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

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:

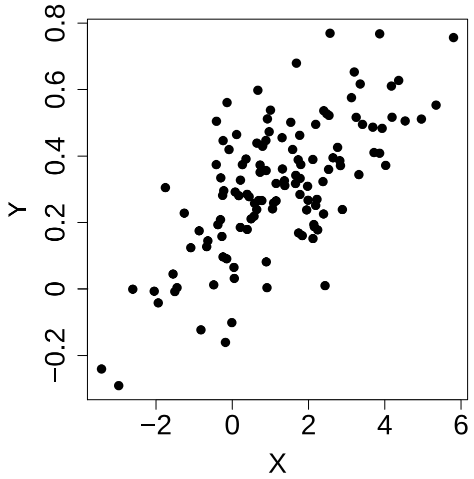
Signed: ______________________________

- You have 3 hours to complete the exam.
- This exam has 17 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) $b_0 = 0,\ b_1 = 0.2$

(ii) $b_0 = 0.2,\ b_1 = 0.1$

(iii) $b_0 = 0,\ b_1 = -1$

(iv) $b_0 = -0.2,\ b_1 = 1$

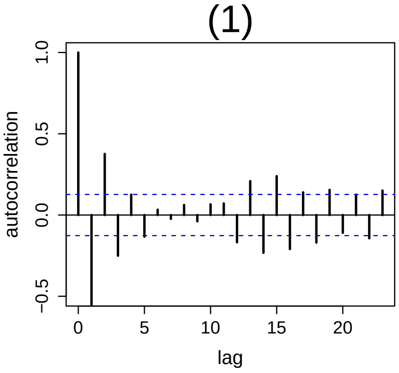
(v) $b_0 = 0.1,\ b_1 = 0.2$

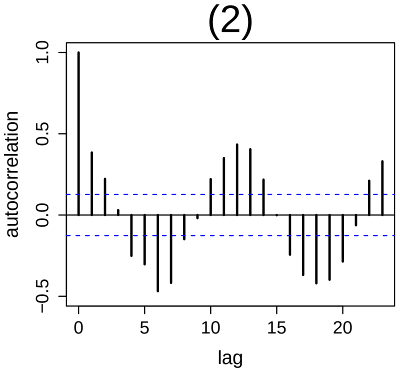
(b) Label each of the time series models listed below as (1) – (6) to match it to the correct `acf` plot of the six shown here. In all cases $\varepsilon_t \sim \mathcal{N}(0,1)$.

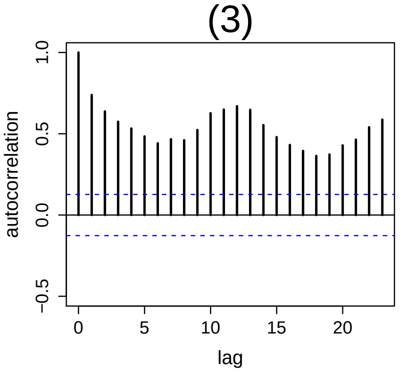
( ) $Y_t = 0.7Y_{t-1} + \varepsilon_t$ ( ) $Y_t = 0.9Y_{t-1} + \varepsilon_t$

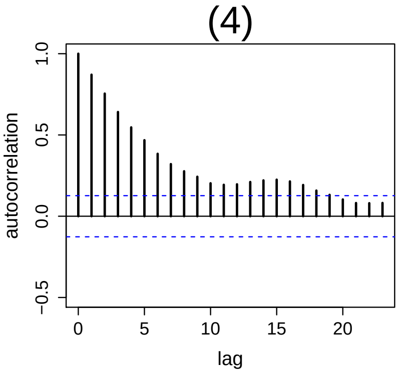
( ) $Y_t = -0.6Y_{t-1} + \varepsilon_t$ ( ) $Y_t = \varepsilon_t$

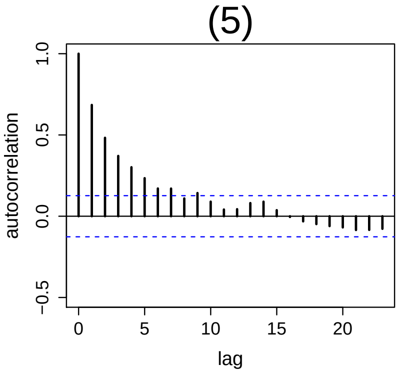
( ) $Y_t = \sin(2\pi t/12) + \cos(2\pi t/12) + \varepsilon_t$ ( ) $Y_t = \sin(2\pi t/12) - t/50 + \varepsilon_t$

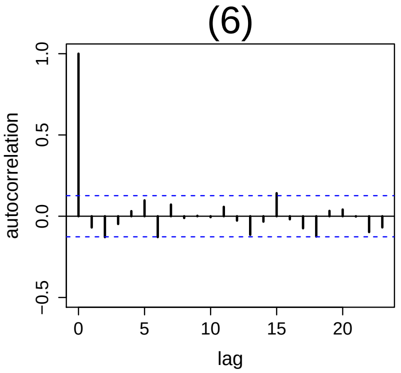












(c) If there is time dependence in a series, the first-order autocorrelation will be statistically significantly different from zero at the 5% level in a large enough sample. TRUE or FALSE? Justify your answer.

(d) Suppose you know that an outcome Y obeys the following model

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 t + \varepsilon_t, \qquad \text{where} \quad \varepsilon_t \overset{iid}{\sim} \mathcal{N}(0,1) \text{ and independent of } X.$$

(Note all the standard regression assumptions are met.) However, you only observe $\Delta Y_t = Y_t - Y_{t-1}$ and $\Delta X_t = X_t - X_{t-1}$, not Y_t or X_t. To estimate β_1 you run a regression of ΔY_t on ΔX_t.

(i) What is the distribution of the error term in the model for the regression of ΔY_t on ΔX_t, i.e. $\Delta\varepsilon_t = \varepsilon_t - \varepsilon_{t-1}$?

(ii) Does the regression of ΔY_t on ΔX_t satisfy all the standard regression assumptions, including those regarding the errors? Justify your answer.

(e) In the context of variable selection:

(i) Explain the difference between `forward` and `backward` stepwise selection.

(ii) Explain the difference between the BIC and AIC selection criteria.

(f) Suppose you run a factory which each day produces an output Y as a function of capital (e.g. machines, computers, trucks, etc) K and labor (people) L according to the production function $Y_t = AK_t^{\beta_1} L_t^{\beta_2} e^{\varepsilon_t}$, where $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$, and represents an idiosyncratic shock to daily productivity (e.g. an employee is less focused, a machine breaks down, etc) which is independent of K_t and L_t.

(i) Provide a regression model specification to estimate the elasticity of production with respect to capital per worker. Write all parameters of this model in terms of those given above: $A, \beta_1, \beta_2, \sigma^2$.

(ii) Does your regression model obey the standard regression assumptions? Justify your answer.

2 Simple Linear Regression

As discussed in class, the *Capital Asset Pricing Model* (CAPM) relates the rate of return of **any** given asset, R_A, to the market return, R_M, through the following model:

$$\mathbb{E}\left[R_A \mid R_M\right] = \alpha + \beta R_M.$$

In this problem, the asset returns R_A will be for either `Ford` or `GE` stock and the market return R_M will be `MarketReturn`. We have monthly data from 1980–2016. Over this period, the average returns from market, `Ford`, and `GE`, are 1.25, 1.75, and 1.53, respectively, while the standard deviations are 4.1, 11.43, and 6.56.

(a) Clearly and concisely state **all** of the simple regression assumptions that we would use to estimate this CAPM from data.

(b) Use the following `summary` output, from separate CAPMs for `Ford` and `GE`, to answer the questions below.

```
Call:
lm(formula = Ford ~ MarketReturn)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.12       0.51     0.2      0.8
MarketReturn     1.32       0.12    10.8   <2e-16

Residual standard error: 10 on 431 degrees of freedom
Multiple R-squared:  0.21,  Adjusted R-squared:  0.21
F-statistic: 1.2e+02 on 1 and 431 DF,  p-value: <2e-16
```

```
Call:
lm(formula = GE ~ MarketReturn)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.136      0.237     0.6      0.6
MarketReturn    1.133      0.057    20.0   <2e-16

Residual standard error: 4.7 on 431 degrees of freedom
Multiple R-squared:  0.48,  Adjusted R-squared:  0.48
F-statistic: 4e+02 on 1 and 431 DF,  p-value: <2e-16
```

(i) For which of `Ford` or `GE` does the CAPM better explain observed stock returns? Cite and interpret specific values from the `summary` output.

(ii) For both stocks, comment on whether it is likely, given this data, that $\alpha = 0$ and separately, if $\beta = 1$. Cite and interpret specific values from the `summary` output.

(c) Suppose that next month the market does not change (so that `MarketReturn` $= 0$).

(i) Give a 95% prediction interval for next month's return for `Ford`.

(ii) Would a 95% prediction interval for next month's return for `GE` be the wider, narrower, or the same length, as the on for `Ford` you gave above? Justify your answer.

3 Multiple Linear Regression and Model Building

The *Fama-French three-factor* model aims to improve on the Capital Asset Pricing Model (CAPM) by adding two additional variables to explain the returns of **any** given asset. [Bonus: it was developed at Booth!] Recall that the CAPM relates the rate of return of **any** given asset, R_A, to the market return, R_M, through the following model:

$$\mathbb{E}\left[R_A \mid R_M\right] = \alpha + \beta R_M.$$

The Fama-French three-factor model adds to this

- S: a measure of the historic excess returns of **S**mall cap[1] stocks over big cap stocks, and
- V: a measure of the historic excess returns of **V**alue stocks[2] over growth stocks.

The model thus becomes

$$\mathbb{E}\left[R_A \mid R_M, S, V\right] = \alpha + \beta_1 R_M + \beta_2 S + \beta_3 V.$$

In this problem we have monthly returns data from 1980–2016, where

- the asset returns R_A will be for either `Ford` or `GE`;
- the market return R_M will be `MarketReturn`;
- the variable S will be `SmallMinusBig`;
- the variable V will be `HighMinusLow`.

(a) Clearly and concisely state **all** of the regression assumptions that we would use to estimate the three-factor model from data.

[1] "cap" is short for market capitalization, the value of all shares outstanding of a company. Companies are divided into three categories: (1) small-cap, the least valuable, (2) mid-cap, and (3) big-cap, the most valuable. The variable S reflects that, historically, small-cap firms have higher returns than big-cap.

[2] "Value" stocks have a high book-to-market ratio, where the "book" value is the value of a company based on its accounting balance sheet and the "market" value is the market capitalization. "Growth" stocks have a low boot-to-market ratio. The variable S reflects that, historically, value firms have higher returns than growth.

(b) Below is a `summary` of the three-factor model for `Ford` and an `anova` table for it vs the CAPM.

```
Call:
lm(formula = Ford ~ MarketReturn + SmallMinusBig
                                 + HighMinusLow)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.516      0.494    -1.0      0.3
MarketReturn    1.563      0.121    12.9   <2e-16
SmallMinusBig   0.081      0.163     0.5      0.6
HighMinusLow    1.193      0.175     6.8    3e-11

Residual standard error: 0.1 on 429 degrees of freedom
Multiple R-squared:  0.29,   Adjusted R-squared:  0.29
F-statistic:  59 on 3 and 429 DF,  p-value: <2e-16
```

```
Analysis of Variance Table

Model 1: Ford ~ MarketReturn
Model 2: Ford ~ MarketReturn + SmallMinusBig + HighMinusLow
  Res.Df   RSS Df Sum of Sq  F Pr(>F)
1    431 44383
2    429 39888  2      4495 24  1e-10
```

Consider the following candidate models for `Ford` returns:

$$\textbf{(i)}\quad \mathbb{E}\left[R_A \mid R_M, S, V\right] = \alpha + \beta_1 R_M$$

$$\textbf{(ii)}\quad \mathbb{E}\left[R_A \mid R_M, S, V\right] = \alpha + \beta_1 R_M + \beta_2 S$$

$$\textbf{(iii)}\quad \mathbb{E}\left[R_A \mid R_M, S, V\right] = \alpha + \beta_1 R_M + \beta_3 V$$

$$\textbf{(iv)}\quad \mathbb{E}\left[R_A \mid R_M, S, V\right] = \alpha + \beta_1 R_M + \beta_2 S + \beta_3 V$$

Is it possible to tell from the output above which of these models best explains the observed variation in returns? Why or why not? Use the output to justify your answer.

(c) For each of the models above, numbered **(i)**, **(ii)**, **(iii)**, and **(iv)**, we compute BIC $= n\log(s^2) + p\log(n)$, which gives the following results:

	Model **(i)**	 **(ii)**	 **(iii)**	 **(iv)**
n	433	433	433	433
s^2	103.0	102.8	92.82	92.98
p	1	2	2	4
BIC	2016.88	2021.46	1976.96	1982.78

Which model does this output suggest is "best"? What does "best" mean here? Explain why the model selected here is the same or different from the one you selected in **part (b)**.

(d) To improve the CAPM, Gene Fama and Ken French chose to add the two variables `SmallMinusBig` and `HighMinusLow`. We do not know why they selected these two variables. We will now develop our own three-factor model by choosing two variables to add to the CAPM. In addition to the variables above, we also have:

- `RiskFree` is the risk free interest rate;
- `ShortTermReversal` measures the tendency of stocks with strong gains and stocks with strong losses to reverse in a short-term time frame (up to one month);
- `LongTermReversal` is the same, but for time frames of 3-5 years;
- `Momentum` measures the tendency for the stock price to continue rising if it is going up and to continue declining if it is going down;
- `Ind_Consumer`, `Ind_Manufact`, `Ind_HighTech`, `Ind_Health`, `Ind_Other` are returns for the consumer goods, manufacturing, high-tech, health care, and other, respectively.

Consider the below R commands and (excerpted) output. The “`~ .`” collects these variables and `MarketReturn`, `SmallMinusBig`, `HighMinusLow`.

```
> full <- lm(Ford ~ .)
> NewThreeFactorModel <- step(CAPM.Ford, scope=formula(full), direction="forward", k=log(n), steps=2)
Start:  AIC=2016.88
Ford ~ MarketReturn

Step:  AIC=1955.71
Ford ~ MarketReturn + Momentum

Step:  AIC=1935.58
Ford ~ MarketReturn + Momentum + HighMinusLow
```

(i) Explicitly list the elements of this process which are controlled by the user as opposed to what is controlled by the computer, and the choices made in this instance.

(ii) We now repeat the above process considering GE instead of Ford returns, according to the below R commands and (excerpted) output.

```
> full <- lm(GE ~ .)
> NewThreeFactorModel <- step(CAPM.GE, scope=formula(full), direction="forward", k=log(n), steps=2)
Start:  AIC=1354.42
GE ~ MarketReturn

Step:  AIC=1327.27
GE ~ MarketReturn + SmallMinusBig

Step:  AIC=1307.95
GE ~ MarketReturn + SmallMinusBig + LongTermReversal
```

Are there any evident differences between this model building and the one above for Ford? Explain why there are, or are not, differences.

(iii) Is it possible, with this output, to assess which three-factor model is better: the Fama-French model, our model for Ford, or our model for GE? If so, which is better and why? If not, what additional information would you like to see in order to make a comparison?

4 Logistic Regression

A credit card company has asked us to build a decision rule for accepting or rejecting future credit card applications. For 1,312 past applications, we have the actual outcome and seven applicant characteristics. Thee outcome of interest is the binary variable `given.card` indicating if the application for a credit card was accepted or not. Of the 1,312 applicants, 1,017 (or 78%) were given a card, 295 were denied. The seven characteristics consist of four binary and three continuous variables:

`home.owner` = {0,1}. Does the individual own his or her home?
`self.employed` = {0,1}. Is the individual self-employed?
`has.negative.reports` = {0,1}. Does the individual have any bad credit reports?
`have.card.already` = {0,1}. Does the individual have any major credit cards already?
`age` = Age in years
`annual.income` = Yearly income (in USD 10,000)
`ratio.spending.income` = Ratio of monthly credit card expenditure to yearly income

We will build a classifier using logistic regression on a training subsample of the data and test it on a held-out validation sample. For training, 500 of the 1312 applications are randomly selected, leaving the remaining 812 for testing.

(a) In our current context, what are the two types of errors that a classifier can make?

(i)

(ii)

In the present decision-making context, which type of mistake do you think is "worse" and why? (*Full credit for good reasoning; either error can be the correct answer.*)

(b) This credit card company previously hired Bad Consulting Group, Inc to develop a decision rule. These consultants came up with the rule to reject applicants if they have bad credit reports (`has.negative.reports` = 1) and accept all others (`has.negative.reports` = 0). In the testing sample of 812 applicants, this decision rule gives the following results:

		BCG classifier	
		Reject	Accept
`given.card`	No	92	81
	Yes	68	571

From this table, compute the error rates for this decision rule for both types of errors you identified in **part (a)**.

(i)

(ii)

We will try to build a better rule by using data. Consider the below `step`wise selection commands and (excerpted) output and the `summary` of the final model.

```
> base <- glm(given.card ~ 1, family=binomial, data=CreditCard[training.samples,])
> full <- glm(given.card ~ has.negative.reports + have.card.already + home.owner + self.employed +
age + annual.income + log(ratio.spending.income), family=binomial, data=CreditCard[training.samples,])
> final <- step(base, scope=formula(full), direction="forward", k=log(length(training.samples)))
Start:  AIC=561.86
given.card ~ 1

Step:  AIC=117.51
given.card ~ log(ratio.spending.income)

Step:  AIC=115.53
given.card ~ log(ratio.spending.income) + annual.income

Step:  AIC=113.8
given.card ~ log(ratio.spending.income) + annual.income + has.negative.reports

> summary(final)

Call:
glm(formula = given.card ~ log(ratio.spending.income) + annual.income + has.negative.reports, family=binomial)

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                   17.81       3.65     4.9    1e-06
log(ratio.spending.income)     2.75       0.53     5.2    2e-07
annual.income                  0.57       0.19     3.0    0.002
has.negative.reportsTRUE      -1.97       0.83    -2.4    0.017
```

(c) Give a precise and numerical explanation of what these results say about the relationship between `ratio.spending.income` and the outcome `given.card`.

(d) Based on this output, how much more likely or less likely to be given a credit card is someone with prior negative reports compared to someone without prior negative reports?

(e) We obtain from the final selected model (in the above `summary`) estimated probabilities for giving each of the 812 testing applicants a credit card, that is we find the estimated $\mathbb{P}[\texttt{given.card} = 1 \mid X]$, for $X = (\texttt{ratio.spending.income}, \texttt{annual.income}, \texttt{has.negative.reports})$. Our decision rule will be to accept applications when this probability is above a cut-off κ. For $\kappa = \{0.15, 0.25, 0.35\}$, this rule gives the following results:

		`Our classifier` for different κ					
		$\kappa = 0.15$		$\kappa = 0.25$		$\kappa = 0.35$	
		Reject	Accept	Reject	Accept	Reject	Accept
`given.card`	No	129	44	159	14	172	1
	Yes	4	635	5	634	8	631

(i) For the first type of error you identified in **part (a)**, examine how this error responds to changes in κ, and give a well-reasoned explanation for any pattern you find.

(ii) Based on this output and your answer in **part (a)**, what is your preferred choice of κ and why? (*Improving on the BCG classifier is not a good reason why!*)

5 Transformations

We have data from 400 stores regarding the `Sales` of a certain carseat model. The variables are:

`Sales` = Units sold (in thousands) at each location
`Price` = The price of the carseat
`CompPrice` = The price of a competing model
`Display.Quality` = {`Bad` or `Good`}, indicating the quality of the shelf position in the store

Using these variables in a linear regression gives the following `summary`.

```
Call:
lm(formula = log(Sales) ~ log(CompPrice) + Display.Quality * log(Price))

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                         2.4        1.0       2    0.019
log(CompPrice)                      1.8        0.2      10   <2e-16
Display.QualityGood                -2.9        0.9      -3    0.002
log(Price)                         -2.0        0.2     -11   <2e-16
Display.QualityGood:log(Price)      0.7        0.2       4    3e-04

Residual standard error: 0.4 on 394 degrees of freedom
Multiple R-squared:  0.5,      Adjusted R-squared:  0.5
F-statistic: 8e+01 on 4 and 394 DF,  p-value: <2e-16
```

(a) Give a precise interpretation of the coefficient for `Display.QualityGood` (which is an indicator for `Display.Quality`=`Good`). Does this value make intuitive sense to you? Explain.

(b) Describe all the values for the **competitor** price elasticity that could be rejected at level $\alpha = 0.05$, against a two-sided alternative.

(c) Compute the **own** price elasticity for each level of `Display.Quality`.

(d) Explain/justify any patterns you see in the **own** price elasticities computed in **part (c)**.

(e) At a certain store, the carseat is in a `Bad` display position, priced at \$150, and the competing model is also \$150.

(i) Which of the these changes would be associated with the largest increase in `Sales` and why? You may find useful that $\log(150) = 5$.

(1) a \$30 decrease in price

(2) a \$30 increase in the competitor's price

(3) moving to a `Good` display position

(4) a \$15 decrease in price *and*
a \$15 increase in the competitor's price

(ii) Is this context, explain what a decision maker would learn from a 95% interval for the change in `Sales` following your chosen action, and why or why not, this information would be useful for decision making.

(iii) Can you give a 95% interval for the predicted `Sales` following your chosen action above? If so, state the interval, or if not, state what other information you would require. (*No need to do a lot of calculations here, just show the correct expression with the correct numbers in the correct places.*)