Fachhochschule Südwestfalen
University of Applied Sciences

Statistik

CH.11 - Regression

SS 2021 || Prof. Dr. Buchwitz, Sommer, Henke

Wir geben Impulse

Lineare Regression

- **Ziel:** Erkennen von Abhängigkeiten und Zusammenhängen zwischen mehrere Merkmalen und Modellierung der Effektgrößen der Zusammenhänge.
- **Beispiele:**
  - Umsatz und Werbeetat einer Supermarktkette: *Hängt der Umsatz von den eingesetzten Werbemitteln ab?*
  - Körpergröße und Gewicht von Personen: *Ist das Gewicht einer Person von dessen Körpergröße abhängig?*
  - Benzinpreis und Mineralölpreis: *Ist der deutsche Benzinpreis eine Funktion des globalen Mineralölpreises?*

Sind die beiden gezeigten Größen voneinander Abhängig?

```
##               x        y
##  [1,]  5.310173 32.24161
##  [2,]  7.442478 35.41568
##  [3,] 11.457067 41.15682
##  [4,] 18.164156 52.87679
##  [5,]  4.033639 28.82066
##  [6,] 17.967794 53.04090
##  [7,] 18.893505 50.24641
##  [8,] 13.215956 44.26326
##  [9,] 12.582281 44.81691
## [10,]  1.235725 27.18693
## [11,]  4.119491 33.18971
## [12,]  3.531135 30.15394
## [13,] 13.740457 46.86971
## [14,]  7.682074 39.23753
## [15,] 15.396828 46.36786
## [16,]  9.953985 36.71948
## [17,] 14.352370 46.64537
## [18,] 19.838122 54.16792
## [19,]  7.600704 35.04383
## [20,] 15.548904 47.24008
```

Datenbeschreibung

- x Berufserfahrung in Jahren
- y Gehalt in Euro

Sind die beiden gezeigten Größen voneinander Abhängig?

Lineare Regression

$$Y = f(X) + \epsilon$$

- Wir betrachten zunächst den einfachen Fall, bei dem die abhängige Variable Y durch **eine** unabhängige Variable X erklärt wird. Unabhängige Variablen bezeichnet man auch als Regressoren oder Prädiktoren.
- ϵ bezeichnet den Anpassungsfehler und wird Fehlerterm oder Residuum genannt.
- Wir verzichten auf die vollständige Herleitung der gezeigten Formeln und fokussieren auf die zugrundeliegenden Mechanismen und die zugehörige Intuition.

Kovarianz

Salary vs. Experience
Salary in Euro (Y)
$\bar{y}$
$\bar{x}$
Experience in Years (X)

Kovarianz

Aufgabe: Bestimmen des Vorzeichens

- $y_i - \bar{y}$ ist die Differenz jeder Beobachtung y_i vom arithmetischen Mittel der abhängigen Variablen
- $x_i - \bar{x}$ ist die Abweichung x_i vom arithmetischen Mittel des Prädiktors
- $(y_i - \bar{y})(x_i - \bar{x})$ ist das Produkt der vorherigen beiden Größen

Quadrant	$y_i - \bar{y}$	$x_i - \bar{x}$	$(y_i - \bar{y})(x_i - \bar{x})$
1 (top right)			
2 (top left)			
3 (bottom left			
4 (bottom right)			

Covariance

Positive Relationship

- Wenn der Zusammenhang zwischen Y und X **positiv** ist (also wenn X größer wird, dann wird auch Y größer), dann sind mehr Datenpunkte im ersten und dritten Quadranten als im zweiten und vierten.
- Die Summe der Elemente in der letzten Spalte der vorherigen Tabelle ist dann mit großer Wahrscheinlichkeit positiv, also $Cov(Y, X) > 0$.

Covariance

Positive Relationship

- Wenn der Zusammenhang zwischen Y und X **positiv** ist (also wenn X größer wird, dann wird auch Y größer), dann sind mehr Datenpunkte im ersten und dritten Quadranten als im zweiten und vierten.
- Die Summe der Elemente in der letzten Spalte der vorherigen Tabelle ist dann mit großer Wahrscheinlichkeit positiv, also $Cov(Y, X) > 0$.

Negative Relationship

- Wenn der lineare Zusammenhang zwischen Y und X **negativ** ist (also wenn X sinkt, sinkt auch Y), dann befinden sich mehr Datenpunkte im zweiten und vierten Quadranten als im ersten und dritten.
- Die Summe der Elemente in der letzten Spalte der vorherigen Tabelle ist dann mit großer Wahrscheinlichkeit negativ, also $Cov(Y, X) < 0$.

Kovarianz

$$Cov(X, Y) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})$$

- Die aufwändig berechnete Größe ist die Kovarianz zwischen Y und X.
- Das Vorzeichen der Kovarianz gibt die Richtung des Zusammenhangs zwischen Y und X an.
- Die Kovarianz gibt **nur die Richtung des Zusammenhangs an** und erlaubt keine Beurteilung der Stärke dieses Zusammenhangs.
- Die Kovarianz verändert sich mit Veränderungen der Einheit der Daten (z.B. von Euro in TEuro).

Your turn

Wie ändert sich die Kovarianz, wenn Sie $Cov(X, Y)$ anstelle von $Cov(Y, X)$ berechnen?

Korrelationskoeffizient

$$Cor(Y, X) = \frac{1}{n-1} \sum_{i=1}^{n} (\frac{y_i - \bar{y}}{s_y})(\frac{x_i - \bar{x}}{s_x}) = \frac{Cov(Y, X)}{s_y s_x}$$

- $Cor(Y, X)$ kann auf zwei Arten interpretiert werden:
  - als Kovarianz der z-Standardisierten Variablen X und Y.
  - als Verhältnis von Kovarianz zum Produkt der Standardweichungen der Variablen.
- Im Gegensatz zur Kovarianz ist $Cor(Y, X)$ skaleninvariant mit einem WErtebereich von $-1 \geq Cor(Y, X) \leq 1$ und erlaubt daher die Beurteilung von **Richtung** und **Stärke** des Zusammenhangs.

Kovarianz und Korrelation

```r
cov(y, x)
```

```
## [1] 49.66758
```

```r
cor(y, x)
```

```
## [1] 0.9810286
```

Verwendung des Zusammenhangs

Kovarianz und Korrelationskoeffizient können nicht für Vorhersagen (X gegeben und Y gesucht) verwendet werden!

Lineare Regression

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Regressionsanalyse ist eine Erweiterung der Korrelationsanalyse und erlaubt es den Zusammehang zwischen abhängiger und unabhängigen Variablen numerisch zu beschreiben.
- β_0 und β_1 sind konstanten die als **Regressionskoeffizienten** bezeichnet werden, ϵ ist der Fehlerterm
  - β_0 ist der Achsenabschnitt und ist der vorhergesagte Wert, wenn $X = 0$.
  - β_1 ist die Steigung und kann interpretiert werden als Änderung in Y, wenn X sich um eine Einheit erhöht.

Wie bestimmen wir Werte für β_0 und β_1?

Parameterschätzung

Paramterschätzung

Residuen: Wieso ist die eingezeichnete Gereade optimal?

Residuen: Wieso ist die eingezeichnete Gereade optimal?

Minimieren: $$S(\beta_0, \beta_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 \cdot x_i))^2$$

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 \cdot x_i))^2$$

- Die quadratische Funktion $S(\beta_0, \beta_1)$ muss minimiert werden und liefert dann die Lösung $\hat{\beta}_0$ und $\hat{\beta}_1$. Diese Werte werden zuweilen auch mit b_0 und b_1 bezeichnet.
- Die Werte $\hat{\beta}_0 = b_0$ und $\hat{\beta}_1 = b_1$ werden **Kleinste-Quadrate-Schätzer** (Ordinary Least Squares Estimates, OLS Estimates) genannt und spezifizieren die Gerade mit der kleinsten möglichen Summe der quadrierten vertikalen Distanzen zu den Beobachtungen.

Lineare Regression

- Die mit der Methode der kleinsten Quadrate bestimmten Regresssionslinie existiert immer und ist gegeben durch:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

- Mit Hilfe der Beobachtungsgleichung können die angepassten Werte (fitted Values) berechnet werden:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad \text{for} \quad i = 1, 2, \ldots, n$$

- Jeder Punkt $(x_i, \hat{y}_i)$ **liegt auf der Regressionsgerade**.
- Die zugehörigen Residuen (Ordinary Least Squares Residuals) geben die vertikale Distanz zwischen Beobachtung und Gerade (Anpassungsfehler) an und können wie folgt berechnet werden:

$$\hat{\epsilon}_i = y_i - \hat{y}_i \quad \text{for} \quad i = 1, 2, \ldots, n$$

Analytische Lösung

- Für die Lösung des Minimierungsproblems gibt es eine analytische Lösung:

$$\hat{\beta}_1 = b_1 = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})} \quad \text{und} \quad \hat{\beta}_0 = b_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Herleitung der Formeln:

- Minimierung der quadratischen Funktion $S(\beta_0, \beta_1)$ mit Hilfe der Differentialrechnung
- Bildung der pratiellen Ableitungen nach b_0 and b_1
- Setzen der Ableitungen = 0
- Lösen des resultierenden Gelichungssystems
- Die gezeigten Fomeln sind die erhaltene Lösung

Lineare Regression

```r
summary(x) # Experience in Years
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.236   5.310  11.457  10.636  15.397  19.838
```

```r
summary(y) # Salary in Euro
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   27.19   33.18   41.16   40.84   47.24   54.17
```

```r
cor(y,x)
```

```
## [1] 0.9810286
```

```r
lm(y ~ 1 + x)
```

```
##
## Call:
## lm(formula = y ~ 1 + x)
##
## Coefficients:
## (Intercept)            x
##      25.601        1.433
```

Hypothesentests

- Die bestimmte Gerade beschreibt die Daten der **Stichprobe**. Interessant ist jedoch die Frage ob der Zusammenhang auch verallgemeinert werden und für die Grundgesamtheit angenommen werden kann.
- Prüfen der Hypothese $\beta_1 = 0$ is equivalent zur Aussage, dass **kein linearer Zusammenhang** vorhanden ist.
- Sollte $\beta_1 > 0$ oder $\beta_1 < 0$ gelten (Annahme der entsprechenden Alternativhypothese) liefert **Evidenz** (keinen Beweis) für die Existenz eines linearen Zusammenhangs.

Hypothesentests

- Unter der Annahme, dass die Residuen **unabhängig und gleich verteilt** (i.i.d.) sind ($\epsilon \sim N(0, \sigma^2)$), kann die Residualvarianz σ^2 geschätzt werden.

$$\hat{\sigma}^2 = \frac{\sum \epsilon_i^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{SSE}{n-2}$$

- Mit Hilfe der geschätzten Residualvarianz $\hat{\sigma}^2$ kann der Standardfehler (s.e.) der Regressionsparamter geschätzt werden.

$$s.e.(\hat{\beta}_0) = \hat{\sigma}^2 \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}} \quad \text{and} \quad s.e.(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sqrt{\sum (x_i - \bar{x})^2}}$$

Hypothesentests

- Unter der Annahme der Normalverteilung kann der t-Tests für die Regressionskoeffizienten durchgeführt werden:

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_1 : \beta_1 \neq 0$$

- Die Teststatistik t folgt einer t-Verteilung mit n-2 Freiheitsgraden. Ergänzend muss noch eine Irrtumswahrscheinlichkeit α für den Test festgelegt werden.

$$t_1 = \frac{\hat{\beta}_1}{s.e.(\hat{\beta}_1)}$$

- Die Nullhypothese $\beta_1 = 0$ kann für eine gegebene Irrtumswahrscheinlichkeit α verworfen werden, wenn gilt:

$$|\, t \,| \quad \geq \quad t_{(n-2,1-\alpha/2)}$$

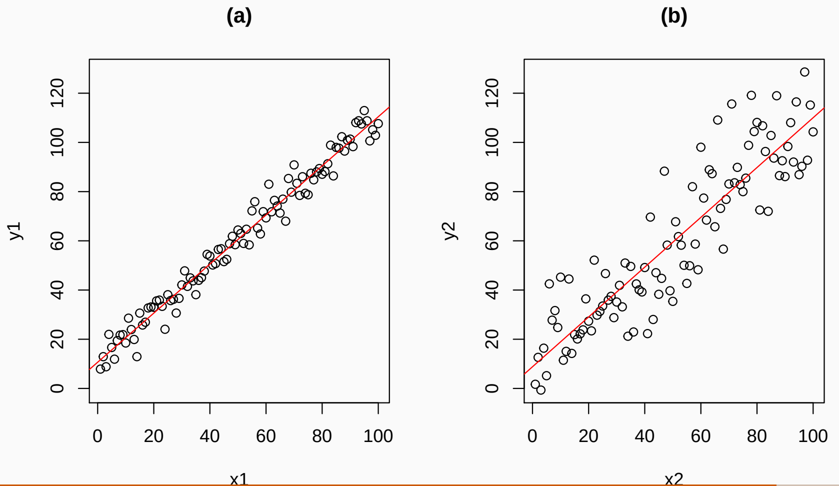
Lineare Regression

```r
summary(lm(y ~ 1 + x))
```

```
##
## Call:
## lm(formula = y ~ 1 + x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1407 -0.9661 -0.2699  1.5024  3.1583
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.60062    0.71422   35.84   <2e-16 ***
## x            1.43255    0.05903   24.27   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.703 on 23 degrees of freedom
## Multiple R-squared:  0.9624, Adjusted R-squared:  0.9608
## F-statistic:   589 on 1 and 23 DF,  p-value: < 2.2e-16
```

Anpassungsgüte

Welche Gerade hat eine höhere Anpassungsgüte und bildet daher den Sachverhalt in den Daten präziser ab?



Anpassungsgüte

Definition von Streugrößen:

$$SST = \sum (y_i - \bar{y})^2 \qquad SSR = \sum (\hat{y}_i - \bar{y})^2 \qquad SSE = \sum (y_i - \hat{y}_i)^2$$

- Sum of Squares Total (SST) ist die gesamte Abweichung von Y vom zugehörigen arithmetischem Mittel $\bar{y}$.
- Sum of Squares Regression (SSR) ist die erklärte Variation, die durch die Regressionsgerade abgebildet werden kann
- Sum of Squares Error (SSE) ist die unerklärte Streuung und die Varianz der Residuen.

Bestimmtheitsmaß

- **SSR**, misst die Qualität von X als Prädiktor für Y
- **SSE**, misst den Fehler in dieser Prädiktion
- Das Verhältnis $R^2 = SSR/SST$ ist der Anteil der durch X erklärten Varianz and der totalen Varianz. Zur Beurteilung der Anpassungsgüte einer Regressionsgerade kann entsprechend das Bestimmtheitsmaß R^2 herangezogen werden.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = [Cor(Y, \hat{Y})]^2$$

- Es gilt $0 \leq R^2 \leq 1$ und je näher R^2 an 1 liegt, desto intensiver ausgeprägt ist der lineare Zusammenhang.

Im Fall von nur einem einzigen Prädiktor gilt zudem $[Cor(Y, X)]^2$!

Lineare Regression

```r
summary(lm(y ~ 1 + x))
```

```
##
## Call:
## lm(formula = y ~ 1 + x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1407 -0.9661 -0.2699  1.5024  3.1583
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.60062    0.71422   35.84   <2e-16 ***
## x            1.43255    0.05903   24.27   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.703 on 23 degrees of freedom
## Multiple R-squared:  0.9624, Adjusted R-squared:  0.9608
## F-statistic:   589 on 1 and 23 DF,  p-value: < 2.2e-16
```

Verständnisfragen

- Wozu wird die Methoden der linearen Regression verwendet?
- Was ist die zugrundeliegende Methodik zu bestimmung der Parameter der Regressionsgerade?
- Wozu dient das Bestimmtheitsmaß R^2.