1

It is believed that the relationship between egg and bacon consumption is well described by the linear regression model. However, there is one complicating factor: Easter, which occurs in April and increases the consumption of eggs. What is not clear is whether the Easter effect is best represented as an intercept shift, a slope shift, or both. To analyze this question, you are given the following three regression equations, which all come from the same data set.

	Model	SSR(regression)	SSE(error)
I.	$\hat{y}_t = 30.69 + 2.15x_t$ $s.e.(4.46)(0.14)$	53230	17564
II.	$\hat{y}_t = 18.89 + 23.60d_t + 2.06x_t$ $s.e(2.83)(1.95)(0.08)$	64651	6143
III.	$\hat{y}_t = 21.03 + 19.33d_t + 2.08x_t + 0.14(d_t \times x_t)$ $s.e.(3.76)(5.32)(0.12)(0.16)$	64709	6085

where y_t: number of eggs consumed in month t; x_t: pounds of bacon consumed in month t; and d_t = 1 if month t is April and 0 otherwise. n = 81 for all regressions. SSR: Regression Sum of Squares and SSE: Sum of Squares Errors.

1.a

For each model, compute the R^2, the adjusted $\bar{R}^2$, the AIC, and the BIC.

Model	SSR	SSE	SST	R^2	$\bar{R}_2$	AIC	BIC
Formula	-	-	$SSR + SSE$	$1 - \frac{SSE}{SST}$	$1 - \frac{SSE/(T-K)}{SST/(T-1)}$	$ln(\frac{SSE}{T}) + \frac{2K}{T}$	$ln(\frac{SSE}{T}) + \frac{Kln(T)}{T}$
I.	53230	17564	70794	0.7519	0.7488	5.4285	5.4877
II.	64651	6143	70794	0.9132	0.9121	4.378	4.4371
III.	64709	6085	70794	0.914	0.913	4.3685	4.4276

1.b

Test at the .05 level the hypothesis that both the intercept and the slope are the same in April as in other months. Write out the null and alternative hypotheses.

$H_0 : \begin{cases} \delta_1 = 0 \\ \delta_2 = 0 \end{cases}$
H_1 : At least one is not true

$$F = \frac{\frac{SSR}{K-1}}{\frac{SSE}{T-K}} \tag{1}$$

$$= \frac{\frac{64709}{2-1}}{\frac{6085}{81-2}} \tag{2}$$

$$= 840.1 \tag{3}$$

$F_c = 3.96$ at $\alpha = 0.05$ with 2 hypotheses and 79 degrees of freedom. Since $f > F_c$ (840.1 $>$ 3.96), we have sufficient evidence to reject the null hypothesis at a 0.05 significance level.

1.c

In equation II, interpret the parameter estimates for for the intercept and the dummy variable.

When it is any month but April ($d_t = 0$) and the consumption of bacon is zero, then the consumption of eggs is predicted to be 18.89, ceteris paribus. If the month is indeed April ($d_t = 1$) and bacon

consumption is zero, then the consumption of eggs is predicted to be 23.6 higher than months other than April, or in other words, egg consumption in April will be $18.89 + 23.6 = 42.49$ when bacon consumption is zero, ceteris paribus.

1.d

According to equation III, what is the impact of an additional 10 pounds of bacon consumed on egg consumption in April? What about non-April months?

In April, according to model III, egg consumption for an additional 10 pounds of bacon will be $21.03 + 19.33 + 2.08(10) + 0.14(10) = 62.56$, ceteris paribus. In the other months, egg consumption will be $21.03 + 2.08(10) = 41.83$, ceteris paribus. This means for the month of April in comparison to the other months, $63.55 - 41.83 = 20.73$ more eggs are consumed, ceteris paribus.

1.e

Test at the .05 level the hypothesis that only the intercept is different in April than in other months.

$H_0 : \delta_1 = 0$
$H_1 : \delta_1 \neq 0$

$$t = \frac{d_1 - \delta_1}{s.e.(d_2)} \tag{1}$$
$$= \frac{23.6 - 0}{1.95} \tag{2}$$
$$= 12.1026 \tag{3}$$

Since our $t_c = 1.99$ at 78 degrees of freedom with $\alpha = 0.05$ and $t > t_c$ ($12.1026 > 1.99$), we have sufficient evidence to reject the null hypothesis.

2

Using a sample of 545 full-time workers in the USA, a researcher is interested in the question whether women are systematically underpaid compared to men. First, she estimates the average hourly wages in the sample for men and women, which are \$5.91 and \$5.09, respectively.

2.a

Do these numbers give an answer to the question of interest? Why not? How could one (at least partially) correct for this?

No, those numbers alone do not answer our question of interest. We do not know the variance and therefore the spread of our data, so the differences in average hourly wages between men and women could simply be explained by a large variance in average hourly wages.

Additionally, we need to isolate the other correlating factors that might produce this differential in outcomes. If we are concerned with systematic underpayment due to only gender, then we would have to control for other relevant variables, such as age, education, work experience, marriage, hours worked and etc.

2.b

The researcher also runs a simple regression of an individual's wage on a male dummy, equal to 1 for males and to 0 for females. This gives the following results:

Variable	Estimate	Standard error	t-ratio
Constant	5.09	0.58	8.78
Male	0.82	0.15	5.47

$N = 545$, $\hat{\sigma}^2 = 2.17$, $R^2 = 0.26$

How can you interpret the coefficient estimate of 0.82? How do you interpret the estimated intercept of 5.09?

When the individual is male, they will earn \$0.82 more on their average hourly wages compared to when the individual is not male, ceteris paribus. The average hourly wage for women is \$5.09, which is also the intercept, ceteris paribus.

2.c

How do you interpret the R^2 of 0.26?

Given that the R^2 is 0.26, about 26% of the variation in average hourly wage can be explained by the gender (male or female) of the individual, ceteris paribus.

2.d

Explain the relationship between the coefficient estimates in the table and the average wage rates of males and females.

From our table, the model can be written out as:

$$\hat{\text{wage}} = 5.09 + 0.82\hat{D}_1$$

Since we included a male dummy variable, this means if the individual is male, $d_1 = 1$, and they will earn an additional \$0.82 on their average hourly wage rate of the \$5.09 ($5.09 + 0.82 = 5.91$). However, if the individual is not male (in this case, they would then be female), then $d_1 = 0$, and their average hourly wage rate would simply be \$5.09.

In other words, this means that \$5.09 is the intercept for the average hourly wage rate and the dummy variable is the intercept shifter to increase the average hourly wage rate for males.

2.e

A student is unhappy with this model as "a female dummy is omitted from the model." Comment upon this criticism.

Males and females are included both in the male dummy variable. Since our dummy variable is binary, 1 for male and 0 for female, when we plug in in 0 for d in our equations, our model would the give us the average female hourly wage. However, if we were to switch our dummy variable to:

$$D = \begin{cases} 1 \text{ for female} \\ 0 \text{ for male} \end{cases},$$

we would get:

Variable	Estimate
Constant	5.91
Female	-0.82

which would yield in the same model where instead of males earning 0.82 higher than female, females would earn 0.82 less than males, and these, of course, are equivalent results compared to using our original male dummy variable and the intercept of 5.91 would be what average hourly wage rates male earns, ceteris paribus.

Lastly, if we added both a male *and* a female dummy variable where:

$$D_1 \begin{cases} 1 \text{ if male} \\ 0 \text{ otherwise} \end{cases} \quad \text{and } D_2 \begin{cases} 1 \text{ if female} \\ 0 \text{ otherwise} \end{cases}$$

We would run into the dummy trap where $\beta_1 = D_1 + D_2$, which violates assumption #2 of full rank.

2.f

Test, using the above results, the hypothesis than men and women have, on average, the same wage rate. State the assumptions required for this test to be valid.

$H_0 : \delta_1 = 0$
$H_1 : \delta_1 \neq 0$

$$t = \frac{d_1 - \delta_1}{s.e.(d_1)} \tag{1}$$
$$= \frac{0.82 - 0}{0.15} \tag{2}$$
$$= 5.47 \tag{3}$$

$t_c = 1.96$ with 543 degrees of freedom at $\alpha = 0.05$. Since $t > t_c$ ($5.47 > 1.96$), we have sufficient evidence to reject the null hypothesis and conclude that men and women, on average, have *different* wage rates, all else equal, at a 0.05 significance level.

The assumptions we are making is that the error terms are normally distributed, homoscedastic (constant variance), and statically independent. The normality of error terms ensures that we can rely on the results of the t-test for the dummy variable. Dummy variables meet the assumption of linearity by assumption, since they create two data points, and two points define a straight line.

2.g

Construct a 95% confidence interval for the average wage differential between males and females in the population. Subsequently, the above model is extended to include differences in age and education, by including the variables *age* (age in years) and *educ* (education level, from 1 to 5). Simultaneously, the researcher used the natural logarithm of the hourly wage rate. The results are reported in the following table:

Variable	Estimate	Standard error	t-ratio
Constant	-1.09	0.38	2.88
Male	0.13	0.03	4.47
Age	0.09	0.02	4.38
Educ	0.18	0.05	3.66

$N = 454$, $\hat{\sigma}^2 = 0.24$, $R^2 = 0.691$, $\bar{R}^2 = 0.682$

Our confidence interval can be calculated as such:

$$\text{Confidence interval} : d_1 \pm t_c s.e.(d_1) \tag{1}$$
$$0.82 \pm 1.96(0.15) \tag{2}$$
$$(0.526, 1.114) \tag{3}$$

We are 95% confident that the true value for δ_1 lies within the interval of (0.526,1.114).

The extended statistical model would be written as:

$$ln(\text{wage}) = \beta_1 + \delta_1 D_{gender} + \beta_2 age + \beta_3 educ + e_t$$

Therefore, our economic model would be:

$$ln(\hat{\text{wage}}) = -1.09 + 0.13\hat{D}_{gender} + 0.09\hat{age} + 0.18\hat{educ}$$

2.h

How do you interpret the coefficients of 0.13 for the male dummy, and 0.09 for age?

The coefficient for the male dummy would mean if the individual was a male, there would be a 0.13 percent increase in their average hourly wage rate, ceteris paribus. For age, every one year increase in age would result in a 0.09 percent increase in their average hourly wage rate, ceteris paribus.

2.i

Test the joint hypothesis that gender, age, and education do not affect a person's wage.

$$H_0 : \begin{cases} \delta_1 = 0 \\ \beta_2 = 0 \\ \beta_3 = 0 \end{cases}$$

H_1 : At least one is not true Since we know:

$$SSE = \sigma^2(T - K) = 0.24(454 - 3) = 108.24$$
$$SST = \frac{SSE}{1 - R^2} = \frac{108.24}{1 - 0.691} = 350.291$$
$$SSR = SST - SSE = 350.291 - 108.24 = 242.051$$

Therefore,

$$F = \frac{\frac{SSR}{K-1}}{\frac{SSE}{T-K}} \tag{1}$$
$$= \frac{\frac{242.051}{3-1}}{\frac{108.24}{454-3}} \tag{2}$$
$$= 504.273 \tag{3}$$

$F_c = 2.6247$ for $\alpha = 0.05$ with 3 hypotheses at 451 degrees of freedom. Since $F > F_c$ ($504.273 > 2.6247$), we have sufficient evidence to reject the null hypothesis at a 0.05 significance level.

2.j

A student is unhappy with this model as the effect of education is rather restrictive. Can you explain this criticism? How could the model be extended to meet the above criticism? How can you test whether the extension has been useful?.

The criticism may be valid since the model treats education at discrete units within 5 different categories. However, even within these categories, there could be explanatory variations. For instance, by categorizing a undergraduate who dropped out as a freshman and another undergraduate who dropped out as a senior as the same level of education may indeed be restrictive.

Furthermore, even within these categories, other factors may be useful to discern, such as the type of school (i.e. home-school, private school, religious school) and for colleges and universities it may be useful to distinguish between Ivy League schools, top state schools, and community colleges.

There are assumptions that the effect of education is the same between males and females as well as remaining constant for all ages. For instance, education may affect women average hourly wage rates more or less than males. Another instance regarding age and education would be if a male had a PhD at the age of 18 compared to a male with a PhD at the age of 30 would presumably have different rates of effect of education depending on age (it would be hypothesized that the PhD attained at age 18 will affect the average wage rate more than the PhD attained at age 30). In other words, education could be correlated with age and gender. Therefore, we would need to include an education slope-shifter for age and gender.

To test, we would need to again set up an F-test and see if these regressions with an education slope-shifter is significant.

2.k

The researcher re-estimates the above model including age^2 as an additional regressor. The t-value on this new variable is -1.14, while $R^2 = 0.699$ and $\bar{R}^2$ increases to 0.683. Could you give a reason why the inclusion of age^2 might be appropriate?

The inclusion of a variable squared is appropriate if that variable exhibits diminishing returns. For age, in the context of hourly wage rates, this may make sense, since presumably productivity increases up until a certain age when the individual accumulates skills, knowledge, and connections but then productivity decreases due to old age with withering physical and mental abilities.

2.l

Could you retain this new variable given the the R^2 and the $\bar{R}^2$ measures? Would you retain age^2 given its t-value? Explain this apparent conflict in conclusions.

We would not be able to justify retaining this new variable using R^2 since by adding a new variable, the original model and this model would have different number of regressors and therefore using R^2 to compare them would be invalid. However, since $\bar{R}^2$ is higher and we can still compare models using $\bar{R}^2$, this gives us reason to retain this new variable.

On the other hand, the t-value of age^2 is -1.14, which suggests age^2 is not significant to our model. However, there is no conflict with this conclusion because $\bar{R}^2$ only increases by 0.001 with the inclusion of the age^2 variable, which the t-value confirms is not a significant amount of explanation for the variation in wages by the variation of age^2.

3

Consider the aggregate production function $q_t = A \times K_t^{\beta_2} \times L_t^{\beta_3}$, where the industry output q is assumed to be a function of the level of the inputs capital K and labor L. You are considering to estimate the production function using the following model

$$ln(q_t) = \beta_1 + \beta_2 ln(K_t) + \beta_3 ln(L_t) + e_t$$

which also can be written as

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + e$$

3.a

What quantities appear in y and X when this model is written in the matrix algebra notation? The data you are using can be summarized by:

$$X'X = \begin{bmatrix} 100 & 123 & 96 \\ 123 & 252 & 125 \\ 96 & 125 & 167 \end{bmatrix}$$

$$X'X^{-1} = \begin{bmatrix} 0.0353 & -0.0114 & -0.0118 \\ -0.0114 & 0.0100 & -0.0009 \\ -0.0118 & -0.0009 & 0.0134 \end{bmatrix}$$

$$X'y = \begin{bmatrix} 460 \\ 810 \\ 615 \end{bmatrix}$$

$$y'y = \begin{bmatrix} 3924 \end{bmatrix}$$

$$y = \begin{bmatrix} ln(q_1) \\ ln(q_2) \\ \vdots \\ ln(q_{100}) \end{bmatrix} \text{ and } X = \begin{bmatrix} 1 & ln(K_1) & ln(L_1) \\ 1 & ln(K_2) & ln(L_2) \\ \vdots & \vdots & \vdots \\ 1 & ln(K_{100}) & ln(L_{100}) \end{bmatrix}$$

3.b

Complete the following table.

Variable	Parameter	Estimate	Std. error	t-value
$Intercept$	β_1	-0.2264	0.568	-0.3986
$ln(K)$	β_2	2.2801	0.3022	7.545
$ln(L)$	β_3	2.1061	0.3504	6.011
T=100	$R^2 = 0.5100$		$\hat{\sigma}^2 = 9.1341$	

3.c

Interpret the parameters for capital and labor.

For every additional one percent increase in capital, industry output increases by 2.2801 percent, ceteris paribus. For every additional one percent increase in labor, industry output increases by 2.1061 percent, ceteris paribus.

3.d

If capital is 4 units and labor is 2 units, what is the production? How confident are you about this estimate?

$$\begin{aligned} ln(q_t) &= \beta_1 + \beta_2 ln(K_t) + \beta_3 ln(L_t) + e_t \\ &= -0.2264 + 2.2801(ln(4)) + 2.1061(ln(2)) \\ ln(q_t) &= 4.3956 \\ q_t &= e^{4.3956} e^{\sigma^2/2} \\ &= e^{4.3956} e^{(9.1341/2)} \\ &= 7806 \end{aligned}$$

$$x_0 = \begin{bmatrix} 1 \\ ln(4) \\ ln(2) \end{bmatrix}$$

$$\text{Confidence interval: } \hat{y}_0 \pm t_c \sqrt{\hat{\sigma}^2 (x_0' (X'X)^{-1} x_0) + 1} \tag{1}$$

$$4.3956 \pm 1.9847 \sqrt{1.1030} \tag{2}$$

$$4.3956 \pm 2.0844 \tag{3}$$

$$(e^{4.3956 - 2.0844} e^{(9.1341/2)}, e^{4.3956 + 2.0844} e^{(9.1341/2)}) \tag{4}$$

$$(971, 62759) \tag{5}$$

I am 95% confident that the estimate is within in interval $(971, 62759)$.

3.e

We say we have increasing returns to scale whenever we double (for example) the inputs (capital and labor in this case), the output more than double; that is $\beta_2 + \beta_3 > 1$. We have decreasing returns to scale whenever we double the inputs, the output less than double; that is $\beta_2 + \beta_3 < 1$. We have constant returns to scale whenever we double the inputs, the output double; that is $\beta_2 + \beta_3 = 1$. By just examining the results, what returns to scale does the production exhibit?

Since $\beta_2 + \beta_3 > 1$ ($2.2801 + 2.1061 = 4.3862 > 1$), we can see our result suggests an increasing returns to scale from capital and labor into the output. However, this result is inconclusive since there could be sampling error, therefore, we would have to formally conduct a statistical test in order to be more confident.

3.f

Formally test whether your answer is consistent with the sample evidence.

$H_0 : \beta_2 + \beta_3 \leq 1$
$H_1 : \beta_2 + \beta_3 > 1$

$$t = \frac{(b_2 + b_3) - (\beta_2 + \beta_3)}{\sqrt{var(b_2, b_3)}} \tag{1}$$
$$= \frac{2.2801 + 2.1061 - 1}{\sqrt{0.0913 + 0.1228 + 2(-0.0084)}} \tag{2}$$
$$= \frac{3.3862}{0.4442} \tag{3}$$
$$= 7.6234 \tag{4}$$

t_c for 97 degrees of freedom at $\alpha = 0.05$ is 1.97. Since $t > t_c$ ($7.6234 > 1.97$) we have sufficient evidence to reject the null hypothesis at a 0.05 significance level.

3.g

Estimate the model by imposing the resulting returns to scale you fail to reject in (f).

Our given equation is:

$$ln(q_t) = \beta_1 + \beta_2 ln(K_t) + \beta_3 ln(L_t) + e_t$$

To minimize our sum of squared errors, we must:

$$\min S = \sum e_t^2 = \sum (ln(q_t) - \beta_1 - \beta_2 ln(K_t) - \beta_3 ln(L_t))^2 \text{ s.t. } \beta_2 + \beta_3 = 1 \tag{1}$$

Therefore, we will use the following equation:

$$b^* = b + (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(r - Rb) \tag{1}$$

With the following:

$$R = \begin{bmatrix} 0 & 1 & 1 \end{bmatrix} \text{ and } r = \begin{bmatrix} 1 \end{bmatrix}$$

Which results in:

$$b^* = \begin{bmatrix} 3.4084 \\ 0.8578 \\ 0.1422 \end{bmatrix}$$

and our economic model is,

$$ln(\hat{q}) = 3.4084 + 0.8578 ln(\hat{K}) + 0.1422 ln(\hat{L})$$

4

Consider the demand for beer model:

$$ln(q_t) = \beta_1 + \beta_2 ln(p_{Bt}) + \beta_3 ln(p_{Lt}) + \beta_4 ln(p_{Rt}) + \beta_5 ln(m_t) + e_t$$

where

$$\beta_2 + \beta_3 + \beta_4 + \beta_5 = 0$$
$$\beta_3 = \beta_4$$
$$\beta_5 = 1$$

Use these three restrictions on the elements of β to eliminate the parameters β_2, β_4, and β_5 from the model and show that the resulting model can be written as:

$$ln(\frac{q_t p_{Bt}}{m_t}) = \beta_1 + \beta_3 ln(\frac{p_{Lt} p_{Rt}}{p_{Bt}^2}) + e_t$$

$$ln(q_t) = \beta_1 + \beta_2 ln(p_{Bt}) + \beta_3 ln(p_{Lt}) + \beta_4 ln(p_{Rt}) + \beta_5 ln(m_t) + e_t \quad (1)$$
$$\beta_2 + \beta_3 + \beta_4 + \beta_5 = 0 \quad (2)$$
$$\beta_3 = \beta_4 \quad (3)$$
$$\beta_5 = 1 \quad (4)$$
$$\beta_2 + \beta_3 + \beta_3 + 1 = 0 \quad (5)$$
$$\beta_2 = -1 - 2\beta_3 \quad (6)$$
$$ln(q_t) = \beta_1 + (-1 - 2\beta_3) ln(p_{Bt}) + \beta_3 ln(p_{Lt}) + \beta_3 ln(p_{Rt}) + ln(m_t) + e_t \quad (7)$$
$$ln(q_t) = \beta_1 - ln(p_{Bt}) - 2\beta_3 ln(p_{bt}) + \beta_3 ln(p_{Lt}) + \beta_3 ln(p_{Rt}) + ln(m_t) + e_t \quad (8)$$
$$ln(q_t) + ln(p_{Bt}) - ln(m_t) = \beta_1 - 2\beta_3 ln(p_{Bt}) + \beta_3 ln(p_{Lt}) + \beta_3 ln(p_{Rt}) + e_t \quad (9)$$
$$ln(q_t) + ln(p_{Bt}) - ln(m_t) = \beta_1 + \beta_3(-2ln(p_{Bt}) + ln(p_{Lt}) + ln(p_{Rt})) + e_t \quad (10)$$
$$ln(q_t) + ln(p_{Bt}) - ln(m_t) = \beta_1 + \beta_3(-ln(p_{Bt}^2) + ln(p_{Lt}) + ln(p_{Rt})) + e_t \quad (11)$$
$$ln(\frac{q_t p_{Bt}}{m_t}) = \beta_1 + \beta_3 ln(\frac{p_{Lt} p_{Rt}}{p_{Bt}^2}) + e_t \quad (12)$$

1) Given
2) Given
3) Given
4) Given
5) Substitution from (3) and (4) into (2)
6) Rearrange to find β_2 from (5)
7) Substitution from (3), (4), and (6) into (1)
8) Distribution
9) Move $ln(p_{Bt})$ and $ln(m_t)$ to the left-hand side
10) Factor out β_3
11) Exponential rule for logarithms
12) Addition and subtraction rules for logarithms of different bases

5

Provide a description of the term paper/project: The topic, the data, and the model if you already have an idea.

For my term project, I will try to determine if there is a relationship between COVID-19 *cases* per capita and economic freedom (EF), inequality (GINI coefficient), GDP per capital, and perceived corruption (PC). Additionally, I will model COVID-19 *deaths* per capita with the same explanatory variables. The objective will be to see if there is any relationship between economic freedom and the wealth of countries with the case and death rates of COVID-19. The statistical models will look something like the following:

$$\text{cases}_t = \beta_1 + \beta_2(\text{EF})_t + \beta_3\text{GDP}_t + \beta_3\text{GINI}_t + \beta_4(\text{PC})_t + e_t$$
$$\text{deaths}_t = \alpha_1 + \alpha_2(\text{EF})_t + \alpha_3\text{GDP}_t + \alpha_3\text{GINI}_t + \alpha_4(\text{PC})_t + e_t$$

These variables are correlated but they may not necessarily be perfectly linearly correlated. For instance, economic freedom may raise GDP per capita, however, there may be countries with low GDP that have initiated economic reforms. Likewise, there may be a correlation between economic freedom and the GINI coefficient since economic freedom allows business-owners and entrepreneurs to earn more money, however, the country may have a redistribution system that lowers income inequality for their citizens. I have included perceived corruption as a variable because citizens that perceives its

government as corrupt may not be as willing to follow government mandates that may or may not help with the spread of COVID-19, which would, of course, impact COVID-19 case and death rates. *This is still a tentative model and I may add more or change these variables as I move forward with this project*.

I will be obtaining the 2020 Economic Freedom Index from the Fraser Institute, the GDP per capita and the GINI coefficients from the World Bank, and perceived corruption from Our World Data.

I may create an OECD dummy variable for OECD countries

$$D = \begin{cases} 1 \text{ if OECD country} \\ 0 \text{ if otherwise} \end{cases}$$

to see the effect of being an OECD country. However, this would also be correlated with all of the variables included in our model. I may also create another dummy variable for the continent these countries are located in to see if that is another factor. The other variables I may be considering are: literacy rates, the Human Development Index, working hours, and taxation, although I may use some of these as instrumental variables as well. If you could suggest any other relevant variables and data sets, I would greatly appreciate it.

From these initial conditions, some problems I may encounter will be multicollinearity, since, as I mentioned above, that the variables I chose may be highly correlated and would most likely affect one another. Another problem could be omission bias. Since governments may categorize things differently and collect data on different years, I can only analyze countries with complete data sets for all the variables I listed, which I suspect will narrow my data set and cause some factor omission bias. It is also the case that wealthier, more developed countries will collect data, which will potentially further bias my analysis. I will have to more formally test these issues and see if I can correct them in order for my model to be accurate.