Konfidenz- und Prognoseintervalle

Konfidenzintervalle für Regressionskoeffizienten

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I}) \text{ und } y_i \sim \mathcal{N}(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}, \sigma^2)$$

$$\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2 C_{jj}) \text{ und } \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \sim t(n-k-1)$$

↪ $\sqrt{\hat{\sigma}^2 C_{jj}}$ wird von R als „standard error“ angegeben

↪ Nach dem in Abschnitt 5 angegebenen Verfahren (vgl. 5-7) wird damit folgendes KI bestimmt:

$(1-\alpha)$-Konfidenzintervall für den Regressionskoeffizienten β_j

$$\hat{\beta}_j - t_{1-\alpha/2, n-k-1} \sqrt{\hat{\sigma}^2 C_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{1-\alpha/2, n-k-1} \sqrt{\hat{\sigma}^2 C_{jj}}$$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.341231   1.096730   2.135 0.044170 *
n.prod      1.615907   0.170735   9.464 3.25e-09 ***
distance    0.014385   0.003613   3.981 0.000631 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.259 on 22 degrees of freedom
Multiple R-squared:  0.9596, Adjusted R-squared:  0.9559
F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16
```

$t_{0.975,22} = 2.074$

<u>95%-Konfidenzintervall:</u>

$$1.615907 - 2.074 \cdot 0.170735 \leq \quad \beta_1 \quad \leq 1.615907 + 2.074 \cdot 0.170735$$

$$1.26181 \leq \quad \beta_1 \quad \leq 1.97001$$

Konfidenzintervalle für den erwarteten Wert am Punkt x_0 (im Datensatz vorhanden)

$$\begin{aligned}
\mathrm{x}_0 &:= (1, x_{01}, \ldots, x_{0k}) \\
\hat{y}_0 &= \mathrm{x}_0^T \hat{\beta} \\
E[\hat{y}_0] &= E(y|\mathrm{x}_0) \\
var(\hat{y}_0) &= \hat{\sigma}^2 \mathrm{x}_0^T (\mathrm{X}^T \mathrm{X})^{-1} \mathrm{x}_0
\end{aligned}$$

$(1-\alpha)$-Konfidenzintervall für den erwarteten Wert

$$\hat{y}_0 - t_{1-\alpha/2,n-k-1}\sqrt{\hat{\sigma}^2 \mathrm{x}_0^T (\mathrm{X}^T \mathrm{X})^{-1} \mathrm{x}_0} \le E(y|\mathrm{x}_0) \le \hat{y}_0 + t_{1-\alpha/2,n-k-1}\sqrt{\cdots}$$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.341231   1.096730   2.135 0.044170 *
n.prod      1.615907   0.170735   9.464 3.25e-09 ***
distance    0.014385   0.003613   3.981 0.000631 ***
```

$t_{0.975,22} = 2.074$

• 95%-Konfidenzintervall für das Geschäft mit 8 nachzufüllenden Produkten und Fußweg von 635 ft (Fall 24)

$$\begin{aligned} x_0 &= (1\ 8\ 635)^T \\ \hat{y}_0 &= x_0^T \hat{\beta} = (1\ 8\ 635) \begin{bmatrix} 2.341231 \\ 1.615907 \\ 0.014385 \end{bmatrix} = 24.40 \text{ Minuten} \end{aligned}$$

```
Residual standard error: 3.259 on 22 degrees of freedom
Multiple R-squared:  0.9596, Adjusted R-squared:  0.9559
F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16
```

$$
\begin{aligned}
var(\hat{y}_0) &= \hat{\sigma}^2 \mathrm{x}_0^T (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{x}_0 \\
&= 10.6239 \cdot (1\ 8\ 635) \times (\mathrm{X}^T\mathrm{X})^{-1} \times (1\ 8\ 635)^T = 1.2814
\end{aligned}
$$

95%-Konfidenzintervall für die erwartete Lieferzeit:

$$
\begin{aligned}
\hat{y}_0 - t_{1-\alpha/2,n-k-1}\sqrt{\hat{\sigma}^2 \mathrm{x}_0^T (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{x}_0} \leq \quad & E(y|\mathrm{x}_0) & \leq \hat{y}_0 + t_{1-\alpha/2,n-k-1}\sqrt{\cdots} \\
24.40 - 2.074 \cdot \sqrt{1.2814} \leq \quad & E(y|\mathrm{x}_0) & \leq 24.40 + 2.074 \cdot \sqrt{1.2814} \\
22.06 \leq \quad & E(y|\mathrm{x}_0) & \leq 26.75
\end{aligned}
$$

95% der so konstruierten Intervalle werden die wahre erwartete Lieferzeit überdecken.

Umsatz an einer Tankstelle

Konstruieren Sie ein 95%-Konfidenzintervall für den erwarteten Gewinn des Tankstellenbetreibers bei einem Kraftstoffumsatz von 7 000€ sowie sonstigen Einnahmen in Höhe von 6 000€.

$$X = \begin{pmatrix} 1 & 6\,000 & 7\,000 \\ 1 & 2\,500 & 6\,500 \\ 1 & 8\,500 & 3\,000 \\ 1 & 6\,500 & 7\,000 \\ 1 & 9\,500 & 7\,500 \end{pmatrix}, \quad Y = \begin{pmatrix} 3\,000 \\ 4\,000 \\ 2\,000 \\ 3\,000 \\ 3\,500 \end{pmatrix}, \quad \hat{\beta} = \begin{pmatrix} 2\,161.889 \\ -0.117 \\ 0.276 \end{pmatrix}$$

$$\begin{aligned} \hat{y}_0 &= x_0^T \cdot \hat{\beta} \\ &= \begin{pmatrix} 1 & 7\,000 & 6\,000 \end{pmatrix} \cdot \begin{pmatrix} 2\,161.889 \\ -0.117 \\ 0.276 \end{pmatrix} = 2\,997.981 \end{aligned}$$

$$\hat{y}_0 = 2\,997.981$$

$$\hat{\sigma} = 510.96$$

$$\Rightarrow \quad \sqrt{\hat{\sigma}^2 \mathrm{x}_0^T (\mathrm{X}^T \mathrm{X})^{-1} \mathrm{x}_0} = 232.53$$

$$t_{1-\alpha/2,n-k-1} = t_{0.975,5-2-1} = 4.3027$$

$$E(y|\mathrm{x}_0) \in [\hat{y}_0 - t_{1-\alpha/2,n-k-1}\sqrt{\cdots}, \hat{y}_0 + t_{1-\alpha/2,n-k-1}\sqrt{\cdots}]$$

$$= [1\,997.483, \quad 3\,998.479]$$

Prognoseintervall für den Wert y_0 an einem neuen Punkt x_0

- das KI für den erwarteten Wert ist nicht geeignet
- benötigt wird eine Wahrscheinlichkeitsaussage für einen konkreten Wert aus der Verteilung

$$\begin{aligned} \kappa &= y_0 - \hat{y}_0 \sim \mathcal{N}(0, var(\kappa)) \\ var(\kappa) &= \underbrace{var(y_0)}_{=\sigma^2} + \underbrace{var(\hat{y}_0)}_{=\sigma^2 \cdot \mathrm{x}_0^T (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{x}_0} = \sigma^2(1 + \mathrm{x}_0^T(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{x}_0) \end{aligned}$$

Durch Schätzen von $var(\kappa)$ Übergang zur t-Verteilung:

$$\begin{aligned} -t_{1-\alpha/2,n-k-1}\sqrt{\hat{\sigma}^2(1 + \mathrm{x}_0^T(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{x}_0)} \leq \quad & y_0 - \hat{y}_0 \quad \leq t_{1-\alpha/2,n-k-1}\sqrt{\cdots} \\ \hat{y}_0 - t_{1-\alpha/2,n-k-1}\sqrt{\hat{\sigma}^2(1 + \mathrm{x}_0^T(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{x}_0)} \leq \quad & y_0 \quad \leq \hat{y}_0 + t_{1-\alpha/2,n-k-1}\sqrt{\cdots} \end{aligned}$$

95%-Prognoseintervall für die Lieferzeit zu einem Geschäft mit 8 nachzufüllenden Produkten und Fußweg von 275 ft.

$$
\begin{aligned}
\hat{y}_0 - t_{1-\alpha/2,n-k-1}\sqrt{\hat{\sigma}^2(1 + x_0^T(X^TX)^{-1}x_0)} \leq \quad & y_0 \quad \leq \hat{y}_0 + t_{1-\alpha/2,n-k-1}\sqrt{\ldots} \\
19.22 - 2.074 \cdot \sqrt{10.6239(1 + 0.05346)} \leq \quad & y_0 \quad \leq 19.22 + 2.074 \cdot \sqrt{\ldots} \\
12.28 \leq \quad & y_0 \quad \leq 26.16
\end{aligned}
$$

- das Prognoseintervall ist immer breiter als das Konfidenzintervall für den erwarteten Wert
- das PI basiert sowohl auf der Unsicherheit des geschätzten Modells als auch auf der Unsicherheit der wahren zukünftigen Werte

Umsatz an einer Tankstelle

Konstruieren Sie nun ein 95%-Prognoseintervall für den Gewinn des Tankstellenbetreibers gegeben den Umsätzen (7 000€und 6 000€) der vorherigen Übungsaufgabe.

$$\hat{y}_0 = 2\,997.981, \quad \hat{\sigma} = 510.96, \quad t_{1-\alpha/2,n-k-1} = t_{0.975,5-2-1} = 4.3027$$

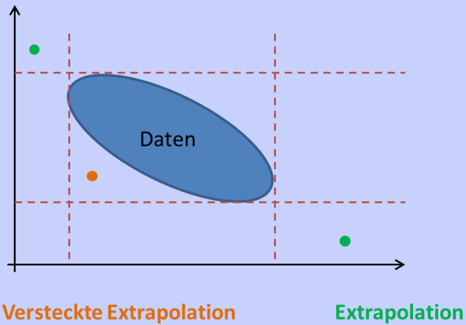
$$\Rightarrow \sqrt{\hat{\sigma}^2 \cdot (1 + x_0^T(X^TX)^{-1}x_0)} = 561.3799$$

$$y \in \left[\hat{y}_0 - t_{1-\alpha/2,n-k-1}\sqrt{\cdots}, \hat{y}_0 + t_{1-\alpha/2,n-k-1}\sqrt{\cdots}\right]$$

$$= [582.56, \quad 5\,413.40]$$

Extrapolation

- Vorsicht bei Extrapolation!



- die Prognosequalität kann außerhalb des Datenbereiches u.U. sehr schlecht sein

Multikollinearität

Vorsicht bei Multikolinearität

- beschreibt lineare bzw. fast lineare Abhängigkeiten zwischen den erklärenden Variablen (d.h. von Spalten der Matrix X)
- exakte lineare Abhängigkeit resultiert in Singularität von $(X^T X)$ $\rightsquigarrow$ $(X^T X)^{-1}$ und Parameterschätzungen können nicht berechnet werden
- Näherungsweise lineare Abhängigkeit resultiert in numerischen Ungenauigkeiten bei der Schätzung
- Solche Modelle haben üblicherweise schlechte Prognosequalität
- Solche Modelle sind sensitiv gegenüber kleinen Änderungen in den erklärenden Variablen
- Einflüsse der Variablen können nicht mehr korrekt unterschieden werden

9.4 Überprüfung der Modellannahmen

Ein lineares Modell kann praktisch immer angepasst werden. Die Frage ist wie gut die Annahmen des Modells – zumindest approximativ – erfüllt sind.

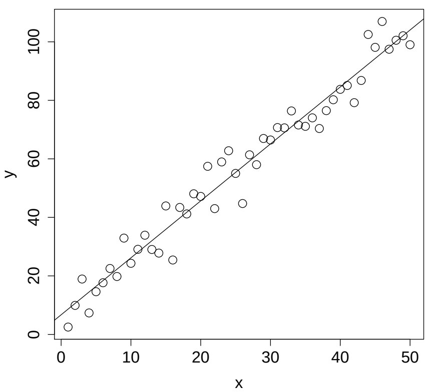
Annahmen:

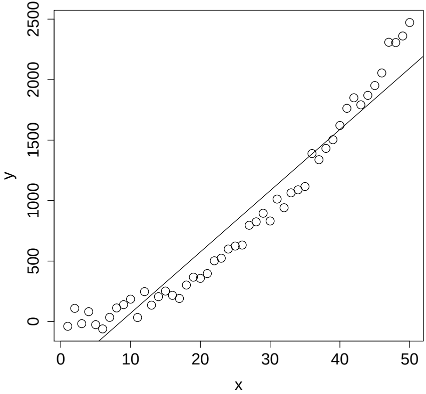
- der Zusammenhang zwischen der abhängigen und den erklärenden Variablen ist linear
- $E(\varepsilon) = 0$
- $var(\varepsilon) = \sigma^2 \mathrm{I}$, d.h. $cov(\varepsilon_i, \varepsilon_j) = 0,\ i \neq j$ und $var(\varepsilon_i) = \sigma^2,\ i = 1, \ldots, n$
- $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I})$

Bei Verletzungen der Modellannahmen können instabile Modelle entstehen sowie wahre Zusammenhänge nicht richtig erkannt werden.

Beispiel

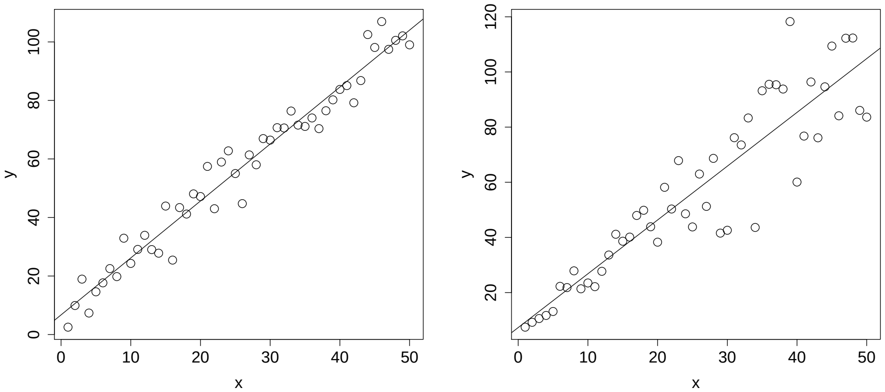
Sind die Modellannahmen in diesen Fällen wohl erfüllt?
Beachte: Wir unterstellen $Y = \beta_0 + \beta_1 X + \varepsilon$.





Beispiel

Sind die Modellannahmen in dem folgenden Fall erfüllt?



Residualanalyse

Residuen werden als Realisationen der Modellfehler ε betrachtet

$$e_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n$$

D.h. die Differenz aus tasächlichem y-Wert und der Vorhersage. Es gilt:

$$E\left(\frac{\sum\limits_{i=1}^{n} e_i^2}{n-k-1}\right) = E\left(\frac{SS_{res}}{n-k-1}\right) = E\left(MS_{res}\right) = \sigma^2$$

- häufig werden Residuen standardisiert
- Ausreisser bzw. extreme Werte können identifiziert werden
- sogenannte Diagnoseplots geben Hinweise auf die Adäquatheit der Modellannahmen

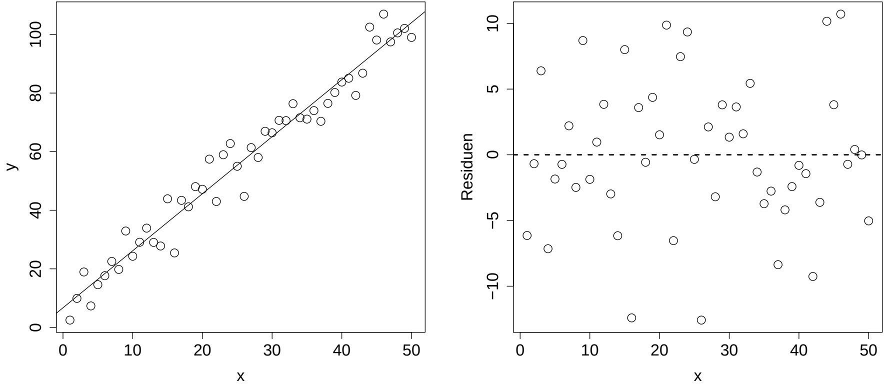
Standardisierte Residuen

Standardisierung mittels der Residualvarianz (siehe reguläre Standardisierung; dadurch ist die Varianz ungefähr bei 1):

$$d_i = \frac{e_i}{\sqrt{MS_{res}}} \quad i = 1, \ldots, n \text{ mit } E(d_i) = 0, var(d_i) \approx 1$$

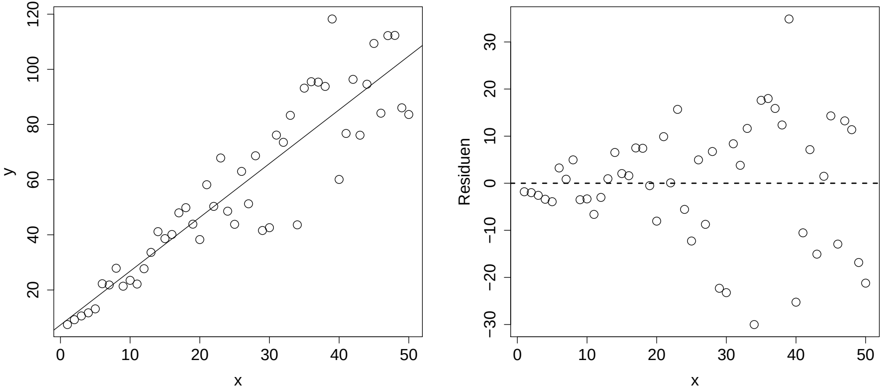
Beispiel - Residuen

Im folgenden Fall ist alles okay:



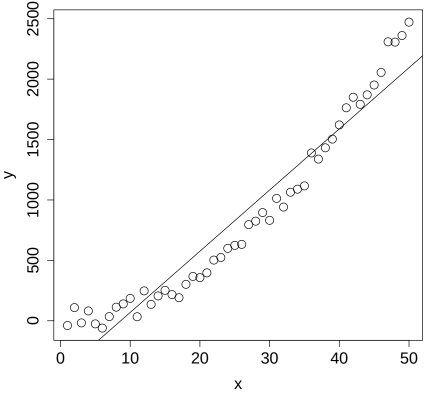
Beispiel - Residuen

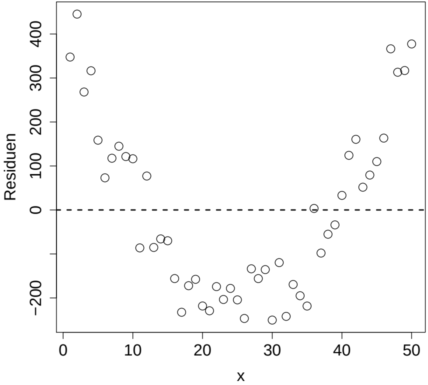
Hier scheint eine Modellannahme nicht erfüllt zu sein. Welche?



Beispiel - Residuen

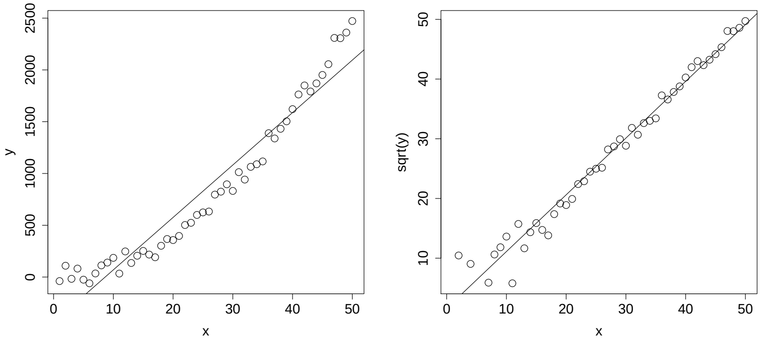
Hier erkennen wir ein eindeutiges Muster was auf nicht-Linearität schließen lässt:





Transformationen als Hilfsmittel

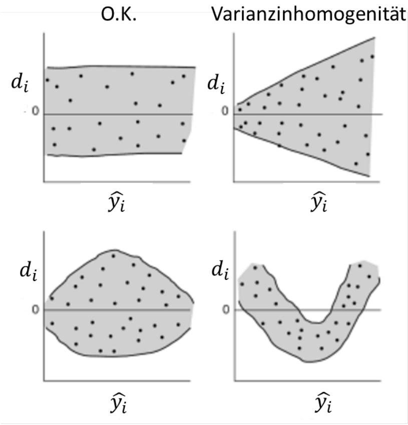
Betrachten wir erneut die offensichtlich nicht-linearen Daten aus dem letzten Beispiel. Hier lässt ein Blick auf die Daten selbst (oder die Residuen) den Schluss zu, dass ein linearer Zusammenhang nicht angebracht. Vielmehr scheint es einen quadratischen Zusammenhang zu geben. Links: originale Daten, rechts: $\tilde{Y} = \sqrt{Y}$



Abschließende Worte zu Residuen

(Standardisierten) Residuen vs. vorhergesagte Werte. Wie hilft uns das?

- verschiedene Abweichungen von den Modellannahmen könnten auffällig werden
- U.U. sind Transformationen oder Hinzunahme weiterer Variablen nötig.
- In der linearen Einfachregression offenbaren Residuen-Plots nicht wirklich mehr als die Plots der Daten selbst. Im multiplen Fall sind sie hingegen ein probates Mittel.



Varianzinhomogenität Nichtlinearität

Einordnung

Mit der Analyse von (standardisierten) Residuen können wir einige Annahmen testen:

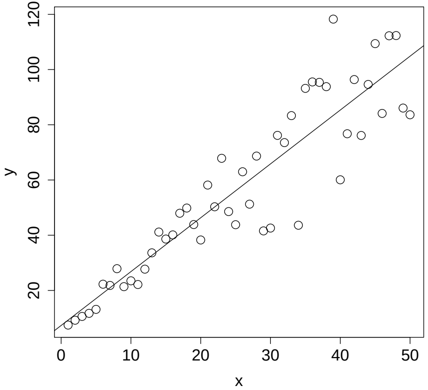
$$E(\varepsilon) = 0 \quad \text{(Im Mittel richtig?)} \quad \checkmark$$
$$var(\varepsilon) = \sigma^2 \mathrm{I} \quad \text{(Unkorreliert und mit gleicher Varianz)} \quad \checkmark$$

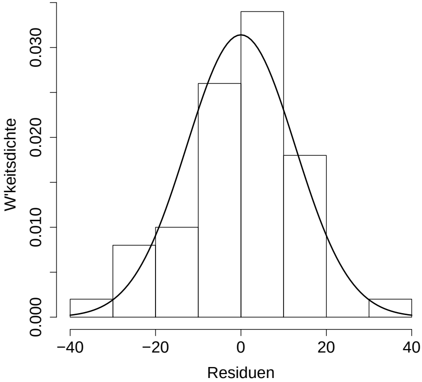
Was fehlt noch? Die Überprüfung auf Normalverteilung:

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I})$$

Erster Ansatz - Histogramme

Wir plotten ein Histogramm der Residuen e_i bzw. d_i, $i = 1, \ldots, n$





Quantile-Quantile-Plots (QQ-Plots)

- Vergleich der realisierten Werte mit Realisationen einer $\mathcal{N}(0,1)$-Verteilung
- Beobachtete Werte d_i werden sortiert: $(d_{(1)} < d_{(2)} < \cdots < d_{(n)})$
- Genau j/n der Werte sind kleiner oder gleich $d_{(j)}$,

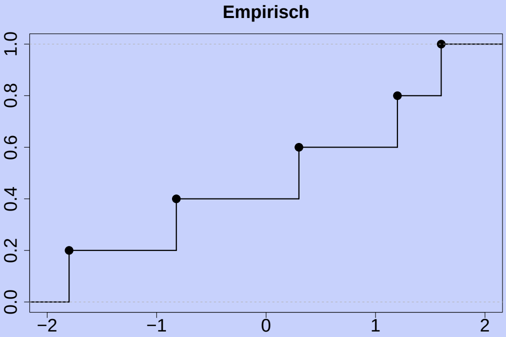
QQ-Plot

- Fasse $d_{(j)}$ als empirisches j/n-Quantil auf und stelle es dem entsprechenden Quantil von $\mathcal{N}(0,1)$ gegenüber
- **Stetigkeitskorrektur**: Anteil j/n wird approximiert durch $j^* = \frac{j-1/2}{n}$.
- Schließlich: n Punktpaare werden gegeneinander aufgetragen:

$$q_j = \left(\Phi^{-1}\left(\frac{j-1/2}{n}\right), d_{(j)}\right)$$

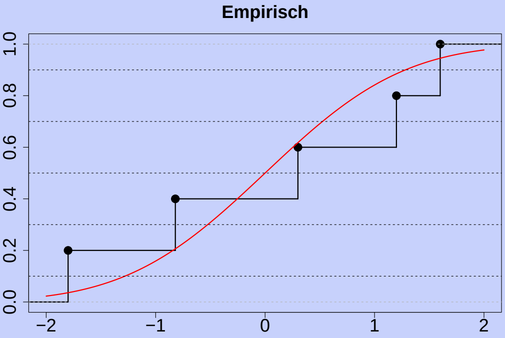
Quantile-Quantile-Plots (QQ-Plots)

d	Ränge	j^*
-1.8	1	0.1
-0.82	2	0.3
0.3	3	0.5
1.2	4	0.7
1.6	5	0.9



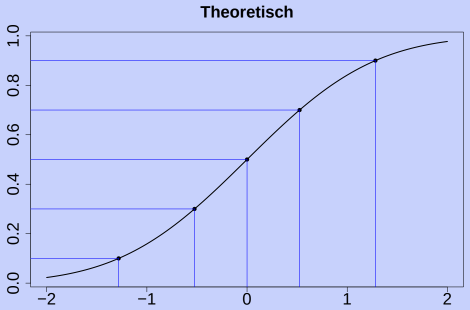
Quantile-Quantile-Plots (QQ-Plots)

d	Ränge	j^*
-1.8	1	0.1
-0.82	2	0.3
0.3	3	0.5
1.2	4	0.7
1.6	5	0.9



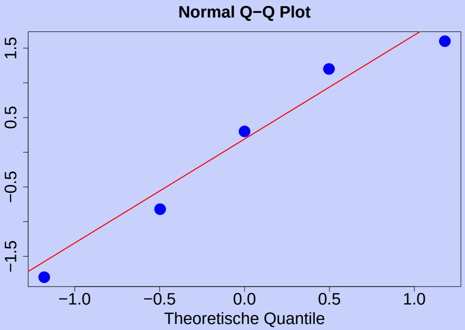
Quantile-Quantile-Plots (QQ-Plots)

d	Ränge	j^*	q (normal)
-1.8	1	0.1	-1.28
-0.82	2	0.3	-0.52
0.3	3	0.5	0.00
1.2	4	0.7	0.52
1.6	5	0.9	1.28



Quantile-Quantile-Plots (QQ-Plots)

d	Ränge	j^*	q (normal)
-1.8	1	0.1	-1.28
-0.82	2	0.3	-0.52
0.3	3	0.5	0.00
1.2	4	0.7	0.52
1.6	5	0.9	1.28

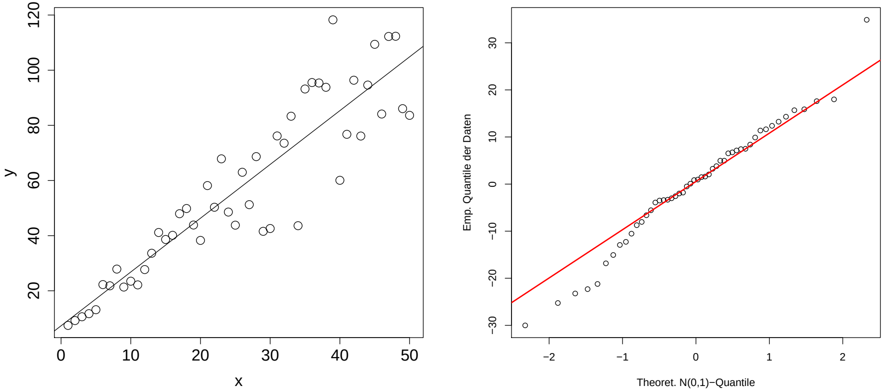


- Plots auf der Originalskala enthalten die folgenden Punktepaare:

$$\left(\Phi^{-1}\left(\frac{j-0.5}{n}\right),\ d_{(j)}\right) \text{ mit der Geraden } y_j = \bar{d} + s_d \cdot d_{(j)}$$

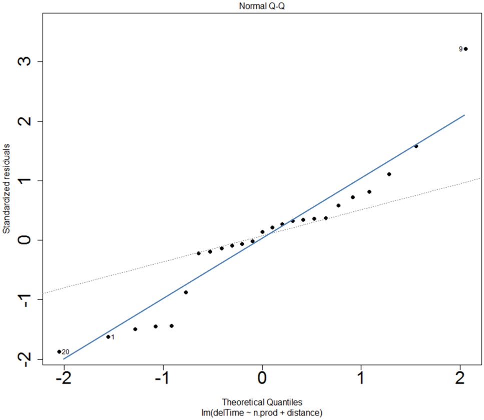
Beispiel: QQ-Plots

Wir zeichnen einen QQ-Plot der Residuen e_i (bzw. d_i), $i = 1, \ldots, n$



Lieferzeiten, QQ-Plot

```
lm.deli = lm(delTime ~ n.prod + distance, data = delivery)

plot(lm.deli)
```



Umsatz an einer Tankstelle

Untersuchen Sie, ob für das lineare Modell, welches den Gewinn des Tankstellenbetreibers in Abhängigkeit der Umsätze modelliert, die Normalverteilungsannahme der Residuen erfüllt ist.

$$X = \begin{pmatrix} 1 & 6\,000 & 7\,000 \\ 1 & 2\,500 & 6\,500 \\ 1 & 8\,500 & 3\,000 \\ 1 & 6\,500 & 7\,000 \\ 1 & 9\,500 & 7\,500 \end{pmatrix}, \quad Y = \begin{pmatrix} 3\,000 \\ 4\,000 \\ 2\,000 \\ 3\,000 \\ 3\,500 \end{pmatrix}, \quad \hat{\beta} = \begin{pmatrix} 2\,161.889 \\ -0.117 \\ 0.276 \end{pmatrix}$$

$$\hat{y} = X \cdot \hat{\beta} = \begin{pmatrix} 3\,390.997 \\ 3\,662.798 \\ 1\,994.550 \\ 3\,332.459 \\ 3\,119.197 \end{pmatrix} \quad \Rightarrow \quad e = y - \hat{y} = \begin{pmatrix} -390.997 \\ 337.202 \\ 5.450 \\ -332.459 \\ 380.803 \end{pmatrix}$$

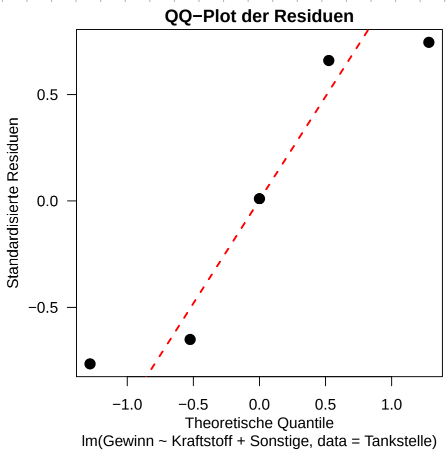
$$e = y - \hat{y} = \begin{pmatrix} -390.997 \\ 337.202 \\ 5.450 \\ -332.459 \\ 380.803 \end{pmatrix}$$

$$\Rightarrow \quad MS_{res} = \frac{\sum_{i=1}^{5} e_i^2}{n-k-1} = \frac{522\,153.8}{5-2-1} = 261\,076.9$$

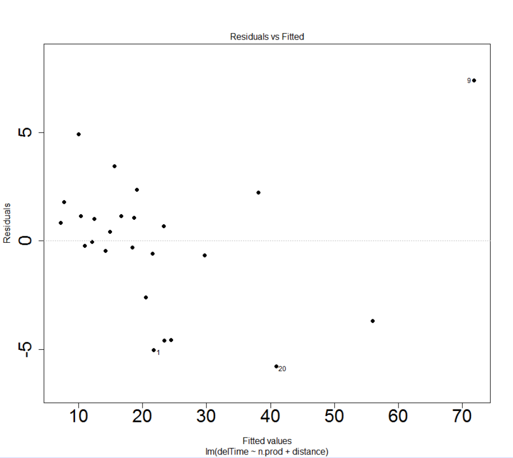
$$d = \frac{e}{\sqrt{MS_{res}}} = \begin{pmatrix} -0.765 \\ 0.660 \\ 0.011 \\ -0.651 \\ 0.745 \end{pmatrix} \quad \rightsquigarrow \quad \text{sortieren: } \tilde{d} = \begin{pmatrix} -0.765 \\ -0.651 \\ 0.011 \\ 0.660 \\ 0.745 \end{pmatrix}$$

Die Werte der theoretischen Quantile q(normal) stimmen mit denen des vorherigen Beispiels überein (da diese lediglich von der Anzahl an Beobachtungen abhängen).

Zugehöriger QQ-Plot:



Lieferzeiten



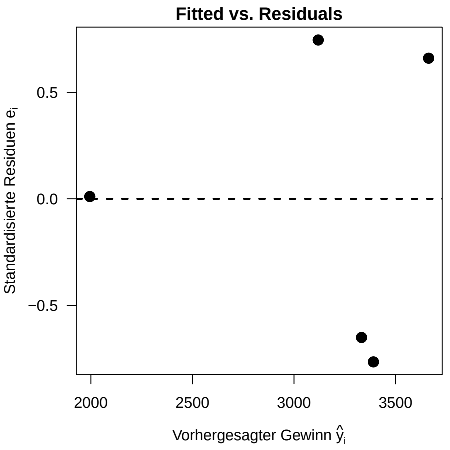
Umsatz an einer Tankstelle

Ist ein Zusammenhang zwischen den prognostizierten Gewinnen ($\hat{y}$) und den (standardisierten) Residuen des linearen Modells (e) erkennbar?

$$\hat{y} = X \cdot \hat{\beta} = \begin{pmatrix} 3\,390.997 \\ 3\,662.798 \\ 1\,994.550 \\ 3\,332.459 \\ 3\,119.197 \end{pmatrix}$$

$$e = y - \hat{y} = \begin{pmatrix} -390.997 \\ 337.202 \\ 5.450 \\ -332.459 \\ 380.803 \end{pmatrix}$$

Zugehöriger Fitted-vs.-Residual-Plot:



9.5 Indikatorvariablen

↪ neben quantitativen erklärenden Variablen (z.B. Temperatur, Gewicht, etc.) können auch qualitative oder kategorielle Variablen auftreten (z.B. Geschlecht, Schicht (morgens, mittags, abends), etc.)

↪ Modellierung mit Hilfe von Indikator- bzw. Dummyvariablen.

Beispiel: Untersuchung des Einflusses der Art von Schneidwerkzeug und der Umdrehungen/min. einer Drehmaschine auf die Lebensdauer des Werkzeugs. [5]

- □ Y: Lebensdauer des Schneidwerkzeuges
- □ X_1: Umdrehungen pro Minute
- □ X_2: $\begin{cases} 0 & \text{Werkzeugtyp } A \\ 1 & \text{Werkzeugtyp } B \end{cases}$

	y	x1	x2		y	x1	x2
1	18.73	610	0	11	30.16	670	1
2	14.52	950	0	12	27.09	770	1
3	17.43	720	0	13	25.40	880	1
4	14.54	840	0	14	26.05	1000	1
5	13.44	980	0	15	33.49	760	1
6	24.39	530	0	16	35.62	590	1
7	13.34	680	0	17	26.07	910	1
8	22.71	540	0	18	36.78	650	1
9	12.68	890	0	19	34.95	810	1
10	19.32	730	0	20	43.67	500	1

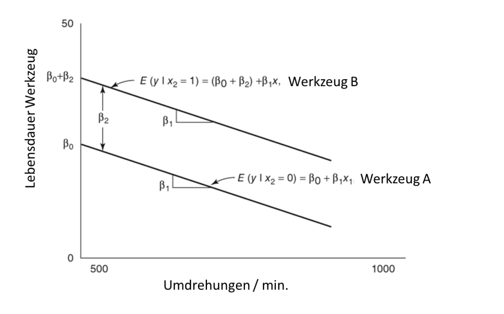
↪ die Reihenfolge der Zuweisung bei X_2 ist unerheblich.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Für die verschiedenen Werkzeugtypen ergibt sich:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 \cdot 0 + \varepsilon = \beta_0 + \beta_1 X_1 + \varepsilon \qquad \text{für Typ A: } X_2 = 0$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 \cdot 1 + \varepsilon = (\beta_0 + \beta_2) + \beta_1 X_1 + \varepsilon \qquad \text{für Typ B: } X_2 = 1$$



[5]

$\hookrightarrow$ parallele Regressionsgeraden

$\hookrightarrow$ Fehlervarianzen werden identisch angenommen

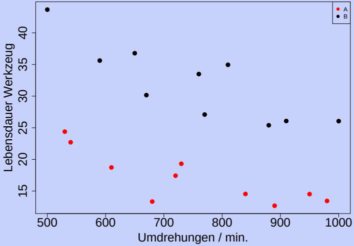
$\hookrightarrow$ β_2 repräsentiert den erwarteten Unterschied in der Lebensdauer, wenn von Typ A auf Typ B gewechselt wird.

$\hookrightarrow$ Generalisierung auf a verschiedene Typen möglich, es werden $a-1$ Indikatorvariablen benötigt.

X_2	X_3	Typ
0	0	Beobachtung von Werkzeugtyp A
1	0	Beobachtung von Werkzeugtyp B
0	1	Beobachtung von Werkzeugtyp C

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

Schätzung des Regressionsmodells:



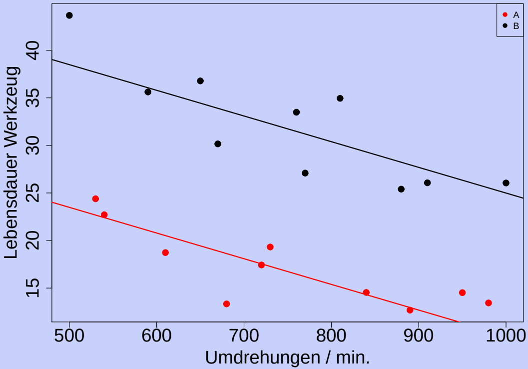

```
17
18 summary(lm(y ~ x1+factor(x2),data=ToolLife))
19
17:1  (Top Level) ÷

Console  Terminal ×  Jobs ×
~/

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.98560    3.51038  10.536 7.16e-09 ***
x1          -0.02661    0.00452  -5.887 1.79e-05 ***
factor(x2)1 15.00425    1.35967  11.035 3.59e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.039 on 17 degrees of freedom
Multiple R-squared:  0.9003,	Adjusted R-squared:  0.8886
F-statistic: 76.75 on 2 and 17 DF,  p-value: 3.086e-09
```

$$\hat{Y} = 36.986 - 0.027X_1 + 15.004X_2$$

$$\hat{Y} = \beta_0 + \beta_1 X_1 = 36.986 - 0.027X_1 \qquad \text{für Typ A: } X_2 = 0$$

$$\hat{Y} = (\beta_0 + \beta_2) + \beta_1 X_1$$

$$= (36.986 + 15.004) - 0.027X_1 = 51.99 - 0.027X_1 \qquad \text{für Typ B: } X_2 = 1$$



$$\hat{Y} = 36.986 - 0.027X_1 + 15.004X_2$$

$$\hat{Y} = 36.986 - 0.027X_1 \quad \text{für Typ A}$$

$$\hat{Y} = 51.99 - 0.027X_1 \quad \text{für Typ B}$$

↪ β_2 repräsentiert den erwarteten Unterschied in der Lebensdauer, wenn von Typ A auf Typ B gewechselt wird.

↪ 95%-Konfidenzintervall für β_2

$$\begin{aligned} \hat{\beta}_2 - t_{0.975,17} \cdot se(\hat{\beta}_2) \leq \quad & \beta_2 \quad \leq \hat{\beta}_2 + t_{0.975,17} \cdot se(\hat{\beta}_2) \\ \Leftrightarrow \quad 15.004 - 2.110 \cdot 1.360 \leq \quad & \beta_2 \quad \leq 15.004 + 2.110 \cdot 1.360 \\ \Leftrightarrow \quad 12.135 \leq \quad & \beta_2 \quad \leq 17.873 \end{aligned}$$

Aspekte der gemeinsamen Modellierung:

\+ einfacher zu interpretieren

↪ Schätzung einer gemeinsamen Fehlervarianz
  \+ mehr Freiheitsgrade für die Schätzung
  − nur sinnvoll, wenn Annahme gleicher Fehlervarianz korrekt ist.

\+ Änderung in der Modellierung möglich, um unterschiedliche Steigungen zu erhalten.

Annahme: Unterschiede im Achsenabschnitt und in der Steigung der Regressionsgeraden je nach Werkzeugtyp.

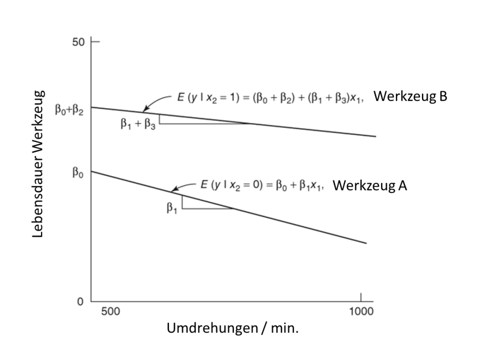
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

Für Typ A (d.h. $X_2 = 0$):

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 \cdot 0 + \beta_3 X_1 \cdot 0 + \varepsilon \\ &= \beta_0 + \beta_1 X_1 + \varepsilon \end{aligned}$$

Für Typ B (d.h. $X_2 = 1$):

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 \cdot 1 + \beta_3 X_1 \cdot 1 + \varepsilon \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 + \varepsilon \end{aligned}$$



$\hookrightarrow$ β_2 repräsentiert den erwarteten Unterschied im Achsenabschnitt, wenn von Typ A auf Typ B gewechselt wird.

$\hookrightarrow$ β_3 repräsentiert den erwarteten Unterschied in der Steigung, wenn von Typ A auf Typ B gewechselt wird.

$\hookrightarrow$ Tests können auf dem gemeinsamen Modell durchgeführt werden

Schätzung des Regressionsmodells:

$$\hat{Y} = 32.775 - 0.021 \cdot X_1 + 23.971 \cdot X_2 - 0.012 \cdot X_1 X_2, \quad \text{d.h.}$$

$$\hat{Y} = 32.775 - 0.021 \cdot X_1 \quad \text{für Typ A}$$

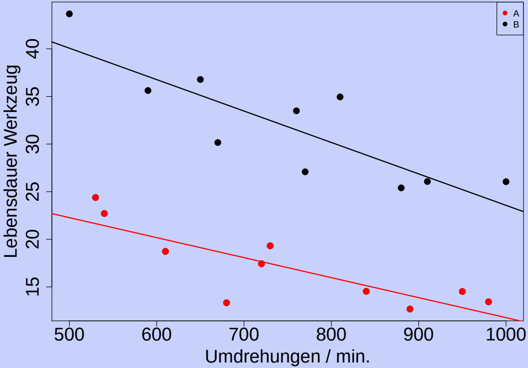
$$\hat{Y} = (32.775 + 23.791) + (-0.021 - 0.012) \cdot X_1$$

$$= 56.566 - 0.033 \cdot X_1 \quad \text{für Typ B}$$

```
summary(lm(y ~ x1*factor(x2),data=ToolLife))

~/
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     32.774760   4.633472   7.073 2.63e-06 ***
x1              -0.020970   0.006074  -3.452  0.00328 **
factor(x2)1     23.970593   6.768973   3.541  0.00272 **
x1:factor(x2)1  -0.011944   0.008842  -1.351  0.19553
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.968 on 16 degrees of freedom
Multiple R-squared:  0.9105,	Adjusted R-squared:  0.8937
F-statistic: 54.25 on 3 and 16 DF,  p-value: 1.319e-08
```

$$\hat{Y} = 32.775 - 0.021 \cdot X_1 \qquad \text{für Typ A}$$

$$\hat{Y} = 56.566 - 0.033 \cdot X_1 \qquad \text{für Typ B}$$



$H_0 : \beta_3 = 0 \quad \text{vs.} \quad \beta_3 \neq 0 \quad \rightarrow$ wird nicht abgelehnt.

Variablen mit mehr als 2 Levels

$\hookrightarrow$ 4 Levels können über 3 Indikatorvariablen realisiert werden.

X_2	X_3	X_4
0	0	0
1	0	0
0	1	0
0	0	1

$\hookrightarrow$ Regressionsmodell:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

$\hookrightarrow$ durch Hinzufügen von Interaktionen können verschiedene Steigungen realisiert werden

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_1 X_2 + \beta_6 X_1 X_3 + \beta_7 X_1 X_4 + \varepsilon$$

Beispiel Klimasysteme, vgl. [5]

Y : Energieverbrauch in den Sommermonaten Juni-September (kw/h)

X_1 : Größe des Hauses (ft^2)

$$X_2 : \text{Klimasystem} = \begin{cases} 1 & \text{kein Klimasystem} \\ 2 & \text{Fensterklimasysteme} \\ 3 & \text{Wärmepumpe} \\ 4 & \text{zentrale Klimaanlage} \end{cases}$$

$\hookrightarrow$ Modellierung von X_2 mit Dummyvariablen:

Klimasystem	X_{22}	X_{23}	X_{24}
kein Klimasystem	0	0	0
Fensterklimasysteme	1	0	0
Wärmepumpe	0	1	0
zentrale Klimaanlage	0	0	1

Eine Variable X_{21} wird nicht benötigt: Mit X_{21} wäre zudem $\mathrm{X}^T\mathrm{X}$ nicht invertierbar (Überparametrisierung). Andere Formen der Kodierung sind möglich, im Falle von Überparametrisierung werden die Lösungen durch zusätzliche (lineare) Nebenbedingungen an die Parameter eindeutig gemacht.

Modellierungsmöglichkeit I

$$Y = \beta_0 + \beta_1 X_1 + \beta_{22} X_{22} + \beta_{23} X_{23} + \beta_{24} X_{24} + \varepsilon$$

Wie sehen die resultierenden individuellen Regressionsmodelle aus?

$$\begin{aligned}
Y &= \beta_0 + \beta_1 X_1 + \varepsilon && \text{für } X_2 = 1 \\
Y &= (\beta_0 + \beta_{22}) + \beta_1 X_1 + \varepsilon && \text{für } X_2 = 2 \\
Y &= (\beta_0 + \beta_{23}) + \beta_1 X_1 + \varepsilon && \text{für } X_2 = 3 \\
Y &= (\beta_0 + \beta_{24}) + \beta_1 X_1 + \varepsilon && \text{für } X_2 = 4
\end{aligned}$$

$\hookrightarrow$ Annahme: linearer Grundzusammenhang, Steigung und Varianz hängen nicht vom Klimasystem ab

$\hookrightarrow$ β_{22}, β_{23} und β_{24} beschreiben den Effekt des jeweiligen Systems

Modellierungsmöglichkeit I

$$Y = \beta_0 + \beta_1 X_1 + \beta_{22} X_{22} + \beta_{23} X_{23} + \beta_{24} X_{24} + \varepsilon$$

Wie sehen die resultierenden individuellen Regressionsmodelle aus?

$$\begin{aligned}
Y &= \beta_0 + \beta_1 X_1 + \varepsilon && \text{für } X_2 = 1\\
Y &= (\beta_0 + \beta_{22}) + \beta_1 X_1 + \varepsilon && \text{für } X_2 = 2\\
Y &= (\beta_0 + \beta_{23}) + \beta_1 X_1 + \varepsilon && \text{für } X_2 = 3\\
Y &= (\beta_0 + \beta_{24}) + \beta_1 X_1 + \varepsilon && \text{für } X_2 = 4
\end{aligned}$$

$\hookrightarrow$ relative Effizienz kann gemessen werden, z.B. misst $\beta_{22} - \beta_{24}$ die relative Effizienz von Fensterklimasystemen im Vergleich mit einer zentralen Klimaanlage

$\hookrightarrow$ Modellgleichung in `R`:
`y ~ x1 + factor(x2)` (R generiert dann die Dummy-Variablen)

Modellierungsmöglichkeit II

↪ Sind die Annahmen realistisch?

↪ wahrscheinlicher ist eine Interaktion zwischen Hausgröße und Typ des Klimasystems

↪ d.h. unterschiedliche Steigungen realistischer

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_{22} X_{22} + \beta_{23} X_{23} + \beta_{24} X_{24} \\ &\quad + \beta_{122} X_1 X_{22} + \beta_{123} X_1 X_{23} + \beta_{124} X_1 X_{24} + \varepsilon \end{aligned}$$

Modellierungsmöglichkeit II

$\hookrightarrow$ d.h. unterschiedliche Steigungen realistischer

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_{22} X_{22} + \beta_{23} X_{23} + \beta_{24} X_{24} \\ &\quad + \beta_{122} X_1 X_{22} + \beta_{123} X_1 X_{23} + \beta_{124} X_1 X_{24} + \varepsilon \end{aligned}$$

Wie sehen die resultierenden individuellen Regressionsmodelle aus?

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \varepsilon && \text{für } X_2 = 1 \\ Y &= (\beta_0 + \beta_{22}) + (\beta_1 + \beta_{122}) X_1 + \varepsilon && \text{für } X_2 = 2 \\ Y &= (\beta_0 + \beta_{23}) + (\beta_1 + \beta_{123}) X_1 + \varepsilon && \text{für } X_2 = 3 \\ Y &= (\beta_0 + \beta_{24}) + (\beta_1 + \beta_{124}) X_1 + \varepsilon && \text{für } X_2 = 4 \end{aligned}$$

Modellgleichung in R:

`y ~ x1 * factor(x2)`

Übung: Im Herzviertel wurden Wohnfläche, Zimmerzahl und Kaltmiete von 1-4-Zimmer-Wohnungen erhoben:

Whg	qm	Zi	KM	Whg	qm	Zi	KM	Whg	qm	Zi	KM	Whg	qm	Zi	KM
1	40	1	590	9	50	2	590	17	115	3	1030	25	120	4	960
2	35	1	580	10	45	2	560	18	115	3	1020	26	105	4	870
3	35	1	540	11	50	2	600	19	100	3	860	27	110	4	960
4	55	1	820	12	45	2	540	20	90	3	810	28	120	4	1010
5	40	1	630	13	80	2	930	21	95	3	860	29	105	4	830
6	40	1	620	14	75	2	860	22	75	3	610	30	120	4	950
7	45	1	720	15	105	3	940	23	105	3	890				
8	45	2	540	16	115	3	1040	24	90	3	850				

Die Kaltmiete soll anhand der Wohnfläche und Zimmerzahl erklärt werden. Wie lautet ein mögliches Regressionsmodell? Stellen Sie sinnvolle Hypothesen auf

Ohne Interaktionen: $KM = \beta_0 + \beta_1 q + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4 + V$
Dabei sind Z_2, Z_3, Z_4 Dummy-Variablen mit $Z_i = 1$ wenn die Wohnung i Zimmer hat und $Z_i = 0$ sonst. Diese Auswahl der Dummy-Variablen wird von R gesetzt.
Mit Interaktionen: $KM = \beta_0 + \beta_1 q + \sum_{i=2}^4 \beta_i * Z_i + \sum_{i=21}^4 \gamma_i Z_i q + V$

$H_0 : \beta_1 = 0$ (Miete unabhängig von Wohnfläche)

$H_0 : \beta_2 = \cdots = \beta_4 = 0$ (Sockelmiete unabhängig von Zimmerzahl)

$H_0 : \gamma_2 = \cdots = \gamma_4 = 0$ (qm-Miete unabhängig von Zimmerzahl)