Chicago Booth BUS 41100
Midterm Sample #2
Instructor: Max H. Farrell

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) Use the space to the right of the plot to list which assumptions required by linear regression, if any, appear to be violated in the data set plotted below.

(b) If \( n = 25 \), \( \bar{Y} = -6 \), \( \bar{X} = 4 \), \( s_Y^2 = 9 \), \( s_X^2 = 16 \), and \( r_{xy} = 0.75 \), what are the least squares estimates of \( b_0 \) and \( b_1 \)? What is the \( R^2 \) from the least squares regression?

(c) Which of the following always results in a wider predictive interval for \( Y_f \) at a new location \( X_f \)? Circle all that apply.

- (a) a larger sample size \( (n) \)
- (b) a larger value of \( \hat{Y}_f \)
- (c) a larger degree of confidence (smaller \( \alpha \))
- (d) an \( X_f \) with lower leverage
- (e) a smaller estimated residual variance \( (s^2) \)
- (f) none of these
2 Understanding regression output #1

From the below summary of the regression of women’s labor force participation (WLFP) in nineteen cities in 1972 (wlfp72) on WLFP in 1968 (wlfp68), answer the questions below.

Call:
  lm(formula = wlfp72 ~ wlfp68)

Residuals:

          Min       1Q   Median       3Q      Max
-0.13086 -0.02797  0.01493  0.03678  0.06837

Coefficients:

                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              0.17435    0.09611  1.814  0.08736 .
wlfp68                   0.60513    0.18088  3.345  0.00383 **
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 0.05433 on 17 degrees of freedom
Multiple R-squared: 0.397, Adjusted R-squared: 0.3615
F-statistic: 11.19 on 1 and 17 DF,  p-value: 0.003835

(a) What is the t statistic for a hypothesis test of whether or not the slope is equal to one? Write out the null and alternative hypotheses, and explain what the test means in terms of WLFP. What do you conclude at significance level $\alpha = 0.05$? Discuss if you think this conclusion is reasonable given the degrees of freedom.

(b) Suppose that the correlation between $\log(\text{wlfp72})$ and $\log(\text{wlfp68})$ was 0.67. Based on this information, would you describe the corresponding log-log model as a better fit? Why?
3 True or False

For each question, circle either T (true) or F (false). Answering “true” implies that the given statement is always true. Statements are made in the context of this class, and the usual SLR/MLR assumptions.

(a) T F Forecast uncertainty for $Y_f$ does not depend on the input $X_f$.

(b) T F A confidence interval for $\beta_1$ is centered at $\beta_1$.

(c) T F Our linear regression model implies an error variance that is the same for all values of the explanatory variable.

(d) T F An observation with a studentized residual of more than 10 is probably an outlier.

(e) T F Least squares residuals are not correlated with the fitted values.

(f) T F All else being equal, a prediction interval is wider if the standard error for $b_0$ is larger.

(g) T F Uncertainty about the regression coefficients depends upon the variance of the residuals.

(h) T F The $R^2$ for a regression of $Y$ onto $X$ is the same as $R^2$ for the regression of $X$ onto $Y$.

(i) T F It is possible to reject a null hypothesis when the null hypothesis is true.

(j) T F Our linear regression model implies that the marginal distribution for $Y$ is normal.

(k) T F Assuming our multiple linear regression model, each least squares coefficient $b_j$ has mean $\beta_j$.

(l) T F In simple linear regression, the slope of the regression line is equal to the correlation between $X$ and $Y$.

(m) T F Least squares estimates of the coefficients $\{b_0,b_1,...\}$ are chosen to maximize $R^2$.  

4
4 Residual Diagnostics

For each of the following plots, separately, use the space to the right to describe (in one sentence) what you think is problematic with each (potential) linear fit. Each plot below considers a different data set. Then, suggest how you would fix the problem (make only one suggestion in each case).

(a) 

(b) 

(c)
5 Multiple Linear Regression 1: Electricity Demand

For an energy company in Alabama, we have the daily total electricity demand (measured in Megawatts) and daily temperature (temp in degrees Fahrenheit above 32) for 364 days, with the day of the week (Sunday, Monday, ...) stored in weekday. Our goal is to predict electricity demand, so that the energy company can operate efficiently.

(a) Running a simple linear regression of Megawatts on temp yields the three plots shown below. Use these plots to answer the questions.

(i) What do you conclude from plots 1 and 2? Justify the answer by referring to each plot.

(ii) Plot 3 indicates a problem with this regression that can be fixed by adding another variable to the model. What variable should be added and why?
(b) Consider a regression of MegaWatts on the categorical weekday. Below are the output results from two different versions of this regression. In the first, Sunday is the baseline category, while in the second, Monday is the baseline. Answer the questions below the output.

### Regression 1

**Call:**
```
lm(formula = MegaWatts ~ weekday)
```

**Coefficients:**

|                | Estimate | Std. Err. | t value | Pr(>|t|) |
|----------------|----------|-----------|---------|----------|
| (Intercept)    | 3162.0   | 62.0      | 51.0    | <2e-16   *** |
| weekday2_Mon   | 288.0    | 87.0      | 3.0     | 0.001 **  |
| weekday3_Tue   | 375.0    | 87.0      | 4.0     | 2e-05 *** |
| weekday4_Wed   | 345.0    | 87.0      | 4.0     | 9e-05 *** |
| weekday5_Thu   | 263.0    | 87.0      | 3.0     | 0.003 **  |
| weekday6_Fri   | 278.0    | 87.0      | 3.0     | 0.002 **  |
| weekday7_Sat   | 174.0    | 87.0      | 2.0     | 0.046 *   |

---

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

Residual standard error: 400 on 357 degrees of freedom
Multiple R-squared: 0.07, Adjusted R-squared: 0.05
F-statistic: 4 on 6 and 357 DF, p-value: 4e-04

### Regression 2

**Call:**
```
lm(formula = MegaWatts ~ weekday)
```

**Coefficients:**

|                | Estimate | Std. Err. | t value | Pr(>|t|) |
|----------------|----------|-----------|---------|----------|
| (Intercept)    | 3451.0   | 62.0      | 56.1    | <2e-16   *** |
| weekday2_Tue   | 87.0     | 87.0      | 1.0     | 0.318    |
| weekday3_Wed   | 56.0     | 87.0      | 0.6     | 0.519    |
| weekday4_Thu   | -26.0    | 87.0      | -0.3    | 0.767    |
| weekday5_Fri   | -10.0    | 87.0      | -0.1    | 0.908    |
| weekday6_Sat   | -114.0   | 87.0      | -1.3    | 0.189    |
| weekday7_Sun   | -288.0   | 87.0      | -3.3    | 0.001 **  |

---

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

Residual standard error: 400 on 357 degrees of freedom
Multiple R-squared: 0.07, Adjusted R-squared: 0.05
F-statistic: 4 on 6 and 357 DF, p-value: 4e-04

(i) Which day of the week on average has the highest electricity demand? The lowest? Justify your answer numerically.

(ii) Discussing both of the regression outputs, what do you learn from the t tests and their associated p-values? Be specific and justify your answer numerically.

(iii) Using both of the about regression outputs, what do you learn from the F test? Conceptually, why do the two regressions have the same F test?
(c) Using the output from the model below, which includes temp and weekday, what is the predicted total electricity demand (measured in MegaWatts) for a Wednesday where the temperature is 52 degrees Fahrenheit? If the standard error of the fitted value is $s_{fit} = 300$ MegaWatts, what is a 95% predictive interval for your answer?

Call:
```
lm(formula = MegaWatts ~ weekday * temp)
```

Coefficients:

|                   | Estimate | Std. Error | t value | Pr(>|t|) |
|-------------------|----------|------------|---------|----------|
| (Intercept)       | 2632     | 243        | 11      | <2e-16 *** |
| weekday2_Tue      | 495      | 345        | 1       | 0.153    |
| weekday3_Wed      | -476     | 381        | -1      | 0.212    |
| weekday4_Thu      | -1352    | 416        | -3      | 0.001 ** |
| weekday5_Fri      | -1012    | 401        | -2      | 0.012 *  |
| weekday6_Sat      | -1001    | 398        | -2      | 0.012 *  |
| weekday7_Sun      | -1330    | 386        | -3      | 6e-04 ***|
| temp              | 19       | 5          | 3       | 7e-04 ***|
| weekday2_Tue:temp | -9       | 8          | -1      | 0.225    |
| weekday3_Wed:temp | 12       | 8          | 1       | 0.173    |
| weekday4_Thu:temp | 29       | 9          | 3       | 0.002 ** |
| weekday5_Fri:temp | 22       | 9          | 2       | 0.014 *  |
| weekday6_Sat:temp | 19       | 9          | 2       | 0.030 *  |
| weekday7_Sun:temp | 23       | 9          | 3       | 0.007 ** |

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

Residual standard error: 400 on 350 degrees of freedom
Multiple R-squared: 0.4, Adjusted R-squared: 0.4
F-statistic: 2e+01 on 13 and 350 DF, p-value: <2e-16

8
6  Multiple Linear Regression 2: Predicting Wages

This problem examines predicting wages based on observed characteristics. The data consists of 550 employed individuals in 1978 and has the following variables:

- \( \text{wage} \): Hourly wage in 1978
- \( \text{educ} \): Years of education completed
- \( \text{exper} \): Years of labor market experience
- \( \text{female} \): 1 if female, 0 if male
- \( \text{kids} \): Number of dependent children.

(a) Comment on the relationship between \( \text{wage} \) and the other four variables using the boxplots below. Note any patterns as well as concerns.
Consider a multiple regression using $\text{educ}$, $\text{exper}$, $\text{female}$, and $\text{kids}$ to predict hourly wage. Below are three pairs of plots (labeled “1”, “2”, and “3”), where in the top plots $\text{wage}$ is used as the output and in the bottom plots $\log(\text{wage})$ is used instead. Use the plots to answer the questions below.

(b) For each pair describe what is being shown and how they are useful as a diagnostic tool.

(i)

(ii)

(iii)

(c) As a whole, what do you conclude based on these plots? What has changed? Do you prefer to use $\text{wage}$ or $\log(\text{wage})$, and why?
Consider the following summary.

Call:
\texttt{lm(formula = log(wage) ~ educ + exper + female + kids)}

Residuals:
\begin{center}
\begin{tabular}{rrrrr}
 Min & 1Q & Median & 3Q & Max \\
-2.39318 & -0.25112 & 0.02287 & 0.24630 & 1.32295 \\
\end{tabular}
\end{center}

Coefficients:
\begin{center}
\begin{tabular}{lrrrrr}
 Estimate & Std. Error & t value & Pr(>|t|) \\
 (Intercept) & 0.590007 & 0.099703 & 5.918 & 5.78e-09 *** \\
 educ & 0.077766 & 0.006601 & 11.781 & < 2e-16 *** \\
 exper & 0.013355 & 0.001378 & 9.694 & < 2e-16 *** \\
 female & -0.337317 & 0.035669 & -9.457 & < 2e-16 *** \\
 kids & -0.007021 & 0.013411 & -0.523 & 0.601 \\
\end{tabular}
\end{center}

---

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

Residual standard error: 0.4007 on 545 degrees of freedom
Multiple R-squared: 0.3367,  Adjusted R-squared: 0.3318
F-statistic: 69.15 on 4 and 545 DF,  p-value: < 2.2e-16

(d) Give a precise, numerical interpretation of what the above output says about the association between \textit{educ} and \textit{wage}.

(e) Using the summary above, give a 95% confidence interval for the coefficient of \textit{kids}. Interpret your answer, referring to part (a).
(f) Define $\text{exper.squared} = (\text{exper})^2$. Using the summary below, should \text{exper.squared} be included in the model? Why or why not? Interpret your answer, referring to part (a).

Call:
\[
\text{lm(formula = log(wage) ~ educ + exper + exper.squared + female + kids)}
\]

Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-2.42121</td>
<td>-0.23810</td>
<td>0.01701</td>
<td>0.24124</td>
<td>1.37470</td>
</tr>
</tbody>
</table>

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|---------|
| (Intercept)    | 0.5615258 | 0.0988398  | 5.681   | 2.18e-08 *** |
| educ           | 0.0719948 | 0.0067054  | 10.737  | < 2e-16 *** |
| exper          | 0.0318576 | 0.0051465  | 6.190   | 1.18e-09 *** |
| exper.squared  | -0.0004211 | 0.0001130  | -3.728  | < 2e-08 *** |
| female         | -0.3419768 | 0.0352764  | -9.694  | < 2e-16 *** |
| kids           | -0.0285579 | 0.0144596  | -1.975  | 0.048772 * |

---

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

Residual standard error: 0.396 on 544 degrees of freedom
Multiple R-squared: 0.3532, Adjusted R-squared: 0.3472
F-statistic: 59.41 on 5 and 544 DF, p-value: < 2.2e-16
7 Regression: Baseball Data

For each Major League Baseball team we have the number of wins ($\text{Wins}$) and the total player salary in millions of dollars ($\text{Salary}$) for 2006. (You don’t need to know anything about baseball for this question.) The total league payroll was $2,326.707$ million. For each team $i$, define

$$\text{SalaryShare}_i = \frac{\text{Salary}_i}{\sum_{j=1}^{n} \text{Salary}_j} = \frac{\text{Salary}_i}{2,326.707}.$$

(a) We use a linear regression to predict $\text{Wins}$ using $\text{SalaryShare}$ and obtain the three plots below. Consider the data point labeled “NYY”. Do you think this observation should be removed? Why or Why not? If it was removed, what would happen to the intercept and slope estimates (discuss the direction of any changes and how large you expect the changes to be)?
Now consider the following summary.

Call:
\texttt{lm(formula = Wins ~ SalaryShare)}

Residuals:

\begin{tabular}{lcccc}
Min & 1Q & Median & 3Q & Max \\
-17.7907 & -4.5503 & 0.3654 & 4.5352 & 17.4042 \\
\end{tabular}

Coefficients:

\begin{tabular}{lcccccc}
Estimate & Std. Error & t value & Pr(>\mid t\mid) \\
(Intercept) & 67.982 & 4.178 & 16.271 & 8.4e-16 & *** \\
SalaryShare & 389.540 & 116.013 & 3.358 & 0.00228 & ** \\
--- \\
Signif. codes: & 0 *** & 0.001 ** & 0.01 * & 0.05 . & 0.1 1 \\
\end{tabular}

Residual standard error: 8.665 on 28 degrees of freedom
Multiple R-squared: 0.2871, Adjusted R-squared: 0.2616
F-statistic: 11.27 on 1 and 28 DF, p-value: 0.002277

(b) But suppose that instead of regressing \texttt{Wins} on \texttt{SalaryShare} we used \texttt{Salary} itself as the input. Use the summary above to compute the estimates of the intercept $b_0$, the slope $b_1$, and the $R^2$ value for this hypothetical regression.

(c) Do we have reason to believe in a linear relationship between \texttt{Wins} and \texttt{Salary}, in the hypothetical regression in part (b)? State a formal hypothesis test, the value of the test statistic, and the conclusion.
(d) In 2006 the Chicago White Sox payroll was $102.75 million and won 90 games. What is the predicted number in wins if they added $10 million to their payroll? If the standard error of the fitted value is $s_{\hat{y}} = 2$, what is a 95% predictive interval for your answer?