

# Statistik

CH.12 - Regression

2024 || Prof. Dr. Buchwitz, Sommer, Henke

Wir geben Impulse

- Nachvollziehen der Grundideen des linearen Modells und verinnerlichen der Existenz einer Wirkungsrichtung in der Modellierung.
- Verstehen der Idee der kleinsten Quadrate als zentrale Optimierungsgröße.
- Verknüpfung der inferenzstatistischen vorgehensweise mit der linearen Regression.

- **Ziel:** Erkennen von Abhängigkeiten und Zusammenhängen zwischen mehreren 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,]	5.310	32.24
##	[2,]	7.442	35.42
##	[3,]	11.457	41.16
##	[4,]	18.164	52.88
##	[5,]	4.034	28.82
##	[6,]	17.968	53.04
##	[7,]	18.894	50.25
##	[8,]	13.216	44.26
##	[9,]	12.582	44.82
##	[10,]	1.236	27.19
##	[11,]	4.119	33.19
##	[12,]	3.531	30.15
##	[13,]	13.740	46.87
##	[14,]	7.682	39.24
##	[15,]	15.397	46.37
##	[16,]	9.954	36.72
##	[17,]	14.352	46.65
##	[18,]	19.838	54.17
##	[19,]	7.601	35.04
##	[20,]	15.549	47.24
##	[21,]	18.694	51.42
##	[22,]	4.243	33.18
##	[23,]	13.033	47.43
##	[24,]	2.511	31.25
##	[25,]	5.344	31.94

## Datenbeschreibung

x Berufserfahrung in Jahren

y Gehalt in Tausend Euro

## Sind die beiden gezeigten Größen voneinander abhängig?

Salary vs. Experience



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

- 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.
- $\epsilon$  bezeichnet den Anpassungsfehler und wird Fehlerterm oder Residuum genannt.
- Wir verzichten auf die vollständige Herleitung der gezeigten Formeln und fokussieren uns auf die zugrundeliegenden Mechanismen und die zugehörige Intuition.

Salary vs. Experience



## 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 (oben rechts)			
2 (oben links)			
3 (unten links)			
4 (unten rechts)			



## Positiver Zusammenhang

- 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  $\text{Cov}(Y, X) > 0$ .

## Positiver Zusammenhang

- 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  $\text{Cov}(Y, X) > 0$ .

## Negativer Zusammenhang

- Wenn der lineare Zusammenhang zwischen  $Y$  und  $X$  **negativ** ist (z.B. wenn  $X$  sinkt, steigt  $Y$ ), 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  $\text{Cov}(Y, X) < 0$ .

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

- 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  $\text{Cov}(X, Y)$  anstelle von  $\text{Cov}(Y, X)$  berechnen?

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

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

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

```
## [1] 49.67
```

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

```
## [1] 0.981
```

## Verwendung des Zusammenhangs

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

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

- Regressionsanalyse ist eine Erweiterung der Korrelationsanalyse und erlaubt es, den Zusammenhang zwischen abhängiger und unabhängigen Variablen numerisch zu beschreiben.
- $\beta_0$  und  $\beta_1$  sind Konstanten, die als **Regressionskoeffizienten** bezeichnet werden,  $\epsilon$  ist der Fehlerterm
  - ▶  $\beta_0$  ist der Achsenabschnitt und ist der vorhergesagte Wert, wenn  $X = 0$ .
  - ▶  $\beta_1$  ist die Steigung und kann interpretiert werden als Änderung in  $Y$ , wenn  $X$  sich um eine Einheit erhöht.

Wie bestimmen wir  
Werte für  $\beta_0$  und  $\beta_1$ ?

Salary vs. Experience





Salary vs. Experience



## Residuen: Wieso ist die eingezeichnete Gerade optimal?

Salary vs. Experience



# Residuen: Wieso ist die eingezeichnete Gerade optimal?

Salary vs. Experience



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

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

- 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.

- Die mit der Methode der kleinsten Quadrate bestimmten Regressionslinie 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:

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

- 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:

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

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

$$\hat{\beta}_1 = b_1 = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \quad \text{und} \quad \hat{\beta}_0 = b_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

## Herleitung der Formeln:

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

# Lineare Regression

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

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      1.24   5.31   11.46   10.64   15.40   19.84
```

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

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      27.2   33.2   41.2   40.8   47.2   54.2
```

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

```
## [1] 0.981
```

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

```
##
## Call:
## lm(formula = y ~ 1 + x)
##
## Coefficients:
## (Intercept)          x
##      25.60       1.43
```

- 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  $\beta_1 = 0$  ist äquivalent 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.



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

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

- Unter der Annahme der Normalverteilung kann der  $t$ -Test 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  $\alpha$  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  $\alpha$  verworfen werden, wenn gilt:

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

# Lineare Regression

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

```
##
## Call:
## lm(formula = y ~ 1 + x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.141 -0.966 -0.270  1.502  3.158
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   25.601      0.714    35.8   <2e-16 ***
## x              1.433      0.059    24.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.7 on 23 degrees of freedom
## Multiple R-squared:  0.962, Adjusted R-squared:  0.961
## F-statistic: 589 on 1 and 23 DF,  p-value: <2e-16
```

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

(a)



(b)



## Definition von Streuungsgrößen:

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

- 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.

- **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 an der totalen Varianz. Zur Beurteilung der Anpassungsgüte einer Regressionsgerade kann entsprechend das Bestimmtheitsmaß  $R^2$  herangezogen werden.

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

- 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  $[\text{Cor}(Y, X)]^2$ !

# Lineare Regression

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

```
##
## Call:
## lm(formula = y ~ 1 + x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.141 -0.966 -0.270  1.502  3.158
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   25.601      0.714    35.8   <2e-16 ***
## x              1.433      0.059    24.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.7 on 23 degrees of freedom
## Multiple R-squared:  0.962, Adjusted R-squared:  0.961
## F-statistic: 589 on 1 and 23 DF, p-value: <2e-16
```

- Wozu wird die Methode der linearen Regression verwendet?
- Was ist die zugrundeliegende Methodik zur Bestimmung der Parameter der Regressionsgeraden?
- Wozu dient das Bestimmtheitsmaß  $R^2$ ?