Lösung 13

1) Die relevanten Parameter sind $n = 35$, $\bar{x} = 18.67$, $\widehat{\sigma}_x = 2$ und $\mu = 18.00$.

(a) Die Teststatistik

$$T = \frac{\bar{X} - \mu}{\widehat{\Sigma}_x/\sqrt{n}}$$

ist nach Annahme t-verteilt mit 34 Freiheitsgraden. Der Annahmebereich ist $[-2.032, 2.032]$, denn aus $P[T \in [-q, q]] = 1 - 0.05 = 0.95$ also nach Umformung $P[T \leq q] = 1 - \frac{0.05}{2} = 0.975$ folgt mit `Mathematica`, dass $q = 2.032$. Der Verwerfungsbereich ist gleich

$$\mathbb{R} \setminus [-2.032, 2.032] = (-\infty, -2.032) \cup (2.032, \infty).$$

Die Beobachtung ist

$$t = \frac{\bar{x} - \mu}{\widehat{\sigma}_x/\sqrt{n}} = \frac{-18.67 - 18}{2/\sqrt{35}} = 1.982.$$

Folglich $t \in [-2.032, 2.032]$, also H_0 wird beibehalten.

(b) Man hat wegen der Symmetrie der t-Verteilung

$$P[T \notin [-t, t]] = 2P[T > t] = 2(1 - P[T \leq t])$$

Aus (a) haben wir $t = 1.982$. Mit Mathematica hat man $P[T \leq t] = 0.972$. Das heisst genau $P[T > t] = 1 - 0.972 = 0.028$, also $P[T \notin [-t, t]] = 0.056$. Der P-Wert für den obigen Test ist also 5.6% (insbesondere wird H_0 für $t = 1.982$ verworfen, solange das Signifikanzniveau α, die Bedingung $\alpha > 0.056$ erfüllt und wird beibehalten wenn $\alpha \leq 0.056$ wie in (a)).

(c) Wir setzen $P[T \in [-q, q]] = 1 - \alpha$, wobei α das Signifikanzniveau bezeichnet, also $P[T \leq q] = 1 - \frac{\alpha}{2} = 0.975$ und wollen q bestimmen. Mit Mathematica hat man wie bei (a): $q = 2.032$. Ausserdem $t = \frac{\bar{x} - \mu}{\widehat{\sigma}_x/\sqrt{n}}$ und H_0 wird beibehalten genau dann wenn $t \in [-q, q]$ oder äquivalent genau dann wenn

$$\mu \in [\bar{x} - \frac{\widehat{\sigma}_x}{\sqrt{n}}q, \bar{x} + \frac{\widehat{\sigma}_x}{\sqrt{n}}q] = [18.67 - \frac{2}{\sqrt{35}} \cdot 2.032, 18.67 + \frac{2}{\sqrt{35}} \cdot 2.032] = [17.983, 19.357].$$

2) Bei einem Test auf Differenz lautet die Nullhypothese $H_0 : \delta = \delta_0$ für die Differenz $\delta = \mu_1 - \mu_2$ zweier unbekannter Mittelwerte. Man setzt $\bar{\Delta} = \bar{X} - \bar{Y}$ und die standardisierte Testvariable dazu ist

$$T = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{\bar{\Delta} - \delta_0}{\sqrt{\frac{(n_1-1)\hat{\sigma}_1^2 + (n_2-1)\hat{\sigma}_2^2}{n_1+n_2-2}}} \sim t_{n_1+n_2-2}$$

unter der Annahme von H_0. Das heisst $T \sim t_{533}$, also mit Mathematica kann man überprüfen, dass $[-q, q] = [-1.964, 1.964]$ der entsprechende beidseitige Annahmebereich auf dem 5%-Niveau ist. Mit $\bar{\delta} = \bar{x}_1 - \bar{x}_2$ ergibt das Einsetzen der Werte und der Nullhypothese $\delta_0 = 0$:

$$t = \sqrt{\frac{35 \cdot 500}{35 + 500}} \frac{(18.67 - 17.9) - 0}{\sqrt{\frac{(35-1)2^2 + (500-1)3^2}{35+500-2}}} = 1.495 \in [-1.964, 1.964] = [-q, q].$$

Die Nullhypothese H_0 wird also beibehalten und wir glauben dementsprechend, dass sich der Zuckergehalt nicht verändert hat. er gesunken ist, auch wenn die Schätzer $\bar{x}_1$ und $\bar{x}_2$ das nahelegen.

3) Der Output ist

```
Daten42 = {{2, -1, 0, 3}, {1, 0, 0, 2}, {2, 1, 1, 4},
  {1, 2, 1, 1}, {2, 3, 0, 0}}


{{2, -1, 0, 3}, {1, 0, 0, 2},
 {2, 1, 1, 4}, {1, 2, 1, 1}, {2, 3, 0, 0}}

M42 = LinearModelFit[Daten42, {x1, x2, x3}, {x1, x2, x3}]

FittedModel[ 0.375 + «19» x1 - «19» x2 + 1.75 x3 ]

M42["BestFit"](* Regressionsgleichung bestimmen *)

0.375 + 1.125 x1 - 0.875 x2 + 1.75 x3

M42["FitResiduals"](* Residuen bestimmen *)

{-0.5, 0.5, 0.5, -0.5, 1.77636 × 10^-15}

M42["ParameterTable"](* Die Parametertabelle *)
```

	Estimate	Standard Error	t Statistic	P-Value
1	0.375	1.65359	0.226779	0.858029
x1	1.125	0.927025	1.21356	0.43877
x2	−0.875	0.330719	−2.64575	0.230053
x3	1.75	0.968246	1.80739	0.321722

Wir haben also eine lineare Regression mit 3 Ausgangsvariablen X_1, X_2, X_3, einer Zielvariable Y die linear von diesen abhängen soll, sowie 5 Datensätze.

Zu a): Die Regressionsgleichung kann man entweder direkt von der Ausgabe des Befehls `BestFit` ablesen, oder aus der `estimate`-Spalte der Parametertabelle:

$$Y \;=\; 0.375 \;+\; 1.125X_1 \;-\; 0.875X_2 \;+\; 1.75X_3\,.$$

Zu b): Die Residuen wurden durch den Befehl `FitResiduals` berechnet:

$$r_1 \;=\; -0.5 \;\;,\;\; r_2 \;=\; 0.5 \;\;,\;\; r_3 \;=\; 0.5 \;\;,\;\; r_4 \;=\; -0.5 \;\;,\;\; r_5 \;=\; 1.77\cdot 10^{-15} \;\approx\; 0\,.$$

Das Residuum für den fünften Datensatz ist in Wirklichkeit Null, der Wert 10^{-15} kommt durch Rechenungenauigkeiten in `Mathematica` zustande. Am besten Geschätzt ist der Parameter $\hat{\beta}_2$, der Koeffizient von X_2, denn er hat in der Parametertabelle die kleinste Standardabweichung bzw. den kleinsten P-Wert. Die Residuen haben damit nichts zu tun: sie geben an wie gut die fünf Datensätze zum Modell passen, und nicht wie gut die 4 Parameter geschätzt sind.

Zu c): Die Vorhersage des Modells entsteht, indem man die Ausgangswerte $x_1 = 2$, $x_2 = 3$ und $x_3 = 4$ in die Regressionsgleichung einsetzt:

$$y \;=\; 0.375 \;+\; 1.125\cdot 2 \;-\; 0.875\cdot 3 \;+\; 1.75\cdot 4 \;=\; 7\,.$$

Zu d): Die Abszisse β_0 der Regressionsgleichung ist eine Zufallsvariable, die von den gezogenen Datenwerten abhängig ist. Für die fünf konkret vorliegenden Datensätze bekommt man den konkreten Schätzwert $\hat{\beta}_0 = 0.375$. Um diesen Wert sollen wir ein Intervall I legen, so dass $P(\beta \in I) = 95\%$ ist bzgl. der Normalverteilung. Die Variable β_0 ist noch nicht standardnormalverteilt, ihre Standardisierung (so dass der Intervallmittelpunkt $\hat{\beta}_0$ ist) lautet

$$Z \;=\; \frac{\beta_0 - \hat{\beta}_0}{\hat{\sigma}} \;=\; \frac{\beta_0 - 0.375}{1.65}\,.$$

Dabei wurde die (geschätzte) Standardabweichung $\hat{\beta}_0 = 1.65$ aus der zweiten Spalte der Parametertabelle abgelesen. Diese berücksichtigt (bei Mathematica) bereits die Stichprobenzahl, man muss also nicht zusätzlich durch $\sqrt{n}$ dividieren. Laut Aufgabe ist die Normalverteilung einzusetzen (was

eigentlich unzulässig ist, da $\hat{\sigma}$ geschätzt wurde und die Stichprobenzahl viel zu klein ist, aber darum geht es in dieser Aufgabe nicht):

$$95\% \;=\; 0.95 \;\stackrel{!}{=}\; P(-a \le Z \le a) \;=\; \Phi(a) - \Phi(-a) \;=\; 2\Phi(a) - 1$$

für die Verteilungsfunktion $\Phi(z)$ der Standardnormalverteilung. Aus der Tabelle (oder Mathematica) liest man den Quantilswert $a = 1.96$ ab. Diesen Wert müssen wir von Z auf β_0 umrechnen:

$$95\% \;=\; P(-1.96 \le Z \le 1.96) \;=\; P\left(-1.96 \le \frac{\beta_0 - 0.375}{1.65} \le 1.96\right) \;=\; P(-2.859 \le \beta_0 \le 3.609)\,.$$

Also ist das gesuchte Intervall $I = [-2.859\,,\, 3.609]$. Gemessen am eigentlichen Wert $\hat{\beta}_0$ ist es sehr breit: die kleine Stichprobenzahl führt dazu, dass der Schätzwert $\hat{\beta}_0$ nicht sonderlich vertrauenswürdig ist.

4) Die entsprechenden Mathematica-Kommandos lauten

```
In[1]:= Daten45 = {{2, -1, 0}, {1, 0, 0}, {2, 1, 1}, {1, 2, 1}, {2, 3, 0}}

Out[1]= {{2, -1, 0}, {1, 0, 0}, {2, 1, 1}, {1, 2, 1}, {2, 3, 0}}

In[2]:= M45 = LinearModelFit[Daten45, {x1, x2}, {x1, x2}]

Out[2]= FittedModel[ 0.566667 - <<20>> x1 + 0.1 x2 ]

In[3]:= M45["BestFit"](* Regressionsebene bestimmen *)
Out[3]= 0.566667 - 0.166667 x1 + 0.1 x2

In[4]:= M45["FitResiduals"](* Residuen bestimmen *)
Out[4]= {-0.133333, -0.4, 0.666667, 0.4, -0.533333}

In[5]:= M45["ParameterTable"](* Die komplette Parametertabelle,
        bis auf Rundungsfehler sind das die per Hand berechneten
         Werte von der Serie *)
```

Out[5]=

	Estimate	Standard Error	t Statistic	P-Value
1	0.566667	1.1392	0.497425	0.668194
x1	−0.166667	0.666667	−0.25	0.825922
x2	0.1	0.23094	0.433013	0.70723

Hier die Rechnung per Hand:

Wir nehmen eine lineare Abhängigkeit der Form

$$y \;=\; \beta^{(0)} + \beta^{(1)} X^{(1)} + \beta^{(2)} X^{(2)}$$

an, und wollen dazu die Schätzer $\hat{\beta}_j$ der Koeffizienten β_j ausrechnen. Die Methode der kleinsten Quadrate bestimmt diese Werte so, dass der (quadratische) Abstand der Messdaten zur gefundenen Gleichung minimal ist. Dazu stellen wir die Matrix X auf: Jede Zeile gehört zu einem Datensatz aus den Messdaten, und ist von der Form $(1\; x_1\; \cdots\; x_k)$, wobei $k = 2$ hier die Anzahl der Steigungskoeffizienten β_j ist. Die 1 am Anfang der Zeile gehört zum Abschnittskoeffizienten β_0. Einsetzen der Messdaten liefert

$$X \;=\; \begin{pmatrix} 1 & 2 & -1 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 3 \end{pmatrix}\,.$$

Die Messwerte der Zufallsvariablen Y tragen wir in den Vektor

$$y \;=\; \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{pmatrix}$$

ein. Angenommen die lineare Abhängigkeit ist richtig, und $\beta_0, \beta_1, \beta_2$ sind die richtigen Koeffzienten, dann gilt $X \cdot \beta = y$ als lineares Gleichungssystem. Nun ist die Abhängigkeit aber nicht sicher, und die Koeffizienten kennen wir nicht, also lösen wir statt des LGS $X\beta = y$ das Kleinste-Quadrate-System $(X^T X)\beta = X^T y$, denn dieses hat immer eine Lösung, die der „echten“ Lösung am nächsten kommt. Das kann man nun machen, indem man $X^T X$ invertiert und auf die andere Seite bringt, oder indem man das System auf Zeilenstufenform bringt und eine spezielle Lösung abliest (die dann, weil $X^T X$ regulär ist, auch die einzige Lösung des Systems ist). Wir berechnen erstmal

$$X^T \cdot X \;=\; \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 1 & 2 & 1 & 2 \\ -1 & 0 & 1 & 2 & 3 \end{pmatrix} \cdot \begin{pmatrix} 1 & 2 & -1 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 3 \end{pmatrix} \;=\; \begin{pmatrix} 5 & 8 & 5 \\ 8 & 14 & 8 \\ 5 & 8 & 15 \end{pmatrix} .$$

Wir rechte Seite des Kleinste-Quadrate-Systems ist

$$X^T \cdot y \;=\; \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 1 & 2 & 1 & 2 \\ -1 & 0 & 1 & 2 & 3 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{pmatrix} \;=\; \begin{pmatrix} 2 \\ 3 \\ 3 \end{pmatrix} ,$$

wir haben also das LGS

$$\begin{pmatrix} 5 & 8 & 5 \\ 8 & 14 & 8 \\ 5 & 8 & 15 \end{pmatrix} \cdot \beta \;=\; \begin{pmatrix} 2 \\ 3 \\ 3 \end{pmatrix}$$

zu bearbeiten, und wir wissen schon dass es nur genau einen Lösungsvektor gibt. Umformen der augmentierten Matrix ergibt

$$\begin{pmatrix} 5 & 8 & 5 & \mathbf{2} \\ 8 & 14 & 8 & \mathbf{3} \\ 5 & 8 & 15 & \mathbf{3} \end{pmatrix} \overset{\substack{\text{Elimination} \\ \longrightarrow \\ \text{Spalte 1}}}{} \begin{pmatrix} 5 & 8 & 5 & \mathbf{2} \\ 0 & \frac{6}{5} & 0 & \mathbf{-\frac{1}{5}} \\ 0 & 0 & 10 & \mathbf{1} \end{pmatrix} .$$

Diese Matrix hat schon Zeilenstufenform, und wir lesen

$$\hat{\beta} \;=\; \begin{pmatrix} \frac{17}{30} \\ -\frac{1}{6} \\ \frac{1}{10} \end{pmatrix}$$

ab. Das sind die Schätzer der linearen Gleichung:

$$\hat{\beta}_0 \;=\; \frac{17}{30} \;,\; \hat{\beta}_1 \;=\; -\frac{1}{6} \;,\; \hat{\beta}_2 \;=\; \frac{1}{10}$$

und die Regressionsebene lautet

$$y \;=\; \frac{17}{30} - \frac{1}{6} x^{(1)} + \frac{1}{10} x^{(2)} .$$

Die Residuen, d.h. die Fehler die wir machen wenn wir diese Gleichung als die richtige annehmen,

sind

$$R \;=\; \begin{pmatrix} R_1 \\ R_2 \\ R_3 \\ R_4 \\ R_5 \end{pmatrix} \;=\; y - X \cdot \hat{\beta} \;=\; \begin{pmatrix} -\frac{2}{15} \\ -\frac{2}{5} \\ \frac{2}{3} \\ \frac{2}{5} \\ -\frac{8}{15} \end{pmatrix} .$$