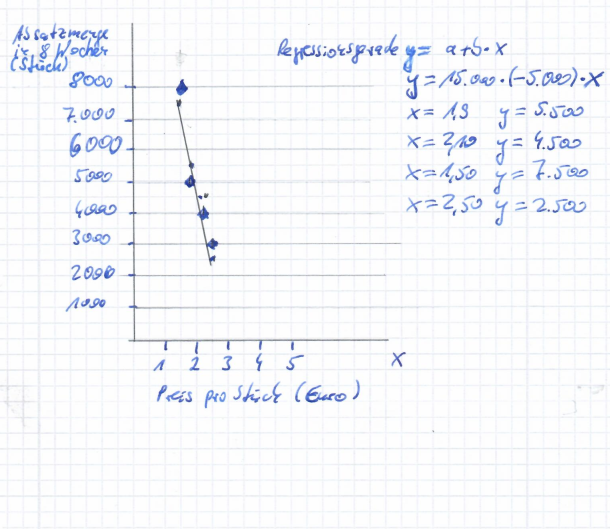
Ŷ = 15.000 + (- 5.000) \* 2,10 = **4.500**

Ŷ = 15.000 + (- 5.000) \* 1,50 = **7.500**

Ŷ = 15.000 + (- 5.000) \* 2,50 = **2.500**

b)

Zeichnen Sie das Streudiagramm und die Regressionsfunktion in das Koordinatensystem ein. Beschriftung der Achsen nicht vergessen!



c)

Interpretieren Sie die beiden Regressionskoeffizienten der in a) berechneten Regressionsfunktion betriebswirtschaftlich

Regressionskoeffizient a = 15.000

Regressionskoeffizient b = -5.000

Regressionsfunktion Ŷ = a + b \* x Ŷ = 15.000 + -5.000x

x = Preis pro Stück

y = Absatzmenge

Bei einem Preis von 0,00 EUR werden 15.000 Stück verkauft

15.000 + -5.000x 🡆 15.000 + (-5000) \* 0,00 € = 15.000 Stück

15.000 Stück sind damit die Obergrenze für den Verkauf im Testmarkt.

Mit zunehmenden Verkaufspreis (unabhängiges Merkmal x) nimmt die Absatzmenge (abhängiges Merkmal y) ab.

Pro 1 EUR Preissteigerung sinkt der Absatz um 5000 Stück.

Ab einem Verkaufspreis von 3,00 € ist der Absatz = 0 Stück

15.000 + -5.000x 🡆 15.000 + (-5000) \* 3,00 € = 0,00 Stück

d)

Berechnen Sie den Korrelationskoeffizienten und das Bestimmtheitsmaß zur obigen Regressionsrechnung

**Korrelationskoeffizienten rx,y**

🡆 Kovarianz

🡆 Standardabweichung der unabhängigen Variable x

🡆 Standardabweichung der abhängigen Variable y

**x̅** = 8 : 4 = **2 x̅² = 4**

**y̅** = 20.000 : 4 = **5.000 y̅² = 25.000.000**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Testmärkte | x  Preis  pro Stück (€) | y  Absatzmenge  in 8 Wochen (Stück) | x² | y² | xi \* yi | x̅ \* y̅ |
| A | 1,90 | 5000 | 3,61 | 25.000.000 | 9.500 | 10.000 |
| B | 2,10 | 4000 | 4,41 | 16.000.000 | 8.400 | 10.000 |
| C | 1,50 | 8000 | 2,25 | 64.000.000 | 12.000 | 10.000 |
| D | 2,50 | 3000 | 6,25 | 9.000.000 | 7.500 | 10.000 |
| **SUMMEN** | **8,00** | **20.000** | **16,52** | **114.000.000** | **37.400** | **-** |
|  |  |  | ¼ \* 16,52 = 4,13 |  |  |  |
|  |  |  |  |  |  |  |

**rx,y = - 0,9636**

**Varianz von y 🡆 sy² 1/n \* ∑(yi² - y̅ ²)**

**∑yi² = 114.000.000** 🡆 Wert aus der Tabelle

**y̅ = 20.000 : 4 = 5.000 🡆 y̅² = 25.000.000**

**sy²** = ¼ \* 114.000.000 – 25.000.000 =

**sy²** = 28.500.000 – 25.000.000

**sy²** = **3.500.000**

**Varianz der Regressionswerte sŷ² = 1/n \* ∑(ŷ² - ŷ̅² )**

Regressionskoeffizient a = 15.000

Regressionskoeffizient b = - 5.000

Regressionswerte (Koordinaten für die Regressionsgeraden): **y = a + b\*x**

a b x = **a + b \* x**

Ŷ = 15.000 + (- 5.000) \* 1,90 = **5.500**

Ŷ = 15.000 + (- 5.000) \* 2,10 = **4.500**

Ŷ = 15.000 + (- 5.000) \* 1,50 = **7.500**

Ŷ = 15.000 + (- 5.000) \* 2,50 = **2.500**

SUMME Ŷ 20.000

**y̅ = 20.000 : 4 = 5.000 🡆 y̅² = 25.000.000**

**WICHTIG**

**Nicht die Summe der Regressionswerte quadrieren** (also 20.000² = 400.000.000) **sondern jeden einzelnen Regressionswert quadrieren und dann summieren.**

sŷ² = 1/n \* ∑(ŷ² - y̅²)

sŷ²= ¼ \* (**5.500² + 4.500² + 7.500² + 2.500²**) - **25.000.000**

sŷ²= ¼ \* 113.000.000 - 25.000.000

sŷ²= 28.250.000 – 25.000.000

**sŷ²= 3.250.000**

**Varianz der Residuen su² 🡆 u = y – ŷ = y – (a + b \* x)**

su² = (1/n \* ∑u²) - u̅²

su² = (1/n \* ∑( y – (a + b \* x)))² - (1/n \* ∑( y – (a + b \* x))²) \* 1/n

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| yi | ŷi | u = yi – ŷ |  | ui² |
| 5.000 | 5.500 | -500 | Das Quadrat einer negativen Zahl ist eine positive Zahl  -500² = 250.000 | 250.000 |
| 4.000 | 4.500 | -500 | 250.000 |
| 8.000 | 7.500 | 500 | 250.000 |
| 3.000 | 2.500 | 500 | 250.000 |
| SUMME |  | 0 |  | 0 |

u̅ = 0 / 4 = 0 🡆 u̅² = 0

su² = 1/n \* ∑u² - u̅²

su² = ¼ \* ( (-500)² + (-500)² + 500² + 500²) - 0

su² = ¼ (250.000 + 250.000 + 250.000 + 250.000) – 0

su² = ¼ \* 1.000.000

**su² = 250.000**

e)

Interpretieren Sie das Bestimmtheitsmaß für diesen Fall

* Bestimmtheitsmaß R² ist Gütemaß der linearen Regression und informiert, **wie gut die unabhängige Variable geeignet ist, die Varianz der abhängigen Variable x zu erkläre**n.
  + - unbrauchbares Modell zur linearen Regression, wenn R² = 0%
    - perfektes Modell zur linearen Regression, wenn R² = 100%
* nutzt das Konzept der Varianzzerlegung 🡆 **Gesamtvarianz des abhängigen Merkmales y wird in nicht erklärte Varianz der Residuen (Residualvarianz) u und erklärte Varianz des Regressionswerts ŷ** zerlegt

**Formel zur Varianzzerlegung**

Gesamt- Residual- Varianz

varianz varianz Regressionswert ŷ

ŷi = a + b \* x

sy² = su² + sŷ²

**sy² = 1/n \* ∑(yi – y̅ )²**

**su² = 1/n \* ∑(ui – u̅ )²** ui = yi - ŷi

**sŷ² = 1/n \* ∑(ŷi – ŷ̅ )²** ŷi = a + b \* x

* **Bestimmtheitsmaß R²**   
  ist der **Anteil Varianz der abhängigen Variable y** (abhängiges Merkmal) an der Gesamtvarianz, **der sich durch die Varianz der unabhängigen Variable x** (unabhängiges Merkmal) **erklären lässt**.

**Bestimmtheitsmaß R² aus dem Verhältnis von Varianz der Regression aus sŷi² zur Gesamtvarianz aus sy²**

oder

**Bestimmtheitsmaß R² aus dem Verhältnis von Varianz der Residuen aus su² zur Gesamtvarianz aus sy²**

**Bei einer einfacher linearer Regression (mit nur einer unabhängigen Variablen y) entspricht das Bestimmtheitsmaß R² dem Quadrat der Korrelationskoeffizienten.**

**R² = (rx,y) = 0,9636² = 0,92852**

92,852 % der Varianz (Unterschiede) in den Absatzmengen lassen sich durch die Preisunterschiede (Varianz der Preise) in den verschiedenen Testmärkten erklären

7,15 % des Varianz zu den Absatzmengen werden durch andere Faktoren) beeinflusst und lassen sich nicht erklären

f)

Erstellen Sie auf der Basis der in a) ermittelten Regressionsfunktion eine Absatzmengen-Prognose für einen Testmarkt M, der den Testmärkten A, B, C und D entspricht.

Welche Absatzmenge ist dort in 8 Wochen zu erwarten bei einem Preis von € 1,70

*Regressionskoeffizient a = 15.000*

*Regressionskoeffizient b = - 5.000*

Test-Markt Preis (€) Absatzmenge (Stück)

xi yi

A 1,90 5000

B 2,10 4000

C 1,50 8000

D 2,50 3000

M 1,70 6.500

Prognose:

ŷ(x) = a + b \* x

y(1,7) = 15.000 + -5.000 \* 1,70

y(1,7) = 15.000 – 8.500

ŷ = 6.500 Stück

Bei einem Preis von 1,70 ist in 8 Wochen eine Absatzmenge von 6.500 Stück zu erwarten.

g)

Berechnen Sie für die Daten die Varianz der empirisch ermittelten Absatzmengen und die Varianz der dazugehörigen Regressionswerte (= entsprechende Absatzmengen auf der Regressionsfunktion).

Was ergibt der Quotient „Varianz der Regressionswerte zur Varianz der Absatzmengen“?

**Varianz der Absatzmengen**

**Varianz der Absatzmengen ist die Varianz zu y (sy²)**

sy² = 1/n \* ∑y² - y̅²

**WICHTIG:** ŷ̅ ist das arithmetische Mitte von y (y̅)

Absatzmengen für y

5.000

4.000

8.000

3.000

**∑ 20.000**

**y̅ 20.000 : 4 = 5.000**

**y̅²**  **5.000² = 25.000.000**

sy² = ¼ \* 114.000.000 – **5.000²**

sy² = 28.500.000 – **25.000.000**

**sy² = 3.500.000**

**Varianz der Regressionswerte sŷ²**

sŷ² = 1/n \* ∑ŷ² - ŷ̅² = 3.250.000

**ŷ = a + b \* x** | *Regressionskoeffizient a = 15.000*

| *Regressionskoeffizient b = - 5.000*

Regressionswerte ŷ

ŷ(1,9) = 15.000 + (- 5.000) \* 1,9 = **5.500**

ŷ(2,1) = 15.000 + (- 5.000) \* 2,1 = **4.500**

ŷ(1,5) = 15.000 + (- 5.000) \* 1,5 = **7.500**

ŷ(2,5) = 15.000 + (- 5.000) \* 2,5 = **2.500**

sŷ² = ¼ \*( **5.500² + 4.500² + 7.500² + 2.500²)** - **5.000²**

sŷ² = ¼ \*( **5.500² + 4.500² + 7.500² + 2.500²)** - **25.000.000**

sŷ² = ¼ \* **113.000.000** - **25.000.000**

sŷ² = **28.250.000** – **25.000.000**

**sŷ² = 3.250.000**

**Quotient „Varianz der Regressionswerte zur Varianz der Absatzmengen“**

**Varianz der Regressionswerte sŷ² = 3.250.000**

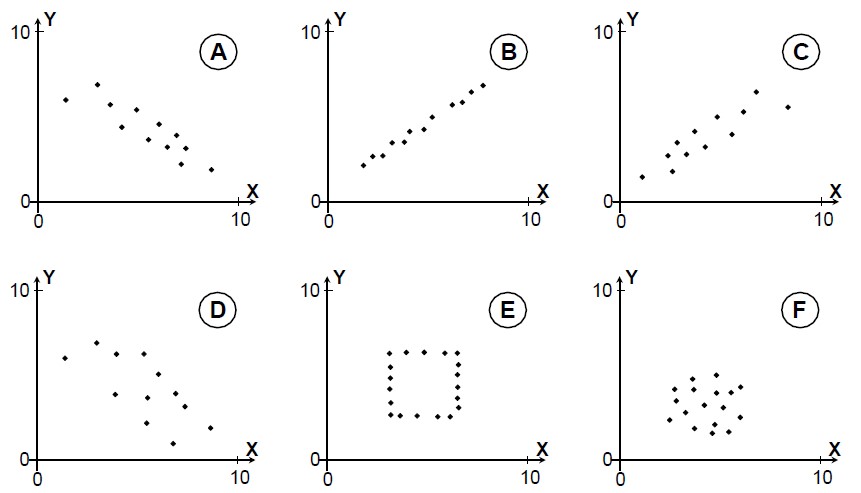
**Varianz der Absatzmengen sy² = 3.500.000**

**Bestimmtheitsmaß R² ist das Verhältnis aus Varianz der Regressionswerte zur Gesamtvarianz (= Varianz der abhängigen Variablen) bzw. das Quadrat der Korrelationskoeffizienten rx,y**

# Aufgabe 4

In den folgenden Abbildungen sind 6 Streudiagramme dargestellt, die den Zusammenhang zwischen den zwei Merkmalen X und Y visualisieren. Bringen Sie die Korrelationskoeffizienten rA, rB, rC, rD, rE, rF nach ihren Werten in eine aufsteigende Rangfolge.

(rA: Korrelationskoeffizient für das Streudiagramm A, usw.)



**Interpretation der Korrelationskoeffizienten r**

* Korrelationskoeffizienten r < 0 🡆 negativer Zusammenhang (negative Korrelation)
  + Punkte sind im Punktdiagramm fallend (von links oben nach rechts unten)
* Korrelationskoeffizienten r = -1 🡆 starker negativer Zusammenhang
  + Punkte liegen dicht und linear beieinander, fast wie in einer Geraden   
    (ähnlich Diagramm A)
* Korrelationskoeffizienten r > 0 🡆 positiver Zusammenhang (positive Korrelation)
  + Punkte sind im Punktdiagramm steigend (von links unten nach rechts oben)
* Korrelationskoeffizienten r = +1 🡆 starker positiver Zusammenhang
  + Punkte liegen dicht und linear beieinander, fast wie in einer Geraden   
    (ähnlich Diagramm B)

**Je dichter die Punkte zusammenliegen, desto stärker ist der (negative oder positive) Zusammenhang**

**Lösung:**

-1 < rA < rD < rF  ≈ 0 ≈ rE < rC < rB < +1

# Aufgabe 5

Es soll den Zusammenhang zwischen Preis P und Absatzmenge X empirisch untersucht werden. Dazu werden in 20 ausgewählten Testmärkten unterschiedliche Preise eingesetzt und Absatzmengen zu diesen Preisen dokumentiert. Die Regressionsrechnung liefert die folgende lineare Regressionsfunktion:

Ŷ = Ŷ(p) = a + b\*p = 2.000 – 500\*p (p in €, Y in Tsd. Stück)

Der Korrelationskoeffizient beträgt r = - 0,75.

Nehmen Sie kurz Stellung zu den beiden folgenden Aussagen a) und b):

1. Aus der Regressionsfunktion kann man erkennen, dass der Zusammenhang zwischen Preis und Absatzmenge sehr stark ist.
2. Die Absatzmengenunterschiede in den Testmärkten lassen sich zu 75% durch die unterschiedliche Preispolitik erklären, d.h. durch die Preisunterschiede in den Testmärkten.

# Aufgabe 6

Für ein neues Produkt wird untersucht, wie die Käufer auf unterschiedliche Preise reagieren. In 4 vergleichbar großen Testgeschäften in verschiedenen Gegenden wird das neue Produkt einen Monat lang zu jeweils unterschiedlichen Preisen angeboten. In der folgenden Tabelle sind für die Testphase die jeweils geforderten Preise (**p**) und die Reaktion der Käufer ausgedrückt in Absatzmengen des Produktes (**Y**) zusammengestellt.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Testgeschäft Nr.** | **1** | **2** | **3** | **4** |
| **x: Preis pro Stück (€)** | 3,00 | 1,50 | 1,00 | 2,50 |
| **Y: Absatzmenge (in Tausend Stück)** | 2 | 6 | 14 | 10 |

Nach der Methode der kleinsten Quadrate wurde mit Hilfe der Berechnungstabelle und den entsprechenden Formeln für die beiden Regressionskoeffizienten die folgende lineare PreisAbsatz-Funktion ermittelt:

Ŷ = Ŷ(x) = a + b\*x = 16 – 4\*x

a)

Erstellen Sie eine Berechnungstabelle. Berechnen Sie die Regressionswerte und Residualwerte und tragen Sie diese in die Tabelle ein. Zeichnen Sie in das Koordinatensystem ein:

* die Beobachtungswertepaare durch ∎
* die berechnete Regressionsfunktion,
* die Regressionswerte Ŷ**i** durch •,
* die Residualwerte (wenn farbiger Stift vorhanden, in Farbe)

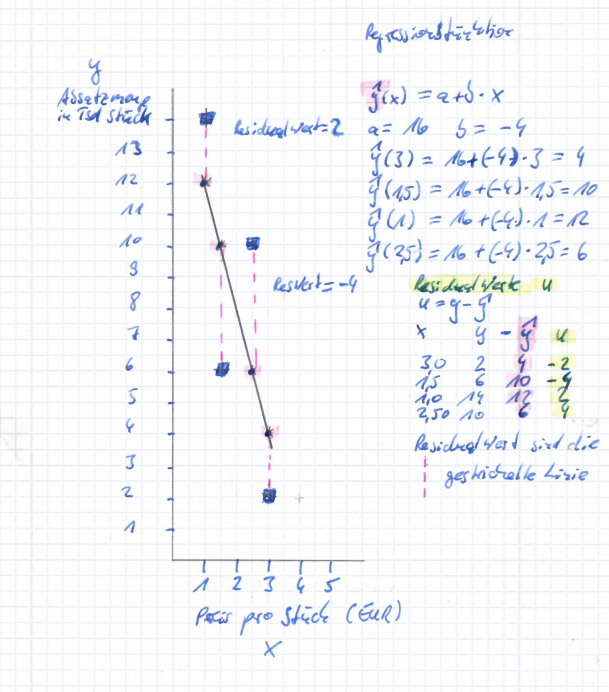
Regressionskoeffizient a = 16

Regressionskoeffizient b = -4

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Testgeschäft Nr | xi  Preis pro Stück (€) | yi  Absatzmenge  (in Tsd. Stück) | ŷ = a + b \* x  ŷ = 16 + (-4) \* x | Residualwerte  u = yi - ŷi | xi² | yi² | xi \* yi |
| 1 | 3,00 | 2 | 4 | -2 | 9,00 | 4,00 | 6 |
| 2 | 1,50 | 6 | 10 | -4 | 2,25 | 36,00 | 9 |
| 3 | 1,00 | 14 | 12 | 2 | 1,00 | 196,00 | 14 |
| 4 | 2,50 | 10 | 6 | 4 | 6,25 | 100,00 | 25 |
| **SUMMEN** | **8,00** | **32** | **32** |  | **18,50** | **336,00** | **54** |

**x̅ = 8 : 4 = 2 x̅² = 2² = 4**

**y̅ = 32 : 4 = 8 y̅² = 8² = 64**



b)

Berechnen Sie den Korrelationskoeffizienten und das Bestimmtheitsmaß für die

Daten in a) und interpretieren Sie die beiden Ergebnisse bezogen auf unser Beispiel

Korrelationskoeffizienten rx,y =

rx,y = 1/n \* ∑(xi \* yi) – (x̅ \* y̅)

rx,y = ¼ \* 54 – 16

rx,y = 13,5 - 16

rx,y = -2,5

rx,y = -2,5

rx,y = -2,5

rx,y = - 0,7072

Korrelationskoeffizient muss zwischen -1 und +1 liegen

- 1,0 bis - 0,8 = „starke“ Korrelation

- 0,8 bis - 0,5 = „mittlere“ Korrelation

- 0,5 bis - 0,0 = „schwache“ Korrelation

0 = „keine“ Korrelation

0 bis 0,5 = „schwache“ Korrelation

0,5 bis 0,8 = „mittlere“ Korrelation

0,8 bis 1,0 = „starke“ Korrelation

Bestimmungsmaß R² = 0,50

50% der Varianz der Absatzmengen (variable Absatzmenge) lassen sich durch die Varianz der Preise bzw. durch die Regressionsfunktion (ŷ = a + b \* x) erklären.

Die weiteren 50% lassen sich durch andere (nicht erklärbare) Einflussfaktoren erklären.

**(Bestimmungsmaß ist Anteil der abhängigen Variablen an der Gesamtvarianz der sich durch die unabhängige Variable erklären lässt.)**

**Verhältnis der Varianz der Regressionswerte (sŷ²) zur Gesamtvarianz (sy²)**

c)

Prognostizieren Sie auf der Basis der Regressionsfunktion die monatliche Absatzmenge für einen den 4 Testgeschäften vergleichbaren Einzelhandelsbetrieb, wenn dort ein Preis von € 2,00 pro Stück verlangt wird

ŷ(x) = a + b \* x | Regressionskoeffizient a = 16

| Regressionskoeffizient b = – 4

ŷ(2) = 16 - 4 \* 2

ŷ(2) = 16 – 8

ŷ(2) = 8 Tsd. Stück

Bei einem Preis von 2,00 € wird die voraussichtliche Absatzmenge ŷ = 8.000 Stück sein.

d)

Was halten Sie von der Güte der obigen Prognose (unter c))?

Die Prognosegüte von 50% ist für die Modellgüte nicht gut.

Nur 50% der Unterschiede (Varianz) in der Absatzmenge sind durch die Varianz in den Preisen zu erklären. Die übrigen 50% der Varianz in der Absatzmenge werden durch andere Einflussgrößen (Faktoren) erklärt.

e)

Berechnen Sie mit Hilfe der Berechnungstabelle aus a) die Varianz 𝑠𝑌2 der Absatzmengen Yi und die Varianz 𝑠𝑌̂2 der entsprechenden Regressionswerte Ŷi.

Zeigen Sie, dass hier die folgende Beziehung gilt:

**Bestimmtheitsmaß =**

**Varianz 𝑠𝑌2 der Absatzmengen Y**

𝑠𝑌2 = 1/n \* ∑yi² - y̅²

𝑠𝑌2 = ¼ \* 336 – 8²

𝑠𝑌2 = ¼ \* 336 – 64

𝑠𝑌2 =87 – 64

𝑠𝑌2 = 20

**Varianz Regressionswerte Ŷi**

𝑠𝑌̂2 = 1/n \*∑ŷ² - y̅²

𝑠𝑌̂2 = ¼ \* (4² + 10² + 12² + 6²) – 8²

𝑠𝑌̂2 = ¼ \* (16 + 100 + 144 + 36) – 64

𝑠𝑌̂2 = ¼ \* 296 – 64

𝑠𝑌̂2 = 74 – 64

𝑠𝑌̂2 = 10

f)

Berechnen Sie außerdem mit Hilfe der Berechnungstabelle aus a) die Varianz 𝑠𝑥2 der Preise xi.

Zeigen Sie, dass die Varianz der Regressionswerte bestimmt wird durch die Varianz des unabhängigen Merkmals, d.h., dass hier folgende Beziehung gilt:

𝑠ŷ² =𝑏²∗𝑠𝑥² (b 🡆 Regressionskoeffizient)

**𝑠x2 = 1/n \* ∑xi² - x̅²**

𝑠x2 = 1/n \* 18,5 - 2²

𝑠x2 = 4,625 - 4

𝑠x2 = 0,625

sŷ² = 10

b = -4 b² = 16

𝑏²∗𝑠𝑥² = (-4)² \* 0,625 = 16 \* 0,625 = 10 = sŷ²

10 = 16 \* 0,625

# Aufgabe 7

Für 4 Monate liegen die Daten über den Hypothekenzinssatz **x** sowie über den saisonbereinigten monatlichen Auftragseingang **Y** im Bauhauptgewerbe vor, der auf den privaten Wohnungsbau entfällt.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Testgeschäft Nr.** | **1** | **2** | **3** | **4** |
| **x: Hypothekenzinssatz in %** | 6 | 7 | 5,5 | 7,5 |
| **Y: Auftragseingang (in Mrd. €)** | 30 | 40 | 50 | 20 |

a)

Berechnen Sie nach der Methode der kleinsten Quadrate die lineare Regressionsfunktion, die die „mittlere“ Abhängigkeit des Auftragseingangs im Bauhauptgewerbe vom Hypothekenzinssatz möglichst gut beschreibt

ŷ(x) = a + b \* x | a und b sind Regressionskoeffizienten

Formel zur Berechnung des Regressionskoeffizienten a

Formel zur Berechnung des Regressionskoeffizienten a

**Hinweis:**

**Der Nenner muss nur einmal berechnet werden, da er für beide Regressionskoeffizienten identisch ist.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Test-  geschäft Nr | xi  Hypotheken-zinssatz in % | yi  Auftrags-eingang  (in Mrd. € | xi² | yi² | xi \* yi | ŷ = a + b \* x |
| 1 | 6,0 | 30 | 36,00 | 900 | 180 | 40 |
| 2 | 7,0 | 40 | 49,00 | 1600 | 280 | 30 |
| 3 | 5,5 | 50 | 30,25 | 2500 | 275 | 45 |
| 4 | 7,5 | 20 | 56,25 | 400 | 150 | 25 |
| **SUMME** | **26,0** | **140** | **171,50** | **5400** | **885** | **140** |

ŷ(x) = a + b \* x

ŷ(6) = 100 + (-10) \* 6 = 100 – 60 = 40 Mrd. EUR

ŷ(7) = 100 + (-10) \* 7 = 100 – 70 = 30 Mrd. EUR

ŷ(5,5) = 100 + (-10) \* 5,5 = 100 – 55 = 45 Mrd. EUR

ŷ(7,5) = 100 + (-10) \* 7,5 = 100 – 75 = 25 Mrd. EUR

b)

Interpretieren Sie für diesen Fall die beiden Regressionskoeffizienten

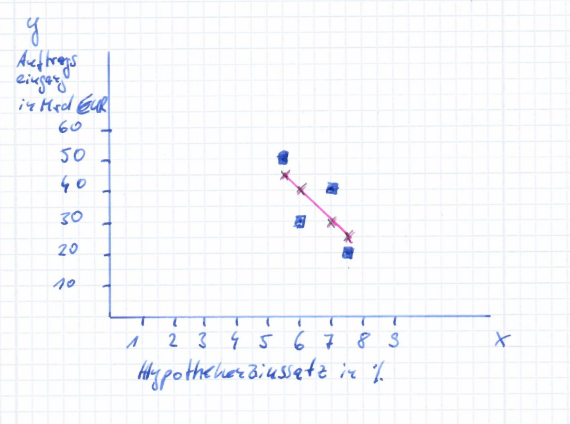
Mit abnehmenden Hypothekenzinssatz (unabhängige Variable x) nimmt der Auftragseingang zu (abhängige Variable y).

Bei einem Zinssatz von 10% hat der Auftragseingang ein Volumen von 0 Mrd. EUR (Untergrenze).

Bei einem Zinssatz von 0% hat der Auftragseingang ein Volumen 100 Mrd. EUR (Obergrenze).

c)

Zeichnen Sie in das Koordinatensystem die Wertepaare des Streudiagramms und die in a) berechnete Regressionsfunktion. Markieren Sie die Regressionswerte Ŷi



d)

Berechnen Sie den Korrelationskoeffizienten und das Bestimmtheitsmaß zur obigen Regressionsrechnung

Korrelationskoeffizienten

**x̅ = 26 : 4 = 6,5 x̅² = 42,25**

**y̅ = 140 : 40 = 35,0 y̅² = 1225**

rx,y = - 0,70

e)

Interpretieren Sie das Bestimmtheitsmaß für diesen Fall

**Bestimmtheitsmaß R² = rx,y² = - 0,70² = 0,50**

50% der Varianz in den Auftragseingängen (Unterschiede in den Auftragseingängen, abhängiges Merkmal) lassen sich durch die Varianz der Hypothekenzinssätze (unabhängiges Merkmal) erklären. Das bedeutet 50% der Streuung bzw. 50% der Unterschiede bei den Auftragseingängen erklären sich durch die unterschiedlich hohen Hypothekenzinssätze im Beobachtungszeitraum.

Die übrigen 50% der Unterschiede in den Auftragseingängen werden durch andere Einflussgrößen (Faktoren) erklärt.

50% ist eine schwache Korrelation.

f)

Prognostizieren Sie auf der Basis der Regressionsanalyse den (saisonbereinigten) monatlichen Auftragseingang im Bauhauptgewerbe, der zu erwarten ist bei einem

Hypothekenzinssatz von 6,6% und von 7,2%

ŷ(6,6) = 100 + (-10) \* 6,6 = 34 Mrd. EUR

ŷ(7,2) = 100 + (-10) \* 7,2 = 28 Mrd. EUR

g)

Welche Kennzahl kann man zur Beurteilung der Güte der obigen Prognose nutzen? Was halten Sie von der in f) erstellten Prognose?

Das Bestimmtheitsmaß