Swarthmore Honors Exam 2012: Statistics

John W. Emerson, Yale University

NAME: ____

Instructions:

This is a closed-book three-hour exam having six questions. You may not refer to notes or textbooks. You may use a calculator that does not do algebra or calculus. Normal, t, and F tables should be supplied with this exam. Please note:

- A few of the questions have a considerable amount of description and background which is intended to help you. Read this material carefully and completely before working on the problem.
- Most questions have multiple parts (only question 4 is a single question). Number the questions and parts clearly in your work, and start each of the questions on a new page.
- Questions that explicitly ask for (or imply the need for) discussion are a chance to demonstrate your understanding of the material. As a guideline, a short paragraph is likely appropriate and preferable to either a single sentence or an longer essay.

1. A basketball coach wants to devise an innovative way to test whether her players have improved by practicing over the summer. The previous season, her star, Julie, made 70% of her free throws. Julie claims to have improved over the summer; she now thinks she is a 80% free throw shooter. The coach doubts Julie has improved; she asks Julie to start shooting free throws. Let X be a random variable corresponding to the number of consecutive free throws Julie makes before her first miss. If we assume Julie's free throws are independent and that she makes each shot with some fixed, unknown probability p, the geometric distribution would seem appropriate:

$$P(X = x) = p^x(1-p), \text{ for } x = 0, 1, 2, \ldots$$

The following table presents these probabilities for two values of p, 0.8 and 0.7. The last row gives the probabilities of making 20 or more shots before the first miss.

x	$P(X = x \vert p = 0.8)$	$P(X = x \vert p = 0.7)$
0	0.2000	0.3000
1	0.1600	0.2100
2	0.1280	0.1470
3	0.1024	0.1029
4	0.0819	0.0720
5	0.0655	0.0504
6	0.0524	0.0353
7	0.0419	0.0247
8	0.0336	0.0173
9	0.0268	0.0121
10	0.0215	0.0085
11	0.0172	0.0059
12	0.0137	0.0042
13	0.0110	0.0029
14	0.0088	0.0020
15	0.0070	0.0014
16	0.0056	0.0010
17	0.0045	0.0007
18	0.0036	0.0005
19	0.0029	0.0003
≥ 20	0.0115	0.0009

a. Consider testing $H_0 : p = 0.7$ versus $H_a : p > 0.7$. Derive the rejection region for the test corresponding to significance level $\alpha = 0.05$.

b. What is the probability of a Type I error? Explain this concept in the context of this problem to Julie's coach, who has never studied statistics.

c. What is the power of this test against Julie's proposal that she improved and is actually a 80% free throw shooter? Again, explain this concept in the context of this problem to Julie's coach, who has never studied statistics.

2. Suppose $\{X_i\}$ is a set of $n \geq 1$ independent identically distributed (*iid*) random variables from the uniform distribution on the interval $(0, \theta)$ for $0 < \theta < \infty$.

a. Find the maximum likelihood estimator (MLE) for θ.

b. Prove that the MLE is a biased estimator for θ.

c. Prove that the MLE is a consistent estimator for θ.

d. Propose an unbiased estimator for θ. Is this estimator preferable to the MLE? Include a discussion of the basis for your preference, as if you were presenting this solution to a fellow student of mathematical statistics.

3. Suppose X_1, X_2, ... are *iid* Bernoulli random variables such that $P(X_i = 1) = p$, with p some fixed value in $(0, 1)$. Define L_1 and L_2 to be the lengths of the first and second "runs," respectively, in the sequence generated by the X_i's. A "run" is a collection of consecutive common outcomes. So, for example, the sequence 0, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0, ... would produce $L_1 = 3$, $L_2 = 2$, $L_3 = 1$, and $L_4 = 5$.

a. What is the distribution of L_1?

b. What is the distribution of L_2?

c. Are L_1 and L_2 independent? Discuss.

4. A coin will be tossed once, and we are interested in estimating the probability of heads, p. A Bayesian will propose using a uniform prior distribution on p,

$$g(p) = 1, \ 0 \leq p \leq 1,$$

and will estimate p using the mean of the posterior distribution. A frequentist will use the maximum likelihood estimator (MLE) for p. Show how to derive both the MLE and the Bayes estimator. Using squared error loss, describe (exactly or approximately, with justification) the range of values of p for which the MLE is preferable to the Bayes estimator.

5. This problem examines data showing the effect of two soporific (sleep-inducing) drugs. The variable `extra` shows the increase in hours of sleep for each of $n = 20$ patients who had been randomly assigned to two groups for the study. A few observations are shown here:

```
  extra group
1   3.4     0
2   0.8     0
3   1.9     1
4   4.4     1
5  -1.2     0
```

There are $n_0 = n_1 = 10$ subjects in each group. In terms of underlying probability models, you may assume $EXTRA_i \sim N(\mu_0, \sigma^2)$ for subject i in group 0, and $EXTRA_j \sim N(\mu_1, \sigma^2)$ for subject j in group 1. The sample variance of all 20 measurements is

$$s^2_{total} = \frac{1}{n-1}\sum_{i=1}^{n}(extra_i - \overline{extra})^2 = 4.0720,$$

where the overall mean is

$$\overline{extra} = \frac{1}{n}\sum_{i=1}^{n} extra_i = 1.5400.$$

The sample variance of measurements in group 0 is s_0^2 =3.2006, and the variance of measurements in group 1 is s_1^2 =4.009. A pooled 2-sample t-test is performed, making use of the pooled estimate of the variance,

$$s^2_{pooled} = \frac{(n_0 - 1) * s_0^2 + (n_1 - 1) * s_1^2}{n_0 + n_1 - 2} = 3.6048,$$

and giving the following result:

```
        Two Sample t-test

data:  extra by group
t = -1.8608, df = 18, p-value = 0.07919
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.363874  0.203874
sample estimates:
mean in group 0 mean in group 1
           0.75            2.33
```

a. Show how to obtain the confidence interval $(-3.363874, 0.203874)$ from the information provided, above. Clearly define any quantities used.

b. Critique the following statement: The 95% confidence interval $(-3.363874, 0.203874)$ contains the true difference in drug effectiveness with probability 0.95.

c. Consider the linear model

$$EXTRA_i = \alpha + \beta * group_i + \varepsilon_i$$

for $i = 1, 2, \ldots, 20$, where $group_i$ takes values 0 and 1 as described above for this data set and standard regression assumptions include that ε_i are *iid* $N(0, \sigma^2)$. We observe $extra_i$ in our data set (and have various helpful summary statistics on the previous page of this exam). We fit the model and obtain estimates $\widehat{\alpha}$ and $\widehat{\beta}$, provided below. Complete the following regression summary and analysis of variance (ANOVA) table that would be associated with fitting this model using these data.

A word of caution: This final part of this problem should not require memorization of general regression or analysis of variance formulae. Almost everything involves means and variances (or associated sums of squares) closely related to quantities shown above. *End of word of caution.*

```
Coefficients:

            Estimate    Std. Error    t value   Pr(>|t|)

(Intercept)     0.75    __________    _______   ________

group           1.58    __________    _______   ________

Residual standard error: _________ on ____ degrees of freedom

Multiple R-squared: ______
```

```
Analysis of Variance Table

Response: extra

              Df    Sum Sq    Mean Sq    F value     Pr(>F)

group       ____    ______    _______    _______    0.07919

Residuals   ____    ______    _______
```

6. This problem examines basketball games from the 2007-08 NBA season. Prior to each game, bookies publish “point spreads” which are predicted point differentials. For instance, if Team A is thought to be stronger than Team B by a margin of 10 points, the point spread (from the perspective of Team A) would be -10. After each game, of course, we can observe the actual point differential (the score for Team B minus the score for Team A in this example, from the perspective of Team A) and compare this value to the bookies’ prediction. Thus, an actual point differential of -9 (Team A in fact winning by 9 points) would be very close to the predicted 10 point advantage (the published point spread of -10).

A few raw data entries from the perspective of the Boston Celtics (thought to be a stronger team) follow, where `-3.5` in the 11/27 game for Boston versus Cleveland (V shows that Boston was visiting Cleveland) indicates that Boston was favored by 3.5 points (and then lost the game, 104-109):

```
Boston versus (3 sample game results):
11/27  Cleveland        -3.5  104-109   V
11/29  New York        -12.5  104-59    H
11/30  Miami            -3.0   95-85    V
```

A few raw data entries from the perspective of the New York Knicks (thought to be a weaker team) follow, where `+5.5` in the 11/26 game for New York versus Utah (H shows that it was a home game for New York) indicates that Utah was favored by 5.5 points (but New York won, 113-109):

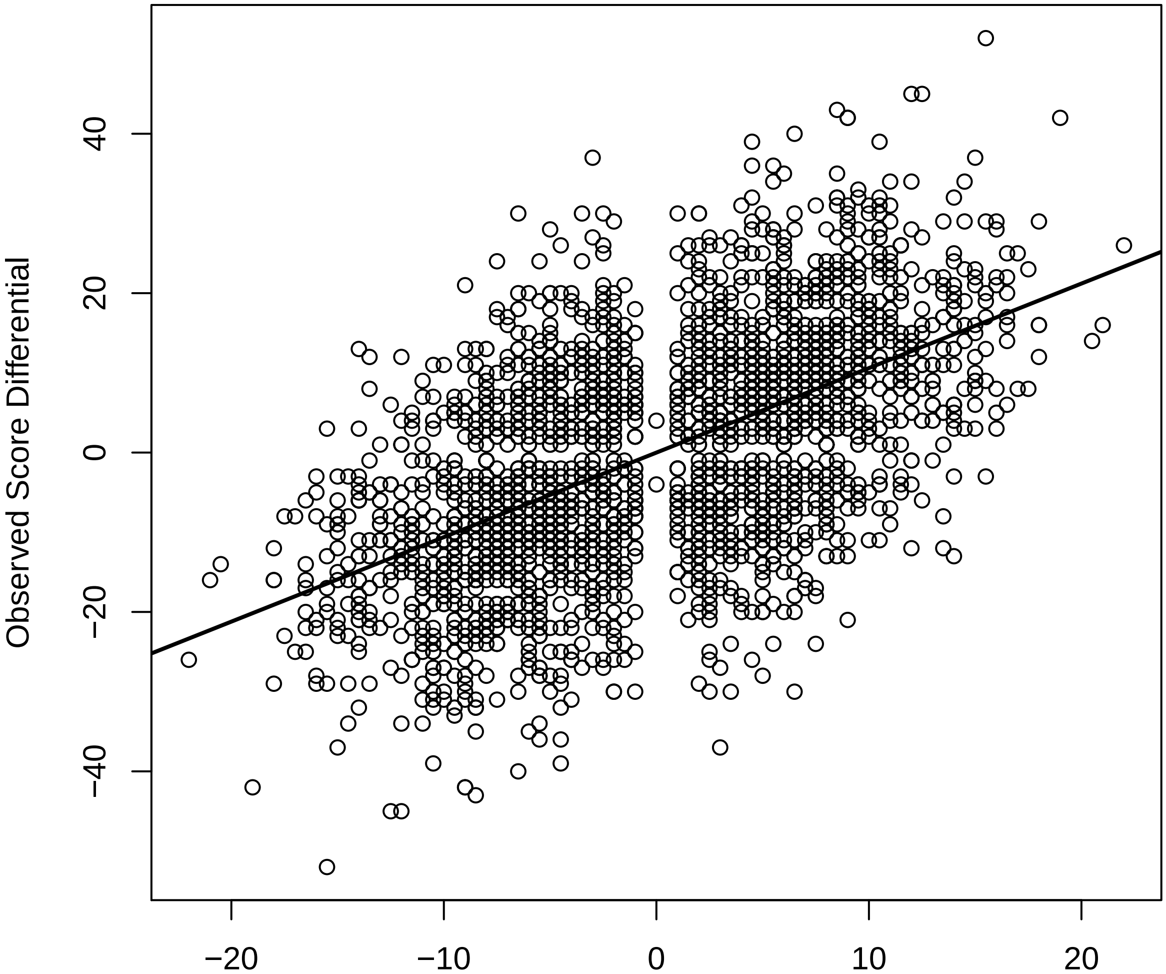
```
New York versus (3 sample game results):
11/26  Utah             +5.5  113-109   H
11/29  Boston          +12.5   59-104   V
11/30  Milwaukee        +2.0   91-88    H
```

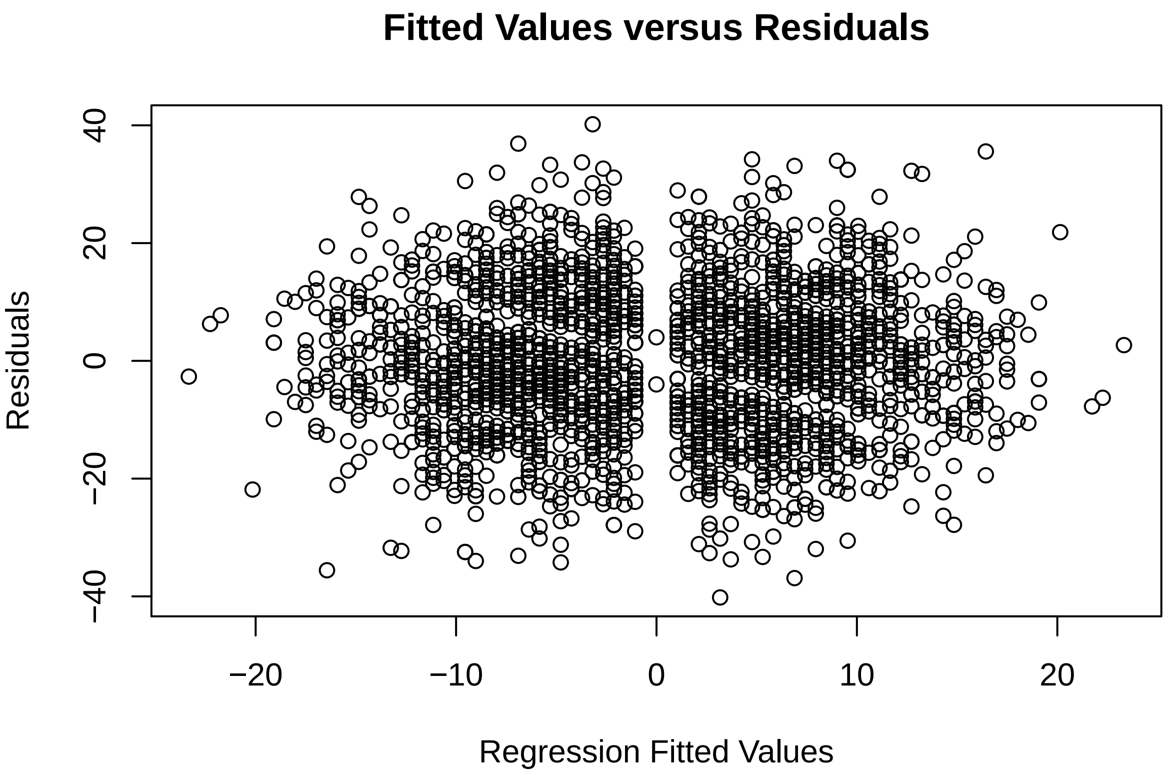
There are 2496 such entries in this data set, yielding the plot on the next page and the regression summary given as “Problem 6, regression summary” two pages later. Two plots of residuals are also provided for your consideration.

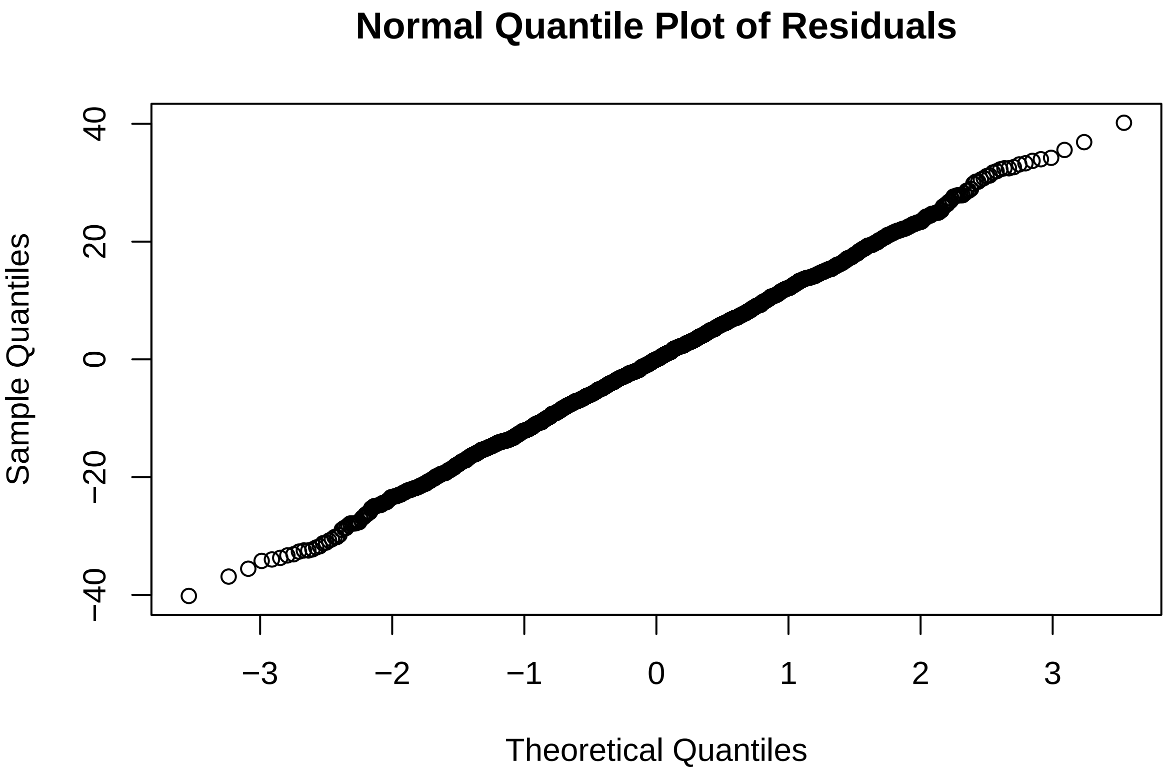
A reporter proposes a two-sided hypothesis test of an “efficient market theory” that the unknown true slope, β, of the relationship between bookies’ point spreads and the observed game point differential is equal to one. That is, if the gambling market were efficient, the bookies’ point spreads prior to the game would be unbiased predictors of the game point differentials.

a. Find the p-value for the reporter’s desired test based on the regression output, “Problem 6, regression summary.” This is not available directly from the regression output; show your work. Explain to the reporter, who has never studied statistics, the meaning of this p-value.

b. In fact, a very sensible solution to part a, receiving full credit above, could still be misleading. What do you think? If there is still a problem, explain whether you think it really matters in terms of the desired test. If you think it is appropriate and feasible (given that you don’t have access to a computer), offer an approximate correction of this p-value and discuss any assumptions you make in doing so. If you don’t feel it is appropriate or feasible, explain why. In either case, defend your reasoning.


Observed Score Differential
−40
−20
0
20
40
−20
−10
0
10
20
Bookie Spread (prior to start of game)


Fitted Values versus Residuals
Residuals
40
20
0
−20
−40
−20
−10
0
10
20
Regression Fitted Values


Normal Quantile Plot of Residuals
Sample Quantiles
40
20
0
−20
−40
−3
−2
−1
0
1
2
3
Theoretical Quantiles

Problem 6, regression summary.

```
Residual summary:
    Min      1Q  Median      3Q     Max
-40.177  -7.836   0.005   7.845  40.186

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.004692   0.239000   -0.02    0.984
pointspread  1.060423   0.029207   36.31   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.94 on 2494 degrees of freedom
Multiple R-squared: 0.3165,       Adjusted R-squared: 0.3162
F-statistic:  1155 on 1 and 2494 DF,  p-value: < 2.2e-16
```

An aside, in case you were curious: You can see evidence of a “crossing” pattern in the scatterplot. Basketball games can’t end in a tie, explaining the lack of 0 values for the observed game score differentials. Less obvious is the fact that bookies don’t publish point spreads of -0.5 or 0.5 (and point spreads equal to 0 are more unusual than you might expect). This creates the vertical gap in the plot.