Methodenlehre II: Regression

Vorlesung WS 15/16

Prof. Dr. Rolf Steyer

Bedingter Erwartungswert

Definition 1 *Ist Y eine reellwertige Zufallsvariable mit endlich vielen Werten $y_1, \ldots, y_n$ und ist $P(X{=}x) > 0$, dann ist der* bedingte Erwartungswert *von Y gegeben $X{=}x$ die mit den bedingten Wahrscheinlichkeiten $P(Y{=}y_i \,|\, X{=}x)$ gewichtete Summe ihrer Werte:*

$$E(Y \,|\, X{=}x) \;:=\; \sum_{i=1}^{n} y_i \cdot P(Y{=}y_i \,|\, X{=}x).$$

Kann Y nur die beiden Werte 1 *und* 0 *annehmen, dann folgt aus obiger Gleichung:*

$$\begin{aligned} E(Y \,|\, X{=}x) &= 1 \cdot P(Y{=}1 \,|\, X{=}x) + 0 \cdot P(Y{=}0 \,|\, X{=}x) \\ &= P(Y{=}1 \,|\, X{=}x) \end{aligned}$$

Bedingter Erwartungswert: Rechenregeln

Rechenregeln Sind X und Z diskrete Zufallsvariablen und Y sowie Y_1 und Y_2 numerische Zufallsvariablen auf $\langle \Omega, \mathscr{A}, P \rangle$ mit endlichen Erwartungswerten sowie α und β reelle Zahlen, dann gelten:

$$\begin{aligned} E(\alpha \,|\, X{=}x) &= \alpha \\ E(\alpha \cdot Y \,|\, X{=}x) &= \alpha \cdot E(Y \,|\, X{=}x) \\ E(\alpha \cdot Y_1 + \beta \cdot Y_2 \,|\, X{=}x) &= \alpha \cdot E(Y_1 \,|\, X{=}x) + \beta \cdot E(Y_2 \,|\, X{=}x) \\ E(Y \,|\, X{=}x) &= \sum_{z} E(Y \,|\, X{=}x, Z{=}z) \cdot P(Z{=}z \,|\, X{=}x) \end{aligned}$$

Regression

Definition 2 *Die Regression $E(Y \mid X)$ kann als diejenige Funktion von X definiert werden, deren Werte die* bedingten Erwartungswerte *$E(Y \mid X{=}x)$ von Y gegeben $X{=}x$ sind. Dabei kann der Regressor X durchaus beliebige Werte annehmen, die nicht unbedingt Zahlen sein müssen.*

Joe and Ann With Random Assignment

Table 1: Joe and Ann With Random Assignment

Elements of Ω		Random variables			Conditional Expectations		
Unit Treatment Success	Probabilities of elementary events $P(\{\omega_i\})$	Observational-unit variable U	Treatment variable X	Outcome variable Y	$E(Y \mid X, U)$	$E(Y \mid X)$	Conditional treatment probability $P(X{=}1 \mid U)$
ω_1 = (*Joe, no, −*)	.09	*Joe*	0	0	.70	.45	.40
ω_2 = (*Joe, no, +*)	.21	*Joe*	0	1	.70	.45	.40
ω_3 = (*Joe, yes, −*)	.04	*Joe*	1	0	.80	.60	.40
ω_4 = (*Joe, yes, +*)	.16	*Joe*	1	1	.80	.60	.40
ω_5 = (*Ann, no, −*)	.24	*Ann*	0	0	.20	.45	.40
ω_6 = (*Ann, no, +*)	.06	*Ann*	0	1	.20	.45	.40
ω_7 = (*Ann, yes, −*)	.12	*Ann*	1	0	.40	.60	.40
ω_8 = (*Ann, yes, +*)	.08	*Ann*	1	1	.40	.60	.40

Joe and Ann With Self-Selection to Treatment

Table 2: Joe and Ann With Self-Selection to Treatment

Outcomes ω				Observables			Conditional Expectations				Residuals	
Unit	Treatment	Success	Probabilities of elementary events $P(\{\omega\})$	Observational-unit variable U	Treatment variable X	Outcome variable Y	$E(Y\|X,U)$	$E(Y\|X)$	$E(Y\|U)$	Conditional treatment probability $P(X=1\|U)$	$\varepsilon_Y = Y - E(Y\|U)$	$\varepsilon_X = X - P(X=1\|U)$
Joe	*no*	−	.144	*Joe*	0	0	.70	.60	.704	.04	−.704	−.04
Joe	*no*	+	.336	*Joe*	0	1	.70	.60	.704	.04	.296	−.04
Joe	*yes*	−	.004	*Joe*	1	0	.80	.42	.704	.04	−.704	.96
Joe	*yes*	+	.016	*Joe*	1	1	.80	.42	.704	.04	.296	.96
Ann	*no*	−	.096	*Ann*	0	0	.20	.60	.352	.76	−.352	−.76
Ann	*no*	+	.024	*Ann*	0	1	.20	.60	.352	.76	.648	−.76
Ann	*yes*	−	.228	*Ann*	1	0	.40	.42	.352	.76	−.352	.24
Ann	*yes*	+	.152	*Ann*	1	1	.40	.42	.352	.76	.648	.24

Note. The probabilities of the elementary events are fictive

Regression: Rechenregeln

Rechenregeln für Regressionen

(i) $E(\alpha \mid X) = \alpha, \qquad \alpha \in \mathbb{R}$

(ii) $E(\alpha\, Y \mid X) = \alpha\, E(Y \mid X), \qquad \alpha \in \mathbb{R}$

(iii) $E(\alpha\, Y_1 + \beta\, Y_2 \mid X) = \alpha\, E(Y_1 \mid X) + \beta\, E(Y_2 \mid X), \qquad \alpha, \beta \in \mathbb{R}$

(iv) $E[E(Y \mid X)] = E(Y)$

(v) $E[f(X) \mid X] = f(X), \qquad$ falls $f(X)$ numerisch ist

(vi) $E[E(Y \mid X) \mid f(X)] = E[Y \mid f(X)]$

(vii) $E[f(X) \cdot Y \mid X] = f(X) \cdot E(Y \mid X), \qquad$ falls $f(X)$ numerisch ist

Residuum

Definition 3 *Das* Residuum ε *bezüglich einer Regression* $E(Y \mid X)$ *ist definiert als* Abweichung der Zufallsvariablen Y *von ihrer Regression* $E(Y \mid X)$ *auf* X.

In Formeln: $\varepsilon := Y - E(Y \mid X)$

Residuum

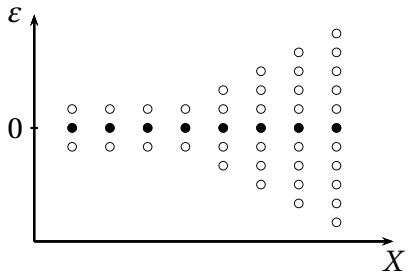

Table 3: Die Regressionen von Y und ε auf einen numerischen Regressor X. In diesem Beispiel sind die bedingten Varianzen des Residuums von X abhängig. Das Zeichen markiert die Werte der Regression $E(\varepsilon \mid X)$ und die Werte des Residuums ε.

Eigenschaften des Residuums

Box 6.3. Die wichtigsten Eigenschaften des Residuums

(i) $Y = E(Y \mid X) + \varepsilon$

(ii) $E(\varepsilon) = 0$

(iii) $E(\varepsilon \mid X) = 0$

(iv) $E[\varepsilon \mid f(X)] = 0$

(v) $E[\varepsilon \mid E(Y \mid X)] = 0$

(vi) $Cov(\varepsilon, X) = 0$, falls X numerisch ist

(vii) $Cov(\varepsilon, X_i) = 0$, $i = 1, \ldots, m$, falls $X = (X_1, \ldots, X_m)$ numerisch ist

(viii) $Cov[\varepsilon, f(X)] = 0$, falls $f(X)$ numerisch ist

(ix) $Cov[\varepsilon, E(Y \mid X)] = 0$

(x) $Var(Y) = Var[E(Y \mid X)] + Var(\varepsilon)$

Beweis von Eigenschaft (vi) des Residuums

$$
\begin{aligned}
Cov(\varepsilon, X) &= E(\varepsilon \cdot X) - E(\varepsilon) \cdot E(X) && \text{[Regel (i) für Kovarianzen]} \\
&= E(\varepsilon \cdot X) && \text{[Regel (ii) für Residuen]} \\
&= E[E(\varepsilon \cdot X \mid X)] && \text{[Regel (iv) für Regressionen]} \\
&= E[X \cdot E(\varepsilon \mid X)] && \text{[Regel (vii) für Regressionen]} \\
&= E(X \cdot 0) && \text{[Regel (iii) für Residuen]} \\
&= 0 \cdot E(X) && \text{[Regel (ii) für Erwartungswerte]} \\
&= 0
\end{aligned}
$$

Determinationskoeffizient

Ein Begriff, der unmittelbar auf den oben behandelten Eigenschaften des Residuums ε [insb. auf Regel (x)] basiert, ist der des Determinationskoeffizienten, der für jede numerische Zufallsvariable Y mit endlichem Erwartungswert $E(Y)$ und endlicher Varianz $Var(Y)$ definiert ist, und zwar durch:

$$R^2_{Y|X} = \frac{Var[E(Y \mid X)]}{Var(Y)}, \quad \text{falls } Var(Y) > 0$$

Determinationskoeffizient (fortgesetzt)

Der Determinationskoeffizient lässt sich als der durch X determinierte Varianzanteil von Y interpretieren. Wie man sehen kann, addiert er sich mit dem Residualvarianzanteil von Y zu 1 auf, falls $Var(Y) > 0$:

$$1 = \frac{Var(Y)}{Var(Y)} = \frac{Var[E(Y \mid X)]}{Var(Y)} + \frac{Var(\varepsilon)}{Var(Y)}$$

Multiple Korrelation

Die Wurzel aus dem Determinationskoeffizienten heißt *multiple Korrelation.* Für diese gilt auch:

$$R_{Y|X} \;=\; Corr\,[Y, E(Y\,|\,X)]$$

Der Regressionsbegriff bei diskretem Regressor

Definition 4 *Seien* $(\Omega, \mathscr{A}, P)$ *ein Wahrscheinlichkeitsraum,* Y *eine numerische Zufallsvariable auf* $(\Omega, \mathscr{A}, P)$ *mit endlichem Erwartungswert [d. h.* $-\infty < E(Y) < +\infty$*] und* X *eine diskrete Zufallsvariable auf* $(\Omega, \mathscr{A}, P)$ *mit einer endlichen Anzahl von Werten* $x_1, \ldots, x_n$*, für die jeweils* $P(X = x_i) > 0$ *gilt. Dann heißt die Zufallsvariable*

$$E(Y\,|\,X) \;:=\; \sum_{i=1}^{n} E(Y\,|\,X{=}x_i) \cdot I_{X=x_i}$$

Regression von Y *auf* X *oder* bedingte Erwartung *von* Y *unter* X.