This exam is designed to be TWICE as long your final.

Name: ___________________________________________________ Section (circle): { 01 – Morning  
                                        02 – Afternoon  
                                        03 – Evening }

I pledge my honor that I have not violated the Chicago Booth Honor Code during this exam:

 Signed: _______________________________________________

• You have 3 hours to complete the exam.
• This exam has 13 pages.
• Do not spend an inordinate amount of time on any one problem. Some questions are harder than others. Many questions on the exam are independent of each other.
• The exam is meant to be too long for everyone to finish. Don’t worry.
• You may use a calculator and one 8.5 × 11 size (both sides) “cheat sheet” of your own notes, otherwise the exam is closed book, closed notes, etc.
• Throughout, when calculating probabilities or intervals, you can assume that:
  – 95% of observations will fall within 2 standard deviations of the mean.
  – 90% of observations will fall within 1.6 standard deviations of the mean.
• Present your answers in a clear and concise manner.
• Do not write your name on any page except this one.

GOOD LUCK!!
1 Simple Linear Regression Mechanics

You run simple linear regression of $Y$ on $X$. The estimated regression line is $\hat{Y} = 1 + 2X$. The $t$-statistic for testing the null hypothesis $H_0 : \beta_1 = 0$ against the alternative $H_1 : \beta_1 \neq 0$ is 2.1, and you reject the null at the 5% level. The sample mean of $Y$ is 7.

(a) What is the standard error of $b_1$?

(i) 1.48
(ii) 0.95
(iii) 1.96
(iv) 0.72
(v) None of these

(b) Suppose you obtain one more data point, and re-estimate the regression using the original data together with the new data point. The new estimated regression line is $\hat{Y} = 1 + 2X$. Using the original data together with the new data point you re-construct the $t$-statistic for testing the null hypothesis $H_0 : \beta_1 = 0$ against the alternative $H_1 : \beta_1 \neq 0$. Call this new $t$-statistic $t'$. Which of the following is TRUE?

(i) $t' > 2.1$
(ii) $t' = 2.1$
(iii) $t' < 2.1$
(iv) Cannot determine the relationship between $t'$ and 2.1 with the information given.

(c) Which of the following could potentially correspond the the new observation in part (e)?

(i) $X = 0, Y = 1$
(ii) $X = 2, Y = 5$
(iii) $X = -3, Y = -5$
(iv) $X = 1, Y = 3$
(v) All of the above
(vi) None of the above
2 Short Answer & Multiple Choice

(a) What would the value of $R^2$ be if you estimated a regression model with only an intercept? Assume $\text{var}(Y) > 0$.

(i) 1
(ii) 0
(iii) 0.5
(iv) We cannot tell unless the sample size is given.
(v) We cannot tell even if the sample size is given.

(b) List all simple linear regression assumptions that might not be satisfied for the following data. You do not need to suggest any fixes.

\[
\begin{array}{c}
\text{X} \\
\hline
0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\
\text{Y} \\
\end{array}
\]

(c) The relationship between $Y$ and $X$ in the data in part (b) can be characterized by three (3) features. List them (order does not matter).

(i)
(ii)
(iii)
3 Simple linear regression

The data for this question includes results for the 2008 presidential election as well as income information for each state/district in the lower continental US. We will be relating $V_{\text{Dem}}$, Obama’s share of the presidential vote in each state, to $\text{INC}$, the median income (in $\$1000$) for each state.

Consider the following plots for the model $\mathbb{E}[V_{\text{Dem}} | \text{INC}] = \beta_0 + \beta_1 \text{INC}$.

(a) What problem do you see? How would you fix it? Justify your changes to the regression.

(b) Do you think that your fix will lead to big change in the least-squares regression coefficients? In other words, would you trust the model we’ve already fit? Why or why not?
After correcting the problem in part (1), the regression output is as follows.

**Call:**
```
lm(formula = VDem ~ INC)
```

**Coefficients:**

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 0.167884 | 0.069429 | 2.418 | 0.0196 * |
| INC | 0.006695 | 0.001336 | 5.010 | 8.51e-06 *** |

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07434 on 46 degrees of freedom
Multiple R-squared: 0.353, Adjusted R-squared: 0.3389
F-statistic: 25.1 on 1 and 46 DF, p-value: 8.509e-06

(c) How do the results from this regression relate to those shown in the graph below, plotting the share of the democratic vote across income groups in three states, and the USA at large? Are these at odds, or in agreement? In either case, do you have a explanation?
4 Multiple Linear Regression

We want to assess the effect of participating in a job training program on future earnings. For 2,675 men, we observe their participation status (job.training = 1 if they received training, 0 if not) and earnings.after the program completed (in dollars, regardless of whether they received training). We also observe their earnings.before the program, education in years, age, and binary indicators for being black, hispanic, married, and a high.school.grad.

The goal of the study is to assess the association between job.training and earnings.after. The question is what else to control for, i.e. we must build a model.

(a) First consider the below output from a simple linear regression of earnings.after on job.training.

```
Call:
  lm(formula = earnings.after ~ job.training)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)  
(Intercept) 21553.9     303.6   70.98  <2e-16 ***
job.trainingYes -15204.8     1154.6  -13.17  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15150 on 2673 degrees of freedom
Multiple R-squared:  0.06092,   Adjusted R-squared:  0.06057
F-statistic: 173.4 on 1 and 2673 DF,  p-value: < 2.2e-16
```

(i) Give a precise interpretation of the coefficient estimate for job.training. Based on this output, what do you conclude about the relationship between participation in the training program and earnings afterward? State a formal hypothesis test, including null and alternative hypotheses, significance level, test statistic value, and conclusion.

(ii) Give a reasonable explanation for why the coefficient is negative.
(b) Now consider the following multiple regression output, which controls for \texttt{earnings.before} the program and years of \texttt{education}.

Call:
\texttt{lm(formula = earnings.after \sim job.training + earnings.before + educ)}

Coefficients:

\begin{tabular}{lcccc}
 & Estimate & Std. Error & t value & Pr(>|t|) \\
(Intercept) & -2.138e+03 & 8.160e+02 & -2.620 & 0.00883 ** \\
job.trainingYes & 4.792e+02 & 8.219e+02 & 0.583 & 0.55992 \\
earnings.before & 8.360e-01 & 1.657e-02 & 50.436 & < 2e-16 *** \\
educ & 6.275e+02 & 6.879e+01 & 9.121 & < 2e-16 *** \\
\end{tabular}

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10160 on 2671 degrees of freedom
Multiple R-squared: 0.5777, Adjusted R-squared: 0.5772
F-statistic: 1218 on 3 and 2671 DF, p-value: < 2.2e-16

(i) Give a precise interpretation of the coefficient estimate for \texttt{job.training}. Based on this output, what do you conclude about the relationship between participation in the training program and earnings afterward? State a formal hypothesis test, including null and alternative hypotheses, significance level, test statistic value, and conclusion.

(ii) What does this conclusion say about the training program?
(iii) Consider the four diagnostic plots below. Referring to each by number, what do you conclude about this regression? What problems do you see, and what fixes do you propose for these problems?
(c) Now consider expanding the regression to include all the observed variables described above. We have the following regression output.

**Call:**
```
lm(formula = earnings.after ~ job.training + earnings.before + educ + age + black + hispanic + married + high.school.grad)
```

**Coefficients:**

|                      | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------------|----------|------------|---------|----------|
| (Intercept)          | -43.50987| 1697.68742 | -0.026  | 0.9796   |
| job.trainingYes      | 714.92999| 920.24279  | 0.777   | 0.4373   |
| earnings.before      | 0.84832  | 0.01743    | 48.658  | < 2e-16 *** |
| educ                 | 597.11088| 103.86503  | 5.749   | 1.0e-08 *** |
| age                  | -88.46638| 20.90856   | -4.231  | 2.4e-05 *** |
| blackYes             | -621.57753| 497.80243  | -1.249  | 0.2119   |
| hispanicYes          | 1904.11397| 1097.24813 | 1.735   | 0.0828 . |
| marriedYes           | 1184.57915| 589.37459  | 2.010   | 0.0445 * |
| high.school.gradNo   | 610.11290 | 650.31497  | 0.938   | 0.3482   |

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10130 on 2666 degrees of freedom
Multiple R-squared: 0.5817, Adjusted R-squared: 0.5804
F-statistic: 463.4 on 8 and 2666 DF, p-value: < 2.2e-16

We also have the Analysis of Variance Table below, where “Model 2” is the current regression and “Model 1” is the regression in part (b).

<table>
<thead>
<tr>
<th>Model</th>
<th>Formula</th>
<th>Res.Df</th>
<th>RSS</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>F</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>earnings.after ~ job.training + earnings.before + educ</td>
<td>2671</td>
<td>2.7597e+11</td>
<td>1</td>
<td>2.622e+09</td>
<td>5.144</td>
<td>0.0001133 ***</td>
</tr>
<tr>
<td>Model 2</td>
<td>earnings.after ~ job.training + earnings.before + educ + age + black + hispanic + married + high.school.grad</td>
<td>2666</td>
<td>2.7335e+11</td>
<td>2</td>
<td>5.622e+09</td>
<td>5.144</td>
<td>0.0001133 ***</td>
</tr>
</tbody>
</table>

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(i) Give a precise description of the p-value in the Analysis of Variance Table above. Include the null and alternative hypotheses of the relevant test, what these mean conceptually for comparing “Model 1” and “Model 2”, and what you conclude based on this output.

(ii) “Model 1” from part (b) has a BIC value of 49390.32, whereas “Model 2” has a BIC value of 49404.24. Based on this information and the Analysis of Variance Table, which model do you prefer? Why?
5 Classification & Model Building

Our goal for this question is to build an email spam filter: based on observed characteristics of an email message we want to build a classification rule for assigning the message either as spam (marked with a “1”) or not spam (“0”). To build our filter we have a data of 4601 emails, and for each message we have a human-assigned label spam (1 for spam, 0 for not spam) and the following characteristics:

- **caps_avg** = the average of the lengths of strings of capital letters used in the email (e.g. “The” = 1, “HELLO” = 5)
- **c(paren), c(exclaim), c(dollar)** = the percentage of characters in the message which are parentheses (“(”, “[”), “)”, “]”), exclamation point (“!”), and dollar sign (“$”) respectively. (Percentages are between 0 and 100.)

(a) In our current context, what are the two types of errors that a classifier can make?

   (i) 

   (ii) 

   In the present context, is one type of mistake “worse” than the other? Explain your reasoning.

Use the following output to answer parts (b) – (f).

Call:
```
glm(formula = spam ~ caps_avg + c(paren) + c(exclaim) + c(dollar),
    family = "binomial", data = spam)
```

Coefficients:

|              | Estimate | Std. Error | z value | Pr(>|z|) |
|--------------|----------|------------|---------|----------|
| (Intercept)  | -1.75    | 0.07       | -25     | <2e-16   *** |
| caps_avg     | 0.21     | 0.02       | 12      | <2e-16   *** |
| c(paren)     | -1.66    | 0.23       | -7      | 2e-13    *** |
| c(exclaim)   | 1.38     | 0.11       | 12      | <2e-16   *** |
| c(dollar)    | 11.86    | 0.62       | 19      | <2e-16   *** |

Null deviance: 6170.2  on 4600 degrees of freedom
Residual deviance: 4160.7  on 4596 degrees of freedom
AIC: 4171

Number of Fisher Scoring iterations: 15
(b) Provide a precise, numerical interpretation of the coefficient estimate for $c_{\text{dollar}}$. Do you find this result credible? Why or why not?

(c) Provide a precise, numerical interpretation of the coefficient estimate for $\text{caps}_{\text{avg}}$. Do you find this result credible? Why or why not?

(d) For a message that is 2% parentheses, 2% exclamation points, has zero dollar signs, and never strings together more than one capital letter, what estimated probability this message is spam?

(e) We will use the regression in above to build a classification rule based on the predicted probabilities. For some number $K$, we will flag a message as spam if the estimated $P[\text{spam} = 1 | X] > K$. Referring to your answer in part (a), would you prefer to choose $K = 1/4$, $K = 1/2$, or $K = 3/4$? Why?
(f) For $K = 0.5$, the table below compares the classification results to the human-assigned labeling. Use the table to compute the rates of the two types of errors in part (a).

<table>
<thead>
<tr>
<th>Classifier</th>
<th>$\leq 0.5$</th>
<th>$&gt; 0.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human-assigned: spam==0</td>
<td>2658</td>
<td>130</td>
</tr>
<tr>
<td>spam==1</td>
<td>628</td>
<td>1185</td>
</tr>
</tbody>
</table>

List error rate results:

(i) 

(ii) 

In addition to the variables defined above, we also have word counts for 43 different words, which are stored in variables like $w_{\text{<word>}}$. For example $w_{\text{credit}}$ is the number of times the word “credit” appears in the message. We want to build a classifier that includes these word counts only if that word is relevant for predicting spam. We will take the train/test approach, holding out 1000 messages for testing. We will choose variables using (i) forward stepwise selection based on AIC, (ii) on BIC, and (iii) selection using the LASSO.

(g) Briefly explain in words the three approaches to model building, and how they differ from each other and what advantages and disadvantages they each have.

(i) Stepwise AIC:

(ii) Stepwise BIC:

(iii) LASSO:
On the 1000 held-out messages, we obtain predictions from each selected model and flag a message as spam if the estimated $P[\text{spam} = 1|X] > 0.5$ (just like part (f)). For each of the three models, we then computed the following for each of the 1000 held-out messages:

$$\text{error} = \text{spam} - 1 \{P[\text{spam} = 1|X] > 0.5\}$$

where $1 \{P[\text{spam} = 1|X] > 0.5\}$ indicates if the estimated $P[\text{spam} = 1|X] > 0.5$. We obtained the following results.

<table>
<thead>
<tr>
<th></th>
<th>error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-1</td>
</tr>
<tr>
<td>AIC</td>
<td>40</td>
</tr>
<tr>
<td>BIC</td>
<td>50</td>
</tr>
<tr>
<td>LASSO</td>
<td>20</td>
</tr>
</tbody>
</table>

(h) Based on these results, which method of model selection do you prefer and why? Refer to your answer in part (a).