3.120 Methoden: Statistik

Übungsblatt 6: Regressionsanalyse

Mathis Mörke
Michael Schürle
und Alois Weigand

Universität St.Gallen (HSG)

Herbstsemester 2024

Aufgabe 1

5 Verkäufer vergleichen ihren Erfolg. Sie wollen den Zusammenhang zwischen der Anzahl von Kundenanrufen (Variable X) und der Anzahl monatlicher Vertragsabschlüsse (Variable Y) berechnen. Hierzu ist folgende Tabelle gegeben:

Verkäufer	Anzahl monatlicher Kundenanrufe (X)	Anzahl monatlicher Vertragsabschlüsse (Y)
1	14	10
2	17	9
3	9	8
4	21	19
5	19	16

1. Bestimmen Sie die Kovarianz. Mit welchem Verfahren kann das Ergebnis normalisiert werden, um eine aussagekräftigere Vergleichbarkeit zu gewährleisten?

Refresher: Kovarianz und Korrelationskoeffizient

- Kovarianz der Stichprobe zwischen den Variablen X und Y:
  - aus n Wertepaaren (x_i, y_i) berechnete Grösse

    $cov_{XY} = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \overline{x})(y_j - \overline{y})$
  - arithmetisches Mittel des Produkts der Abweichungen der einzelnen Beobachtungswerte vom jeweiligen Mittel
  - oder umgeformt:

    $$cov_{XY} = \frac{1}{n-1} (\sum_{j=1}^{n} x_j y_j - n\bar{x}\bar{y})$$

Refresher: Kovarianz und Korrelationskoeffizient

- Kovarianz zwischen den Variablen X und Y:
  - Kovarianz lässt **keine Aussage über kausale Zusammenhänge** zu!
  - Kovarianz bestimmt positive oder negative lineare Zusammenhänge zwischen den Werten x_i und y_i
  - Wir können aber nicht bestimmen, wie stark dieser Zusammenhang ist.
  - X und Y sind unabhängig $\Rightarrow cov_{XY} = 0$
  - **Aber:** Aus $cov_{XY} = 0$ folgt nicht zwingend stochastische Unabhängigkeit

Refresher: Kovarianz und Korrelationskoeffizient

- Kovarianz zwischen den Variablen X und Y
  - Wir können nicht bestimmen, wie stark dieser Zusammenhang ist.

- Normalisierung der Kovarianz
  - Bravais-Pearson Korrelationskoeffizient $\rho_{XY} = \frac{cov_{XY}}{s_X \cdot s_Y}$ mit den Standardabweichungen s_X und s_Y
    $\Rightarrow$ nimmt Werte zwischen -1 und $+1$ an
    $\Rightarrow$ kann entsprechend seiner Stärke interpretiert werden

Aufgabe 1

Verkäufer	x_j	y_j	$x_j y_j$	x_j^2	$x_j - \bar{x}$	$(x_j - \bar{x})^2$	$y_j - \bar{y}$	$(y_j - \bar{y})^2$
1	14	10	140	196	-2	4	-2.4	5.76
2	17	9	153	289	1	1	-3.4	11.56
3	9	8	72	81	-7	49	-4.4	19.36
4	21	19	399	441	5	25	6.6	43.56
5	19	16	304	361	3	9	3.6	12.96
$\sum$	80	62	1068	1368		88		93.2

- Berechnen Sie die Kovarianz:
  - Mittelwert $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j = \frac{80}{5} = 16$
  - Mittelwert $\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j = \frac{62}{5} = 12.4$
  - Kovarianz:
    $$c_{XY} = \frac{1}{n-1}\left(\sum_{j=1}^{n} x_j y_j - n\bar{x}\bar{y}\right) = \frac{1}{4} \cdot 1068 - 5 \cdot 16 \cdot 12.4 = 19.0$$

Aufgabe 1

Verkäufer	x_j	y_j	$x_j y_j$	x_j^2	$x_j - \bar{x}$	$(x_j - \bar{x})^2$	$y_j - \bar{y}$	$(y_j - \bar{y})^2$
1	14	10	140	196	-2	4	-2.4	5.76
2	17	9	153	289	1	1	-3.4	11.56
3	9	8	72	81	-7	49	-4.4	19.36
4	21	19	399	441	5	25	6.6	43.56
5	19	16	304	361	3	9	3.6	12.96
$\sum$	80	62	1068	1368		88		93.2

- Mit welchem Verfahren kann das Ergebnis normalisiert werden, um eine aussagekräftigere Vergleichbarkeit zu gewährleisten?

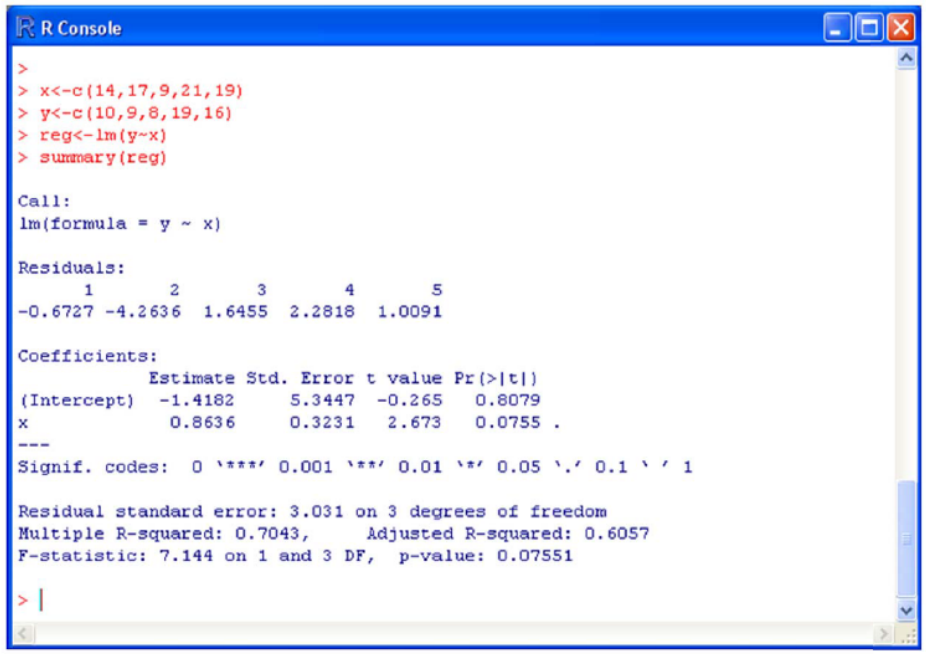
  $\Rightarrow$ Bravais-Pearson Korrelationskoeffizient

  - Kovarianz: $cov_{XY} = \frac{1}{4} \cdot 1068 - 5 \cdot 16 \cdot 12.4 = 19.0$
  - Varianz von X und Y:
    $s_x^2 = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x})^2 = \frac{1}{4} \times 88 = 22.0 \Rightarrow s_x = \sqrt{s_x^2} = 4.690$

    $s_y^2 = \frac{1}{n-1} \sum_{j=1}^{n} (y_j - \bar{y})^2 = \frac{1}{4} \times 93.2 = 23.3 \Rightarrow s_y = 4.827$
  - Korrelationskoeffizient $\rho_{XY} = \frac{c_{XY}}{s_X \cdot s_Y} = \frac{19.0}{4.690 \cdot 4.827} = 0.839$

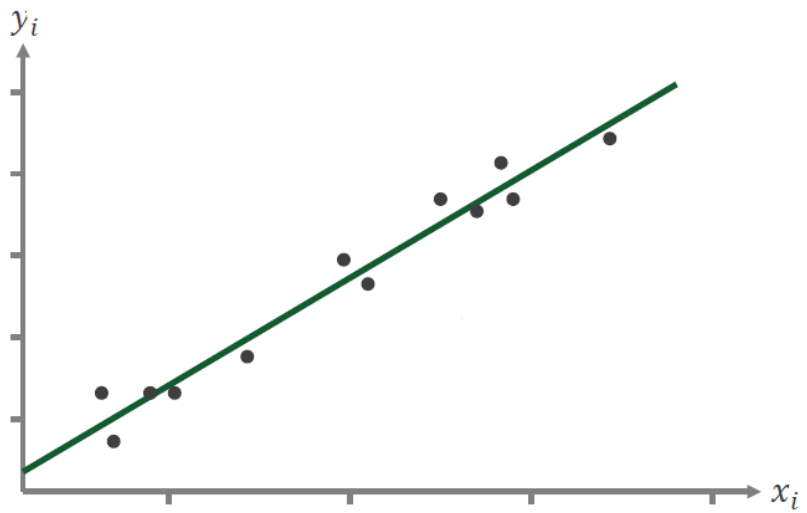
Aufgabe 1

2. Nun soll folgendes Regressionsmodell geschätzt werden: $\textit{Vertragsabschluss} = \alpha + \beta \times \textit{Kundenanruf} + \varepsilon$ Sie erhalten folgendes Ergebnis mit der Statistiksoftware R. Zeigen Sie, dass die folgende Werte richtig sind: Die geschätzten Koeffizienten, der Standardfehler des Koeffizienten $\hat{\beta}$, t-Statistik des Koeffizienten $\hat{\beta}$, und der p-Wert des Koeffizienten $\hat{\beta}$. Interpretieren Sie die Werte.
Beachten Sie, dass der Prozentualwert von 2.6729 in der t-Verteilung mit 3 Freiheitsgraden in etwa 0.96225 entspricht.



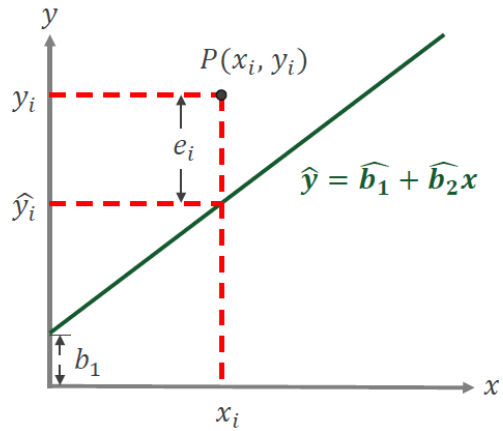
Refresher: Regressionsanalyse

- Was ist der lineare Zusammenhang zwischen X und Y in der Grundgesamtheit?
- **Idee:** Schätze eine Regressionsgerade mit Hilfe einer Stichprobe $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$



Refresher: Regressionsanalyse

- Finde $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$



$y_i = b_1 + b_2 x_i + e_i$ für alle i = 1, …, n

$\hat{e}_i = y_i - \hat{y}_i$

- $\hat{y}_i = \beta_0 + \beta_1 x_i$ mit $\varepsilon_i = y_i - \hat{y}_i$ als Abweichung des tatsächlichen Wert y_i vom best-fit $\hat{y}_i$

- Idee: Finde $\hat{\beta}_0$ und $\hat{\beta}_1 \Rightarrow$ Minimiere die Summe der quadrierten Abweichungen! $\Rightarrow \varepsilon_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$
- $Min_{\beta_0,\beta_1} \sum_{i=1}^{n} (\varepsilon_i)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

Refresher: Regressionsanalyse

- Intuition: Nach β_0 und β_1 ableiten:
  $Min_{\beta_0,\beta_1} \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
  nach β_0: $2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1) = 0$
  nach β_1: $2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i) = 0$

- $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$ und $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(x_i) = 0$

- $\sum_{i=1}^{n} y_i - \hat{\beta}_0 n - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0$
  $\frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}_0 - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i = 0$
  $\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i$
  $\Rightarrow$ y-Achsenabschnitt $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$

Refresher: Regressionsanalyse

- $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$ in $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(x_i) = 0$ einsetzen

- $\sum_{i=1}^{n}(y_i - \overline{y} + \hat{\beta}_1 \overline{x} - \hat{\beta}_1 x_i)x_i = 0$
  $\sum_{i=1}^{n} y_i x_i - \overline{y}\sum_{i=1}^{n} x_i + \hat{\beta}_1 \overline{x}\sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i x_i = 0$
  $\sum_{i=1}^{n} y_i x_i - \overline{y}\sum_{i=1}^{n} x_i + \hat{\beta}_1 \overline{x}\sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$
  $\frac{1}{n}\sum_{i=1}^{n} y_i x_i - \overline{y}\frac{1}{n}\sum_{i=1}^{n} x_i + \hat{\beta}_1 \overline{x}\frac{1}{n}\sum_{i=1}^{n} x_i - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i^2 = 0$
  $\frac{1}{n}\sum_{i=1}^{n} y_i x_i - \overline{yx} + \hat{\beta}_1 \overline{xx} - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i^2 = 0$

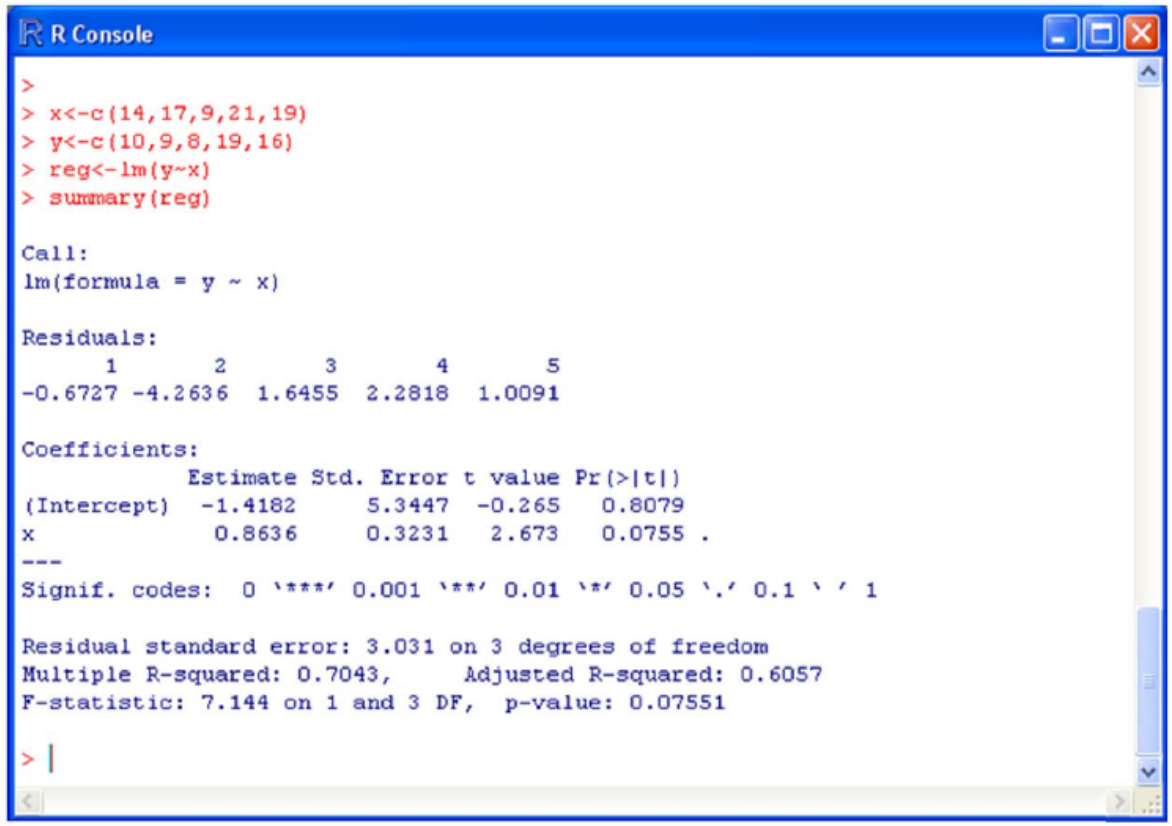
- $\hat{\beta}_1(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2) = \frac{1}{n}\sum_{i=1}^{n} y_i x_i - \overline{yx}$

- $\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} y_i x_i - \overline{yx}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2} \Rightarrow \hat{\beta}_1 = \frac{Cov(X,Y)}{Var(X)}$

Bei kleinen Stichproben soll jedoch die Stichprobenvarianz bzw. die Covarianz der Stichproble verwendet werden. Das Ergebnis bleibt jedoch dasselbe.

Aufgabe 1

Zeigen Sie, dass die folgende Werte richtig sind: Die **geschätzten Koeffizienten**, der **Standardfehler des Koeffizienten $\hat{\beta}$, t-Statistik und den p-Wert des Koeffizienten $\hat{\beta}$**. Interpretieren Sie die Werte. Beachten Sie, dass der Prozentualwert von 2.6729 in der t-Verteilung mit 3 Freiheitsgraden in etwa 0.96225 entspricht.



Aufgabe 1

- Refresher: Bestimmtheitsmass R^2:
  - Erklärungsgehalt des linearen Modells (in Prozent an Gesamtvarianz)
  - aus: $y_i - \overline{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \overline{y})$
  - $\sum_i^n (y_i - \bar{y})^2 = \sum_i^n (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2 + \sum_i^n \varepsilon_i^2$
  - $\sum_i^n (y_i - \bar{y})^2 = \sum_i^n (\hat{y}_i - \bar{y})^2 + \sum_i^n (y_i - \hat{y}_i)^2$
  - *Gesamtvarianz* = erklärte Varianz + *Reststreuung*
    $\Rightarrow R^2 = \frac{\sum_i^n (\hat{y}_i - \bar{y})^2}{\sum_i^n (y_i - \bar{y})^2} = \frac{\text{erklärte Varianz}}{\textit{Gesamtvarianz}}$
  - korrigiertes Bestimmtheitsmass $R^2 = \frac{\text{erklärte Varianz}}{\textit{Gesamtvarianz}} \times \frac{N-1}{N-k}$
  - N Gesamtanzahl and Beobachtungen
  - k = Anzahl der zu schätzenden Parameter im Modell

Aufgabe 1

Verkäufer	x_j	y_j	$x_j y_j$	x_j^2	$x_j - \bar{x}$	$(x_j - \bar{x})^2$	$y_j - \bar{y}$	$(y_j - \bar{y})^2$
1	14	10	140	196	-2	4	-2.4	5.76
2	17	9	153	289	1	1	-3.4	11.56
3	9	8	72	81	-7	49	-4.4	19.36
4	21	19	399	441	5	25	6.6	43.56
5	19	16	304	361	3	9	3.6	12.96
$\sum$	80	62	1068	1368		88		93.2

- **Schätzer des Koeffizienten β:**

$$\hat{\beta} = \frac{Cov(X,Y)}{Var(X)} = \frac{\frac{1}{n-1}(\sum x_j y_j - n\bar{x}\bar{y})}{\frac{1}{n-1}(\sum x_j^2 - \frac{(\sum x_j)^2}{n})} = \frac{\frac{1}{4}\times 1068 - 5\times 16\times 12.4}{\frac{1}{4}\times(1368 - 80^2/5)} = \frac{19}{22}$$

$$\hat{\beta} = 0.8636$$

- **Schätzer des Koeffizienten α:**

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 12.4 - 0.8636 \times 16 = -1.4182$$

$\Rightarrow$ Lineare Regressionsfunktion: $y_i = -1.4182 + 0.8636 x_i + \varepsilon_i$

Aufgabe 1

Verkäufer	x_j	y_j	$\hat{y}$	$\hat{\varepsilon} = y - \hat{y}$	$\hat{\varepsilon}^2$
1	14	10	10.6722	-0.6722	0.4519
2	17	9	13.263	-4.263	18.1732
3	9	8	6.3542	1.6458	2.7087
4	21	19	16.7174	2.2826	5.2103
5	19	16	14.9902	1.0098	1.0197
$\sum$	80	62			27.5638

- **Standardfehler des Parameters** $\hat{\beta}$:
  - Berechne Störvarianz

    $$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{n-2}} = \sqrt{\frac{27.5638}{5-2}} = 3.0312$$

  - Standardfehler (Streuung) des Parameters $\hat{\beta}$:

    $$SE(\hat{\beta}) = s\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = 3.0312\sqrt{\frac{1}{88}} = 0.3231$$

  - **Achtung:** Parameter abhängig von der Stichprobe
  - Koeffizienten liefern im Durchschnitt genaue Schätzung (Erwartungstreue)
  - kleiner Standardfehler $\rightarrow$ geringe Streuung um erwarteten Wert

Aufgabe 1

Verkäufer	x_j	y_j	$\hat{y}$	$\hat{\varepsilon} = y - \hat{y}$	$\hat{\varepsilon}^2$
1	14	10	10.6722	-0.6722	0.4519
2	17	9	13.263	-4.263	18.1732
3	9	8	6.3542	1.6458	2.7087
4	21	19	16.7174	2.2826	5.2103
5	19	16	14.9902	1.0098	1.0197
$\sum$	80	62			27.5638

- **Standardfehler des Parameters $\hat{\alpha}$:**
  - Berechne Störvarianz

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{n-2}} = \sqrt{\frac{27.5638}{5-2}} = 3.0312$$

  - Standardfehler (Streuung) des Parameters $\hat{\alpha}$:

$$SE(\hat{\alpha}) = s\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \overline{x})^2}} = 3.0312\sqrt{\frac{1368}{5 \times 88}} = 5.3448$$

Aufgabe 1

- ***t*-Statistik und *p*-Wert des Steigungskoeffizienten β**
- **Idee:** Ist Parameter β in der Grundgesamtheit signifikant von Null verschieden? $\Rightarrow$ analog zu Hypothesentest
  - unsere Nullhypothese $H_0 : \beta = 0$ vs $H_A : \beta \neq 0$
  - **Achtung:** Wir wollen die Nullhypothese verwerfen, da sonst empirisch keine Beziehung zwischen X und Y nachweisbar wäre
  - Abweichung des Steigungskoeffizienten wäre zufällig, wenn wir die Nullyhpothese nicht verwerfen können.

Aufgabe 1

- **t-Statistik und p-Wert des Steigungskoeffizienten β**

- Teststatistik (eigentlich z-Test)
  - $t = \frac{\hat{\beta} - 0}{SE(\hat{\beta})} \Rightarrow t = \frac{|0.8636 - 0|}{0.3231} = 2.6729$
    $\Rightarrow$ siehe R-output
  - Vergleich mit $\alpha = 5\%$ (zweiseitig, aus t-Tabelle mit 5-2=3 Freiheitsgraden, da (sehr!) geringe Stichprobe)
  - kritischer Wert von 3.182
  - Vergleich t-Statistik (2.6729) mit kritischem Wert
    $\Rightarrow$ nicht in Ablehnungsbereich
    $\rightarrow$ Nullhypothese wird nicht verworfen
  - Steigungsparameter β scheinbar nicht signifikant von null verschieden

Aufgabe 1

- **t-Statistik und p-Wert des Steigungskoeffizienten β**

- Würden wir die Nullhypothese auch verwerfen, wenn wir uns den p-Wert anschauen?

- **p-Wert**
  - Wahrscheinlichkeit, einen Wert von 2.6729 (oder einen extremeren) zu beobachten, wenn die Nullhypothese korrekt ist
  - **Achtung:** Prozentualwert von 2.6729 in der t-Verteilung mit 3 $(n-2)$ Freiheitsgraden in etwa 0.96225
    $\Rightarrow 1 - 0.96225 = 0.03775$
  - p-Wert von $\hat{\beta}$: $2 \times (1 - t_3(2.6729)) = 2 \times (0.03775) = 0.0755$

Aufgabe 1

- ***t*-Statistik und *p*-Wert des Steigungskoeffizienten** β
- Würden wir die Nullhypothese auch verwerfen, wenn wir uns den *p*-Wert anschauen?
- ***p*-Wert** $2 \times (1 - t_3(2.6729)) = 2 \times (0.03775) = 0.0755$
- Unter der Annahme, dass der Koeffizient tatsächlich dem Wert null entspricht, besteht eine Wahrscheinlichkeit von 7.55% in der Stichprobe eine Teststatistik mit mindestens diesem Wert zu erhalten.
- Je kleiner der *p*-Wert, desto geringer wäre die Eintrittswahrscheinlichkeit der ermittelten Teststatistik.
- Nullhypothese kann bei Signifikanzniveau von 10% verworfen werden ($7.55\% < 10\%$ oder $2.6729 > 2.353$ bei $\alpha = 0.10$ (zweiseitig, aus *t*-Tabelle mit 3 Freiheitsgraden)).
- siehe R-output