Statistics 216
Homework 2, due Wednesday Feb 12, 2014.

1. Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right) \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

for a particular value of s. For parts (a) through (e), indicate which of i. through v. is correct. Justify your answer.

   (a) As we increase s from 0, the training RSS will:

      i. Increase initially, and then eventually start decreasing in an inverted U shape.

      ii. Decrease initially, and then eventually start increasing in a U shape.

      iii. Steadily increase.

      iv. Steadily decrease.

      v. Remain constant.

   (b) Repeat (a) for test RSS.

   (c) Repeat (a) for variance.

   (d) Repeat (a) for (squared) bias.

   (e) Repeat (a) for the irreducible error.

2. We will now derive the probability that a given observation is part of a bootstrap sample. Suppose that we obtain a bootstrap sample from a set of n observations.

   (a) What is the probability that the first bootstrap observation is *not* the jth observation from the original sample? Justify your answer.

   (b) What is the probability that the second bootstrap observation is *not* the jth observation from the original sample?

   (c) Argue that the probability that the jth observation is *not* in the bootstrap sample is $(1 - 1/n)^n$.

   (d) When $n = 5$, what is the probability that the jth observation is in the bootstrap sample?

   (e) When $n = 100$, what is the probability that the jth observation is in the bootstrap sample?

   (f) When $n = 10,000$, what is the probability that the jth observation is in the bootstrap sample?

   (g) Create a plot that displays, for each integer value of n from 1 to $100,000$, the probability that the jth observation is in the bootstrap sample. Comment on what you observe.

   (h) We will now investigate numerically the probability that a bootstrap sample of size $n = 100$ contains the jth observation. Here $j = 4$. We repeatedly create bootstrap samples, and each time we record whether or not the fourth observation is contained in the bootstrap sample.

```
> store=rep(NA, 10000)
> for(i in 1:10000){
    store[i]=sum(sample(1:100, rep=TRUE)==4)>0
}
> mean(store)
```

   Comment on the results obtained.

3. Logistic regression can give poor results when the two classes can be perfectly separated by a linear decision boundary. Consider just a logistic regression with a single predictor, X, so that our model is

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

Remember that the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ maximize the likelihood function

$$L(\beta_0, \beta_1) = \prod_{i:\, y_i=1} p(x_i) \prod_{i':\, y_{i'}=0} (1 - p(x_{i'}))$$

for the observed data $\{(x_i, y_i)\}_{i=1}^n$.

(a) Show that the likelihood function $L(\beta_0, \beta_1)$ is always strictly less than 1.

(b) Suppose that all of the x_i corresponding to $y_i = 0$ are negative, and all of the other x_i are positive. In that case, show that we can get $L(\beta_0, \beta_1)$ arbitrarily close to 1. That is, show that for any value $a < 1$, no matter how close to 1, you can always find values β_0 and β_1 for which $L(\beta_0, \beta_1) > a$.

Explain why this means that $\hat{\beta}_0$ and $\hat{\beta}_1$ are undefined.

(c) Show that $\hat{\beta}_0$ and $\hat{\beta}_1$ are similarly undefined if there is *any* value c for which every x_i corresponding to $y_i = 0$ is less than c and every other x_i is larger than c.

In fact, the same is true for multivariate logistic regression. Whenever there is a linear decision boundary that perfectly separates the two classes, the maximum likelihood logistic regression coefficients are undefined (but you don't have to prove this last fact).

(d) Come up with your own data set of the form in (c) and fit a logistic regression to it in R. Plot your data, as well as the logistic regression fit $\hat{p}(x)$.

You will probably get warning messages that the fit didn't converge, and that you have numerically 0 or 1 fitted probabilities. The first message usually signals that you have fit a logistic regression to perfectly separable classes.

4. **You may work in groups up to size 4 on this problem. If you do work in groups, write the names of all your group members on your problem set.**

In class, we fit a linear regression model to the NCAA data using the scores as a continuous response. y_i represented the (possibly negative) margin of victory for the home team, and (including home court advantage) we modeled:

$$y_i = \beta_0 h_i + \beta_{\text{home}(i)} - \beta_{\text{away}(i)} + \epsilon_i$$

One criticism of this model is that it may give teams too much credit for running up the score on their weaker opponents. Increasing one's margin of victory from 30 to 35 points is assigned the same importance as changing a 2-point loss to 3-point win.

To answer this criticism, we could change our model so that it only takes win-loss outcomes into account. Defining z_i now as a 0/1 indicator of whether the home team won the game, the "logistic regression version" of our model for $P(z_i = 1)$ is of the form:

$$p(\text{home}(i), \text{away}(i), h_i) = \frac{e^{\beta_0 h_i + \beta_{\text{home}(i)} - \beta_{\text{away}(i)}}}{1 + e^{\beta_0 h_i + \beta_{\text{home}(i)} - \beta_{\text{away}(i)}}}$$

Except for the home-team advantage term, this is a version of the famous *Bradley-Terry Model*, used in college football, chess, and elsewhere to compute automatic team rankings.

(a) Fit the logistic regression model above to the data and examine the rankings. What happened to make the teams `saint-mary-saint-mary` and `st.-thomas-(tx)-celts` look so good? Can you explain it in terms of your answer to the previous question?

(b) Get rid of teams that played less than five games and refit the model. Make a rank table like the ones we made in class, where you compare the logistic regression rankings to the linear regression rankings, the AP Rankings, and the USA Today rankings. Which model seems to correspond better to the voters' decisions, the linear regression or logistic regression?

(c) When we ignore the actual value of y_i and instead only use whether $y_i > 0$, we are discarding information, so we might expect our model standard errors to be larger relative to the effect sizes. If we use the linear regression model, for what fraction of teams are we confident ($p < 0.05$) that the team is better (or worse) than Stanford? For what fraction are we confident if we instead use the logistic regression model?

(d) Each model makes a prediction about which team will win any given game. We can use ten-fold cross-validation to estimate the test error rate for these predictions, and also try to determine whether one model is better than the other.

For each game in a given test set, there are four possible outcomes: both models are right in their prediction, both are wrong, only logistic regression is right, or only linear regression is right. Make a 2×2 contingency table displaying how often each of these four outcomes occur, over all the test examples in all ten folds of cross-validation. Your table should look like:

	logistic right	logistic wrong
linear right	n_{11}	n_{12}
linear wrong	n_{21}	n_{22}

(e) n_{11} and n_{22} don't tell us anything about which model is better, because they correspond to games where both models agree with each other. So to compare the two models, we need to look at n_{12} and n_{21}. Let $D = n_{12} + n_{21}$ be the number of test games in which the two models *disagreed*.

If both models are equally good and the test set games are independent, then every time the models disagree, each model is equally likely to be right. Then, conditional on D,

$$n_{12} \sim \text{Binom}(D, 1/2)$$

For large D, the above binomial distribution is approximately normal with mean $D/2$ and variance $D/4$ (hence standard deviation $\sqrt{D}/2$). **You do not have to prove any of the above statements, just take them as given.**

Use the normal approximation to carry out a test of the hypothesis that both models are equally good at predicting games. What is the conclusion of your test?

What you just did is called *McNemar's Test*, and it is the correct way of comparing the performance of two classifiers on a test set.