class_3_10

Lindsey Greenhill

3/10/2021

Exercises 5.7 and 5.8

Question 7

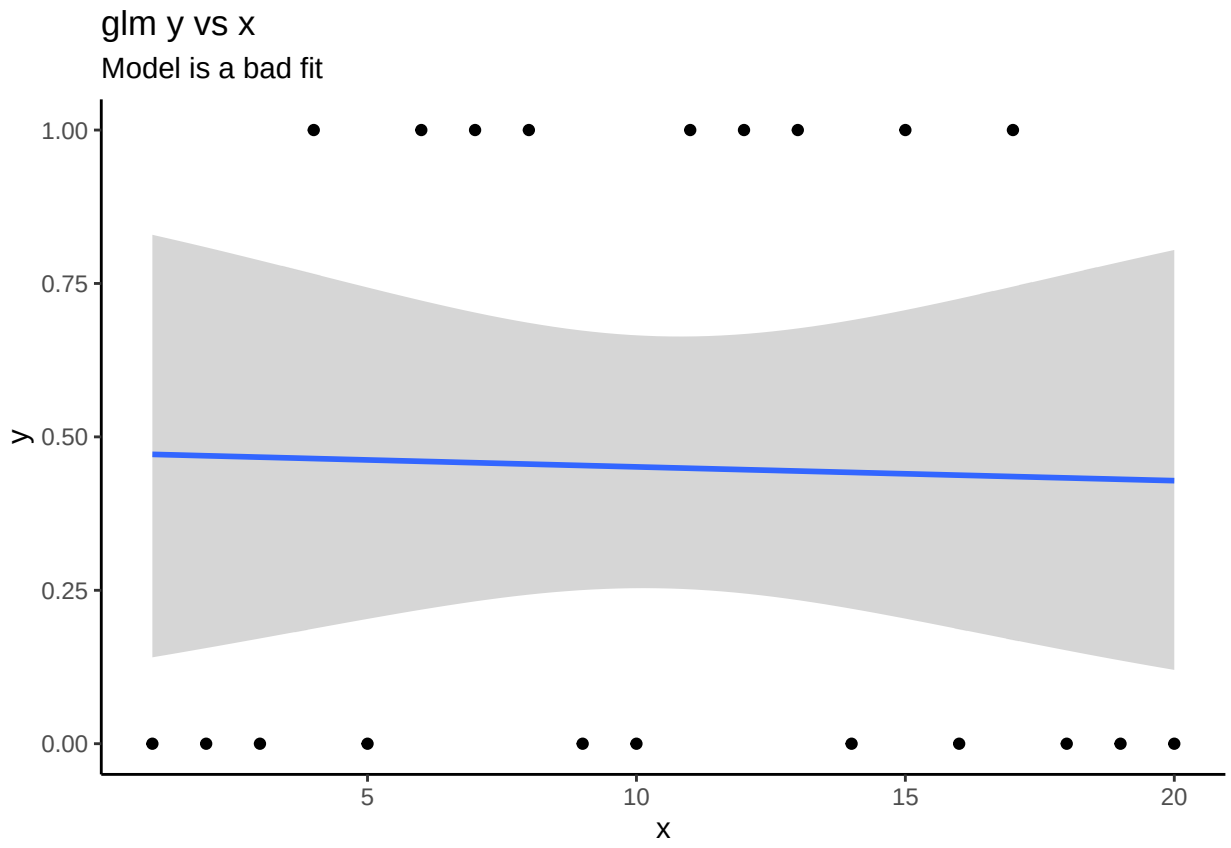
7. Limitations of logistic regression: consider a dataset with n = 20 points, a single predictor x that takes on the values 1,. . . , 20, and binary data y. Construct data values y1,. . . , y20 that are inconsistent with any logistic regression on x. Fit a logistic regression to these data, plot the data and fitted curve, and explain why you can say that the model does not fit the data.

```
set.seed(123)

# assigning values randomly for y so there isn't a relationship

df_7 <- tibble(x = 1:20,
               y = sample(x = c(0,1),
                          size = 20,
                          replace = T))

# doing link = logit so it makes results easier to interpret

fit <- glm(y ~ x, data = df_7, family = binomial(link = "logit"))

an <- anova(fit)

tibble(
  "Null Dev" = 27.53,
  "Resid. Dev" = an$"Resid. Dev"[1]
) %>%
  gt()
```

Null Dev	Resid. Dev
27.53	27.52555

```
# it is not a good fit because the residual deviance is only 1.6 lower, so x is
# only random noise and not a predictor

ggplot(df_7, aes(x = x, y = y)) +
```

```
  geom_point() +
  geom_smooth(method = glm,
              method.args = list(family = binomial)) +
  labs(title = "glm y vs x",
       subtitle = "Model is a bad fit") +
  theme_classic()
```




```
# the plot shows that the model does not fit the data. This makes sense

pchisq(deviance(fit), df.residual(fit), lower = FALSE)
```

```
## [1] 0.06988241
```

```
# we are concerned with the upper tail. That is 95% and above. The value is so
# low of the p chis sq so that indicates a bad fit.
```

- The pchisq value for the model is .067. Because the significance comes from the higher tail of the chisq distribution, this low score indicates that the model is not a good fit.
- The plot of the data and fit visualizes the poor relationship between the model curve and the data points.
- The null deviance is 27.53 and the residual deviance is 27.51. The difference in these two values is less than 1, implying that the x variable has no explanatory power.

Question 8

8. Building a logistic regression model: the folder rodents contains data on rodents in a sample of New York City apartments. (a) Build a logistic regression model to predict the presence of rodents (the variable rodent2 in the dataset) given indicators for the ethnic groups (race). Combine categories as appropriate. Discuss the estimated coefficients in the model. (b) Add to your model some other potentially relevant predictors describing the apartment, building, and community district. Build your model using the general principles explained in Section 4.6. Discuss the coefficients for the ethnicity indicators in your model.

Part a

```
# recoding the values so they will be easier to interpret.

rodents$race <-
  dplyr::recode(
    rodents$race,
    "1" = "White",
    "2" = "Black",
    "3" = "Puerto Rican",
    "4" = "Other Hispanic",
    "5" = "Asian/Pacific Islander",
    "6" = "Amer-Indian/Native Alaskan",
    "7" = "Two or More Races"
  )

# Not sure what categories would be appropriate to combine so I'm not going to
# combine any.

rodents <- rodents %>%
  mutate(race = as_factor(race))

# making white the first level in the factor variable

rodents$race <- relevel(rodents$race, "White")

# creating model with only race predictor

# residual deviance: 1518.7, df 1515. null deviance 1672.2

mod_1 <- glm(rodent2 ~ race, data = rodents,
             family = binomial(link = "logit"))


stargazer(mod_1, type = "latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Mar 14, 2021 - 23:19:04

Coefficients

I will be using the divide by 4 rule, which is a rough but quick interpretation of the coefficients.

	Dependent variable:
	rodent2
racePuerto Rican	1.449***
	(0.214)
raceBlack	1.536***
	(0.169)
raceOther Hispanic	1.867***
	(0.187)
raceAsian/Pacific Islander	0.400
	(0.292)
raceAmer-Indian/Native Alaskan	2.152***
	(0.826)
raceTwo or More Races	0.766
	(0.801)
Constant	−2.152***
	(0.128)
Observations	1,522
Log Likelihood	−759.350
Akaike Inf. Crit.	1,532.700
Note:	*p<0.1; **p<0.05; ***p<0.01

- Intercept: The intercept is the baseline of comparison for the other races vs white households on average, but it doesn't tell us how likely it is that white households have rodents in their house.
- Black: compared to white households, black households are roughly 38% more likely than white households to live in a rodent-infested building on average. This coefficient is statistically significant.
- Puerto Rican: compared to white households, puerto rican households are roughly 36% more likely than white households to live in a rodent-infested building on average. This coefficient is statistically significant.
- Other Hispanic: compared to white households, other hispanic households are roughly 46% more likely than white households to live in a rodent-infested building on average. This coefficient is statistically significant.
- Pacific Islander: compared to white households, pacific islander households are roughly 10% more likely than white households to live in a rodent-infested building on average. This coefficient is not statistically significant.
- American Indian/Native Alaskan: compared to white households, Native Alanskan or American Indian households are roughly 54% more likely than white households to live in a rodent-infested building on average. This coefficient is statistically significant.
- Two or More Races: compared to white households, two or more race households are roughly 19% more likely than white households to live in a rodent-infested building on average. This coefficient is not statistically significant.

Part b

```
# creating new model that takes race, dilapidated status of building, income
# level, and regext, which tells us whether there is a regular exterminator
# service

mod_2 <- glm(rodent2 ~ race + dilap + regext + totincom2 , data = rodents,
             family = binomial(link = "logit"))

# residual deviance: 1326, df = 1331. null deviance 1672.2

stargazer(mod_1, mod_2, type = "latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Mar 14, 2021 - 23:19:04

Discussion of Ethnicity Coefficients

Adding new explanatory variables changes the values of the ethnicity coefficients in varying degrees. I am using the divide by 4 rule for convenience.

- Black:compared to white households, black households are roughly 35% more likely than white households to live in a rodent-infested building on average, holding all else constant. This is a smaller effect than the one we saw in model 1 (38%). This coefficient is statistically significant.
- Puerto Rican: compared to white households, puerto rican households are roughly 30% more likely than white households to live in a rodent-infested building on average, holding all else constant. This is a smaller effect than the one we saw in model 1 (36%). This coefficient is statistically significant.

	Dependent variable:	
	rodent2	
	(1)	(2)
racePuerto Rican	1.449***	1.195***
	(0.214)	(0.230)
raceBlack	1.536***	1.408***
	(0.169)	(0.181)
raceOther Hispanic	1.867***	1.767***
	(0.187)	(0.204)
raceAsian/Pacific Islander	0.400	0.305
	(0.292)	(0.303)
raceAmer-Indian/Native Alaskan	2.152***	1.947**
	(0.826)	(0.830)
raceTwo or More Races	0.766	0.840
	(0.801)	(0.815)
dilap		1.202***
		(0.297)
regext		−0.219
		(0.138)
totincom2		−0.00001***
		(0.00000)
Constant	−2.152***	−1.661***
	(0.128)	(0.176)
Observations	1,522	1,341
Log Likelihood	−759.350	−662.990
Akaike Inf. Crit.	1,532.700	1,345.981
Note:	*p<0.1; **p<0.05; ***p<0.01	

- Other Hispanic: compared to white households, other hispanic households are roughly 44% more likely than white households to live in a rodent-infested building on average, holding all else constant. This is a smaller effect than the one we saw in model 1 (46%). This coefficient is statistically significant.
- Pacific Islander: compared to white households, pacific islander households are roughly 7% more likely than white households to live in a rodent-infested building on average, holding all else constant. This is a smaller effect than the one we saw in model 1 (10%). This coefficient is not statistically significant.
- American Indian/Native Alaskan: compared to white households, Native Alaskan or American Indian households are roughly 48% more likely than white households to live in a rodent-infested building on average, holding all else constant. This is a smaller effect than the one we saw in model 1 (54%). This coefficient is statistically significant.
- Two or More Races: compared to white households, two or more race households are roughly 21% more likely than white households to live in a rodent-infested building on average, holding all else constant. This is a larger effect than the one we saw in model 1 (19%). This coefficient is not statistically significant.

Overall, with the exception of the two or more races variable, the ethnicity coefficients are smaller in model 2 compared to model 1. This suggests that other predictors in model 2 added explanatory power. The statistical significance did not change for any of the coefficients.

Discussion of Deviances

The table below shows the null and residual deviances for both models. The residual deviance is significantly lower for model 2 than model 1, also suggesting that the new variables have explanatory power and are not just "noise."

```
tibble("Model" = c("mod_1", "mod_2"),
       "Null Deviance" = 1672.2,
       "Residual Deviance" = c(1518.7, 1326)) %>%
  gt()
```

Model	Null Deviance	Residual Deviance
mod_1	1672.2	1518.7
mod_2	1672.2	1326.0

Interpretation of other coefficients in mod_2

- dilap: holding all else constant, a dilapidated building is roughly 30% more likely than a non dilapidated building to have rodents on average. This coefficient is statistically significant.
- regext: holding all else constant, buildings with regular exterminators are roughly 5% less likely to have rodents than buildings without regular exterminators on average. This coefficient is not statistically significant
- totincom2: this coefficient is very close to 0 (.00001), meaning that an increase in 1 dollar has a tiny effect in the likeliness to have rodents in a building.