# STAT-S 432: Homework 2

*Due February 15, 2019*

**Instructions:** You must submit this homework by pushing a file named "hw2.Rmd" file to your team's repo. Note that that is the **only** file you will be allowed to push. Commit early and often.

**Data named uval.csv**

"Gross domestic product" is a standard measure of the size of an economy; it's the total value of all goods and services bought and solid in a country over the course of a year. It's not a perfect measure of prosperity, but it is a very common one, and many important questions in economics turn on what leads GDP to grow faster or slower.

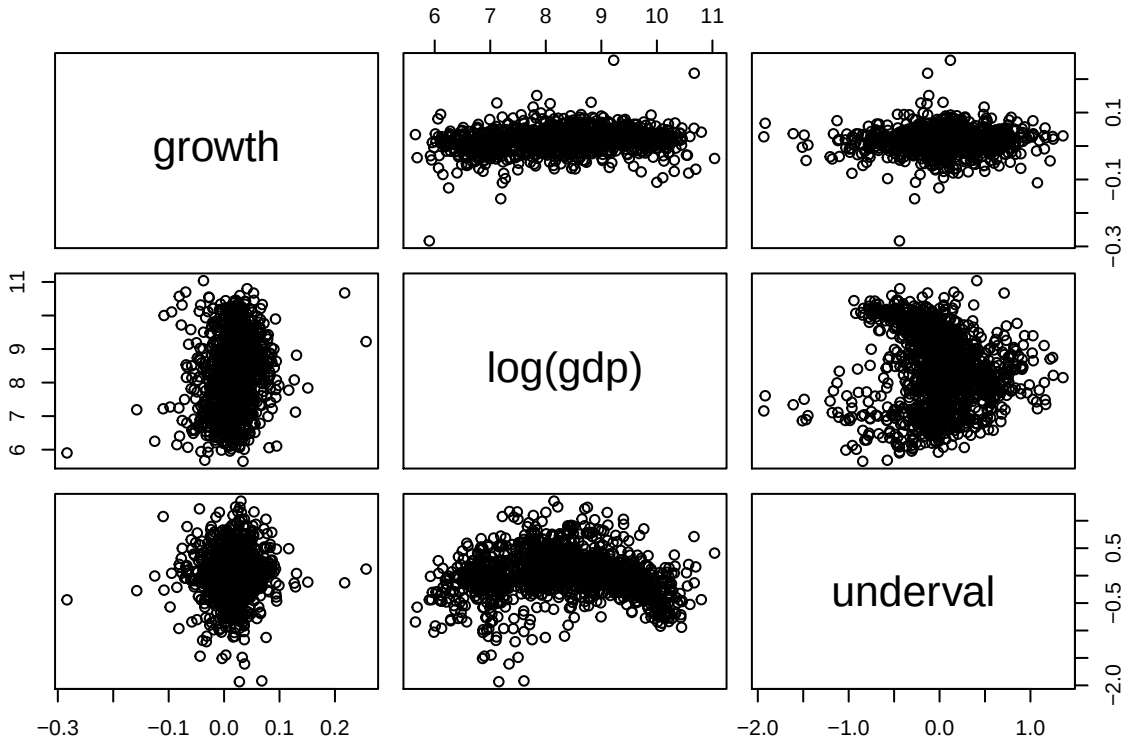
One common idea is that poorer economies, those with lower initial GDPs, should grow faster than richer ones. The reasoning behind this "catching up" is that poor economies can copy technologies and procedures from richer ones, but already-developed countries can only grow as technology advances. A second, separate idea is that countries can boost their growth rate by under-valuing their currency, making the goods and services they export cheaper. This week's data set contains the following variables:

- Country, in a three-letter code (see http://en.wikipedia.org/wiki/ISO_ 3166-1_alpha-3).
- Year (in five-year increments).
- Per-capita GDP, in dollars per person per year ("real" or inflation-adjusted).
- Average percentage growth rate in GDP over the next five years.
- An index of currency under-valuation
  - The index is 0 if the currency is neither over- nor under- valued, positive if under-valued, negative if it is over-valued.Note that not all countries have data for all years. However, there are no missing values in the data table.

1. Perform and Exploratory Data Analysis on the data. Write sentences describing the data. You should address (or do) all the things listed:
   - What is the data and where does it come from?
   - How many observations do we have? Are there any missing?
   - What and which kind of predictors are available?
   - Produce a pairwise scatterplot of all the continuous variables. Comment on what types of relationships you see.
   - How does this plot inform whether a linear regression model is reasonable or not?
   - Are there any obvious outliers? Why do you say that? Should we remove them?

The dataset presented provides a look into the economies of various countries through economic indicators over 10 five year periods from 1955 to 2000. There are a total of 179 countries for which we are provided the GDP, and a currency valuation index across the time periods, though not all countries have observations for each time period. The objective is to use the provided data to predict the growth rate of the countries across the time periods using the GDP and currency valuation.

The main objective is to see if growth rate is affected by GDP and currency valuation; specifically, we want to see if economies with low GDPs tend to have higher economic growth. As the data stands, there are no missing values beyond certain countries not having observations for all time periods. In total, there are 1301 observations.




A pairwise scatterplot is given above. Note that the log of GDP is used instead of the raw GDP values as GDP is a highly right-skewed variable. Very often, the log of financial data must used due to the inherently skewed nature. From the plots it is difficult to see a direct relation with either the log of GDP or the undervaluation index. If there is a relationship, it appears that there is a slight positive relatinoship between growth and both `log(gdp)` and the undervaluation index, though this pattern may change when breaking down the data by year.

There is a very strange relationship between undervaluation and `log(gdp)`. High and low values of the `log(gdp)` are associated with overvalued (negative `underval`) currencies.

There appear to be three outliers in terms of growth which are presented in the following table. The two highest growths from Equatorial Guinea (GNQ) in 2000, and United Arab Emirates (ARE) in 1975. It should be noted that the UAE was founded in 1971 and this may be the reason for its high growth in 1975. A cursory examination of Equatorial Guinea's history shows no extraordinary events around that time period. The lowest growth was found to be from Liberia (LBR) in 1990. This year in Liberia may merit removal from the dataset as this was the time of the First Liberian Civil War, which took place from 1989 to 1996. For now, all observations will remain in the dataset.

| country | year | gdp | growth | underval |
|---|---|---|---|---|
| LBR | 1990 | 366.0304 | -0.2834164 | -0.4411074 |
| ARE | 1975 | 43165.5977 | 0.2175430 | -0.1314420 |
| GNQ | 2000 | 10118.0615 | 0.2563137 | 0.1195629 |

2. Linearly regress the growth rate on the under-valuation index and the log of GDP.
   - Report the coefficients and their standard errors.
   - Do the coefficients support the idea of "catching up"? Do they support the idea that under-valuing

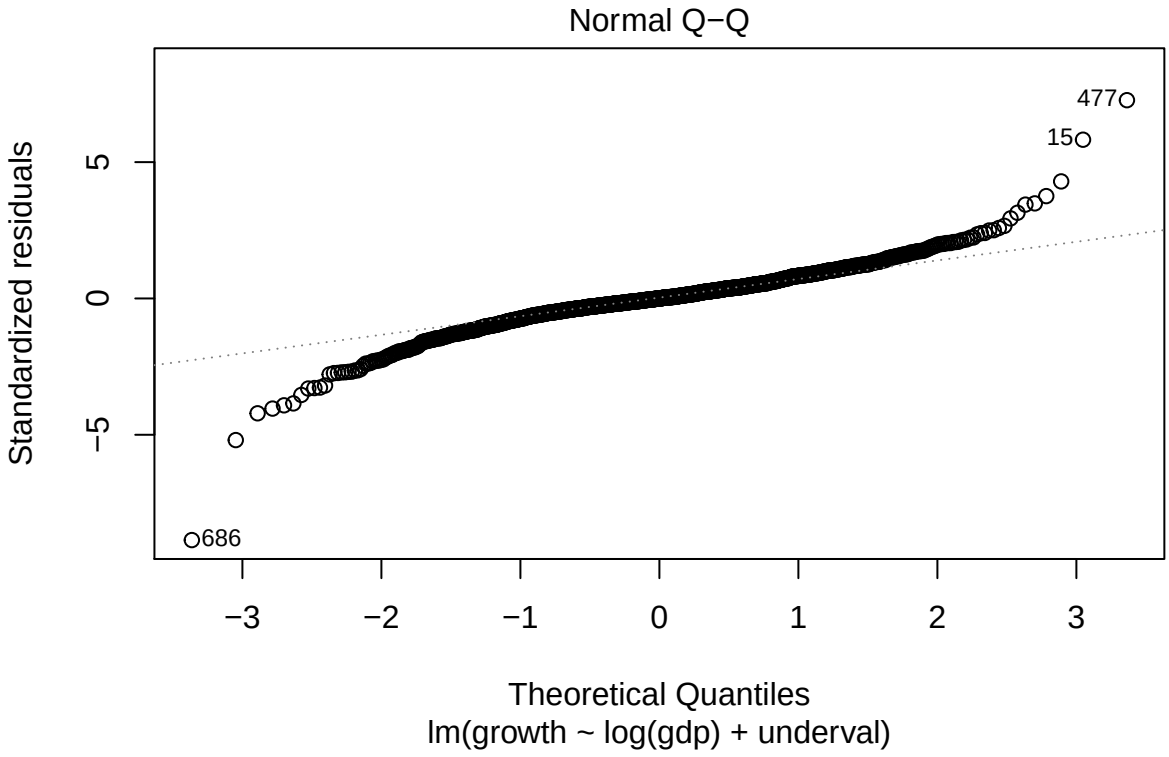
a currency boosts economic growth?
- Check the residuals and report your findings.

The first model to be fit to the data used `log(gdp)` and the undervaluation index to predict the growth rate. The coefficients and their standard errors are presented in the following table.

The coefficient for the `log(gdp)` is 0.006 which means that higher GDP is associated with higher growth. Therefore we do not have support of the idea that lower GDP countries "catch up", i.e., have higher growth on average.

The coefficient for the undervaluation index is 0.005. This indicates that higher values of this index are associated with higher growth rates. Higher values of this index indicate a more undervalued currency. Therefore, we have support of the idea that, in general, countries with undervalued currencies will have higher rates of growth on average.

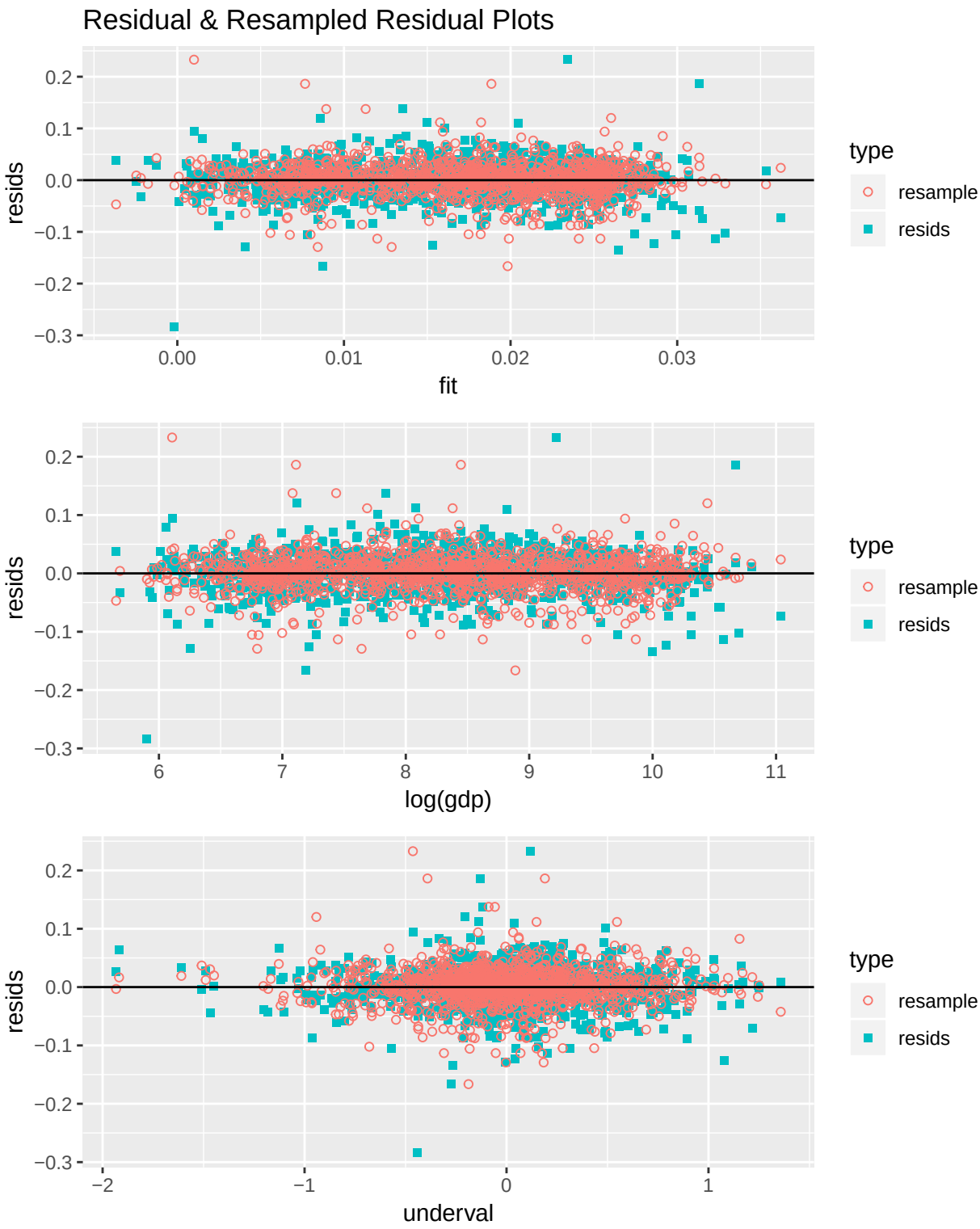
| | Estimate | Std. Error |
|---|---|---|
| (Intercept) | -0.035 | 0.007 |
| log(gdp) | 0.006 | 0.001 |
| underval | 0.005 | 0.002 |




To assess the residuals, a Q-Q plot was produced. A strong deviation from normality is evident due to the heavy tails.

The following residual plots show, in descending order, the residuals versus the fitted values, the residuals versus the predictor `log(gdp)`, and the residuals versus the undervaluation index. Included in each plot is also a random resampling of the reisiduals. If the distribution of the residuals is homoskedastic and unbiased, then both sets of points should have the same distribution. It is difficult to tell, but there is not strong

evidence of assumption violations. The residuals in all three plots are centered about 0 with no trends, and the variability is consistent. The resampled residuals seem to match the distribution of the original residuals.




(It could be argued that there is an indication of heteroskedacity. Particularly in the Undervaluation plot. It just seems so subtle... Maybe my eyes are going bad.)

3. Repeat the linear regression but add as predictors the country, and the year. Use `factor(year)`, not year, in the regression formula.
    - Report the coefficients for log GDP and undervaluation, and their standard errors in a table. Interpret the coefficients.
    - Explain why it is more appropriate to use `factor(year)` in the formula instead of just year.
    - Plot the coefficients of `factor(year)` over the years.
    - Check the residuals and report your findings. How do the residuals differ from before
    - Does this model support the idea that low GDP countries 'catch up'? Of undervaluation boosting growth?

The following table displays our coefficients for the `log(gdp)` and undervaluation index for the new model that includes the year and country. Interpreting the `log(gdp)` coefficient is a bit tricky. With logs, we should consider a percentage increase in the predictor. The increase in growth rate for a 10% increase in GDP is 0.00276 or 0.276%; this is assuming that undervaluation is held constant and we are looking at a single country within a single time period. The undervaluation index indicates a 0.014 increase in the growth rate should all other factors (GDP, country, year) be held constant.

| | Estimate | Std. Error |
|---|---|---|
| (Intercept) | -0.236 | 0.024 |
| log(gdp) | 0.029 | 0.003 |
| underval | 0.014 | 0.003 |

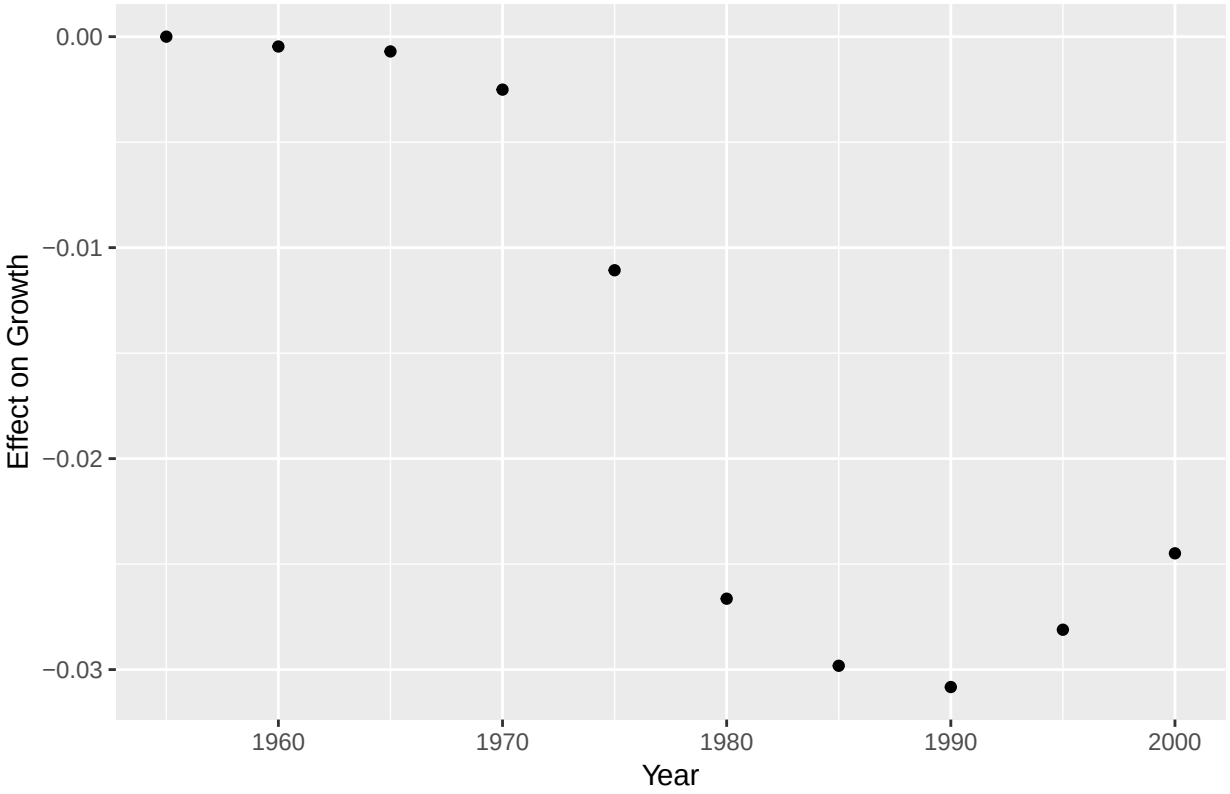
As far as the question of low GDP countries "catching up", we must again answer negatively to that. The `log(gdp)` coefficent was positive indicating that smaller GDPs are associated with less growth, and higher GDPs are associated with higher growth. I suppose this supports the idea that the more money you have, the more your money can grow. (It takes money to make money.)

For the model calculations, we are calculating the model such that each year is a different group. This allows the effect of the year to vary from year to year. If we treated year as a continuous variable, we be assuming that the effect of time constant despite the year. To my knowledge, economies unversally grow and decline over time, so it would be folly to assume that effect is constant. Most likely the effect of year would be averaged out.

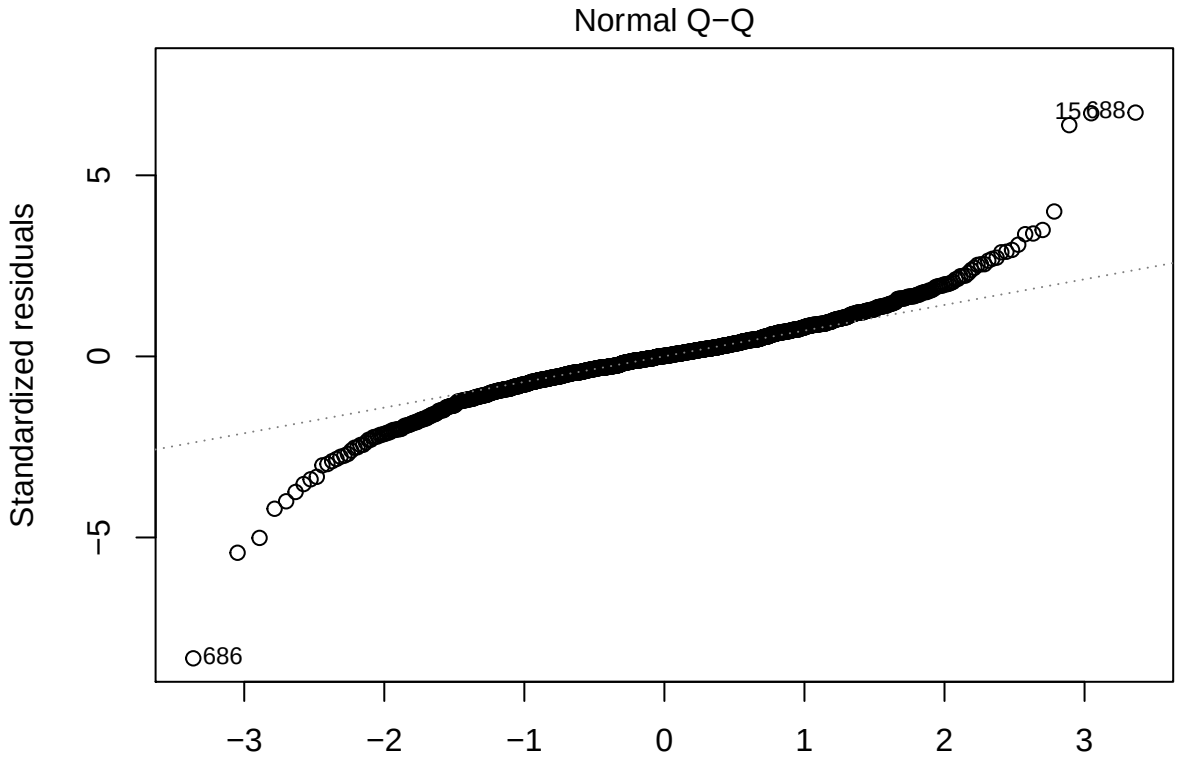
Actually... lets check that! The following table is the coefficients for a model that is the same is model 2, but year is treated as a continuous variable. In the table, we see the coefficient for year is -0.0007, a fairly minor effect, even if it were significant.

| | Estimate | Std. Error |
|---|---|---|
| (Intercept) | 1.2348 | 0.1536 |
| log(gdp) | 0.0281 | 0.0031 |
| underval | 0.0128 | 0.0030 |
| year | -0.0007 | 0.0001 |

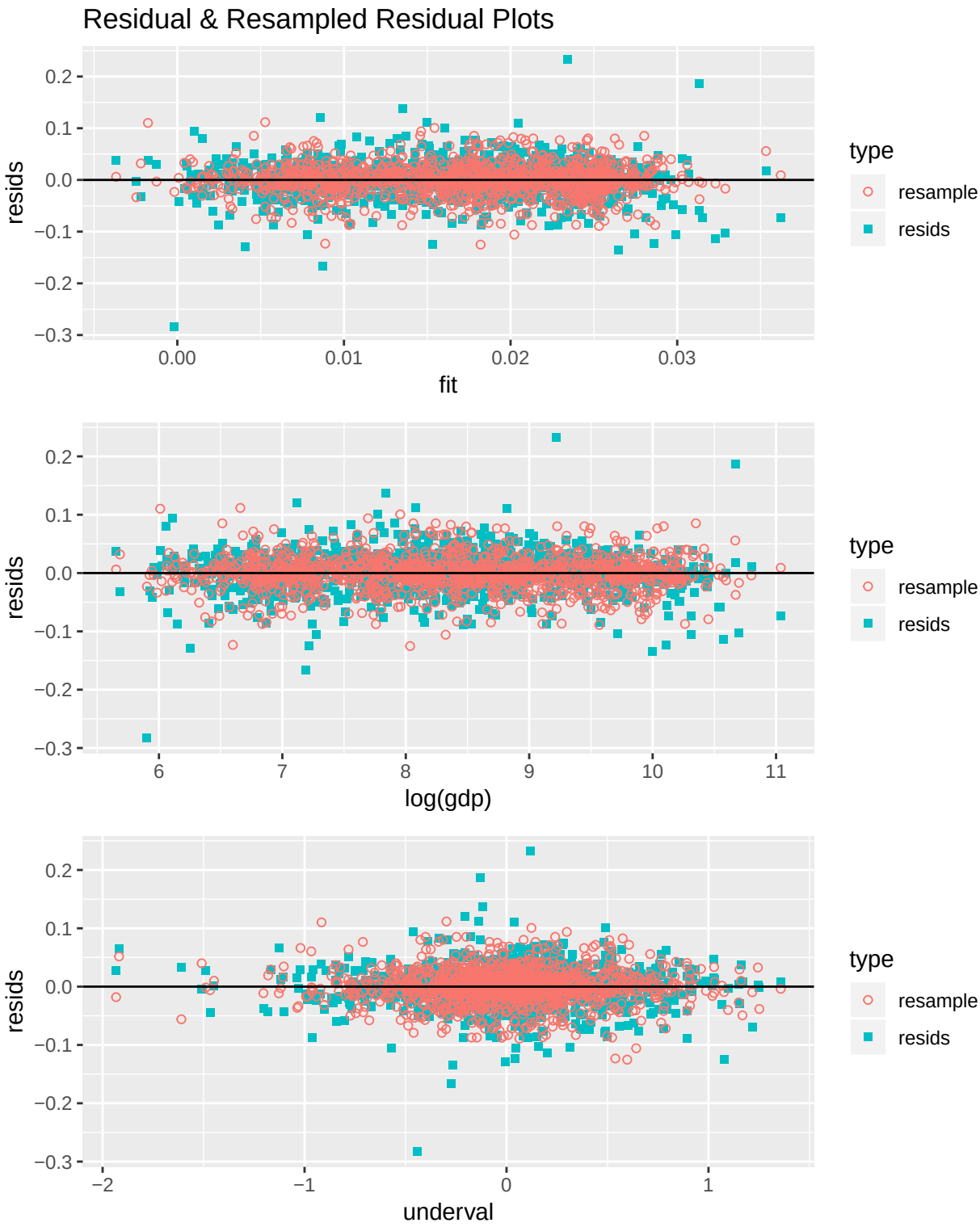
In comparison to that mode, look to the following graph which depicts the effect that each 5 year time period is estimated to have on the growth rate, assuming all other variables (GDP, undervaluation, country) are held constant. From 1955 to 1970, there is a fairly minor effect (near 0), but in 1975, that is now 0.01 decrease (-0.01) on the growth rate. This effect is at the highest magnitude of -0.03 in 1990. This coincides with a early 1990s recession that affected a large part of the Western world.




The following plots are used to check assumptions on residuals.




First, the Q-Q plot above, and then the reisdual plots below: residuals vs fitted, and residuals vs predictors. From the Q-Q plot, we have the same issue, heavy tails indicating non-normality. In the proceeding residual plots original residuals and resampled residuals are given for each residual plot. The results are the same as before, no issues are visible in the plots; the variance appears constant and the residuals are centered at 0.


Residual & Resampled Residual Plots
resids
fit
log(gdp)
underval
type
resample
resids

4. Does adding in year and country as covariates improve the predictability of a linear model which includes log GDP and under-valuation?

- What are $R^2$ and adjusted $R^2$ of the two models? Does this tell us anything about which model is better? Explain.
- Use leave-one-out cross-validation to find the mean squared errors of the two models. Which one actually predicts better?
- Explain why 5-fold cross validation may be inappropirate here. You do not need to actually do it.

Model 1 (the one without year and country) has an $R^2$ and Adjusted $R^2$ of 0.049 and 0.047 respectively. Compared to the values for Model 2 (year and country included as factors), 0.429 and 0.332 respectively. This, at first glance, tells indicates that Model 2 is much better than Model 1 by a fair margin, but to be sure, we will check using LOO-CV.

Cross validation gives an estimated risk of 0.00103 for model 1, and estimated risk of 0.00095 for model 2. This indicates that model 2 is slightly better than model 1. The increase in $R^2$ from model 1 to model 2 was quite large, but the benefit of model 2 when considering the estimated risk is much smaller.

As for the question about 5-fold cross validation, the main issue has to do with the countries. Some countries have less than 5 observations (minimum 2). This means that unless the folds were carefully chosen, sometimes a model would be estimated without a particular country in the training data, then when it comes to the validation data, we would not be able to get a prediction since there was no coefficient for that country.

## Rubric

Probelm 1 (6/6): Each question or bullet point has been addressed. Questions are answered accurately. Any plot produced is explained and variable relationships are described. Outliers are addressed.

For problems 2 & 3, residuals do not need to be resampled. Just discussed.

Problem 2 (7/7): Coefficients and their standard errors are reported. Coefficients are correctly interpreted. Residuals are analyzed through plots: Q-Q Plot, Residuals vs Fitted, and Residuals vs Predictors. The assumptions of normality, bias, and homoskedasticity are discussed and any issues (or lack thereof) with these assumptions are explained.

Problem 3 (7/7): Requested oefficients and their standard errors are reported. Coefficients are correctly interpreted. Question regarding `factor(year)` is correctly answered and plot of `year` coefficients is given. Residuals are analyzed through plots: Q-Q Plot, Residuals vs Fitted, and Residuals vs continuous predictors. The assumptions of normality, bias, and homoskedasticity are discussed and any issues (or lack thereof) with these assumptions are explained.

Problem 4 (5/5): Question regarding $R^2$ is answered and a reasonable explanation is given. Cross validation is correctly applied the correct model is chosen. Explanation for 5-fold CV is given.