Vorlesung Mathematik und Statistik
Teilgebiet Statistik

Regression

Ziel: Untersuchung der Beziehung zweier Variablen x und y. Die notwendigen Parameter zur Berechnung der Regression werden im folgenden beschrieben.

Kovarianz:
Wie wir bereits wissen, lässt sich die Beziehung zwischen 2 Zufallsvariablen durch die Kovarianz spezifizieren.

$$\sigma_{x,y} = \sum_{x,y} x \cdot y \cdot p(x,y) - \mu_x \mu_y$$

Für die Ableitung der Kovarianz aus der Stichprobe, gehen wir analog zur Berechnung der Stichprobenvarianz vor. Wir kennen die gemeinsame Verteilung der beiden Variablen nicht. Daher geben wir jeder Beobachtung dasselbe Gewicht. Wir berechnen die Kovarianz als Summe der Produkte der Abweichungen der beiden Variablen von ihren Mittelwerten:

$$s_{x,y} = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

Beachten Sie die Analogie zur Berechnung der Stichprobenvarianz. Wir hatten bereits in Kapitel 5 festgestellt, daß die Varianz ein Sonderfall der Kovarianz ist: nämlich die Kovarianz einer Variablen mit sich selbst. Auch für die Kovarianz gilt, daß bei der Berechnung ein Freiheitsgrad verlorengeht. Daher wird durch $(n-1)$ dividiert. Die Summe der Produkte der Abweichungen wird oft auch kurz als Summenprodukt (SP_{xy}) bezeichnet. Daher können wir schreiben:

$$s_{x,y} = \frac{SP_{xy}}{n-1}$$

Für die Berechnung des Summenprodukts kann man die vereinfachte Formel anwenden, die die Berechnung der Mittelwerte vermeidet:

$$SP_{xy} = \sum_{i=1}^{n} x_i \cdot y_i - \frac{\sum x_i \cdot \sum y_i}{n}$$

Korrelation:

$$r_{x,y} = \frac{s_{x,y}}{s_x \cdot s_y}$$

bzw.

$$r_{x,y} = \frac{SP_{xy}}{\sqrt{SQ_x \cdot SQ_y}}$$

Regression von y auf x:

Die Korrelation ist eine gute und dimensionslose Größe zur Messung der Enge der Beziehung zweier Variablen. Gerade dieser abstrakte Charakter macht sie jedoch für bestimmte Fragestellungen ungeeignet. Oftmals will man die Beziehung zwischen zwei Variablen dazu benutzen, die Werte der einen Variablen durch Korrektur für die Unterschiede in der anderen Variablen besser vergleichbar zu machen. Hierzu muß man wissen, um welchen Betrag sich die eine Variable ändert, wenn sich die andere Variable um eine Einheit ändert. Die Größe, die diese Veränderung quantifiziert, nennt man den *Regressionskoeffizienten*.

Beispiel

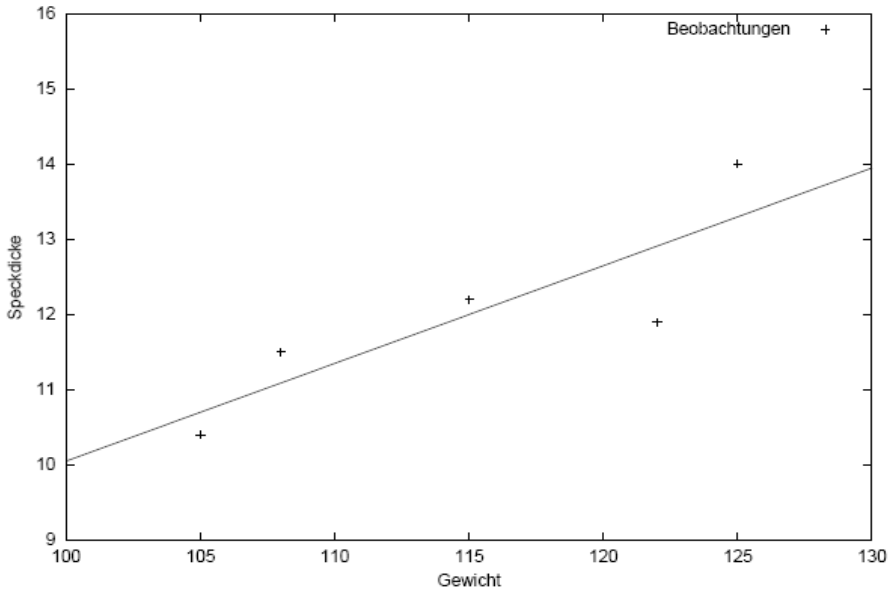

Abbildung 18: Beobachtungen von Gewicht und Rückenspeckdicke bei Auktionsebern

Es gilt nun, eine lineare Beziehung zwischen dem Merkmal Gewicht und Speckdicke (Regressionsgerade) zu schätzen. Diese Regressionsgerade hat die allgemeine Form:

$$\hat{y} = a + b \cdot x$$

hierbei sind $\hat{y}$ = der geschätzte Wert der Variablen y
a = der Ordinatenschnittpunkt der Geraden,
also der Wert von $\hat{y}$, wenn $x = 0$
b = der Regressionskoeffizient, also die Steigung der Geraden
x = der beobachtete Wert der Variablen x

Will man den Abstand der Punkte zur Regressionsgeraden minimieren, verbietet sich die einfache Summe der Abweichungen der tatsächlichen und der geschätzten Werte $d = Y - \hat{Y}$. Der Grund dafür ist, daß sich positive und negative Abweichungen ausgleichen könnten. Um dies zu vermeiden, kann man den mittleren Betrag der Abweichungen ($|d|$) oder die mittlere quadrierte Abweichung verwenden. Traditionell verwendet man letztere, was als Methode der kleinsten Quadrate[1] bekannt ist. Die Bedingung für die optimale Regressionsgerade lautet also:

$$\text{minimiere} \quad \sum d^2 = \sum (Y - \hat{Y})^2$$

Die Minimierung dieses Ausdrucks erfordert, daß wir das Maximum der ersten Ableitung dieses Ausdrucks finden. Hierzu formulieren wir zunächst die Regressionsgerade in einer anderen Form:

$$\hat{y} = (a + b\overline{x}) + b(x - \overline{x})$$

Den ersten Term bezeichnen wir als $a_* = (a + b\overline{x})$. Die zu minimierende Größe ergibt sich somit als:

$$\min \quad \sum [y - (a_* + b(x - \overline{x}))]^2$$

Die Funktion hängt von a_* und b ab und folglich müssen wir die beiden partiellen Ableitungen Null setzen, um das Minimum zu finden. Wir beginnen mit der partiellen Ableitung nach a_*:

$$\frac{\partial}{\partial a_*} \sum d^2 = \sum 2(y - a_* - b(x - \overline{x}))^1 \cdot (-1) = 0$$

Hieraus folgt:

$$\sum(y - a_* - b(x - \overline{x})) = 0$$

und

$$\sum y - na_* - b\sum(x - \overline{x})$$

Die Summe $(x - \overline{x})$ ist bekanntlich Null und daher erhalten wir den einfachen Ausdruck:

$$\sum y = na_*$$

und daraus:

$$a_* = \overline{y}$$

Setzen wir für a_* wieder den alten Ausdruck ein, so ergibt sich folgende Beziehung:

$$\hat{a} = \overline{y} - b\overline{x}$$

Für die partielle Ableitung nach dem Parameter b setzen wir den gefundenen Wert für $a_*(=\overline{y})$ in die zu minimierende Funktion ein. Dies ergibt Folgendes:

$$\frac{\partial}{\partial b}\sum d^2 = \sum 2(y - \overline{y} - b(x - \overline{x}))^1 \cdot (-(x - \overline{x})) = 0$$

und hieraus:

$$\sum(y - \overline{y})(x - \overline{x}) - b\sum(x - \overline{x})^2 = 0$$

Damit ergibt sich die Lösung:

$$\hat{b} = \frac{\sum(y - \overline{y})(x - \overline{x})}{\sum(x - \overline{x})^2} = \frac{SP_{xy}}{SQ_x}$$

Dividiert man Zähler und Nenner durch die FG, so erhält man die Lösung in Form von Varianzen und Kovarianzen:

$$\hat{b} = \frac{s_{x,y}}{s_x^2}$$

Aufgrund unserer Kenntnisse über den Parameter a können wir die Regressionsgerade in einer etwas anderen, aber ebenfalls häufig gebrauchten Form schreiben:

$$\hat{y} = \overline{y} + b \cdot (x - \overline{x})$$

Hieraus folgt unmittelbar die Form eines *linearen Modells*, in dem die Beobachtung ausschließlich durch den Wert von x bestimmt wird:

$$y_i = \overline{y} + b \cdot (x_i - \overline{x}) + e_i$$

Der Regressionskoeffizient gibt an, um wieviel Einheiten sich y ändert, wenn x sich um eine Einheit ändert.

Beachten Sie, daß diese Beziehung *nicht umgekehrt* gilt! Aus diesem Grund bezeichnet man den so ermittelten Regressionskoeffizienten auch als $b_{y,x}$, gelesen als "Regression von Y auf X". Y ist in diesem Fall die *abhängige* Variable und X die *unabhängige* Variable. $\hat{y}$ ist also eine Funktion von x. Aus der Form der Regressionsgleichung wird weiterhin ersichtlich, daß die Regressionsgerade *immer* durch den Punkt $(\overline{y},\overline{x})$ verläuft. Es muß also nur *ein* weiterer Punkt auf der Geraden bekannt sein, um die Regressionsgerade eindeutig zu bestimmen. Aus diesem Grund verbraucht das lineare Modell der Regression nur einen Freiheitsgrad.

Ein Beispiel mag die Berechnungsweise verdeutlichen. Das Zahlenmaterial in der Abbildung besteht aus den folgenden Werten:

Gewicht (x)	Speck (y)	$(x-\overline{x})^2$	$(y-\overline{y})^2$	$(x-\overline{x})(y-\overline{y})$
105	10.4	100	2.56	16.0
108	11.5	49	.25	3.5
115	12.2	0	.04	0.0
122	11.9	49	.01	-.7
125	14.0	100	4.00	20.0
$\sum$ 575	60.0	298	6.86	38.8

SP_{xy} = 38.8 (kg*mm)
SQ_x = 298 (kg*kg)
SQ_y = 6.86 (mm*mm)

$$b_{y,x} = \frac{SP_{xy}}{SQ_x} = \frac{38.80}{298} = 0.13\,(\text{mm/kg})$$

$$a = 12.0 - .13 \cdot 115 = -2.95\,(\text{mm})$$

Die Regressionsgerade hat also die Form:

$$\hat{y} = -2.95 + .13 \cdot x = 12 + .13 \cdot (x - 115)$$

Obwohl wir oben erwähnt haben, daß es nicht zulässig ist, den Regressionskoeffizienten $b_{y,x}$ zur Schätzung von x bei gegebenem y zu verwenden ist die Regression natürlich keine Einbahnstraße. Man muß in dem Fall nur die Regressionsgerade neu berechnen. Hierbei wird X mit Y vertauscht und folglich ergibt sich als Regressionskoeffizient:

$$b_{x,y} = \frac{SP_{xy}}{SQ_y}$$

In diesem Fall steht also das Summenquadrat von y im Nenner und für unser Beispiel ergibt sich:

$$b_{x,y} = \frac{38.8}{6.86} = 5.66\,(\text{kg/mm})$$

und die Regressionsgerade nimmt die Form

$$\hat{x} = 115 + 5.66 \cdot (y - 12.0)$$

an. Wenn sich die Rückenspeckdicke um einen Millimeter erhöht, steigt das Gewicht um 5.66 kg.

Merkregel: Bei der Regression von y **auf** x steht das SQ_x **unten**. Bei der Regression von x **auf** y steht das SQ_y **unten**.

Die beiden Regressionsgleichungen sind also nicht identisch. Die einzige Gemeinsamkeit ist, daß beide durch den Punkt $(\overline{x},\overline{y})$ gehen.

Korrektur mittels Regression

Wir kommen nun zurück zum eigentlichen Ziel unserer Analyse, nämlich, den Vergleich der Speckdicken der Auktionseber unbeeinflußt von Unterschieden im Gewicht zu ermöglichen. Hierzu werden die Speckdicken mit Hilfe der Regression auf ein einheitliches Gewicht korrigiert.[2] Das Regressionsmodell lautet:

$$y_i = \overline{y} + b \cdot (x_i - \overline{x}) + e_i$$

Die Korrektur erfolgt, indem wir von *jeder Beobachtung* den durch die Regression erklärten Teil subtrahieren:

$$y_i^* = y_i - b \cdot (x_i - \overline{x}) = \overline{y}_i + e_i$$

Aus dieser Gleichung wird ersichtlich, wie die Korrektur wirkt: überdurchschnittlich schwere Schweine haben eine positive Abweichung $(x_i - \overline{x})$. Bei ihnen erfolgt eine *Verminderung* der gemessenen Speckdicke. Leichte Schweine dagegen erhalten einen *Zuschlag* zur gemessenen Speckdicke.

Bei negativen Regressionskoeffizienten wirkt die Korrektur in umgekehrter Weise. In unserem Fall wurde auf das mittlere Gewicht korrigiert. Dies ist jedoch nicht unbedingt notwendig. Man kann auch auf ein willkürlich festgelegtes Standardgewicht korrigieren, solange es im Bereich der Gewichte liegt, an denen die Regression geschätzt wurde. In diesem Fall wird anstelle von $\overline{x}$ das Standardgewicht in die Korrekturformel eingesetzt.