Case Control Study: Death on the Roads

Timothy Lee

11/04/2020

Context

The dataset below is a subset of the data from www.gov.uk/government/statistical-datasets/ras30-reported-casualtie with all of the road traffic accidents in the UK from 1979 to 2015. The data below consist of all pedestrians involved in motor vehicle accidents with either fatal or slight injuries (pedestrians with moderate injuries have been removed). We want to investigate whether the UK road accident data are consistent with the hypothesis that women tend to be, on average, safer as pedestrians than men, particularly as teenagers and in early adulthood.

```
pedestrainFile = Pmisc::downloadIfOld(
'http://pbrown.ca/teaching/303/data/pedestrians.rds')
```

```
## Loading required namespace: R.utils
```

```
pedestrians = readRDS(pedestrainFile)
pedestrians = pedestrians[!is.na(pedestrians$time), ]
pedestrians$y = pedestrians$Casualty_Severity == 'Fatal'
```

Dimensions of data

```
dim(pedestrians)
```

```
## [1] 1159453       7
```

Head of dataframe

```
pedestrians[1:3, ]
```

```
##                   time     age  sex Casualty_Severity      Light_Conditions
## 54 1979-01-01 22:40:00 26 - 35 Male            Slight Darkness - lights lit
## 65 1979-01-02 10:40:00 26 - 35 Male            Slight              Daylight
## 79 1979-01-02 14:25:00 46 - 55 Male            Slight              Daylight
##       Weather_Conditions     y
## 54 Snowing no high winds FALSE
## 65 Raining no high winds FALSE
## 79 Raining no high winds FALSE
```

Table of injuries by sex.

```
table(pedestrians$Casualty_Severity, pedestrians$sex)
```

```
##
##            Male Female
##   Slight 637977 481832
##   Fatal   24432  15212
```

Range of time

```
range(pedestrians$time)
```

```
## [1] "1979-01-01 01:00:00 AST" "2015-12-31 23:35:00 AST"
```

Preliminary Analysis

Notice that men are involved in accidents more than women, and the proportion of accidents which are fatal is higher for men than for women. This might be due in part to women being more reluctant than men to walk outdoors late at night or in poor weather, and could also reflect men being on average more likely to engage in risky behaviour than women.

A `GLM` adjusting for weather and light conditions is below.

```
theGlm = glm(y ~ sex + age + Light_Conditions + Weather_Conditions, data = pedestrians, family = binomia
knitr::kable(summary(theGlm)$coef, digits = 3)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.177	0.020	-203.938	0.000
sexFemale	-0.275	0.011	-24.670	0.000
age0 - 5	0.186	0.032	5.839	0.000
age6 - 10	-0.357	0.030	-12.021	0.000
age11 - 15	-0.504	0.029	-17.659	0.000
age16 - 20	-0.338	0.027	-12.302	0.000
age21 - 25	-0.158	0.029	-5.430	0.000
age36 - 45	0.325	0.027	12.218	0.000
age46 - 55	0.660	0.026	25.050	0.000
age56 - 65	1.138	0.025	45.365	0.000
age66 - 75	1.760	0.023	75.245	0.000
ageOver 75	2.328	0.022	104.312	0.000
Light_ConditionsDarkness - lights lit	0.995	0.012	81.217	0.000
Light_ConditionsDarkness - lights unlit	1.175	0.052	22.413	0.000
Light_ConditionsDarkness - no lighting	2.766	0.021	131.317	0.000
Light_ConditionsDarkness - lighting unknown	0.259	0.068	3.787	0.000
Weather_ConditionsRaining no high winds	-0.214	0.017	-12.961	0.000
Weather_ConditionsSnowing no high winds	-0.751	0.092	-8.136	0.000
Weather_ConditionsFine + high winds	0.176	0.037	4.810	0.000
Weather_ConditionsRaining + high winds	-0.066	0.040	-1.648	0.099
Weather_ConditionsSnowing + high winds	-0.550	0.172	-3.193	0.001
Weather_ConditionsFog or mist	0.069	0.069	0.990	0.322

Another `GLM` with interactions of sex and age is fitted. This model accounts for the significance of (the `interaction`) between both the person's `age` and `sex`.

```
theGlmInt = glm(y ~ sex * age + Light_Conditions + Weather_Conditions,
data = pedestrians, family = binomial(link = "logit"))
knitr::kable(summary(theGlmInt)$coef, digits = 3)
```

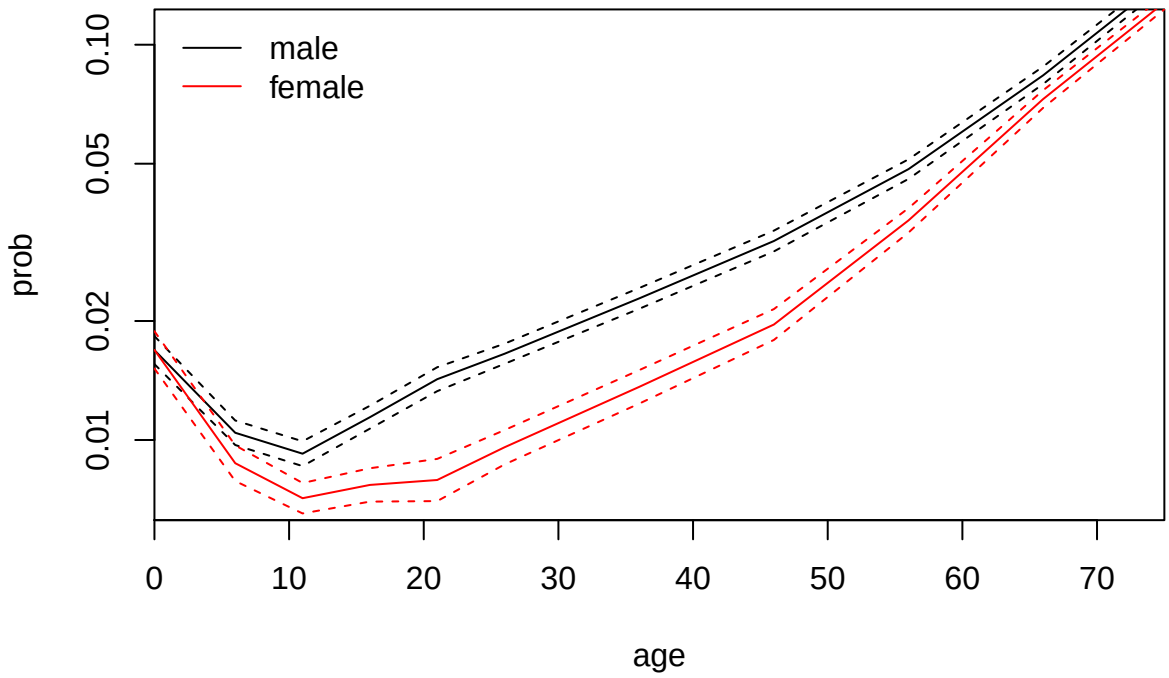
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.103	0.023	-179.897	0.000
sexFemale	-0.545	0.044	-12.427	0.000
age0 - 5	0.021	0.039	0.550	0.583
age6 - 10	-0.460	0.035	-13.099	0.000
age11 - 15	-0.582	0.035	-16.619	0.000

	Estimate	Std. Error	z value	Pr(>|z|)
age16 - 20	-0.370	0.032	-11.468	0.000
age21 - 25	-0.148	0.033	-4.470	0.000
age36 - 45	0.322	0.031	10.511	0.000
age46 - 55	0.657	0.031	21.301	0.000
age56 - 65	1.075	0.030	35.736	0.000
age66 - 75	1.622	0.029	56.324	0.000
ageOver 75	2.180	0.027	79.603	0.000
Light_ConditionsDarkness - lights lit	0.990	0.012	80.672	0.000
Light_ConditionsDarkness - lights unlit	1.174	0.052	22.397	0.000
Light_ConditionsDarkness - no lighting	2.746	0.021	130.179	0.000
Light_ConditionsDarkness - lighting unknown	0.257	0.068	3.758	0.000
Weather_ConditionsRaining no high winds	-0.211	0.017	-12.768	0.000
Weather_ConditionsSnowing no high winds	-0.746	0.092	-8.075	0.000
Weather_ConditionsFine + high winds	0.177	0.037	4.839	0.000
Weather_ConditionsRaining + high winds	-0.062	0.040	-1.544	0.123
Weather_ConditionsSnowing + high winds	-0.548	0.172	-3.189	0.001
Weather_ConditionsFog or mist	0.065	0.069	0.943	0.345
sexFemale:age0 - 5	0.546	0.068	7.971	0.000
sexFemale:age6 - 10	0.367	0.066	5.607	0.000
sexFemale:age11 - 15	0.285	0.062	4.605	0.000
sexFemale:age16 - 20	0.150	0.062	2.416	0.016
sexFemale:age21 - 25	-0.042	0.069	-0.612	0.541
sexFemale:age36 - 45	0.029	0.062	0.477	0.634
sexFemale:age46 - 55	0.058	0.060	0.973	0.331
sexFemale:age56 - 65	0.246	0.056	4.419	0.000
sexFemale:age66 - 75	0.406	0.052	7.879	0.000
sexFemale:ageOver 75	0.411	0.049	8.351	0.000

Plot of tredicted probability of being a case in baseline conditions (daylight, fine no wind) with 99% CI using `theGlmInt`.

```
newData = expand.grid(
age = levels(pedestrians$age),
sex = c('Male', 'Female'),
Light_Conditions = levels(pedestrians$Light_Conditions)[1],
Weather_Conditions = levels(pedestrians$Weather_Conditions)[1])
thePred = as.matrix(as.data.frame(
predict(theGlmInt, newData, se.fit=TRUE)[1:2])) %*% Pmisc::ciMat(0.99)
thePred = as.data.frame(thePred)
thePred$sex =newData$sex
thePred$age = as.numeric(gsub("[[:punct:]].*|[[:alpha:]]", "", newData$age))
toPlot2 = reshape2::melt(thePred, id.vars = c('age','sex'))
toPlot3 = reshape2::dcast(toPlot2, age ~ sex + variable)

matplot(toPlot3$age, exp(toPlot3[,-1]),
type='l', log='y', col=rep(c('black','red'), each=3),
lty=rep(c(1,2,2),2),
ylim = c(0.007, 0.11), xaxs='i',
xlab= 'age', ylab='prob')
legend('topleft', lty=1, col=c('black','red'), legend = c('male','female'), bty='n')
```



Discussion

For *theGlm* object, the "case" group is defined to be the pedestrians involved in motor vehicle accidents with fatal injuries and the "control" group is defined to be ones with slight injuries. The intercept represents 26-35 year-old, male pedestrians under day light condition with fine weather and no high winds who are involved in motor vehicle accidents with either fatal or slight injuries. The covariates are sex (categorical with levels Female and Male, male is the reference level), age (categorical with levels 0to5, 6to10, ... etc and 26to35 is the reference level), Light Conditions (categorical with levels darkness or daylight with lights lit, unlit, no lighting or lighting unknown and the daylight coniditon is our reference), and weather conditions (categorical with levels of raining, snowing, fine, fog or mist, high winds, or no high winds, and fine weather and no high winds is our reference).

For *theGlmInt* object, the "case" group is defined to be the pedestrians involved in motor vehicle accidents with fatal injuries and the "control" group is defined to be ones with slight injuries. The intercept represents 26-35 year-old, male pedestrians under day light condition with fine weather and no high winds who are involved in motor vehicle accidents with either fatal or slight injuries. The covariates are sex (categorical with levels Female and Male, male is the reference level), age (categorical with levels 0to5, 6to10, ... etc and 26to35 is the reference level), Light Conditions (categorical with levels darkness or daylight with lights lit, unlit, no lighting or lighting unknown and the daylight coniditon is our reference), weather conditions (categorical with levels of raining, snowing, fine, fog or mist, high winds, or no high winds, and fine weather and no high winds is our reference), and also the interaction terms of sex and age.

Analysis of Hypothesis

In order to address the question of interest, we should consider model 2 as this model accounts for the potential interaction between age and sex (i.e., rather than only considering a particular age group or a

particular sex individually, it makes more sense to consider a given age along with that age group's sex since different sexes in the same age group might vary greatly). This approach is also reinforced by the fact that only 3 of the age and sex interaction groups have "1" in the C.I., indicating a potential to have no effect.

Hence, based on the data and results from model 2, it is possible to conclude that our findings are consistent to the fact that women tend to be, on average, safer as pedestrians than men. This is also shown by model 2's point estimate of an odds ratio 0.58 when sex = female, which is lower compared to the implicit reference level of sex=male with the odds ratio of 1. The fact that this estimate's confidence interval doesn't contain a "1", it is possible to conclude that there is a significant effect of sex (of females) to be less likely to be involved in an accident, making them "safer" compared to men.

However, this is not true/inconclusive for particularly women as teenagers and in early adulthood. If we define teenagers to early adulthood by the age of 16 to 35, we can see females in the first age group of 16 to 20 has a higher odds ratio of 1.16 (which is greater than "1" for the reference group, and "1" is also not in its CI, indicating that we are 95% confident that it has a significant effect). For the age group of 21 to 25, since "1" is in our confidence interval, we cannot be certain that this factor is significant. Fianlly, for the age group of 26 to 35 (which is our reference level), the odds ratio estimate is implicit to be 1. All in all, our findings do not support the fact that females are (significantly) safer particularly as teenagers and in early adulthood, but women do tend to be safer on average.

Hypothesis 2: It is well established that women are generally more willing to seek medical attention for health problems than men, and it is hypothesized that men are less likely than women to report minor injuries caused by road accidents. We want to test that whether or not the control group is a valid one for assessing whether women are on average better at road safety than man.

Firstly, the hypothesis given is that men are less likely than women to report minor injuries caused by road accidents. Hence, based on figure 2, we can see that male does have a higher probability of being a case in baseline conditions (daylight, fine no winds) in *theGlmInt* model. In other words, the data given supports the fact that men are more likely to be involved in accidents which are fatal, which could possibly reflect the fact that men are less likely than women to report minor injuries. However, there is a problem in the control group for assessing whether women are on average better at road safety than man, since some of the underlying assumptions of case-control studies could be violated such as the inclusion in the study might depend on the covariates of sex. For example, men and women might have different, subjective opinions on what is considered "fatal" and "slight" for injuries (due to perhaps physical reasons, where men might be more likely to describe more injuries as "slight" compared to females), which could in turn affect the results as our case control study could be based on the subject's reporting. Also, men might also be less likely to go directly to a hospital after either a fatal or slight injury (so these men won't be in our data in the first place), resulting in an under-representation of such men. Hence, due to these reasons of the control group, it might be inaccurate to generalize our findings of whether or not women are on average better at road safety than man.