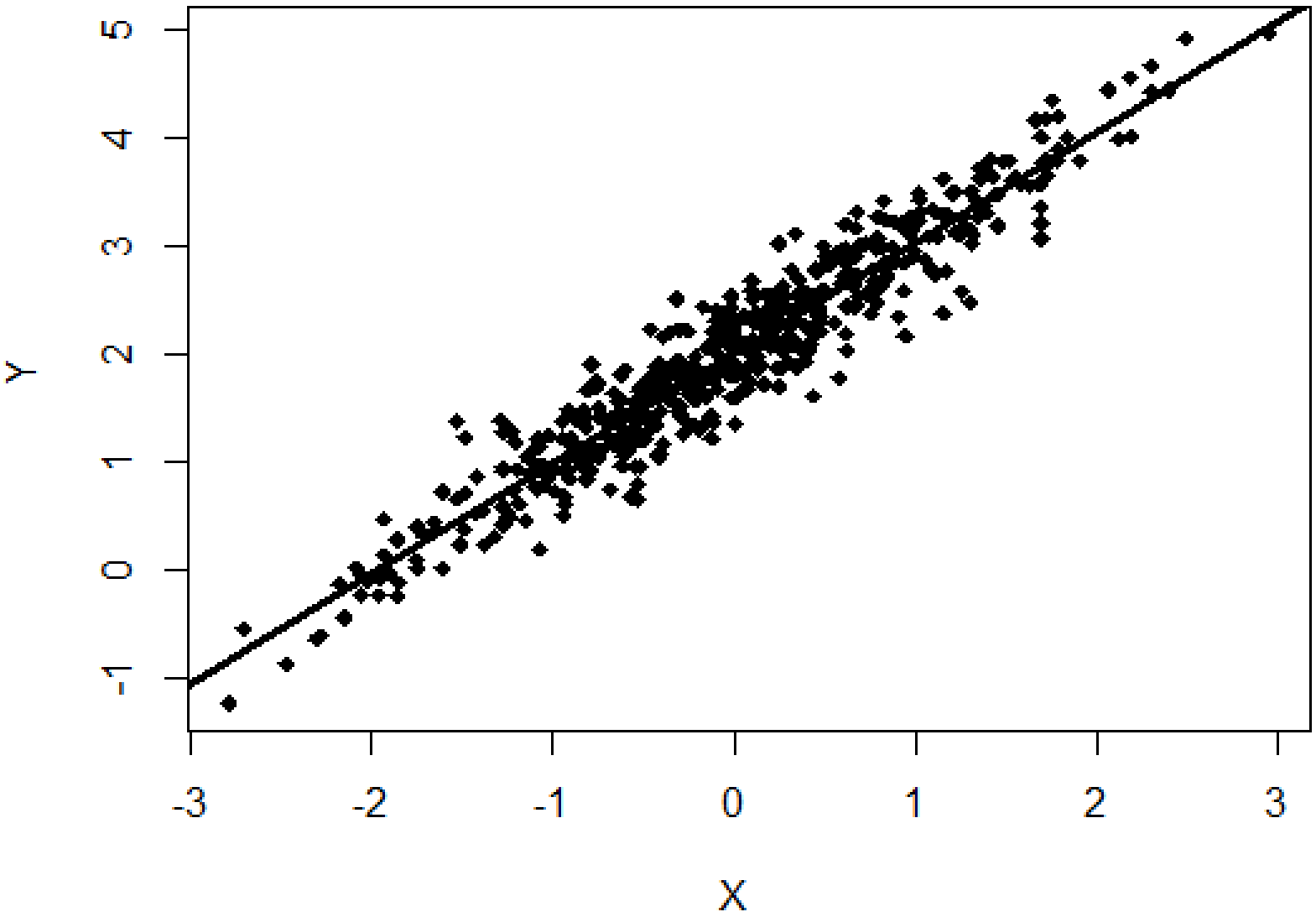
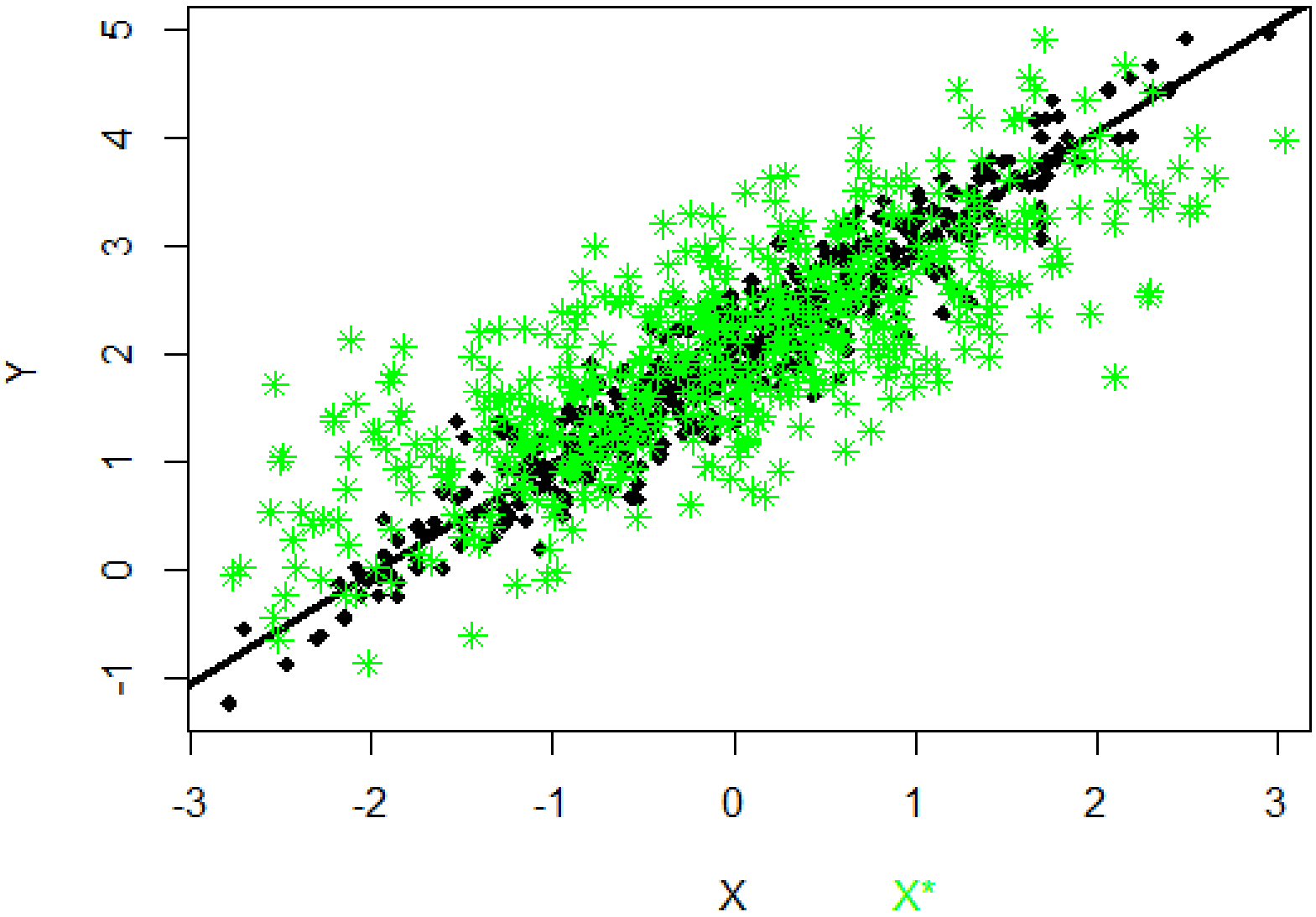
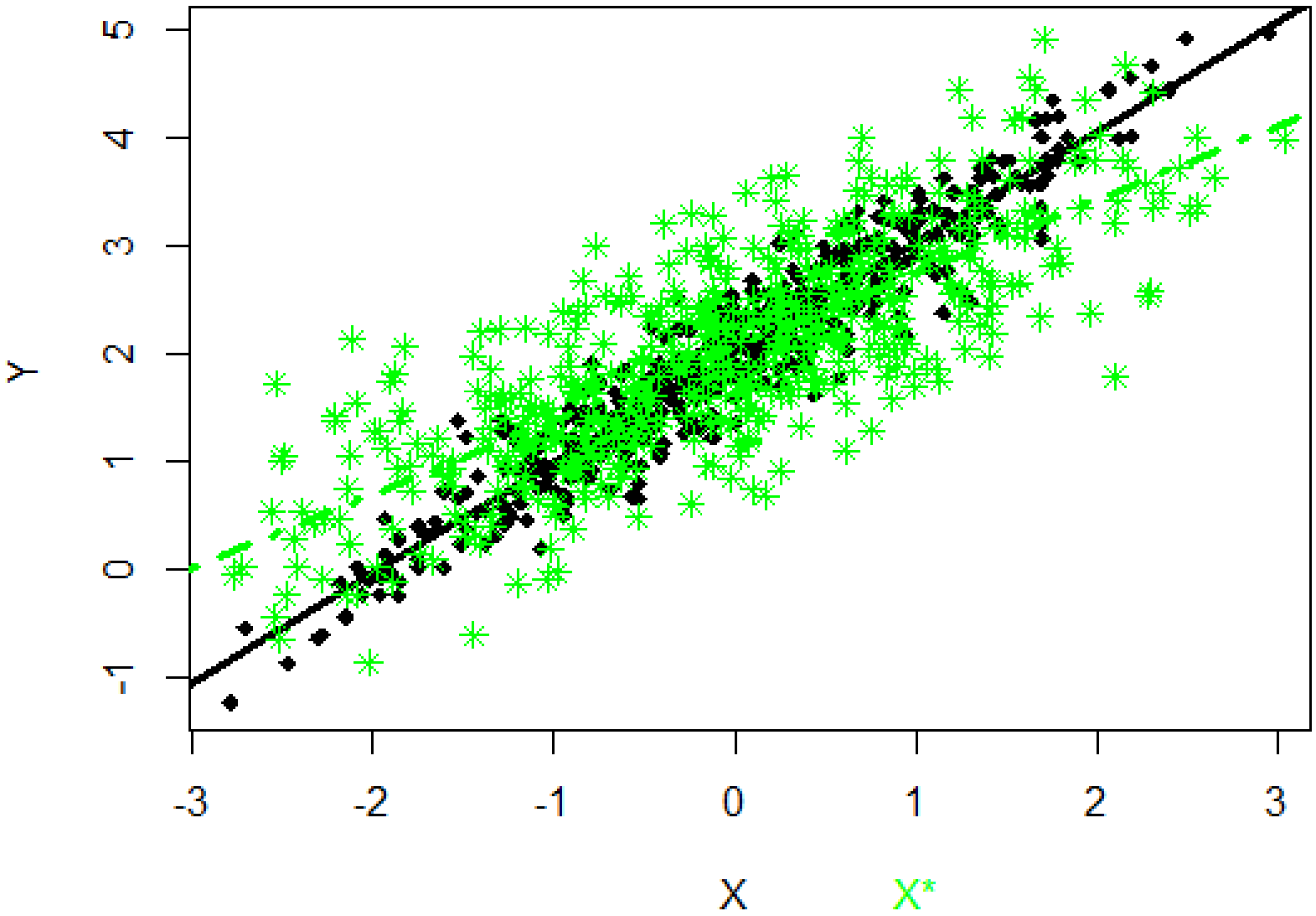
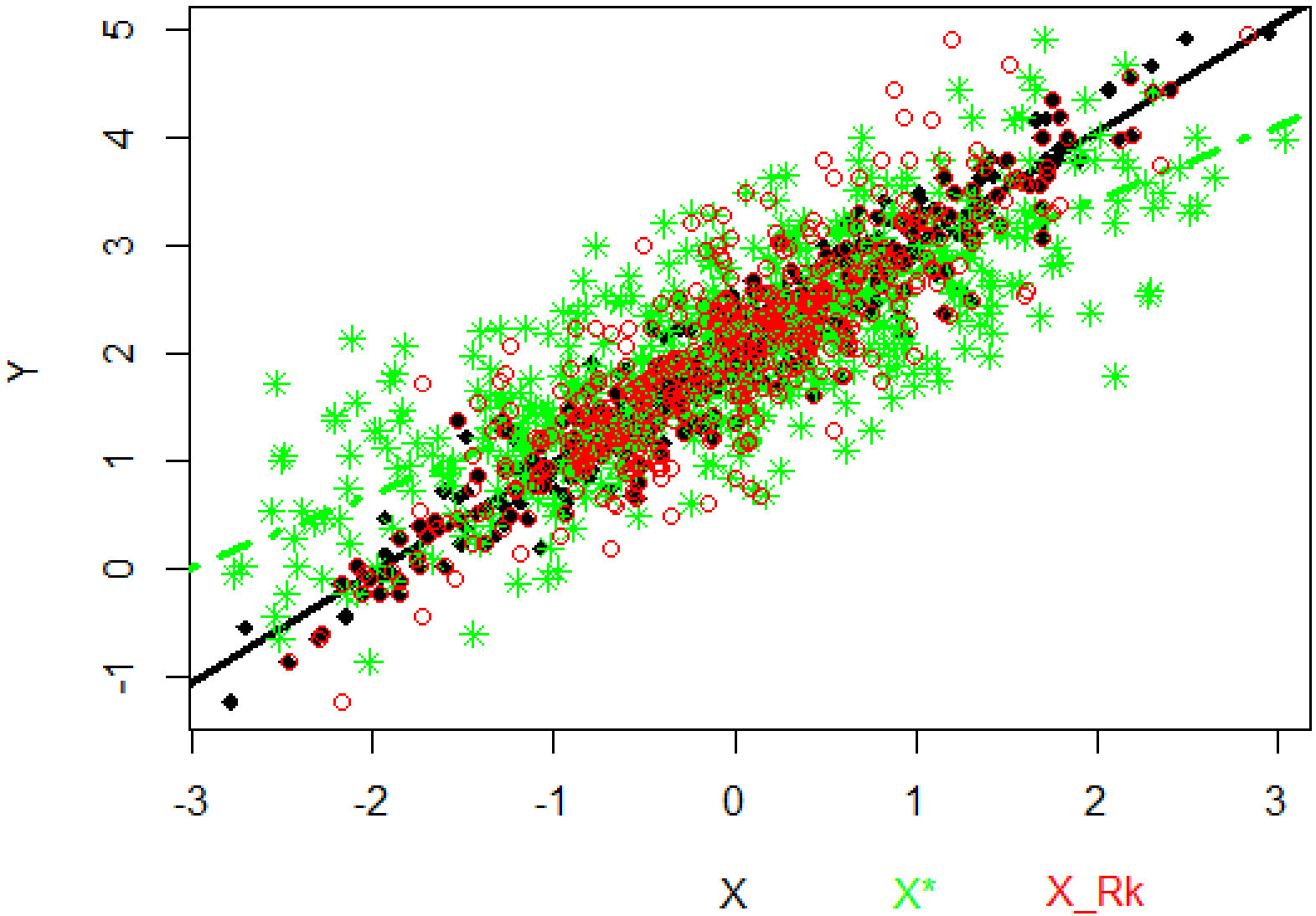
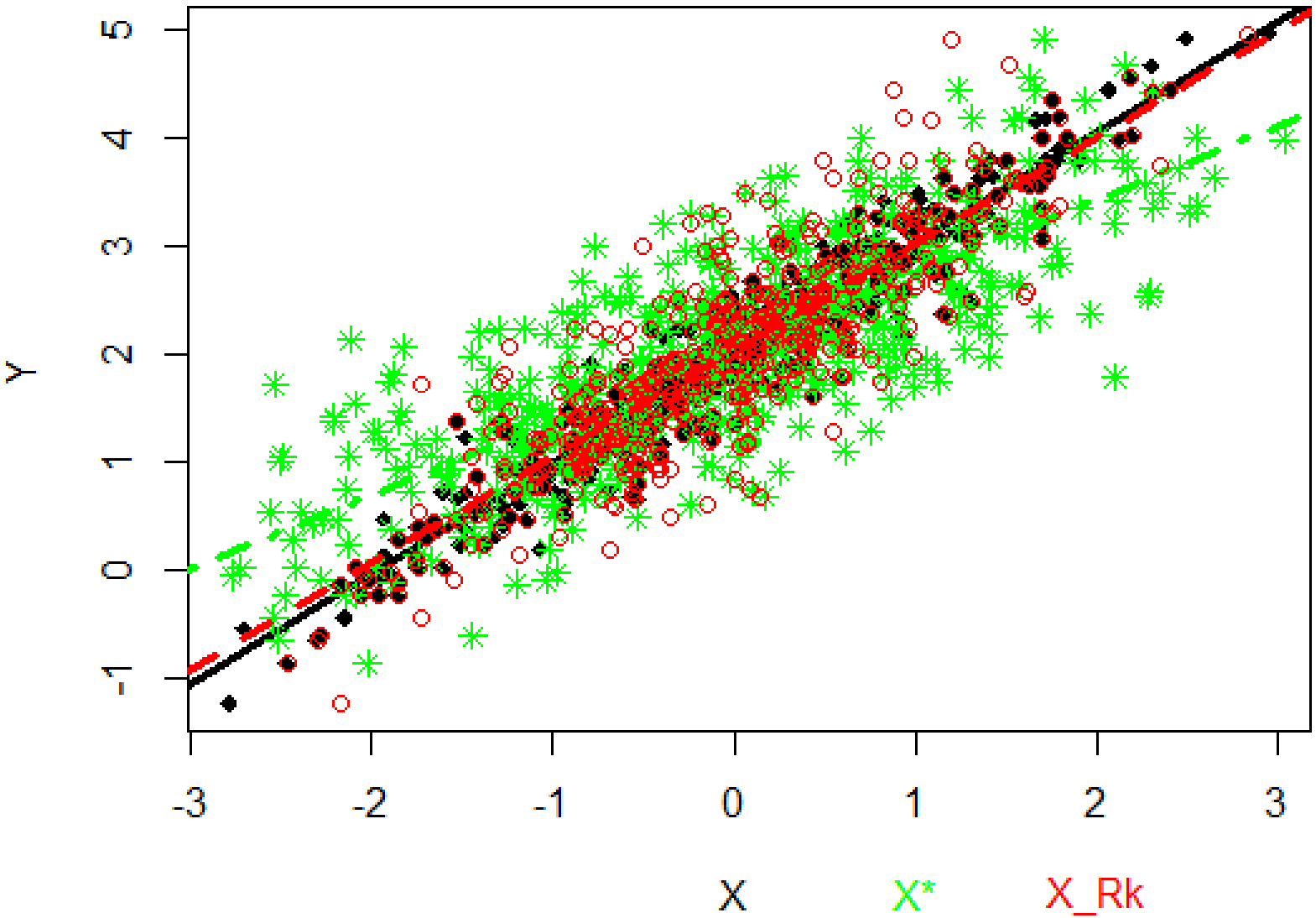

Y
5
4
3
2
1
0
-1
-3
-2
-1
0
1
2
3
X


y
5
4
3
2
1
0
-1
-3
-2
-1
0
1
2
3
X
X*


y
5
4
3
2
1
0
-1
-3
-2
-1
0
1
2
3
X
X*


y
5
4
3
2
1
0
-1
-3
-2
-1
0
1
2
3
X
X*
X_Rk


y
5
4
3
2
1
0
-1
-3
-2
-1
0
1
2
3
X
X*
X_Rk

Regressionskalibrierung

Statistische Herausforderungen im Umgang mit fehlenden bzw. fehlerbehafteten Daten

Le Minh-Anh

Ludwig Maximilians-Universität
Institut für Statistik
Bachelor-Seminar
Betreuer: Prof. Dr. Thomas Augustin

5. Dezember, 2014

Gliederung

Gliederung

Problemstellung

fehlerbehaftete Daten

Szenario:
Man interessiert sich für den Einfluss von $\boldsymbol{X}, \boldsymbol{Z}$ auf $\boldsymbol{Y}$.
Problem: $\boldsymbol{X_i}$ wird fehlerhaft gemessen $\rightarrow \boldsymbol{X_i^*} = \boldsymbol{X_i} + \boldsymbol{U_i}$
Eine naive Regression von $\boldsymbol{Y}$ auf $(\boldsymbol{X}^*, \boldsymbol{Z})$ führt zu verzerrte Inferenzen.

$\Rightarrow$ Anwendung von Messfehlerkorrektur-Verfahren

Für $\boldsymbol{U_i}$

- [klassischer] Fehler
- nicht- differentieller Fehler

ist die *Regressionskalibierung* anwendbar.

Gliederung

Regressionskalibrierung-Algorithmus

Von Interesse ist $E[Y|X,Z] = m_y(X,Z,\beta)$, aber wahre X nicht vorhanden

- Schritt 1: $E[X|X^*,Z] = m_x(X^*,Z,\gamma)$
- Schritt 2: $E[Y|Z,X^*] \approx m_y(m_x(X^*,Z,\widehat{\gamma}),Z,\beta_{RK})$
  mit wahre $\beta \approx \beta_{RK}$
- Schritt 3: Schätze Standardabweichung von $\widehat{\beta}_{RK}$ durch Bootstrapping (oder andere Methoden)

Daten

Schritt 1 & 2

Möglichkeiten für Parameterschätzung in Schritt 1 ist abhängig von der vorliegenden Datenstruktur.

- interne Validierungsdaten
- Wiederholungsdaten
- Instrumentaldaten

Gliederung

Validierungsdaten

Schritt 1 & 2

Für einen Teil der Daten liegen wahre X-Werte vor

i	Y_i	X_i	X_i^*	Z_i
1	Y_1	X_1	X_1^*	Z_1
.	.	.	.	.
k	Y_k	X_k	X_k^*	Z_k
k+1	Y_{k+1}	NA	X_{k+1}^*	Z_{k+1}
.	.	.	.	.
n	Y_n	NA	X_n^*	Z_n

z.B. Untersuchung auf Prostatakrebs durch Stanzbiopsie ist teuer und schmerzhaft. Alternativ kann man eine einfache Blutprobe entnehmen, die aber weniger genau ist.

Validierungsdaten

Schritt 1 & 2

Für einen Teil der Daten liegen wahre X-Werte vor

i	Y_i	X_i	X_i^*	Z_i	$\widehat{X}_i$
1	Y_1	X_1	X_1^*	Z_1	NA
.	.	.	.	.	.
k	Y_k	X_k	X_k^*	Z_k	NA
k+1	Y_{k+1}	NA	X_{k+1}^*	Z_{k+1}	$\widehat{X}_{k+1}$
.	.	.	.	.	.
n	Y_n	NA	X_n^*	Z_n	$\widehat{X}_n$

- Schritt 1: Rechne für $i \in \{1,..k\}$ eine Regression X $\sim X^*$+Z

  $\rightarrow E[X|X^*, Z] = m_x(X^*, Z, \gamma)$

  Berechne $\widehat{X}_i = (\underline{1}, \underline{X_i^*}, \underline{Z_i})\ \widehat{\gamma}$ für $i \in \{k+1,...,n\}$

Validierungsdaten

Schritt 1 & 2

Für einen Teil der Daten liegen wahre X-Werte vor

i	Y_i	X_i	X_i^*	Z_i	$\widehat{X}_i$	X_{reg_i}	V_i
1	Y_1	X_1	X_1^*	Z_1	NA	X_1	1
.	.	.	.	.	.	.	.
k	Y_k	X_k	X_k^*	Z_k	NA	X_k	1
k+1	Y_{k+1}	NA	X_{k+1}^*	Z_{k+1}	$\widehat{X}_{k+1}$	$\widehat{X}_{k+1}$	0
.	.	.	.	.	.	.	.
n	Y_n	NA	X_n^*	Z_n	$\widehat{X}_n$	$\widehat{X}_n$	0

- Schritt 1: Rechne für $i \in \{1,..k\}$ eine Regression X $\sim X^*$+Z

  $\rightarrow E[X|X^*, Z] = m_x(X^*, Z, \gamma)$

  Berechne $\widehat{X}_i = (\underline{1}, \underline{X_i^*}, \underline{Z_i})\ \widehat{\gamma}$ für $i \in \{k+1,...,n\}$

Validierungsdaten

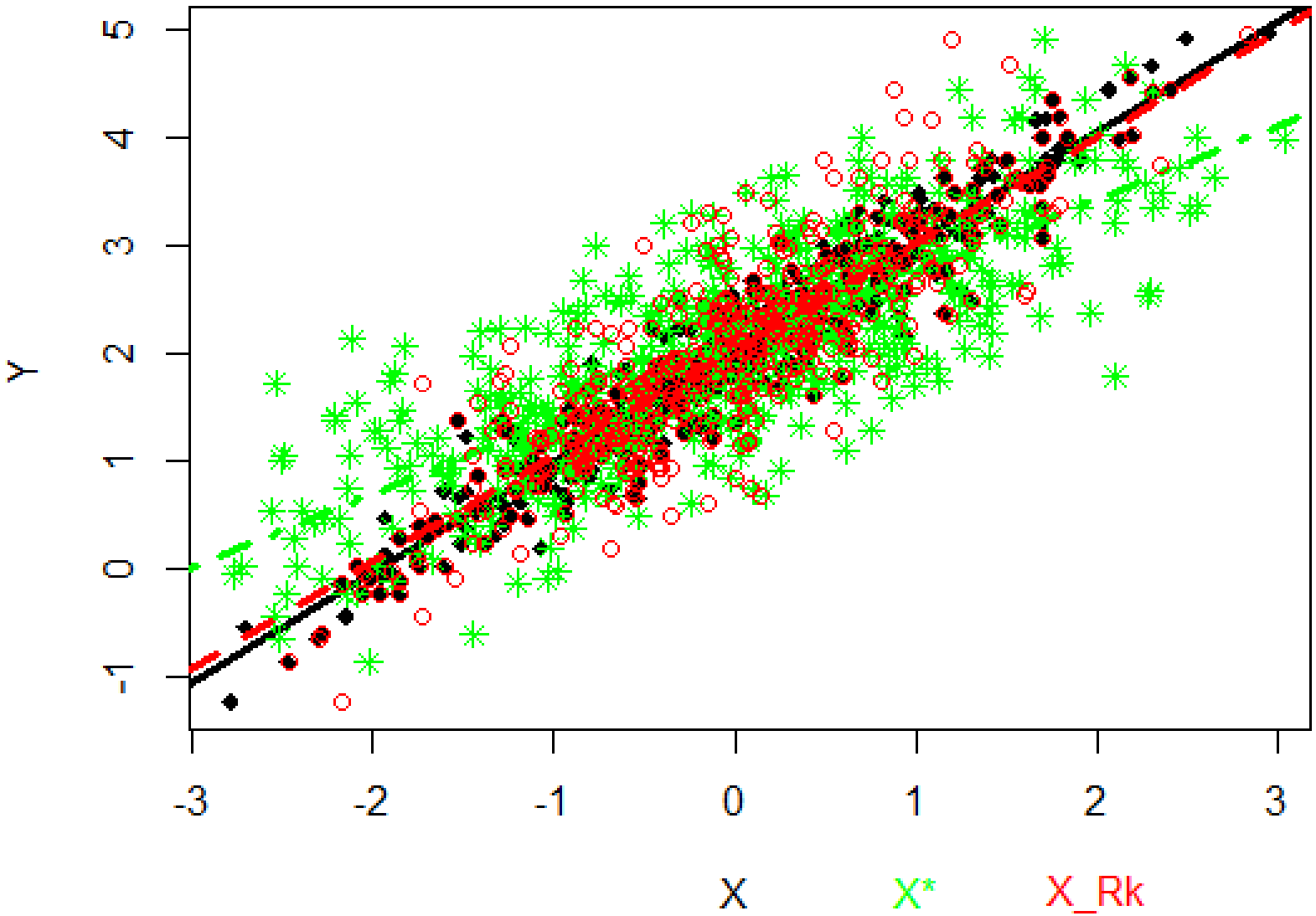
Schritt 1 & 2

Für einen Teil der Daten liegen wahre X-Werte vor

i	Y_i	X_i	X_i^*	Z_i	$\widehat{X}_i$	X_{reg_i}	V_i
1	Y_1	X_1	X_1^*	Z_1	NA	X_1	1
.	.	.	.	.	.	.	.
k	Y_k	X_k	X_k^*	Z_k	NA	X_k	1
k+1	Y_{k+1}	NA	X_{k+1}^*	Z_{k+1}	$\widehat{X}_{k+1}$	$\widehat{X}_{k+1}$	0
.	.	.	.	.	.	.	.
n	Y_n	NA	X_n^*	Z_n	$\widehat{X}_n$	$\widehat{X}_n$	0

- Schritt 2: Rechne für alle i $Y \sim X_{reg} + Z + V$

  $\rightarrow E[Y|X_{reg}, Z, V] = m_y(X_{reg}, Z, V, \beta_{RK})$


y
5
4
3
2
1
0
-1
-3
-2
-1
0
1
2
3
X
X*
X_Rk

Gliederung

Wiederholungsdaten

Schritt 1 & 2

Es liegen keine wahren Werte vor, aber Messwiederholungen (hier: $k_i = 4$)
Weitere Annahmen:
eine Einflussgröße X und eine Zielgröße Y, $E(\boldsymbol{Y}|\boldsymbol{X}) = \beta_0 + \beta_1 \boldsymbol{X}$

i	Y_i	$X^*_{i\mathbf{1}}$	$X^*_{i\mathbf{2}}$	$X^*_{i\mathbf{3}}$	$X^*_{i\mathbf{4}}$
1	$Y_{\mathbf{1}}$	$X^*_{\mathbf{11}}$	$X^*_{\mathbf{12}}$	$X^*_{\mathbf{13}}$	$X^*_{\mathbf{14}}$
.	.	.	.	.	.
k	Y_k	$X^*_{k\mathbf{1}}$	$X^*_{k\mathbf{2}}$	$X^*_{k\mathbf{3}}$	$X^*_{k\mathbf{4}}$
.	.	.	.	.	.
n	Y_n	$X^*_{n\mathbf{1}}$	$X^*_{n\mathbf{2}}$	$X^*_{n\mathbf{3}}$	$X^*_{n\mathbf{4}}$

z.B. Wiederholte Messung von Angst anhand von Herzfrequenz, wahrer Angstwert nicht beobachtbar.

Wiederholungsdaten

Schritt 1 & 2

Es liegen keine wahren Werte vor, aber Messwiederholungen (hier: $k_i = 4$)
Weitere Annahmen:
eine Einflussgröße X und eine Zielgröße Y, $E(\boldsymbol{Y}|\boldsymbol{X}) = \beta_0 + \beta_1 \boldsymbol{X}$

i	Y_i	$X^*_{i\mathbf{1}}$	$X^*_{i\mathbf{2}}$	$X^*_{i\mathbf{3}}$	$X^*_{i\mathbf{4}}$	$\overline{X}^*_{i\cdot}$
1	Y_1	$X^*_{1\mathbf{1}}$	$X^*_{1\mathbf{2}}$	$X^*_{1\mathbf{3}}$	$X^*_{1\mathbf{4}}$	$\overline{X}^*_{1\cdot}$
.	.	.	.	.	.	.
k	Y_k	$X^*_{k\mathbf{1}}$	$X^*_{k\mathbf{2}}$	$X^*_{k\mathbf{3}}$	$X^*_{k\mathbf{4}}$	$\overline{X}^*_{k\cdot}$
.	.	.	.	.	.	.
n	Y_n	$X^*_{n\mathbf{1}}$	$X^*_{n\mathbf{2}}$	$X^*_{n\mathbf{3}}$	$X^*_{n\mathbf{4}}$	$\overline{X}^*_{n\cdot}$

- Schritt 0: $\overline{X}^*_{i\cdot} = \frac{X^*_{i1}+X^*_{i2}+X^*_{i3}+X^*_{i4}}{4}$

Wiederholungsdaten

Schritt 1 & 2

Es liegen keine wahren Werte vor, aber Messwiederholungen (hier: $k_i = 4$)
Weitere Annahmen:
eine Einflussgröße X und eine Zielgröße Y, $E(\boldsymbol{Y}|\boldsymbol{X}) = \beta_0 + \beta_1 \boldsymbol{X}$

i	Y_i	$X^*_{i\mathbf{1}}$	$X^*_{i\mathbf{2}}$	$X^*_{i\mathbf{3}}$	$X^*_{i\mathbf{4}}$	$\overline{X}^*_{i\cdot}$	$\widehat{X}_i$
1	$Y_{\mathbf{1}}$	$X^*_{\mathbf{11}}$	$X^*_{\mathbf{12}}$	$X^*_{\mathbf{13}}$	$X^*_{\mathbf{14}}$	$\overline{X}^*_{\mathbf{1}\cdot}$	$\widehat{X}_{\mathbf{1}}$
.	.	.	.	.	.	.	.
k	Y_k	$X^*_{k\mathbf{1}}$	$X^*_{k\mathbf{2}}$	$X^*_{k\mathbf{3}}$	$X^*_{k\mathbf{4}}$	$\overline{X}^*_{k\cdot}$	$\widehat{X}_k$
.	.	.	.	.	.	.	.
n	Y_n	$X^*_{n\mathbf{1}}$	$X^*_{n\mathbf{2}}$	$X^*_{n\mathbf{3}}$	$X^*_{n\mathbf{4}}$	$\overline{X}^*_{n\cdot}$	$\widehat{X}_n$

- Schritt 1: Berechne

$$\widehat{E[\boldsymbol{X_i}|\overline{\boldsymbol{X_i}}^*]} \approx \frac{4\widehat{\sigma}_x^2}{4\widehat{\sigma}_x^2+\widehat{\sigma}_u^2}\overline{\boldsymbol{X_i}}^* + \hat{\mu}_{x^*}\left(1 - \frac{4\widehat{\sigma}_x^2}{4\widehat{\sigma}_x+\widehat{\sigma}_u}\right) = \widehat{\boldsymbol{X_i}}$$

Wiederholungsdaten

Schritt 1 & 2

Es liegen keine wahren Werte vor, aber Messwiederholungen (hier: $k_i = 4$)
Weitere Annahmen:
eine Einflussgröße X und eine Zielgröße Y, $E(\boldsymbol{Y}|\boldsymbol{X}) = \beta_0 + \beta_1 \boldsymbol{X}$

i	Y_i	$X^*_{i\mathbf{1}}$	$X^*_{i\mathbf{2}}$	$X^*_{i\mathbf{3}}$	$X^*_{i\mathbf{4}}$	$\overline{X}^*_{i\cdot}$	$\widehat{X}_i$
1	$Y_{\mathbf{1}}$	$X^*_{\mathbf{11}}$	$X^*_{\mathbf{12}}$	$X^*_{\mathbf{13}}$	$X^*_{\mathbf{14}}$	$\overline{X}^*_{\mathbf{1}\cdot}$	$\widehat{X}_{\mathbf{1}}$
.	.	.	.	.	.	.	.
k	Y_k	$X^*_{k\mathbf{1}}$	$X^*_{k\mathbf{2}}$	$X^*_{k\mathbf{3}}$	$X^*_{k\mathbf{4}}$	$\overline{X}^*_{k\cdot}$	$\widehat{X}_k$
.	.	.	.	.	.	.	.
n	Y_n	$X^*_{n\mathbf{1}}$	$X^*_{n\mathbf{2}}$	$X^*_{n\mathbf{3}}$	$X^*_{n\mathbf{4}}$	$\overline{X}^*_{n\cdot}$	$\widehat{X}_n$

- Schritt 1: Berechne

$$\widehat{E[\boldsymbol{X_i}|\overline{\boldsymbol{X_i}}^*]} \approx \frac{4\widehat{\sigma}_x^2}{4\widehat{\sigma}_x^2+\widehat{\sigma}_u^2}\overline{\boldsymbol{X_i}}^* + \hat{\mu}_{x^*}\left(1 - \frac{4\widehat{\sigma}_x^2}{4\widehat{\sigma}_x+\widehat{\sigma}_u}\right) = \widehat{\boldsymbol{X_i}}$$

$$\widehat{\Sigma}_{uu} = \widehat{\sigma}_u^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{4}(\boldsymbol{X}^*_{ij}-\overline{\boldsymbol{X}}^*_{i\cdot})(\boldsymbol{X}^*_{ij}-\overline{\boldsymbol{X}}^*_{i\cdot})^t}{3n}$$

Wiederholungsdaten

Schritt 1 & 2

Es liegen keine wahren Werte vor, aber Messwiederholungen (hier: $k_i = 4$)
Weitere Annahmen:
eine Einflussgröße X und eine Zielgröße Y, $E(\boldsymbol{Y}|\boldsymbol{X}) = \beta_0 + \beta_1 \boldsymbol{X}$

i	Y_i	$X^*_{i\mathbf{1}}$	$X^*_{i\mathbf{2}}$	$X^*_{i\mathbf{3}}$	$X^*_{i\mathbf{4}}$	$\overline{X}^*_{i\cdot}$	$\widehat{X}_i$
1	$Y_{\mathbf{1}}$	$X^*_{\mathbf{11}}$	$X^*_{\mathbf{12}}$	$X^*_{\mathbf{13}}$	$X^*_{\mathbf{14}}$	$\overline{X}^*_{\mathbf{1}\cdot}$	$\widehat{X}_{\mathbf{1}}$
.	.	.	.	.	.	.	.
k	Y_k	$X^*_{k\mathbf{1}}$	$X^*_{k\mathbf{2}}$	$X^*_{k\mathbf{3}}$	$X^*_{k\mathbf{4}}$	$\overline{X}^*_{k\cdot}$	$\widehat{X}_k$
.	.	.	.	.	.	.	.
n	Y_n	$X^*_{n\mathbf{1}}$	$X^*_{n\mathbf{2}}$	$X^*_{n\mathbf{3}}$	$X^*_{n\mathbf{4}}$	$\overline{X}^*_{n\cdot}$	$\widehat{X}_n$

- Schritt 1: Berechne

$$\widehat{\mu}_X = \widehat{\mu}_{X^*} = \frac{\sum_{i=1}^{n} \overline{\boldsymbol{X}^*_{i\cdot}}}{n}$$

Wiederholungsdaten

Schritt 1 & 2

Es liegen keine wahren Werte vor, aber Messwiederholungen (hier: $k_i = 4$)
Weitere Annahmen:
eine Einflussgröße X und eine Zielgröße Y, $E(\boldsymbol{Y}|\boldsymbol{X}) = \beta_0 + \beta_1 \boldsymbol{X}$

i	Y_i	$X^*_{i\mathbf{1}}$	$X^*_{i\mathbf{2}}$	$X^*_{i\mathbf{3}}$	$X^*_{i\mathbf{4}}$	$\overline{X}^*_{i\cdot}$	$\widehat{X}_i$
1	$Y_{\mathbf{1}}$	$X^*_{\mathbf{11}}$	$X^*_{\mathbf{12}}$	$X^*_{\mathbf{13}}$	$X^*_{\mathbf{14}}$	$\overline{X}^*_{\mathbf{1}\cdot}$	$\widehat{X}_{\mathbf{1}}$
.	.	.	.	.	.	.	.
k	Y_k	$X^*_{k\mathbf{1}}$	$X^*_{k\mathbf{2}}$	$X^*_{k\mathbf{3}}$	$X^*_{k\mathbf{4}}$	$\overline{X}^*_{k\cdot}$	$\widehat{X}_k$
.	.	.	.	.	.	.	.
n	Y_n	$X^*_{n\mathbf{1}}$	$X^*_{n\mathbf{2}}$	$X^*_{n\mathbf{3}}$	$X^*_{n\mathbf{4}}$	$\overline{X}^*_{n\cdot}$	$\widehat{X}_n$

- Schritt 1: Berechne

$$\widehat{\mu}_X = \widehat{\mu}_{X^*} = \frac{\sum_{i=1}^n \overline{\boldsymbol{X}^*_{i\cdot}}}{n}$$

$$\widehat{\Sigma}_{xx} = \widehat{\sigma}^2_x = \frac{\sum_{i=1}^n (\overline{\boldsymbol{X}^*_{i\cdot}} - \widehat{\mu}_{x^*})(\overline{\boldsymbol{X}^*_{i\cdot}} - \widehat{\mu}_{x^*})^t}{(n-1)} - \frac{\widehat{\Sigma}_{uu}}{4}$$

Wiederholungsdaten

Schritt 1 & 2

- Wiederholungsmessungen ermöglichen die Schätzung von Σ_{uu}
- auch ohne einen einzigen wahren Wert von X kann X geschätzt werden

$$\widehat{E[\boldsymbol{X_i}|\overline{\boldsymbol{X_i}}^*]} \approx \underbrace{\frac{4\widehat{\sigma}_x^2}{4\widehat{\sigma}_x^2 + \widehat{\sigma}_u^2}}_{\widehat{\gamma}_1} \overline{\boldsymbol{X_i}}^* + \underbrace{\hat{\mu}_{x^*}(1 - \frac{4\widehat{\sigma}_x^2}{4\widehat{\sigma}_x + \widehat{\sigma}_u})}_{\widehat{\gamma}_0} = \widehat{\boldsymbol{X_i}}$$

Approximations-Schritt

- Schritt 2:
  Ersetze die nicht beobachtete Variable $\mathbf{X}$ durch die im vorherigen Schritt durchgeführte Schätzung, d.h. ersetze im Hauptmodell $\mathbf{X}$ durch $m_{\boldsymbol{x}}(\boldsymbol{X}^*, \widehat{\gamma})$. Führe anschließend eine Standardanalyse durch, um die Parameterschätzer zu erhalten. Somit erhält man:

$$E[\boldsymbol{Y}|\overline{\boldsymbol{X}}^*] \approx m_{\boldsymbol{y}}(\underbrace{\boldsymbol{m_x}(\overline{\boldsymbol{X}}^*, \widehat{\gamma})}_{\widehat{\boldsymbol{X}}}, \boldsymbol{\beta}_{RK})$$

Approximations-Schritt

- Schritt 2:

$$
\begin{aligned}
E(Y|\overline{X}^*) &= E(\ \{E(Y|X,\overline{X}^*)\}\ |\overline{X}^*) \\
&= E(\ \{E(Y|X)\}\ |\overline{X}^*) \\
&= E(\ \{\beta_0 + \beta_1 X\}\ |\overline{X}^*) \\
&= \beta_0 + \beta_1 E(X|\overline{X}^*)
\end{aligned}
$$

Approximations-Schritt

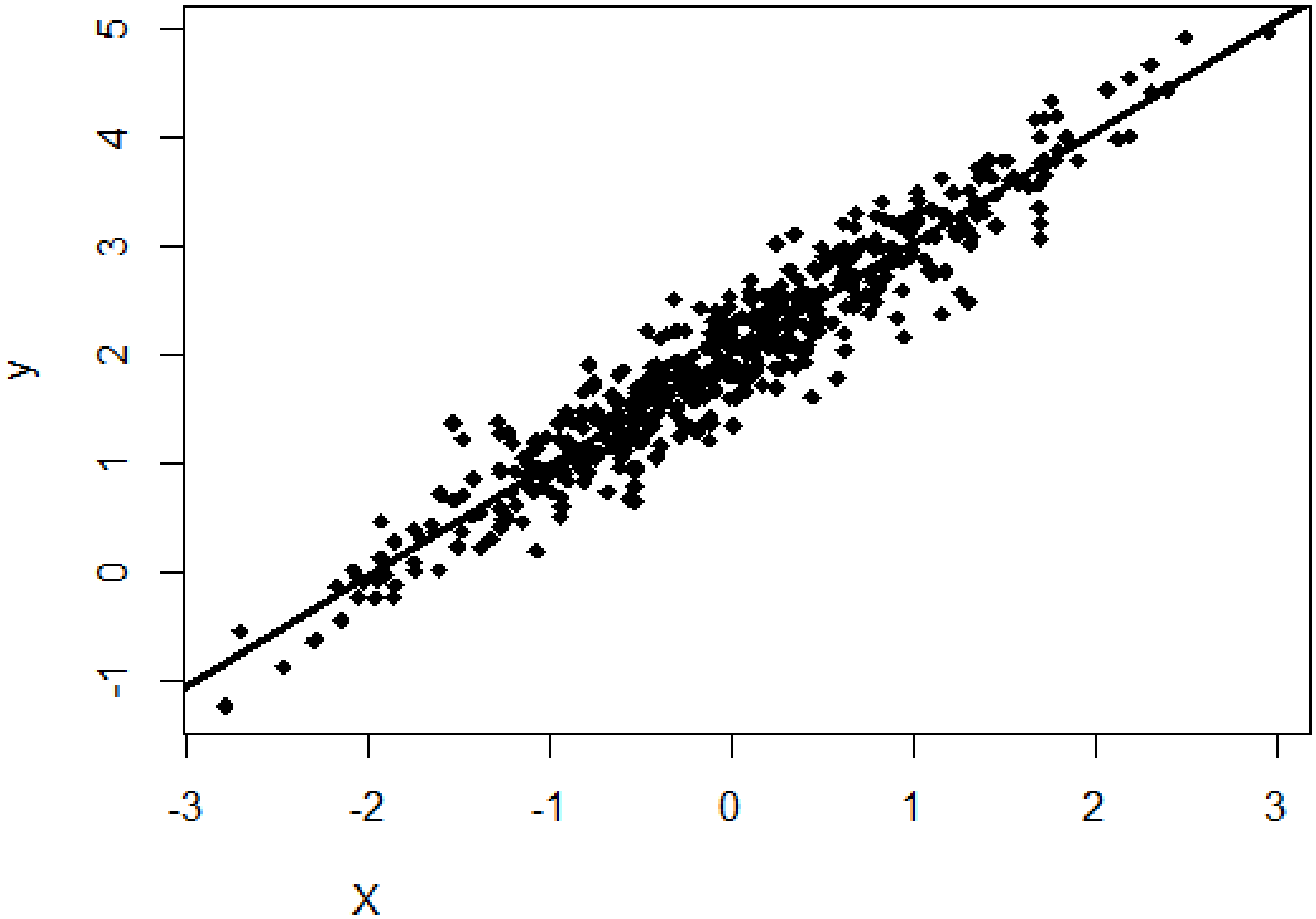
- Schritt 2:

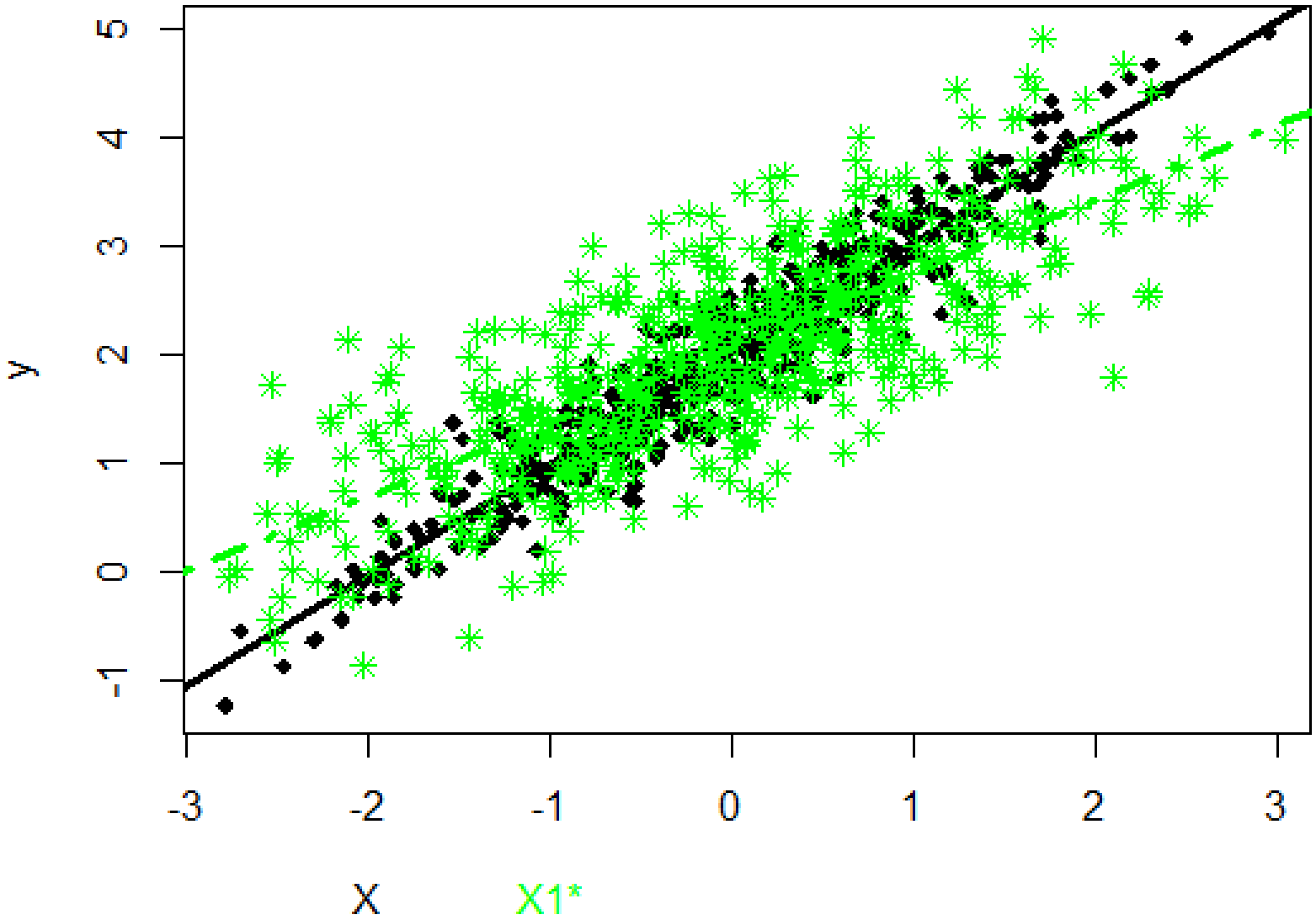
$$
\begin{aligned}
E(Y|\overline{X}^*) &= E(\ \{E(Y|X,\overline{X}^*)\}\ |\overline{X}^*) \\
&= E(\ \{E(Y|X)\}\ |\overline{X}^*) \\
&= E(\ \{\beta_0 + \beta_1 X\}\ |\overline{X}^*) \\
&= \beta_0 + \beta_1 E(X|\overline{X}^*) \\
&\approx \beta_{RK_0} + \beta_{RK_1}[\ \underbrace{(\frac{4\widehat{\sigma}_x^2}{4\widehat{\sigma}_x + \widehat{\sigma}_u})}_{\widehat{\gamma}_1}\overline{\boldsymbol{X}}^* + \hat{\mu}_{x^*}\underbrace{(1 - \frac{4\widehat{\sigma}_x^2}{4\widehat{\sigma}_x + \widehat{\sigma}_u})}_{\widehat{\gamma}_0}] \\
&\approx \underbrace{\beta_{RK_0}\beta_{RK_1}\widehat{\gamma}_0}_{\beta_{naiv_0}} + \underbrace{\beta_{RK_1}\widehat{\gamma}_1}_{\beta_{naiv_1}}\overline{\boldsymbol{X}}^*
\end{aligned}
$$

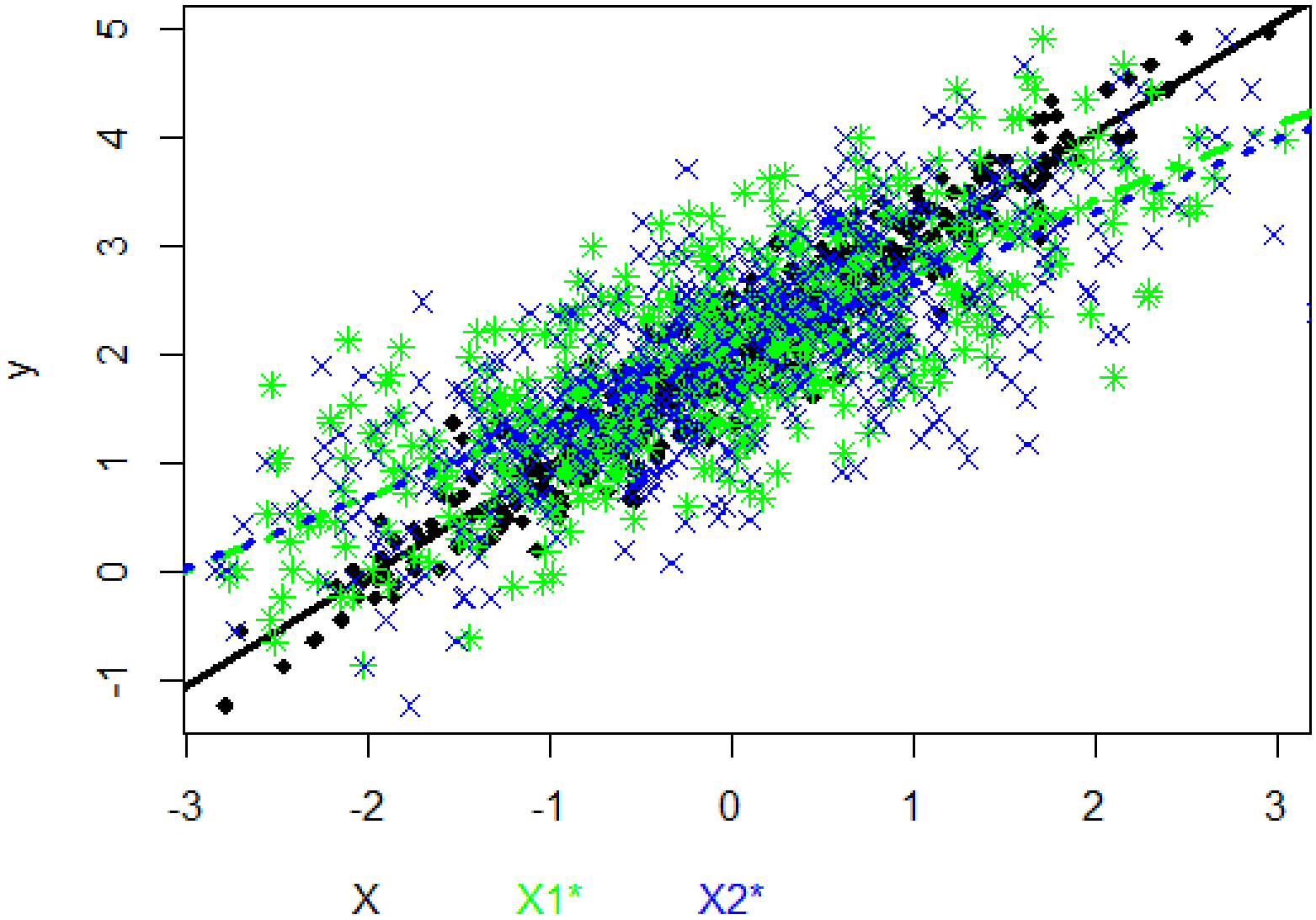
Approximations-Schritt

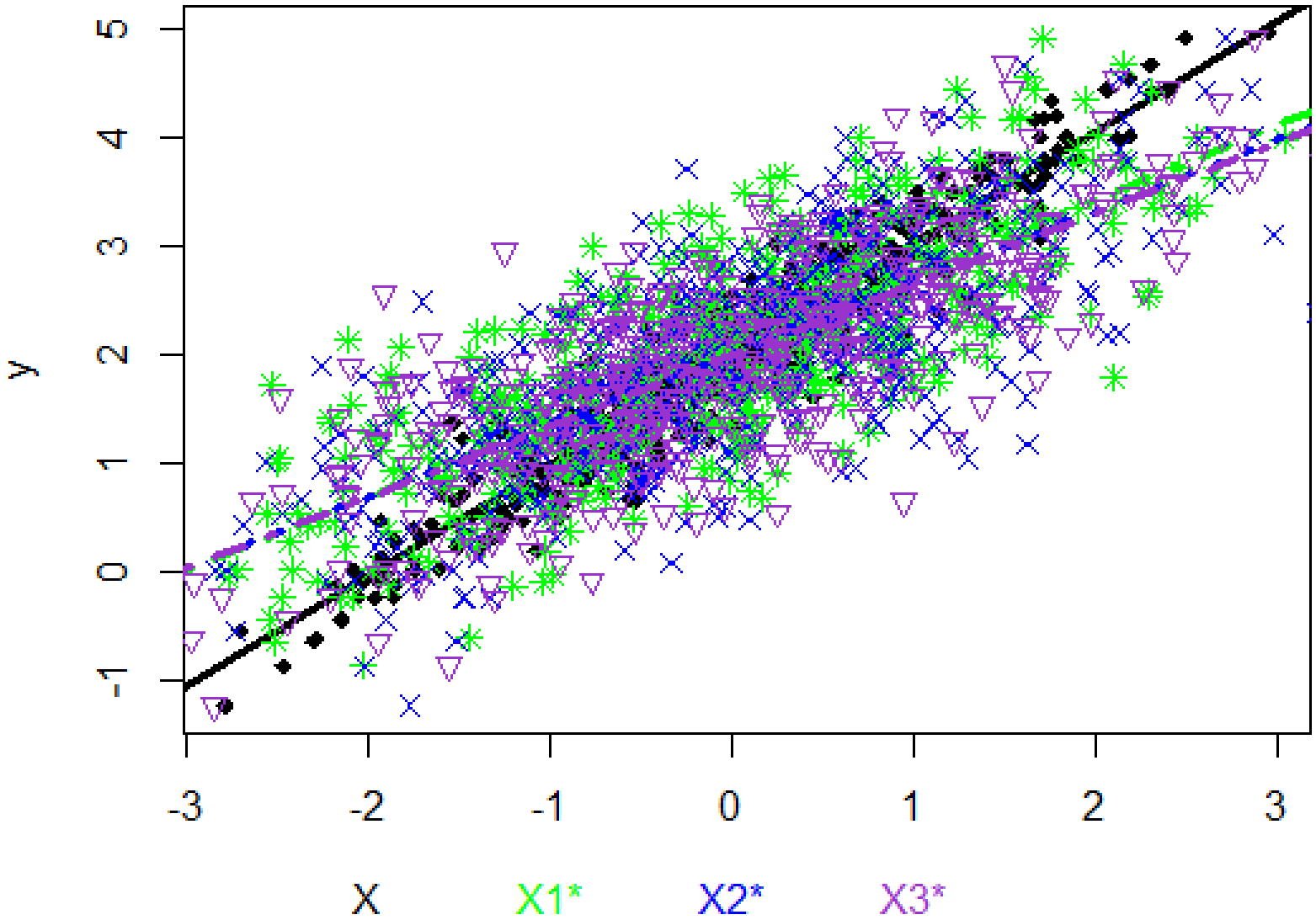
Die Schätzer für β_0 & β_1 können extrahiert werden, vorausgesetzt die Schätzung von $\boldsymbol{X}$ in Schritt 1 der RK ist gültig.

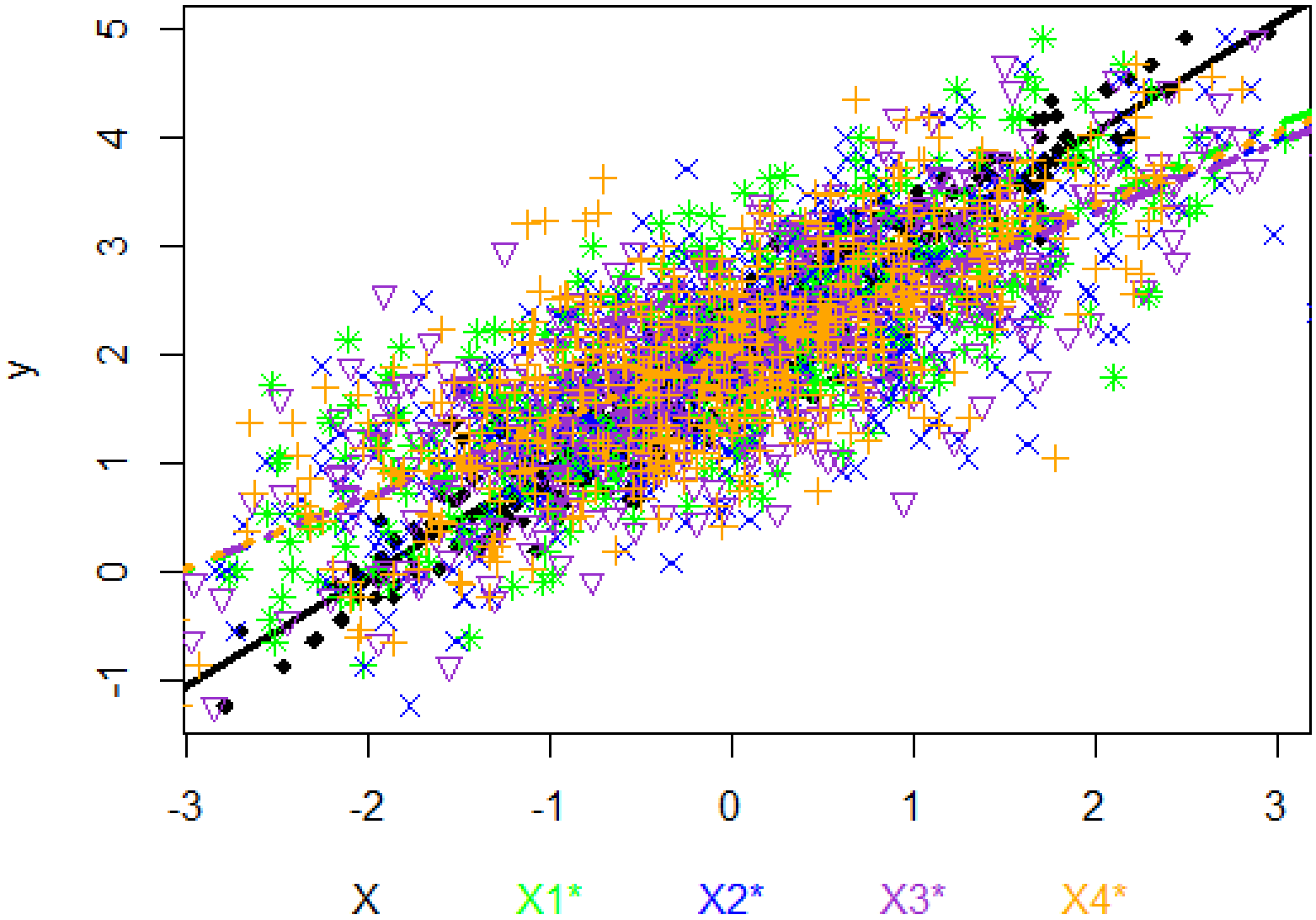
$$\widehat{\beta}_1 \approx \widehat{\beta}_{RK_1} = \frac{\widehat{\beta}_{naiv_1}}{\widehat{\gamma}_1}, \qquad \widehat{\beta}_0 \approx \widehat{\beta}_{RK_0} = \frac{\widehat{\beta}_{naiv_0}}{\widehat{\gamma}_0 \widehat{\beta}_1},$$

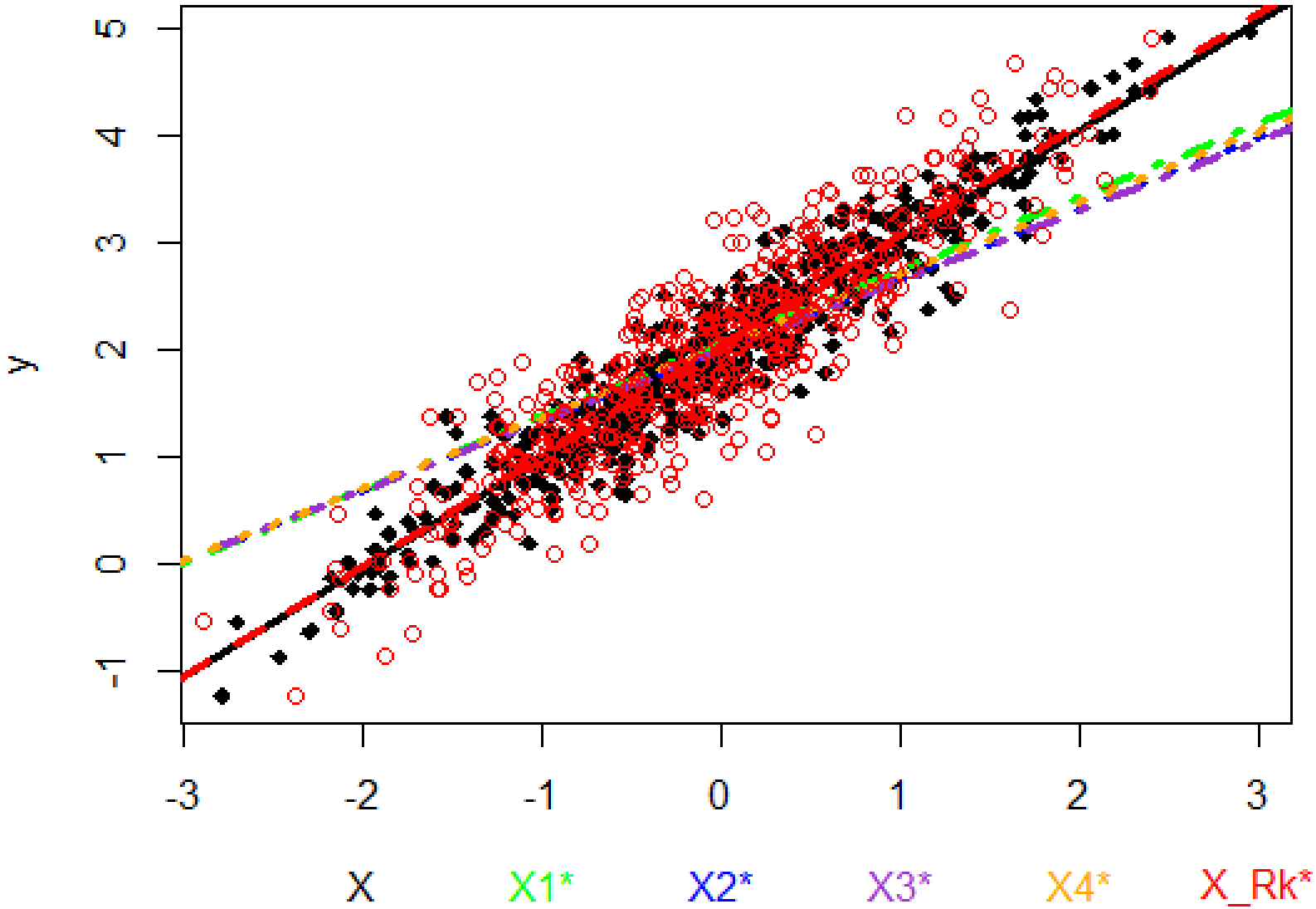

y
5
4
3
2
1
0
-1
-3
-2
-1
0
1
2
3
X


y
5
4
3
2
1
0
-1
-3
-2
-1
0
1
2
3
X
X1*


y
5
4
3
2
1
0
-1
-3
-2
-1
0
1
2
3
X
X1*
X2*


y
5
4
3
2
1
0
-1
-3
-2
-1
0
1
2
3
X
X1*
X2*
X3*


y
5
4
3
2
1
0
-1
-3
-2
-1
0
1
2
3
X
X1*
X2*
X3*
X4*


y
5
4
3
2
1
0
-1
-3
-2
-1
0
1
2
3
X
X1*
X2*
X3*
X4*
X_Rk*

Bootstrapping

- Schritt 3: Bootstrapping
  Für die resultierende $\widehat{\beta}_{RK}$ soll im dritten Schritt der Standardfehler mit Bootstrapping (oder andere Methoden) geschätzt werden.
- Warum?
  Da für die Schätzung von $\boldsymbol{Y}$ wiederum eine Schätzung von $\boldsymbol{X}$ eingesetzt wurde.
  $\Rightarrow$ resultierende p-Werte und Standardabweichungen in statistischen Programmen nur approximativ als "ersten Eindruck "anzusehen.

Bootstrapping

Schritt 3

Bootstrapping

- parametrische Verfahren
  Ziehen aus einer angenommenen Verteilung
- nonparametrische Verfahren
  Ziehen mit zurücklegen aus den vorliegenden Daten.
  (Stichprobenumfang $\equiv$ Umfang vorliegende Daten)
  - resampling Vectors
  - resamping Residuals

Vorteil von Bootstrapping ist, dass bekannte statistische Verfahren auf Bootstrapp-Stichproben angewendet werden kann.
z.B. Regressionskalibrierung.

Bootstrapping bei Messfehlerkorrektur

Schritt 3

- Daten können in verschiedenen Formen vorliegen.
  z.B. Mischung aus interne Validierungsdaten, Instrumentaldaten, Wiederholungsdaten (2,3,4,...Messwiederholungen)
- Unterschiedliche Datenstrukturen $\rightarrow$ Unterschiedliche Informationen
- Für Bootstrapping bei Messfehlerkorrektur gilt:
  - Teildatensätze bilden (gruppieren nach den verschiedenen Strukturen)
    aus Teildatensätze Bootstrap-Stichproben ziehen
    $\rightarrow$ ermöglicht Ziehungen aus homogener Umgebung
    $\rightarrow$ geringere Varianz

Gliederung

Resampling Vectors

Validierungsdaten:

i	Y_i	X_i	X_i^*	Z_i
1	Y_1	X_1	X_1^*	Z_1
.	.	.	.	.
k	Y_k	X_k	X_k^*	Z_k
k+1	Y_{k+1}	NA	X_{k+1}^*	Z_{k+1}
.	.	.	.	.
n	Y_n	NA	X_n^*	Z_n

Vektorweise ziehen mit zurücklegen aus
$\{(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{X}_i^*, \boldsymbol{Z}_i)\}_{i=1}^k$ bzw. $\{(\boldsymbol{Y}_i, \boldsymbol{X}_i^*, \boldsymbol{Z}_i)\}_{i=k+1}^n$

- Vorteil: Kaum Annahmen müssen getroffen werden.
  besondere Beziehungen müssen nicht explizit berücksichtigt werden.
  z.B. wenn ϵ_i von Z_i abhängt
- Nachteil: Bootstrapstichprobe enthält nicht die gleiche Variablenmenge.
  z.B. High-Leverage Point mehrmals oder gar nicht enthalten.

Gliederung

Resampling Residuals- Validierungsdaten

Validierungsdaten

i	Y_i	X_i	X_i^*	Z_i
1	Y_1	X_1	X_1^*	Z_1
.	.	.	.	.
k	Y_k	X_k	X_k^*	Z_k
k+1	Y_{k+1}	NA	X_{k+1}^*	Z_{k+1}
.	.	.	.	.
n	Y_n	NA	X_n^*	Z_n

- Annahme: Zwei Regressionsmodelle liegen vor
  - $Y_i \sim (Z_i, X_i), \epsilon_i \sim$ iid mit $\Sigma_i \approx \Sigma$
  - $X_j^* \sim (Z_j, X_j), \epsilon_j \sim$ iid mit $\Sigma_j \approx \Sigma'$

Resampling Residuals- Validierungsdaten

Validierungsdaten

i	Y_i	X_i	X_i^*	Z_i	$Y_i^{(1)}$	.	$Y_i^{(M)}$
1	Y_1	X_1	X_1^*	Z_1	$Y_1^{(1)}$	.	$Y_1^{(M)}$
.	.	.	.	.	.	.	.
k	Y_k	X_k	X_k^*	Z_k	$Y_k^{(1)}$	.	$Y_k^{(M)}$
k+1	Y_{k+1}	NA	X_{k+1}^*	Z_{k+1}	.	.	.
.	.	.	.	.	.	.	.
n	Y_n	NA	X_n^*	Z_n	.	.	.

- Bootstrapping $Y_i^{(m)}$, M = # Bootstrapstichprobe und m $\in \{1,..,M\}$
  - 1. $\epsilon_i = Y_i - m_y(Z_i, X_i, \widehat{\mathcal{B}})$ für i $\in \{1,..,k\}$
  - 2. $B = \{(\epsilon_i - \bar{\epsilon})\}_i^k$
  - 3. k mal Ziehen mit zurücklegen aus B $\rightarrow \{\epsilon_i^{(m)}\}_i^k$
  - 4. $Y_i^{(m)} = m_y(Z_i, X_i, \widehat{\mathcal{B}}) + \epsilon_i^{(m)}$ für i $\in \{1,..,k\}$

Resampling Residuals- Validierungsdaten

Validierungsdaten

i	Y_i	X_i	X_i^*	Z_i	$Y_i^{(1)}$	.	$Y_i^{(M)}$	$X_i^{*(1)}$	.	$X_i^{*(M)}$
1	Y_1	X_1	X_1^*	Z_1	$Y_1^{(1)}$	.	$Y_1^{(M)}$	$X_1^{*(1)}$	.	$X_1^{*(M)}$
.	.	.	.	.	.	.	.	.	.	.
k	Y_k	X_k	X_k^*	Z_k	$Y_k^{(1)}$	.	$Y_k^{(M)}$	$X_k^{*(1)}$	.	$X_k^{*(M)}$
k+1	Y_{k+1}	NA	X_{k+1}^*	Z_{k+1}	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
n	Y_n	NA	X_n^*	Z_n	.	.	.	.	.	.

- Bootstrapping $Y_i^{(m)}$, M = # Bootstrapstichprobe und m $\in$ {1,..,M}
  - 1. $\epsilon_i = Y_i - m_y(Z_i, X_i, \widehat{\mathcal{B}})$ für i $\in$ {1,..,k}
  - 2. $B = \{(\epsilon_i - \bar{\epsilon})\}_i^k$
  - 3. k mal Ziehen mit zurücklegen aus B $\rightarrow \{\epsilon_i^{(m)}\}_i^k$
  - 4. $Y_i^{(m)} = m_y(Z_i, X_i, \widehat{\mathcal{B}}) + \epsilon_i^{(m)}$ für i $\in$ {1,..,k}
- Bootstraping $X_i^{*(m)}$, analog

Gliederung

Bootstrap- Algorithmus

- Schritt 1: M Bootstrapstichproben ziehen
- Schritt 2: Schritt 1 & 2 des Regressionskalibrierungs-Algorithmus anwenden

  Man erhält somit nach M Durchläufen die Parameter

  $$\widehat{\beta}_{RK_k}^{(1)}, .., \widehat{\beta}_{RK_k}^{(M)}$$

- Schritt 3: Aus den vorliegenden $\widehat{\beta}_{RK_k}$s kann nun die Standardabweichung $\widehat{\sigma}_{\beta_{RK_k}}$ geschätzt werden

$$\widehat{\sigma}^2_{RK_k} = \widehat{var(\widehat{\beta}_{RK_k})} = \frac{1}{M-1} \sum_{m=1}^{M} (\widehat{\beta}_{RK_k}^{(m)} - \overline{\widehat{\beta}_{RK_k}})(\widehat{\beta}_{RK_k}^{(m)} - \overline{\widehat{\beta}_{RK_k}})^t.$$

Nachteile-Regressionskalibrierung

Nachteile

- nur approximatives Verfahren
- die Berechnung der Regression von $\mathbf{X}$ auf $(\mathbf{X}^*, \mathbf{Z})$ stellt eine Herausforderung dar, da $\mathbf{X}$ nicht beobachtbar ist
- Schätzer sind nicht unbedingt konsistent (abhängig vom Modell)

Vorteile-Regressionskalibrierung

Vorteile I

- auf viele Modelle anwendbar (GLM)
- effektive Methode im Umgang mit fehlerhaften gemeinsam Einflussgrößen
- einfache Berechnung
- Regression Y $\sim$ X+Z möglich, obwohl wahres **X** nicht beobachtet
- anschließende Standardanalysen noch möglich
- Reduzierung der Bias
- keine extra Implementierung in statistische Programme nötig

Vielen Dank für eure Aufmerksamkeit.

Anhang

- *RK mit Instrumentaldaten*
- *Vergleich Valid Wdh1 Wdh4*
- *Vergleich RK Valid mit/ohne Dummyvariable*
- *Überprüfen der Schätzung in Schritt 1 der RK*

Instrumentaldaten

Schritt 1 & 2

Instrumentaldaten $\boldsymbol{T}$

- $\boldsymbol{T}$ ist abhängig von $\boldsymbol{X}$
- $\boldsymbol{T}$ ist unkorreliert mit Fehler $\boldsymbol{U} = \boldsymbol{X}^* - \boldsymbol{X}$
- $\boldsymbol{T}$ unkorreliert mit $\epsilon = \boldsymbol{Y} - E[\boldsymbol{Y}|\boldsymbol{Z}, \boldsymbol{X}]$

Außerdem soll gelten $\boldsymbol{T}$ ist *unverzerrt* für $\boldsymbol{X}$ d.h. eine Regression von T $\sim$ Z+X^* entspricht einer Regression von X $\sim$ Z+X^*

$$E[\boldsymbol{T}|\boldsymbol{X}^*, \boldsymbol{Z}] = E[\boldsymbol{X}|\boldsymbol{X}^*, \boldsymbol{Z}]$$

Instrumentaldaten

Schritt 1 & 2

Beispiel:

- $\boldsymbol{Y}$: Brustkrebs
- $\boldsymbol{X}$: langfristige durchschnittliche Aufnahme von Nährstoffen
- $\boldsymbol{T}$: durchschnittliche Aufnahme von Nährstoffen, extrahiert aus einem Ernhährungs-Tagebuch über vier Wochen (professionell dokumentiert)
- $\boldsymbol{X}^*$: durchschnittliche Aufnahme von Nährstoffen, extrahiert aus einem Fragebogen über Ernährung

Instrumentaldaten

Schritt 1 & 2

Für einen Teil der Daten liegen T-Werte vor,
Wobei T unverzerrt für X ist.

i	Y_i	T_i	X_i^*	Z_i
1	Y_1	T_1	X_1^*	Z_1
.	.	.	.	.
k	Y_k	T_k	X_k^*	Z_k
k+1	Y_{k+1}	NA	X_{k+1}^*	Z_{k+1}
.	.	.	.	.
n	Y_n	NA	X_n^*	Z_n

- Schritt 1: Rechne für $i \in \{1,..k\}$ eine Regression T $\sim X^*$+Z

  $\rightarrow E[T|X^*, Z] = m_T(X^*, Z, \gamma) = E[X|X^*, Z]$

  Berechne $\widehat{X}_i = (\underline{1}, \underline{X_i^*}, \underline{Z_i})\ \widehat{\gamma}$ für $i \in \{k+1,...,n\}$

Instrumentaldaten

Schritt 1 & 2

Für einen Teil der Daten liegen T-Werte vor,
Wobei T unverzerrt für X ist.

i	Y_i	T_i	X_i^*	Z_i	$\widehat{X}_i$
1	Y_1	T_1	X_1^*	Z_1	NA
.	.	.	.	.	.
k	Y_k	T_k	X_k^*	Z_k	NA
k+1	Y_{k+1}	NA	X_{k+1}^*	Z_{k+1}	$\widehat{X}_{k+1}$
.	.	.	.	.	.
n	Y_n	NA	X_n^*	Z_n	$\widehat{X}_n$

- Schritt 1: Rechne für $i \in \{1,..k\}$ eine Regression T $\sim X^*$+Z

  $\rightarrow E[T|X^*, Z] = m_T(X^*, Z, \gamma) = E[X|X^*, Z]$

  Berechne $\widehat{X}_i = (\underline{1}, \underline{X_i^*}, \underline{Z_i})\ \widehat{\gamma}$ für $i \in \{$k+1,...,n$\}$

Instrumentaldaten

Schritt 1 & 2

Für einen Teil der Daten liegen T-Werte vor,
Wobei T unverzerrt für X ist.

i	Y_i	T_i	X_i^*	Z_i	$\widehat{X}_i$	X_{reg_i}	VT_i
1	Y_1	T_1	X_1^*	Z_1	NA	T_1	1
.	.	.	.	.	.	.	.
k	Y_k	T_k	X_k^*	Z_k	NA	T_k	1
k+1	Y_{k+1}	NA	X_{k+1}^*	Z_{k+1}	$\widehat{X}_{k+1}$	$\widehat{X}_{k+1}$	0
.	.	.	.	.	.	.	.
n	Y_n	NA	X_n^*	Z_n	$\widehat{X}_n$	$\widehat{X}_n$	0

- Schritt 1: Rechne für $i \in \{1,..k\}$ eine Regression T $\sim X^*$+Z

  $\rightarrow E[T|X^*, Z] = m_T(X^*, Z, \gamma) = E[X|X^*, Z]$

  Berechne $\widehat{X}_i = (\underline{1}, \underline{X_i^*}, \underline{Z_i})\ \widehat{\gamma}$ für $i \in \{k+1,...,n\}$

Instrumentaldaten

Schritt 1 & 2

Für einen Teil der Daten liegen T-Werte vor,
Wobei T unverzerrt für X ist.

i	Y_i	T_i	X_i^*	Z_i	$\widehat{X}_i$	X_{reg_i}	VT_i
1	Y_1	T_1	X_1^*	Z_1	NA	T_1	1
.	.	.	.	.	.	.	.
k	Y_k	T_k	X_k^*	Z_k	NA	T_k	1
k+1	Y_{k+1}	NA	X_{k+1}^*	Z_{k+1}	$\widehat{X_{k+1}}$	$\widehat{X}_{k+1}$	0
.	.	.	.	.	.	.	.
n	Y_n	NA	X_n^*	Z_n	$\widehat{X_n}$	$\widehat{X}_n$	0

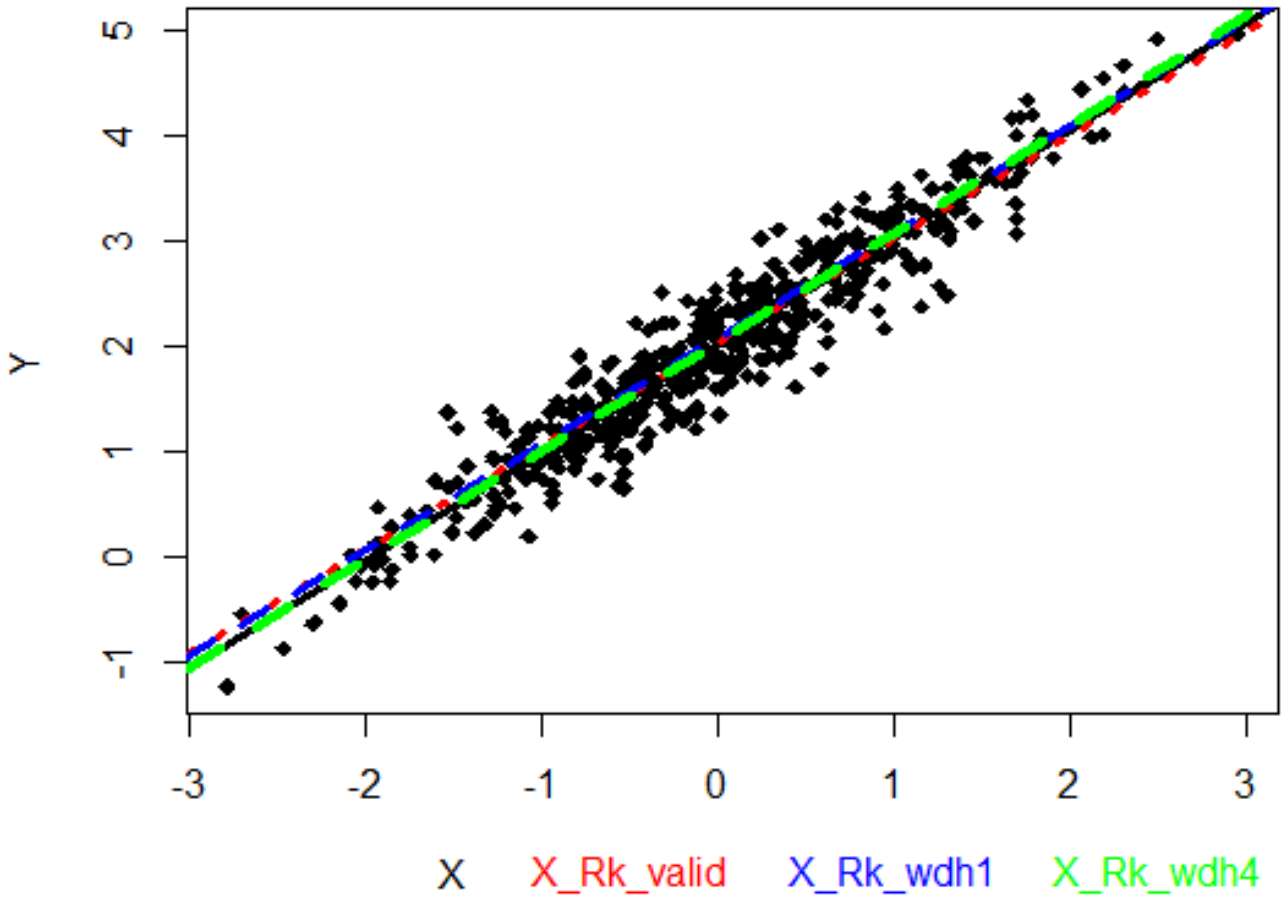
- Schritt 2: Rechne für alle i $Y \sim X_{reg} + Z + VT$

  $\rightarrow E[Y|X_{reg}, Z, VT] = m_y(X_{reg}, Z, VT, \beta_{RK})$

Anhang Übersicht

Vergleich: Valid $\sim$ Wdh1 $\sim$ Wdh4

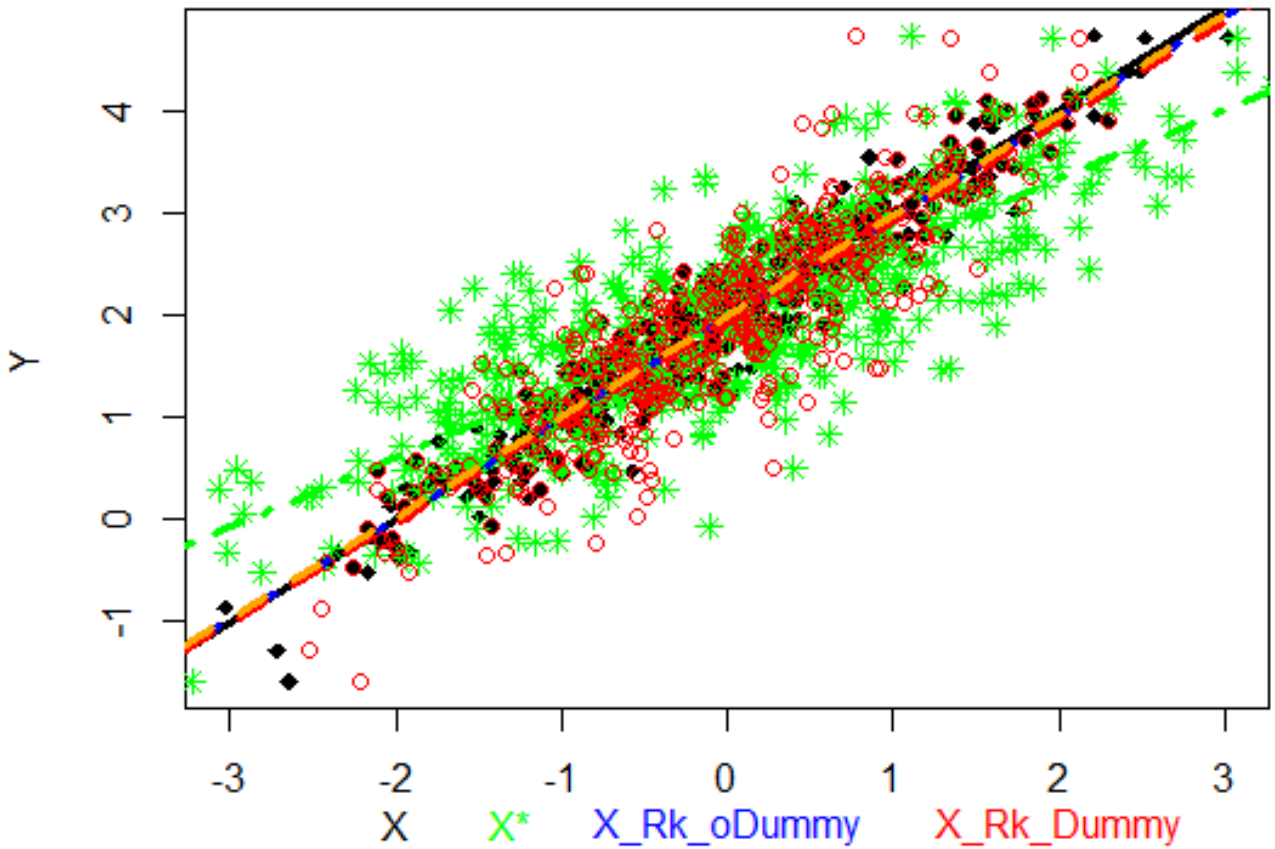
Kennzahl	Valid	Wdh1	Wdh4
MSE(β_0)	0.0009120709	1.149171367	1.0528207133
MSE(β_1)	0.0015110924	0.004843075	0.0019111721



Anhang Übersicht

Vergleich: Validierungsdaten mit/ohne Dummy

Modell	ohne Dummy $Y \sim \widehat{X}$	mit Dummy $Y \sim \widehat{X} + Valid(Ja/Nein)$
MSE(β_0)	0.0010143196	0.9939534458
MSE(β_1)	1.0529453460	0.0013049964



Anhang Übersicht

Überprüfung der Schätzung in Schritt 1

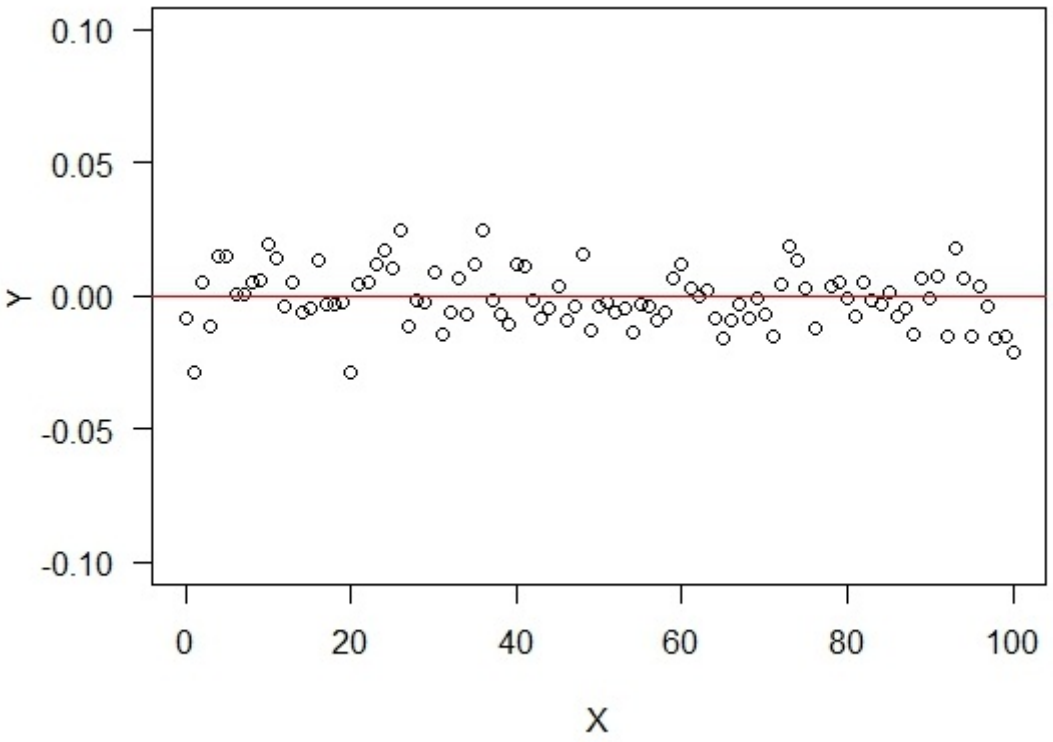
Schätzung von $\boldsymbol{X}$ in Schritt 1 der RK kann überprüft werden

- durch gewöhnliche Regressionsdiagnosen bei
  - Validierungsdaten
  - Instrumentaldaten
- mithilfe von [partielle] Wiederholungsdaten
  - Wiederholungsdaten

Überprüfung der Schätzung in Schritt 1

Validierungsdaten & Instrumentaldaten

Residuenplot



Validierungsdaten

- X $\equiv$ wahre X_i
- Y
  $\equiv \epsilon_i = X_i - E[X_i|X_i^*, Z_i]$

Instrumentaldaten

- X $\equiv$ T
- Y
  $\equiv \epsilon_i = T_i - E[T_i|X_i^*, Z_i]$

Überprüfung der Schätzung in Schritt 1

[partielle] Wiederholungsdaten

Überprüfen der Schätzung in Schritt 1 der RK anhand von [partiellen] Wiederholungsdaten

i	Y_i	$X^*_{i\mathbf{1}}$	$X^*_{i\mathbf{2}}$
1	$Y_{\mathbf{1}}$	$X^*_{\mathbf{11}}$	$X^*_{\mathbf{12}}$
.	.	.	.
k	Y_k	$X^*_{k\mathbf{1}}$	$X^*_{k\mathbf{2}}$
k+1	$Y_{k+\mathbf{1}}$	$X^*_{k+\mathbf{1}\ \mathbf{1}}$	NA
.	.	.	.
n	Y_n	$X^*_{n\mathbf{1}}$	NA

- Wie überprüft man ob das eine gute Schätzung für $\mathbf{X}$ ist ohne die wahren $\mathbf{X}$ beobachtet zu haben?

Überprüfung der Schätzung in Schritt 1

[partielle] Wiederholungsdaten

Es liegen klassische Fehler vor

$$\boldsymbol{X}_{i1}^* = \boldsymbol{X}_i + \boldsymbol{U}_{i1}$$
$$\boldsymbol{X}_{i2}^* = \boldsymbol{X}_i + \boldsymbol{U}_{i2}$$

und somit zunächst

$$\begin{aligned} E[\boldsymbol{X}_{i2}^* | \boldsymbol{Z_i}, \boldsymbol{X_{i1}^*}] &= E[\boldsymbol{X}_i + \boldsymbol{U}_{i2} | \boldsymbol{Z_i}, \boldsymbol{X_{i1}^*}] \\ &= \underbrace{E[\boldsymbol{X}_i | \boldsymbol{Z_i}, \boldsymbol{X_{i1}^*}]}_{Schritt\ 1\ der\ RK} + \underbrace{E[\boldsymbol{U}_{i2} | \boldsymbol{Z_i}, \boldsymbol{X_{i1}^*}]}_{V_i}. \end{aligned}$$

Im klassischem Fehlermodell gilt für die bedingte Zufallsvariable V_i $E[V_i]=0$.

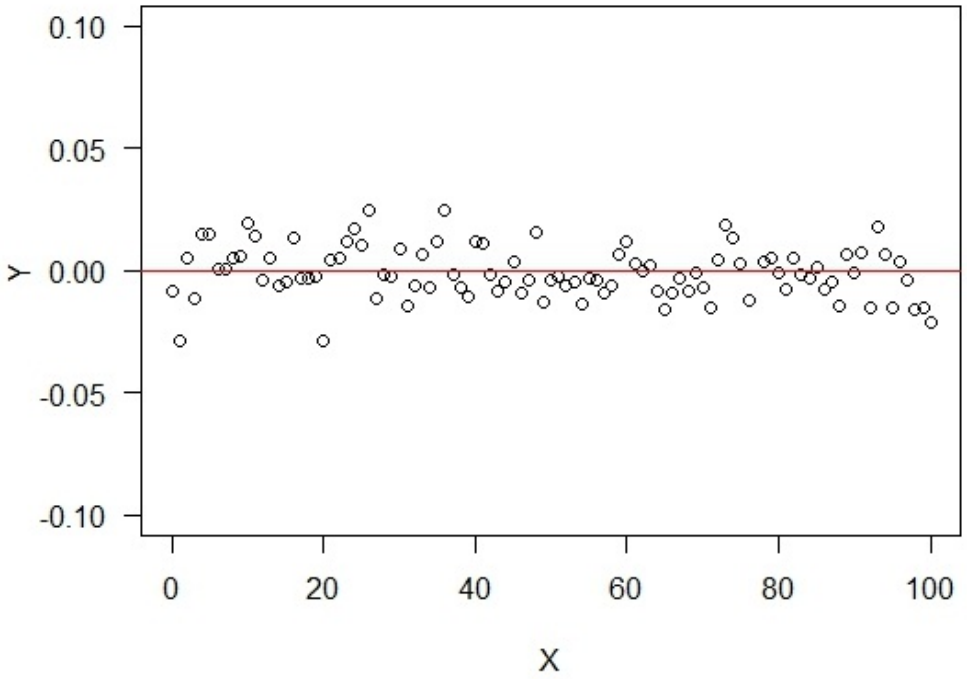
Überprüfung der Schätzung in Schritt 1

[partielle] Wiederholungsdaten

$$\epsilon_i^* = \boldsymbol{X_{i2}^*} - E[\boldsymbol{X}_{i2}^* | \boldsymbol{Z_i}, \boldsymbol{X_{i1}^*}] =$$

$$(\boldsymbol{X}_i + \boldsymbol{U}_{i2}) - (E[\boldsymbol{X}_{i2} | \boldsymbol{Z}_i, \boldsymbol{X}^*{}_{i1}] + E[\boldsymbol{U}_{i2} | \boldsymbol{Z}_i, \boldsymbol{X}_{i1}^*])$$

$$(\boldsymbol{X}_i - E[\boldsymbol{X}_i | \boldsymbol{Z}_i, \boldsymbol{X}_{i1}^*]) \; + \; (\boldsymbol{U}_{i1} - E[\boldsymbol{U}_{i2} | \boldsymbol{Z}_i, \boldsymbol{X}_{i1}^*]) = \epsilon_i + \tilde{\epsilon}_i$$



[partielle] Wiederholungsdaten

- $X \equiv X_{i2}^* = X_i + U_{i2}$
- $Y \equiv \epsilon_i^* = \epsilon_i + \tilde{\epsilon}_i$
  $\rightarrow$ Aus Residuenplot Tendenz ersichtlich ob Schätzung in Schritt 1 der RK eine gute Schätzung ist.

Anhang Übersicht