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

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 15 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 Short Answer/Multiple Choice

(a) Below are 4 scatter plots of an outcome $y$ versus predictor $x$ followed by four regression fit summaries labeled A, B, C and D. Label each plot according to the corresponding summary.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>intercept</th>
<th>slope</th>
<th>residual standard error</th>
<th>SSR/SST</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$b_0 = 8.1$, $s_{b_0} = 0.11$</td>
<td>$b_1 = 2.1$, $s_{b_1} = 0.066$</td>
<td>$s = 1.08$</td>
<td>$R^2 = 0.90$</td>
</tr>
<tr>
<td>B</td>
<td>$b_0 = 8.0$, $s_{b_0} = 0.10$</td>
<td>$b_1 = 2.0$, $s_{b_1} = 0.017$</td>
<td>$s = 1.01$</td>
<td>$R^2 = 0.99$</td>
</tr>
<tr>
<td>C</td>
<td>$b_0 = 1.0$, $s_{b_0} = 0.10$</td>
<td>$b_1 = 2.0$, $s_{b_1} = 0.060$</td>
<td>$s = 0.97$</td>
<td>$R^2 = 0.93$</td>
</tr>
<tr>
<td>D</td>
<td>$b_0 = 0.9$, $s_{b_0} = 0.20$</td>
<td>$b_1 = 1.9$, $s_{b_1} = 0.120$</td>
<td>$s = 2.09$</td>
<td>$R^2 = 0.71$</td>
</tr>
</tbody>
</table>

(b) The following quantities summarize a least-squares regression: $n = 15$, $\bar{Y} = 3$, $\bar{X} = 4$, $s_X = 2$ and $b_1 = -2$, and $s_{b_1} = 3$. Give a prediction 95% interval for a new input $X_f = 1$.

(c) Suppose data are gathered on a $Y$ that is linearly related to $X$, possibly with noise. Which of the following is always true about a least-squares fit? Circle all that apply.

(a) The 95% confidence interval for $b_0$ contains zero.
(b) The 95% confidence interval for $b_1$ does not contain zero.
(c) The marginal variance of $Y$ is larger than the conditional variance of $Y$ given $X$.
(d) $0 < R^2 \leq 1$.
(e) We will reject the null hypothesis that $\beta_1 = 0$.
(f) none of these
(d) 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.

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

(f) 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)
2 Understanding regression output

Call:
`lm(formula = keystne ~ valmrkt)`

Residuals:

<table>
<thead>
<tr>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.32706</td>
<td>-0.02290</td>
<td>0.00202</td>
<td>0.02220</td>
<td>0.18405</td>
</tr>
</tbody>
</table>

Coefficients:

| Estimate  | Std. Error | t value | Pr(>|t|) |
|-----------|------------|---------|---------|
| (Intercept) | -0.003770  | 0.003218 |  (??)  |
| valmrkt   | 1.513719   | 0.066552 | 22.745 | <2e-16 *** |

---

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

Residual standard error: 0.04274 on 178 degrees of freedom
Multiple R-squared: 0.744, Adjusted R-squared: 0.7426
F-statistic: 517.3 on 1 and 178 DF, p-value: < 2.2e-16

From the above summary of the regression of returns for a Keystone mutual fund (keystne) onto the value weighted market index return (valmrkt), answer the following:

(a) What is the correlation between the Keystone and market returns?

(b) What is a 95% confidence interval for the regression intercept?

(c) What is the t-statistic for a hypothesis test of whether or not the intercept is equal to zero? What do you conclude at significance level $\alpha = 0.05$?

(d) What is the t-statistic for a hypothesis test of whether or not the slope is equal to one? What do you conclude at significance level $\alpha = 0.05$?
3 The Regression Model

Assume the following simple linear regression model:

\[ Y = 6 + \beta_1 X + \epsilon, \quad \epsilon \overset{iid}{\sim} \mathcal{N}(0, 3^2), \quad X \overset{iid}{\sim} \mathcal{N}(2, 4^2) \]

(a) If \( \beta_1 = 2 \), what \( X \) value will give us \( \mathbb{E}[Y|X] = 0 \)?

(b) Suppose \( X = 3 \). For what value of \( \beta_1 \) will the marginal variance of \( Y \) equal the conditional variance of \( Y|X \)? How does your answer change when \( X = 4 \)?

(c) If \( \beta_1 = 2 \), around what value would you expect for the \( R^2 \) from such a regression?

(d) Consider instead the log-log model \( \log(Y) = 6 + \beta_1 \log(X) + \epsilon \). What does this imply as a model for \( Y \) (i.e., \( Y \cdots \))? What is the approximate expected percentage change in \( Y \) per 1% increase in \( X \)?
4 Regression and Description: Chicago Restaurants

This question considers the 2008 Zagat survey of restaurants in the Chicago River North neighborhood. The data contains observations of 95 restaurants including ratings of price (in $), food, decor and service (on discrete and ordered scales).

(a) Comment on the relationship between price and the groupings created by the other three ratings. Be concise but thorough, noting patterns as well as concerns.
(b) Consider the summary output below.

Call:
  lm(formula = price ~ decor + service + food)

Residuals:
   Min     1Q   Median     3Q    Max
-18.6788 -4.0725  -0.1874  3.0993  22.2666

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept) -57.7114    5.4348  -10.619 < 2e-16 ***
decor        1.3011     0.2438   5.336  6.91e-07 ***
service      2.6620     0.4633   5.746  1.20e-07 ***
food         0.8979     0.3608   2.489   0.0146 *

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

Residual standard error: 7.03 on 91 degrees of freedom
Multiple R-squared: 0.7963, Adjusted R-squared: 0.7896
F-statistic: 118.6 on 3 and 91 DF,  p-value: < 2.2e-16

Briefly describe the model being fit and how the statistical test(s), and other information contained in the summary support/refute your initial impression(s) from part (i).

(c) Again referring to the summary, why might you want to use the transformed response: log(price)?
5 Regression: Baseball Data

This question considers performance statistics from the 2000 season for all Major League Baseball (MLB) teams. In particular, we want to determine the effect of a team’s total number of hits \((H)\) on their total number of runs scored \((R)\). Consider the following summary.

Call: lm(formula = R ~ H)

Residuals:

<table>
<thead>
<tr>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
</table>

Coefficients:

| Estimate  | Std. Error | t value | Pr(>|t|) |
|-----------|------------|---------|----------|
| (Intercept) | -309.7154 | 190.9087 | -1.622   | 0.116   |
| H          | 0.7572     | 0.1264  | 5.990    | 1.88e-06*** |

---

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

Residual standard error: 53.18 on 28 degrees of freedom
Multiple R-squared: 0.5617, Adjusted R-squared: 0.546
F-statistic: 35.88 on 1 and 28 DF,  p-value: 1.879e-06

(a) What percentage of the variation in team runs is explained by regression onto hits? Do we have reason to believe in a linear relationship between hits and runs? State the formal hypothesis test.
(b) What is a 95% confidence interval for the expected increase in runs scored corresponding to a single extra hit?

(e) What is the predicted number of runs \( \hat{R}_f \) for a team with 1440 hits \( H_f \)? If the standard error of this fitted value is \( sd(\hat{R}_f) \equiv s_{\text{fit}}(1440) = 10 \), what is a 95% prediction interval for runs of a 1440 hit team?
Consider the summary output of the following two regressions where the analysis is separated for the American and National leagues. (Some of the output is omitted to save space.)

American League:
Call:
`lm(formula = R[League == "American"] ~ H[League == "American"]`)

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | -159.1593 | 415.5425 | -0.383 | 0.7084 |
| H[League == "American"] | 0.6569 | 0.2685 | 2.447 | 0.0308 * |

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

National League:
Call:
`lm(formula = R[League == "National"] ~ H[League == "National"]`)

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | -486.0854 | 233.7741 | -2.079 | 0.0565 . |
| H[League == "National"] | 0.8796 | 0.1583 | 5.555 | 7.09e-05 *** |

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

Based on these summaries, is the change in the number of runs scored for each extra hit (statistically) different in the two leagues? State your hypothesis, formally, and give the test statistic and your conclusion.
(e) In all three regressions above, the estimated intercept is negative. Why might this be problematic? Argue, based on evidence given/calculated in parts (i–iv), that one reasonable fix is to set $\beta_0 = 0$ in each case. Does this completely solve the problem? If not, what might be a better fix?
Consider results for an expanded model which includes both hits and strike-outs as covariates.

Call: lm(formula = R ~ H + SO)

Coefficients:

|               | Estimate | Std. Error | t value | Pr(>|t|) |
|---------------|----------|------------|---------|----------|
| (Intercept)   | -1125.05 | 315.20     | -3.57   | 0.0014   ** |
| H             | 1.05     | 0.15       | 7.16    | 1.05e-07 *** |
| SO            | 0.36     | 0.12       | 3.05    | 0.0050   ** |

Residual standard error: 46.69 on 27 degrees of freedom
Multiple R-squared: 0.67, Adjusted R-squared: 0.65
F-statistic: 27.9 on 2 and 27 DF, p-value: 2.66e-07

Are strike-outs a useful predictor for runs, given that you already know the number of hits? What is the proportion of variability explained by our expanded model? What proportion of variability is explained by the introduction of SO as a covariate? Based on this summary information, which variables might you try transforming (if any)?
6 Classification

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:

(Intercept)  -1.75   0.07  -25  <2e-16 ***
caps_avg     0.21   0.02   12  <2e-16 ***
c_paren     -1.66   0.23  -7.0 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 $\text{caps_{avg}}$. 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: $\text{spam}==0$</td>
<td>2658</td>
<td>130</td>
</tr>
<tr>
<td>Human-assigned: $\text{spam}==1$</td>
<td>628</td>
<td>1185</td>
</tr>
</tbody>
</table>

List error rate results:

(i) 
(ii)