Vorlesung: Lineare Modelle

Prof. Dr. Helmut Küchenhoff

Institut für Statistik, LMU München

SoSe 2014

- Einführung und Beispiele
1. Das einfache lineare Regressionsmodell
- Das multiple lineare Regressionsmodell

Quadratsummenzerlegung und statistische Inferenz im multiplen linearen Regressionsmodell

Quadratsummenzerlegung

Gegeben sei das Modell (2.1) mit Design-Matrix X und

$$rg(X) = p + 1 =: p'. \tag{3.1}$$

Dann gilt:

$$\underbrace{(Y-\bar{Y})'(Y-\bar{Y})}_{\text{SST}} = \underbrace{(Y-\hat{Y})'(Y-\hat{Y})}_{\text{SSE}} + \underbrace{(\hat{Y}-\bar{Y})'(\hat{Y}-\bar{Y})}_{\text{SSM}} \tag{3.2}$$

Interpretation:

SST	:	Gesamt-Streuung, (korrigierte) Gesamt-Quadratsumme, "Total"
SSE	:	Fehler-Quadratsumme, "Error"
SSM	:	Modell-Quadratsumme, "Model"

Zerlegung ohne Absolutglied

Die Zerlegung (3.2) setzt ein Absolutglied in der Regression voraus.

Weitere Zerlegung, die nicht notwendig ein Absolutglied in der Regression voraussetzt:

$$\underbrace{Y'Y}_{\text{SST*}} = \underbrace{(Y-\hat{Y})'(Y-\hat{Y})}_{\text{SSE}} + \underbrace{\hat{Y}'\hat{Y}}_{\text{SSM*}} \tag{3.3}$$

SST^*	:	nicht korrigierte Gesamt-Quadratsumme Erfasst auch Abweichungen von $Y = 0$ und nicht nur von $\bar{Y}$.
SSE	:	Fehler-Quadratsumme, wie bei (3.2)
SSM^*	:	nicht korrigierte Modell-Quadratsumme

Nachweis der Quadratsummenzerlegung (3.3)

$$\begin{aligned} Y'Y &= Y'IY \\ &= Y'(Q+P)Y \\ &= Y'QY + Y'PY \\ &= Y'Q'QY + Y'P'PY \\ &= \hat{\varepsilon}'\hat{\varepsilon} + \hat{Y}'\hat{Y} \qquad \text{qed} \end{aligned}$$

P und Q Projektionsmatrizen.

Nachweis der Quadratsummenzerlegung (3.2) I

Wir definieren:

$$P_e = e(e'e)^{-1}e', \quad Q_e = I - P_e \text{ mit } e = (1,1,\ldots,1)'. \tag{3.4}$$

P_e entspricht der Regression auf eine Konstante ($E(Y) = \beta_0$)

$P_e Y =$	$\bar{Y}$	(„Mittelwert bilden“)
$Q_e Y =$	$Y - \bar{Y}$	(„Mittelwertsbereinigung“)

$$\begin{aligned} (Y-\bar{Y})'(Y-\bar{Y}) &= Y'Q_e'Q_eY = Y'Q_eY \\ &= Y'Q_e(Q+P)Y = Y'Q_eQY + Y'Q_ePY \\ &\overset{(*)}{=} \ldots \end{aligned}$$

Nachweis der Quadratsummenzerlegung (3.2) II

$$(*) \quad Q_e QY = Q_e\hat{\varepsilon} \underset{\substack{\uparrow \\ \text{„Residuen haben Mittelwert 0“}}}{=} \hat{\varepsilon} = QY$$

$$\begin{aligned}(*) \quad Q_e P Q_e PY &= Q_e P Q_e \hat{Y} = Q_e P(\hat{Y} - \bar{Y}) \\ = Q_e P\hat{Y} - Q_e P\bar{Y} &\underset{\substack{\uparrow \\ \text{„Regression auf } \hat{Y} \text{ und } \bar{Y} \text{ liefert Identität“}}}{=} Q_e\hat{Y} - Q_e\bar{Y} = Q_e\hat{Y} - 0\end{aligned}$$

$$\begin{aligned}(Y - \bar{Y})'(Y - \bar{Y}) &= Y'Q_e'Q_eY = Y'Q_eY \\ &= Y'Q_e(Q + P)Y = Y'Q_eQY + Y'Q_ePY \\ &\overset{(*)}{=} Y'QY + Y'Q_ePQ_ePY \\ &= \hat{\varepsilon}'\hat{\varepsilon} + Y'P'Q_ePY \\ &= (\hat{Y} - \bar{Y})'(\hat{Y} - \bar{Y})\end{aligned}$$

Erwartungswerte der Quadratsummen

Wir betrachten das multiple Regressionsmodell (2.1) mit (2.2) bis (2.4) und

$$P_e = e(e'e)^{-1}e', \quad Q_e = I - P_e \text{ mit } e = (1, 1, \ldots, 1)'.$$

Dann gilt für die Erwartungswerte der Quadratsummen:

$$\begin{aligned}E(SST^*) = E(Y'Y) &= \sigma^2 n + \beta'X'X\beta && (3.5) \\ E(SST) = E(Y - \bar{Y})'(Y - \bar{Y}) &= \sigma^2(n - 1) + \beta'(Q_eX)'(Q_eX)\beta && (3.6) \\ E(SSE) = E(\hat{\varepsilon}'\hat{\varepsilon}) &= \sigma^2(n - p') && (3.7) \\ E(SSM^*) = E(\hat{Y}'\hat{Y}) &= \sigma^2 p' + \beta'X'X\beta && (3.8) \\ E(SSM) = E(\hat{Y} - \bar{Y})'(\hat{Y} - \bar{Y}) &= \sigma^2 p + \beta'(Q_eX)'(Q_eX)\beta && (3.9)\end{aligned}$$

Bemerkungen

1. Die stochastischen Eigenschaften der Quadratsummen werden zur Konstruktion von Tests bezüglich β genutzt
2. $\beta = 0 \Longrightarrow E(Y'Y) = \sigma^2 n$
3. $\beta_1, \ldots\beta_p = 0 \Longrightarrow E(SSM) = \sigma^2 p$

Zum Nachweis von 3. benutzt man, dass Q_e der „Mittelwertsbereinigungs- Operator“ ist:

$$Q_e x = (I - P_e)x = x - \bar{x}$$

Die erste Spalte von Q_eX ist also der Nullvektor.

Beweisidee

Allgemein berechnet man den Erwartungswert von quadratischen Formen wie folgt:

$$\begin{aligned}E(Y'AY) &= E\left(\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}Y_iY_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}E(Y_iY_j) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}E(Y_i)E(Y_j) + \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}Cov(Y_i, Y_j) \\ &= E(Y)'AE(Y) + Sp(AV(Y)) \\ V(Y) &:= \text{Varianz-Kovarianzmatrix von Y}\end{aligned}$$

Unter Benutzung von $Sp(AV(Y)) = sp(A) * \sigma^2 = rang(A) * \sigma^2$ für Projektionsmatrizen erhält man obige Identitäten.

Mittlere Quadratsummen

Wir definieren entsprechend der Zahl der Freiheitsgrade die **mittleren Quadratsummen**:

$$MST^* := \frac{SST^*}{n} \quad (3.10)$$

$$MST := \frac{SST}{n-1} \quad (3.11)$$

$$MSE := \frac{SSE}{n-p'} \quad (3.12)$$

$$MSM^* := \frac{SSM^*}{p'} \quad (3.13)$$

$$MSM := \frac{SSM}{p} \quad (3.14)$$

Verteilungsdefinitionen

Normalverteilung

Ein n-dimensionaler Zufallsvektor Z heißt multivariat normalverteilt, falls für seine Dichtefunktion gilt:

$$f_z(z) = \frac{1}{\sqrt{\det\Sigma}\sqrt{(2\pi)^n}} \exp\left[-\frac{1}{2}(z-\mu)'\Sigma^{-1}(z-\mu)\right] \quad (3.15)$$

mit positiv definiter, symmetrischer Matrix Σ.

Bezeichnung: $Z \sim N_n(\mu, \Sigma)$

Eigenschaften der Normalverteilung

Ist $Z \sim N_n(\mu, \Sigma)$, so gilt:

1. **Momente:**

$$E(Z) = \mu$$
$$V(Z) = \Sigma$$

2. **Lineare Transformationen:**
Ist $A : R^n \longrightarrow R^m$ eine lineare Tranformation mit $rg(A) = m$

$$\Longrightarrow AZ \sim N_m(A\mu, A\Sigma A') \quad (3.16)$$

3. **Orthogonale Transformation in unabhängige Komponenten (Spektralzerlegung)**
Es existiert eine Matrix $T \in R^{n\times n}$ mit $T'T = I$ und
$T\Sigma T' =$ diag $(\lambda_1, \dots, \lambda_n)$, so dass

$$TZ \sim N_n(T\mu, \text{ diag } (\lambda_1, \dots, \lambda_n)) \text{ gilt.} \quad (3.17)$$

Chi-Quadrat-Verteilung I

Ist $Z \sim N_n(\mu, I)$ so heißt $X = Z'Z$ (nicht-zentral) Chi-Quadrat-verteilt.

Bezeichnung: $X \sim \chi^2(n, \delta)$

n:	Zahl der Freiheitsgrade
$\delta := \mu'\mu$:	Nichtzentralitätsparameter

Im Fall $\delta = 0$ erhält man die zentrale $\chi^2(n)$-Verteilung.

Chi-Quadrat-Verteilung II

Eigenschaften: Ist $X \sim \chi^2(n, \delta)$, so gilt:

1. **Momente:**

$$E(X) = n + \delta$$
$$V(X) = 2n + 4\delta$$

2. **Allgemeiner Bezug zur Normalverteilung**

$$Z \sim N_n(\mu, \Sigma) \implies Z'\Sigma^{-1}Z \sim \chi^2(n, \mu'\Sigma^{-1}\mu) \tag{3.18}$$

t-Verteilung

Seien Z und W voneinander unabhängige Zufallsgrößen mit

$$Z \sim N(\delta, 1),$$
$$W \sim \chi^2(n).$$

Dann heißt $X = \frac{Z}{\sqrt{\frac{W}{n}}}$ (nicht-zentral) t-verteilt.

Bezeichnung: $X \sim t(n, \delta)$

n: Zahl der Freiheitsgrade,

δ: Nicht-Zentralitätsparameter

Erwartungswert: Ist $X \sim t(n, \delta)$,so gilt:

$$E(X) = \delta \sqrt{\frac{n}{2}} \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})} \quad \text{für } n > 1. \tag{3.19}$$

F-Verteilung

Seien W_1 und W_2 voneinander unabhängige Zufallsgrößen mit

$$W_1 \sim \chi^2(n_1, \delta)$$
$$W_2 \sim \chi^2(n_2)$$

Dann heißt $X = \frac{W_1/n_1}{W_2/n_2}$ (nicht-zentral) F-verteilt.

Bezeichnung: $X \sim F(n_1, n_2, \delta)$

n_1: Zählerfreiheitsgrade

n_2: Nennerfreiheitsgrade

δ: Nichtzentralitätsparameter

Erwartungswert: Ist $X \sim F(n_1, n_2, \delta)$, so gilt:

$$E(X) = \frac{n_2}{n_2 - 2}(1 + \delta/n_1) \text{ für } n_2 > 2. \tag{3.20}$$

Unabhängigkeit von Quadratsummen (Satz von Cochran)

Sei $Z \sim N(\mu, I)$, dim Z = n,
A $\in R^{n \times n}$, $A' = A$, $rg(A) = r$, $A^2 = A$,
$B \in R^{n \times n}$, $B^2 = B$, $B = B'$,
$C \in R^{m \times n}$.

Dann gilt:

$$Z'AZ \sim \chi^2(r, \mu'A\mu) \tag{3.21}$$
$$CA = 0 \implies CZ \text{ und } Z'AZ \text{ sind unabh.} \tag{3.22}$$
$$AB = 0 \implies Z'AZ \text{ und } Z'BZ \text{ sind unabh.} \tag{3.23}$$

Beweis von (3.21)

Spektralzerlegung von A (3.17)

$A = T\Sigma T'$

$$\Sigma = \underbrace{\begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & & 0 & \\ & & & & 0 \end{pmatrix}}_{r} \qquad \text{da } A^2 = A$$

$\Rightarrow Z'AZ = Z'T\Sigma T'Z$

$Var(T'Z) = T'IT = I$

$\Rightarrow Z'AZ = \widetilde{Z}\Sigma Z = \sum_{i=1}^{r} \widetilde{z}_i^2 \sim \chi^2(r, \mu' A\mu)$