Understanding Eye Color Inheritance through Logistic Regression

Funfay Jen

2020-12-10

Research Question

In this project, we set out to investigate the patterns that govern the inheritance of eye colors. The data are from Sir Francis Galton, which contain the eye colors (light, hazel and dark) of the parents and grandparents, as well as the numbers of total children and light-eyed children. Here's the first 6 rows of the 78 families on which information is collected:

P_Light	P_Hazel	P_Dark	G_Light	G_Hazel	G_Dark	Total	Light
2	0	0	4	0	0	6	6
2	0	0	4	0	0	6	6
2	0	0	4	0	0	6	6
2	0	0	4	0	0	6	5
2	0	0	4	0	0	7	7
2	0	0	4	0	0	7	7

Summary of Conlusions

Using a logistic regression model with parents' eye colors as a factor, we find convincing evidence that children of 2 dark-eyed parents is unlikely to have light eye colors (estimated probability = 0.244), whereas children of one light-eyed and one hazel-eyed parent is likely to have light eye colors (estimated probability = 0.733) and children of 2 light-eyed parents is almost sure to have light eye colors (estimated probability = 0.944). On the other hand, we do not find conclusive evidence for how other eye color combinations of the parents affect the inheritance in their children. For these children, we might as well assume that they have a 50-50 chance of having light eye colors. A complete table understanding the chance of having light eye colors given the eye colors of the parents is given below. Here, the code "002" in the 1st column means no light eyed parents, no hazel-eyed parents, and two dark-eyed parents, and likewise for other codes. The 2nd column shows the naive way of averaging the proportions of light-eyed children of the same parent category, which can be ignored but is included for comparison. The 3rd column shows the sober way of computing the grand proportion of light-eyed children of the same parent category, which we show to be the same as the point estimate of the logistic regression model. The last 3 columns show the estimated probability of having light-eyed children and the associated confidence intervals for all 6 possible parent eye-color combinations.

P	light_freq_naive	light_freq_sober	light_prob	light_prob_lo	light_prob_hi
002	0.243	0.244	0.244	0.135	0.382
011	0.461	0.486	0.486	0.326	0.648
020	0.598	0.600	0.600	0.351	0.817
101	0.547	0.544	0.544	0.463	0.622
110	0.755	0.733	0.733	0.659	0.800
200	0.949	0.944	0.944	0.910	0.969

Analysis

Before analyzing the data, We first convert the parent and grandparent eye color data into a form suitable for analysis. We want to treat the first 3 columns corresponding to parent eye colors as one factor, and similarly the second 3 columns corresponding to grandparent eye colors as one. But we also don't want to lose the descriptive information in these columns (i.e., we don't want to call one level of the "parent" factor "orange" and another "apple"). The most natural way is to concatenate the numbers into 3 digits, and treat that as non-numeric. This results in the eye-color codes in the summary table in the previous section.

We then propose the following three models, and ultimately the best will be chosen to produce the conclusions:

$$\textit{Model P:} \quad \text{logit}(p) = \beta_1 P_{002} + \beta_2 P_{011} + \beta_3 P_{020} + \beta_4 P_{101} + \beta_5 P_{110} + \beta_6 P_{200} = \mathbf{P}$$

$$\textit{Model G:} \quad \text{logit}(p) = \mathbf{G}$$

$$\textit{Model PG:} \quad \text{logit}(p) = \mathbf{P} + \mathbf{G}$$

where P_{002}, etc are indicator variables for the levels of the parent eye colors, and we can further abbreviate them into one single letter $\mathbf{P}$. We choose to not include an intercept, so that the coefficient directly represent the effect (log odds) of a particular factor level. The same conventions are adopted for models $\mathbf{G}$ and $\mathbf{P} + \mathbf{G}$.

We find that model G fails the assumptions; models P and P + G both satisfy the assumptions, but the former is superior for its parsimony and better AIC value. The residual deviance of model P divided by the degrees of freedom is 1.654, so it's possible that we can improve the over-dispersion issue, but it is not excessive and should not bother us. The rest of the analysis proceeds with model P in mind. The ANOVA tables comparing the three models, the model AIC values, and residual diagnostics plots follow.

Table 2: ANOVA of Model P and Model PG

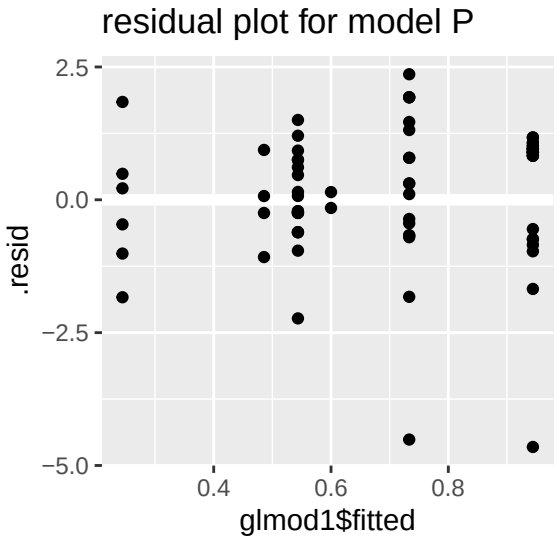
Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
72	119.1	NA	NA	NA
61	106.6	11	12.54	0.3241

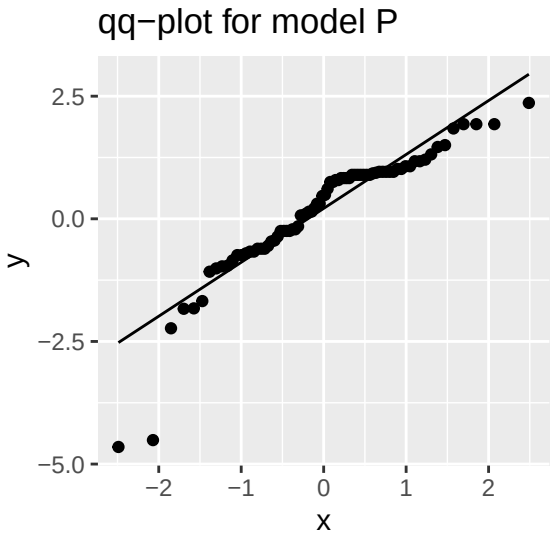
Table 3: ANOVA of Model G and Model PG

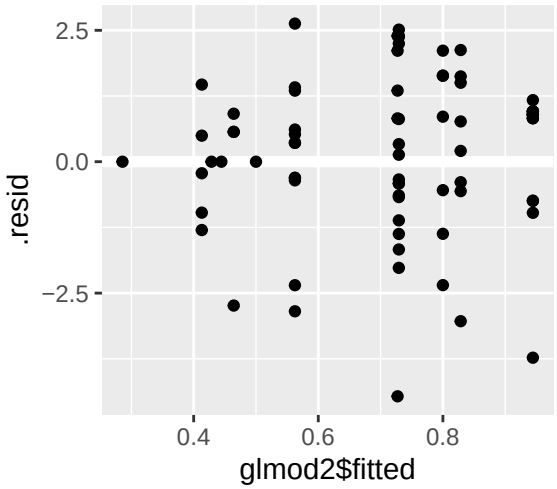
Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
66	172.1	NA	NA	NA
61	106.6	5	65.5	8.824e-13

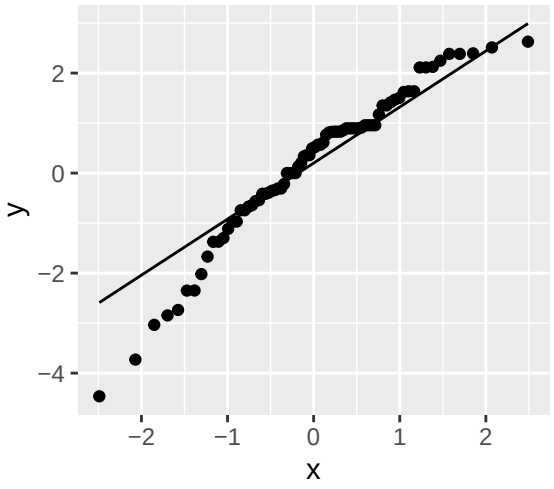
Table 4: AIC Values

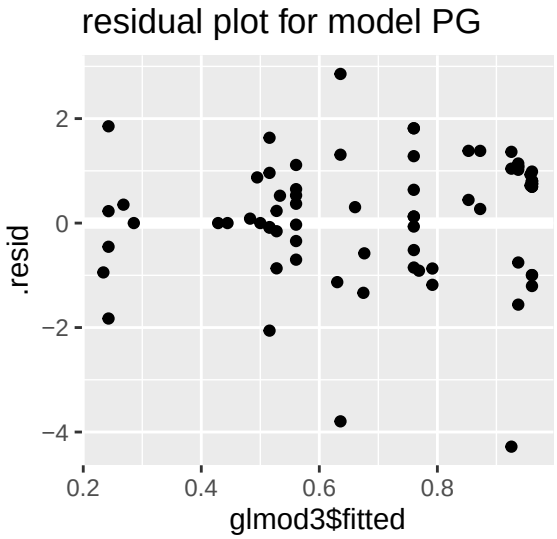
Model P	Model G	Model PG
250.5	315.4	259.9

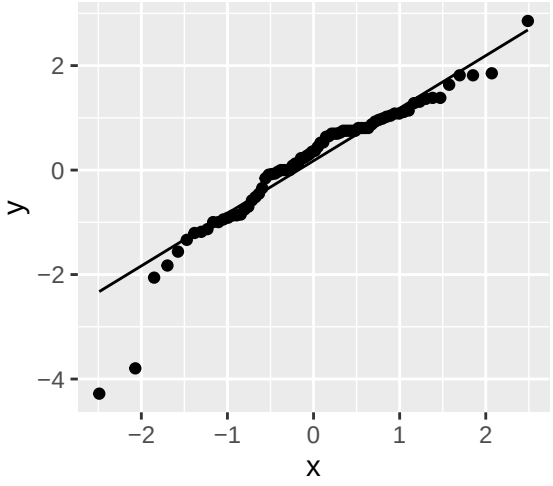

residual plot for model P
.resid
glmod1$fitted
2.5
0.0
−2.5
−5.0
0.4
0.6
0.8


qq−plot for model P
y
x
2.5
0.0
−2.5
−5.0
−2
−1
0
1
2


residual plot for model G
.resid
glmod2$fitted
2.5
0.0
−2.5
0.4
0.6
0.8


qq−plot for model G
y
x
2
0
−2
−4
−2
−1
0
1
2


residual plot for model PG
.resid
glmod3$fitted
2
0
−2
−4
0.2
0.4
0.6
0.8


qq−plot for model PG
y
x
2
0
−2
−4
−2
−1
0
1
2

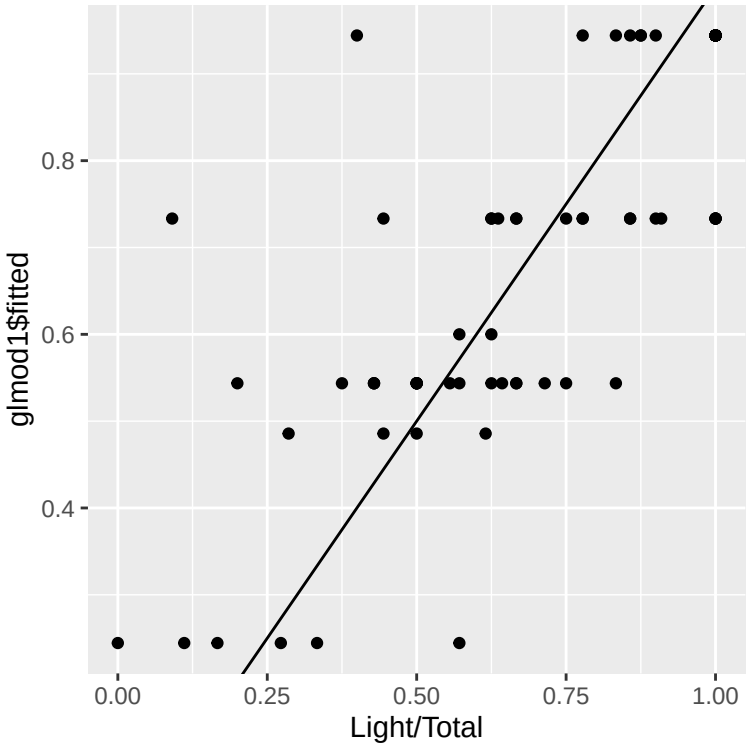
A summary of the fitted coefficients of the chosen model, Model P, is given below:

	Estimate	Std. Error	z value	Pr(>|z|)
P002	-1.128	0.3469	-3.253	0.001141
P011	-0.05716	0.3382	-0.169	0.8658
P020	0.4055	0.527	0.7693	0.4417
P101	0.1749	0.1645	1.064	0.2875
P110	1.012	0.1846	5.479	4.281e-08
P200	2.829	0.2854	9.91	3.754e-23

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	389.9 on 78 degrees of freedom
Residual deviance:	119.1 on 72 degrees of freedom

We can plot the fitted probabilities of having light eye colors against the observed probabilities. The ideal $y = x$ line is set against them and we should be reassured that things are going well. The way to read this plot is by recognizing the 6 horizontal lines as corresponding to the 6 distinct parent eye color combinations, and we do see, visually, that some groups of the observed probabilities are significantly different from 0.5. Hence, estimated confidence intervals constructed from these data points should reflect well on the visual appearance.



This takes us to the conclusion at the beginning of this report. We can calculate the point estimate and the associated confidence intervals of the probability of having light-eyed children for a given parent eye color combination by transforming those on the logit scale by the inverse logistic function. It is worth noting that this is a simplified problem of the byssinosis problem on the midterm. Here, the covariates are simpler (they are one-hot) and hence the regression coefficients are not "shaped" by the covariate vector, and we are not

computing any risk difference.

One thing to comment on is the calculations of the "naive" and "sober" proportions. The naive way takes the average of observed proportions of light-eyed children for each parent eye color combination, which gives slightly different point estimates than those given by the logistic model. This is because in this averaging we fail to incorporate the sample sizes. The sober way instead calculates the grand proportion of light-eyed children for each parent eye color combination in one go. The remarkable fact is that this gives estimates that coincide with the logistic model. And this is exactly expected, since the MLE of a probability is the grand proportion.

In general, the point estimate of a mean parameter can usually be obtained by algebraic/arithmetic formulas like this (which is also a simplistic model in disguise). However, the construction of CIs require assumptions and/or models. The main contribution of using a logistic model is that it provides its own confidence intervals.

Future Work

We can try looking into the over-dispersion issue a bit. But of greater interest is whether other models can provide better insight. As well, we should construct the CIs of the algebraic point estimates and compare.

Codes

```
eye_colors <- read_table2("eye_color.txt", col_types=cols(
  P_Light = col_integer(),
  P_Hazel = col_integer(),
  P_Dark = col_integer(),
  G_Light = col_integer(),
  G_Hazel = col_integer(),
  G_Dark = col_integer(),
  Total = col_integer(),
  Light = col_integer()
), col_names = c("P_Light", "P_Hazel", "P_Dark",
                 "G_Light", "G_Hazel", "G_Dark",
                 "Total", "Light"),
skip = 1)


eye_colors <- eye_colors %>% mutate(
  P = factor(sprintf("%03d", 100*P_Light + 10 * P_Hazel + P_Dark)),
  G = factor(sprintf("%03d", 100*G_Light + 10 * G_Hazel + G_Dark))
) %>% select(P, G, Total, Light)

pander(eye_colors)

glmod1 <- glm(cbind(Light, Total - Light) ~ P - 1, family = binomial, data=eye_colors)
glmod2 <- glm(cbind(Light, Total - Light) ~ G - 1, family = binomial, data=eye_colors)
glmod3 <- glm(cbind(Light, Total - Light) ~ P + G - 1, family = binomial, data=eye_colors)

anova(glmod1, glmod3, test="Chisq")
anova(glmod2, glmod3, test="Chisq")


glmod_aics <- cbind(glmod1$aic, glmod2$aic, glmod3$aic)
colnames(glmod_aics) <- c("Model P", "Model G", "Model PG")
pander(glmod_aics)
```

```r
# AIC only makes sense if the model is valid; so check residuals

# Apart from two outliers, the Model P is very satisfying!
eye_colors_diags1 <- augment(glmod1)
ggplot(eye_colors_diags1, aes(glmod1$fitted, .resid)) +
  geom_hline(yintercept = 0, size = 2, color = "white") +
  geom_point()

ggplot(eye_colors_diags1, aes(sample = .resid)) +
  geom_qq_line() +
  geom_qq()

# Model G fails to satisfy assumptions
eye_colors_diags2 <- augment(glmod2)
ggplot(eye_colors_diags2, aes(glmod2$fitted, .resid)) +
  geom_hline(yintercept = 0, size = 2, color = "white") +
  geom_point()

ggplot(eye_colors_diags2, aes(sample = .resid)) +
  geom_qq_line() +
  geom_qq()

# Model PG satisfies the assumptions, but is not parsimonious
eye_colors_diags3 <- augment(glmod3)
ggplot(eye_colors_diags3, aes(glmod3$fitted, .resid)) +
  geom_hline(yintercept = 0, size = 2, color = "white") +
  geom_point()

ggplot(eye_colors_diags3, aes(sample = .resid)) +
  geom_qq_line() +
  geom_qq()

# Check for over-dispersion; we are content with 1.65 for now
glmod1$deviance/glmod1$df.residual

# Plot observed vs fitted probabilities, under Model P
ggplot(eye_colors) +
  geom_point(aes(x = Light/Total, y = glmod1$fitted)) +
  geom_abline(slope = 1, intercept = 0)

# Let's understand the data
eye_colors_understanding <- eye_colors %>%
                              group_by(P) %>%
                              summarize(
                                light_freq_naive = mean(Light/Total),
                                light_freq_sober = sum(Light)/sum(Total))

inv_logit <- function(x) {exp(x) / (1 + exp(x))}

light_prob <- inv_logit(coef(glmod1))
light_prob_lo <- inv_logit(confint(glmod1))[, 1]
light_prob_hi <- inv_logit(confint(glmod1))[, 2]
```

```
eye_colors_understanding <- eye_colors_understanding %>%
                                mutate(
                                  light_prob = light_prob,
                                  light_prob_lo = light_prob_lo,
                                  light_prob_hi = light_prob_hi
                                )

pander(summary(glmod1))

kable(eye_colors_understanding, digits = 3)
```