Statistik II für Betriebswirte
Vorlesung 9

Dr. Andreas Wünsche

TU Bergakademie Freiberg
Institut für Stochastik

9. Dezember 2019

Einfache lineare Regression durch den Koordinatenursprung

- Bei bestimmten Problemstellungen ist es sinnvoll zu fordern, dass die Regressionsgerade durch den Koordinatenursprung geht. Man spricht dann auch von einer Regression ohne Absolutglied oder einer eigentlich-linearen Regression.
- Man erhält nun als Modellansatz

$$Y_i = b_1 x_i + \varepsilon_i\,, \qquad i = 1, \ldots, n\,;$$

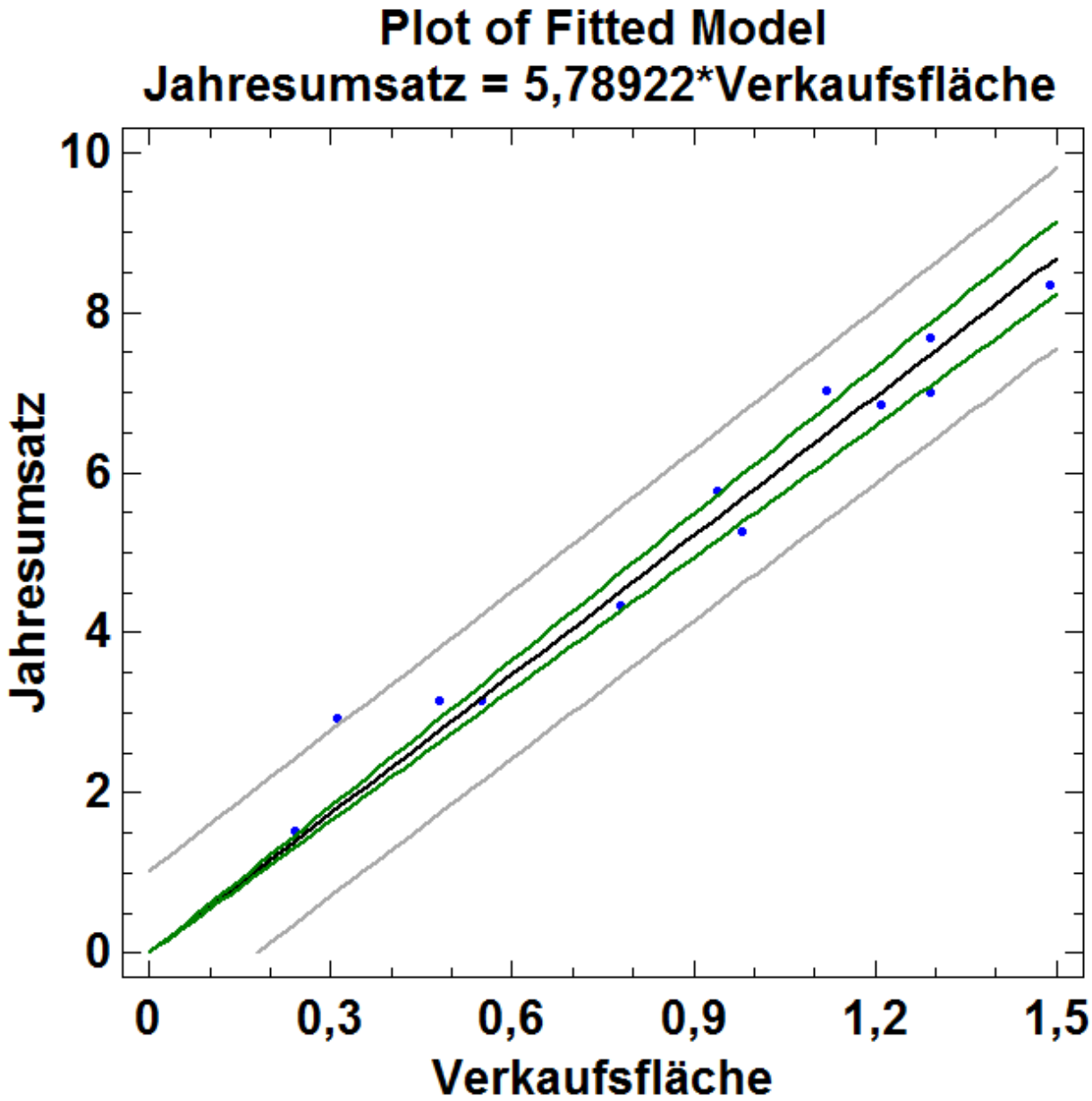
als Schätzung für den Parameter b_1

$$\hat{b}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

und als Schätzung für die Varianz der zufälligen Fehler

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad \text{mit} \quad \hat{y}_i = \hat{b}_1 x_i\,.$$

Regressionsgerade ohne Absolutglied im Beispiel 7.1



Transformationen auf Linearität

- Ist die gesuchte Abhängigkeitsfunktion eine nichtlineare Funktion (eine Gerade ist schlecht an die Daten anpassbar), kann man mitunter durch geeignete Variablentransformationen die Aufgabenstellung in eine der einfachen linearen Regression transformieren. Diese ist dann aber nicht äquivalent zur ursprünglichen Aufgabenstellung.
- Nichtlineare, in lineare transformierbare Funktionen sind z.B.

$$
\begin{aligned}
y &= \alpha x^{\beta} && \Rightarrow && \ln y = \ln \alpha + \beta \ln x \\
y &= \alpha \mathrm{e}^{\beta x} && \Rightarrow && \ln y = \ln \alpha + \beta x \\
y &= (\alpha + \beta x)^{-1} && \Rightarrow && y^{-1} = \alpha + \beta x \\
y &= x(\alpha + \beta x)^{-1} && \Rightarrow && y^{-1} = \alpha x^{-1} + \beta \\
y &= \alpha \mathrm{e}^{\beta / x} && \Rightarrow && \ln y = \ln \alpha + \beta x^{-1} \\
y &= \left(\alpha + \beta \mathrm{e}^{-x}\right)^{-1} && \Rightarrow && y^{-1} = \alpha + \beta \mathrm{e}^{-x}
\end{aligned}
$$

7.2. Multiple parameterlineare Regression

- Im Folgenden soll die Abhängigkeit eines Regressanden (einer Wirkungsgröße oder einer endogenenen Variablen) Y von mehreren Regressoren (Einflussgrößen oder exogenen Variablen) $X_1, \ldots, X_k$ beschrieben werden, d.h. es soll gelten

$$Y \approx f(X_1, \ldots, X_k)$$

  mit einer geeigneten Funktion $f : \mathbb{R}^k \to \mathbb{R}$.
- Wir werden größtenteils wieder annehmen, dass die Regressoren deterministisch sind (z.B. mit exakt einstellbaren Werten), und dies durch kleine Buchstaben $x_1, \ldots, x_k$ in den Gleichungen kennzeichnen.
- Man erhält dann als Modellgleichung

$$Y(x_1, \ldots, x_k) = f(x_1, \ldots, x_k) + \varepsilon$$

  mit einem zufälligen Fehler ε.

Parameterlineare Ansätze

- Häufig werden bei solchen Aufgabenstellungen parameterlineare Ansätze verwendet, d.h. man setzt eine Beziehung

$$Y(x_1, \ldots, x_k) = a_1 f_1(x_1, \ldots, x_k) + \ldots + a_r f_r(x_1, \ldots, x_k) + \varepsilon$$

mit speziell gewählten, bekannten Funktionen $f_1, \ldots, f_r$ und zu bestimmenden Koeffizienten (Parametern) $a_1, \ldots, a_r$ (die linear in die Gleichung eingehen) voraus.

- Die einfache lineare Regression mit der Modellgleichung

$$Y(x) = b_0 + b_1 x + \varepsilon$$

ist ein Spezialfall davon, dort gelten $k = 1$, $r = 2$, $f_1(x) = 1$, $f_2(x) = x$, $a_1 = b_0$, $a_2 = b_1$.

Beispiele für parameterlineare Ansätze

- Im eigentlich nichtmultiplen Fall $k = 1$ (nur eine Einflussgröße) gilt bei der quadratischen Regression

$$Y(x) = b_0 + b_1 x + b_2 x^2 + \varepsilon$$

und allgemeiner bei der polynomialen Regression vom Grade m

$$Y(x) = b_0 + b_1 x + \ldots + b_m x^m + \varepsilon\,.$$

- Der $k-$faktorielle Ansatz ohne Wechselwirkungen

$$Y(x_1, \ldots, x_k) = b_0 + b_1 x_1 + \ldots + b_k x_k + \varepsilon$$

wird zur Bestimmung der Ausgleichsebene (für die ebene Regression) genutzt.

- **Bemerkung:** Eine Gleichung der Form $y = b_0 + b_1 x_1 + \ldots + b_k x_k$ definiert eine (Hyper-)Ebene im $(k+1)-$dimensionalen Raum von Punkten mit Koordinaten $(x_1, \ldots, x_k, y)$.

Weitere Beispiele für parameterlineare Ansätze

- Als Beispiel eines $k-$faktoriellen Ansatzes mit Wechselwirkungen werde hier noch der Fall einer multiplen quadratischen Regression vorgestellt:

$$\begin{aligned} Y(x_1, \ldots, x_k) = b_0 + b_1 x_1 + \ldots + b_k x_k \\ + b_{12}\, x_1 x_2 + \ldots + b_{k-1,k}\, x_{k-1} x_k \\ + b_{11}\, x_1^2 + \ldots + b_{kk}\, x_k^2 \\ + \varepsilon. \end{aligned}$$

- Auch höhere Polynomgrade oder andere Funktionen der Variablen $x_1, \ldots, x_k$ sind möglich und werden auch verwendet.

Regressionsansatz in Vektorschreibweise

- Es ist vorteilhaft, die Vektorschreibweise zu nutzen. Es seien

$$\underline{x} = (x_1, \ldots, x_k)^T = \begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix}, \quad \underline{a} = (a_1, \ldots, a_r)^T = \begin{pmatrix} a_1 \\ \vdots \\ a_r \end{pmatrix},$$

$$\underline{f}(\underline{x}) = (f_1(\underline{x}), \ldots, f_r(\underline{x}))^T = \begin{pmatrix} f_1(\underline{x}) \\ \vdots \\ f_r(\underline{x}) \end{pmatrix}.$$

- Der parameterlineare Ansatz kann dann geschrieben werden als

$$Y(\underline{x}) = \underline{a}^T \underline{f}(\underline{x}) + \varepsilon\,, \tag{1}$$

wobei die Definition der Multiplikation von Vektoren (als Spezialfall der Matrixmultiplikation oder als Skalarprodukt) genutzt wird.

Die Methode der kleinsten Quadrate

- Sind die (zufallsbeeinflussten) „Wirkungen“ y_i für $i = 1, \ldots, n$ an den „Einflussstellen“ $\underline{x}_i = (x_{1i}, \ldots, x_{ki})^T$ durch Messungen bestimmt worden, kann man mit Hilfe der Methode der kleinsten Quadrate eine geeignete Schätzung $\hat{\underline{a}}$ des Vektors $\underline{a}$ der Regressionskoeffizienten im parameterlinearen Ansatz (1) finden.
- Die Schätzung $\hat{\underline{a}}$ ist ein Vektor von Regressionskoeffizienten $\underline{a}$, für den $\sum_{i=1}^{n} \left(y_i - \underline{a}^T \underline{f}(\underline{x}_i) \right)^2$ minimal wird.
- Die geschätzte Regressionsfunktion ist dann

$$\hat{y}(\underline{x}) = \hat{a}_1 f_1(\underline{x}) + \ldots + \hat{a}_r f_r(\underline{x}) = \hat{\underline{a}}^T \underline{f}(\underline{x}) = \underline{f}(\underline{x})^T \hat{\underline{a}} \,.$$

- Im Weiteren genutzte Bezeichnungen sind $\underline{y} = (y_1, \ldots, y_n)^T$ und

$$F = (\underline{f}(\underline{x}_1), \ldots, \underline{f}(\underline{x}_n))^T = \begin{pmatrix} f_1(\underline{x}_1) & \ldots & f_r(\underline{x}_1) \\ \vdots & \ddots & \vdots \\ f_1(\underline{x}_n) & \ldots & f_r(\underline{x}_n) \end{pmatrix} .$$

Das Normalgleichungssystem

- Die Schätzung $\underline{\hat{a}}$ des Vektors $\underline{a}$ der Regressionskoeffizienten kann dann mit Hilfe des sogenannten Normalgleichungssystems gefunden werden:

$$F^T F \, \underline{\hat{a}} = F^T \underline{y} \, . \tag{2}$$

  Dies ist ein lineares Gleichungssystem zur Bestimmung der Komponenten von $\underline{\hat{a}}$.

- Ist die Matrix $F^T F$ regulär, dann ist (2) eindeutig auflösbar und es gilt

$$\underline{\hat{a}} = \left(F^T F \right)^{-1} F^T \underline{y} \, . \tag{3}$$

Eigenschaften der Schätzung

- Unter der Annahme, dass die beobachteten Werte y_i Realisierungen der Zufallsgrößen

$$Y_i = a_1 f_1(\underline{x}_i) + \ldots + a_r f_r(\underline{x}_i) + \varepsilon_i$$

sind, wobei die zufälligen Fehler ε_i unabhängige normalverteilte Zufallsgrößen mit Erwartungswert 0 und konstanter Varianz σ^2 sind, ist die Schätzung $\hat{\underline{a}}$ aus (3) erwartungstreu und konsistent.

Beispiel 7.2 Jahresumsatz

- Fortsetzung Beispiel 6.4 (und 7.1)
  Daten aus Bleymüller et al, Statistik für Wirtschaftswissenschaftler, 2004, Kap. 20 .
- i Filiale
  x_{1i} Verkaufsfläche in Tsd. qm
  x_{2i} Passantenfrequenz in Tsd. Passanten pro Tag
  y_i Jahresumsatz in Mio. €

i	1	2	3	4	5	6
x_{1i}	0.31	0.98	1.21	1.29	1.12	1.49
x_{2i}	10.24	7.51	10.81	9.89	13.72	13.92
y_i	2.93	5.27	6.85	7.01	7.02	8.35

i	7	8	9	10	11	12
x_{1i}	0.78	0.94	1.29	0.48	0.24	0.55
x_{2i}	8.54	12.36	12.27	11.01	8.25	9.31
y_i	4.33	5.77	7.68	3.16	1.52	3.15

Fortsetzung Beispiel 7.2 Jahresumsatz

- Wir wählen als Ansatz mit $\underline{x} = (x_1, x_2)^T$

$$Y(\underline{x}) = a_1 + a_2 x_1 + a_3 x_2 + \varepsilon = (a_1, a_2, a_3) \begin{pmatrix} 1 \\ x_1 \\ x_2 \end{pmatrix} + \varepsilon.$$

- Dann erhält man

$$F^T = \begin{pmatrix} 1 & 1 & 1 & \dots \\ 0.31 & 0.98 & 1.21 & \dots \\ 10.24 & 7.51 & 10.81 & \dots \end{pmatrix},$$

$$F^T F = \begin{pmatrix} 12.00 & 10.68 & 127.83 \\ 10.68 & 11.41 & 118.97 \\ 127.83 & 118.97 & 1410.14 \end{pmatrix},$$

$$\underline{\hat{a}}^T = (-0.83\,, 4.74\,, 0.175)$$

und so $\quad \hat{y}(x_1, x_2) = -0.83 + 4.74\, x_1 + 0.175\, x_2$

als geschätzte Regressionsfunktion.

Beispiel 7.2 in Statgraphics

Multiple Regression - Jahresumsatz

Dependent variable: Jahresumsatz (Mio Euro)
Independent variables:
Verkaufsfläche (1000 qm)
Passantenfrequenz (1000/Tag)

		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	-0,831946	0,412663	-2,01604	0,0746
Verkaufsfläche	4,74295	0,226498	20,9403	0,0000
Passantenfrequenz	0,174988	0,0448676	3,90009	0,0036

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	52,8765	2	26,4383	384,18	0,0000
Residual	0,619359	9	0,0688177		
Total (Corr.)	53,4959	11			

R-squared = 98,8422 percent
Standard Error of Est. = 0,262331

Streuungszerlegung und Bestimmtheitsmaß

- Wie im Fall der einfachen linearen Regression gilt für den parameterlinearen Ansatz die Quadratsummenzerlegung (Streuungszerlegung) $\mathrm{SST} = \mathrm{SSE} + \mathrm{SSR}$ (bei Schätzung der Regressionskoeffizienten mit der Methode der kleinsten Quadrate).
- Dabei sind wieder

  $\mathrm{SST} = \sum_{i=1}^{n} (y_i - \overline{y})^2$, die Totalvariabilität (Totalvarianz);

  $\mathrm{SSE} = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$, die „erklärte“ Variabilität (erklärte Varianz);

  $\mathrm{SSR} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$, die Restvariabilität (Restvarianz).
- Das Bestimmtheitsmaß ist

  $$B = \frac{\mathrm{SSE}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSR}}{\mathrm{SST}} = r^2_{Y|(f_1(\underline{X}),\ldots,f_r(\underline{X}))} .$$

Streuungszerlegung und Bestimmtheitsmaß im Beispiel 7.2 in Statgraphics

Analysis of Variance

Source	***Sum of Squares***	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	**52,8765**	2	26,4383	384,18	0,0000
Residual	**0,619359**	9	0,0688177		
Total (Corr.)	**53,4959**	11			

R-squared = **98,8422** percent
Standard Error of Est. = 0,262331

- $\mathrm{SSE} = 52.8765$
- $\mathrm{SSR} = 0.619359$
- $\mathrm{SST} = 53.4959$
-

$$B = \frac{\mathrm{SSE}}{\mathrm{SST}} = \frac{52.8765}{53.4959} = 0.9884215$$

Schätzung der Fehlervarianz

- Eine konstante Varianz der zufälligen Fehler ε_i (und damit der Zufallsgrößen $Y(x_i)$ kann analog zum Fall der einfachen linearen Regression durch

$$\hat{\sigma}^2 = s^2_{Rest} = \frac{\mathrm{SSR}}{n-r}$$

geschätzt werden. Der Nenner $n-r$ ist durch die Schätzung von r Parametern bedingt.

- Im Beispiel 7.2:

$$\begin{aligned} \hat{\sigma}^2 &= s^2_{Rest} = \frac{\mathrm{SSR}}{n-r} = \frac{0.619359}{12-3} = 0.0688177 \\ \hat{\sigma} &= \sqrt{0.0688177} = 0.262331 \end{aligned}$$

- Statgraphics im Beispiel 7.2:

Standard Error of Est. = 0,262331

Konfidenzschätzungen I

- Für die folgenden Aussagen zu Konfidenzschätzungen und Tests setzen wir wieder voraus, dass die zufälligen Fehler ε_i unabhängige normalverteilte Zufallsgrößen mit Erwartungswert 0 und konstanter Varianz σ^2 sind.
- Mit m_i wird das $i-$te Diagonalelement der Matrix $(F^T F)^{-1}$ bezeichnet.
- Die Schätzung für die Varianz von $\hat{a}_i$ ist damit: $s^2_{a_i} = s^2_{Rest} m_i$.
- Konfidenzintervall zum Niveau $1-\alpha$ für die Komponente a_i von $\underline{a}$:
  $I = \left[\hat{a}_i - s_{a_i} t_{n-r;1-\alpha/2} \; ; \; \hat{a}_i + s_{a_i} t_{n-r;1-\alpha/2}, \right]$.
- Konfidenzintervall zum Niveau 0.95 für a_3 (vgl. Folie 15):
  $\hat{a}_3 = 0.175\,; \quad s_{a_3} = 0.0448676\,;$
  $t_{9;0.975} = 2.26 \quad \Rightarrow \quad I = [0.074; 0.276]\,.$

Konfidenzschätzungen II

- Ein Konfidenzintervall für die Fehlervarianz σ^2 ist
$$\left[\frac{(n-r)\hat{\sigma}^2}{\chi^2_{n-r;1-\alpha/2}} \;;\; \frac{(n-r)\hat{\sigma}^2}{\chi^2_{n-r;\alpha/2}}\right] = \left[\frac{\mathrm{SSR}}{\chi^2_{n-r;1-\alpha/2}} \;;\; \frac{\mathrm{SSR}}{\chi^2_{n-r;\alpha/2}}\right].$$
- Konfidenzintervall zum Niveau $1-\alpha$ für die Regressionsfunktion $\underline{f}(\underline{x})^T\underline{a}$:
$$I = \Bigg[\underline{f}(\underline{x})^T\hat{\underline{a}} - t_{n-r;1-\alpha/2}\sqrt{s^2_{Rest}\underline{f}(\underline{x})^T(F^TF)^{-1}\underline{f}(\underline{x})}\;;$$
$$\underline{f}(\underline{x})^T\hat{\underline{a}} + t_{n-r;1-\alpha/2}\sqrt{s^2_{Rest}\underline{f}(\underline{x})^T(F^TF)^{-1}\underline{f}(\underline{x})}\Bigg].$$
- Auch Prognoseintervalle können konstruiert werden.

$t-$Test für einzelne Parameter

- **Hypothesen:** $H_0 : a_i = a_i^{(0)}$, $H_A : a_i \neq a_i^{(0)}$.
- **Testgröße:** $T = \dfrac{\hat{a}_i - a_i^{(0)}}{s_{a_i}}$.

  Diese Testgröße ist unter H_0 $t-$verteilt mit $n - r$ Freiheitsgraden.
- **Kritischer Bereich zum Niveau α :**
  $K = \{t \in \mathbb{R} : |t| > t_{n-r;1-\alpha/2}\}$.
- Analog können einseitige Tests durchgeführt werden.
- Test mit $H_0 : a_3 = 0$, $H_A : a_3 \neq 0$, $\alpha = 0.05$ im Beispiel 7.2.
  $t = \dfrac{0.175}{0.0448676} = 3.9 > 2.26 = t_{9;0.975}$
  $\Rightarrow$ H_0 wird abgelehnt, d.h. der Koeffizient a_3 (der die Abhängigkeit des Umsatzes von der Passantenfrequenz beschreibt) ist signifikant verschieden von 0 .

$F-$Test für das Modell (Varianzanalyse)

- Wir setzen voraus, dass $f_1(x) = 1$ gilt, d.h. a_1 die Konstante im Modell ist.
- **Hypothesen:** $H_0 : a_2 = \ldots = a_r = 0\,, \quad H_A : a_i \neq 0$ für ein $i > 1\,.$
- **Testgröße:** $T = \dfrac{\mathrm{MSE}}{\mathrm{MSR}}$ mit $\mathrm{MSE} = \dfrac{\mathrm{SSE}}{r-1}\,,\ \mathrm{MSR} = \dfrac{\mathrm{SSR}}{n-r}\,.$

  Diese Testgröße ist unter H_0 $F-$verteilt mit $(r-1; n-r)$ Freiheitsgraden.
- **Kritischer Bereich zum Niveau α:**
  $K = \{t \in \mathbb{R} : t > F_{r-1;n-r;1-\alpha}\}\,.$

$t-$ und $F-$Tests in Statgraphics im Beispiel 7.2

$t-$Test für a_3 hervorgehoben:

		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	-0,831946	0,412663	-2,01604	0,0746
Verkaufsfläche	4,74295	0,226498	20,9403	0,0000
Passantenfrequenz	0,174988	0,0448676	**3,90009**	**0,0036**

$F-$Test:

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	52,8765	2	26,4383	**384,18**	**0,0000**
Residual	0,619359	9	0,0688177		
Total (Corr.)	53,4959	11			