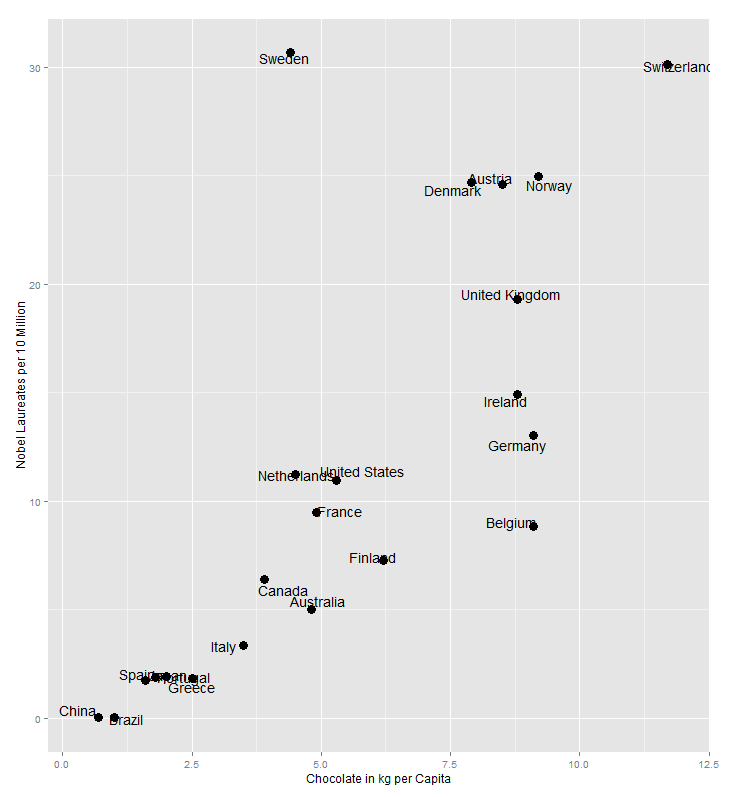
**Assignment 6**

**Katherine Rodgers and John Merranko**

We first tracked down updated Nobel prize and chocolate consumption data to replicate the analyses performed in the original paper. In addition to scatterplotting and calculating the correlation between the two variables, we regressed Nobel prizes per 10 million on chocolate consumption per capita. As was the case in the original paper, correlation between the two variables was high (nearly 0.8), and chocolate consumption had a significant positive association with Nobel prizes (p=3.63e-05, R2=0.58) as demonstrated by linear regression.



> cor(nobel$nobel.per.10.million, nobel$choc)

[1] 0.7630346

lm(formula = nobel.per.10.million ~ choc, data = nobel)

Residuals:

Min 1Q Median 3Q Max

-11.2573 -3.2452 -0.7373 1.5147 21.7347

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.510 2.837 -0.532 0.6

choc 2.376 0.450 5.279 3.63e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.646 on 20 degrees of freedom

Multiple R-squared: 0.5822, Adjusted R-squared: 0.5613

F-statistic: 27.87 on 1 and 20 DF, p-value: 3.628e-05

While these results are seemingly conclusive, we figured it may be the case that chocolate consumption does not literally increase the number of Nobel laureates in a nation. Rather, chocolate consumption serves as a proxy variable for the relative economic and educational prosperity of a country. Thus, we added the following variables to the analysis that we believed more directly measure the prosperity of nations:

Median Household Income

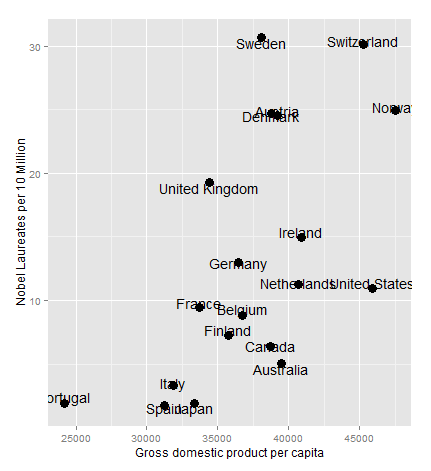
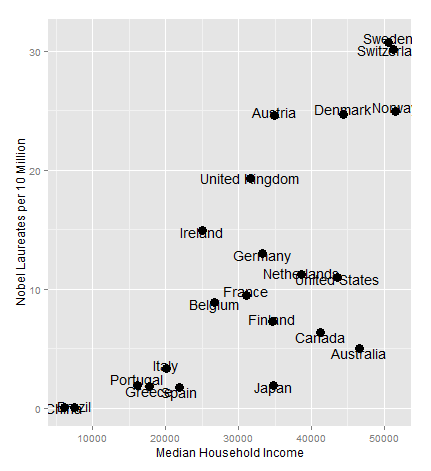
Gross domestic product per capita

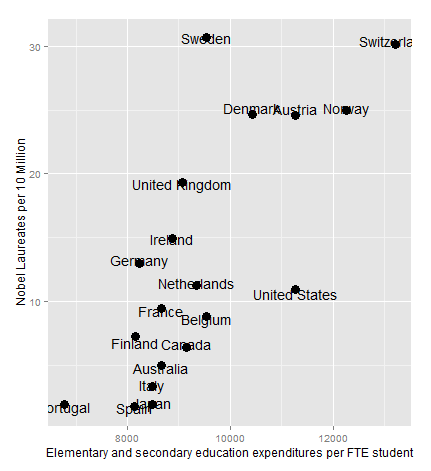
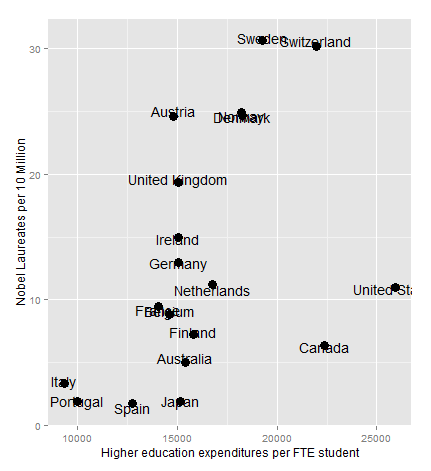
Elementary and secondary education expenditures per full time equivalent student

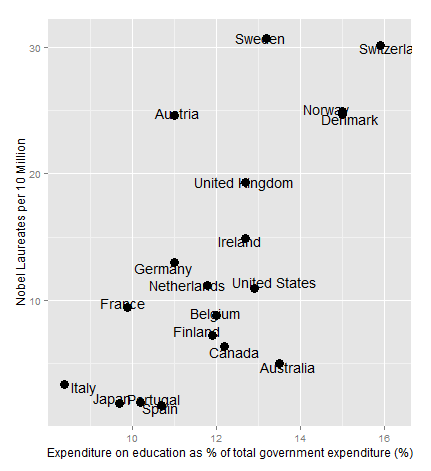
Higher education expenditures per full time equivalent student

Expenditure on education as percent of total government expenditure

Below we include scatterplots of Nobel laureates per 10 million against each of the variables above along with the correlation and simple linear regression between each pair:





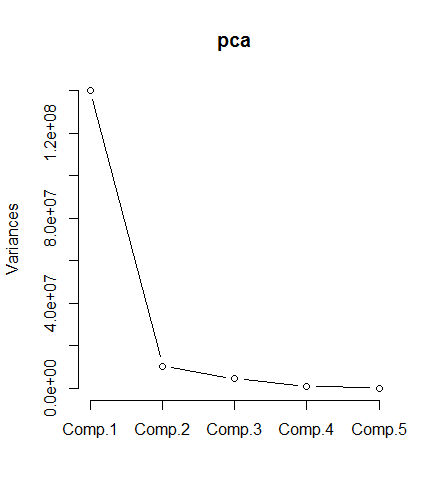
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Correlation and Simple Linear Regression of Nobel Prizes per 10 Million** | | | | | |
| **Predictor** | **Correlation** | **Regression  Coefficient** | **t-stat** | **p-value** | **R2** |
| Median household income | 0.72 | 0.0005 | 4.65 | 0.0002 | 0.52 |
| Gross domestic product per capita | 0.60 | 0.001 | 3.07 | 0.007 | 0.36 |
| Elementary and secondary  education expenditures per student | 0.75 | 0.005 | 4.63 | 0.0002 | 0.56 |
| Higher education  expenditures per student | 0.45 | 0.001 | 2.1 | 0.05 | 0.21 |
| Expenditure on education as  percent of total government expenditure | 0.71 | 3.53 | 4.11 | 0.0007 | 0.50 |

Overall, all variables analyzed were positively associated with Nobel laureates per 10 million. The most strongly associated correlates appeared to be median household income and elementary and secondary education expenditures per student. We next sought to assemble multiple regression models using all the variables analyzed above including chocolate consumption, however the correlation matrix below showed substantial intercorrelation between the potential regressors:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | choc | edu.gov | gdp | elem.sec.edu | higher.edu |
| edu.gov | 0.41 |  |  |  |  |
| gdp | 0.63 | 0.37 |  |  |  |
| elem.sec.edu | 0.67 | 0.36 | 0.94 |  |  |
| higher.edu | 0.38 | 0.37 | 0.85 | 0.81 |  |
| income | 0.58 | 0.41 | 0.71 | 0.75 | 0.81 |

To avoid introducing multicollinearity to our regression models (which would have the consequence of biasing inference on individual regressors), we decided to reduce the full set of variables to one variable representing each of the following domains: chocolate consumption, national wealth (GDP and median household income), and education (expenditures on elementary and secondary education, higher education, and education as percent of total government expenditure). We explored the possibility of using principal components analysis to reduce the national wealth and education variables to orthogonal factors, but as you can see below, PCA did not effectively discriminate between the domains. The scree plot indicates that the optimal solution is a one-factor solution (which of course wouldn’t help us), and the two-factor solution we were hoping would help the most does not discriminate between the domains.

|  |
| --- |
| Factor Loadings:  PC1 PC2  edu.gov 0.62 0.61  gdp 0.57 0.75  elem.sec.edu 0.34 0.91  higher.edu 0.87 0.35  income 0.81 0.47 |



Our next step to reduce within our regressor domains was then to try and evaluate which of the variables within each domain best regressed the outcome Nobel laureates per 10 million. The previously presented correlation and simple linear regression results indicated that median household income was the best national wealth regressor, and elementary and secondary education expenditures per student was the best education regressor. However, we also wanted to run multiple regression models with regressors within each domain to see how they perform when competing against one another in the same model. We note that multicollinearity prevents us from putting too much weight behind results from this stage of the analysis, but results (shown below) supported the conclusions drawn from the simple linear regressions.

lm(formula = nobel.per.10.million ~ median.hh.income + GDP.per.capita,

data = master)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.280e+01 4.806e+00 -2.663 0.0142 \*

median.hh.income 4.262e-04 1.742e-04 2.446 0.0229 \*

GDP.per.capita 2.966e-04 1.680e-04 1.766 0.0913 .

lm(formula = nobel.per.10.million ~ elem.sec.edu + higher.edu + edu.gov,

data = master)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.725e+01 8.976e+00 -3.036 0.0068 \*\*

elem.sec.edu 2.934e-03 1.151e-03 2.549 0.0196 \*

higher.edu -2.491e-04 5.184e-04 -0.481 0.6363

edu.gov 1.409e+00 7.077e-01 1.991 0.0611 .

Thus, we opted to use median household income as the measure of national wealth and Elementary and secondary education expenditures per student as the measure of education. We then regressed Nobel laureates per 10 million on all possible two-variable combinations between these two variables as well as chocolate consumption, and then a final full three-variable model was assembled. Our hope was to find which variable(s) of the three best regressed the primary outcome.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Multiple Linear Regression of Nobel Prizes per 10 Million** | | | |
|  | **Predictors in Model** | **t-stat** | **p-value** | **Adj. R2** |
| **Model 1** | Chocolate consumption | 3.38 | 0.003 | 0.67 |
| Median household income | 2.74 | 0.01 |
| **Model 2** | Chocolate consumption | 2.01 | 0.06 | 0.61 |
| Elem. and sec. edu. | 2.56 | 0.02 |
| **Model 3** | Median household income | 1.8 | 0.09 | 0.6 |
| Elem. and sec. edu. | 2.57 | 0.02 |
| **Model 4** | Chocolate consumption | 2.34 | 0.03 | 0.64 |
| Median household income | 1.58 | 0.1 |
| Elem. and sec. edu. | 0.82 | 0.4 |

As you can see above, the only variable with a more significant effect than chocolate consumption in any regression was elementary and secondary education expenditures per student (Model 2). The model with the largest adjusted R2 was Model 1, which included chocolate consumption and median household income.

As a final stage of the analysis, we sought to select a model with the strongest possible adjusted R2. Since this doesn’t depend on parametric tests on individual regressors, collinearity between predictors is inconsequential. Thus, we used our full subset of regressors in this stage of the analysis. We then wrote a program to perform an exhaustive search over all possible subsets for the regressor subset that regressed Nobel laureates per 10 million with the strongest adjusted R2. The resulting “best” model included chocolate consumption, GDP, elementary and secondary education expenditures per student, and median household income, and the resulting adjusted R2 was 0.6831 (model output shown below).

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -6.8515084 11.0125606 -0.622 0.5438

choc 1.6972717 0.5978977 2.839 0.0131 \*

gdp -0.0010028 0.0005167 -1.941 0.0727 .

elem.sec.edu 0.0030935 0.0016320 1.896 0.0789 .

income 0.0005052 0.0002064 2.447 0.0282 \*

Residual standard error: 5.479 on 14 degrees of freedom

Multiple R-squared: 0.7535, Adjusted R-squared: 0.6831

F-statistic: 10.7 on 4 and 14 DF, p-value: 0.0003467

To summarize, we first replicated the analyses performed in the paper that showed the correlation and significant positive association of chocolate consumption with winning Nobel prizes. Because the author states “a correlation between X and Y does not prove causation but indicates that either X influences Y, Y influences X, or X and Y are influenced by a common underlying mechanism”, we decided to investigate possible underlying mechanisms. We decided that consumption of a luxury good such as chocolate is probably more prevalent in well developed countries, and, thus, decided to look at other qualities that well developed countries would possess, such as economic and educational variables. Thus, we added two domains to our analysis, national wealth and education. All variables within these two domains showed a positive association with Nobel laureates per 10 million with the strongest correlations being median household income and elementary and secondary education expenditures per student. We explored using principal components analysis, but PCA did not effectively discriminate between the domains. Because of the substantial intercorrelation between the variables in these two domains and to avoid biasing the significance of individual predictors, at this point we decided to use the best regressor in each domain: median household income for wealth and elementary and secondary education expenditures per student for education. Along with chocolate consumption, we regressed Nobel laureates per 10 million on all possible two-variable combinations plus a full three-variable model. One noteworthy finding is, in a head-to-head matchup, elem.sec.edu.expenditures.per.FTE.student actually predicts better than chocolate consumption. In the full three-variable model, chocolate consumption is the only significant predictor, although it should be noted that the other two predictors are much more intercorrelated which results in a dampening one another’s individual effects in the model. Lastly, we decided to run an exhaustive search program on the full subset of predictors to find which model produces the higher adjusted r-squared value. Since this doesn’t depend on the internal parametric tests, collinearity between predictors is inconsequential. This model included chocolate consumption, GDP, elementary and secondary education expenditures per student, and median household income showing income and chocolate consumption as significant with chocolate consumption being the most significant.