**Abstract**

Knowing the relationship between the productivity of a worker's labor and how long they work is essential to understanding whether putting more hours into working is an efficient use of labor and capital. This study investigated productivity ($/hour) and average annual hours of work per worker from 67 countries. After looking at two simple linear model, one with transformed variables and one without, we found that the simple linear model of the natural log of productivity versus average annual hours of work per worker does a slightly better job of representing the relationship between the data. The sampling variability assumptions, multiple R squared values, model summaries, and correlation values in this study go deeper into the reasoning behind the model selection. In the end, there is a negative linear relationship between the natural log of productivity and average annual hours of work per worker, meaning when people work more, they are less productive, on average.

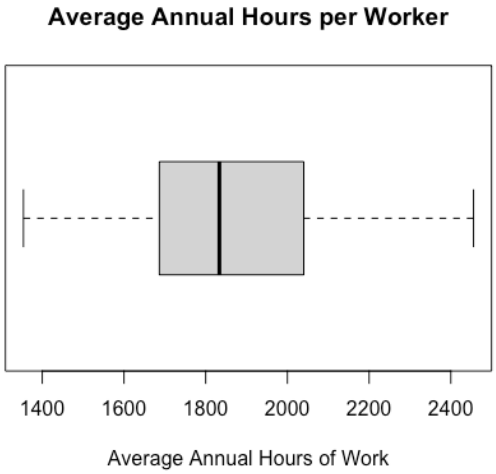
The goal of this project is to see whether there is a linear relationship between productivity ($/hour) and average annual hours of work per worker. Understanding and identifying the relationship the relationship between the productivity of a worker's labor and their hours of work is important to knowing if working more hours is an efficient use of labor, capital, and resources. A positive linear relationship between these two variables would mean that when people work more, they are more productive, and a negative linear relationship would mean when people work more, they are less productive.

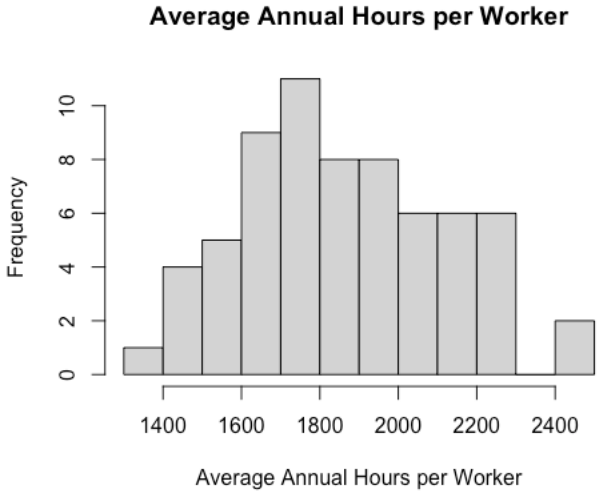
We initially found the data set from ourworldindata.org, which is a nonprofit with a team of researchers who have the goal to use data to help solve world problems (Our World in Data, 2017). The researchers from ourworldindata.org obtained the sources for productivity and average annual working hours per worker from the University of Groningen's Growth and Development Center in the Netherlands. The data is updated yearly, but the updates are not equal across all countries, meaning some countries are not updated for stretches of years.

## Data

To be clear, the explanatory variable that will be explored via the simple linear models is the average annual hours of work per worker and the response variable is productivity because productivity is dependent on how many hours workers spend working. Looking at the raw data, "Averagehrs" average annual hours of work per worker while "Productivity" is the productivity ($/hour) of a worker. According to the University of Groningen, average annual hours of work per worker is calculated by adding up all the hours worked by all workers in a particular country and dividing by the number of employed persons. Productivity is calculated by taking total output-side real GDP at chained PPPs (in millions of 2011US$) and dividing that by the number of persons engaged (in millions) multiplied by the average annual hours worked by employed people (Feenstra et al., 2015). All data analyzed in this study is selected from 2017.

Numerical (explanatory) variable #1: Average Annual Hours of Work per Worker




```
summary(current$Averagehrs)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 1354    1688    1833    1861    2037    2456     176
```

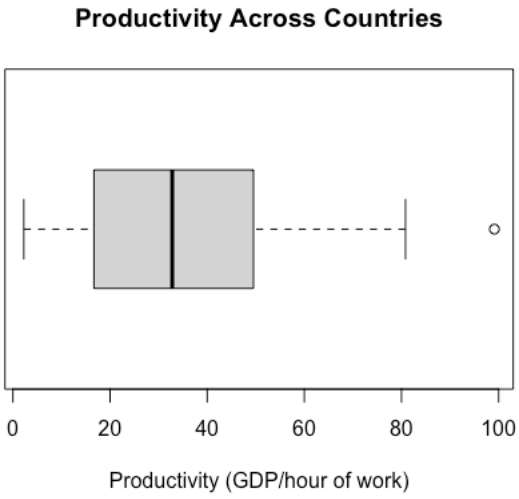
Standard deviation:

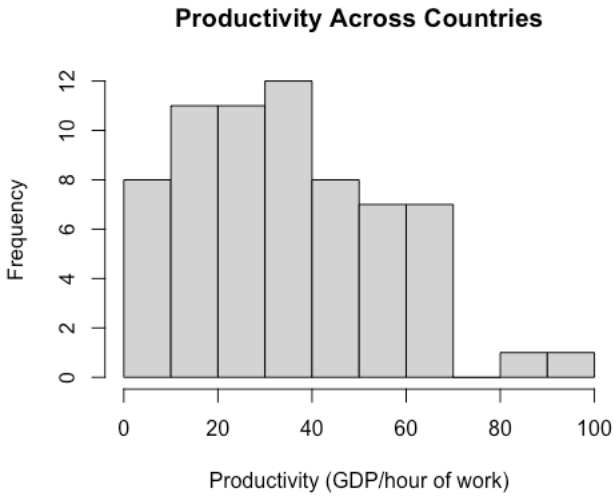
```
[1] 259.7861
```

The shape of the data looks to be slightly skewed right since the peak of the histogram veers towards the left a bit. There are no outliers by looking at the boxplot. The center of the data is a median of 1833 average annual hours per worker and a mean of 1861 average annual hours per

worker. This helps support the suggestion that the distribution is skewed right because the mean is larger than the median. Looking at the summary statistics, the spread goes from a minimum of 1354 hours to a maximum of 2456 hours.

Numerical (response) variable #2: Productivity



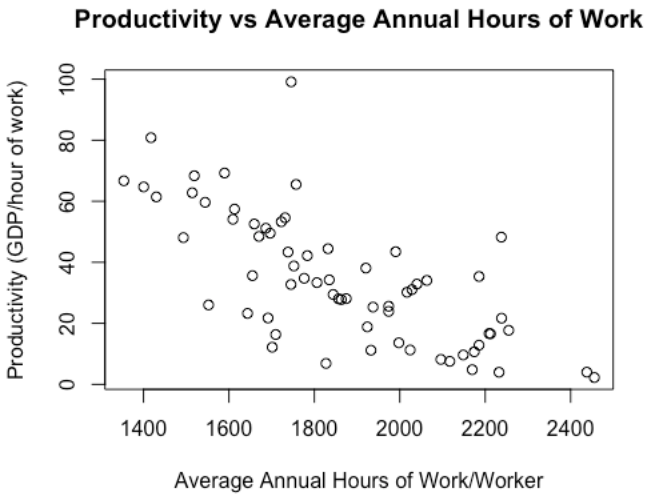




```
summary(current$Productivity)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
2.244  16.935  32.802  34.646  49.270  99.129     176
```

Standard deviation:

```
[1] 21.22618
```

The shape of the distribution looks to be skewed right slightly because the peak of the histogram veers a bit towards the left. By looking at the boxplot, there is one high outlier at a productivity of 99.129 $/hour of work. The center of the spread is a median of 32.802 $/hour of work and a mean of 34.646 $/ hour of work. Because the mean is slightly larger than the median, this helps support the idea that the distribution is skewed right. Looking at the summary statistics, the spread goes from a minimum of 2.244 $/hour of work to a maximum of 99.129 $/hour of work.




Correlation: `[1] -0.7121495`

There looks to be a negative and moderately strong linear relationship between the average annual hours of work per worker and productivity.

## Summary of Statistical Findings

Model #1: Untransformed Variables (Productivity~Averagehrs):

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 142.92702   13.47015  10.611 9.69e-16 ***
Averagehrs   -0.05819    0.00717  -8.115 2.02e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.02 on 64 degrees of freedom
  (176 observations deleted due to missingness)
Multiple R-squared:  0.5072,    Adjusted R-squared:  0.4995
F-statistic: 65.86 on 1 and 64 DF,  p-value: 2.019e-11
```

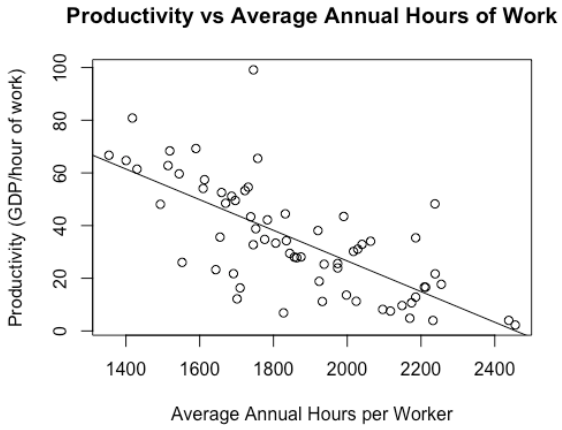
Fitted model equation:

$\mu$[Productivity|Averagehrs] = 142.92702 – 0.05819(Averagehrs)

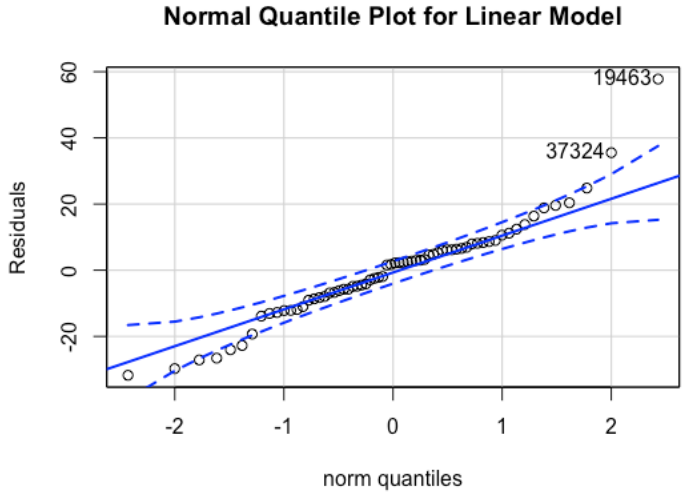
For every 1 hour increase in average annual hours of work per worker, productivity decreases by 0.05819 $/hour.

95% CI: (-0.07251094, -0.04386341). We are 95% confidence that the slope of the regression line of the linear relationship between average annual hours of work per worker and productivity lies between -0.07251094 and -0.04386341 $/hour. The confidence interval does not contain zero, indicating that the model is useful for this data set and that a negative slope exists between average annual hours of work per worker and productivity.

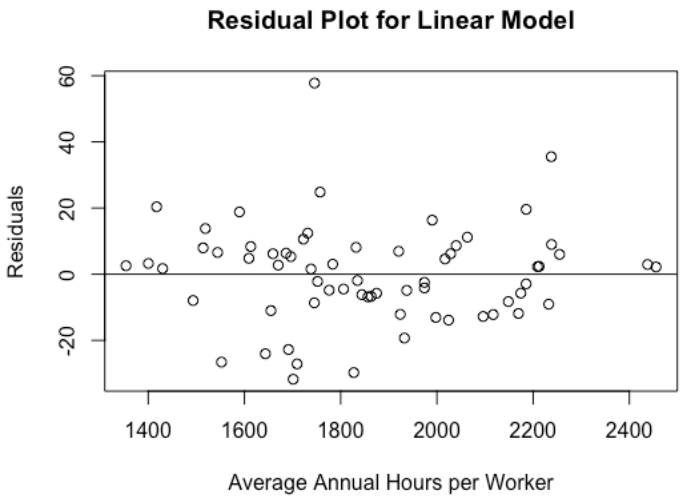
| **Coefficient** | **Estimate** | **Null Hypothesis** | **Results** |
|---|---|---|---|
| Intercept | 142.92702 | H0: B0 = 0 (The predicted productivity when average annual working hours per worker is 0 equals 0) | Test statistic = 10.611<br><br>P-Value = <9.69e^-16<br><br>P-value < $\alpha$→ Statistically significant → reject null hypothesis |
| Averagehrs<br><br>(Slope) | -0.05819 | Ha: B1 = 0<br><br>(Predicted decrease in productivity when average annual working hours per worker increases by 1 hour is equal to 0 or does not change) | Test statistic = -8.115<br><br>P-Value = 2.02e^-11<br><br>P-Value < $\alpha$ → statistically significant → reject null hypothesis |




The Sampling Variability Assumptions (SVAs) were checked for this model. The main SVAs were considered: first, that the error terms are normally distributed and second, that there is equal variance with a linear relationship.




There seem to be a handful of data points that lie outside the dashed boundary lines, especially towards the very low and very high values of norm quantiles. However, because the data set has a sample size greater than 40, we can safely proceed with our statistical methods.




Looking at the residual plot, the error terms seem to have equal variances because the data points are scattered equally above the horizontal line at 0. There also is no systematic pattern with the data points, meaning the relationship is linear. The model is appropriate for the data set.

Conclusion:

Based on the scatterplot, statistics, model summary output, and SVAs, there seems to be a linear relationship between productivity and average annual hours of work per worker. The scatterplot has a moderately strong correlation with negative direction. The model summary output indicates that both the intercept and slope coefficients were not zero since the p-value was less than 0.05 for both, meaning the simple linear model is useful for the untransformed variables. The multiple R squared value for this model is 0.5072, meaning that 50.72% of the variability in productivity is explained by the relationship between productivity and average annual hours of work per worker. The SVAs are satisfied for both normality and equal variance/linearity. Therefore, a linear relationship can be used to describe productivity vs annual hours of work per worker. That being said, a logarithmic transformation could be explored to see if an even closer relationship between the variables exists.

```
> predict(linearmodel, new = data.frame(Averagehrs = 1600))
       1
49.82754
> predict(linearmodel, new = data.frame(Averagehrs = 1600), interval = "confidence")
       fit      lwr      upr
1 49.82754 44.57369 55.08139
> predict(linearmodel, new = data.frame(Averagehrs = 1600), interval = "prediction")
       fit      lwr      upr
1 49.82754 19.37039 80.28469
```

The estimated mean productivity for a country with average annual working hours per worker of 1600 hours is 49.82754 $/hour.

**Confidence interval:** The estimated mean productivity of all countries with average annual working hours per worker of 1600 hours is 49.82754 $/hour. 95% CI: (44.57369, 55.08139)

**Prediction interval:** The estimated mean productivity of a country with average annual working hours per worker of 1600 hours is 49.82754 $/hour. 95% PI: (19.37039, 80.28469)

Due to the slightly right skewed nature of the variables in the data exploration section, a natural log transformation was performed with respect to productivity.

Model #2: Transformed Variables (Log-Linear → ln (Productivity)~Averagehrs):

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         7.4913872  0.5046671  14.844  < 2e-16 ***
current$Averagehrs -0.0022556  0.0002686  -8.397 6.46e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5626 on 64 degrees of freedom
  (176 observations deleted due to missingness)
Multiple R-squared:  0.5242,    Adjusted R-squared:  0.5167
F-statistic:  70.5 on 1 and 64 DF,  p-value: 6.458e-12
```

Fitted model equations:

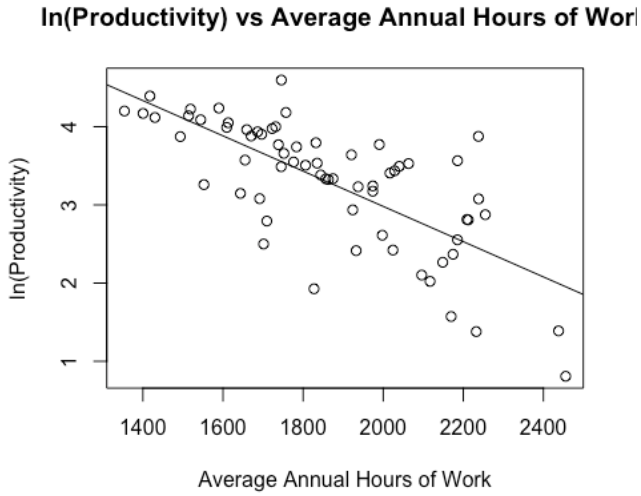
Mu [ln (Productivity) | Averagehrs] =7.4913872-0.0022556(Averagehrs)

Median [Productivity | Averagehrs] = e^(7.4913872-0.002256(Averagehrs))

When the number of average annual hours of work per worker increases by 1 hour, the median productivity decreases by a factor of .0033% (e^-0.0022556 = 0.9977)

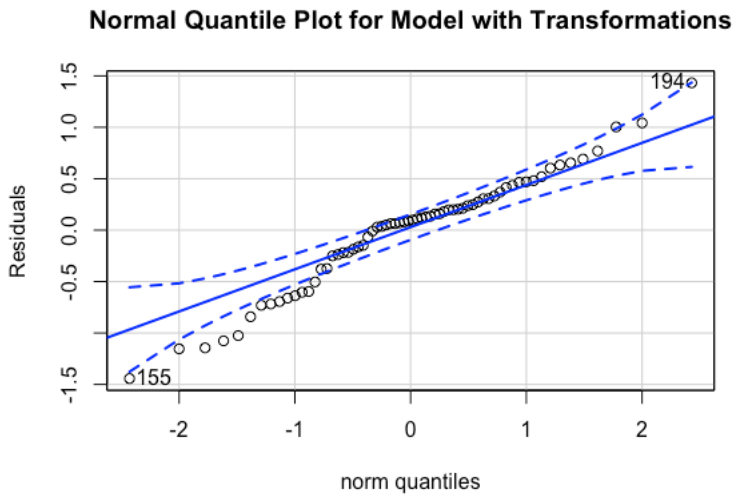
95% CI: ( -0.002792205, -0.001718908). We are 95% confidence that the slope of the log-linear relationship between ln(productivity) and average annual hours of work per worker lies between -0.002792205 -0.001718908 $/hour. The confidence interval does not contain zero, indicating that the model is useful for this data set and that a negative slope exists between average annual hours of work per worker and ln(productivity).

| Coefficient | Estimate | Null Hypothesis | Results |
|---|---|---|---|
| ln. Intercept | 7.4913872 | H0: B0 = 0 (The median productivity when average annual hours of work is 0 equals 0) | Test statistic = 14.844<br><br>P-Value = <2e^-16<br><br>P-value < $\alpha$→ Statistically significant → reject null hypothesis |
| Averagehrs (Slope) | -0.002256 | Ha: B1 = 0<br><br>(The slope for this coefficient is 0, representing a horizontal line) | Test statistic = -8.397<br><br>P-Value = 6.46e^-12<br><br>P-Value < $\alpha$ → statistically significant → reject null hypothesis |

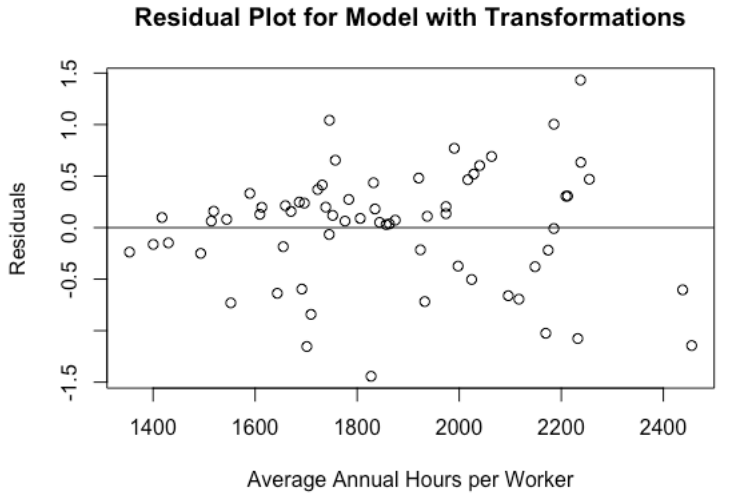



Correlation: `[1] -0.7239963`

The same SVAs were checked for the log-linear model (ln (Productivity)~Averagehrs): 1) the error terms are normally distributed and, 2) there is equal variance.




There are a few more data points that lie outside the dashed boundary lines compared to the linear model, especially towards the very low values of norm quantiles. However, because the data set has a sample size greater than 40, we can safely proceed with our statistical methods.




Looking at the residual plot, the error terms seem to have equal variances because the data points are scattered more or less equally above the horizontal line at 0. There also does not seem to be any systematic pattern with the data points. This model is also appropriate for the data set.

Conclusion:

Based on the scatterplot, statistics, model summary output, and SVAs, there seems to be a linear relationship between the natural log of productivity ($/hour) and average annual hours of work per worker. The scatterplot has a moderately strong correlation of-0.7239963 with negative direction. The model summary output indicates that the null hypothesis for both the intercept and slope coefficients could be rejected since the p-value was less than 0.05 for both, meaning the transformed linear model of ln (Productivity)~Averagehrs is useful. The multiple R squared value for this model is 0.5242, meaning that 52.42% of the variability in ln (Productivity) is explained by the relationship between ln (productivity) and average annual hours of work per worker. The SVAs are for the most part satisfied just like with the simple linear model with untransformed variables.

```
> exp(predict(loglinearmodel, new = data.frame(Averagehrs = 1600)))
       1
48.54527
> exp(predict(loglinearmodel, new = data.frame(Averagehrs = 1600), interval = "confidence"))
       fit      lwr      upr
1 48.54527 39.87136 59.10616
> exp(predict(loglinearmodel, new = data.frame(Averagehrs= 1600), interval = "prediction"))
       fit      lwr      upr
1 48.54527 15.50871 151.9561
```

The estimated median productivity for a country with average annual working hours per worker of 1600 hours is 48.54527 $/hour.

**Prediction interval:** The estimated median productivity of a country with average annual working hours per worker of 1600 hours is 48.54527 $/hour. 95% PI: (15.50871, 151.9561)

**Confidence interval:** The estimated median productivity of all countries with average annual working hours per worker of 1600 hours is 48.54527 $/hour. 95% CI: (39.87136, 59.10616)

**Comparison of Models**

Although both models were quite similar in terms of best fit, the best fitting model was determined to be a simple linear model of a log-linear transformation of productivity ($/hour) vs average annual working hours per worker (ln (Productivity)~Averagehrs). The multiple R squared values, correlation values, and SVAs were compared to determine which model best fit the dataset. First, the multiple R squared value for the untransformed variables was 0.5072. The multiple R squared value for the simple linear model of the log-linear transformed variables was 0.5242. The larger R squared value for the log-linear model indicates that more of the variability in ln (productivity) is explained by linear relationship between ln (Productivity) and Averagehrs rather than the simple linear regression line from the untransformed variables. Additionally, the correlation value for the untransformed simple linear model is -0.7121495 while the correlation

for the log-linear simple linear model is  -0.7121495. Because -0.7239963 is closer to -1 than -0.7121495, the natural log of productivity has a closer relationship to average annual hours of work per worker than the untransformed productivity variable. Lastly, the SVAs were the virtually the same between the two models, therefore, there is no difference between the models based on whether the SVAs were satisfied for each model or not.

**Bibliography**

Feenstra, Robert C., Robert Inklaar and Marcel P. Timmer (2015), "The Next Generation of the Penn World Table" *American Economic Review*, 105(10), 3150-3182

Our World in Data. (2017). *Productivity vs. Annual hours of work*. Our World in Data. https://ourworldindata.org/grapher/productivity-vs-annual-hours-worked