

Statistik

CH.13 - Multiple Regression

SS 2022 | | Prof. Dr. Buchwitz, Sommer, Henke

Wir geben Impulse

- Ziel 1
- Ziel 2
- Ziel 3

1 Multiple Lineare Regression

2 Hypothesentests

3 Residualdiagnostik

4 Multikollinearität

5 Nichtlinearität

$$Y = f(X_1, X_2, \dots, X_p) + \epsilon$$

- Mit Hilfe der **einfachen linearen Regression** kann der Zusammenhang einer abhängigen Variablen Y mit **einer** unabhängigen Variablen X modelliert werden.
- Die **multiple lineare Regression** erlaubt das Modellieren des Zusammenhangs einer abhängigen Variablen Y mit **mehreren** unabhängigen Variablen X_1, X_2, \dots, X_p .

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon$$

- Die zuvor diskutierte einfache lineare Regression kann als **Spezialfall** der multiplen linearen Regression aufgefasst werden, bei der gilt $p = 1$.
- Wir nehmen weiterhin an, dass *innerhalb des Wertebereichs* der Daten der wahre Zusammenhang zwischen Y und den Prädiktoren durch ein lineares Modell **approximiert** werden kann.
- Jeder Regressor geht mit einem eigenen Koeffizienten $\beta_1, \beta_2, \dots, \beta_p$ in die Gleichung ein. Der Fehlerterm ϵ enthält zudem keine **systematischen Informationen** zur Erklärung der Streuung von Y , die nicht bereits durch die Regressoren abgebildet wurden.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon_i$$

$$i = 1, 2, \dots, n$$

- Aus der Modellgleichung folgt die obige Darstellung für jede Beobachtung. Dabei repräsentiert y_i die i -te Beobachtung der abhängigen Variablen Y . Die Werte $x_{i1}, x_{i2}, \dots, x_{ip}$ sind die Werte der zugehörigen Regressoren für die i -te Beobachtung in der Stichprobe (üblicherweise i -te Zeile im Datensatz).
- Der Wert ϵ_i ist der Anpassungsfehler (Fehlerterm) der linearen Approximation für die i -te Beobachtungseinheit.

Beispiel: Autodaten

d

| | 1100km | weight | hp | cyl | hub |
|------------------------|--------|--------|-----|-----|-------|
| ## Mazda RX4 | 11.201 | 1.1884 | 110 | 6 | 2.622 |
| ## Mazda RX4 Wag | 11.201 | 1.3041 | 110 | 6 | 2.622 |
| ## Datsun 710 | 10.316 | 1.0523 | 93 | 4 | 1.770 |
| ## Hornet 4 Drive | 10.991 | 1.4583 | 110 | 6 | 4.228 |
| ## Hornet Sportabout | 12.578 | 1.5604 | 175 | 8 | 5.899 |
| ## Valiant | 12.995 | 1.5694 | 105 | 6 | 3.687 |
| ## Duster 360 | 16.449 | 1.6193 | 245 | 8 | 5.899 |
| ## Merc 240D | 9.640 | 1.4470 | 62 | 4 | 2.404 |
| ## Merc 230 | 10.316 | 1.4288 | 95 | 4 | 2.307 |
| ## Merc 280 | 12.251 | 1.5604 | 123 | 6 | 2.746 |
| ## Merc 280C | 13.214 | 1.5604 | 123 | 6 | 2.746 |
| ## Merc 450SE | 14.342 | 1.8461 | 180 | 8 | 4.520 |
| ## Merc 450SL | 13.596 | 1.6919 | 180 | 8 | 4.520 |
| ## Merc 450SLC | 15.475 | 1.7146 | 180 | 8 | 4.520 |
| ## Cadillac Fleetwood | 22.617 | 2.3814 | 205 | 8 | 7.735 |
| ## Lincoln Continental | 22.617 | 2.4603 | 215 | 8 | 7.538 |
| ## Chrysler Imperial | 16.001 | 2.4244 | 230 | 8 | 7.210 |
| ## Fiat 128 | 7.260 | 0.9979 | 66 | 4 | 1.290 |
| ## Honda Civic | 7.737 | 0.7326 | 52 | 4 | 1.241 |
| ## Toyota Corolla | 6.938 | 0.8323 | 65 | 4 | 1.165 |
| ## Toyota Corona | 10.940 | 1.1181 | 97 | 4 | 1.968 |
| ## Dodge Challenger | 15.175 | 1.5966 | 150 | 8 | 5.211 |
| ## AMC Javelin | 15.475 | 1.5581 | 150 | 8 | 4.982 |
| ## Camaro Z28 | 17.685 | 1.7418 | 245 | 8 | 5.735 |
| ## Pontiac Firebird | 12.251 | 1.7441 | 175 | 8 | 6.555 |
| ## Fiat X1-9 | 8.616 | 0.8777 | 66 | 4 | 1.295 |
| ## Porsche 914-2 | 9.047 | 0.9707 | 91 | 4 | 1.971 |
| ## Lotus Europa | 7.737 | 0.6863 | 113 | 4 | 1.558 |
| ## Ford Pantera L | 14.887 | 1.4379 | 264 | 8 | 5.752 |
| ## Ferrari Dino | 11.940 | 1.2564 | 175 | 6 | 2.376 |
| ## Maserati Bora | 15.681 | 1.6193 | 335 | 8 | 4.933 |

Datenbeschreibung

1100km Kraftstoffverbrauch in Litern pro 100km bei normaler Fahrweise.

weight Fahrzeuggewicht in Tonnen.

hp Motorleistung in PS.

cyl Anzahl der Zylinder des Fahrzeugmotors.

hub Hubraum des Motors in Litern.

Beispiel: Autodaten

```
dim(d) # Anzahl Beobachtungen und Anzahl Variablen
```

```
## [1] 32 5
```

```
t(sapply(d, summary)) # Deskriptive Statistik für alle Variablen
```

| ## | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|-----------|---------|---------|---------|---------|---------|---------|
| ## l100km | 6.9385 | 10.316 | 12.251 | 12.755 | 15.250 | 22.617 |
| ## weight | 0.6863 | 1.171 | 1.508 | 1.459 | 1.637 | 2.460 |
| ## hp | 52.0000 | 96.500 | 123.000 | 146.688 | 180.000 | 335.000 |
| ## cyl | 4.0000 | 4.000 | 6.000 | 6.188 | 8.000 | 8.000 |
| ## hub | 1.1651 | 1.980 | 3.217 | 3.781 | 5.342 | 7.735 |

Beispiel: Autodaten

```
round(var(d),4)           # Varianz-Kovarianz-Matrix
```

```
##          l100km weight      hp      cyl      hub
## l100km  14.925  1.5258  202.09   5.6144   6.9033
## weight   1.526  0.1970   20.05   0.6202   0.8004
## hp       202.086 20.0454 4700.87 101.9315 110.1403
## cyl       5.614  0.6202  101.93   3.1895   3.2719
## hub       6.903  0.8004  110.14   3.2719   4.1249
```

```
round(cor(d),4)           # Paarweise Korrelationskoeffizienten
```

```
##          l100km weight      hp      cyl      hub
## l100km  1.0000  0.8899  0.7629  0.8137  0.8798
## weight  0.8899  1.0000  0.6587  0.7825  0.8880
## hp      0.7629  0.6587  1.0000  0.8324  0.7909
## cyl     0.8137  0.7825  0.8324  1.0000  0.9020
## hub     0.8798  0.8880  0.7909  0.9020  1.0000
```

Multiple Linear Regression

Wovon ist die Größe 1100km abhängig?



$$\begin{array}{rclclclclcl} Y & = & \beta_0 & + & \beta_1 X_1 & + & \beta_2 X_2 & + & \epsilon \\ \text{Kraftstoffverbrauch} & = & \beta_0 & + & \beta_1 \text{Gewicht} & + & \beta_2 \text{Motorleistung} & + & \epsilon \end{array}$$

- Wir nehmen an, dass Y linear von (mindestens) zwei erklärenden Variablen abhängig ist.
- Diese Annahme muss verifiziert werden (was wir zunächst ignorieren), da sonst nicht sichergestellt ist, dass diese Variablen einen Einfluss haben oder entscheidende Variablen im Modell fehlen.

$$\begin{array}{rclclclclcl} Y & = & \beta_0 & + & \beta_1 X_1 & + & \beta_2 X_2 & + & \epsilon \\ \text{Kraftstoffverbrauch} & = & \beta_0 & + & \beta_1 \text{Gewicht} & + & \beta_2 \text{Motorleistung} & + & \epsilon \end{array}$$

- Wir nehmen an, dass Y linear von (mindestens) zwei erklärenden Variablen abhängig ist.
- Diese Annahme muss verifiziert werden (was wir zunächst ignorieren), da sonst nicht sichergestellt ist, dass diese Variablen einen Einfluss haben oder entscheidende Variablen im Modell fehlen.

Wie können die Parameter $\beta_0, \beta_1, \dots, \beta_p$ bei der multiplen linearen Regression bestimmt werden?

Wie können die Regressionsparameter $\beta_0, \beta_1, \dots, \beta_p$ bestimmt werden?



- **Lösung:** Minimieren der Fehlerquadratsumme nach dem Prinzip der kleinsten Quadrate (Kleinste-Quadrate-Schätzung).
- Der Anpassungsfehler für jede Beobachtung ergibt sich aus der umgestellten Beobachtungsgleichung:

$$\epsilon_i = y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \dots - \beta_p x_{ip}$$

- Die zu minimierende Funktion in Abhängigkeit der Parameter $\beta_0, \beta_1, \dots, \beta_p$ ergibt sich damit wie folgt:

$$S(\beta_0, \beta_1, \dots, \beta_p) = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \dots - \beta_p x_{ip})^2$$

- $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ bzw. b_0, b_1, \dots, b_p sind die Werte, die die Funktion $S()$ minimieren.

Your Turn

Schreiben Sie die zugehörige Regressionsgleichung auf.

```
mod <- lm(l100km ~ 1 + weight + hp, data=d)
mod

##
## Call:
## lm(formula = l100km ~ 1 + weight + hp, data = d)
##
## Coefficients:
## (Intercept)      weight          hp
##      1.4831       5.9558       0.0176
```

- Multiple lineare Regressionsmodelle können in R ebenfalls mit Hilfe der Funktion `lm()` geschätzt werden.
- R greift für die Bestimmung der Parameterschätzer ebenfalls auf die Methode der kleinsten Quadrate zurück.

- Einfache Regressionsmodelle (nur X_1) können als Gerade dargestellt werden. Multiple Regressionsmodelle (X_1 und X_2) können mit einer Ebene oder als Hyperebene (mehr als zwei Prädiktoren) dargestellt werden. Diese Darstellung wird sehr schnell unübersichtlich.
- β_0 ist der Achsenabschnitt und der abgebildete Wert von Y , wenn $X_1 = X_2 = \dots = X_p = 0$.
- Die Steigungskoeffizienten β_j haben mehrere Interpretationen:
 - ▶ β_j ist die **Veränderung** in Y , wenn sich X_j um eine Einheit erhöht und alle anderen Prädiktoren konstant gehalten werden (ceteris paribus).
 - ▶ β_j wird also als **Partialeffekt** bezeichnet, weil er den Effekt von X_j auf Y abbildet, nachdem die Zielvariable um die Effekte der anderen Variablen adjustiert wurde.


```
mod <- lm(l100km ~ 1 + weight + hp, data=d)
summary(mod)

##
## Call:
## lm(formula = l100km ~ 1 + weight + hp, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.968 -1.167  0.180  0.941  3.344
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.48306    0.96288    1.54   0.134
## weight      5.95580    0.84011    7.09 8.4e-08 ***
## hp          0.01759    0.00544    3.23  0.003 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.56 on 29 degrees of freedom
## Multiple R-squared:  0.847, Adjusted R-squared:  0.837
## F-statistic: 80.3 on 2 and 29 DF,  p-value: 1.49e-12
```

Your Turn

Interpretieren Sie die Schätzergebnisse ($\alpha = 0.05$) und das Gütemaß des Regressionsmodells.

- In wissenschaftlichen Aufsätzen werden Regressionsmodelle häufig schrittweise aufgebaut und übersichtlich in Tabellen dargestellt.

| | Model 1 | Model 2 | Model 3 |
|---------------------|-------------------|-------------------|-------------------|
| (Intercept) | 1.45 (1.10) | 6.45*** (1.07) | 1.48 (0.96) |
| weight | 7.75*** (0.72) | | 5.96*** (0.84) |
| hp | | 0.04*** (0.01) | 0.02** (0.01) |
| R ² | 0.79 | 0.58 | 0.85 |
| Adj. R ² | 0.78 | 0.57 | 0.84 |
| Num. obs. | 32 | 32 | 32 |

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Table 1: Statistical models

1 Multiple Lineare Regression

2 Hypothesentests

3 Residualdiagnostik

4 Multikollinearität

5 Nichtlinearität

- Ergänzend zum t-Test für die einzelnen Koeffizienten ($H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$) gibt es auch die Möglichkeit, **alle Koeffizienten auf einmal** einem Hypothesentest zu unterziehen.
- Das Szenario, ob alle Regressoren zusammen genommen einen Effekt auf die abhängige Variable Y haben, kann mit Hilfe des **F-Tests** untersucht werden.
- Die Idee dieses simultanen Testens ist, zu prüfen, ob mit hoher Wahrscheinlichkeit davon auszugehen ist, dass nicht alle Parameter $\beta_1, \beta_2, \dots, \beta_p$ gleich 0 sind, also zu prüfen:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0 \text{ für min. ein } j$$

$$\text{FM: } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon$$

$$\text{RM: } Y = \beta_0 + \epsilon$$

- Die Nullhypothese ist gleichbedeutend mit der Tatsache, dass auch ein **reduziertes Modell** (RM) ohne Regressoren den gleichen Erklärungsgehalt liefert wie das volle Modell (FM) mit allen p Regressoren.
- Dieser fehlende **Fit** kann mit Hilfe der Fehlerquadratsumme (SSE) für die beiden Modelle messbar gemacht werden.

$$SSE(FM) = \sum (y_i - \hat{y}_i)^2 \quad SSE(RM) = \sum (y_i - \hat{y}_i^*)^2$$

$$F = \frac{[SSE(RM) - SSE(FM)] / (p + 1 - k)}{SSE(FM) / (n - p - 1)}$$

- Die Differenz $SSE(RM) - SSE(FM)$ gibt die Erhöhung der Residualstreuung durch Rückgriff auf das reduzierte Modell an. Wenn diese Differenz groß ist, ist das RM mit k Parametern **nicht adäquat**.
- Wenn der beobachtete F-Wert größer ist als der kritische Wert, ist der F-Test signifikant zum Level α .
- Das bedeutet, dass das reduzierte Modell (RM) nicht zufriedenstellend ist und die Nullhypothese (und die entsprechenden Werte für die β 's) verworfen werden kann.
- Verwerfe H_0 , wenn gilt:

$$F \geq F_{(p+1-k, n-p-1; 1-\alpha)} \quad \text{oder} \quad p(F) \leq \alpha$$

Hypothesentests

```
mod <- lm(l100km ~ 1 + weight + hp, data=d)
summary(mod)
```

```
##
## Call:
## lm(formula = l100km ~ 1 + weight + hp, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.968 -1.167  0.180  0.941  3.344
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.48306    0.96288    1.54   0.134
## weight        5.95580    0.84011    7.09 8.4e-08 ***
## hp            0.01759    0.00544    3.23  0.003 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.56 on 29 degrees of freedom
## Multiple R-squared:  0.847, Adjusted R-squared:  0.837
## F-statistic: 80.3 on 2 and 29 DF,  p-value: 1.49e-12
```

1 Multiple Lineare Regression

2 Hypothesentests

3 Residualdiagnostik

4 Multikollinearität

5 Nichtlinearität

- Modellierungsprobleme wie inkorrekt spezifizierte Modelle, fehlende und vergessene Variablen **äußern sich häufig in einer Verletzung der Annahmen der Residuen.**
- Um zu überprüfen, ob das ausgewählte Regressionsmodell den theoretischen Anforderungen genügt, müssen daher die Residuen inspiziert werden. Diesen Prozess nennt man Residualdiagnostik.
- Residuen sollten (annähernd) normalverteilt sein, keine Zusammenhangsstruktur aufweisen (i.i.d.) und frei von Ausreißern sein. Diese Eigenschaften werden häufig in **grafischen Darstellungen** der Residuen sichtbar.
- **R-Funktion:** `residuals()` erlaubt das Extrahieren von Residuen aus der Rückgabe der `lm()` Funktion.

- Darstellung Residuen der Regression Kraftstoffverbrauch Y erklärt durch Fahrzeuggewicht (X_1) und Motorleistung (X_2).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Indexplot der Residuen



Histogramm der Residuen



1 Multiple Lineare Regression

2 Hypothesentests

3 Residualdiagnostik

4 Multikollinearität

5 Nichtlinearität

- Die Interpretation der Koeffizienten eines multiplen Regressionsmodells setzt voraus, dass die Prädiktoren keinen ausgeprägten Zusammenhang untereinander haben, da die (ceteris paribus) Interpretation der Koeffizienten dann nicht mehr greift.
- Wenn eine starke Abhängigkeitsstruktur zwischen den Prädiktoren vorhanden ist, dann bezeichnet man dieses Problem als **Multikollinearität**. Multikollinearität ist ein Problem in den Daten und kein Problem der Modellierung.
- Multikollinearität führt zu unplausiblen Werten der Koeffizientenschätzer und wird durch spezielle Maßzahlen wie Varianzinflationsfaktoren (VIF) messbar.

```
cor(d)
```

```
##          l100km weight      hp    cyl    hub
## l100km 1.0000 0.8899 0.7629 0.8137 0.8798
## weight 0.8899 1.0000 0.6587 0.7825 0.8880
## hp      0.7629 0.6587 1.0000 0.8324 0.7909
## cyl     0.8137 0.7825 0.8324 1.0000 0.9020
## hub     0.8798 0.8880 0.7909 0.9020 1.0000
```

- Um Multikollinearitätsprobleme zu diagnostizieren, müssen die Zusammenhangsstrukturen zwischen den Prädiktoren untersucht werden. Das beinhaltet die Analyse des R^2 , das aus der Regression jedes Prädiktors auf alle verbleibenden Prädiktoren resultiert.

$$\text{VIF}_j = \frac{1}{1 - R_j^2} \quad \text{mit } j = 1, \dots, p$$

- R_j^2 bezeichnet das Bestimmtheitsmaß bei der Erklärung von X_j durch alle verbleibenden $p - 1$ Prädiktoren. Wenn X_j gut durch die anderen Variablen erklärt werden kann, wird das R_j^2 nah bei 1 sein und in einem großen Wert des VIF_j resultieren.

Ein Wert von $\text{VIF} > 10$ wird oft als Grenzwert gesehen, ab dem man von Multikollinearität in problematischem Ausmaß ausgeht.

- Varianzinflationsfaktoren können mit der **R-Funktion** `vif()` aus dem Zusatzpaket `car` berechnet werden.

```
mod1 <- lm(l100km ~ 1 + weight + hp, data=d)
car::vif(mod1)
```

```
## weight      hp
##  1.767  1.767
```

```
mod2 <- lm(l100km ~ 1 + weight + hp + hub + cyl, data=d)
car::vif(mod2)
```

```
## weight      hp      hub      cyl
##  4.848  3.406 10.373  6.738
```

1 Multiple Lineare Regression

2 Hypothesentests

3 Residualdiagnostik

4 Multikollinearität

5 Nichtlinearität

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \epsilon$$

- Die lineare Regression ist linear in Bezug auf die Tatsache, dass die Parameter $\beta_0, \beta_1, \dots, \beta_p$ **linear** in das Modell eingehen.
- Mit der linearen Regression können dennoch **nicht-lineare** Zusammenhänge modelliert werden, indem **nicht-lineare Transformationen** als zusätzliche unabhängige Variablen in das Modell integriert werden.

Welches Modell passt besser zu den gezeigten Daten?



```
mod1
```

```
##  
## Call:  
## lm(formula = y ~ 1 + x)  
##  
## Coefficients:  
## (Intercept)          x  
##          5.33          6.78
```

```
mod2
```

```
##  
## Call:  
## lm(formula = y ~ 1 + x + x_sq)  
##  
## Coefficients:  
## (Intercept)          x          x_sq  
##          25.58          1.44          0.25
```

- Wie ist das Bestimmtheitsmaß R^2 bei der multiplen Regression zu interpretieren?
- Was ist damit gemeint, dass die diskutierten Regressionsmodelle lineare Modelle sind?
- Was ist Multikollinearität?