Final2

Harry Woo

2020 6 12

### Problem 2

<https://socialsciences.mcmaster.ca/jfox/Courses/SPIDA/> <https://sms.wgtn.ac.nz/foswiki/pub/Courses/DATA303_2020T1/Lab2/lab2.html> <https://sms.wgtn.ac.nz/foswiki/pub/Courses/DATA303_2020T1/Lab3/lab3.html> <https://sms.wgtn.ac.nz/foswiki/pub/Courses/DATA303_2020T1/Lab4/lab4.html> <http://www.jpstats.org/Regression/ch_05_05.html>

*After exploring the data graphically, perform a linear least‐squares regression of the total fertility rate (tfr) on GDP per capita (GDPperCapita), the female illiteracy rate (illiteracyFemale), and the rate of contraceptive use by married women (contraception).*

*Introduce the factor region into the regression, first fitting an additive model to the data and then considering the possibility that region interacts with the other explanatory variables.*

#### 2.0 Data loading

un <- read.table("http://socialsciences.mcmaster.ca/jfox/Books/Applied-Regression-2E/datasets/UnitedNations.txt", header = TRUE)  
  
str(un)

## 'data.frame': 207 obs. of 13 variables:  
## $ region : chr "Asia" "Europe" "Africa" "Asia" ...  
## $ tfr : num 6.9 2.6 3.81 NA NA 6.69 NA 2.62 1.7 1.89 ...  
## $ contraception : int NA NA 52 NA NA NA 53 NA 22 76 ...  
## $ educationMale : num NA NA 11.1 NA NA NA NA NA NA 16.3 ...  
## $ educationFemale : num NA NA 9.9 NA NA NA NA NA NA 16.1 ...  
## $ lifeMale : num 45 68 67.5 68 NA 44.9 NA 69.6 67.2 75.4 ...  
## $ lifeFemale : num 46 74 70.3 73 NA 48.1 NA 76.8 74 81.2 ...  
## $ infantMortality : int 154 32 44 11 NA 124 24 22 25 6 ...  
## $ GDPperCapita : int 2848 863 1531 NA NA 355 6966 8055 354 20046 ...  
## $ economicActivityMale : num 87.5 NA 76.4 58.8 NA NA 74.4 76.2 65 74 ...  
## $ economicActivityFemale: num 7.2 NA 7.8 42.4 NA NA 56.2 41.3 52 53.8 ...  
## $ illiteracyMale : num 52.8 NA 26.1 0.264 NA NA NA 3.8 0.3 NA ...  
## $ illiteracyFemale : num 85 NA 51 0.36 NA NA NA 3.8 0.5 NA ...

summary(un)

## region tfr contraception educationMale   
## Length:207 Min. :1.190 Min. : 2.00 Min. : 3.30   
## Class :character 1st Qu.:1.950 1st Qu.:21.00 1st Qu.: 9.75   
## Mode :character Median :3.070 Median :47.00 Median :11.25   
## Mean :3.529 Mean :43.43 Mean :11.41   
## 3rd Qu.:4.980 3rd Qu.:64.00 3rd Qu.:13.90   
## Max. :8.000 Max. :86.00 Max. :17.20   
## NA's :10 NA's :63 NA's :131   
## educationFemale lifeMale lifeFemale infantMortality   
## Min. : 2.000 Min. :36.00 Min. :39.10 Min. : 2.00   
## 1st Qu.: 9.325 1st Qu.:57.38 1st Qu.:59.60 1st Qu.: 12.00   
## Median :11.650 Median :66.50 Median :72.15 Median : 30.00   
## Mean :11.275 Mean :63.63 Mean :68.39 Mean : 43.48   
## 3rd Qu.:13.650 3rd Qu.:70.90 3rd Qu.:76.42 3rd Qu.: 66.00   
## Max. :17.800 Max. :77.40 Max. :82.90 Max. :169.00   
## NA's :131 NA's :11 NA's :11 NA's :6   
## GDPperCapita economicActivityMale economicActivityFemale illiteracyMale   
## Min. : 36 Min. :51.20 Min. : 1.90 Min. : 0.200   
## 1st Qu.: 442 1st Qu.:72.30 1st Qu.:37.00 1st Qu.: 2.952   
## Median : 1779 Median :76.80 Median :48.40 Median :10.829   
## Mean : 6262 Mean :76.46 Mean :46.79 Mean :17.555   
## 3rd Qu.: 7272 3rd Qu.:81.20 3rd Qu.:56.40 3rd Qu.:27.575   
## Max. :42416 Max. :93.00 Max. :90.60 Max. :79.100   
## NA's :10 NA's :42 NA's :42 NA's :47   
## illiteracyFemale  
## Min. : 0.200   
## 1st Qu.: 4.847   
## Median :20.100   
## Mean :27.906   
## 3rd Qu.:48.025   
## Max. :93.400   
## NA's :47

데이터를 un으로 로딩한 후 기술통계량을 확인한 결과, 상당수의 결측치가 존재하는 것으로 확인되는 바, 우선 결측치 처리를 진행해야 하는 것으로 판단된다.

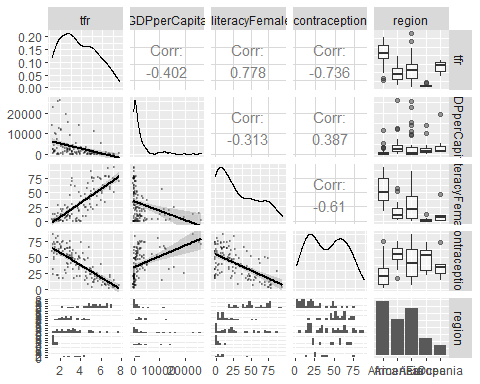
# NA 처리, 변수 선택  
un2 <- na.omit(un[,c("tfr", "GDPperCapita", "illiteracyFemale", "contraception", "region")])

na.omit으로 결측치를 제거하고, 문제에서 제시한 변수인 tfr, GDPperCapita, illiteracyFemale, contraception, region 을 선택하여 un2로 저장하였다.

#### 2.1 Exploring the data graphically

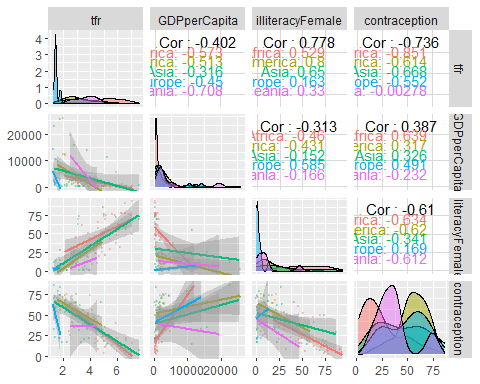
ggpairs(un2,  
 lower = list(continuous = wrap("smooth", alpha = 0.3, size = 0.1)),  
 diag = list(discrete="barDiag", continuous = wrap("densityDiag", alpha = 0.5 )),  
 upper = list(combo = wrap("box\_no\_facet", alpha = 0.5),  
 continuous = wrap("cor", size = 4,alignPercent = 0.8))) +  
 theme(legend.position = "bottom")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



산점도/상관계수 행렬을 통해 살펴본 결과, GPDperCapita와 illiteracyFemale 두 변수는 치우침을 고려하여 log transformation을 시도할 필요가 있는 것으로 보인다. 종속변수 tfr은 GDPperCapita 및 Contraception과 음의 상관관계를, illiteracyFemale과는 양의 상관관계를 가지며, 적합된 회귀선의 양상으로는 다른 변수들 간에도 상관관계가 존재하는 것으로 보인다.

# by region  
ggpairs(un2[,1:4], aes(colour = un2$region),  
 lower = list(continuous = wrap("smooth", alpha = 0.3, size = 0.1)),  
 diag = list(discrete="barDiag", continuous = wrap("densityDiag", alpha = 0.5 )),  
 upper = list(combo = wrap("box\_no\_facet", alpha = 0.5),  
 continuous = wrap("cor", size = 4,alignPercent = 0.8))) +  
 theme(legend.position = "bottom")



region 변수의 영향을 확인하기 위해 새로운 산점도/상관계수 행렬을 작성하였다. 특히 치우친 형태의 분포를 보이던 변수들의 분포를 지역을 기준으로 다시 그리자 지역별로 양상이 크게 다르게 나타나며, 특히 유럽이 다른 지역과 차이를 보이는 것을 확인할 수 있다. 향후 모델에 region 변수를 추가하는 것이 합리적인 것으로 판단된다.

#### 2.2 Linear least-squares regression

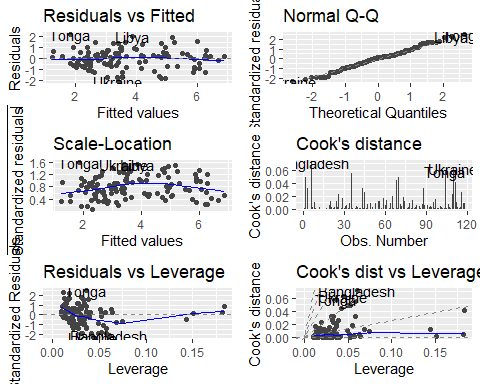
un\_lm <- lm(tfr ~ GDPperCapita + illiteracyFemale + contraception, data = un2)   
summary(un\_lm)

##   
## Call:  
## lm(formula = tfr ~ GDPperCapita + illiteracyFemale + contraception,   
## data = un2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.02594 -0.58644 0.08299 0.51946 2.17032   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.021e+00 2.837e-01 14.175 < 2e-16 \*\*\*  
## GDPperCapita -2.998e-05 1.771e-05 -1.693 0.0933 .   
## illiteracyFemale 3.188e-02 3.905e-03 8.165 4.83e-13 \*\*\*  
## contraception -2.884e-02 4.800e-03 -6.008 2.30e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8952 on 114 degrees of freedom  
## Multiple R-squared: 0.7212, Adjusted R-squared: 0.7138   
## F-statistic: 98.28 on 3 and 114 DF, p-value: < 2.2e-16

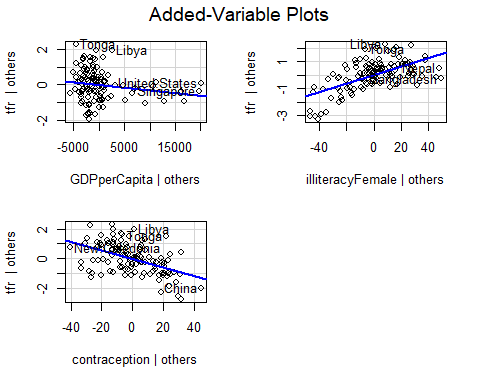
문제에서 주어진 바와 같이 region을 제외한 설명변수를 기준으로 회귀모형 un\_lm을 적합하였다. illiteracyFemale과 contraception이 매우 유의하게 나타났고, 의아하게도 GDPperCapita는 유의수준 0.05에서 유의하지 않은 것으로 나타났다. 결정계수는 72% 수준으로 양호하다.

##### Diagnosis

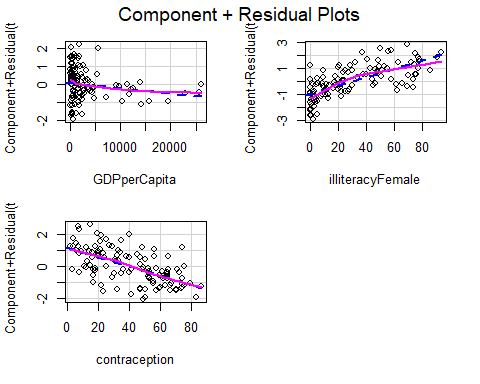
autoplot(un\_lm, which = c(1:6))



avPlots(un\_lm, ask=FALSE)



crPlots(un\_lm, ask=FALSE)



Regression diagnostic plot을 통해 확인한 결과, 오차의 선형성, 정규성, 등분산성에는 큰 문제가 없는 것으로 판단된다. Outlier 및 influential point와 관련하여서도 Tonga, Bangladesh 등의 데이터가 반복적으로 튀고 있으나 cook’s distance 수치가 양호하며 added variable plot 상에서도 큰 이슈는 없는 것으로 보인다.

그러나 Component+residual plot을 통해 확인한 결과 GDPperCapita와 illiteracyFemale에 nonlinearity 이슈가 있는 것으로 판단되는 바, 최초 graphical 분석에서 확인한 바와 같이 두 변수에 log transformation을 시도하였다.

##### Log transformation

un\_lm2 <- lm(tfr ~ log(GDPperCapita) + log(illiteracyFemale) + contraception, data = un2)   
summary(un\_lm2)

##   
## Call:  
## lm(formula = tfr ~ log(GDPperCapita) + log(illiteracyFemale) +   
## contraception, data = un2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.31858 -0.54751 -0.00501 0.50938 3.05329   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.520983 0.477317 11.567 < 2e-16 \*\*\*  
## log(GDPperCapita) -0.237220 0.070972 -3.342 0.00112 \*\*   
## log(illiteracyFemale) 0.465621 0.052299 8.903 9.85e-15 \*\*\*  
## contraception -0.033228 0.004569 -7.272 4.79e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8528 on 114 degrees of freedom  
## Multiple R-squared: 0.7469, Adjusted R-squared: 0.7402   
## F-statistic: 112.1 on 3 and 114 DF, p-value: < 2.2e-16

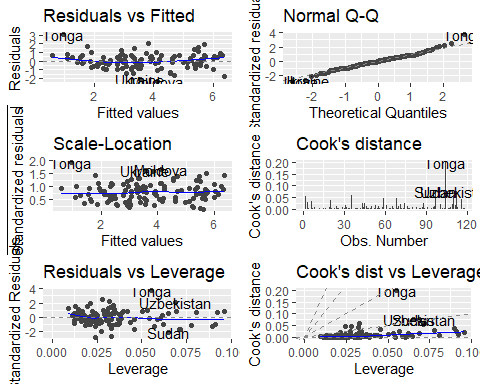
anova(un\_lm, un\_lm2)

## Analysis of Variance Table  
##   
## Model 1: tfr ~ GDPperCapita + illiteracyFemale + contraception  
## Model 2: tfr ~ log(GDPperCapita) + log(illiteracyFemale) + contraception  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 114 91.348   
## 2 114 82.916 0 8.4326

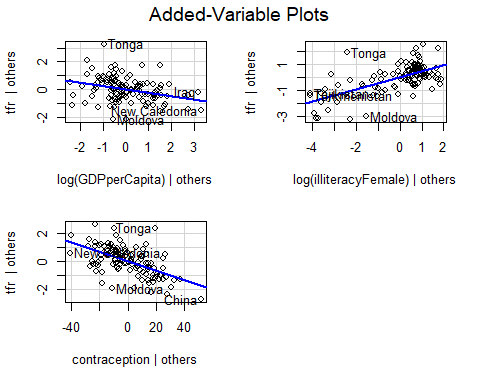
로그변환 후 회귀모형을 un\_lm2로 적합하여 summary를 확인한 결과, 세 설명변수가 모두 유의하며, 특히 인당 GDP 변수가 통계적으로 유의한 것으로 확인되었다. 결정계수 또한 소폭 향상되었다.

##### Diagnosis (2)

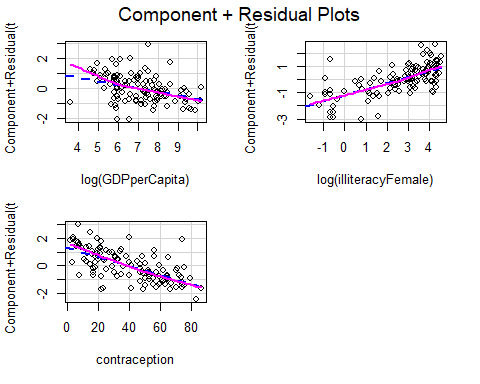
autoplot(un\_lm2, which = c(1:6))



avPlots(un\_lm2, ask=FALSE)



crPlots(un\_lm2, ask=FALSE)

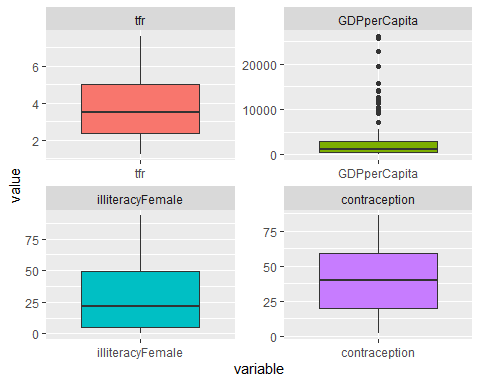


un2[row.names(un2) == "Tonga",]

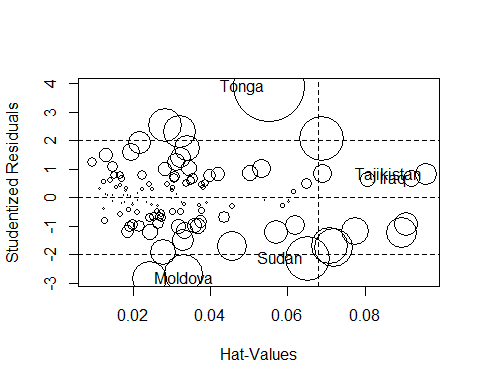
## tfr GDPperCapita illiteracyFemale contraception region  
## Tonga 4.02 1787 0.504 74 Oceania

# multiple boxplot using reshape2 package  
ggplot(data = melt(un2[,1:4]), aes(x = variable, y = value)) +   
 geom\_boxplot(aes(fill = variable)) + facet\_wrap(~variable, scale="free") +  
 theme(legend.position = "none")

## No id variables; using all as measure variables



influencePlot(un\_lm2)



## StudRes Hat CookD  
## Iraq 0.6657241 0.09196411 0.01127640  
## Moldova -2.8360846 0.02431565 0.04719748  
## Sudan -2.1450482 0.06501132 0.07753357  
## Tajikistan 0.8288009 0.09577489 0.01823938  
## Tonga 3.9070559 0.05526690 0.19842281

# exluding Tonga?  
summary(lm(tfr ~ log(GDPperCapita) + log(illiteracyFemale) + contraception, data = un2,  
 subset = row.names(un2) != "Tonga"))

##   
## Call:  
## lm(formula = tfr ~ log(GDPperCapita) + log(illiteracyFemale) +   
## contraception, data = un2, subset = row.names(un2) != "Tonga")  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.26968 -0.48479 -0.00185 0.49727 2.05735   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.342560 0.452303 11.812 < 2e-16 \*\*\*  
## log(GDPperCapita) -0.216454 0.067120 -3.225 0.00165 \*\*   
## log(illiteracyFemale) 0.494802 0.049867 9.922 < 2e-16 \*\*\*  
## contraception -0.035062 0.004333 -8.091 7.44e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.804 on 113 degrees of freedom  
## Multiple R-squared: 0.777, Adjusted R-squared: 0.771   
## F-statistic: 131.2 on 3 and 113 DF, p-value: < 2.2e-16

모형진단 결과, 오차의 선형성, 정규성과 등분산성에 문제가 없으며 Component+residual plot 또한 크게 개선된 것으로 보인다. 다만 cook’s distance 수치가 전반적으로 다소 상승하였으며, 지속적으로 Tonga 데이터가 튀고 있는 것으로 보인다. 이에 Tonga 를 제외한 모형을 적합해 본 결과, 다소 수치가 개선되었으나 본래 모형과 큰 차이는 없는 것으로 보여 기존 모형 un\_lm2를 유지하였다.

#### 2.3 Introducing the factor region

##### 2.3.1 Fitting and additive model

un\_lm3 <- lm(tfr ~ log(GDPperCapita) + log(illiteracyFemale) + contraception + region, data = un2)   
summary(un\_lm3)

##   
## Call:  
## lm(formula = tfr ~ log(GDPperCapita) + log(illiteracyFemale) +   
## contraception + region, data = un2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.79793 -0.44651 0.01122 0.32643 2.34340   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.194247 0.463040 13.377 < 2e-16 \*\*\*  
## log(GDPperCapita) -0.220621 0.067832 -3.252 0.00152 \*\*   
## log(illiteracyFemale) 0.299507 0.061401 4.878 3.64e-06 \*\*\*  
## contraception -0.030977 0.004491 -6.897 3.53e-10 \*\*\*  
## regionAmerica -0.505058 0.245030 -2.061 0.04164 \*   
## regionAsia -0.536170 0.212904 -2.518 0.01323 \*   
## regionEurope -1.656766 0.346862 -4.776 5.54e-06 \*\*\*  
## regionOceania -0.194681 0.360201 -0.540 0.58996   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7829 on 110 degrees of freedom  
## Multiple R-squared: 0.7942, Adjusted R-squared: 0.7811   
## F-statistic: 60.64 on 7 and 110 DF, p-value: < 2.2e-16

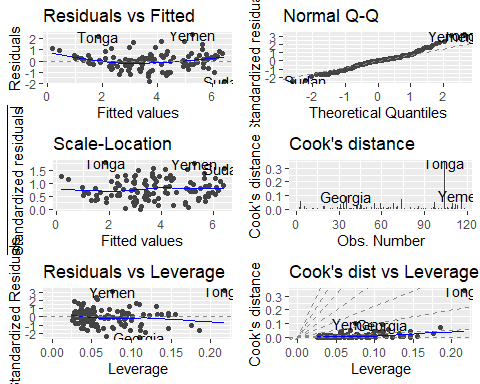
anova(un\_lm2, un\_lm3)

## Analysis of Variance Table  
##   
## Model 1: tfr ~ log(GDPperCapita) + log(illiteracyFemale) + contraception  
## Model 2: tfr ~ log(GDPperCapita) + log(illiteracyFemale) + contraception +   
## region  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 114 82.916   
## 2 110 67.424 4 15.491 6.3183 0.0001294 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

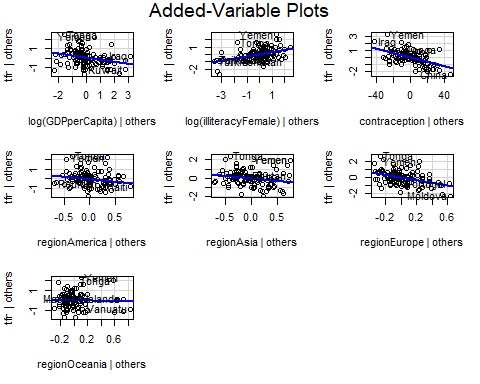
문제에서 요구한 바와 같이 region을 추가한 회귀모형을 적합하여 un\_lm3로 저장하였다. summary 확인 결과, 산점도/상관계수 행렬에서 확인한 바와 같이 지역이라는 factor 변수가 Oceania를 제외하면 통계적으로 유의한 것으로 나타났다. 결정계수는 79% 수즌으로 향상되었으며, anova 결과 또한 유의하게 나타났다.

##### Diagnosis

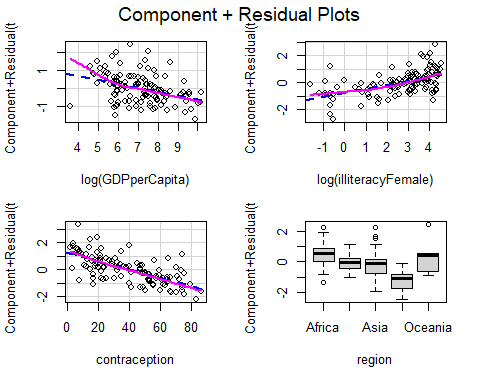
autoplot(un\_lm3, which = c(1:6))



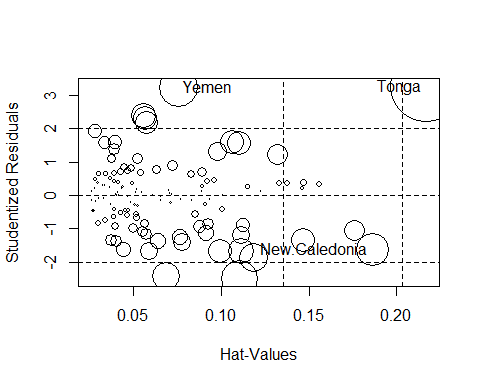
avPlots(un\_lm3, ask=FALSE)



crPlots(un\_lm3, ask=FALSE)



influencePlot(un\_lm3)



## StudRes Hat CookD  
## New.Caledonia -1.612424 0.1861082 0.07324809  
## Tonga 3.266682 0.2167754 0.33935192  
## Yemen 3.245419 0.0756837 0.09920627

모형 진단결과, 기존에는 넘어갔던 Tonga의 영향력이 region을 추가한 모형에서는 특히 Oceania 지역에 지나친 영향력을 보이는 사실상의 outlier로 모형의 설명력을 저해하고 있는 것으로 판단되었다.

# Tonga out  
un3 <- un2[row.names(un2) != "Tonga",]  
  
un\_lm4 <- lm(tfr ~ log(GDPperCapita) + log(illiteracyFemale) + contraception + region, data = un3)   
summary(un\_lm4)

##   
## Call:  
## lm(formula = tfr ~ log(GDPperCapita) + log(illiteracyFemale) +   
## contraception + region, data = un3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.77138 -0.44043 0.00588 0.40694 2.18851   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.953066 0.450033 13.228 < 2e-16 \*\*\*  
## log(GDPperCapita) -0.188978 0.065751 -2.874 0.00487 \*\*   
## log(illiteracyFemale) 0.329775 0.059593 5.534 2.18e-07 \*\*\*  
## contraception -0.034139 0.004413 -7.735 5.55e-12 \*\*\*  
## regionAmerica -0.409801 0.236723 -1.731 0.08626 .   
## regionAsia -0.455080 0.205624 -2.213 0.02897 \*   
## regionEurope -1.507121 0.335691 -4.490 1.78e-05 \*\*\*  
## regionOceania -0.536326 0.360829 -1.486 0.14007   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7506 on 109 degrees of freedom  
## Multiple R-squared: 0.8125, Adjusted R-squared: 0.8005   
## F-statistic: 67.47 on 7 and 109 DF, p-value: < 2.2e-16

이에 Tonga를 제외한 새로운 데이터셋 un3를 바탕으로 모형을 재수립하였다. 비록 기존에 유의하였던 regionAmerica 의 p-value가 증가하였으나 regionOceania는 개선되었고, 결정계수 또한 소폭 개선되었다.

##### 2.3.2 Considering the possibility that region interacts with the other

un\_lm5 <- lm(tfr ~ log(GDPperCapita) \* region + log(illiteracyFemale) \* region +   
 contraception \* region, data = un3)   
summary(un\_lm5)

##   
## Call:  
## lm(formula = tfr ~ log(GDPperCapita) \* region + log(illiteracyFemale) \*   
## region + contraception \* region, data = un3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.72581 -0.33063 0.00319 0.24322 1.91494   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.403614 1.682028 4.402 2.76e-05 \*\*\*  
## log(GDPperCapita) -0.087217 0.141297 -0.617 0.53851   
## regionAmerica -2.304996 2.860935 -0.806 0.42240   
## regionAsia -1.873552 1.780214 -1.052 0.29522   
## regionEurope -5.060743 2.434529 -2.079 0.04028 \*   
## regionOceania 2.060569 3.528968 0.584 0.56064   
## log(illiteracyFemale) -0.085470 0.332109 -0.257 0.79745   
## contraception -0.054180 0.009879 -5.484 3.30e-07 \*\*\*  
## log(GDPperCapita):regionAmerica -0.166863 0.277354 -0.602 0.54883   
## log(GDPperCapita):regionAsia -0.091390 0.166373 -0.549 0.58406   
## log(GDPperCapita):regionEurope -0.041443 0.345472 -0.120 0.90476   
## log(GDPperCapita):regionOceania -0.450172 0.325509 -1.383 0.16985   
## regionAmerica:log(illiteracyFemale) 0.509575 0.408033 1.249 0.21472   
## regionAsia:log(illiteracyFemale) 0.464549 0.339471 1.368 0.17433   
## regionEurope:log(illiteracyFemale) 0.144240 0.383469 0.376 0.70763   
## regionOceania:log(illiteracyFemale) -0.046910 0.515415 -0.091 0.92767   
## regionAmerica:contraception 0.033406 0.014983 2.230 0.02808 \*   
## regionAsia:contraception 0.014889 0.011469 1.298 0.19729   
## regionEurope:contraception 0.055350 0.020713 2.672 0.00884 \*\*   
## regionOceania:contraception 0.009775 0.047646 0.205 0.83788   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7111 on 97 degrees of freedom  
## Multiple R-squared: 0.8502, Adjusted R-squared: 0.8209   
## F-statistic: 28.99 on 19 and 97 DF, p-value: < 2.2e-16

anova(un\_lm4, un\_lm5)

## Analysis of Variance Table  
##   
## Model 1: tfr ~ log(GDPperCapita) + log(illiteracyFemale) + contraception +   
## region  
## Model 2: tfr ~ log(GDPperCapita) \* region + log(illiteracyFemale) \* region +   
## contraception \* region  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 109 61.412   
## 2 97 49.047 12 12.365 2.0378 0.02861 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

변수 간의 interaction 을 확인하기 위하여 먼저 교호작용을 고려한 회귀모형 un\_lm5를 적합하여 확인한 결과, GDP, 문맹률이 더 이상 유의하지 않게 되었으며 사실상 피임율만이 합계출산율에 영향을 주는 부적절한 모형이 만들어졌다. anova 결과 또한 유의하지 않았다.

un\_lm6 <- lm(tfr ~ log(GDPperCapita) + log(illiteracyFemale) + contraception \* region, data = un3)   
summary(un\_lm6)

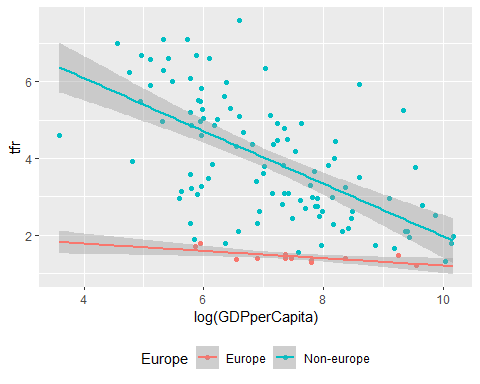
##   
## Call:  
## lm(formula = tfr ~ log(GDPperCapita) + log(illiteracyFemale) +   
## contraception \* region, data = un3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.92271 -0.34648 0.03285 0.34970 2.00771   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.180752 0.439897 14.050 < 2e-16 \*\*\*  
## log(GDPperCapita) -0.193799 0.063635 -3.045 0.002938 \*\*   
## log(illiteracyFemale) 0.332035 0.057881 5.736 9.41e-08 \*\*\*  
## contraception -0.042705 0.006802 -6.278 7.85e-09 \*\*\*  
## regionAmerica -1.098633 0.604473 -1.818 0.071992 .   
## regionAsia -0.446484 0.347572 -1.285 0.201766   
## regionEurope -3.557876 0.654648 -5.435 3.58e-07 \*\*\*  
## regionOceania -1.443123 1.084499 -1.331 0.186176   
## contraception:regionAmerica 0.017549 0.011812 1.486 0.140365   
## contraception:regionAsia 0.003813 0.008307 0.459 0.647215   
## contraception:regionEurope 0.047025 0.012833 3.664 0.000391 \*\*\*  
## contraception:regionOceania 0.032462 0.034853 0.931 0.353774   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7134 on 105 degrees of freedom  
## Multiple R-squared: 0.8368, Adjusted R-squared: 0.8197   
## F-statistic: 48.95 on 11 and 105 DF, p-value: < 2.2e-16

anova(un\_lm4, un\_lm6)

## Analysis of Variance Table  
##   
## Model 1: tfr ~ log(GDPperCapita) + log(illiteracyFemale) + contraception +   
## region  
## Model 2: tfr ~ log(GDPperCapita) + log(illiteracyFemale) + contraception \*   
## region  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 109 61.412   
## 2 105 53.444 4 7.9685 3.9139 0.005296 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# interaction?  
un3 %>% mutate(Europe = ifelse(region == "Europe", "Europe", "Non-europe")) %>%   
 ggplot(aes(x = log(GDPperCapita), y = tfr, colour = Europe)) +  
 geom\_point() + stat\_smooth(method = lm, fullrange = TRUE, se = TRUE) +  
 theme(legend.position = "bottom")

## `geom\_smooth()` using formula 'y ~ x'



vif(un\_lm6)

## GVIF Df GVIF^(1/(2\*Df))  
## log(GDPperCapita) 1.788595 1 1.337384  
## log(illiteracyFemale) 1.970992 1 1.403920  
## contraception 5.292330 1 2.300506  
## region 6179.314521 4 2.977608  
## contraception:region 11200.652920 4 3.207417

위 모형을 참고하여 피임율에만 region과의 교호작용을 고려하여 모델을 재설계한 결과, region은 Europe인 경우에 한하여 통계적으로 매우 유의하였고, contraception와 regionEurope의 교호작용 또한 매우 유의한 것으로 나타났으며, anova 결과 또한 유의하였다.

Europe 지역인 경우와 그렇지 않은 경우의 contraception 대 tfr의 관계를 산점도로 관찰한 결과, 회귀선이 완전히 평행하지는 않으나 일정 정도 interaction이 존재함을 유추할 수 있다. 공선성에는 문제가 존재하지 않았다.

# what if brought Tonga?  
summary(lm(tfr ~ log(GDPperCapita) + log(illiteracyFemale) + contraception \* region, data = un2))

##   
## Call:  
## lm(formula = tfr ~ log(GDPperCapita) + log(illiteracyFemale) +   
## contraception \* region, data = un2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.92759 -0.34654 0.04233 0.34401 2.00533   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.185691 0.439937 14.060 < 2e-16 \*\*\*  
## log(GDPperCapita) -0.196207 0.063601 -3.085 0.002597 \*\*   
## log(illiteracyFemale) 0.333818 0.057863 5.769 7.98e-08 \*\*\*  
## contraception -0.042566 0.006802 -6.258 8.42e-09 \*\*\*  
## regionAmerica -1.095554 0.604558 -1.812 0.072793 .   
## regionAsia -0.441123 0.347584 -1.269 0.207182   
## regionEurope -3.550404 0.654707 -5.423 3.72e-07 \*\*\*  
## regionOceania -2.304383 0.676556 -3.406 0.000932 \*\*\*  
## contraception:regionAmerica 0.017532 0.011814 1.484 0.140779   
## contraception:regionAsia 0.003731 0.008308 0.449 0.654277   
## contraception:regionEurope 0.046996 0.012835 3.661 0.000393 \*\*\*  
## contraception:regionOceania 0.063255 0.017213 3.675 0.000375 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7135 on 106 degrees of freedom  
## Multiple R-squared: 0.8353, Adjusted R-squared: 0.8182   
## F-statistic: 48.86 on 11 and 106 DF, p-value: < 2.2e-16

마지막으로 기존에 제외한 Tonga의 영향을 확인하기 위해 Tonga를 추가한 모형을 적합했을 때, regionOceania가 매우 유의하고 contraception과도 교호작용함을 확인하는 것으로 나타나고 있으나, 이는 unusual data인 Tonga로 인해 과대평가 된 것이라 판단된다. 그러므로 **기존의 모형인 un\_lm6이 최적의 모형이며, 지역과 피임률 간에는 일부 상호작용이 존재하는 것으로 판단된다.**

### Problem 3

*Using the Canadian occupational prestige data (the Prestige data frame in the car package – see ?Prestige), replicate the linear regression of prestige on education, income, and women (i.e., percent women in the occupation).*

*Use “regression diagnostics” to check for problems in this regression and attempt to correct any problems that you discover.*

#### Data Loading and EDA

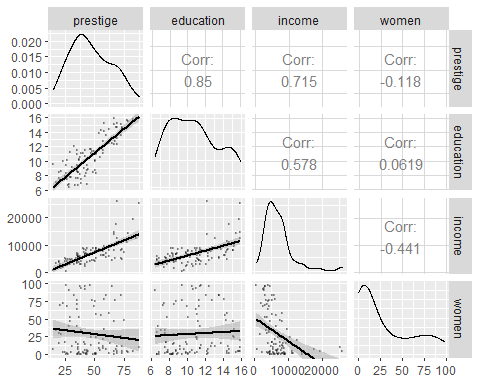
data("Prestige")  
summary(Prestige)

## education income women prestige   
## Min. : 6.380 Min. : 611 Min. : 0.000 Min. :14.80   
## 1st Qu.: 8.445 1st Qu.: 4106 1st Qu.: 3.592 1st Qu.:35.23   
## Median :10.540 Median : 5930 Median :13.600 Median :43.60   
## Mean :10.738 Mean : 6798 Mean :28.979 Mean :46.83   
## 3rd Qu.:12.648 3rd Qu.: 8187 3rd Qu.:52.203 3rd Qu.:59.27   
## Max. :15.970 Max. :25879 Max. :97.510 Max. :87.20   
## census type   
## Min. :1113 bc :44   
## 1st Qu.:3120 prof:31   
## Median :5135 wc :23   
## Mean :5402 NA's: 4   
## 3rd Qu.:8312   
## Max. :9517

str(Prestige)

## 'data.frame': 102 obs. of 6 variables:  
## $ education: num 13.1 12.3 12.8 11.4 14.6 ...  
## $ income : int 12351 25879 9271 8865 8403 11030 8258 14163 11377 11023 ...  
## $ women : num 11.16 4.02 15.7 9.11 11.68 ...  
## $ prestige : num 68.8 69.1 63.4 56.8 73.5 77.6 72.6 78.1 73.1 68.8 ...  
## $ census : int 1113 1130 1171 1175 2111 2113 2133 2141 2143 2153 ...  
## $ type : Factor w/ 3 levels "bc","prof","wc": 2 2 2 2 2 2 2 2 2 2 ...

pt <- Prestige %>% select("prestige", "education", "income", "women")   
  
ggpairs(pt, lower = list(continuous = wrap("smooth", alpha = 0.3, size = 0.1)),  
 diag = list(discrete="barDiag", continuous = wrap("densityDiag", alpha = 0.5 )),  
 upper = list(combo = wrap("box\_no\_facet", alpha = 0.5),  
 continuous = wrap("cor", size = 4,alignPercent = 0.8))) +  
 theme(legend.position = "bottom")



car 패키지의 데이터 Prestige를 로드한 후, 문제에서 주어진 대로 prestige, education, income, women 네 변수를 선택해 pt로 저장하였다.

산점도/상관계수 행렬을 통해 살펴본 결과, prestige와 education, income 사이에 강한 선형 관계가 관측되며, women은 income과 음의 상관관계가 있는 것으로 보인다. 데이터의 분포 형태 및 소득 변수 자체의 특성을 감안할 때, income은 log transformation이 필요할 것으로 추측되며, women 또한 log transformation을 시도할 필요가 있을 것으로 보인다.

#### Model Fitting

*linear regression of prestige on education, income, and women*

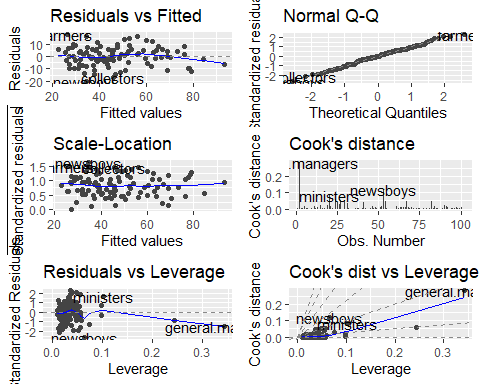
pt\_lm <- lm(prestige ~ income + education + women, data = pt)   
summary(pt\_lm)

##   
## Call:  
## lm(formula = prestige ~ income + education + women, data = pt)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19.8246 -5.3332 -0.1364 5.1587 17.5045   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6.7943342 3.2390886 -2.098 0.0385 \*   
## income 0.0013136 0.0002778 4.729 7.58e-06 \*\*\*  
## education 4.1866373 0.3887013 10.771 < 2e-16 \*\*\*  
## women -0.0089052 0.0304071 -0.293 0.7702   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.846 on 98 degrees of freedom  
## Multiple R-squared: 0.7982, Adjusted R-squared: 0.792   
## F-statistic: 129.2 on 3 and 98 DF, p-value: < 2.2e-16

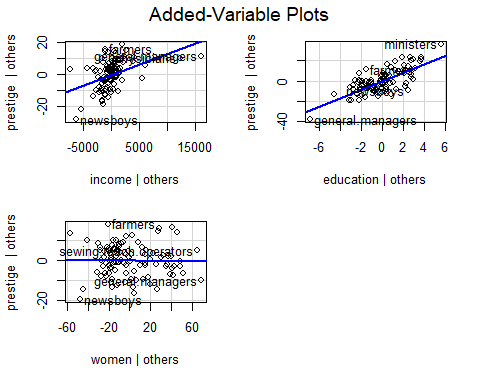
문제에서 주어진 대로, prestige를 종속변수로, income, education, women을 설명변수로 회귀모형 pt\_lm을 적합하였다. summary 확인 결과, income과 education은 유의수준 0.05에서 통계적으로 유의하며, F 통계량의 p-value는 매우 유의, 결정계수는 79.8% 수준으로 양호한 것으로 나타났다.

#### Regressioni Diagnostic

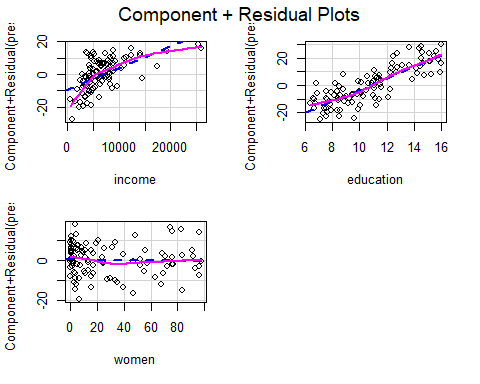
autoplot(pt\_lm, which = c(1:6))



avPlots(pt\_lm, ask=FALSE)



crPlots(pt\_lm, ask=FALSE)



회귀모형에 대한 diagnostic plot을 통해 검토한 결과, qqplot 상에서 정규성 문제는 없는 것으로 보이며, 스튜던트화잔차 plot에서 오차의 등분산성 문제도 없는 것으로 보인다.

선형성과 관련하여서는 추가적인 검토가 필요한 것으로 보여 Component Residual Plot을 통해 변수별 비선형성 이슈를 검토하였으며, 그 결과 education은 큰 이슈가 없으나, income은 monotone nonlinearity, women은 non-monotone (slight) nonlinearity 이슈가 존재하는 것으로 보인다. 이를 개선하기 위해 두 변수에 대한 변환이 필요한 것으로 판단된다.

Outlier 및 influential 관측치의 경우 general manager가 다소 눈에 띄고 있으나 모델에 영향을 주는 수치는 아닌 것으로 판단된다.

#### Correction

pt\_lm2 <- lm(prestige ~ log(income) + education + women, data = pt)   
summary(pt\_lm2)

##   
## Call:  
## lm(formula = prestige ~ log(income) + education + women, data = pt)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -17.364 -4.429 -0.101 4.316 19.179   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -110.9658 14.8429 -7.476 3.27e-11 \*\*\*  
## log(income) 13.4382 1.9138 7.022 2.90e-10 \*\*\*  
## education 3.7305 0.3544 10.527 < 2e-16 \*\*\*  
## women 0.0469 0.0299 1.568 0.12   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.093 on 98 degrees of freedom  
## Multiple R-squared: 0.8351, Adjusted R-squared: 0.83   
## F-statistic: 165.4 on 3 and 98 DF, p-value: < 2.2e-16

pt\_lm3 <- lm(prestige ~ log(income) + education + poly(women, 2, raw = TRUE), data = pt)   
summary(pt\_lm3)

##   
## Call:  
## lm(formula = prestige ~ log(income) + education + poly(women,   
## 2, raw = TRUE), data = pt)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.8191 -5.5418 0.6296 4.0416 18.5616   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.105e+02 1.454e+01 -7.602 1.86e-11 \*\*\*  
## log(income) 1.350e+01 1.874e+00 7.204 1.26e-10 \*\*\*  
## education 3.770e+00 3.475e-01 10.850 < 2e-16 \*\*\*  
## poly(women, 2, raw = TRUE)1 -1.391e-01 8.678e-02 -1.603 0.112   
## poly(women, 2, raw = TRUE)2 2.149e-03 9.438e-04 2.277 0.025 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.946 on 97 degrees of freedom  
## Multiple R-squared: 0.8435, Adjusted R-squared: 0.837   
## F-statistic: 130.7 on 4 and 97 DF, p-value: < 2.2e-16

anova(pt\_lm, pt\_lm2, pt\_lm3)

## Analysis of Variance Table  
##   
## Model 1: prestige ~ income + education + women  
## Model 2: prestige ~ log(income) + education + women  
## Model 3: prestige ~ log(income) + education + poly(women, 2, raw = TRUE)  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 98 6033.6   
## 2 98 4929.9 0 1103.68   
## 3 97 4679.8 1 250.13 5.1845 0.02499 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

step(pt\_lm3, direction = "both")

## Start: AIC=400.25  
## prestige ~ log(income) + education + poly(women, 2, raw = TRUE)  
##   
## Df Sum of Sq RSS AIC  
## <none> 4679.8 400.25  
## - poly(women, 2, raw = TRUE) 2 373.9 5053.6 404.09  
## - log(income) 1 2504.1 7183.8 441.97  
## - education 1 5679.1 10358.8 479.30

##   
## Call:  
## lm(formula = prestige ~ log(income) + education + poly(women,   
## 2, raw = TRUE), data = pt)  
##   
## Coefficients:  
## (Intercept) log(income)   
## -1.105e+02 1.350e+01   
## education poly(women, 2, raw = TRUE)1   
## 3.770e+00 -1.391e-01   
## poly(women, 2, raw = TRUE)2   
## 2.149e-03

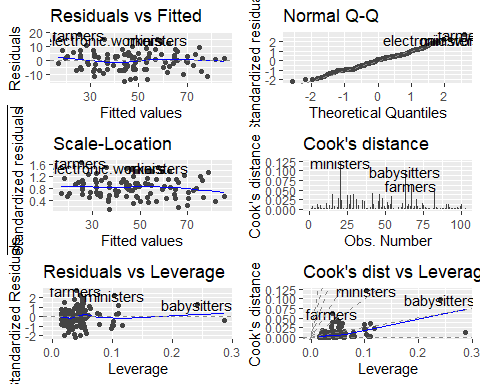
진단결과를 반영하여 모형을 다음과 같이 개선하였다.

1. 먼저 income 변수를 log transformaion 하여 새로운 모형 pt\_lm2를 적합하였다.
2. 추가로 비선형성 이슈를 감안, women 변수에 polynomial regression을 시도한 모형 pt\_lm3을 적합하였다.

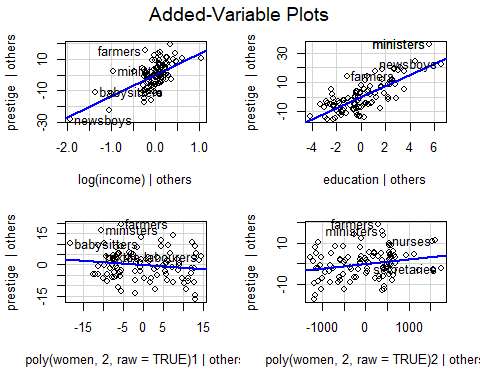
pt\_lm 대비 두 모형 모두 결정계수가 개선되었고, pt\_lm3에서는 특히 women의 2차항이 유의한 것으로 나타나면서 women 변수가 처음으로 설명력을 갖게 되었다.

anova 결과에서도 pt\_lm3가 유의함을 확인할 수 있으며, AIC 기준의 변수 선택 결과에서도 변화가 없었다. 시대적 상황을 고려할 때 women의 영향력을 제거하는 모형이 설득력을 갖기 어려운 바, **pt\_lm3를 최종 모형으로 선택**하였다.

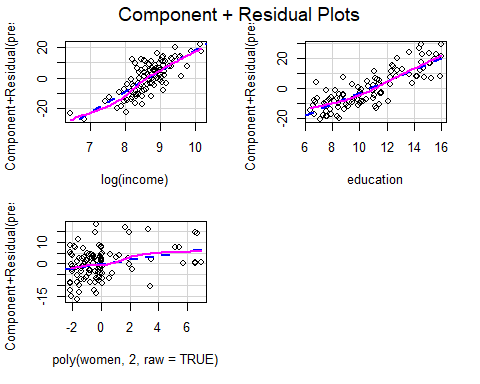
autoplot(pt\_lm3, which = c(1:6))



avPlots(pt\_lm3, ask=FALSE)



crPlots(pt\_lm3, ask=FALSE)



vif(pt\_lm3)

## GVIF Df GVIF^(1/(2\*Df))  
## log(income) 2.572891 1 1.604023  
## education 1.881801 1 1.371788  
## poly(women, 2, raw = TRUE) 1.819240 2 1.161375

Regression diagnostic plot 에서 선형성, 정규성, 등분산성에 문제가 없으며 cook’s distance 수치도 개선되고, outlier, influential 관측치 이슈도 없는 것을 판단된다. Component+residual plot에서도 income의 비선형성이 개선되었고, women의 nonmonotone 비선형성 이슈도 상대적으로 개선된 것으로 판단된다.