ST 518 Homework 11

Eric Warren

December 5, 2023

Contents

1 Problem 1

The file "kickerdata.txt" contains information on $n = 4774$ field goal attempts from games played in the National Football League (NFL). Consider a logistic regression model in which the log odds of a successful field goal is linear in the distance (yards) of the attempt. This model is fit with `PROC GLIMMIX` with code and output on the next page.

1.1 Part A

Let $\pi(x)$ denote the chance of success for a kick made from distance x. Write out the probability model for $\pi(x)$, including specification an appropriate probability distribution for random variables.

The probability model we can write is that $P(Y_i = y) = \pi^y(1-\pi)^{1-y}$ where $y = 0, 1$ (and 0 is for a failure and 1 for a success). Sometimes we call π as p so you can see this equation as $P(Y_i = y) = p^y(1-p)^{1-y}$ with the same conditions for y.

1.2 Part B

Give the equation for the fitted model.

We know that the equation for the fitted model is $\hat{y} = \frac{1}{e^{-(\hat{\beta}_0+\hat{\beta}_1 x)}}$. Since in our output $\hat{beta}_0 = 5.7665$ and $\hat{\beta}_1 = -0.1077$ our fitted model is $\hat{p(x)} = \frac{1}{1+e^{-(5.7665-0.1077x)}}$.

1.3 Part C

Estimate the change in the log-odds of a success as the distance is increased by 2 yard and also by 10 yards. Report standard errors for both estimated changes.

We know that the estimated change in log odds is just found from finding $\hat{\beta}_1(x_1 - x_0)$ and the standard error is found from $\hat{SE(\ \beta_1)}(x_1 - x_0)$. Note, we know from the output that $\hat{\beta}_1 = -0.1077$ and $\hat{SE(\ \beta_1)} = 0.004788$.

- For a change in two yards we know that $x_1 - x_0 = 2$ so we can find the estimated change of log odds to be $\hat{\beta}_1(x_1 - x_0) = -0.1077 * 2 = -0.2154$ and the standard error to be $\hat{SE(\ \beta_1)}(x_1 - x_0) = 0.004788 * 2 = 0.009576$.
- For a change in ten yards we know that $x_1 - x_0 = 10$ so we can find the estimated change of log odds to be $\hat{\beta}_1(x_1 - x_0) = -0.1077 * 10 = -1.077$ and the standard error to be $\hat{SE(\ \beta_1)}(x_1 - x_0) = 0.004788 * 10 = 0.04788$.

1.4 Part D

Complete the table below:

Distance	log($\hat{P}$ (success)/(1 - $\hat{P}$ (success)))	$\hat{P}$ (success)
40	1.460	0.81
42	1.245	0.78
Diff		-0.03
50	0.383	0.59
52	0.168	
Diff		

Let us fill in the table with one thing at a time.

- We can get the first difference to be 1.460 - 1.245 = 0.215.
- Now let us get the $\hat{P}$ (success) value when $x = 52$. We can solve this by plugging in our equation from **Part B** of $\hat{p(x)} = \frac{1}{1+e^{-(5.7665-0.1077x)}} = \frac{1}{1+e^{-(5.7665-0.1077(52))}} \approx 0.54$. So $\hat{P}$ (success) value when $x = 52$ is about 0.54.

- Lastly, let us get our last differences. The change in log(odds) (column 2) is 0.383 - 0.168 = 0.215. The change in the probability of success (column 3) is 0.59 - 0.54 = 0.05.

So now after calculating all of these values, we can fill in the table to get the values of:

Distance	log($\hat{P}$ (success)/(1 - $\hat{P}$ (success)))	$\hat{P}$ (success)
40	1.460	0.81
42	1.245	0.78
Diff	0.215	-0.03
50	0.383	0.59
52	0.168	0.54
Diff	0.215	-0.05

1.5 Part E

Fill in the blanks: On the _____ scale, the effect of increasing the distance is _____, while on the probability scale, the effect of increasing the distance depends on _____.

The question will be answered in the same order it is asked. We can answer using the information from **Part D**. Note that column 2 is the log(odds) and does not change in difference as x is increasing by the same margin with higher x values. Thus, the first blank is **log(odds)** and the second blank is **a proportional reduction** (can also say decreasing linearly with distance). As we can see, on the probability scale it is not proportional because it is decreasing exponentially (really logistically). We can also see that this difference depends on the denominator of what $1 + e^{-(\hat{\beta_0}+\hat{\beta_1}x)}$ for both values. I am not 100% sure how the last blank should be filled in but I will say that it depends on the **probability density function of success** (or really what the denominator of $1 + e^{-(\hat{\beta_0}+\hat{\beta_1}x)}$ gets us since it is not linear).

1.6 Part F

Estimate the ratio of the odds of a successful field goal at a distance of x + 1 relative to the odds at distance x. (This is called the odds ratio.)

So we can calculate this by finding what $e^{-\beta_1}$ is. In our case this is $e^{-\beta_1} = e^{-0.1077} \approx .898$, which is also what our output table value says as well.

1.7 Part G

Give the distance at which the estimated success probability is 0.5, (where a miss is as probable as a make.)

We can find this by saying that $\hat{p(x)} = 0.5 = \frac{1}{1+e^{-(5.7665-0.1077x)}} <=> 1 + e^{-(5.7665-0.1077x)} = \frac{1}{0.5} = 2 <=> e^{-(5.7665-0.1077x)} = 1 <=> -5.7665 + 0.1077x = 0 <=> x = \frac{5.7665}{0.1077} = 53.54225$. So when a kicker attempts a field goal at about 53.5 yards, they have a 50-50 chance of making (or missing) it.

2 Problem 2

(Rao, (1998) eg 16.5) An experiment investigates three formulations of a diet for rats. The response is absorption of a certain chemical by the kidneys. The design involves four litters, each with three rats (a total of 12 rats). Within each litter, rats are randomized to diet formulation. Another factor of interest is the method of measuring this absorption. There are three methods up for investigation, with differences that may be subtle compared to diet formulation effects. So, three specimens are sampled from each rat, and these specimens are randomized to the methods. Data are available as "absorb.dat".

2.1 Part A

Identify the name of the experimental design used here.

This design seems to be a randomized complete block split plot design with the whole plot factor being diet. The whole plot unit seems to be rats. The factor seems to be litter. The split plot factor seems to be the method of measurement. And the split plot unit seems to be the specimens. For purposes of data I am going to assume column i is diet, column j is measurement method and column k is litter.

2.2 Part B

Propose a statistical model.

For RCBSPDs, we know the model is $Y_{ijk} = \mu + \alpha_i + S_k + (\alpha S)_{ik} + \beta_j + (\alpha\beta)_{ij} + E_{ijk}$ so this is our model with i being diet (so i = 1, 2, 3 = a), j being measurement method (j = 1, 2, 3 = b) and k being the litter block (k = 1, 2, 3, 4 = r).

2.3 Part C

Sketch an ANOVA table, with columns for source, df and EMS.

Here we can make an ANOVA table in R using our data.

```
library(tidyverse)
rats <- read_table("absorb.dat") %>%
  dplyr::mutate(i = as.factor(i),
                j = as.factor(j),
                k = as.factor(k)) %>%
  dplyr::rename(diet = i,
                measurement_method = j,
                litter = k)

(anova_problem2 <- anova(lm(y ~ diet*litter + measurement_method + diet:measurement_method, rats)))
```

```
## Analysis of Variance Table
##
## Response: y
##                          Df  Sum Sq Mean Sq  F value             Pr(>F)
## diet                      2 314.232 157.116 275.6270 0.0000000000000316 ***
## litter                    3   0.340   0.113   0.1989             0.8958
## measurement_method        2 260.574 130.287 228.5611 0.0000000000001608 ***
## diet:litter               6   6.264   1.044   1.8314             0.1493
## diet:measurement_method   4  98.308  24.577  43.1154 0.0000000055027389 ***
## Residuals                18  10.261   0.570
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From this ANOVA table we can plug in the source and degrees of freedom for each value. Now we have to find the E(MS) values for each. This will be done in steps below, noting that $a = 3, b = 3, r = 4, ab = 9, ar = 12, br = 12, abr = 36$.

- diet: We know this is $(br)\psi_a^2 + b\sigma_{as}^2 + \sigma^2 = 12\psi_a^2 + 3\sigma_{as}^2 + \sigma^2$
- litter: We know this is $(ab)\sigma_s^2 + b\sigma_{as}^2 + \sigma^2 = 9\sigma_s^2 + 3\sigma_{as}^2 + \sigma^2$

- measurement_method: We know this is $(ar)\psi_b^2 + \sigma^2 = 12\psi_b^2 + \sigma^2$ because both α and β are fixed effects
- diet*litter: We know this is $b\sigma_{as}^2 + \sigma^2 = 3\sigma_{as}^2 + \sigma^2$
- diet*measurement_method: We know this is $(r)\psi_{ab}^2 + \sigma^2 = 4\psi_{ab}^2 + \sigma^2$
- Residuals: We know this is just σ^2.

Now we can make a table of our different values.

Source	DF	Expected(MS)	Estimated(MS) from ANOVA table
diet	2	$12\psi_a^2 + 3\sigma_{as}^2 + \sigma^2$	157.1158361
litter	3	$9\sigma_s^2 + 3\sigma_{as}^2 + \sigma^2$	0.1133815
measurement_method	2	$12\psi_b^2 + \sigma^2$	130.2867861
diet:litter	6	$3\sigma_{as}^2 + \sigma^2$	1.0439509
diet:measurement_method	4	$4\psi_{ab}^2 + \sigma^2$	24.5771028
Residuals	18	σ^2	0.5700306

2.4 Part D

Test for an interaction between formulation and method. Use $\alpha = 0.05$.

Here our $H_0 : \alpha\beta = 0$ and $H_A : \alpha\beta \neq 0$. To test this we are going to get the F-statistic of $\frac{MA(AB)}{MS(E)} = \frac{24.577}{0.57} = 43.1$ on degrees of freedom 4, 18 (which can be found from our table above). We can find the p-value to be `pf(43.1, 4, 18, lower.tail = F)` which gets us the p-value0. Since this value is less than our alpha level set at 0.05, we can reject our null hypothesis and say that the interaction between formulation and method is significant.

2.5 Part E

Test ($\alpha = 0.05$) for simple method effects separately at each level of formulation.

Here we are saying that $H_0 : \mu_{i1.} = \mu_{i2.} = \mu_{i3.}$ and the alternative is that they all do not. We can test all of this using SAS's `PROC MIXED` and `LSMEANS` statement. The code and attached information is shown below.

```
proc mixed data = rats method = type3;
  class diet method litter;
  model y = diet|method;
  random litter diet*litter;
  lsmeans method*diet / slice = diet;
  run;
```

Figure 1: Code to Run

Least Squares Means							
Effect	diet	method	Estimate	Standard Error	DF	t Value	Pr > |t|
diet*method	1	1	27.0975	0.3952	18	68.57	<.0001
diet*method	1	2	21.7425	0.3952	18	55.02	<.0001
diet*method	1	3	29.0925	0.3952	18	73.62	<.0001
diet*method	2	1	17.8950	0.3952	18	45.29	<.0001
diet*method	2	2	16.4150	0.3952	18	41.54	<.0001
diet*method	2	3	23.5200	0.3952	18	59.52	<.0001
diet*method	3	1	20.8700	0.3952	18	52.81	<.0001
diet*method	3	2	25.1575	0.3952	18	63.66	<.0001
diet*method	3	3	28.9550	0.3952	18	73.27	<.0001

Tests of Effect Slices					
Effect	diet	Num DF	Den DF	F Value	Pr > F
diet*method	1	2	18	101.37	<.0001
diet*method	2	2	18	98.61	<.0001
diet*method	3	2	18	114.81	<.0001

Figure 2: Hypothesis Test Results

As we can see for all $i = 1, 2, 3$ the p-value is less than 0.0001 which means we have statistically significant evidence to say that simple method effects separately at each level of formulation produces different results. In conclusion this means that at each level of diet formulation, we can see that the statement $\mu_{i1.} = \mu_{i2.} = \mu_{i3.}$ is **not** true.

2.6 Part F

Test for the simple effects of formulation using method 3. Carry out all three pairwise comparisons among formulations using this method, identifying any significant differences.

Here we are saying that $H_0 : \mu_{13.} = \mu_{23.} = \mu_{33.}$ and the alternative is that they all do not. We can test

all of this using SAS's `PROC GLIMMIX` and `LSMEANS` statement. The code and attached information is shown below.

```
proc glimmix data = rats;
 class diet method litter;
 model y = diet|method;
 random litter diet*litter;
 lsmeans method*diet / slice = diet slicediff = method;
 run;
```

Figure 3: Code to Run

diet*method Least Squares Means

diet	method	Estimate	Standard Error	DF	t Value	Pr > |t|
1	1	27.0975	0.3952	18	68.57	<.0001
1	2	21.7425	0.3952	18	55.02	<.0001
1	3	29.0925	0.3952	18	73.62	<.0001
2	1	17.8950	0.3952	18	45.29	<.0001
2	2	16.4150	0.3952	18	41.54	<.0001
2	3	23.5200	0.3952	18	59.52	<.0001
3	1	20.8700	0.3952	18	52.81	<.0001
3	2	25.1575	0.3952	18	63.66	<.0001
3	3	28.9550	0.3952	18	73.27	<.0001

Figure 4: Estimates of Means

Simple Effect Comparisons of diet*method Least Squares Means By method							
Simple Effect Level	diet	_diet	Estimate	Standard Error	DF	t Value	Pr > |t|
method 1	1	2	9.2025	0.5588	18	16.47	<.0001
method 1	1	3	6.2275	0.5588	18	11.14	<.0001
method 1	2	3	-2.9750	0.5588	18	-5.32	<.0001
method 2	1	2	5.3275	0.5588	18	9.53	<.0001
method 2	1	3	-3.4150	0.5588	18	-6.11	<.0001
method 2	2	3	-8.7425	0.5588	18	-15.64	<.0001
method 3	1	2	5.5725	0.5588	18	9.97	<.0001
method 3	1	3	0.1375	0.5588	18	0.25	0.8084
method 3	2	3	-5.4350	0.5588	18	-9.73	<.0001

Figure 5: Hypothesis Test Results

As we can see the pairwise comparisons when only looking at method 3 is that diets 1 and 2 and diets 2 and 3 are significant but diets 1 and 3 are not. Therefore, when looking only at method 3 we can say that diet formulation 2 is the only statistically significant different diet.

3 Problem 3

The production equipment is such that temperature and pressure must be held constant for the entire day. The experiment is conducted over 8 days, which are randomly assigned to the four treatment combinations of temperature and pressure. Three batches are produced each day, one of each recipe, for a total of 24 batches. Each batch will have a single viscosity measurement taken. Pertinent SAS code and output are given on the next page.

3.1 Part A

Propose a mixed model for the viscosity measurements with random day effects. Take the effects of all three treatment factors, A, B, C and their first and second order interactions to be fixed. Assume normal distributions for all random effects.

We can write our model as $Y_{ijkl} = \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + D_{l(ij)} + E_{ijkl}$ where $i = 1, 2; j = 1, 2; k = 1, 2, 3; l = 1, ..., 8$.

3.2 Part B

Carry out F-tests for the four hypotheses listed below. For each, report the F-ratio, the associated degrees of freedom, a conclusion regarding the hypothesis, and where available from the output, a p-value.

- No 2nd order interaction between temperature, pressure and recipe. To find this we can say that our F-statistic = $\frac{MS(abc)}{MS(E)} = \frac{0.13041667}{0.25833333} = 0.5048387$ on degrees of freedom 2, 8. We can find the p-value to be `pf(0.5048387, 2, 8, lower.tail = F)` which gets us the p-value0.6216171. Since this value is greater than our alpha level set at 0.05, we can fail to reject our null hypothesis and say that the interaction between the three variables is not significant. It is plausible to say that there is no interaction between all three variables.
- No 1st order interaction between temperature and pressure. To find this we can say that our F-statistic = $\frac{MS(ab)}{MS(E)} = \frac{1.00041667}{0.25833333} = 3.872581$ on degrees of freedom 1, 8. We can find the p-value to be `pf(3.872581, 1, 8, lower.tail = F)` which gets us the p-value0.0846197. Since this value is less than our alpha level set at 0.05, we can reject our null hypothesis and say that the first order interaction between the temperature and pressure is significant.
- No main effect of pressure. To find this we can say that our F-statistic = $\frac{MS(b)}{MS(day(ab))} = \frac{26.67041667}{3.00458333} = 8.876577$ on degrees of freedom 1, 4. We can find the p-value to be `pf(8.876577, 1, 4, lower.tail = F)` which gets us the p-value0.0407647. Since this value is less than our alpha level set at 0.05, we can reject our null hypothesis and say that the main effect of pressure is significant.
- No main effect of recipe. To find this we can say that our F-statistic = $\frac{MS(c)}{MS(E)} = \frac{2.15291667}{0.25833333} = 8.333871$ on degrees of freedom 2, 8. We can find the p-value to be `pf(8.333871, 2, 8, lower.tail = F)` which gets us the p-value0.0110622. Since this value is less than our alpha level set at 0.05, we can reject our null hypothesis and say that the main effect of pressure is significant.

3.3 Part C

Estimate the variance component for the random day effect.

We know that this can be calculated from $\sigma_D^2 = \frac{MS(Day) - MS(E)}{k} = \frac{3.00458333 - 0.25833333}{3} = 0.9154167$ since k is not included as a subscript in the calculations. So we can see that the variance component for the random day effect is 0.9154167.

3.4 Part D

Report an estimate and standard error for:

- i. the marginal mean of recipe 1: To estimate the marginal mean for recipe 1, $\bar{Y_{..1.}} = \frac{2*5.5+2*3.7+2*6.8+2*4.4}{8} = 5.1$. We know that the standard error has to be $\sqrt{\frac{1}{nab}MS(E)} = \sqrt{\frac{1}{k}MS(E)} = \sqrt{\frac{1}{8}0.25833333} = \sqrt{0.03229167} = 0.1796988$.
- ii. the difference between recipes 1 and 2: To estimate the difference between recipes 1 and 2, $\bar{Y_{..1.}} - \bar{Y_{..2.}} = \frac{2*5.5+2*3.7+2*6.8+2*4.4}{8} - \frac{2*5.1+2*2.85+2*6.55+2*3.85}{8} = 5.1 - 4.5875 = 0.5125$. We know that the standard error has to be $\sqrt{\frac{2}{nab}MS(E)} = \sqrt{\frac{2}{k}MS(E)} = \sqrt{\frac{2}{8}0.25833333} = \sqrt{0.06458333} = 0.2541325$.
- iii. the marginal mean of temperature 1: To estimate the marginal mean for temperature 1, $\bar{Y_{1...}} = \frac{2*5.5+2*5.1+2*5.35+2*3.7+2*2.85+2*4.3}{12} = 4.466667$. We know that the standard error has to be $\sqrt{\frac{1}{nbc}MS(Day)} = \sqrt{\frac{1}{2*2*3}MS(Day)} = \sqrt{\frac{1}{12}3.00458333} = \sqrt{0.2503819} = 0.5003818$.

- iv. the difference between temperatures 1 and 2: To estimate the difference between temperatures 1 and 2, $\bar{Y_{1...}} - \bar{Y_{2...}} = \frac{2*5.5+2*5.1+2*5.35+2*3.7+2*2.85+2*4.3}{12} - \frac{2*6.8+2*6.55+2*7.65+2*4.4+2*3.85+2*5.2}{12} = 4.466667 - 5.741667 = -1.275$. We know that the standard error has to be $\sqrt{\frac{2}{nbc}MS(Day)} = \sqrt{\frac{2}{2*2*3}MS(Day)} = \sqrt{\frac{2}{12}3.00458333} = \sqrt{0.5007639} = 0.7076467$.