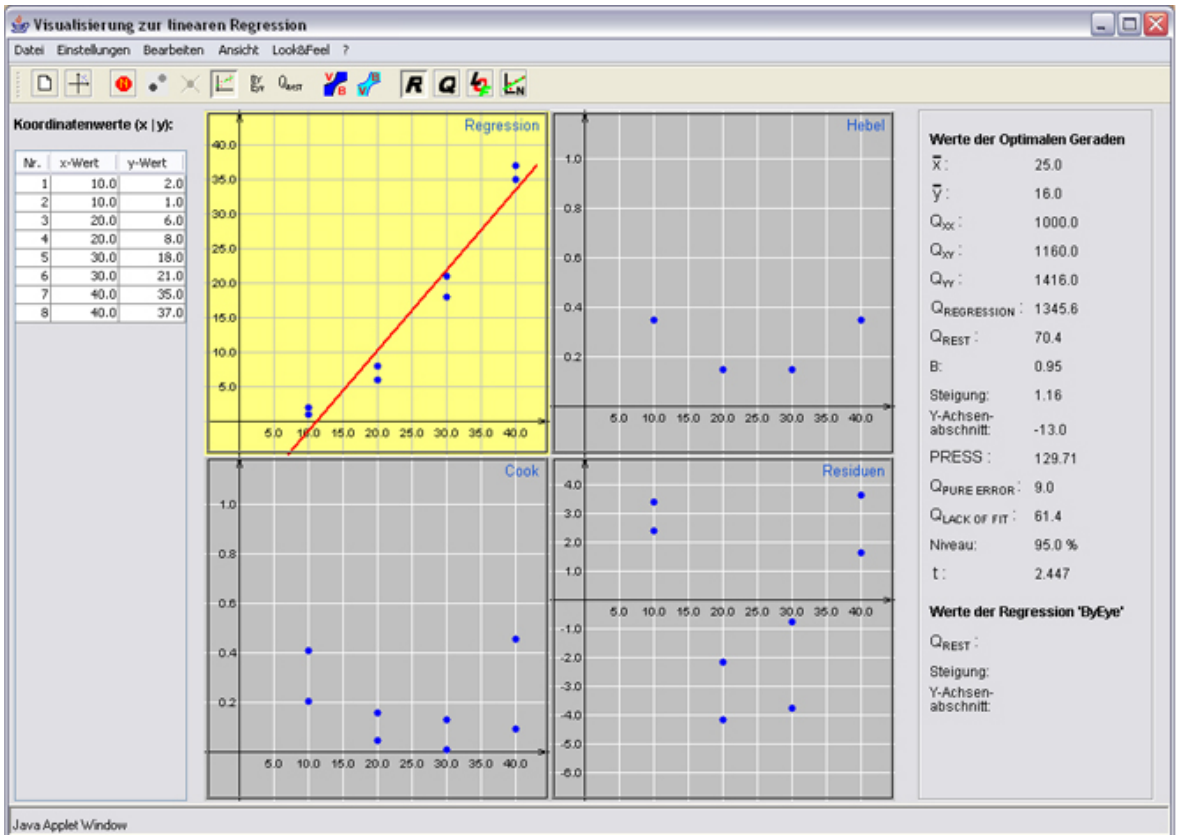
Übungen mit dem Applet



by
Michael Gärtner

Betreuer: Prof. Dr. Wilhelm Kleppmann
Abgabe: 20. October 2006

Inhaltsverzeichnis

1 Prinzip der kleinsten Quadrate

Bei der linearen Regression werden die Koeffizienten b_0 und b_1 der Regressionsgeraden $\widehat{y} = b_1 \cdot x + b_0$ so bestimmt, dass die Summe der quadrierten Abweichungen der Punkte von der Geraden in y-Richtung so klein wie möglich ist. b_0 ist der Achsenabschnitt auf der y-Achse und b_1 die Steigung.
Öffnen Sie eines der vorgegebenen Beispiele (Menüpunkt Datei>Beispiele), überlegen Sie, wie die Regressionsgerade aussehen könnte, wählen Sie das Icon "ByEyeünd tragen Sie Ihre Gerade ein. Im Ausgabebereich rechts finden Sie als Q_{REST} unten die Summe der quadrierten Abweichungen von Ihrer Geraden, weiter oben Q_{REST} der besten Geraden. Nehmen Sie die Herausforderung an und versuchen Sie, dem Bestwert möglichst nahe zu kommen. Als Hilfe können Sie mit dem Icon  die Quadrate anzeigen lassen: Ihre Aufgabe ist es, die Summe der Quadratflächen zu minimieren. Mit diesem Icon  erhalten Sie die bestmögliche Gerade zum Vergleich. Wenn Sie die "ByEyeGerade ausschalten, sehen Sie die Summe der quadrierten Abweichungen von der besten Gerade.

2 Quadrierte Abweichungen und Bestimmtheitsmaß

Das Bestimmtheitsmaß B gibt an, welcher Anteil der Summe der quadrierten Abweichungen Q_{YY} der y-Werte von ihrem Mittelwert durch die Regressionsgerade erklärt wird. Es gilt immer $0 \leq B \leq 1$. Je näher B bei 1 liegt, desto besser passt die Gerade. Wählen Sie ein Beispiel aus und tragen Sie die Regressionsgerade ein. Mit dem Icon  erhalten Sie vier verschiedene Darstellungen:

- Links oben finden Sie die Quadrate der Abweichungen vom Mittelwert aller y-Werte - die Summe dieser Quadratflächen ist Q_{YY} .
- Rechts oben finden Sie den Anteil $Q_{REGRESSION}$, den die Regressionsgerade erklärt (die Quadrate beginnen immer auf der Geraden). Es gilt:

  Bestimmtheitsmaß $B = r^2 = \dfrac{Q_{REGRESSION}}{Q_{YY}}$
- Links unten finden Sie Q_{REST} , die Summe der quadrierten Abweichungen von der Regressionsgeraden. Es gilt immer: $Q_{YY} = Q_{REST} + Q_{REGRESSION}$ - überzeugen Sie sich mit Hilfe der Ausgaben in der rechten Spalte davon. Die Minimierung von Q_{REST} ist damit gleichzeitig eine Maximierung von $Q_{REGRESSION}$ und B.
- Rechts unten finden Sie PRESS (**p**redicted **re**sidual **s**um of **s**quares). Wie Q_{REST} ist auch PRESS eine Summe der quadrierten Abweichungen der y-Werte von der Regressionsgeraden. Der Unterschied besteht nur darin, dass jetzt bei

der Berechnung der Regressionsgeraden jeweils der betrachtete Punkt nicht berücksichtigt wird. Damit erhält man für jeden Punkt eine andere Regressionsgerade und die Vorhersagekraft der Geraden wird getestet. Die Quadrate können überlappen und PRESS ist immer größer als Q_{REST} - der Wert wird in der rechten Spalte angezeigt.

Fassen Sie nun mit der Maus einen Punkt, der besonders weit von der Regressionsgeraden entfernt liegt und schieben Sie ihn näher an die Gerade heran. Beobachten Sie, wie Q_{REST} kleiner wird (am besten grafisch) und B größer. Verschieben Sie immer mehr Punkte zur Geraden hin und beobachten Sie, wie B sich 1 annähert und Q_{REST} immer kleiner wird. PRESS sollte nicht wesentlich größer sein als Q_{REST} . Wenn ein einzelner Punkt von der Geraden abweicht, ist die quadratische Abweichung dieses Punktes bei PRESS jedoch wesentlich größer als bei Q_{REST} - insbesondere wenn sich der x-Wert deutlich von denen der anderen Punkte unterscheidet. Testen Sie dies, indem Sie einen Punkt weit von den anderen wegschieben.

3 Einfluss einzelner Punkte auf die Gerade

Wählen Sie ein Beispiel aus und tragen Sie die Regressionsgerade ein. Wählen Sie einen Punkt aus der Mitte des x-Bereichs und verschieben Sie ihn in y-Richtung. Die Regressionsgerade wird parallel verschoben, d.h. der Achsenabschnitt verändert sich, aber nicht die Steigung. Wählen Sie nun einen Punkt vom Rand des x-Bereichs und verschieben Sie ihn in y-Richtung. Die Regressionsgerade ändert die Steigung. Fügen Sie nun einen Punkt hinzu mit wesentlich größerem x-Wert als alle anderen Punkte. Wenn Sie diesen Punkt verschieben, folgt ihm die Gerade fast völlig - er bestimmt die Steigung, weil er einen großen Hebel hat (Maßzahl Hebel, leverage). Schalten Sie nun mit dem Icon  auf die Diagnosedarstellungen um.

- Das Bild oben links zeigt wie bisher die Punkte und die daran angepasste Regressionsgerade.
- Das Bild rechts oben zeigt den Hebel: Die in x-Richtung dicht beieinander liegenden Punkte haben einen kleinen Hebel, der weit entfernt liegende Punkt hat einen Hebel von fast 1 (er bestimmt die Steigung fast für sich allein"). Der Hebel hängt nur von den x-Werten ab.
- Das Bild rechts unten zeigt die Residuen (die Abweichungen von der Regressionsgeraden) - der Punkt mit dem großen Hebel hat meist ein relativ kleines Residuum, das Residuum ist unauffällig. Darin liegt ein großes Risiko, wenn zur Diagnose nur die Residuen betrachtet werden.
- Das Bild links unten zeigt die Cookdistanz - sie berücksichtigt x- und y-Werte gleichzeitig. Punkte mit großem Hebel haben einen großen Einfluss auf das

angepasste Modell - besonders, wenn die y-Werte nicht zu denen der anderen Punkte passen. Solche Punkte haben dann eine große Cookdistanz. Punkte mit großer Cookdistanz sollten daher besonders sorgfältig überprüft werden.

Verschieben Sie nun einzelne Punkte und beobachten Sie die Wirkung dieser Verschiebungen auf Regressionsgerade, Hebel, Cookdistanz und Residuen. Beachten Sie insbesondere, dass auch ein Punkt mit einem großen Hebel eine kleine Cookdistanz haben kann, wenn er gut zu der Geraden durch die restlichen Punkte passt. Wenn Sie eine solche Anordnung der Punkte gefunden haben, schalten Sie mit dem Icon  zur Darstellung der quadrierten Abweichungen und überzeugen Sie sich, dass der Beitrag dieses Punktes zu PRESS klein ist.

4 Vertrauensbereich und Vorhersagebereich

Wenn der tatsächliche Zusammenhang zwischen x- und y-Werten linear ist und die Abweichungen der Einzelwerte von der Geraden zufällig und normalverteilt sind mit fester Standardabweichung, ist die Regressionsgerade die beste Schätzgerade für den tatsächlichen Zusammenhang. Der Vertrauensbereich enthält dann den wahren Mittelwert der Verteilung der y-Werte mit vorgegebenem Vertrauensniveau, der Vorhersagebereich enthält die Einzelwerte (und ist natürlich immer breiter, weil er zusätzlich die Streuung der Einzelwerte um den Mittelwert berücksichtigt (vgl. das Applet zum Vertrauensbereich für den Mittelwert). Wählen Sie ein Beispiel aus und tragen Sie die Regressionsgerade ein. Schalten Sie mit dem Icon 

den Vertrauensbereich ein. In der Mitte ist er am schmälsten, die Breite dort ergibt sich ähnlich zum Vertrauensbereich für den Mittelwert. Nach außen wird er immer breiter, da sich dort die Unsicherheit in der Steigung zusätzlich bemerkbar macht. Grundeinstellung ist ein Vertrauensniveau von 95%. Mit dem Menüpunkt Einstellungen>Niveaueinstellungen können Sie Werte zwischen 90% und 99,9% einstellen. Beachten Sie, dass die Bereiche umso breiter sind, je höher das gewählte Vertrauensniveau ist. Je sicherer Sie sein wollen, dass der tatsächliche Mittelwert im Vertrauensbereich liegt, desto breiter wird der Bereich. Schalten Sie mit dem Icon 

zusätzlich den Vorhersagebereich ein - er ist breiter, weil er zusätzlich die Streuung der Einzelwerte berücksichtigt, hat aber im Prinzip dieselbe Trichterform wie der Vertrauensbereich. Verschieben Sie nun einzelne Punkte auf die Regressionsgerade zu - Vertrauensbereich und Vorhersagebereich sind umso schmäler, je weniger die Einzelwerte von der Geraden abweichen.

5 Nichtlineare Punktewolke und quasilineare Regression

Wählen Sie das Beispiel "Betonäus. Es zeigt die Zugfestigkeit von Betonproben, die nach unterschiedlichen Trockenzeiten (in Tagen) entnommen und gemessen wurden (aus Graf/ Henning/Stange/Wilrich). Der Zusammenhang ist deutlich nichtlinear: Am Anfang nimmt die Zugfestigkeit schnell zu, dann immer langsamer und nach einigen Wochen nähert sie sich einem Grenzwert. Tragen Sie die Regressionsgerade ein. Obwohl der Zusammenhang offensichtlich nichtlinear ist, erhält man eine Gerade, weil es so vorgegeben wurde. Das bedeutet: Die Form der Funktion wird vorgegeben und nicht angepasst. Wenn eine falsche Funktion vorgegeben wird, erhält man auch ein falsches Ergebnis. Die Form der Funktion muss aufgrund inhaltlich/technischer Überlegungen festgelegt werden. Im Fall der Zugfestigkeit hat z.B. die Funktion $Zug = \alpha \cdot e^{\frac{-\beta}{t}}$ die inhaltlich richtige Form (Sättigungsverhalten). Logarithmiert man diesen Zusammenhang, so erhält man $\ln(Zug) = \ln \alpha - \frac{\beta}{t}$ d.h. trägt man $\ln(Zug)$ gegen $\frac{1}{t}$ auf, so erwartet man einen linearen Zusammenhang, dessen Parameter man mit linearer Regression in diesen transformierten Variablen bestimmen kann, so genannter quasilinearer Regression. Die transformierten Daten finden Sie im Beispiel "Beton transformiert". Hier ist lineare Regression sinnvoll. Hinweis: Auch $Zug = \alpha \cdot (1 - e^{-\lambda \cdot t})$ beschreibt ein Sättigungsverhalten. Für diese Funktion gibt es jedoch keine Transformation, die sie linearisiert. Sie muss daher mit einem iterativen Verfahren angepasst werden (nichtlineare Regression).

6 Prüfung auf Nichtlinearität

Hierzu eignen sich besonders die Beispiele "Betonünd "Bremsweg gegen Geschwindigkeit". In beiden Beispielen ist der tatsächliche Zusammenhang offensichtlich nicht linear. Wie kann dies erkannt werden? Neben der grafischen Darstellung der Daten selbst eignen sich die Darstellung der Residuen und der Linearitätstest.

Darstellung der Residuen

Schalten Sie mit dem Icon  die Diagnosegrafiken ein. In der Darstellung links oben ist die Abweichung von der Linearität erkennbar, in der Darstellung der Residuen gegen den x-Wert rechts unten ist dieselbe Abweichung noch deutlicher erkennbar. Residuen zeigen Strukturen in der Abweichung von der Geraden deutlicher - man betrachtet die Abweichungen mit einer Lupe". Diese Darstellung der Residuen ist immer möglich und sinnvoll.

Lack of Fit (Linearitätstest)
Der Lack-of-Fit-Test (Linearitätstest) ist nur möglich, wenn für mindestens einen

x-Wert mehrere y-Werte vorliegen. Mit dem Icon  schalten Sie die zugehörigen Grafiken ein. Beim Lack-of-Fit-Test (Linearitätstest) wird die Summe der quadrierten Abweichungen von der Regressionsgeraden Q_{REST} zerlegt in einen Anteil Lack of Fit"(Abweichung der Gruppenmittelwerte von der Regressionsgeraden) und einen Anteil Pure Error"(Abweichung der Einzelwerte vom jeweiligen Gruppenmittelwert). Die rechte Spalte zeigt die Werte - beachten Sie, dass $Q_{REST} = Q_{LACK\ OF\ FIT} + Q_{PURE\ ERROR}$. Der Lack-of-Fit-Test besteht dann darin, dass der Lack-of-Fit pro Freiheitsgrad

$$s^2_{LACK\ OF\ FIT} = \frac{Q_{LACK\ OF\ FIT}}{f_{LACK\ OF\ FIT}}$$

verglichen wird mit dem Pure Error pro Freiheitsgrad

$$s^2_{PURE\ ERROR} = \frac{Q_{PURE\ ERROR}}{f_{PURE\ ERROR}}$$

.
Ist der Lack-of-Fit wesentlich größer (um einen kritischen Faktor) als der Pure Error, so ist die Abweichung von der Linearität statistisch signifikant.