Lösungsskizze Serie 1

1. **a)** Setze $Z := (1, X_1, \ldots, X_p)$ und $\theta := (\theta_0, \ldots, \theta_p)^t$. Durch Ableiten und Nullsetzen von $\mathcal{E}[(Y - Z\theta)^2]$ nach θ erhalten wir $-2\,\mathcal{E}[Z^t(Y - Z\theta)] = 0$ und daraus das gesuchte Gleichungssystem

$$\mathcal{E}[Z^t Y] = \mathcal{E}[Z^t Z]\theta \qquad (*)$$

(Erwartungswerte von Vektoren und Matrizen sind komponentenweise zu verstehen.)
In der obigen Herleitung wurden Ableitung und Integral (Erwartungswert) vertauscht. Dies müsste noch begründet werden.
Eine Variante wäre, $(Y - Z\theta)^2$ auszumultiplizieren, θ aus dem Erwartungswert zu ziehen und dann erst abzuleiten. Dies führt zum gleichen Ergebnis und benötigt keine Vertauschung von Ableitung und Integral. Allerdings sind die Rechnungen etwas aufwendiger.

b) Die erste Gleichung im System (*) ist gerade $\mathcal{E}[Y] = \theta_0 + \theta_1\,\mathcal{E}[X_1] + \ldots + \theta_p\,\mathcal{E}[X_p]$; auflösen nach θ_0 ergibt die Bedingung für θ_0^{opt}.
Um das Gleichungssystem für $\theta_1, \ldots, \theta_p$ zu finden, führen wir neue Variablen $\tilde{Y} = Y - \mathcal{E}[Y]$ und $\tilde{X}_i = X_i - \mathcal{E}[X_i]$ ein. Beachte, dass die "Regression" mit den Erklärenden $(\tilde{X}_i)$ und Zielgrösse $\tilde{Y}$ zu den gleichen Koeffizienten $\theta_1, \ldots, \theta_p$ führt, wie die ursprüngliche "Regression"; θ_0 ist jetzt aber 0. Das Gleichungssystem aus a) für die neuen Variablen lautet nun (mit der Bezeichnung $\tilde{Z} := (1, \tilde{X}_1, \ldots, \tilde{X}_p)$): $\mathcal{E}[\tilde{Z}^t\tilde{Y}] = \mathcal{E}[\tilde{Z}^t\tilde{Z}]\cdot(0, \theta_1, \ldots, \theta_p)^t$.
Die letzten p Gleichungen dieses Systems liefern das gesuchte Gleichungssystem:

$$\mathcal{E}[\tilde{X}^t\tilde{Y}] = \mathcal{E}[\tilde{X}^t\tilde{X}]\cdot(\theta_1, \ldots, \theta_p)^t \quad \text{bzw.} \quad \mathrm{Cov}(X, Y) = \mathrm{Cov}(X)\cdot(\theta_1, \ldots, \theta_p)^t,$$

wobei $\tilde{X} := (\tilde{X}_1, \ldots, \tilde{X}_p)$; Cov bezeichnet die Kovarianzmatrix definiert durch $\mathrm{Cov}(X, Y) := \mathcal{E}[(X - E(X))^t(Y - E(Y))]$, $\mathrm{Cov}(X) := \mathrm{Cov}(X, X)$.

2. **a)** Es ist $\Sigma^{-1} = \frac{1}{(1-\rho^2)\sigma_x^2\sigma_y^2}\begin{pmatrix} \sigma_y^2 & -\rho\sigma_x\sigma_y \\ -\rho\sigma_x\sigma_y & \sigma_x^2 \end{pmatrix}$. Wenn wir $\tilde{x} := x - \mu_x$ und $\tilde{y} := y - \mu_y$ setzen ergibt sich

$$\begin{aligned}(\tilde{x}, \tilde{y})\Sigma^{-1}\begin{pmatrix}\tilde{x}\\ \tilde{y}\end{pmatrix} &= \frac{1}{(1-\rho^2)\sigma_y^2}(\tilde{y}^2 - 2\rho\frac{\sigma_x}{\sigma_y}\tilde{y}\tilde{x} + \frac{\sigma_y^2}{\sigma_x^2}\tilde{x}^2) \\ &= \frac{1}{(1-\rho^2)\sigma_y^2}(\tilde{y} - \rho\frac{\sigma_x}{\sigma_y}\tilde{x})^2 + \frac{\tilde{x}^2}{\sigma_x^2}\end{aligned}$$

b) Die Dichte der Randerteilung von X ist

$$\begin{aligned}f_X(x) = \int_{\mathbb{R}} f(x, y)dy &= \frac{1}{\sqrt{2\pi}\sigma_x}\exp(-\frac{1}{2}\frac{(x-\mu_x)^2}{\sigma_x^2})\cdot \\ &\quad \frac{1}{\sqrt{2\pi}\sqrt{(1-\rho)^2}\sigma_y}\int_{\mathbb{R}}\exp(-\frac{1}{2}\frac{1}{(1-\rho)^2\sigma_y^2}(y - \mu_y - \rho\frac{\sigma_y}{\sigma_x}(x-\mu_x))^2)dy\end{aligned}$$

Die Behauptung folgt, weil die zweite Zeile gerade 1 ist.

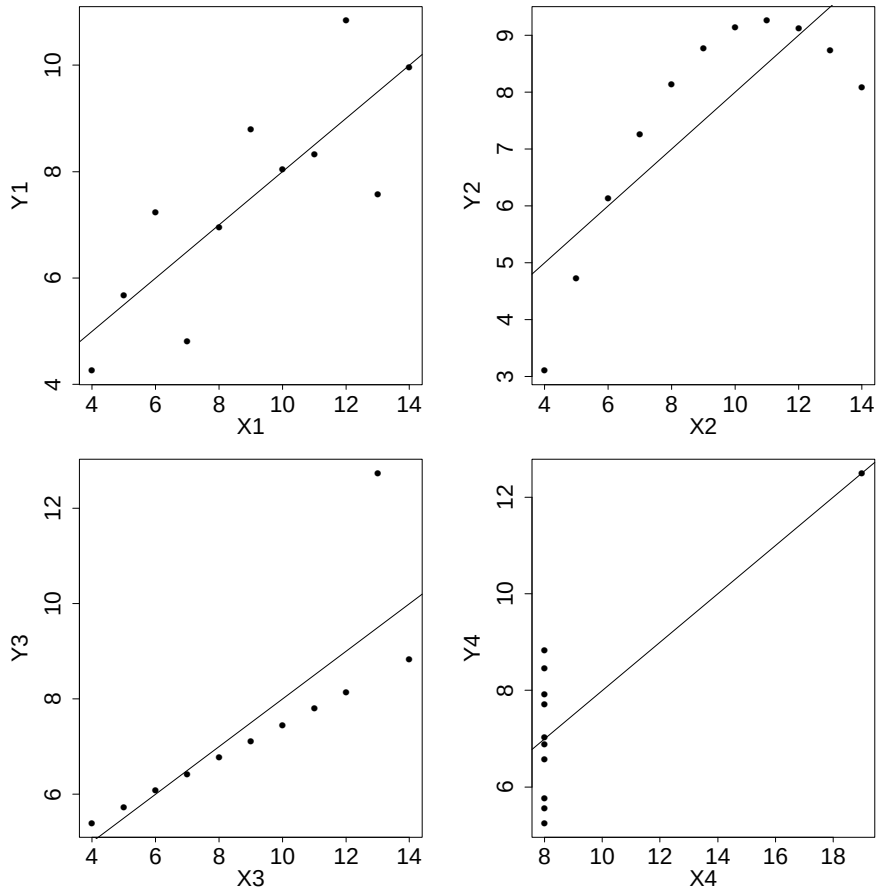
c) Es gilt $P[Y \le y | x \le X \le x + h] = \frac{P[y \le y, x \le X \le x+h]}{P[x \le X \le x+h]} \approx \frac{\int_{-\infty}^{y} f_{X,Y}(x,v)\cdot h\,dv}{f_X(x)\cdot h}$. Also ist der gesuchte Limes gleich $\frac{\int_{-\infty}^{y} f_{X,Y}(x,v)dv}{f_X(x)}$. Mit Hilfe von a) ergibt sich, dass $\frac{f_{X,Y}(x,v)dv}{f_X(x)}$ gerade die Dichte einer Normalverteilung der Form $\mathcal{N}(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x), \sigma_y^2(1-\rho^2))$ ist. Der gesuchte Limes ist somit

$$\Phi(\frac{y - (\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x))}{\sigma_y\sqrt{1-\rho^2}}).$$

3.

Funktion	Transformation	Lineare Form
$y = \alpha x^{\beta}$	$y' = \log(y),\ x' = \log(x)$	$y' = \log(\alpha) + \beta \cdot x'$
$y = \alpha e^{\beta \cdot x}$	$y' = \log(y)$	$y' = \log(\alpha) + \beta \cdot x$
$y = \alpha + \beta \cdot \log(x)$	$x' = \log(x)$	$y = \alpha + \beta \cdot x'$
$y = x/(\alpha \cdot x - \beta)$	$y' = \frac{1}{y},\ x' = -\frac{1}{x}$	$y' = \alpha + \beta \cdot x'$
$y = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta \cdot x}}$	$y' = \frac{1}{y},\ y'' = y' - 1,\ y''' = -\log(y'')$	$y''' = \alpha + \beta \cdot x$
$y = \alpha e^{\beta/x}$	$y' = \log(y),\ x' = \frac{1}{x}$	$y' = \log(\alpha) + \beta \cdot x'$
$y = 1/(\alpha + \beta e^{-x})$	$y' = \frac{1}{y},\ x' = e^{-x}$	$y' = \alpha + \beta \cdot x'$

4. Betrachtet man die vier Streudiagramme, so sieht man, dass nur im ersten Fall eine lineare Regression angebracht ist. Im zweiten Fall ist die Beziehung zwischen X und Y nicht linear, sondern eher quadratisch. Im dritten Fall gibt es einen Ausreisser, welcher die geschätzten Parameter stark beeinflusst. Im vierten Fall wird die Regressionsgerade durch einen einzigen Punkt bestimmt.



Ergänzung: Bei allen vier Modellen sind der Achsenabschnitt, die Steigung und die zugehörigen Standardfehler, sowie $\widehat{\sigma}^2$ und R^2 praktisch identisch:

	Modell 1	Modell 2	Modell 3	Modell 4
Achsenabschnitt ($\widehat{\alpha}$)	3.000	3.001	3.002	3.002
Steigung ($\widehat{\beta}$)	0.500	0.500	0.500	0.500
$\text{se}^{(\hat{\alpha})}$	1.125	1.125	1.124	1.124
$\text{se}^{(\hat{\beta})}$	0.118	0.118	0.118	0.118
$\widehat{\sigma}^2$	1.529	1.531	1.528	1.527
R^2	0.667	0.666	0.666	0.667

Fazit: Es genügt **nicht**, nur $\widehat{\alpha}$, $\widehat{\beta}$, $\widehat{\sigma}(\widehat{\alpha})$, $\widehat{\sigma}(\widehat{\beta})$, R^2 und $\widehat{\sigma}$ anzuschauen. In allen Modellen sind diese Schätzungen fast gleich, aber die Datensätze sehen ganz unterschiedlich aus. Eine (graphische) Überprüfung der Modellannahmen ist also unumgänglich.

5. **a)** Die Originaldaten zeigen einen globalen Trend (die Anzahl Passagiere wird immer grösser) und monatliche Schwankungen. Der globale Trend scheint zwar mehr oder weniger linear, aber die monatlichen Schwankungen werden immer stärker. Dies deutet eher auf ein multiplikatives Modell als auf ein lineares Modell hin.

b) Bei den logarithmierten Daten scheint jetzt der saisonale Trend additiv zu sein. Auch der globale Trend ist ziemlich linear. Es ist also sinnvoll, ein lineares Modell anzupassen.

c) Den globalen Trend modellieren wir linear, d.h. wir wählen $f_1(t) := t$. Für den saisonalen Trend wählen wir Monatseffekte, d.h. wir definieren z.B.

$$f_2(t) := \begin{cases} 1 & \text{falls } t\text{='Januar'} \\ 0 & \text{sonst} \end{cases}$$

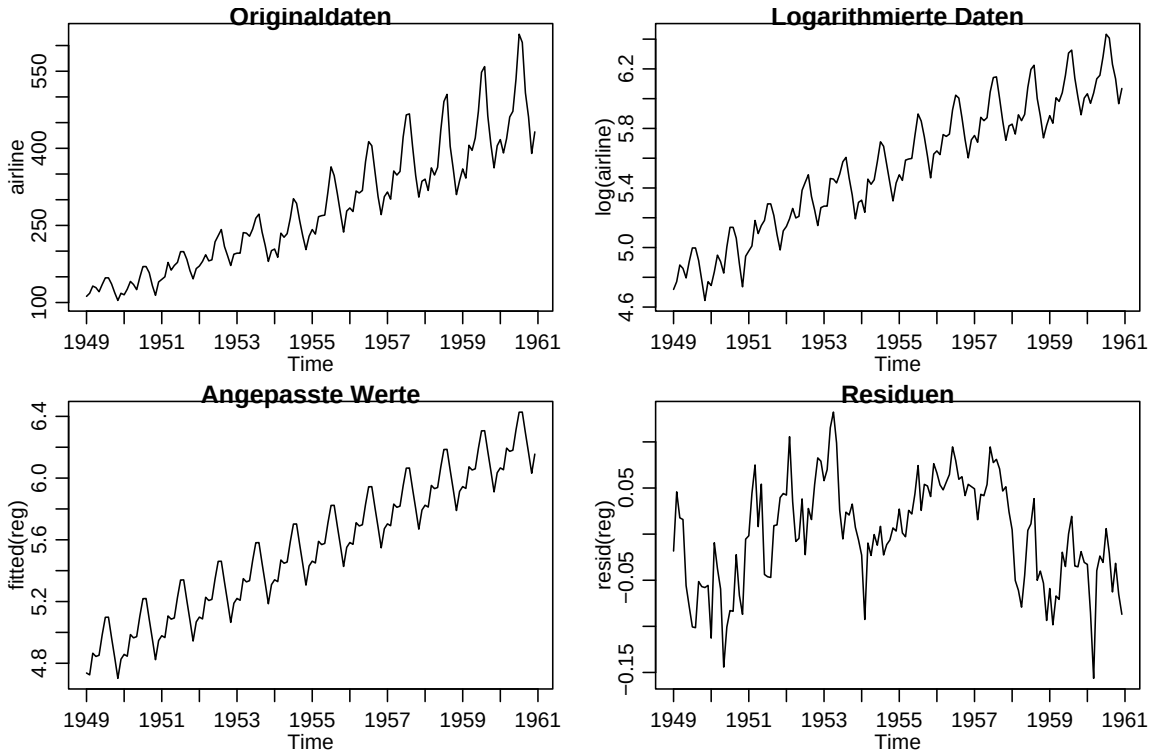
und entsprechend die Funktionen $f_3, \ldots, f_{13}$ für die anderen Monate. Beachte, dass ein Achsenabschnitt bereits mit den Monatseffekten mitmodelliert ist.

d) Entweder beschreibt man alle Funktionen mit 'R' separat. Eine elegantere Variante ist, den Monatseffekt direkt als Faktor zu modellieren. Dies wurde aber noch nicht behandelt.

```
## Modell anpassen
t <- 1:144                                    #linearer Trend
janTrend <- rep(c(1,rep(0,11)),12)            #Monatseffekt Januar
febTrend <- rep(c(0,1,rep(0,10)),12)          #Monatseffekt Februar
#etc.
reg <- lm(log(airline)~t+janTrend+febTrend+...-1)

## Modell anpassen (elegantere Variante)
t <- 1:144                                    #Linearer Trend
monthTrend <- as.factor(rep(month.name,12))   #Monatseffekt
reg <- lm(log(airline)~t+monthTrend-1)
```

Die angepassten Werte passen ziemlich gut. Allerdings haben zeigen die Residuen immer noch eine Struktur, sie sind korreliert.


Originaldaten
airline
100
250
400
550
1949
1951
1953
1955
1957
1959
1961
Time
Logarithmierte Daten
log(airline)
4.6
5.0
5.4
5.8
6.2
1949
1951
1953
1955
1957
1959
1961
Time
Angepasste Werte
fitted(reg)
4.8
5.2
5.6
6.0
6.4
1949
1951
1953
1955
1957
1959
1961
Time
Residuen
resid(reg)
−0.15
−0.05
0.05
1949
1951
1953
1955
1957
1959
1961
Time