Analysis of Incidences of Tuberculosis (TB)

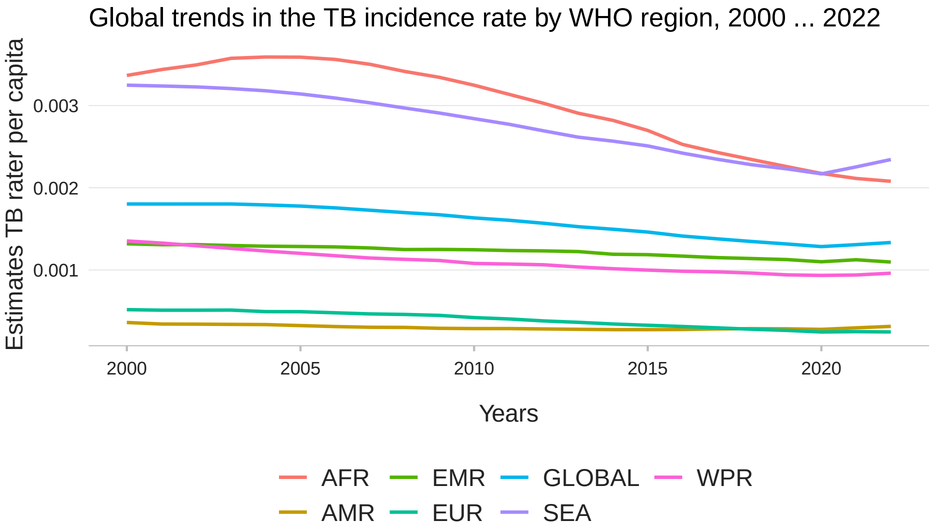
Bich Na Choi

2023-12-15

Using data from WHO TB estimates (https://www.who.int/teams/global-tuberculosis-programme/data) and any other public data sources:

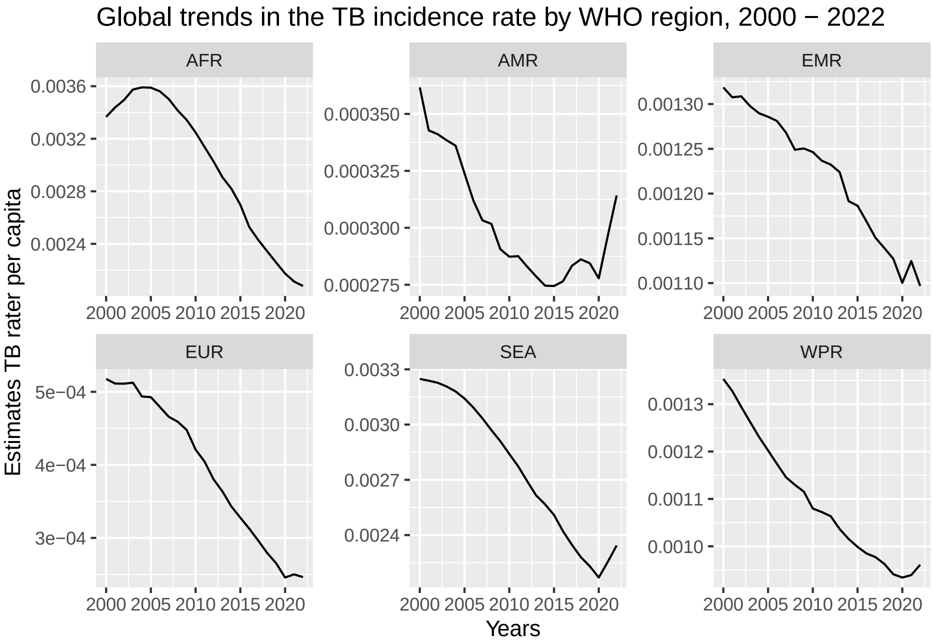
```
ggplot(tb_total,(aes(x=year,y=tb_inc_rate, group=g_whoregion,color=g_whoregion)))+
       geom_line(linewidth=0.8) +
       theme_gtb() +
       labs(x="Years",y="Estimates TB rater per capita",
            title = "Global trends in the TB incidence rate by WHO region, 2000 - 2022",
            colour = "Region")
```

1. Graph per capita TB incidence rates over time for all countries, grouped by WHO region;



The global Tuberculosis (TB) incidence rate is shown by the blue line. The global trend in TB incidence per capita is gradually decreasing. The per capita incidence rates in Africa and South-East Asia are above the global trend.The other regions are below the global trend. The Americas and Europe regions have the lowest TB incidence rates in all years. Trends in TB incidence rates in each region tend to be decreasing; however, some regions have shown an increase in the last 3 years. To explore the trend at regional level, we provides plots separately.

```r
ggplot(tb_region,(aes(x=year,y=tb_inc_rate)))+
    geom_line()+
    labs(x="Years",y="Estimates TB rater per capita",
            title = "Global trends in the TB incidence rate by WHO region, 2000 - 2022") +
    facet_wrap(~g_whoregion, scales = "free")
```



At the regional level, the TB incidence rates show an overall downward trend in all six regions, but the three (Americas, South-East Asia and Western Pacific) regions show an increasing trend over the past three years. The Eastern Mediterranean and European regions show a fluctuating trend over the last 3 years. In contrast, the Africa region has shown a continuous downward trend since 2005.

```r
fig.2.a <- anlys_dat_2022 %>%
    ggplot(aes(x=gdp/1e3,y=e_inc_100k)) +
    geom_point() +
    xlab('GDP per capita (US$ thousands)') +
    ylab('Incidence per 100k population') +
    ggtitle("The relationship between GDP per capita and TB incidence, 2022")+
    theme_gtb()


fig.2.b <- anlys_dat_2022 %>%
    ggplot(aes(x=gdp/1e3,y=e_inc_100k)) +
    geom_point() +
    scale_x_log10() + scale_y_log10(limits=c(1,1000)) +
    xlab('GDP per capita (US$ thousands)') +
    ylab('Incidence per 100k population (log scale)') +
    ggtitle("The relationship between GDP per capita and TB incidence, 2022")+
    theme_gtb() +
    geom_smooth(method='lm', formula = y~x)
```

```
#grid.arrange(fig.2.a, fig.2.b)
fig.2.a  + fig.2.b
```

2. graph the relationship between per capita TB incidence and per capita GDP;

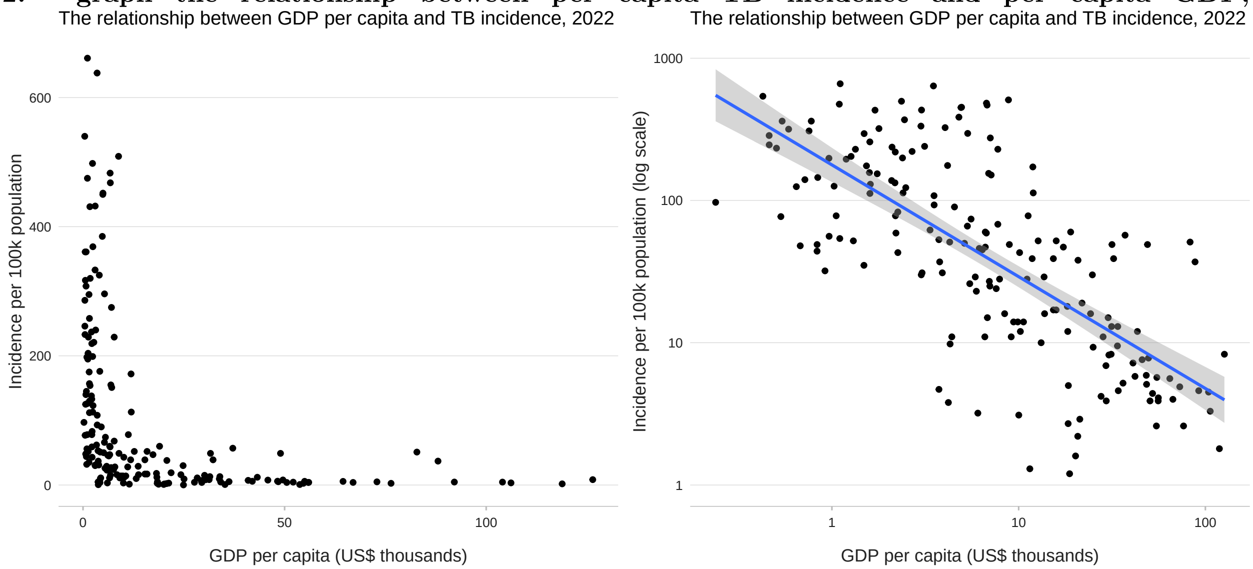
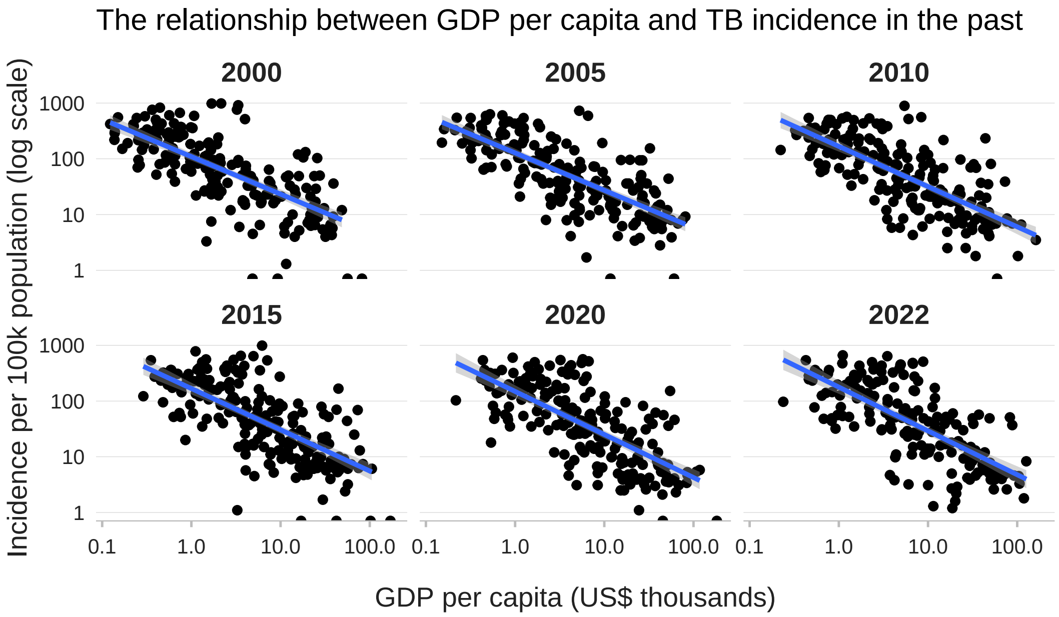


Figure 2.a (left) shows the relationship between GDP per capita and TB incidence per 100 000. It shows that as GDP per capita increases, the incidence of TB decreases exponentially. Figure 2.b (right) shows the negative linear relationship between GDP per capita and TB incidence, with a logarithmic scale. In conclusion, the wealth of a country is very closely related to the incidence of TB. To explore the relationship between GDP and TB incidence rates in the past, scatter plots are provided for the years 2000, 2005, 2010, 2015, 2020 and 2022.

```
selected_year <- c(2000, 2005, 2010, 2015, 2020, 2022)
anlys_dat_recent <-  anlys_dat %>% filter(year %in% selected_year)
fig.2.c <- anlys_dat_recent %>%
    ggplot(aes(x=gdp/1e3,y=e_inc_100k)) +
    geom_point() +
    scale_x_log10() + scale_y_log10(limits=c(1,1000)) +
    xlab('GDP per capita (US$ thousands)') +
    ylab('Incidence per 100k population (log scale)') +
    ggtitle("The relationship between GDP per capita and TB incidence in the past")+
    theme_gtb() +
    geom_smooth(method='lm', formula = y~x)+
    facet_wrap(~year)
fig.2.c
```



The association between GDP per capita and TB incidence (with logarithmic scale) in other years.

3. Using regression, develop a model that could predict TB incidence in a country without using data from any of the TB estimates files

This task aimed to predict TB incidence using the log regression prediction model in 2022. Predictor variables were collected from two primary sources. First, the TB data (not TB estimates files) was collected from WHO. Second, the GDP and population data were collected from World Bank. For developing a regression model to predict TB incidence in a country, a outcome variable and predictor variables are defined in the next step.

Outcome variable

The total of new and relapse cases and cases with unknown previous TB treatment history (`c_newinc`) variable is the outcome variable in a regression model to predict an incidence in a country. The outcome variable can find in the notification data. The values of the outcome that were missing (14 countries) in 2022 were removed.

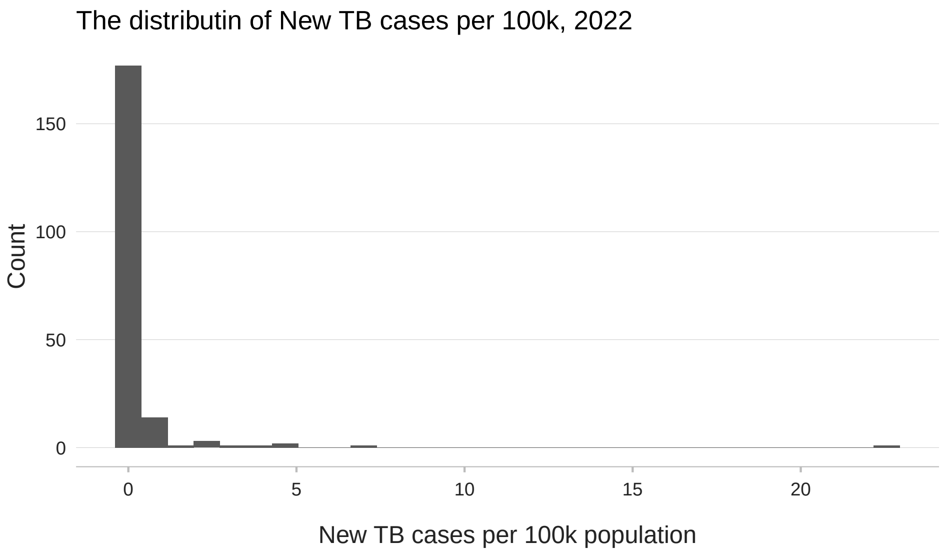
Predictor variables

All variables from WHO TB data (data provided by countries and territories are selected as predictor variables), GDP per capita and number of population are potentially considered as predictor variables. Please refer to the README.md for details of the analysis datasets.

The variables that are not available in 2022 have been dropped. In addition, variables with 50% missing values are deleted. Categorical variables are not considered as predictor variables except WHO region variable (`g_whoregion`). The `g_whoregion` values were transformed one-hot encoded. Continuous variables were re-scaled into z-score standardization.

Finally, the 201 countries and 79 predictor variables are considered in the log regression model.

```r
ggplot(tb_imputed, aes(c_newinc/1e5)) +
    geom_histogram(bins=30)+
    ylab('Count') +
    xlab('New TB cases per 100k population') +
    ggtitle("The distributin of New TB cases per 100k, 2022")+
    theme_gtb()
```



The above figure shows the distribution of the new TB cases in 2022. The outcome variable is count data and the distribution is positive skewed. Hence, the poisson canonical link function is used for developing a generalized linear model.

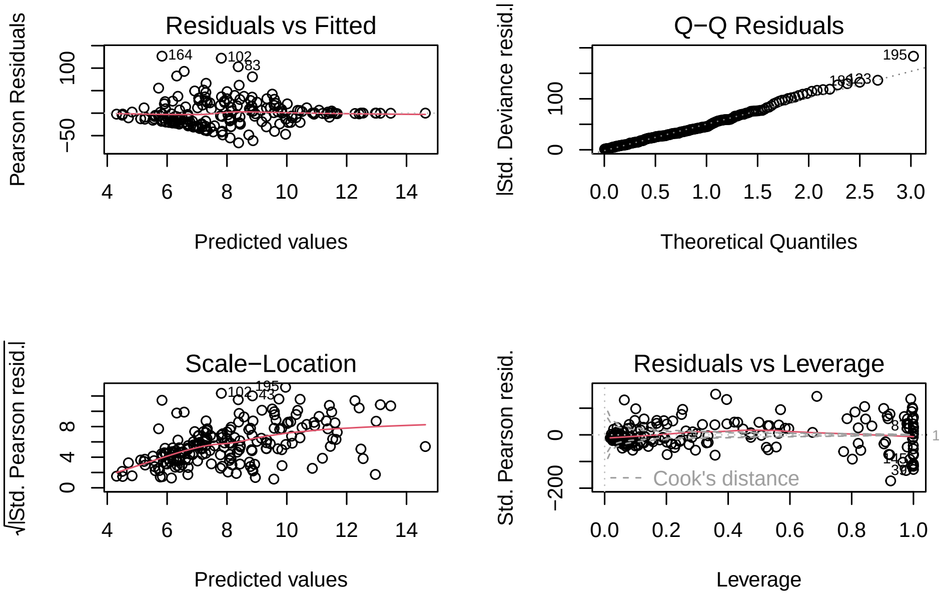
```
fit_tb <- glm(c_newinc ~ ., family="poisson", data=tb_dat)
summary(fit_tb)
```

```
##
## Call:
## glm(formula = c_newinc ~ ., family = "poisson", data = tb_dat)
##
## Coefficients: (1 not defined because of singularities)
##                         Estimate Std. Error  z value Pr(>|z|)
## (Intercept)            8.959e+00  1.123e-02  797.627  < 2e-16 ***
## budget_tot            -4.210e+00  3.576e-02 -117.736  < 2e-16 ***
## newinc_con             3.430e+00  4.477e-02   76.616  < 2e-16 ***
## newinc_con_screen     -4.138e+00  8.238e-02  -50.238  < 2e-16 ***
## newinc_con_tb          4.637e-01  9.480e-03   48.919  < 2e-16 ***
## newinc_con_prevtx     -5.119e-01  4.174e-02  -12.264  < 2e-16 ***
## newinc_con04_prevtx   -1.242e+00  1.516e-02  -81.927  < 2e-16 ***
## dx_test_sites         -5.090e+00  2.013e-01  -25.279  < 2e-16 ***
## smear                  2.925e+00  1.672e-01   17.498  < 2e-16 ***
## culture                2.267e+00  1.959e-02  115.703  < 2e-16 ***
## m_wrd_tests_performed -3.480e+00  6.348e-02  -54.821  < 2e-16 ***
## m_wrd_tests_positive   1.512e+00  9.257e-02   16.336  < 2e-16 ***
## m_wrd                  5.192e+00  4.117e-02  126.111  < 2e-16 ***
## m_inh                 -6.850e-02  1.035e-02   -6.620 3.60e-11 ***
## m_fq                   1.962e-02  1.020e-02    1.923 0.054468 .
## dst_naat_pza          -3.470e-01  6.058e-03  -57.270  < 2e-16 ***
## dst_moxlev            -5.101e+00  2.831e-02 -180.143  < 2e-16 ***
## dst_bdq               -3.855e+00  3.154e-02 -122.226  < 2e-16 ***
## dst_lzd                5.110e+00  4.162e-02  122.777  < 2e-16 ***
## iso15189_accredited    3.809e-01  4.993e-03   76.291  < 2e-16 ***
## qms_pending            3.774e-01  8.359e-03   45.150  < 2e-16 ***
## lmis                  -3.247e+00  6.880e-02  -47.205  < 2e-16 ***
## m_wrd_etrans          -2.033e-01  1.487e-02  -13.671  < 2e-16 ***
## exp_tot                2.580e+00  2.993e-02   86.209  < 2e-16 ***
## hcfvisit_dstb          1.175e-01  2.545e-03   46.183  < 2e-16 ***
## hcfvisit_mdr          -1.778e-02  2.326e-03   -7.646 2.08e-14 ***
## hospd_dstb_prct       -4.267e-01  3.698e-03 -115.382  < 2e-16 ***
## hospd_mdr_prct        -1.230e-01  2.708e-03  -45.414  < 2e-16 ***
## hospd_dstb_dur         5.482e-02  2.444e-03   22.433  < 2e-16 ***
## hospd_mdr_dur          9.804e-02  2.254e-03   43.496  < 2e-16 ***
## new_ep                 7.102e+00  7.946e-02   89.374  < 2e-16 ***
## new_labconf            9.056e+00  2.189e-01   41.378  < 2e-16 ***
## new_clindx            -3.597e-01  1.846e-01   -1.949 0.051344 .
## ret_rel_labconf       -7.173e-01  8.652e-02   -8.290  < 2e-16 ***
## ret_rel_clindx        -1.524e-01  4.843e-02   -3.146 0.001654 **
## ret_rel_ep             3.642e-01  2.511e-02   14.505  < 2e-16 ***
## ret_nrel              -5.846e+00  1.085e-01  -53.869  < 2e-16 ***
## notif_foreign         -4.754e-02  3.136e-03  -15.160  < 2e-16 ***
## newrel_m04            -2.356e+02  4.742e+00  -49.682  < 2e-16 ***
## newrel_m514           -2.972e+02  6.156e+00  -48.284  < 2e-16 ***
## newrel_m014            5.346e+02  1.052e+01   50.828  < 2e-16 ***
## newrel_m1524          -5.310e-01  7.362e-01   -0.721 0.470771
## newrel_m2534           2.057e+01  7.968e-01   25.813  < 2e-16 ***
## newrel_m3544          -9.263e+00  8.202e-01  -11.294  < 2e-16 ***
## newrel_m4554           1.585e+01  7.889e-01   20.091  < 2e-16 ***
## newrel_m5564           3.446e+00  8.475e-01    4.066 4.78e-05 ***
## newrel_m65             4.474e+00  6.614e-01    6.765 1.33e-11 ***
## newrel_m15plus        -3.205e+01  4.307e+00   -7.440 1.01e-13 ***
## newrel_mu              7.995e-01  7.624e-02   10.486  < 2e-16 ***
## newrel_f04             1.267e+02  4.198e+00   30.180  < 2e-16 ***
## newrel_f514            2.019e+02  7.367e+00   27.404  < 2e-16 ***
## newrel_f014           -3.319e+02  1.090e+01  -30.446  < 2e-16 ***
## newrel_f1524          -1.130e+01  9.618e-01  -11.749  < 2e-16 ***
## newrel_f2534          -1.832e+01  7.790e-01  -23.523  < 2e-16 ***
## newrel_f3544           1.827e+01  6.296e-01   29.019  < 2e-16 ***
## newrel_f4554          -2.594e+01  4.810e-01  -53.937  < 2e-16 ***
## newrel_f5564          -2.893e+00  4.328e-01   -6.684 2.32e-11 ***
## newrel_f65            -5.690e+00  3.016e-01  -18.868  < 2e-16 ***
## newrel_f15plus         3.331e+01  2.967e+00   11.228  < 2e-16 ***
## newrel_fu             -1.262e+00  4.578e-02  -27.576  < 2e-16 ***
## nrr                    8.020e+00  2.517e-01   31.869  < 2e-16 ***
## nrr_tx                -3.017e+00  8.633e-02  -34.952  < 2e-16 ***
## conf_rr_nfqr           4.705e+00  3.300e-01   14.259  < 2e-16 ***
## conf_rr_fqr           -1.560e+00  9.789e-02  -15.933  < 2e-16 ***
## rr_nfqr_014_tx         7.760e-01  2.455e-02   31.603  < 2e-16 ***
## unconf_rr_nfqr_tx      3.687e-01  2.195e-02   16.797  < 2e-16 ***
## conf_rr_nfqr_tx        1.202e+00  3.236e-01    3.715 0.000203 ***
## conf_rr_fqr_tx         2.685e+00  8.805e-02   30.498  < 2e-16 ***
## mdr_tx_adsm            1.797e-01  2.602e-02    6.906 4.99e-12 ***
## newrel_hivtest        -4.737e+00  1.091e-01  -43.410  < 2e-16 ***
## newrel_hivpos         -5.435e-01  1.304e-01   -4.167 3.09e-05 ***
## newrel_art             1.715e-01  1.278e-01    1.342 0.179698
## gdp                   -2.708e-01  4.289e-03  -63.144  < 2e-16 ***
## population             2.821e+00  1.932e-02  145.993  < 2e-16 ***
## EMR                   -6.118e-01  1.309e-02  -46.741  < 2e-16 ***
## EUR                   -1.317e+00  1.331e-02  -98.953  < 2e-16 ***
## AFR                    -4.214e-01  1.203e-02  -35.043  < 2e-16 ***
## AMR                    -9.986e-01  1.369e-02  -72.964  < 2e-16 ***
## WPR                    -1.361e+00  1.321e-02 -103.016  < 2e-16 ***
## SEA                            NA         NA       NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 33530386  on 200  degrees of freedom
## Residual deviance:   175452  on 122  degrees of freedom
## AIC: 177439
##
## Number of Fisher Scoring iterations: 6
```

The results show the coefficients of the log regression model and whether predictor variables are statistically significant or not under the common alpha (α) level of 0.05.

The residual plots are provided to evaluate the log regression model.

```r
par(mfrow = c(2, 2))
plot(fit_tb)
```



From the above residual plots, the four assumptions (Linearity, Independence, Homoscedasticity, and Normality) for a linear regression model are satisfied.

```r
R2_tb    <- with(summary(fit_tb), 1 - deviance/null.deviance) #R-squared
adjR2_tb <- adjR2(fit_tb);  # adjusted R-squared

R2_tab<- cbind(R2_tb, adjR2_tb)
colnames(R2_tab) <- c("R-squared", "Adj R-squared")
knitr::kable(R2_tab)
```

R-squared	Adj R-squared
0.9947674	0.9914

The R^2 and Adjusted R^2 values are very closed to 1. The values imply that most of the variation (99%) of the outcome variable is explained by predictors variables in the log regression model.

In conclusion, the log regression model is developed to predict TB incidence (new TB cases) in a country. The model will be useful for estimating TB incidence. However, the model was developed using 2022 data. The validation of the developed model should be explored in the future.