Chicago Booth BUS 41100
Solutions to Final Exam **SAMPLE #1**

Instructor: Max H. Farrell

These solutions are a guide only! Your answers should show more work/detail/reasoning.

1 Simple Linear Regression Mechanics

(a) **(ii)** because the t statistic is $\frac{b_1 - 0}{s_{b_1}} = \frac{2}{s_{b_1}} = 2.1$, and solving yields 0.95.

(b) Only **(i)** because you have more data, but the exact same estimate for b_1, which is still different from zero. Hence you must be more sure that $\beta_1 \neq 0$.

(c) **(v)** because they all lie on the line $y = 1 + 2x$.

2 Multiple linear regression

(a) **(i)** The coefficient means that on average people that underwent training made \$-15204.80 **less** than those that did no training. This difference is signifcant at the 0.05 level. The null is $H_0 : \beta_1 = 0$ and the alternative is $H_1 : \beta_1 \neq 0$; the t test statistic value is -13.17 and the p value is given as $< 2e - 16$.

(ii) One possibility (and the later parts verify that this is correct) is that the people that underwent training make much less to begin with. Notice that in this part, we are not modeling the **change** in earnings, just the level afterward. If the people that did the training program had very low earnings relative to those that did not do training, then even if the training raised their income, it would still be lower than the no-training sample.

(b) **(i)** The coefficient means that on average people that underwent training made \$479.20 **more** than those that did no training, *holding fixed education and pre-training earnings.* So if you have two people with the same education level and the same earnings, and you put only one through training, then on average that one will earn \$479.20 more than the untrained person. This difference is **not** signifcant at the 0.05 level. The null is $H_0 : \beta_1 = 0$ and the alternative is $H_1 : \beta_1 \neq 0$; the t test statistic value is 0.583 and the p value is given as 0.55992.

(ii) Statistically, the training does not add significant value, once controling for pre-training earnings and education level. This doesn't rule out the possibility that some sub-groups could benefit, such as specific earnings levels or demographic groups.

(c) Clearly there is non-constant variance and there may also be outliers.

The other problem is the weird linear pattern in the bottom left corner of plot 1. Remember that plotting residuals against X should show a nice cloud of points. So what's going on here? This is caused by the fact that many people have zero earnings, i.e. they are unemployed. This is true both before and after training, and for people that did training and people that didn't.

To fix all these a lot needs to be done actually. Most obviously, taking a log transform of education or using dummy variables for certain levels of attainment may help. Removing the outiers may help. We can't take the log of earnings either (before or after) because it takes the value zero very often. So the best strategy would be to treat these points separately, or to remove them entirely. Full credit would be given for a well-reasoned answer.

(d) **(i)** The null hypothesis is $H_0 : \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$ and the alternative is $H_1 : \beta_j \neq 0$ for at least one $j \in \{4, 5, 6, 7, 8\}$. Conceptually, the null hypothesis is that the additional variables, as a whole, do not add anything to the model, while the alternative is that "Model 2" is prefered to "Model 1", i.e. the new variables add something. The p-value is less than 0.05, so at the 5% level, we reject the null hypothesis and decide that the new model is prefered.

(ii) According to BIC, the old model is better, but according to the partial F test, the new model is. Which is better is largely a matter of taste at this point. Full credit would be given for any correctly justified answer.

Remember that our goal is to estimate β_1, and the estimates have different interpretations in the different models because they control for different things. If we do not want to condition on demographics, like race, then the first model is prefered.

3 Understanding regression output

We have 70 observations of flat panel TV `price` collected from an online retailer, and we also know the `size` in inches of the diagonal length of the viewing area, the `brand` (indicating LG, Panasonic, or Samsung), and the `type` (indicating an LED or plasma). Our goal is to build a model to predict `price` using the other three variables.

Consider the following regression output:

```
Call:
lm(formula = log(price) ~ size + type + brand)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.975399   0.190335  26.140  < 2e-16 ***
size           0.045329   0.003923  11.554  < 2e-16 ***
typeplasma    -0.266354   0.070914  -3.756 0.000371 ***
brandPanasonic -0.017702   0.085992  -0.206 0.837546
```

```
brandSamsung    0.174208   0.064959   2.682 0.009272 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.231 on 65 degrees of freedom
Multiple R-squared:  0.7309,    Adjusted R-squared:  0.7143
F-statistic: 44.13 on 4 and 65 DF,  p-value: < 2.2e-16
```

(a) Conceptually, what do the coefficient estimates for `typeplasma`, `brandPanasonic`, and `brandSamsung` add to our understanding of the relationship between `price` and `size`? That is, visually, what do these variables represent if the relationship were plotted?

Solution. *These allow the* **intercept** *of the line of* `price` *on* `size` *to be different by* `brand` *and by* `type`*.*

(b) Numerically, give an interpretation of the results for `brandPanasonic`. Interpret both the coefficient estimate itself and the associated significance testing.

Solution. *The omitted brand is LG. The intercept says that Panasonic TVs are 1.7% cheaper than LGs on average (their line is shifted down by this amount), however, this difference is not statistically significant (the p-value = 0.84 ¿ 0.05).*

(c) What would be the p-value for the partial F test of whether `type` is worthwhile to add to the model beyond `size` and `brand`? What do you conclude based on this p-value?

Solution. *Remember that for one variable, the partial F test and the t test are identical, so the p-value is 0.000371, and we conclude that* `type` *is a worthwhile addition.*

For parts **(d) - (f)** below, consider the following expanded regression.

```
Call:
lm(formula = log(price) ~ size * type + brand)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.992840   0.223091  22.380  < 2e-16 ***
size             0.044945   0.004680   9.603 5.07e-14 ***
typeplasma      -0.331737   0.433194  -0.766  0.44661
brandPanasonic  -0.013657   0.090589  -0.151  0.88064
brandSamsung     0.174483   0.065477   2.665  0.00974 **
size:typeplasma  0.001255   0.008200   0.153  0.87886
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2327 on 64 degrees of freedom
Multiple R-squared:  0.731,     Adjusted R-squared:   0.71
F-statistic: 34.78 on 5 and 64 DF,  p-value: < 2.2e-16
```

(d) Conceptually, what does the coefficient estimate for `size:typeplasma` add to our understanding of the relationship between `price` and `size`? That is, visually, what does this variable represent if the relationship were plotted? How is this different from part **(a)**?

Solution. *The interaction of* `size` *and* `type` *allows for a different* **slope** *for different types of TVs, as well as different intercepts, whereas part* **(a)** *it is the intercepts are different.*

(e) What would be the p-value for the partial F test of whether this expanded model is worthwhile beyond the original regression? What do you conclude based on this p-value?

Solution. *Even though it seems like we're adding more than one variable, the logic of part* **(c)** *still applies. The p-value is 0.87886 and we conclude that the different slopes are not worthwhile.*

(f) Comparing this regression output to the original, why does your conclusion in part **(e)** make sense? Cite specific values from both outputs.

Solution. *The fact that the different slopes are not worth doing makes senses when you look at the R^2 values: 0.7309 in the first model and 0.7310 in the second, a difference of 0.0001. The partial F test is a formal test of whether this is a statistically "big" difference, and with a sample size of 70, there is no way it is significant.*

4 Classification & Model Building

(a) We can (i) flag an email as spam when it is not, or (ii) we can fail to flag a message as spam when it is in fact spam. Which of these is worse is a matter of taste. Here's one idea. Suppose that people don't ever shift through their spam folder (probably true). Then, I really worry about the first error, because it means people could be missing important emails, whereas the second one is only an inconvenience in having to delete some junk from you inbox. To reason for the opposite, if on average each person receives thousands of spam messages per day, and only a couple important messages, then you might want the computer to throw most messages away, because people won't want to scan through a lot of spam.

(b) A one percent increase in the characters that are dollar signs, holding everything else fixed, is associated with an increase in the odds ratio of $e^{11.86} = 141,492$! That's huge. According to this model, just a small increase in the percent of characters that are dollar signs is a very strong signal that the message is spam. The increase is so big that I'm not convinced this is a good model.

Here's an example. The answer to part (a) above has 716 characters, and currently zero dollar signs, and hence zero percent dollar signs. If I added just a single dollar sign, it would then have $100 \times 1/717 = 0.13947\%$ dollar signs. The change in the odds ratio would be $e^{0.13947 \times 11.86} \approx 5$, so the message is already 5 times more likely to be spam. If I added two dollar signs, the message would be 30 times more likely to be spam!

(c) Having an average string of caps that is one character longer, holding everything else fixed, is associated with an increase in the odds ratio of $e^{0.21} = 1.23$. Also, having an average two longer means an increase of $e^{2 \times 0.21} \approx 1.5$, that is, 50% more likely to be spam. This seems credible to me. Most messages just have one capital letter at a time (the start of the sentence, a name, etc), with maybe a few acronyms thrown in.

(d) For a message which contains 2 parentheses, 2 exclamation points, zero dollar signs, and never strings together more than one capital letter, what estimated probability this message is spam?

The linear log odds ratio at this point is: $-1.75 + 1 \times 0.21 - 2 \times 1.66 + 2 \times 1.38 + 0 \times 11.86 = -2.1$.

Taking the exponential gives 0.12. Thus we need to solve

$$\frac{p}{1-p} = 0.12,$$

which gives $p = 3/28 \approx 1/10$.

(e) If you are worried about error (i), you prefer $K = 3/4$, so you flag fewer messages as spam. For error (ii), $K = 1/4$. If you are indifferent between the two errors, set $K = 1/2$.

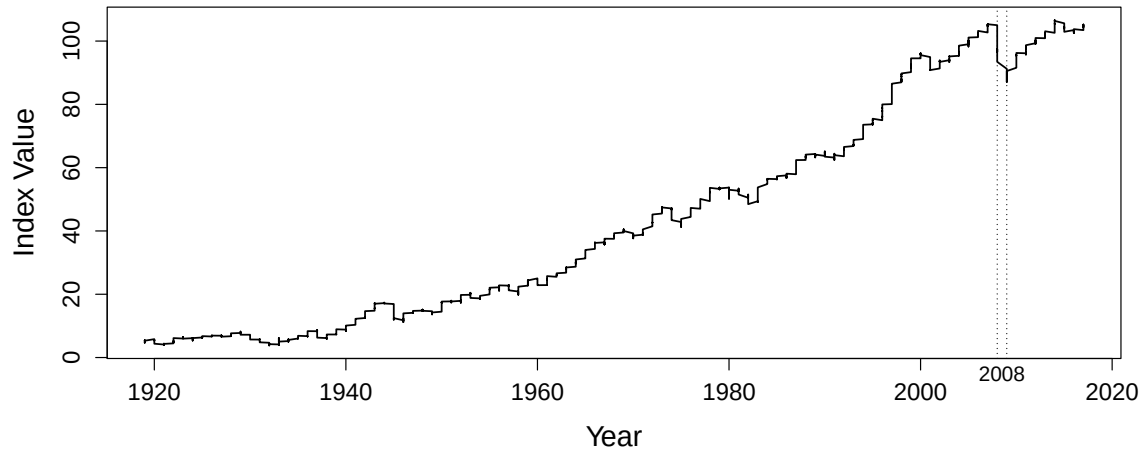
(f) The rate for error (i) is 130/(2658+130) and for error (ii) it is 628/(628+1185).

(g) See class discussion.

(h) The -1 column corresponds to errors of type (i), the +1 column is type (ii). So you might prefer LASSO to avoid type (i) errors and AIC to avoid type (ii) errors. BIC makes more errors of both kinds than does AIC, so you should never prefer BIC according to these results.

5 Time Series #1

The Industrial Production Index (`index.prod`) is an economic indicator that measures real output for all facilities located in the United States, including manufacturing, mining, and electric, and gas utilities. It measures movements in production output and highlights structural developments in the economy. The index is compiled on a monthly basis to bring attention to short-term changes in industrial production. We have the index value from January 1919 to September 2017. The series is plotted below, with 2008 between the dotted lines.



Consider the autoregressive model of order 1, AR(1): $\texttt{index.prod}_t = \beta_0 + \beta_1 \texttt{index.prod}_{t-1} + \varepsilon_t$. The (abbreviated) `summary` output from this regression is below.

```
Call:
lm(formula = index.prod[2:T] ~ index.prod[1:(T - 1)])

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.051139   0.019251    2.66    0.008
```

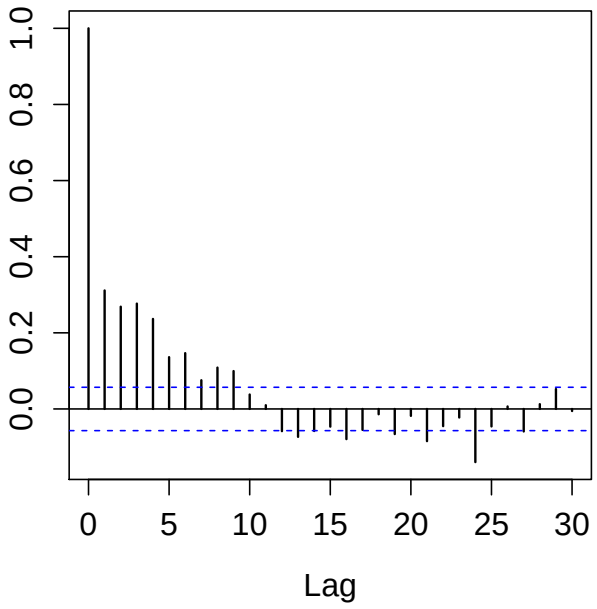
```
index.prod[1:(T - 1)] 1.000758   0.000349 2864.51   <2e-16
---

Residual standard error: 0.406 on 1182 degrees of freedom
Multiple R-squared:     1,      Adjusted R-squared:     1
F-statistic: 8.21e+06 on 1 and 1182 DF,  p-value: <2e-16
```

(a) Interpret the coefficient, standard error, and p-value corresponding to `index.prod[1:(T - 1)]`, in the context of this specific data set.

Solution. *The coefficient in the AR(1) model is $b_1 = 1.000758$, which is the estimate of β_1 in the above model. The next three entries would be the measures of uncertainty, but because the coefficient is almost exactly one, we have a unit root, and we can not trust these numbers.*

(b) Below is a plot based on the residuals from the AR(1). Use the space below, describe what is being shown, how this plot is useful as a diagnostic tool, and what you conclude in this example.



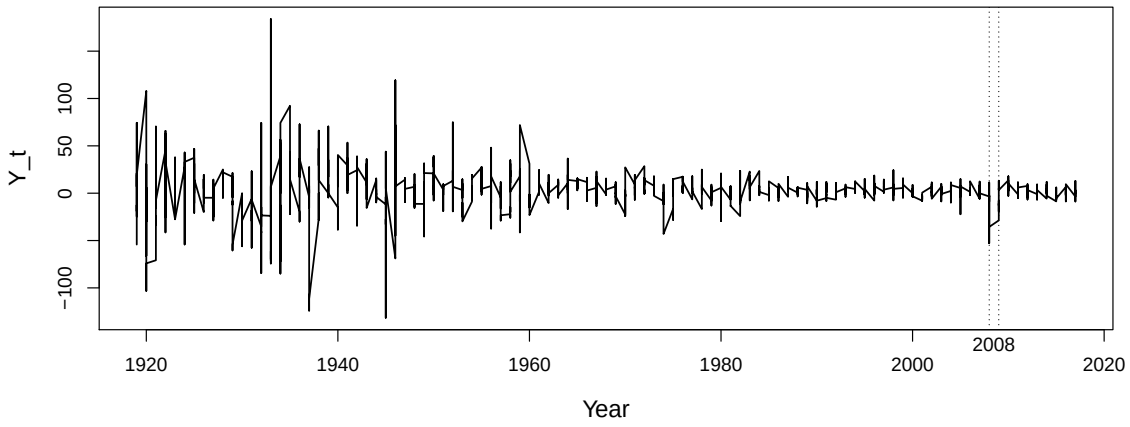
Solution. *This is the autocorrelation function of the residuals. We are looking for remaining information, i.e. patterns, in the "X" variable here, Y_{t-1}. We are trying to diagnose if we've captured the time series dependence pattern in this series. In this example, we conclude that the AR(1) model does not capture all the time dependence. There is still a large of amount of correlation in the short lags.*

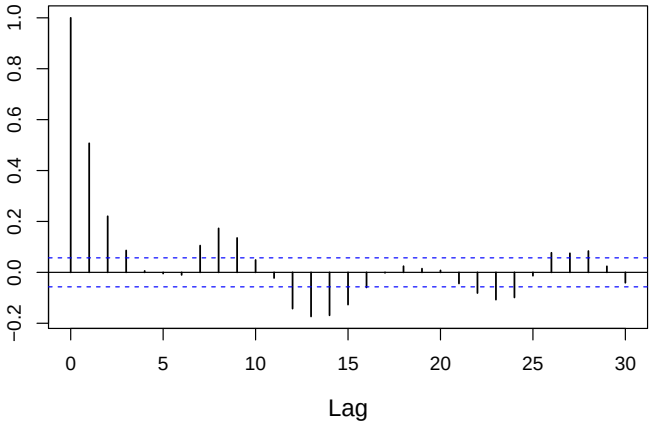
6 Time Series #2

Continuing our study of the Industrial Production Index, we will now consider a transformed series, defined by

$$Y_t = 1200 \times \log\left(\frac{\texttt{index.prod}_t}{\texttt{index.prod}_{t-1}}\right).$$

This new series is plotted below, with 2008 between the dotted lines. Below that is the autocorrelation function (ACF) for Y_t.





(a) A forecaster states that Y_t measures the monthly percentage change in industrial production, measured in percentage points per annum. Is this correct? Explain.

Solution. *This is correct, roughly. The log gives the change, multiplying by 100 turns it into a percentage, and scaling it by 12 turns it into annual change.*

(b) Motivated by the ACF plot, we run the regression model given by

$$Y_t = \beta_0 + \beta_1 \sin(2\pi t/12) + \beta_2 \cos(2\pi t/12) + \varepsilon_t.$$

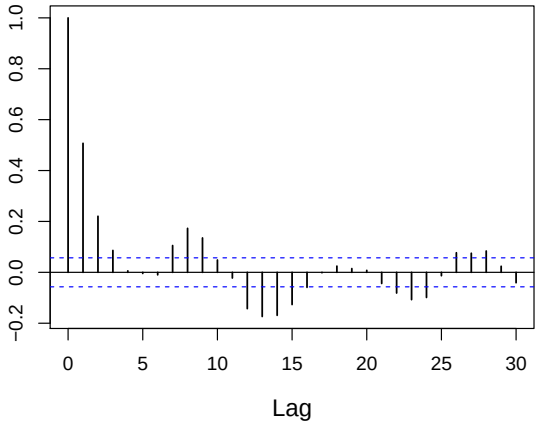
(i) Give a precise explanation of how this regression models the time series behavior of Y_t.

Solution. *We are fitting a cyclical/periodic pattern, with a period of 12 time periods, that is, an annual cycle.*

Abbreviated `summary` output and the ACF plot of the residuals from this model are below.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.07       0.67     4.6    4e-06
sin12           0.30       0.94     0.3      0.8
cos12          -0.54       0.94    -0.6      0.6
---

Residual standard error: 23 on 1181 degrees of freedom
Multiple R-squared: 0.00036,  Adjusted R-squared: -0.0013
F-statistic: 0.22 on 2 and 1181 DF,  p-value: 0.81
```



(ii) Interpret the output and ACF plot. What is the pattern in the plot and why is it there?

Solution. *The coefficients are not significant, and neither is the F test. The ACF plot shows periodicity in the residuals. The problem is that the original cycle is not annual, so adding the sin and cos functions don't help model the series at all.*

(c) We now consider an AR(1) model of Y_t, given by $Y_t = \beta_0 + \beta_1 Y_{t-1} + \varepsilon_t$. Motivated by the idea that the Great Recession in 2008 may have affected the time series, we also try the model $Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 \mathbb{1}\{\texttt{year} = 2008\} + \varepsilon_t$. Below is abbreviated `summary` output from each.

AR(1) Model

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.564      0.578    2.71   0.0069
Y[1:(TY - 1)]    0.507      0.025   20.29   <2e-16
---

Residual standard error: 19.7 on 1181 degrees of freedom
Multiple R-squared: 0.259, Adjusted R-squared: 0.258
F-statistic:  412 on 1 and 1181 DF,  p-value: <2e-16
```

AR(1) + 2008

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.663      0.581    2.86   0.0043
Y[1:(TY - 1)]        0.505      0.025   20.19   <2e-16
year2008[3:T]TRUE   -9.081      5.724   -1.59   0.1129
---

Residual standard error: 19.7 on 1180 degrees of freedom
Multiple R-squared: 0.26,  Adjusted R-squared: 0.259
F-statistic:  207 on 2 and 1180 DF,  p-value: <2e-16
```

(i) Interpret the p-value for `year2008[3:T]TRUE`. What does it mean for these two models?

Solution. *It's the p-value for the partial-F test, and being larger than 0.05, it means that the additional complexity is not worthwhile to add.*

(ii) Consider the following R commands and output.

```
> extractAIC(AR1.model)
[1]    2.000 7055.461
> extractAIC(AR1.plus.2008.model)
[1]    3.00 7054.94
```

Explain what these commands do, i.e. what information is being computed and why is it useful. Then, what does this specific output mean for these two models? Compare your answer to part **(i)** above.

Solution. *The commands are computing the AIC for the two models. The second one is lower, indicating that according to AIC, the larger model is better. This is the opposite of what we found above, but the criteria are different, so it shouldn't being surprising that the answers are different. In particular, we know from class that AIC favors larger models.*

(d) Concerned that the AR(1) model may not be adequate, we consider AR(p) models for lag orders p=1, 2, 3, and 4.

(i) Write down the AR(4) model, complete with assumptions.

Solution. $Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \beta_4 Y_{t-4} + \varepsilon_t$, *where the errors are independent and identically distributed as* $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$. *That's all we need at this level. To go further, and do inference, we need them to be less than one in absolute value.*

(ii) The s^2 from each model is given in the table below. Use these numbers and AIC to estimate the number of lags that should be in the model. Circle your choice in the table. Show your work.

AR order:	1	2	3	4
$\mathbf{s^2}$	388.5	388.2	387.7	387.8

Solution. *The formula for AIC is* $n \log(SSE/n) + p2$. *The tricky part is that both* n *and* p *change for each model. The sample size changes because you have few observations on longer lags. You also have to convert* s^2 *to SSE. If* T *is the total sample size for the index, then* $T-1$ *is the total sample size for* Y. *Let* $k = 1, 2, 3, 4$ *be the lag order. The final formula is*

$$(T-1-k)\log\left(\frac{(T-1-k)-(k+1)}{T-1-k}s^2\right) + (k+1)2.$$

The AIC numbers for the AR models are 7055.391, 7049.514, 7043.027, and 7038.367, so the AR(4) is the winner.