Notes on the case study data and case study replication attempts in HRG (2005) using the R Apollo package

(and the file CaseStudy.doc received in DH’s March 2, 2025 email which is Chapter 9 of the book but with differing section, table and graph numbering)

1 Files in Github repo hrg05repl 1

2 Data screening and cleaning 1

2.1 Travel mode labelling (variables ALTIJ, CN,…) 1

2.1.1 From the text 1

2.1.2 From the data 2

2.2 Levels of TOLLPRED 3

2.3 Cleaning 3

2.4 Replicated data set descriptions 3

3 Multicollinearity checks 3

4 MNL estimation 5

4.1 preparations 5

4.2 results 5

# Files in Github repo hrg05repl

References in this text to r-scripts (.r), and “listing” files (.lis) refer to files in the directories “R/cmd”, and “R/lis” in the repo “hrg05repl”.

The r-scipt *\_hrg\_steer.r* allows calling all analysis scripts (by removing the comment-qualifyier “#”) and provides brief hints on what each analysis script does.

# Data screening and cleaning

## Travel mode labelling (variables ALTIJ, CN,…)

### From the text

Different parts in Chapter 9 of HRG05 imply two different interpretations of choice set compositions (public transport mode combinations) in the SP study.

**Interpretation A** is that the combinations **Bus+LR** and **Train+BW** do appear in the choice cards while **Bus+BW** and **Train+LR** never appear:

1. *Four alternatives appear in each mode choice scenario; CAR NO TOLL, CAR TOLL ROAD, BUS or BUSWAY, and TRAIN or LIGHT RAIL.* (Section 9.4.1 (p.268)
2. … types of showcards described scenarios involving … public transport pairs (4): bus vs. light rail, bus vs. train (heavy rail), busway vs. light rail, and busway vs. train. (Section 9.4.1 (p.268)

**Interpretation B** is that the combinations **Bus+BW** and **Train+LR** do appear while the combinations **Bus+LR** and **Train+BW** never appear:

1. Given the method used to determine which public transport methods were present within a choice set, the bus and LR and train and busway alternatives never appeared within the same choice set. (Section 9.4.1 (p.268)
2. N.b. To determine whether the choice set has bus/busway or train/LR present, the switch factors must be considered concurrently  
   Switch1 = 0, Switch2 = 0: bus and train present; Switch1 = 0, Switch2 = 1: bus and busway present;   
   Switch1 = 1, Switch2 = 0: train and Light Rail present; Swicth1 = 1, Switch2 = 1: busway and Light Rail present (note to Table 9.5, p. 273)
3. Variable CN (“Alternatives present within choice set”) indicates public transport mode combinations available in each SP choice set. The codebook (Appendix 9C, p. 302) indicates “Bus – Busway” and Train – Light Rail”.

According to the code book (Appendix 9C, p. 302) the ALTIJ numeric values are labelled as follows:

1 cart Car with toll

2 carnt car no toll

3 bus Bus

4 train Train

5 lr Light Rail

6 bw busway

### From the data

#### SPXXX (names of variables indicating choice)

The variable ALTIJ indicates the transport mode in each alternative (row) with numeric category codes 1:6.[[1]](#footnote-1)

The 1/0-dummy variable CHOICE indicates the alternative among the four alternatives in each choice set which was chosen by the respondent.

The variable SPCHOICE has the same numeric value for each row (alternative) in each respective choice set: the ALTIJ-value of the chosen alternative, i.e. of the row with CHOICE=1.

Each of the six 1/0 dummy variables SPCART, SPCARNT, SPBUS, SPTN, SPBW, and SPLR also has the same numeric value for each row (alternative) in each respective choice: it indicates the mode chosen in the choice situation. From their column names the following labelling of ALTIJ (and SPCHOICE) can be derived:

SPCHOICE SPdummy=1

1. SPCART
2. SPCARNT
3. SPBUS
4. SPTN
5. SPBW
6. SPLR

Labelling according to the codebook and labelling derived from the data are consistent with respect the first four codes (CART=1, CARNT=2, BUS=3, TRAIN=4) but inconsistent w respect to codes 5 and 6 (BW=?, LR=?).

#### Considering mode combinations in choice sets

The frequency of combinations of public transport modes in choice situations in the SP survey data are as follows:

|  |  |
| --- | --- |
| 3+4 | 823 |
| 3+5 | 771 |
| 3+6 | 0 |
| 4+5 | 0 |
| 4+6 | 818 |
| 5+6 | 1202 |

If interpretation A is true (Bus+LR and Train+BW occuring) 3+5 means Bus+LR and 4+6 means Train+BW, implying the labelling LR=5 and BW=6 consistent with the codebook (App. 9C).

If interpretation B is true (Bus+BW and Train+LR occuring) 3+5 means Bus+BW and 4+6 means Train+LR, implying the labelling LR=6 and BW=5, consistent with labels derived from the SPXXX-data.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Appearing on choice cards | NEVER on choice cards | Resulting labelling |
| **Interpretation A**: all choice sets combine one “rail tracks mode” and one “street mode” | **Bus+LR**  **Train+BW** | **Bus+BW**  **Train+LR** | **LR=5**  **BW=6** |
| **Interpretation B**: some choice sets combine two rail tracks modes or two street modes | **Bus+BW**  **Train+LR** | **Bus+LR**  **Train+BW** | **LR=6**  **BW=5** |

The codebook specifies for variable CN (“Alternatives present within choice set”):

0 = …

1 = …

2 = Bus – Busway (SP)

3 = Train – Light Rail (SP)

4 = …

This is consistent with interpretation B as the data show that:

* Choice sets with CN=2 = “Bus – Busway (SP)” contain (1,2,3 and) 5 for variable ALTIJ
* Choice sets with CN=3 = “Train – Light Rail (SP)” contain (1,2,4 and) 6 for variable ALTIJ

The DC model estimation results in chapter 10 of HRG05 can be replicated if labelling according to interpretation B is implemented.

## Levels of TOLLPRED

Table 9.4 in HRG2005 (p.269) indicates three levels of the TOLLPRED variable (“Pay toll if you leave at this time (otherwise free)”

|  |  |
| --- | --- |
| Level code | Label |
| 0 | 06:00 - 10:00 |
| 1 | 06:30 – 08:30 |
| 2 | 06:30 – 09:00 |

A consistent ranking regarding restrictiveness would imply different labels, e.g. by exchanging the second and third. Which labels for levels 0, 1, and 2have actually been used in the choice cards?

## Cleaning

The assignment of choice cards is in some cases not in accordance with stated usual commuting time (SPLENGTH). Respondents stating SPLENGTH level 3 (30-45min) sometimes received choice cards with travel time attribute levels below 15 or above 45 minutes. I did not find any problems regarding the assignment of choice cards to respondents with SPLENGTH level 2 (<30).

## Replicating data set descriptions

The data set received has 2 observations more than indicated in the book. ID 6207 has one choice (4 rows) and two RP-rows. These are the only two rows with splength=”RP”. I removed these two rows (in *hrg01a\_imp\_repl.r*) to see if this entails consistency between data an book. It does exactly replicate descriptives such as the mean of variable ID.

# Multicollinearity checks

For including other alternatives (as suggested in the last paragraph of p. 295) requires data rearrangements (as explained on p. 247). Implemented using apollo\_longToWide() form apollo package (hrg02\_descr.r).

Supplementary regressions on data with splength <30min and for first (=CART) alternative (p. 293-95 in the book) exactly replicated: two supplementary regressions suggest multicollinearity.

Regressions for second alternative (CARNT) find no collinearity.

Regressions for remaining alternatives (BUS, TR, BW,LR) indicate collinearity throughout.

This corresponds with correlation matrix:



Image 1: Correlations (SPLENGTH = <30min)

Is this in the data or due to a bug?

Large correlations occur within each of the public transport alternatives as well as between the attributes in Train and Busway alternatives and

For respondents with longer commuting times the pattern is similar:



Image 2: Correlations (SPLENGTH = 30-45min)



Image 3: Correlations (SPLENGTH = >45min)

# MNL estimation

## preparations

## results

* For the first model (6 modes, 5 ASC, one cost parameter, CST, for fuel and fare) HRG05 Ch. 10.3 (p.316) I get identical results for final LL and parameter estimates (except for LL of no-data base model and the resulting fit statistics): hrg03\_mnl1.r
* ASC-only model (p.329): hrg03\_mnl2
* LR-Test comparing two models (335ff): hrg03\_mnl3
* HRG05 10.4 (p.344) Dummy coded attributes (nonlinearity) hrg03\_mnl4
* HRG05 10.5 (p.352ff) interactions1 hrg03\_mnl5  
  results in the book (p.353) result from the code on p.352. (with erroneously using the categorical variable *drivlic* (respondent’s drivers licence) rather than *ndrivlic* (number of licensed drivers licenses in household). The result from using *ndrivlic* (hrg03\_mnl5.lis) suggest the interaction term vehlic non significant, however the LR-test indicates improvement over base model. Using *ndrivlic* rather than *drivlic* requires changes to Table 10.4 and Figure 10.11.
* HRG05 10.5 (p.355-57) Interactions w categorical vars: exact replication: hrg03\_mnl6
* HRG05 10.6 (p.357-59) WTP (Value of Travel Time Savings, VTTS): replication in *hrg03\_hrg\_mnl07*
* HRG05 10.7-8 (p.360-371) replicated in *hrg03\_hrg\_mnl08*:

- Choice probabilities/predictions and aggregates (sums) of pred probs

- Utilities (366ff)

* HRG05 11.2.3 Crosstab (p.381ff) replication in *hrg03\_hrg\_mnl08*:

HRG15 13.2.4 ;Crosstab mentions two possible ways of „counting” the number of times each level was predicted: a) the level-wise summation of choice probabilities, and b) counting the numbrer of times the respective level had highest predicted probability. The tables on p. 382/83 are produced using method a) while the text menions method b.

1. ALTIJ labelling according to the codebook differs between SP and RP rows but I only consider only SP rows: SPRP=2 [↑](#footnote-ref-1)