CatData HW5

Matthew Vanaman

03-31-2019

3.19

(a)

Confirm the sizes of the deviances and degrees of freedom, respectively, for each model:

```
TrainCollisions ~ 1
```

```
35.1 28
```

```
TrainCollisions ~ Year
```

```
23.5 27
```

The difference in deviances and residual df, respectively, between the models is:

```
11.6
```

```
1
```

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

```
0.000659
```

With this evidence, the intercept-only term has poorer fit. Collision rates are probably not at a constant rate, but instead vary as a function of time.

(b)

The Wald test is the difference between $\hat{\beta}$ and 0 in standard error units. The wald statistic z^2 is:

```
6.72
```

The z^2 statistic approximates χ^2. Thus the p-value for a chi-squared statistic this large with 1 degree of freedom is:

```
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 and get:

```
( 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.

(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:

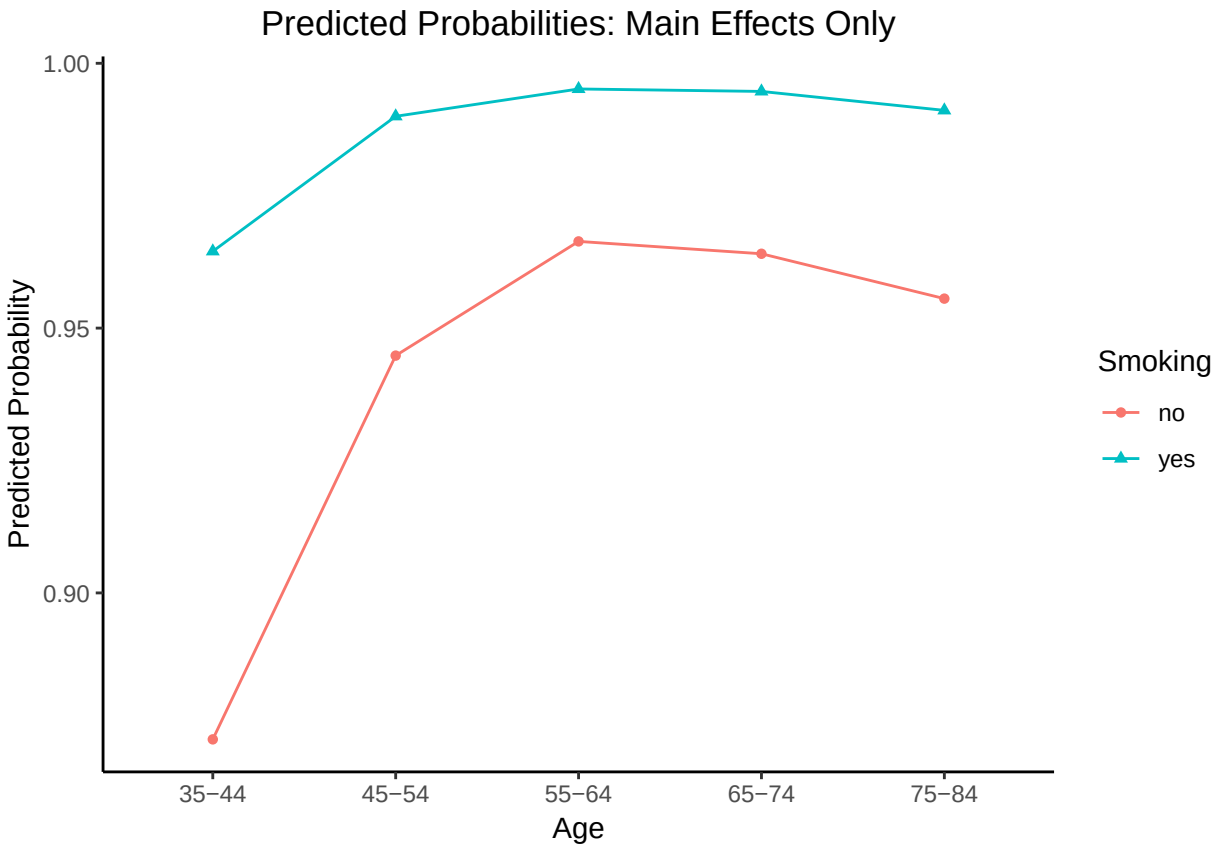
```
Deaths ~ Age * Smoking

59.9

6

Deaths ~ Age + Smoking

12.1

4
```

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

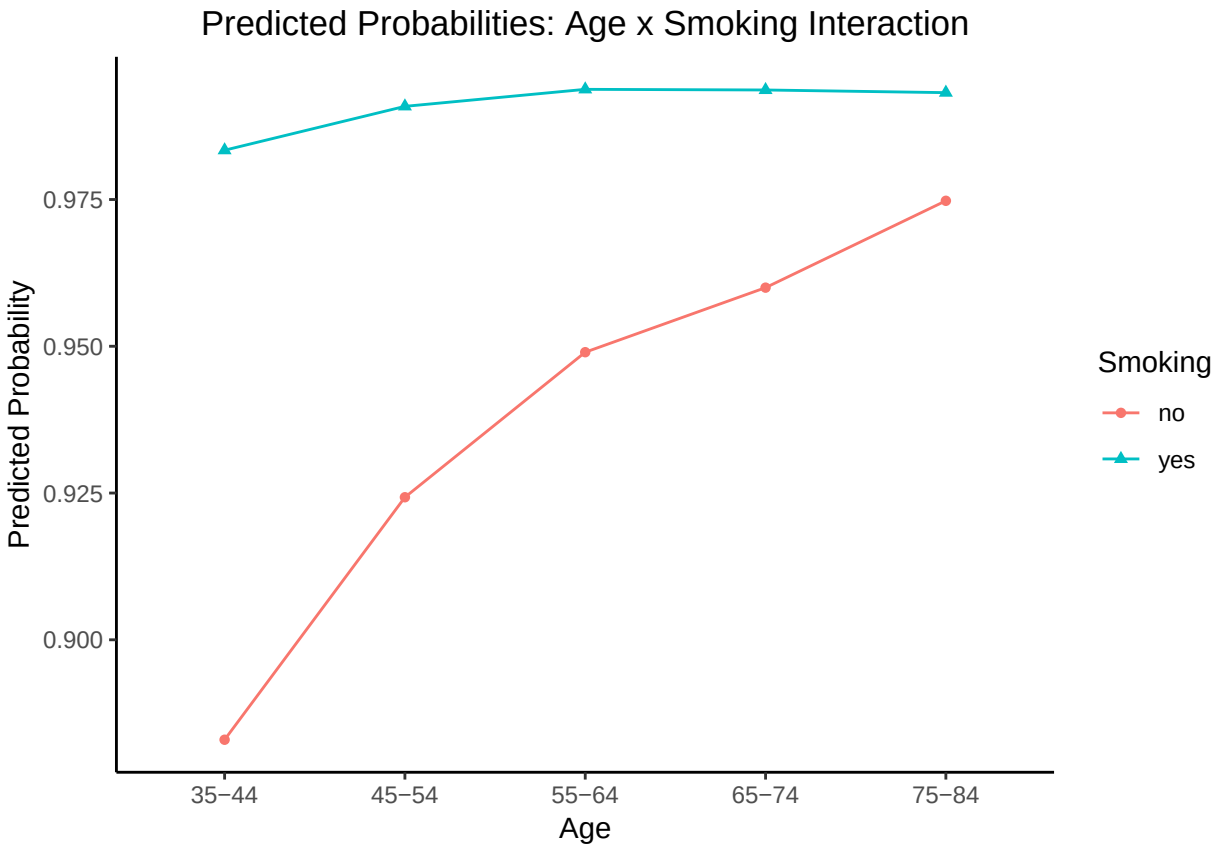
```
0.0000000000485
```

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.



3.21

Fitting model with t: you could fit this model because holding x constant, μ will increase (or decrease) proportionately with t. If μ were gun deaths in a city and t were population size of a city, gun deaths will rise with population size. If one is interested in this effect only, one doesn't need any other variables in the model. This model would be $\mu = t$, where the mean is modeled directly from an explanatory variable.

Fitting a model with tx: If μ were gun deaths, t were population size, and x were gun ownership, you could fit a gun ownership * population size model where there is a multiplicative effect such that as population size increases, so too does the number of gun owners. In this case, you'd have the effect of population itself as the reference variable and a population by gun ownership interaction term. This might be the model you'd use if you wanted a gun ownership variable that was "adjusted" for population size: $\mu = t + \beta(tx)$.

3.22

(a)

True. With ANOVA models, the researcher is interested in modeling the mean itself as a direct function of the explanatory variables, which is exactly what the identity link does (models the mean directly).

(b)

It is true that GLMs allow Y to have non-normal distributions, and also true that one can model functions of Y instead of just Y itself (e.g. the logit function). However, it is false that Y must be constant at all values of x_j in order to get ML estimates. In fact, GLMs are often working with data that came from exponential distributions, which by definition means the magnitude unit change in Y for every change x_{ij} varies across levels of x_j. This is why we take derivatives: if the regression were a straight line, we would not need to determine the slope at each level of x_j (instantaneous rate of change) in order to predict a given Y. In GLMs, the slope itself is a determined by the levels of x_{ij}.

(c)

False. In principle yes, the Pearson residual distribution approximates the standard normal. In practice though, the residuals of the *estimated* means are less variable than the standard normal.

Appendix

3.19

(a)

```
trainData <- as.data.frame(matrix(c(2003, 2002, 2001, 2000, 1999, 1998, 1997,
                                    1996, 1995, 1994, 1993, 1992, 1991, 1990,
                                    1989, 1988, 1987, 1986, 1985, 1984, 1983,
                                    1982, 1981, 1980, 1979, 1978, 1977, 1976,
                                    1975,

                                    518, 516, 508, 503, 505, 487, 463, 437,
                                    423, 415, 425, 430, 439, 431, 436, 443,
                                    397, 414, 418, 389, 401, 372, 417, 430,
                                    426, 430, 425, 426, 436,

                                    0, 1, 0, 1, 1, 0, 1, 2, 1, 2, 0, 1, 2,
                                    1, 4, 2, 1, 2, 0, 5,2, 2, 2, 2, 3, 2, 1,
                                    2, 5,

                                    3, 3, 4, 3, 2, 4, 1, 2, 2, 4, 4, 4, 6, 2,
                                    4, 4, 6, 13, 5, 3, 7, 3, 2, 2, 3, 4, 8,
                                    12, 2),

                                  nrow = 29, ncol = 4))
colnames(trainData) <- c("Year", "TrainKm",
                         "TrainCollisions",
                         "TrainRoadCollisions")

mod_intercept_only <- glm(TrainCollisions ~ 1,
                          offset = log(TrainKm),
                          data=trainData,
                          family = poisson)

mod_with_time <- glm(TrainCollisions ~ Year,
                     offset = log(TrainKm),
                     data = trainData,
                     family = poisson)
```

Confirm the sizes of the deviances and degrees of freedom, respectively, for each model:

```
mod_intercept_only$formula
```

```
TrainCollisions ~ 1
```

```
cat(mod_intercept_only$deviance, mod_intercept_only$df.residual)
```

```
35.1 28
```

```
mod_with_time$formula
```

```
TrainCollisions ~ Year
```

```
cat(mod_with_time$deviance, mod_with_time$df.residual)
```

```
23.5 27
```

The difference in deviances and residual df, respectively, between the models is:

```
cat(mod_intercept_only$deviance - mod_with_time$deviance)
```

```
11.6
```

```
cat(mod_intercept_only$df.residual - mod_with_time$df.residual)
```

```
1
```

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

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

```
0.000659
```

With this evidence, the intercept-only term has poorer fit. Collision rates are probably not at a constant rate, but instead vary as a function of time.

(b)

Work:

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

```
6.72
```

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

```
0.00953
```

(c)

Work:

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

```
0.942 0.992
```

3.20

(a)

Work:

```
deathrates <- cbind(data.frame(c("35-44", "45-54", "55-64",
                                 "65-74", "75-84")),

                    matrix(c(0.1064226, 1.124332, 4.903678,
                             10.83172, 21.20383,

                             0.6106055, 2.404735 ,7.199776
                             ,14.68846 ,19.18375,

                             0.1742903 ,0.4675493 ,0.6810875,
                             0.7374306, 1.105302),
                    ncol = 3, nrow = 5))
colnames(deathrates) <- c("Age",
                          "DeathRate_nonSmokers",
                          "DeathRate_Smokers",
                          "Ratio")
deathrates[,-1] <- round(deathrates[,-1], 2)
knitr::kable(deathrates)
```

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)

Work:

```
deadData <- data.frame(Age = c("35-44", "35-44", "45-54", "45-54",
                               "55-64", "55-64", "65-74", "65-74",
                               "75-84", "75-84"),

                           Smoking = c("no", "yes", "no", "yes", "no",
                                       "yes", "no", "yes", "no", "yes"))

mat <- as.data.frame(matrix(c(18793, 52407, 10673, 43248, 5710,
                              28612, 2585, 12663, 1462, 5317,

                              2, 32, 12, 104, 28,
                              206, 28, 186, 31, 102),
```

```
                              nrow = 10, ncol = 2))

colnames(mat) <- c("PersonYears", "Deaths")
deadData <- as.data.frame(cbind(deadData, mat))
knitr::kable(deadData)
main_effects_only <- glm(Deaths ~ Age + Smoking,
                         offset = log(PersonYears),
                         data = deadData,
                         family = poisson)
```

The specified model would be:

```
print(main_effects_only$formula)
```

```
Deaths ~ Age + Smoking
```

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

```
knitr::kable(tidy(model.matrix(main_effects_only)))
```

(c)

Work:

```
interaction <- glm(Deaths ~ Age * Smoking,
                   offset = log(PersonYears),
                   data = deadData,
                   family = poisson)
print(interaction$formula)
knitr::kable(tidy(interaction))
```

(d)

Work:

```
print(main_effects_only$formula)
knitr::kable(tidy(main_effects_only))
```

Scored Age model:

```
deadData2 <- data.frame(Age = c(1, 1,2, 2, 3, 3, 4, 4, 5, 5),

                                Smoking = c("no", "yes", "no", "yes", "no",
                                            "yes", "no", "yes", "no", "yes"))

mat2 <- as.data.frame(matrix(c(18793, 52407, 10673, 43248, 5710, 28612,
                               2585, 12663, 1462, 5317,

                               2, 32, 12, 104, 28, 206, 28, 186, 31, 102
                               ), nrow = 10, ncol = 2))
colnames(mat2) <- c("PersonYears", "Deaths")
deadData2 <- as.data.frame(cbind(deadData2, mat2))
```

```r
# get model
scored_interaction <- glm(Deaths ~ Age * Smoking,
                          offset = log(PersonYears),
                          data = deadData2,
                          family = poisson)

print(scored_interaction$formula)
```

```r
# get output
knitr::kable(tidy(scored_interaction))
```

```r
scored_interaction$formula
```

```
Deaths ~ Age * Smoking
```

```r
cat(scored_interaction$deviance)
```

```
59.9
```

```r
cat(scored_interaction$df.residual)
```

```
6
```

```r
main_effects_only$formula
```

```
Deaths ~ Age + Smoking
```

```r
cat(main_effects_only$deviance)
```

```
12.1
```

```r
cat(main_effects_only$df.residual)
```

```
4
```

```r
# get p-value
cat(pchisq(47.5, df=2, lower.tail = FALSE))
```

```
0.0000000000485
```

```r
# Main-Effects Only Plot
poisson.link.pred <- predict(main_effects_only, type="link")
poisson.prob.pred <- gtools::inv.logit(poisson.link.pred)
scored_interaction %>%
  ggplot() +
  aes(x = Age,
      y = poisson.prob.pred,
      color = Smoking,
      shape = Smoking) +
  stat_summary(fun.y = mean, geom = "point") +
```

```
  stat_summary(fun.y = mean, geom = "line") +
  ggtitle("Predicted Probabilities: Main Effects Only") +
  ylab("Predicted Probability")  +
  scale_x_discrete("Age",
                   labels = c("35-44", "45-54", "55-64", "65-74", "75-84"),
                   limits = 1:5) +
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5))
```

```
# Model with interaction term
poisson.link.pred <- predict(scored_interaction, type="link")
poisson.prob.pred <- gtools::inv.logit(poisson.link.pred)
scored_interaction %>%
  ggplot() +
  aes(x = Age,
      color = Smoking,
      group = Smoking,
      shape = Smoking,
      y = poisson.prob.pred) +
  stat_summary(fun.y = mean, geom = "point") +
  stat_summary(fun.y = mean, geom = "line") +
  ggtitle("Predicted Probabilities: Age x Smoking Interaction") +
  ylab("Predicted Probability") +
  scale_x_discrete("Age",
                   labels = c("35-44", "45-54", "55-64", "65-74", "75-84"),
                   limits = 1:5) +
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5))
```

3.21

Work N/A

3.22

(a)

Work N/A

(b)

Work N/A

(c)

Work N/A