Statistics Project 2

David Rodden, Michael McIntosh, Khang Tran, Jeffrey Lazatin

April 23, 2017

```
sheet = read_excel("GDP.xlsx")
attach(sheet)
#Country, GDP, LEB, NLLEB, NLGDP
```

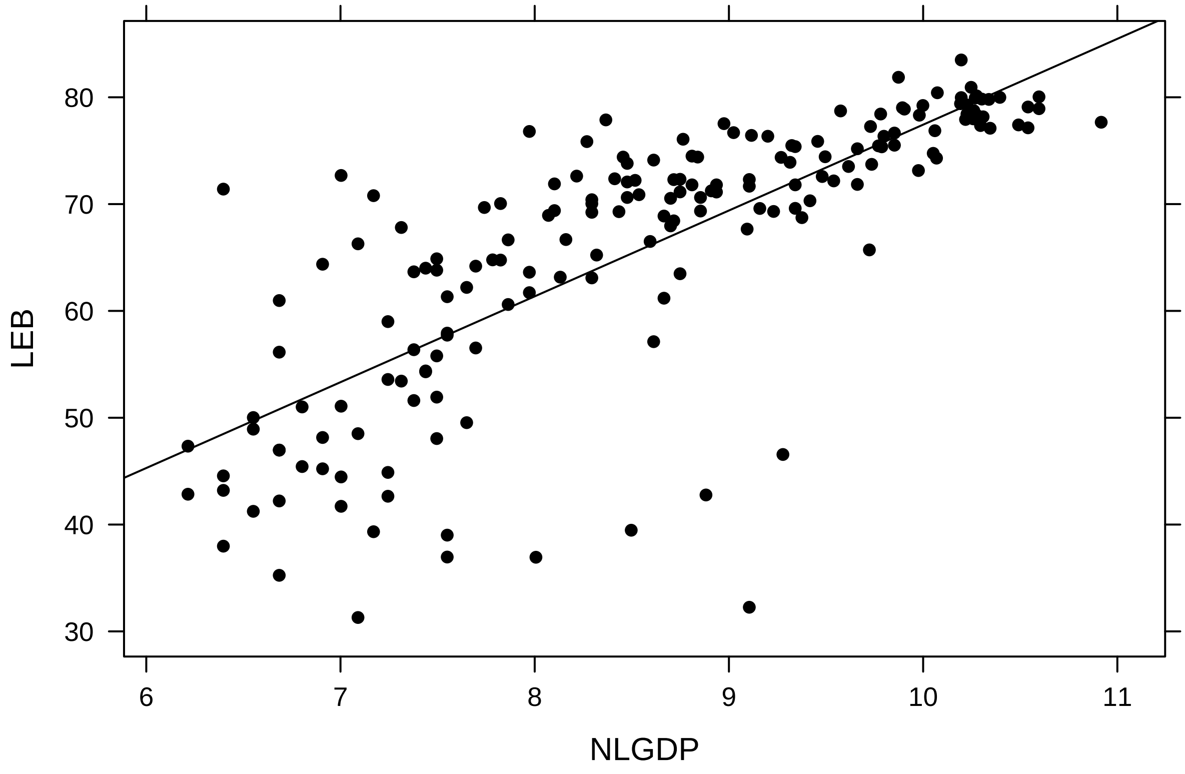
Introduction

We are investigating the relationship between Life Expectancy of a country based upon its GDP. The data we are using was collected in 2003 from the CIA Factbook; the data is across 180 countries. The investigation is looking to see if there is a positive correlation between life expectancy (LEB) and a country's GDP (NLGDP); using GDP as a predictor. In order to normalize the data, we use the natural log of the GDP. The data considers a country's life expectancy at birth and the GDP per capita (PPP). The data was collected from official reports that each nation compiles. We found the data from *Index Mundi*, who pulled from the CIA Factbook. The data is a sample of the world's countries, and is an observational study.

$H_0 : \rho = 0$ vs $H_a : \rho \neq 0$

Summary & Visualization

```
favs = favstats(LEB ~ NLGDP)
anova.b = anova(lm(LEB ~ NLGDP))
xyplot(LEB ~ NLGDP, type = c("p", "r"), pch=16, col="black")
```



The scatter plot shows that there are a few outliers which will influence the overall model. The outliers will impact the regression which we use to model and predict, based upon the data. There is a slight departure from linearity, a subtle curve in the data, but still increasing overall. The data does possesses changing variability, a fanning trend, wide to narrow from left to right. There appears to be a positive linear association between the two quantitative variables, LEB ~ NLGDP.

```
#five number summary
sum.sheet = summary(sheet); sum.sheet
```

```
##    Country               GDP             LEB             NLGDP
##  Length:180         Min.   :  500   Min.   :31.30   Min.   : 6.215
##  Class :character   1st Qu.: 1800   1st Qu.:57.87   1st Qu.: 7.496
##  Mode  :character   Median : 5650   Median :70.47   Median : 8.639
##                     Mean   :10051   Mean   :65.95   Mean   : 8.571
##                     3rd Qu.:15700   3rd Qu.:75.86   3rd Qu.: 9.661
##                     Max.   :55100   Max.   :83.49   Max.   :10.917
```

```
#standard deviation
gdp.sd = sd(sheet$GDP); gdp.sd
```

```
## [1] 10757.43
```

```
leb.sd = sd(sheet$LEB); leb.sd
```

```
## [1] 12.75888
```

```
nlgdp.sd = sd(sheet$NLGDP); nlgdp.sd
```

```
## [1] 1.223437
```

The sample size is 180 countries. The means for GDP, Life expectancy at birth, and Natural log of GDP are Mean :10051 , Mean :65.95 , Mean : 8.571 , respectively.

The standard deviation for GDP, Life expectancy, at birth and Natural log of GDP are 1.0757429×10^4, 12.7588816, 1.2234365 respectively.

Correlation Test

$H_0 : \rho = 0$ vs $H_a : \rho \neq 0$

```
corr = cor.test(NLGDP, LEB)
corr.p.value = corr$p.value
```

The p-value $\approx 1.1930935 \times 10^{-36}$. As the p-value is very small, we reject the null hypothesis in favor of the alternative hypothesis that there is a non-zero correlation between Life expectancy at birth and the natural log of GDP.

Regression

```
mod = lm(LEB ~ NLGDP)
y.hat = makeFun(mod); y.hat
```

```
## function (NLGDP, ..., transformation = function (x)
## x)
## return(transformation(predict(model, newdata = data.frame(NLGDP = NLGDP),
##     ...)))
## <environment: 0x000000000db62140>
## attr(,"coefficients")
## (Intercept)       NLGDP
##   -2.920003    8.035419
```

```
mod.prediction = predict(mod); head(mod.prediction)
```

```
##        1        2        3        4        5        6 
## 47.01698 47.01698 48.48201 48.48201 48.48201 48.48201
```

```
mod.resid = resid(mod); head(mod.resid)
```

```
##           1           2           3           4           5           6 
##  -4.1769791   0.3230209  -3.9220093  -5.2820093  22.9179907 -10.5020093
```

```
mod.coef = unname(coef(mod)); mod.coef
```

```
## [1] -2.920003  8.035419
```

Using the regression model, $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x = -2.9200029 + 8.0354193x$ We predict a value of $\hat{y} \approx 47.0169791$ using x $\approx$ 6.2146081 with a residual ≈ -4.1769791.

```
sd.verify = max(na.omit(favs$sd)) > 2 * min(na.omit(favs$sd)); sd.verify
```

```
## [1] TRUE
```

```
n.verify = max(na.omit(favs$n)) >= 30; n.verify
```

```
## [1] FALSE
```

Verifying conditions, we see that our data fails the standard deviation check for ANOVA.

Despite the conditions for ANOVA testing not meeting their requirements, we will proceed with the ANOVA test anyhow.

```
anova(mod)
```

```
## Analysis of Variance Table
## 
## Response: LEB
##            Df Sum Sq Mean Sq F value    Pr(>F)    
## NLGDP       1  17300 17299.5  260.08 < 2.2e-16 ***
## Residuals 178  11840    66.5                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
r = cor(LEB ~ NLGDP); r
```

```
## [1] 0.7705084
```

```
n = length(NLGDP)
t = r * sqrt((n - 2) / (1 - r ^ 2)); t
```

```
## [1] 16.12705
```

```
p.value = 2 * pt(-abs(t), df = n - 2)
```

Our evidence allows us to reject the null hypothesis in favor of the alternative hypothesis. We have a t-statistic value . Given the t statistic, ≈ 16.1270513, and the p-value, $\approx 1.1930935 \times 10^{-36}$, we reject the null hypothesis with evidence for the alternative hypothesis. Our evidence from the data collected implies that there is a correlation between a country's life expectancy at birth and GDP.

Teamwork