Chicago Booth BUS 41100
Midterm **SAMPLE #1**

Instructor: Max H. Farrell

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

Name: ______________________________________ **Section (circle):** $\begin{cases} \text{01 – Morning} \\ \text{02 – Afternoon} \\ \text{81 – Evening} \end{cases}$

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

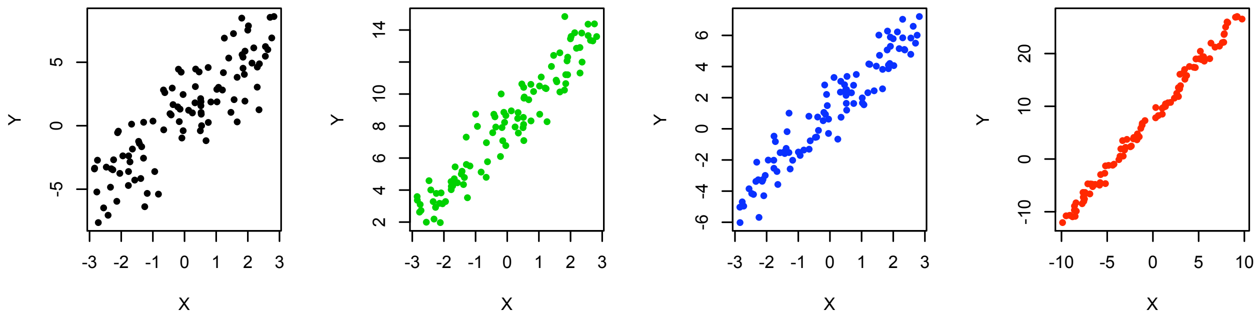
Signed: ______________________________________

- You have 3 hours to complete the exam.
- This exam has 14 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.



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

(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

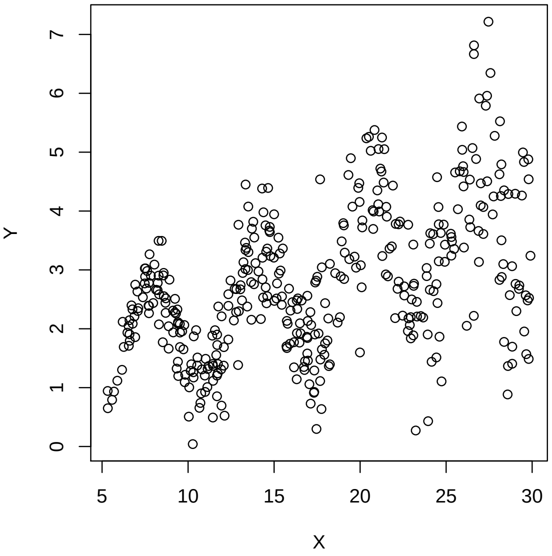
(ii) 0

(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 **(e)** 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:
     Min       1Q   Median       3Q      Max
-0.32706 -0.02290  0.00202  0.02220  0.18405

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 $\alpha = .05$?

(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 $\alpha = .05$?

3 The Regression Model

Assume the following simple linear regression model:

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

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

(b) Suppose $X = 3$. For what value of β_1 will the marginal variance of Y equal the conditional variance of $Y|X$? How does your answer change when $X = 4$?

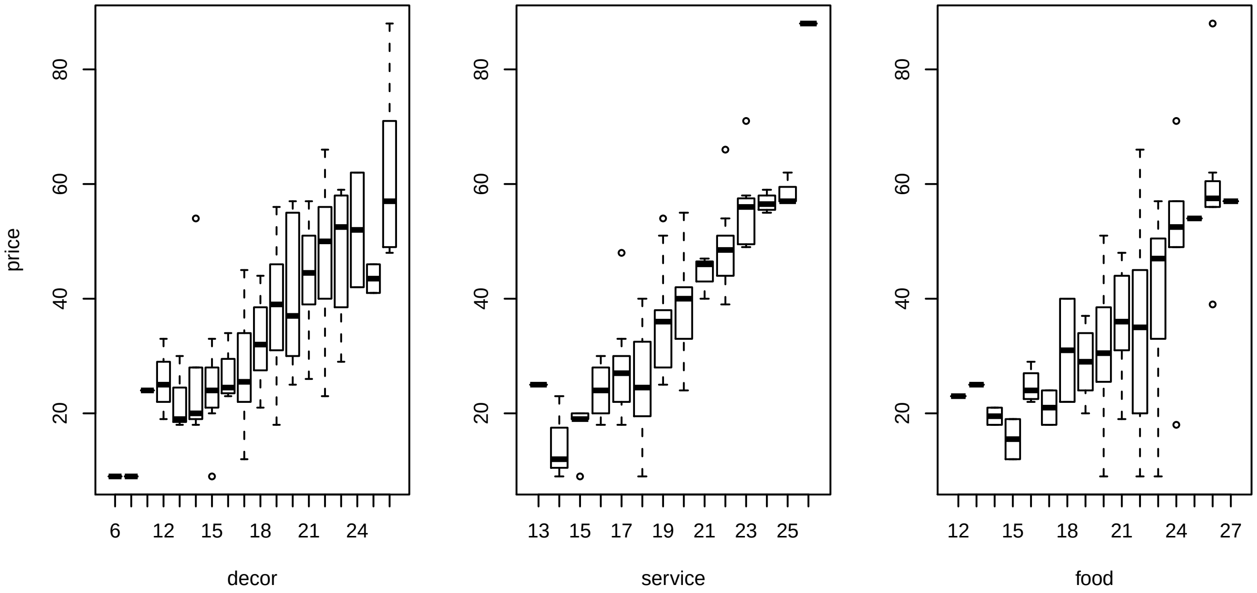
(c) If $\beta_1 = 2$, around what value would you expect for the R^2 from such a regression?

(d) Consider instead the log-log model $\log(Y) = 6 + \beta_1 \log(X) + \varepsilon$. 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:
    Min      1Q  Median      3Q     Max
-90.273 -37.383  -8.498  21.515 120.086

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?

(c) 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 $\mathrm{sd}(\hat{R}_f) \equiv s_{\text{fit}}(1440) = 10$, what is a 95% prediction interval for runs of a 1440 hit team?

(d) 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?

(f) 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) -1125.0466   315.2035  -3.569  0.00137 **
H               1.0491     0.1464   7.164 1.05e-07 ***
SO              0.3590     0.1175   3.054  0.00503 **
---

Residual standard error: 46.69 on 27 degrees of freedom
Multiple R-squared: 0.6742,Adjusted R-squared: 0.6501
F-statistic: 27.94 on 2 and 27 DF,  p-value: 2.656e-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 Panel and Clustered Data

We have fuel economy data for 82 different car `model`s from `year` 2001–2010. We want to study the relationship between engine size (measured by `displ`acement) and miles per gallon in the city (`cty`). We also observe each car's `make` (the company making it, e.g. Honda is a make, Accord is a model).

Consider the following regression model, where i indexes models and t indexes years:

$$\texttt{cty}_{i,t} = \alpha_{0,i} + \gamma_{0,t} + \beta_1 \texttt{displ}_{i,t} + \varepsilon_{i,t}.$$

(a) What do the terms $\alpha_{0,i}$ and $\gamma_{0,t}$ represent and what do they control for? That is, explain conceptually why it is important that each is in the model.

(b) Explain the difference between *fixed* effects and *random* effects.

(c) Why is it impossible to add a term for `make` to the model above? That is, explain what is wrong with assuming

$$\texttt{cty}_{i,t} = \alpha_{0,i} + \gamma_{0,t} + \beta_1 \texttt{displ}_{i,t} + \varepsilon_{i,t}.$$

(d) Recall the model is $\texttt{cty}_{i,t} = \alpha_{0,i} + \gamma_{0,t} + \beta_1 \texttt{displ}_{i,t} + \varepsilon_{i,t}$. In R we obtain the following output:

```
Call:
plm(formula = cty ~ displ, data = fuel.panel, model = "within",
    index = c("model", "year"))

Balanced Panel: n = 82, T = 10, N = 820

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-4.510962 -0.570767 -0.026918  0.440195  9.030572

Coefficients:
      Estimate Std. Error t-value Pr(>|t|)
displ -0.36541    0.17381 -2.1024  0.03586
```

Using the notation from the model, list and explain **all** the assumptions we need to make for this output to be valid/useful.

(e) We now use each car's `make` in the following R code, where `mpg.model` refers to the regression in part **(d)**.

```
> vcov <- cluster.vcov(mpg.model, fuel.panel$make)
> coeftest(mpg.model, vcov)
  Estimate Std. Error    t value   Pr(>|t|)
    -0.365      0.353     -1.035      0.301
```

Explain how, and *why* we are using `make` by referring to the model notation above. Compare the results numerically to those in part **(d)**, and explain any differences.