Vorhersagemodellierung Gehalt

Horst Vozeigestudent, Matrikelnummer 007007

2021-02-10

# Einleitung

Stellen Sie hier die Forschungsfrage vor und skizzieren Sie Ihr Vorgehen.

Ungefährer Umfang: 0,5–1,5 Seiten.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

# Explorative Datenanalyse und Datenvorverarbeitung

Daten einlesen:

train <- read\_csv("https://raw.githubusercontent.com/sebastiansauer/Statistiklehre/main/data/tips/tips\_train.csv")

##   
## ── Column specification ────────────────────────────────────────────────────────  
## cols(  
## id = col\_double(),  
## total\_bill = col\_double(),  
## tip = col\_double(),  
## sex = col\_character(),  
## smoker = col\_character(),  
## day = col\_character(),  
## time = col\_character(),  
## size = col\_double()  
## )

test <- read\_csv("https://raw.githubusercontent.com/sebastiansauer/Statistiklehre/main/data/tips/tips\_test.csv")

##   
## ── Column specification ────────────────────────────────────────────────────────  
## cols(  
## id = col\_double(),  
## total\_bill = col\_double(),  
## tip = col\_double(),  
## sex = col\_character(),  
## smoker = col\_character(),  
## day = col\_character(),  
## time = col\_character(),  
## size = col\_double()  
## )

Testen, ob das Einlesen funktioniert hat:

glimpse(train)

## Rows: 122  
## Columns: 8  
## $ id <dbl> 1, 3, 5, 8, 11, 12, 13, 14, 15, 17, 18, 20, 21, 23, 25, 27…  
## $ total\_bill <dbl> 16.99, 21.01, 24.59, 26.88, 10.27, 35.26, 15.42, 18.43, 14…  
## $ tip <dbl> 1.01, 3.50, 3.61, 3.12, 1.71, 5.00, 1.57, 3.00, 3.02, 1.67…  
## $ sex <chr> "Female", "Male", "Female", "Male", "Male", "Female", "Mal…  
## $ smoker <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No", "No"…  
## $ day <chr> "Sun", "Sun", "Sun", "Sun", "Sun", "Sun", "Sun", "Sun", "S…  
## $ time <chr> "Dinner", "Dinner", "Dinner", "Dinner", "Dinner", "Dinner"…  
## $ size <dbl> 2, 3, 4, 4, 2, 4, 2, 4, 2, 3, 3, 3, 2, 2, 2, 2, 2, 2, 4, 2…

glimpse(test)

## Rows: 122  
## Columns: 8  
## $ id <dbl> 2, 4, 6, 7, 9, 10, 16, 19, 22, 24, 26, 29, 30, 36, 37, 39,…  
## $ total\_bill <dbl> 10.34, 23.68, 25.29, 8.77, 15.04, 14.78, 21.58, 16.97, 20.…  
## $ tip <dbl> 1.66, 3.31, 4.71, 2.00, 1.96, 3.23, 3.92, 3.50, 2.75, 7.58…  
## $ sex <chr> "Male", "Male", "Male", "Male", "Male", "Male", "Male", "F…  
## $ smoker <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No", "No"…  
## $ day <chr> "Sun", "Sun", "Sun", "Sun", "Sun", "Sun", "Sun", "Sun", "S…  
## $ time <chr> "Dinner", "Dinner", "Dinner", "Dinner", "Dinner", "Dinner"…  
## $ size <dbl> 3, 2, 4, 2, 2, 2, 2, 3, 2, 4, 4, 2, 2, 3, 3, 3, 3, 3, 2, 4…

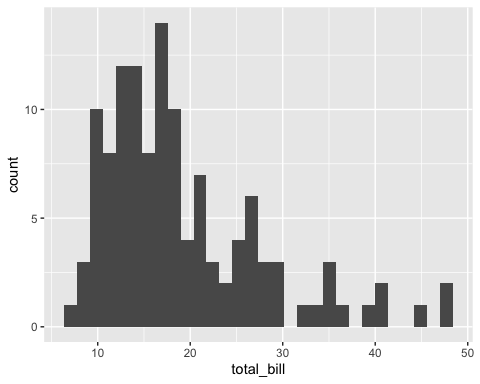
Sieht alles gut aus.

Führen Sie in diesem Abschnitt eine Explorative Datenanalyse der von Ihnen verwendeten Variablen der Trainingsdaten durch. Auch eine evtl. Datenvorverarbeitung erfolgt in diesem Abschnitt.

Bei Einzelarbeiten sollte der reine Text (ohne Code, Abbildungen etc.) einen Umfang von ca. 1–2 Seiten haben, bei Gruppenarbeiten einen von ca. 2–4 Seiten.

train %>%   
 ggplot(aes(x = total\_bill)) +   
 geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

# Analyse und Ergebnis

Wenden Sie hier Ihr Modell an und interpretieren Sie Ihr Ergebnis. Bei Einzelarbeiten sollte der reine Text (ohne Code, Abbildungen etc.) einen Umfang von ca. 1–2 Seiten haben.

Mein Lieblingsmodell bla bla bla:

lm.model <- lm(total\_bill ~ size, data = train)  
summary(lm.model)

##   
## Call:  
## lm(formula = total\_bill ~ size, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.590 -4.478 -1.335 2.420 22.380   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.6671 1.6755 2.189 0.0306 \*   
## size 6.0842 0.6193 9.825 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.506 on 120 degrees of freedom  
## Multiple R-squared: 0.4458, Adjusted R-squared: 0.4412   
## F-statistic: 96.52 on 1 and 120 DF, p-value: < 2.2e-16

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

# Vorhersage

Modell zur Vorhersage auf Testdatensatz anwenden und als neue Spalte im Test-Datensatz speichern:

tips\_test\_mit\_vorhersage <-  
 test %>%   
 mutate(pred = predict(lm.model, newdata = test))

Kontrolle, ob alles gut aussieht:

glimpse(tips\_test\_mit\_vorhersage)

## Rows: 122  
## Columns: 9  
## $ id <dbl> 2, 4, 6, 7, 9, 10, 16, 19, 22, 24, 26, 29, 30, 36, 37, 39,…  
## $ total\_bill <dbl> 10.34, 23.68, 25.29, 8.77, 15.04, 14.78, 21.58, 16.97, 20.…  
## $ tip <dbl> 1.66, 3.31, 4.71, 2.00, 1.96, 3.23, 3.92, 3.50, 2.75, 7.58…  
## $ sex <chr> "Male", "Male", "Male", "Male", "Male", "Male", "Male", "F…  
## $ smoker <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No", "No"…  
## $ day <chr> "Sun", "Sun", "Sun", "Sun", "Sun", "Sun", "Sun", "Sun", "S…  
## $ time <chr> "Dinner", "Dinner", "Dinner", "Dinner", "Dinner", "Dinner"…  
## $ size <dbl> 3, 2, 4, 2, 2, 2, 2, 3, 2, 4, 4, 2, 2, 3, 3, 3, 3, 3, 2, 4…  
## $ pred <dbl> 21.91975, 15.83555, 28.00395, 15.83555, 15.83555, 15.83555…

Prognosedaten als CSV-Datei schreiben:

tips\_test\_mit\_vorhersage %>%   
 select(id, pred) %>%   
 write\_csv("Raetsel\_Rudi\_0123456\_Prognose.csv")

# Zusammenfassung

Fassen Sie gemeinsam kurz die zentralen Ergebnisse zusammen (0,5–1 Seite). Gehen Sie auch auf die Grenzen Ihrer Analyse ein.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

# Literatur

Sie müssen keine Quellen zitieren, können aber.