Faraway\_ch6

Harry Woo

2020 5 26

## Faraway Chapter 6 Diagnostics

### Problem 1

*Using the sat dataset, fit a model with the total SAT score as the response and expend, salary, ratio and takers as predictors. Perform regression diagnostics on this model to answer the following questions. Display any plots that are relevant. Do not provide any plots about which you have nothing to say. Suggest possible improvements or corrections to the model where appropriate.*

library(faraway)  
library(ggplot2)  
library(GGally)  
library(dplyr)  
library(knitr)  
library(MASS)  
library(car)  
library(tibble)  
library(gridExtra)

data("sat")  
head(sat)

## expend ratio salary takers verbal math total  
## Alabama 4.405 17.2 31.144 8 491 538 1029  
## Alaska 8.963 17.6 47.951 47 445 489 934  
## Arizona 4.778 19.3 32.175 27 448 496 944  
## Arkansas 4.459 17.1 28.934 6 482 523 1005  
## California 4.992 24.0 41.078 45 417 485 902  
## Colorado 5.443 18.4 34.571 29 462 518 980

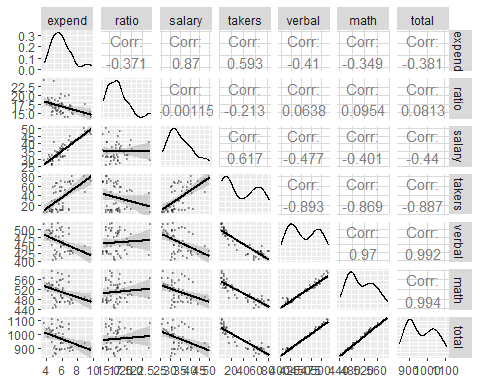
str(sat)

## 'data.frame': 50 obs. of 7 variables:  
## $ expend: num 4.41 8.96 4.78 4.46 4.99 ...  
## $ ratio : num 17.2 17.6 19.3 17.1 24 18.4 14.4 16.6 19.1 16.3 ...  
## $ salary: num 31.1 48 32.2 28.9 41.1 ...  
## $ takers: int 8 47 27 6 45 29 81 68 48 65 ...  
## $ verbal: int 491 445 448 482 417 462 431 429 420 406 ...  
## $ math : int 538 489 496 523 485 518 477 468 469 448 ...  
## $ total : int 1029 934 944 1005 902 980 908 897 889 854 ...

summary(sat)

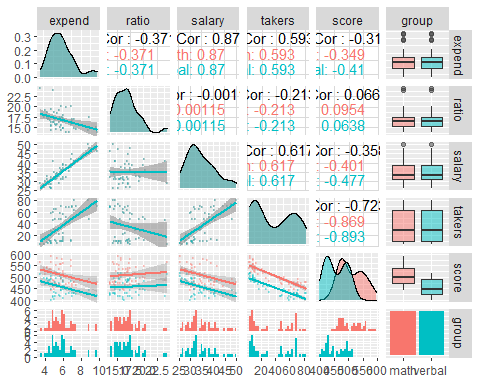
## expend ratio salary takers   
## Min. :3.656 Min. :13.80 Min. :25.99 Min. : 4.00   
## 1st Qu.:4.882 1st Qu.:15.22 1st Qu.:30.98 1st Qu.: 9.00   
## Median :5.768 Median :16.60 Median :33.29 Median :28.00   
## Mean :5.905 Mean :16.86 Mean :34.83 Mean :35.24   
## 3rd Qu.:6.434 3rd Qu.:17.57 3rd Qu.:38.55 3rd Qu.:63.00   
## Max. :9.774 Max. :24.30 Max. :50.05 Max. :81.00   
## verbal math total   
## Min. :401.0 Min. :443.0 Min. : 844.0   
## 1st Qu.:427.2 1st Qu.:474.8 1st Qu.: 897.2   
## Median :448.0 Median :497.5 Median : 945.5   
## Mean :457.1 Mean :508.8 Mean : 965.9   
## 3rd Qu.:490.2 3rd Qu.:539.5 3rd Qu.:1032.0   
## Max. :516.0 Max. :592.0 Max. :1107.0

ggpairs(sat,   
 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")



sat\_math <- sat %>%   
 dplyr::select(-c("verbal", "total")) %>%   
 rename(score = math) %>%   
 mutate(group = "math")  
  
sat\_verbal <- sat %>%   
 dplyr::select(-c("math", "total")) %>%   
 rename(score = verbal) %>%   
 mutate(group = "verbal")  
  
sat2 <- rbind(sat\_math, sat\_verbal)  
  
ggpairs(sat2, aes(colour = group),   
 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`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



#### Model Fitting

문제에서 요구한 바와 같이 total을 종속변수로, 다른 변수를 설명변수로 하여 회귀모형 sat\_lm을 적합하였다. 전체 모형의 결정계수는 0.825 수준에 p-value는 유의수준 0.05에서 매우 유의하다. 회귀계수 중 takers만 통계적으로 유의수준 0.05에서 유의한 것으로 나타난다.

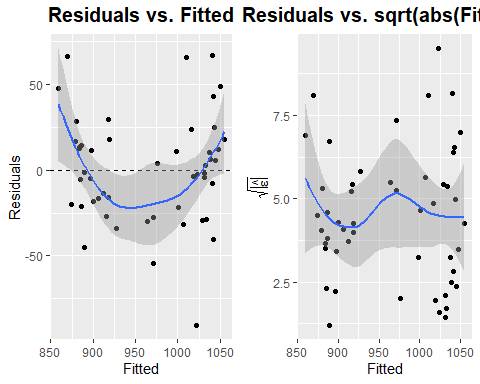
sat\_lm <- lm(total ~ expend + salary + ratio + takers, data = sat)  
summary(sat\_lm)

##   
## Call:  
## lm(formula = total ~ expend + salary + ratio + takers, data = sat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -90.531 -20.855 -1.746 15.979 66.571   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1045.9715 52.8698 19.784 < 2e-16 \*\*\*  
## expend 4.4626 10.5465 0.423 0.674   
## salary 1.6379 2.3872 0.686 0.496   
## ratio -3.6242 3.2154 -1.127 0.266   
## takers -2.9045 0.2313 -12.559 2.61e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 32.7 on 45 degrees of freedom  
## Multiple R-squared: 0.8246, Adjusted R-squared: 0.809   
## F-statistic: 52.88 on 4 and 45 DF, p-value: < 2.2e-16

#### (a) Check the constant variance assumption for the errors.

#par(mfrow = c(1, 2))  
#plot(sat\_lm, which = 1)  
#plot(sat\_lm, which = 3)  
  
diag1 <- ggplot(data = sat\_lm, aes(x = .fitted, y = .resid)) +  
 geom\_point() + geom\_hline(yintercept = 0, linetype = 2) +  
 geom\_smooth(method = "loess") +  
 xlab("Fitted") + ylab("Residuals") +   
 ggtitle(label = "Residuals vs. Fitted") +  
 theme(plot.title = element\_text(size = 15, hjust = 0.5, vjust = 1.5, face = "bold"))  
  
diag2 <- ggplot(data = sat\_lm, aes(x = .fitted, y = sqrt(abs(.resid)))) +  
 geom\_point() + geom\_smooth(method = "loess") +  
 xlab("Fitted") + ylab(expression(sqrt(abs(hat(epsilon))))) +   
 ggtitle(label = "Residuals vs. sqrt(abs(Fitted))") +  
 theme(plot.title = element\_text(size = 15, hjust = 0.5, vjust = 1.5, face = "bold"))  
  
grid.arrange(diag1, diag2, ncol = 2)

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



잔차와 추정치 사이의 관계를 산점도로 표현하여 검토한 결과, 잔차의 약한 비선형성이 관측되며, 950점을 전후로 두 그룹으로 분포가 나뉘어 등분산성에 문제가 있을 수 있는 것으로 보인다. 그러나 두 번째 plot인 과 추정치 간의 관계에서는 등분산성에 큰 문제가 없는 것으로 관측된다.

summary(lm(abs(residuals(sat\_lm)) ~ fitted(sat\_lm)))

##   
## Call:  
## lm(formula = abs(residuals(sat\_lm)) ~ fitted(sat\_lm))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -22.827 -14.323 -5.001 6.204 65.754   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 10.60384 40.82315 0.260 0.796  
## fitted(sat\_lm) 0.01386 0.04216 0.329 0.744  
##   
## Residual standard error: 20.05 on 48 degrees of freedom  
## Multiple R-squared: 0.002247, Adjusted R-squared: -0.01854   
## F-statistic: 0.1081 on 1 and 48 DF, p-value: 0.7438

summary(lm(sqrt(abs(residuals(sat\_lm))) ~ fitted(sat\_lm)))

##   
## Call:  
## lm(formula = sqrt(abs(residuals(sat\_lm))) ~ fitted(sat\_lm))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.3097 -1.2366 -0.2234 0.9929 5.0337   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 4.6484524 4.0660807 1.143 0.259  
## fitted(sat\_lm) -0.0001637 0.0041994 -0.039 0.969  
##   
## Residual standard error: 1.997 on 48 degrees of freedom  
## Multiple R-squared: 3.166e-05, Adjusted R-squared: -0.0208   
## F-statistic: 0.00152 on 1 and 48 DF, p-value: 0.9691

잔차의 절대값 및 제곱근과 추정치 간의 관계를 분석한 결과, p-value 가 모두 매우 높아 등분산성 가정을 기각할 수 없다.

var.test(residuals(sat\_lm)[fitted(sat\_lm) >= 950],   
 residuals(sat\_lm)[fitted(sat\_lm) < 950])

##   
## F test to compare two variances  
##   
## data: residuals(sat\_lm)[fitted(sat\_lm) >= 950] and residuals(sat\_lm)[fitted(sat\_lm) < 950]  
## F = 1.651, num df = 27, denom df = 21, p-value = 0.2419  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.7071032 3.6928150  
## sample estimates:  
## ratio of variances   
## 1.65098

ncvTest(sat\_lm)

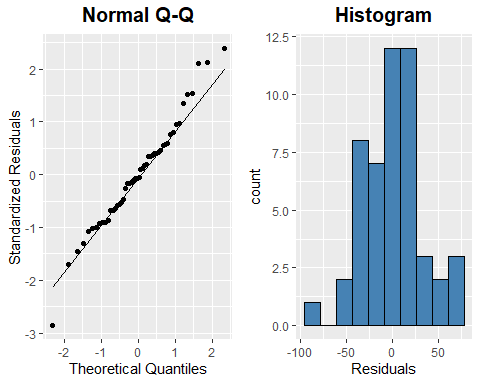
## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 0.6972119, Df = 1, p = 0.40372

첫 번째 plot에서 관측된 집단화 가능성과 관련하여, 950점을 전후로 잔차를 두 그룹으로 나누어 F test를 진행한 결과, p-value가 0.24로 등분산성 가정을 기각할 수 없다. 위의 분석에 따라 sat 데이터에는 non-constant variance 이슈가 있다고 할 수 없는 것으로 판단된다.

추가로 car 패키지의 ncvTest 함수를 통해서도 검토한 결과, p = 0.40 으로 등분산성 가정을 기각할 수 없다.

#### (b) Check the normality assumption.

#plot(sat\_lm, which = 2)  
  
diag3 <- ggplot(data = sat\_lm, aes(sample = .stdresid)) +  
 stat\_qq() + stat\_qq\_line() +  
 xlab("Theoretical Quantiles") + ylab("Standardized Residuals") +  
 ggtitle(label = "Normal Q-Q") +   
 theme(plot.title = element\_text(size = 15, hjust = 0.5, vjust = 1.5, face = "bold"))  
  
diag4 <- ggplot(data = sat\_lm, aes(x = .resid)) +  
 geom\_histogram(bins = 10, fill = "steelblue", colour = "black") +  
 xlab("Residuals") + ggtitle(label = "Histogram") +   
 theme(plot.title = element\_text(size = 15, hjust = 0.5, vjust = 1.5, face = "bold"))  
  
grid.arrange(diag3, diag4, ncol = 2)



정규성 검토를 위하여 Normal Q-Q plot을 검토한 결과, 거의 직선형태의 분포가 관측되었으며, 우상단에서 일부 직선을 벗어나는 값이 관측되나 정규성을 기각할 증거는 없는 것으로 판단된다. 잔차의 히스토그램 또한 종 형태에 가까운 모양이 관측되며 정규성 가정을 뒷받침하고 있는 것으로 보인다.

shapiro.test(sat\_lm$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: sat\_lm$residuals  
## W = 0.97691, p-value = 0.4304

Shapiro-Wilk’s Test 실시 결과 으로, 잔차의 정규성 가정(귀무가설)을 기각하지 못한다.

#### (c) Check for large leverage points.

sat\_hatv <- hatvalues(sat\_lm)  
head(sat\_hatv)

## Alabama Alaska Arizona Arkansas California Colorado   
## 0.09537668 0.18030612 0.04931612 0.05382878 0.28211791 0.03014533

sum(sat\_hatv)

## [1] 5

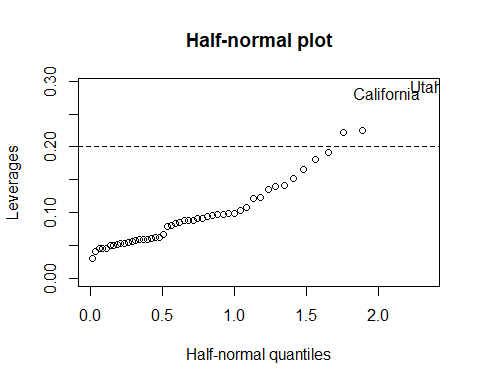
hatvalues를 통해 leverage를 sat\_hatv로 생성하였다. 일반적으로 large leverage는 인 지점을 나타내고, p = sum(sat\_hatv) 이므로, 간단히 large leverage point를 추려보면 다음과 같다.

sat\_large <- 2 \* sum(sat\_hatv) / length(sat\_lm$fitted.values)  
  
sat %>%   
 rownames\_to\_column() %>% rename(state = rowname) %>%   
 mutate(hatv = hatvalues(sat\_lm)) %>%   
 filter(hatv > sat\_large) %>%   
 arrange(desc(hatv))

## state expend ratio salary takers verbal math total hatv  
## 1 Utah 3.656 24.3 29.082 4 513 563 1076 0.2921128  
## 2 California 4.992 24.0 41.078 45 417 485 902 0.2821179  
## 3 Connecticut 8.817 14.4 50.045 81 431 477 908 0.2254519  
## 4 New Jersey 9.774 13.8 46.087 70 420 478 898 0.2220978

위의 4개 주가 large leverage points로 추정된다.

sat\_states <- row.names(sat)  
halfnorm(sat\_hatv, labs = sat\_states, ylab = "Leverages")  
abline(h = (2 \* sum(sat\_hatv) / length(sat\_lm$fitted.values)), lty = 2)  
title(main = "Half-normal plot")



Half-normal plot을 통해 확인한 결과, Utah와 California가 명확히 Large leverage로 확인되며, Connecticut과 New Jersey는 large leverage로 분류할 수 있으나 그 정도는 상대적으로 낮은 것으로 보인다.

#### (d) Check for outliers.

set.seed(123)  
sat\_stud <- rstudent(sat\_lm)  
  
sat\_stud2 <- as.data.frame(sat\_stud) %>%   
 rownames\_to\_column() %>% rename(state = rowname) %>%   
 rename(stdresid = sat\_stud)

Bonferroni correction을 통한 outlier 확인을 위해 스튜던트 잔차를 sat\_stud2 라는 데이터프레임으로 생성하였다.

sat\_stud2 %>% arrange(stdresid) %>% head(3)

## state stdresid  
## 1 West Virginia -3.124428  
## 2 Nevada -1.732004  
## 3 South Carolina -1.468832

sat\_stud2 %>% arrange(stdresid) %>% tail(3)

## state stdresid  
## 48 New Hampshire 2.190006  
## 49 North Dakota 2.213686  
## 50 Utah 2.529587

#Bonferroni critical value  
sat\_bcv <- qt(0.05 / (50\*2), 44)  
sat\_bcv

## [1] -3.525801

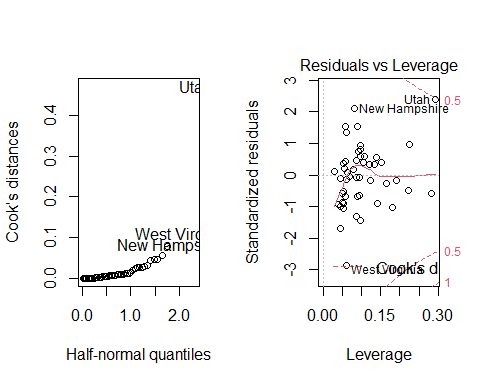
sat\_stud2 %>% filter(abs(stdresid) > abs(sat\_bcv))

## [1] state stdresid  
## <0 rows> (or 0-length row.names)

유의수준 0.05, 자유도 44 (50 - 변수 5 - 1 = 44) 기준으로 계산한 결과 에서 Bonferroni critical value = -3.53이며, 절대값을 통해 양측에서 outlier를 확인한 결과, 해당하는 outlier가 없음을 확인할 수 있다.

#### (e) Check for influential points.

sat\_cook <- cooks.distance(sat\_lm)  
  
par(mfrow = c(1, 2))  
halfnorm(sat\_cook, 3, labs = sat\_states, ylab="Cook’s distances")  
plot(sat\_lm, which = 5)



Influential point 는 잔차와 leverage가 모두 큰 관측치로, Cooks’s distance를 통해 확인할 수 있는 것처럼 Utah가 가장 큰 Influential point로 확인된다.

모델 개선을 위하여 Utah를 제외한 모형을 적합해보면 다음과 같다.

sat\_noUtah <- sat %>%   
 rownames\_to\_column() %>% rename(state = rowname) %>%   
 filter(state != "Utah")  
  
sat\_lm2 <- lm(total ~ expend + salary + ratio + takers, data = sat\_noUtah)  
  
summary(sat\_lm2)

##   
## Call:  
## lm(formula = total ~ expend + salary + ratio + takers, data = sat\_noUtah)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -92.118 -18.402 1.808 14.890 67.669   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1093.8460 53.4226 20.475 <2e-16 \*\*\*  
## expend -0.9427 10.1922 -0.092 0.927   
## salary 3.0964 2.3283 1.330 0.190   
## ratio -7.6391 3.4279 -2.229 0.031 \*   
## takers -2.9308 0.2188 -13.397 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 30.9 on 44 degrees of freedom  
## Multiple R-squared: 0.8396, Adjusted R-squared: 0.825   
## F-statistic: 57.58 on 4 and 44 DF, p-value: < 2.2e-16

(sat\_lm$coefficients - sat\_lm2$coefficients) / sat\_lm$coefficients

## (Intercept) expend salary ratio takers   
## -0.045770306 1.211253663 -0.890467644 -1.107796537 -0.009063192

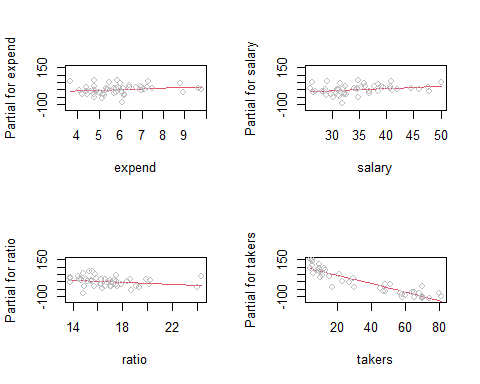
전체 모형의 결정계수는 0.84로 약간 증가하였고 p-value는 유사하다. 가장 주목할만한 점으로 회귀계수 중 takers에 이어 ratio가 추가로 유의해졌음을 확인할 수 있다. 한편, 회귀계수의 변동에서는 expend가 100% 이상 증가하고, salary는 약 90% 감소, ratio는 100% 이상 감소한 것을 확인할 수 있다.

sat[row.names(sat) == "Utah",]

## expend ratio salary takers verbal math total  
## Utah 3.656 24.3 29.082 4 513 563 1076

#### (f) Check the structure of the relationship between the predictors and the response.

par(mfrow = c(2, 2))  
termplot(sat\_lm, partial.resid = TRUE)



termplot 함수를 통해 각 설명변수별로 종속변수와의 관계에 대한 partial residual plot을 작성하였다. takers를 제외한 설명변수들의 영향력은 회귀모형의 summary에서 보인 바와 같이 유의하지 않은 것으로 보이며, 산포의 정도에 있어서도 특이할 만한 점은 보이지 않는다.

takers는 상대적으로 밀집된 분포와 강한 관계성을 보이고 있다. 또한 40을 기준으로 두개의 그룹으로 구분할 수 있는 것으로 보인다.

# takers >= 40  
  
summary(lm(total ~ expend + salary + ratio + takers,   
 data = sat, subset = (takers >= 40)))

##   
## Call:  
## lm(formula = total ~ expend + salary + ratio + takers, data = sat,   
## subset = (takers >= 40))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -38.434 -15.938 -2.509 11.444 50.289   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 801.4329 105.6773 7.584 5.2e-07 \*\*\*  
## expend 11.1444 10.8359 1.028 0.317   
## salary -0.6354 2.7190 -0.234 0.818   
## ratio 3.9147 4.8627 0.805 0.431   
## takers -0.3003 0.8869 -0.339 0.739   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 23.74 on 18 degrees of freedom  
## Multiple R-squared: 0.2627, Adjusted R-squared: 0.09882   
## F-statistic: 1.603 on 4 and 18 DF, p-value: 0.2167

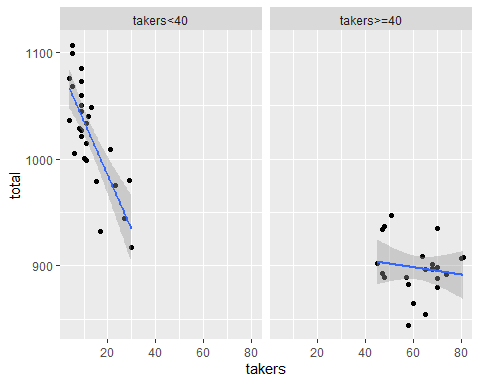
# takers < 40  
  
summary(lm(total ~ expend + salary + ratio + takers,   
 data = sat, subset = (takers < 40)))

##   
## Call:  
## lm(formula = total ~ expend + salary + ratio + takers, data = sat,   
## subset = (takers < 40))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -69.155 -16.672 1.747 15.654 54.956   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 993.7178 84.5010 11.760 5.86e-11 \*\*\*  
## expend 7.7581 16.4329 0.472 0.641   
## salary 1.0293 3.3058 0.311 0.758   
## ratio 1.4251 4.6111 0.309 0.760   
## takers -5.5242 0.8706 -6.345 2.19e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 30.95 on 22 degrees of freedom  
## Multiple R-squared: 0.656, Adjusted R-squared: 0.5935   
## F-statistic: 10.49 on 4 and 22 DF, p-value: 6.549e-05

takers == 40을 기준으로 데이터를 분리하여 회귀모형을 적합, summary를 도출한 결과, takers < 40에서 모형의 설명력, 유의성이 크게 나타나는 반면, takers >= 40에서는 모형의 설명력, 유의성이 크게 떨어지는 것으로 나타난다.

sat\_takers <- sat %>%   
 mutate(group = ifelse(takers >= 40, "takers>=40", "takers<40"))  
  
ggplot(data = sat\_takers, aes(x = takers, y = total)) +  
 geom\_point() + facet\_grid(~ group) + stat\_smooth(method = "lm")

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



takers == 40을 기준으로 데이터를 분리하여 scatter plot을 작성한 결과에서도, takers < 40 에서 takers에 따른 total의 변화가 강하게 나타남을 확인할 수 있다.

### Problem 2

*For the swiss data, fit a model with Fertility as the response and the other variables as predictors. Answer the questions posed in the first question.*

data("swiss")  
head(swiss)

## Fertility Agriculture Examination Education Catholic  
## Courtelary 80.2 17.0 15 12 9.96  
## Delemont 83.1 45.1 6 9 84.84  
## Franches-Mnt 92.5 39.7 5 5 93.40  
## Moutier 85.8 36.5 12 7 33.77  
## Neuveville 76.9 43.5 17 15 5.16  
## Porrentruy 76.1 35.3 9 7 90.57  
## Infant.Mortality  
## Courtelary 22.2  
## Delemont 22.2  
## Franches-Mnt 20.2  
## Moutier 20.3  
## Neuveville 20.6  
## Porrentruy 26.6

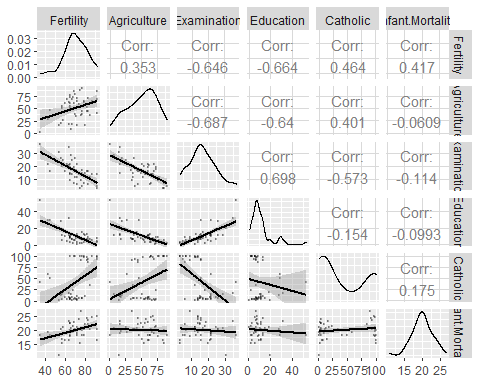
str(swiss)

## 'data.frame': 47 obs. of 6 variables:  
## $ Fertility : num 80.2 83.1 92.5 85.8 76.9 76.1 83.8 92.4 82.4 82.9 ...  
## $ Agriculture : num 17 45.1 39.7 36.5 43.5 35.3 70.2 67.8 53.3 45.2 ...  
## $ Examination : int 15 6 5 12 17 9 16 14 12 16 ...  
## $ Education : int 12 9 5 7 15 7 7 8 7 13 ...  
## $ Catholic : num 9.96 84.84 93.4 33.77 5.16 ...  
## $ Infant.Mortality: num 22.2 22.2 20.2 20.3 20.6 26.6 23.6 24.9 21 24.4 ...

summary(swiss)

## Fertility Agriculture Examination Education   
## Min. :35.00 Min. : 1.20 Min. : 3.00 Min. : 1.00   
## 1st Qu.:64.70 1st Qu.:35.90 1st Qu.:12.00 1st Qu.: 6.00   
## Median :70.40 Median :54.10 Median :16.00 Median : 8.00   
## Mean :70.14 Mean :50.66 Mean :16.49 Mean :10.98   
## 3rd Qu.:78.45 3rd Qu.:67.65 3rd Qu.:22.00 3rd Qu.:12.00   
## Max. :92.50 Max. :89.70 Max. :37.00 Max. :53.00   
## Catholic Infant.Mortality  
## Min. : 2.150 Min. :10.80   
## 1st Qu.: 5.195 1st Qu.:18.15   
## Median : 15.140 Median :20.00   
## Mean : 41.144 Mean :19.94   
## 3rd Qu.: 93.125 3rd Qu.:21.70   
## Max. :100.000 Max. :26.60

ggpairs(swiss,   
 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")



#### Model Fitting

문제에서 요구한 바와 같이 Fertility을 종속변수로, 다른 변수를 설명변수로 하여 회귀모형 sat\_lm을 적합하였다. 전체 모형의 결정계수는 0.71 수준에 p-value는 유의수준 0.05에서 매우 유의하다. 회귀계수의 경우, Examination을 제외한 모든 설명변수에서 통계적으로 유의한 것으로 나타난다.

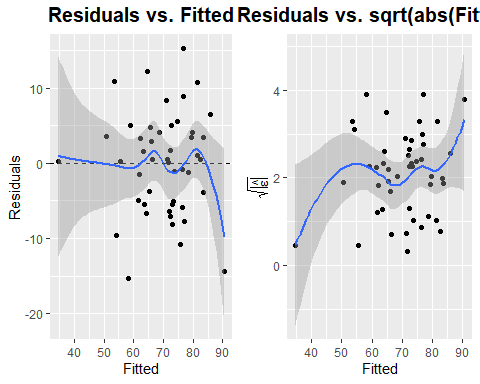
swiss\_lm <- lm(Fertility ~ Agriculture + Examination + Education +   
 Catholic + Infant.Mortality, data = swiss)  
summary(swiss\_lm)

##   
## Call:  
## lm(formula = Fertility ~ Agriculture + Examination + Education +   
## Catholic + Infant.Mortality, data = swiss)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.2743 -5.2617 0.5032 4.1198 15.3213   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 66.91518 10.70604 6.250 1.91e-07 \*\*\*  
## Agriculture -0.17211 0.07030 -2.448 0.01873 \*   
## Examination -0.25801 0.25388 -1.016 0.31546   
## Education -0.87094 0.18303 -4.758 2.43e-05 \*\*\*  
## Catholic 0.10412 0.03526 2.953 0.00519 \*\*   
## Infant.Mortality 1.07705 0.38172 2.822 0.00734 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.165 on 41 degrees of freedom  
## Multiple R-squared: 0.7067, Adjusted R-squared: 0.671   
## F-statistic: 19.76 on 5 and 41 DF, p-value: 5.594e-10

#### (a) Check the constant variance assumption for the errors.

sw\_diag1 <- ggplot(data = swiss\_lm, aes(x = .fitted, y = .resid)) +  
 geom\_point() + geom\_hline(yintercept = 0, linetype = 2) +  
 geom\_smooth(method = "loess") +  
 xlab("Fitted") + ylab("Residuals") +   
 ggtitle(label = "Residuals vs. Fitted") +  
 theme(plot.title = element\_text(size = 15, hjust = 0.5, vjust = 1.5, face = "bold"))  
  
sw\_diag2 <- ggplot(data = swiss\_lm, aes(x = .fitted, y = sqrt(abs(.resid)))) +  
 geom\_point() + geom\_smooth(method = "loess") +  
 xlab("Fitted") + ylab(expression(sqrt(abs(hat(epsilon))))) +   
 ggtitle(label = "Residuals vs. sqrt(abs(Fitted))") +  
 theme(plot.title = element\_text(size = 15, hjust = 0.5, vjust = 1.5, face = "bold"))  
  
grid.arrange(sw\_diag1, sw\_diag2, ncol = 2)

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



잔차와 추정치 사이의 관계를 산점도로 표현하여 검토한 결과 및 과 추정치 간의 관계 모두에서 잔차가 매우 고르게 분포하고 있는 것으로 나타나는 바, 등분산성 및 선형성의 가정에 큰 문제가 없는 것으로 관측된다.

summary(lm(abs(residuals(swiss\_lm)) ~ fitted(swiss\_lm)))

##   
## Call:  
## lm(formula = abs(residuals(swiss\_lm)) ~ fitted(swiss\_lm))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.2880 -3.4239 -0.3645 1.8645 10.4961   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 2.16400 4.10172 0.528 0.600  
## fitted(swiss\_lm) 0.04501 0.05785 0.778 0.441  
##   
## Residual standard error: 4.12 on 45 degrees of freedom  
## Multiple R-squared: 0.01328, Adjusted R-squared: -0.008649   
## F-statistic: 0.6055 on 1 and 45 DF, p-value: 0.4405

summary(lm(sqrt(abs(residuals(swiss\_lm))) ~ fitted(swiss\_lm)))

##   
## Call:  
## lm(formula = sqrt(abs(residuals(swiss\_lm))) ~ fitted(swiss\_lm))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.7992 -0.7539 0.1132 0.5724 1.9678   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 1.15010 0.94993 1.211 0.232  
## fitted(swiss\_lm) 0.01361 0.01340 1.016 0.315  
##   
## Residual standard error: 0.9542 on 45 degrees of freedom  
## Multiple R-squared: 0.02242, Adjusted R-squared: 0.0006947   
## F-statistic: 1.032 on 1 and 45 DF, p-value: 0.3151

잔차의 절대값 및 제곱근과 추정치 간의 관계를 분석한 결과에서도, p-value 가 높아 등분산성 가정을 기각할 수 없다.

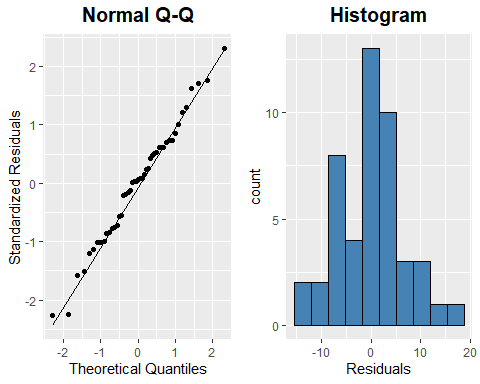
ncvTest(swiss\_lm)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 0.3077424, Df = 1, p = 0.57907

추가로 car 패키지의 ncvTest 함수를 통해서도 검토한 결과, p = 0.58 로 나타나는 바 등분산성 가정을 기각할 수 없으며, swiss 데이터에는 non-constant variance 이슈가 없는 것으로 판단된다.

#### (b) Check the normality assumption.

sw\_diag3 <- ggplot(data = swiss\_lm, aes(sample = .stdresid)) +  
 stat\_qq() + stat\_qq\_line() +  
 xlab("Theoretical Quantiles") + ylab("Standardized Residuals") +  
 ggtitle(label = "Normal Q-Q") +   
 theme(plot.title = element\_text(size = 15, hjust = 0.5, vjust = 1.5, face = "bold"))  
  
sw\_diag4 <- ggplot(data = swiss\_lm, aes(x = .resid)) +  
 geom\_histogram(bins = 10, fill = "steelblue", colour = "black") +  
 xlab("Residuals") + ggtitle(label = "Histogram") +   
 theme(plot.title = element\_text(size = 15, hjust = 0.5, vjust = 1.5, face = "bold"))  
  
grid.arrange(sw\_diag3, sw\_diag4, ncol = 2)



정규성 검토를 위하여 Normal Q-Q plot을 검토한 결과, 거의 직선형태의 분포가 관측되었으며, 직선을 벗어나는 값은 거의 관측되지 않는다. 잔차의 히스토그램 또한 종 형태에 가까운 모양이 관측되며 정규성 가정을 잘 뒷받침하고 있는 것으로 보인다.

shapiro.test(swiss\_lm$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: swiss\_lm$residuals  
## W = 0.98892, p-value = 0.9318

Shapiro-Wilk’s Test 실시 결과 으로, 잔차의 정규성 가정(귀무가설)을 기각하지 못한다.

#### (c) Check for large leverage points.

swiss\_hatv <- hatvalues(swiss\_lm)  
head(swiss\_hatv)

## Courtelary Delemont Franches-Mnt Moutier Neuveville Porrentruy   
## 0.15681744 0.12258494 0.17368296 0.07961648 0.07219003 0.19833238

sum(swiss\_hatv)

## [1] 6

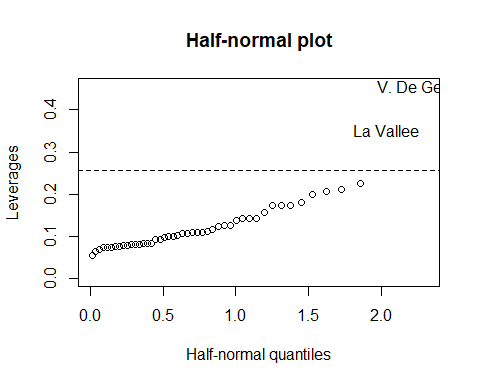
hatvalues를 통해 leverage를 swiss\_hatv로 생성하였다. 일반적으로 large leverage는 인 지점을 나타내고, p = sum(swiss\_hatv) 이므로, 간단히 large leverage point를 추려보면 다음과 같다.

swiss\_large <- 2 \* sum(swiss\_hatv) / length(swiss\_lm$fitted.values)  
  
swiss %>%   
 rownames\_to\_column() %>% rename(province = rowname) %>%   
 mutate(hatv = hatvalues(swiss\_lm)) %>%   
 filter(hatv > swiss\_large) %>%   
 arrange(desc(hatv))

## province Fertility Agriculture Examination Education Catholic  
## 1 V. De Geneve 35.0 1.2 37 53 42.34  
## 2 La Vallee 54.3 15.2 31 20 2.15  
## Infant.Mortality hatv  
## 1 18.0 0.4558363  
## 2 10.8 0.3512078

위의 2개 주가 large leverage points로 추정된다.

swiss\_provinces <- row.names(swiss)  
halfnorm(swiss\_hatv, labs = swiss\_provinces, ylab = "Leverages")  
abline(h = (2 \* sum(swiss\_hatv) / length(swiss\_lm$fitted.values)), lty = 2)  
title(main = "Half-normal plot")



Half-normal plot을 통해 확인한 결과, V. De Geneve와 La Vallee가 Large leverage로 확인된다. 다만, 앞선 sat 데이터에서의 high leverage에 비했을 때 그 정도는 상대적으로 낮은 것으로 보인다.

#### (d) Check for outliers.

set.seed(123)  
swiss\_stud <- rstudent(swiss\_lm)  
  
length(swiss\_lm$fitted.values)

## [1] 47

swiss\_stud2 <- as.data.frame(swiss\_stud) %>%   
 rownames\_to\_column() %>% rename(province = rowname) %>%   
 rename(stdresid = swiss\_stud)

Bonferroni correction을 통한 outlier 확인을 위해 스튜던트 잔차를 swiss\_stud2 라는 데이터프레임으로 생성하였다.

swiss\_stud2 %>% arrange(stdresid) %>% head(3)

## province stdresid  
## 1 Rive Gauche -2.394471  
## 2 Porrentruy -2.367218  
## 3 Moudon -1.604183

swiss\_stud2 %>% arrange(stdresid) %>% tail(3)

## province stdresid  
## 45 Neuchatel 1.745412  
## 46 Neuveville 1.816142  
## 47 Sierre 2.445227

#Bonferroni critical value  
swiss\_bcv <- qt(0.05 / (47\*2), 40)  
swiss\_bcv

## [1] -3.529468

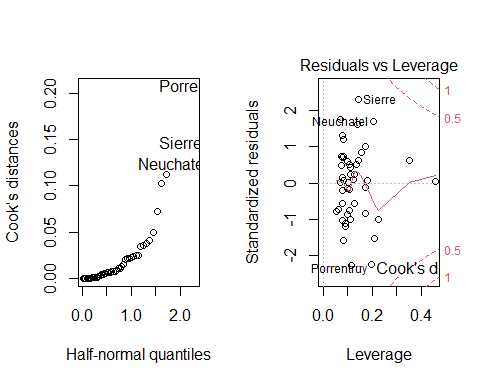
swiss\_stud2 %>% filter(abs(stdresid) > abs(swiss\_bcv))

## [1] province stdresid  
## <0 rows> (or 0-length row.names)

유의수준 0.05, 자유도 40 (47 - 변수 6 - 1 = 40) 기준으로 계산한 결과 에서 Bonferroni critical value = -3.53이며, 절대값을 통해 양측에서 outlier를 확인한 결과, 해당하는 outlier가 없음을 확인할 수 있다.

#### (e) Check for influential points.

swiss\_cook <- cooks.distance(swiss\_lm)  
  
par(mfrow = c(1, 2))  
halfnorm(swiss\_cook, 3, labs = swiss\_provinces, ylab="Cook’s distances")  
plot(swiss\_lm, which = 5)



Influential point 는 잔차와 leverage가 모두 큰 관측치로, Cook’s distance를 통해 확인할 수 있는 것처럼 Porrentruy가 가장 큰 Influential point로 확인되나, Cook’s distance가 0.5 미만이다.

모델 개선 여부 확인을 위하여 Porrentruy를 제외한 모형을 적합해보면 다음과 같다.

swiss\_noPor <- swiss %>%   
 rownames\_to\_column() %>% rename(province = rowname) %>%   
 filter(province != "Porrentruy")  
  
swiss\_lm2 <- lm(Fertility ~ Agriculture + Examination + Education +   
 Catholic + Infant.Mortality, data = swiss\_noPor)  
  
summary(swiss\_lm2)

##   
## Call:  
## lm(formula = Fertility ~ Agriculture + Examination + Education +   
## Catholic + Infant.Mortality, data = swiss\_noPor)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.7365 -5.0540 0.1953 4.1084 15.5399   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 65.45554 10.16998 6.436 1.15e-07 \*\*\*  
## Agriculture -0.21034 0.06859 -3.067 0.00387 \*\*   
## Examination -0.32278 0.24227 -1.332 0.19031   
## Education -0.89506 0.17384 -5.149 7.36e-06 \*\*\*  
## Catholic 0.11269 0.03363 3.351 0.00177 \*\*   
## Infant.Mortality 1.31567 0.37571 3.502 0.00115 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.794 on 40 degrees of freedom  
## Multiple R-squared: 0.7415, Adjusted R-squared: 0.7091   
## F-statistic: 22.94 on 5 and 40 DF, p-value: 8.583e-11

(swiss\_lm$coefficients - swiss\_lm2$coefficients) / swiss\_lm$coefficients

## (Intercept) Agriculture Examination Education   
## 0.02181330 -0.22211255 -0.25102845 -0.02769452   
## Catholic Infant.Mortality   
## -0.08231754 -0.22154724

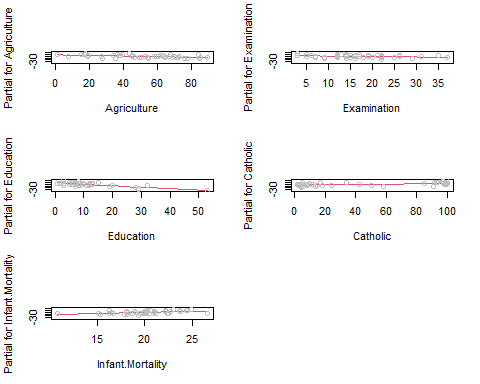
전체 모형의 결정계수는 0.74로 약간 증가하였고 p-value는 더욱 낮아졌다. 회귀계수의 유의성에는 큰 변화가 없었다. 다만, 회귀계수 값이 Agriculture, Examination, Infant.Mortality에서 20% 이상 감소한 것이 확인된다.

swiss[row.names(swiss) == "Porrentruy",]

## Fertility Agriculture Examination Education Catholic  
## Porrentruy 76.1 35.3 9 7 90.57  
## Infant.Mortality  
## Porrentruy 26.6

#### (f) Check the structure of the relationship between the predictors and the response.

par(mfrow = c(3, 2))  
termplot(swiss\_lm, partial.resid = TRUE)



termplot 함수를 통해 각 설명변수별로 종속변수와의 관계에 대한 partial residual plot을 작성하였다. Examination을 제외한 설명변수들이 모두 앞서 살펴본 바와 같이 관계에 영향력을 가지고 유의함을 확인할 수 있다.

한편, Catholic의 경우 60을 기준으로 두개의 그룹으로 구분할 수 있는 것으로 보인다.

# Catholic >= 60  
  
summary(lm(Fertility ~ Agriculture + Examination + Education +   
 Catholic + Infant.Mortality, data = swiss, subset = (Catholic >= 60)))

##   
## Call:  
## lm(formula = Fertility ~ Agriculture + Examination + Education +   
## Catholic + Infant.Mortality, data = swiss, subset = (Catholic >=   
## 60))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.2801 -4.3747 -0.4216 3.9807 14.9933   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 128.9720 109.7559 1.175 0.267  
## Agriculture -0.2002 0.2452 -0.817 0.433  
## Examination 0.6356 1.1732 0.542 0.600  
## Education -1.2951 1.3355 -0.970 0.355  
## Catholic -0.2986 1.0061 -0.297 0.773  
## Infant.Mortality -0.1929 1.4966 -0.129 0.900  
##   
## Residual standard error: 8.964 on 10 degrees of freedom  
## Multiple R-squared: 0.2091, Adjusted R-squared: -0.1863   
## F-statistic: 0.5289 on 5 and 10 DF, p-value: 0.7502

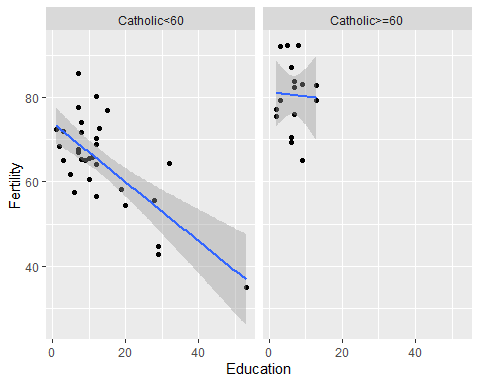
# Catholic < 60  
  
summary(lm(Fertility ~ Agriculture + Examination + Education +   
 Catholic + Infant.Mortality, data = swiss, subset = (Catholic < 60)))

##   
## Call:  
## lm(formula = Fertility ~ Agriculture + Examination + Education +   
## Catholic + Infant.Mortality, data = swiss, subset = (Catholic <   
## 60))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.0880 -3.7290 -0.7957 3.3144 15.6641   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 69.28042 11.79454 5.874 3.97e-06 \*\*\*  
## Agriculture -0.21970 0.07049 -3.117 0.00455 \*\*   
## Examination -0.49787 0.25861 -1.925 0.06566 .   
## Education -0.53880 0.19044 -2.829 0.00906 \*\*   
## Catholic -0.21234 0.10512 -2.020 0.05423 .   
## Infant.Mortality 1.27050 0.42439 2.994 0.00613 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.909 on 25 degrees of freedom  
## Multiple R-squared: 0.7531, Adjusted R-squared: 0.7037   
## F-statistic: 15.25 on 5 and 25 DF, p-value: 6.637e-07

Catholic == 60을 기준으로 데이터를 분리하여 회귀모형을 적합, summary를 도출한 결과, Catholic < 60에서 모형의 설명력, 유의성이 크게 나타나는 반면, Catholic >= 60에서는 모형의 설명력, 유의성이 크게 떨어지는 것으로 나타난다.

swiss\_cath <- swiss %>%   
 mutate(group = ifelse(Catholic >= 60, "Catholic>=60", "Catholic<60"))  
  
ggplot(data = swiss\_cath, aes(x = Education, y = Fertility)) +  
 geom\_point() + facet\_grid(~ group) + stat\_smooth(method = "lm")

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



Catholic == 60을 기준으로 데이터를 분리하여 scatter plot을 작성한 결과에서도, Catholic < 60 에서 Education에 따른 Fertility의 변화가 강하게 나타나는 등, 그룹의 분류에 따른 차이를 확인할 수 있어 그룹별 분석을 수행하는 것도 모델을 개선하는 데 도움이 될 수 있을 것으로 판단된다.