**Übungsblatt Regression**

**Aufgabe 1)**

1. Beschreibe in eigenen Worten das Vorgehen bei einer linearen Regressionsanalyse und erkläre das Prinzip der „Ordinary-Least-Squares“

**Lösung:**

In der linearen Regression wir die Merkmalsausprägung einer abhängigen Variable Y auf Basis einer oder mehrerer unabhängiger Variablen X vorhergesagt.

Y (Explanandum) soll durch X (Explanans) erklärt/vorhergesagt werden

Dabei wird angenommen, dass der Zusammenhang zwischen X und Y linear ist und die Richtung des Zusammenhangs in jedem beliebigen Werteintervall der Variable Y gleich ist

Gleichung für die vorhergesagten Werte lautet: ŷ = a + b\*x

* Für jeden X-Wert gibt es einen beobachteten und einen vorhergesagten Y-Wert
* Differenz zwischen beobachtetem und vorhergesagtem Wert ist das Residuum *e*

Ordinary-Least –Squares: Regressionsgrade wird so durch Punktewolke gelegt, dass Summe der quadrierten Residuen *e* möglichst klein ausfällt

1. Beschreiben Sie kurz die einzelnen Elemente der Gleichung.

y =β0+β1∗x

**Lösung:**

β0 (auch oft als α bezeichnet) ist der Wert, den Y annimmt, wenn alle X-Variablen den Wert 0 annehmen.

β1 ist das Regressionsgewicht, wenn sich x um eine Einheit erhöht, dann erhöht sich Y um β1 Einheiten.

X ist die Ausprägung der unabhängigen Variable

**Aufgabe 2)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | xi | yi | xi- x̄ | yi- ȳ |
| 1 | 3 | 4,25 | -1,08 | -0,81 |
| 2 | 2 | 3,5 | -2,08 | -1,56 |
| 3 | 5,5 | 6,125 | 1,42 | 1,065 |
| 4 | 7 | 7,25 | 2,92 | 2,19 |
| 5 | 1 | 2,75 | -3,08 | -2,31 |
| 6 | 6 | 6,5 | 1,92 | 1,44 |
| Mittelwerte (≈) | 4,08 | 5,06 |  |  |

Begründen Sie, ob die Regressionsgleichung mit der angegebenen Tabelle zusammenpasst. Falls nein, geben Sie bitte die richtige Gleichung an.

y=2,5+0,5∗x

**=** (-1,08)\*(-0,81) + (-2,08)\*(-1,56) + 1,42\*1,065 + 2,92\*2,19 + (-3,08)\*(-2,31) + 1,92\*1,44] / (n-1)

= [0,8748 + 3,2448 + 1,5123 + 6,3948 + 7,1148 + 2,7648] / 5

=21,9063 / 5 **= 4,38126**

**s2**x= [(-1,08)² + (-2,08)² + 1,42² + 2,92² + (-3,08)2 + 1,922] / (n-1)

= [1,1664 + 4,3264 + 2,0164 + 8,5264 + 9,4864 + 3,6864] / 5

= 29,2084 / 5 = **5,84168**

β1 = = = **0,75**

β0 = ȳ - β1 \* x̄ **=** 5,06 – 0,75\*(4,08) = 5,06 – 3,06 **= 2**

**Regressionsgleichung: ŷ = 2 + 0,75\*x**

**Aufgabe 3)**

Wir haben folgende Datenmatrix gegeben:

|  |  |  |
| --- | --- | --- |
| **Person** | **X-Wert** | **Y-Wert** |
| 1 | 0 | 8 |
| 2 | 2 | 3 |
| 3 | 3 | 1 |
| 4 | 5 | -2 |
| 5 | 4 | 1 |
| 6 | 1 | 1,5 |
| 7 | 7 | -4 |

Bestimme die Mittelwerte von X und Y, die Varianzen, die Kovarianz und die Gleichung der Regressionsgeraden (🡪 y =β0+β1∗x)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Person** | **xi** | **yi** | **xi- x̄** | **yi- ȳ** |
| 1 | 0 | 8 | -3,14 | 6,79 |
| 2 | 2 | 3 | -1,14 | 1,79 |
| 3 | 3 | 1 | -0,14 | -0,21 |
| 4 | 5 | -2 | 1,86 | -3,21 |
| 5 | 4 | 1 | 0,86 | -0,21 |
| 6 | 1 | 1,5 | -2,14 | 0,29 |
| 7 | 7 | -4 | 3,86 | -5,21 |
| Mittelwert (≈) | 3,14 | 1,21 |  |  |

**=** [(-3,14)\*6,79 + (-1,14)\*1,79 + (-0,14)\*(-0,21) + 1,86\*(-3,21) + 0,86\*(-0,21) + ( -2,14)\*0,29 + 3,86\*(-5,21)] / (n-1)

= [-21,3206 + (-2,0406) + 0,0294 + (-5,9706) + (-0,1806) + (-0,6206) + (-20,1106)] / 6

=-50,2142 / 6 **= -8,3690**

**s2**x = [(-3,14)² + (-1,14)² + (-0,14)² + 1,86² + 0,86² + (-2,14)² + 3,86] / (n-1)

= [9,8596 + 1,2996 + 0,0196 + 3,4596 + 0,7396 + 4,5796 + 14,8996] / 6

= 34,8572 / 6 = **5,8095**

β1 = = ≈ **-1,44**

β0 = ȳ - β1 \* x̄ **=** 1,21 – (-1,44)\*(3,14) = 1,21 – (-4,5216) ≈ **5,73**

**Regressionsgleichung: ŷ = 5,73 – 1,44\*x**

**Aufgabe 4)**

1. Berechne die Regressionsgleichung:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Schüler** | **IQ(xi)** | **Abipunkte(yi)** | **xi- x̄** | **yi- ȳ** |
| 1 | 105 | 514 | -1,43 | -111,57 |
| 2 | 104 | 574 | -2,43 | -51,57 |
| 3 | 112 | 774 | 5,57 | 148,43 |
| 4 | 97 | 431 | -9,43 | -194,57 |
| 5 | 115 | 626 | 8,57 | 0,43 |
| 6 | 102 | 811 | -4,42 | 185,43 |
| 7 | 110 | 649 | 3,57 | 23,43 |
| Mittelwert (≈) | 106,43 | 625,57 |  |  |

= [(-1,43)\*(-111,47) + (-2,43)\*(-51,57) + 5,57\*148,43 + (-9,43)\*(-194,57) + 8,57\*0,43 + (-4,42)\*185,43 + 3,57\*23,43] / (n-1)

= [159,4021 + 125,3151 + 826,7551 + 1.834,7951 + 3,6851 + (-819,6006) + 83,6451] / 6

= 2.213,997 / 6 = **368,9995**

**s2**x = [(-1,43)² + (-2,43)² + 5,57² + (-9,43)² + 8,57² + (-4,42)² + 3,57²] / n-1

= [2,0449 + 5,9049 + 31,0249 + 88,9249 + 73,4449 + 19,5364 + 12,7449] / 6

= 233,6258 / 6 **= 38,9376333**

β1 = = ≈ **9,48**

β0 = ȳ - β1 \* x̄ **=** 625,57 – 9,48\*106,43 = 625,57 – 1008,9564 ≈ **-383,39**

**Regressionsgleichung: ŷ = -383,39 +9,48\*x**

1. Wie viele Abiturpunkte würden für eine Person mit einem durchschnittlichen IQ (von 100) erwartet

= -383,39 + 9,48\*100

= -383,39 + 948

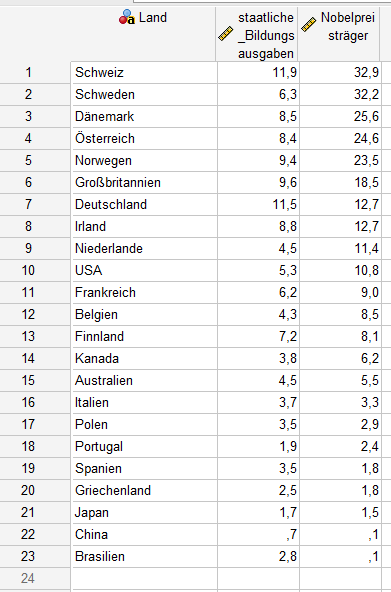
= 564,61

**Aufgabe 5)**

Du hast in einem Buch gelesen, dass es einen Zusammenhang zwischen den staatlichen Bildungsausgaben pro Kind im Jahr (gemessen in Tausendeuro) und der Anzahl der Nobelpreisträger\*innen gibt und hast das Ganze in SPSS überprüft.

**Informationen**

|  |
| --- |
| **Messung**  Es handelt sich um staatliche Bildungsausgaben pro Kind im Jahr in Tausendeuro.  Die Nobelpreisträger\*innen sind als Anzahl pro 10 Millionen Einwohner gemessen.  Achtung: Es handelt sich um ein fiktives Beispiel!  In der folgenden Tabelle kann man die Daten sehen, für den Fall, dass jemand das Beispiel gerne selber in SPSS umsetzen will. |



Folgendes Regressionstabelle kommt heraus

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Unstandardisierter Koeffizienten** | **Standardisierte**  **Koeffizienten** | **T** | **Sig.** |
| Staatliche Bildungsausgaben | 2,526 | 0,789 | 5,880 | ,000 |
| Konstante | -3,191 |  | -1,146 | ,264 |
| r² | 0,622 |  |  |  |

Abhängige Variable: Anzahl der Nobelpreisträger

1. Interpretiere die Koeffizienten und die Signifikanz

Die Konstante (-3,191) ist der Wert, wenn die unabhängige Variable den Wert 0 annimmt, d.h. dass für ein Land ohne staatliche Bildungsausgaben -3,191 Nobelpreisträger pro 10 Millionen Einwohner vorhergesagt werden. Da negative Nobelpreisträger nicht möglich sind ist die Konstante hier inhaltlich nicht zu interpretieren.

β1 beträgt 2,526, d.h. dass für je Tausendeuro an staatlichen Bildungsausgaben, die im Jahr in ein Kind investiert werden kommen 2,526 Nobelpreisträger pro 10 Millionen Einwohner hinzu. Der Effekt ist signifikant, wie man am t-Wert vom 5,880 (>1,96) und dem p-Wert (0.000) erkennen kann.

1. Stelle die Regressionsgleichung auf.

Y = -3,191 + 2,526\*x

1. Wie viele Nobelpreisträger werden jeweils für ein Land mit von 10, 20 und 50 Tausendeuro Bildungsausgaben pro Kind im Jahr vorhergesagt?

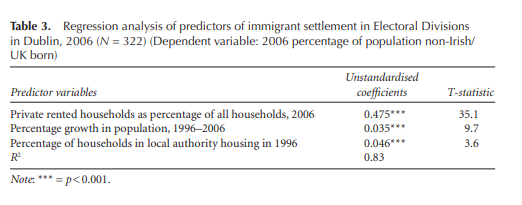
= -3,191 + 2,526\*10 = 22,069

= -3,191 + 2,526\*20 = 47,329

= -3,191 + 2,526\*50 = 123,109

**Aufgabe 6)**

Du hast in einem Aufsatz folgende Regressionstabelle zu den verschiedenen Erklärungen des Migrantenanteils an der lokalen Bevölkerung in verschiedenen Stadteilen Dublins gefunden.



Quelle: Fahey and Fanning 2010

1. Interpretiere die Koeffizienten
2. Welche der Koeffizienten sind signifikant? Woran kann man das erkennen?
3. Ist das Modell geeignet, um den Anteil der Personen mit Migrationshintergrund in den verschiedenen Stadtteil Dublins zu erklären?
4. Welche Angabe fehlt?

**Lösung:**

* Wenn der Anteil der Mietwohnungen um 1 Prozentpunkt steigt, dann steigt der Anteil der Anteil der Migranten an der lokalen Bevölkerung um 0,475 Prozentpunkte
* Für jeden Prozentpunkt an Bevölkerungswachstum zwischen 1996 und 2006 steigt der Anteil der von Migranten an der lokalen Bevölkerung um 0,035 Prozentpunkte
* Für jeden Prozentpunkt an Sozialwohnungen im Jahr 1996 steigt der Anteil der Migranten an der lokalen Bevölkerung um 0,046 Prozentpunkte
* Alle drei Koeffizienten sind mindestens auf dem 5%-Niveau statistisch signifikant, dass erkennt man daran, dass der jeweilige T-Wert über 1,96 liegt
* Der Determinationskoeffizient r² beträgt 0,83, d.h. unser Modell erklärt 83% der Varianz unserer unabhängigen Variable. Das Modell ist also ziemlich gut geeignet, um den Anteil der Personen mit Migrationshintergrund in den verschiedenen Stadtteil Dublins zu erklären.
* Es fehlt die Konstante, ohne diese können wir keine Regressionsgleichung aufstellen