Proseminar Wissenschaftliches Arbeiten: Ökonometrie und Statistik

Woche 4: Simulationen

Elias Wolf

30. April 2023

Bewertung eines geschätzten Modells

- Beobachtungen $(Y_t, X_{1,t}, \ldots, X_{k,t})$ für $t = 1, \ldots, T$
- Geschätzte/angepasste/prognostizierte Werte:

$$\widehat{Y}_t = \widehat{f}(X_{1,t}, \ldots, X_{k,t}), \quad t = 1, \ldots, T$$

- Im Beispiel einer linearen Regression:

$$\widehat{f}(X_{1,t}, \ldots, X_{k,t}) = \widehat{\beta}_0 + \widehat{\beta}_1 X_{1,t} + \ldots + \widehat{\beta}_k X_{k,t},$$

  wobei $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \ldots, \widehat{\beta}_k)' = \left(\sum_{t=1}^T \boldsymbol{x}_t \boldsymbol{x}_t'\right)^{-1} \sum_{t=1}^T (Y_t \cdot \boldsymbol{x}_t)$ mit $\boldsymbol{x}_t = (1, X_{1,t}, \ldots, X_{k,t})'$.
- Fragestellungen:
  - Wie gut ist meine Parameterschätzung $\widehat{\boldsymbol{\beta}}$?
    $\rightarrow$ Vergleich mit „wahrem“ Parameter $\boldsymbol{\beta}$ (der in der Praxis jedoch unbekannt ist)
  - Wie gut ist der Fit? Wie gut ist die Prognose? Mean Squarred Error:

$$MSE = \frac{1}{T} \sum_{t=1}^{T} (\widehat{Y}_t - Y_t)^2$$

Bewertung eines geschätzten Modells

Zwei Ansätze um den Fit/die Prognosegüte von Modellen zu vergleichen:

1.) In-Sample Fit: Wir schätzen und bewerten das Modell mit demselben Datensatz (Training-MSE):

$$MSE = \frac{1}{T} \sum_{t=1}^{T} \left(Y_t - \widehat{f}(X_{1,t}, \ldots, X_{k,t}) \right)^2$$

- Je besser der Fit desto niedriger der MSE
- Problem Overfitting: Je komplexer das Modell (je mehr Parameter) desto besser der Fit im Datensatz
- Für lineare Regression: Besser adjusted R^2 oder Informationskriterien AIC und BIC verwenden

Güte eines geschätzten Modells

2.) Out-of-Sample Fit:

- Die Beobachtungen werden in ein **Trainingssample** und ein **Evaluationssample** aufgeteilt:

$$
\begin{aligned}
&(Y_t^{(tr)}, X_{1,t}^{(tr)}, \dots, X_{k,t}^{(tr)}), \quad t = 1, \dots, T^{(tr)}, \\
&(Y_t^{(ev)}, X_{1,t}^{(ev)}, \dots, X_{k,t}^{(ev)}), \quad t = 1, \dots, T^{(ev)}, \quad T^{(tr)} + T^{(ev)} = T.
\end{aligned}
$$

- Wir schätzen $\widehat{f}_{(tr)}$ mithilfe des Trainigssamples und prognostizieren damit die abhängige Variable im Evaluationssample:

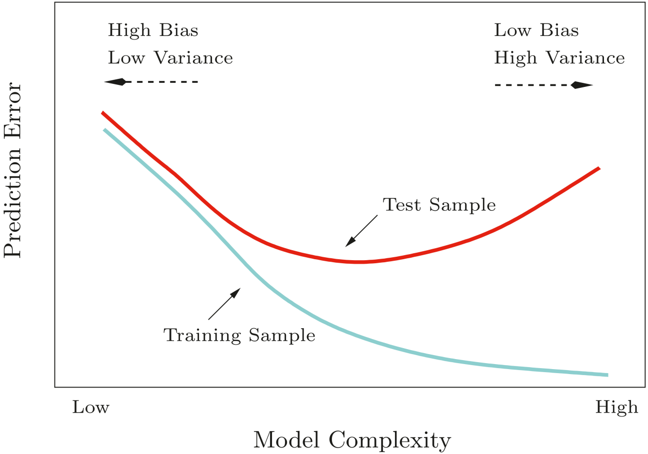
$$\widehat{Y}_t = \widehat{f}_{(tr)}(X_{1,t}^{(ev)}, \dots, X_{k,t}^{(ev)}), \quad t = 1, \dots, T^{(ev)}$$

- Out-of-Sample MSE (Test-MSE):

$$MSE = \frac{1}{T^{(ev)}} \sum_{t=1}^{T^{(ev)}} \left(Y_t - \widehat{f}_{(tr)}(X_{1,t}^{(ev)}, \dots, X_{k,t}^{(ev)}) \right)^2$$

- Kein Overfitting, MSE sinkt nicht zwingend mit der Modellkomplexität
- Problematisch mit kleinen Datensätzen da nicht alle Daten zur Schätzung verwendet werden können (Alternative: Cross-Validation).

Güte eines geschätzten Modells



Simulationen

Zur Beurteilung eines statistischen Verfahrens bieten Simulationen einen entscheidenden Vorteil gegenüber echten Daten

1.) Wir simulieren die Daten selbst und kennen daher den datengenerierenden Prozess inklusive aller Modellparameter.

2.) Wir können beliebig viele neue Beobachtungen simulieren.

Für das Beispiel OLS-Methode können wir

- die Genauigkeit des geschätzten Parametervektors $\widehat{\boldsymbol{\beta}}$ analysieren.
- die Genauigkeit von Prognosewerten $\widehat{Y}_t$ analysieren, welche wir beliebig oft replizieren können.

Simulation 1: Wie präzise ist der OLS-Schätzer

Aufbau einer Simulationsstudie: Schritt 1 – Simulieren von Daten

- Wir definieren einen datengenerierenden Prozess, z.B.:

$$Y_t = 1 + 2X_{t,1} - 3X_{t,2} + u_t, \quad X_{t,1}, X_{t,2} \overset{i.i.d.}{\sim} \mathcal{N}(1,2),\ u_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1),$$

  mit Parametern $\beta_0 = 1$, $\beta_1 = 2$ und $\beta_2 = -3$

- Wir erzeugen mithilfe eines Zufallsgenerators auf dem Computer M Datensätze

$$\begin{aligned} D^{(1)} &= (X_{1,t}^{(1)}, X_{2,t}^{(1)}, u_t^{(1)}) \quad t = 1, \ldots, T \\ D^{(2)} &= (X_{1,t}^{(2)}, X_{2,t}^{(2)}, u_t^{(2)}) \quad t = 1, \ldots, T \\ &\vdots \\ D^{(M)} &= (X_{1,t}^{(M)}, X_{2,t}^{(M)}, u_t^{(M)}) \quad t = 1, \ldots, T. \end{aligned}$$

  M wird dabei „groß“ gewählt, z.B. $M = 1000$.

- Für jeden Datensatz $k = 1, \ldots, M$ ergibt sich

$$Y_t^{(k)} = 1 + 2X_{t,1}^{(k)} - 3X_{t,2}^{(k)} + u_t^{(k)}, \quad t = 1, \ldots, T.$$

Simulation 1: Wie präzise ist der OLS-Schätzer

Aufbau einer Simulationsstudie

Schritt 2: Für jeden Datensatz $D^{(k)}$ berechnen wir den Schätzer $\widehat{\boldsymbol{\beta}}$ und bezeichnen den resultierenden Schätzwert mit $\widehat{\boldsymbol{\beta}}^{(k)} = (\widehat{\beta}_0^{(k)}, \widehat{\beta}_1^{(k)}, \widehat{\beta}_2^{(k)})'$. Wir erhalten damit M Realisierungen des Schätzers.

Schritt 3: Analyse der Schätzwerte $\widehat{\beta}_i^{(k)}$ für $k = 1, \ldots, M$.

Wie diese Analyse aussieht, hängt davon ab, welches Ziel Sie mit der Simulation verfolgen. Z.B. könnte die Analyse darin bestehen, dass man ein Gütekriterium wie z.B.

- den Mittelwert: $\overline{\widehat{\beta}_i} = \frac{1}{M} \sum_{k=1}^{M} \widehat{\beta}_i^{(k)}$
- die Varianz: $\frac{1}{M} \sum_{k=1}^{M} \left(\widehat{\beta}_i^{(k)} - \overline{\widehat{\beta}_i}\right)^2$
- den mittleren quadratischen Fehler: $MSE = \frac{1}{M} \sum_{k=1}^{M} (\beta_i - \hat{\beta}_i{}^{(k)})^2$

berechnet.

Simulation 2: Wie präzise ist die OLS-Prognose

- Wenn man nicht an dem Schätzwert $\widehat{\beta}_i$ an sich interessiert ist, sondern an der Prognosegüte des geschätzten Modells, können wir auch den In-Sample Fit und die Out-of-Sample Prognose evaluieren
- Für die Out-of-Sample Evaluation wird zu jedem Trainingsdatensatz $D^{(k)}$ ein neuer Evaluation-Datensatz simuliert,

$$
\begin{aligned}
D^{(1,ev)} &= (X_{1,t}^{(1,ev)}, X_{2,t}^{(1,ev)}, u_t^{(1,ev)}) \quad t = 1, \dots, T \\
D^{(2,ev)} &= (X_{1,t}^{(2,ev)}, X_{2,t}^{(2,ev)}, u_t^{(2,ev)}) \quad t = 1, \dots, T \\
&\vdots \\
D^{(M,ev)} &= (X_{1,t}^{(M,ev)}, X_{2,t}^{(M,ev)}, u_t^{(M,ev)}) \quad t = 1, \dots, T,
\end{aligned}
$$

mit $Y_t^{(k,ev)} = 1 + 2X_{t,1}^{(k,ev)} - 3X_{t,2}^{(k,ev)} + u_t^{(k,ev)}$, $k = 1, \dots, M$.

- Für jeden Datensatz (k) berechnen wir die Out-of-Sample Prognosen:

$$\widehat{Y}_t^{(k,ev)} = \widehat{\beta}_0^{(k)} + \widehat{\beta}_1^{(k)} X_{t,1}^{(k,ev)} - \widehat{\beta}_2^{(k)} X_{t,2}^{(k,ev)}$$

und bestimmen den MSE:

$$MSE^{(k)} = \frac{1}{M} \sum_{k=1}^{M} (Y_t^{(k,ev)} - \widehat{Y}_t^{(k,ev)})^2$$

Simulationen

Bevor Sie eine Simulation beginnen, sollten Sie sich über das Simulationsdesign Gedanken machen:

- Welche Eigenschaften des Schätzers sollen untersucht werden?
- Wie sollen die Daten erzeugt werden?
- Welche Rahmenbedingungen sollen variiert werden (z.B. Anzahl der Beobachtungen T, Anzahl der Variablen k, Stärke der Korrelation)?
- Wie weit stimmt meine Simulation mit der Situation in der Anwendung überein, die mich interessiert?