Assignment 1

Linear Regression with R

Juan Carlos Villaseñor-Derbez

January 26, 2018

Contents

1 Descriptive statistics of `SmallHHfile.csv` data

Using the `describe()` function from the `psych` package (Revelle 2017) we can get a table of summary statistics (Table 1).

Table 1: Descriptive summary statistics of the SmallHHfile data.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
SAMPN	1	42431	2588378.634	1641345.144	1971814.000	2195483.363	847148.744	1031985	7212388.000	6180403.000	2.039	3.092	7968.164
INCOM	2	42431	13.179	26.285	5.000	5.509	2.965	1	99.000	98.000	2.922	6.620	0.128
HHSIZ	3	42431	2.572	1.374	2.000	2.405	1.483	1	8.000	7.000	1.030	0.898	0.007
HHEMP	4	42431	1.222	0.883	1.000	1.185	1.483	0	6.000	6.000	0.471	0.328	0.004
HHSTU	5	42431	0.644	1.023	0.000	0.441	0.000	0	8.000	8.000	1.664	2.516	0.005
HHLIC	6	42431	1.861	0.848	2.000	1.810	0.000	0	8.000	8.000	0.604	1.705	0.004
DOW	7	42431	4.020	1.994	4.000	4.025	2.965	1	7.000	6.000	0.002	-1.241	0.010
HTRIPS	8	42431	8.290	7.776	6.000	7.142	5.930	0	99.000	99.000	1.721	4.880	0.038
Mon	9	42431	0.136	0.343	0.000	0.046	0.000	0	1.000	1.000	2.119	2.489	0.002
Tue	10	42431	0.144	0.351	0.000	0.055	0.000	0	1.000	1.000	2.025	2.103	0.002
Wed	11	42431	0.145	0.352	0.000	0.056	0.000	0	1.000	1.000	2.020	2.081	0.002
Thu	12	42431	0.147	0.354	0.000	0.059	0.000	0	1.000	1.000	1.990	1.959	0.002
Fri	13	42431	0.139	0.346	0.000	0.049	0.000	0	1.000	1.000	2.081	2.332	0.002
Sat	14	42431	0.141	0.348	0.000	0.051	0.000	0	1.000	1.000	2.064	2.260	0.002
Sun	15	42431	0.147	0.354	0.000	0.059	0.000	0	1.000	1.000	1.994	1.976	0.002
TotDist	16	42431	68.093	118.516	33.894	45.435	45.128	0	5838.261	5838.261	8.379	196.692	0.575
center	17	42431	0.281	0.449	0.000	0.226	0.000	0	1.000	1.000	0.975	-1.050	0.002
suburb	18	42431	0.288	0.453	0.000	0.235	0.000	0	1.000	1.000	0.938	-1.121	0.002
exurb	19	42431	0.229	0.420	0.000	0.161	0.000	0	1.000	1.000	1.290	-0.335	0.002
rural	20	42431	0.202	0.401	0.000	0.127	0.000	0	1.000	1.000	1.486	0.208	0.002
other	21	42431	0.000	0.000	0.000	0.000	0.000	0	0.000	0.000	NaN	NaN	0.000
highinc	22	42431	0.413	0.492	0.000	0.391	0.000	0	1.000	1.000	0.353	-1.876	0.002
HHVEH	23	42431	1.862	0.997	2.000	1.809	1.483	0	8.000	8.000	0.803	2.258	0.005
HHBIC	24	42431	1.584	3.786	1.000	1.196	1.483	0	99.000	99.000	20.405	513.753	0.018
VEHNEW	25	42431	2.153	2.016	2.000	1.566	1.483	1	9.000	8.000	2.380	4.200	0.010
OWN	26	42431	1.244	0.555	1.000	1.158	0.000	1	9.000	8.000	5.964	67.494	0.003
CarBuy	27	42431	0.453	0.498	0.000	0.441	0.000	0	1.000	1.000	0.191	-1.964	0.002
snglhm	28	42431	0.818	0.386	1.000	0.897	0.000	0	1.000	1.000	-1.646	0.708	0.002
ownhm	29	42431	0.773	0.419	1.000	0.842	0.000	0	1.000	1.000	-1.306	-0.295	0.002
MilesPr	30	42431	27.122	43.460	14.503	18.400	18.191	0	1167.652	1167.652	5.149	47.236	0.211
TrpPrs	31	42431	3.280	2.579	3.000	3.020	2.224	0	32.000	32.000	1.266	3.675	0.013

2 Estimate model 1

Coefficients in model 1 were estimated via ordinary least squares with the `lm()` function in R (R Core Team 2017).

```
model1 <- lm(formula = MilesPr ~ Mon + Tue + Wed + Thu + Fri + Sat
             + HHVEH + HHSIZ + suburb + exurb + rural, data = SmallHHfile)
```

2.1 Report model 1

Estimated coefficients and their respective standard error and t-statistics, as well as information on model fit are presented in Table 2.

Table 2: Ordinary least square estimates for coefficients in model 1.

Variable	Estimate	t-statistic
(Intercept)	19.077 (0.771)	24.728 ***
Mon	-0.680 (0.786)	-0.865
Tue	0.715 (0.775)	0.922
Wed	0.959 (0.774)	1.238
Thu	0.666 (0.771)	0.864
Fri	3.786 (0.781)	4.845 ***
Sat	3.569 (0.779)	4.579 ***
HHVEH	4.894 (0.231)	21.148 ***
HHSIZ	-2.360 (0.166)	-14.242 ***
suburb	3.118 (0.558)	5.585 ***
exurb	6.251 (0.595)	10.513 ***
rural	6.911 (0.617)	11.199 ***
Obs.	42431	
RSE	43.054 (df = 42419)	
R2	0.019	
F Statistic	74.030*** (df = 11; 42419)	

Note:
*p<0.1; **p<0.05; ***p<0.01
Numbers in parentheses represent standard errors

2.2 Equation for model 1

$$MilesPr = \beta_0 + \beta_1 Mon + \beta_2 Tue + \beta_3 Wed + \beta_4 Thu + \beta_5 Fri + \beta_6 Sat + \beta_7 HHVEH + \beta_8 HHSIZ + \beta_9 suburb + \beta_{10} exurb + \beta_{11} rural + \epsilon \quad (1)$$

2.3 Summay of model 1

Model 1 describes the relationship between `MilesPr` and a number of explanatory variables. On its current specification, the included variables explain 1.9% of the variance in `MilesPr` (R^2 = 0.019; $F(11, 42419) = 74.030$; $p < 0.01$). Among the included variables, the day of the week (Sun - Sat) captures the average change in `MilesPr` depending on the day. From table 1, we observe that Friday and Saturday significantly increase `MilesPr` by 3.78 and 3.56 miles, respectively ($p < 0.01$), as compared to the reference level (Sunday, living in the center of the city). The estimates for all other days of the week imply that they are not significantly different from zero, and thus have no significant effect on `MilesPr` ($p > 0.05$). The number of vehicles in the household (`HHVEH`) had a strong positive effect on `MilesPr`, indicating a marginal increase of 4.89 miles per vehicle ($p < 0.01$). Conversely, household size (`HHSIZ`) had a mild negative effect on `MilesPr`, indicating a reduction in `MilesPr` for the addition of a person to the household ($p < 0.01$). Households further away from the city center also tend to travel more miles. Consistently, miles traveled per person follow `rural` > `exurb` > `suburb` ($p < 0.01$).

Under this model specification, the model suggests the following:

- People travel larger distances Friday and Saturdays (likely as leisure)
- The positive relationship observed between miles per person and number of cars in the household suggest that more cars enable a individuals from a household to travel more.
- However, an increase in household size resulted in a decrease in miles traveled per person. Without accounting for income or interactions with other variables, we are unable to draw conclusions from the observed model.
- Finally, people further away from the city center have larger per capita travel distances.

3 Estimate model 2

$$HTRIPS = \beta_0 + \beta_2 HHSIZ + \beta_3 HHVEH + \beta_4 TrpPrs + \beta_5 INCOM + \beta_6 Mon + \beta_7 Tue + \beta_8 Wed + \beta_9 Thu + \beta_{10} Fri + \beta_{11} Sat + \epsilon \quad (2)$$

3.1 Report model 2

Table 3: Ordinary least square estimates for coefficients in model 2.

Variable	Estimate	t-statistic
(Intercept)	-7.402 (0.059)	-126.072 ***
HHSIZ	3.230 (0.013)	254.274 ***
HHVEH	0.009 (0.018)	0.490
TrpPrs	2.181 (0.006)	347.124 ***
INCOM	-0.000 (0.001)	-0.572
Mon	0.127 (0.060)	2.101 **
Tue	0.378 (0.060)	6.329 ***
Wed	0.380 (0.060)	6.371 ***
Thu	0.323 (0.059)	5.438 ***
Fri	0.326 (0.060)	5.423 ***
Sat	0.021 (0.060)	0.349
Obs.	42431	
RSE	3.300 (df = 42420)	
R2	0.820	
F Statistic	19321.927*** (df = 10; 42420)	

Note:
*p<0.1; **p<0.05; ***p<0.01
Numbers in parentheses represent standard errors

3.2 Comparing model 1 and model 2

Model 1 focused on identifying how *per capita* distance travaled by household depended on a series of explanatory variables. In this case, model 2 estimates the effect that a set of variables have on the number of trips (ignoring the length of the trip) that a household does. This model explains 82% of the variance in `HTRIPS` ($R^2 = 0.82$; F (10, 42420) = 19321.927; $p < 0.01$).

The intercept estimate explains the mean effect of a Sunday, indicating that people perform less trips this day (perhaps less but longer trips, considering results of 1). In this case, we identify that people do more trips Tue - Fri, with similar coefficient estimates, as compared to the baseline of a Sunday. Perhaps this reflects a monotonous behaviour in day-to-day activities during week days. Finally, household size and the number of trips per person had a positive effect on the number of trips that the household did (p , 0.01), while income and number of household vehicles had no apparent effect ($p > 0.01$).

4 Other packages used

- `broom` (Robinson 2017)
- `lmtest` (Zeileis and Hothorn 2002)
- `tidyverse` (Wickham 2017)
- `knitr` (Xie 2017)
- `kableExtra` (Zhu 2018)

References

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Revelle, William. 2017. *Psych: Procedures for Psychological, Psychometric, and Personality Research*. Evanston, Illinois: Northwestern University. https://CRAN.R-project.org/package=psych.

Robinson, David. 2017. *Broom: Convert Statistical Analysis Objects into Tidy Data Frames*. https://CRAN.R-project.org/package=broom.

Wickham, Hadley. 2017. *Tidyverse: Easily Install and Load the 'Tidyverse'*. https://CRAN.R-project.org/package=tidyverse.

Xie, Yihui. 2017. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://yihui.name/knitr/.

Zeileis, Achim, and Torsten Hothorn. 2002. "Diagnostic Checking in Regression Relationships." *R News* 2 (3): 7–10. https://CRAN.R-project.org/doc/Rnews/.

Zhu, Hao. 2018. *KableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.