Homework 5

Wenda Zhou

June 19, 2017

1. (4 points) Linear regression

   The LPGA (Ladies' Professional Golf Association) collects statistics on the performance of its members on the course. We fit a linear model attempting to predict the average performance of a golfer (expressed in terms of her average tournament percentile) as a function of some statistics.

   We have collected the following:

   GreensPct The percent of greens under regulation

   GreenPutts Average putts on green per round

   Drive Average drive length (yards)

   FairwayPct Percentage of fairway hits

   SavePct Percentage of sand saves

   The output of the regression in R is as follows:

```
Call:
lm(formula = Percentile ~ GreensPct + GreenPutts + Drive + FairwayPct +
    SavePct, data = lpga)

Residuals:
     Min       1Q   Median       3Q      Max
-15.9915  -2.9872   0.2517   3.4501  15.1468

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  222.03690   35.75779   6.209 5.69e-09 ***
GreensPct      2.13857    0.15495  13.801  < 2e-16 ***
GreenPutts  -182.13728   13.00530 -14.005  < 2e-16 ***
Drive          0.03221    0.08077   0.399 0.690633
FairwayPct     0.09782    0.10942   0.894 0.372859
SavePct        0.22432    0.06082   3.688 0.000322 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.475 on 140 degrees of freedom
Multiple R-squared:  0.8738,Adjusted R-squared:  0.8693
F-statistic: 193.9 on 5 and 140 DF,  p-value: < 2.2e-16
```

   (a) For each covariate, indicate whether and how they affect the performance of a golfer.

   (b) Compute a confidence interval for the estimate of the coefficient of GreensPct.

   (c) Suppose a player could choose to work on either improving their greens under regulation percentage by 2 percent, or reduce the average number of putts by 0.02. Which would improve their performance the most?

(d) The model seems to indicate that the percentage of fairway hits or average drive length does not impact the player's placement. Would that extend to a model for amateur golfers? Explain your answer.

2. (3 points) Logistic regression

Let $\pi = p/(1-p)$ be the odds ratio. In logistic regression, we write (c.f. notes):

$$\log \pi = \alpha + \beta x \tag{1}$$

(a) Compute $\pi(x+1)$ as a function of $\pi(x)$ and the coefficients.

(b) We model the probability of a field goal attempt in the NFL as a logistic regression on the distance. Fitting the model in R from data gives (distance given in yards):

```
Call:
glm(formula = success ~ distance, family = binomial(), data = data.fga)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9526   0.2039   0.3478   0.5826   1.2309

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.76271    0.54443  12.422   <2e-16 ***
distance    -0.12084    0.01229  -9.836   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 817.72  on 1038  degrees of freedom
Residual deviance: 686.93  on 1037  degrees of freedom
AIC: 690.93

Number of Fisher Scoring iterations: 6
```

Describe the relationship between the odds of field goal attempt and the distance.

(c) Compute the probability of success of a field goal attempt at 40 yards.

3. (3 points) Confounding variables

This exercise considers statistics in the PGA and LPGA in the year of 2008. We have obtained a record of the average driving distance and average fairway accuracy (chance of the drive hitting the fairway) for each golfer. We wish to model the accuracy as a function of the driving distance and the gender.

We have done two regressions, one including only the gender of the golfer (male or female), and one including both the gender and the driving distance (in yards), and included the output below.

```
Call:
lm(formula = accuracy ~ gender, data = pga_accuracy)

Residuals:
    Min      1Q  Median      3Q     Max
-18.291  -3.540   0.035   3.590  17.035

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  67.5911     0.4469 151.248  < 2e-16 ***
genderM      -4.2261     0.5991  -7.055 9.24e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.6 on 352 degrees of freedom
Multiple R-squared:  0.1239,Adjusted R-squared:  0.1214
F-statistic: 49.77 on 1 and 352 DF,  p-value: 9.239e-12

Call:
lm(formula = accuracy ~ distance + gender, data = pga_accuracy)

Residuals:
     Min       1Q   Median       3Q      Max
-25.0712  -2.8263   0.4867   3.3494  12.0275

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 147.26894    7.03492  20.934  < 2e-16 ***
distance     -0.32284    0.02846 -11.343  < 2e-16 ***
genderM       8.94888    1.26984   7.047 9.72e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.797 on 351 degrees of freedom
Multiple R-squared:  0.3589,Adjusted R-squared:  0.3552
F-statistic: 98.24 on 2 and 351 DF,  p-value: < 2.2e-16
```

(a) Interpret the result of the first regression.

(b) Interpret the result of the second regression, and give a possible explanation as to why the sign of the gender variable changed.

(c) Write down the equation of the accuracy against the driving distance for women, and another equation for men.

(d) Plot a graph of the accuracy against the driving distance for men and for women on the same plot.