Vorlesung: Statistik I für Studierende der Statistik, Mathematik & Informatik

Dozent: Fabian Scheipl
Material: H. Küchenhoff
LMU München

Regression

Einfache lineare Regression

- Linearer Zusammenhang zwischen zwei metrischen Größen Y, X wird als Gerade visualisiert
- Finde Gerade $Y \approx \alpha + \beta \cdot X$
- β Steigung der Geraden: erhöht sich X um eine Einheit, so erhöht sich Y um $\approx \beta$ Einheiten.
- α : Achsenabschnitt, d.h. Wert von Y für $X = 0$

Bestimmung der Regressionsgerade

Welche Gerade ist die "Beste"?

- sollte etwa in der "Mitte" der Punktwolke liegen
- Abweichungen der Wertepaare (x_i, y_i) (Punkte) von der Geraden möglichst "klein" (minimal)

Methode der kleinsten Quadrate

- Y ist **Ziel**größe und X **Einfluss**größe
- $\Rightarrow$ Y soll mit Hilfe von X erklärt oder vorhergesagt werden
- Lineares Modell $Y = \alpha + \beta X + \varepsilon$
- Minimierung der Abstände **in** Y**-Richtung**
- Wähle $\hat{\alpha}$ und $\hat{\beta}$ so, dass $\sum_{i=1}^{n} \left(y_i - (\hat{\alpha} + \hat{\beta} x_i) \right)^2$ minimal wird

Lineare Einfachregression und Kleinste-Quadrate-Schätzer

Seien $(x_1, y_1), \ldots, (x_n, y_n)$ Beobachtungen der Merkmale X und Y, dann heißt

$$y_i = \alpha + \beta x_i + \varepsilon_i, \; i = 1, \ldots, n,$$

lineare Einfachregression, wobei α den **Achsenabschnitt** (*intercept*), β die **Steigung** (*slope*) und ε den **Fehler** (Residuum, *error*, *residual*) bezeichnet.

Die Kleinste-Quadrate-Schätzer für $\hat{\alpha}$ und $\hat{\beta}$ sind gegeben durch

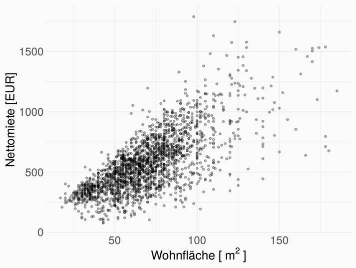
$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \qquad \hat{\beta} = \frac{S_{xy}}{S_x^2}.$$

Die Residuen berechnen sich durch

$$\varepsilon_i = y_i - \hat{y}_i, \; i = 1, \ldots, n,$$

mit $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$.

Beispiel: Nettomiete in Abhängigkeit von der Wohnfläche

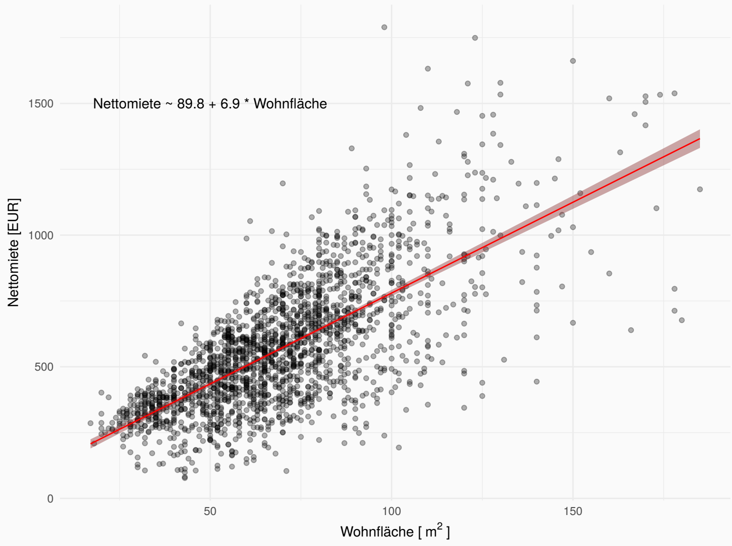



```
model_nm <- lm(nm ~ wfl, data = mietspiegel)
summary(model_nm)$coef
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     89.8      11.26       8  2.5e-15
## wfl              6.9       0.15      45 1.7e-311
```

Interpretation:
Mit einer Steigerung der Wohnfläche um 1m^2 ist durchschnittlich eine Steigerung der Miete um 6.9€ verbunden.

Beispiel: Nettomiete in Abhängigkeit von der Wohnfläche



Standardabweichung des Störterms

Die geschätzte Standardabweichung der y-Werte von der geschätzten Geraden ergibt sich zu:

$$
\begin{aligned}
s_\varepsilon &= \sqrt{\frac{1}{n-2}\sum \varepsilon_i^2} \\
\text{mit } \varepsilon_i &= y_i - \hat{y}_i
\end{aligned}
$$

Wichtiges intuitives Maß zur Modellanpassung, hier:

```r
sqrt(summary(model_nm)$sigma)
```

```
## [1] 13
```


Streuungs- und Quadratsummenzerlegung

Ziel:
Erklärung der Streuung von Y durch X

$\sum_{i=1}^n (y_i - \bar{y})^2$	$=$	$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	$+$	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$
Streuung von Y	$=$	Erklärte Streuung	$+$	Rest
SST	$=$	SSM	$+$	SSE
Quadratsumme Gesamt (Total)	$=$	Quadratsumme Regression (Model)	$=$	Quadratsumme Residuen (Error)

Das Bestimmtheitsmaß R^2

Anteil der durch die lineare Regression auf X erklärten Varianz:

$$
\begin{aligned}
R^2 &= \frac{SSM}{SST} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\
&= \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}
\end{aligned}
$$

Es gilt: Bestimmtheitsmaß = Quadrat der Bravais-Pearson-Korrelation zwischen X und Y:

$$R^2 = \frac{S_{xy}^2}{S_x^2 S_y^2} = r_{XY}^2$$

Wichtiges Maß zur Güte der Modellanpassung, hier:

```r
summary(model_nm)$r.squared
```

```
## [1] 0.5
```


Nachweis von $R^2 = r_{XY}^2$

$$\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^n \hat{y}_i = \frac{1}{n}\sum_{i=1}^n (\hat{\alpha} + \hat{\beta}x_i) = \hat{\alpha} + \hat{\beta}\bar{x} = (\bar{y} - \hat{\beta}\bar{x}) + \hat{\beta}\bar{x} = \bar{y}$$

Daraus folgt:

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2 = \sum_{i=1}^n (\hat{\alpha} + \hat{\beta}x_i - \hat{\alpha} + \hat{\beta}\bar{x})^2 = \hat{\beta}^2 \sum_{i=1}^n (x_i - \bar{x})^2$$

somit für R^2:

$$
\begin{aligned}
R^2 &= \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\hat{\beta}^2 \sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\
&= \frac{s_{XY}^2 \cdot s_X^2}{(s_X^2)^2 \cdot s_Y^2} = \left(\frac{s_{XY}}{s_X s_Y}\right)^2 = r_{XY}^2
\end{aligned}
$$

Umkehrregression

Vertauscht man die Rollen von X und Y, so erhält man die **Umkehrregression**.

Daten$(X_i, Y_i), i = 1, \ldots, n$

Regression:	$Y = \alpha + \beta X$	$\beta = \frac{S_{XY}}{S_X^2}$
Umkehrregression:	$X = \gamma + \delta Y$	$\delta = \frac{S_{XY}}{S_Y^2}$

Im XY-Koordinatensystem hat die Gerade der Umkehrregression die Darstellung

$$Y = -\frac{\gamma}{\delta} + \frac{1}{\delta} X$$

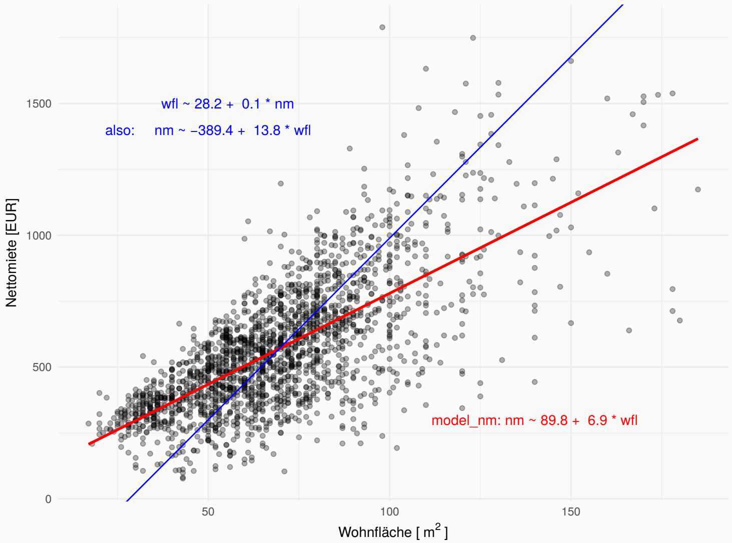
Umkehrregression

Es gilt:

$$\beta \cdot \delta = \frac{S_{XY}^2}{S_X^2 S_Y^2} = r^2 \leq 1$$

- $\Rightarrow |\beta| \leq \frac{1}{|\delta|}$, also Gerade der Umkehrregression steiler
- $\Rightarrow \beta \cdot \delta \geq 0$, also β undδ haben gleiches Vorzeichen

Beispiel: Umkehrregression



Orthogonale Regression

Falls man die orthogonalen quadratischen Abstände zur Gerade minimiert, erhält man eine Gerade zwischen Regression und Umkehrregression. Löse Minimierungsproblem in α, β:

$$(\alpha_{ORR}, \beta_{ORR}) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} \underbrace{\frac{(y_i - \alpha - \beta x_i)^2}{1 + \beta^2}}_{\text{Orthog. Abstand}}$$

$$\begin{aligned}
\hat{\beta}_{ORR} &= \frac{1}{2S_{XY}} \left[(S_Y^2 - S_X^2) + \sqrt{4S_{XY}^2 + (S_Y^2 - S_X^2)^2} \right] \\
\hat{\alpha}_{ORR} &= \bar{y} - \hat{\beta}_{ORR} \cdot \bar{x}
\end{aligned}$$

Wichtige Eigenschaften der linearen Regression

- Asymmetrie: Regressionsgerade von Y auf X verschieden von Regressionsgerade von X auf Y
- Die Regressionsgerade geht durch $(\bar{x}, \bar{y})$
- Interpretation der Steigung b steht im Mittelpunkt der Interpretation
- R^2-Wert gibt den Varianz-Erklärungsanteil wieder
- R^2 ist Quadrat der Korrelation zwischen X und Y
- s_ε gibt durchschnittliche Abweichung der Werte von der Regressionsgeraden an

Partielle Korrelation

Ziel:
Bestimmung der Korrelation zweier Merkmale unter "konstant halten" eines dritten Merkmals (auch: "kontrollieren für" ein drittes Merkmal).

Beispiel:
Korrelation der Quadratmetermiete und der Wohnfläche bei *konstanter* Zimmerzahl

Idee:
"Herausrechnen" des Zusammenhangs mit dem dritten Merkmal durch lineare Regression auf letzteres.

Partieller Korrelationskoeffizient (Definition)

Es sei:

$$\begin{aligned} X &= \hat{x} + E \\ &= a + bZ + E \\ Y &= \hat{y} + F \\ &= c + dZ + F \end{aligned}$$

(also: Residuen E bzw F)

Dann heißt die Maßzahl

$$r_{XY|Z} = r_{EF}$$

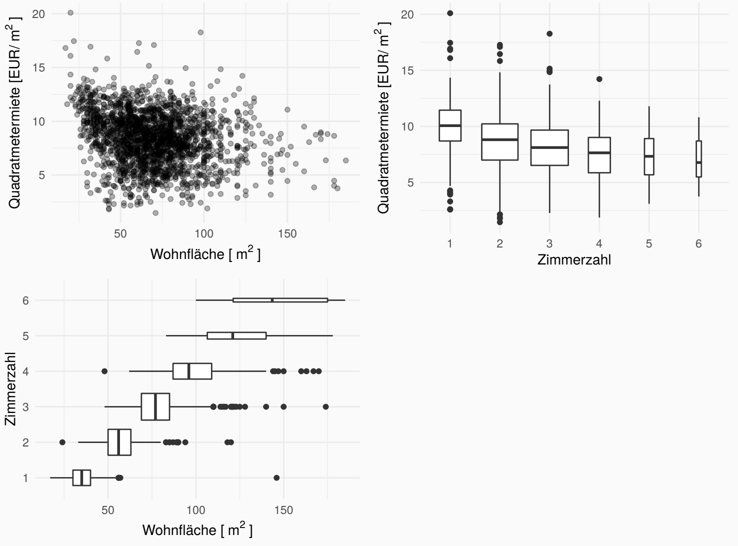
partieller Korrelationskoeffizient zwischen X und Y unter Z (auch "adjustiert/kontrolliert für Z").

Partieller Korrelationskoeffizient:

Es gilt:

$$r_{XY|Z} = \frac{r_{XY} - r_{XZ} r_{YZ}}{\sqrt{1 - r_{XZ}^2}\sqrt{1 - r_{YZ}^2}}$$

Beispiel: (Partielle) Korrelation der Quadratmetermiete und der Wohnfläche



Beispiel: (Partielle) Korrelation der Quadratmetermiete und der Wohnfläche

```
(marginal_correlations <- cor(mietspiegel[, c("nmqm", "wfl", "rooms")]))
```

```
##        nmqm   wfl rooms
## nmqm   1.00 -0.23 -0.27
## wfl   -0.23  1.00  0.84
## rooms -0.27  0.84  1.00
```

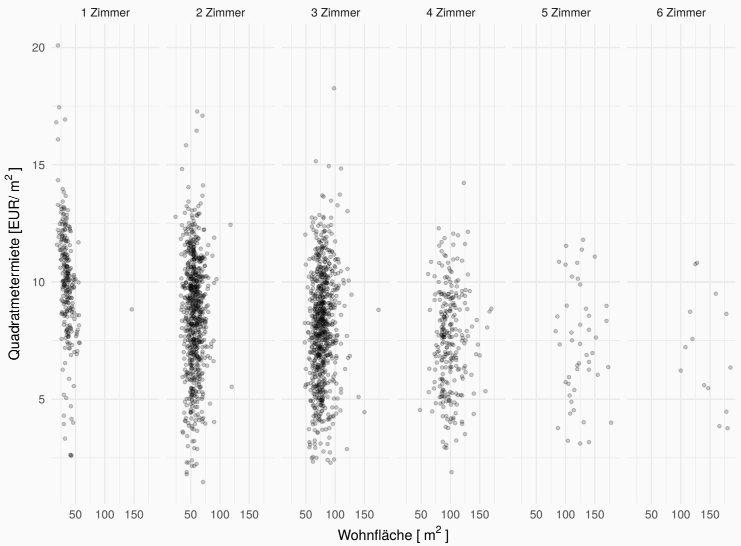
```
nmqm_residuals <- resid(lm(nmqm ~ rooms, data = mietspiegel))
wfl_residuals <- resid(lm(wfl ~ rooms, data = mietspiegel))
(partial_correlation <- cor(nmqm_residuals, wfl_residuals))
```

```
## [1] 0.005
```

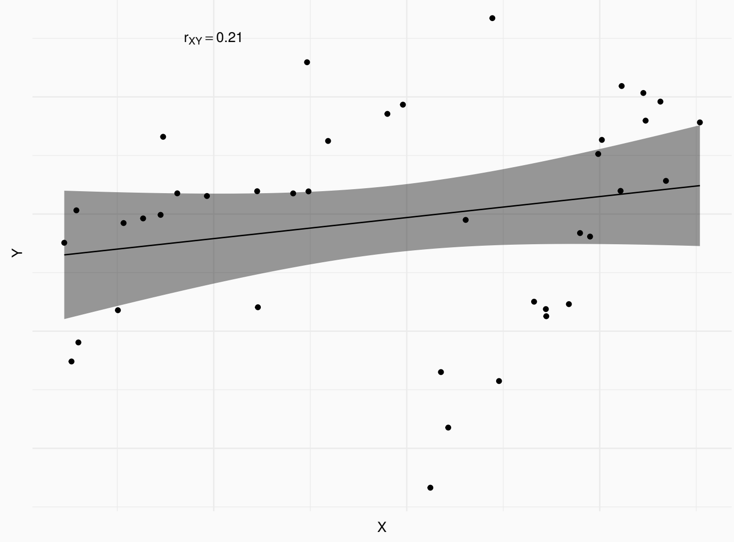
Marginale Korrelation zwischen Quadratmetermiete (`nmqm`) und Wohnfläche (`wfl`): $r_{\texttt{nmqm},\texttt{wfl}} \approx -0.23$

Partielle Korrelation zwischen Quadratmetermiete (`nmqm`) und Wohnfläche (`wfl`) adjustiert für Wohnfläche: $r_{\texttt{nmqm},\texttt{wfl}|\texttt{rooms}} \approx 0$

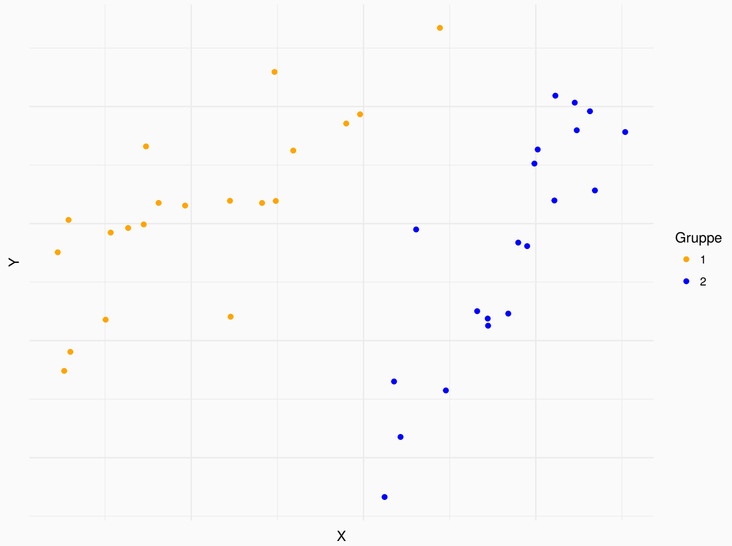
Beispiel: (Partielle) Korrelation der Quadratmetermiete und der Wohnfläche



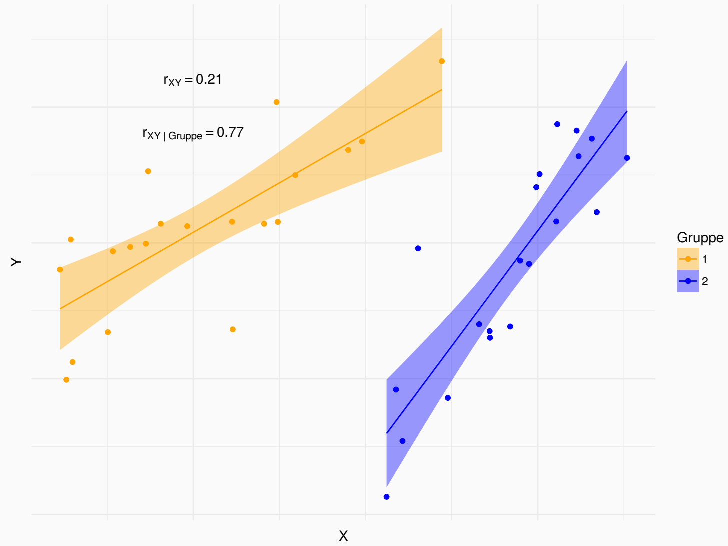
Beispiel: (Partielle) Korrelation



Beispiel: (Partielle) Korrelation



Beispiel: (Partielle) Korrelation



Multiples Regressionsmodell

Gegeben sind ein **Zielmerkmal** Y und die **Einflussgrößen** X_k

$$y = a + b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_p \cdot x_p + \varepsilon$$

Das Modell kann aus den entsprechenden Daten mit Hilfe der **KQ-Methode** geschätzt werden. Analog zum linearen Modell ist das **Bestimmtheitsmaß** r^2 ein zentrales Kriterium für die Modellanpassung.

Die Parameter b_k haben folgende Interpretation:
Steigt das Merkmal X_k um eine Einheit und bleiben die anderen konstant ("*ceteris paribus*-Bedingung"), so verändert sich Y im Durchschnitt um b_k Einheiten.

"Zielmerkmal" auch: *response, outcome, target*, abhängige Variable;
"Einflussgrößen" auch: Prädiktoren, Kovariablen, *features, inputs*, unabhängige Variablen.

Beispiel: Quadratmetermiete

Quadratmetermiete $=a+b_1\cdot$ Zimmerzahl $+b_2\cdot$ Baujahr $+b_3\cdot$ Zentralheizung $+\varepsilon$

```
multiple_m_nmqm <- lm(nmqm ~ rooms + bj + zh0, data = mietspiegel)
summary(multiple_m_nmqm)
```

```
##
## Call:
## lm(formula = nmqm ~ rooms + bj + zh0, data = mietspiegel)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.191 -1.472 -0.046  1.483 10.420
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -23.70299    4.16309   -5.69  1.4e-08 ***
## rooms        -0.60130    0.05068  -11.86  < 2e-16 ***
## bj            0.01728    0.00211    8.18  4.9e-16 ***
## zh0          -2.07650    0.18602  -11.16  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.2 on 2049 degrees of freedom
## Multiple R-squared:  0.185,  Adjusted R-squared:  0.184
## F-statistic:  155 on 3 and 2049 DF,  p-value: <2e-16
```

Beispiel: Quadratmetermiete

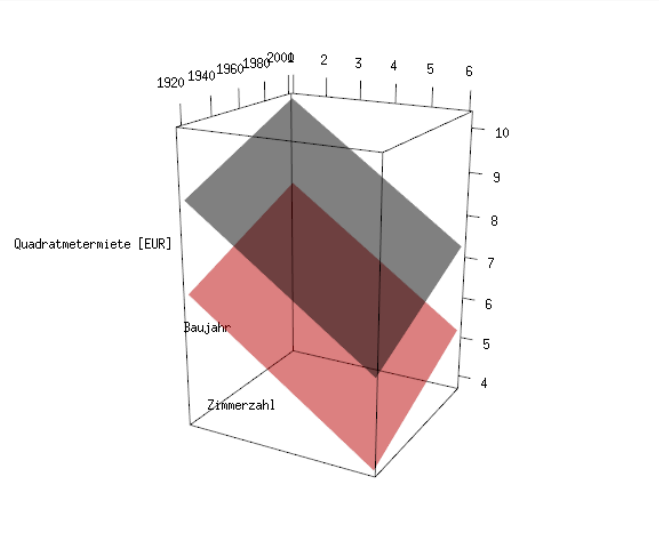


Abbildung 1:

Beispiel: Quadratmetermiete

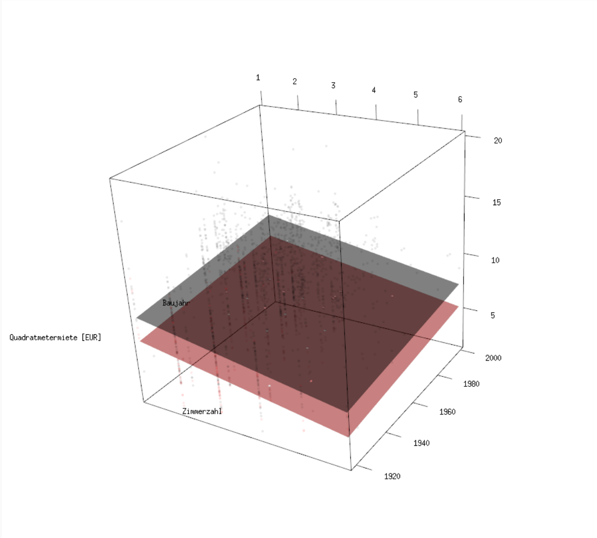


Abbildung 2:

Zusammenfassung multiples Regressionsmodell

Das multiple Regressionsmodell ist nützlich, um Zusammenhänge zwischen Merkmalen zu analysieren.

Es ermöglicht:

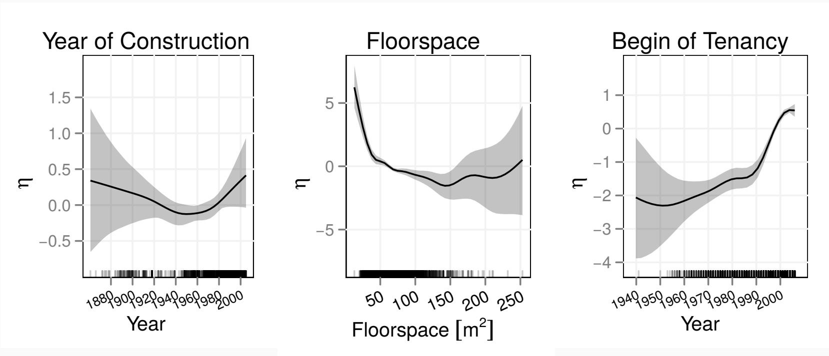
- Quantifizierung des Zusammenhangs
- Herausrechnen von Störgrößen
- Auswahl von relevanten Einflussgrößen

Erweiterungen

Erweiterungen des Modells beinhalten:

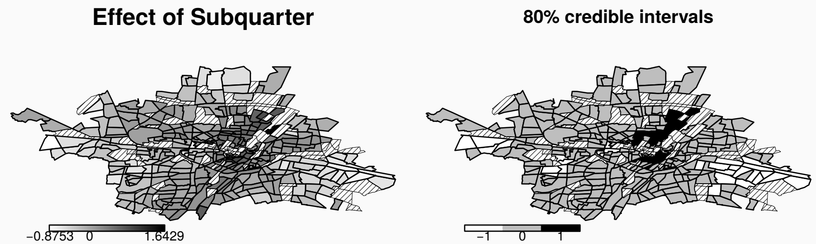
- Nichtlineare Zusammenhänge: $y = a + f_1(x_1) + f_2(x_2) + \cdots + \varepsilon$
- nominale/ordinale Merkmale als Einflussgrößen (z.B. Stadtbezirk, Geschlecht, Nationalität, etc.)
- Einbeziehung von Interaktionseffekten (z.B. geschlechterspezifische Effekte)
- Binäre Zielgrößen (krank/gesund), ordinale Zielgrößen, etc.

Beispiel: Räumliches additives Mietspiegelmodell



Quadratmetermiete = a + f(Baujahr) + f(Wohnfläche) + f(Bezugsdatum) + f(Stadtbezirk) $+ \cdots + \varepsilon$

Beispiel: Räumliches additives Mietspiegelmodell



Quadratmetermiete = a + f(Baujahr) + f(Wohnfläche) + f(Bezugsdatum) + f(Stadtbezirk) $+ \cdots + \varepsilon$

Nichtlineare Regressionsmodelle

Nichtlinearer Zusammenhang zwischen X und Y, β kann Vektor sein.

$$Y = f(X, \beta) + \varepsilon$$

KQ- Schätzer aus Daten Y_i, X_i:

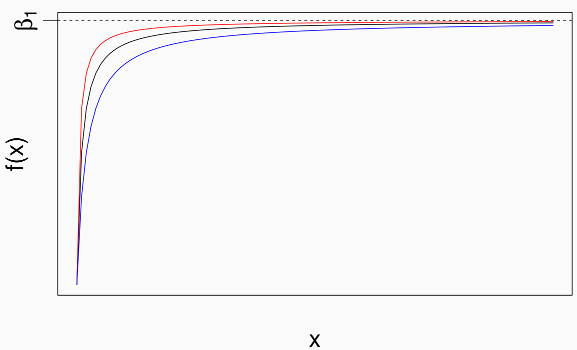
$$\hat{\beta} := \arg\min_{\beta} \sum_{i=1}^{n} (y_i - f(x_i, \beta))^2$$

Lösung meist nicht analytisch, nur numerisch möglich, z.B. mit Paket `nls` in R.

Beispiel: Michaelis-Menten-Modell

Beispiel: Michaelis-Menten-Modell zur Beschreibung von chemischen Reaktionsraten bei Konzentration x mit Obergrenze β_1 und Rate β_2.

$$y = \frac{\beta_1 \cdot x}{1/\beta_2 + x} + \varepsilon$$

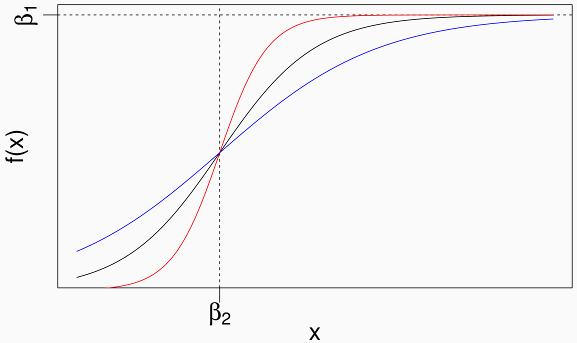


rotes $\beta_2 >$ schwarzes $\beta_2 >$ blaues β_2

Beispiel: Logistisches Wachstumsmodell

Beispiel: Logistisches Modell zur Beschreibung von limitierten Wachstumsprozessen mit Obergrenze β_1, Wendepunkt β_2 und Rate β_3

$$y = \frac{\beta_1}{1 + \exp((\beta_2 - x)\beta_3)} + \varepsilon$$



rotes β_3 > schwarzes β_3 > blaues β_3