13. Statistische Tests und p-Werte

13. Statistische Tests: das allgemeine Vorgehen

1. Formuliere eine **Nullhypothese H_0**, sodass ihre *Ablehnung* Erkenntnisse liefert, z.B. $\beta_j = \beta_{j0}$ (Punkthypothese) bzw. $\beta_j \le \beta_0$ (Intervallhypothese): *Man kann keine H_0 bestätigen!*
2. Wähle eine **Test-Statistik T**
   - mit bekannter Verteilung am Rande H_0^* der Nullhypothese ($H_0^* = H_0$ bei Punkyhypothesen)

     Was tun, wenn die Statistik eine bekannte Verteilung hat, aber die zugehörige Varianz unbekannt ist? Normalisierung auf die *geschätzte* Standardabweichung
   - mit klar definierten **Ablehnungsbereichen $R(\alpha)$** welche bei H_0 selten erreicht werden, aber häufiger bei $H_1 = \overline{H_0}$
3. Berechne eine Realisierung t_{data} von T aus den Daten
4. Teste, ob $t_{\text{data}} \in R(\alpha)$. Falls ja, kann H_0 bei einer **Fehlerwahrscheinlichkeit** bzw einem **Significanzniveau** α verworfen werden. Andernfalls *kann keine Aussage getroffen werden* (Maskenbeispiel mit H_0: "Maske nutzlos").

4a Alternativ und genauer: Berechne den **p-Wert** als kleinste Fehlerwahrscheinlichkeit α, bei der H_0 abgelehnt werden kann.

13.1.1 Wahl der Nullhypothese H_0: Fehler 1. und 2. Art

	H_0 not rejected	H_0 rejected
H_0 is true	✓	Type I error
H_0 is not true	Type II error	✓

- Ein Signifikanztest hat binären Input und Output. Zwei Kombinationen davon sind korrekt
- Kontrollierbar: **Fehler 1. Art bzw α-Fehler** $P(H_0 \text{ abgelehnt}|H_0) \leq \alpha$
- Unbekannt und nicht kontrollierbar: **Fehler 2. Art bzw. β-Fehler** $P(H_0 \text{ nicht abgelehnt}|\overline{H_0})$. Da dessen Wahrscheinlichkeit unbekannt ist, legt man die ernsthaftere Konsequenz auf den α-Fehler
- Fundamentales Problem: Gesucht: $P(H_0|\text{abgelehnt})$ and $P(H_0|\overline{\text{abgelehnt}})$, kann aber nur $P(\text{abgelehnt}|H_0) \leq P(\text{abgelehnt}|H_0^*)$ kontrollieren $\Rightarrow$ **Bayes'sche Statistik**

13.1.2 Schritte 2 und 3: Test-Statistik

- (i) Direkter Parametertest wie H_0: $\beta_j = \beta_{j0}$ or $\beta_j \geq \beta_{j0}$ or $\beta_j \leq \beta_{j0}$: Die Test-Statistik ist der geschätzte Abstand von H_0^* in Einheiten der *geschätzten* Standardabweichung. Sie ist **student-t** verteilt mit #dataPoints − #parameters **Freiheitsgraden (df)**:

$$T = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\hat{V}_{jj}}} \sim T(n-1-J)$$

- (ii) Test einer Funktion der Parameter, z.B. H_0: $\beta_1/\beta_2 = 2$, ≤ 2 or ≥ 2: Transformiere in eine Linearkombination, deren Normalisierung unter H_0^* wieder student-t verteilt ist. Im Beispiel gilt unter H_0^* die Bedingung $b = \beta_1 - 2\beta_2 = 0$:

$$\begin{aligned} \hat{b} &= \hat{\beta}_1 - 2\hat{\beta}_2, \\ \hat{V}(\hat{b}) &= \hat{V}_{11} + 4\hat{V}_{22} - 4\hat{V}_{12}, \\ T &= \frac{\hat{b}}{\sqrt{\hat{V}(\hat{b})}} \sim T(n-1-J) \end{aligned}$$

Teststatistiken II

- (iii) Test des Korrelationskoefficients von xy-Daten

$$\hat{\rho} = \frac{s_{xy}}{s_x s_y}, \quad H_0 : \rho = 0, \quad T = \frac{\hat{\rho}}{\sqrt{1-\hat{\rho}^2}}\sqrt{n-2} \sim T(n-2)$$

Herleitung: $\rho = 0$ genau dann, wenn $\beta_1 = 0$ in einer linearen Einfachregression $y = \beta_0 + \beta_1 x + \epsilon$, läuf also auf einen Test von $\beta_1 = 0$ hinaus. Under H_0 ist die Teststatistik

$$T = \hat{\beta}_1 / \sqrt{\hat{V}_{11}} = \frac{s_{xy}}{\hat{\sigma}\ s_x}\sqrt{n} \sim T(n-2)$$

Setzt man $\hat{\sigma}$ ein, welcher hier analytisch berechnet werden kann: $\hat{\sigma}^2 = n(s_y^2 - s_{xy}^2/s_x^2)/(n-2)$, bekommt man obiges Ergebnis

- (iv) Test der Varianz von ϵ, H_0: $\sigma^2 = \sigma_0^2$, $\sigma^2 \geq \sigma_0^2$, and $\sigma^2 \leq \sigma_0^2$:

$$T = \frac{\hat{\sigma}^2}{\sigma_0^2}\ (n-1-J) \sim \chi^2(n-1-J)$$

Die **chi-Quadrat-Verteilung mit m Freiheitsgraden** $\chi^2(m) = \sum_{i=1}^m Z_i^2$ ist die Summe von m *quadrierten* unabhängigen standardnormalverteilten Zufallsvariablen *Dichte nicht symmetrisch* $\Rightarrow$ *benötige Tabellen für die α and $1-\alpha$-Quantile*

Teststatistiken III

- (v) Tests simultaner Punkthypothesen, e.g., H_0: $(\beta_1 = 0)$ AND $(\beta_2 = 2)$ mit dem **Fisher-F test**:

$$T = \frac{(S_0 - S)/(M - M_0)}{S/(n-M)} \sim F(M - M_0, n - M)$$

  - S: SSE des geschätzten vollen Modells mit $M = J + 1$ Parametern
  - S_0: SSE des geschätzten *restringierten* Modells unter H_0 mit M_0 freien Parametern

- Die **Fisher-F** distribution ist i.W. der Quotient zweier unabhängiger χ^2 verteilter Zufallsgrößen

$$F(n, d) = \frac{\chi_n^2/n}{\chi_d^2/d},$$

n: Zähler- und d: Nenner-Zahl der Freiheitsgrade

? Warum gilt immer $S_0 \geq S$?

? Warum ist der F-Test mit $M_0 = M - 1$ äquivalent zum T-Test?

13.1.3 Schritt 4: Test-Entscheidung

- Die Testentscheidung basiert auf den *Ablehnungsbereich* der Realisierungen t_{data}:

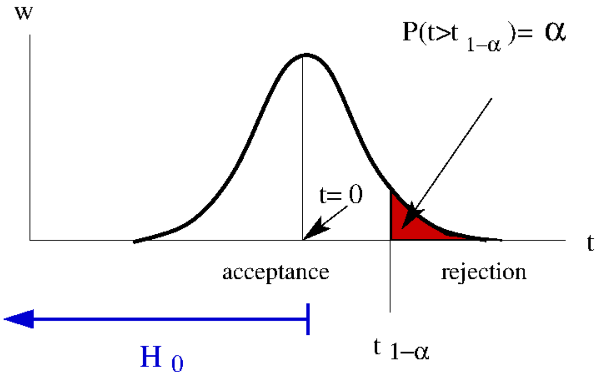
> Der **Ablehnungsbereich $R^{(H_0)}(\alpha)$** zur Fehlerwahrscheinlichkeit α enthält unter H_0^* den Anteil α der Realisierungen t von T, welcher am weitesten von H_0 entfernt ist

- Entscheidung:

> H_0 is abgelehnt zur Fehlerwahrscheinlichkeit α, falls $t_{\text{data}} \in R^{(H_0)}(\alpha)$

- Für gute Teststatistiken ist $R(\alpha)$ intuitiv definiert und die Ablehnungsregion ist so nahe an H_0^* wie möglich
- Im Gegensatz zu T und t_{data}, die nur vom Rand H_0^* abhängen ($\Rightarrow$ dieselben für alle Arten von Punkt- und einseitigen Intervallhypothesen) hängt $R^{(H_0)}(\alpha)$ von der Art der Nullhypothese und ggf vom Vergleichsoperator ($=$, $\geq$, $\leq$) ab

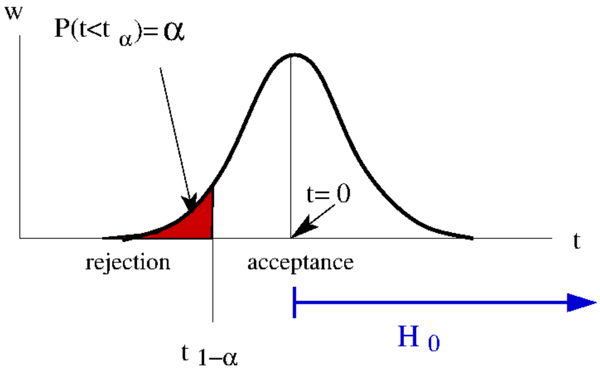
1. Ablehnungsbereich für H_0: "$<$" or "$\leq$" (Intervallhypothese)



- H_0 is abgelehnt zum Signifikanzniveau bzw. zur Fehlerwahrscheinlichkeit α falls

$$t_{\text{data}} > t_{1-\alpha}$$

2. Ablehnungsbereich für H_0: “$>$” or “$\geq$” (Intervallhypothese)

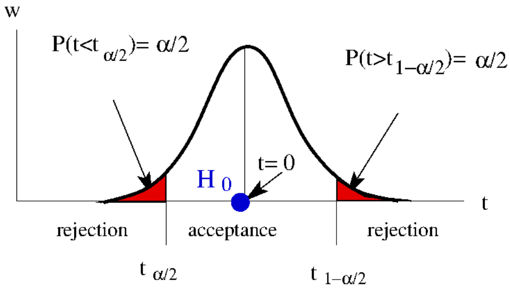


- H_0 ist abgelehnt zur Fehlerwahrscheinlichkeit α falls

$$t_{\text{data}} < t_{\alpha} = 1 - t_{1-\alpha}$$

- (Das Gleichheitszeichen gilt nur für symmetrisch verteilte Statistiken)

3. Ablehnungsbereich für H_0: "=" (Punkthypothese)



- Symmetrisch Teststatistiken: H_0 zum Niveau α abgelehnt, falls

$$\boxed{|t_{\text{data}}| > t_{1-\alpha/2}}$$

- Nichtsymmetrische Teststatistiken wie die χ^2 Verteilung des Varianztests: Definition "was ist am weitesten entfernt" nicht eindeutig $\Rightarrow$ nehme auf beiden Seiten gleiche statistische Gewichte $\alpha/2$ an:

$$\text{abgelehnt} \quad \Leftrightarrow (t_{\text{data}} < t_{\alpha/2}) \cup (t_{\text{data}} > t_{1-\alpha/2})$$

Hotel-Beispiel

Zweifachregression der Hotelbelegung $y(\boldsymbol{x})$ [%]:

$$y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

$x_0 = 1$, x_1: Qualitäts-Proxy [# Sterne]; x_2: Preis [€/night],

$$\hat{\beta}_0 = 25.5, \quad \hat{\beta}_1 = 38.2, \quad \hat{\beta}_2 = -0.952$$

mit der geschätzten Varianz-Kovarianz-Matrix

$$\hat{\mathbf{V}} = \begin{pmatrix} 28.0 & -6.40 & -0.119 \\ -6.40 & 26.0 & -0.941 \\ -0.119 & -0.941 & 0.0397 \end{pmatrix}$$

? Formuliere und teste H_{01}: "Die Sterne sind egal"

! $H_{01} : \beta_1 = 0$, $T = \hat{\beta}_1 / \sqrt{\hat{V}_{11}} \sim T(12-3)$, also $df = 9$ Freiheitsgrade, $t_{\text{data}} = 7.49$, $t^{(9)}_{0.975} = 2.26 < |t_{\text{data}}| \Rightarrow H_0$ abgelehnt, die Sterne sind relevant

? Bevorzugen die Leute mehr Sterne? (bei $\alpha = 5\,\%$)?

! $H_{02} : \beta_1 <= 0$ (wähle H_0 so, dass eine Ablehnung eine Aussage ergibt!), Intervalltest mit T und t_{data} wie oben, aber neue Entscheidung: $t^{(9)}_{0.95} = 1.83 < t_{\text{data}} \Rightarrow H_{02}$ abgelehnt, mehr Sterne sind besser als weniger Sterne

Hotelbeispiel (Fortsetzung)

? Reduziert eine Preiserhöhung um einen € pro nacht die Auslastung um höchstens 1 %?

! $H_{03}: \beta_2 < -1$ (H_{03} is wieder das Gegenereignis!),
$t_\text{data} = (\hat{\beta}_2 + 1)/\sqrt{\hat{V}_{22}} = 0.24 \overset{!}{>} t_{0.95}^{(9)} = 1.83 \Rightarrow H_{03}$ nicht abgelehnt
$\Rightarrow$ der Hotelmanager riskiert einen Verlust von möglicherweise mehr als einen Prozentpunkt

? Lohnt sich die Investition für einen zusätzlichen Stern sodass ich 30 € mehr verlangen kann, ohne Gäste zu verlieren?

! Definiere wieder das Gegenereignis $H_{04}: \beta_1 \le -30\beta_2$ bzw. $\gamma = \beta_1 + 30\beta_2 \le 0$

$$\begin{aligned} \hat{\gamma} &= \hat{\beta}_1 + 30\hat{\beta}_2 = 9.63, \\ \hat{V}(\hat{\gamma}) &= \hat{V}_{11} + 900\hat{V}_{22} + 2 * 1 * 30\hat{V}_{12} = 5.27 \end{aligned}$$

$t_\text{data} = \hat{\gamma}/\sqrt{\hat{V}(\hat{\gamma})} = 4.20 > t_{0.95}^{(9)} = 1.83 \Rightarrow H_{04}$ abgelehnt bei 5 % $\Rightarrow$ Das Risiko einer geringeren Belegung liegt unter 5 %

? Kann simultan $\beta_1 = 30$ und $\beta_2 = -1$ gelten?

! Volles Modell: $\hat{\boldsymbol{\beta}} = (25.5, 38.2, -0.952)'$, $S(\hat{\boldsymbol{\beta}}) = 498.2$;
Restringiertes Modell mit festem $\beta_1 = 30$, $\beta_2 = 1 \Rightarrow \hat{\beta}_0 = 49.0$:
$\hat{\boldsymbol{\beta}}_r = (49.0, 30, -1)'$, $S_0 = S(\hat{\boldsymbol{\beta}}_r) = 1808$; $M - M_0 = 2\,\text{df}$, $n - M = 9\,\text{df}$,
$T \sim F(2,9)$, $t_\text{data} = 9/2\ (S_0 - S)/S = 11.8 > f_{0.95}^{(2.9)} = 4.26 \Rightarrow H_0$ abgelehnt

13.1.4 Der p-Wert

- Ein Test auf eine feste Fehlerwahrscheinlichkeit α ist nicht sonderlich effizient, da man im Dunkeln über die *wirkliche Signifikanz* stochert:
- Ich will eigentlich die *minimale* Fehlerwahrscheinlichkeit wissen, die eine Ablehnung ergibt: Das ist der **p-Wert**.
- Allgemeinste Definition:

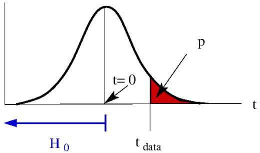
$$\boxed{p = \text{Prob}(T \in E_{\text{data}} | H_0^*)}$$

Die *Extremregion* E_{data} enthält alle möglichen Realisierungen von T, welche *weiter* von H_0 entfernt sind als t_{data}, also liegt t_{data} auf dem Rand von E_{data} Zusammenhang mit Ablehnungsbereich?
p is so definiert, dass $E_{\text{data}} = R(p)$

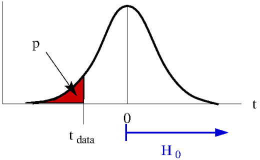
  - $p \geq 5\,\%$: nicht signifikant (kein Stern beim Wert für β, manchmal "+", falls zwischen 5 % und 10 %, z.B. $\beta_1 = 4.2^+$)
  - $p < 5\,\%$: signifikant (ein Stern, z.B. $\beta_1 = 4.2^*$)
  - $p < 1\,\%$: sehr signifikant (zwei Sterne, $\beta_1 = 4.2^{**}$)
  - $p < 0.001$: hochgradig signifikant (drei Sterne, $\beta_1 = 4.2^{***}$)

Berechnung von p für einige Tests

- Intervalltest $H_0 : \beta \leq \beta_0$ or $\beta < \beta_0$
  $p = P(T > t_{\text{data}} | \beta = \beta_0) = 1 - F_T(t_{\text{data}})$

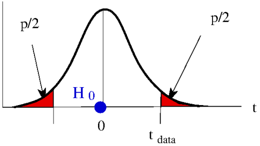


- Intervalltest $H_0 : \beta \geq \beta_0$ or $\beta > \beta_0$
  $p = P(T < t_{\text{data}} | \beta = \beta_0) = F_T(t_{\text{data}})$



- Punkttest $H_0 : \beta = \beta_0$ (Symmetrie von f_T beim 3. Gleichheitszeichen angenommen)

$$\begin{aligned} p &= P\big((T > |t_{\text{data}}|) \cup (T < -|t_{\text{data}}|)\big) \\ &= (1 - F_T(|t_{\text{data}}|)) + F_T(-|t_{\text{data}}|) \\ &= 1 - F_T(|t_{\text{data}}|) + 1 - F_T(|t_{\text{data}}|) \\ &= 2(1 - F_T(|t_{\text{data}}|)) \end{aligned}$$



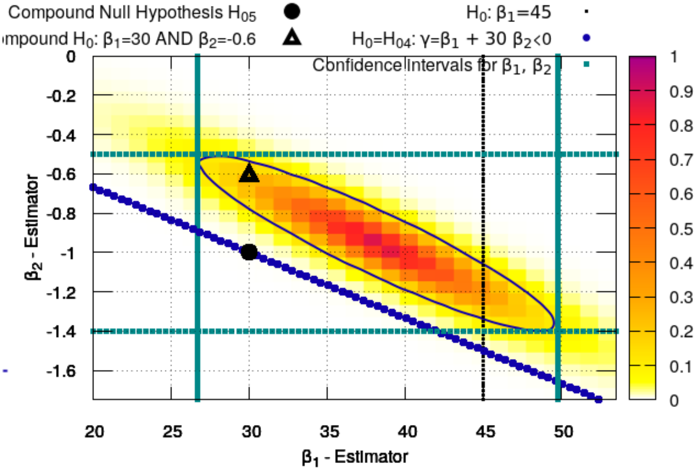
p-Werte der Nullhypothesen des Hotelbeispiels

$$y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

$x_0 = 1$, x_1: Qualitäts-Proxy [# Sterne]; x_2: Preis [€/night],

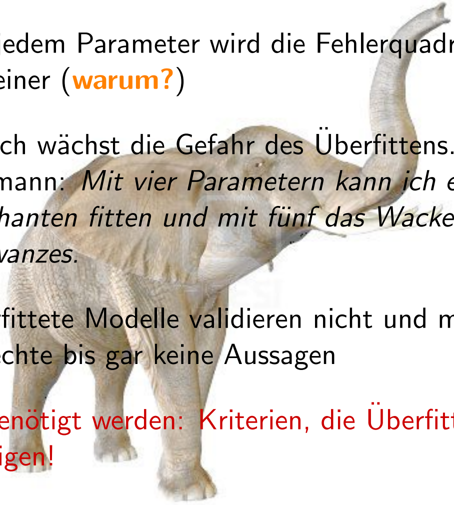
- H_{01} "Sterne egal": Punkt-Hypothese $\beta_1 = 0$
  $t_{\text{data}} = 7.49,\ p = 2(1 - F_T^{(9)})(|t_{\text{data}}|) = 3.7E-5^{***}$
- H_{02} "Mehr Sterne sind besser": Intervall-Hypothese $\beta_1 < 0$
  $t_{\text{data}} = 7.49,\ p = 1 - F_T^{(9)}(t_{\text{data}}) = 1.9E-5^{***}$
- H_{03} "Belegung geht um ≥ 1 Prozentpunkte pro € Preisanstieg zurück": Intervallhypothese $\beta_2 < -1$
  $t_{\text{data}} = 0.24,\ p = 1 - F_T^{(9)}(t_{\text{data}}) = 40\,\%$
- H_{04} Ein Stern ist mindestens 30 € wert": Funktions-Intervallhypothese $\gamma = \beta_1 + 30\beta_2 < 0$
  $t_{\text{data}} = 4.20,\ p = 1 - F_T^{(9)}(t_{\text{data}}) = 0.12\,\%^{**}$
- H_{05} "Stern- und Preissensitivität ist gleichzeitig gegeben": Simultane Nullhypothese $(\beta_1 = 30) \cap (\beta_2 = -1)$
  $t_{\text{data}} = 11.8,\ p = 1 - F_F^{(2,9)}(t_{\text{data}}) = 0.30\,\%^{**}$

Visualisierung



- Türkise Linien: Grenzen der $\alpha = 5\,\%$-KI von β_1 und β_2
- Schwarze Linie: Punkthypothese (t-test): $H_0 : \beta_1 = 45$
- Blaue Punkte: Grenze der funktionalen Intervallhypothese (t-test) $H_0 = H_{04} : \gamma = \beta_1 + 30\beta_2 < 0$
- Schwarze Symbole: Simultane Nullhypothese (F-test)
  $\bullet$: $H_0 = H_{05} : (\beta_1 = 30) \cap \beta_2 = -1)$, $\triangle$: $H_0 : (\beta_1 = 30) \cap (\beta_2 = -0.6)$.

13.2 Strategien zur Modellauswahl
Problemstellung



- Mit jedem Parameter wird die Fehlerquadratsumme S kleiner (**warum?**)
- Jedoch wächst die Gefahr des Überfittens. John Neumann: *Mit vier Parametern kann ich einen Elephanten fitten und mit fünf das Wackeln seines Schwanzes.*
- Überfittete Modelle validieren nicht und machen schlechte bis gar keine Aussagen
- $\Rightarrow$ benötigt werden: Kriterien, die Überfitten anzeigen!

Standard-Kriterien zur Modellauswahl

- **(1) Korrigiertes R^2:**

$$\bar{R}^2 = 1 - \frac{n-1}{n-J-1}\left(1 - R^2\right), \quad R^2 = 1 - \frac{S}{S_0},$$

$S = \text{SSE(kalibr. volles Modell)}, \quad S_0 = \text{SSE(kalibr. Modell } \hat{y} = \beta_0).$

- **(2) Akaike Informationskriterium AIC:**

$$\text{AIC} = \ln \hat{\sigma}^2_{\text{descr}} + J\,\frac{2}{n},$$

- **(3) Bayes'sches Informationskriterium BIC:**

$$\text{BIC} = \ln \hat{\sigma}^2_{\text{descr}} + J\,\frac{\ln n}{n}\,.$$

Hinweis: Bei AIC und BIC wird der deskriptive Varianzschätzer $\hat{\sigma}^2_{\text{descr}} = S/n$ anstelle des Unverzerrten, $\hat{\sigma}^2 = S/(n-1-J)$, genommen

Modellauswahl: Strategy à la "Occam's Rasiermesser"

- Identifiziere J möglicherweise relevante exogene Faktoren (einschließlich der Konstanten) und berechne $\bar{R}^2$, AIC, oder BIC für alle 2^J möglichen Modelle (jeder der Faktoren j ist entweder enthalten oder nicht) durch *rohe Gewalt*.
- Das beste Modell maximiert $\bar{R}^2$ bzw minimiert AIC oder BIC.
- Da AIC (und auch $\bar{R}^2$ komplexe Modelle mit vielen Parametern zu gering "bestrafen", ist üblicherweise BIC das Kriterium der Wahl
- Neben dem Ansatz der *rohen Gewalt* gibt es zwei "schnellere" Strategien, die jedoch nicht garantieren, das beste Modell zu finden (BIC etc sind nicht transitiv)
  - **Top-down**: Starte mit allen J Faktoren. Eliminiere in jeder Runde den Faktor, der den höchsten Zuwachs in $\bar{R}^2$ bzw die höchste Abnahme von AIC oder BIC bewirkt. Stoppe, falls es keine Verbesserung mehr gibt
  - **Bottom-up**: Starte mit dem Konstantenmodell $\hat{y} = \beta_0$ und füge jeweils den Faktor mit der höchsten inkrementellen Verbesserung hinzu, solange es Verbesserungen gibt
- Standard Statistik-Software enthält diese Strategien

13.3. Logistische Regression

- Normale lineare Modelle der Form $Y = \boldsymbol{\beta}'\boldsymbol{x} + \epsilon$ haben zwingend eine stetige endogene Variable (**Warum?**)
- Verkettung mit einem nichtbeobachtbaren Zwischenzustand Y^* erlaubt es, dichotome endogene Variable zu moellieren:

$$Y(\boldsymbol{x}) = \begin{cases} 1 & Y^*(\boldsymbol{x}) > 0 \\ 0 & \text{otherwise,} \end{cases} \quad Y^*(\boldsymbol{x}) = \hat{y}^*(\boldsymbol{x}) + \epsilon = \boldsymbol{\beta}'\boldsymbol{x} + \epsilon$$

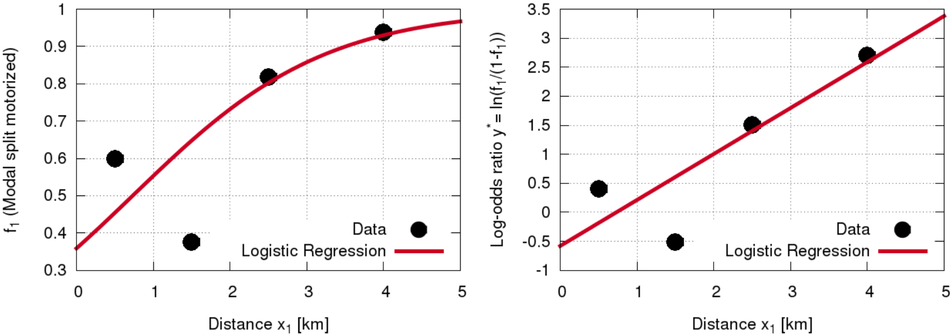
  ϵ gehorcht der **logistischen Verteilung** mit $F_\epsilon(x) = e^x/(e^x + 1)$
- Wahrscheinlichkeit P_1 des Ergebnisses $Y = 1$:

$$P_1 = P(Y^*(\boldsymbol{x}) > 0) = F_\epsilon(\boldsymbol{\beta}'\boldsymbol{x}) = \frac{e^{\boldsymbol{\beta}'\boldsymbol{x}}}{e^{\boldsymbol{\beta}'\boldsymbol{x}} + 1}$$

- Formal ist dies ein normales lineares Regressionsmodell für das **log-odds ratio** $\ln(P_1/P_0) = \ln(P1/(1 - P_1))$:

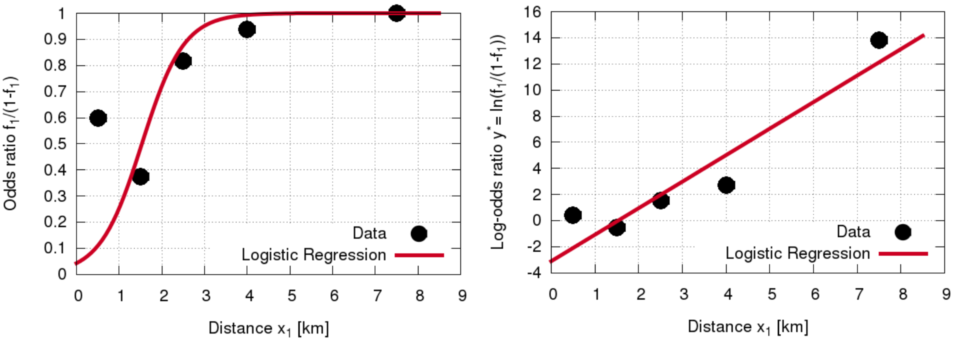
$$\hat{y}^*(\boldsymbol{x}) = \boldsymbol{\beta}'\boldsymbol{x} = \ln\left(\frac{P_1}{P_0}\right)$$

Example: naive OLS-estimation (RP student interviews)



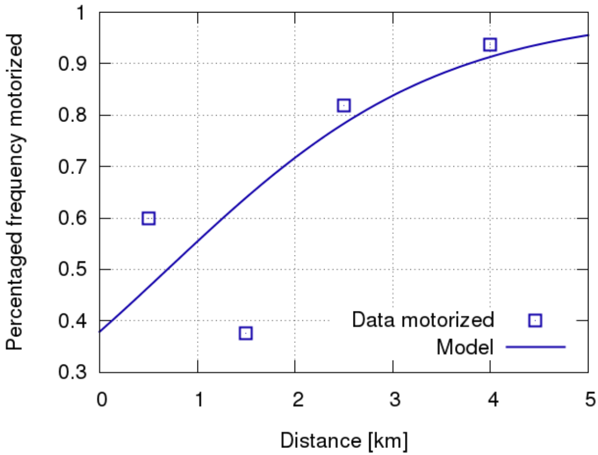
- Alternatives: $i = 1$: motorized and $i = 2$ (not)
- Intermediate variable estimated by percentaged choices: $y^* = \ln(f_1/(1 - f_1))$
- Model: Log. regression, $\hat{y}^*(x_1) = \beta_0 + \beta_1 x_1$
- OLS Estimation: $\beta_0 = -0.58, \quad \beta_1 = 0.79$

Method consistent? added 5th data point with f=0.9999



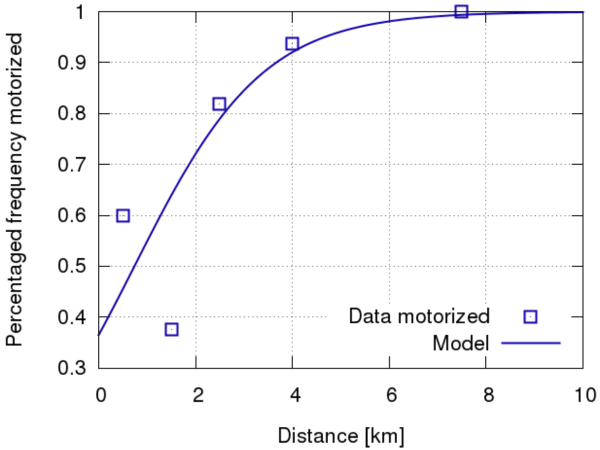
- Same model: $\hat{y}^*(x_1) = \beta_0 + \beta_1 x_1$
- New estimation: $\beta_0 = -3.12, \quad \beta_1 = 2.03$
- Estimation would fail if $f_1 = 0$ or $=1 \Rightarrow$ real discrete-choice model necessary!

Comparison: real Maximum-Likelihood (ML) estimation



- Model: Logit, $V_i(x_1) = \beta_0\delta_{i1} + \beta_1 x_1 \delta_{i1}$, $V_2 = 0$.
- Estimation: $\beta_0 = -0.50 \pm 0.65$, $\beta_1 = +0.71 \pm 0.30$

Comparison: real ML estimation with added 5th data point



- Same logit model, $V_i(x_1) = \beta_0\delta_{i1} + \beta_1 x_1 \delta_{i1}$, $V_2 = 0$.
- New estimation: $\beta_0 = -0.55 \pm 0.63$, $\beta_1 = +0.75 \pm 0.27$