Predicting Labor Force Participation Rate

Christine Legge (19778125), Jody Li (32498115), Estella Dong (51006147), Darren Matis (94897071)

Abstract

Our goal is to create a prediction equation for the labor force participation rate among 15-64 year olds in a specified country. The response variable is the country's labor force participation rate and the explanatory variables being considered are: GDP per capita in US dollars, GDP growth, total GDP in international dollars, total energy use, urban population growth, female labor force participation rate, services value added, and total available food supply. After various analyses, we determined that the best prediction equation involves female labor force participation, log of total energy use, urban population growth, total available food supply, log of GDP per capita and GDP growth.

1 Variables

Using data from www.gapminder.org, we had a sample size of 123 different countries from the year 2007. We removed countries that were missing data from any of the selected variables in 2007.

Response variable:

- LFP: Labor force participation rate for ages 15-64 (%)

Explanatory variables:

- GDPC: Gross domestic product per capita (constant 2000 US dollars)
- GDPG: Gross domestic product growth (%)
- GDPI: Total gross domestic product (constant international dollars)
- PEU: Primary energy use (tonnes of oil equivalent)
- UPG: Annual urban population growth (%)
- FLP: Female labor force participation rate for ages 15-64 (%)
- SVA: Net output value added to the GDP due to services (% of GDP)
- TFA: Total available food supply (kilocalories per person per day)

2 Data Analysis and Results

	LFP ($)	GDPC (%)	GDPG (%)	GDPI ($)	PEU (tonnes)	UPG (%)	FLP (%)	SVA (% GDP)	TFA (Kcal)
Min.	44.90	97.91	-10.97	8.776e8	1.9e4	-3.23	16.40	21.50	1605
1st Qu.	63.25	1116.66	2.31	2.512e10	3.217e6	0.86	49.10	48.78	2438
Median	69.40	2725.82	4.15	9.602e10	1.212e7	1.84	59.10	58.02	2880
Mean	68.41	8135.35	4.58	6.424e11	8.971e7	2.00	56.76	56.90	2840
3rd Qu.	74.25	9376.59	6.64	4.195e11	5.449e7	2.80	67.55	67.79	3224
Max.	90.20	56285.28	23.64	1.550e13	2.337e9	15.21	89.30	84.26	3819
SD	8.72	1.48	4.12	2.12	29.10	1.87	15.49	14.26	499.72

Table 1: Summary statistics for explanatory variables.

	LFP	GDPC	GDPG	GDPI	PEU	UPG	FLP	SVA	TFA
LFP	1.000	0.281	0.031	0.116	0.133	0.100	0.916	0.070	-0.093
GDPC	0.281	1.000	-0.296	0.280	0.224	-0.096	0.275	0.455	0.593
GDPG	0.031	-0.296	1.000	0.029	0.070	-0.303	0.148	-0.179	-0.080
GDPI	0.116	0.280	0.029	1.000	0.984	-0.045	0.097	0.103	0.254
PEU	0.133	0.224	0.070	0.984	1.000	-0.028	0.118	0.062	0.223
UPG	0.100	-0.096	-0.303	-0.045	-0.028	1.000	-0.108	-0.373	-0.384
FLP	0.916	0.275	0.148	0.097	0.118	-0.108	1.000	0.140	-0.021
SVA	0.070	0.455	-0.179	0.103	0.062	-0.373	0.140	1.000	0.420
TFA	-0.093	0.593	-0.080	0.254	0.223	-0.384	-0.021	0.420	1.000

Table 2: Summary of sample correlations for all variables

Based on Table 2, FLP is the only explanatory variable that is highly correlated with LFP. Also note that there are high correlations between the following pairs of explanatory variables: LFA/GDPC and PEU/GDPI, which might suggest that we do not need to include all the explanatory variables in our optimal prediction equation. We explore which variables to include via selection methods presented in Section 3.

Figures 1-3 suggest that a log transform of variables PEU, GDPC and GDPI display an approximately linear relationship with the response variable. As shown in FIgure 4, none of the variables (except for FLP) have a clear increasing or decreasing linear relationship with LFP. However, our analysis shows that we are still able to obtain a good model for the LFP when these variables are considered together.

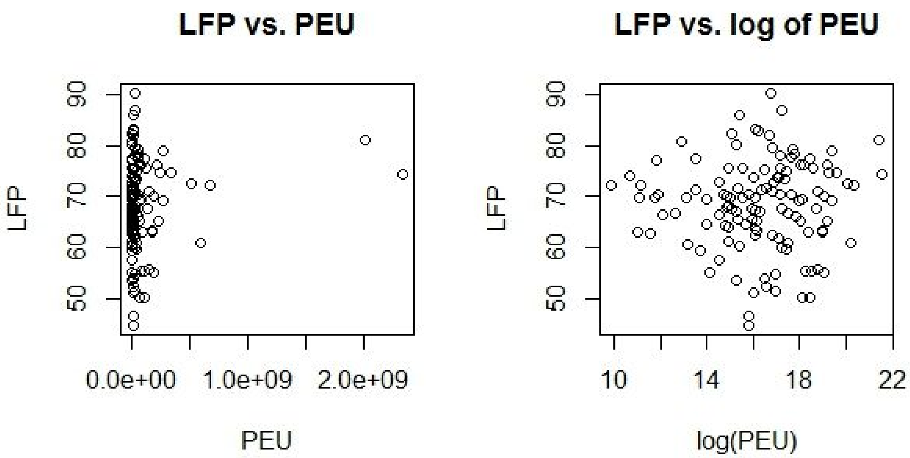

Figure 1: plots of LFP vs. PEU and LFP vs. log of PEU

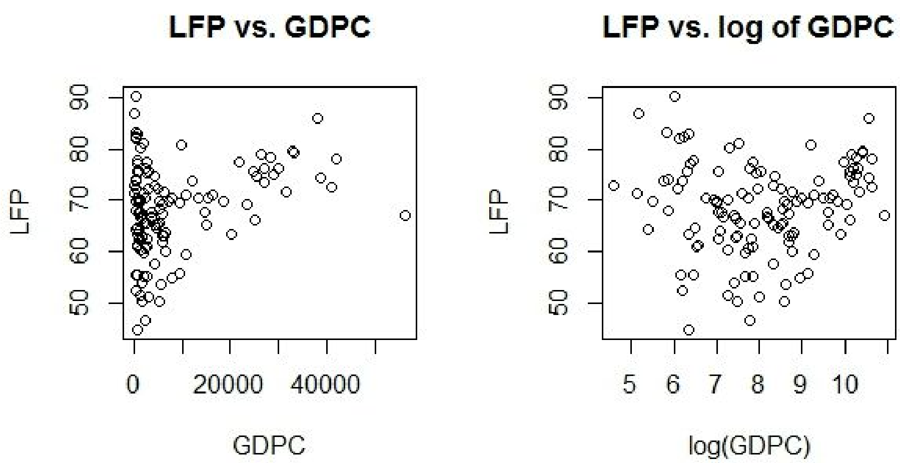

Figure 2: plots of LFP vs. GDPC and LFP vs. log of GDPC

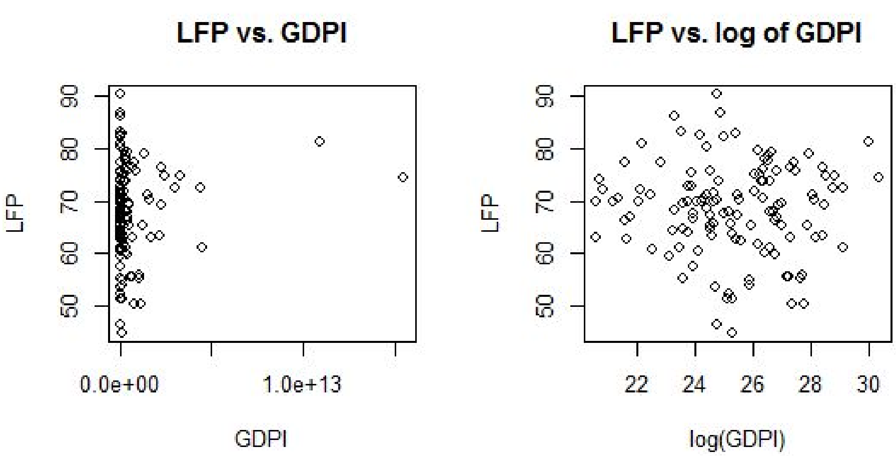

Figure 3: plots of LFP vs. GDPI and LFP vs. log of GDPI

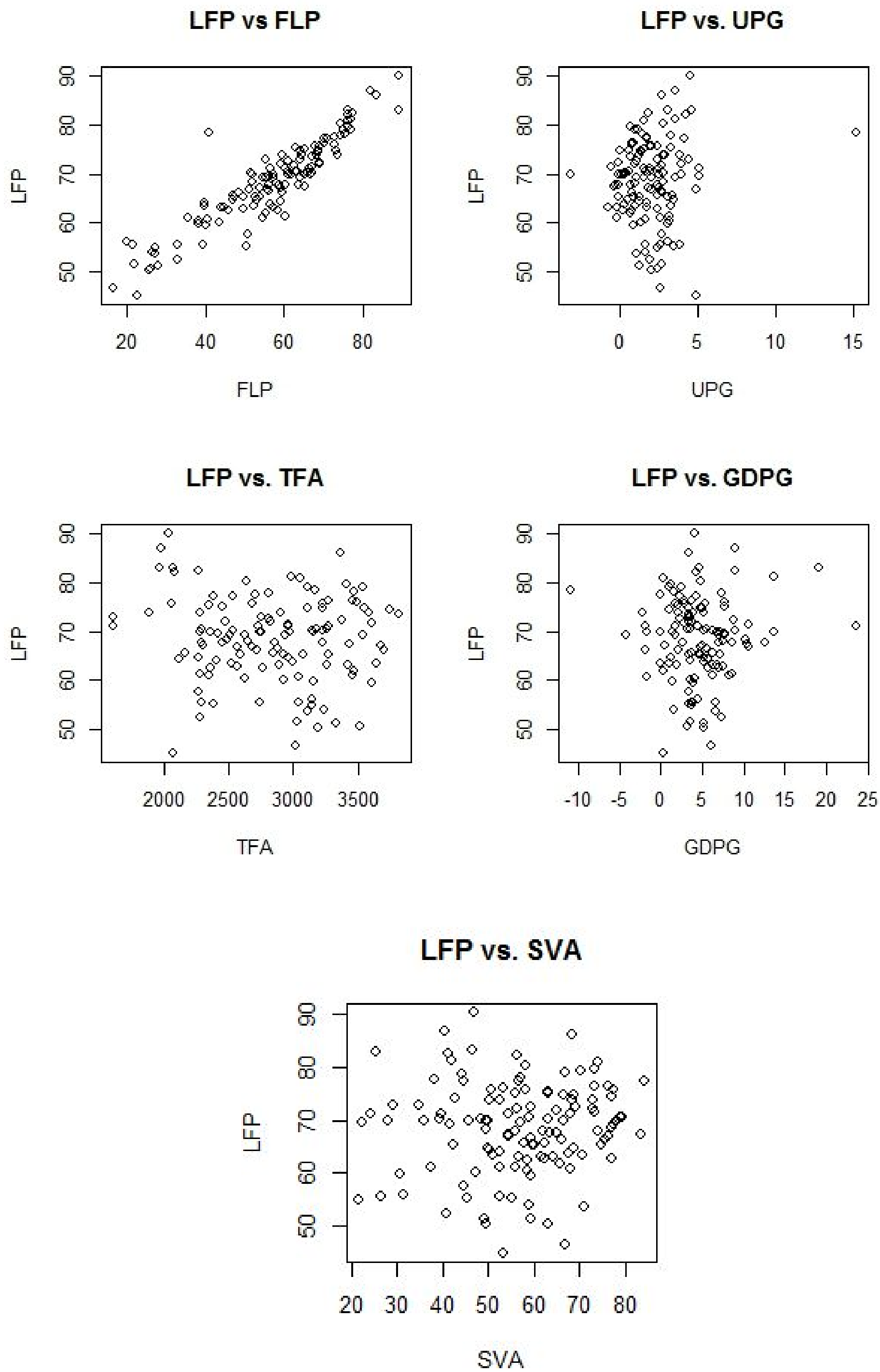

Figure 4: plots of LFP versus the explanatory variables we chose not to transform

3 Variable Selection Methods

Using the leaps package in R, we applied the exhaustive and backward selection methods for variable selection. The model that gives us the maximum $adjR^2$ and minimum Cp values will determine the best model size. Initial analysis showed that the best candidate models contain 3 variables, 4 variables, 5 variables, and 6 variables.

We notate Model3, Model4, Model5 and Model6 to be the models containing 3-6 variables, respectively.
Model3: FLP, UPG, GDPG (exhaustive); FLP, UPG, log(GDPC) (backward)
Model4: FLP, UPG, TFA, log(GDPC)
Model5: FLP, UPG, TFA, log(GDPC), GDPG
Model6: FLP, log(PEU), UPG, TFA, log(GDPC) and GDPG

Exhaustive Method	Model3	Model4	Model5	Model6
$adjR^2$	0.878	0.879	0.879	0.880
Cp	3.480	3.826	4.760	5.319
Backward Selection				
$adjR^2$	0.878	0.879	0.879	0.880
Cp	3.870	3.826	4.760	5.319
AIC	277.740	278.010	278.882	279.342
Resid. Std Dev.	3.044	3.035	3.034	3.028

Table 3: Comparison of Model3, Model4, Model5 and Model6 after using exhaustive and backward selection methods

Based on Table 3, it's difficult to select one optimal model based on the $adjR^2$, Cp, AIC and residual standard deviation values. All the models have very similar $adjR^2$ values for both selection methods. The least Cp value is Model3 and Model4, when using the exhaustive and backward selection methods respectively. The Akaike information criterion (AIC) assesses how well a model fits the data and the ideal model has the least value. The AIC values among all the selected models are very close but Model3 has the smallest AIC value.

Furthermore, the residual standard errors, beta estimates and their corresponding standard errors were also very similar. The signs of the beta estimates did not switch from positive to negative or vice versa when explanatory variables were added. Intuitively, it makes sense that total food supply, GDP growth (%), primary energy use, and urban population growth are not strongly correlated with each other. The signs of the beta estimates match the signs of the corresponding correlation. Thus, adding the explanatory variables sequentially to existing 3 variable models will not lead to a worse prediction.

4 Cross Validation

We used a two-fold cross validation with a training and holdout set to compare Model3, Model4, Model5 and Model6.

We began by randomizing our initial data-set and splitting it equally into the training and holdout sets. In each of our models, we constructed a linear model on the training data, performed a prediction on the holdout, and compared the models by calculating the cross-validated root mean square prediction errors.

	Model3	Model4	Model5	Model6
$CVRMSE_{holdout}$	2.971	2.973	2.944	3.028

Table 4: $CVRMSE_{holdout}$ of the four candidate models based on cross validation

Model5 resulted in the lowest $CVRMSE_{holdout}$ value and was marginally smaller than the $CVRMSE_{holdout}$ values of Model3 and Model4. Combined with our findings from Table 3 and based on the principle of parsimony, Model3 appears to be the best predictor models due to high $adjR^2$ and low Cp, AIC, and residual standard deviation values.

5 Validity of Linear Regression

We analyzed the validity of the linear regression for the best model: Model3. Based on the normal Q-Q plot, the residuals are considered approximately normal since they appear close to the qqline. Moreover, we verify the homoscedastic assumption of the data via a fitted versus residual plot, which shows only four extreme values. Four extreme values is reasonable and a considerably small amount when based on a sample size of 123. Also, the residual plot indicates that we do not need to add quadratic or interaction terms in our prediction equation because the spread of residuals is random.

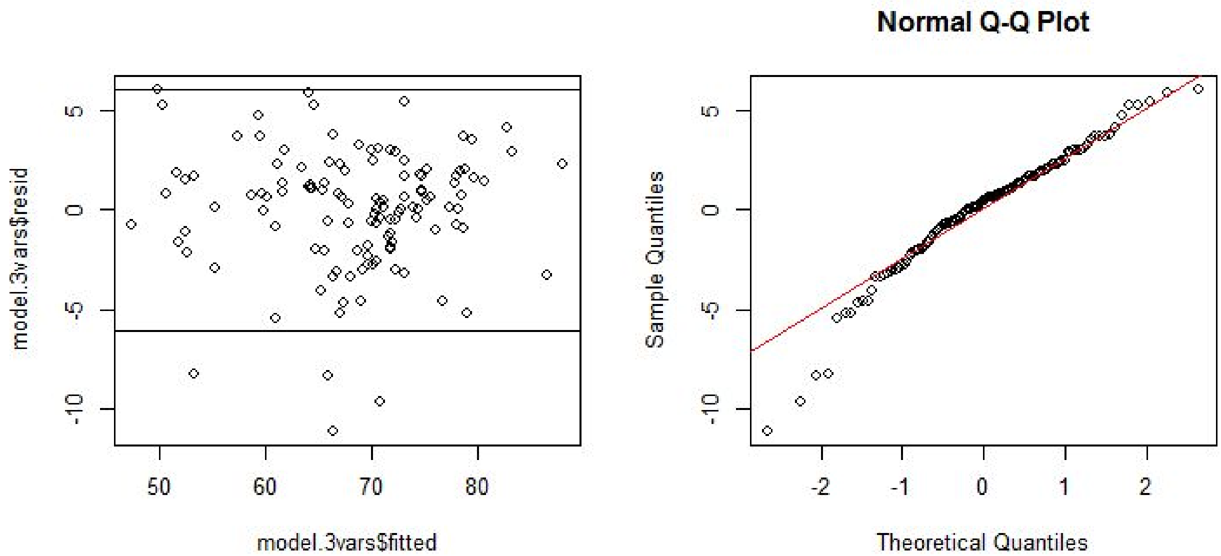

Figure 5. Fitted values vs residual plot and normal Q-Q plot of residuals

Partial Residual Plots

We plotted each dependent explanatory variable against the residuals of the fitted model to verify whether or not there is a linear relationship. All the residual plots show only a handful of values that are beyond the 95% confidence intervals (indicated by the horizontal lines), illustrating that the explanatory variables have a linear relationship with the residuals of the fitted model.

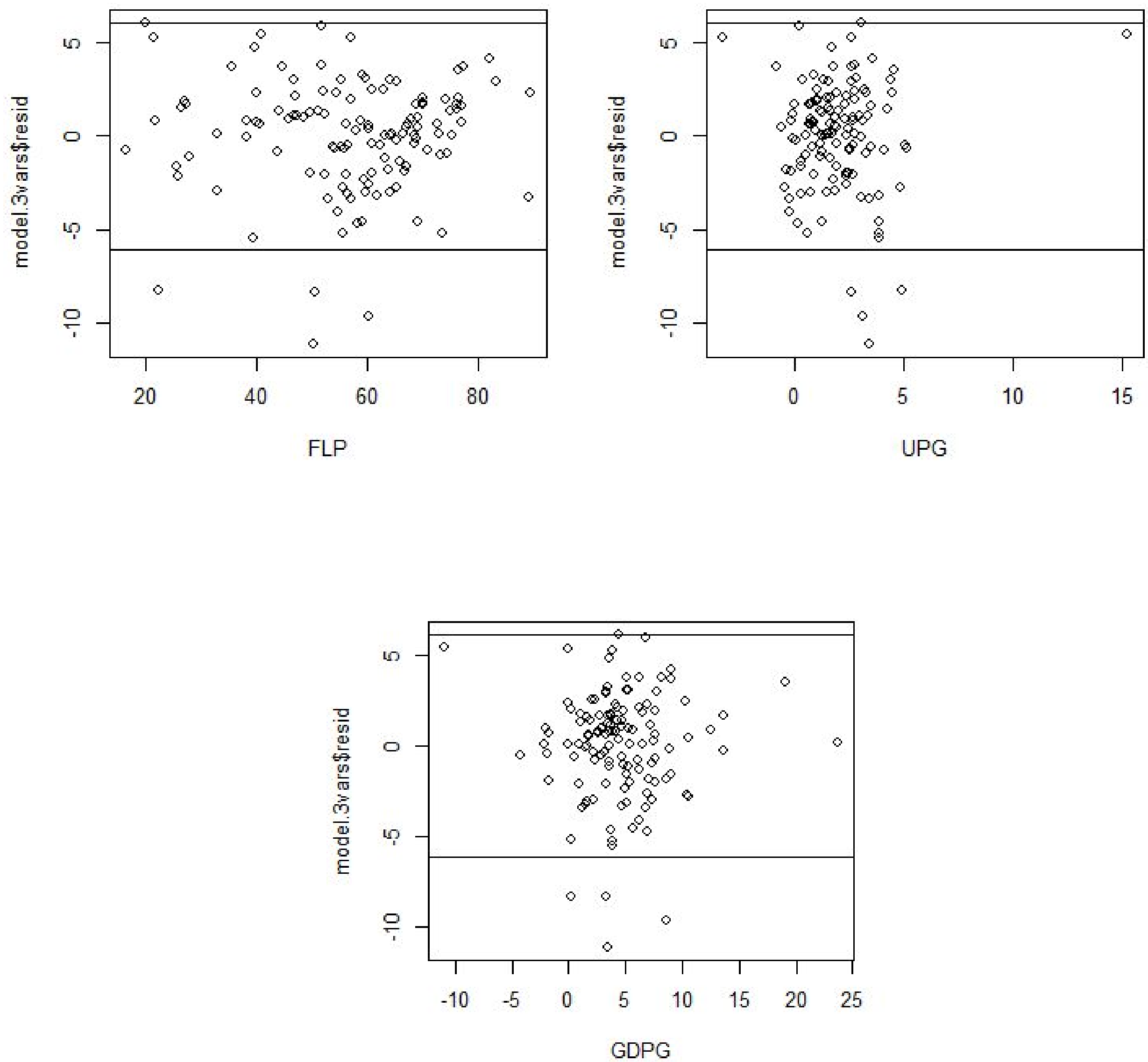

Figure 6. Partial residual plots for each explanatory variable in Model3

Diagnostics

We checked for influential observations using the ls.diag() function to obtain Cook's distance and dfits values. Absolute values of Cook's distances and dfits are considered to be significant if they exceed 0.05, which is defined by the thresholds: $\frac{4}{n}$ and $2\sqrt{\frac{p}{n}}$, respectively (where n is the sample size and p is the number of parameters). We identified five countries that have statistically significant Cook's distance and dfits values: Angola, St. Lucia, Samoa, Saudi Arabia, and United Arab Emirates which had the largest value. We analyzed the raw data and found that United Arab Emirates provided the minimum value of GDPG and maximum value of UPG, thus resulting to be an influential observation.

For future analysis, we can remove this country in the dataset and determine whether the prediction equation would be any different.

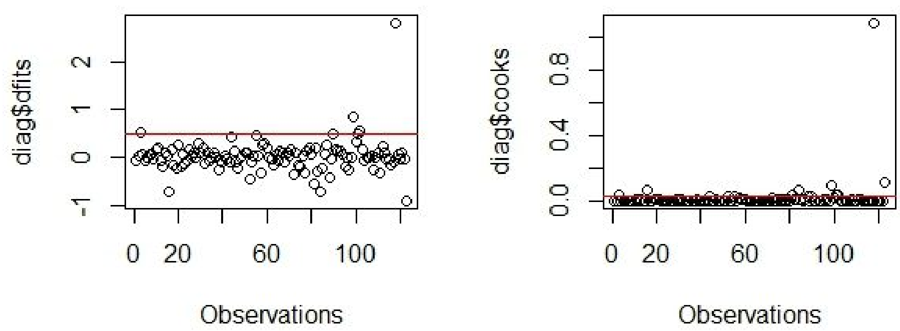

Figure 7. dfits and Cook's distance plots of the observations.

Conclusion

Due to the principle of parsimony, we can conclude that the best prediction equation for LFP is Model3:

LFP = 37.0 + 0.87*UPG + 0.53*FLP -0.11*GDPG

This model resulted in high $adjR^2$ and low Cp, AIC, and residual standard deviation values, which indicates that this is a good predictor of the labor force participation rate. The poor linear relationships in the original scatterplots (Figure 4) could be due to the different economies and unpredictable events that occur in each country. It's difficult to predict the labor force participation of a country based on other countries. However, the combined effect of economic measures of a country is a better predictor.

The signs of UPG and FLP in the predictor equation match our expectations, but GDPG doesn't. It would seem reasonable for GDPG to increase LFP because as the country's economy grows, more jobs are created. Our data model has 5 countries that appear to be outliers. The economies of these countries probably stray farther from the norm compared to the other data points. By adding more explanatory variables, we can account for some of the factors that make these countries outliers and make our model better predictors. Other explanatory variables that we could further explore are: total fertility, children out of school, income per person, industry workers, and mean years in school.

Appendix:

```
> exh.sum
Subset selection object
Call: regsubsets.formula(LFP ~ ., data = lfpr.transform)
8 Variables  (and intercept)
        Forced in Forced out
FLP         FALSE      FALSE
logPEU      FALSE      FALSE
UPG         FALSE      FALSE
SVA         FALSE      FALSE
TFA         FALSE      FALSE
logGDPI     FALSE      FALSE
logGDPC     FALSE      FALSE
GDPG        FALSE      FALSE
1 subsets of each size up to 8
Selection Algorithm: exhaustive
         FLP logPEU UPG SVA TFA logGDPI logGDPC GDPG
1  ( 1 ) "*" " "    " " " " " " " "     " "     " " 
2  ( 1 ) "*" " "    "*" " " " " " "     " "     " " 
3  ( 1 ) "*" " "    "*" " " " " " "     " "     "*" 
4  ( 1 ) "*" " "    "*" " " "*" " "     "*"     " " 
5  ( 1 ) "*" " "    "*" " " "*" " "     "*"     "*" 
6  ( 1 ) "*" "*"    "*" " " "*" " "     "*"     "*" 
7  ( 1 ) "*" "*"    "*" " " "*" "*"     "*"     "*" 
8  ( 1 ) "*" "*"    "*" "*" "*" "*"     "*"     "*" 
```

Figure 8: Exhaustive selection summary table

```
> back.sum
Subset selection object
Call: regsubsets.formula(LFP ~ ., data = lfpr.transform, method = "backward")
8 Variables  (and intercept)
        Forced in Forced out
FLP         FALSE      FALSE
logPEU      FALSE      FALSE
UPG         FALSE      FALSE
SVA         FALSE      FALSE
TFA         FALSE      FALSE
logGDPI     FALSE      FALSE
logGDPC     FALSE      FALSE
GDPG        FALSE      FALSE
1 subsets of each size up to 8
Selection Algorithm: backward
         FLP logPEU UPG SVA TFA logGDPI logGDPC GDPG
1  ( 1 ) "*" " "    " " " " " " " "     " "     " " 
2  ( 1 ) "*" " "    "*" " " " " " "     " "     " " 
3  ( 1 ) "*" " "    "*" " " " " " "     "*"     " " 
4  ( 1 ) "*" " "    "*" " " "*" " "     "*"     " " 
5  ( 1 ) "*" " "    "*" " " "*" " "     "*"     "*" 
6  ( 1 ) "*" "*"    "*" " " "*" " "     "*"     "*" 
7  ( 1 ) "*" "*"    "*" " " "*" "*"     "*"     "*" 
8  ( 1 ) "*" "*"    "*" "*" "*" "*"     "*"     "*" 
```

Figure 9: Backward selection summary table

```
Model3
lm(formula = LFP ~ UPG + FLP + GDPG, data = training.set)

Residuals:
     Min       1Q   Median       3Q      Max
-10.4779  -2.0384   0.3185   1.8577   6.4760

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.54968    1.80535  20.799   <2e-16 ***
UPG          0.59672    0.28432   2.099   0.0402 *
FLP          0.52584    0.02815  18.679   <2e-16 ***
GDPG        -0.10860    0.09936  -1.093   0.2789
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.236 on 58 degrees of freedom
Multiple R-squared:  0.8583,	Adjusted R-squared:  0.851
F-statistic: 117.1 on 3 and 58 DF,  p-value: < 2.2e-16
```

Figure 10: results of cross validation on the model with 3 variables

```
Model4
lm(formula = LFP ~ FLP + UPG + TFA + logGDPC, data = training.set)

Residuals:
     Min       1Q   Median       3Q      Max
-11.0247  -1.6568   0.4921   1.6812   5.9649

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.989680   3.396913  10.889 1.51e-15 ***
FLP          0.517118   0.028152  18.369  < 2e-16 ***
UPG          0.615797   0.313611   1.964   0.0545 .
TFA         -0.001539   0.001266  -1.215   0.2292
logGDPC      0.607604   0.424510   1.431   0.1578
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.238 on 57 degrees of freedom
Multiple R-squared:  0.8605,	Adjusted R-squared:  0.8507
F-statistic: 87.92 on 4 and 57 DF,  p-value: < 2.2e-16
```

Figure 11: results of cross validation on the model with 4 variables

```
Model6
lm(formula = LFP ~ FLP + logPEU + UPG + TFA + logGDPC + GDPG,
    data = training.set)

Residuals:
     Min       1Q   Median       3Q      Max
-10.5479  -1.7501   0.3511   1.6474   6.7588

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 35.450451   3.929944   9.021 1.96e-12 ***
FLP          0.520880   0.028346  18.376  < 2e-16 ***
logPEU       0.302379   0.228393   1.324    0.191
UPG          0.461751   0.327325   1.411    0.164
TFA         -0.002151   0.001363  -1.578    0.120
logGDPC      0.479807   0.435008   1.103    0.275
GDPG        -0.118148   0.105345  -1.122    0.267
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.225 on 55 degrees of freedom
Multiple R-squared:  0.8665,  Adjusted R-squared:  0.852
F-statistic: 59.51 on 6 and 55 DF,  p-value: < 2.2e-16
```

Figure 12: results of cross validation on the model with 6 variables

```
> #model with 3 variables: GDP, FLP, UPG
> model.3vars<- lm(LFP~UPG+FLP+GDPG, data = lfpr.transform)
> summary(model.3vars)

Call:
lm(formula = LFP ~ UPG + FLP + GDPG, data = lfpr.transform)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.1420  -1.5931   0.4602   1.8068   6.0856 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 36.99552    1.15927  31.913  < 2e-16 ***
UPG          0.86797    0.15467   5.612 1.33e-07 ***
FLP          0.53162    0.01803  29.481  < 2e-16 ***
GDPG        -0.10972    0.07075  -1.551    0.124    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.044 on 119 degrees of freedom
Multiple R-squared:  0.8813,	Adjusted R-squared:  0.8783 
F-statistic: 294.5 on 3 and 119 DF,  p-value: < 2.2e-16
```

Figure 13: Summary Table for Model with 3 variables

```
> #model with 4 variables: FLP, UPG, TFA, logGDPC
> model.4vars<- lm(LFP~FLP+UPG+TFA+logGDPC, data = lfpr.transform)
> summary(model.4vars)

Call:
lm(formula = LFP ~ FLP + UPG + TFA + logGDPC, data = lfpr.transform)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.6596  -1.5547   0.4367   1.6139   6.2345 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 35.8837428  2.2048996  16.275  < 2e-16 ***
FLP          0.5210484  0.0182536  28.545  < 2e-16 ***
UPG          0.9314222  0.1602842   5.811 5.38e-08 ***
TFA         -0.0013044  0.0009078  -1.437   0.1534    
logGDPC      0.5949443  0.2947703   2.018   0.0458 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.035 on 118 degrees of freedom
Multiple R-squared:  0.8829,	Adjusted R-squared:  0.879 
F-statistic: 222.5 on 4 and 118 DF,  p-value: < 2.2e-16
```

Figure 14: Summary Table for Model with 4 variables

```
#model with 5 variables: FLP, UPG, TFA, logGDPC, GDPG
> model.5vars<- lm(LFP~FLP+UPG+TFA+logGDPC+GDPG, data = lfpr.transform)
> summary(model.5vars)

Call:
lm(formula = LFP ~ FLP + UPG + TFA + logGDPC + GDPG, data = lfpr.transform)

Residuals:
     Min       1Q   Median       3Q      Max
-11.6274  -1.5262   0.2934   1.6097   6.2220

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.7209177  2.3469905  15.646  < 2e-16 ***
FLP          0.5242899  0.0185126  28.321  < 2e-16 ***
UPG          0.8688315  0.1711938   5.075 1.47e-06 ***
TFA         -0.0012391  0.0009097  -1.362    0.176
logGDPC      0.5050712  0.3071205   1.645    0.103
GDPG        -0.0781851  0.0752950  -1.038    0.301
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.034 on 117 degrees of freedom
Multiple R-squared:  0.884,	Adjusted R-squared:  0.8791
F-statistic: 178.4 on 5 and 117 DF,  p-value: < 2.2e-16
```

Figure 15: Summary Table for Model with 5 variables

```
> #model with 6 variables: FLP, logPEU, UPG, TFA, logGDPC, GDPG
> model.6vars<- lm(LFP~FLP+logPEU+UPG+TFA+logGDPC+GDPG, data = lfpr.transform)
> summary(model.6vars)

Call:
lm(formula = LFP ~ FLP + logPEU + UPG + TFA + logGDPC + GDPG,
    data = lfpr.transform)

Residuals:
     Min       1Q   Median       3Q      Max
-11.4049  -1.4145   0.2854   1.5377   6.6526

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 35.2815869  2.6276091  13.427  < 2e-16 ***
FLP          0.5243740  0.0184764  28.381  < 2e-16 ***
logPEU       0.1582379  0.1308962   1.209    0.229
UPG          0.8366988  0.1729125   4.839 4.06e-06 ***
TFA         -0.0015479  0.0009432  -1.641    0.103
logGDPC      0.4895575  0.3067853   1.596    0.113
GDPG        -0.0936102  0.0762225  -1.228    0.222
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.028 on 116 degrees of freedom
Multiple R-squared:  0.8855,	Adjusted R-squared:  0.8795
F-statistic: 149.5 on 6 and 116 DF,  p-value: < 2.2e-16
```

Figure 16: Summary Table for Model with 6 variables