Lab 7 solutions

Michael Lopez, Skidmore College

```
library(RCurl); library(tidyverse)
gURL <- "https://raw.githubusercontent.com/statsbylopez/StatsSports/master/Data/pbp_data_hockey.rds"
pbp_data <- readRDS(gzcon(url(gURL)))
names(pbp_data)

pbp_data <- pbp_data %>%
  mutate(coords_x_adj = ifelse(event_team == home_team,
                               -1*abs(coords_x), abs(coords_x)),
         coords_y_adj = ifelse(event_team == home_team & coords_x < 0,
                               coords_y, -1*coords_y),
         coords_y_adj = ifelse(event_team == away_team & coords_x > 0,
                               coords_y, -1*coords_y))
```

Logistic regression modeling

We can try a few logistic regression models

```
library(broom)
pbp_data <- pbp_data %>%
  mutate(is_home = event_team == home_team)

fit_1 <- glm(event_type == "GOAL" ~ event_distance +
               event_angle + event_detail ,
             family = "binomial", data = pbp_data)
tidy(fit_1)

fit_2 <- glm(event_type == "GOAL" ~ is_home, data = pbp_data,
          family = "binomial")
tidy(fit_2)

fit_3 <- glm(event_type == "GOAL" ~ event_distance +
               event_angle + event_detail + is_home, data = pbp_data,
          family = "binomial")
tidy(fit_3)
```

3. Interpret the coefficient on `is_home` in `fit_2`. What does this entail about shooting the puck at home?

SOLUTIONS: The odds of a shot going in at home are 1.05 times higher (or 5 percent higher) relative to a shot on the road

4. Interpret the coefficient on `is_home` in `fit_3`. What does this entail about shooting the puck at home?

SOLUTIONS: The odds of a shot going in at home are 1.025 times higher (2.5 percent higher) relative to a shot on the road, given a model with event_detail, distance, and angle

5. Provide one possible explanation for your findings above.

The effect of shooting at home is smaller when accounting for distance, angle, and event detail. Perhaps shots at home are generally easier shots

Hosmer Lemeshow

We'll use the Hosmer Lemeshow test – review your notes from last class – as one approach to assess the fits above. Specifically, we create the goal probability `shot_prob_fit_3` as the predicted goal likelihood.

```
pbp_data$shot_prob_fit_3 <- predict(fit_3, pbp_data, type = "response")

tab_check <- pbp_data %>%
  filter(!is.na(shot_prob_fit_3)) %>%
  mutate(shot_prob_cat = cut(shot_prob_fit_3, 10)) %>%
  group_by(shot_prob_cat) %>%
  summarise(ave_exp_goals = sum(shot_prob),
            ave_act_goals = sum(event_type == "GOAL"),
            n_shots = n())

tab_check <- tab_check %>%
  mutate(diff_sq = (ave_exp_goals - ave_act_goals)^2/
           ((ave_exp_goals)*(1-ave_exp_goals/n_shots)))

tab_check
```

Hosmer Lemeshow

```
hm_test <- tab_check %>%
  summarise(test_stat = sum(diff_sq))
hm_test

1-pchisq(hm_test$test_stat, df = 8, lower.tail = TRUE)
```

7. State the null and alternative hypotheses for the HL test. Would we reject or fail to reject the null hypothesis? What does this say about `fit_3`?

Solutions: The null hypothesis is there is no lack of fit in the shot probabilities. The alternative is there is a lack of fit. Given the test statistic, we would reject the null hypothesis – there is evidence that the probabilities in `fit_3` are not well fitting of the actual data.