Data 621 Final Project: Predicting Fertility Rates

Authors: Vikas Sinha
Luisa Velasco
Dan Wigodsky
Sarah Wigodsky
Zhenni Xie

Introduction

- Why is fertility rates projects important?

- Total Fertility Rate (TFR)

Literature Review

A number of different future fertility rates projections are produced, corresponding with each underlying assumption.

- Medium-fertility Assumption
- High-fertility Assumption
- Low-fertility Assumption
- Constant-fertility Assumption
- Instant-replacement Assumption

Literature Review

- *Modelling Fertility: A Semi-Parametric Approach*

  - Oberhofer and Reichsthaler, 2004

- *The author present a categorical model of fertility based on Generalized Linear Model.*
- *Only one factor was used – the age of the mother*
- *Bernoulli Random Variable*
- *Local Likelihood Estimation*

Data Source

- UNData - A web-based data service for the global user community. It brings international statistical databases within easy reach of users through a single-entry point. Users can search and download a variety of statistical resources compiled by the United Nations (UN) statistical system and other international agencies.

- Data Link: http://data.un.org/Explorer.aspx?d=WHO

Data Exploration

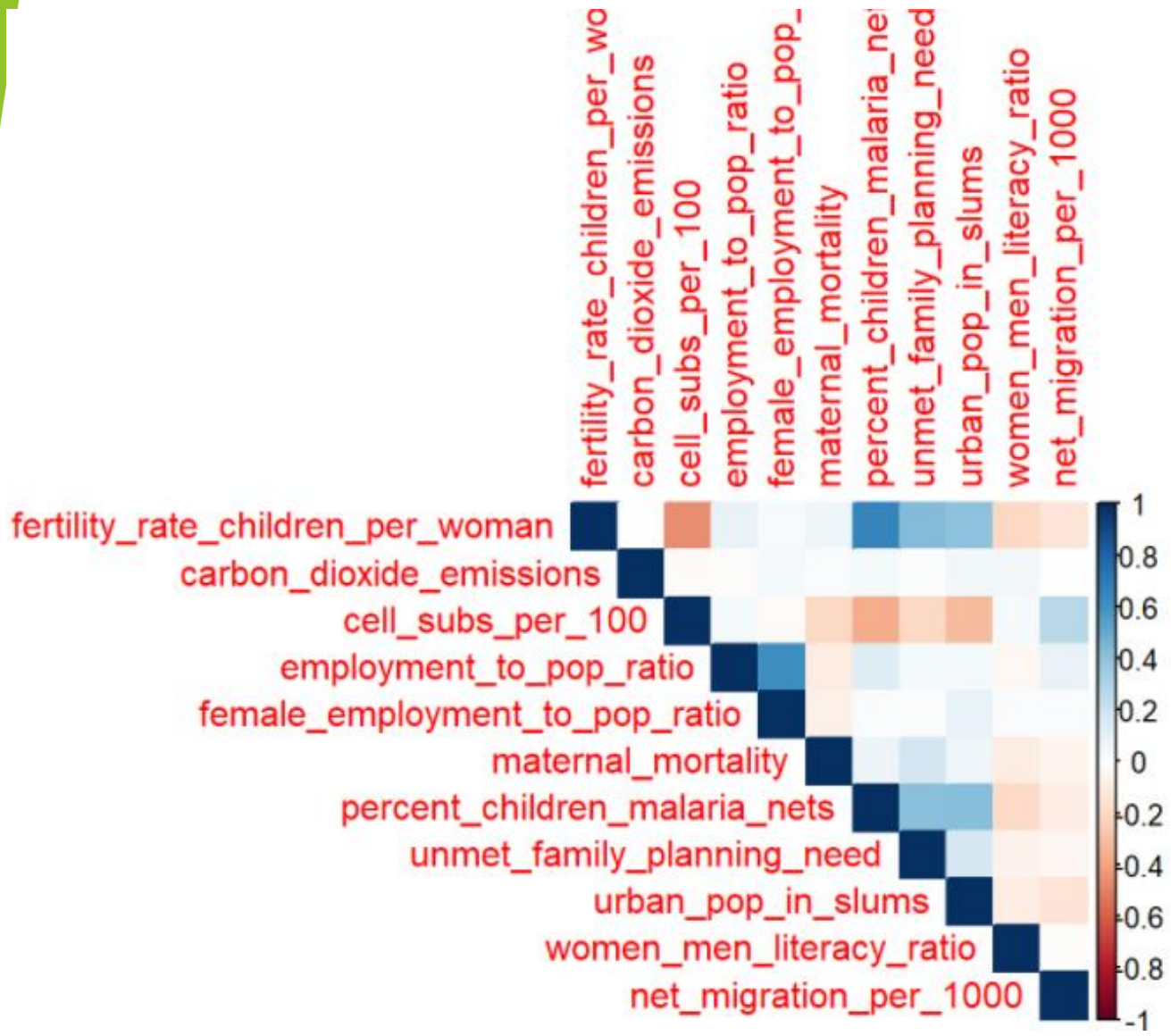
- The data set contains 214 rows, where each row represents from the data form a different country.
- The following variables are used to build the model.

- carbon_dioxide_emissions - carbon dioxide emissions in kilotonnes
- cell_subs_per_100 - Cell subscriptions per 100 population(2014)
- employment_to_pop_ratio - employment to total population ratio
- female_employment_to_pop_ratio - female employee to population ratio
- lowest_quint_income_share - poorest quintile's share in income
- maternal_mortality - maternal mortality per 100,000 live births
- percent_children_malaria_nets - percent of children sleeping under insecticide-treated bed nets
- unmet_family_planning_need - percent unmet family planning need
- urban_pop_in_slums - percent urban population living in slums
- women_men_literacy_ratio - women to men parity index, as ratio of literacy rates
- net_migration_per_1000 - net migration rate per 1000
- region_num - number that signifies the region the country resides in
    - 0 - Antarctica
    - 1 - Asia
    - 2 - Caribbean
    - 3 - Central America
    - 4 - Eastern Africa
    - 5 - Europe
    - 6 - European Union
    - 7 - Middle Africa
    - 8 - Middle East
    - 9 - North America
    - 10 - Northern Africa
    - 11 - Oceania
    - 12 - South America
    - 13 - South Africa
    - 14 - Western Africa

Challenge

How to handle the large number of missing values?

- Percent of Children Sleeping Under Insecticide Treated Bed Nets
- Carbon Dioxide Emissions
- Cell Subscriptions per 100 Population
- Employment to Population Ratio
- Female Employment to Population Ratio
- Lowest Quintile Income Share
- Maternal Mortality
- Unmet Family Planning Need
- Percentage of Urban Population Living In Slums
- The Literacy Ratio of Women to Men
- Net Migration Per 1000



Data Correlation

Build Models

- Backward Elimination - Linear Regression Model - Model 1

```
##
## Call:
## lm(formula = fertility_rate_children_per_woman ~ cell_subs_per_100 +
##     percent_children_malaria_nets + unmet_family_planning_need +
##     women_men_literacy_ratio, data = train1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8656 -0.6466 -0.1245  0.5997  2.9027
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   4.310632   0.618005   6.975 1.66e-10 ***
## cell_subs_per_100            -0.007759   0.002017  -3.847 0.000191 ***
## percent_children_malaria_nets 0.039203   0.005690   6.889 2.57e-10 ***
## unmet_family_planning_need    0.026109   0.011569   2.257 0.025791 *
## women_men_literacy_ratio     -1.546420   0.557318  -2.775 0.006387 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8735 on 123 degrees of freedom
## Multiple R-squared:  0.5824, Adjusted R-squared:  0.5688
## F-statistic: 42.89 on 4 and 123 DF,  p-value: < 2.2e-16
```

- Variables will be removed until every predictor has a p value below 0.05.
- Prediction from Model 1

```
## [1] 1.149661
```

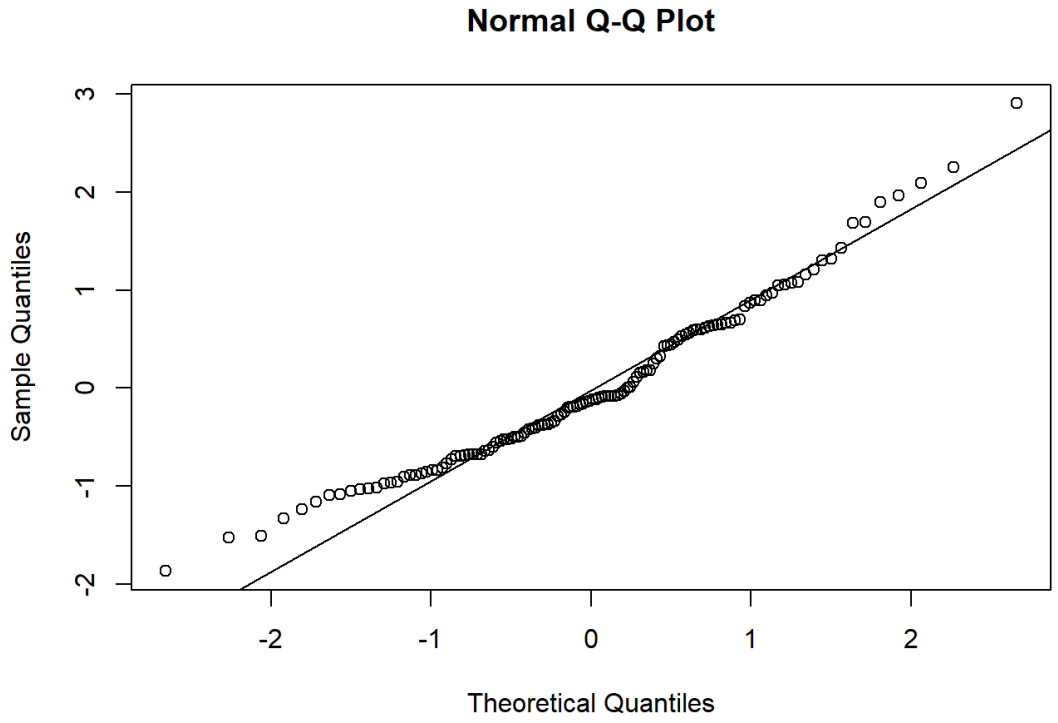
On average, the prediction for the fertility rate, is off by 1.15.

Model 1

- Muliticolinearity Test

```
##            cell_subs_per_100 percent_children_malaria_nets
##                     1.121764                     1.590792
##    unmet_family_planning_need      women_men_literacy_ratio
##                     1.404217                     1.066811
```

- Q-Q Plot



Model 1

- Cook's Distance

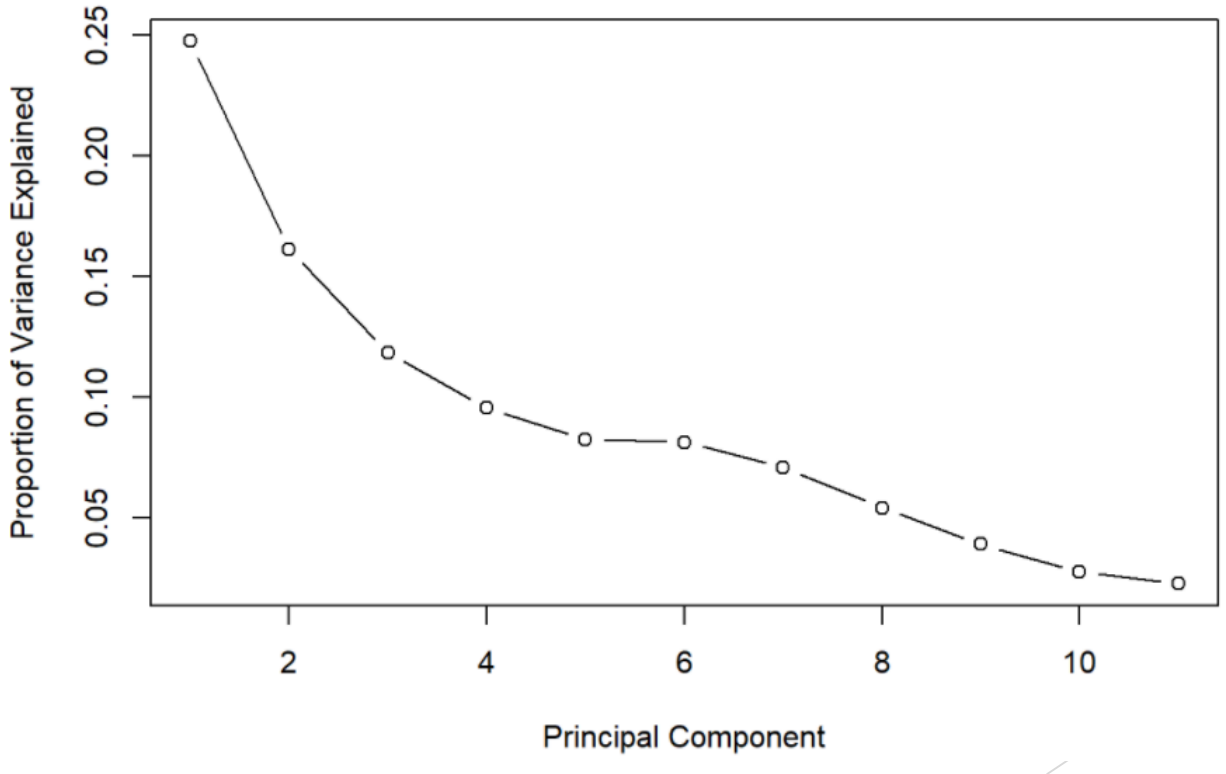
Model 2 - Principal Component Analysis

- Root Mean Square Error

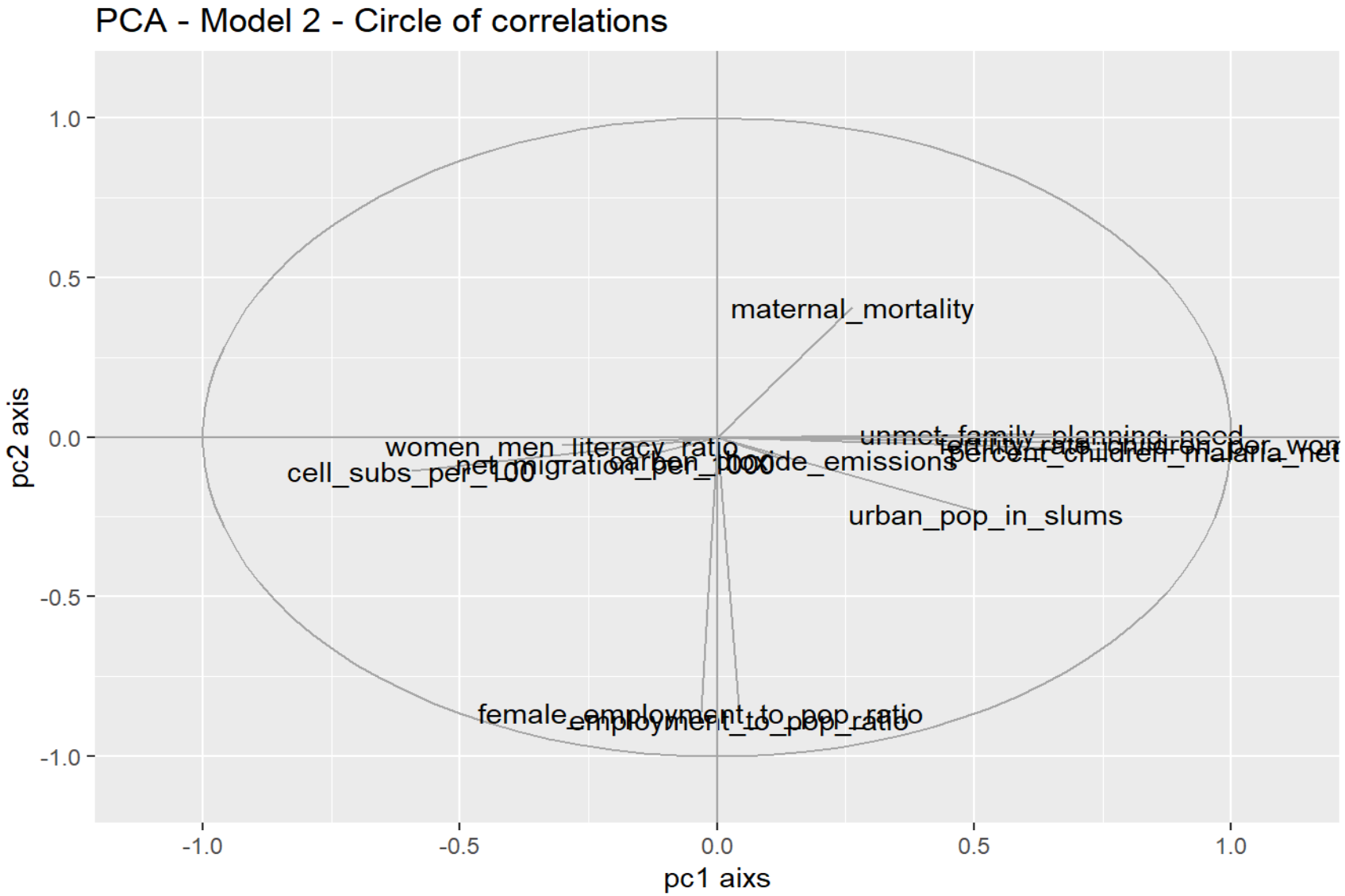
```
## [1] 0.8292908
```

On average, a prediction of the fertility rate is off by 0.83.

- Scree Plot



Model 2 - Principal Component Analysis



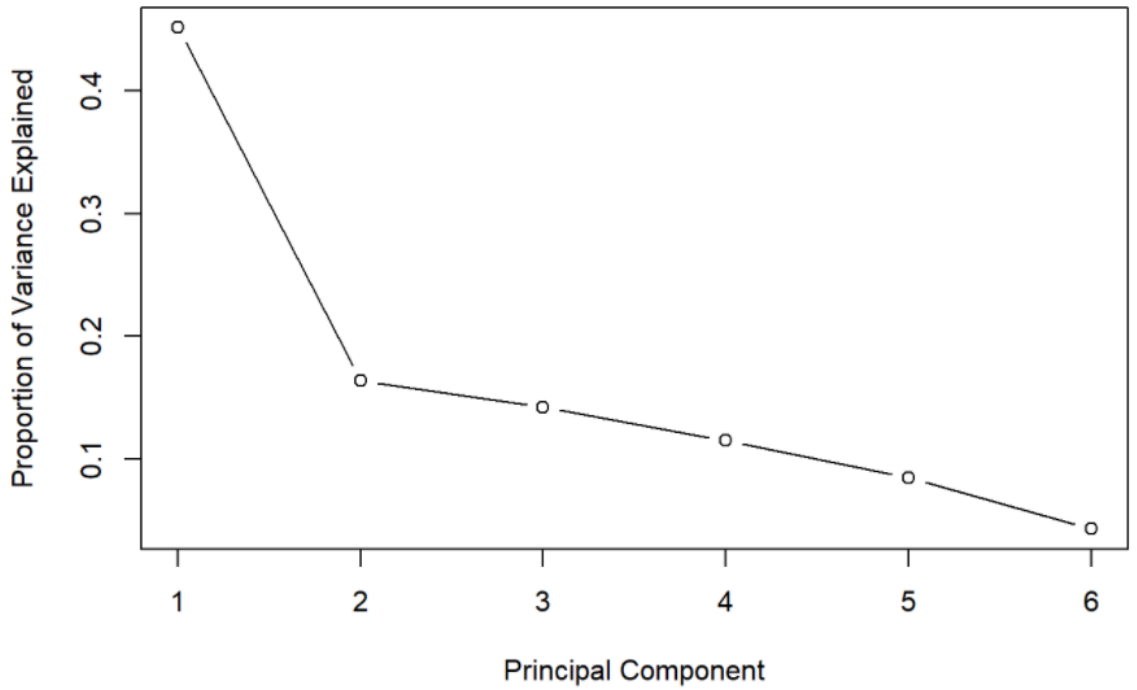
Model 3 - Principal Component Analysis (With only the variables correlated with fertility rate)

- Root Mean Square Error

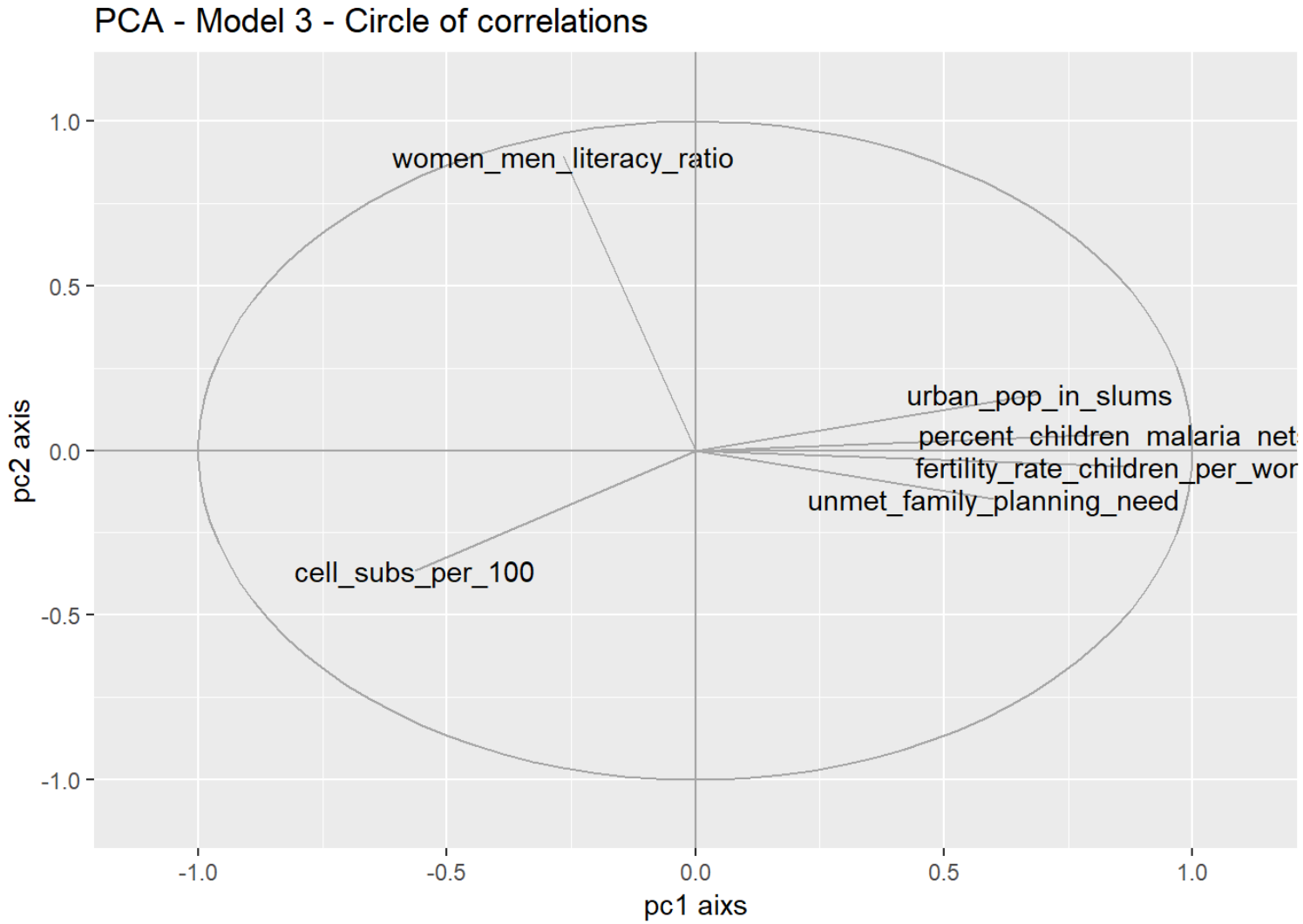
```
## [1] 0.6996821
```

On average, a prediction of the fertility rate is off by 0.70.

- Scree Plot



Model 3 - Principal Component Analysis (With only the variables correlated with fertility rate)



Model 4 - Regression Subset Selection

- Number of variables with highest adjusted (R^2). Variables marked with TRUE are the ones that will be chosen.

```
##                   (Intercept)       carbon_dioxide_emissions
##                          TRUE                          FALSE
##              cell_subs_per_100        employment_to_pop_ratio
##                          TRUE                          FALSE
## female_employment_to_pop_ratio             maternal_mortality
##                         FALSE                          FALSE
##  percent_children_malaria_nets     unmet_family_planning_need
##                          TRUE                           TRUE
##             urban_pop_in_slums       women_men_literacy_ratio
##                          TRUE                           TRUE
##         net_migration_per_1000
##                         FALSE
```

Model 4 - Regression Subset Selection

```
##
## Call:
## lm(formula = fertility_rate_children_per_woman ~ cell_subs_per_100 +
##     percent_children_malaria_nets + urban_pop_in_slums + unmet_family_planning_need +
##     women_men_literacy_ratio, data = train1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7201 -0.6103 -0.1320  0.5978  2.8103
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   4.142134   0.639026   6.482 2.01e-09 ***
## cell_subs_per_100            -0.007439   0.002040  -3.647 0.000392 ***
## percent_children_malaria_nets 0.037642   0.005886   6.395 3.09e-09 ***
## urban_pop_in_slums            0.003866   0.003744   1.033 0.303852
## unmet_family_planning_need    0.026379   0.011569   2.280 0.024339 *
## women_men_literacy_ratio     -1.540002   0.557203  -2.764 0.006599 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8733 on 122 degrees of freedom
## Multiple R-squared:  0.586,  Adjusted R-squared:  0.5691
## F-statistic: 34.54 on 5 and 122 DF,  p-value: < 2.2e-16
```

Model 4 - Regression Subset Selection

- Prediction from Model 4

```
## [1] 1.136879
```

- On average, the prediction for the fertility rate is off by 1.14.

Model 5- Count Models (Poisson, Negative Binomial, Zero Inflated)

```
##
## Call:
## glm(formula = fertility_rate_children_per_woman ~ ., family = "poisson",
##     data = count.train1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -44.194  -21.225   -3.254   13.892   62.403
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    8.005e+00  2.144e-02  373.37   <2e-16 ***
## carbon_dioxide_emissions      -3.881e-08  2.300e-09  -16.88   <2e-16 ***
## cell_subs_per_100             -5.350e-03  6.919e-05  -77.32   <2e-16 ***
## employment_to_pop_ratio        5.300e-03  2.941e-04   18.02   <2e-16 ***
## female_employment_to_pop_ratio -2.839e-03  1.949e-04  -14.56   <2e-16 ***
## maternal_mortality            -2.657e-04  1.176e-05  -22.59   <2e-16 ***
## percent_children_malaria_nets  1.246e-02  1.441e-04   86.50   <2e-16 ***
## unmet_family_planning_need     2.430e-02  3.754e-04   64.73   <2e-16 ***
## urban_pop_in_slums             2.097e-03  1.144e-04   18.33   <2e-16 ***
## women_men_literacy_ratio      -9.832e-01  1.356e-02  -72.53   <2e-16 ***
## net_migration_per_1000        -1.086e-02  3.666e-04  -29.62   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 136492  on 127  degrees of freedom
## Residual deviance:  64540  on 117  degrees of freedom
## AIC: 65674
##
## Number of Fisher Scoring iterations: 5
```

Model 5- Count Models (Poisson, Negative Binomial, Zero Inflated)

```
## Call:
## glm.nb(formula = fertility_rate_children_per_woman ~ cell_subs_per_100 +
##     employment_to_pop_ratio + female_employment_to_pop_ratio +
##     percent_children_malaria_nets + unmet_family_planning_need +
##     women_men_literacy_ratio, data = count.train1, init.theta = 1.864811808,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.1992  -0.9390  -0.1246   0.5274   2.1100
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     8.032977   0.636982  12.611  < 2e-16 ***
## cell_subs_per_100              -0.007210   0.001719  -4.194 2.75e-05 ***
## employment_to_pop_ratio         0.014971   0.009020   1.660  0.09697 .
## female_employment_to_pop_ratio -0.014110   0.006294  -2.242  0.02497 *
## percent_children_malaria_nets   0.016008   0.004972   3.219  0.00128 **
## unmet_family_planning_need      0.020704   0.009869   2.098  0.03592 *
## women_men_literacy_ratio       -0.760415   0.468137  -1.624  0.10430
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.8648) family taken to be 1)
##
##     Null deviance: 226.95  on 127  degrees of freedom
## Residual deviance: 141.75  on 121  degrees of freedom
## AIC: 2078.9
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.865
##           Std. Err.:  0.220
##
##  2 x log-likelihood:  -2062.949
```

Model 5- Count Models (Poisson, Negative Binomial, Zero Inflated)

```
## Call:
## zeroinfl(formula = fertility_rate_children_per_woman ~ . | percent_children_malaria_nets,
##     data = count.train1)
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -28.0672  -6.0188  -0.9492   5.2420  49.3347
##
## Count model coefficients (poisson with log link):
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    7.955e+00         NA      NA       NA
## carbon_dioxide_emissions      -3.777e-08  0.000e+00     Inf   <2e-16 ***
## cell_subs_per_100             -5.048e-03         NA      NA       NA
## employment_to_pop_ratio        5.279e-03         NA      NA       NA
## female_employment_to_pop_ratio -2.649e-03         NA      NA       NA
## maternal_mortality            -2.571e-04         NA      NA       NA
## percent_children_malaria_nets  1.255e-02         NA      NA       NA
## unmet_family_planning_need     2.417e-02         NA      NA       NA
## urban_pop_in_slums             2.302e-03         NA      NA       NA
## women_men_literacy_ratio      -9.758e-01         NA      NA       NA
## net_migration_per_1000        -1.065e-02         NA      NA       NA
##
## Zero-inflation model coefficients (binomial with logit link):
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -4.554         NA      NA       NA
## percent_children_malaria_nets  -55.149         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 28
## Log-likelihood: -3.252e+04 on 13 Df
```

Model 5- Count Models (Poisson, Negative Binomial, Zero Inflated)

- Prediction from Count Models

```
##        count.models       count.rmse
## 1           Poisson 1.20702379052436
## 2 Negative Binomial 1.20716261676315
## 3     Zero Inflated 1.20701084749699
```

- On average, the prediction for the fertility rate is off by 1.21.

Discussion and Conclusion

- It is possible to predict the fertility rate in a country based on the percent of children sleeping under insecticide treated bed nets, the percentage of the urban population in slums, unmet family planning need, the number of cellular subscriptions and the ratio of the literacy rate between women and men.
- Seven different models were created and gave relatively similar results, and the model with lowest root mean square error was model 3.
- Higher fertility rates are associated with having more children sleeping under insecticide treated bed nets, higher percentages of the urban population living in slums and higher unmet family planning need.
- From our analysis, it is not possible to ascertain whether a high fertility rate is the effect of these variables, the cause of these variables or simply correlated with them.
- High fertility rates are associated with higher poverty levels and lower levels of education between women and men.

What to do next?

- A source of further research would involve uncovering the nature of these connections to identify causation for fertility rates, not just correlations.

References

- United Nations, Department of Economic and Social Affairs. World Population Prospects, The 2017 Revision. https://esa.un.org/unpd/wpp/Publications/Files/WPP2017_Methodology.pdf
- Alkema et al (2011). Probabilistic Projections of the Total Fertility Rate for All Countries. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3367999/.
- National Institute of Health (2011). NIH-funded study proposes new method to predict fertility rates. https://www.nih.gov/news-events/news-releases/nih-funded-study-proposes-new-method-predict-fertility-rates
- United Nations, Department of Economic and Social Affairs. Population Trends. http://www.un.org/en/development/desa/population/theme/trends/index.shtml
- Walter Oberhofer and Thomas Reichsthaler (2004). Modelling Fertility: A Semi-Parametric Approach.https://epub.uni-regensburg.de/4511/1/rdisb396.pdf

Appendix

- Rpub Link: http://rpubs.com/sew/451479

- Github Link: https://github.com/swigodsky/Data621/blob/master/finalProject_fertility.Rmd