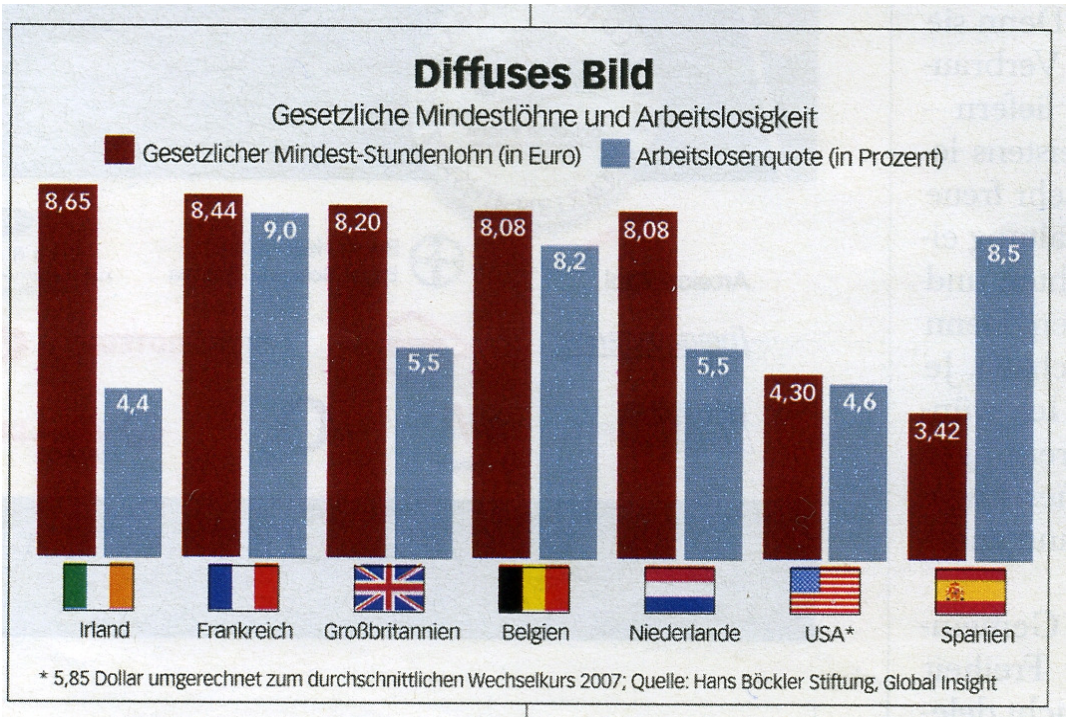
Zwei Rechenbeispiele für die einfache lineare Regression

1. Mindestlöhne – Beispiel[1]

1.1. Daten

Entnommen aus Rolf Ackermann, Spielball des Lobbyisten, Mindestlöhne schaden nicht nur bei Postdiensten sondern in allen Branchen, in: Wirtschaftswoche Nr. 50 (10.12.2007)



Es soll gelten x_i = Höhe des Mindestlohns (x ist später bei einer Erweiterung der Aufgabe x_1), y_i = Arbeitslosenquote. Wir haben hier Querschnitts-, nicht Zeitreihendaten, daher der Laufindex i = 1, 2,…,n statt t = 1, 2,…,T

1.2. Normalgleichungen

Die Zahlen sind leicht der auf der nächsten Seite wiedergegebenen Excel-Tabelle zu entnehmen

Allgemein	Mit den Daten	In Matrixschreibweise
$\alpha n + \beta\Sigma x = \Sigma y$ $\alpha\Sigma x + \beta\Sigma x^2 = \Sigma xy$	$7\cdot\alpha + 49{,}17\cdot\beta = 45{,}7$ $49{,}17\cdot\alpha + 374{,}06\cdot\beta = 318{,}67$	$\begin{bmatrix} n & \Sigma x \\ \Sigma x & \Sigma x^2 \end{bmatrix} \cdot \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \Sigma y \\ \Sigma xy \end{bmatrix}$

Exkurs

(kein Muss für Hörer, die diese Darstellungsart nicht mögen)

Die relevanten Matrizen und Vektoren in diesem Beispiel sind die Momentenmatrix

$$\mathbf{X'X} = \begin{bmatrix} n & \Sigma x \\ \Sigma x & \Sigma x^2 \end{bmatrix}$$ und die Datenmatrix $$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$ sowie der Datenvektor $$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

[1] Anders als die zweite Aufgabe (Affenaufgabe) wird diese Aufgabe wieder aufgegriffen bei der multiplen Regression.

Das Modell lautet somit in Matrixschreibweise **y** = **Xβ** + **u** und die Normalgleichungen als Ergebnis der Methode der kleinsten Quadrate sind $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$ mit $\hat{\boldsymbol{\beta}}' = \begin{bmatrix}\hat{\alpha} & \hat{\beta}\end{bmatrix}$ und **u** analog zu **y**.

1.3. Excel-Tabelle, Berechnung der Regressionskoeffizienten und ihrer Varianzen

A	B	C	D	E	F
	x	**y**	**xy**	**x^2**	**y^2**
Irland	8,65	4,4	38,06	74,8225	19,36
Frankreich	8,44	9	75,96	71,2336	81
Großbritannien	8,2	5,5	45,1	67,24	30,25
Belgien	8,08	8,2	66,256	65,2864	67,24
Niederlande	8,08	5,5	44,44	65,2864	30,25
USA	4,3	4,6	19,78	18,49	21,16
Spanien	3,42	8,5	29,07	11,6964	72,25
Berechnungen Spaltensumme S	49,17	45,7	318,666	374,0553	321,5100
Mittelwert S/n	7,0243	6,5286	45,524	53,4365	45,9300

Varianz von x	4,7785*	⇐ Excel	4,0959**	Steigung	-0,08175
Varianz von y	3,8590*	⇐ Excel	3,3078**	Ordinate	7,102804
Kovarianz	-0,3348	⇐ Excel		Korrelation	-0,09097

Die Regressionsfunktion lautet 7,1028 – 0,08175*x				
Die Variablen sind praktisch nicht miteinander korreliert				
Die Bestimmtheit ist nur 0,00828				

** Diese Werte errechnet Excel als Varianzen (durch n-1 statt durch n geteilt)*

*** Excel-Werte mit (n-1)/n = 6/7 multipliziert*

Die eingegebenen Daten sind in helltürkis markiert. Man kann bestimmte Werte durch Eingeben einer Berechnungsformel berechnen, etwa die Spaltensummen oder die Mittelwerte, um hiermit weiter zu rechnen, etwa um $s_x^2 = \frac{1}{n}\sum_i x_i^2 - \bar{x}^2$ = 53,4365-(7,0243)^2 = 4,0957 (Rundungsfehler,[2] vgl. oben 4,0959) zu bestimmen, oder man lässt dies mit der Excel Funktion (mit f_x wählen!) Mittelwert bzw. Varianz (oder auch Kovarianz) "automatisch" berechnen. Bei den Varianzen wird von Excel jedoch durch n - 1 = 6 geteilt (gelb markierte Felder). Um auf die bekannten Formeln $s_x^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$ und s_y^2 umzurechen ist mit 6/7 zu multiplizieren. Hieraus lassen sich Größen $\hat{\alpha}$ (Ordinatenabschnitt), $\hat{\beta}$ (Steigung), r und r^2 bestimmen. Ferner gilt (Cramerschen Regel)

$$\hat{\alpha} = \frac{\begin{vmatrix}\Sigma y & \Sigma x\\ \Sigma xy & \Sigma x^2\end{vmatrix}}{\begin{vmatrix} n & \Sigma x\\ \Sigma x & \Sigma x^2\end{vmatrix}} = \frac{\begin{vmatrix} 45,7 & 49,17\\ 318,666 & 374,0553\end{vmatrix}}{\begin{vmatrix} 7 & 49,17\\ 49,17 & 374,0553\end{vmatrix}} = \frac{1425,52}{200,6982} = 7,1028$$

und entsprechend für die Steigung

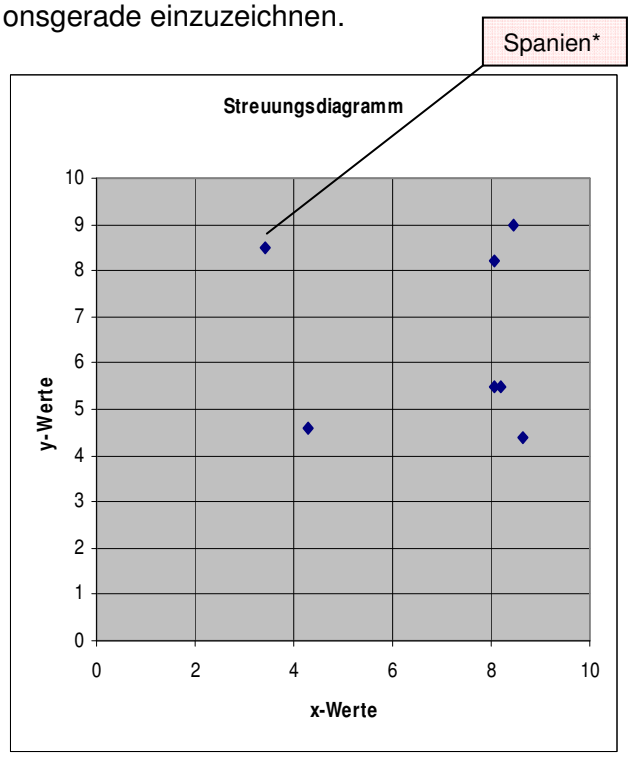
[2] Die Berechnung ist offenbar sehr fehleranfällig Rundungsfehler!) zumal n sehr klein ist und sie erfolgt am besten ausgehend von den Normalgleichungen mit der Cramerschen Regel. In dieser Hinsicht ist das zweite Rechenbeispiel (Affenaufgabe) sehr viel angenehmer, weil hier zwar mit noch weniger, dafür aber "glatteren" Zahlen gearbeitet wird.

$$\hat{\beta} = \frac{\begin{vmatrix} n & \Sigma y \\ \Sigma x & \Sigma xx \end{vmatrix}}{\begin{vmatrix} n & \Sigma x \\ \Sigma x & \Sigma x^2 \end{vmatrix}} = \frac{\begin{vmatrix} 7 & 45{,}7 \\ 49{,}17 & 318{,}666 \end{vmatrix}}{\begin{vmatrix} 7 & 49{,}17 \\ 49{,}17 & 374{,}0553 \end{vmatrix}} = \frac{-16407}{200{,}6982} = -0{,}08175\,.$$

Die Parameter $\hat{\alpha}$, $\hat{\beta}$, r und r^2 kann man auch direkt mit den Excelfunktionen f_x bestimmen. Man erhält dann:

Parameter	Excel Funktion f_x	Ergebnis
Ordinatenabschnitt $\hat{\alpha}$	"Achsenabschnitt"	7,102804
Steigung $\hat{\beta}$	"Steigung"	-0,08174961
Korrelation r	"Pearson" oder "Korrel"	-0,09096885
Bestimmtheit r^2	"Bestimmtheitsmaß"	0,008275332

Das Streuungsdiagramm erhält man als Graphik vom Typ "Punkt (XY)"[3]. Anders als beim nächsten Beispiel (Affenaufgabe) ist hier darauf verzichtet worden, die von Excel bestimmte Regressionsgerade einzuzeichnen.



* Spanien scheint ein Ausreißer zu sein. Rechnet man ohne Spanien, so ist der Korrelationskoeffizient 0,37482 statt – 0,09097 (allerdings ist n dann auch nur noch 6).

Man kann nun auch die Größen bestimmen, die wichtig sind für das Schätzen und Testen von Regressionskoeffizienten. Man erhält (in der Symbolik des Buches von v. Auer) die folgenden Werte:

$S_{yy} = 45{,}93-(6{,}5286)^2 = 3{,}3077551$

$S_{\hat{y}\hat{y}} = r^2 S_{yy} = 0{,}027372771$

$S_{\hat{u}\hat{u}} = S_{yy} - S_{\hat{y}\hat{y}} = 3{,}28038233$

Die geschätzte Varianz der Störgröße ist danach $\hat{\sigma}^2 = \frac{S_{\hat{u}\bar{u}}}{n-2} = \frac{3{,}2804}{5} = 0{,}656076$

Ferner ist $S_{xx} = 53{,}4365-(7{,}0243)^2 = 4{,}09571$ und mit $\hat{\sigma}^2_{\hat{\beta}} = \frac{\hat{\sigma}^2}{S_{xx}} = \frac{0{,}656}{4{,}096} = 0{,}16018$ erhält man die geschätzte Varianz von $\hat{\beta}$ und somit für die Standardabweichung von $\hat{\beta}$ den Wert 0,400223.

Für die geschätzte Varianz und die (geschätzte) Standardabweichung (standard deviation S.D. oder "Std. Error") von $\hat{\alpha}$ ergibt sich daraus $\hat{\sigma}^2_{\hat{\alpha}} = \overline{x^2}\hat{\sigma}^2_{\hat{\beta}}$ mit dem quadrierten quadratischen Mittel $\overline{x^2} = \sum x_1^2 / n$ = 53,4365, so dass die Varianz $\hat{\sigma}^2_{\hat{\alpha}}$ den Wert 53,4365·0,16018 = 8,5595 annimmt ($\hat{\sigma}_{\hat{\alpha}}$ = 2,9256). Die t-Werte sind demnach t = 7,1028/2,9256 = 2,4283 bei der Hypothesen H_0: α = 0 und t = - 0,08175/0,400223 = - 0,20426 bei der H_0: β = 0. Somit ist zwar α, nicht aber β signifikant von 0 verschieden. Berechnungen dieser Art (das Schätzen und Testen betreffend) und vor allem eine multiple Regression lassen sich besser mit **EViews**, statt mit Excel durchführen. Die Erweiterung des Beispiels Mindestlöhne zu einer Aufgabe der multiplen Regression mit den ent-

[3] Wenn man den Bereich markiert, auf den sich die Grafik beziehen soll, dann sollte man auch die Felder x und y mit markieren.

sprechenden Berechnungen findet sich in einem weiteren Download. Das dort mit EViews ermittelte Ergebnis (y in Abhängigkeit von $x = x_1$) sei hier jedoch bereits (verkürzt) wiedergegeben (Ergebnisse, die mit den oben [z.T. mit Excel] berechneten Ergebnissen verglichen werden können sind gelb unterlegt):

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	7.102804	2.925650	2.427769	0.0595
X1	-0.081750	0.400224	-0.204259	0.8462
R-squared	0.008275	Mean dependent var		6.528571
Adjusted R-squared	-0.190070	S.D. dependent var		1.964446*
S.E. of regression	2.143020	Akaike info criterion		4.597266
Sum squared resid	22.96268	Schwarz criterion		4.581811

* das ist die Wurzel aus 3,8590 in der Excel-Tabelle auf Seite 2 oben.

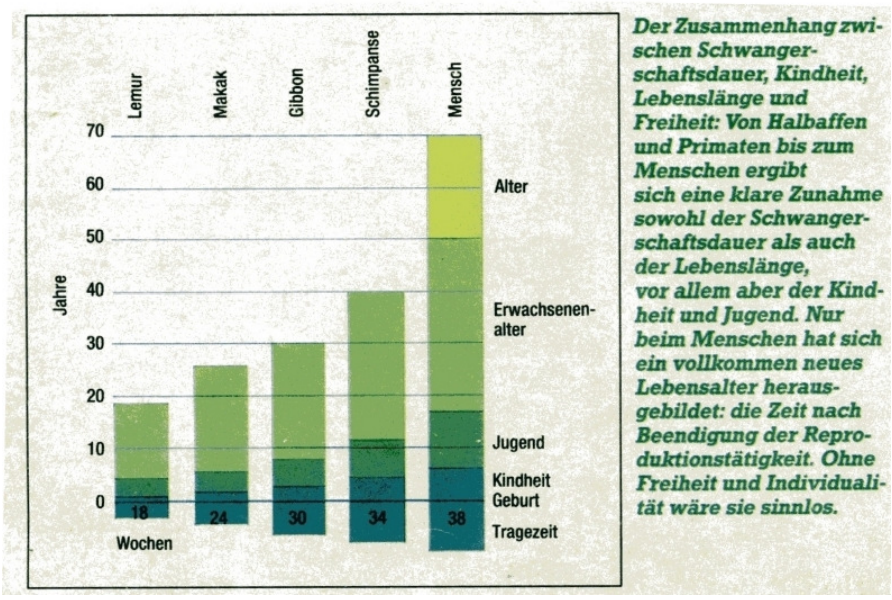
2. "Affenaufgabe"[4]

2.1. Daten

Die folgenden Daten über den Zusammenhang zwischen Dauer der Schwangerschaft und Lebenserwartung (der Mensch als "Ausreißer") sind entnommen aus der Zeitschrift Focus

	X = Dauer der Schwangerschaft	Y = Lebens-erwartung
Lemur	18	18
Makak	24	26
Gibbon	30	30
Schimpanse	34	40
Mensch	38	70
Summe	144	184

Die hier wiedergegebene Abbildung aus FOCUS zeigt, wie schwierig es ist, die Daten verständlich graphisch darzustellen, wenn man glaubt, bei den statistisch nicht vorgebildeten Lesern nicht Gebrauch machen zu können von der Möglichkeit eines Streuungsdiagramms. Man muss dann wohl mit den verschiedensten Farben operieren und es ist sehr fraglich, ob die Dinge so klarer und leichter verständlich werden als mit einem Streuungsdiagramm.



Der Zusammenhang zwischen Schwangerschaftsdauer, Kindheit, Lebenslänge und Freiheit: Von Halbaffen und Primaten bis zum Menschen ergibt sich eine klare Zunahme sowohl der Schwangerschaftsdauer als auch der Lebenslänge, vor allem aber der Kindheit und Jugend. Nur beim Menschen hat sich ein vollkommen neues Lebensalter herausgebildet: die Zeit nach Beendigung der Reproduktionstätigkeit. Ohne Freiheit und Individualität wäre sie sinnlos.

[4] Vorteil dieses Beispiels:
sehr wenige und zudem glatte Zahlen als Daten, so dass es leicht möglich ist, alles mit dem Taschenrechner nachzurechnen.

2.2. Berechnungen zur Deskriptiven Statistik, Excel Tabelle/Grafik und Normalgleichungen

Man sieht hier den Bildschirm und die Eingabe der Daten, wobei in den Spalten D, E und F einige Einfacher Berechnungen durchgeführt werden, die zur Bestimmung der Normalgleichungen notwendig sind:

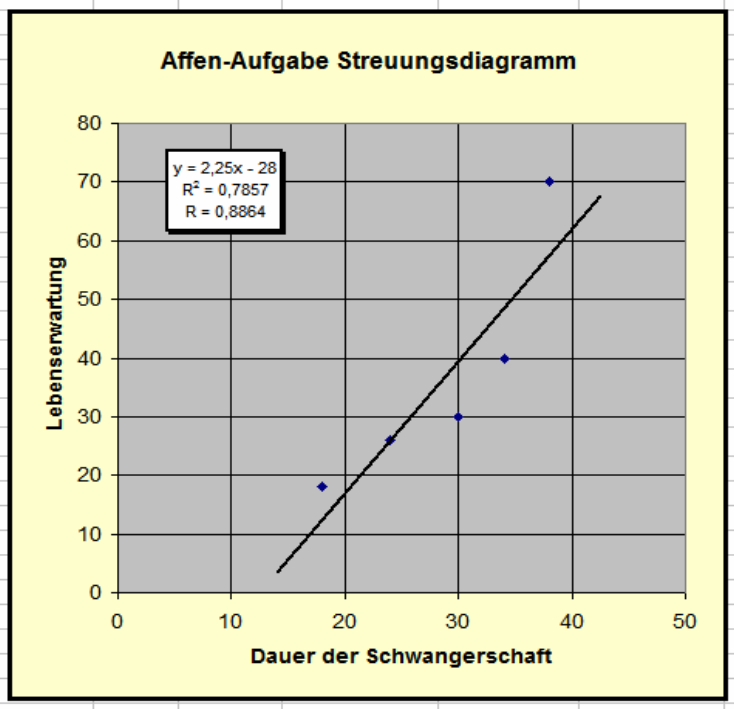
Microsoft Excel

Datei Bearbeiten Ansicht Einfügen Format Extras Daten Fenster

Arial 10

Affenaufgabe

	A	B	C	D	E	F
1						
2		X	Y	XY	X*X	Y*Y
3	Lemur	18	18	324	324	324
4	Makak	24	26	624	576	676
5	Gibbon	30	30	900	900	900
6	Schimpa	34	40	1360	1156	1600
7	Mensch	38	70	2660	1444	4900
8	Summe	144	184	5868	4400	8400



Man kann mit diesen Angaben leicht die Normalgleichungen zusammenstellen und erhält so

$5\cdot\alpha + 144\cdot\beta = 184$
$144\cdot\alpha + 4400\cdot\beta = 5868$

Die Schätzwerte $\hat{\alpha}$ und $\hat{\beta}$ erhält man nach der Cramerschen Regel mit drei Determinanten wie folgt

$$\hat{\alpha} = \frac{\begin{vmatrix} 184 & 144 \\ 5868 & 4400 \end{vmatrix}}{\begin{vmatrix} 5 & 144 \\ 144 & 4400 \end{vmatrix}} = -28$$

$$\hat{\beta} = \frac{\begin{vmatrix} 5 & 184 \\ 144 & 5868 \end{vmatrix}}{\begin{vmatrix} 5 & 144 \\ 144 & 4400 \end{vmatrix}} = 2{,}25$$

Die Regressionsgerade lautet mithin

- 28 + 2,25·x.

Für den Korrelationskoeffizienten erhält man mit Excel den Wert + **0,7857**, so dass die Bestimmtheit r^2 = **0,8864**, also 88,64%.

Die Regressionsgerade erhält man als "Trend" wenn man in der Graphik einen y Punkt anklickt und die rechte Maustaste drückt. Dann kommt ein Menü, mit dem man die Trendlinie (mit denen Typen linear, logarithmisch, gleitende Mittelwerte etc.) wählen kann. Wenn keine Verlängerung "vorwärts" oder "rückwärts" gewählt wird, zeichnet Excel nur die Gerade im Bereich zwischen x_{min} und x_{max}. Den eingezeichneten Trend anklicken → jetzt kann man die Trendlinie formatieren (dicker, farbig etc). Man kann auch – wie hier geschehen – die Option " Parameter anzeigen " wählen und das dazu gehörige Textfeld mit der Regressionsgleichung, der Bestimmtheit R^2 (entspricht r^2) bearbeiten und auch nachträglich den Korrelationskoeffizienten (Funktion "Pearson") eintragen.

Ohne Excel kann man die Parameter $\hat{\alpha}$ und $\hat{\beta}$ sowie r und $\hat{\sigma}^2$ auch wie folgt berechnen:

Mittelwerte	$\overline{x} = 144/5 = 28{,}8$	$\overline{y} = 184/5 = 36{,}8$
Varianzen (und Summen der Abweichungsquadrate)	$s_x^2 = 4400/5 - (28{,}8)^2 = 50{,}56$ $S_{xx} = Ts_x^2 = 5 \cdot 50{,}56 = 252{,}8$	$s_y^2 = 8400/5 - (367{,}8)^2 = 325{,}76$ $S_{yy} = 1628{,}8$
Kovarianz	$s_{xy} = S_{xy}/T - \overline{x} \cdot \overline{y}$?5868/5 − 28,8 · 36,8 = 113,76	
Steigung	$\hat{\beta} = s_{xy}/s_x^2$ = 113,76/50,56 = 2,25	
Ordinatenabschnitt	$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$ = 36,8 – 2,25·28,8 = -28	
Sum squared resid.	$S_{\hat{u}\hat{u}}$ = 349 (verschiedene Berechnungsmöglichkeiten siehe unten)	
S.E. of regression $\hat{\sigma}$	$\hat{\sigma}^2 = \hat{\sigma}_u^2 = S_{\hat{u}\hat{u}}/(T-2)$ = 349/3 = 116,33 → $\hat{\sigma} = \sqrt{116{,}33}$ =10,786	
Korrelation	$r = s_{xy}/\sqrt{s_x^2 s_y^2}$ = 0,8864 → Bestimmtheit r^2 = 0,78573	
Zweites Anfangsmoment	$\overline{x^2} = \sum x_t^2/T$ = 4400/5 = 880	

Es gilt $S_{xx} = T \cdot s_x^2 = \frac{1}{T} \cdot \begin{vmatrix} T & \sum x_t \\ \sum x_t & \sum x_t^2 \end{vmatrix}$, ferner da $\hat{y}_t - \overline{y} = \hat{\beta}(x_t - \overline{x})$ ist auch $S_{\hat{y}\hat{y}} = \hat{\beta}^2 S_{xx} = \hat{\beta}^2 T s_x^2$ und $S_{\hat{u}\hat{u}} = S_{yy} - S_{\hat{y}\hat{y}}$ =1628,8 - 1279,8 = 349 (oder $S_{\hat{u}\hat{u}} = (1-r^2) \cdot T \cdot s_y^2$).

2.3. Berechnungen zur Induktiven Statistik

a) Konfidenzintervall für die Streuung

Mit dem $\alpha/2$ sowie dem 1-$\alpha/2$ Quantil (d.h. mit der unteren und obere Signifikanzschranke) der χ^2 Verteilung bei α = 0,05 und 5-2 = 3 Freiheitsgraden z_u = 0,216 und z_o = 9,348 erhält man die folgenden Grenzen des 95% Konfidenzintervalls für die unbekannte Varianz σ^2 in der Grundgesamtheit: Untergrenze $\frac{S_{\hat{u}\hat{u}}}{z_o} = \frac{349}{9{,}348} = 37{,}33$ und Obergrenze $\frac{S_{\hat{u}\hat{u}}}{z_u} = \frac{349}{0{,}216} = 1615{,}74$.

Es wird gerne vergessen, dass[5] $\hat{\sigma}_u^2 = \hat{\sigma}^2$ = 349/3 = 116,333 genauso ein zu schätzender Parameter des Modells der einfachen Regression ist wie die Parameter $\hat{\alpha}$ und $\hat{\beta}$.

b) Regressionskoeffizienten, H_0 ρ = 0 (Varianzanalyse)

Die weiteren Größen sind für die Bestimmung von Konfidenzintervallen und zur Durchführung von Tests wichtig

$\hat{\sigma}_{\hat{\beta}}^2 = \frac{\hat{\sigma}^2}{S_{xx}} = \frac{116{,}33}{252{,}8} = 0{,}4600475$	$\hat{\sigma}_{\hat{\alpha}}^2 = \overline{x^2}\hat{\sigma}_{\hat{\beta}}^2 = 880 \cdot 0{,}46 = 404{,}842$	$\hat{\sigma}_{\alpha\beta} = -\overline{x} \cdot \hat{\sigma}_{\hat{\beta}}^2 = -13{,}249$

Die Matrix der (geschätzten) Varianzen und Kovarianzen der Regressionskoeffizienten ist demnach

$$\mathbf{V} = \hat{\sigma}^2 \mathbf{X}'\mathbf{X} = \begin{bmatrix} 404{,}842 & -13{,}249 \\ -13{,}249 & 0{,}46005 \end{bmatrix}.$$

Alle weiteren Berechnungen (Konfidenzintervalle für α und β Prognoseintervall für y_0 und sowie Varianzanalyse) vgl. Vorlesung.

[5] Man beachte, dass 116,333 nicht in der Mitte liegt zwischen 37,33 und 1615,74. Das liegt daran, dass die χ^2 Verteilung nicht symmetrisch ist.