7. Vorlesung Statistik II
Regressionsanalytisches Modell für zwei Prädiktoren und Regressionsdiagnostik

Plan für heute

- Modell der MLR mit zwei Prädiktoren
  - Parameterschätzung
  - Hypothesentests
  - Vorhersage und Konfidenzintervall für y_i
- Regressionsdiagnostik
  - Annahmen der ELR
  - Annahmen der MLR

Vorbemerkung zur multiplen linearen Regression (MLR)

- Man spricht von einer multiplen Regression, wenn das Regressionsmodell mehr als einen Prädiktor enthält.
- Ziel ist es, über Hinzunahme von mehr als einem Prädiktor Vorhersagen bezogen auf die Kriteriumsvariable zu verbessern.
- Außerdem interessieren unter anderem folgende Fragen:
  - Wie viel Varianz bezogen auf die Gesamtvarianz des Kriteriums können die Prädiktoren gemeinsam erklären?
  - Welcher Prädiktor weist den größten Vorhersagebeitrag auf?
  - Wie groß ist der eigenständige Vorhersagebeitrag eines Prädiktors, wenn die beiden Prädiktoren miteinander korrelieren?
  - Verändert sich die Stärke, Richtung und damit die Interpretation des Effekts des Prädiktors unter Berücksichtigung eines anderen Prädiktors im Vergleich zur ELR?
- Wir beschränken uns zunächst auf die einfachste Form der multiplen Regression: Die Beschreibung des AV-Werts Y_i durch zwei stetige Prädiktoren und die Fehlervariable. Hat man den Fall mit zwei Prädiktoren verstanden, ist die Erweiterung auf k Prädiktoren einfach.

MLR mit zwei Prädiktoren: Beispiel

- Inhaltliche Untersuchungsfrage: Können irrationale Einstellungen das Phänomen Depression, bzw. die Depressionsschwere erklären?
  - AV („BDI-Gesamt“): Anzahl der Punkte im Beck-Depressions-Inventar (Testverfahren, das die Schwere depressiver Symptomatik im klinischen Bereich erfasst)
  - UV 1 („fie_nsb“): Einschätzung der negativen Selbstbewertung
    Beispielitem: „Ich denke oft, ich bin ein Versager“
  - UV 2 („fie_abk“): Einschätzung der Abhängigkeitskognitionen
    Beispielitem: „Ich brauche es, dass Leute mich mögen“
- Es ist davon auszugehen, dass die beiden Prädiktoren miteinander korrelieren.

MLR mit zwei Prädiktoren

- **Bislang**: Es wurde für jeden Prädiktor eine einfache lineare Regressionsanalyse durchgeführt.

- **Jetzt**: Die Modellgleichung der einfachen linearen Regression wird durch einen zusätzlichen Prädiktor erweitert.

Ausgangslage

- Zur Überprüfung der Untersuchungsfrage wird aus der Population eine einfache Zufallsstichprobe vom Umfang n gezogen.
- Bei jeder Person werden die Depressionsausprägung (AV), die Stärke der negativen Selbstbewertung (UV1) und die Höhe der Abhängigkeitskognition (UV2) gemessen.
- Alle Variablen können durch Zufallsvariablen modelliert werden.

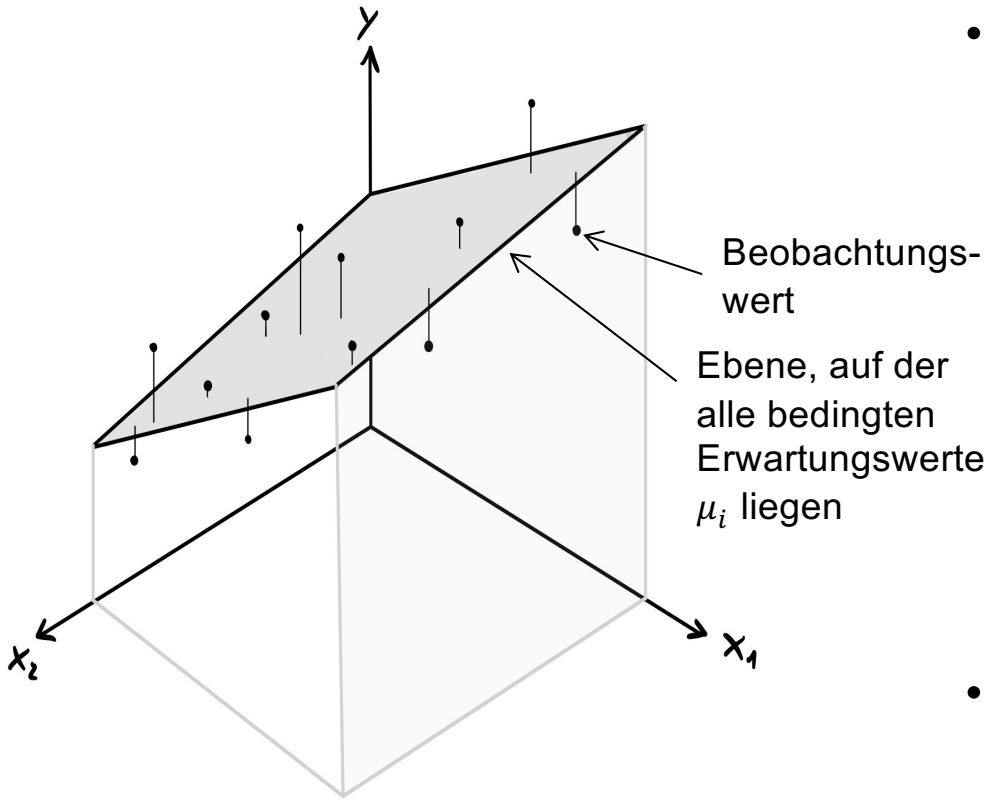
Statistisches Modell bei einer MLR mit zwei Prädiktoren

- Die Zusammensetzung der Zufallsvariable Y_i, deren Realisation der AV-Wert der zufällig gezogenen Person i ist, lässt sich darstellen als

$$Y_i = \alpha + \beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2} + \varepsilon_i, \qquad \text{wobei } \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

- X_{i1} und X_{i2} sind **Zufallsvariablen**. Ihre Realisationen sind jeweils die Werte der zufällig gezogenen Person *i* bezüglich der UV 1 und der UV 2.
- α, β_1, β_2 und σ^2 sind die **Modellparameter** und damit **unbekannte Konstanten**.
- Es gilt: $E(Y_i | X_{i1} = x_{i1}, X_{i2} = x_{i2}) = \mu_i = \alpha + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2}$
- In der Modellgleichung wird formuliert, dass...
  - ...die Abweichung der Zufallsvariable Y_i von ihrem bedingten Erwartungswert an der Stelle x_{i1} und x_{i2} einzig durch die Fehlervariable ε_i verursacht wird.
  - ...ein linearer Zusammenhang zwischen den Prädiktoren X_1 und X_2 und dem Kriterium besteht.

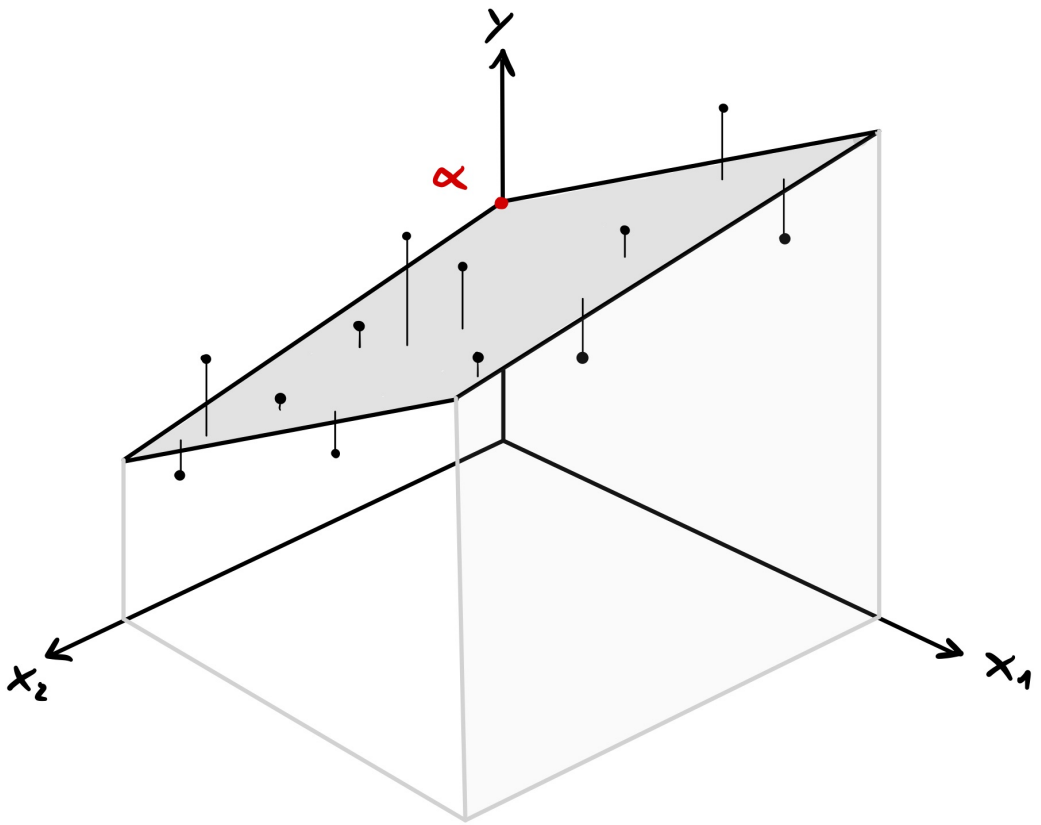
Graphische Veranschaulichung der Regressionsebene in der Population



- Im Falle einer MLR mit zwei Prädiktoren kann der lineare Zusammenhang zwischen dem Kriterium und den Prädiktoren durch eine Regressionsebene dargestellt werden.

- Die nebenstehende Graphik veranschaulicht eine MLR mit zwei Prädiktoren auf der Populationsebene.

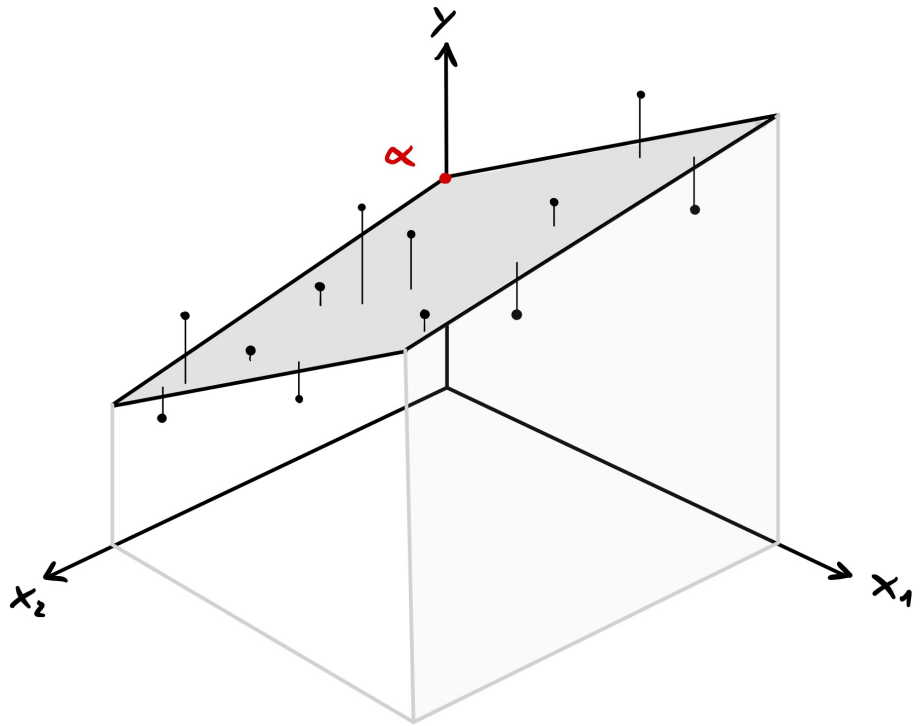
Modellparameter α



Der Modellparameter α stellt den **Achsenabschnitt** dar; α gibt an, in welchem Punkt die Regressionsebene die y-Achse schneidet. Man erhält α, indem man in der Ebenengleichung die Zufallsvariablen X_1 und X_2 gleich null setzt:

$$E(Y_i|X_{i1} = 0, X_{i2} = 0) = \alpha + \beta_1 \cdot 0 + \beta_2 \cdot 0 = \alpha$$

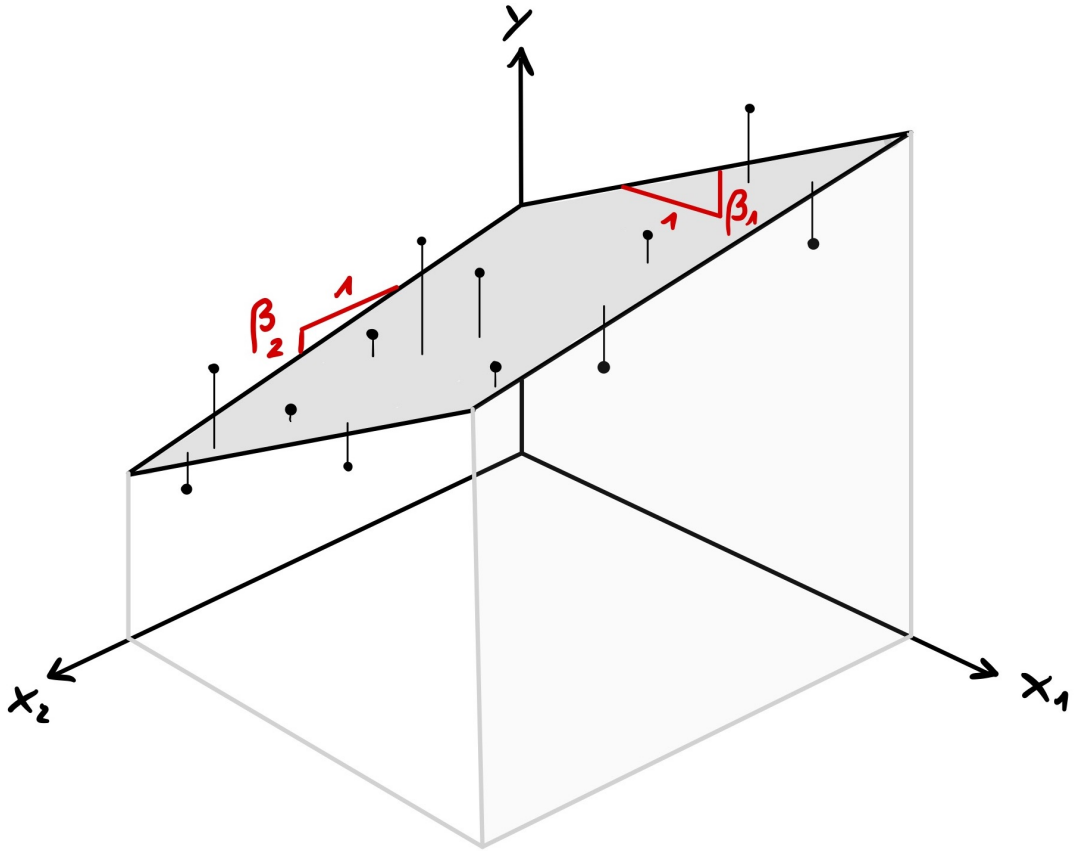
Modellparameter α



- Ob der Modellparameter α sinnvoll interpretiert werden kann, hängt davon ab, ob $x_1 = 0$ und $x_2 = 0$ inhaltlich sinnvolle Werte darstellen.

- In unserem Beispiel ist die Interpretation zum Beispiel wenig sinnvoll, da α hier die mittlere Depressionsausprägung repräsentiert im Falle, dass Personen überhaupt keine negative Selbstbewertung und keine Abhängigkeitskognitionen aufweisen.

Modellparameter β_1 und β_2



- Der Steigungsparameter β_1 gibt an, wie stark die Regressionsebene auf der x_1y-Gerade steigt bzw. fällt. An jeder Stelle $X_{i2} = x_{i2}$ ist die Steigung der Geraden bezogen auf den Prädiktor X_1 konstant.
- Der Steigungsparameter β_2 gibt an, wie stark die Regressionsebene auf der x_2y-Gerade steigt bzw. fällt. An jeder Stelle $X_{i1} = x_{i1}$ ist die Steigung der Geraden bezogen auf den Prädiktor X_2 konstant.

Allgemeine Interpretation der Modellparameter α, β_1 und β_2

Es gilt: $E(Y_i|X_{i1} = x_{i1}, X_{i2} = x_{i2}) = \mu_i = \alpha + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2}$

- Der durchschnittliche Wert der AV in der Population für Personen mit einem Wert von 0 auf UV1 und UV2 beträgt α Einheiten.
- Erhöht sich der Wert auf UV1 um eine Einheit und der Wert auf UV2 bleibt konstant, erhöht sich der durchschnittliche Wert der AV in der Population um β_1 Einheiten.
- Erhöht sich der Wert auf UV2 um eine Einheit und der Wert auf UV1 bleibt konstant, erhöht sich der durchschnittliche Wert der AV in der Population um β_2 Einheiten.

Kausale Interpretation von β_1 und β_2

Wichtig: Der sprachliche Ausdruck „erhöht sich der Wert auf UV1 (oder UV2) um eine Einheit“ meint nicht, dass der UV-Wert einer konkreten Person tatsächlich ansteigt. Eine präzisere (aber umständlichere) Formulierung, bei der es offensichtlicher wird, dass das Modell keine Aussagen bezüglich eines kausalen Zusammenhang (im Sinne von „wenn ich die UV manipuliere, verändert sich dadurch auch die AV“), wäre:

> „Vergleicht man Personen aus der Population, die sich in ihren Werten auf UV1 (oder UV2) um genau eine Einheit unterscheiden, aber alle den gleichen Wert auf dem anderen Prädiktor aufweisen, dann haben die Personen mit dem höheren UV-Wert im Mittel einen um β_1 (oder β_2) Einheiten höheren Wert auf der AV.“

Ausblick: Mit Methoden der „kausalen Inferenz“ ist es durchaus möglich, mithilfe von Regressionsanalysen kausale Aussagen zu treffen. Dies bedeutet jedoch immer bestimmte Zusatzannahmen zu treffen, die niemals durch die Daten alleine überprüft werden können. Einen kurzen Einblick dazu werden wir in der nächsten Vorlesung geben.
Die grundsätzliche Logik ist dabei immer: Wenn ich bereit bin, konkrete Annahmen zu treffen, welche Variablen sich kausal auf andere auswirken können und welche nicht, kann ich unter Berücksichtigung bestimmter Regeln Steigungsparameter in der Regressionsanalyse als kausalen Effekt interpretieren.

MLR mit zwei Prädiktoren: Annahmen

Das Modell der MLR mit zwei stetigen Prädiktoren basiert auf den folgenden Annahmen:

1. Das Kriterium und die Prädiktoren hängen **linear** zusammen:

$$Y_i = \alpha + \beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2} + \varepsilon_i$$

2. Die Fehlervariablen ε_i sind **unabhängig** voneinander und folgen einer **Normalverteilung** mit Erwartungswert 0 und **konstanter Varianz** σ^2:

$$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

MLR mit zwei Prädiktoren: Parameter

- Das Modell der ELR enthält die folgenden Parameter:
  - α
  - β_1
  - β_2
  - σ^2

- Alle diese Parameter können geschätzt werden und für alle diese Parameter können statistische Hypothesentests konstruiert werden. Dies gilt auch für Kombinationen der Parameter.

- Je nach konkreter Fragestellung muss entschieden werden, welche Parameter geschätzt werden sollen bzw. welche Hypothesen getestet werden sollen.

Gliederung

- Bislang:
  - Modell der MLR mit zwei Prädiktoren
- Jetzt:
  - Parameterschätzung

Notation und Bezeichnungen bei der MLR

Die in der ELR festgelegte Verwendung der griechischen und lateinischen sowie der Groß- und Kleinschreibweise wird für die MLR übernommen:

- Notation für unbekannte Parameter: α, β_1, β_2, σ^2
- Notation für Schätzer: $\hat{\alpha} = A$, $\hat{\beta}_1 = B_1$, $\hat{\beta}_2 = B_2$, $\hat{\sigma}^2 = S^2$
- Notation für Schätzwerte: a, b_1, b_2, s^2

Schätzung der Parameter α, β_1 und β_2

- Analog zur ELR können die Schätzfunktionen für die unbekannten Modellparameter α, β_1 und β_2 mithilfe der **Methode der kleinsten Quadrate** („Least Squares“) bestimmt werden.

- Die Formeln sind etwas aufwendig, weswegen wir uns auf die Berechnung in R beschränken.

Standardfehler der Schätzfunktionen für β_1 und β_2

- Die Standardfehler der Schätzfunktionen B_1 und B_2 für β_1 und β_2 sind:

$$SE(B_1) = \sqrt{Var(B_1)} = \sqrt{\frac{1}{1 - r_{x1x2}^2} \cdot \frac{\sigma^2}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}}$$

$$SE(B_2) = \sqrt{Var(B_2)} = \sqrt{\frac{1}{1 - r_{x1x2}^2} \cdot \frac{\sigma^2}{\sum_{i=1}^{n}(x_{i2} - \bar{x}_2)^2}}$$

wobei r_{x1x2}^2 die quadrierte Korrelation zwischen den beiden Prädiktoren ist.

- Die geschätzten Standardfehler $\widehat{SE}(B_j)$ erhalten wir jeweils durch Ersetzen von σ^2 mit der Schätzfunktion S^2.

Konfidenzintervalle für β_1 und β_2

- Die Konfidenzintervalle für β_1 und β_2 haben jeweils die Form

$$\left[B_j \pm t_{1-\frac{\alpha}{2}} \cdot \widehat{SE}(B_j)\right]$$

- Hierbei ist $t_{1-\frac{\alpha}{2}}$ das $1-\frac{\alpha}{2}$- Quantil einer t-Verteilung mit $\nu = n-3$

Schätzwerte in R und Interpretation für α, β_1 und β_2


```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.22761    1.83006  -1.217    0.225
fie_nsb      0.77115    0.06325  12.192   <2e-16 ***
fie_abk      0.11357    0.07638   1.487    0.139
```

Interpretation der konkreten Schätzwerte:

- a: Die mittlere Depressionsschwere in der Population für Personen mit einer negativen Selbstbewertung und einer Abhängigkeitskognition von 0 Punkten beträgt -2,23 Punkte. (Diese Interpretation ist hier nicht sinnvoll.)
- b_1: Erhöht sich die negative Selbstbewertung um einen Punkt und die Abhängigkeitskognition bleibt konstant, erhöht sich die mittlere Depressionsschwere in der Population um 0,77 Punkte.
- b_2: Erhöht sich die Abhängigkeitskognition um einen Punkt und die negative Selbstbewertung bleibt konstant, erhöht sich die mittlere Depressionsschwere in der Population um 0,11 Punkte.

Konfidenzintervalle für α, β_1 und β_2

```
                  2.5 %    97.5 %
(Intercept) -5.83771259 1.3824841
fie_nsb      0.64637580 0.8959172
fie_abk     -0.03709342 0.2642380
```

- Der Parameter α hat in unserem Beispiel keine sinnvolle Interpretation, weswegen wir auch das KI nicht sinnvoll interpretieren können.
- Interpretation des konkreten KI für β_1:

  Wir gehen davon aus, dass die mittlere Depressionsausprägung zwischen 0,65 und 0,90 Punkte steigt, wenn die negative Selbstbewertung um einen Punkt steigt und die Abhängigkeitskognition konstant bleibt.
- Interpretation des konkreten KI für β_2:

  Wir gehen davon aus, dass die mittlere Depressionsausprägung zwischen -0,04 und 0,26 Punkte steigt, wenn die Abhängigkeitskognition um einen Punkt steigt und die negative Selbstbewertung konstant bleibt.

Schätzung der unbekannten Fehlervarianz σ^2

- Die **Schätzfunktion** für die **unbekannte Fehlervarianz** σ^2 lässt sich für eine MLR mit zwei Prädiktoren darstellen als:

$$S^2 = \hat{\sigma}^2 = \frac{1}{n-3}\sum_{i=1}^{n}(Y_i - (A + B_1 \cdot X_{i1} + B_2 \cdot X_{i2}))^2 = \frac{1}{n-3}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \frac{1}{n-3}\sum_{i=1}^{n}E_i^2$$

- Zieht man die Wurzel aus der Schätzfunktion S^2, erhält man (wie in der ELR) eine Schätzfunktion für $\sigma = \sqrt{\sigma^2}$:

$$S = \sqrt{\frac{\sum_{i=1}^{n}E_i^2}{n-3}}$$

- Dessen Realisation (der Standardschätzfehler) ist

$$s = \sqrt{\frac{\sum_{i=1}^{n}e_i^2}{n-3}}$$

- In R:

```
Residual standard error: 7.854 on 188 degrees of freedom
Multiple R-squared:  0.5176,  Adjusted R-squared:  0.5125
F-statistic: 100.9 on 2 and 188 DF,  p-value: < 0.00000000000000022
```

Schätzung der Modellgleichung

$$Y_i = \alpha + \beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2} + \varepsilon_i$$


```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.22761    1.83006  -1.217    0.225
fie_nsb      0.77115    0.06325  12.192   <2e-16 ***
fie_abk      0.11357    0.07638   1.487    0.139
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.854 on 188 degrees of freedom
Multiple R-squared:  0.5176,  Adjusted R-squared:  0.5125
F-statistic: 100.9 on 2 and 188 DF,  p-value: < 2.2e-16
```

- Die geschätzte Modellgleichung ist also

$$Y_i = -2{,}2 + 0{,}77 \cdot X_{i1} + 0{,}11 \cdot X_{i2} + \varepsilon_i, \qquad \text{mit } \varepsilon_i \sim N(0, 7.85^2)$$

Gliederung

- Bislang:
  - Modell der MLR mit zwei Prädiktoren
  - Parameterschätzung
- Jetzt:
  - Hypothesentests

Hypothesentests bei der MLR mit zwei Prädiktoren

Wir besprechen (zunächst) zwei Arten von Hypothesentests bei der MLR:

1. Hypothesentests für $H_0: \beta_j = 0$: Mithilfe der Hypothesentests lässt sich jeweils die Nullhypothese, dass der Regressionsparameter gleich Null ist, überprüfen.
2. Omnibus-Test: Mithilfe eines Omnibus-Tests lässt sich die Nullhypothese, dass alle Steigungsparameter β_j gleich Null sind, überprüfen.

Hypothesentests für einzelne Prädiktoren: Statistische Hypothesen

- Bei einer MLR mit zwei Prädiktoren können (u.a.) die folgenden zwei statistischen Hypothesen überprüft werden:

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

$$H_0: \beta_2 = 0$$
$$H_1: \beta_2 \neq 0$$

- Auf unser Beispiel bezogen: Es soll überprüft werden, ob die Depressionsschwere mit negativer Selbstbewertung bzw. mit Abhängigkeitskognition linear zusammenhängt, wenn der jeweils andere Prädiktor konstant gehalten wird.

Hypothesentests für einzelne Prädiktoren: Teststatistiken

- Die Teststatistiken sind

$$T_1 = \frac{B_1}{\sqrt{\hat{V}ar(B_1)}}$$

$$T_2 = \frac{B_2}{\sqrt{\hat{V}ar(B_2)}}$$

- Unter der Geltung der jeweiligen Nullhypothese folgen diese Teststatistiken jeweils einer t-Verteilung mit $\nu = n - 3$.
- Allgemein gilt für den Parameter ν im Rahmen dieses Hypothesentests im Regressionsmodell $\nu = n - k - 1$ wobei k der Anzahl der Prädiktoren im Modell entspricht (also im vorliegenden Fall $k = 2$).
- Der kritische Bereich ist jeweils beidseitig.

Hypothesentests für einzelne Prädiktoren: Beispiel in R


```
> fit = lm(bdi_ges ~ fie_nsb + fie_abk, data = daten)
> summary(fit)

Call:
lm(formula = bdi_ges ~ fie_nsb + fie_abk, data = daten)

Residuals:
    Min      1Q  Median      3Q     Max
-16.317  -5.359  -1.184   5.426  27.563

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.22761    1.83006  -1.217    0.225
fie_nsb      0.77115    0.06325  12.192   <2e-16 ***
fie_abk      0.11357    0.07638   1.487    0.139
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.854 on 188 degrees of freedom
Multiple R-squared:  0.5176,      Adjusted R-squared:  0.5125
F-statistic: 100.9 on 2 and 188 DF,  p-value: < 2.2e-16
```

- Testentscheidung?

Omnibus-Test: Statistische Hypothesen

- Die folgenden Hypothesen können mit einem Omnibus-Test überprüft werden:

H_0: $\beta_1 = \beta_2 = 0$

H_1: $\beta_j \neq 0$ für mindestens ein j mit j $= 1{,}2$

- Mithilfe dieses Omnibus-Tests kann also überprüft werden, ob bei mindestens einem der Prädiktoren der lineare Zusammenhang mit dem Kriterium ungleich Null ist (bei Konstanthaltung des jeweils anderen Prädiktors).

Omnibus-Test: Teststatistik

- Die Teststatistik des Omnibus-Tests ist

$$F = \frac{\frac{1}{2}\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}{\frac{1}{n-3}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2} = \frac{\frac{1}{2}\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}{S^2}$$

- Unter der Geltung der Nullhypothese folgt diese Teststatistik einer F-Verteilung.
- Der kritische Bereich liegt auf der rechten Seite.

Omnibus-Test: Beispiel in R

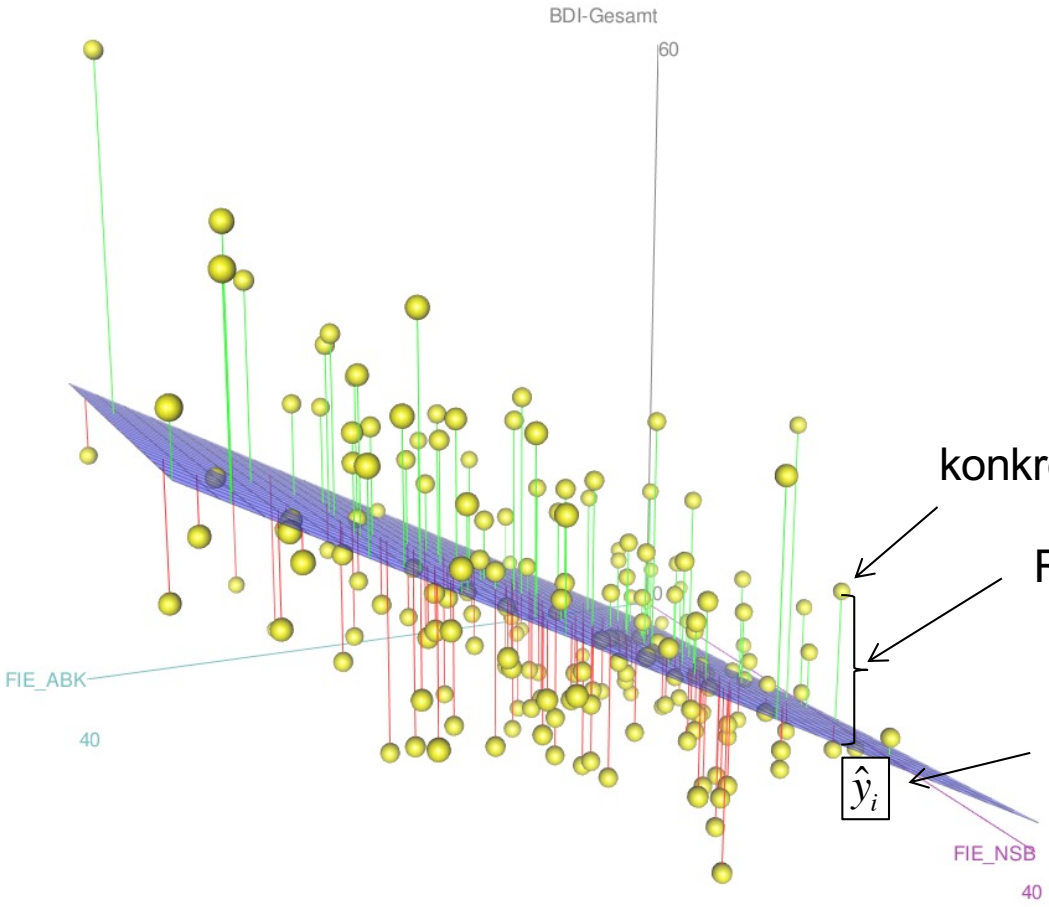

```
> fit = lm(bdi_ges ~ fie_nsb + fie_abk, data = daten)
> summary(fit)

Call:
lm(formula = bdi_ges ~ fie_nsb + fie_abk, data = daten)

Residuals:
    Min      1Q  Median      3Q     Max
-16.317  -5.359  -1.184   5.426  27.563

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.22761    1.83006  -1.217    0.225
fie_nsb      0.77115    0.06325  12.192   <2e-16 ***
fie_abk      0.11357    0.07638   1.487    0.139
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.854 on 188 degrees of freedom
Multiple R-squared:  0.5176,      Adjusted R-squared:  0.5125
F-statistic: 100.9 on 2 and 188 DF,  p-value: < 2.2e-16
```

Gliederung

- Bislang:
  - Modell der MLR mit zwei Prädiktoren
  - Parameterschätzung
  - Hypothesentests
- Jetzt:
  - Vorhersage und Konfidenzintervall für y_i

Graphische Veranschaulichung: Vorhersagewert und Residuum

- Auch bei einer multiplen Regression wird zwischen dem konkreten AV-Wert, dem Vorhersagewert und dem Residualwert unterschieden.



konkreter AV-Wert y_i

Residualwert $e_i = y_i - \hat{y}_i$

Vorhersagewert $\hat{y}_i$ einer Person *i* mit einer negativen Selbstbewertung x_{i1} und einer Abhängigkeitskognition x_{i2}

Vorhersage für y_i

- Genau wie in der ELR können wir bei gegebenen UV-Werten einen unbekannten AV-Wert vorhersagen.

- Konkreter Vorhersagewert:

$$\hat{y}_i = a + b_1 \cdot x_{i1} + b_2 \cdot x_{i2}$$

- Bsp.: Berechnung des Vorhersagewerts $\hat{y}_i$ für eine Person i mit einer negativen Selbstbeurteilung von $x_{i1} = 5$ und einer Abhängigkeitskognition von $x_{i2} = 13$:

$$\hat{y}_i = -2{,}2 + 0{,}77 \cdot 5 + 0{,}11 \cdot 13 = 3{,}08$$

- Man sagt vorher, dass die Depressionsschwere einer Person mit einer negativen Selbstbewertung von $x_1 = 5$ Punkten und mit einer Abhängigkeitskognition von $x_2 = 13$ Punkten, bei ca. 3 Punkten liegt.

Neue Intuition für die Interpretation von β_1 und β_2

- Zur Erinnerung: $b_1 = 0.77$
  Steigt die negative Selbstbewertung um einen Punkt und die Abhängigkeitskognition bleibt konstant, so erhöht sich im Mittel die Depressionsschwere um 0.77 Punkte
- Betrachte zwei Personen, die sich um genau einen Punkt in der negativen Selbstbewertung unterscheiden, aber die gleiche Abhängigkeitskognition aufweisen:

  Person 1: $x_{11} = 20, x_{12} = 25$
  Person 2: $x_{21} = 21, x_{22} = 25$
- Berechne die Vorhersagewerte:

$$\hat{y}_1 = -2{,}2 + 0{,}77 \cdot 20 + 0{,}11 \cdot 25 = 16{,}03$$

$$\hat{y}_2 = -2{,}2 + 0{,}77 \cdot 21 + 0{,}11 \cdot 25 = 16{,}81$$

- Berechne die Differenz der beiden Vorhersagewerte:

$$\hat{y}_2 - \hat{y}_1 = 0{,}77$$

- Die Differenz entspricht genau dem Schätzwert $b_1 = 0{,}77$.
  An diesem Beispiel sieht man, warum die Interpretation von β_1 bzw. b_1 genau so lautet wie besprochen. Sie gilt eben nur dann, wenn der andere Prädiktor konstant bleibt.

KI für y_i

```
> daten_neu <- data.frame(fie_nsb = 33, fie_abk = 29)
> daten_neu
  fie_nsb fie_abk
1      33      29
> predict(fit, daten_neu, interval = 'prediction', level = 0.95)
       fit      lwr      upr
1 26.51382 10.88901 42.13862
```

- Wir gehen davon aus, dass die Depressionsschwere einer Person mit einer negativen Selbstbewertung von $x_{i1} = 33$ Punkten und mit einer Abhängigkeitskognition von $x_{i2} = 29$ Punkten zwischen 10,89 und 42,14 Punkten liegt.

Gliederung

- Bislang:
  - Modell der MLR mit zwei Prädiktoren
  - Parameterschätzung
  - Hypothesentests
  - Vorhersage und Konfidenzintervall für y_i
- Jetzt:
  - Regressionsdiagnostik

Regressionsdiagnostik für ELR und MLR

- Mithilfe der Regressionsdiagnostik werden die Annahmen des Regressionsmodells überprüft.
- Zusätzlich wird im Allgemeinen eine Ausreißeranalyse durchgeführt.

- Wichtig: Die Regressionsdiagnostik ist **vor** den inferenzstatistischen Verfahren durchzuführen.

Wdh: Annahmen der Linearen Regression

Im Modell der Linearen Regression werden die folgenden Annahmen getroffen:

1. Die **UV** und die **AV** hängen **linear** zusammen:

$$Y_i = \alpha + \beta \cdot X_i + \varepsilon_i$$

2. Alle **Fehler** ε_i sind **unabhängig voneinander** und folgen einer **Normalverteilung** mit **Erwartungswert Null** und **konstanter Varianz** σ^2:

$$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

Bemerkung: Die konstante Varianz bedeutet hier, dass die Fehler ε_i überall die gleiche Varianz aufweisen, egal welcher Wert $\hat{y}$ in Abhängigkeit der UV(s) vorhergesagt wird.

Regressionsdiagnostik: Überblick

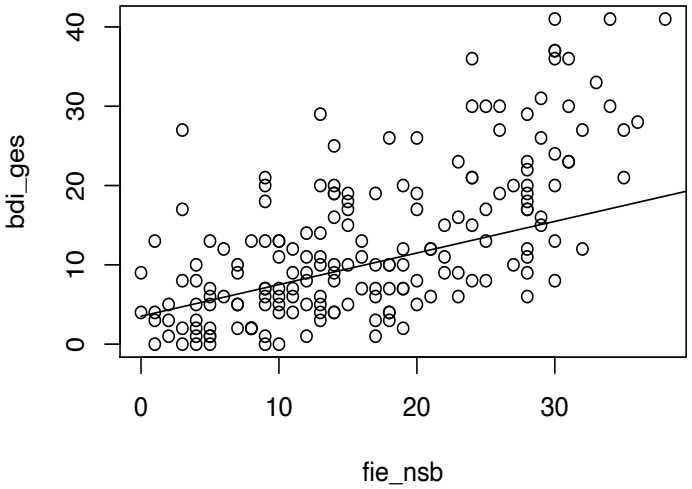
- Die Annahme der Linearität, die Annahme der **Normalverteilung der Fehler** und die Annahme der **konstanten Fehlervarianz** können verletzt sein.
- Die Unabhängigkeit der Fehler kann durch die Stichprobenziehung sichergestellt werden.
- Bemerkung: In Regressionsmodellen werden häufig statt konstanter Fehlervarianz die Begriffe Varianzhomogenität oder Homoskedastizität verwendet. Das Gegenteil bezeichnet man als Varianzheterogenität oder Heteroskedastizität.
- Die Regressionsdiagnostik umfasst also:
  - Überprüfung der Linearitätsannahme
  - Überprüfung der Normalverteilungsannahme
  - Überprüfung der Homoskedastizitätsannahme
  - Identifizierung von Ausreißern

Regressionsdiagnostik: Vorbemerkung

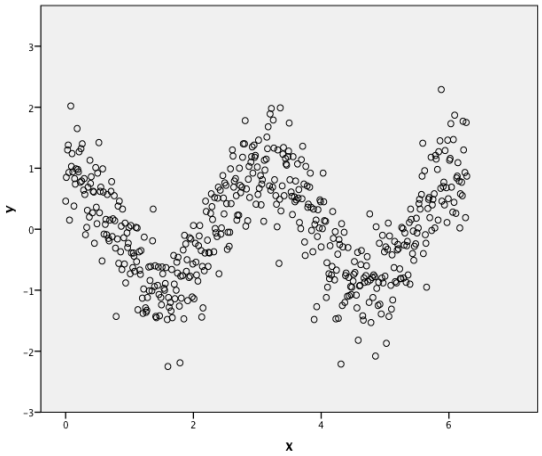
- Für die Regressionsdiagnostik müssen notwendigerweise die **Residuen** verwendet werden, obwohl sich die Annahmen eigentlich auf die Fehlervariablen beziehen.
- Für die deskriptiv-statistische Überprüfung werden häufig so genannte **Residuen-Diagramme** erstellt, die die **Verteilung der Residualwerte graphisch darstellen**.
- Es empfiehlt sich aus mathematischen Gründen, hierbei sogenannte **standardisierte Residuen** zu verwenden.
- Mathematischer Hintergrund: Man kann zeigen, dass die Varianz der Zufallsvariable E_i, also $Var(E_i) = Var(Y_i - \hat{Y}_i)$, vom Prädiktor X_i abhängig ist. Somit ist die Varianz der E_i für unterschiedliche Realisationen von X_i **nicht konstant,** selbst wenn die eigentlich interessante Varianz der Fehler ε_i tatsächlich konstant ist. Dieses Problem wird durch die Standardisierung der Residuen gelöst.

Überprüfung der Linearitätsannahme

linear (zumindest annähernd)



nicht linear



- Die Linearitätsannahme kann für eine ELR deskriptiv-statistisch mithilfe des Streudiagramms der AV gegen die UV überprüft werden.
- Dabei gelten die gleichen Einschränkungen wie bei anderen deskriptiv-statistischen Methoden, die wir in der Vergangenheit im Kontext der Überprüfung von Modellannahmen kennen gelernt haben.
- Ob die Annahme verletzt ist, ist oft nur subjektiv einzuschätzen und nicht immer so klar wie im hier gezeigten Beispiel.

Verletzung der Linearitätsannahme

- Falls man davon ausgehen muss, dass die Linearitätsannahme stark verletzt ist, sind die Parameter des Modells nicht mehr sinnvoll interpretierbar: Falls es nicht zumindest einen monotonen „je – desto“ Zusammenhang zwischen UV und AV gibt (Gegenbeispiel: unteres Streudiagramm auf der letzten Folie), lässt sich der Zusammenhang nicht sinnvoll durch einen Steigungsparameter quantifizieren.
- Die Verletzung der Linearitätsannahme beeinträchtigt also die Interpretation der Modellparameter selbst und nicht nur der inferenzstatistischen Verfahren zur Schätzung/Testung dieser Parameter. Daran kann auch eine große Stichprobe nichts ändern.
- Die Linearitätsannahme sollte daher in jedem Fall überprüft werden. Falls wir davon ausgehen müssen, dass sie stark verletzt ist, müssen andere (nonlineare) Regressionsmodelle verwendet werden.

Überprüfung der Linearitätsannahme in der MLR

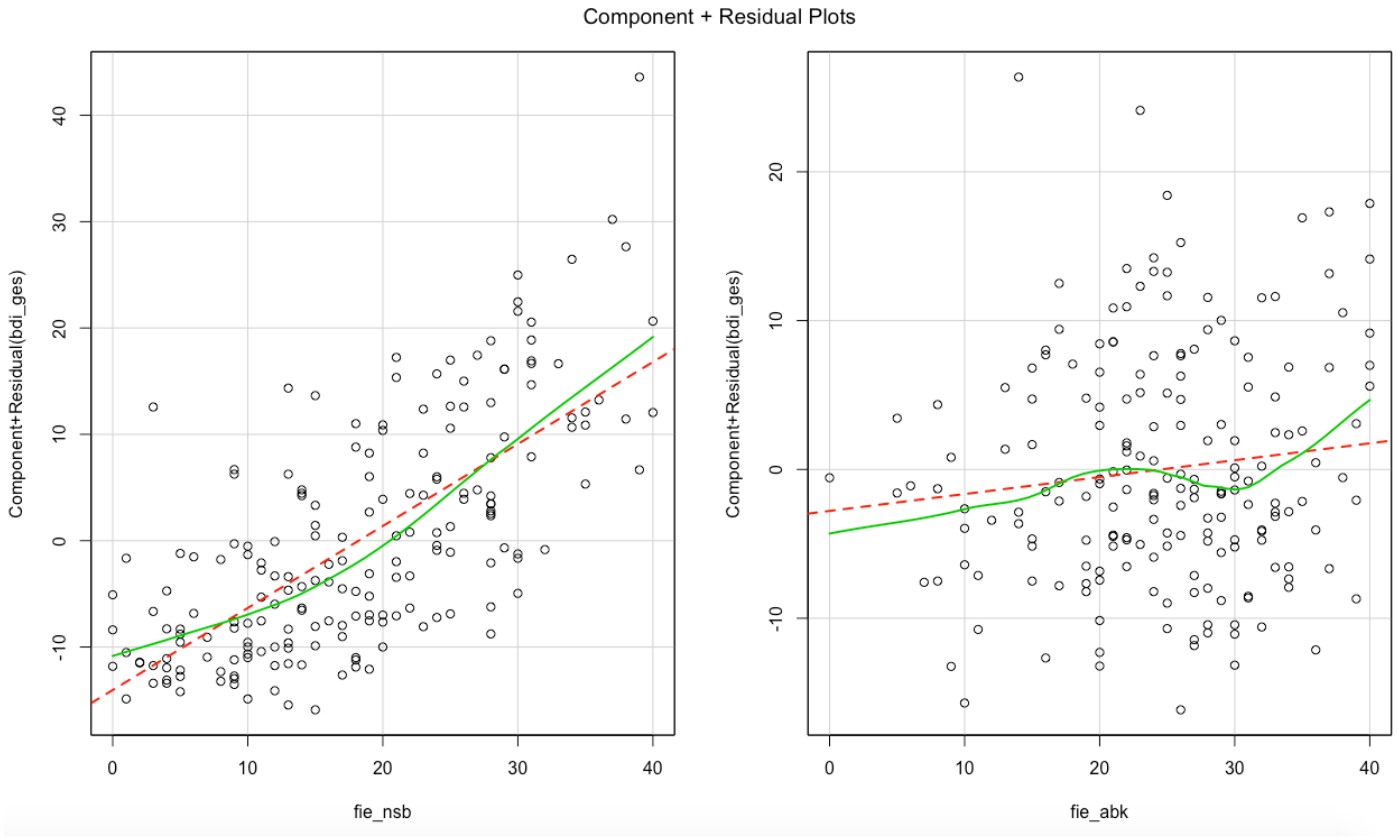
- In der Multiplen Linearen Regression muss die Linearitätsannahme für jeden Prädiktor einzeln überprüft werden.
- Hierbei müssen die Einflüsse der jeweils anderen Prädiktoren für die deskriptiv-statistische Betrachtung herausgerechnet (= „partialisiert“) werden.
- Zur Überprüfung der Linearitätsannahme stellt R sogenannte **partielle Residuen-Plots** zur Verfügung.
- Auf der y-Achse werden die **partiellen Residuen** aufgetragen:
  Die AV-Werte, von denen die allein aus den jeweils anderen Prädiktoren vorhergesagten AV-Werte abgezogen wurden.
- Prinzipielles Vorgehen für eine MLR mit zwei UVs:
  - Für die erste UV:

$$y_i - (a + b_2 x_{i2})$$

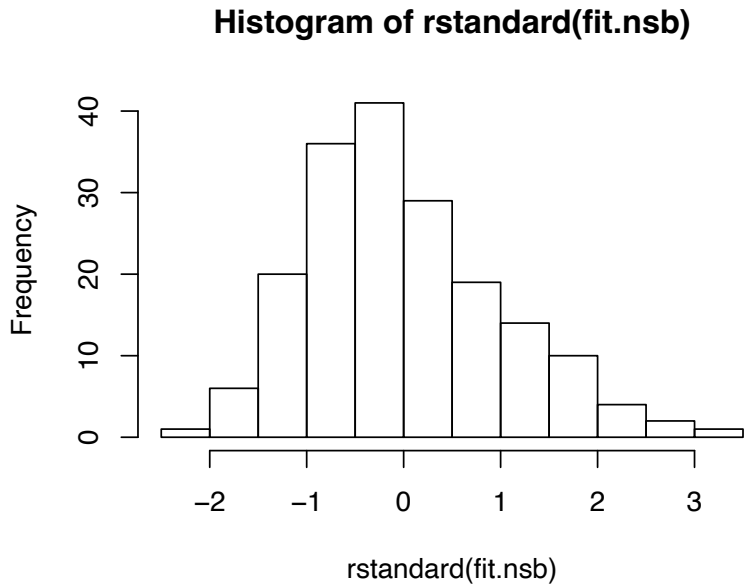
  - Für die zweite UV:

$$y_i - (a + b_1 x_{i1})$$

Partielle Residuen-Plots zur Überprüfung der Linearitätsannahme in der MLR

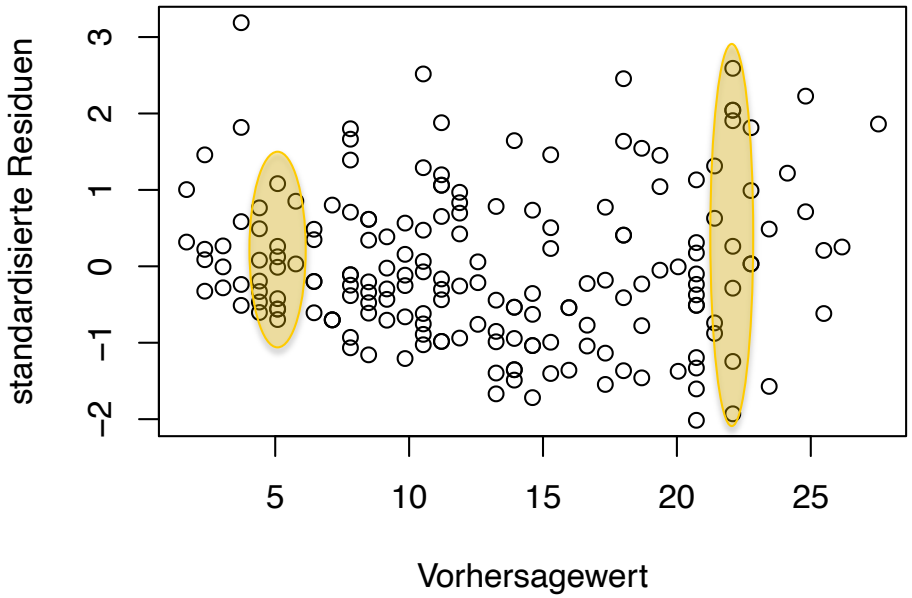


Überprüfung der Normalverteilungsannahme



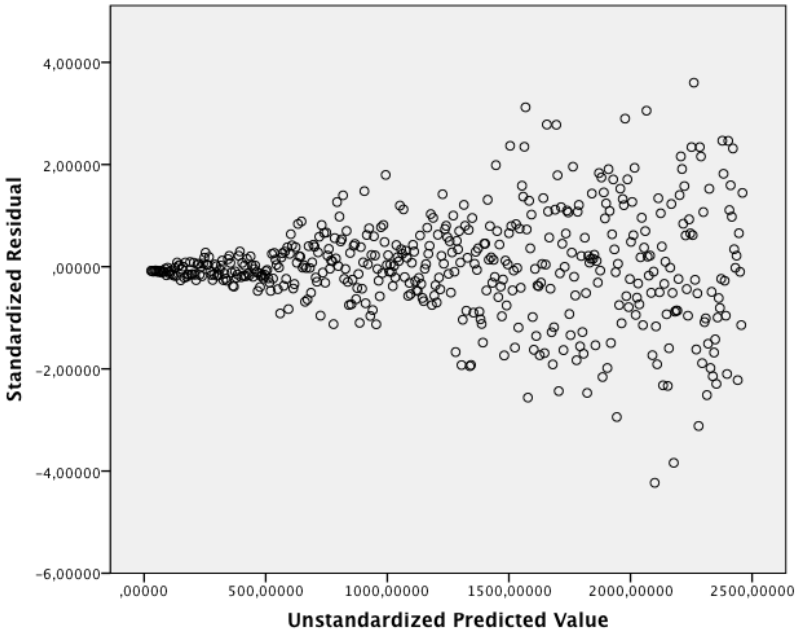
- Die Prüfung der Normalverteilungsannahme der Fehlervariablen kann z.B. mithilfe eines Histogramms der (standardisierten) Residuen erfolgen.
- Auch hier ist die Einschätzung wieder sehr subjektiv.
- Es lässt sich zeigen, dass die von uns besprochenen inferenzstatistischen Verfahren im Rahmen des ELR-Modells in großen Stichproben auch bei Verletzung der Normalverteilungsannahme approximativ die von uns vorausgesetzten Eigenschaften erfüllen.
- Eine Verletzung der Normalverteilungsannahme ist daher in großen Stichproben nicht so schlimm.

Überprüfung der Homoskedastizität



- Homoskedastizität = Die Streuung σ^2 der Fehler ϵ_i ist für alle vorhergesagte Werte $\hat{y}_i$ identisch.
- In diesem Beispiel lässt sich folgender Trend beobachten: Die Streuung der standardisierten Residualwerte e_i ist bei geringen Vorhersagewerten $\hat{y}_i$ kleiner.
- Bei einem vorhergesagten Depressionswert von 5 BDI-Punkten gibt es z.B. kleinere Abweichungen der beobachteten Werte vom vorhergesagten Wert als bei einem Depressionswert von 22 BDI-Punkten.
- Dies deutet eventuell auf eine Verletzung der Homoskedastizität hin.

Verletzung der Homoskedastizität



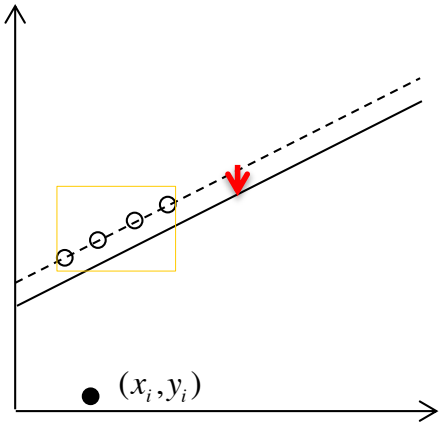
- Bei Vorliegen von Heteroskedastizität (Gegenteil von Homoskedastizität) können die inferenzstatistischen Verfahren auch in großen Stichproben unter Umständen stark beeinträchtigt sein.
- Falls nach graphischer Überprüfung davon ausgegangen werden muss, dass die Homoskedastizität stark verletzt ist, sollte auf alternative inferenzstatistische Verfahren (allgemeines lineares Modell) zurückgegriffen werden.
- Wie immer gilt: Selten ist das deskriptiv-statistische Bild so eindeutig wie links…

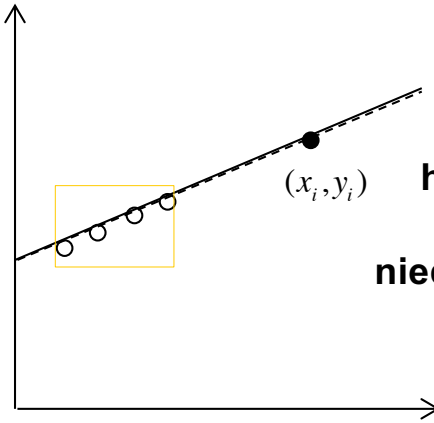
Ausreißeranalyse: Klassifizierung von Ausreißerwerten

- In Regressionsmodellen können ungewöhnlich große oder kleine Werte von AV oder UVs (= Ausreißer) unter Umständen eine stark verzerrende Wirkung haben.
- Entsprechend wichtig ist es zu überprüfen, welche Werte einen solchen Einfluss hätten und ob sie dabei valide Information enthalten oder beispielsweise tatsächlich eher auf Fehler der Datenerhebung zurückzuführen sind.
- Im Rahmen von Regressionsmodellen werden die Ausreißer entsprechend ihrer Wirkung auf die Regressionsgerade klassifiziert. Man unterscheidet dabei zwischen einem
  - **Hebelwert** (Leverage): Vom Betrag her ungewöhnlich große Abweichung des UV-Werts x_i von $\bar{x}_{ohne\ Ausreißer}$
  - **Diskrepanzwert** (Discrepancy): Vom Betrag her ungewöhnlich große Abweichung des AV-Werts y_i von der geschätzten Regressionsgerade (ohne Ausreißer)
  - und einem **Einflusswert** (Influence): Kombination aus Hebel- und Diskrepanzwert.
- Vor allem Einflusswerte können die inferenzstatistischen Verfahren erheblich verzerren.

Illustration der Wirkung von Ausreißern auf die Regressionskoeffizienten

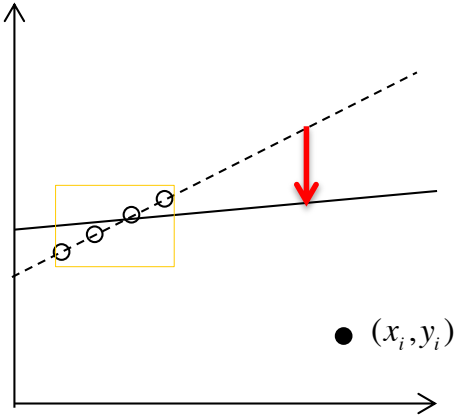
hoher Diskrepanzwert & niedriger Hebelwert





hoher Hebelwert & niedriger Diskrepanz-wert

hoher Einflusswert = hoher Hebelwert & hoher Diskrepanzwert



--------- Regressionsgerade ohne Ausreißer

——— Regressionsgerade mit Ausreißer

Identifizierung von Ausreißern mit R

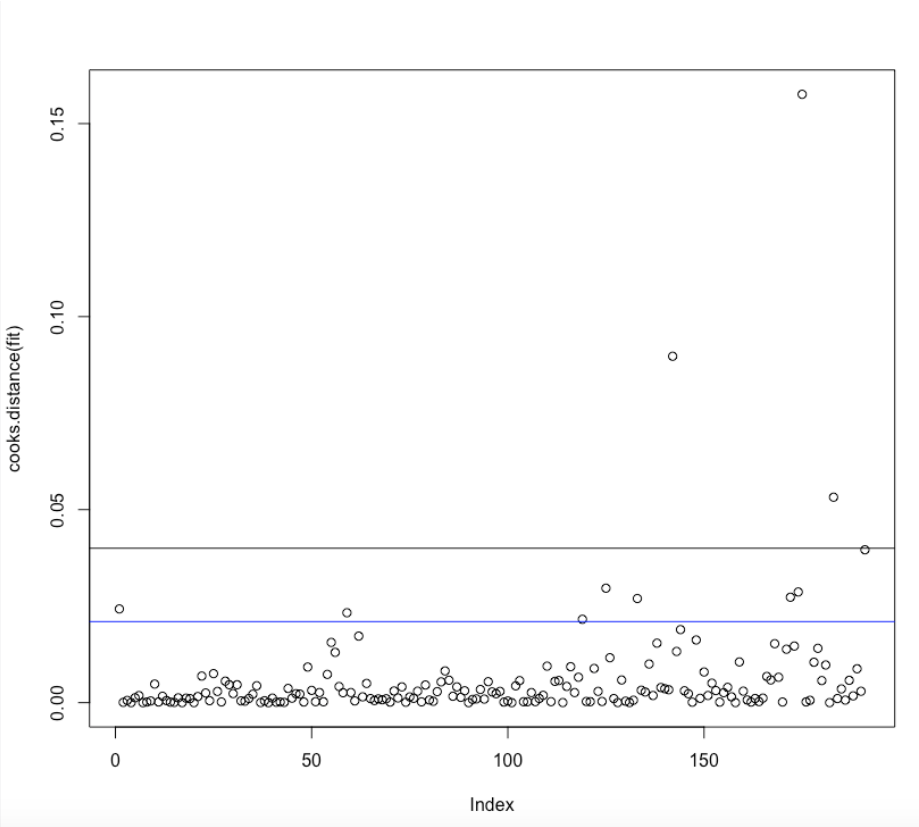
R bietet eine Reihe von Methoden zur Identifizierung von Ausreißerwerten bei regressionsanalytischen Verfahren.

- Erkennung von Hebelwerten: z.B. deskriptiv-statistische Analyse der UV
- Erkennung von Diskrepanzwerten: z.B. standardisierte Residuen
- Erkennung von Einflusswerten: z.B. Cook‘s Distanz

Cook‘s Distanz

- Die Cook‘s Distanz gibt für jede Beobachtung an, wie stark sich die Vorhersagewerte aller Beobachtungen ändern würden, wenn diese Beobachtung aus der Analyse ausgeschlossen wird.
- Je größer die Cook‘s Distanz, desto höher der Einflusswert der Beobachtung.
- Als Cut-off-Wert wird vorgeschlagen, alle Beobachtungen als kritisch zu betrachten, die eine Cook‘s Distanz von über $\frac{4}{n}$ haben, wobei n die Stichprobengröße ist.
- Eine andere Möglichkeit wäre, sich die Beobachtungen mit den extremsten Cook‘s Distanzen genauer anzusehen.

Diagramm für Cook‘s Distanz



- In diesem Diagramm wird für jede Versuchsperson die Cook‘s Distanz in ein Koordinatensystem eingetragen.
- Hier liegen 11 Personen über dem Cut-Off-Wert von $\frac{4}{n} = \frac{4}{191} = 0.021$. Drei Personen haben besonders auffallende Werte von über 0.04.

Inspektion der Beobachtungen mit hohen Einflusswerten

- Welche Personen haben einen hohen Einflusswert?
- 1. Möglichkeit:

```
> which(cooks.distance(fit) > 0.021)
  1  59 119 125 133 142 172 174 175 183 191
  1  59 119 125 133 142 172 174 175 183 191
```

- 2. Möglichkeit: selbstgewählter Cut-off-Wert anhand der Grafik

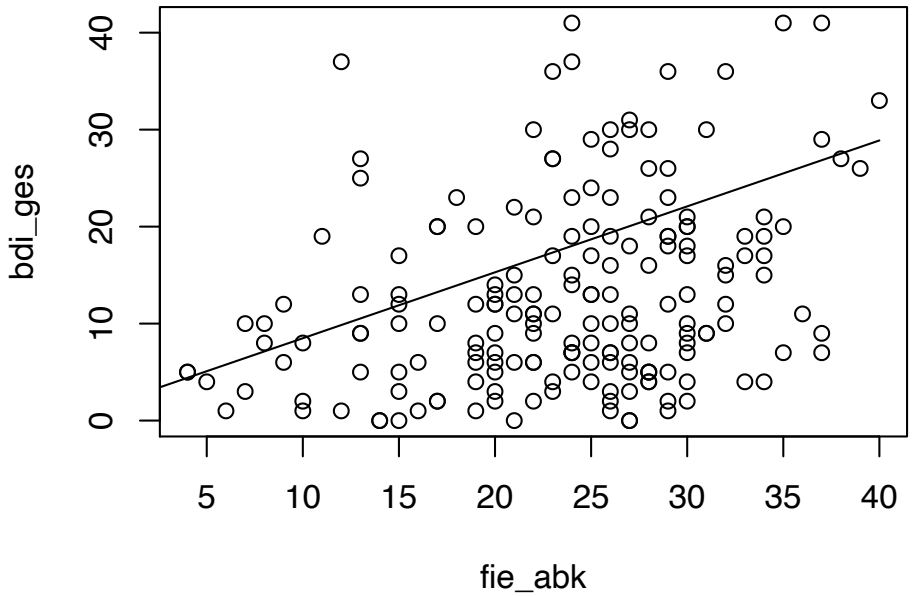
```
> which(cooks.distance(fit) > 0.04)
142 175 183
142 175 183
```

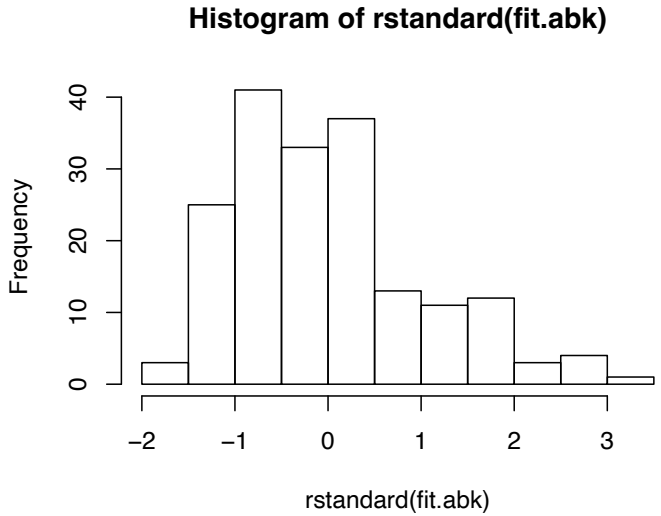
- Die Personen mit den Nummern 142, 175 und 183 haben besonders starke Einflusswerte.
- Die Daten dieser Personen sollten genauer inspiziert werden. Was ist die Ursache für die Ausreißer?
- **Wichtig: Ausreißer sollten nur aus den Daten entfernt werden, wenn es einen guten Grund dafür gibt (z.B. Fehler bei der Dateneingabe).**

2. Beispiel der letzten VL – „fie_abk“

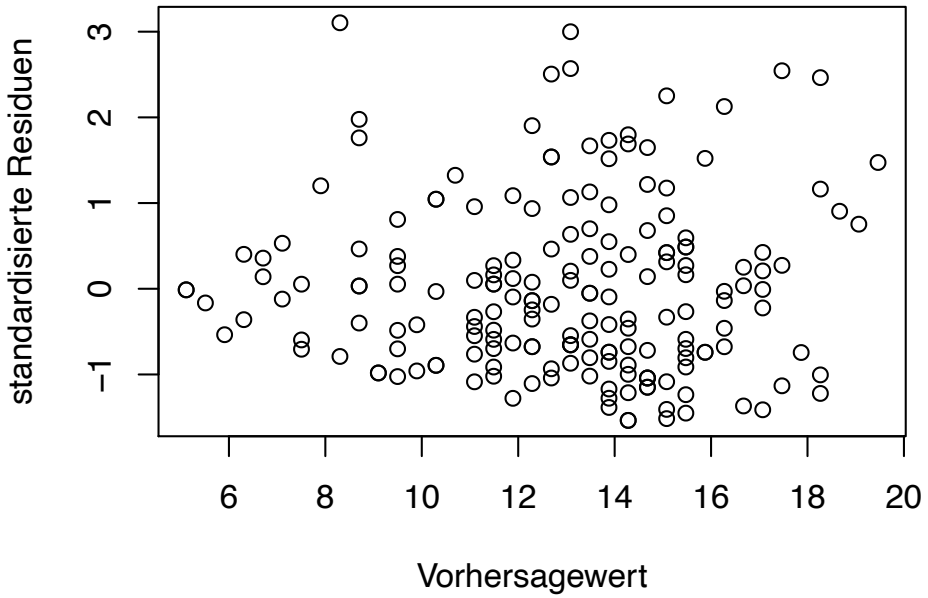
- Im Folgenden werden wir eine Regressionsdiagnostik bezogen auf das zweite Beispiel der letzten VL durchführen.
- UV („fie_abk“): Einschätzung der Abhängigkeitskognition (Skala des Fragebogens zur Erfassung irrationaler Einstellungen)

Überprüfung der Linearitäts- und Normalverteilungsannahme

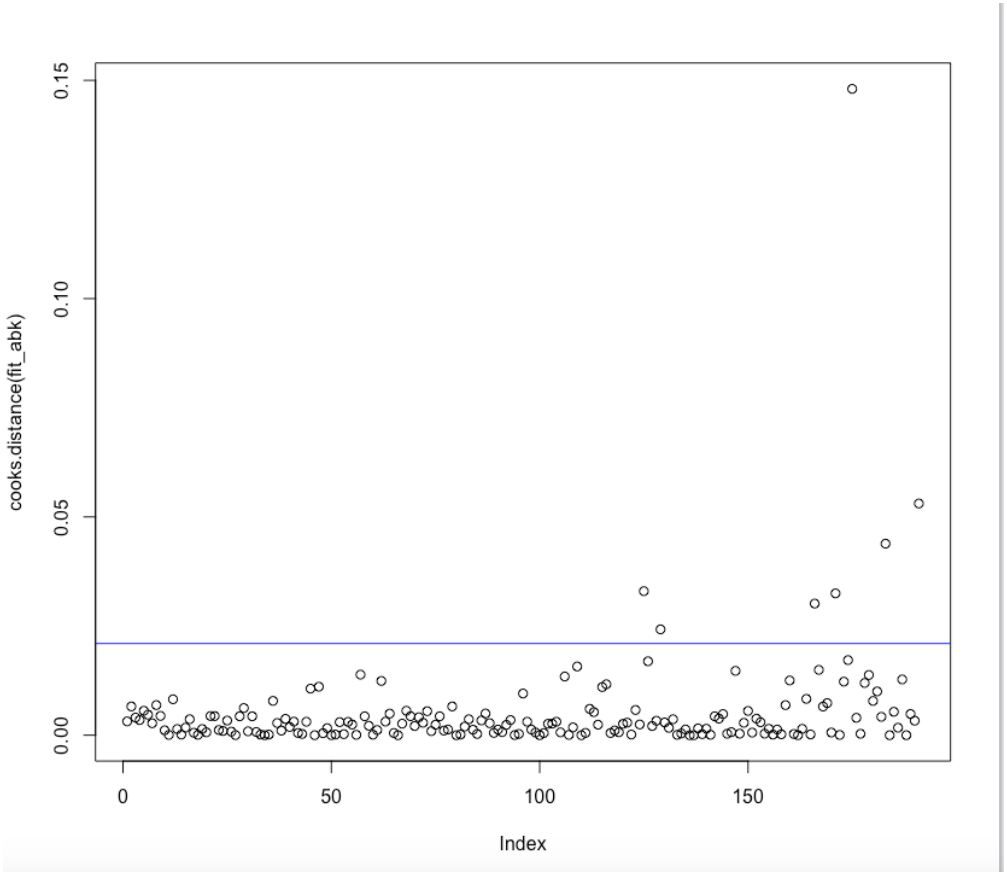




Überprüfung der Homoskedastizität



Identifizierung von Ausreißern



$$\frac{4}{191} = 0.021$$