

Einfache lineare Regression

Ein begleiteter Einstieg in angepasste Werte, Steigungen und Residualvariation

Ratiomera Statistics

Inhaltsverzeichnis

Zusammenfassung der Kursseite	2
Die angepasste Gerade	2
Inferenz und Diagnostik	2
Grenzen der Schlussfolgerung	3
Erweiterte Referenz	4
Zweck und Grundlagen	4
Zentrale Ideen	4
Formelleitfaden	5
Die erklärende Abbildung lesen	7
Checkliste zur Interpretation	8
Verbindung zu anderen Themen	8

Zusammenfassung der Kursseite

Die einfache lineare Regression modelliert den Mittelwert einer quantitativen Ergebnisvariable aus einem Prädiktor. Sie sagt nicht jeden Fall exakt voraus. Die angepasste Gerade beschreibt den bedingten Mittelwert. Jedes Residuum hält die beobachtete Abweichung eines Falls von seinem angepassten Wert fest.

Die angepasste Gerade

Das Populationsmodell lautet

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

und die angepasste Stichprobengerade lautet

$$\hat{Y}_i = b_0 + b_1 X_i.$$

Tabelle 1: Jede Regressionsgrösse beantwortet einen anderen Teil der Modellfrage.

Grösse	Bedeutung
b_0	Angepasster mittlerer Ergebniswert bei $X = 0$; nur interpretieren, wenn null sinnvoll und durch die Daten gestützt ist
b_1	Angepasste Änderung des mittleren Ergebnisses bei einer Erhöhung von X um eine Einheit
$\hat{Y}_i$	Angepasster mittlerer Ergebniswert beim Prädiktorwert von Fall i
$e_i = Y_i - \hat{Y}_i$	Vertikales Residuum; positiv bedeutet, dass der beobachtete Wert oberhalb der Geraden liegt
R^2	Anteil der Stichprobenvariation des Ergebnisses, den die angepasste Gerade darstellt
Residualstandardfehler	Typische verbleibende Ergebnisvariation um die Gerade in Ergebniseinheiten

Die Steigung knüpft direkt an Thema 4 an:

$$b_1 = \frac{s_{XY}}{s_X^2} = r_{XY} \frac{s_Y}{s_X}.$$

Die Kovarianz liefert die gemeinsame Variation. Die Division durch die Prädiktorvarianz verwandelt sie in Ergebniseinheiten pro Prädiktoreinheit. Die Korrelationsform zeigt, wie der einheitenfreie Zusammenhang auf die ursprünglichen Skalen zurückkehrt. In einer einfachen Regression mit Achsenabschnitt gilt $R^2 = r^2$.

Inferenz und Diagnostik

Ein Konfidenzintervall und ein Test für die Steigung verwenden

$$t = \frac{b_1 - 0}{SE(b_1)}, \quad df = n - 2.$$

Prüfe vor der Inferenz, ob eine Gerade das Muster angemessen zusammenfasst, die Residuen eine ungefähr stabile Streuung haben, die Fehlerverteilung mit dem Modell vereinbar ist und die

Fälle gemäss Design unabhängig sind. Ein Residuen-gegen-Anpassungswerte-Diagramm prüft Form und Streuung. Ein Normal-Q-Q-Diagramm prüft die grobe Fehlerform. Hebelwerte und Cook-Distanzen markieren Fälle zur Untersuchung, nicht zur automatischen Löschung.

Grenzen der Schlussfolgerung

- Interpolation bleibt im gestützten Prädiktorbereich. Extrapolation geht darüber hinaus und benötigt stärkere Annahmen.
- Eine Vorhersage für eine Einzelperson braucht mehr Unsicherheit als ein Intervall für den mittleren Ergebniswert beim selben Prädiktorwert.
- Ein kleiner p-Wert beweist weder Kausalität noch korrekte Messung oder gute Vorhersage in einer neuen Population.
- Ein grosses R^2 beweist keine korrekte Modellspezifikation.

Thema 6 verwendet die Residuenidee zweimal: Beide interessierenden Variablen werden für eine dritte Variable bereinigt, danach wird das Verbleibende korreliert. Dadurch ist die partielle Korrelation der natürliche nächste Schritt nach der Ein-Prädiktor-Geraden.

Erweiterte Referenz

Zweck und Grundlagen

Die einfache lineare Regression beschreibt, wie sich der bedingte Mittelwert einer numerischen Zielvariable über die Werte eines Prädiktors verändert. Die **Zielvariable**, geschrieben als Y , ist die Variable, die das Modell erklären oder vorhersagen soll. Der **Prädiktor**, geschrieben als X , liefert die erklärende Information. «Einfach» bedeutet, dass das Modell einen Prädiktor enthält. «Linear» bedeutet, dass sich der modellierte Mittelwert von Y entlang einer Geraden verändert, wenn sich X verändert.

Jeder beobachtete Fall besitzt einen Prädiktorwert x_i und einen Wert der Zielvariable y_i . Das Modell liefert einen angepassten Wert $\hat{y}_i$. Dies ist der von der angepassten Geraden vorhergesagte Wert der Zielvariable beim Prädiktorwert dieses Falles. Die Differenz zwischen beobachtetem und angepasstem Wert ist das **Residuum** e_i . Ein positives Residuum liegt oberhalb der Geraden, ein negatives darunter. Residuen behalten die Einheit der Zielvariable und zeigen den Teil jedes beobachteten Werts, den die angepasste Gerade nicht wiedergegeben hat.

Regression ist eng mit Kovarianz und Korrelation verbunden, doch die Rollen der Variablen unterscheiden sich. Die Korrelation behandelt beide Variablen symmetrisch und fasst ihren standardisierten linearen Zusammenhang zusammen. Die Regression weist ihnen verschiedene Rollen als Prädiktor und Zielvariable zu, behält die Einheit der Zielvariable und liefert eine Gleichung für angepasste Werte. Das Vertauschen von X und Y erzeugt deshalb ein anderes Regressionsproblem, obwohl ihre Korrelation unverändert bleibt.

Bestandteil	Bedeutung	Einheit
Achsenabschnitt b_0	Angepasster Mittelwert der Zielvariable bei $X = 0$	Einheiten der Zielvariable
Steigung b_1	Veränderung des angepassten Werts der Zielvariable bei einem Anstieg von X um eine Einheit	Einheiten der Zielvariable pro Prädiktoreinheit
Angepasster Wert $\hat{y}_i$	Punkt auf der angepassten Geraden bei x_i	Einheiten der Zielvariable
Residuum e_i	Beobachteter minus angepasster Wert der Zielvariable	Einheiten der Zielvariable

Zentrale Ideen

Der Achsenabschnitt ist mathematisch erforderlich, um die Gerade zu positionieren. Seine inhaltliche Interpretation hängt jedoch davon ab, ob null sinnvoll und in den Daten vertreten ist. Liegen die beobachteten Prädiktorwerte weit von null entfernt, ist der Achsenabschnitt eine Extrapolation. Berichte ihn in diesem Fall als Modellkoeffizienten, ohne ihm eine reale Ausgangsbedeutung zuzuschreiben, die der beobachtete Prädiktorbereich nicht stützen kann.

Die Steigung ist der zentrale Koeffizient. Eine positive Steigung bedeutet, dass der angepasste Mittelwert der Zielvariable mit steigendem Prädiktor zunimmt. Eine negative Steigung bedeutet, dass er abnimmt. Ihr Betrag muss zusammen mit beiden Einheiten gelesen werden. Eine Steigung von zwei bedeutet zwei Einheiten der Zielvariable pro Prädiktoreinheit und nicht eine Korrelation von zwei. Die Steigung beschreibt ein durchschnittliches bedingtes Muster und keine garantierte Veränderung bei jedem Fall.

Die gewöhnliche Methode der kleinsten Quadrate wählt Achsenabschnitt und Steigung so, dass die Summe der quadrierten Residuen möglichst klein wird. Durch das Quadrieren können sich positive

und negative Residuen nicht gegenseitig aufheben. Zugleich erhalten grössere Abweichungen mehr Gewicht. Wenn ein Achsenabschnitt enthalten ist, verläuft die angepasste Gerade durch $(\bar{x}, \bar{y})$. Abgesehen von numerischer Rundung addieren sich die Residuen dann ungefähr zu null.

Diagnosemerkmal	Erwünschtes Muster	Mögliche Sorge bei einem sichtbaren Muster
Residuen gegen angepasste Werte	Ungeordnetes Band um null	Krümmung, veränderliche Streuung oder ausgelassene Struktur
Residualstreuung	Über die angepassten Werte annähernd gleich	Nicht konstante bedingte Varianz
Vergleich mit Normalquantilen	Annähernd geradliniges Muster, wenn Inferenz mit normalverteilten Fehlern verwendet wird	Starke Abweichungen an den Rändern oder ungewöhnliche Residuen
Hebelwirkung und Einfluss	Kein einzelner Fall dominiert die angepasste Gerade	Ein Fall an einer ungewöhnlichen Prädiktorposition kann die Koeffizienten stark verändern

R^2 vergleicht die nach Anpassung der Geraden verbleibende Variation mit der Gesamtvariation um den Mittelwert der Zielvariable. Bei einem gewöhnlichen Modell mit Achsenabschnitt liegt es zwischen null und eins. Ein grösserer Wert bedeutet, dass die angepasste Gerade mehr Stichprobenvariation in Y darstellt. Er belegt jedoch keine Kausalität, garantiert keine genauen Vorhersagen für einzelne Personen und beweist nicht, dass die Modellform geeignet ist. Ein hohes R^2 kann zusammen mit einem systematischen Residualmuster auftreten.

Die Inferenz für die Steigung fragt, ob ein linearer Zusammenhang in der Grundgesamtheit unter dem Modell mit null vereinbar ist. Ein Konfidenzintervall zeigt, welche Steigungswerte mit der Schätzung und ihrem Standardfehler vereinbar sind. Die Annahmen betreffen die bedingte Beziehung: eine lineare Mittelwertstruktur, unabhängige Beobachtungen, eine geeignete Residualvarianz und eine für die beabsichtigte Inferenz angemessene Residualverteilung. Der Prädiktor selbst muss nicht normalverteilt sein.

Punktvorhersagen erfordern Vorsicht. Ein angepasster Wert schätzt den bedingten Mittelwert der Zielvariable bei einem gewählten Prädiktorwert. Er garantiert nicht das Ergebnis einer einzelnen Person. Eine Vorhersage innerhalb des beobachteten Bereichs ist eine Interpolation. Eine Vorhersage ausserhalb dieses Bereichs ist eine Extrapolation und setzt die ungeprüfte Fortsetzung der angepassten Geraden voraus. Der nicht durch Beobachtungen gestützte Bereich muss daher ausdrücklich genannt werden.

Formelleitfaden

Das Populationsmodell trennt eine systematische Gerade von einem individuellen Fehlerterm:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

β_0 und β_1 sind Populationskoeffizienten. Der Fehler ε_i stellt die Differenz zwischen Fall i und dem bedingten Mittelwert der Grundgesamtheit dar. Nach der Anpassung an Stichprobendaten lautet die geschätzte Gerade:

$$\hat{y}_i = b_0 + b_1 x_i$$

Das Residuum vergleicht den beobachteten Wert mit diesem angepassten Wert:

$$e_i = y_i - \hat{y}_i$$

Die Kleinste-Quadrate-Steigung kann mit den bereits bei der Kovarianz eingeführten Kreuzprodukten geschrieben werden:

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Der Achsenabschnitt positioniert die Gerade anschliessend so, dass sie durch die beiden Stichprobenmittelwerte verläuft:

$$b_0 = \bar{y} - b_1 \bar{x}$$

Das Bestimmtheitsmass ist der Anteil der Gesamtvariation der Zielvariable, der durch das angepasste Modell dargestellt wird:

$$R^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Bei einem Prädiktor und einem Achsenabschnitt gilt $R^2 = r_{xy}^2$. Die Steigung kann auch als $b_1 = r_{xy}(s_y/s_x)$ geschrieben werden. Diese Darstellung zeigt, wie der standardisierte Zusammenhang in die ursprünglichen Einheiten der Variablen zurückübersetzt wird.

Die angepasste Gerade zerlegt jede beobachtete Abweichung in eine Modellkomponente und eine Residualkomponente:

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i).$$

Bei einem OLS-Modell mit Achsenabschnitt ergibt das Quadrieren und Summieren über alle Fälle:

$$SS_{\text{total}} = SS_{\text{model}} + SS_{\text{error}},$$

mit

$$SS_{\text{total}} = \sum_i (y_i - \bar{y})^2, \quad SS_{\text{model}} = \sum_i (\hat{y}_i - \bar{y})^2, \quad SS_{\text{error}} = \sum_i (y_i - \hat{y}_i)^2.$$

Die exakte Zerlegung der Quadratsummen ist ein Ergebnis der OLS-Anpassung über die gesamte Stichprobe. Quadriere nicht die drei vorzeichenbehafteten Abstände eines einzelnen Falles in der Erwartung, dass dieselbe Identität Fall für Fall gilt.

Der Residualstandardfehler schätzt die typische Streuung der Fehler in Einheiten der Zielvariable:

$$s_e = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n-2}}.$$

Der Nenner verwendet $n-2$, weil für die Gerade ein Achsenabschnitt und eine Steigung geschätzt werden. Diese Grösse unterscheidet sich vom Standardfehler der Steigung. Dieser misst die Unsicherheit von b_1 über hypothetisch wiederholte Stichproben:

$$SE(b_1) = \frac{s_e}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}.$$

Für $H_0 : \beta_1 = 0$ lauten der Koeffiziententest und das zugehörige zweiseitige Intervall:

$$t = \frac{b_1}{SE(b_1)}, \quad df = n - 2,$$

$$b_1 \pm t_{1-\alpha/2, n-2} SE(b_1).$$

Schliesst das zugehörige Intervall null aus, verwirft der entsprechende zweiseitige Test die Nullhypothese einer Steigung von null. Der Einschluss von null bedeutet, dass die Nullhypothese nicht verworfen wird. Er beweist nicht, dass die Populationssteigung genau null ist.

Die Modelltabelle teilt die Modell- und Fehlerquadratsummen durch ihre Freiheitsgrade:

$$F = \frac{MS_{\text{model}}}{MS_{\text{error}}}.$$

Bei einem Prädiktor stellen der globale Modelltest und der zweiseitige Steigungstest dieselbe Nullfrage. Unter demselben Modell gilt daher $F = t^2$, und ihre p-Werte stimmen überein.

Vorhersagefrage	Bedeutung	Erforderliche Vorsicht
Angepasster Wert bei x_0	Geschätzter bedingter Mittelwert $\hat{y} = b_0 + b_1 x_0$	Er ist kein garantiertes Ergebnis für einen einzelnen Fall
Interpolation	x_0 liegt innerhalb des beobachteten Prädiktorbereichs	Modellform und Diagnosen bleiben wichtig
Extrapolation	x_0 liegt ausserhalb des beobachteten Prädiktorbereichs	Die Beziehung kann anders aussehen, wo keine Daten beobachtet wurden

Die Messqualität gehört zur Interpretation. Im klassischen Messfehlermodell des Prädiktors, das im bereitgestellten Material entwickelt wird, zieht Rauschen in X die Steigung der einfachen Regression im Allgemeinen gegen null. Diese Abschwächung kann einen zugrunde liegenden Zusammenhang schwächer erscheinen lassen. Daraus folgt nicht, dass jeder mögliche Messfehlerprozess dieselbe Verzerrung erzeugt.

Die erklärende Abbildung lesen

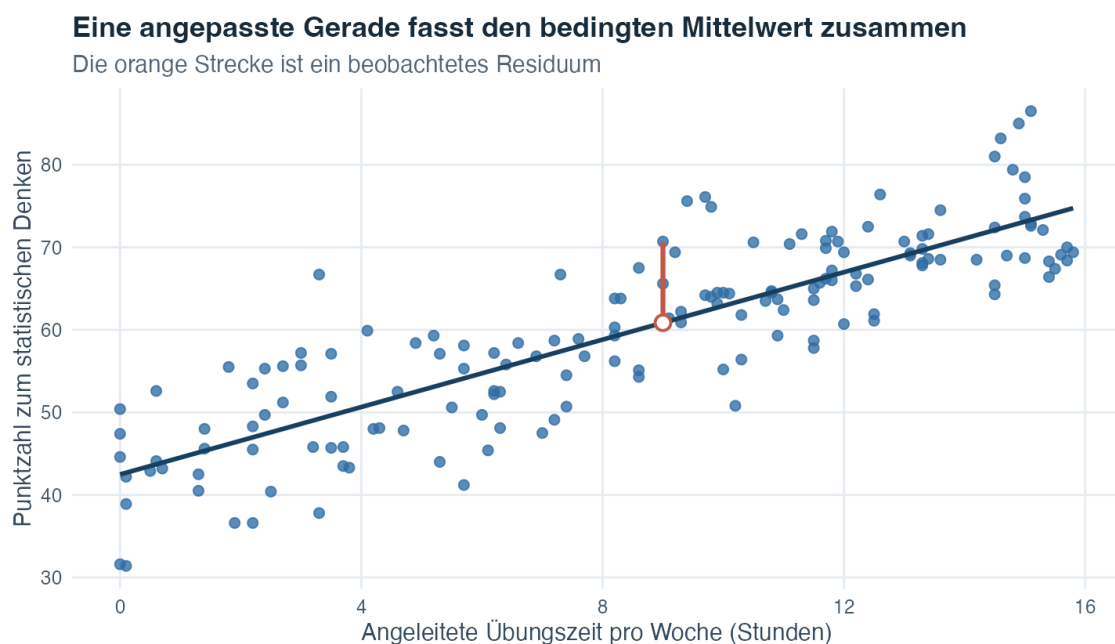


Abbildung 1: Streudiagramm der Stunden angeleiteter Übung und der Punktwerte zum statistischen Denken mit einer ansteigenden angepassten Geraden und einem orangefarbenen vertikalen Residualegment.

Die horizontale Achse zeigt die wöchentlichen Stunden angeleiteter Übung, die vertikale Achse die Punktwerte zum statistischen Denken. Jeder blaue Punkt ist ein simulierter Fall. Die dunkle Gerade steigt an. In diesem Datensatz ist der angepasste mittlere Punktwert bei grösseren Übungswerten also höher. Die Gerade beschreibt das durchschnittliche modellierte Muster. Einzelne Punkte bleiben oberhalb und unterhalb der Geraden verstreut. Dies erinnert daran, dass die Information eines Prädiktors nicht das Ergebnis jeder Person festlegt.

Bei ungefähr neun Übungsstunden markiert der hohle Kreis den angepassten Punktwert auf der Geraden. Der beobachtete Punkt dieses Falles liegt höher. Das orangefarbene vertikale Segment ist das Residuum: beobachteter minus angepasster Punktwert. Seine Richtung ist positiv und seine Länge wird in Punktwerten gemessen. Die Methode der kleinsten Quadrate führt diesen Vergleich für jeden Punkt durch und wählt die Gerade mit der kleinsten Gesamtsumme der quadrierten Residuallängen.

Die Grafik stützt eine Aussage über einen linearen Zusammenhang in den simulierten Daten. Sie zeigt nicht, dass zusätzliche Übung höhere Punktwerte verursacht hat. Vorwissen, Auswahl, Tutoriumsformat oder andere Variablen können mit beidem verbunden sein. Die Grafik zeigt auch nicht, ob die Residualannahmen erfüllt sind. Für Krümmung, veränderliche Varianz, ungewöhnliche Residuen und Einfluss werden getrennte Diagnosegrafiken benötigt.

Checkliste zur Interpretation

Benenne Zielvariable und Prädiktor und gib ihre Einheiten an. Untersuche ihre Verteilungen und das Streudiagramm. Prüfe, ob eine geradlinige Zusammenfassung über den beobachteten Bereich geeignet ist. Berichte die angepasste Gleichung und übersetze die Steigung in einen vollständigen Satz, der beide Einheiten enthält. Interpretiere den Achsenabschnitt nur, wenn sein Bezugswert inhaltlich sinnvoll ist. Berichte R^2 als vom Modell dargestellte Stichprobenvariation und nicht als kausalen Prozentsatz.

Untersuche Residual- und Einflussdiagnosen, bevor du dich auf Inferenz stützt. Berichte, wenn relevant, Steigungsschätzung, Standardfehler, Konfidenzintervall, Teststatistik, Freiheitsgrade und p-Wert. Unterscheide einen geschätzten bedingten Mittelwert von einem garantierten individuellen Ergebnis und kennzeichne Extrapolationen. Beschreibe simulierte Ergebnisse als simuliert und trenne die Sprache des Zusammenhangs von kausaler Sprache.

Verbindung zu anderen Themen

Dieses Modell macht die Verbindung zu Kovarianz und Korrelation konkret. Die Kovarianz liefert den Zähler der Steigung, die Prädiktorvarianz ihren Nenner und die Korrelation standardisiert dieselbe lineare Paarung. Die Regression ergänzt Richtung und Einheiten: Sie fragt, wie sich der angepasste Mittelwert einer festgelegten Zielvariable über den Prädiktor verändert.

Die nächsten Schritte behandeln eine zentrale Begrenzung der Geraden mit einem Prädiktor. Eine dritte Variable kann einen Teil des beobachteten Zusammenhangs erklären. Die partielle Korrelation entfernt aus beiden interessierenden Variablen jene linearen Bestandteile, die mit dieser dritten Variable verbunden sind, und korreliert das Verbleibende. Die multiple Regression nimmt mehrere Prädiktoren in ein Modell auf. Dadurch beschreibt jeder Koeffizient einen bedingten Zusammenhang, während die anderen Prädiktoren konstant gehalten werden. Die hier eingeführte Residualidee wird zur gemeinsamen Sprache beider Erweiterungen.

Notizen 1 von 3

Einfache lineare Regression | Thema 5

Notizen 2 von 3

Einfache lineare Regression | Thema 5

Notizen 3 von 3

Einfache lineare Regression | Thema 5