Chicago Booth BUS 41100
Midterm **SAMPLE #3**

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:

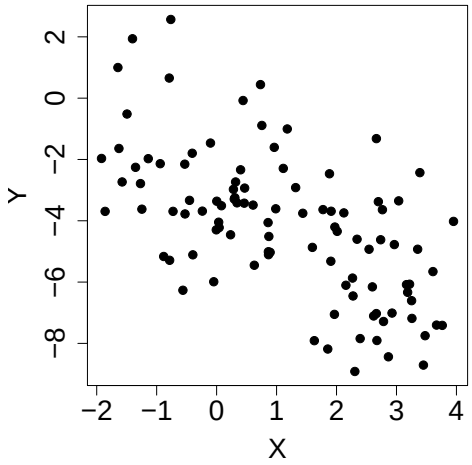
Signed: _____

- 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) Which of the following best describes the least-squares line fit to the data shown in the plot?



(i) $b_0 = 0$, $b_1 = -1$

(ii) $b_0 = -3$, $b_1 = 1$

(iii) $b_0 = -5$, $b_1 = 2$

(iv) $b_0 = -3$, $b_1 = -1$

(v) $b_0 = 0$, $b_1 = -3$

(b) Suppose you estimate a simple linear regression model and the slope coefficient has a `t-value` equal to `-3.1`. Based on this, which of the following statements are WRONG? (Circle all that apply.)

(i) A 95% confidence interval for the true slope would exclude 0.

(ii) It is *possible* that the point estimate for the slope is $b_1 = 4$.

(iii) At the 10% level you fail to reject the null hypothesis that the true slope is equal to 0.

(iv) The probability that the true slope is negative is greater than the probability that the true slope is positive.

(c) Suppose we form a 95% confidence interval based on a simple linear regression. Explain in words what is meant by "95%".

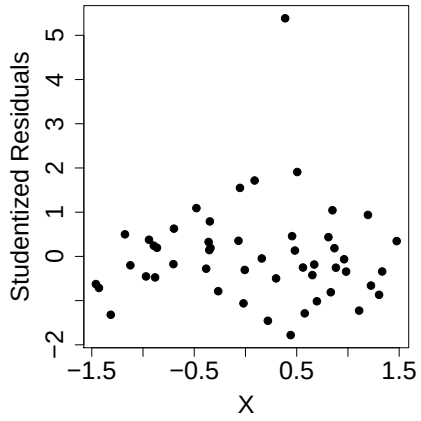
(d) Suppose that in the population $Y = 2 - X - 3X^3 + \log(X) + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 5X^2)$, so the true conditional expectation is $\mathbb{E}[Y|X] = 2 - X - 3X^3 + \log(X)$. In your sample, you run a simple linear regression of Y on X and obtain $\hat{Y}_i = b_0 + b_1 X_i$ and residuals e_i for each point. Then you will have $\texttt{corr}(X, e) \neq 0$ because the linear regression model is incorrect. TRUE or FALSE? Justify your answer.

(e) We have data on three variables: Y, X_1, and X_2. Then we create a new variable by adding together the two X variables: $X_3 = X_1 + X_2$. We now regress Y on X_1, X_2, and X_3. Explain why we can't find the regression coefficient on X_3 even though $\texttt{corr}(X_1, X_3)$ and $\texttt{corr}(X_2, X_3)$ are not 1.

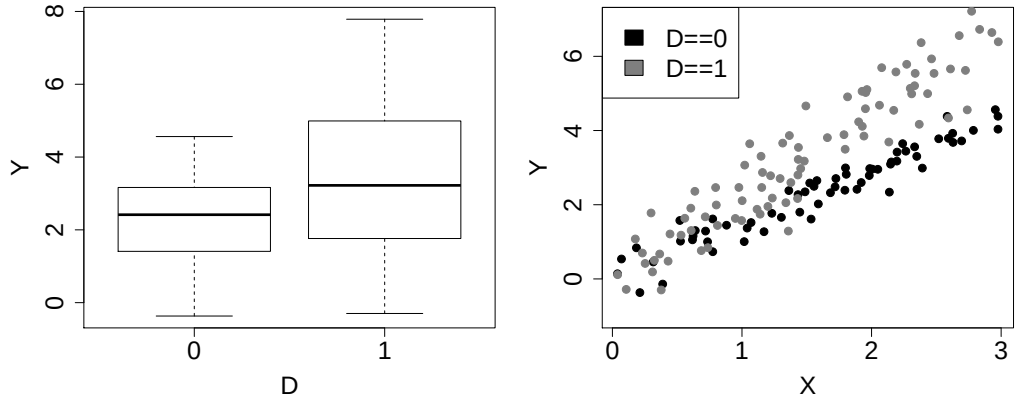
(f) We have data on two variables: Y and X. A positive correlation between X and the residuals from a simple linear fit would indicate that the slope is too small relative to the least squares regression line. TRUE or FALSE? Justify your answer.

2 Plots & Assumptions

(a) For the plot below, use the space to the right to describe what you think is problematic. Which of our standard linear regression assumptions are violated, if any? If there is a potential problem, what would your next step(s) be?



(b) Plotted below are data on 3 variables: a continuous Y and X and a binary variable $D \in \{0, 1\}$.



(i) Based on these two plots, what model specification would you recommend where Y is the outcome and X and D are the independent variables?

(ii) Based on these two plots, which, if any, of our standard linear regression assumptions are violated?

3 Simple Linear Regression

The Capital Asset Pricing Model relates the returns of an asset, given by R_A, to the returns of the market, R_M through a linear regression equation

$$R_A = \alpha + \beta R_M + \varepsilon.$$

Assume that all our standard simple linear regression assumptions are met. We estimate the regression and obtain

$$R_A = 0.03 - 0.6 R_M + e,$$

where the sample mean of $R_M = 0.5$, the standard error of the intercept is 0.01, the standard error of the slope is 0.25, the sample variance of $R_M = 1$, and the residual standard error is 0.02.

(a) Is the slope coefficient statistically significant at the 0.05 level?

(b) Compute the correlation between R_M and e.

(c) What is the average return of the asset in this sample?

(d) Estimate the variance of R_A.

(e) Estimate the R^2 for this regression.

4 Multiple Linear Regression

Researchers at General Motors collected data on 60 U.S. Standard Metropolitan Statistical Areas (SMSA's) in a study of whether or not air pollution contributes to mortality. The dependent variable for analysis is age adjusted `Mortality` rate and the explanatory variable is `HCPot`, the HydroCarbons pollution potential index.

(a) Consider first a linear regression of `log(Mortality)` on `log(HCPot)`, given by the `summary` below.

```
Call:
lm(formula = log(Mortality) ~ log(HCPot))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.8222     0.0219   311.2   <2e-16 ***
log(HCPot)    0.0079     0.0073     1.1      0.3

Residual standard error: 0.066 on 58 degrees of freedom
Multiple R-squared:  0.02,      Adjusted R-squared:  0.0029
F-statistic: 1.2 on 1 and 58 DF,  p-value: 0.28
```

(i) Provide a complete description of the linear regression model and assumptions that are implicitly used by this output.

(ii) Explain in detail what the parameter estimated by the coefficient on `log(HCPot)` represents. Then, in that context, interpret the estimate numerically.

(b) We now consider the second pollution measure, `NOxPot`. Consider the following `summary`.

```
Call:
lm(formula = log(Mortality) ~ log(NOxPot))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.8081     0.0178   382.1   <2e-16 ***
log(NOxPot)   0.0156     0.0069     2.3     0.03 *

Residual standard error: 0.064 on 58 degrees of freedom
Multiple R-squared:  0.082,     Adjusted R-squared:  0.066
F-statistic: 5.2 on 1 and 58 DF,  p-value: 0.026
```

Compare to the `summary` from **part (a)**. Which pollution potential is a better predictor of `Mortality` and why? Can you tell by comparing the two regression outputs?

(c) Consider now using both pollution measures in the same regression model.

```
Call:
lm(formula = log(Mortality) ~ log(HCPot) + log(NOxPot))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.850      0.021   328.1   <2e-16 ***
log(HCPot)    -0.064      0.019    -3.3    0.002 **
log(NOxPot)    0.074      0.019     3.9    2e-04 ***

Residual standard error: 0.059 on 57 degrees of freedom
Multiple R-squared:  0.23,      Adjusted R-squared:  0.2
F-statistic: 8.5 on 2 and 57 DF,  p-value: 6e-04
```

(i) Explain in detail what the parameter estimated by the coefficient on `log(HCPot)` represents. Explicitly compare to **part (a)(ii)**. Then interpret the estimate numerically, again comparing to **(a)(ii)**.

(ii) Explain the change in the standard errors for the coefficients on `log(HCPot)` (comparing to **part (a)**) and `log(NOxPot)` (comparing to **part (b)**).

(d) Let us now consider demographic information, beginning with `HighSchool`, for which a simple linear regression gives the `summary` below.

```
Call:
lm(formula = log(Mortality) ~ HighSchool)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.8602     0.0084   815.5   <2e-16 ***
HighSchoolTRUE  -0.0806     0.0188    -4.3    7e-05 ***

Residual standard error: 0.058 on 58 degrees of freedom
Multiple R-squared:  0.24,      Adjusted R-squared:  0.23
F-statistic:  18 on 1 and 58 DF,  p-value: 6.9e-05
```

(i) Provide a complete description of the linear regression model and assumptions that are implicitly used by this output.

(ii) Explain in detail what the parameter estimated by the coefficient on `HighSchool` represents. Then, in that context, interpret the estimate numerically.

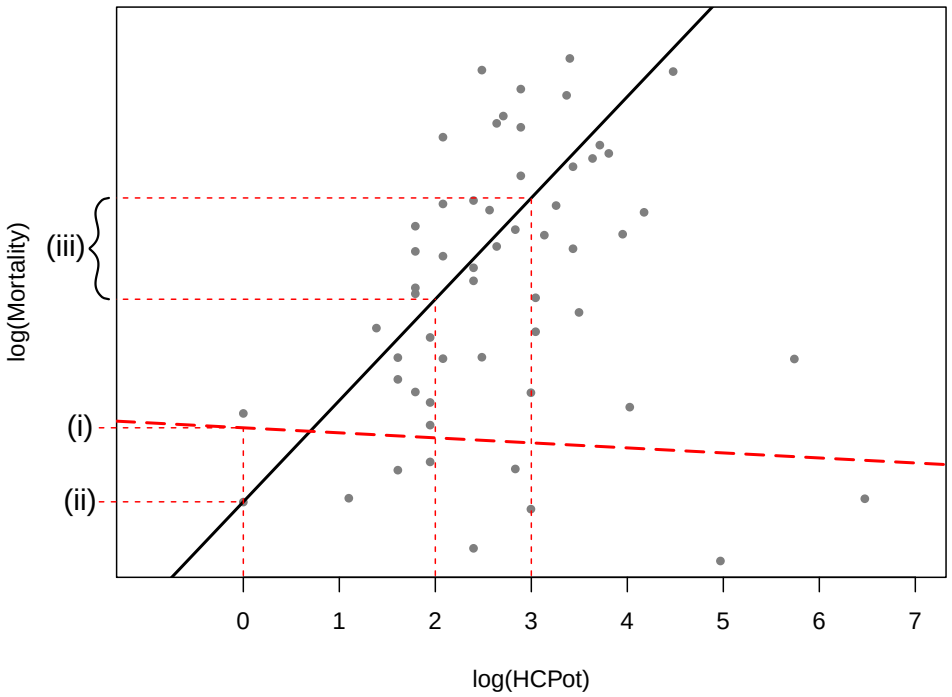
(e) Consider the `summary` below.

```
Call:
lm(formula = log(Mortality) ~ HighSchool * log(HCPot))

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                  6.757      0.020   339.7   <2e-16 ***
HighSchoolTRUE               0.030      0.043     0.7      0.5
log(HCPot)                   0.041      0.007     5.5    9e-07 ***
HighSchoolTRUE:log(HCPot)   -0.043      0.012    -3.6    7e-04 ***

Residual standard error: 0.05 on 56 degrees of freedom
Multiple R-squared:  0.5,       Adjusted R-squared:  0.5
F-statistic: 2e+01 on 3 and 56 DF,  p-value: 1e-08
```

(i) Provide a complete description of the linear regression model and assumptions that are implicitly used by this output.

(ii) Explain in detail what the parameter estimated by the coefficient on `HighSchoolTRUE:log(HCPot)` represents. Then, in that context, interpret the estimate numerically.

(f) Consider the plot below which is based on the regression reported in **part (e)**.



Explain briefly what is shown in this plot. Describe both what information the plot represents and what you conclude from it.

Using the `summary` output from **part (e)** above, compute the numeric values for the points on the Y-axis labeled on the graph as:

(i) = ?

(ii) = ?

(iii) (the difference) = ?

(g) Now consider adding `PopDensity` to the regression in **part (e)**, which yields this `summary`.

```
Call:
lm(formula = log(Mortality) ~ HighSchool * log(HCPot) + log(PopDensity))

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                  6.664      0.153    43.5   <2e-16 ***
HighSchoolTRUE               0.025      0.044     0.6    0.574
log(HCPot)                   0.039      0.009     4.5    4e-05 ***
log(PopDensity)              0.012      0.020     0.6    0.542
HighSchoolTRUE:log(HCPot)   -0.041      0.013    -3.2    0.002 **

Residual standard error: 0.05 on 55 degrees of freedom
Multiple R-squared:  0.5,       Adjusted R-squared:  0.5
F-statistic: 1e+01 on 4 and 55 DF,  p-value: 4e-08
```

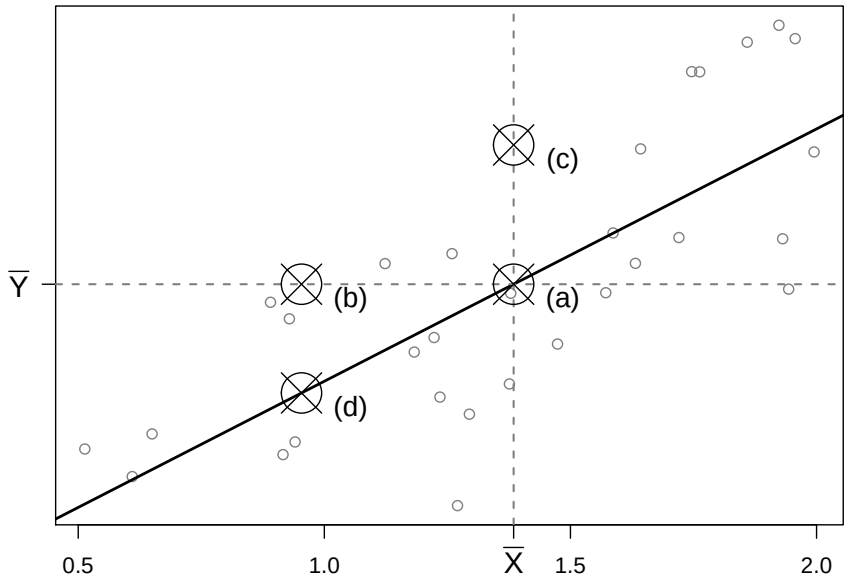
Based on this output, should we add `log(PopDensity)` to the regression? Why or why not? Cite specific numerical values from the output to support your argument.

5 Understanding Simple Linear Regression

This question is about how parameter estimates change as the data changes, but the underlying model is fixed. Throughout, assume that $Y = \beta_0 + \beta_1 X + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. The parameters β_0, β_1, and σ^2 are fixed but our estimates of them change as the data changes. We start with an i.i.d. sample of $n = 30$ points and obtain the following `summary`:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.81       0.57     3.2    0.004 **
X               2.59       0.39     6.6    4e-07 ***

Residual standard error: 0.91 on 28 degrees of freedom
Multiple R-squared:  0.61,      Adjusted R-squared:  0.6
F-statistic:  44 on 1 and 28 DF,  p-value: 3.6e-07
```

The 30 data points (grey circles) and the least squares fit (black line) are plotted below. **One at a time** we add a new data point at one of the four spots labeled **(a)**, **(b)**, **(c)**, and **(d)**. X marks the spot! E.g., point **(a)** is at $(\bar{X}, \bar{Y})$.



Using the plot and the `summary` above, indicate the effect of adding the new data point on each parameter estimate by circling **exactly one** of {up / dn / same / ?} in each cell of the table below. Circle "up" if the estimate will be higher with the new point, "dn" if it will be lower, "same" for unchanged, and "?" if the effect is uncertain given the information.

	New Data Point			
Parameter	**(a)**	**(b)**	**(c)**	**(d)**
b_0	up / dn / same / ?	up / dn / same / ?	up / dn / same / ?	up / dn / same / ?
$\texttt{se}(b_0)$	up / dn / same / ?	up / dn / same / ?	up / dn / same / ?	up / dn / same / ?
b_1	up / dn / same / ?	up / dn / same / ?	up / dn / same / ?	up / dn / same / ?
$\texttt{se}(b_1)$	up / dn / same / ?	up / dn / same / ?	up / dn / same / ?	up / dn / same / ?

Justify/explain your answers for point **(c)**:

b_0 :

$\texttt{se}(b_0)$:

b_1 :

$\texttt{se}(b_1)$:

Consider your answers for points **(a)** and **(d)**. For each parameter estimate:

- If you marked **different** answers, explain why.

  [*e.g. the slope estimate goes up in (a) but down in (d) because ....*]

- If you marked **identical** answers, explain why. In addition, if you marked "up" or "dn", explain which change would be larger, or that the changes would be the same, and why.

  [*e.g. se(b_0) stays the same in both because ....*]

  [*e.g. se(b_0) goes up in both, but increases by more in (a) because ....*]

b_0 :

$\texttt{se}(b_0)$:

b_1 :

$\texttt{se}(b_1)$: