Lösungsskizze Übung 6

1. **a)** Die Vierfeldertafel hat hier folgende Form:

	günstig	ungünstig	Zeilentotal
Aspiringruppe	63	15	78
Kontrollgruppe	43	34	77
Spaltentotal	106	49	155

b) Wären die Verteilungen der günstigen und der ungünstigen Entwicklungen in jeder Gruppe (also Zeile der Vierfeldertafel) gleich, dann müssten sich beispielsweise die 78 Patienten der Aspiringruppe im Verhältnis der beiden Spaltentotale auf die Fälle „günstig“ und „ungünstig“ aufteilen. Unter dieser Annahme (i.e. unsere H_0: „Gruppen unterscheiden sich nicht“) würde man

$$78 \cdot \frac{106}{155} = 53.34$$

günstige Entwicklungen in der Aspiringruppe erwarten. Der Beitrag $d^2_{A/g}$ an die χ^2-Testgrösse aus diesem Feld ist dann (wie immer beim χ^2-Test)

$$\begin{aligned} d^2_{A/g} &= \frac{(\text{beobachtete} - \text{erwartete Anzahl (im Feld } A/g))^2}{\text{erwartete Anzahl (im Feld } A/g)} \\ &= \frac{(63 - 53.34)^2}{53.34} = 1.749\,. \end{aligned}$$

Auf analoge Art berechnet man die erwarteten Anzahlen in den anderen Feldern und deren Beiträge zum χ^2. Als Testgrösse erhält man dann

$$\chi^2 = d^2_{A/g} + d^2_{A/u} + d^2_{K/g} + d^2_{K/u} = 1.749 + 3.783 + 1.771 + 3.832 = 11.13\,.$$

Der Verwerfungsbereich ist gemäss Tabelle der χ^2-Verteilung (hier mit 1 Freiheitsgrad) $K =]3.841, \infty[$. Wir verwerfen H_0.
Der Unterschied zwischen Behandlungs- und Kontrollgruppe ist also gesichert.

2. **a)** Aus den gegebenen Quadratsummen berechnen wir zuerst die geschätzten Regressionsparameter.

$$\begin{aligned} \widehat{\beta} &= \frac{\sum_{i=1}^{26}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{26}(x_i - \overline{x})^2} = \frac{8569.77}{9852.62} = 0.87 \\ \widehat{\alpha} &= \overline{y} - \widehat{\beta}\overline{x} = 51.04 - 0.87 \cdot 53.23 = 4.74 \end{aligned}$$

Man erhält also die Regressionsgerade $x \mapsto y = 4.74 + 0.87x$. Der prognostizierte y-Wert bei $x = 80$ ist $\widehat{y}(80) = 4.74 + 0.87 \cdot 80 = 74.32$.

b) Wir testen hier die Nullhypothese, dass die wahre Steigung β der Regressionsgeraden Null ist, also H_0: $\beta = 0$. Als Alternative wählen wir H_A: $\beta \neq 0$, d. h. wir testen zweiseitig.
Die Teststatistik hat den Wert

$$\begin{aligned} t &= \frac{\widehat{\beta} - 0}{s/\sqrt{S_{xx}}} = \frac{\widehat{\beta}\sqrt{S_{xx}}}{\sqrt{\frac{1}{n-2}(S_{yy} - \widehat{\beta}^2 S_{xx})}} \\ &= \frac{0.87\sqrt{9852.62}}{\sqrt{\frac{1}{26-2}(7836.96 - 0.87^2 \cdot 9852.62)}} = 21.61\,. \end{aligned}$$

Wegen $|t| >> t_{24,0.975} = 2.064$ verwerfen wir die Nullhypothese deutlich. Die Partikelkonzentration in der Stadt beeinflusst die Partikelkonzentration ausserhalb der Stadt. Das sieht man sogar von Auge....

3. **a)** Es ist ein linearer Zusammenhang zwischen der (logarithmierten) Erschütterung und der (logarithmierten) Distanz ersichtlich. Zwischen der (logarithmierten) Erschütterung und der (logarithmierten) Ladung ist kein offensichtlicher Zusammenhang zu sehen, ebensowenig zwischen den beiden erklärenden Variablen.

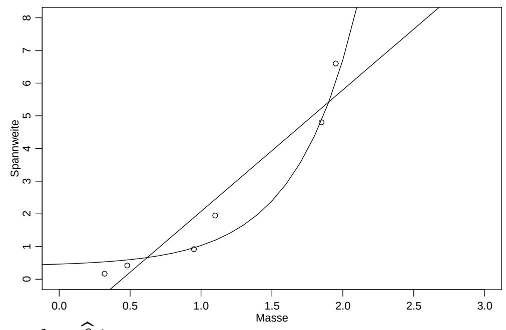
b) `LOGERSCH` $= 6.53 - 1.57 \cdot$ `LOGDIST` $+ 1.07 \cdot$ `LOGLAD`.

c) Ein Mass für die Güte der Modellanpassung ist R^2. Gemäss Systat-Output ist $R^2 = 0.65$, d. h. nur 65% der totalen Varianz in den Daten werden durch das Modell erklärt. Das Modell passt nicht sehr gut.

d) `LOGDIST` hat einen signifikanten Einfluss (P-Wert $= 0.00 < 0.05$), ebenso `LOGLAD` (P-Wert $= 0.00 < 0.05$).

e) **Normalplot:** Im Normalplot ist (mit einiger Erfahrung, die ihr vielleicht noch nicht habt) eine leichte Krümmung zu sehen, die auf eine geringfügige Abweichung der Verteilung der Residuen von der Normalverteilung hinweist.
Tukey-Anscombe-Plot: Die Varianz der Fehler zeigt im Tukey-Anscombe-Plot eine unerwünschte Struktur. Sie scheint zunächst mit zunehmendem Schätzwert zuzunehmen und für Schätzwerte über eins wieder abzunehmen. Ohne genauer zu analysieren, was der Grund für diese Struktur ist, lässt sich festhalten, dass der Tukey-Anscombe Hinweise auf die Verletzung von Modellannahmen zeigt. Die Residuen liegen nicht in einem konstanten Band.

4. a)



Die Steigung $b = \widehat{\beta}$ ist:

$$\widehat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{6}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{6}(x_i-\bar{x})^2} = \frac{1.82+1.29+0.25+0+1.72+3.47}{0.62+0.39+0.03+0+0.55+0.71} = \frac{8.55}{2.3} = 3.72$$

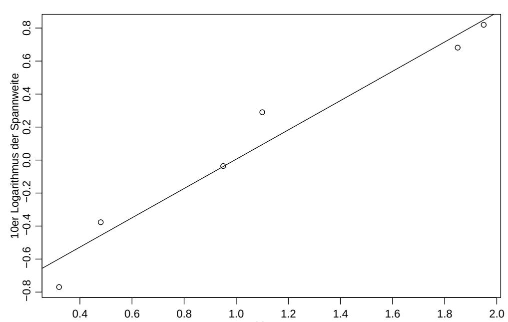
Wobei $\bar{x} = 1.11$ und $\bar{y} = 2.48$.
Und $a = \widehat{\alpha}$ ist:

$$\widehat{\alpha} = \bar{y} - b\bar{x} = 2.48 - 3.72 \cdot 1.11 = -1.65$$

Wir erhalten also die Regressionsgerade $y = -1.65 + 3.72x$.

b) In der ersten Abbildung ist die Regressionsgerade und eine besser passende Kurve $y = 0.063 \cdot 10^x + 0.4$ eingezeichnet. Aufgrund dieser Abbildung werden wir die Y-Werte $\log_{10}$-transformieren.

c)



Die $z_i = \log_{10}(y_i)$ haben die Werte:

$$z_1 = -0.77,\ z_2 = -0.38,\ z_3 = -0.04,\ z_4 = 0.29,\ z_5 = 0.68,\ z_6 = 0.82$$

Mit $\bar{z} = 0.10$ und $\bar{x} = 1.11$. Somit folgt:

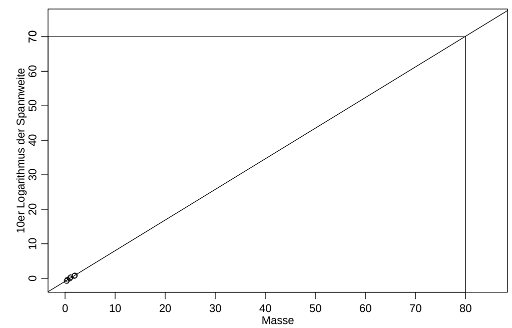
$$d = \widehat{\beta}_z = \frac{S_{xz}}{S_{xx}} = \frac{\sum_{i=1}^{6}(x_i-\bar{x})(z_i-\bar{z})}{\sum_{i=1}^{6}(x_i-\bar{x})^2} = \frac{0.69+0.30+0.02+0+0.43+0.60}{2.3} = \frac{2.04}{2.3} = 0.89$$

Und $a = \widehat{\alpha_z}$ ist:

$$c = \widehat{\alpha_z} = \bar{z} - b\bar{x} = 0.10 - 0.89 \cdot 1.11 = -0.89$$

In der zweiten Abbildung sieht man die nun ziemlich gut passende Regressionsgerade.

d)



Aus der letzten Abbildung liest man ab für $x = 80$: $z = 70$ und somit ist die Spannweite $y = 10^z = 10^{70}$ Meter!
Oder man berechnet: $z = -0.89 + 0.89 * x$. In diese Geradengleichung setzt man den Wert $x = 80$ ein und bekommt $z \approx 70$ und somit $y = 10^z = 10^{70}$.