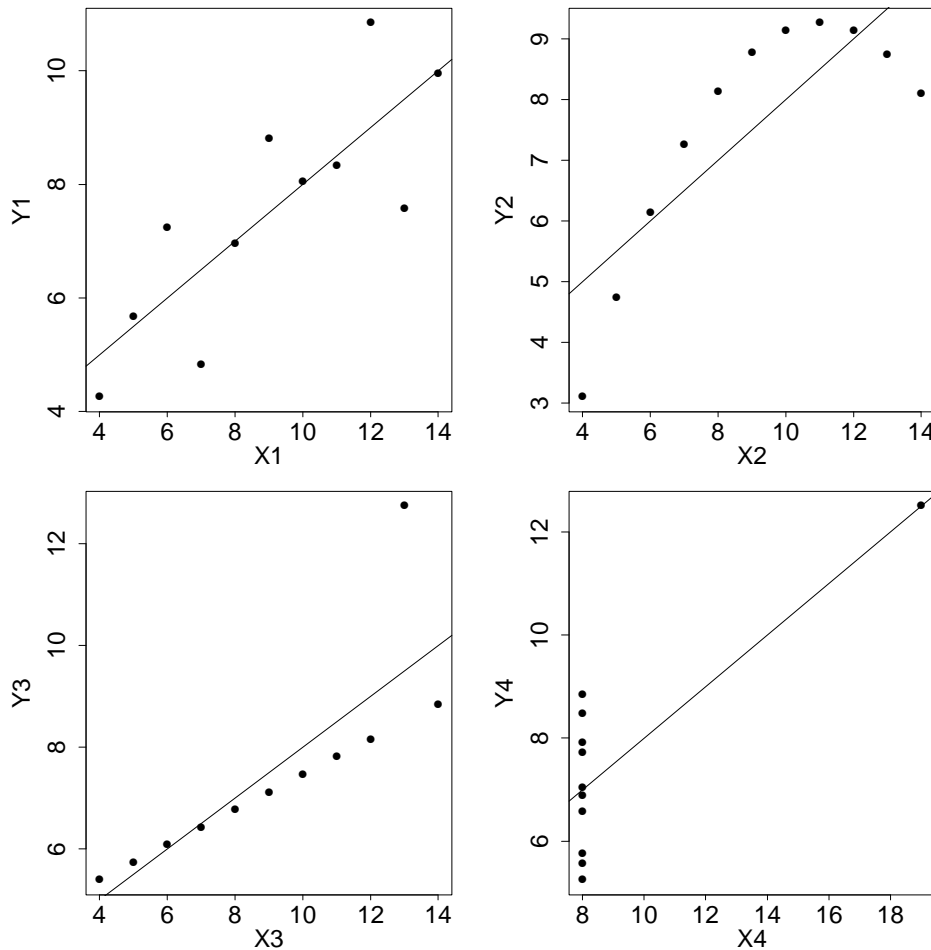


Musterlösung zu Serie 12

1. a) Betrachtet man die vier Streudiagramme, so sieht man, dass nur im ersten Fall eine lineare Regression korrekt ist. Im zweiten Fall ist die Beziehung zwischen X und Y nicht linear, sondern 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.



- b) Bei allen vier Modellen sind die Schätzungen des Achsenabschnitts β_0 , der Steigung β_1 und der Fehlervarianz σ^2 , sowie das Gütemass R^2 fast identisch:

	Modell 1	Modell 2	Modell 3	Modell 4
Achsenabschnitt ($\widehat{\beta}_0$)	3.000	3.001	3.002	3.002
Steigung ($\widehat{\beta}_1$)	0.500	0.500	0.500	0.500
$\widehat{\sigma}^2$	1.529	1.531	1.528	1.527
R^2	0.667	0.666	0.666	0.667

Fazit: Es genügt **nicht**, nur $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\sigma}$ und R^2 anzuschauen. In allen Modellen sind diese Schätzungen fast gleich, aber die Datensätze sehen ganz unterschiedlich aus. Eine (graphische) Überprüfung der Modellannahmen ist also unumgänglich.

2. a), b), c) Während ein Blick auf die Originaldaten suggeriert, dass sie ziemlich schön auf einer Geraden liegen, zeigt der Tukey-Anscombe-Plot, dass dem nicht so ist: die Daten weisen eine Krümmung auf. Zusätzlich fällt der Ausreisser in der Mitte oben auf, den wir auch im Normalplot rechts oben wiederfinden. Ansonsten lässt sich über den Normalplot nicht viel sagen. Im Regressions-Output findet sich keine Hinweis darauf, dass das Modell unpassend sein könnte. Im Gegenteil könnte man meinen, dass das hohe R^2 (0.99) darauf hinweist, dass das Modell gut passt. Es ist aber eben sehr wichtig, sich nicht nur auf den Regressionsoutput zu verlassen, sondern die Daten und die Residuen anzuschauen, um Modellabweichungen festzustellen.

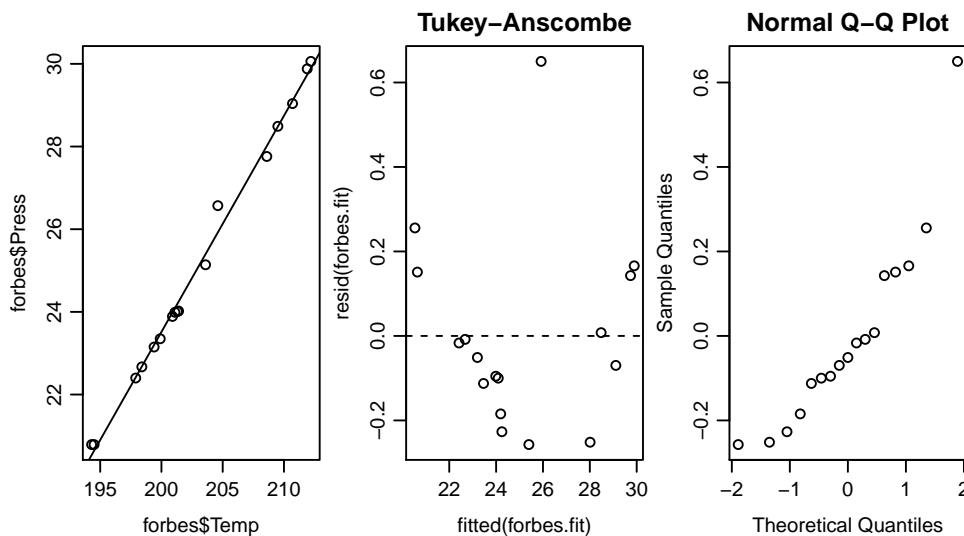
```
> summary(forbes.fit)

Call:
lm(formula = Press ~ Temp, data = forbes)

Residuals:
    Min       1Q   Median       3Q      Max
-0.25717 -0.11246 -0.05102  0.14283  0.64994

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -81.06373   2.05182  -39.51  <2e-16 ***
Temp         0.52289   0.01011   51.74  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2328 on 15 degrees of freedom
Multiple R-squared:  0.9944,    Adjusted R-squared:  0.9941
F-statistic: 2677 on 1 and 15 DF,  p-value: < 2.2e-16
```



- d), e) Die logarithmierten Werte liegen schön auf einer Geraden. Sowohl im Tukey-Anscombe wie auch im Normal-Plot ist ausser dem Ausreisser keine Abweichung von den Modellvoraussetzungen festzustellen.

```
Call:
lm(formula = Logpress ~ Temp, data = forbes)

Residuals:
    Min       1Q   Median       3Q      Max
-0.0073622 -0.0033863 -0.0015865  0.0004322  0.0313139

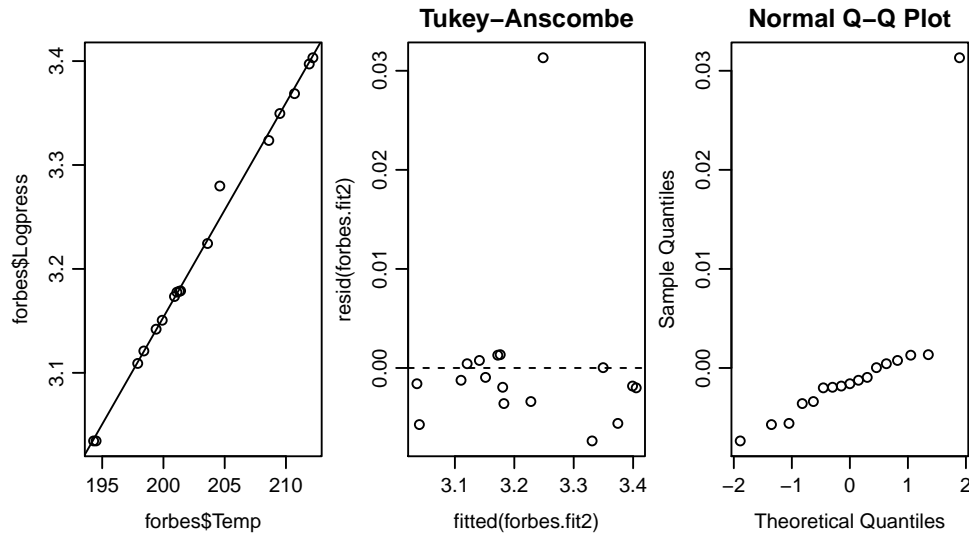
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.9708662  0.0769377  -12.62 2.17e-09 ***
Temp         0.0206224  0.0003789   54.42 < 2e-16 ***
```

```

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.00873 on 15 degrees of freedom
Multiple R-squared:  0.995,    Adjusted R-squared:  0.9946
F-statistic: 2962 on 1 and 15 DF,  p-value: < 2.2e-16

```



- f) Nach Weglassen des Ausreissers sehen alle Plots wunderbar aus. Wie auch in den vorherigen Regressionen ist die erklärende Variable hochsignifikant.

```

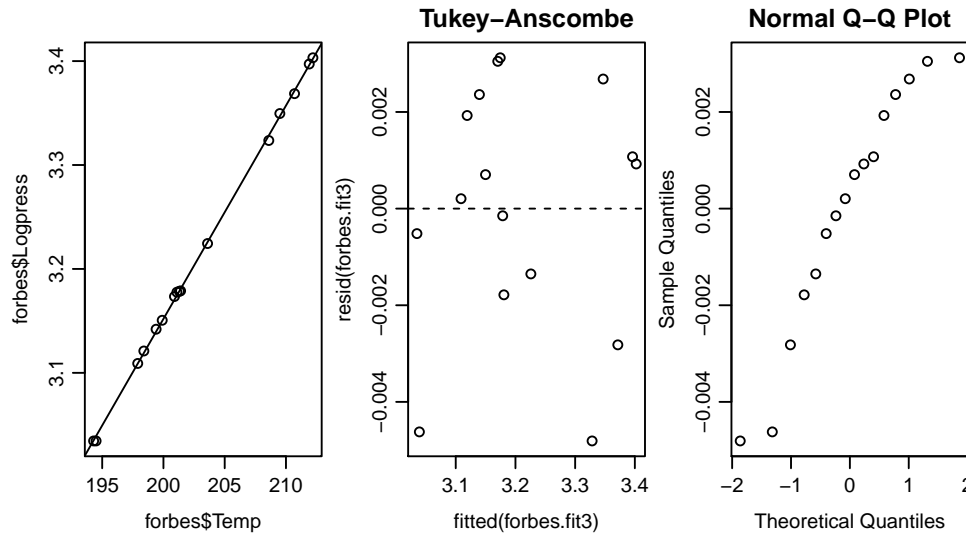
Call:
lm(formula = Logpress ~ Temp, data = forbes[-12, ])

Residuals:
    Min       1Q   Median       3Q      Max
-0.0048082 -0.0014595  0.0004546  0.0020358  0.0031219

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.9517662  0.0231021  -41.2 5.16e-16 ***
Temp         0.0205186  0.0001138  180.2 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.002616 on 14 degrees of freedom
Multiple R-squared:  0.9996,    Adjusted R-squared:  0.9995
F-statistic: 3.249e+04 on 1 and 14 DF,  p-value: < 2.2e-16

```



3. a) Der p-Wert ist extrem klein ($< 2 \cdot 10^{-16}$), also ist β_1 signifikant von Null verschieden.
- b) Das Vertrauensintervall hat die Form $0.18170 \pm t_{11,0.975} \cdot 0.00173$. Es ist $t_{11,0.975} = 2.201$, also haben wir $[0.1779, 0.1855]$. Dies ist Vertrauensintervall ii).
- c) Für 1000m liefert das Modell den Wert $-62.6 + 0.18170 \cdot 1000 = 119.1$. Das Residuum ist daher $136 - 119.1 = 16.9$.
- d) Nein, denn für 100km müssten wir eine Extrapolation verwenden. In diesem Bereich haben wir keine Daten.
- e) Im Output können wir unter **Residual standard error** 62.68 ablesen. Für kleine Distanzen hat das Modell also einen viel zu grossen relativen Fehler!
- f) Man sieht einen sehr deutlichen Trend. Also stimmt das Modell nicht; wir haben einen systematischen Effekt nicht modelliert.
- g) Wir müssen den quadratischen Effekt noch berücksichtigen:

$$\text{Zeit}_i = \beta_0 + \beta_1 \cdot \text{Distanz}_i + \beta_2 \cdot \text{Distanz}_i^2 + \varepsilon_i$$