Income Disparity as a Function of Other Socioeconomic Factors

Background

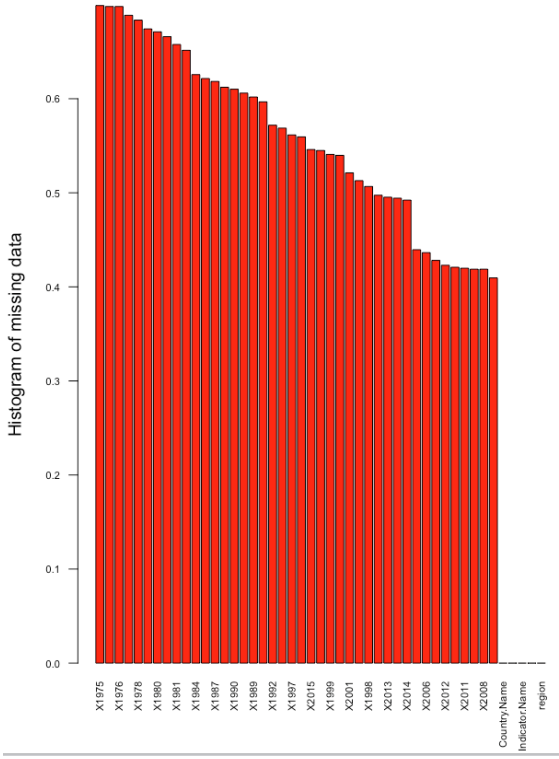
Income disparity can be modeled as a linear function of multiple factors. This study investigates the impact of gender disparity, infant mortality rate and adult literacy on income disparity. Specifically, income disparity (ID) is modeled by taking the top 10% income share minus the lowest 10% income share.

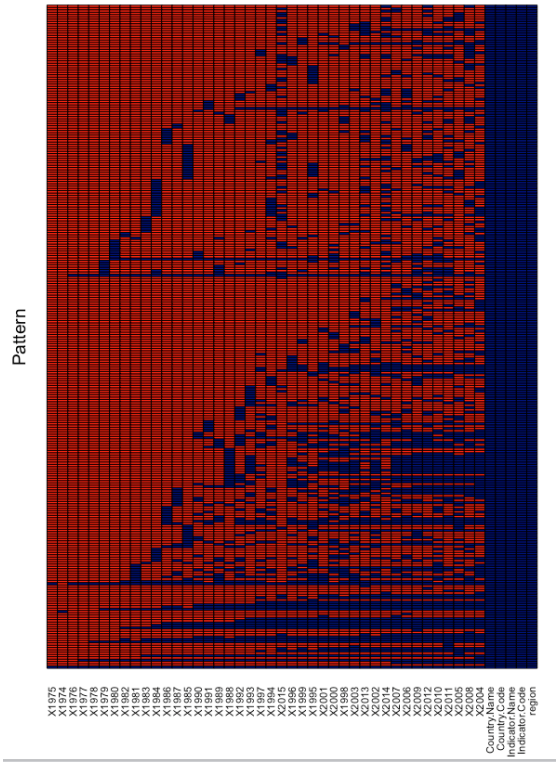
Problem Statement

This study answers the question, “Do gender disparity, education, and health factors influence income disparity?”, and “If so, use multivariable linear regression to model the relationship.”

Constraints and Limitations

The biggest constraints on the study come from the fact that there is a TON of missing data from the original data sets. These graphs show the amount of missing data overall. The histogram to the right shows the percentage of missing data by factor level. Almost 80% of the data are missing from the years 1975 to 1981.





The sparsity graph to the left, the red rectangles represent missing data, and the blue squares show existing data.

There are very strange omissions -- such as all literacy data from countries classified as being in the region “North America.” and “Western Europe”. Ultimately, we imputed the missing data where possible (more on the technique below), and where not possible, we focused on the areas where we DO have data. It allowed us to complete the analysis and also

reinforced the observational nature of this study; the results will only apply to those years and regions studied.

Data Set Description

There were 4 world bank data sources:

- World Bank EdStats All Indicator Query
  - https://data.worldbank.org/data-catalog/ed-stats
- World Bank Gender Statistics Database
  - https://data.worldbank.org/data-catalog/gender-statistics
- World Bank Poverty and Inequality Indicators
  - https://data.worldbank.org/data-catalog/poverty-and-equity-database
- World Bank Health Stats
  - https://data.worldbank.org/data-catalog/health-nutrition-and-population-statistics

These data sets contained many variables that could be used to explain income disparity. In the effort to maintain as small a model as we could and still represent the data, we narrowed the variables to three broad indicators. Our explanatory variables, from the datasets above, are:

- Education – we are using adult literacy rate (percentage) to represent education levels in each region
- Gender disparity – what percentage of the population is female in each region
- Health – the infant mortality rate (# of infant deaths per 1000 live births; we converted this into percentages) is our measure of the level of health in the region

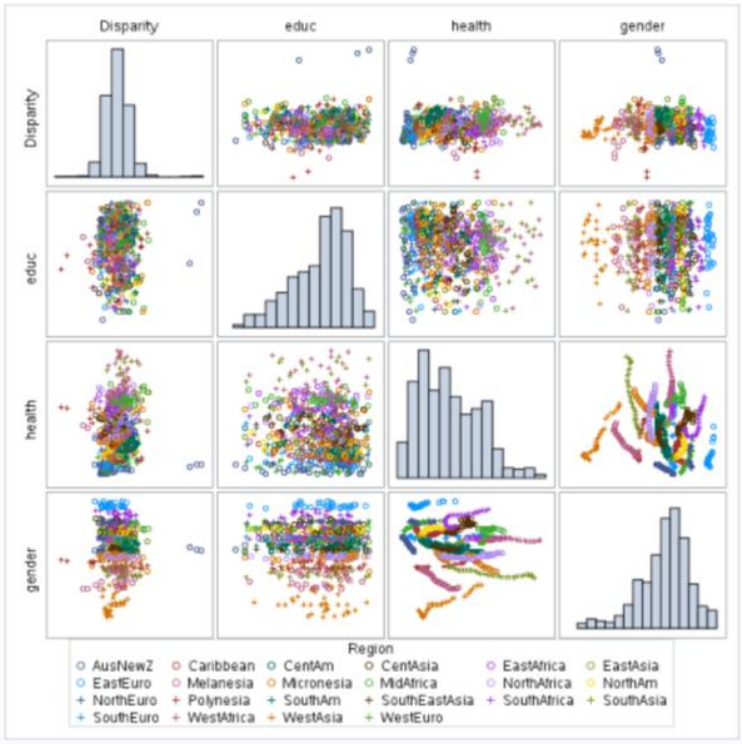
For the response variable we chose two factors in the Poverty and Inequality data, highest ten percent income share and lowest ten percent income share. Using these factors, we calculated the difference and used that value for our response variable (disparity).

After merging the four data sets by year, we then imputed the missing values, taking the average of the previous and next values per indicator type per country, across years. We then labeled the countries by regions using the Maxmind database, then we took the average of the data per year per region and used those as our final data set.

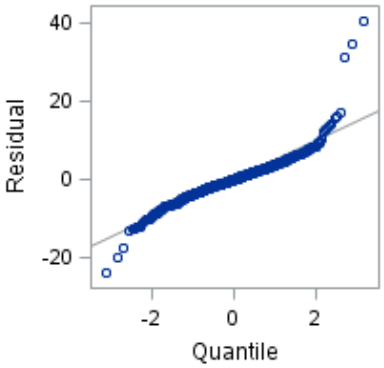
Note: There is some risk in taking the average of average-imputed data, in that it tends to significantly decrease the within-group variance of the data, without moving the grand mean too much. By dropping the within-group variance we decrease the power of the model which increases our chance of rejecting H_0 when H_a is true is lower. We decided that it would be better to err on the conservative side, where a rejected null is more likely to be the accurate choice. Additionally, due to the sparsity of data prior to 1981, we decided to focus on the 33-year period from 1981 to 2014.

Exploratory Data Analysis

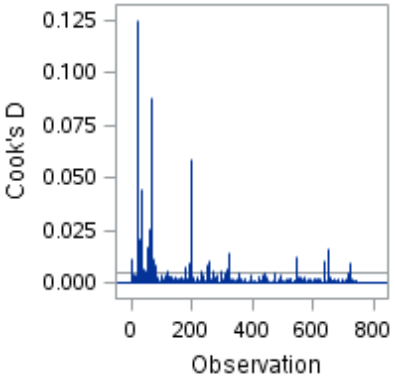
Looking at the initial scatterplots, we see some linear relationships between disparity and gender, but it's hard to see a consistent discernable pattern due to the number of regions. In general, however, there don't appear to be any non-linear relationships that are cause for concern.



The original data looked okay with respect to homoscedasticity, linearity and independence (more below), but had some issues with outliers and normality.



See the deviation in the tails of the qqplot to the left, and the Cook's D outliers to the right, that illustrate our concerns with normality and outliers.



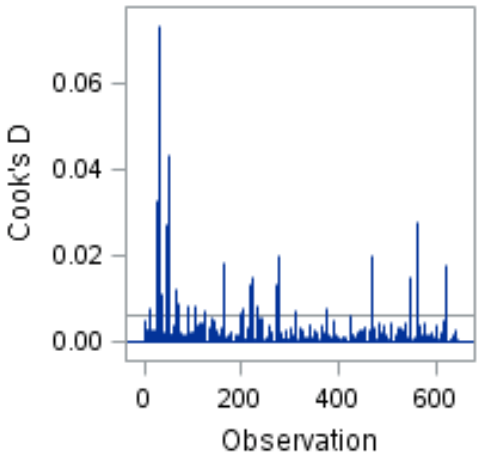
We identified 6 or so outliers that were worth investigating. When we analyzed the data (project2, in our code) on those outliers, we discovered that the strange outliers were mostly from 3 regions:

- Australia / New Zealand
- North America
- Western Europe

	Country.Name	Country.Code	Indicator.Name	Indicator.C	X1974	X1975	X1976	X1977	X1978	X1979	X1980	X1981	X1982	X1983	X1984	X1985	X1986	X1987
11	Australia	AUS	Adult literacy rate, population 15+ years, both sexes (%)	SE.ADT.LITR.	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
148	New Zealand	NZL	Adult literacy rate, population 15+ years, both sexes (%)	SE.ADT.LITR.	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
226	Australia	AUS	Population, female (% of total)	SP.POP.TOTL	49.477271	49.508548	49.541596	49.574293	49.604197	49.62972	49.649948	49.665697	49.679266	49.693762	49.711401	49.731856	49.754621	49.781201
355	New Zealand	NZL	Population, female (% of total)	SP.POP.TOTL	49.737669	49.75503	49.771084	49.788189	49.809535	49.837044	49.871655	49.911859	49.954041	49.993119	50.025642	50.050143	50.068334	50.083529
441	Australia	AUS	Mortality rate, infant (per 1,000 live births)	SP.DYN.IMR1	20.3	20	19.5	19.2	18.8	18.6	18.3	18.3	18.2	18.1	17.8	17.4	16.8	16.2
570	New Zealand	NZL	Mortality rate, infant (per 1,000 live births)	SP.DYN.IMR1	22.6	21.7	20.9	20	19.3	18.7	18.3	17.9	17.5	17.2	16.9	16.5	16.2	15.8
652	Australia	AUS	Income share held by highest 10%	SI.DST.10TH.	NA	NA	NA	NA	NA	NA	NA	23.33	NA	NA	NA	24.29	NA	NA
813	Australia	AUS	Income share held by lowest 10%	SI.DST.FRST.	NA	NA	NA	NA	NA	NA	NA	2.78	NA	NA	NA	2.71	NA	NA

We referenced the original data, regarding those 3 regions, and found that countries in those regions were completely devoid of data for some indicators – most commonly Adult Literacy Rate.

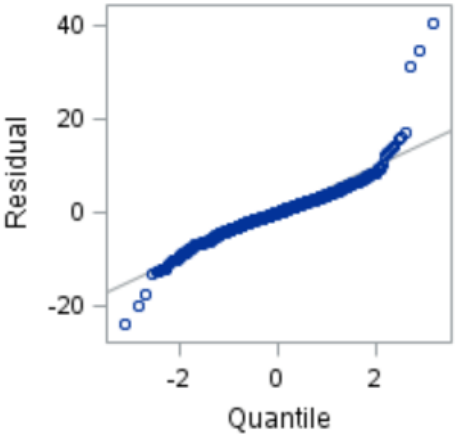
We felt that was sufficient cause to remove any outlier data associated with those 3 regions, and removing them resulted in a large improvement in both normality (see below, right) and outliers (see right).

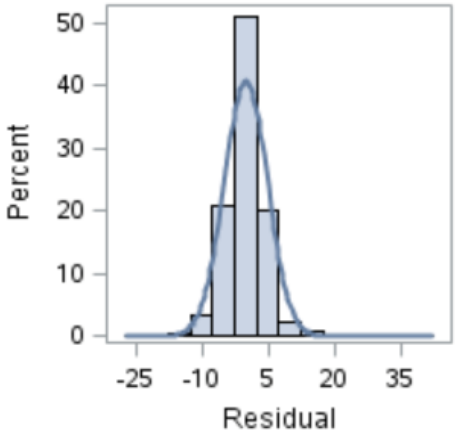


Addressing MLR Assumptions

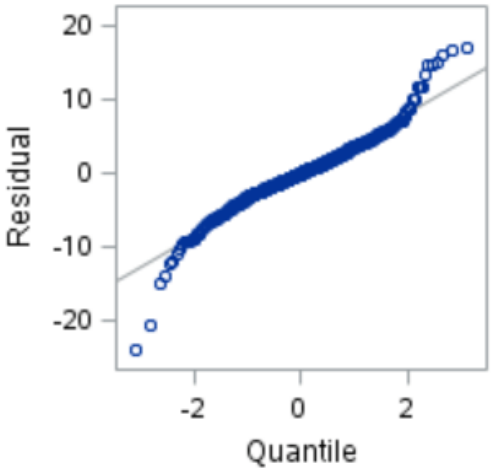
Normality

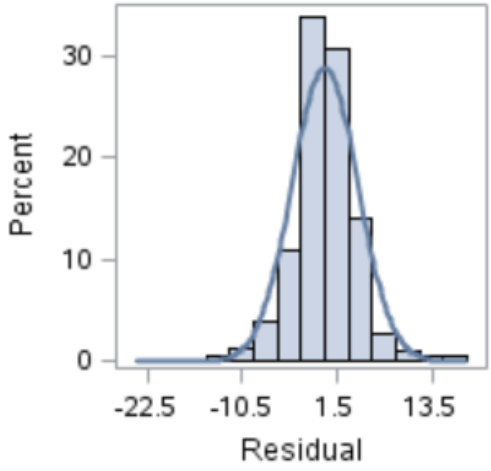
As mentioned above, with the untransformed, unpruned data (below, left), the normality looks okay along most the data, except for some outliers at the upper and lower ends of the qqplot.





We attempted to transform the variables (log, square, root transform of all 3 independent variables) to correct the last bit of deviation on the qqplot, but none of our transformations yielded better results. Removing the outliers, though, resulted in a slightly better qqplot (right), and a much higher R-squared (0.451, from 0.386)



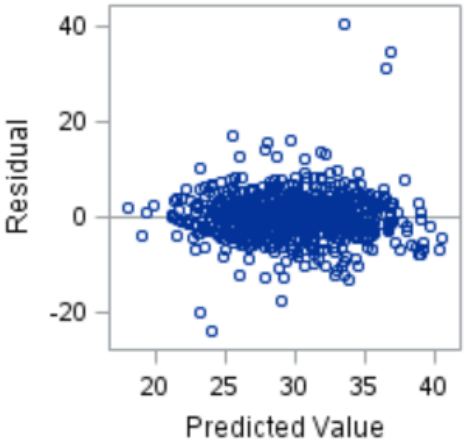


Moving forward with the analysis, we will assume normality of the residuals. Although there is more variation than we'd like, we suspect much of the variation comes from imputing a large portion of a sparse data set. As such, we're willing to accept some amount of uncertainty about the distribution of the data, since there are so many data points.

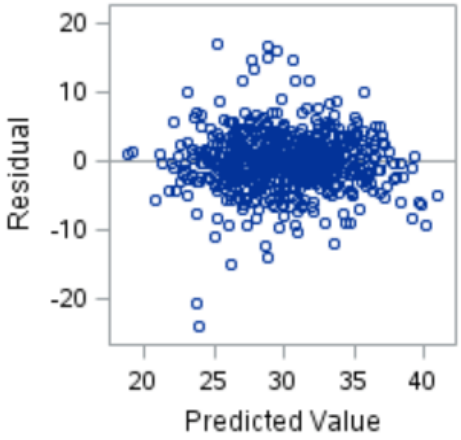
Independence

We assume independence of observations, but there may be latent serial effects at play, since these data are time-series data.

Homoscedasticity

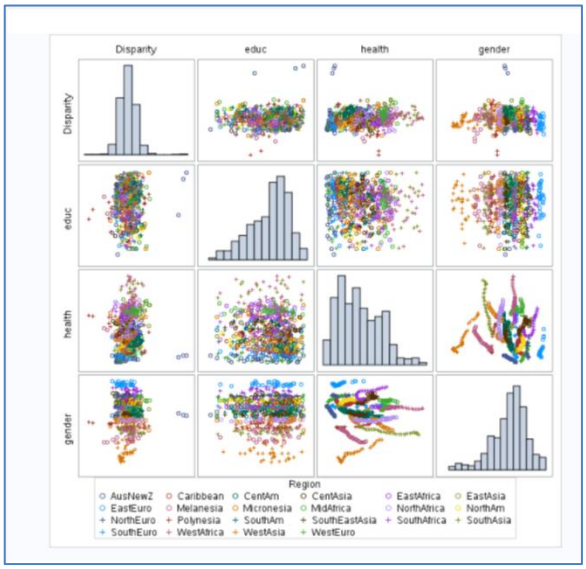


Variance prior to region pruning (left) looked relatively constant, with a few outliers. After the pruning (right) variance of residuals looked a little more consistent across predicted values, with much more well-behaved outliers.



Linearity

Aside from the strange, quasi-linear relationship between gender and health, there is no clear indication of the violation of the linearity assumptions in the scatterplot.



Since this study is focused on variables that inform income disparity, we consider the interaction of those two variables as part of the model, but not as a factor that violates the linearity assumption.

The low VIF, indicated in the table to the right, further confirms linearity and reinforces our assertion that correlation between these variables isn't going to be a problem.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	32.12527	9.89446	3.25	0.0012	0
educ	1	0.01711	0.01489	1.15	0.2510	1.02164
health	1	0.03156	0.00678	4.66	<.0001	1.02928
gender	1	-0.10849	0.19529	-0.56	0.5787	1.00760

Model Selection

We are now able to move forward with the following first model:

$$Y_{disparity} = \beta_0 + \beta_{Region} + \beta_{Year} + \beta_{educ} + \beta_{health} + \beta_{gender} + \varepsilon$$

The model results yielded a Rsquared of about 0.385 and gives us a baseline to work from while computationally selecting a better combination of predictive variables.

R-Square	Coeff Var	Root MSE	disparity Mean
0.385540	17.05998	5.094386	29.86161

To select a more accurate model, we used PROC GLMSELECT In SAS with LASSO to identify the most influential variables.

Specifically, we split the data into 50% train and 50% test sets, and then used AIC to choose and CV to stop.

After 15 steps, the lowest CV PRESS and highest R-squared was achieved, and the effects to the right were the most impactful in the model.

Taking this into account, we adjusted the model to reflect the best variables.

LASSO Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	AIC	ASE	Test ASE	CV PRESS
0	Intercept		1	1437.0353	29.8389	33.4576	9758.1902
1	region_SouthAm		2	1438.1480	29.7578	33.3789	9268.2200
2	year_avg1988		3	1436.5426	29.4305	33.1073	8837.4468
3	region_CentAsia		4	1433.0755	28.9411	32.7818	8432.2871
4	region_EastEuro		5	1422.5519	27.8504	32.0243	8029.2136
5	year_avg1994		6	1421.8599	27.6213	31.8390	7719.5241
6	region_MidAfric		7	1422.6515	27.5191	31.7535	7463.0237
7	region_CentAm		8	1409.9991	26.3097	30.7715	7216.9775
8	region_Polynesi		9	1398.6841	25.2567	30.0152	7105.6306
9	year_avg2007		10	1398.6585	25.1003	29.9086	6955.9022
10	year_avg1996		11	1399.0238	24.9747	29.8224	6771.8504
11	region_WestAsia		12	1399.8869	24.8878	29.7638	6633.5436
12	year_avg1983		13	1399.3551	24.6952	29.6300	6526.2196
13	health		14	1400.2382	24.6108	29.5652	6470.7651
14	year_avg1995		15	1394.7082	24.0488	29.1455	6362.0945
15	year_avg2012		16	1367.4608*	21.9852	27.6782	6256.7185*

* Optimal Value of Criterion

Selection stopped at a local minimum of the cross validation PRESS.

The new model, based on optimized variable selection, is:

$$\begin{aligned} Y_{disparity} = \beta_0 &+ \beta_{region\_SouthAm} + +\beta_{year\_avg1988} + \beta_{region\_CentAsia} + \beta_{region\_EastEuro} \\ &+ \beta_{year\_avg1994} + \beta_{region\_MidAfric} + \beta_{region\_CentAm} + \beta_{region\_Polynesi} + \beta_{year\_avg2007} \\ &+ \beta_{year\_avg1996} + \beta_{region\_WestAsia} + \beta_{year\_avg1983} + \beta_{health} + +\beta_{year\_avg1995} \\ &+ \beta_{year\_avg2012} + \varepsilon \end{aligned}$$

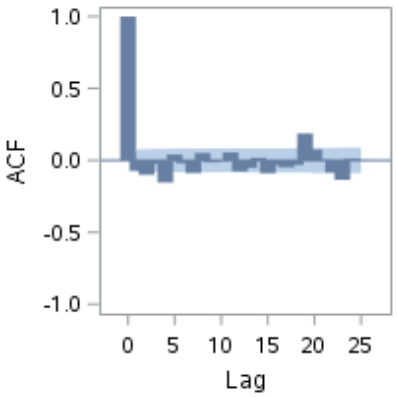
Using this model, we're able to explain much more (65.7%) of the disparity using these effects.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	1849.217561	142.247505	5.15	<.0001
Error	35	966.708775	27.620251		
Corrected Total	48	2815.926336			

R-Square	Coeff Var	Root MSE	Disparity Mean
0.656700	17.89681	5.255497	29.36555

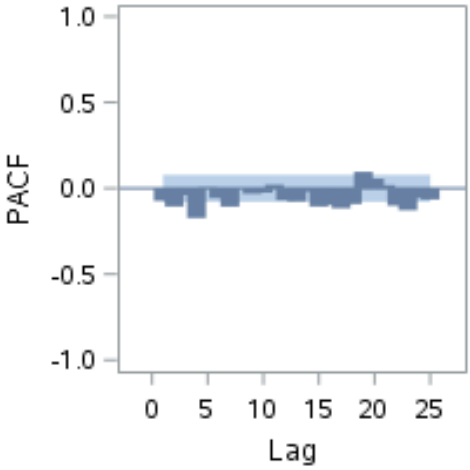
Furthermore, we show that the above model is statistically significant at the alpha = 0.05 level (F value = 5.15, $p < 0.0001$) as seen in the ANOVA table

Serial Correlation



It is possible that there is some serial correlation in the data that we did not address. We decided not to tackle the serial correlation and instead treat the year as an indicator variable instead.

This was partially because we had talked in class about not doing serial effects for this project, but also because a quick glance at the ACF and PACF plots in our initial model don't show any immediately recognizable cyclic behavior that was too concerning.



That being said, we find it unlikely that there is no serial correlation at all, especially for years that ended up in our model (such as the mid-90's) where 3 or more years are strung together and all strongly correlated with the dependent variable. This might be a good study for another project, but was not addressed here.

Conclusions

This is a cross-sectional, observational study. It is cross-sectional in the sense that we did not account for the time-series nature of the data – instead treating year as a categorical variable, and not performing serial correlation analysis.

Since this is an observational study, the results apply to only those countries and years observed, and does not indicate any level of causality.

The final solution equation is:

```
Disparity = 32.148
      - (0.229)health
      + (7.716)avg1983
      + (4.075)avg1988
      + (11.871)avg1994
```

```
+ (13.055)avg1995
+ (10.939)avg1996
+ (1.643)avg2007
+ (5.900)CentAm
+ (2.548)CentAsia
- (7.795)EastEuro
+ (17.915)MidAfrica
+ (0.957)Polynesia
+ (8.445)SouthAm
```

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	32.14841916	B	3.92576929	8.19	<.0001
health	-0.22874653		0.09036042	-2.53	0.0160
Year avg1983	7.71629650	B	4.91746056	1.57	0.1256
Year avg1988	4.70493912	B	4.21755470	1.12	0.2722
Year avg1994	11.87101508	B	3.52957231	3.36	0.0019
Year avg1995	13.05528808	B	3.41630087	3.82	0.0005
Year avg1996	10.93931520	B	3.33453303	3.28	0.0023
Year avg2007	1.64274604	B	2.83578869	0.58	0.5661
Year avg2012	0.00000000	B	.	.	.
Region CentAm	5.89952383	B	2.83911354	2.08	0.0451
Region CentAsia	2.54798557	B	3.44088023	0.74	0.4639
Region EastEuro	-7.79485723	B	3.65237420	-2.13	0.0399
Region MidAfrica	17.91460654	B	5.18858450	3.45	0.0015
Region Polynesia	0.95672049	B	2.84166048	0.34	0.7384
Region SouthAm	8.44542685	B	2.82145033	2.99	0.0050
Region WestAsia	0.00000000	B	.	.	.

Using this equation with the coefficient estimates, we can see that the most highly correlated region with income disparity is MidAfrica, followed by South America. The Polynesian region has the smallest effect of any region in our custom model. Certain years also very much informed income disparity -- 1983, 1988, 1994-1996, and 2007.

Effects on income disparity holding all other values at 0:

Based on our results, this model can be interpreted as follows:

As infant mortality increases by 1% → income disparity decreases by 31.919%

During year 1983 → income disparity increased by 39.867%

During year 1988 → income disparity increased by 36.223%

During year 1994 → income disparity increased by 44.019%

During year 1995 → income disparity increased by 45.203%

During year 1996 → income disparity increased by 43.087%

During year 2007 → income disparity increased by 33.791%

Living in Central America increases income disparity by 38.048%

Living in Central Asia increases income disparity by 34.696%

Living in Eastern Europe increases income disparity by 39.943%

Living in Middle Africa increases income disparity by 50.063%

Living in Polynesia increases income disparity by 33.105%

Living in South America increases income disparity by 40.593%

This study has given an overview of income disparity around most of the world and attempted to describe the differences with gender disparity, infant mortality and literacy rate. With an R-square of .66 we can explain 66% of the measured income disparity with these variables and accounting for year and region.

There were a couple of clear lessons from the study. Living in Middle Africa in 1995, there was an enormous difference between the very poor and the very rich. This isn't necessarily surprising and confirms the common idea of an incredibly destitute population and the upper classes living a life of luxury. We also have the strange situation where as infant mortality increases the economic differences decrease. We have no explanation for this result and it could be a topic for further research.

APPENDIX

EDA / Data Munging R Code

```
library(dplyr)
library(countrycode)
p1file <- "/Users/dyer/Documents/smu/Stats2-6372-
401/projects/project1/data/project1_clean.csv"
p1<-read.csv(p1file, stringsAsFactors = FALSE)
codes.of.origin <- p1$Country.Code
p1$region <- countrycode(codes.of.origin, "iso3c", "region")
weirdos <- p1[is.na(p1$region),]      # just curious
p2 <- p1[!is.na(p1$region),]      # Get rid of everything that doesn't classify into
a region

f1 <- function(dat) {
    N <- length(dat)
    na.pos <- which(is.na(dat))
    if (length(na.pos) %in% c(0, N)) {
        return(dat)
    }
    non.na.pos <- which(!is.na(dat))
    intervals  <- findInterval(na.pos, non.na.pos,
                               all.inside = TRUE)
    left.pos   <- non.na.pos[pmax(1, intervals)]
    right.pos  <- non.na.pos[pmin(N, intervals+1)]
    left.dist  <- na.pos - left.pos
    right.dist <- right.pos - na.pos

    dat[na.pos] <- ifelse(left.dist <= right.dist,
                          dat[left.pos], dat[right.pos])
    return(dat)
}
```

```
# Just keep relevant columns and remove 'countries' without associated regions
p3 <- p2[,c(1:46,62)] %>%
    filter(!is.na(region))
names(p3)

# f1 is the imputation function above.
p4 <- lapply(p3, FUN=f1)
p4 <- as.data.frame(p4)
p4$Country.Name <- p3$Country.Name
p4$Country.Code <- p3$Country.Code
p4$Indicator.Name <- p3$Indicator.Name
p4$Indicator.Code <- p3$Indicator.Code
write.csv(p3, "/Users/dyer/Documents/smu/Stats2-6372-
401/projects/project1/data/p3.csv")
write.csv(p4, "/Users/dyer/Documents/smu/Stats2-6372-
401/projects/project1/data/p4.csv")

# using dplyr to group by region and indicator and then compute the means.
p5 <- p4 %>% group_by(region, Indicator.Name) %>%            # group by region and
indicator.
    summarise(
        avg1974 = mean(as.numeric(X1974), na.rm=TRUE),
        avg1975 = mean(as.numeric(X1975), na.rm=TRUE),
        avg1976 = mean(as.numeric(X1976), na.rm=TRUE),
        avg1977 = mean(as.numeric(X1977), na.rm=TRUE),
        avg1978 = mean(as.numeric(X1978), na.rm=TRUE),
        avg1979 = mean(as.numeric(X1979), na.rm=TRUE),
        avg1980 = mean(as.numeric(X1980), na.rm=TRUE),
        avg1981 = mean(as.numeric(X1981), na.rm=TRUE),
        avg1982 = mean(as.numeric(X1982), na.rm=TRUE),
        avg1983 = mean(as.numeric(X1983), na.rm=TRUE),
        avg1984 = mean(as.numeric(X1984), na.rm=TRUE),
        avg1985 = mean(as.numeric(X1985), na.rm=TRUE),
        avg1986 = mean(as.numeric(X1986), na.rm=TRUE),
        avg1987 = mean(as.numeric(X1987), na.rm=TRUE),
        avg1988 = mean(as.numeric(X1988), na.rm=TRUE),
        avg1989 = mean(as.numeric(X1989), na.rm=TRUE),
        avg1990 = mean(as.numeric(X1990), na.rm=TRUE),
        avg1991 = mean(as.numeric(X1991), na.rm=TRUE),
        avg1992 = mean(as.numeric(X1992), na.rm=TRUE),
        avg1993 = mean(as.numeric(X1993), na.rm=TRUE),
        avg1994 = mean(as.numeric(X1994), na.rm=TRUE),
        avg1995 = mean(as.numeric(X1995), na.rm=TRUE),
        avg1996 = mean(as.numeric(X1996), na.rm=TRUE),
        avg1997 = mean(as.numeric(X1997), na.rm=TRUE),
        avg1998 = mean(as.numeric(X1998), na.rm=TRUE),
        avg1999 = mean(as.numeric(X1999), na.rm=TRUE),
        avg2000 = mean(as.numeric(X2000), na.rm=TRUE),
        avg2001 = mean(as.numeric(X2001), na.rm=TRUE),
        avg2002 = mean(as.numeric(X2002), na.rm=TRUE),
        avg2003 = mean(as.numeric(X2003), na.rm=TRUE),
        avg2004 = mean(as.numeric(X2004), na.rm=TRUE),
        avg2005 = mean(as.numeric(X2005), na.rm=TRUE),
        avg2006 = mean(as.numeric(X2006), na.rm=TRUE),
        avg2007 = mean(as.numeric(X2007), na.rm=TRUE),
```

```
        avg2008 = mean(as.numeric(X2008), na.rm=TRUE),
        avg2009 = mean(as.numeric(X2009), na.rm=TRUE),
        avg2010 = mean(as.numeric(X2010), na.rm=TRUE),
        avg2011 = mean(as.numeric(X2011), na.rm=TRUE),
        avg2012 = mean(as.numeric(X2012), na.rm=TRUE),
        avg2013 = mean(as.numeric(X2013), na.rm=TRUE),
        avg2014 = mean(as.numeric(X2014), na.rm=TRUE),
        avg2015 = mean(as.numeric(X2015), na.rm=TRUE))


write.csv(p5, "/Users/dyer/Documents/smu/Stats2-6372-
401/projects/project1/data/p5.csv")
```

SAS Code;

```
data project;
infile '/home/ddyer0/sasuser.v94/Stats2/Data/p5_new.csv' dlm=',' firstobs=2;
input year $ educ region $ top10 low10 health gender disparity;
run;

data project2;
set project;
healthperc = health / 1000.0;
run;

data project3; set project2;
if region = "AusNewZ" then delete;
if region = "NorthAm" then delete;
if region = "WestEuro" then delete;
log_health = log10(health);
log_disparity = log10(disparity);
log_educ = log10(educ);
log_gender = log10(gender);
sq_educ = educ*educ;
sq_gender = gender*gender;
RandNumber=ranuni(1);
run;

proc glm data = project2 PLOTS=(diagnostics residuals);
class year region;
model disparity = year region educ health gender;
run;

proc glm data = project3 PLOTS=(diagnostics residuals);
class year region;
model disparity = year region educ health gender;
run;

proc reg data = project3  plots(label) = (rstudentbyleverage cooksd)  ;
model disparity = educ health gender /  VIF; *CORRB INFLUENCE CLB;
```

```
run;

*breaking data into test & train sets;
data train;
set project3;
if RandNumber > 1/2 then delete;
run;

data test;
set project3;
if RandNumber < 1/2 then delete;
run;

*plot scatterplot;
proc sgscatter data=project3 ;
matrix Disparity  educ health gender / diagonal=(histogram) group= region;
run;

*check autocorrelation;
proc corr data=project3;
var disparity educ health gender ;
run;


/*Using Where statement to limit the values of region and year to the values
chosen in the model by GLMSELECT procedure but still maintain the power of
using all of the data */

data model; set project3;
where year in ('avg1983', 'avg1988', 'avg1994', 'avg1995', 'avg1996',
'avg2007', 'avg2012');
run;

proc print data=model; run;

proc glm data= model ;
where region in ('CentAm', 'CentAsia', 'EastEuro', 'MidAfrica', 'Polynesia',
'SouthAm', 'WestAsia');
class region year;
model disparity = health year region / solution;
run;
```