HW3\_Template

Lina Cao , Thursdays 3:30PM

SDGB 7844; Prof. Nagaraja; Fall 2017

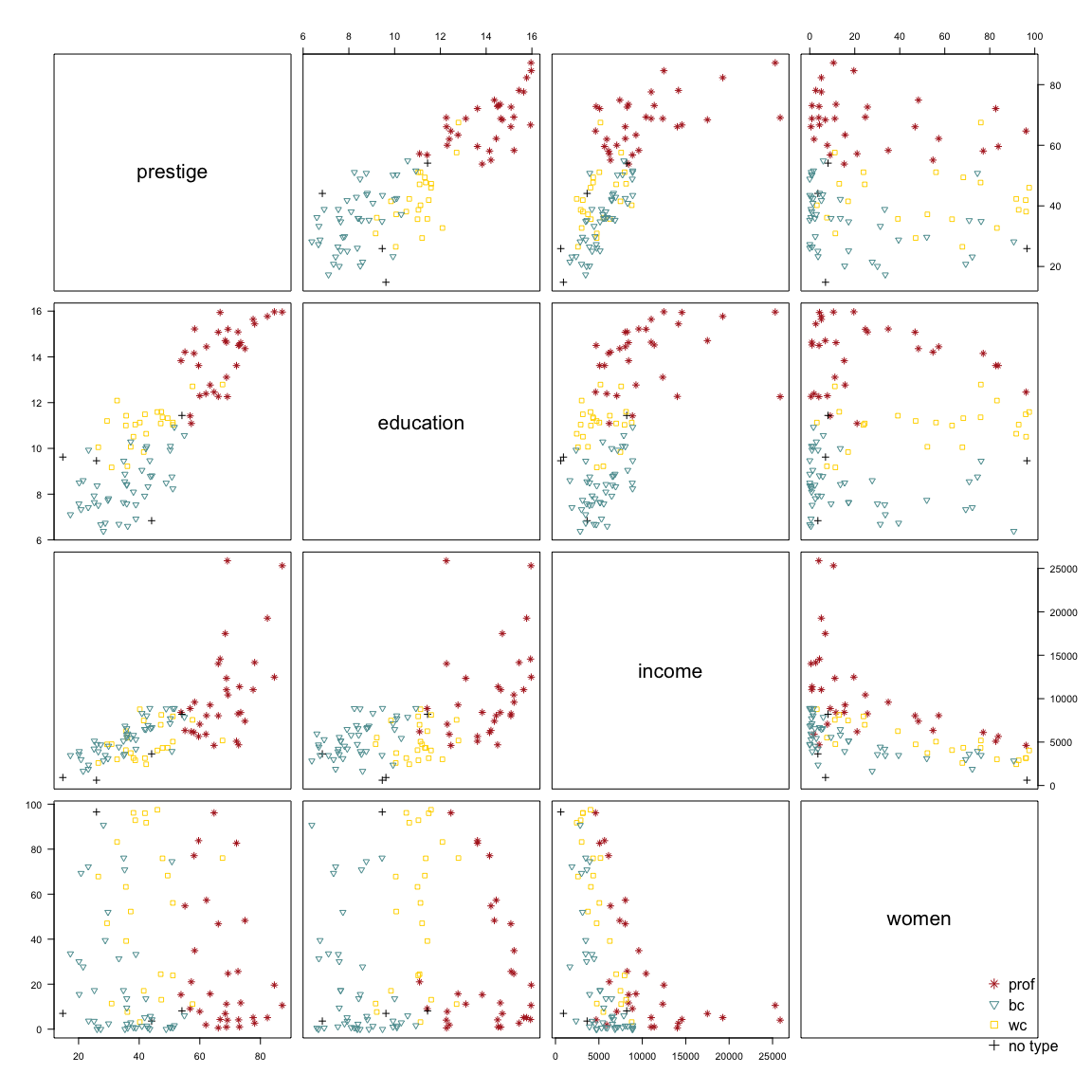
# PROBLEM 1

### (a) Do some internet research and write a short paragraph in your own words about how the Pineo-Porter prestige score is computed. Include the reference(s) you used. Do you think this score is a reliable measure? Justify your answer.

Pineo-Porter prestige score is a direct measure using the average evaluation made of an occupational title by a national sample to census occupational titles. It consists in constucting a regression equation which has the dependent variables as Pinio-Porter scores for the 88 occupations which overlaps the census list, and has the independent variables—the corresponding income level and educational level indices. The regression weights are then applied to all census occupations(343 in 1951 index and 320 in 1961 index) to derive the prestige scores.

Source:  
(1) Deonandan, Raywat, et al. “A comparison of methods for measuring socio-economic status by occupation or postal area.” Chronic Diseases and Injuries in Canada 21.3 (2000): 114.  
(2) Blishen, Bernard R. “A socio‐economic index for occupations in Canada.” Canadian Review of Sociology/Revue canadienne de sociologie 4.1 (1967): 41-53.

### (b) Create a scatterplot matrix of all the quantitative variables and describe what you see. Use a different symbol for each profession type: no type (pch=3), “bc” (pch=6), “prof” (pch=8), and “wc” (pch=0) when making your plot. For the remainder of this question, we will use the explanatory variables: income, education, and type. Does restricting our regression to only these variables make sense given your exploratory analysis? Justify your answer.



PS:To make it clearer when consider the pattern of “type”, besides using different symbols, I also use different color in the plot.

All variables except for “type” are numeric values, while “census” is only a code representing an area, therefore the quantitative variables I used are “prestige”, “education”, “income” and “women”.

From the plot, “prestige” and “education” have clearly positive linear relationship, especially when education is not too large; “prestige” and “income” also show a positive linear relationship, but with some outliers. The relationship can also imagined as convex curve, therefore further investigation is needed to decide which relationship should be accept; “prestige” and “women” have little linear relationship because of no pattern of their scatter plot. The scatter plots between (1)“education” and “income” and (2)“income” and “women” also indicate their linear relationship while ignoring outliers, the first two variables may have positive linear relationship while the second two may be slightly negative linear related. The scatter plot of “education” and “women” hardly shows a pattern, which implies their low chance for having linear relationship.

When take color(‘type’) into consideration, we can observe that different profession types separate apart from each other in each scatter plot, the order of each “layer” in the plots tends to be (from higher to lower) “prof”, “wc”, “bc”, which is consistent with intuition on the prestige of ranking “professor”, “white collar” and “blue collar”.

Therefore, it is wise to include “type” in further study. From the former illustration, “women” shows least contribution to linear relationship, thus the restriction to “income”, “education” and “type” as explanatory variables makes sense.

### (c)Which professions are missing “type”? Since the other variables for these observations are available, we could group them together as a fourth professional category to include them in the analysis. Is this advisable or should we remove them from our data set? Justify your answer.

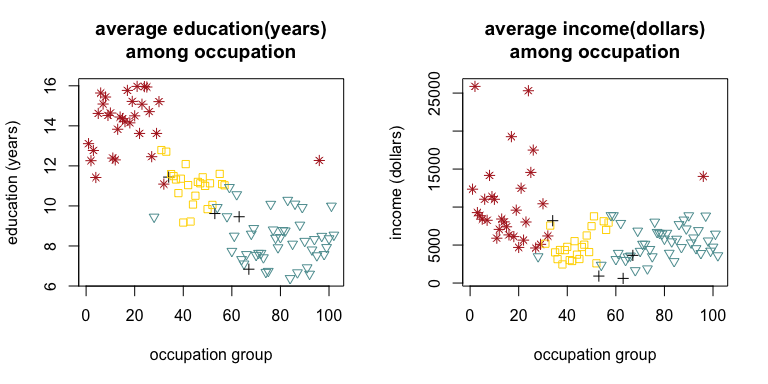
## [1] 34 53 63 67

## [1] "athletes" "newsboys" "babysitters" "farmers"

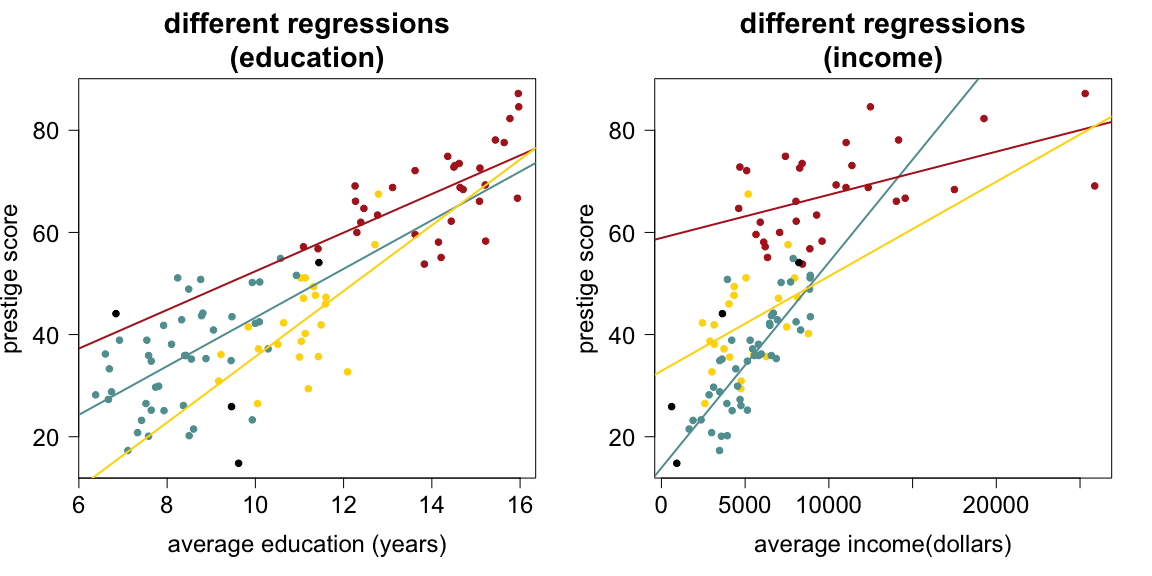
The 34th, 53rd, 63rd and 67th objects have missing “type”. The corresponding occupation groups are “atheletes”, “newsboys”, “babysitters” and “farmers” respectively.

I don’t think we should group them into a new profession category. Because from the plot, the black “+” symbols do not cluster together far away from colorful symbols, instead, some of them are in the cluster of “bc”. Additionally, four data will be a small sample amount for a new group in regression. Therefore, I would prefer regarding them as observations with missing data and delete them from our dataset.

### (d)Visually, does there seem to be an interaction between type and education and/or type and income? Justify your answer.



Yes. EDUCATION: For professional occupations, the education years tend to be higher than other two types, and white collar occupations also have longer education years than blue collars.  
INCOME: Although not as significant as shown in the scatter plot of “education”. Professional occupations earn higher salary than white/blue collar occupations. While the difference between white collar and blue collar on “income” is not recognizable.



From the second picture that fits separate lines for different profession types, the fact that lines are not parallel indicates interaction terms of “Categorical\*Numerical" in the regression. But further discussion is required to determine whether the two interaction term should be both included in the fianl model.

### (e)Fit a model to predict prestige using: income, education, type, and any interaction terms based on your answer to part (d). Evaluate your model by checking the regression assumptions (including collinearity/multicollinearity) and provide details and conclusions for any hypothesis tests. Include relevant output. Use your answer to part (c) to determine which observations to use in your analysis.

##   
## Call:  
## lm(formula = prestige ~ education + income + type + education \*   
## type + income \* type, data = clean)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.462 -4.225 1.346 3.826 19.631   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.276e+00 7.057e+00 0.323 0.7478   
## education 1.713e+00 9.572e-01 1.790 0.0769 .   
## income 3.522e-03 5.563e-04 6.332 9.62e-09 \*\*\*  
## typeprof 1.535e+01 1.372e+01 1.119 0.2660   
## typewc -3.354e+01 1.765e+01 -1.900 0.0607 .   
## education:typeprof 1.388e+00 1.289e+00 1.077 0.2844   
## education:typewc 4.291e+00 1.757e+00 2.442 0.0166 \*   
## income:typeprof -2.903e-03 5.989e-04 -4.847 5.28e-06 \*\*\*  
## income:typewc -2.072e-03 8.940e-04 -2.318 0.0228 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.318 on 89 degrees of freedom  
## Multiple R-squared: 0.8747, Adjusted R-squared: 0.8634   
## F-statistic: 77.64 on 8 and 89 DF, p-value: < 2.2e-16

# CHECK ASSUMPTIONS

(1)Muiticollinearity check

## Loading required package: usdm

## Loading required package: sp

## Loading required package: raster

## Variables VIF  
## 1 education 1.491621  
## 2 income 1.491621

Two VIFs are close to 1, from the regression estimation, education and income both have positive beta-hat, which is consistent to the positive linear relationship from the scatterplots. There might not be multicollinearity in first model.

##   
## Call:  
## lm(formula = prestige ~ education, data = clean)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.605 -6.151 0.366 6.565 17.540   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -10.8409 3.5285 -3.072 0.00276 \*\*   
## education 5.3884 0.3168 17.006 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.578 on 96 degrees of freedom  
## Multiple R-squared: 0.7508, Adjusted R-squared: 0.7482   
## F-statistic: 289.2 on 1 and 96 DF, p-value: < 2.2e-16

However, “education” in lm1 is not significant at alpha=0.05 level, while when we only include “education” in the regression, it is significant. Therefore I decided to do the following steps. (And since income’s interaction term is significant under 0.05 significance level, “income” and “type” will be in the model.)

(1.1) Try to see if we remove interaction term related to education

## Analysis of Variance Table  
##   
## Model 1: prestige ~ education + income + type + income \* type  
## Model 2: prestige ~ education + income + type + education \* type + income \*   
## type  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 91 3791.3   
## 2 89 3552.9 2 238.4 2.9859 0.05557 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Partial F-test has its p-value greater than 0.05, which means we should not reject H0 (the coefficients of variables related to education is all zero), and omit education’s interaction term.

(1.2)New regression result

##   
## Call:  
## lm(formula = prestige ~ education + income + type + income \*   
## type, data = clean)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.8720 -4.8321 0.8534 4.1425 19.6710   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6.7272633 4.9515480 -1.359 0.1776   
## education 3.0396961 0.6003699 5.063 2.14e-06 \*\*\*  
## income 0.0031344 0.0005215 6.010 3.79e-08 \*\*\*  
## typeprof 25.1723873 5.4669586 4.604 1.34e-05 \*\*\*  
## typewc 7.1375093 5.2898177 1.349 0.1806   
## income:typeprof -0.0025102 0.0005530 -4.539 1.72e-05 \*\*\*  
## income:typewc -0.0014856 0.0008720 -1.704 0.0919 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.455 on 91 degrees of freedom  
## Multiple R-squared: 0.8663, Adjusted R-squared: 0.8574   
## F-statistic: 98.23 on 6 and 91 DF, p-value: < 2.2e-16

This time, education is, as expected, significant under 0.05 significant level. The insignificance of “’wc” type and the interaction term of “income\*typewc" is not problematic if we have decided to use “income” and “type” in the model.

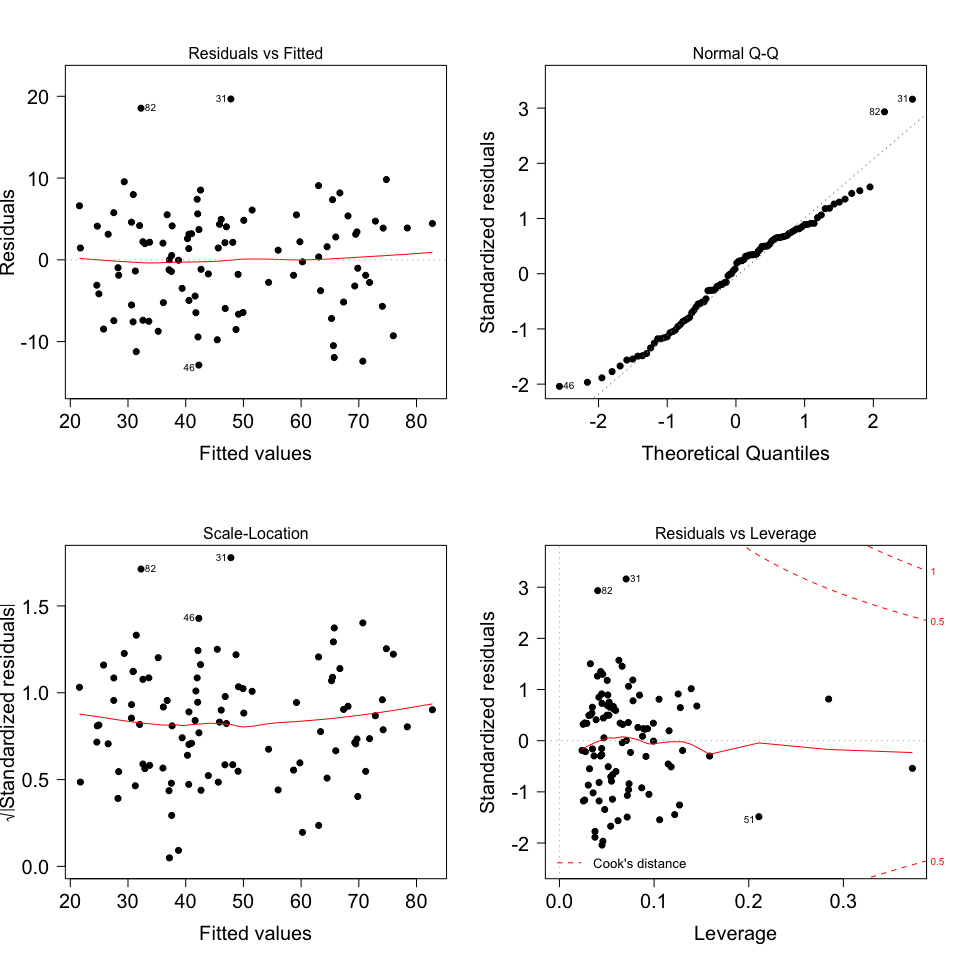
Next, do the linear regression assumption check based on “lm2”. (1)Multicollinearity check

Same as the result above, the VIFs are close to 1 and the estimates are positive, consistent with the relationship shown in scatter plot. There is no colinearity problem.

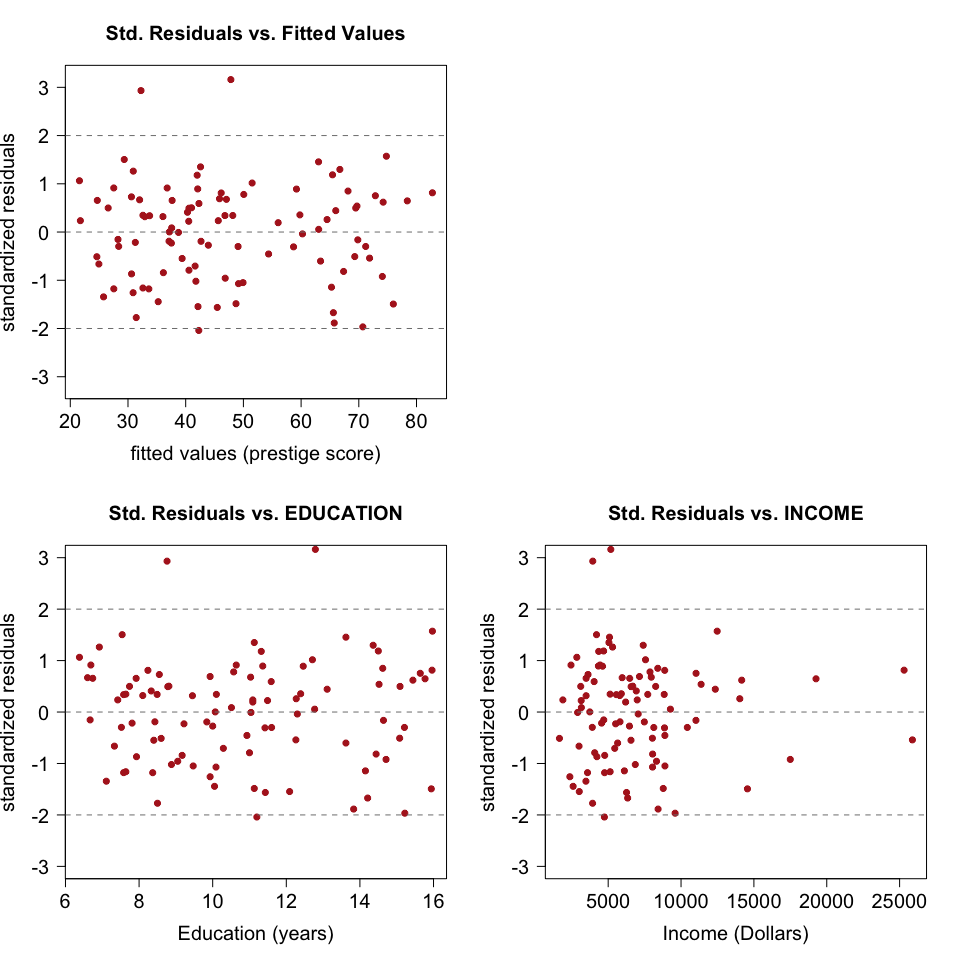
(2)Fixed “X” and measurement error: it is acknowledged for the original data since it is from countries’ Bureau;

(3)Linearity: from the scatterplot matrix in (a), the two quantitative variables are linearly related to response variable;

(4)Zero conditional mean This is an assumption when we do the multi-linear regression.

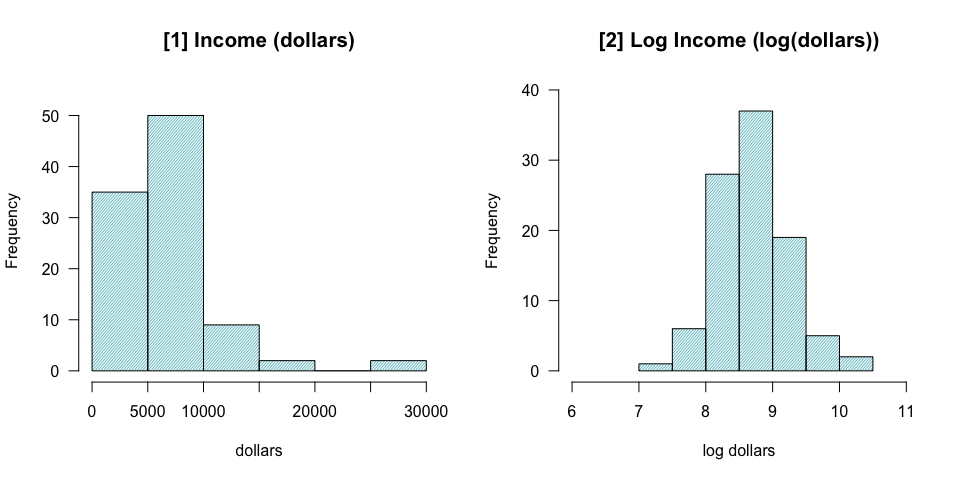
1. Normally distributed residuals 

From the top right plot, residuals are slightly s-shaped indicating that the residual distribution is a little thick-tailed. But it is not severe enough to be worried.

(6)Constant variance of the error term (regardless of “X” values); Neither the residual versus fitted (top left) nor scale-location (bottom left) plot shows the variability of the residuals getting larger as one moves from one side of the plot to the other. When focusing on y-hat, from the top left plot below, residuals almost vary among (-2,2), which implies a constant variance when “X” changes. Therefore constant variance assumption meets. 

From the bottom residual plots in the graph above, “education” has no pattern of its residuals, with its residuals ranging from (-2,2), indicating a relatively good fit in this model. While the residuals of “income” has larger deviation when the x-value becomes smaller, with several significantly large value. Making logarithm of “income” can be next step(as in part(f)) for generating a better fitted model.

### (f)Create a histogram of income and a second histogram of log(income) (i.e., natural logarithm). How does the distribution change?



From [1], income shows a right-skewed distribution, however, after taking logarithm in [2], log of income shows a close to symmetric distribution, slightly skewing to left. Therefore, the distribution turns more close to symmetric after taking log.

### (g)Fit the model in (e) but this time use log(income) (i.e., natural logarithm) instead of income. Evaluate your model by checking the regression assumptions and details and conclusions for any hypothesis tests. Include relevant output.

After trying different interaction in the model, the final fitted model is the same as that in (e), except for using log(income) instead of income.

##   
## Call:  
## lm(formula = prestige ~ logincome + education + type + logincome \*   
## type, data = clean)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.484 -4.453 1.122 4.123 18.737   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -118.4325 20.3728 -5.813 8.97e-08 \*\*\*  
## logincome 14.9336 2.4928 5.991 4.12e-08 \*\*\*  
## education 3.2107 0.5993 5.357 6.31e-07 \*\*\*  
## typeprof 82.7757 31.5059 2.627 0.0101 \*   
## typewc 51.3717 36.8521 1.394 0.1667   
## logincome:typeprof -8.5690 3.5251 -2.431 0.0170 \*   
## logincome:typewc -6.1925 4.3172 -1.434 0.1549   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.491 on 91 degrees of freedom  
## Multiple R-squared: 0.8647, Adjusted R-squared: 0.8558   
## F-statistic: 96.96 on 6 and 91 DF, p-value: < 2.2e-16

As is shown above, the p-values of “logincome” and “education” are both significant, while the categorial term and interaction term both have one situation that has insignificant sign.

## adj.R^2(w/income) adj.R^2(w/logincome)   
## 0.8574365 0.8558233

The adjusted R-squared slightly decreases after transformation. Do assumption check before evaluating.

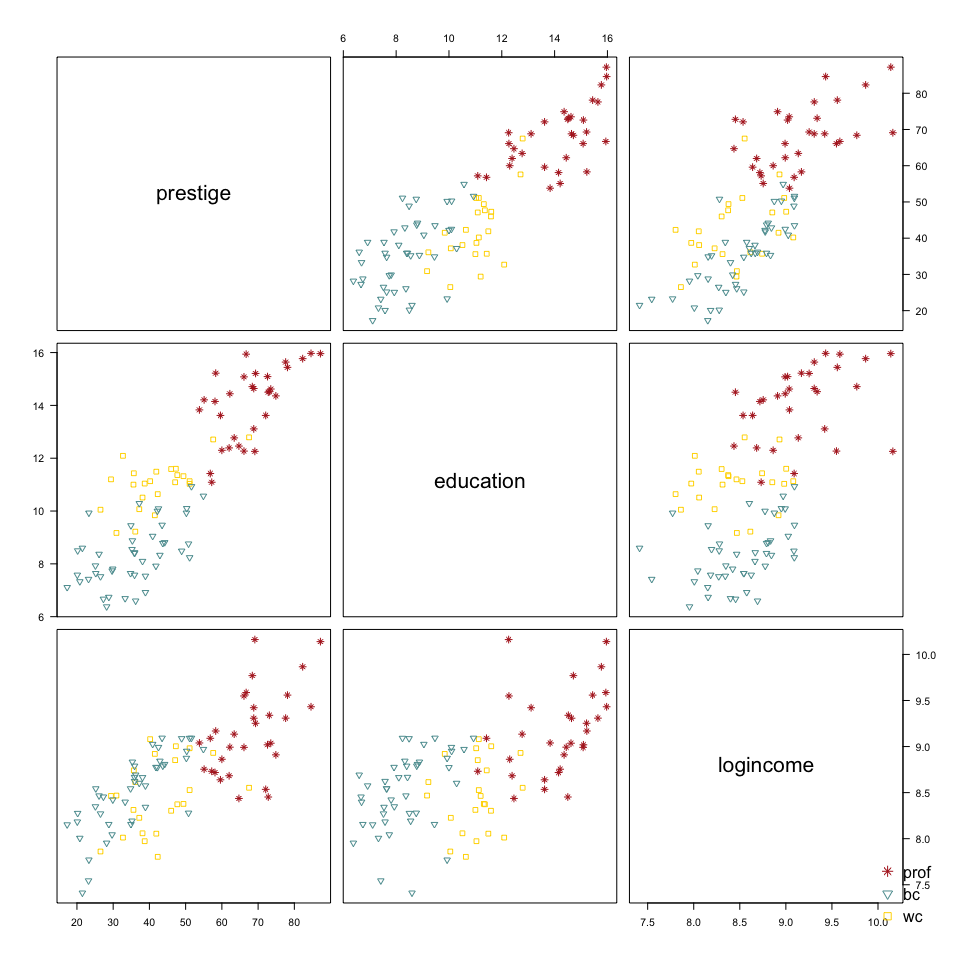
# ASSUMPTION CHECK

(1)Multicollinearity check

## Variables VIF  
## 1 education 1.535256  
## 2 logincome 1.535256

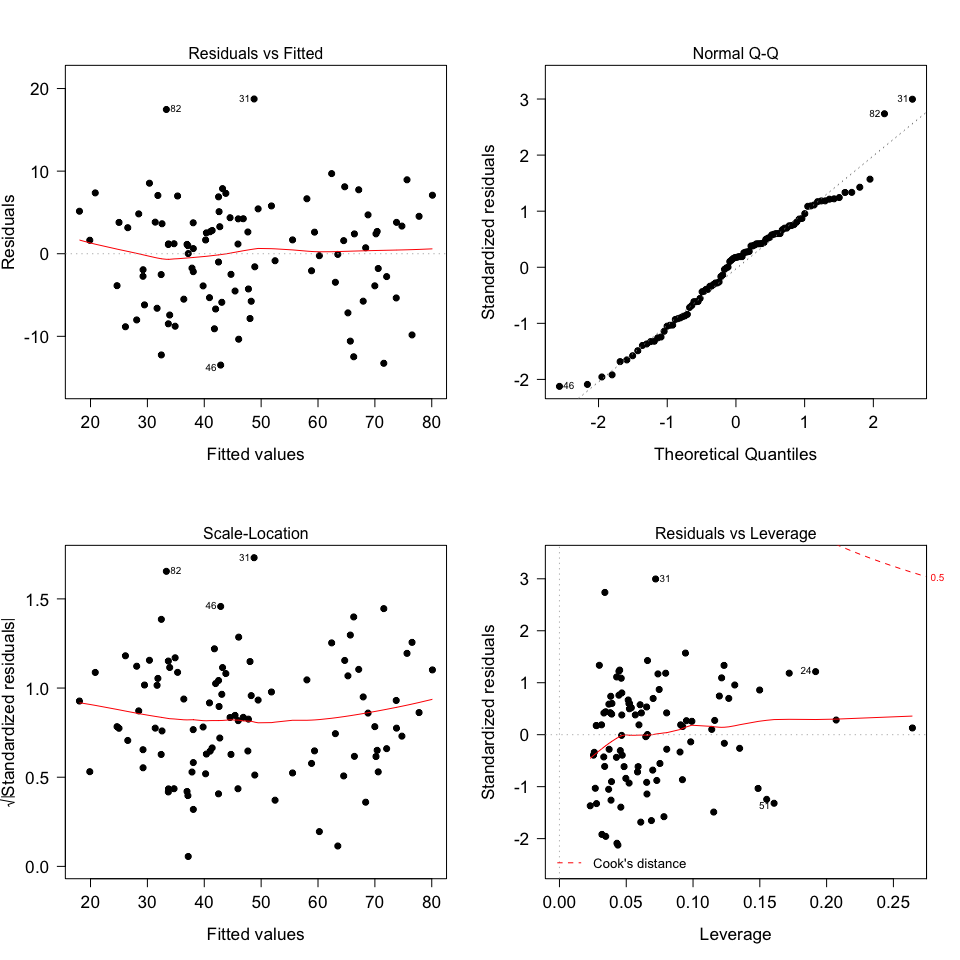
Similar with the result above, the VIFs are close to 1 and the estimates are positive, consistent with the relationship shown in scatter plot. There is no multicollinearity problem.

(2)Fixed “X” and measurement error: it is acknowledged for the original data since it is from countries’ Bureau;

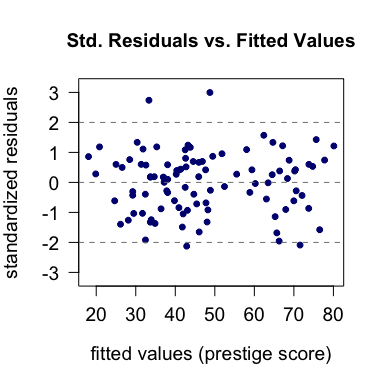
(3)Linearity 

From the scatterplot matrix in above, quantitative variables are linearly related to response variable. Especially “prestige-logincome”, it has a more close to linear relationship compared to the former upward-curved shape in “prestige-income”;

(4)Zero conditional mean This is an assumption when we do the multi-linear regression.

1. Normally distributed residuals 

From the top right plot, residuals are slightly s-shaped on both sides indicating that the residual distribution is a little thick-tailed. But it is not severe enough to be worried.

(6)Constant variance of the error term (regardless of “X” values); Neither the residual versus fitted (top left) nor scale-location (bottom left) plot shows the variability of the residuals getting larger as one moves from one side of the plot to the other. When focusing on y-hat, from the topleft plot below, residuals almost vary among (-2,2), which implies a same variance when “X” changes. Therefore constant variance assumption meets. 

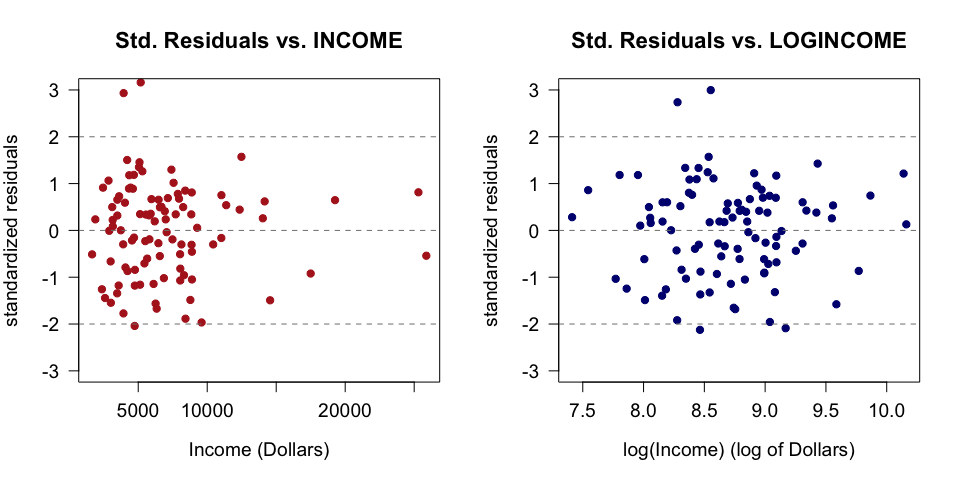
### (h)Is the model in (e) or (g) better? Justify your answer. Why can’t we use a partial F-test here?

|  |  |  |  |
| --- | --- | --- | --- |
|  | a.r.squared | AIC | BIC |
| lm2-with INCOME | 0.8574 | 369.8730 | 393.1377 |
| lm3-with logINCOME | 0.8558 | 372.2376 | 390.3324 |

I think the one with log(income) is better. The difference between the “criterion table” is not significant between two models: lm2 has better adjusted R-squared(larger) and AIC(smaller), but BIC of lm3 is better(smaller).

## $`coefficient(INCOME)`  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -6.7273 4.9515 -1.3586 0.1776  
## education 3.0397 0.6004 5.0630 0.0000  
## income 0.0031 0.0005 6.0098 0.0000  
## typeprof 25.1724 5.4670 4.6045 0.0000  
## typewc 7.1375 5.2898 1.3493 0.1806  
## income:typeprof -0.0025 0.0006 -4.5392 0.0000  
## income:typewc -0.0015 0.0009 -1.7035 0.0919  
##   
## $`coefficient(logINCOME)`  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -118.4325 20.3728 -5.8133 0.0000  
## logincome 14.9336 2.4928 5.9908 0.0000  
## education 3.2107 0.5993 5.3575 0.0000  
## typeprof 82.7757 31.5059 2.6273 0.0101  
## typewc 51.3717 36.8521 1.3940 0.1667  
## logincome:typeprof -8.5690 3.5251 -2.4309 0.0170  
## logincome:typewc -6.1925 4.3172 -1.4344 0.1549

The absolute value differences between the two model’s value are quite small, thus, further investigation is needed to make decision.



The residuals plot of the transformed variable shows a better, random pattern after taking log, therefore, I think lm3 is better.

Partial F test is used when comparing nested models, so this is the main reason we don’t use it to compare models. Here we change variables in the model and we use same number of varialbes in the model, therefore the two models are not nested at all.

# PROBLEM 2

### (a)Clean your data. Randomly assign your data into training and test sets. Half of the data should be in the training set and half should be in the test set. How many observations are in the training set? Is there any missing data?

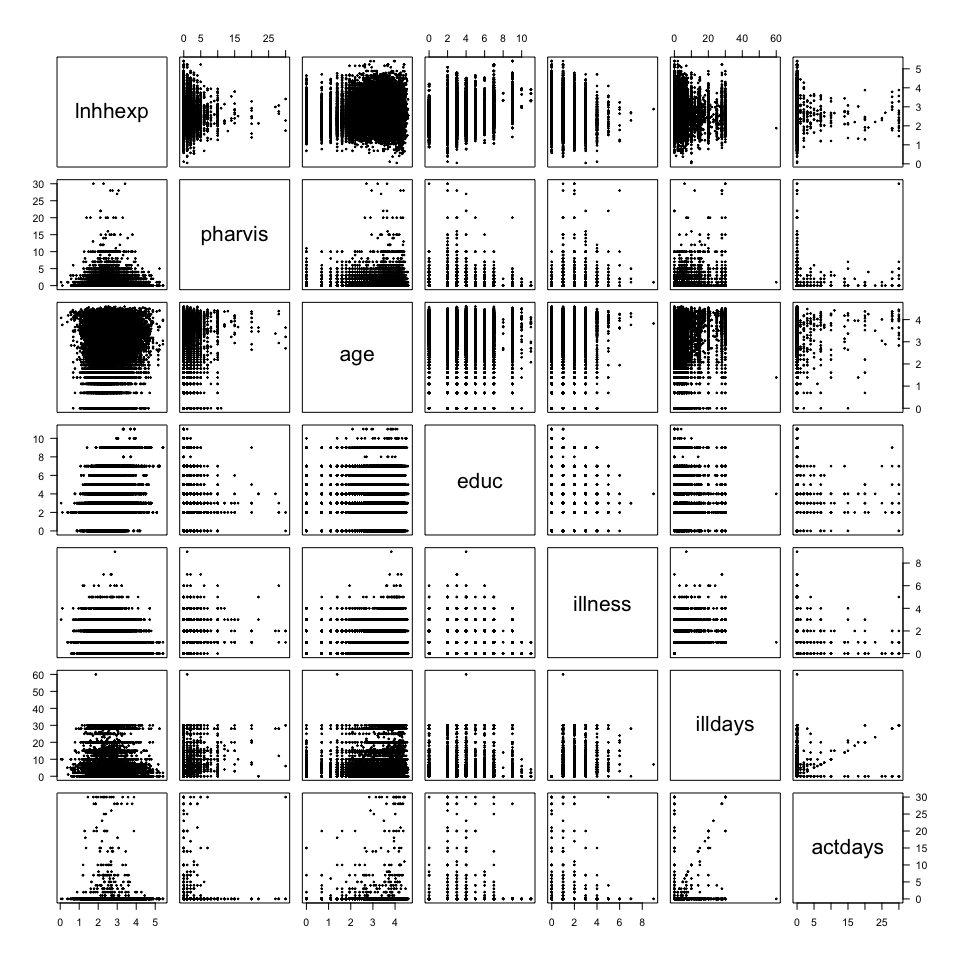
## missing data? numeric as numeric? character as character?   
## FALSE TRUE TRUE

## [1] 13883 12

There are 13,883 observations in training set. No missing data is found.

### (b)Create a table of summary statistics (mean, median, standard deviation, minimum, and maximum) and make scatterplots for the pairs of numerical variables. If you need to transform any data because of potential linearity issues, do that now. Also, if you see any interaction terms which may be useful while doing your exploratory data analysis, you can add them in when you do questions 2c, 2d, and 2e. (Note: since the response variable has already been transformed, do not transform it again.)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | units | mean | median | sd | min | max |
| pharvis | times | 0.5118 | 0.0000 | 1.3134 | 0.0000 | 30.0000 |
| lnhhexp | log(dollar) | 2.6026 | 2.5349 | 0.6244 | 0.0467 | 5.4055 |
| age | years old | 2.9775 | 3.1355 | 0.9671 | 0.0000 | 4.5951 |
| educ | years | 3.3907 | 3.0000 | 1.9311 | 0.0000 | 11.0000 |
| illness | times | 0.6220 | 0.0000 | 0.8995 | 0.0000 | 9.0000 |
| illdays | days | 2.8040 | 0.0000 | 5.4582 | 0.0000 | 60.0000 |
| actdays | days | 0.0657 | 0.0000 | 1.1159 | 0.0000 | 30.0000 |



The dots are very dense, therefore no idea of transformation right now. PS: I did the following attempts for transformation. <1>AGE After the first scatterplots, age seems be strange with range of (0,4.60), I tried transformation with exp(age) to let the number make sense. The dots are more randomly distributed after the transformation, but the regression does not have a larger r-squared. Therefore, I decided not to transform in the end. <2>COMMUNE

## [1] 1 194

Since there are 194 categories in “commune”, it will be messy if directly include “commune” as a categorial variable.

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 2.918874716 1.050170e-02 277.94306 0.000000e+00  
## commune -0.003141364 9.046372e-05 -34.72512 1.949922e-253

And from the result of regression with commune, commune turns out to be a significant variable, therefore I decide to group communes into a small number of groups.  
After doing research on what ward each number represents, I found communes are separate into 10 main areas.  
(source: <http://web.worldbank.org/archive/website00002/WEB/PDF/VN98BI-2.PDF>)

## CLUSTERS COMMUNE.RANGE  
## 1 Major Urban Areas 1~20  
## 2 Medium Urban Areas 21~36  
## 3 Minor Urban Areas 37~58  
## 4 Rural Northern Mountains 59~79  
## 5 Rural Red River Delta 80~104  
## 6 Rural North Central Coast 105~123  
## 7 Rural South Central Coast 124~139  
## 8 Rural Central Highlands 140~151  
## 9 Rual Southeast 152~168  
## 10 Rural Mekong Delta 169~194

Because of lacking knowledge of Vietnam’s economic situations and other specific development in differenct province, I just simply group the part into “urban” and “rural” areas in my study.

<3>MULTICOLLINEARITY

## lnhhexp pharvis age educ illness illdays actdays  
## lnhhexp 1.0000 -0.0313 0.0617 0.2556 -0.1007 -0.0650 -0.0099  
## pharvis -0.0313 1.0000 0.0834 -0.0528 0.4263 0.3545 0.0457  
## age 0.0617 0.0834 1.0000 0.0251 0.0811 0.1466 0.0315  
## educ 0.2556 -0.0528 0.0251 1.0000 -0.0451 -0.0221 -0.0044  
## illness -0.1007 0.4263 0.0811 -0.0451 1.0000 0.5825 0.0149  
## illdays -0.0650 0.3545 0.1466 -0.0221 0.5825 1.0000 0.0818  
## actdays -0.0099 0.0457 0.0315 -0.0044 0.0149 0.0818 1.0000

## Variables VIF  
## 1 lnhhexp 1.091508  
## 2 pharvis 1.290613  
## 3 age 1.036410  
## 4 educ 1.081936  
## 5 illness 1.709331  
## 6 illdays 1.544885  
## 7 actdays 1.007867

The VIFs are all close to 1, from VIF value, multicollinearity is not a big problem so far, so further regression is needed to see whether the correlation signal is consistent with the estimates.

### (c) Use forward selection to choose the “best” model. Explain how you decided which model was “best.”

The criterion of deciding the better model using forward selection is stopping adding variable when its p-value is not significant once added. With the help of the result, I compare RMSE, adjusted R-squared, CP value and BIC to decide the “best” two models, and do partial F test in the end to make final decision.

## Loading required package: leaps

## Subset selection object  
## Call: regsubsets.formula(lnhhexp ~ ., data = train, method = "forward",   
## nvmax = 11)  
## 11 Variables (and intercept)  
## Forced in Forced out  
## pharvis FALSE FALSE  
## age FALSE FALSE  
## sexmale FALSE FALSE  
## educ FALSE FALSE  
## illness FALSE FALSE  
## illdays FALSE FALSE  
## actdays FALSE FALSE  
## communeurban FALSE FALSE  
## ifmarriedyes FALSE FALSE  
## ifinjuredyes FALSE FALSE  
## ifinsuredyes FALSE FALSE  
## 1 subsets of each size up to 11  
## Selection Algorithm: forward  
## pharvis age sexmale educ illness illdays actdays communeurban  
## 1 ( 1 ) " " " " " " " " " " " " " " "\*"   
## 2 ( 1 ) " " " " " " "\*" " " " " " " "\*"   
## 3 ( 1 ) " " " " " " "\*" "\*" " " " " "\*"   
## 4 ( 1 ) " " " " " " "\*" "\*" " " " " "\*"   
## 5 ( 1 ) "\*" " " " " "\*" "\*" " " " " "\*"   
## 6 ( 1 ) "\*" " " " " "\*" "\*" " " " " "\*"   
## 7 ( 1 ) "\*" "\*" " " "\*" "\*" " " " " "\*"   
## 8 ( 1 ) "\*" "\*" " " "\*" "\*" " " "\*" "\*"   
## 9 ( 1 ) "\*" "\*" "\*" "\*" "\*" " " "\*" "\*"   
## 10 ( 1 ) "\*" "\*" "\*" "\*" "\*" " " "\*" "\*"   
## 11 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*" "\*"   
## ifmarriedyes ifinjuredyes ifinsuredyes  
## 1 ( 1 ) " " " " " "   
## 2 ( 1 ) " " " " " "   
## 3 ( 1 ) " " " " " "   
## 4 ( 1 ) " " " " "\*"   
## 5 ( 1 ) " " " " "\*"   
## 6 ( 1 ) "\*" " " "\*"   
## 7 ( 1 ) "\*" " " "\*"   
## 8 ( 1 ) "\*" " " "\*"   
## 9 ( 1 ) "\*" " " "\*"   
## 10 ( 1 ) "\*" "\*" "\*"   
## 11 ( 1 ) "\*" "\*" "\*"

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| regression | RMSE | adj.R.2 | C.P | BIC |
| trial\_1 | 64.8076 | 0.2257 | 524.1566 | -3533.244 |
| trial\_2 | 63.9708 | 0.2455 | 156.6165 | -3884.549 |
| trial\_3 | 63.7122 | 0.2516 | 45.3747 | -3987.491 |
| trial\_4 | 63.6806 | 0.2522 | 33.5684 | -3991.729 |
| trial\_5 | 63.6471 | 0.2530 | 20.9537 | -3996.788 |
| trial\_6 | 63.6382 | 0.2531 | 19.0645 | -3991.137 |
| trial\_7 | 63.6082 | 0.2538 | 7.9798 | -3994.684 |
| trial\_8 | 63.6026 | 0.2539 | 7.5285 | -3987.599 |
| trial\_9 | 63.6009 | 0.2538 | 8.7733 | -3978.816 |
| trial\_10 | 63.5999 | 0.2538 | 10.3394 | -3969.712 |
| trial\_11 | 63.5991 | 0.2538 | 12.0000 | -3960.513 |

A smaller CP and BIC value is a more preferable sign for the corresponding model to be chosen. The smallest two CPs are from trial\_7 and trial\_8, while BIC’s smallest two numbers are in 5th and 7th trial.  
The adjusted R-squared stable after the 7th model. One point to be noticed is that the R-squared increase after adding two variables after trail\_5, therefore although trial\_5 has least BIC, 7th model has a slightly greater adjusted R-squared.  
I decided to use the 7th model. Also try manually add variables using the result above to see among trail\_5 to trail\_8, which to regard as the “best”.

##   
## Call:  
## lm(formula = lnhhexp ~ commune + educ + illness + ifinsured +   
## pharvis + ifmarried, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.94411 -0.35937 -0.02916 0.34448 2.54670   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.313420 0.010406 222.305 < 2e-16 \*\*\*  
## communeurban 0.610644 0.010721 56.956 < 2e-16 \*\*\*  
## educ 0.045000 0.002528 17.802 < 2e-16 \*\*\*  
## illness -0.063256 0.005671 -11.154 < 2e-16 \*\*\*  
## ifinsuredyes 0.049796 0.013006 3.829 0.000129 \*\*\*  
## pharvis 0.015222 0.003882 3.921 8.85e-05 \*\*\*  
## ifmarriedyes -0.018638 0.009455 -1.971 0.048716 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5402 on 13876 degrees of freedom  
## Multiple R-squared: 0.2535, Adjusted R-squared: 0.2531   
## F-statistic: 785.1 on 6 and 13876 DF, p-value: < 2.2e-16

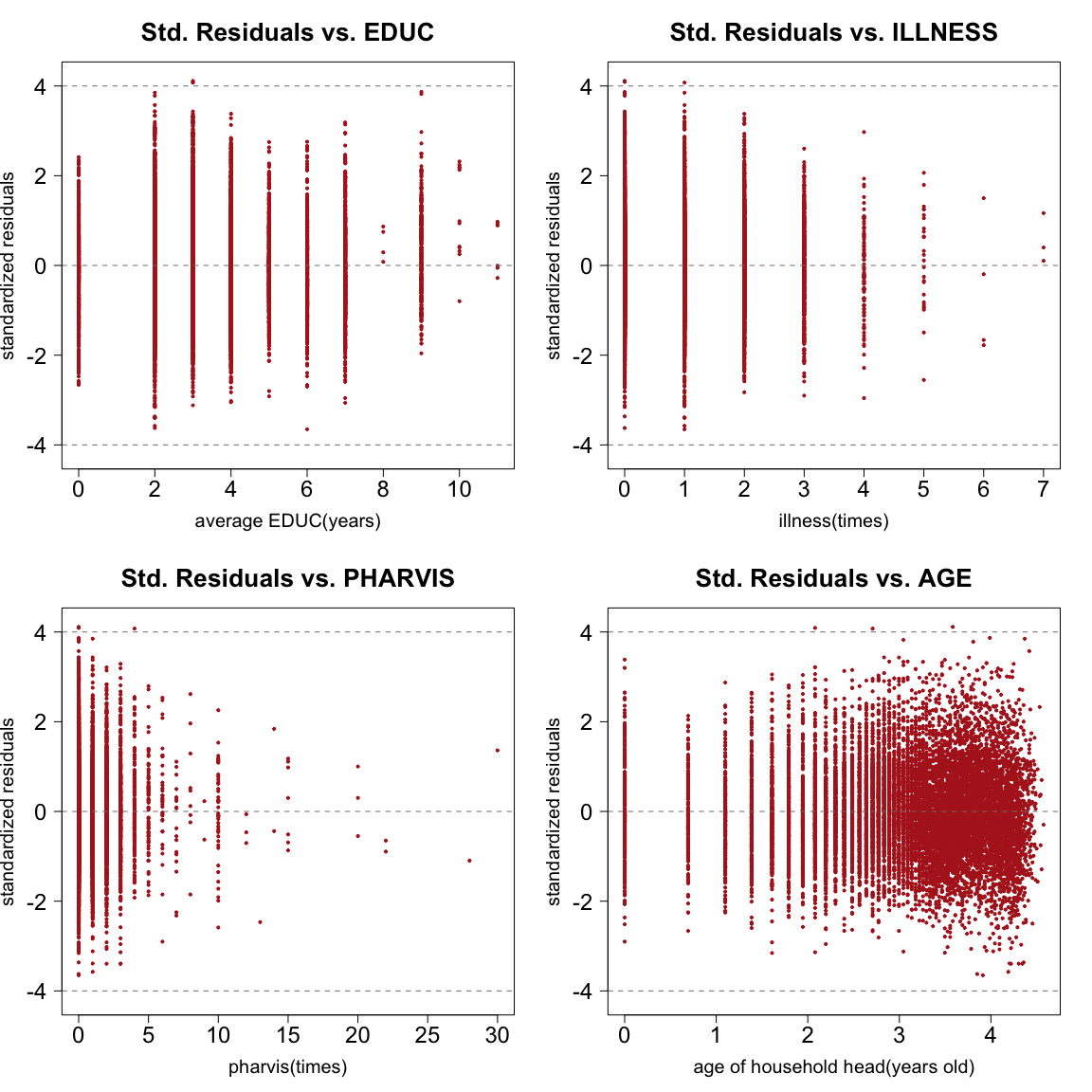
##   
## Call:  
## lm(formula = lnhhexp ~ commune + educ + illness + ifinsured +   
## pharvis + ifmarried + age, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.97042 -0.35881 -0.02796 0.34345 2.54647   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.259181 0.018249 123.795 < 2e-16 \*\*\*  
## communeurban 0.606552 0.010776 56.286 < 2e-16 \*\*\*  
## educ 0.045486 0.002530 17.976 < 2e-16 \*\*\*  
## illness -0.063635 0.005669 -11.224 < 2e-16 \*\*\*  
## ifinsuredyes 0.046247 0.013037 3.547 0.000390 \*\*\*  
## pharvis 0.014606 0.003884 3.761 0.000170 \*\*\*  
## ifmarriedyes -0.044788 0.011899 -3.764 0.000168 \*\*\*  
## age 0.021897 0.006054 3.617 0.000299 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.54 on 13875 degrees of freedom  
## Multiple R-squared: 0.2542, Adjusted R-squared: 0.2538   
## F-statistic: 675.4 on 7 and 13875 DF, p-value: < 2.2e-16

Apparently, after adding age into the model from trail\_6 to trial\_7, “age” is of more significance. Therefore I decide trial\_7 as the “best” from forward selection method.

Then using partial F test to compare model lm17 with the model without dropped variables.

## Analysis of Variance Table  
##   
## Model 1: lnhhexp ~ commune + educ + illness + ifinsured + pharvis + ifmarried +   
## age  
## Model 2: lnhhexp ~ pharvis + age + sex + educ + illness + illdays + actdays +   
## commune + ifmarried + ifinjured + ifinsured  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 13875 4046.0   
## 2 13871 4044.8 4 1.1605 0.995 0.4088

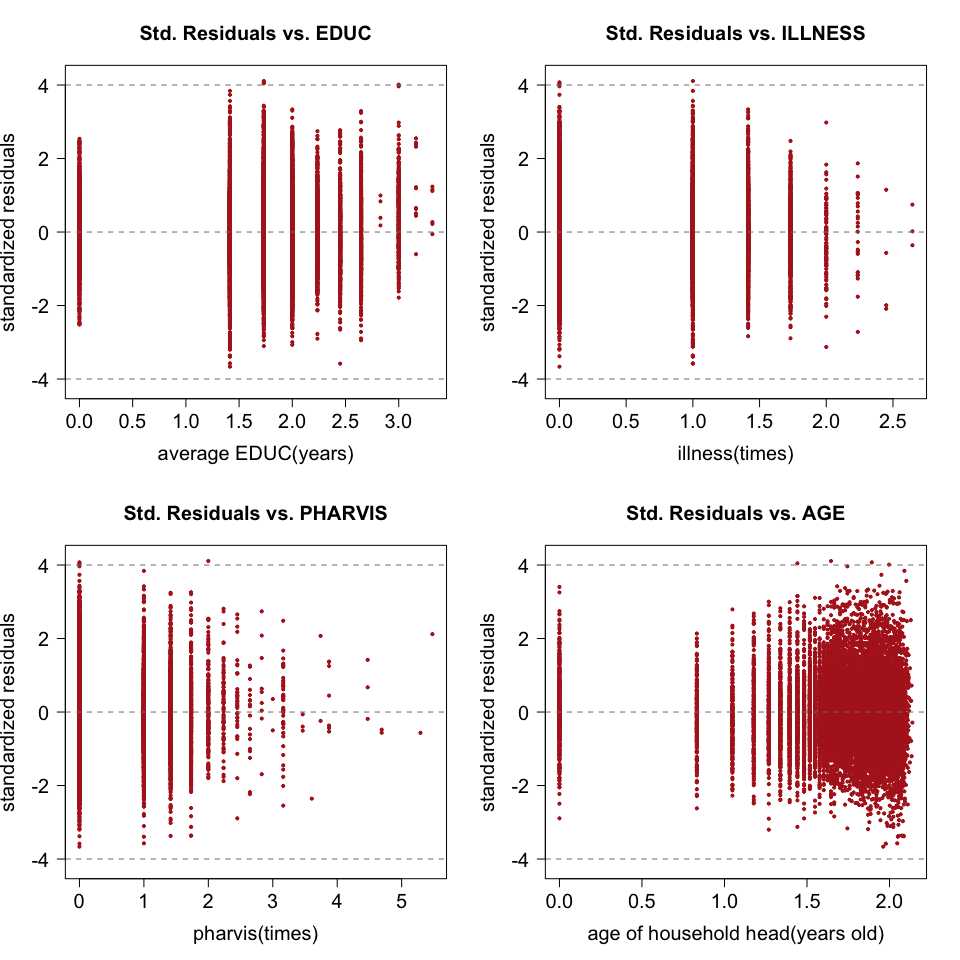
The p-value of partial F test is much greater than 0.05, so that we should not reject H0( all variables that is to be removed has zero beta estimate). It is reasonable to drop the rest four variables out of the model.

Then I check the residual plots for each numerical variables to see if there’s any transformation can be made. 

All four residual plots have many observations locates between 3 and 4. From summary table in part (b), they all have zero value, try new model with all these four variables taking square root instead of log.

##   
## Call:  
## lm(formula = lnhhexp ~ commune + sqrt(educ) + sqrt(illness) +   
## ifinsured + sqrt(pharvis) + ifmarried + sqrt(age), data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.97940 -0.35690 -0.02686 0.34466 2.58713   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.153672 0.025752 83.631 < 2e-16 \*\*\*  
## communeurban 0.615055 0.010692 57.526 < 2e-16 \*\*\*  
## sqrt(educ) 0.133213 0.007527 17.699 < 2e-16 \*\*\*  
## sqrt(illness) -0.086008 0.009399 -9.150 < 2e-16 \*\*\*  
## ifinsuredyes 0.054109 0.012964 4.174 3.02e-05 \*\*\*  
## sqrt(pharvis) 0.020484 0.009159 2.237 0.025327 \*   
## ifmarriedyes -0.039429 0.010977 -3.592 0.000329 \*\*\*  
## sqrt(age) 0.054668 0.014023 3.899 9.72e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5405 on 13875 degrees of freedom  
## Multiple R-squared: 0.2527, Adjusted R-squared: 0.2524   
## F-statistic: 670.4 on 7 and 13875 DF, p-value: < 2.2e-16

##   
## Call:  
## lm(formula = lnhhexp ~ commune + educ + illness + ifinsured +   
## pharvis + ifmarried + age, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.97042 -0.35881 -0.02796 0.34345 2.54647   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.259181 0.018249 123.795 < 2e-16 \*\*\*  
## communeurban 0.606552 0.010776 56.286 < 2e-16 \*\*\*  
## educ 0.045486 0.002530 17.976 < 2e-16 \*\*\*  
## illness -0.063635 0.005669 -11.224 < 2e-16 \*\*\*  
## ifinsuredyes 0.046247 0.013037 3.547 0.000390 \*\*\*  
## pharvis 0.014606 0.003884 3.761 0.000170 \*\*\*  
## ifmarriedyes -0.044788 0.011899 -3.764 0.000168 \*\*\*  
## age 0.021897 0.006054 3.617 0.000299 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.54 on 13875 degrees of freedom  
## Multiple R-squared: 0.2542, Adjusted R-squared: 0.2538   
## F-statistic: 675.4 on 7 and 13875 DF, p-value: < 2.2e-16



The shape of residual plots seems not change much after transformation. compare AIC and BIC of two models.

|  |  |  |  |
| --- | --- | --- | --- |
|  | a.r.squared | AIC | BIC |
| lm17 | 0.2538 | -17100.83 | -17040.52 |
| lm17.new-with transformation | 0.2524 | -17074.59 | -17014.28 |

## $`coefficient(lm17)`  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 2.2592 0.0182 123.7954 0e+00  
## communeurban 0.6066 0.0108 56.2859 0e+00  
## educ 0.0455 0.0025 17.9764 0e+00  
## illness -0.0636 0.0057 -11.2241 0e+00  
## ifinsuredyes 0.0462 0.0130 3.5474 4e-04  
## pharvis 0.0146 0.0039 3.7605 2e-04  
## ifmarriedyes -0.0448 0.0119 -3.7641 2e-04  
## age 0.0219 0.0061 3.6173 3e-04  
##   
## $`coefficient(lm17.new)`  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 2.1537 0.0258 83.6313 0.0000  
## communeurban 0.6151 0.0107 57.5261 0.0000  
## sqrt(educ) 0.1332 0.0075 17.6988 0.0000  
## sqrt(illness) -0.0860 0.0094 -9.1503 0.0000  
## ifinsuredyes 0.0541 0.0130 4.1736 0.0000  
## sqrt(pharvis) 0.0205 0.0092 2.2366 0.0253  
## ifmarriedyes -0.0394 0.0110 -3.5919 0.0003  
## sqrt(age) 0.0547 0.0140 3.8986 0.0001

From the table above, the difference is insignificant. lm17 has a slightly larger adjusted R-squared and a smaller AIC than lm17.new. Additionally, the p-value of lm17 performs better (all smaller than 0.001) than lm17.new. There may be other ways to get a better fit, but my finial decision is still lm17. The final model using “FORWARD SELECTION” method is lm17.

## lm(formula = lnhhexp ~ commune + educ + illness + ifinsured +   
## pharvis + ifmarried + age, data = train)

### (d) Use backward elimination to choose the “best” model. Explain how you decided which model was “best.”

The criterion of deciding the better model using backward elimination is delete variables with least p-value one at a time. With the help of the result, I compare RMSE, adjusted R-squared, CP value and BIC to decide the “best” two models, and do partial F test in the end to make final decision.

## Subset selection object  
## Call: regsubsets.formula(lnhhexp ~ ., data = train, method = "backward",   
## nvmax = 11)  
## 11 Variables (and intercept)  
## Forced in Forced out  
## pharvis FALSE FALSE  
## age FALSE FALSE  
## sexmale FALSE FALSE  
## educ FALSE FALSE  
## illness FALSE FALSE  
## illdays FALSE FALSE  
## actdays FALSE FALSE  
## communeurban FALSE FALSE  
## ifmarriedyes FALSE FALSE  
## ifinjuredyes FALSE FALSE  
## ifinsuredyes FALSE FALSE  
## 1 subsets of each size up to 11  
## Selection Algorithm: backward  
## pharvis age sexmale educ illness illdays actdays communeurban  
## 1 ( 1 ) " " " " " " " " " " " " " " "\*"   
## 2 ( 1 ) " " " " " " "\*" " " " " " " "\*"   
## 3 ( 1 ) " " " " " " "\*" "\*" " " " " "\*"   
## 4 ( 1 ) " " "\*" " " "\*" "\*" " " " " "\*"   
## 5 ( 1 ) " " "\*" " " "\*" "\*" " " " " "\*"   
## 6 ( 1 ) "\*" "\*" " " "\*" "\*" " " " " "\*"   
## 7 ( 1 ) "\*" "\*" " " "\*" "\*" " " " " "\*"   
## 8 ( 1 ) "\*" "\*" " " "\*" "\*" " " "\*" "\*"   
## 9 ( 1 ) "\*" "\*" "\*" "\*" "\*" " " "\*" "\*"   
## 10 ( 1 ) "\*" "\*" "\*" "\*" "\*" " " "\*" "\*"   
## 11 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*" "\*"   
## ifmarriedyes ifinjuredyes ifinsuredyes  
## 1 ( 1 ) " " " " " "   
## 2 ( 1 ) " " " " " "   
## 3 ( 1 ) " " " " " "   
## 4 ( 1 ) " " " " " "   
## 5 ( 1 ) "\*" " " " "   
## 6 ( 1 ) "\*" " " " "   
## 7 ( 1 ) "\*" " " "\*"   
## 8 ( 1 ) "\*" " " "\*"   
## 9 ( 1 ) "\*" " " "\*"   
## 10 ( 1 ) "\*" "\*" "\*"   
## 11 ( 1 ) "\*" "\*" "\*"

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| regression | RMSE | adj.R.2 | C.P | BIC |
| trial\_1 | 64.8076 | 0.2257 | 524.1566 | -3533.244 |
| trial\_2 | 63.9708 | 0.2455 | 156.6165 | -3884.549 |
| trial\_3 | 63.7122 | 0.2516 | 45.3747 | -3987.491 |
| trial\_4 | 63.7020 | 0.2517 | 42.8996 | -3982.416 |
| trial\_5 | 63.6672 | 0.2525 | 29.7115 | -3988.039 |
| trial\_6 | 63.6371 | 0.2532 | 18.5641 | -3991.637 |
| trial\_7 | 63.6082 | 0.2538 | 7.9798 | -3994.684 |
| trial\_8 | 63.6026 | 0.2539 | 7.5285 | -3987.599 |
| trial\_9 | 63.6009 | 0.2538 | 8.7733 | -3978.816 |
| trial\_10 | 63.5999 | 0.2538 | 10.3394 | -3969.712 |
| trial\_11 | 63.5991 | 0.2538 | 12.0000 | -3960.513 |

RMSE and adjusted R-squared stable after the 7th model. A smaller CP and BIC value is a more preferable sign for the corresponding model to be chosen. The smallest two CPs are from trial\_7 and trial\_8, while BIC is smallest (-3994.68) in trial\_7. Same as in (c), I manually check the significance level of the regression variables in trial\_6 to trial\_8 to make final decision.

##   
## Call:  
## lm(formula = lnhhexp ~ commune + educ + illness + age + ifmarried +   
## pharvis + ifinsured, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.97042 -0.35881 -0.02796 0.34345 2.54647   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.259181 0.018249 123.795 < 2e-16 \*\*\*  
## communeurban 0.606552 0.010776 56.286 < 2e-16 \*\*\*  
## educ 0.045486 0.002530 17.976 < 2e-16 \*\*\*  
## illness -0.063635 0.005669 -11.224 < 2e-16 \*\*\*  
## age 0.021897 0.006054 3.617 0.000299 \*\*\*  
## ifmarriedyes -0.044788 0.011899 -3.764 0.000168 \*\*\*  
## pharvis 0.014606 0.003884 3.761 0.000170 \*\*\*  
## ifinsuredyes 0.046247 0.013037 3.547 0.000390 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.54 on 13875 degrees of freedom  
## Multiple R-squared: 0.2542, Adjusted R-squared: 0.2538   
## F-statistic: 675.4 on 7 and 13875 DF, p-value: < 2.2e-16

##   
## Call:  
## lm(formula = lnhhexp ~ commune + educ + illness + age + ifmarried +   
## pharvis + ifinsured + actdays, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.9711 -0.3590 -0.0279 0.3435 2.5459   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.258881 0.018249 123.779 < 2e-16 \*\*\*  
## communeurban 0.606498 0.010776 56.284 < 2e-16 \*\*\*  
## educ 0.045520 0.002530 17.990 < 2e-16 \*\*\*  
## illness -0.063570 0.005669 -11.213 < 2e-16 \*\*\*  
## age 0.022082 0.006054 3.647 0.000266 \*\*\*  
## ifmarriedyes -0.044833 0.011898 -3.768 0.000165 \*\*\*  
## pharvis 0.014718 0.003884 3.789 0.000152 \*\*\*  
## ifinsuredyes 0.046253 0.013036 3.548 0.000389 \*\*\*  
## actdays -0.006299 0.004023 -1.566 0.117429   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.54 on 13874 degrees of freedom  
## Multiple R-squared: 0.2543, Adjusted R-squared: 0.2539   
## F-statistic: 591.4 on 8 and 13874 DF, p-value: < 2.2e-16

“actdays” are not significant after being added in the model, so I decided to stop at trail\_7, which happens to have the same variables as that in (c), with small difference in the order of variables added in the model. So I omit the process of partial F-test here.

## lm(formula = lnhhexp ~ commune + educ + illness + age + ifmarried +   
## pharvis + ifinsured, data = train)

### (e) Use the sequential replacement(or called forward and backward stepwise)method to choose the “best” model (seqrep in the regsubsets() function). Explain how you decided which model was “best.”

## Subset selection object  
## Call: regsubsets.formula(lnhhexp ~ ., data = train, method = "seqrep",   
## nvmax = 11)  
## 11 Variables (and intercept)  
## Forced in Forced out  
## pharvis FALSE FALSE  
## age FALSE FALSE  
## sexmale FALSE FALSE  
## educ FALSE FALSE  
## illness FALSE FALSE  
## illdays FALSE FALSE  
## actdays FALSE FALSE  
## communeurban FALSE FALSE  
## ifmarriedyes FALSE FALSE  
## ifinjuredyes FALSE FALSE  
## ifinsuredyes FALSE FALSE  
## 1 subsets of each size up to 11  
## Selection Algorithm: 'sequential replacement'  
## pharvis age sexmale educ illness illdays actdays communeurban  
## 1 ( 1 ) " " " " " " " " " " " " " " "\*"   
## 2 ( 1 ) " " " " " " "\*" " " " " " " "\*"   
## 3 ( 1 ) " " " " " " "\*" "\*" " " " " "\*"   
## 4 ( 1 ) " " " " " " "\*" "\*" " " " " "\*"   
## 5 ( 1 ) "\*" "\*" "\*" "\*" "\*" " " " " " "   
## 6 ( 1 ) "\*" "\*" " " "\*" "\*" " " " " "\*"   
## 7 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*" " "   
## 8 ( 1 ) "\*" "\*" " " "\*" "\*" " " "\*" "\*"   
## 9 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*" "\*"   
## 10 ( 1 ) "\*" "\*" "\*" "\*" "\*" " " "\*" "\*"   
## 11 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*" "\*"   
## ifmarriedyes ifinjuredyes ifinsuredyes  
## 1 ( 1 ) " " " " " "   
## 2 ( 1 ) " " " " " "   
## 3 ( 1 ) " " " " " "   
## 4 ( 1 ) " " " " "\*"   
## 5 ( 1 ) " " " " " "   
## 6 ( 1 ) "\*" " " " "   
## 7 ( 1 ) " " " " " "   
## 8 ( 1 ) "\*" " " "\*"   
## 9 ( 1 ) "\*" " " " "   
## 10 ( 1 ) "\*" "\*" "\*"   
## 11 ( 1 ) "\*" "\*" "\*"

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| regression | RMSE | adj.R.2 | C.P | BIC |
| trial\_1 | 64.8076 | 0.2257 | 524.1566 | -3533.244 |
| trial\_2 | 63.9708 | 0.2455 | 156.6165 | -3884.549 |
| trial\_3 | 63.7122 | 0.2516 | 45.3747 | -3987.491 |
| trial\_4 | 63.6806 | 0.2522 | 33.5684 | -3991.729 |
| trial\_5 | 70.8504 | 0.0743 | 3343.3203 | -1019.831 |
| trial\_6 | 63.6371 | 0.2532 | 18.5641 | -3991.637 |
| trial\_7 | 70.8348 | 0.0746 | 3339.7702 | -1006.845 |
| trial\_8 | 63.6026 | 0.2539 | 7.5285 | -3987.599 |
| trial\_9 | 63.6283 | 0.2532 | 20.7313 | -3966.854 |
| trial\_10 | 63.5999 | 0.2538 | 10.3394 | -3969.712 |
| trial\_11 | 63.5991 | 0.2538 | 12.0000 | -3960.513 |

Same as the criterion using above, the smallest RMSE(63.6026) and CP(7.53) and largest adjusted R-squared is from trial\_8, and BIC of it is the third smallest. So I chose trail\_7, which contains “pharvis”, “age”, “educ”, “illness”, “actdays”, “commune”, “ifmarried”, “ifinsured” in the model. It is the same as model in lm28, from which we’ve proved that “actdays” will be insignificant if it is in the model. Hence the model becomes, again, same as that in (c). I also tried the 6th result, with variables “pharvis”, “age”, “educ”, “illness”, “commune” and “ifmarried”, which I think also performs well in BIC and adjusted R-squared. But it cannot pass the partial F-test. The partial F-test has p-value smaller than 0.05, and results in the rejection of all variables tending to remove has zero coefficient.

## Analysis of Variance Table  
##   
## Model 1: lnhhexp ~ pharvis + age + educ + illness + commune + ifmarried  
## Model 2: lnhhexp ~ pharvis + age + sex + educ + illness + illdays + actdays +   
## commune + ifmarried + ifinjured + ifinsured  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 13876 4049.7   
## 2 13871 4044.8 5 4.8302 3.3128 0.005424 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

My final decision of the model is still

## lm(formula = lnhhexp ~ commune + educ + illness + ifinsured +   
## pharvis + ifmarried + age, data = train)

### (f) Compute the test set RMSE for the models you chose in 2c, 2d, and 2e. Compare them to the RMSE values from the training set. Which model performs best? Justify your answer.

## train.RMSE test.RMSE   
## 0.5400 0.5355

The result in (c), (d) and (e) is the same model, no comparision is needed here. As for the model’s performance, lm17 is a good fit, because the RMSE of the test set is close to that from train set.

### (g) Check your regression assumptions (including collinearity/multicollinearity) for the model you chose in question 2f. Are the assumptions satisfied? Justify your answer.

# ASSUMPTION CHECK

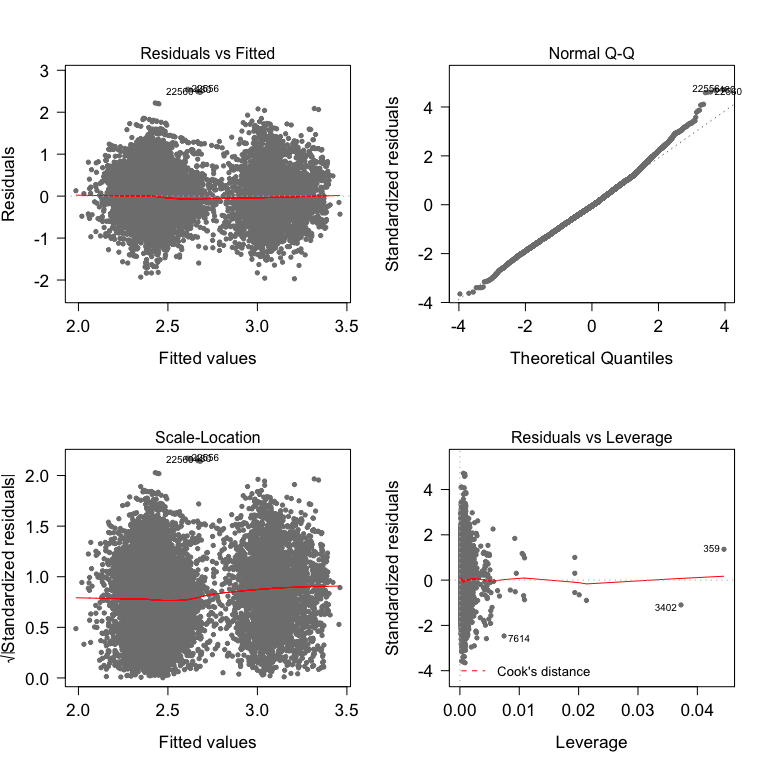
(1)Multicollinearity check

## Variables VIF  
## 1 educ 1.003942  
## 2 illness 1.630615  
## 3 age 1.023534  
## 4 pharvis 1.210104  
## 5 illdays 1.532340

Same as the result above, the VIFs are close to 1 and the estimates are positive, consistent with the relationship shown in scatter plot. There is no colinearity problem.

(2)Fixed “X” and measurement error: it is acknowledged for the original data since it is from Vietnam World Bank Survey;

(3)Zero conditional mean This is an assumption when we do the multi-linear regression.

1. Normally distributed residuals  From the top right plot, residuals has slightly upward trend when x-axis move to a larger amount. But a good sign is the well-fitted part in the middle. Since the observation number is quite large, it is not severe enough to be worried about the extreme large values since the normally distributed observations take up a large absolute number. Data can still be concluded as normally distributed. But it is recommended to deal with extreme values (sicne taking square root of large number seems not a good choice) in later study.

(6)Constant variance of the error term (regardless of “X” values); Neither the residual versus fitted (top left) nor scale-location (bottom left) plot shows the variability of the residuals getting larger as one moves from one side of the plot to the other.

From the bottom right plot, there is an observation that has a relatively “large”" leverage on the regression. CHECK COOK’S DISTANCE

## [1] FALSE

However, there’s no observations whose cook’s distance is greater than 0.1, therefore no further remove of observations is needed. Hence the assumptions are all satisfied.

### (h) Interpret the partial slopes and y-intercept of the model you chose in question 2f.

My model in 2(f) is

##   
## Call:  
## lm(formula = lnhhexp ~ commune + educ + illness + ifinsured +   
## pharvis + ifmarried + age, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.97042 -0.35881 -0.02796 0.34345 2.54647   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.259181 0.018249 123.795 < 2e-16 \*\*\*  
## communeurban 0.606552 0.010776 56.286 < 2e-16 \*\*\*  
## educ 0.045486 0.002530 17.976 < 2e-16 \*\*\*  
## illness -0.063635 0.005669 -11.224 < 2e-16 \*\*\*  
## ifinsuredyes 0.046247 0.013037 3.547 0.000390 \*\*\*  
## pharvis 0.014606 0.003884 3.761 0.000170 \*\*\*  
## ifmarriedyes -0.044788 0.011899 -3.764 0.000168 \*\*\*  
## age 0.021897 0.006054 3.617 0.000299 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.54 on 13875 degrees of freedom  
## Multiple R-squared: 0.2542, Adjusted R-squared: 0.2538   
## F-statistic: 675.4 on 7 and 13875 DF, p-value: < 2.2e-16

1. interpreting y-intercept

## (Intercept)   
## 9.575241

Uninsured (“ifinsured”=“no”) single(“ifmarried”=“no”) citizens who live in rural area(“commune”=0) of Vietnam with zero years of education(“educ”=0), zero times of illnesses experience(“illness”), direct pharmacy visits (“pharvis”=0) within the past year and household head age is zero (“age”=0) will on average spend 9.58 (dollar) as total medical expenditures (exponential of original intercept).

## educ illness pharvis age   
## Min. : 0.000 Min. :0.0000 Min. : 0.0000 Min. :0.000   
## 1st Qu.: 2.000 1st Qu.:0.0000 1st Qu.: 0.0000 1st Qu.:2.485   
## Median : 3.000 Median :0.0000 Median : 0.0000 Median :3.135   
## Mean : 3.374 Mean :0.6182 Mean : 0.5162 Mean :2.979   
## 3rd Qu.: 4.000 3rd Qu.:1.0000 3rd Qu.: 1.0000 3rd Qu.:3.714   
## Max. :11.000 Max. :7.0000 Max. :30.0000 Max. :4.575

All the numerical variables have sample at zero, therefore in this problem, the situation hypothesized above can be reached, even age=0, so this intercept has meaning.

1. interpreting beta for COMMUNE

## [1] 83.40959

Holding all other variables constant, people living in urban areas will on average spend approximately 83.41% in total medical expenditures more than those who live in rural areas. This can be easily explained by (1)“high expenditure on everything in urban area” and (2)“high awareness of go to doctors and spending money on healthcare” for urban citizens.

1. interpreting beta for EDUC

## [1] 4.653607

Holding all other variables constant, one year increase of education years is associated with approximately 4.65% increase in the total medical expenditures. This can be expalined by (1) a longer term of education will increase the awareness of self health care, and be more likely to spend money for medical care and (2) longer years of education tends to imply a higher age, which causea a higher likelyhood of medicare.

1. interpreting beta for ILLNESS

## [1] -6.165229

Holding all other variables constant, one time increase of illness experienced in the past year is associated with approximately 6.17% decrease in the total medical expenditures. It seems obey the common sense. Since the more times you experienced illness, usually the higher you will spend for treatments. Further investigation is needed.

1. interpreting beta for IFINSURED

## [1] 4.733329

Holding all other variables constant, if one has health insurance coverage, he/she will spend approximately 4.73% in total medical expenditures more than those who don’t have insurance coverage. If the expenditure means the part people have to pay before compensation, it makes sense. Because people would be more likely to use a better and more expensive treatment if they were covered by insurance.

1. interpreting beta for PHARVIS

## [1] 0.01460567

## [1] 1.471285

Holding all other variables constant, one time increase of direct pharmacy visit is associated with approximately 1.47% increase in total medical expenditures. This can be intuitively understood, because everytime you visit the pharmacy you must have something wrong and thus expenditure will be made.

1. interpreting beta for IFMARRIED

## [1] -4.380015

Holding all other variables constant, married status (not single) is associated with approximately 4.38% decrease in the total medical expenditures. This may be caused by married people can take care of each other and have less chance to pay medical expenditure.

1. interpreting beta for AGE

## [1] 2.213877

Holding all other variables constant, one year increase of the age of household head is associated with approximately 2.21% increase in the total medical expenditures. This can also be explained same as the situation in “EDUC”, the older person is, the higher chance he/she will spend a high amount of medical expenditure.