CatData HW5

Matthew Vanaman

03-29-2019

3,20, 3,21, 3.22

3.19

(a)

Confirm the sizes of the deviances:

```
TrainCollisions ~ 1
```

```
[1] 35.1
```

```
TrainCollisions ~ Year
```

```
[1] 23.5
```

Get the difference in deviances and residual df between the models:

```
mod_intercept_only$deviance - mod_with_time$deviance
```

```
[1] 11.6
```

```
mod_intercept_only$df.residual - mod_with_time$df.residual
```

```
[1] 1
```

The deviance approximates a chi-squared distribution when samples are large. Get the p-value for $\Delta\chi^2$:

```
pchisq(11.60185, df=1, lower.tail = FALSE)
```

```
[1] 0.000659
```

(b)

Get the likelihood χ^2, or z^2, and p-value with 1 degree of freedom:

```
(-0.0337/ 0.0130)^2
```

```
[1] 6.72
```

```
pchisq(6.720059, df=1, lower.tail = FALSE)
```

```
[1] 0.00953
```

(c)

The confidence intervals from the negative binomial model are around an additive, linear effect (because negative binomials are a special mixture of Poisson distributions). To get confidence intervals around a multiplicative effect, you exponentiate the intervals:

```
c(exp(-0.060), exp(-0.008))
```

```
[1] 0.942 0.992
```

3.20

(a)

Answer:

Age	DeathRate_nonSmokers	DeathRate_Smokers	Ratio
35-44	0.11	0.61	0.17
45-54	1.12	2.40	0.47
55-64	4.90	7.20	0.68
65-74	10.83	14.69	0.74
75-84	21.20	19.18	1.11

Deathrates for smokers and non-smokers alike increase with age, as one would expect. Of interest is the fact that the ratio of death rats also increases, such that as people age, the death rates between those who smoke and those who don't approaches a 1:1 ratio (which would mean no effect of smoking on death rate beyond age) before exceeding a 1:1 ratio after age 75. This implies curvilinear relationship between smoking and age. You stand greater risk of death up until you are 75, but if you make it to 75 you may benefit from being smoker.

Work:
Calculations for deathrates:

nonsmokers

```
2/(18793/1000) # age 35-44
```

```
[1] 0.106
```

```
12/(10673/1000) # age 45-54
```

```
[1] 1.12
```

```
28/(5710/1000) # age 55 - 64
```

```
[1] 4.9
```

```
28/(2585/1000) # age 65 - 74
```

```
[1] 10.8
```

```
31/(1462/1000) # age 75 - 84
```

```
[1] 21.2
```

smokers

```
32/(52407/1000) # age 35-44
```

```
[1] 0.611
```

```
104/(43248/1000) # age 45-54
```

```
[1] 2.4
```

```
206/(28612/1000) # age 55 - 64
```

```
[1] 7.2
```

```
186/(12663/1000)  # age 65 - 74
```

```
[1] 14.7
```

```
102/(5317/1000) # age 75 - 84
```

```
[1] 19.2
```

ratio

```
2/(18793/1000)/(32/(52407/1000))  # age 35-44
```

```
[1] 0.174
```

```
12/(10673/1000)/(104/(43248/1000)) # age 45-54
```

```
[1] 0.468
```

```
28/(5710/1000)/(206/(28612/1000)) # age 55 - 64
```

```
[1] 0.681
```

```
28/(2585/1000)/(186/(12663/1000)) # age 65 - 74
```

```
[1] 0.737
```

```
31/(1462/1000)/(102/(5317/1000)) # age 75 - 84
```

```
[1] 1.11
```

(b)

Age	Smoking	PersonYears	Deaths
35-44	no	18793	2
35-44	yes	52407	32
45-54	no	10673	12
45-54	yes	43248	104
55-64	no	5710	28
55-64	yes	28612	206
65-74	no	2585	28
65-74	yes	12663	186
75-84	no	1462	31
75-84	yes	5317	102

The specified model would be:

```
Deaths ~ Age + Smoking
```

with Age and Smoking as dummy variables. In other words:

X.Intercept.	Age45.54	Age55.64	Age65.74	Age75.84	Smokingyes
1	0	0	0	0	0
1	0	0	0	0	1
1	1	0	0	0	0
1	1	0	0	0	1
1	0	1	0	0	0
1	0	1	0	0	1
1	0	0	1	0	0
1	0	0	1	0	1
1	0	0	0	1	0
1	0	0	0	1	1

A change in the ratio of deaths between smokers and non smokers across levels of age would suggest an interaction (the effect of smoking does vary by level of age). This goes the other way too: main-effects only model assumes that the effect of age does not vary across categories of smoking (yes or no). By leaving out the interaction term, you are assuming the model is complete without it, which in our case is probably false because the ratio does change across age levels.

(c)

You can see from the ratios in part (a) that they do linearly increase with age, which suggests an interaction.

```
Deaths ~ Age * Smoking
```

term	estimate	std.error	statistic	p.value
(Intercept)	-9.148	0.707	-12.94	0.000
Age45-54	2.358	0.764	3.09	0.002
Age55-64	3.830	0.732	5.23	0.000
Age65-74	4.623	0.732	6.32	0.000
Age75-84	5.295	0.730	7.26	0.000
Smokingyes	1.747	0.729	2.40	0.017
Age45-54:Smokingyes	-0.987	0.790	-1.25	0.212
Age55-64:Smokingyes	-1.363	0.756	-1.80	0.071
Age65-74:Smokingyes	-1.442	0.757	-1.91	0.057
Age75-84:Smokingyes	-1.847	0.757	-2.44	0.015

In poisson regression with dummy variable predictors, the coeficient of each level of age shows the unique effect of that level of age against the reference category. Each level of age (x) is coded as 1 when all the others are 0. As you move up in age, the beta coefficient increases, meaning the log ratio (the output of this model) will also increase. The interaction terms also linearly increase with age. Therefore, if you aggregate the effect of age, of smoking, and of being both a certain age and smoking, you see that as you move up each level of age, the output of the model will increase linearly.
You can also estimate for the non-smokers as well. In that case, you would add the effect of each age level while omitting the effect of smoking and smoking + age (they would be cancel out because they are coded as 0). This will also be linear because the effect of age by itself is linear.

(d)

```
Deaths ~ Age + Smoking
```

term	estimate	std.error	statistic	p.value
(Intercept)	-7.919	0.192	-41.30	0.000
Age45-54	1.484	0.195	7.61	0.000
Age55-64	2.628	0.184	14.30	0.000
Age65-74	3.351	0.185	18.13	0.000
Age75-84	3.700	0.192	19.25	0.000
Smokingyes	0.355	0.107	3.30	0.001

We could rank ages by scoring them as 1, 2, 3, 4, and 5.

```
Deaths ~ Age * Smoking
```

term	estimate	std.error	statistic	p.value
(Intercept)	-8.867	0.306	-29.01	0.000
Age	1.047	0.077	13.52	0.000
Smokingyes	1.284	0.326	3.94	0.000
Age:Smokingyes	-0.249	0.084	-2.98	0.003

The deviances and degrees of freedom, respectively, for each model are:

```
Deaths ~ Age * Smoking
```

```
[1] 59.9
```

```
[1] 6
```

```
Deaths ~ Age + Smoking
```

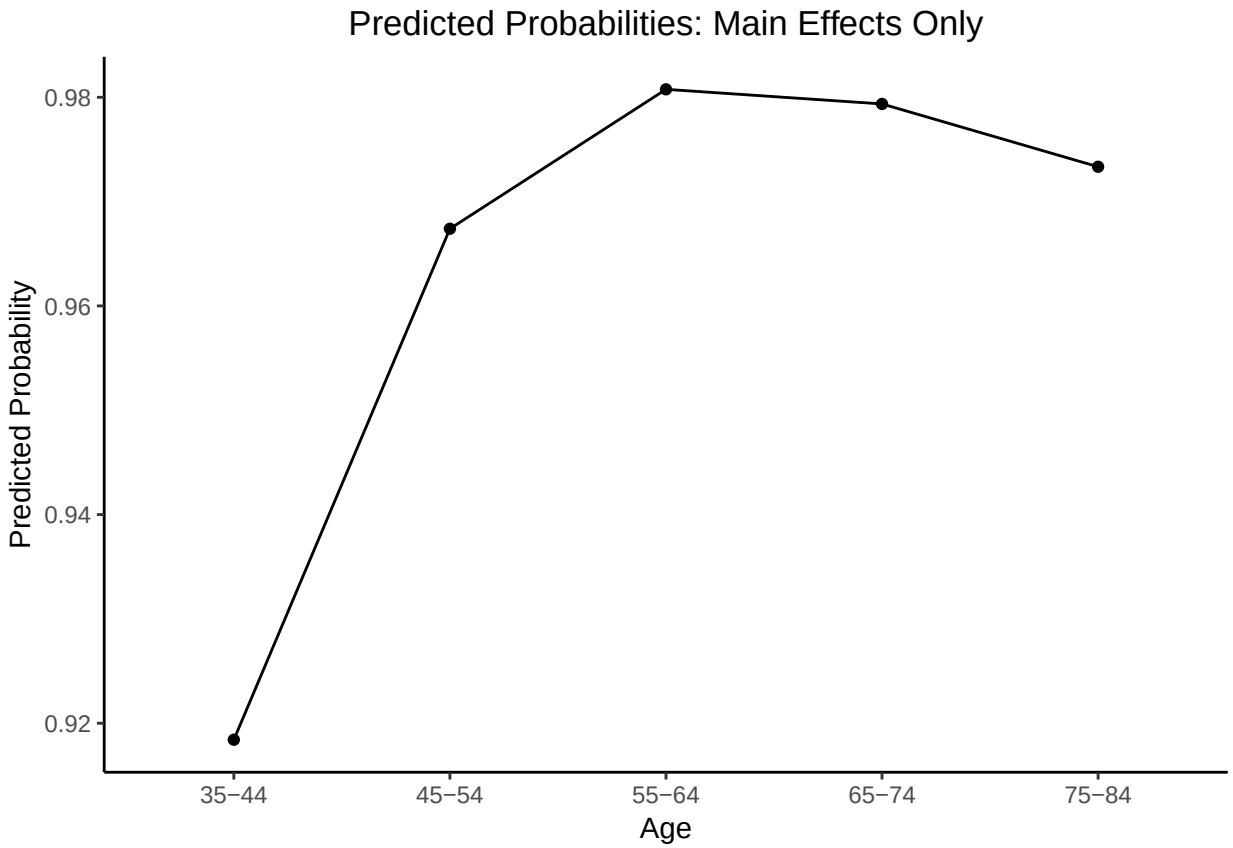
```
[1] 12.1
```

```
[1] 4
```

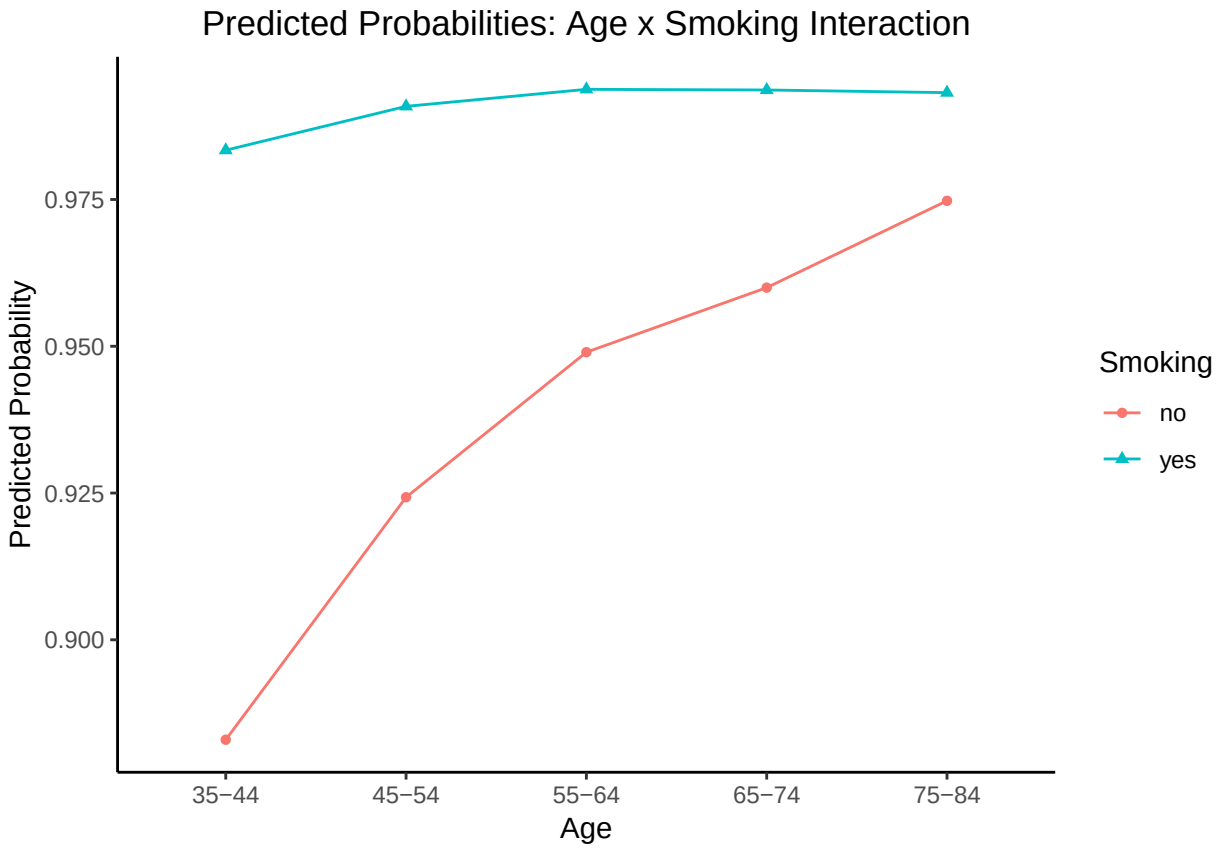
The deviance approximates a chi-squared distribution when samples are large. The p-value for $\Delta\chi^2$ is:

```
[1] 0.0000000000485
```

We're firmly rejecting the null here, meaning that the interaction model has better fit to the data. It also makes some sense of a weird trend in the data. Before you include smoking, it seems Age has a quadratic relationship with the predicted probability of death:



That is not what one would expect. The data make more sense when we include smoking: there is a clean linear relationship between age and the probability of death for non-smokers, but the probability of death is high for all smokers regardless of age group. This is much more likely.


Predicted Probabilities: Age x Smoking Interaction
Predicted Probability
0.975
0.950
0.925
0.900
35−44
45−54
55−64
65−74
75−84
Age
Smoking
no
yes