

Inlämningsuppgift - Statistisk analys av komplexa data, 7.5 hp

Moment - kategoriska data, trunkerade data och Tobitregression

Inlämningsuppgiften skall lösas individuellt. Valfritt program får användas för att lösa uppgifterna nedan, men mina instruktioner och ledtrådar gäller för programmeringsspråket R. **Labbrapport** med lösningar på nedanstående uppgifter samt relevanta datorutskrifter bifogade **mailas till mig bertil.wegmann@liu.se senast torsdagen den 9:e november kl. 23:59.**

1. (9.5 poäng) Wegmann och Villani (2011) analyserade förpackningar av samlarmynt från 1000 auktioner på auktionssajten eBay. Datamaterialet "eBaydata.rda" finner ni på kurshemsidan och består av
 - beroende variabel *UnOpen*: en dummyvariabel med värdet 1 om myntförpackningen levereras sluten och öppen i dess originalförpackning.
 - förklaringsvariabel *PowerSeller*: en dummyvariabel med värdet 1 om säljaren är rankad bland de mest lyckosamma försäljarna på eBay.
 - förklaringsvariabel *LogBookValue*: naturliga logaritmen av priset på myntförpackningen från den stora mynthanblaren Golden Eagle Coins på Internet (<http://www.goldeneaglecoin.com>).
 - förklaringsvariabel *StartBudKvot*: en lägre gräns för tillåtna bud i auktionen (satt av säljaren) dividerat med priset på myntförpackningen från den stora mynthanblaren Golden Eagle Coins.
- (a) Beräkna med både logit- och probitmodellen den skattade sannolikheten för att *UnOpen* = 1 i auktion nr. 22 (rad 22 i data). Jämför dessa skattade sannolikheter med det observerade värdet för auktion nr. 22 på *UnOpen* och kommentera rimligheten i respektive skattad sannolikhet från modellerna. (**Tips** : ?Normal(stats) ger info om funktioner för normalfördelningen.)
- (b) **Endast** för logitmodellen: Beräkna den förväntade förändringen av oddset då förklaringsvariabeln *LogBookValue* ökar med ln 2 enheter, givet att dom övriga förklaringsvariablerna hålls konstanta. Tolka resultatet i ord.
- (c) **Endast** för logitmodellen: Givet medelvärden av respektive förklaringsvariabel *LogBookValue* och *StartBudKvot*, beräkna hur många gånger mer sannolikt det är för *UnOpen* = 1 om säljaren är lyckosam (*PowerSeller* = 1) jämfört med om säljaren inte är lyckosam (*PowerSeller* = 0). Tolka resultatet i ord.
- (d) **Endast** för probitmodellen: Givet medelvärden av förklaringsvariabeln *StartBudKvot* och att *PowerSeller* = 0, beräkna den marginella effekten av förklaringsvariabeln *LogBookValue* på $P(\text{UnOpen} = 1)$, utvärderat i medelvärdet för *LogBookValue*. Tolka resultatet i ord. (**Tips** : ?Normal(stats) ger info om funktioner för normalfördelningen.)
- (e) **Endast** för probitmodellen: Utför beräkning av den marginella effekten i (d), men nu utvärderad för alla observerade värden i *LogBookValue*. Redovisa den marginella effekten som funktion av *LogBookValue* i en figur och kommentera utförligt vad figuren visar. (**Tips** : använd exempelvis funktionen plot(x,y) för att plotta y som funktion av x i R.)
- (f) **Endast** för probitmodellen: Utför både ett Wald-test och ett likelihood-ratio-test för om förklaringsvariablerna *LogBookValue* och *StartBudKvot* tillsammans bidrar till att förklara sannolikheten för att myntförpackningen levereras sluten och öppen i dess originalförpackning, givet att *PowerSeller* redan finns med i modellen (**tips** : använd funktionerna wald.test och

lrtest i R genom att installera paketen *aod*: `install.packages('aod');library(aod)` och *lmtest*: `install.packages('lmtest');library(lmtest)`. Jämför resultaten av testen utifrån olika val av signifikansnivåer och tolka dem i ord. Ger testen samma resultat? Varför eller varför inte? Motivera ditt svar i ord på dessa två frågor utifrån följande dokumentation:

<https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-are-the-likelihood-ratio-wald-and-lagrange-multiplier-score-tests-different-and-or-similar/>

- (g) **Endast** för logitmodellen: Beräkna **för hand** värdet på Walds teststatistika under nollhypotesen för Waldtestet i uppgift (f) med hjälp av skattningsresultaten från logitmodellen. Avgör sedan om förklaringsvariablerna *LogBookValue* och *StartBudKvot* tillsammans bidrar till att förklara sannolikheten för att myntförpackningen levereras sluten och öppen i dess originalförpackning, givet att *PowerSeller* redan finns med i modellen. **Obs!** Var noga med att redovisa alla dina beräkningar för att räkna ut värdet på Walds teststatistika.
 - (h) Utför ett z-test med både logit- och probitmodellen för att testa om förklaringsvariabeln *LogBookValue* bidrar till att förklara sannolikheten för att myntförpackningen levereras sluten och öppen i dess originalförpackning, givet att *PowerSeller* redan finns med i modellen. Drag fullständiga slutsatser av dina test i ord utifrån olika val av signifikansnivåer för testen och jämför testresultaten mellan modellerna.
2. (4 poäng) I denna uppgift ska den multinomiala logistiska regressionsmodellen användas för att analysera datamaterialet i 'Example 3' från följande dokumentation av dataanalys i R:

<https://stats.idre.ucla.edu/r/dae/multinomial-logistic-regression/>

I dokumentationen ger funktionen `summary(test)` en tabell med skattade koefficienter och skattade standardavvikelser för de skattade koefficienterna. Antag att vi kodar om värden för den beroende variabeln 'prog' enligt ("general"=0, "vocation"=1, "academic"=2). Då följer att parameterskattningarna för exempelvis rad "vocation" i dokumentationen motsvarar skattningen av parametervektorn β_1 i föreläsningsunderlaget på kategoriska data.

Använd parameterskattningarna från den skattade multinomiala logistiska regressionsmodellen och redovisa dina beräkningar för hand i uppgift (a) till (c). Bifoga de utskrifter du använt från R.

- (a) Skatta sannolikheten för att en student som har 'writing score = 49' och socio-ekonomisk status 'high' väljer program "academic".
 - (b) Vad är odds-kvoten för att vara i program "vocation" gentemot program "general" för en student med socio-ekonomisk status 'low' och 'writing score = "medelvärde av writing score"' ? Motivera dina beräkningar och tolka odds-kvoten i ord.
 - (c) Vad är det genomsnittliga värdet på odds-kvoten för att vara i program "academic" gentemot program "vocation" för en student med socio-ekonomisk status 'middle' sett över alla odds-kvoter för alla observerade värden på 'writing score' i datamaterialet? Motivera dina beräkningar och tolka det genomsnittliga värdet på odds-kvoten i ord.
3. (3.5 poäng) Använd datamaterialet från följande dokumentation av dataanalys i R:

<https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/>

- (a) Skatta sannolikheten för att en student i en modell med $gpa = 3.1$ och $pared = 1$ är "very likely" att ansöka till *Graduate school*.
- (b) Skatta sannolikheten för att en student i en modell med $gpa = 3.4$ och $pared = 0$ och $public = 1$ är "somewhat likely" att ansöka till *Graduate school*.

- (c) Plotta de marginella effekterna från förklaringsvariabeln *gpa* på sannolikheten att en student är “unlikely” att ansöka till *Graduate school*, givet *public* = 1.
4. (3 poäng) Antag att intelligensen för en population (Y), mätt i IQ utifrån ett intelligenstest, är normalfördelad med medelvärdet 100 och att en IQ över 140 uppnås av 1 % av populationen.
- (a) Skapa ett histogram som speglar fördelningen för den del av populationen som har en IQ över 120. Var noga med att redovisa hur du tog fram histogrammet. (**Tips** : Använd funktion `rnorm` för att slumpa fram värden från en normalfördelning. Använd funktion `hist(x)` för att ta fram ett histogram för en variabel X .)
- (b) Beräkna det exakta väntevärdet och variansen för den del av populationen som har en IQ över 120 och jämför med skattningar av väntevärde och varians från simulerade värden för denna subpopulation. Vad illustrerar väntevärdet och variansen i subpopulationen jämfört med väntevärdet och variansen i hela populationen? Motivera utförligt!