Stichprobenverfahren

Modellbasierte Stichprobenverfahren

Willi Mutschler (willi@mutschler.eu)

Sommersemester 2017

Motivation (1)

Wählerumfrage

- Bei Wählerumfrage ist es naheliegend, Informationen von der letzten Wahl zu nutzen, da Ergebnisse hierbei bekannt sind. Typischerweise stellt man neben der aktuellen Wahlabsicht auch die Frage nach dem Verhalten bei der letzten Wahl.
- Wichtige Überlegung: Großteil der Wähler bleibt einer Partei treu
- Annahme: in der gezogenen Stichprobe geben 40% der Befragten an, bei der letzten Wahl die Partei ABC gewählt zu haben, während diese Partei jedoch bei der letzten Wahl tatsächlich nur 35% der Stimmen erhalten hat, d.h. Menge der Individuen, die bei der letzten Wahl für Partei ABC gestimmt hat, ist überrepräsentiert. Überproportionale Anteil der Wähler von Partei ABC in der Stichprobe lässt vermuten, dass die Partei ABC auch bei der aktuellen Wahlabsicht in der Stichprobe besser abschneidet als in der Grundgesamtheit.
- Annahme: in der Stichprobe haben 44% ihre aktuelle Präferenz für Partei ABC angegeben. Rein intuitiv sofort Zweifel, ob die Partei ABC bei der nächsten Wahl tatsächlich mit 44% abschneiden wird. Ist dies wirklich Überschätzung?
- Es gibt nun verschiedene Möglichkeiten, dieses Ergebnis zu korrigieren.
    - Korrigiere das Ergebnis von 44 auf 39%, da Anteil um 5 Prozentpunkte überschätzt wird
    - Partei hat in der Stichprobe ihre Anhängerschaft um 10% (von 40 auf 44%) gesteigert. Übertragung auf Grundgesamtheit gibt Schätzung von 38, 5%

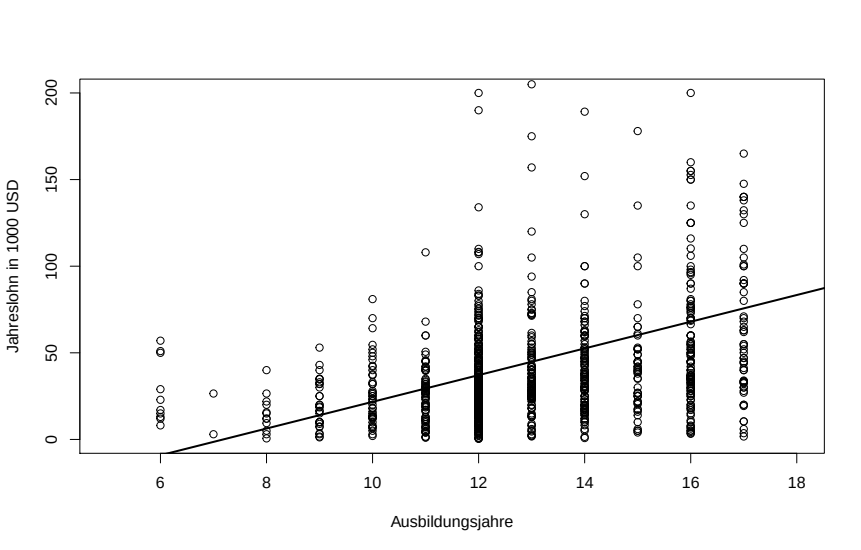
Motivation (2)

Beispiel: Marktanalyse

Ein Unternehmen plant, ein neues Produkt einzuführen und möchte hierzu eine Marktanalyse vornehmen. Es wird vermutet, dass das Produkt in verschiedenen Altersgruppen unterschiedlich angenommen wird. Ferner scheint es plausibel, dass Frauen dem Produkt anders gegenüberstehen als Männer. Das Unternehmen besitzt Sekundärinformationen über die Population (beispielsweise aus statistischen Jahrbüchern). Insbesondere ist die Altersverteilung je Geschlecht bekannt. Das Unternehmen wählt Individuen aus der Population zufällig aus (einfache Zufallsstichprobe) und befragt diese nach der Produktakzeptanz. Ist nun die Alters- oder Geschlechtsstruktur in der Stichprobe, bedingt durch die zufällige Auswahl der Individuen, anders als in der Population, so kann und sollte das Stichprobenergebnis diesbezüglich korrigiert werden. Liegt zum Beispiel der Frauenanteil in der Stichprobe unter dem Frauenanteil in der Bevölkerung und sind Frauen im Mittel dem Produkt mehr abgeneigt als Männer, so sollte das Stichprobenergebnis korrigiert werden.

Motivation (3)

- Ausnutzung von zusätzlichen Informationen: Verbesserung der Effizienz sowohl eines Stichproben Designs als auch von Schätzfunktionen möglich
- Beispiel: PSID Datensatz
  - Schätzung Durchschnittlichen Jahreslohns $\bar{Y}_U$ in der Grundgesamtheit mit zusätzliche Informationen über die Anzahl an Ausbildungsjahre X_k für alle $k \in U$



  - Korrelation zwischen Y und X ist 0.2866
  - Diese Information kann z.B. genutzt werden um
    - anhand von Ausbildungsjahren zu schichten
    - Einschlusswahrscheinlichkeiten in Abhängigkeit von Ausbildungsjahren zu definieren
    - die Schätzung effizienter zu machen

Motivation (4)

- Sekundärinformation kann sinnvoll eingesetzt werden, wenn sie folgende Eigenschaften aufweist:
  - Sekundärinformation X steht in einem engen Zusammenhang zur Primärinformation und interessierenden Größe Y
  - Von der Sekundärinformation X ist die Merkmalssumme oder der Mittelwert in der Population bekannt
- Basierend auf einer Stichprobe, bekommen wir Schätzwerte für z.B. die Merkmalssummen von X, $\hat{t}_x$, und von Y, $\hat{t}_y$
- Intuition: Wenn wir in der Stichprobe Schätzfehler (Über- oder Unterschätzung) bei $\hat{t}_x$ machen, so werden diese auch bei $\hat{t}_y$ auftreten
- Der Vergleich mit dem bekannten wahren Wert t_x gibt uns prinzipiell drei Möglichkeiten den Schätzer für Y zu verbessern:
  - Quotientenschätzer vermutet einen proportionalen Relation zwischen X und Y
  - Differenzenschätzer vermutet einen konstanten Shift, unabhängig vom Level von X
  - Regressionsschätzer vermutet einen (nicht)linearen Zusammenhang zwischen X und Y
- Allgemein benutzen wir ein statistisches Modell für den Zusammenhang zwischen X und Y

Quotientenschätzung (1)

- Der Quotient zweier Merkmalssummen in U ist

$$r = \frac{t_y}{t_x}$$

- t_x ist bekannt, aber t_y und somit auch r sind unbekannt
- Basierend auf einer Stichprobe bekommen wir $\hat{t}_y$ und $\hat{t}_x$, folglich:

$$\hat{r} = \frac{\hat{t}_y}{\hat{t}_x}$$

- Für den Quotientenschätzer von t_y gilt nun:

$$\hat{t}_y^{ra} = \hat{r} t_x = \hat{t}_y \frac{t_x}{\hat{t}_x} = \sum_s \frac{y_k}{\pi_k} \frac{t_x}{\hat{t}_x}$$

- Korrektur des freien Schätzers $\hat{t}_y$ um Über- oder Unterschätzung von t_x durch $\hat{t}_x$
- Idee: Wenn X und Y korreliert sind und t_x in der Stichprobe überschätzt wird, so gilt dies auch für t_y
- $\hat{t}_y^{ra}$ ist allerdings nur asymptotisch erwartungstreu und die Varianz lässt sich auch nur approximativ berechnen

Quotientenschätzung (2)

- Der Quotientenschätzer ist eine Funktion von zwei Variablen $\hat{t}_y$ und $\hat{t}_x$
- Approximiere $f(\hat{t}_y, \hat{t}_x) = \hat{t}_y / \hat{t}_x$ mithilfe einer Taylor Approximation:

$$\frac{\hat{t}_y}{\hat{t}_x} \approx \frac{t_y}{t_x} + \frac{1}{t_x}(\hat{t}_y - r\hat{t}_x)$$

- Einsetzen in die Quotientenschätzfunktion

$$\hat{t}_y^{ra} = t_y + \hat{t}_y - r\hat{t}_x$$

- $\hat{t}_y^{ra}$ ist daher approximativ erwartungstreu, da

$$E(t_y + \hat{t}_y - r\hat{t}_x) = t_y + t_y - \frac{t_y}{t_x} t_x = t_y$$

- Die approximierte Varianz ist dann

$$AV(\hat{t}_y^{ra}) = V(\hat{t}_y) + r^2 V(\hat{t}_x) - 2rCov(\hat{t}_y, \hat{t}_x)$$

Quotientenschätzung (3)

- Quotientenschätzer bei einfacher Zufallsauswahl:

$$AV(\hat{t}_y^{ra}) = N^2 \frac{1}{n} \left(1 - \frac{n}{N}\right) \left(S_{y_U}^2 + r^2 S_{x_U}^2 - 2rS_{xy_U}\right)$$

- Schätzung basierend auf einer Stichprobe:

$$\hat{AV}(\hat{t}_y^{ra}) = N^2 \frac{1}{n} \left(1 - \frac{n}{N}\right) \left(S_{y_s}^2 + \hat{r}^2 S_{x_s}^2 - 2\hat{r}S_{xy_s}\right)$$

- Effizienter als eine einfache Zufallsauswahl, falls $r^2 S_{x_U}^2 - 2rS_{xy_U} < 0$ oder

$$\rho_{xy} > \frac{1}{2} \frac{CV_{x_U}}{CV_{y_U}}$$

Differenzenschätzung (1)

- Bei der Differenzenschätzung nehmen wir an, dass folgender Zusammenhang gilt

$$y_k - x_k = \beta + \varepsilon_k$$

- Der Differenzenschätzer ist

$$\hat{t}_y^{diff} = \hat{t}_y + (t_x - \hat{t}_x)$$

  bei einfacher Zufallsauswahl

$$\hat{t}_y^{diff} = N\bar{y}_s + N(\bar{x}_U - \bar{x}_s)$$

- Der freie Schätzer $\hat{t}_y$ wird korrigiert, wenn die Stichprobe „zu große“ ($t_x - \hat{t}_x < 0$) bzw. „zu kleine“ ($t_x - \hat{t}_x > 0$) Merkmalsträger hat

Differenzenschätzung (2)

- Der Differenzenschätzer ist erwartungstreu:

$$E(\hat{t}_y^{diff}) = E(\hat{t}_y) + E(t_x) - E(\hat{t}_x) = t_y + t_x - t_x = t_y$$

- Die Varianz (bei einfacher Zufallsauswahl) ist

$$V(\hat{t}_y^{diff}) = N^2 \frac{1}{n} \left(1 - \frac{1}{N}\right) (S_{y_U}^2 + S_{x_U}^2 - 2S_{xy_U})$$

  und kann geschätzt werden mit

$$\hat{V}(\hat{t}_y^{diff}) = N^2 \left(\frac{1}{n} - \frac{1}{N}\right) (S_{y_s}^2 + S_{x_s}^2 - 2S_{xy_s})$$

- Effizienter als eine einfache Zufallsauswahl, falls $S_{x_U}^2 - 2S_{xy_U} < 0$ oder

$$\rho_{xy} > \frac{1}{2} \frac{S_{x_U}}{S_{y_U}}$$

Regressionsschätzung (1)

- Regressionszusammenhang als Approximation für Y: $y_k = B_1 + B_2 x_k + E_k$
- y_k und x_k sind für alle N Elemente in U fix
- Notation:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}, y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}$$

- Das Kleinste-Quadrate-Kriterium ist $\sum_U (y_k - x_k' B)^2$, die zugehörigen Parameter

$$B_2 = \frac{N\left(\sum_U x_k y_k\right) - \left(\sum_U x_k\right)\left(\sum_U y_k\right)}{N\left(\sum_U x_k^2\right) - \left(\sum_U x_k\right)^2} = \frac{\sum_U (x_k - \bar{x}_U)(y_k - \bar{y}_U)}{\sum_U (x_k - \bar{x}_U)^2}$$

$$B_1 = \bar{y}_U - B_2 \bar{x}_U$$

- Der π-Schätzer basierend auf einer Stichprobe s ist definiert als

$$\hat{B}_2 = \frac{\left(\sum_s \frac{1}{\pi_k}\right)\left(\sum_s \frac{x_k y_k}{\pi_k}\right) - \left(\sum_s \frac{x_k}{\pi_k}\right)\left(\sum_s \frac{y_k}{\pi_k}\right)}{\left(\sum_s \frac{1}{\pi_k}\right)\left(\sum_s \frac{x_k^2}{\pi_k}\right) - \left(\sum_s \frac{x_k}{\pi_k}\right)^2} = \frac{\sum_s \frac{(x_k - \tilde{x}_s)(y_k - \tilde{y}_s)}{\pi_k}}{\sum_s \frac{(x_k - \tilde{x}_s)^2}{\pi_k}}$$

$$\hat{B}_1 = \tilde{y}_s - \hat{B}_2 \tilde{x}_s$$

mit $\tilde{y}_s = \frac{\hat{t}_{y,\pi}}{\hat{N}}$ und $\tilde{x}_s = \frac{\hat{t}_{x,\pi}}{\hat{N}}$

- Bei bekannter Populationsgröße N verwendet man $\tilde{y}_s = \hat{\bar{y}}$ und $\tilde{x}_s = \hat{\bar{x}}$

Regressionsschätzung (2)

- Die Schätzfunktion von $\hat{B}$ basiert auf Quotienten von Zufallsgrößen, daher kein analytischer Ausdruck für Varianzen möglich
- Lösung: Taylor-Approximation 1. Ordnung:

$$AV(\hat{B}_2) = \frac{\sum\sum_U \Delta_{kl} \frac{(x_k - \bar{x}_U)E_k}{\pi_k} \frac{(x_l - \bar{x}_U)E_l}{\pi_l}}{\left[\sum_U (x_k - \bar{x}_U)^2\right]^2}$$

  mit $E_k = y_k - B_1 - B_2 x_k$

- Dies kann geschätzt werden basierend auf einer Stichprobe:

$$\hat{AV}(\hat{B}_2) = \frac{\sum\sum_s \breve{\Delta}_{kl} \frac{(x_k - \tilde{x}_s)e_k}{\pi_k} \frac{(x_l - \tilde{x}_s)e_l}{\pi_l}}{\left[\sum_s \frac{(x_k - \tilde{x}_s)^2}{\pi_k}\right]^2}$$

  mit $e_k = y_k - \hat{B}_1 - \hat{B}_2 x_k$

Regressionsschätzung (3)

- Im Fall der einfachen Zufallsauswahl vereinfacht sich hier einiges:

$$\hat{B}_2 = \frac{\sum_s (x_k - \bar{x}_s)(y_k - \bar{y}_s)}{\sum_s (x_k - \bar{x}_s)^2}$$

  und $\hat{B}_1 = \bar{y}_s - \hat{B}_2 \bar{x}_s$

- Die approximierte Varianz ist

$$AV(\hat{B}_2) = \frac{N^2 \frac{1-f}{n} \frac{1}{N-1} \sum_U (x_k - \bar{x}_U)^2 E_k^2}{[\sum_U (x_k - \bar{x}_U)^2]^2}$$

  und schätzbar basierend auf einer Stichprobe mit

$$\hat{AV}(\hat{B}_2) = \frac{(1-f) \frac{n}{n-1} \sum_s (x_k - \bar{x}_s)^2 e_k^2}{[\sum_s (x_k - \bar{x}_s)^2]^2}$$

- Achtung: im stochastischen Regressionsmodel ist die Varianz des Steigungsparameter $\frac{\frac{1}{n-2} \sum_s e_k^2}{\sum_s (x_k - \bar{x}_s)^2}$ unterschiedlich groß

Fazit

- Modellbasierte Stichprobenverfahren bauen auf Modellen auf, die den Einfluss der Sekundärinformation X auf Y beschreiben.
- Das globale Modell ist dabei ein Regressionsmodell, d.h. Y ergibt sich als lineare Approximation von X
- Die zu Grunde liegenden Modelle der einzelnen Schätzer lassen sich somit wie folgt schreiben:
  - Regressionsschätzer: $y_k = B_1 + B_2 x_k$
  - Quotientenschätzer: $y_k = B_1 + B_2 x_k$ mit $B_1 = 0$
  - Differenzenschätzer : $y_k = B_1 + B_2 x_k$ mit $B_2 = 1$
- Es ist zu beachten, dass die Herleitung der Schätzverfahren nicht auf der Gültigkeit eines linearen Regressionsmodells als datengenerierendem Prozess basiert, sondern nur die Regressionsgerade der Grundgesamtheit als Hilfsmittel benutzt.
- Lineare Regressionsmodell wird hier nicht mit den üblichen Modellannahmen verwendet wird, sonder für die „modellunterstützte (model assisted)“ Schätzung.