Einführung in die Induktive Statistik: Regressionsanalyse

Jan Gertheiss
LMU München

Sommersemester 2011

Vielen Dank an Christian Heumann für das Überlassen von TEX-Code!

Regressionsanalyse

- Ziel: Analyse des Einflusses einer oder mehrerer Variablen $X_1, \ldots, X_p$ auf eine Zielvariable Y.
- Bezeichnungen:

  $X_1, \ldots, X_p$ erklärende Variablen (exogene Variablen, Kovariablen, Regressoren, Prädiktoren)

  Y Zielvariable (abhängige Variable, endogene Variable, Regressand, Response)
- Verschiedene Arten von Regressionsmodellen, abhängig vom Typ der Zielvariable Y und der Art des Einflusses von $X_1, \ldots, X_p$.
- Hier: Y metrisch/stetig.

Regressionsanalyse

- Lineare Einfachregression
- Das multiple lineare Regressionsmodell
- Ausblick: Varianzanalyse, nichtlineare und nichtparametrische Regression, generalisierte Regression.

Lineare Einfachregression

Einführung

Datensituation wie beim Streudiagramm (Deskriptive Statistik):

(y_i, x_i), $i = 1, \ldots, n$, Beobachtungen für stetige bzw. metrische Merkmale Y und X.

Beispiel: Mietspiegel

Y Nettomiete bzw. Nettomiete/qm, X Wohnfläche.

Lineare Einfachregression

Einführung

- Zusammenhang zwischen Y und X nicht deterministisch, sondern durch (zufällige) Fehler additiv überlagert.

$$Y = f(x) + \epsilon,$$

  wobei f deterministische Funktion, ϵ additiver Fehler.

- Lineare Einfachregression: f linear, d.h.

$$Y = \alpha + \beta x + \epsilon.$$

- Primäres Ziel: Schätze α und β aus Daten (y_i, x_i), $i = 1, \ldots, n$. Unterstelle dabei lineare Beziehung

$$y_i = \alpha + \beta x_i + \epsilon_i,$$

  wobei $\alpha + \beta x_i$ systematische Komponente, ϵ_i zufällige Fehler mit $E(\epsilon_i) = 0$.

  Weitere Annahmen an die Fehler ϵ_i:

$$\epsilon_i \text{ i.i.d. mit } \sigma^2 = Var(\epsilon_i)$$

Lineare Einfachregression

Einführung

Standardmodell der linearen Einfachregression:

Es gilt

$$Y_i = \alpha + \beta x_i + \epsilon_i\,, \quad i = 1, \ldots, n.$$

Dabei sind:

$Y_1, \ldots, Y_n$	beobachtbare metrische Zufallsvariablen,
$x_1, \ldots, x_n$	gegebene deterministische Werte oder Realisierungen einer metrischen Zufallsvariable X.
$\epsilon_1, \ldots, \epsilon_n$	unbeobachtbare Zufallsvariablen, die unabhängig und identisch verteilt sind mit $E(\epsilon_i) = 0$ und $Var(\epsilon_i) = \sigma^2$.

Die Regressionskoeffizienten α, β und die Varianz σ^2 sind unbekannte Parameter, die aus den Daten $(y_i, x_i), i = 1, \ldots, n$, zu schätzen sind.

Lineare Einfachregression

Einführung

Bemerkungen:

- Falls Regressoren nicht deterministisch sondern stochastisch, bedingte Betrachtungsweise, d.h. Modell und Annahmen unter der Bedingung $X_i = x_i$, $i = 1, \ldots, n$.
- Eigenschaften der Zielvariablen:

$$E(Y_i \mid x_i) = E(\alpha + \beta x_i + \epsilon_i) = \alpha + \beta x_i$$

$$Var(Y_i \mid x_i) = Var(\alpha + \beta x_i + \epsilon_i) = Var(\epsilon_i) = \sigma^2$$

$$Y_i \mid x_i,\ i = 1, \ldots, n, \text{ unabhängig}$$

- Oft zusätzlich Normalverteilungsannahme:

$$\epsilon_i \sim N(0, \sigma^2) \quad \text{bzw.} \quad Y_i \mid x_i \sim N(\alpha + \beta x_i, \sigma^2)$$

Schätzen, Testen und Prognose

Ziele:

- Punkt- bzw. Intervallschätzer für α, β und σ^2.
- Testen von Hypothesen über α und v.a. β.
- Prognose von Y für neuen Wert x des Regressors X.

Schätzen:

KQ-(Kleinste-Quadrate-)Methode: Bestimme Schätzer $\hat{\alpha}, \hat{\beta}$ so, dass

$$\sum_{i=1}^{n} (Y_i - \alpha - \beta x_i)^2 \to \min_{\alpha,\beta}.$$

Lösung:

KQ-Schätzer

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x} \ , \qquad \hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Schätzer für die Varianz σ^2:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\epsilon}_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - (\hat{\alpha} + \hat{\beta}x_i))^2$$

Lineare Einfachregression

Schätzen, Testen und Prognose

Geschätzte Regressionsgerade (Ausgleichsgerade):

$$\hat{Y} = \hat{\alpha} + \hat{\beta}x$$

Geschätzte Fehler, Residuen:

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\alpha} - \hat{\beta}x_i$$

Streuungszerlegung und Bestimmtheitsmaß

Streuungszerlegung (Quadratsummenzerlegung):

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{SQT} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{SQE} + \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{SQR}$$

SQT: Gesamtabweichungsquadratsumme in Y-Richtung
SQE: Durch die Regression erklärter Teil von SQT
SQR: Trotz der Regression unerklärt bleibender Teil von SQT

Lineare Einfachregression

Schätzen, Testen und Prognose

Bestimmtheitsmaß:

- Definition:

$$R^2 = \frac{SQE}{SQT} = 1 - \frac{SQR}{SQT}$$

- Berechnung:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{\sum_{i=1}^{n}\hat{Y}_i^2 - n\bar{Y}^2}{\sum_{i=1}^{n}Y_i^2 - n\bar{Y}^2}$$

Verteilungseigenschaften der KQ-Schätzer

Verteilung der geschätzten Regressionskoeffizienten unter Normalverteilungsannahme bzgl. ϵ_i bzw. Y_i:

$$\hat{\alpha} \sim N(\alpha, \sigma^2_{\hat{\alpha}}) \quad \text{mit} \quad Var(\hat{\alpha}) = \sigma^2_{\hat{\alpha}} = \sigma^2 \frac{\sum_i x_i^2}{n \sum_i (x_i - \bar{x})^2} = \sigma^2 \frac{\sum_i x_i^2}{n(\sum_i x_i^2 - n\bar{x}^2)}$$

$$\hat{\beta} \sim N(\beta, \sigma^2_{\hat{\beta}}) \quad \text{mit} \quad Var(\hat{\beta}) = \sigma^2_{\hat{\beta}} = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} = \frac{\sigma^2}{\sum_i x_i^2 - n\bar{x}^2}$$

Verteilung der standardisierten Schätzfunktionen (unter NV-Annahme):

$$\frac{\hat{\alpha} - \alpha}{\hat{\sigma}_{\hat{\alpha}}} \sim t(n-2) \quad \text{mit} \quad \hat{\sigma}_{\hat{\alpha}} = \hat{\sigma} \frac{\sqrt{\sum_i x_i^2}}{\sqrt{n \sum_i (x_i - \bar{x})^2}} = \hat{\sigma} \frac{\sqrt{\sum_i x_i^2}}{\sqrt{n(\sum_i x_i^2 - n\bar{x}^2)}}$$

$$\frac{\hat{\beta} - \beta}{\hat{\sigma}_{\hat{\beta}}} \sim t(n-2) \quad \text{mit} \quad \hat{\sigma}_{\hat{\beta}} = \frac{\hat{\sigma}}{\sqrt{\sum_i (x_i - \bar{x})^2}} = \frac{\hat{\sigma}}{\sqrt{\sum_i x_i^2 - n\bar{x}^2}}$$

Lineare Einfachregression

Schätzen, Testen und Prognose

Aus den Verteilungseigenschaften folgen:

- $(1-\alpha)$-Konfidenzintervalle für α und β:

  für α: $\quad \left[\hat{\alpha} - \hat{\sigma}_{\hat{\alpha}} t_{1-\alpha/2}(n-2),\ \hat{\alpha} + \hat{\sigma}_{\hat{\alpha}} t_{1-\alpha/2}(n-2)\right]$

  für β: $\quad \left[\hat{\beta} - \hat{\sigma}_{\hat{\beta}} t_{1-\alpha/2}(n-2),\ \hat{\beta} + \hat{\sigma}_{\hat{\beta}} t_{1-\alpha/2}(n-2)\right]$

- Teststatistiken T_{α_0} und T_{β_0} zum Testen von Hypothesen bzgl. α und β:

$$T_{\alpha_0} = \frac{\hat{\alpha} - \alpha_0}{\hat{\sigma}_{\hat{\alpha}}} \quad \text{und} \quad T_{\beta_0} = \frac{\hat{\beta} - \beta_0}{\hat{\sigma}_{\hat{\beta}}}$$

Lineare Einfachregression

Schätzen, Testen und Prognose

Hypothesen			Ablehnbereich
$H_0 : \alpha = \alpha_0$	vs.	$H_1 : \alpha \neq \alpha_0$	$\lvert T_{\alpha_0} \rvert > t_{1-\alpha/2}(n-2)$
$H_0 : \beta = \beta_0$	vs.	$H_1 : \beta \neq \beta_0$	$\lvert T_{\beta_0} \rvert > t_{1-\alpha/2}(n-2)$
$H_0 : \alpha \geq \alpha_0$	vs.	$H_1 : \alpha < \alpha_0$	$T_{\alpha_0} < -t_{1-\alpha}(n-2)$
$H_0 : \beta \geq \beta_0$	vs.	$H_1 : \beta < \beta_0$	$T_{\beta_0} < -t_{1-\alpha}(n-2)$
$H_0 : \alpha \leq \alpha_0$	vs.	$H_1 : \alpha > \alpha_0$	$T_{\alpha_0} > t_{1-\alpha}(n-2)$
$H_0 : \beta \leq \beta_0$	vs.	$H_1 : \beta > \beta_0$	$T_{\beta_0} > t_{1-\alpha}(n-2)$

Lineare Einfachregression

Schätzen, Testen und Prognose

- Prognose:

$$\hat{Y}_0 = \hat{\alpha} + \hat{\beta} x_0$$

mit Konfidenzintervall für Y_0:

$$\left[\hat{Y}_0 \pm t_{1-\alpha/2}(n-2)\, \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum x_i^2 - n\bar{x}^2}} \right]$$

Multiple lineare Regression

Einführung

Ziel: Erweiterung der linearen Einfachregression für mehrere Kovariablen $X_1, \ldots, X_p$

Daten: $(y_i, x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$

Zielvariable Y: metrisch bzw. stetig

Kovariablen: metrisch oder kategorial

- Metrische Kovariable x kann auch Transformation $x = f(z)$ einer ursprünglichen erklärenden Variablen z sein, z.B. $x = z^2$, $x = \log(z)$, usw.
- Kategorialer Regressor mit k Kategorien $1, \ldots, k$ durch $k - 1$ Dummy-Variablen $x^{(1)}, \ldots, x^{(k-1)}$ kodiert; mit k als Referenzkategorie.

Dummy-Kodierung

$$x^{(j)} = \begin{cases} 1, & \text{falls Kategorie } j \text{ vorliegt,} \\ 0, & \text{sonst,} \end{cases}$$

wobei $j = 1, \ldots, k-1$.

$x^{(1)} = \ldots = x^{(k-1)} = 0 \quad \Leftrightarrow \quad$ Referenzkategorie k liegt vor.

Multiple lineare Regression

Einführung

Standardmodell der linearen multiplen Regression

Es gilt

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i\,, \quad i = 1, \ldots, n\,.$$

Dabei sind

$Y_1, \ldots, Y_n$	beobachtbare metrische Zufallsvariablen,
$x_{1j}, \ldots, x_{nj}$	deterministische Werte der Variablen X_j oder Realisierungen von Zufallsvariablen X_j,
$\epsilon_1, \ldots, \epsilon_n$	unbeobachtbare Zufallsvariablen, die unabhängig und identisch verteilt sind mit $E(\epsilon_i) = 0$ und $Var(\epsilon_i) = \sigma^2$.

Bei Normalverteilungsannahme:

$$\epsilon_i \sim N(0, \sigma^2) \Leftrightarrow Y_i \mid x_{i1}, \ldots, x_{ip} \sim N(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, \sigma^2)$$

Multiple lineare Regression

Einführung

Matrixnotation

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \; X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}, \; \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \; \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Y Beobachtungsvektor der Zielvariablen, X Designmatrix

$Y = X\beta + \epsilon, \;\; E(\epsilon) = 0$; Annahme: Rang von $X = p + 1$

Schätzen, Testen und Prognose

Schätzer $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)'$ nach dem KQ-Prinzip

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 = (Y - X\beta)'(Y - X\beta) \to \min_{\beta}$$

Lösung: KQ-Schätzer

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Multiple lineare Regression

Schätzen, Testen und Prognose

Gefittete Werte:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip}$$

Residuen:

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i \ , \quad i = 1, \ldots, n.$$

Schätzer für die Varianz σ^2:

$$\hat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Multiple lineare Regression

Schätzen, Testen und Prognose

Erwartungstreue:

$$E(\hat{\beta}_j) = \beta_j, \;\; j = 0, \ldots, p; \;\; E(\hat{\sigma}^2) = \sigma^2$$

Varianz:

$$\sigma_j^2 := Var(\hat{\beta}_j) = \sigma^2 v_j; \;\; v_j \;\; j\text{-tes Diagonalelement von } (X'X)^{-1}$$

Geschätzte Varianz:

$$\hat{\sigma}_j^2 = \hat{\sigma}^2 v_j$$

Multiple lineare Regression

Schätzen, Testen und Prognose

Verteilung der standardisierten Schätzfunktionen:

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_j} \sim t(n-p-1) \ , \quad j = 0, \ldots, p$$

$(1-\alpha)$-Konfidenzintervalle für β_j:

$$\left[\hat{\beta}_j - \hat{\sigma}_j t_{1-\alpha/2}(n-p-1),\ \hat{\beta}_j + \hat{\sigma}_j t_{1-\alpha/2}(n-p-1)\right]$$

Multiple lineare Regression

Schätzen, Testen und Prognose

Einfache Teststatistiken:

$$T_j = \frac{\hat{\beta}_j - \beta_{0j}}{\hat{\sigma}_j} \ , \quad j = 0, \ldots, p$$

Hypothesen und Ablehnbereiche:

Hypothesen			Ablehnbereich
$H_0 : \beta_j = \beta_{0j}$	vs.	$H_1 : \beta_j \neq \beta_{0j}$	$\lvert T_j \rvert > t_{1-\frac{\alpha}{2}}(n-p-1)$
$H_0 : \beta_j \geq \beta_{0j}$	vs.	$H_1 : \beta_j < \beta_{0j}$	$T_j < -t_{1-\alpha}(n-p-1)$
$H_0 : \beta_j \leq \beta_{0j}$	vs.	$H_1 : \beta_j > \beta_{0j}$	$T_j > t_{1-\alpha}(n-p-1)$

Multiple lineare Regression

Schätzen, Testen und Prognose

Overall–F–Test:

- Hypothesen:

$$\begin{aligned} H_0: &\quad \beta_1 = \ldots = \beta_p = 0 \\ H_1: &\quad \beta_j \neq 0 \quad \text{für mindestens ein } j \end{aligned}$$

- Teststatistik:

$$F = \frac{R^2}{1-R^2}\,\frac{n-p-1}{p} = \frac{SQE}{SQR}\,\frac{n-p-1}{p}$$

- Ablehnungsbereich:

$$F > F_{1-\alpha}(p, n-p-1)$$

Multiple lineare Regression

Schätzen, Testen und Prognose

Prognose:

$$\hat{Y}_0 = x_0' \hat{\beta}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \ldots + \hat{\beta}_p x_{0p},$$

mit $x_0 = (1, x_{01}, \ldots, x_{0p})'$ als neuem Kovariablenvektor.

Ausblick

Varianzanalyse (ANOVA)

Situation: Alle unabhängigen Variablen sind kategorial, die Zielgröße Y ist metrisch/stetig.

- **Einfaktorielle Varianzanalyse**: Eine unabhängige Variable (Faktor) mit Stufen $i = 1, \ldots, I$.

  Modell:

  $$Y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, n_i,$$

  wobei $\epsilon_{ij} \sim N(0, \sigma^3)$.

  Frage: Unterscheidet sich der Erwartungswert von Y zwischen den Faktorstufen, d.h.

  $$\mu_1 = \mu_2 = \ldots = \mu_I \ ?$$

- **Mehrfaktorielle Varianzanalyse**: Betrachte nicht nur einen Faktor sondern mehrere.

Ausblick

Nichtlineare und nichtparametrische Regression

Nichtlineare parametrische Regression

Bisher: Regressionsmodell $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon$ linear in den Parametern $\beta_0, \ldots, \beta_p$ bzw. in $X_1, \ldots, X_p$.

Nichtlineares Modell:

$$Y = f(X_1, \ldots, X_p; \theta) + \epsilon$$

f nichtlinear, parametrisiert über θ.

Aber: Spezifikation einer parametrischen Regressionsfunktion $f(X; \theta)$ a priori oft schwierig.

Ausblick

Nichtlineare und nichtparametrische Regression

Nichtparametrische Regression

Nichtparametrische Regression flexibler als parametrische: Keine parametrische funktionale Form postuliert; nur qualitativ-strukturelle Annahmen.

Beispiel: Additives Modell

$$Y = f_1(X_1) + f_2(X_2) + \beta_1 Z_1 + \ldots + \beta_p Z_p + \epsilon$$

mit $f_1, f_2, \ldots$ als glatte, unbekannte Funktionen, die aus den Daten "nichtparametrisch" geschätzt werden.

Ausblick

Generalisierte Regression

Generalisierte Regressionsmodelle

Situation: Y ist nicht mehr normalverteilt, sondern z.B. binär.

- Lineares Modell wie bisher nicht mehr tauglich.
- Spezifiziere (generalisiertes lineares Modell)

$$E(Y|X_1 = x_1, \ldots, X_p = x_p) = h(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p),$$

mit bekannter "Responsefunktion" h.

Weitere Flexibilisierung z.B. durch generalisierte additive Modelle.