HTW Chur
Hochschule für Technik und Wirtschaft
University of Applied Sciences

Regressionsdiagnostik

Quantitative Forschungsmethoden

Prof. Dr. Franz Kronthaler

Lernziele

- verstehen was unverzerrte und effiziente Schätzergebnisse sind
- kennen der Gauss-Markov-Annahmen
- wissen wie die Gauss-Markov-Annahmen getestet werden
- kennen einfacher Lösungsmöglichkeiten bei Verletzung der Gauss-Markov-Annahmen
- wissen wie Ausreisser identifiziert werden
- wissen wie die Ergebnisse der Regressionsanalyse validiert werden

Einführung

- in der letzten Vorlesung haben wir die Regressionsfunktion aus einer Stichprobe geschätzt

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_J X_J$$

- die geschätzten Koeffizienten b_j können als Ausprägungen von Zufallsvariablen aufgefasst werden, sie werden mit Hilfe einer von vielen möglichen Stichproben geschätzt
- entsprechend variieren die Koeffizienten b_j um ihren wahren Wert β_j, wobei zwei Eigenschaften erfüllt sein müssen, damit bestmögliche Ergebnisse erzielt werden
  - b_j sollen unverzerrt geschätzt werden, d. h. der Erwartungswert von b_j ist gleich β_j $E\left(b_j\right) = \beta_j$
  - b_j sollen effizient geschätzt werden, d. h. die Koeffizienten werden mit der kleinstmöglichen Streuung geschätzt
- die geschätzte Funktion lässt sich als Realisation einer «wahren» Funktion interpretieren mit den unbekannten Parametern β_1, β_2,..., β_J und der **Zufallsgrösse** u

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_J X_J + u$$

Gauss-Markov-Annahmen GM

- die Methode der Kleinsten Quadrate ist der **B**este **L**ineare **U**nverzerrte und **E**ffiziente Schätzer, wenn eine Reihe von Annahmen erfüllt ist (BLUE)
- die Annahmen sind die sogenannten Gauss-Markow-Annahmen (GM)
- wir haben insgesamt sieben Annahmen, die erfüllt sein müssen
  - **GM 1:** Die Regressionsfunktion ist richtig spezifiziert!

    $$y_i = \beta_0 + \sum \beta_j \times x_{ji} + u_i$$

    - sie enthält alle relevanten unabhängigen Variablen
    - sie enthält keine nicht-relevanten Variablen
    - sie ist linear in ihren Parametern β_j
    - Ausreisser sind berücksichtigt

Gauss-Markov-Annahmen GM

- **GM 2:** Die Störgrössen haben den Erwartungswert Null!

  $E(u_i) = 0$

- **GM 3:** Die Störgrösse ist nicht korreliert mit den unabhängigen Variablen!

  $Cov\left(u_i, x_{ji}\right) = 0$

- **GM 4:** Die Varianz der Störgrösse ist konstant (keine Heteroskedastizität)!

  $Var(u_i) = \sigma^2$

- **GM 5:** Die Störgrössen sind unkorreliert!

  $Cov(u_i, u_{i+r}) = 0$

- **GM 6:** Zwischen den unabhängigen Variablen X_j besteht keine lineare Abhängigkeit (keine Multikollinearität)!

- **GM 7:** Die Störgrössen u_i sind normalverteilt!

Gauss-Markov-Annahmen GM

- die meisten dieser Annahmen analysieren wir mit Hilfe der Abweichung der beobachteten Y-Werte von den geschätzten $\hat{Y}$-Werten, kurz auch Residuen genannt
- wir weisen den R Commander an, uns die Werte zu unserem Datensatz hinzuzufügen
- im Folgenden (GM 1 bis GM 7) greifen wir auf das Beispiel der letzten Vorlesung reg_sales.RData zurück

  # lade den Datensatz «reg_sales.RData»

  # schätze das Regressionsmodell aus der letzten Vorlesung

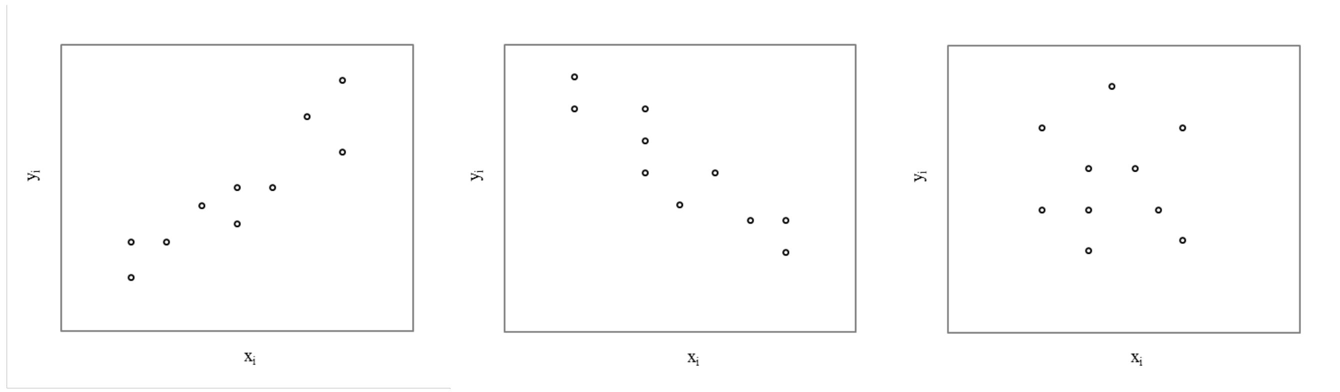
  # füge dem Datensatz die vorhergesagten Werte, die Residuen, die studentisierten Residuen, die Hebelwerte (hat values), Cook's Distanz und die Fallbeschriftung hinzu (nutze am Besten die Oberfläche des R Commanders)

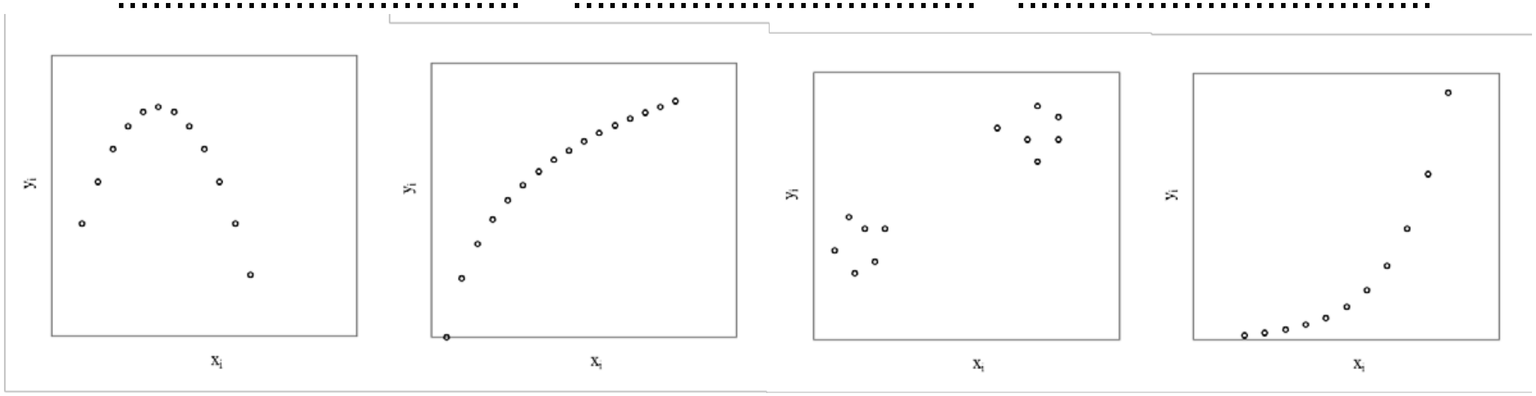
zu GM 1

- die Regressionsfunktion ist richtig spezifiziert
  - richtig spezifiziert bedeutet, dass alle relevanten unabhängigen Variablen in der Regressionsfunktion aufgenommen sind
  - dies erfordert, dass bei der Aufstellung der Regressionsfunktion theoretisch sauber gearbeitet wurde
- die Regressionsfunktion ist linear in ihren Parametern β_j
  - die Linearität können wir mit Hilfe von Streudiagrammen zwischen der abhängigen Variable Y und den unabhängigen Variablen X_j begutachten
- ist GM 1 verletzt sind die geschätzten Koeffizienten b_j möglicherweise verzerrt
- zur Überprüfung der Linearität werden die Streudiagramme zwischen abhängiger Variable und unabhängigen Variablen genutzt

zu GM 1

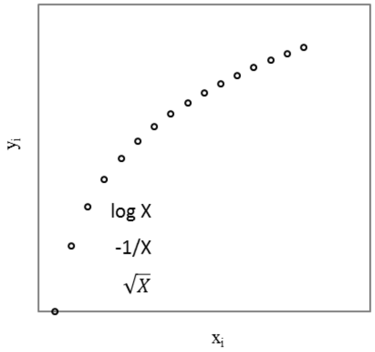
- Beispiele für lineare und nicht-lineare Beziehungen

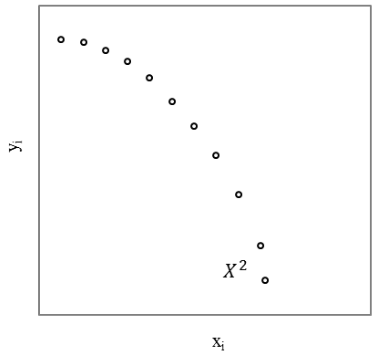


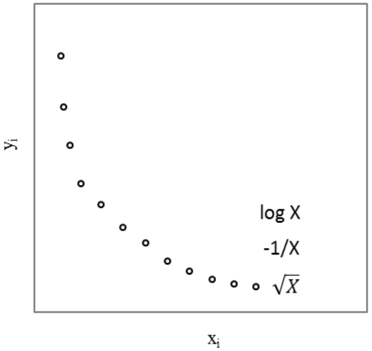


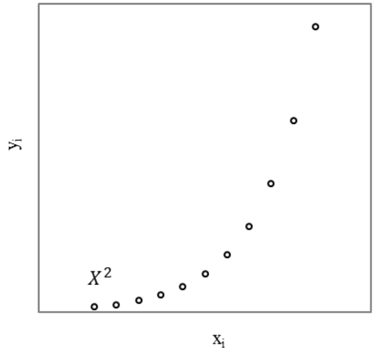
zu GM 1

- Lösungsmöglichkeiten bei Verletzung der Linearität
  - Transformation der unabhängigen Variable
  - Versuch und Irrtum

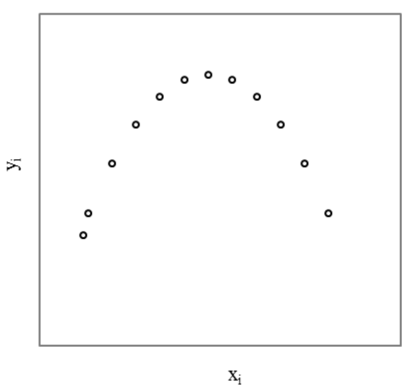


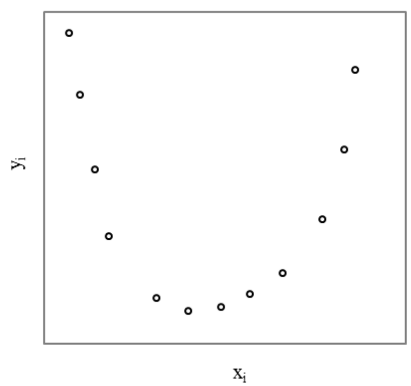






$z.B.\ \hat{Y} = b_0 + b_1 \times \sqrt{X}$ ....





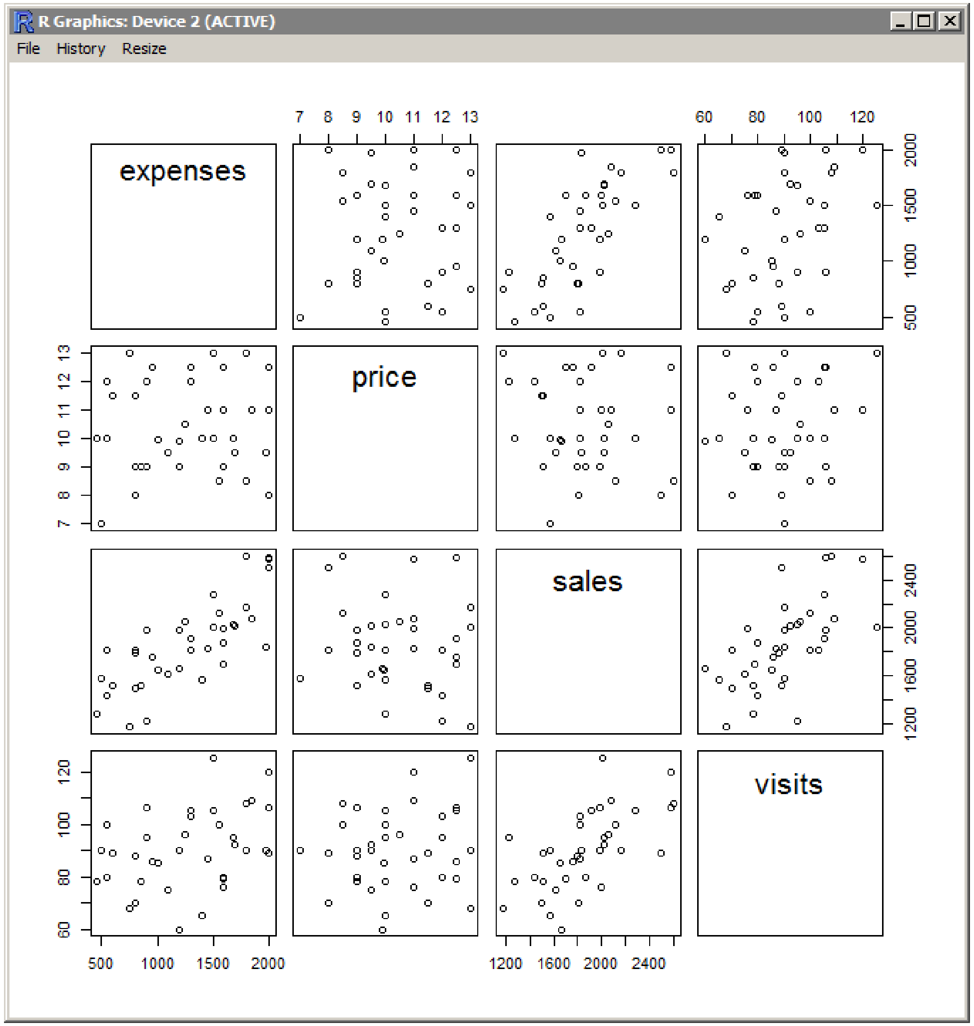
....

Beispiel GM 1

- zeichne die Streudiagramme zwischen abhängiger und unabhängiger Variable (nutze ggf. den Befehl «Scatterplot matrix» (achte darauf die reinen Streudiagramme zu zeichnen ohne Zusatzinformationen)

```
# nutze die Benutzeroberfläche des R Commanders oder die folgenden Befehle

scatterplot(sales~price, reg.line=FALSE, smooth=FALSE, spread=FALSE, id.method='mahal', id.n = 2, boxplots=FALSE, span=0.5, data=reg_sales)

scatterplot(sales~visits, reg.line=FALSE, smooth=FALSE, spread=FALSE, id.method='mahal', id.n = 2, boxplots=FALSE, span=0.5, data=reg_sales)

scatterplot(sales~expenses, reg.line=FALSE, smooth=FALSE, spread=FALSE, id.method='mahal', id.n = 2, boxplots=FALSE, span=0.5, data=reg_sales)

# oder

scatterplotMatrix(~expenses+price+sales+visits, reg.line=FALSE, smooth=FALSE, spread=FALSE, span=0.5, id.n=0, diagonal = 'none', data=reg_sales)
```

Beispiel GM 1



zu GM 2

- die Störgrössen haben den Erwartungswert Null

  $E(u_i) = 0$

  - wenn alle systematischen Einflussgrößen (unabhängige Variablen) im Modell erfasst sind, enthält die Störvariable nur zufällige Effekte und die Schwankungen gleichen sich im Mittel aus
  - die Annahme erfordert, dass die Schwankungen nicht nur global sondern auch lokal null sind
- ist GM 2 verletzt ist b_0 verzerrt
- zur Aufdeckung tragen wir die studentisierten Residuen e_i gegenüber den geschätzten Werten $\hat{y}_i$ ab und teilen die Grafik in Intervalle ein

zu GM 2

- Skizze Verletzung von GM 2

Skizze keine Verletzung von GM 2

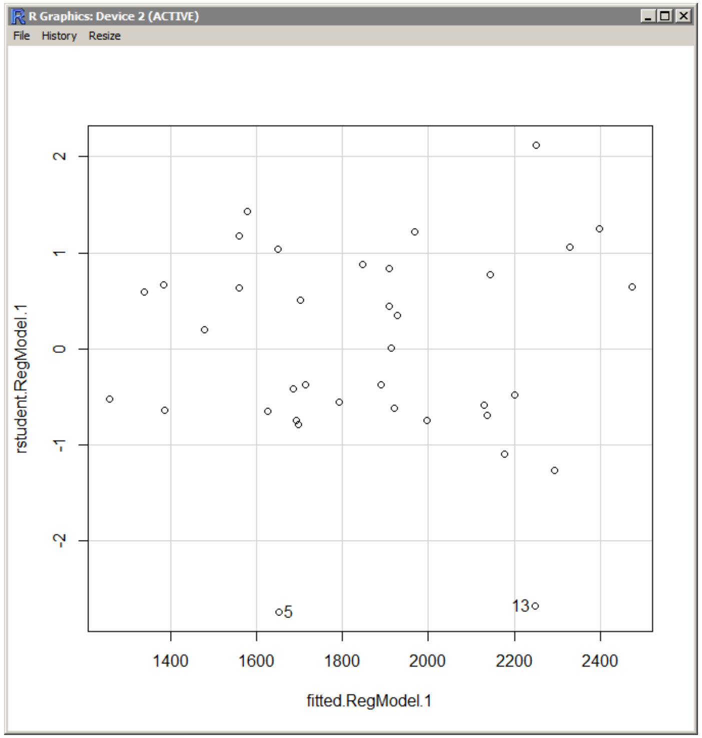
- eine Verletzung von GM 2 kann in der Regel auf das Fehlen wichtiger unabhängiger Variablen zurückgeführt werden
- ggf. ist eine Verletzung von GM 2 auch auf Ausreisser zurückzuführen

Beispiel GM 2

- zeichne das Streudiagramm mit studentisierten Residuen und geschätzten Werten

  # nutze die Benutzeroberfläche des R Commanders oder den folgenden Befehl

  scatterplot(rstudent.RegModel.1~fitted.RegModel.1, reg.line=FALSE, smooth=FALSE, spread=FALSE, id.method='mahal', id.n = 2, boxplots=FALSE, span=0.5, data=reg_sales)



zu GM 3

- die Störgrösse ist nicht korreliert mit den unabhängigen Variablen

  $Cov\left(u_i, x_{ji}\right) = 0$

  - besteht eine Korrelation zwischen der Störgrösse und einer unabängigen Variable und ist die unabhängige Variable mit der abhängigen Variable korreliert, dann ist auch die Störgrösse mit Y korreliert
  - in einem solchen Fall würde die Variation von Y, die von u kommt X_j zugeordnet
- eine Verletzung von GM 3 führt dazu, dass die Koeffizienten b_j verzerrt geschätzt werden
- evaluieren können wir die Voraussetzung mit dem Streudiagramm
  - wir tragen die Residuen oder die studentisierten Residuen gegen die unabhängigen Variablen ab
  - wenn die Voraussetzung nicht verletzt ist, sollte keine Beziehung zwischen den Variablen erkennbar sein

zu GM 3

- ist GM 3 verletzt, so liegt ein Lösungsansatz in der Überprüfung der Modellspezifikation und dem Hinzufügen möglicher fehlender unabhängiger Variablen
- ein anderer Lösungsansatz ist die sogenannte Instrument-Variablen-Schätzung (nicht weiter diskutiert in der Vorlesung)
  - eine Instrumentvariable ist eine Variable, die mit der jeweiligen unabhängigen Variable hochkorreliert ist, aber nicht mit dem Störterm
  - Hauptproblem bei diesem Ansatz ist das Finden geeigneter Instrumentvariablen

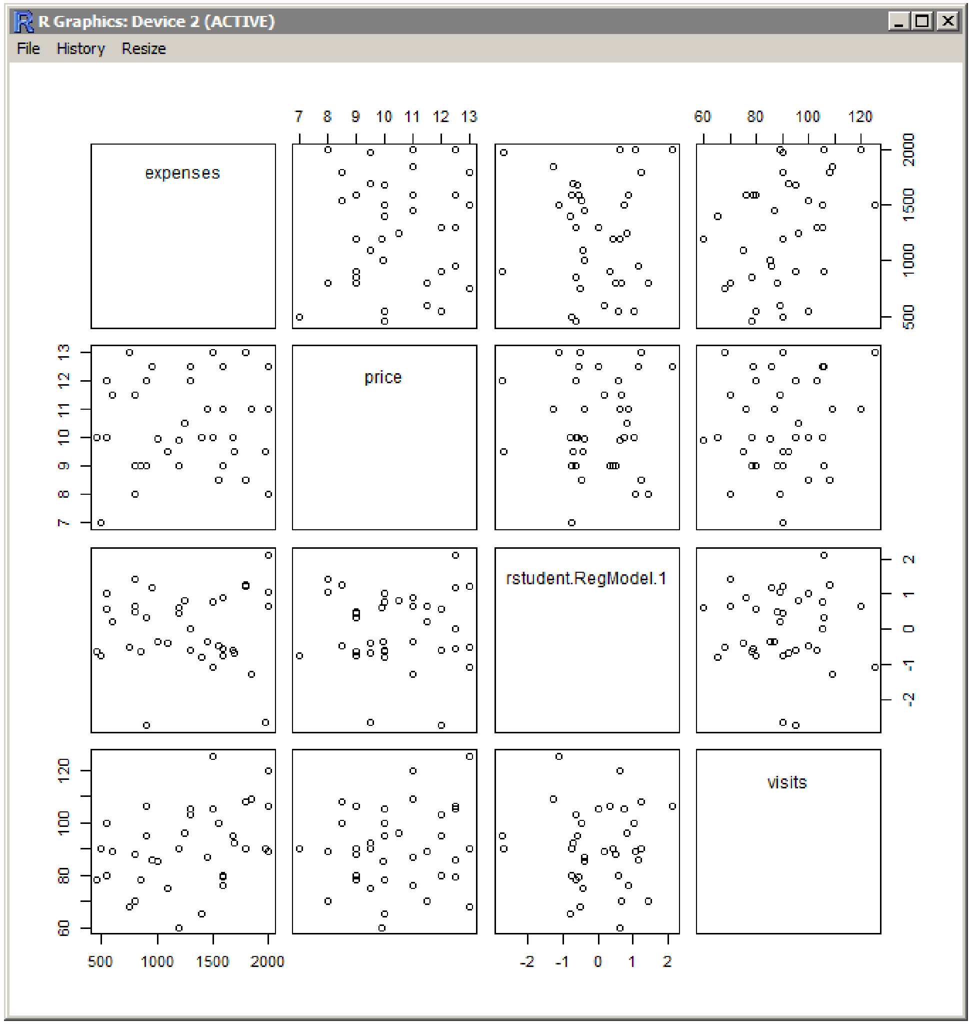
Beispiel GM 3

- zeichne die Streudiagramme zwischen den unabhängigen Variablen und dem Störterm

  # nutze die Benutzeroberfläche des R Commanders oder die folgenden Befehle

  ```
  scatterplot(rstudent.RegModel.1~price, reg.line=FALSE, smooth=FALSE,
  spread=FALSE,   id.method='mahal', id.n = 2, boxplots=FALSE, span=0.5,
  data=reg_sales)

  scatterplot(rstudent.RegModel.1~visits, reg.line=FALSE, smooth=FALSE,
  spread=FALSE,   id.method='mahal', id.n = 2, boxplots=FALSE, span=0.5,
  data=reg_sales)

  scatterplot(rstudent.RegModel.1~expenses, reg.line=FALSE, smooth=FALSE,
  spread=FALSE,   id.method='mahal', id.n = 2, boxplots=FALSE, span=0.5,
  data=reg_sales)

  # oder

  scatterplotMatrix(~expenses+price+rstudent.RegModel.1+visits, reg.line=FALSE,
  smooth=FALSE, spread=FALSE, span=0.5, id.n=0, diagonal = 'none', data=reg_sales)
  ```

Beispiel GM 3



zu GM 4

- die Varianz der Störgrösse ist konstant

  $Var(u_i) = \sigma^2$

  - die vierte Voraussetzung fordert, dass die Varianz der Abweichungen konstant ist, technisch sprechen wir von Homo- bzw. Heteroskedastizität
  - Homoskedastizität bedeutet, dass über den Wertebereich der geschätzten $\hat{Y}$-Werte die Varianz konstant ist
  - Heteroskedastizität liegt vor, wenn die Varianz nicht konstant ist
- eine Verletzung von GM 4 führt zu Ineffizienz und entsprechend dazu, dass die Ergebnisse der Signifikanzstests nicht mehr zuverlässig sind
- evaluieren können wir die vierte Annahme mit Hilfe der Residuen
  - wir zeichnen das Streudiagramm für die studentisierten Residuen oder der quadrierten studentisierten Residuen und der geschätzten $\hat{Y}$-Werte
  - wenn die Residuen in derselben Bandbreite streuen, d. h. die Abweichungen nicht systematisch kleiner oder grösser werden, dann ist die Voraussetzung erfüllt

zu GM 4

- Skizze keine Verletzung von GM 4 (Homoskedastizität)

- Skizzen Verletzung von GM 4 (Heteroskedastizität)

- neben der graphischen Methode sind der Breusch-Pagan-Test und der White-Test beliebte Methoden der Aufdeckung von Heteroskedastiztiät (nicht weiter diskutiert)

zu GM 4

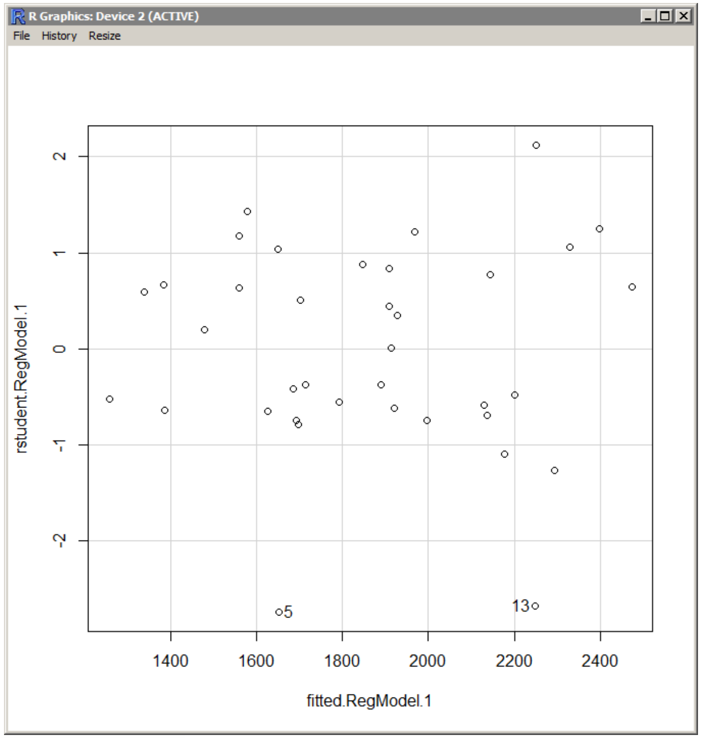
- ein bei Heteroskedastizität in vielen Fällen möglicher Lösungsansatz ist die Transformation der abhängigen Variable Y, so dass diese möglichst symmetrisch wird
  - Y ist flachgipflig dann $\frac{1}{Y}$
  - Y ist rechtsschief dann $\log(Y)$
  - Y ist linksschief dann $\sqrt{Y}$
  - Abweichung der Residuen wird grösser mit zunehmenden $\hat{Y}$ dann $\frac{1}{Y}$
  - Abweichung der Residuen wird kleiner mit zunehmenden $\hat{Y}$ dann $\sqrt{Y}$
  - bei keiner Information führt die sogenannte Box-Cox-Transformation zu einer möglichst symmetrischen Variable (nicht weiter diskutiert)
- ein weiterer Lösungsansatz ist die sogenannte Gewichtete-Kleinste-Quadrate Schätzung WLS (nicht weiter diskutiert)

Beispiel GM 4

- zeichne das Streudiagramm mit studentisierten Residuen und geschätzten Werten

  # nutze die Benutzeroberfläche des R Commanders oder den folgenden Befehl

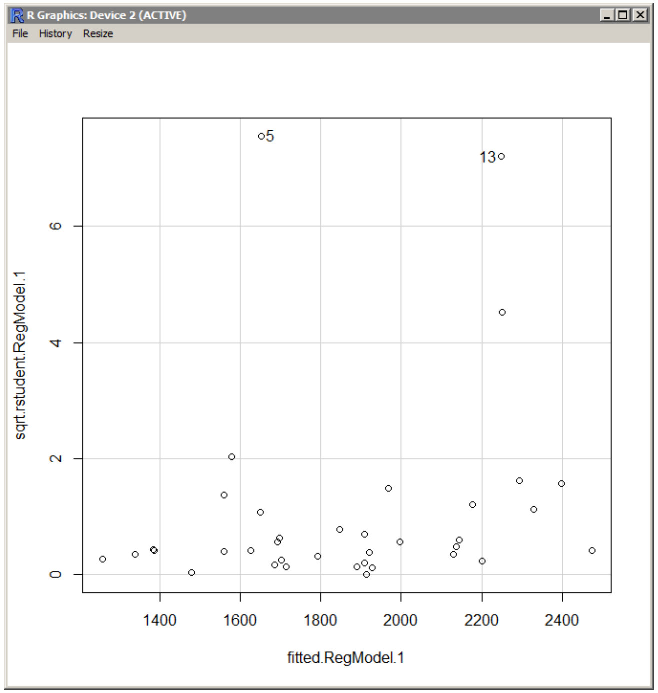
  scatterplot(rstudent.RegModel.1~fitted.RegModel.1, reg.line=FALSE, smooth=FALSE, spread=FALSE, id.method='mahal', id.n = 2, boxplots=FALSE, span=0.5, data=reg_sales)



Beispiel GM 4

- oder zeichne das Streudiagramm mit den quadrierten studentisierten Residuen und geschätzten Werten

```
# nutze die Benutzeroberfläche des R Commanders oder die folgenden Befehle

reg_sales$sqrt.rstudent.RegModel.1 <- with(reg_sales, rstudent.RegModel.1* rstudent.RegModel.1)

scatterplot(sqrt.rstudent.RegModel.1~fitted.RegModel.1, reg.line=FALSE, smooth=FALSE, spread=FALSE,   id.method='mahal', id.n = 2, boxplots=FALSE, span=0.5, data=reg_sales)
```



zu GM 5

- die Störgrössen sind unkorreliert

  $$Cov(u_i, u_{i+r}) = 0$$

  - die fünfte Voraussetzung fordert, dass die Störgrössen nicht voneinander abhängig sind
  - technisch sprechen wir von keine Autokorrelation
- eine Verletzung von GM 5 führt wie bei Heteroskedastitzität zu Ineffizienz und damit dazu, dass die Ergebnisse der Signifikanztests nicht mehr zuverlässig sind
- evaluieren können wir die fünfte Annahme wieder mit Hilfe der Residuen
  - wir zeichnen das Streudiagramm für die studentisierten Residuen und tragen diese gegenüber der Reihenfolge der Werte ab
  - wenn wir keinen systematischen Zusammenhang sehen, dann liegt keine Autokorrelation vor
- das Problem der Autokorrelation existiert vor allem bei der Analyse von Zeitreihen

zu GM 5

- Skizze keine Verletzung von GM 5

- Skizzen Verletzung von GM 5

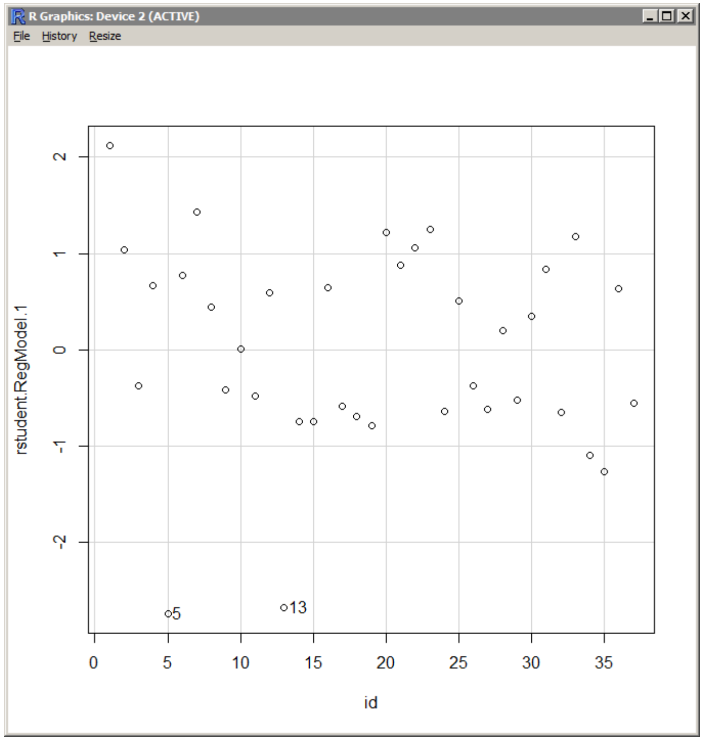
- neben der graphischen Methode ist der Durbin-Watson-Test eine beliebte Methode zur Aufdeckung von Autokorrelation (nicht weiter diskutiert)

Beispiel GM 5

- zeichne das Streudiagramm mit studentisierten Residuen und Wertereihenfolge

  # nutze die Benutzeroberfläche des R Commanders oder den folgenden Befehl

  scatterplot(rstudent.RegModel.1~id, reg.line=FALSE, smooth=FALSE, spread=FALSE, id.method='mahal', id.n = 2, boxplots=FALSE, span=0.5, data=reg_sales)



zu GM 6

- zwischen den unabhängigen Variablen X_j besteht keine lineare Abhängigkeit
  - die Annahme fordert, dass die unabhängigen Variablen zueinander nicht in Beziehung stehen
  - technisch sprechen wir von Multikollinearität bzw. keine Multikollinearität
- eine Verletzung von GM 6 führt dazu, dass die Koeffizienten b_j nicht mehr unverzerrt geschätzt werden
- zur Überprüfung stehen verschiedene Möglichkeiten zur Verfügung
  - Streudiagramme zwischen den unabhängigen Variablen X_i, X_j und prüfen, ob eine lineare Beziehung zwischen ihnen besteht
  - bivariate Korrelationskoeffizienten zwischen den unabhängigen Variablen X_i, X_j, Korrelationskoeffizienten grösser als 0.8 bzw. 0.9 gelten als problematisch

zu GM 6

- Varianzinflationsfaktor (VIF)
  - $VIF_j = \frac{1}{1-R_j^2}$
  - R_j^2 gleich dem Bestimmtheitsmass der unabhängigen Variable X_j auf die übrigen unabhängigen Variablen
  - Varianzinflationsfaktorenwerte grösser als 4 gelten als problematisch

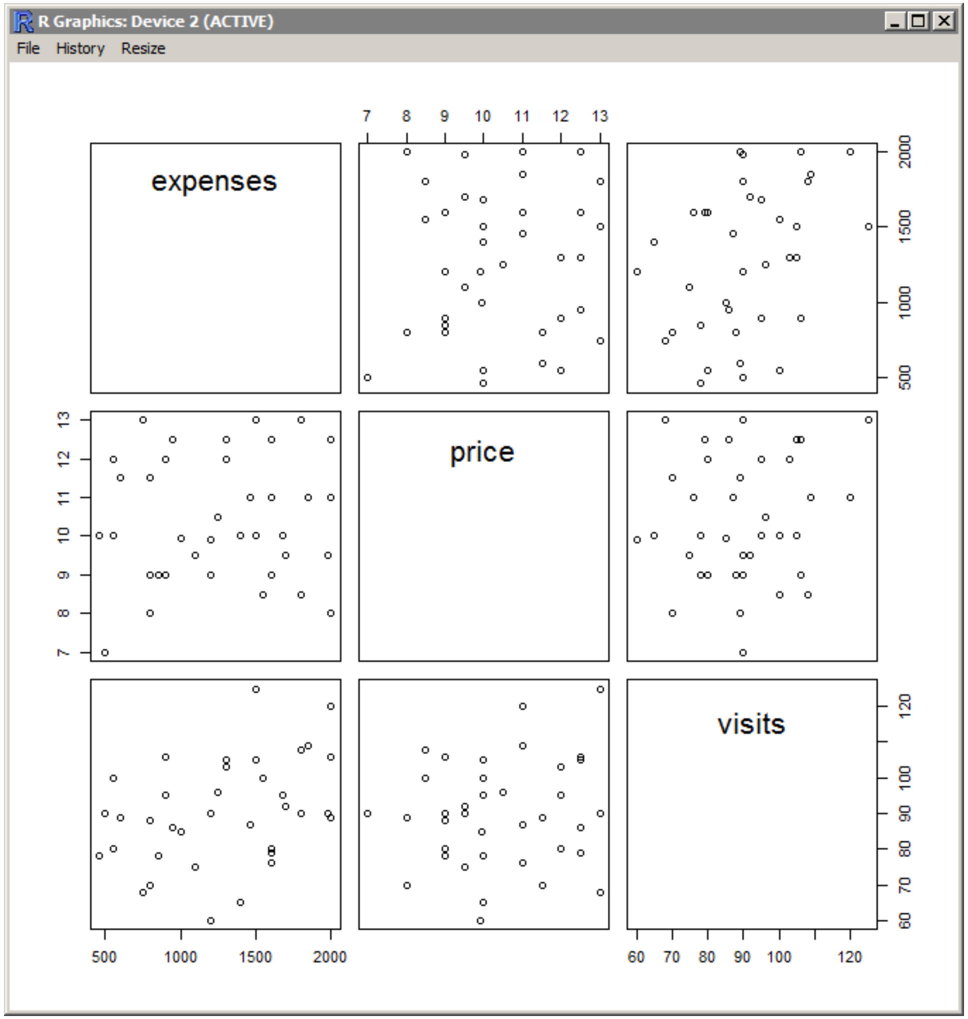
- ein oft verwendeter Lösungsansatz bei Multikollinearität ist das Weglassen der für die Multikollinearität verantwortlichen unabhängigen Variable
- ein anderer Lösungsansatz ist das Durchführen einer Faktorenanalyse bevor die Regressionsanalyse geschätzt wird (nicht weiter diskutiert)

Beispiel GM 6

- zeichne die Streudiagramme zwischen den unabhängigen Variablen

```
# nutze die Benutzeroberfläche des R Commanders oder die folgenden Befehle
scatterplot(expenses~price, reg.line=FALSE, smooth=FALSE, spread=FALSE,
id.method='mahal', id.n = 2, boxplots=FALSE, span=0.5, data=reg_sales)
scatterplot(expenses~visits, reg.line=FALSE, smooth=FALSE, spread=FALSE,
id.method='mahal', id.n = 2, boxplots=FALSE, span=0.5, data=reg_sales)
scatterplot(price~visits, reg.line=FALSE, smooth=FALSE, spread=FALSE,
id.method='mahal', id.n = 2, boxplots=FALSE, span=0.5, data=reg_sales)
# oder
scatterplotMatrix(~expenses+price+visits, reg.line=FALSE, smooth=FALSE,
spread=FALSE, span=0.5, id.n=0, diagonal = 'none', data=reg_sales)
```

Beispiel GM 6

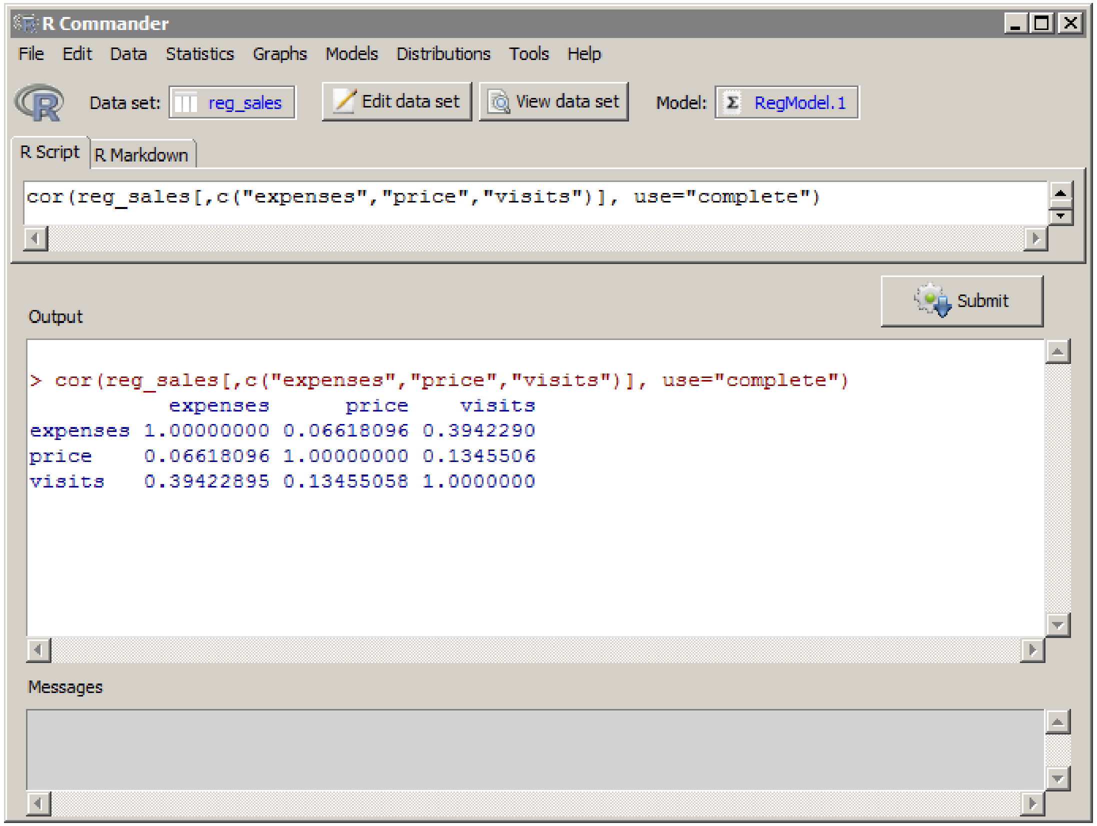


Beispiel GM 6

- berechne die bivariaten Korrelationskoeffizienten

  # nutze die Benutzeroberfläche des R Commanders oder den folgenden Befehl

  cor(reg_sales[,c("expenses","price","visits")], use="complete")

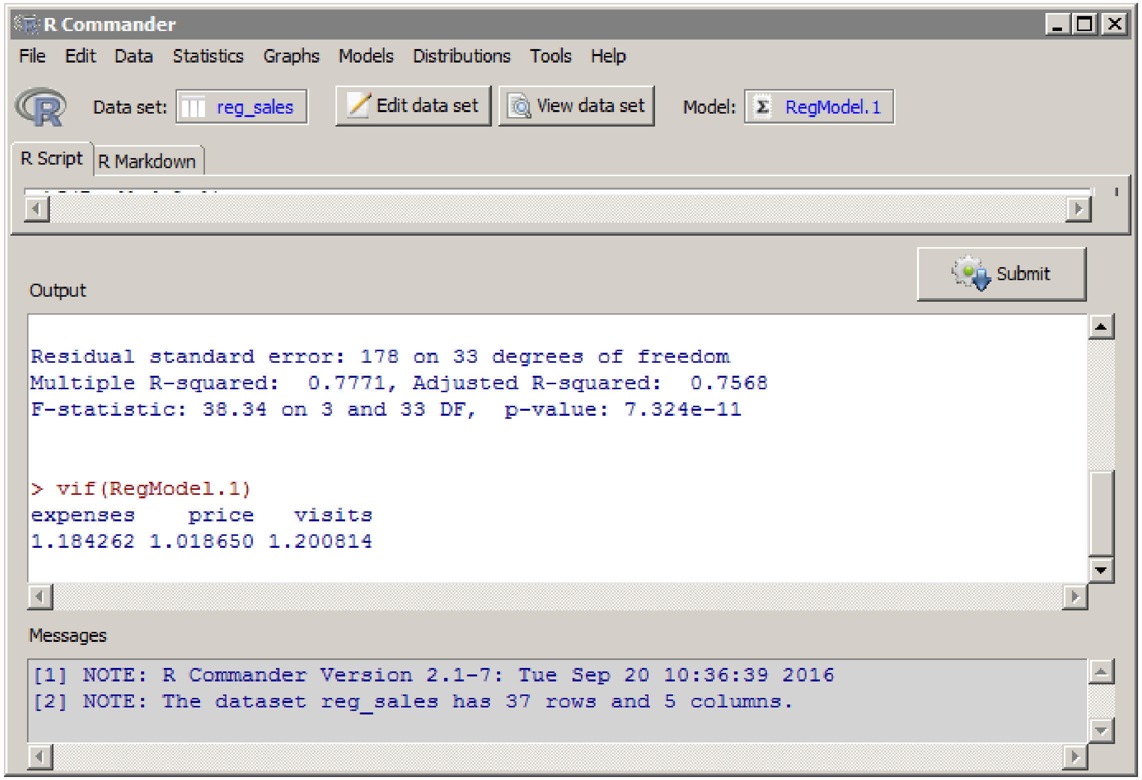


Beispiel GM 6

- berechne den Varianzinflationsfaktor

nutze die Benutzeroberfläche des R Commanders oder den folgenden Befehl (der Befehl erfordert, dass das Regressionsmodell geschätzt wurde)

vif(RegModel.1)



zu GM 7

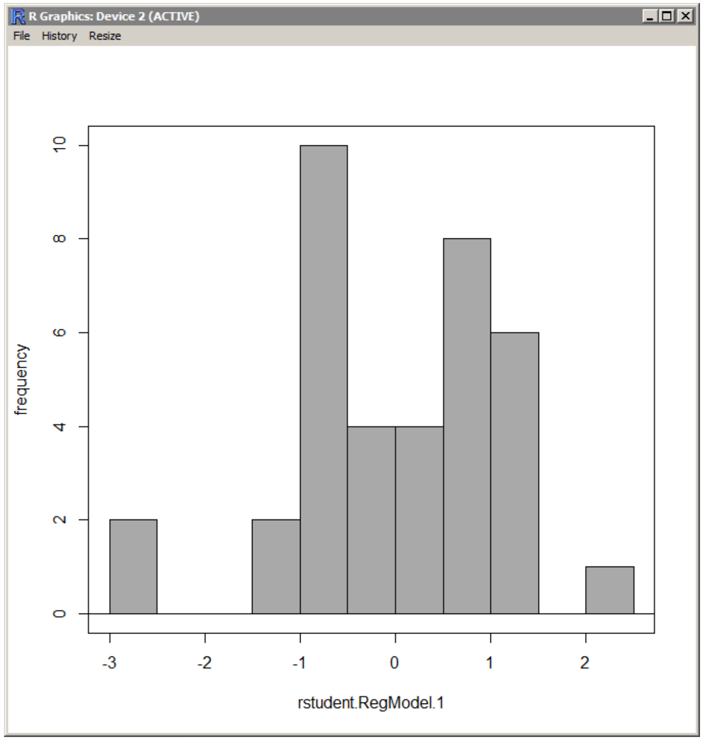
- die Störgrössen u_i sind normalverteilt
- eine Verletzung von GM 7 führt dazu, dass die Signifikanztests nicht mehr vertrauenswürdig sind
- evaluieren können wir die siebte Annahme mit den bereits erlernten Techniken der Prüfung auf Normalverteilung
  - Histogramm
  - Quantil-Quantil-Plot
  - Shapiro-Wilk-Test
- die Voraussetzung ist insbesondere bei kleinen Stichproben von Bedeutung
- ist die Anzahl an Beobachtungen gross (z. B. $n > 40$) sind die Signifikanztests unabhängig von der Verteilung der Störgrössen vertrauenswürdig

Beispiel GM 7

- zeichne das Histogramm, den Quantil-Quantil-Plot und berechne den Shapiro-Wilk-Test

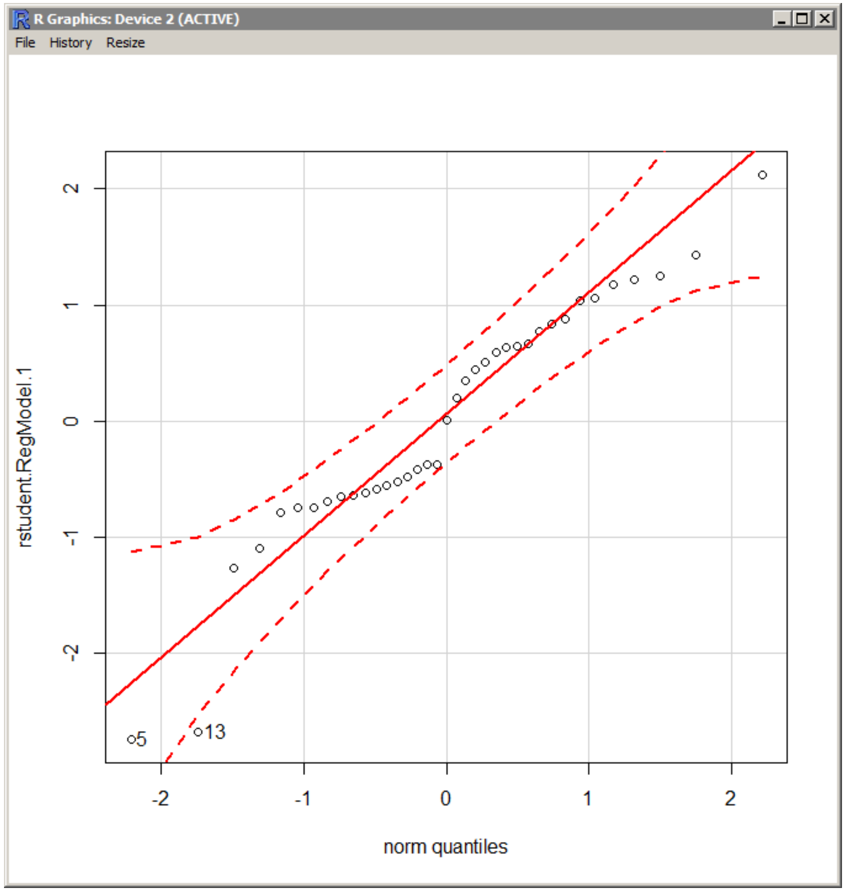
nutze die Benutzeroberfläche des R Commanders oder die folgenden Befehle

```
with(reg_sales, Hist(rstudent.RegModel.1, scale="frequency", breaks="Sturges", col="darkgray"))
```



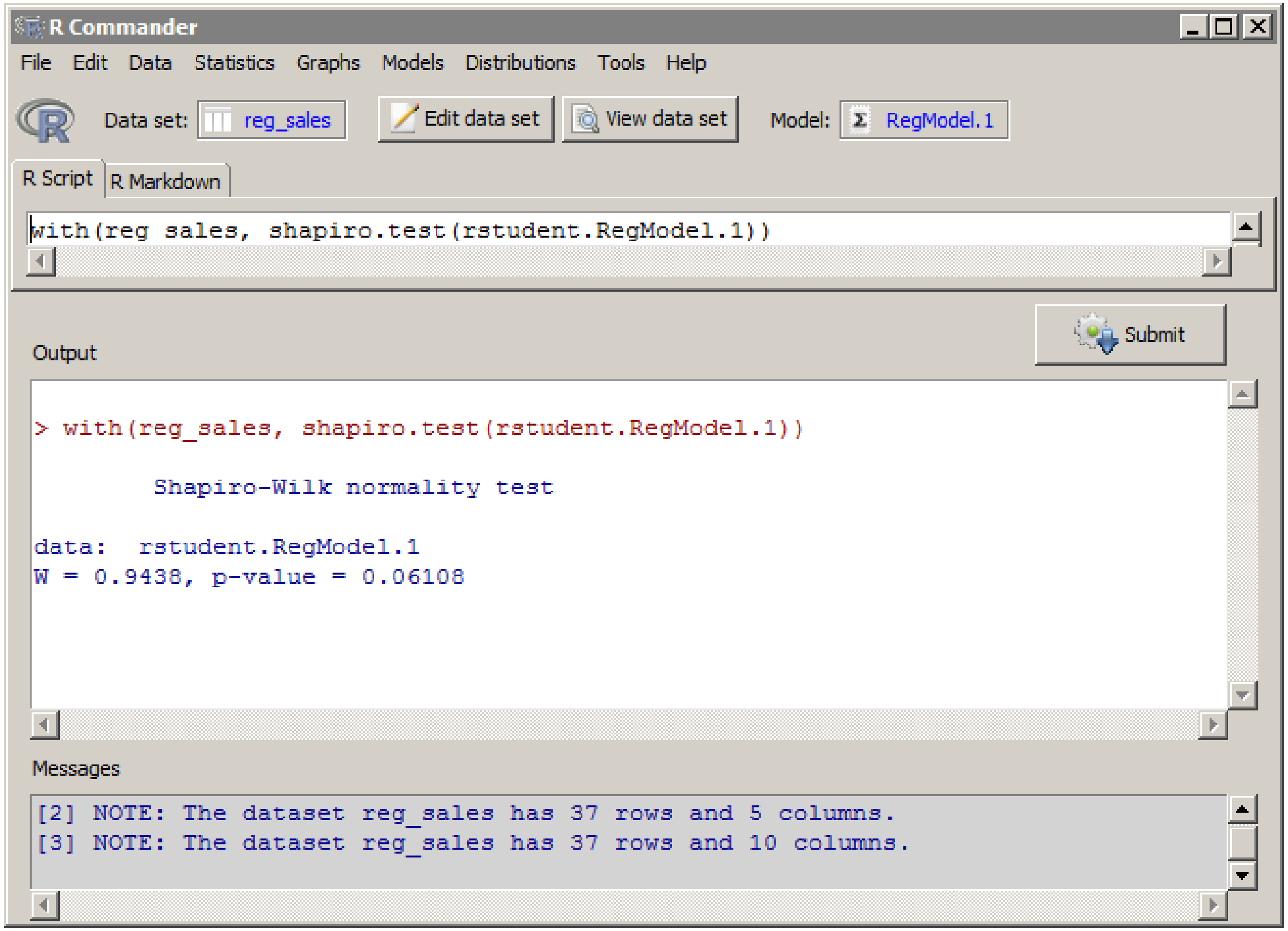
Beispiel GM 7

with(reg_sales, qqPlot(rstudent.RegModel.1, dist="norm", id.method="y", id.n=2, labels=rownames(reg_sales)))



Beispiel GM 7

with(reg_sales, shapiro.test(rstudent.RegModel.1))



Gauss-Markov-Annahmen GM

- Zusammenfassung: Auswirkungen bei Verletzung der Gauss-Markov-Annahmen

Annahme	Verletzung	Konsequenz
Linearität	Nichtlinearität	Verzerrung von b_j
Vollständigkeit des Modells	unabhängige Variablen fehlen	Verzerrung von b_j
Erwartungswert der Störgrössen ist gleich 0	Erwartungswert ist ungleich 0	Verzerrung von b_0
Störgrösse ist nicht korreliert mit X_j	Störgrösse ist korreliert mit X_j	Verzerrung von b_j
Homoskedastizität	Heteroskedastitzität	Ineffizienz
Unabhängigkeit der Störgrössen	Autokorrelation	Ineffizienz
Keine lineare Abhängigkeit zwischen X_j	Multikollinearität	Verzerrung von b_j
Normalverteilung der Störgrössen	keine Normalverteilung	ungültige Signifikanztests bei kleinem n

Gauss-Markov-Annahmen GM

- bei einer ernsthaften Verletzung einer dieser Annahmen, GM 1 bis GM 6, sind die Regressionsergebnisse nicht mehr BLUE
- eine Verletzung von GM 7 hat keine Auswirkung auf die BLUE Eigenschaft, die Signifikanztests sind aber nicht mehr vertrauenswürdig
- wenn die Annahmen verletzt sind können die diskutierten Lösungsmöglichkeiten helfen, ggf. muss auf weiterführende Verfahren zurückgegriffen werden

Ausreisser

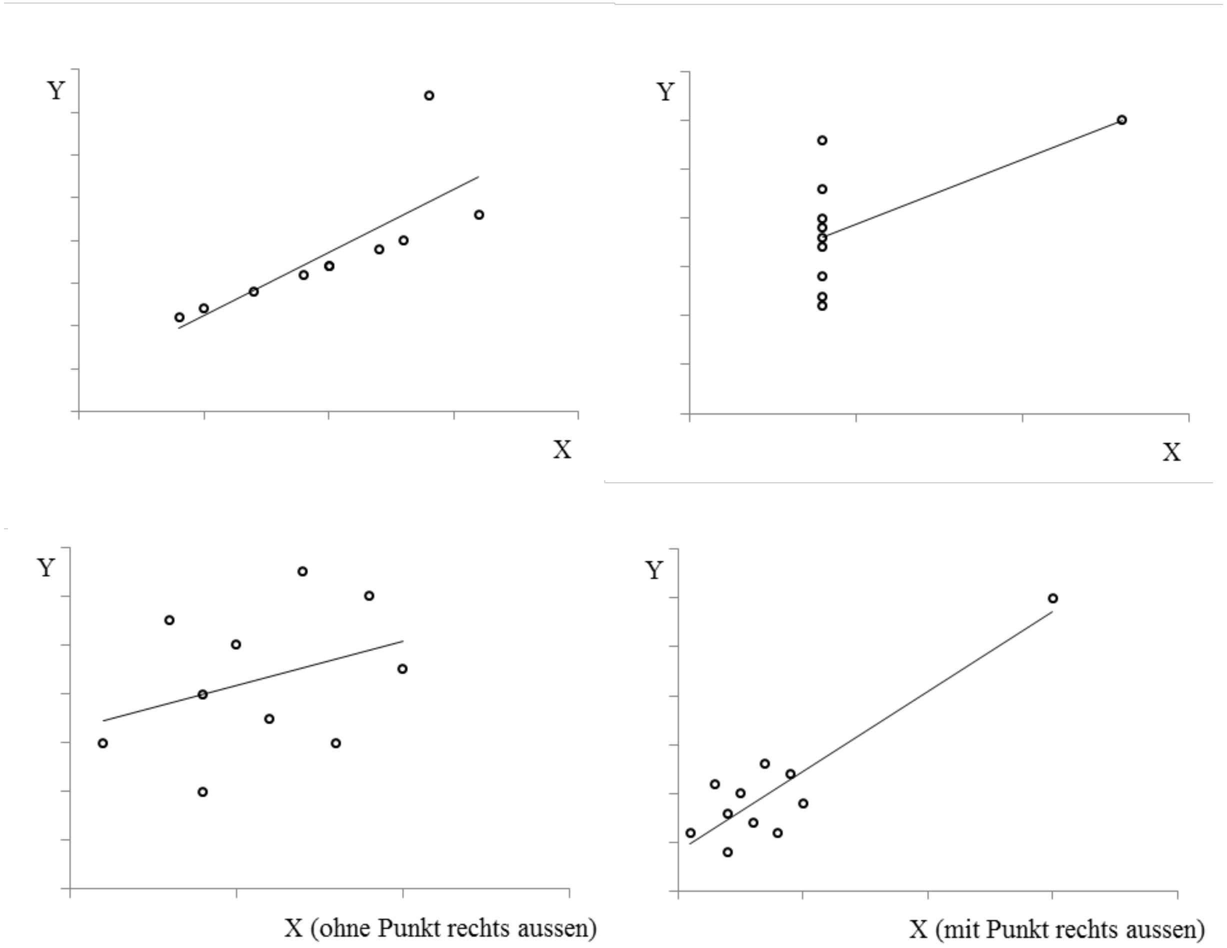
- Ausreisser und einflussreiche Beobachtungen können die Ergebnisse der Regressionsanalyse erheblich beeinflussen
- Ausreisser und einflussreiche Beobachtungen können zu einer Verletzung der Annahmen führen
- die Daten sind daher auf Ausreisser und einflussreiche Beobachtungen zu prüfen
- es gibt drei Ursachen für Ausreisser und einflussreiche Beobachtungen

  1. Fehler in den Daten, z. B. bei Datenerfassung
  2. ungewöhnliche Beobachtung, die erklärbar ist
  3. ungewöhnliche Beobachtung, die nicht erklärbar ist

- bei 1 und 2 ist die Vorgehensweise einfach
  - bei 1 wird der Fehler verbessert oder bereinigt
  - bei 2 wird die Erklärung genutzt, um zu entscheiden ob die Beobachtung wichtig für das Modell ist
- bei 3 liegt es im Ermessen des Forschers, ob mit der Beobachtung gearbeitet oder diese Weggelassen wird

Ausreisser

- Beispiele für Einfluss von Ausreissern



Ausreisser

- zur Identifikation von Ausreissern und einflussreichen Beobachtungen stehen uns verschiedene Hilfsmittel zur Verfügung
  - Streudiagramme zwischen abhängiger Variable Y und unabhängigen Variablen X_j
    - univariate Ausreisser
    - ungewöhnliche Beobachtungen für eine der unabhängigen Variablen X_j auf Y
  - Streudiagramm zwischen studentisierten Residuen e_i und geschätzter abhängiger Variable $\hat{Y}$
    - Regressionsausreisser
    - Wert von Y bezüglich alle X_j ist ungewöhnlich
    - Abweichungen grösser ± 2 Standardabweichungen gelten als potentielle Ausreisser
    - Achtung, ca. 5% der Werte sind aufgrund der Verteilungseigenschaften immer ausserhalb des ± 2 Intervalls

Ausreisser

- Hebelwerte (hat values)
  - Hebelwirkung wird mit den Hebelwerten h_i gemessen
  - der Definitionsbereich liegt zwischen ${}^{1}/_{n} \leq h_i \leq 1$
  - die Hebelwirkung ist gross wenn $h_i > 2(k+1)/n$ mit k gleich Anzahl unabhängiger Variablen und n gleich Anzahl Beobachtungen
  - wir tragen die Hebelwerte gegenüber der Beobachtungsnummer ab und ziehen eine Linie bei der kritischen Grösse
  - zur Evaluation wird oft auch das Hebelarm-Diagramm verwendet
    - ❖ es werden die studentisierten Residuen gegenüber den Hebelwerten abgetragen
    - ❖ es wird gleichzeitig die Residualgrösse und die Grösse der Hebelwerte evaluiert

Ausreisser

- Cook's Distanz
  - Cook's Distanz misst den Einfluss einer Beobachtung bzw. was passiert, wenn eine Beobachtung weggelassen wird
  - Cook's Distanz wird als gross betrachtet, wenn $D_i > 4/(n-k-1)$, mit n gleich Anzahl an Beobachtungen und k gleich Anzahl an unabhängigen Variablen
    - ❖ wir tragen Cook's Distanz gegenüber der Beobachtungsnummer ab und zeichnen den kritischen Wert horizontal ein
    - ❖ Werte gelten als Einflussreich, welche die kritische Linie überschreiten

Beispiel Ausreisser

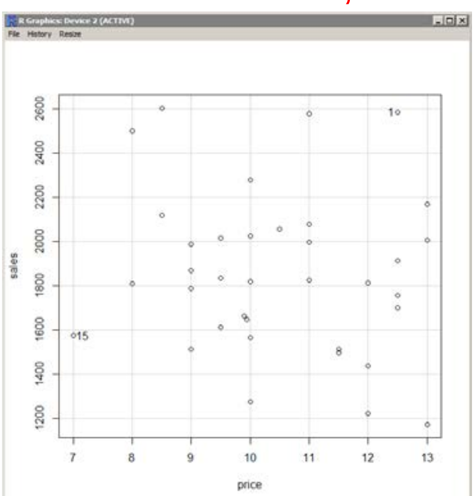
- zeichne die Streudiagramme zwischen unabhängigen Variablen und den abhängigen Variablen

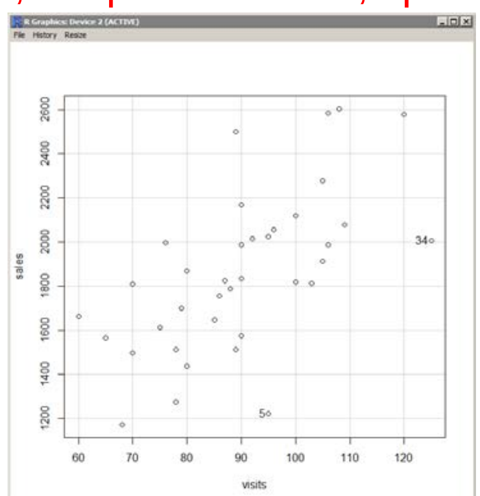
  # nutze die Benutzeroberfläche des R Commanders oder die folgenden Befehle

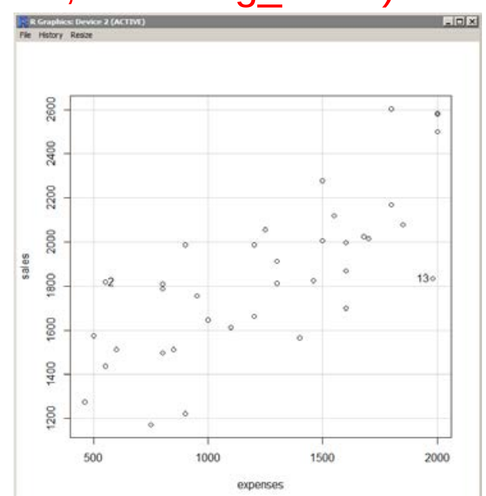
  ```
  scatterplot(sales~price, reg.line=FALSE, smooth=FALSE, spread=FALSE, id.method='mahal', id.n = 2, boxplots=FALSE, span=0.5, data=reg_sales)
  ```

  ```
  scatterplot(sales~visits, reg.line=FALSE, smooth=FALSE, spread=FALSE, id.method='mahal', id.n = 2, boxplots=FALSE, span=0.5, data=reg_sales)
  ```

  ```
  scatterplot(sales~expenses, reg.line=FALSE, smooth=FALSE, spread=FALSE, id.method='mahal', id.n = 2, boxplots=FALSE, span=0.5, data=reg_sales)
  ```





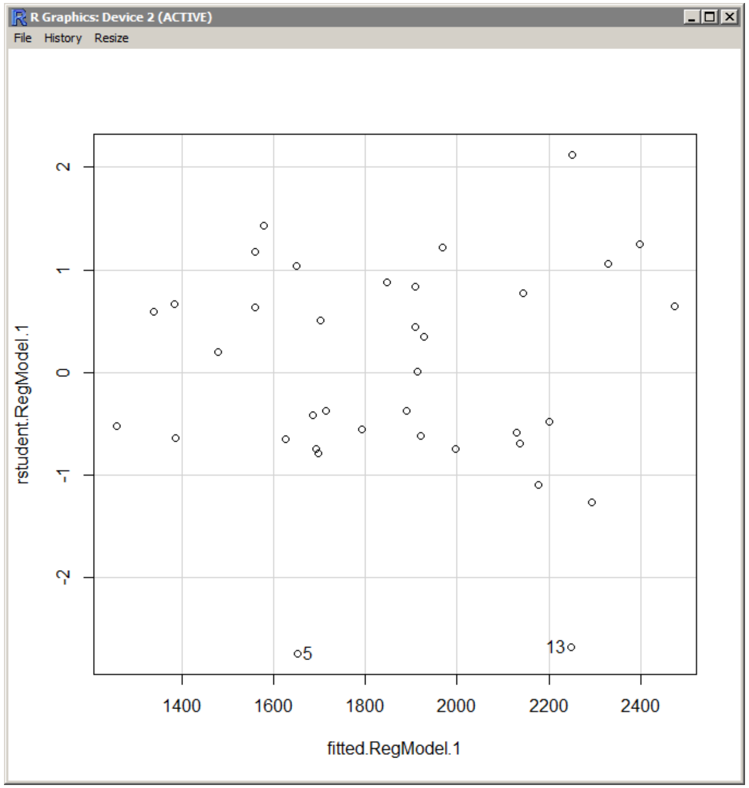


Beispiel Ausreisser

- zeichne das Streudiagramm zwischen studentisierten Residuen und geschätzter abhängiger Variable

  # nutze die Benutzeroberfläche des R Commanders oder den folgenden Befehl

  scatterplot(rstudent.RegModel.1~fitted.RegModel.1, reg.line=FALSE, smooth=FALSE, spread=FALSE, id.method='mahal', id.n = 2, boxplots=FALSE, span=0.5, data=reg_sales)

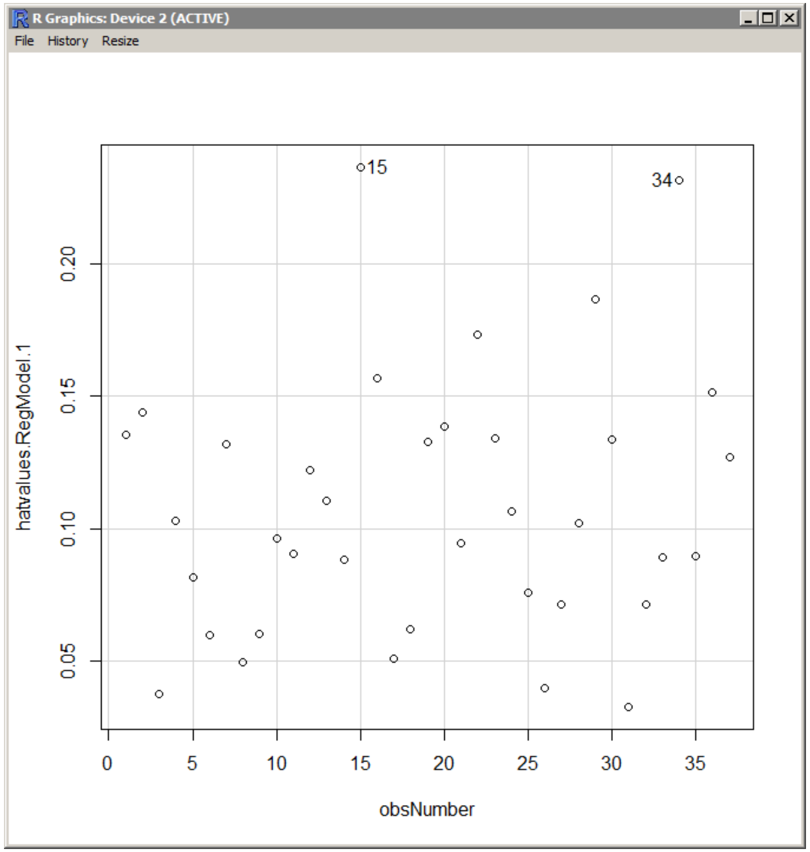


Beispiel Ausreisser

- zeichne das Streudiagramm zwischen Hebelwerten und Beobachtungswerten

  # nutze die Benutzeroberfläche des R Commanders oder den folgenden Befehl

  scatterplot(hatvalues.RegModel.1~obsNumber, reg.line=FALSE, smooth=FALSE, spread=FALSE, id.method='mahal', id.n = 2, boxplots=FALSE, span=0.5, data=reg_sales)

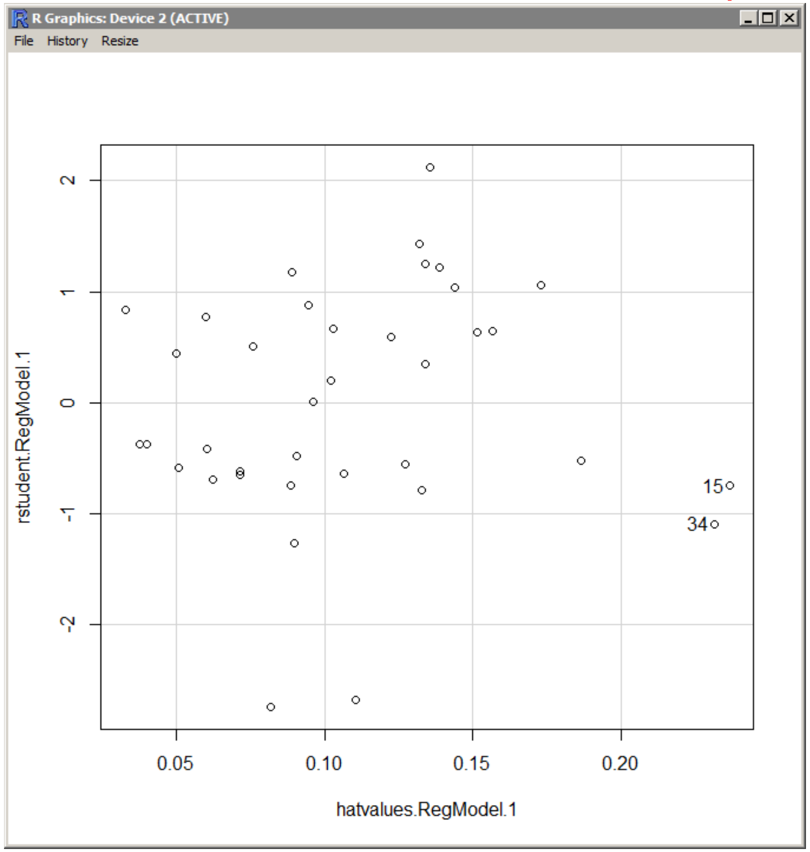


Beispiel Ausreisser

- zeichne das Hebelarm-Diagramm

  # nutze die Benutzeroberfläche des R Commanders oder den folgenden Befehl

  scatterplot(rstudent.RegModel.1~hatvalues.RegModel.1, reg.line=FALSE, smooth=FALSE, spread=FALSE, id.method='mahal', id.n = 2, boxplots=FALSE, span=0.5, data=reg_sales)

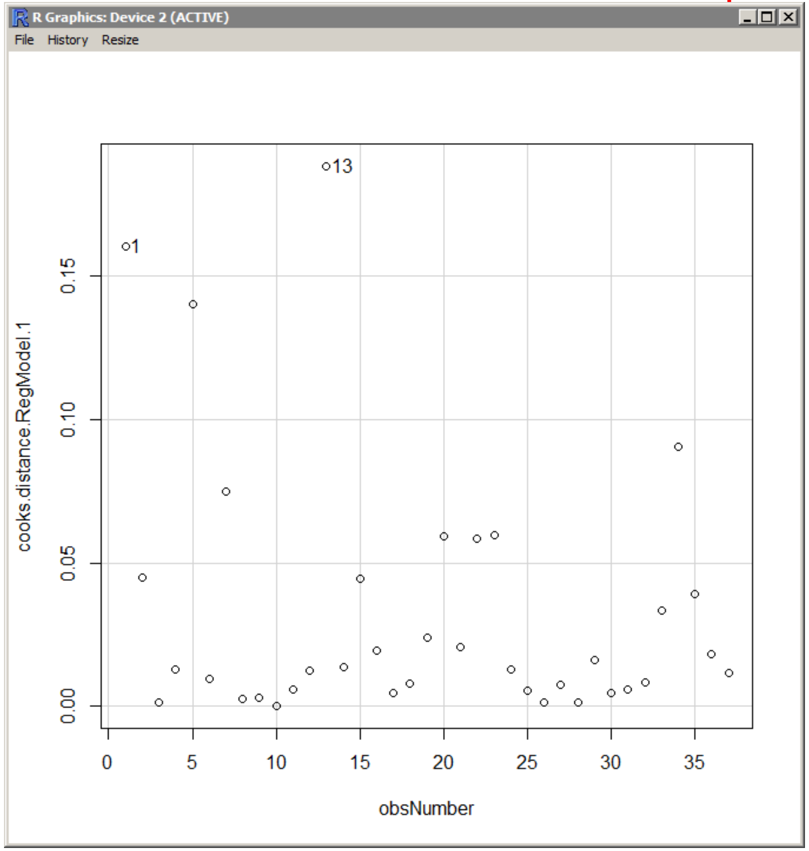


Beispiel Ausreisser

- zeichne das Streudiagramm für Cook’s Distanz gegenüber den Beobachtungswerten

  # nutze die Benutzeroberfläche des R Commanders oder den folgenden Befehl

  scatterplot(cooks.distance.RegModel.1~obsNumber, reg.line=FALSE, smooth=FALSE, spread=FALSE, id.method='mahal', id.n = 2, boxplots=FALSE, span=0.5, data=reg_sales)



Zusammenfassung Diskussion der Annahmen und Ausreisser

Validierung der Ergebnisse

- am Ende des Prozesses steht ein fertiges Regressionsmodell
- bevor dieses weiter verwendet wird, sollten die Ergebnisse validiert werden
- für die Validierung der Ergebnisse bieten sich folgende Verfahren an
  - Schätzen des Modell mit neuen/ anderen Daten
  - Schätzen des Modells mit systematischer Auswahl an Beobachtungen aus der Stichprobe
  - Schätzen des Modells mit zufälliger Auswahl an Beobachtungen aus der Stichprobe
- wenn sich die Schätzergebnisse gegenüber den ursprünglich gefundenen Werten nicht gross unterscheiden gelten die Resultate als valide

Anwendungen

- Nutze den Datensatz smartphone.Rdata (simuliert). Schätze das Regressionsmodell mit der unabhängigen Variable «price» und der abhängigen Variable «sales».
  - Prüfe das Regressionsmodell auf Verletzung der Annahmen.
  - Suche Lösungen für Verletzungen der Annahmen.
  - Zu welchen Ergebnis kommen Sie mit Blick auf die verfügbaren Daten.
- Nutze den Datensatz reg_country.RData (aus dem Lehrbuch von Norusis 2008). Schätze das Regressionsmodell mit den unabhängigen Variablen «urban, doc, bed, gdp, rad» und der abhängigen Variable «expect».
  - Prüfe das Regressionsmodell auf Verletzung der Annahmen.
  - Suche Lösungen für Verletzungen der Annahmen.
  - Schätze das fertige Modell und interpretiere die Ergebnisse.

Anwendungen

- Nutze den Datensatz abortion.RData (zur Verfügung gestellt von Prof. Kahane). Schätze das Regressionsmodell mit den unabhängigen Variablen «religion, price, laws, funds, educ, income, picket» und der abhängigen Variable «abortion».
  - Prüfe das Regressionsmodell auf Verletzung der Annahmen.
  - Suche Lösungen für Verletzungen der Annahmen.
  - Schätze das fertige Modell und interpretiere die Ergebnisse.