Statistics 216
Homework 2 Solutions, 65 total points

1. *15 points*
   Grading notes: Each part is worth 3 points, 2 for the correct answer and 1 for a valid explanation. Student explanations need not match the explanations provided here, but they must be correct, and they must match the answer given. If the incorrect answer is given, but the explanation matches the answer, award 1 point out of 3 for that part, even though the explanation cannot be correct because it led to the incorrect conclusion

   (a) **iv.** The training RSS must always decrease (or at least not increase) since larger s means we are minimizing over a strictly larger set of possible β values

   (b) **ii.** Initially, increasing s and fitting a more complex model will improve the test error, because we will reduce bias. But eventually, we will start to overfit and the test RSS will increase due to variance.

   (c) **iii.** Increasing s increases complexity and thus gives a more variable fit.

   (d) **iv.** Increasing complexity gives a less biased fit because the β vector is less shrunken.

   (e) **v.** Irreducible error by definition is error that we can't do anything about.

2. *10 points: (a, c) 2 points each, other parts 1 point each*

   (a) The first observation is equally likely to come from any of the n original observations, so its probability is being the jth is $1/n$ and its probability of not being the jth is $1 - 1/n$.

   Grading notes for part (a): 1 point for correct answer, 1 for correct explanation as in problem 1.

   (b) Same, $1-1/n$. We're sampling with replacement, so the first draw does not affect the second draw at all.

   (c) We need n independent events to all occur, each of which has probability $1 - 1/n$. Hence, the overall probability is $(1 - 1/n)^n$. Equivalently, the probability that the jth sample *does* occur is $1 - (1 - 1/n)^n$.

   Grading notes for part (c): 1 point for multiplying the probabilites to get $(1 - 1/n)^n$, 1 point for saying somewhere it's because the events are independent.

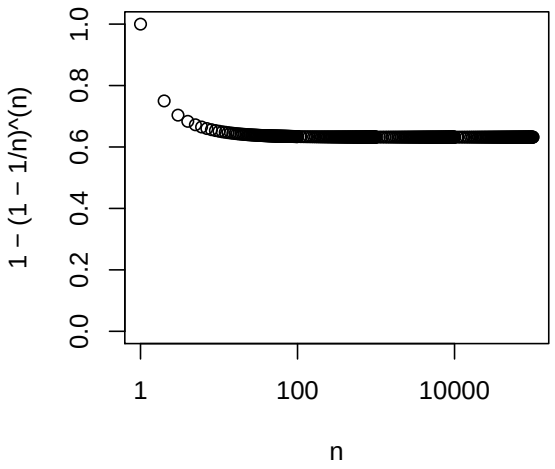
   (d) $1 - (1 - 1/5)^5 \approx 0.672$

   (e) $1 - (1 - 1/100)^{100} \approx 0.634$

   (f) $1 - (1 - 1/10^4)^{100^4} \approx 0.632$

   (g) The probability asymptotes to $1 - e^{-1} \approx 0.632$.

Grading notes for part (g): 1 points for the plot being correct. For the comment, anything goes.



(h) The result should be a slightly noisy estimate of 0.634, which we computed analytically a moment ago.

Grading notes for part (h): anything goes, as long as it is correct.

3. *20 points: (a, c, d) 4 points each. (b) 8 points.*
Grading notes: (a-c) are proofs, so the students should be expected to have correct, rigorous arguments for full credit. Many of these students will be unfamiliar with proofs, so be relatively lenient (e.g. 3/4 points for "right idea," but incomplete proof).

(a) Because $\frac{e^z}{1+e^z} \in (0,1)$ for any $z \in \mathbb{R}$ (and so is $1 - \frac{e^z}{1+e^z}$), any product of n factors of that form must also be less than 1.

(b) The basic idea here is that by setting $\beta_0 = 0$ and $\beta_1 > 0$, all the model's predictions will be right. The larger β_1 is, our model makes all the correct predictions with an ever-increasing certitude, driving the likelihood function ever closer to 1. The easiest way to say this rigorously is probably by a constructive argument like the one below:

Let $K = \min_i |x_i| > 0$. Then taking $\beta_0 = 0$ and $\beta_1 = B$, the factors in the first product are all greater than $\frac{e^{KB}}{1+e^{KB}}$ and all the factors in the second product are greater than $1 - \frac{e^{-KB}}{1+e^{-KB}} = \frac{e^{KB}}{1+e^{KB}}$. There are n factors, so it must be the case that

$$L(0, B) \geq \left(\frac{e^{KB}}{1 + e^{KB}} \right)^n$$

For any $a < 1$, we just need to choose B large enough that

$$\left(\frac{e^{KB}}{1 + e^{KB}} \right)^n > a$$

We can manipulate the inequality to obtain the requirement that

$$B > K^{-1} \log \frac{a^{1/n}}{1 - a^{1/n}} \tag{1}$$

Then $L(0, B) > a$ for any B satisfying (1)

This means $\hat{\beta}_0$ and $\hat{\beta}_1$ are undefined, simply because *any* particular choice of $\hat{\beta}$ is less than 1, and thus can be improved upon.

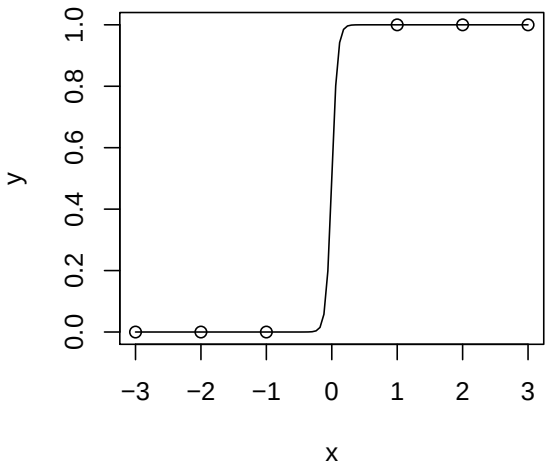
Grading notes for part (b): 4 points for a valid proof that L can get arbitrarily close to 1, 4 for a valid explanation of why that makes the MLE undefined.

(c) Repeat the same argument above, but with $K = \min_i |x_i - c|$. This time, we want the x-intercept of $\beta_0 + \beta_1 x$ to be c, so we set $\beta_0 = -\beta_1 c$ and drive $\beta_1 \to \infty$ as before.

For B satisfying (1), we will have as before

$$L(-Bc, B) \geq \left(\frac{e^{KB}}{1 + e^{KB}}\right)^n > a$$

(d) Here is an example:



Grading notes for part (d): It is OK if the student didn't plot the full prediction curve as I have, because we didn't explicitly ask them too. As long as they have clearly plotted the predictions for the training x's, the answer should be marked as correct.

4. *20 points: 4 points each.*

Grading notes: See the attached R code for demonstration of how to compute these things. Grading this problem will require an element of judgment. Answers can vary a little because there were some choices I made that the students could have reasonably made differently. Some student choices are NOT valid: in part (b) they should NOT be fitting the logistic regression to postseason data, but it is up to them whether to do so for part (c-e). The students SHOULD be using home court advantage as a predictor for all their models.

Students should show their R code for all parts. If a student gets one part wrong, but their answers to subsequent parts are consistent with the part they got wrong, then they should not keep on losing points.

(a) Examining the data, we can see that those two teams (which come out on top when we first fit a logistic regression to the data) each only played (and won) a single game. Thus, the model that *best* explains the data will actually want to give those two teams coefficients that are as large as possible.

By increasing one of these teams' coefficients, we always increase the model's prediction of how likely they were to win their only game — without changing the predictions for any other games. Thus, sending both teams' β_j to ∞ is what makes the training data most likely. These degenerate

predictions are akin to the kind of degenerate predictions we saw in the previous question.

Grading notes for part (a): 2 points for noticing that each team only played and won a single game, 2 points for explaining why that means we can always make the likelihood bigger by increasing those teams' fitted coefficients.

(b) The table is in Table 1. The logistic regression model is much more in agreement with the AP and USA Today rankings, suggesting that the voters were doing something like the logistic regression model "in their heads."

Grading notes for part (b): 3 points for the students' logistic regression fits and rankings matching mine, 1 points for some accurate comment on which model matches the voter rankings better (it's OK if for some reason they don't agree with my assessment that the LR model matches better). If they get the wrong rankings but their comment is consistent with their rankings, then they should get a point.

	Linear Score	Linear Rank	Logistic Score	Logistic Rank	AP Rank	USAT Rank
gonzaga-bulldogs	10.01	4	2.77	1	1	1
louisville-cardinals	12.13	3	2.41	2	2	2
kansas-jayhawks	8.63	6	2.27	3	3	3
indiana-hoosiers	13.26	1	2.26	4	4	5
new-mexico-lobos	3.01	23	2.24	5	11	10
ohio-state-buckeyes	7.89	7	2.14	6	7	6
duke-blue-devils	9.40	5	2.11	7	6	7
georgetown-hoyas	4.32	18	2.03	8	8	8
michigan-state-spartans	6.17	12	1.98	9	9	9
michigan-wolverines	7.50	8	1.96	10	10	11
miami-(fl)-hurricanes	5.20	14	1.76	11	5	4
kansas-state-wildcats	1.65	36	1.75	12	12	14
syracuse-orange	7.14	10	1.60	13	16	18
memphis-tigers	0.95	43	1.53	14	19	15
saint-louis-billikens	2.64	26	1.52	15	13	13
marquette-golden-eagles	2.35	30	1.51	16	15	16
butler-bulldogs	-0.73	51	1.46	17		
wisconsin-badgers	6.83	11	1.45	18	18	17
florida-gators	12.65	2	1.32	19	14	12
oklahoma-state-cowboys	3.70	20	1.30	20	17	19
unlv-rebels	2.18	32	1.29	21		
arizona-wildcats	4.70	16	1.29	22	21	20
pittsburgh-panthers	7.40	9	1.27	23	20	22
notre-dame-fighting-irish	1.95	34	1.25	24	23	
colorado-state-rams	3.17	22	1.24	25		

TABLE 1. Rank table for logistic regression vs other rankings. The logistic regression appears to agree much better than the linear regression.

(c) Because Stanford is the baseline, model coefficients for other teams represent comparisons to Stanford and we can read off p-values from the summary table.

With the linear regression model, we are confident about 268, or about 77% of the other 346 teams. With logistic regression, we are only confident about 214, or about 62% of them. We can determine this by loking at the fourth column of the coefficient table for each model's summary.

Grading notes for part (c): Students should say why the p-values from the R summary table represent p-values for comparing each team to Stanford.

(d) The table for McNemar's test is

	logistic right	logistic wrong
linear right	3818	268
linear wrong	215	1165

Grading notes for part (d): I used all the postseason games as well as regular season to do my cross-validation, but it is fine if the students used regular season only. Also there is some randomness involved in the cross-validation, but the four counts shouldn't be drastically different from mine and McNemar's test should still reject or nearly reject. Give 2 points for carrying out the cross-validation correctly and 2 points for constructing the table correctly based on the results of cross-validation.

(e) Thus $D = 268 + 215 = 483$, and $n_{12} = 268$. This is significant. The p-value is the (two-tailed) probability of a truly standard normal random variable having absolute value larger than the observed $Z = \frac{n_{1,2} - D/2}{\sqrt{D}/2} \approx 2.41$. The two-tailed p-value is 0.016.

Appendix: Relevant R code (for Question 4)

```
options(width=160, digits=3)
games <- read.csv("http://www.stanford.edu/~wfithian/games.csv",as.is=TRUE)
teams <- read.csv("http://www.stanford.edu/~wfithian/teams.csv",as.is=TRUE)
all.teams <- sort(unique(c(teams$team,games$home,games$away)))

y <- with(games, homeScore-awayScore)

## Construct a data frame of the right dimensions, with all zeros
X0 <- as.data.frame(matrix(0,nrow(games),length(all.teams)))
names(X0) <- all.teams

## Fill in the columns, one by one
for(tm in all.teams) {
  X0[[tm]] <- 1*(games$home==tm) - 1*(games$away==tm)
}

## Get rid of Stanford's column
X <- X0[,names(X0) != "stanford-cardinal"]
reg.season.games <- which(games$gameType=="REG")
homeAdv <- 1 - games$neutralLocation

## 4a) Fit logistic regression model and check out coefficients
y.bin <- 1*(y>0)
glm.mod <- glm(y.bin ~ 0 + homeAdv + ., family=binomial, data=X, subset=reg.season.games)
sort(coef(glm.mod),decreasing=TRUE)[1:10]

## What happened to the two teams that look so good?
games[games$home == "saint-mary-saint-mary" | games$away == "saint-mary-saint-mary",]
games[games$home == "st.-thomas-(tx)-celts" | games$away == "st.-thomas-(tx)-celts",]

## 4b) Get rid of teams that played fewer than five games
num.games <- table(c(games$home,games$away))
bad.teams <- names(num.games)[num.games < 5]
bad.games <- games$home %in% bad.teams | games$away %in% bad.teams

games <- games[!bad.games,]
X <- X[!bad.games, !(colnames(X) %in% bad.teams)]
```

```
y <- y[!bad.games]
y.bin <- y.bin[!bad.games]
homeAdv <- homeAdv[!bad.games]
reg.season.games <- which(games$gameType=="REG")

glm.mod <- glm(y.bin ~ 0 + homeAdv + ., family=binomial, data=X, subset=reg.season.games)
lm.mod <- lm(y ~ 0 + homeAdv + ., data=X, subset=reg.season.games)

glm.coef <- coef(glm.mod)[paste("`",teams$team,"`",sep="")]
lm.coef <- coef(lm.mod)[paste("`",teams$team,"`",sep="")]
glm.coef[teams$team == "stanford-cardinal"] <- lm.coef[teams$team=="stanford-cardinal"] <- 0
names(glm.coef) <- names(lm.coef) <- teams$team

## The logistic model is in almost perfect agreement with the USAT and AP ranks.
rank.table <- cbind("Linear Score" = lm.coef,
                    "Linear Rank"  = rank(-lm.coef,ties="min"),
                    "Logistic Score" = glm.coef,
                    "Logistic Rank"  = rank(-glm.coef,ties="min"),
                    "AP Rank"     = teams$apRank,
                    "USAT Rank"   = teams$usaTodayRank)
library(xtable)
xtable(rank.table[order(glm.coef,decreasing=TRUE)[1:25],],digits=c(1,2,0,2,0,0,0))

sum(summary(glm.mod)$coef[-1,4]<0.05)
sum(summary(lm.mod)$coef[-1,4]<0.05)

## 4d) We use both regular and postseason games for the cross-validation.
set.seed(1)
folds <- sample(1:10,nrow(X),replace=TRUE)
mcnemar.tab <- matrix(0,2,2)
dat <- data.frame(homeAdv, X)
for (fld in unique(folds)) {
  train <- folds != fld
  test <- folds == fld
  lm.mod <- lm(y ~ 0 + ., data=dat, subset=train)
  lm.correct <- sign(predict(lm.mod,newdata=dat[test,])) == sign(y[test])
  glm.mod <- glm(y.bin ~ 0 + ., data=dat, family=binomial, subset=train)
  glm.correct <- sign(predict(glm.mod,newdata=dat[test,])) == sign(y.bin[test] - 0.5)

  mcnemar.tab <- mcnemar.tab + table(lm.correct,glm.correct)
}
mcnemar.tab

2*pnorm(268,(268+215)/2,sqrt(268+215)/2,lower.tail=F)
```