STATS 500 HOMEWORK 9 YUAN YIN

First we regress our model in a simple linear model, summary the model and the result is as follows:

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             130.91725   15.55534   8.416 4.55e-13 ***
regionEurope            -96.26275   47.63292  -2.021   0.0462 *
regionAsia              -30.28530   24.14769  -1.254   0.2129
regionAmericas          -66.30647   26.15577  -2.535   0.0129 *
income                    0.04163    0.02702   1.541   0.1268
regionEurope:income      -0.04669    0.03018  -1.547   0.1253
regionAsia:income        -0.04842    0.03121  -1.552   0.1242
regionAmericas:income    -0.05133    0.02982  -1.721   0.0886 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 79.83 on 93 degrees of freedom
Multiple R-squared:  0.2811,    Adjusted R-squared:  0.227
F-statistic: 5.194 on 7 and 93 DF,  p-value: 5.055e-05
```

Finding that many of the predictors are not significant (p - value is larger than $\alpha = 0.01$), also we can see that $R^2 = 0.2811$ which is far from 1, it shows that this model doesn't fit very well to our data, so we should find some method to improve our model.

As the region is a categorical predictors, we can clearly analysis our data in a plot as follows (left):

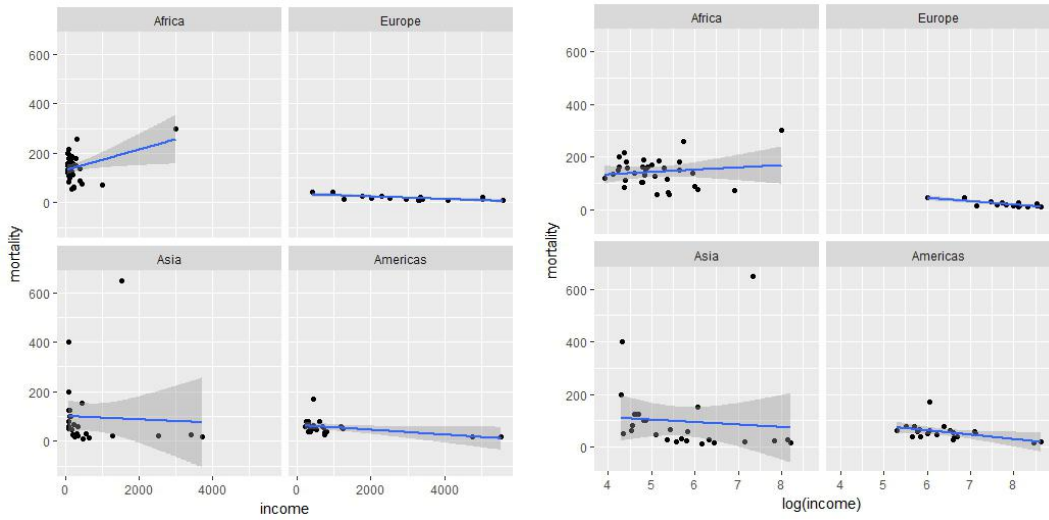

Fig 1 summary of data

We can see that for data in Europe and Americas, it seems that the model fits well, but for data in Africa and Asia, apparently that linear model doesn't fit well.

What if we transform the predictors? We can see that for Asia, the variance of mortality seems like to spread when income goes up, to maintain the constant variance, it seems like that we can use log(income) to solve this problem, the new ggplot is as above (right).

Apparently, the new model fits better than the original model. Changing our model with new predictor: log(income). Also we want to use Box-cox method to transform the response, the result is as follows:

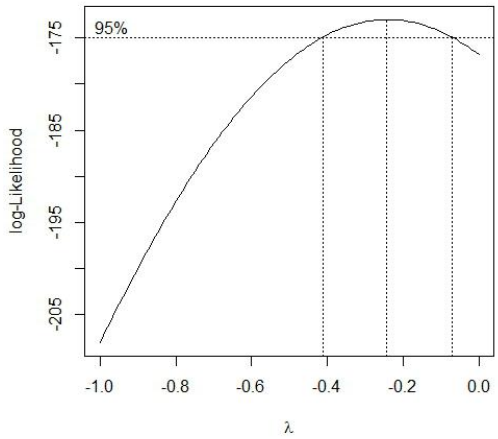

Fig 2 transformation of response

Finding that the confidence interval of λ is as above, we choose $\lambda = -0.25$, thus our transformation is like:

$$y^{-1/4} = \beta_0 + \beta_{\log(income)} x_{\log(income)} + \beta_{region} x_{region} \text{(three terms)+"interaction"(three terms)}$$

Use anova to test the significant of interaction terms:

```
> anova(ybc)
Analysis of Variance Table

Response: mortality^(-0.25)
                 Df  Sum Sq Mean Sq  F value    Pr(>F)
log(income)       1 0.45371 0.45371 166.9567 < 2.2e-16 ***
region            3 0.09272 0.03091  11.3734 2.016e-06 ***
log(income):region  3 0.03091 0.01030   3.7912   0.01294 *
Residuals        93 0.25273 0.00272
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that all terms are significant but interaction term has higher p-value which is larger than $\alpha = 0.01$. Finding that if we drop the interaction term, the model has better p-value than before.

Look at F-test and it seems that all the term are significant in our final model.

```
> drop1(ydr, test="F")
Single term deletions

Model:
mortality^(-0.25) ~ log(income) + region
           Df Sum of Sq     RSS     AIC F value    Pr(>F)
<none>                  0.28364 -583.39
log(income)  1  0.080831 0.36447 -560.07  27.358 9.892e-07 ***
region       3  0.092722 0.37636 -560.83  10.461 5.116e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To make diagnostics to check our assumptions, first we plot the residuals vs fitted values to see if we have constant variance, also we plot QQ-plot to see if residuals are normal distribution

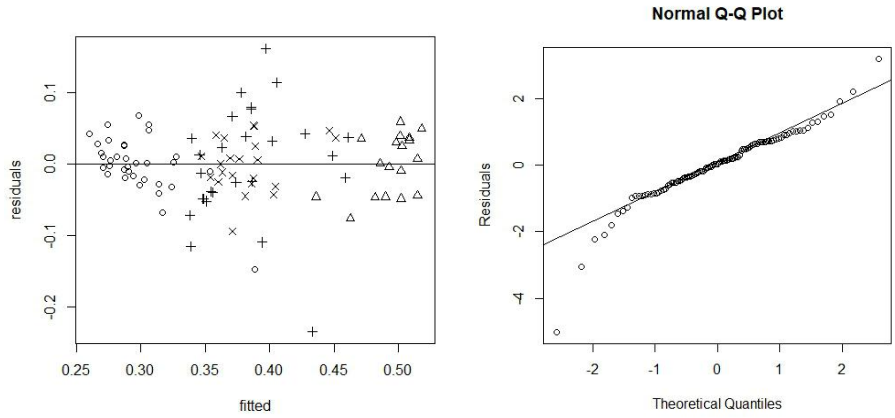

Fig 3 residuals vs fitted values and regression lines and QQ-plot

We can see that the residuals are equally distributed on the two side of zero and also they seem to have no relation with fitted values, we can assume there is constant variance.

Also looking at influential points, we use cook's distance and finding that point 25 and 27 is two influential points. We refit our model without these two points. Our result:

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.111766   0.031670   3.529 0.000645 ***
log(income)    0.037367   0.006077   6.149 1.83e-08 ***
regionEurope   0.088341   0.023185   3.810 0.000246 ***
regionAsia     0.057372   0.014201   4.040 0.000108 ***
regionAmericas 0.030698   0.016767   1.831 0.070263 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05215 on 95 degrees of freedom
Multiple R-squared:  0.6819,    Adjusted R-squared:  0.6685
F-statistic: 50.91 on 4 and 95 DF,  p-value: < 2.2e-16
```

All the terms are significant and $R^2 = 0.6819$ which is much larger than before. In this way, we get our final model which is like:

$$y^{-1/4} = 0.112x_{Africa} + 0.037x_{\log(income)} + 0.200x_{Europe} + 0.169x_{Asia} + 0.143x_{Americas}$$

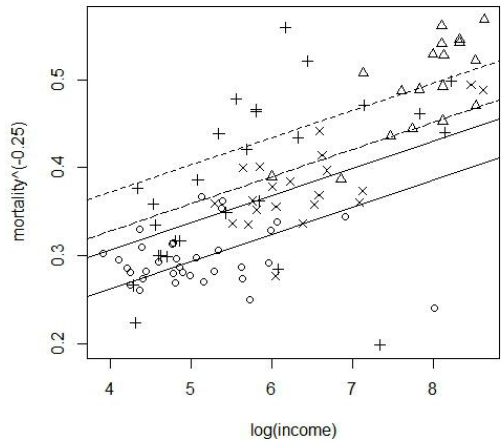

Fig 4 final regression model

The four line in the model represents, with the different region, how the mortality changes when log(income) goes up, it seems like when log(income) goes up with 1, the $mortality^{-1/4}$ of same region will go up in 0.037. Also at this time, the increase of

true income is different with what the previous income is. That is to say, when true income goes up with Δx, log(income) will go up with $\log(income + \Delta x) - \log(income)$. Also comparing different region, with the Africa is the reference level, comparing with it, we find that $mortality^{-1/4}$ of Europe will go up 0.088 on average, and $mortality^{-1/4}$ of Asia will go up 0.057 on average, and $mortality^{-1/4}$ of Americas will go up 0.031 on average.

Appendix

```
library(faraway)
data(infmort)
infmortm = na.omit(infmort)
library(MASS)
y = lm(mortality ~ region * income, data = infmortm)
summary(y)
require(ggplot2)
ggplot(aes(x=income,y=mortality),data=infmortm)+geom_point()+facet_wrap(~
region)+geom_smooth(method="lm")
ggplot(aes(x=log(income),y=mortality),data=infmortm)+geom_point()+facet_wrap(~
region)+geom_smooth(method="lm")
ylog = lm(mortality ~ log(income) + region + log(income):region, data = infmortm)
summary(ylog)
anova(ylog)
# Box-Cox method for data
boxcox(ylog, plotit = T)
boxcox(ylog, plotit = T, lambda = seq(-1, 0, by = 0.1))
ybc = lm(mortality^(-0.25) ~ log(income) + region + log(income):region, data = infmortm)
summary(ybc)
anova(ybc)
# drop interaction term
ydr = lm(mortality^(-0.25) ~ log(income) + region, data = infmortm)
summary(ydr)
# after the other has been taken into account
drop1(ydr, test="F")
# diagnostics
plot(residuals(ydr) ~ fitted(ydr), pch = unclass(infmortm$region), xlab = "fitted", ylab = "residuals")
abline(h = 0)
# regression lines
plot(mortality^(-0.25) ~ log(income), infmortm, pch = as.numeric(infmortm$region))
abline(0.138, 0.031)
abline(0.138+0.110, 0.031, lty = 2)
abline(0.138+0.066, 0.031, lty = 5)
abline(0.138+0.044, 0.031, lty = 7)
```