Homework 08 solutions (draft)

11/6/2020

Problem 1

EDA (not required)

The below EDA shows us few things, but the plot visualization is clearest relative to future exploration we will be doing. We can observe the neither the rows nor the columns inform the viewer of the full relationship between such class structure and the number of breaks of the wool. Specifically note that Wool A appears to break more with loose (L) tension (compared to Wool B), but that this structure (that Wool A breaks more than Wool B) isn't observed / as strong across all levels of tension.

```
library(tidyverse)
#table(warpbreaks$wool, warpbreaks$tension) # all 9 values
xtabs(breaks ~ wool + tension, data = warpbreaks)
##
       tension
##
              М
                  Η
  wool
          L
##
      A 401 216 221
##
      B 254 259 169
warpbreaks %>% head
##
     breaks wool tension
## 1
         26
               А
                        L
## 2
         30
               А
                        L
## 3
         54
                        L
                A
## 4
         25
                А
                        L
## 5
         70
                        L
                А
## 6
         52
                A
                        L
warpbreaks %>%
  ggplot() +
  geom_histogram(aes(x = breaks)) +
 facet_wrap(wool ~ tension)
```



(a)

(a) No interaction term

Our anova model without an interaction term between wool and tension in predicting number of breaks doesn't fit the data to well. Although this anova model only provides us with (at most) 6 different fitted values for 54 observations, we can see that the lack of interaction terms impacts the our residual vs fitted plot, where we find that the true average value for each of the Wool + tension combinations does not exactly equal the predicted average (this can be seen by the smoothed regression line of the residual vs fitted not being 0 for all fitted values). For this anova, it's probably not the best to *over-interpret* the standard summary plots (due to the 6 fitted values). A similar observation relative to the residuals vs fitted can be seen in marginal plot, where the fitted curve doesn't match the raw average of the number of breaks verse predicted number of breaks. Note that the x-axis values are of the log(breaks) scale, whereas the y-axis is on the breaks scale.

Our "DHARMa" residual plots present a different way to see that our residuals still relationships with our predicted values (aka the 6 classes between Wool and tension), suggesting that the model that we fit doesn't create / simulate residuals similar to the truely observed value. The over-dispersion seen in the QQ plot of the ranks of the residuals makes sense and it suggests that the true residuals are more variable than the simulation model would expect (which does make sense relative to our observation that the model doesn't well capture the mean of each group).

It should be noted that I did not use the binned residual plot as it would show almost the same information as the regular residual vs fitted plot (relative to 6 bins). The binned residual plot may have captured the information in the summary's residual vs fitted and absolute residual vs fitted plots. We include the code to do so below just in case that would be useful - just commented out.

```
p_glm1 <- glm(breaks~wool+tension, data = warpbreaks,</pre>
             family = poisson)
summary(p_glm1)
##
## Call:
## glm(formula = breaks ~ wool + tension, family = poisson, data = warpbreaks)
##
## Deviance Residuals:
##
      Min
                1Q
                    Median
                                 ЗQ
                                          Max
## -3.6871 -1.6503 -0.4269 1.1902
                                      4.2616
##
## Coefficients:
##
              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 3.69196 0.04541 81.302 < 2e-16 ***
## woolB
              -0.20599
                          0.05157 -3.994 6.49e-05 ***
## tensionM
              -0.32132
                        0.06027 -5.332 9.73e-08 ***
## tensionH -0.51849
                        0.06396 -8.107 5.21e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
      Null deviance: 297.37 on 53 degrees of freedom
##
## Residual deviance: 210.39 on 50 degrees of freedom
## AIC: 493.06
##
## Number of Fisher Scoring iterations: 4
par(mfrow = c(2,2))
plot(p_glm1)
```





sim_resid1 <- simulateResiduals(p_glm1, plot = F)
plot(sim_resid1)</pre>





Residual vs. predicted Quantile deviations detected (red curves) Combined adjusted quantile test significa



Model predictions (rank transformed)

```
# library(arm)
# binnedplot(predict(p_glm1), resid(p_glm1))
```

(a) Interaction term

For the model with an interaction term (who's diagnostic plots are below), we find the model is able to capture all the different groupings of wool an tensions mean (but design). This is seen in the residual vs fitted and the fact that both smoothed fitted and true lines are very close to each other in overall marginal plot. Although this TA would discourage from interpreting the standard summary plots too much, we can observe that distribution of errors for the wool + tension group with average log(breaks) of 3.8 has a log-based residuals that are larger than other wool + tension groups (as seen in the residual vs fitted and absolute residual vs fitted plots). The difference between the variability of the residuals per each wool + tension group also seems to lead to overdispersion of the residuals relative to simulations draw from out fitted models (visualized in the "DHARMA" plots).

```
summary(p_glm2)
```

```
##
## Call:
##
  glm(formula = breaks ~ wool * tension, family = poisson, data = warpbreaks)
##
## Deviance Residuals:
##
       Min
                 1Q
                      Median
                                    ЗQ
                                           Max
## -3.3383 -1.4844 -0.1291
                               1.1725
                                         3.5153
##
  Coefficients:
##
##
                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)
                   3.79674
                              0.04994 76.030 < 2e-16 ***
                              0.08019 -5.694 1.24e-08 ***
## woolB
                  -0.45663
                                       -7.330 2.30e-13 ***
## tensionM
                  -0.61868
                              0.08440
## tensionH
                  -0.59580
                              0.08378
                                       -7.112 1.15e-12 ***
## woolB:tensionM 0.63818
                              0.12215
                                         5.224 1.75e-07 ***
## woolB:tensionH 0.18836
                              0.12990
                                         1.450
                                                  0.147
##
  ___
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##
       Null deviance: 297.37 on 53
                                     degrees of freedom
## Residual deviance: 182.31 on 48 degrees of freedom
## AIC: 468.97
##
## Number of Fisher Scoring iterations: 4
anova(p_glm1, p_glm2)
## Analysis of Deviance Table
##
## Model 1: breaks ~ wool + tension
## Model 2: breaks ~ wool * tension
##
    Resid. Df Resid. Dev Df Deviance
```





#library(arm)
#binnedplot(predict(p_glm2), resid(p_glm2))

simulateResiduals(p_glm2, plot = T)

DHARMa residual diagnostics



Expected

Model predictions (rank transformed)

Object of Class DHARMa with simulated residuals based on 250 simulations with refit = FALSE . See ?DI
##
Scaled residual values: 0 0.005054066 0.9226526 0 1 0.8865403 0.8222991 0 0.9974159 0.1130634 0.20563

(b)

(i)

As can be seen in the below table, the second poisson model with interaction terms has a lower AIC and BIC than the first model. If we remember that both of these score functions look like

 $-2 \cdot (\log \text{likelihood}) + \text{constant} \cdot (\text{number of parameters})$

then we can conclude the both AIC and BIC recommend going with the poisson model with interaction terms (aka lower AIC/BIC is better).

knitr::kable(df_aic_info)

	poisson model, no interaction	poisson model, interaction
AIC	493.0560	468.9692
BIC	501.0119	480.9031

(ii)

The LRT also suggests we should reject the null hypothesis that the model with an interaction only explains as much information as the the model without the interaction term. This would lead us to select the second model like AIC and BIC did.

```
diff_log_like <- -2 * (logLik(p_glm1) - logLik(p_glm2))
1-pchisq(diff_log_like, df = 4)</pre>
```

```
## 'log Lik.' 1.197798e-05 (df=4)
```

(c)

The below p-values suggest that the residuals in model 2 are over-dispersed and match the simulation test from DHARMa.

```
n <- nrow(warpbreaks)
pp2 <- length(coef(p_glm2))
e_y_2 <- predict(p_glm2, type = "response")
z_2 <- (warpbreaks$breaks - e_y_2)/ sqrt(e_y_2)
test_stat_2 <- sum(z_2^2)
pchisq(test_stat_2, n-pp2, lower.tail = F)
## [1] 2.926195e-17</pre>
```

```
testDispersion(p_glm2, alternative = "greater")
```



Simulated values, red line = fitted model. p-value (greater) = 0

```
##
   DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##
##
   simulated
##
## data: simulationOutput
## ratioObsSim = 1.9846, p-value < 2.2e-16</pre>
## alternative hypothesis: greater
(d)
p_glm2d <- glm(breaks~wool*tension, data = warpbreaks,</pre>
              family = quasipoisson)
summary(p_glm2d)
##
## Call:
##
  glm(formula = breaks ~ wool * tension, family = quasipoisson,
##
       data = warpbreaks)
##
## Deviance Residuals:
##
       Min
                 1Q
                      Median
                                    ЗQ
                                            Max
## -3.3383 -1.4844 -0.1291
                                1.1725
                                         3.5153
##
## Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
##
## (Intercept)
                   3.79674
                              0.09688 39.189 < 2e-16 ***
## woolB
                  -0.45663
                              0.15558 -2.935 0.005105 **
## tensionM
                  -0.61868
                              0.16374 -3.778 0.000436 ***
## tensionH
                  -0.59580
                              0.16253 -3.666 0.000616 ***
## woolB:tensionM 0.63818
                              0.23699
                                         2.693 0.009727 **
## woolB:tensionH 0.18836
                              0.25201
                                        0.747 0.458436
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
  (Dispersion parameter for quasipoisson family taken to be 3.76389)
##
##
##
       Null deviance: 297.37 on 53
                                      degrees of freedom
## Residual deviance: 182.31 on 48 degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

Fitting the second model using quasipoisson allows us to have a dispersion parameter that is not 1. The new model's dispersion parameter is now 3.7639, which is much larger than 1 - which does suggest that we had a problem with residuals being overdispersed in the original model.

Problem 2

(a)

Nothing is required to be turned in here.

l_glm_1976 <- glm(rep_pres_intent ~ ., data = nes_1976, family = "binomial")</pre>

(b) Fit assessment

For standard logistic regression, where the data are assumed to be **Bernoulli** random variables (aka 0/1), the standard summary plots are not very helpful in letting us understand the fit. The summary does suggest that the model captures some of the variability that we see. Interestly, the binned residual plot we see that the average logit predicted value is less extreme than it should be at either end the spectrum (extremely likely to vote Democrat or Republican). At the same time, the DHARMa residual diagnostics suggest that the ranking and ranking conditional on model predictions fit simulation expectations well.

```
binnedplot(predict(l_glm_1976), resid(l_glm_1976))
```

Binned residual plot

Expected Values



DHARMa residual diagnostics

(b) model interpretation

female

presapprov2. disapprove

```
summary(l_glm_1976)
##
## Call:
## glm(formula = rep_pres_intent ~ ., family = "binomial", data = nes_1976)
##
## Deviance Residuals:
##
       Min
                 1Q
                      Median
                                    ЗQ
                                            Max
## -2.9003
           -0.2796
                      0.2682
                                0.5318
                                         2.6967
##
## Coefficients:
##
                                                   Estimate Std. Error z value
## (Intercept)
                                                   0.830985
                                                               1.026176
## ideo72. liberal
                                                  -2.327337
                                                              0.845434
## ideo73. slightly liberal
                                                  -2.886738
                                                              0.811851
## ideo74. moderate, middle of the road
                                                  -1.833549
                                                              0.772496
## ideo75. slightly conservative
                                                  -0.531175
                                                              0.777492
## ideo76. conservative
                                                   0.023279
                                                              0.791822
## ideo77. extremely conservative
                                                   0.892863
                                                              1.171559
## black
                                                  -2.021512
                                                              0.692902 -2.917
```

0.343980

-3.940278

0.810

-2.753

-3.556

-2.374

-0.683

0.029

0.762

1.597

0.215341

0.295717 -13.324

```
0.004966
                                                              0.007755
                                                                         0.640
## age
## educ12. high school (12 grades or fewer, incl
                                                                         2.339
                                                   1.019021
                                                              0.435615
## educ13. some college(13 grades or more,but no
                                                   1.077248
                                                              0.471817
                                                                         2.283
## educ14. college or advanced degree (no cases
                                                                         2.654
                                                   1.288787
                                                              0.485539
## urban2. suburban areas
                                                   0.500890
                                                              0.273666
                                                                         1.830
## urban3. rural, small towns, outlying and adja 0.074872
                                                              0.284912
                                                                         0.263
## income2. 17 to 33 percentile
                                                   0.253838
                                                              0.534507
                                                                         0.475
## income3. 34 to 67 percentile
                                                  -0.113254
                                                              0.497111
                                                                        -0.228
## income4. 68 to 95 percentile
                                                   0.027157
                                                              0.495444
                                                                         0.055
## income5. 96 to 100 percentile
                                                   0.354393
                                                              0.603023
                                                                         0.588
## union2. no, no one in the household belongs t 0.602304
                                                              0.248466
                                                                         2.424
## perfin12. same
                                                                        -0.278
                                                  -0.075231
                                                              0.271044
## perfin13. worse now
                                                  -0.314517
                                                              0.262847
                                                                        -1.197
##
                                                  Pr(|z|)
## (Intercept)
                                                  0.418062
## ideo72. liberal
                                                  0.005908 **
## ideo73. slightly liberal
                                                  0.000377 ***
## ideo74. moderate, middle of the road
                                                  0.017619 *
## ideo75. slightly conservative
                                                  0.494487
## ideo76. conservative
                                                  0.976546
## ideo77. extremely conservative
                                                  0.445991
## black
                                                  0.003529 **
## female
                                                  0.110182
## presapprov2. disapprove
                                                   < 2e-16 ***
## age
                                                  0.521980
## educ12. high school (12 grades or fewer, incl 0.019322 *
## educ13. some college(13 grades or more,but no 0.022419 *
## educ14. college or advanced degree (no cases 0.007946 **
## urban2. suburban areas
                                                  0.067205 .
## urban3. rural, small towns, outlying and adja 0.792714
## income2. 17 to 33 percentile
                                                  0.634858
## income3. 34 to 67 percentile
                                                  0.819783
## income4. 68 to 95 percentile
                                                  0.956287
## income5. 96 to 100 percentile
                                                  0.556738
## union2. no, no one in the household belongs t 0.015347 *
## perfin12. same
                                                  0.781350
## perfin13. worse now
                                                  0.231471
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##
       Null deviance: 1245.55 on 908 degrees of freedom
## Residual deviance: 597.62 on 886
                                       degrees of freedom
##
     (903 observations deleted due to missingness)
## AIC: 643.62
##
## Number of Fisher Scoring iterations: 6
```

To predict if a voter with vote for a Republican (or Democrat) we see some interesting trends. First, holding all other features equal, if an individual is **black** (not white), we should expect $e^{\beta_{\text{black}}}$ (0.132455) multiplicative decrease in the odds ratio of voting for a Republican over a Democrat. Similarly, but in reverse, **not having any household member being in a union** (instead of the reverse) is observed as increasing the odds ratio of voting for a Republican over a Democrat by 1.8263209. Not surprisingly, given that the current president is a Republican, if **one disapproves of the current president** (as opposed to approving of the current president), would lead to a multiplicative decrease in the odds ratio of voting for a Republican over a Democrat by multiplication by 0.0194428.

There are also some non-binary variable that were significant, but they more nuanced interpretations. For different **ideological** views we see impacts on whether the individual will vote Republican or Democrat. In this case, the base case is "extremely liberal" and interestedly we see that if one is moves from "extremely liberal to "liberal or "slightly liberal" we'd expect a multiplicative decrease in the odds ratio of voting for a Republican over a Democrat by 0.0975551 or 0.0557578 respectively. We don't observe significant differences (holding all other variables equal) between the odds ratio between "extremely liberal" vs any time of conservative. In context this is surprising, but one could hypothesis that this may be related to the fact that certain differences between "extremely liberal" and "conservatives" are already captured in other variables. Another significant categorical variable is **maximum education level attained**, where the base case is only some grade school (0-8 grades). Compared with this base level we do see that as we increase education levels the estimated coefficient for voting Republican over Democrat increases. We can't directly use the p-values of the higher education levels to test if a high school education vs college education significantly impacts a voter's likelihood of voting for a Republican, but looking at the difference of the coefficients and the standard error we might guess (without directly doing the test) that they are not given the standard error for each β value is larger than the difference between the β values.

(c)

(c) Fit assessment

The fit of the above model is worse (in terms of shrinking the residual deviance). From the marginal plot, this may be related to the less optimal fit relative to ideology and age. Other than that - it does look like a similar fit to the factor based model.

binnedplot(predict(l_glm_1976_numeric), resid(l_glm_1976_numeric))

Binned residual plot



Expected Values

```
plot(simulateResiduals(l_glm_1976_numeric))
```









(c) model interpretation

This numerical model suggests similar parameters are useful. A similar interpretation of **black** (vs white), not having any household member being in a union (instead of the reverse) and presidental approval exists for this model as the last. Specifically, holding all other features equal, if an individual is **black** (not white), we should expect $e^{\beta_{\text{black}}}$ (0.1258773) multiplicative decrease in the odds ratio of voting for a Republican over a Democrat, not having any household member being in a union (instead of the reverse) is observed as increasing the odds ratio of voting for a Republican over a Democrat by a multiplication of 1.814176, and if one disapproves of the current president (as opposed to approving of the current president), would lead to a decrease in the odds ratio of voting for a Republican over a Democrat by a multiplication of0.0268055.

We now can only fit a linear trend (that ranges from "extremely liberal" to "extremely conservative"), but it is significant, a one "unit" increase (between the levels) of ideology towards conservatism, we should expect $e^{\beta_{ideo7}}$ (2.0066553) multiplicative decrease in the odds ratio of voting for a Republican over a Democrat. Education attainment is no longer a significant variable.

```
summary(l_glm_1976_numeric)
```

```
##
## Call:
   glm(formula = rep_pres_intent ~ ideo7 + black + female + presapprov +
##
##
       age + educ1 + urban + income + union + perfin1, family = binomial,
##
       data = nes 1976 numeric)
##
## Deviance Residuals:
##
       Min
                  1Q
                       Median
                                    3Q
                                             Max
##
   -2.5791
            -0.3257
                       0.3169
                                0.5721
                                          3.1081
```

```
##
## Coefficients:
##
                Estimate Std. Error z value Pr(>|z|)
                2.239570
                           1.281720
                                      1.747 0.08058
## (Intercept)
## ideo7
                0.696469
                           0.094070
                                      7.404 1.32e-13 ***
## black
               -2.072448
                                     -3.096 0.00196 **
                           0.669428
## female
                0.288426
                           0.206065
                                      1.400
                                             0.16161
## presapprov
               -3.619149
                           0.267075 -13.551
                                              < 2e-16 ***
## age
                0.005650
                           0.006706
                                      0.842
                                              0.39952
## educ1
                0.239311
                           0.123796
                                      1.933
                                              0.05322
## urban
               -0.067092
                           0.138109
                                      -0.486
                                              0.62712
                0.072922
                           0.110340
                                      0.661
                                             0.50868
## income
## union
                0.595631
                           0.236410
                                      2.519 0.01175 *
## perfin1
               -0.103289
                                     -0.811 0.41750
                           0.127397
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
   (Dispersion parameter for binomial family taken to be 1)
##
##
       Null deviance: 1245.55
                               on 908 degrees of freedom
## Residual deviance: 630.72 on 898
                                        degrees of freedom
     (903 observations deleted due to missingness)
##
## AIC: 652.72
##
## Number of Fisher Scoring iterations: 6
```

(d)

Both models contain similar information, and, in terms of significant variables, all variables that are significant in the first model (except education) are significant in the second model as well. Naturally, the "linearizing" of the ordered factor variables reduced the interpretability of the model. Additionally, although not significant, the interpretation of the linearization of the levels of income is probably a worse assumption. Requiring any of these ordered factors to be linear requires (1) a strictly increasing or decreasing trend, and (2) that the change between levels should be the same. In the interpretation of the first model we saw (1) violated with ideological levels and (2) violated with education levels. As such I wouldn't recommend converting all the variables to numeric in this case.

(e)

l_glm_2000 <- glm(rep_pres_intent ~ ., data = nes_2000, family = "binomial")</pre>

(e) Fitted assessment

Interestingly, this model (relative to the model for 1976) seems a bit better fitted. Most of the binned residuals still fit in the expected variability range, with much fewer bins (still at the extremes) that are under-estimated in the extremes. Some of this may actually just be related to the fact that the 2000 sample has 1458 individuals whereas the 1976 sample had more (1812).

binnedplot(predict(l_glm_2000), resid(l_glm_2000))

Binned residual plot



DHARMa residual diagnostics



(e) model interpretation

Some similarities exist between the two models (in terms of which variables are important), yet we see new trends emerging and some naturally being reverse based on intuition.

For binary variables, **black** and **previous president disapproval** still is significant (though the impact of previous president disapproval is not the reverse - favoring a vote for a Republican over a Democrat)¹. Specifically, holding all other features equal, if an individual is **black** (not white), we should expect $e^{\beta_{\text{black}}}$ (0.2491609) multiplicative decrease in the odds ratio of voting for a Republican over a Democrat, and if **one disapproves of the current president** (as opposed to approving of the current president), would lead to an increase in the odds ratio of voting for a Republican over a Democrat by a multiplication of 13.0170957.

We see that ideology and education level is not longer a significant variable (and unlikely and comparison due to the very high standard errors estimated for each β value). On the other hand, **suburban** (vs urban) looks to cause a significant to a 2.5716781 multiplicative decline in the odds ratio of voting for a Republican over a Democrat. We also find the age is now a significant variable (with a 1 year increase in age) related to a 0.9697891 multiplicative decrease in the odds ratio of voting for a Republican over a Democrat.

summary(l_glm_2000)

Call: ## glm(formula = rep_pres_intent ~ ., family = "binomial", data = nes_2000) ## **##** Deviance Residuals: ## Min 1Q Median ЗQ Max -0.2218 ## -2.1156-0.6513 0.5124 2.4382

 $^{^{1}}$ This makes sense become the previous president is now a Democrat, while in 1976 it was a Republican.

```
## Coefficients:
##
                                                  Estimate Std. Error z value
## (Intercept)
                                                  -34.48395 2621.43463 -0.013
## ideo72. liberal
                                                  15.01763 2248.12503 0.007
## ideo73. slightly liberal
                                                   14.66998 2248.12506
                                                                         0.007
## ideo74. moderate, middle of the road
                                                 16.12093 2248.12497
                                                                         0.007
## ideo75. slightly conservative
                                                  16.20928 2248.12500
                                                                         0.007
## ideo76. conservative
                                                  17.36908 2248.12497
                                                                         0.008
## ideo77. extremely conservative
                                                   17.46256 2248.12535
                                                                         0.008
## black
                                                   -1.38966
                                                               0.68005 -2.043
## female
                                                   -0.31696
                                                               0.38111 -0.832
## presapprov2. disapprove
                                                    2.56626
                                                               0.43950
                                                                        5.839
## age
                                                   -0.03068
                                                               0.01223 -2.509
## educ12. high school (12 grades or fewer, incl
                                                   16.58022 1348.27773
                                                                        0.012
## educ13. some college(13 grades or more,but no
                                                   16.97609 1348.27775
                                                                         0.013
## educ14. college or advanced degree (no cases
                                                   16.80642 1348.27776
                                                                         0.012
## urban2. suburban areas
                                                    0.94456
                                                               0.46955
                                                                         2.012
## urban3. rural, small towns, outlying and adja
                                                    0.82538
                                                               0.50495
                                                                        1.635
## income2. 17 to 33 percentile
                                                   -0.05480
                                                               0.65949 -0.083
## income3. 34 to 67 percentile
                                                    0.04054
                                                               0.60877
                                                                        0.067
## income4. 68 to 95 percentile
                                                    0.44415
                                                               0.64855
                                                                        0.685
## income5. 96 to 100 percentile
                                                    0.36119
                                                               0.97648
                                                                        0.370
## union2. no, no one in the household belongs t
                                                    0.79305
                                                               0.54209
                                                                         1.463
## perfin12. same
                                                    1.11384
                                                               0.41241
                                                                         2.701
## perfin13. worse now
                                                               0.64879
                                                    0.51878
                                                                         0.800
##
                                                 Pr(|z|)
## (Intercept)
                                                  0.98950
## ideo72. liberal
                                                  0.99467
## ideo73. slightly liberal
                                                  0.99479
## ideo74. moderate, middle of the road
                                                  0.99428
## ideo75. slightly conservative
                                                  0.99425
## ideo76. conservative
                                                  0.99384
## ideo77. extremely conservative
                                                  0.99380
## black
                                                  0.04101 *
## female
                                                  0.40559
## presapprov2. disapprove
                                                 5.25e-09 ***
## age
                                                  0.01211 *
## educ12. high school (12 grades or fewer, incl 0.99019
## educ13. some college(13 grades or more,but no 0.98995
## educ14. college or advanced degree (no cases
                                                  0.99005
## urban2. suburban areas
                                                  0.04426 *
## urban3. rural, small towns, outlying and adja 0.10214
## income2. 17 to 33 percentile
                                                  0.93378
## income3. 34 to 67 percentile
                                                  0.94691
## income4. 68 to 95 percentile
                                                  0.49345
## income5. 96 to 100 percentile
                                                  0.71146
## union2. no, no one in the household belongs t 0.14348
## perfin12. same
                                                  0.00692 **
## perfin13. worse now
                                                  0.42394
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

(e) numeric, fitted assessment Again, this model looks a bit worse than the non-linear transformation of factor variables. We see at differences in education levels are not well captured, and maybe a bit suboptial fit in the linearized ideological variable (via the marginal plot). Interestly, the binned residuals look like a better fit (with only the most extreme bin for Republican voting to be highly underestimated).

binnedplot(predict(l_glm_2000_numeric), resid(l_glm_2000_numeric))



Binned residual plot

Expected Values



DHARMa residual diagnostics



(e) numeric, model interpretation The TA leaves this up to the student to interpret, but we note that ideology of the individual is now significant, and urban is no longer significant. Additionally similar comments as in (d) relate to the linearized and factor based models.

summary(l_glm_2000_numeric)

```
##
## Call:
##
   glm(formula = rep_pres_intent ~ ideo7 + black + female + presapprov +
##
       age + educ1 + urban + income + union + perfin1, family = binomial,
##
       data = nes_2000_numeric)
##
## Deviance Residuals:
##
       Min
                  1Q
                      Median
                                    ЗQ
                                             Max
##
   -2.2661
           -0.6780
                     -0.3307
                                0.5783
                                          2.3440
##
##
  Coefficients:
                Estimate Std. Error z value Pr(>|z|)
##
                                      -4.687 2.77e-06 ***
##
   (Intercept) -11.96238
                             2.55218
## ideo7
                 0.56281
                             0.13832
                                       4.069 4.72e-05 ***
## black
                -1.28042
                             0.63813
                                      -2.007
                                               0.04480 *
                                       -1.225
                                               0.22075
## female
                -0.43189
                             0.35270
                 2.42242
                             0.39637
                                       6.112 9.87e-10 ***
##
  presapprov
## age
                -0.02921
                             0.01088
                                       -2.684
                                               0.00727 **
                                       0.655
## educ1
                 0.14627
                             0.22337
                                               0.51258
## urban
                  0.39483
                             0.23649
                                        1.670
                                               0.09501
## income
                 0.11936
                             0.18156
                                       0.657
                                              0.51092
```

union 0.45653 0.49674 0.919 0.35807 0.43815 0.27943 1.568 0.11688 ## perfin1 ## ___ 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 ## Signif. codes: ## (Dispersion parameter for binomial family taken to be 1) ## ## ## Null deviance: 350.34 on 256 degrees of freedom ## Residual deviance: 223.97 on 246 degrees of freedom ## (1201 observations deleted due to missingness) ## AIC: 245.97 ## ## Number of Fisher Scoring iterations: 5

(f)

This exercise suggests that different voter characteristics effected voter decisions across the years (although some like previous president approval² and race seem pretty constant). Additionally, ideology seems to have some impact (but maybe not linear in all cases). Interestingly, the model fit on the 2000 data seems to better conform with the linear model assumptions. The differences between the models fit on the different years relate to the impact of education level (significant for 1976 but not 2000) and urban vs suburban (significant for 2000 but not 1976). As captured in the discussion about interested 1976 trend for ideology, we should make sure **not** to claim causal structure related for voter preferences.

 $^{^{2}}$ With the party in power flip.