DigStat

Digitale Lerneinheiten in der Statistik

Vorstellung und Evaluation des Projekts

Farhad Razeghpour
farhad.razeghpour@rub.de

Christian Müller
c.mueller@hhu.de

ORCA.nrw-Tagung, 18.11.2024

Ein Kooperationsvorhaben empfohlen durch die:
DIGITALE HOCHSCHULE NRW
INNOVATION DURCH KOOPERATION

Gefördert durch:
Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen

Warum brauchen wir noch eine Einführung in die Statistik?

- Studierende wünschen sich mehr digitale Lernmaterialien in mathematischen Lehrveranstaltungen. (CHE, 2022)
- Dozierende haben nicht genügend zeitliche Ressourcen, alle Lernmaterialien eigenständig zu konzipieren. (ORCA, 2023)
- Elaboriertes Feedback ist ein entscheidender Faktor zur Unterstützung der Lern- und Motivationsprozesse von Studierenden. (Hattie, 2014)
- Studierende erhalten aus zeitlichen Gründen selten elaboriertes Feedback bei der Bearbeitung von Aufgaben. (Mai et al., 2021)

Lernziele von DigStat: Studierende können ...

- Daten aus verschiedenen Datenquellen in R aufbereiten, indem sie mithilfe entsprechender Befehle ein R-Skript schreiben, um die anschließende statistische Analyse durchzuführen.
- verschiedene statistische Verfahren zur Lösung eines Problems gegenüberstellen, indem sie die Rahmenbedingungen und mathematischen Modellannahmen vergleichen, um entsprechende Grenzen dieser Verfahren abzuschätzen und ein für den vorliegenden Datensatz geeignetes Verfahren zu identifizieren.
- einschätzen, inwiefern der Einsatz eines statistischen Verfahrens gerechtfertigt ist, indem sie die mathematischen Modellannahmen bei einer konkreten Datenlage überprüfen, um die Aussagekraft der Ergebnisse, beispielsweise einer Studie, im Sinne der Statistical Literacy kritisch zu beurteilen.

Der Moodle-Kurs behandelt die Grundlagen der Statistik mit R

Jede Lerneinheit besteht aus Skripten und STACK-Aufgaben

Lerneinheit: Lineare Regression

Allgemeine Informationen zur Lerneinheit

▸ Klicken Sie hier für mehr Informationen zu den Inhalten, Lernzielen und Voraussetzungen.

Kapitel 1: Einfache lineare Regression

Skript: Einfache lineare Regression

In diesem Kapitel behandeln wir die einfache lineare Regression. Wir erklären das zugrunde liegende statistische Modell und wie mithilfe der Kleinste-Quadrate-Methode Schätzwerte für die Parameter der Regressionsgeraden ermittelt werden können.

Aufgabe: Energieverbrauch in Abhängigkeit der Außentemperatur

In dieser Aufgabe lernen Sie, wie Sie eine Regressionsgerade bestimmen, indem Sie die Steigung und den y-Achsenabschnitt aus zusammengefassten Daten schätzen. Weiter lernen Sie, wie man Konfidenzintervalle für die Parameter bestimmt.

Aufgabe: Lineare Regression mit R

In dieser Aufgabe lernen Sie, wie Sie eine lineare Regression mithilfe von R ausführen. Zur Bearbeitung der Aufgabe ist R nicht notwendig.

Kapitel 2: Hypothesentest und Konfidenzintervalle bei der einfachen linearen Regression

Skript: Hypothesentest und Konfidenzintervall

Die Skripte verknüpfen die mathematische Theorie . . .

1 Einleitung und Motivation

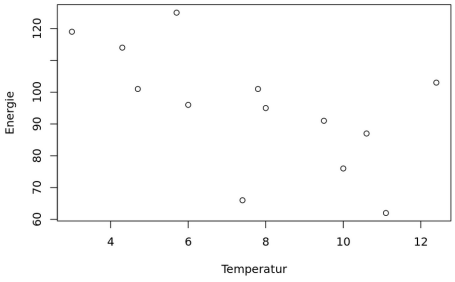
Neben Zufallsexperimenten, die bei Wiederholungen unabhängig und unter identischen Bedingungen ausgeführt wurden und entsprechend als Realisierungen unabhängiger und identisch verteilter Zufallsvariablen aufgefasst werden können, kann der Ausgang eines Experiments außer vom Zufall auch noch vom Wert einer *erklärenden Variablen* abhängen. Das Ziel einer statistischen Analyse ist es dann, die funktionale Abhängigkeit zwischen der erklärenden Variablen und dem Ergebnis des Experiments zu beschreiben.

Beispiel 1 Professor D. interessiert sich für die Abhängigkeit des Energieverbrauchs seiner Fernwärmeheizung von der morgendlichen Außentemperatur. Im November 2019 hat er dazu an 13 aufeinanderfolgenden Tagen die Außentemperatur (in Grad Celsius) um 7:00 morgens sowie den Energieverbrauch (in kWh) an diesem Tag abgelesen. Er liest die Daten wie folgt in R ein:

```
Daten <- data.frame(
  Temperatur = c(7.4, 7.8, 4.7, 3.0, 5.7, 12.4, 10.6, 11.1, 10.0, 9.5, 8.0, 6.0, 4.3),
  Energie = c(66, 101, 101, 119, 125, 103, 87, 62, 76, 91, 95, 96, 114)
  )
Daten
```

Temperatur <dbl>	Energie <dbl>
7.4	66
7.8	101
4.7	101

Hier sind die Daten aus Beispiel 1 grafisch in einem *Scatterplot* dargestellt:



Es ist offensichtlich, dass die Abhängigkeit zwischen der morgendlichen Außentemperatur x und dem Fernwärmeverbrauch y keiner einfachen funktionalen Beziehung folgt, d. h. dass wir keine einfache Funktion angeben können, sodass $y = f(x)$.

Theorem 1 Die Kleinste Quadrate Schätzer für die Regressionskoeffizienten α und β sind gegeben durch

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{3}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{4}$$

Beweis. Wir bestimmen den Kleinste Quadrate Schätzer für α und β, indem wir die partiellen Ableitungen von $Q(\alpha, \beta) = \sum_{i=1}^{n}(y_i - \alpha - \beta\, x_i)^2$ nach α und β gleich 0 setzen. Dies führt uns auf das lineare Gleichungssystem

$$\sum_{i=1}^{n}(y_i - \alpha - \beta x_i) = 0$$

$$\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)\, x_i = 0.$$

Mit den Abkürzungen $\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$ und $\bar{y} := \frac{1}{n}\sum_{i=1}^{n} y_i$ folgt aus der ersten Gleichung $\alpha = \bar{y} - \beta\bar{x}$ und damit Gleichung 3. Wir setzen dies in die zweite Gleichung ein, erhalten $\sum_{i=1}^{n}(y_i - \bar{y} - \beta(x_i - \bar{x}))\, x_i = 0$ und bestimmen daraus $\hat{\beta}$,

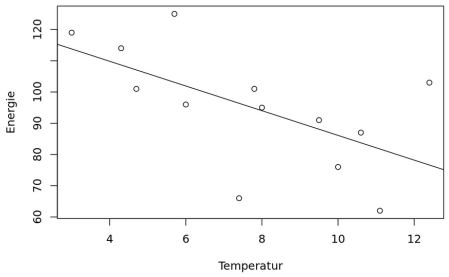
$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})x_i} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\, y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{5}$$

Für die beiden letzten Identitäten haben wir benutzt, dass $\sum(x_i - \bar{x}) = \sum(y_i - \bar{y}) = 0$.

Aufgabe 1 Berechnen Sie mithilfe der obigen Formeln aus den Daten von Beispiel 1 Schätzwerte für die Parameter der Regressionsgerade. Verwenden Sie dann die Regressionsgerade, um den Energieverbrauch bei einer morgendlichen Außentemperatur von 10 °C vorherzusagen.

Lösung zu Aufgabe 1

Einsetzen der Werte in die obigen Formeln liefert $\hat{\beta} = -3.95$, $\hat{\alpha} = 125.64$ und $s_{y|x} = 15.89$. Hier haben wir die Datenwolke zusammen mit der Kleinste-Quadrate-Regressionsgerade in einem Koordinatensystem dargestellt. Unter allen Geraden ist diese dadurch ausgezeichnet, dass sie die Summe der vertikalen Abstandsquadrate minimiert.



Mithilfe der Regressionsgerade kann man jetzt den Energieverbrauch bei beliebigen Temperaturen vorhersagen. Bei einer morgendlichen Außentemperatur von x Grad Celsius erwarten wir den Verbrauch $\widehat{Y}(x) = 125.64 - 3.95\, x$ (in kWh). Bei $x = 10$ erwarten wir zum Beispiel den Energieverbrauch $\widehat{Y}(10) = 86.14$.

3 Lineare Regression mit R

Die Schätzwerte für die Parameter α, β und σ können wir mit Hilfe des `R`-Befehls `lm` berechnen. Für die Daten aus Beispiel 1 sehen die Eingabe und die Ausgabe so aus:

```
summary(lm(Energie ~ Temperatur, data = Daten))
```

```
Call:
lm(formula = Energie ~ Temperatur, data = Daten)

Residuals:
    Min      1Q  Median      3Q     Max
-30.384  -6.058   2.917   5.361  26.381

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  125.638     13.033   9.640 1.07e-06 ***
Temperatur    -3.953      1.587  -2.492   0.0299 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.89 on 11 degrees of freedom
Multiple R-squared:  0.3608,    Adjusted R-squared:  0.3027
F-statistic: 6.208 on 1 and 11 DF,  p-value: 0.02995
```

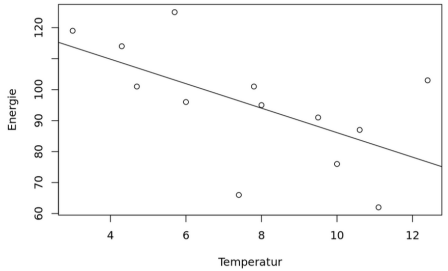
Die Schätzwerte für die Parameter α und β können im Abschnitt `Coefficients` in der Spalte `Estimate` abgelesen werden. Der Wert für α steht in der Zeile `(Intercept)`, der Wert für β steht in der Zeile `Temperatur`. Diese Zeile ist immer nach der erklärenden Variable benannt.

Eine grafische Darstellung der Daten zusammen mit der Regressionsgeraden erhält man mit den R-Befehlen `plot` und `ablines`. Zu den Daten aus Beispiel 1 erhält man mit den Befehlen

```
attach(Daten)
plot(Temperatur,Energie)
abline(lm(Energie~Temperatur))
```

einen Plot des täglichen Energieverbrauchs gegen die morgendliche Termperatur zusammen mit der von R berechneten Regressionsgeraden

$$y = 125.638 - 3.953\,x.$$



Alle Skripte sind auch in einem Quarto-Buch zusammengefasst

Die STACK-Aufgaben vertiefen das Wissen aus den Skripten

Aufgabe: Lineare Regression mit R

Frage-Tests und eingesetzte Varianten

▼ **Inhalt und Umfang dieser Aufgabe**

In dieser Aufgabe lernen Sie, wie Sie eine lineare Regression mithilfe von R ausführen. Zur Bearbeitung der Aufgabe ist R nicht notwendig. Sie besteht aus den folgenden vier Aufgabenteilen:

- In **(a)** geben Sie den R-Befehl der linearen Regression an.
- In **(b)** treffen Sie eine Vorhersage mithilfe des linearen Modells.
- In **(c)** schätzen Sie die Varianz der Residuen.
- In **(d)** geben Sie den durch das Modell erklärten Anteil der Streuung an.

In dem vorinstallierten Datensatz `faithful` befinden sich Messungen zu dem Geysir „Old Faithful" des Yellowstone-Nationalparks in den USA. Um den Einfluss der Eruptionsdauer auf die Wartezeit zu untersuchen, haben wir in den Datenvektoren `eruptions` und `waiting` die Eruptionsdauern des Geysirs und die anschließenden Wartezeiten bis zur nächsten Eruptionen *(in Minuten)* gespeichert. Für die Analyse der Abhängigkeit der Daten wird ein lineares Modell der Form

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i \in \{1, \cdots, 272\}$$

angenommen, wobei x_i die Eruptionsdauer und Y_i die Wartezeit bezeichnen. Der Zufallsterm ϵ_i wird als $N(0, \sigma^2)$-verteilt angenommen.

Hinweis: Bitte geben Sie alle Zahlenwerte auf mindestens eine Nachkommastelle genau an.

(a) Geben sie den R-Befehl für die Analyse dieses linearen Modells an. Verwenden Sie hierfür die R-Befehle `summary` und `lm`, sowie die Namen der Datenvektoren.

Prüfen

(a) Geben sie den R-Befehl für die Analyse dieses linearen Modells an. Verwenden Sie hierfür die R-Befehle `summary` und `lm`, sowie die Namen der Datenvektoren.

lm(eruptions ~ waiting)

✗ **Falsche Antwort.**

Diese Antwort ist nicht korrekt. Machen Sie sich noch einmal mit der Verwendung der R-Befehle `summary` und `lm` vertraut. Sie können in R eine lineare Regression mit Hilfe des Befehls `lm` durchführen. Hierbei muss innerhalb der Klammern angegeben werden, welche Abhängigkeitsverhältnisse wir in unserem Modell annehmen. Dabei können die Namen der Datenvektoren, sowie das Symbol `~` verwendet werden, welches die Bedeutung „in Abhängigkeit von" hat. Der Befehl `summary` kann verwendet werden, um die wichtigsten Ergebnisse einer linearen Regression auszugeben.

Bewertung für diese Einreichung: 0,00/1,00.

(a) Geben sie den R-Befehl für die Analyse dieses linearen Modells an. Verwenden Sie hierfür die R-Befehle `summary` und `lm`, sowie die Namen der Datenvektoren.

lm(waiting ~ eruptions)

✗ **Falsche Antwort.**

Diese Antwort ist nicht vollständig. Sie haben den korrekten Befehl eingegeben, um das oben beschriebene Regressionsmodell an die Daten anzupassen. Verwenden Sie den Befehl `summary` um eine Zusammenfassung der wichtigsten Parameter der Regression auszugeben.

Bewertung für diese Einreichung: 0,00/1,00.

(a) Geben sie den R-Befehl für die Analyse dieses linearen Modells an. Verwenden Sie hierfür die R-Befehle `summary` und `lm`, sowie die Namen der Datenvektoren.

summary(lm(eruptions ~ waiting))

✗ **Falsche Antwort.**

Diese Antwort ist nicht korrekt. Haben Sie die Datenvektoren in der Eingabe vertauscht? Beachten Sie welche die abhängige und welche die erklärende Variable in dem Modell ist.

Bewertung für diese Einreichung: 0,00/1,00.

(a) Geben sie den R-Befehl für die Analyse dieses linearen Modells an. Verwenden Sie hierfür die R-Befehle `summary` und `lm`, sowie die Namen der Datenvektoren.

summary(lm(waiting ~ eruptions))

✓ **Richtige Antwort, gut gemacht!**

Der Befehl, um eine lineare Regression mit R durchzuführen, lautet `lm`. Mithilfe des Befehls `summary` können Sie dann eine Zusammenfassung der Ergebnisse der Regressionsrechnung bekommen. Um die Regression durchzuführen, müssen Sie zunächst die erklärende Variable und die abhängige Variable festlegen. Der Vektor `waiting` enthält hier die Daten der erklärenden Variable und der Vektor `eruptions` die Daten der abhängigen Variable. Also lautet der vollständige R-Befehl `summary(lm(waiting~eruptions))`.

Klicken Sie auf Weiter, um zum nächsten Aufgabenteil zu gelangen.

Bewertung für diese Einreichung: 1,00/1,00.

Sie erhalten folgende Ausgabe:

```
> summary(lm(waiting~eruptions))

Call:
lm(formula = waiting ~ eruptions)

Residuals:
     Min       1Q   Median       3Q      Max
-12.0796  -4.4831   0.2122   3.9246  15.9719

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  33.4744     1.1549   28.98   <2e-16 ***
eruptions    10.7296     0.3148   34.09   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.914 on 270 degrees of freedom
Multiple R-squared:  0.8115,    Adjusted R-squared:  0.8108
F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

Hinweis: Bitte geben Sie alle Zahlenwerte auf mindestens eine Nachkommastelle genau an.

(b) Der Geysir kommt nach einer 2.5 Minuten langen Eruption zum erliegen. Wie viele Minuten beträgt die geschätzte Wartezeit bis zum nächsten Ausbruch?

Prüfen

(b) Der Geysir kommt nach einer 3 Minuten langen Eruption zum erliegen. Wie viele Minuten beträgt die geschätzte Wartezeit bis zum nächsten Ausbruch?

Prüfen

(b) Der Geysir kommt nach einer 4 Minuten langen Eruption zum erliegen. Wie viele Minuten beträgt die geschätzte Wartezeit bis zum nächsten Ausbruch?

Prüfen

Zur Evaluation wurden Fragebögen eingesetzt

Diese Aussage trifft ...	überhaupt nicht zu	eher nicht zu	eher zu	völlig zu
Die Aufgabenstellungen der digitalen Aufgaben waren verständlich.	0%	0%	61%	39%
Ich fand die Unterteilung in Zwischenschritte nach einer falschen Antwort hilfreich.	0%	7%	36%	57%
Dass ich die digitalen Aufgaben mehrfach mit anderen Werten beantworten konnte, fand ich gut.	0%	11%	46%	43%
Ich fand das Feedback hilfreich.	0%	4%	39%	57%

Außerdem wurde eine Interviewstudie durchgeführt

Studierende äußerten konstruktiv-kritische Aspekte ...

Eingabe:
„Ein Hinweis dazu, dass man hier nicht konkrete Ergebnisse eintragen muss, sondern dass auch die Rechnungsvorschrift schon genügt.“
(S9)

Aufgabenschleifen:
„Einen Zurück-Button vielleicht noch hinzufügen, um auch innerhalb der gleichen Aufgabe zurückkehren zu können, falls man sehen möchte, was man da nochmal gemacht hätte.“
(S8)

. . . und hoben positive Aspekte hervor

Randomisierung:
„Was ich auch noch gut fand, war auf jeden Fall, dass immer wenn man den Test neu gestartet hat, dass man dann auch […] andere Werte in der Stichprobe hatte“ (S10)

Aufgabenschleifen:
„Ich fand ganz gut, dass einem da nochmal so Mut gemacht wird, dass man jetzt durch die Zwischenschritte dann auf die richtige Lösung kommen kann.“ (S5)

Feedback:
„Die Hilfestellungen, die kamen, waren sehr hilfreich. Sie haben die Aufgabe gut erläutert, und da hab ich dann auch besser verstanden was zu tun war.“ (S6)

Tipps:
„Dann kam ein Tipp, dass ich vergessen habe, dass der Roboter auch wieder zurück muss. Das war ein sehr hilfreicher Tipp, weil das ja genau der Denkfehler war von mir.“ (S1)

Unsere Projektwebseite ist `oer-stochastik-nrw.de`

RUHR UNIVERSITÄT BOCHUM RUB

technische universität dortmund

Heinrich Heine Universität Düsseldorf

Universität Siegen

CHE Centrum für Hochschulentwicklung (2022). *Studierende wünschen sich auch nach der Pandemie mehr digitales Lernen*.
https://www.che.de/2022/studierende-wuenschen-sich-auch-nach-der-pandemie-mehr-digitales-lernen/

Hattie, J. (2014). *Lernen sichtbar machen. Überarbeitete deutschsprachige Ausgabe von "'Visible Learning"' besorgt von Wolfgang Beywl und Klaus Zierer* (2. korrigierte Auflage). Schneider Verlag Hohengehren.

Mai, T., Wassong, T., & Becher, S. (2021). Über das Potenzial computergestützter Aufgaben zur Mathematik am Beispiel eines auf Blended Learning basierenden Vorkurses. In R. Biehler, A. Eichler, R. Hochmuth, S. Rach, & N. Schaper (Hrsg.), *Lehrinnovationen in der Hochschulmathematik* (S. 291-320). Springer.

ORCA (2023). *Ein hochwertiger Grundstock schafft mehr Zeit für innovative Lehre*.
https://www.orca.nrw/blog/WILMO