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

Instructor: Max H. Farrell

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

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:

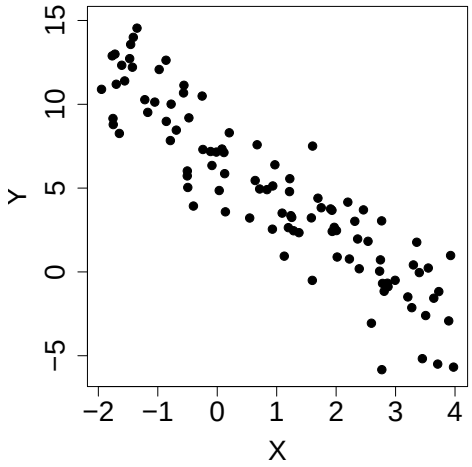
Signed: _____

- You have 3 hours to complete the exam.
- This exam has 16 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? (Circle one only.)



(i) $b_0 = 5$, $b_1 = 3$

(ii) $b_0 = 7$, $b_1 = -2.5$

(iii) $b_0 = 12$, $b_1 = -3$

(iv) $b_0 = 7$, $b_1 = -1$

(v) $b_0 = -5$, $b_1 = 6$

(b) Which of the following statements are FALSE? (Circle all that apply.)

(i) In the simple linear regression model, R^2 is equal to the square of the correlation between Y and $\hat{Y}$ (the fitted values).

(ii) In the multiple regression model, R^2 is equal to the square of the correlation between Y and $\hat{Y}$ (the fitted values).

(iii) In the simple linear regression model, R^2 is equal to the square of the correlation between Y and X.

(iv) In the multiple regression model, R^2 is equal to the square of the correlation between Y and all of the X variables.

(v) None of the statements above are FALSE; all are true.

(c) Which of the following statements are TRUE for logistic regression? (Circle all that apply.)

(i) The residuals are uncorrelated with the fitted values.

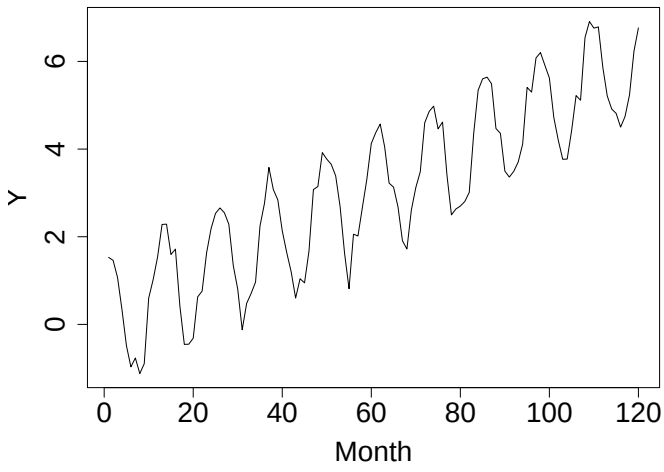
(ii) The odds ratio is always above 1.

(iii) The log odds ratio is always negative.

(iv) Different regressions can be compared using BIC.

(v) All of the above statements are true.

(d) What terms would you put in a time series regression model for the monthly time-series Y shown in the plot? (Circle one only.)



(i) Lagged outcome value(s)

(ii) sin and cos terms

(iii) A time trend

(iv) (i) and (ii)

(v) (ii) and (iii)

(vi) (i), (ii), and (iii)

(e) Which is the correct interpretation for the coefficient estimate for `x1` in the logistic regression below? (Circle all that apply.)

```
Call:
glm(formula = Y ~ x1 + x2, family = "binomial")

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.07       0.22    -0.3      0.7
x1              2.3        0.38     6.2    5e-10 ***
x2             -2.0        0.32    -6.4    2e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 277.18  on 199  degrees of freedom
Residual deviance: 131.32  on 197  degrees of freedom
AIC: 137.3

Number of Fisher Scoring iterations: 6
```

(i) $\log(\mathbb{P}[\texttt{Y} = 1])$ increases 2.3 units for a one unit increase in `x1`.

(ii) $\log\left(\mathbb{P}[\texttt{Y} = 1]/\mathbb{P}[\texttt{Y} = 0]\right)$ increases tenfold for a one unit increase in `x1`.

(iii) The odds of $\texttt{Y} = 1$ increase tenfold for a one unit increase in `x1`.

(iv) `Y` increases 2.3% for a one unit increase in `x1`.

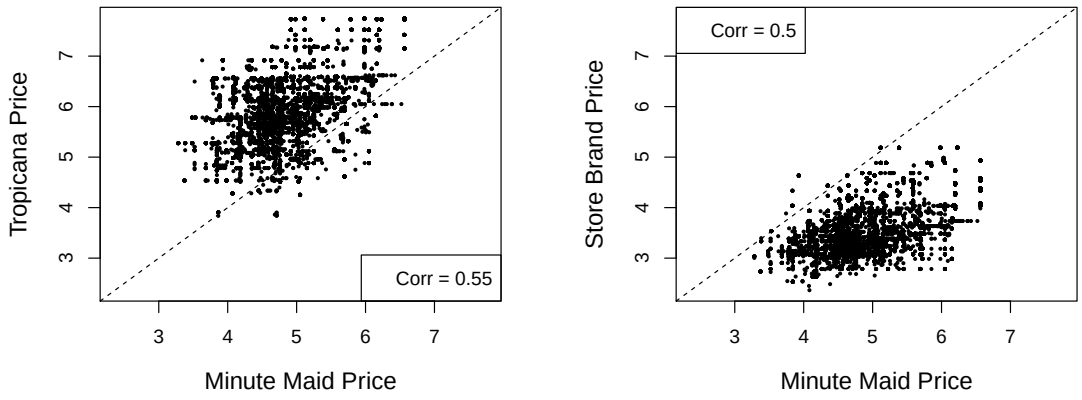
(v) (i) and (iii) are both correct.

(vi) (ii) and (iv) are both correct.

2 Multiple Linear Regression

We have 9649 observations of weekly orange juice prices and sales volume from different locations of a supermarket chain. The data has the sales volume and price for Minute Maid orange juice (respectively `minute.maid.sales` in gallons and `minute.maid.price` in dollars per gallon), as well as the prices per gallon for Tropicana (`tropicana.price`) and the store brand (`store.brand.price`). We also have indicators for whether each brand was featured in the store's marketing for that week (`minute.maid.ad`, `tropicana.ad`, and `store.brand.ad`, all binary indicators).

(a) Using the plots below, comment on the pricing strategy of each brand. What position in the market does each brand occupy? ("market position" can be defined as "the place a brand occupies in consumers' minds relative to competing offerings".)



(b) Describe the regression being run in the (abbreviated) output below. In terms of Minute Maid sales and price, what does the slope coefficient represent? Interpret the numerical value (both the sign and magnitude).

```
Call:
lm(formula = log(minute.maid.sales) ~ log(minute.maid.price))

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)                5.33       0.16      33   <2e-16 ***
log(minute.maid.price)    -2.64       0.05     -54   <2e-16 ***
```

(c) Consider the (abbreviated) regression output below. In terms of Minute Maid sales, what do the coefficients of `log(tropicana.price)` and `log(store.brand.price)` represent? Interpret the numerical values (both the sign and magnitude).

```
Call:
lm(formula = log(minute.maid.sales) ~ log(minute.maid.price) +
    log(tropicana.price) + log(store.brand.price))

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)               9.94       0.17      58   <2e-16 ***
log(minute.maid.price)   -4.15       0.05     -77   <2e-16 ***
log(tropicana.price)      1.60       0.06      28   <2e-16 ***
log(store.brand.price)    1.29       0.05      26   <2e-16 ***
```

(d) Consider now the expanded model shown below.

```
Call:
lm(formula = log(minute.maid.sales) ~ log(minute.maid.price) *
    minute.maid.ad + log(tropicana.price) + log(store.brand.price))

Coefficients:
                                         Estimate Std. Error t value Pr(>|t|)
(Intercept)                                10.73       0.09     122   <2e-16 ***
log(minute.maid.price)                     -3.14       0.06     -53   <2e-16 ***
minute.maid.adTRUE                          1.89       0.13      14   <2e-16 ***
log(tropicana.price)                        0.98       0.05      18   <2e-16 ***
log(store.brand.price)                      1.10       0.04      24   <2e-16 ***
log(minute.maid.price):minute.maid.adTRUE  -0.86       0.08     -10   <2e-16 ***
```

(i) Give a precise interpretation of the coefficient estimate for `minute.maid.adTRUE`. Does this value make sense to you? Why or why not?

(ii) What does the coefficient estimate on `log(minute.maid.price):minute.maid.adTRUE` represent? That is, visually, what would we learn from this estimate if the data were plotted? Does the estimate value make sense to you (both the sign and the magnitude)?

3 Model Building

This problem examines predicting wages based on observed characteristics. The data consists of 550 employed individuals in 1978 and has the following variables: log of hourly wage (`log.wage`), years of labor market experience (`exper` and its square `exper2`), years of education (`educ`), `age`, number of dependent `kids` (coded as 0, 1, 2, or ≥ 3), and binary variables for `sex` (Male or Female), `race` (White or Nonwhite), `married` (Yes or No), and being a `union.member` (Yes or No).

(a) Consider the below sequence of progressively more complicated models.

```
Model 1: log.wage ~ exper + educ + sex
Model 2: log.wage ~ exper + exper2 + educ + sex
Model 3: log.wage ~ exper + exper2 + educ + sex + race + married + age + kids
Model 4: log.wage ~ exper + exper2 + educ + sex + race + married + age + kids + union.member
Model 5: log.wage ~ exper + exper2 + educ + sex + race + married + age + kids + union.member
    + sex * race
Model 6: log.wage ~ exper + exper2 + educ + sex + race + married + age + kids + union.member
    + sex * race + union.member * educ
```

For this sequence of models we have the following results:

```
Analysis of Variance Table

                   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
Model 1 vs. Model 2    545 85.927  1    1.6118 11.1951 0.0008773 ***
Model 2 vs. Model 3    542 84.372  3    1.5549  3.6000 0.0134441 *
Model 3 vs. Model 4    541 79.475  1    4.8974 34.0160  9.43e-09 ***
Model 4 vs. Model 5    540 78.548  1    0.9274  6.4417 0.0114275 *
Model 5 vs. Model 6    539 77.601  1    0.9469  6.5768 0.0106014 *
```

(i) Briefly describe what is being presented in each row of the "Analysis of Variance Table" and what you conclude from the results of each row.

(ii) Briefly explain what is wrong with this approach to model building.

(b) Instead of the above, consider the following R commands and (excerpted) output from automated method to build a model based on BIC.

```
> null <- lm(log.wage ~ 1, data=cps)
> full <- lm(log.wage ~ . + .^2, data=cps)
> fwdBIC.part_b <- step(null, scope=formula(full), direction="forward", k=log(n))
Start:  AIC=-779.02
log.wage ~ 1

Step:  AIC=-843.17
log.wage ~ sex

Step:  AIC=-903.21
log.wage ~ sex + educ

Step:  AIC=-985.57
log.wage ~ sex + educ + age

Step:  AIC=-1011.71
log.wage ~ sex + educ + age + union.member

Step:  AIC=-1013.29
log.wage ~ sex + educ + age + union.member + exper2

Step:  AIC=-1015.49
log.wage ~ sex + educ + age + union.member + exper2 + educ:union.member

Step:  AIC=-1016.83
log.wage ~ sex + educ + age + union.member + exper2 + race +
    educ:union.member
```

(i) Briefly describe the process being used here to build a model.

(ii) Explicitly list the elements of this process which are controlled by the user as opposed to what is controlled by the computer.

(c) Next, consider a slightly altered version of the same process.

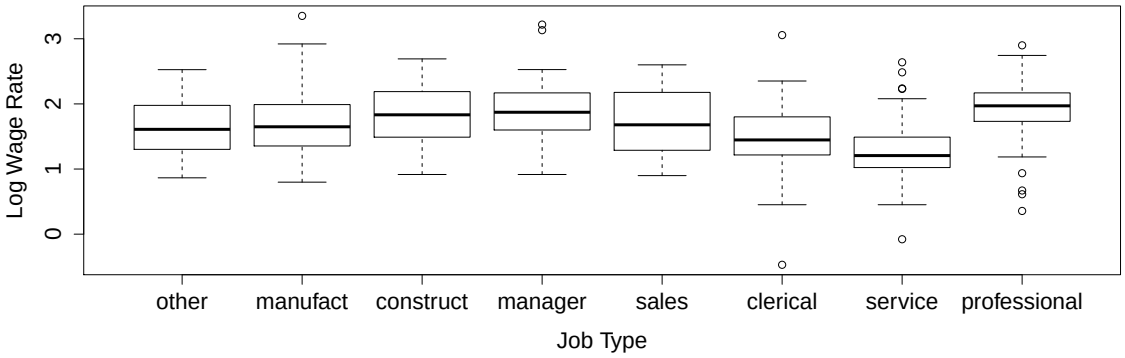
```
> model1 <- lm(log.wage ~ exper + educ + sex)
> full <- lm(log.wage ~ . + .^2, data=cps)
> fwdBIC.part_c <- step(model1, scope=formula(full), direction="forward", k=log(n))
Start:  AIC=-985.57
log.wage ~ exper + educ + sex

Step:  AIC=-1011.71
log.wage ~ exper + educ + sex + union.member

Step:  AIC=-1013.29
log.wage ~ exper + educ + sex + union.member + exper2

Step:  AIC=-1015.49
log.wage ~ exper + educ + sex + union.member + exper2 + educ:union.member

Step:  AIC=-1016.83
log.wage ~ exper + educ + sex + union.member + exper2 + race +
    educ:union.member
```

(i) Explain any differences between this approach and the one in part **(b)**.

(ii) The final selected models contain different variables but the final BIC values are identical (it's not rounding error). How do you explain this?

(iii) Comparing the two models of parts **(b)** and **(c)**, which model do you prefer, and why?

(d) The data also contains indicators for `job.type`, grouped into 8 categories. These categories appear on the x-axis of the plot below.



We also construct a new variable `job.level` that collapses `job.type` into three categors: `LOW` = {`clerical, service`}, `MEDIUM` = {`other, manufact, sales`}, and `HIGH` = {`construct, manager, professional`}.

Our goal is to select one of these two measures of job description to add to the final model selected in part **(e)**. We have the following commands and output.

```
> reg.job.type <- lm(log.wage ~ exper + educ + sex + union.member
+ exper2 + race + educ:union.member + job.type)
> reg.job.level <- lm(log.wage ~ exper + educ + sex + union.member
+ exper2 + race + educ:union.member + job.level)
> c(extractAIC(reg.job.type, k=2)[2], extractAIC(reg.job.level, k=2)[2])
[1] -1079.569 -1076.706
> c(extractAIC(reg.job.type, k=log(n))[2], extractAIC(reg.job.level, k=log(n))[2])
[1] -1014.920 -1033.607
```

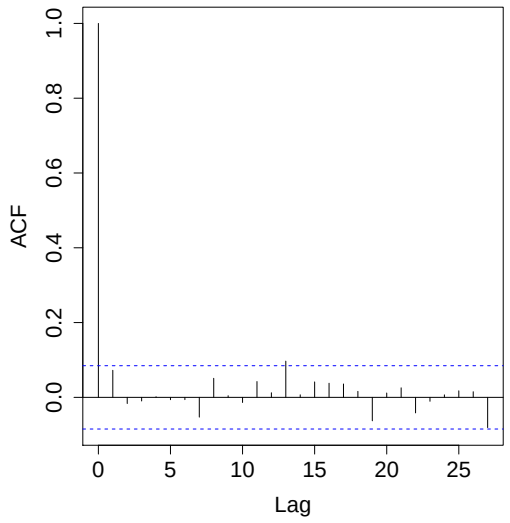
(i) Explain what is being computed in the final two lines and how to evaluate the numerical output.

(ii) Identify the models chosen by each method. Explain these choices, and why they are the same or different.

4 Time Series Regression

This problem studies the time series of returns for the Istanbul Stock Exchange. We have daily returns for an index variable for this exchange (`ISE`) from January 5th, 2009 to February 22, 2011, along with returns for indexes for the following markets: US (`SP`), Germany (`DAX`), the UK (`FTSE`), Japan (`NIKKEI`), Brazil (`BOVESPA`), and the MSCI Emerging markets index (`EM`).

(a) Below is the ACF plot for the time series `ISE` as well as a simple linear regression.




```
Call:
lm(formula = ISE[2:T] ~ ISE[1:(T - 1)])

Residuals:
   Min     1Q Median     3Q    Max
-0.084 -0.012  0.001  0.011  0.105

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.00137    0.00091     1.5     0.13
ISE[1:(T - 1)]  0.07254    0.04312     1.7     0.09 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 0.021 on 533 degrees of freedom
Multiple R-squared:  0.0053,    Adjusted R-squared:  0.0034
F-statistic: 2.8 on 1 and 533 DF,  p-value: 0.093
```

(i) Explain in words what is shown in the ACF plot.

(ii) Write down the model that is implicitly behind the regression on the right above? Identify the independent and dependent variables.

(iii) Interpret the results of the regression output making specific reference to the ACF plot.

(b) Consider simple linear regressions of the ISE returns on each of the other indexes, individually. We have the following pairwise correlations

```
          ISE   SP  DAX FTSE NIKKEI BOVESPA   EM
ISE     1.00 0.45 0.63 0.65   0.39    0.45 0.70
```

For the individual linear regressions, we have the following output.

```
> c(extractAIC(lm(ISE ~ SP), k=log(536))[2],
+   extractAIC(lm(ISE ~ DAX), k=log(536))[2],
+   extractAIC(lm(ISE ~ FTSE), k=log(536))[2],
+   extractAIC(lm(ISE ~ NIKKEI), k=log(536))[2],
+   extractAIC(lm(ISE ~ BOVESPA), k=log(536))[2],
+   extractAIC(lm(ISE ~ EM), k=log(536))[2])
[1] -4244.630 -4393.779 -4416.384 -4213.648 -4243.023 -4487.411
```

(i) Do the pairwise correlations agree with the second set of output? How so?

(ii) Based on this output alone, can you tell which, if any, of the individual regressions will yield statistically significant slope coefficients (i.e. different from zero)? If so, which ones and how? If not, why not?

5 Transformations

The following questions pertain to a regression model fit, summarized below, to data on the brain weight (`brain.weight`, in grams) and the body weight (`body.weight`, in kilograms) of 62 mammals.

```
Call:
lm(formula = log(brain.weight) ~ log(body.weight))

Residuals:
     Min       1Q   Median       3Q      Max
-1.71550 -0.49228 -0.06162  0.43597  1.94829

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.13479    0.09604   22.23   <2e-16 ***
log(body.weight)  0.75169    0.02846   26.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6943 on 60 degrees of freedom
Multiple R-squared:  0.9208,    Adjusted R-squared:  0.9195
F-statistic: 697.4 on 1 and 60 DF,  p-value: < 2.2e-16
```

(a) How does the model assume brain weight and body weight are related in their original scale?

(b) How would you justify using the log transform here? What additional evidence, if any, would you like to base your decision on?

(c) What percentage of the variation in the response is explained by regressing onto the explanatory variable? Do we have reason to believe in a linear relationship between these variables? State the formal hypothesis test (and the conclusions).

(d) What would the model predict for brain weight for a mammal with a body weight of 110 kgs? If $s_{\text{fit}}(\log 110) = 0.145$ (in units of log grams), give a 95% predictive interval for the brain weight.

6 Logistic Regression

For 546 homes we observe the following information

`AC` = 1 if it has air conditioning, 0 if not,
`price` = sale price, and
`bedrooms` = number of bedrooms.

Our goal is to predict which houses have air condition and which do not.

(a) First we use a logistic regression with only `bedrooms` and obtain the following output.

```
Call:
glm(formula = AC ~ bedrooms, family = "binomial")

Coefficients:
            Estimate Std. Error z value    Pr(>|z|)
(Intercept)   -2.179      0.399   -5.46 0.000000047
bedrooms       0.469      0.127    3.68     0.00023
```

Give a precise, numerical interpretation of the relationship between `bedrooms` and `AC` by:

(i) Interpreting the coefficient estimate, 0.469, in this context.

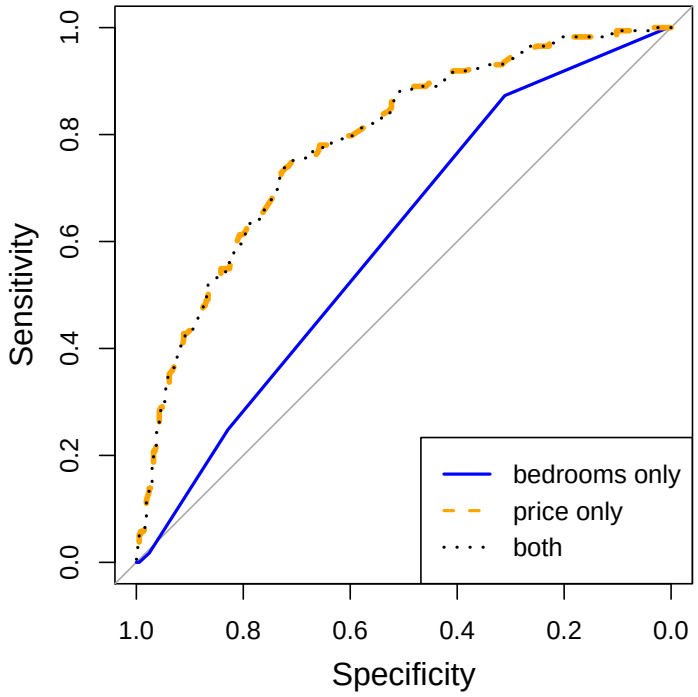
(ii) Interpreting the statistical significance as best you can with this output.

(b) We now switch to using `price` and obtain the following output.

```
Call:
glm(formula = AC ~ price, family = "binomial")

Coefficients:
               Estimate  Std. Error z value            Pr(>|z|)
(Intercept) -3.76328283  0.34599029   -10.9 <0.0000000000000002
price        0.00004223  0.00000459     9.2 <0.0000000000000002
```

Based on this output, is `price` or `bedrooms` better at predicting whether or not a house has AC? Cite specific numeric evidence in favor or your argument.

Use the plot below to answer the next two questions.



(c) Provide an intuitive explanation of what is being plotted here.

(d) Based on this plot, comment on how `bedrooms` contribute to predicting `AC` when used in conjunction with `price`.