Chicago Booth BUS 41100
Midterm **SAMPLE #2**

Instructor: Max H. Farrell

This exam is designed to be **50% longer** than your midterm will be.

Name: _____ **Section (circle):** 01 – Morning / 02 – Afternoon / 81 – Evening

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

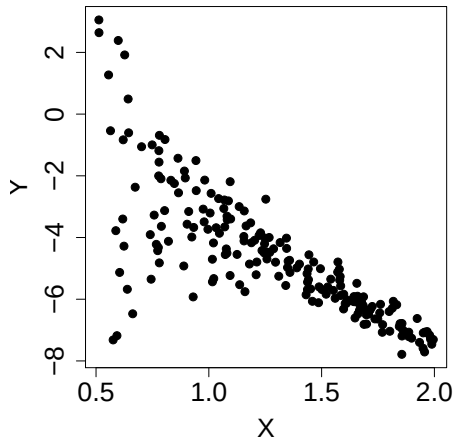
Signed: _____

- 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 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 α)
(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(`wlfp72`) and log(`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 β_1 is centered at β_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 Least squares residuals are not correlated with the fitted values.

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

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

(g) 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.

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

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

(j) T F Assuming our multiple linear regression model, each least squares coefficient b_j has mean β_j.

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

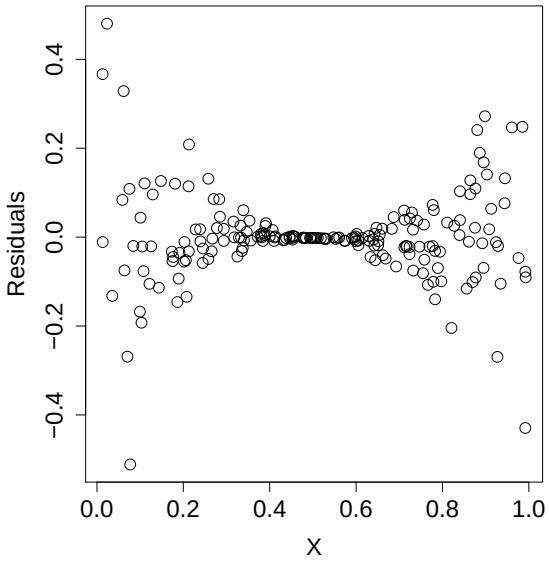
(l) T F Least squares estimates of the coefficients $\{b_0,b_1,...,\}$ are chosen to maximize R^2.

4 Analyzing Plots

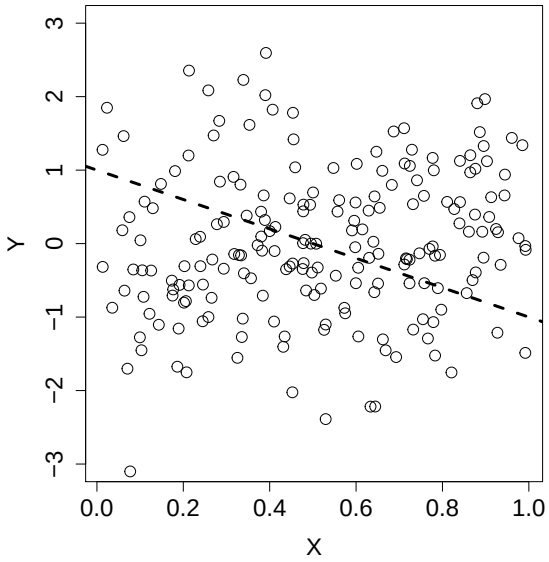
The questions in this section are based on analyzing a single plot related to some regression model. The models and the data are different for each part.

Use the space to the right of each plot to answer the question.

(a) Based on this plot, are any simple linear regression assumptions violated? If so, propose a single fix per violation.



(b) Assume that $Y = 1 - 2X + \varepsilon$ is true. Based on this information and the plot below, which assumption or assumptions of the simple linear regression model are violated? Be as specific as you can.



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) 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         62      51   <2e-16 ***
weekday2_Mon      288         87       3    0.001 **
weekday3_Tue      375         87       4    2e-05 ***
weekday4_Wed      345         87       4    9e-05 ***
weekday5_Thu      263         87       3    0.003 **
weekday6_Fri      278         87       3    0.002 **
weekday7_Sat      174         87       2    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         62    56.1   <2e-16 ***
weekday2_Tue       87         87     1.0    0.318
weekday3_Wed       56         87     0.6    0.519
weekday4_Thu      -26         87    -0.3    0.767
weekday5_Fri      -10         87    -0.1    0.908
weekday6_Sat     -114         87    -1.3    0.189
weekday7_Sun     -288         87    -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?

(b) 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 $s_{\text{fit}} = 300$ MegaWatts, what is a 95% predictive interval for your answer?

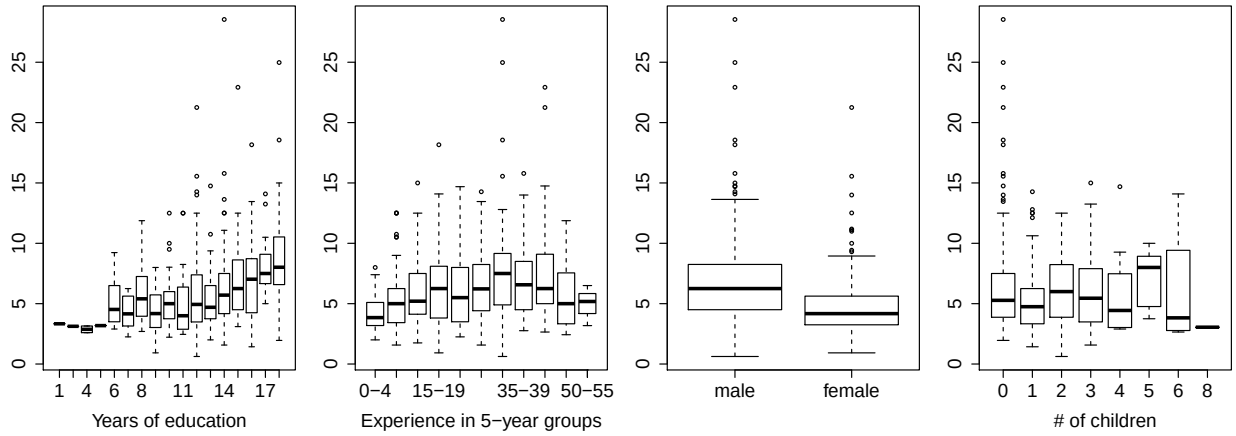
```
Call:
lm(formula = MegaWatts ~ weekday * temp)

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
```

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:

`wage` = Hourly wage in 1978
`educ` = Years of education completed
`exper` = years of labor market experience
`female` = 1 if female, 0 if male
`kids` = number of dependent children.

(a) Comment on the relationship between `wage` and the other four variables using the boxplots below. Note any patterns as well as concerns.



Consider the following `summary`.

```
Call:
lm(formula = log(wage) ~ educ + exper + female + kids)

Residuals:
     Min       1Q   Median       3Q      Max
-2.39318 -0.25112  0.02287  0.24630  1.32295

Coefficients:
             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
---
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
```

(b) Give a precise, numerical interpretation of what the above output says about the association between `educ` and `wage`.

(c) Using the summary above, give a 95% confidence interval for the coefficient of `kids`. Interpret your answer, referring to part **(a)**.

(d) Define `exper.squared` = (`exper`)^2. Using the summary below, should `exper.squared` be included in the model? Why or why not? Interpret your answer, referring to part **(a)**.

```
Call:
lm(formula = log(wage) ~ educ + exper + exper.squared + female +
    kids)

Residuals:
     Min       1Q   Median       3Q      Max
-2.42121 -0.23810  0.01701  0.24124  1.37470

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 0.000213 ***
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 (`Wins`) and the total player salary in millions of dollars (`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

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

Now consider the following `summary`.

```
Call:
lm(formula = Wins ~ SalaryShare)

Residuals:
     Min       1Q   Median       3Q      Max
-17.7907  -4.5503   0.3654   4.5352  17.4042

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

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
```

(a) But suppose that instead of regressing `Wins` on `SalaryShare` we used `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.

(b) Do we have reason to believe in a linear relationship between `Wins` and `Salary`, in the hypothetical regression in part **(b)**? State a formal hypothesis test, the value of the test statistic, and the conclusion.

(c) 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 $s_{\text{fit}} = 2$, what is a 95% predictive interval for your answer?