Statistik1

Geradenmodelle 1

Sebastian Sauer

Einstieg

Lernziele

- Sie können ein Punktmodell von einem Geradenmodell begrifflich unterscheiden.
- Sie können die Bestandteile eines Geradenmodells aufzählen und erläutern.
- Sie können die Güte eines Geradenmodells anhand von Kennzahlen bestimmen.
- Sie können Geradenmodelle sowie ihre Modellgüte in R berechnen.

Vorhersagen

Was macht eine gute Vorhersage aus?

Vorhersagen sind nützlich, unter (mindestens) folgenden Voraussetzungen:

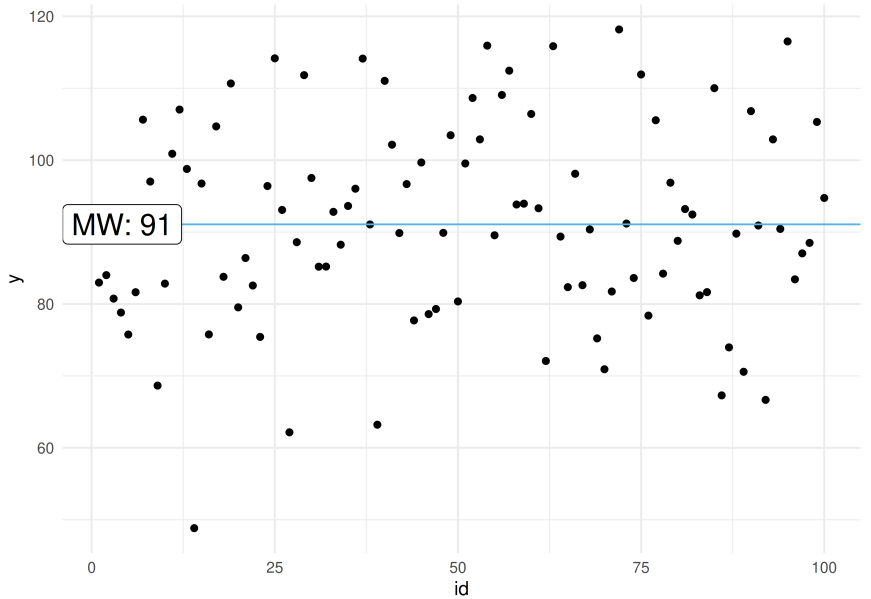
1. Sie sind präzise.
2. Wir wissen, wie präzise.
3. Jemand interessiert sich für die Vorhersage.

Die Methode des Vorhersagens, die wir hier betrachten, nennt man auch *lineare Regression*.

Vorhersage ohne Prädiktor: Tonis Frage

Toni fragt: “Welche Statistiknote kann ich in der Klausur erwarten?” — ohne weitere Angaben.

Beste Antwort: der Notenschnitt (Mittelwert) der letzten Klausur — wir nutzen den Mittelwert als *Punktmodell*.



Vorhersagen

Definition: Nullmodell (Punktmodell)

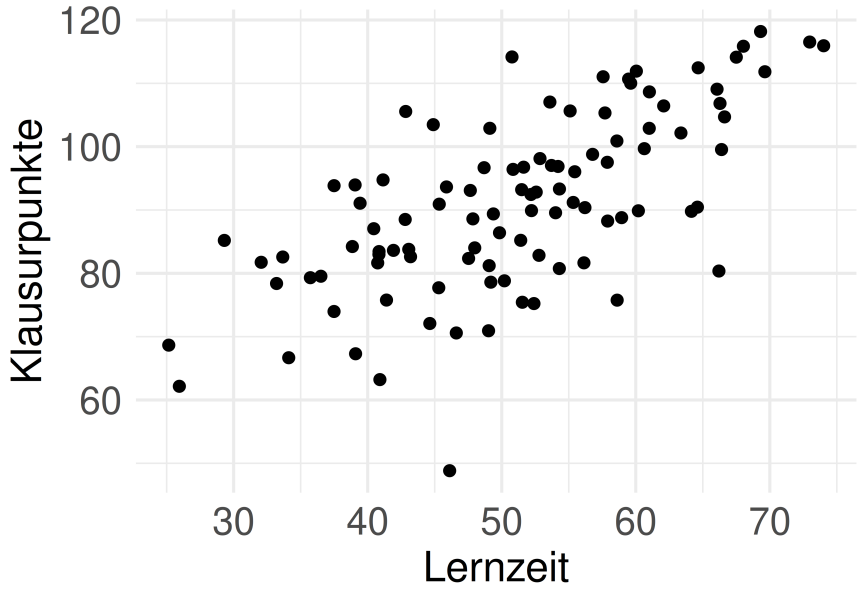
Definition: Nullmodell (Punktmodell)

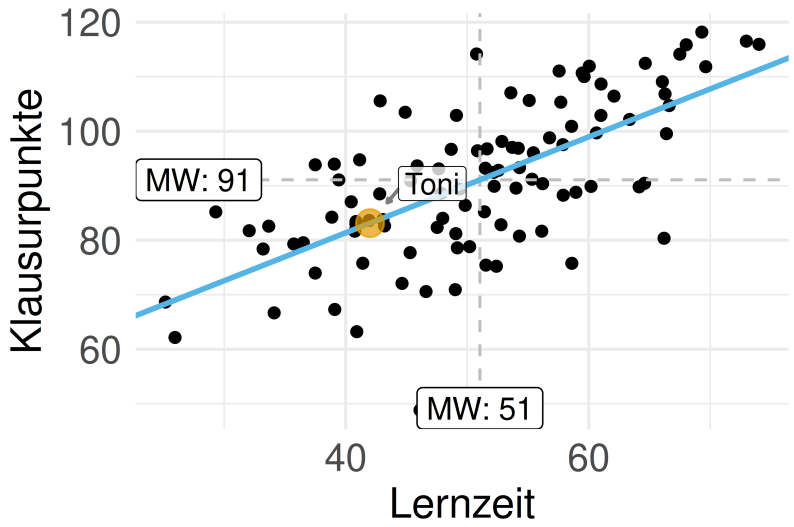
Modelle ohne UV, Punktmodelle also, kann man so bezeichnen: `y ~ 1`. Da das Modell null UV hat, nennt man es auch manchmal *Nullmodell*.

Auf Errisch: `lm(y ~ 1, data = noten2)` — der zurückgemeldete Koeffizient `(Intercept)` ist in diesem Fall der Mittelwert.

Toni verrät die Lernzeit

Toni: “Ich hab insgesamt 42 Stunden gelernt.” Der Trend im Streudiagramm ist deutlich zu erkennen: Je mehr Lernzeit, desto mehr Klausurpunkte.





Geradenmodelle

Von Punktmodell zu Geradenmodell

Ein *Geradenmodell* ist eine Verallgemeinerung des Punktmodells: Ein Punktmodell sagt für alle Beobachtungen den gleichen Wert vorher.

In einem Geradenmodell wird nicht mehr (notwendig) für jede Beobachtung die gleiche Vorhersage $\hat{y}$ gemacht.

Definition: Gerade

Definition: Gerade

Eine Gerade ist das, was man bekommt, wenn man eine lineare Funktion in ein Koordinatensystem einzeichnet. Man kann sie durch zwei *Koeffizienten* festlegen: Achsenabschnitt (engl. *intercept*) und Steigung (engl. *slope*).

Die Geradengleichung

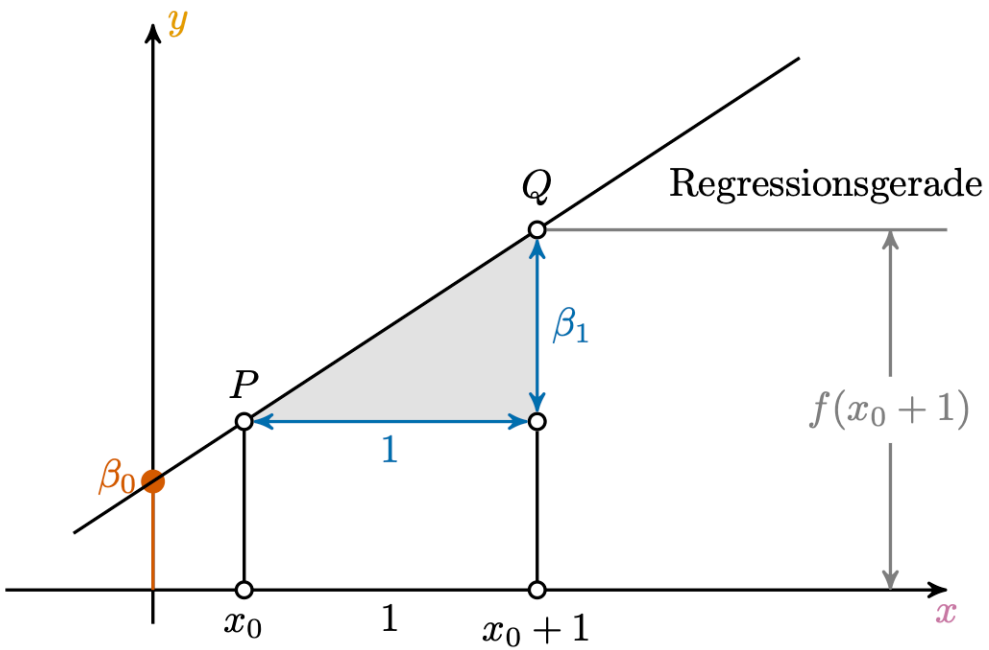
In der Statistik wird folgende Nomenklatur bevorzugt:

$$f(x) = \hat{y} = \beta_0 + \beta_1 x$$

Die Nomenklatur mit b_0, b_1 hat den Vorteil, dass man das Modell einfach erweitern kann: $b_2, b_3, \ldots$. Anstelle von b liest man auch oft β.

Das “Dach” über y, $\hat{y}$, drückt aus, dass es sich um den vom Modell vorhergesagten Wert handelt — nicht den tatsächlich beobachteten Wert von y.

Elemente einer Regressionsgeraden



Achsenabschnitt (β_0) und Steigung (β_1) einer Regressionsgeraden (Menk, 2014)

Definition: Das einfache lineare Modell

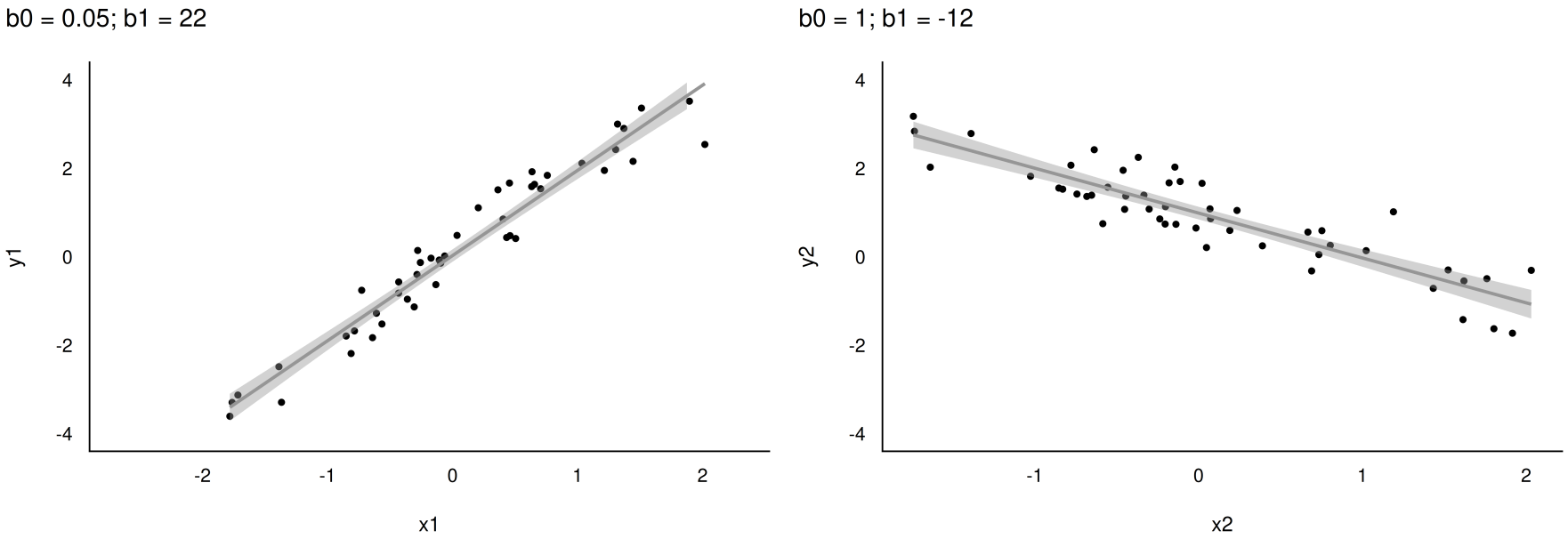
Definition: Das einfache lineare Modell

Beschreibt den Wert einer abhängigen metrischen Variablen, y, als lineare Funktion von einer unabhängigen Variablen, x, plus einem Fehlerterm, ϵ.

$$y = f(x) + \epsilon$$
$$y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$$

- β_0: geschätzter y-Achsenabschnitt laut Modell (*intercept*)
- β_1: geschätzte Steigung laut Modell (*slope*)
- ϵ: Fehler des Modells

Verschiedene Daten, verschiedene Geraden



Je nach Datenlage können sich Regressionsgeraden in Steigung oder Achsenabschnitt unterscheiden.

Spezifikation eines Geradenmodells

$$\hat{y} \sim \mathrm{x}$$

Lies: “Laut meinem Modell ist mein vorhergesagtes $\hat{y}$ irgendeine Funktion von x.” Wir erinnern uns: $AV \sim UV$.

In Errisch: `lm(y ~ x, data = meine_daten)`. `lm` steht für “lineares Modell” — die Gerade nennt man auch *Regressionsgerade*.

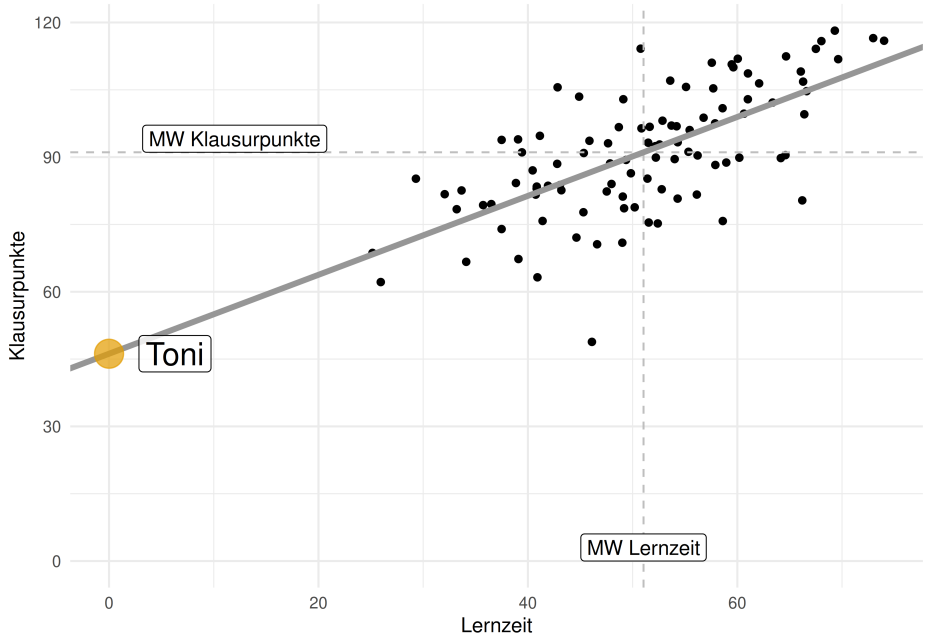
Koeffizienten in R berechnen

```
1 lm_toni <- lm(y ~ x, data = noten2)
2 lm_toni
```

Die Regressionsgleichung lautet demnach: `y_pred = 8.6 + 0.88*x`.

`8.6` ist der Achsenabschnitt (Wert von y bei $x = 0$). `0.88` ist die Steigung: Für jede Stunde Lernzeit steigt der vorhergesagte Klausurerfolg um `0.88` Punkte.

Der Achsenabschnitt

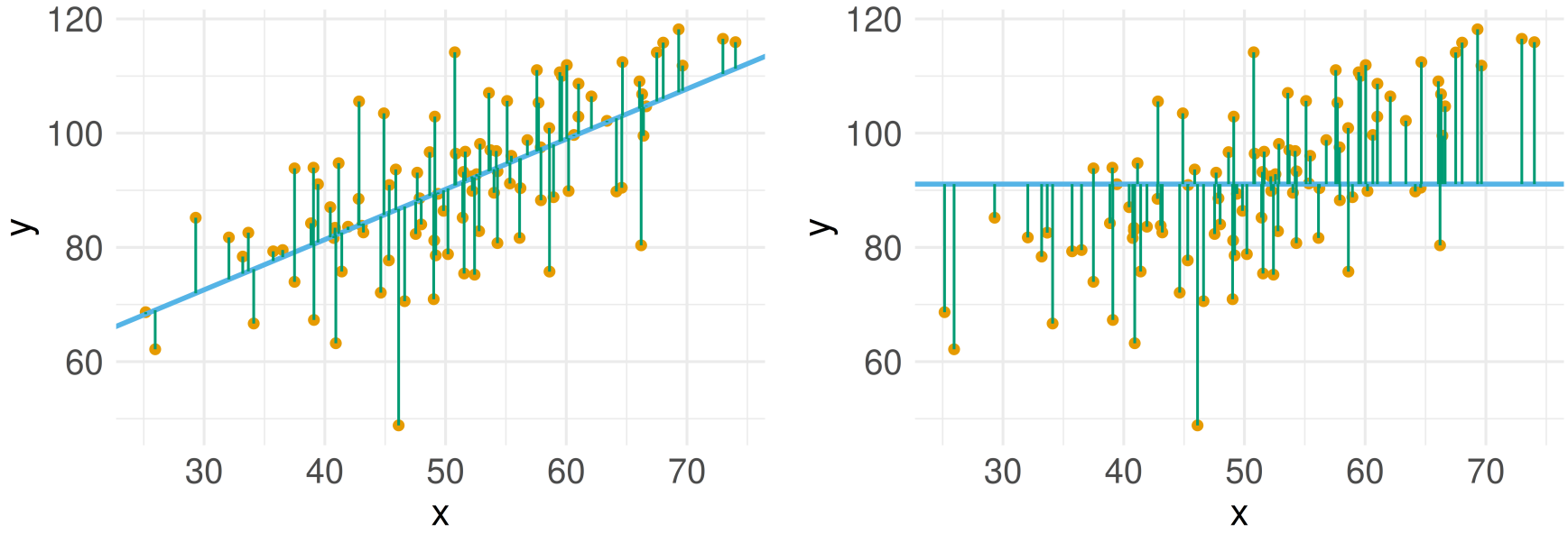


Vorhersagefehler (Residuum)

Die Differenz zwischen vorhergesagtem Wert und tatsächlichem Wert nennt man Vorhersagefehler (*error*) oder *Residuum*:

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

Vorhersagefehler im Vergleich



Beim Geradenmodell (`lm_toni`) ist die mittlere Absolutabweichung (MAE) deutlich kleiner als beim Nullmodell (`lm0`) — das Geradenmodell ist viel besser.

Definition: Fehlerstreuung

Definition: Fehlerstreuung

Als Fehlerstreuung bezeichnen wir die Verteilung der Abweichungen der beobachteten Werte (y_i) vom vorhergesagten Wert ($\hat{y}_i$).

Kenngrößen dafür: MAE oder MSE. Ein Geradenmodell ist immer besser als ein Punktmodell, solange X mit Y korreliert ist.

Berechnung der Modellkoeffizienten

Wie legt man die Regressionsgerade in das Streudiagramm?
Die Koeffizienten β_0 und β_1 wählt man so, dass die Residuen *minimal* sind — genauer: die Summe der quadrierten Residuen wird minimiert.

$$\min \sum_i e_i^2$$

R^2 als Maß der Modellgüte

R^2 als Verringerung der Fehlerstreuung

Wir vergleichen die Fehlerstreuung (MSE) des Modells mit der Fehlerstreuung (MSE) des Nullmodells:

$$R^2 = 1 - \frac{\text{MSE}_m}{\text{MSE}_{m0}}$$

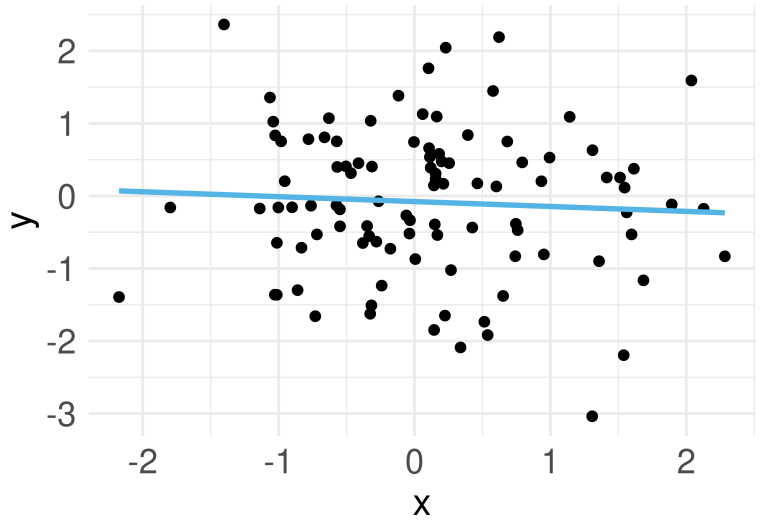
```
1 r2(lm_toni)
```

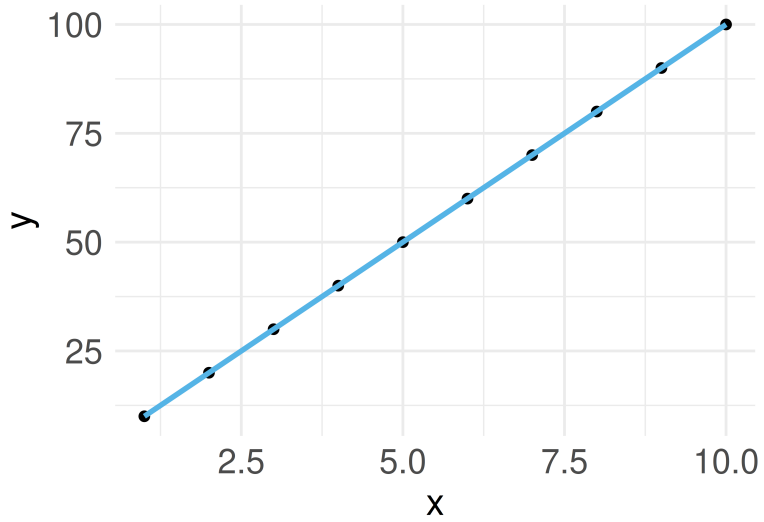
Definition: R-Quadrat

Definition: R-Quadrat

Der Anteil der Verringerung der Fehlerstreuung zwischen `lm0` und dem gerade untersuchten Modell nennt man *R-Quadrat* (R^2). R^2 ist ein Maß der Modellgüte: Je größer R^2, desto besser die Vorhersage. Wertebereich: 0 bis 1. Im Nullmodell beträgt R^2 per Definition Null. Für Modelle des Typs $y \sim x$ gilt: $R^2 = r^2_{xy}$.

Extremfälle von R^2





Je größer R-Quadrat, desto besser passt das Modell zu den Daten.

Interpretation eines Regressionsmodells

Modellgüte

Die Residuen (Vorhersagefehler) bestimmen die Modellgüte: Sind die Residuen im Schnitt groß, ist die Modellgüte gering, und umgekehrt.

Verschiedene Koeffizienten stehen zur Verfügung: R^2, r (Korrelation von tatsächlichem y und vorhergesagtem $\hat{y}$), MSE, RMSE, MAE, …

Koeffizienten interpretieren

- Achsenabschnitt (β_0) und Steigung (β_1) sind nur eingeschränkt zu interpretieren, wenn man die zugrundeliegenden kausalen Abhängigkeiten nicht kennt.
- Allein aufgrund eines statistischen Zusammenhangs darf man keine kausalen Abhängigkeiten annehmen.
- Beispiel: Bei einer Regression von Gewicht auf Körpergröße ist der Achsenabschnitt wenig nützlich — es gibt keine Menschen der Größe 0.

Die Steigung interpretieren

"Im Modell `lm_toni` beträgt der Regressionskoeffizient $\beta_1 = 0{,}88$. Zwei Studentinnen, deren Lernzeit sich um eine Stunde unterscheidet, unterscheiden sich *laut Modell* um den Wert von β_1."

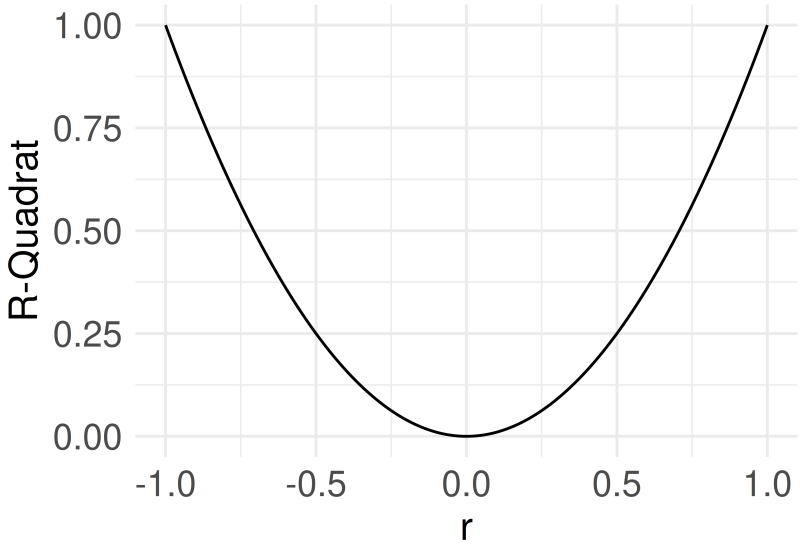
Vorsicht

"Effekt" klingt kausal. Ohne weitere Absicherung darf man Regressionskoeffizienten *nicht* kausal verstehen — besser: *statistischer* Effekt.

Wie man mit Statistik lügt

r vs. R^2: eine nichtlineare Beziehung

Der Unterschied in Modellgüte zwischen $r = .1$ und $r = .2$ ist *viel kleiner* als zwischen $r = .7$ und $r = .8$ – da $r = \sqrt{R^2}$, dürfen Unterschiede in r nicht wie Unterschiede in R^2 interpretiert werden.



Vorsicht

Unterschiede zwischen Korrelationsdifferenzen dürfen nicht linear interpretiert werden.

Fallbeispiel Mariokart

Vorhersage des Verkaufspreises

Ziel: eine genaue Vorhersage der Verkaufspreise (`total_pr`) von Mariokart-Auktionen.

```
lm2 <- lm(total_pr ~ start_pr, data = mariokart)
r2(lm2)
```

Oh nein! Unterirdisch schlecht. Anstelle von bloßem Rumprobieren: Welche Variable korreliert am stärksten mit `total_pr`?

Ein besseres Modell: `ship_pr`

```
lm3 <- lm(total_pr ~ ship_pr, data = mariokart)
parameters(lm3)
```

- Achsenabschnitt: erwarteter Verkaufspreis bei 0 Dollar Versandkosten (“Sockelbetrag”).
- Steigung: Anstieg des erwarteten Verkaufspreises *pro Dollar* Versandkosten.
- Regressionsgleichung: `total_pr_pred = b0 + b1 * ship_pr`

Fallstudie Immobilienpreise

Ziel: Häuserpreise vorhersagen

In dieser Fallstudie geht es darum, Immobilienpreise vorherzusagen — Datenbasis ist ein Kaggle-Wettbewerb (“House Prices”).

- `d_train`: Trainingsdaten zum Modellbau (enthält `SalePrice`)
- `d_test`: Testdaten, für die der Preis vorhergesagt werden soll

Mehrere Prädiktoren in einem Modell

```
mein_modell <- lm(av ~ uv1 + uv2 + ... + uv_n, data = meine_daten)
```

Das Pluszeichen ist hier kein arithmetischer Operator, sondern bedeutet: “nimm UV1 *und* UV2 *und* … als Prädiktoren.”

```
lm_immo2 <- lm(SalePrice ~ OverallQual + GrLivArea + GarageCars, data = d_t
```

Modellvergleich per RMSE

- `lm_immo1`: nur `OverallQual` als Prädiktor
- `lm_immo2`: `OverallQual + GrLivArea + GarageCars`
- `m0`: Nullmodell (nur der Mittelwert)

```
1 rmse(lm_immo1)
2 rmse(lm_immo2)
3 rmse(m0)
```

Ob die Modellgüte “gut” ist, beurteilt man am besten *relativ* — im Vergleich zu anderen Modellen.

Vom Modell zur eingereichten Prognose

```
1 lm_immo2_pred <- predict(lm_immo2, newdata = d_test)
```

Die Vorhersagen werden mit den Spalten `Id` und `SalePrice` als CSV-Datei eingereicht — das allgemeine Muster jeder Prognose auf neuen (unbeobachteten) Daten.

Zusammenfassung

- Ein **Geradenmodell** verallgemeinert das Punktmodell: $\hat{y} = \beta_0 + \beta_1 x$
- Achsenabschnitt (β_0) und Steigung (β_1) legen die Regressionsgerade fest
- Die Koeffizienten werden per *Methode der kleinsten Quadrate* geschätzt — die Summe der quadrierten Residuen wird minimiert
- **R^2** misst die Verringerung der Fehlerstreuung gegenüber dem Nullmodell (Wertebereich 0–1)
- Regressionskoeffizienten sind ohne zugrundeliegende Theorie *nicht* kausal interpretierbar
- Mehrere Prädiktoren lassen sich mit `+` kombinieren; Modellgüte ist über RMSE/MAE/R^2 vergleichbar

Literaturhinweise

- Gelman et al. (2021) — eine deutlich umfassendere Einführung in die Regressionsanalyse
- Çetinkaya-Runde & Hardin (2021) — moderne, R-orientierte Einführung inklusive Regressionsanalyse
- Cohen et al. (2003) — ein Klassiker mit viel Aha-Potenzial

Referenzen

Çetinkaya-Runde, M., & Hardin, J. (2021). *Introduction to Modern Statistics*. https://openintro-ims.netlify.app/
Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, 3rd Ed*. Lawrence Erlbaum.
Gelman, A., Hill, J., & Vehtari, A. (2021). *Regression and Other Stories*. Cambridge University Press.
Menk. (2014, Juli 29). *Linear Regression [Computer Code]*. https://texample.net/tikz/examples/linear-regression/