

Introduction to R for Social Scientists

Exercises

– Tables, Regression Tables and Coefficient Plots –

Benjamin Elbers

be2239@columbia.edu

In-Class Exercises

Exercises A

Use the `flights` dataset from the `nycflights13` package to recreate the following table. Use `datasummary()` from the `'modelsummary'` package. Output the table in `tex`, `html`, and `docx` format.

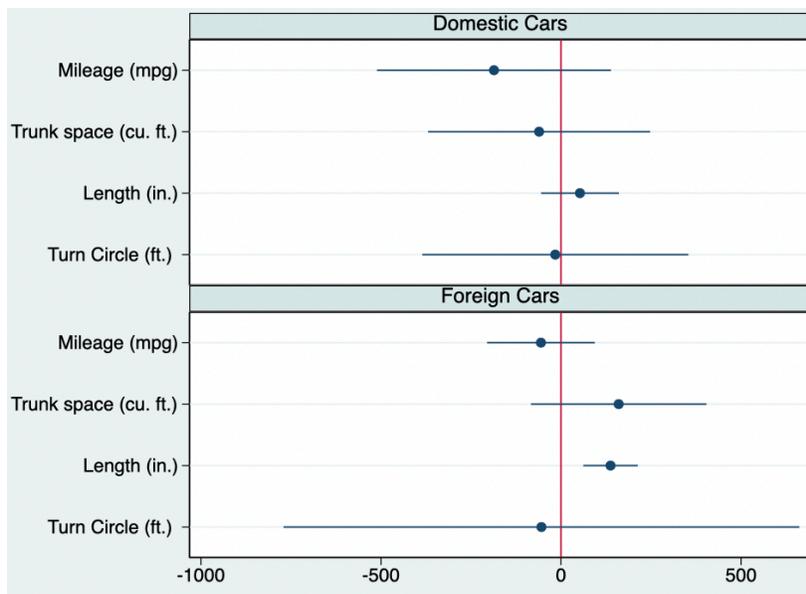
origin		Dep. Delay (mins.)	Proportion Dep. Delay >30min
EWR	Mean	15.11	0.17
	SD	41.32	0.38
	N	117596	117596
	N cancelled	3239	3239
JFK	Mean	12.11	0.14
	SD	39.04	0.35
	N	109416	109416
	N cancelled	1863	1863
LGA	Mean	10.35	0.13
	SD	39.99	0.34
	N	101509	101509
	N cancelled	3153	3153

Exercises B

1. Use `webuse::webuse("margex")` to load the `margex` dataset. Estimate two logistic regression models, using `outcome` and a number of predictors, including interactions and/or polynomials. Create a polished regression table.
2. Using the same models, explore what happens when you run `margins()` on the models and pass these objects to `modelsummary()`.

Exercises C

The following plot was produced using Stata's `coefplot`. Recreate this plot in R. The dataset uses `is auto`, and the two regressions that are shown are estimated in Stata by `reg price mpg trunk length turn if foreign==0` (or 1). You can go through `parmbly` or estimate the regressions in R.



Additional exercises

1. Use `webuse::webuse("margex")` to load the `margex` dataset. Estimate two logistic regression models, using `outcome` and a number of predictors, including interactions and/or polynomials. Create a polished plot of the marginal effects.
2. Use `library("gapminder")` to load the `gapminder` dataset. Using only countries in Europe, regress `life expectancy` on `year` fixed effects and `log(GDP per Capita)`.
 - (a) Using `augment()` from the 'broom' package and 'ggplot', plot the standardized residuals against the fitted values.
 - (b) Plot the standardized residuals against `log(GDP per Capita)`.
 - (c) (*Hard*) Manually create a `qqplot`, comparing the standardized residuals against the expected residuals from a $N(0, 1)$ (i.e. standard normal) distribution. This requires sorting the data frame by `.std.resid`, and computing the standard normal quantiles based on

$$\frac{i - 1/2}{n}$$

where i the row number and n is the number of cases (this gives the proportion of the data that falls below a certain value). The plot should have a diagonal line indicating perfect fit, should be square (check `coord_fixed()`), and the x and y axis limits should be identical.

- (d) Remove the problematic data points, re-estimate the model and repeat (a)-(c).