Faraway Ch14

Harry Woo

2020-5-18

이 과제물은 1차적으로 R Markdown 으로 작성되었으며, Word 문서로 knit 후 가독성을 위하여 일부 내용을 추가로 편집하였습니다.

## Faraway Chapter 14 Categorical Predictors

### Problem 1

*Using the teengamb data, model gamble as the response and the other variables as predictors, investigate the possibility of interactions between sex and the other predictors. Interpret your final model.*

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

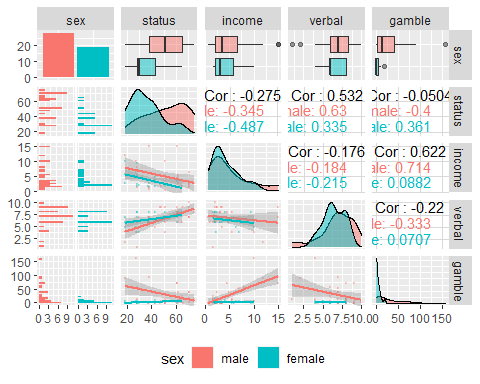
teengamb 데이터에서 실제 factor 변수인 sex를 factor로 변환하고, ggpairs를 통해 성별 기준으로 데이터의 분포형태를 관찰하였다. income과 gamble은 반드시 0 이상의 값을 가지게 된다는 점과 치우침의 형태를 감안하여 추후 로그변환을 시도할 필요가 있는 것으로 보이며, 셩별에 따른 status, gamble의 차이가 두드러지게 나타난다.

data(teengamb)  
head(teengamb)

## sex status income verbal gamble  
## 1 1 51 2.00 8 0.0  
## 2 1 28 2.50 8 0.0  
## 3 1 37 2.00 6 0.0  
## 4 1 28 7.00 4 7.3  
## 5 1 65 2.00 8 19.6  
## 6 1 61 3.47 6 0.1

tg <- teengamb %>%   
 mutate(sex = factor(sex, labels = c("male", "female")))  
  
ggpairs(tg, legend = 1,   
 mapping = aes(colour = sex),  
 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`.

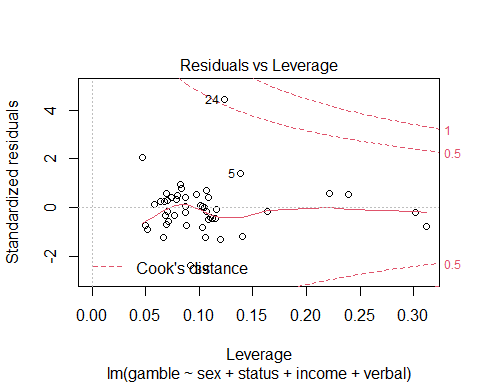
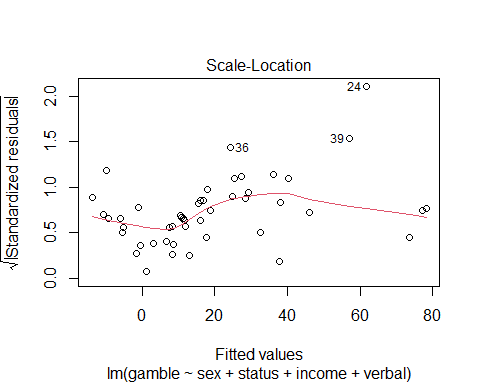
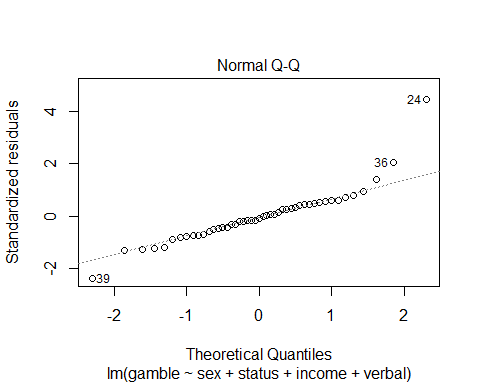
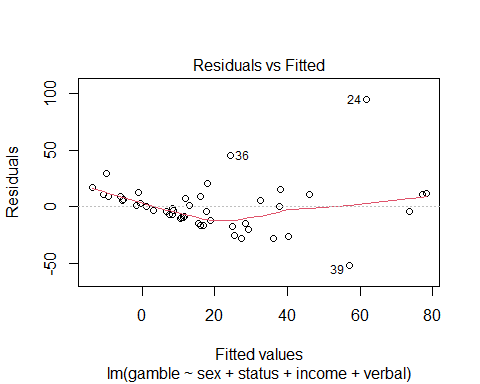


문제에서 요구한 바와 같이 gamble를 종속변수로, 다른 변수를 설명변수로 하여 회귀모형 tg\_lm을 적합하였다.성별과 소득만 통계적으로 유의하며, 결정계수는 0.53 수준에 p-value는 유의수준 0.05에서 매우 유의하다. 그러나 모형진단 결과 및 shapiro test 결과 잔차의 정규성 및 등분산성 가정을 충족하지 못하는 문제가 존재하여, gamble 및 income의 로그변환을 시도하였다.

tg\_lm <- lm(gamble ~ sex + status + income + verbal, data = tg)  
summary(tg\_lm)

##   
## Call:  
## lm(formula = gamble ~ sex + status + income + verbal, data = tg)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -51.082 -11.320 -1.451 9.452 94.252   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 22.55565 17.19680 1.312 0.1968   
## sexfemale -22.11833 8.21111 -2.694 0.0101 \*   
## status 0.05223 0.28111 0.186 0.8535   
## income 4.96198 1.02539 4.839 1.79e-05 \*\*\*  
## verbal -2.95949 2.17215 -1.362 0.1803   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 22.69 on 42 degrees of freedom  
## Multiple R-squared: 0.5267, Adjusted R-squared: 0.4816   
## F-statistic: 11.69 on 4 and 42 DF, p-value: 1.815e-06

plot(tg\_lm)



shapiro.test(tg\_lm$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: tg\_lm$residuals  
## W = 0.86839, p-value = 8.16e-05

ncvTest(tg\_lm)

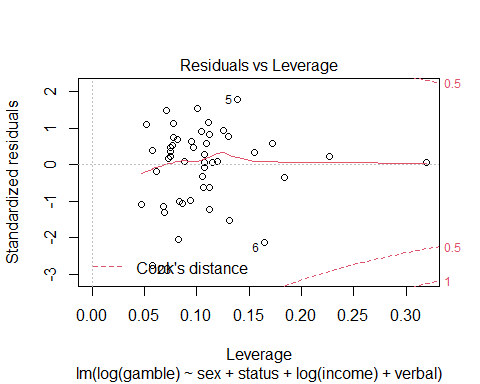
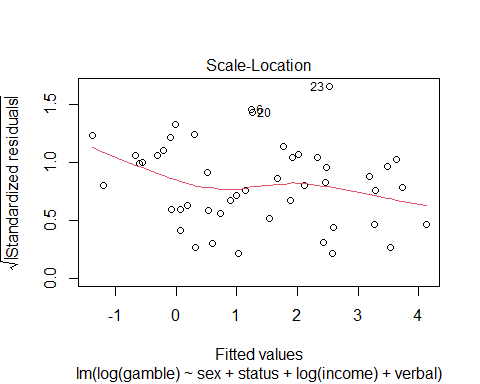
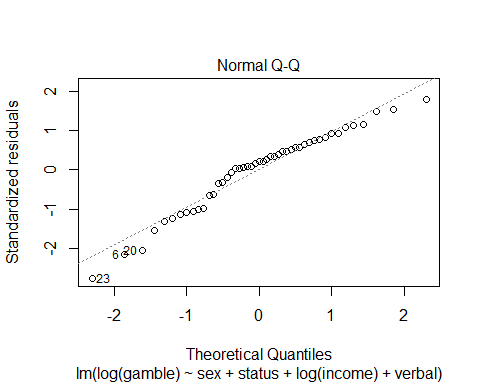
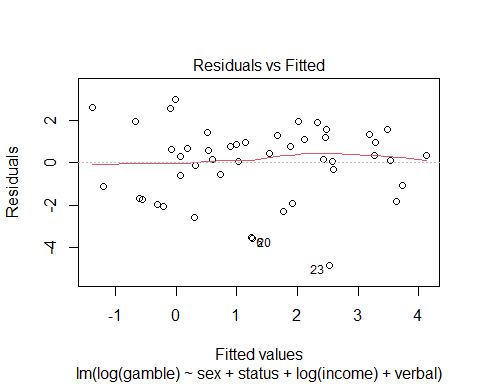
## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 24.29051, Df = 1, p = 8.2846e-07

로그변환을 위하여 0값이 gamble 및 income에 0.1을 더해주고 해당 데이터를 tg2로 저장하였다. tg2에 대하여 gamble 및 income에 로그를 취해 회귀모형을 적합한 결과, 결정계수는 다소 줄어들었으나 잔차의 정규성, 등분산성 문제가 크게 개선된 회귀모형이 적합되었다.

tg2 <- tg %>%   
 mutate(gamble = ifelse(gamble == 0, gamble + 0.1, gamble),  
 income = ifelse(income == 0, income + 0.1, income))  
  
tg2\_lm2 <- lm(log(gamble) ~ sex + status + log(income) + verbal, data = tg2)  
summary(tg2\_lm2)

##   
## Call:  
## lm(formula = log(gamble) ~ sex + status + log(income) + verbal,   
## data = tg2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.8341 -1.0776 0.3515 1.1479 2.9882   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.41888 1.37844 1.029 0.30922   
## sexfemale -1.58871 0.63533 -2.501 0.01639 \*   
## status 0.03912 0.02189 1.787 0.08118 .   
## log(income) 1.12441 0.38220 2.942 0.00529 \*\*  
## verbal -0.39565 0.17306 -2.286 0.02735 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.808 on 42 degrees of freedom  
## Multiple R-squared: 0.412, Adjusted R-squared: 0.356   
## F-statistic: 7.357 on 4 and 42 DF, p-value: 0.0001386

plot(tg2\_lm2)



shapiro.test(tg2\_lm2$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: tg2\_lm2$residuals  
## W = 0.95287, p-value = 0.05615

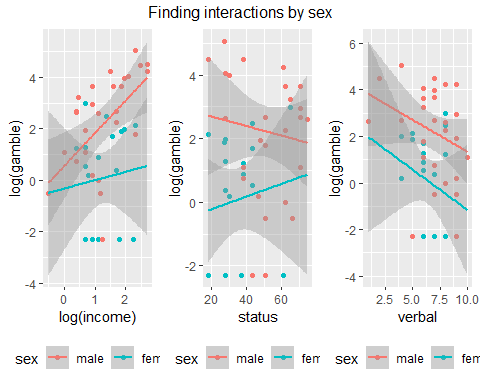
ncvTest(tg2\_lm2)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 0.3507095, Df = 1, p = 0.55371

문제에서 요구한 바와 같이 sex와 다른 설명변수 간의 교호작용을 검토해보았다. 아래의 plot에서 확인되는 바와 같이, sex에 따라 적합되는 회귀선의 기울기가 크게 변화하고 있음을 고려했을 때, sex에 따라 단순히 절편의 차이만 있는 것으로 보이는 verbal을 제외하고, income 및 status에 대해서는 sex와의 교호작용을 고려해 볼 필요가 있는 것으로 보인다.

inter1 <- ggplot(data = tg2, aes(x = log(income), y = log(gamble), colour = sex)) +  
 geom\_point() + stat\_smooth(method = lm, fullrange = TRUE, se = TRUE) +  
 theme(legend.position = "bottom")  
inter2 <- ggplot(data = tg2, aes(x = status, y = log(gamble), colour = sex)) +  
 geom\_point() + stat\_smooth(method = lm, fullrange = TRUE, se = TRUE) +  
 theme(legend.position = "bottom")  
inter3 <- ggplot(data = tg2, aes(x = verbal, y = log(gamble), colour = sex)) +  
 geom\_point() + stat\_smooth(method = lm, fullrange = TRUE, se = TRUE) +  
 theme(legend.position = "bottom")  
grid.arrange(inter1, inter2, inter3, ncol = 3, top = "Finding interactions by sex")

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



verbal을 제외한 교호작용을 추가하여 회귀모형을 tg2\_lm3로 적합한 결과, 예상한 바와 달리 income과 verbal만이 유의한 것으로 관측되며, 탐험적 분석에서 관측되었던 sex의 영향력이 오히려 유의성을 잃는 결과가 나왔다.

tg2\_lm3 <- lm(log(gamble) ~ sex + status + log(income) + verbal +   
 sex:status + sex:log(income), data = tg2)  
summary(tg2\_lm3)

##   
## Call:  
## lm(formula = log(gamble) ~ sex + status + log(income) + verbal +   
## sex:status + sex:log(income), data = tg2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.8571 -1.1741 0.3359 1.2800 2.5774   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.43158 1.50887 0.949 0.34843   
## sexfemale -1.25024 2.56908 -0.487 0.62916   
## status 0.03218 0.02666 1.207 0.23444   
## log(income) 1.29834 0.43897 2.958 0.00518 \*\*  
## verbal -0.37749 0.17989 -2.098 0.04223 \*   
## sexfemale:status 0.01252 0.04370 0.286 0.77603   
## sexfemale:log(income) -0.70399 0.94205 -0.747 0.45926   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.83 on 40 degrees of freedom  
## Multiple R-squared: 0.4262, Adjusted R-squared: 0.3401   
## F-statistic: 4.952 on 6 and 40 DF, p-value: 0.0007137

모형의 개선을 위하여 stepAIC를 통해 변수선택을 진행한 결과, 교호작용을 고려한 변수들이 제거되고 오히려 기존의 모델인 tg2\_lm2이 최종 모형으로 채택되었다.

tg2\_lm4 <- stepAIC(tg2\_lm3, direction = "both")

## Start: AIC=63.24  
## log(gamble) ~ sex + status + log(income) + verbal + sex:status +   
## sex:log(income)  
##   
## Df Sum of Sq RSS AIC  
## - sex:status 1 0.2748 134.28 61.340  
## - sex:log(income) 1 1.8709 135.88 61.895  
## <none> 134.01 63.244  
## - verbal 1 14.7518 148.76 66.152  
##   
## Step: AIC=61.34  
## log(gamble) ~ sex + status + log(income) + verbal + sex:log(income)  
##   
## Df Sum of Sq RSS AIC  
## - sex:log(income) 1 3.0441 137.32 60.393  
## <none> 134.28 61.340  
## - status 1 8.8574 143.14 62.342  
## + sex:status 1 0.2748 134.01 63.244  
## - verbal 1 16.4978 150.78 64.786  
##   
## Step: AIC=60.39  
## log(gamble) ~ sex + status + log(income) + verbal  
##   
## Df Sum of Sq RSS AIC  
## <none> 137.32 60.393  
## + sex:log(income) 1 3.0441 134.28 61.340  
## - status 1 10.4398 147.76 61.837  
## + sex:status 1 1.4480 135.88 61.895  
## - verbal 1 17.0896 154.41 63.906  
## - sex 1 20.4449 157.77 64.916  
## - log(income) 1 28.2980 165.62 67.199

tg2\_lm4$anova

## Stepwise Model Path   
## Analysis of Deviance Table  
##   
## Initial Model:  
## log(gamble) ~ sex + status + log(income) + verbal + sex:status +   
## sex:log(income)  
##   
## Final Model:  
## log(gamble) ~ sex + status + log(income) + verbal  
##   
##   
## Step Df Deviance Resid. Df Resid. Dev AIC  
## 1 40 134.0057 63.24352  
## 2 - sex:status 1 0.2748349 41 134.2805 61.33982  
## 3 - sex:log(income) 1 3.0440629 42 137.3246 60.39338

Log를 취하지 않은 모형에 비해 모형의 설명력이 다소 저하되었고 설명에 다소 어려움이 존재하는 측면이 있으나, 음수가 아닌 gamble, income의 수치를 기반으로 하는 정상적이고 자연스러운 모델일 뿐만 아니라, 잔차의 등분산성 및 정규성 문제를 크게 개선한다는 측면에서 tg2\_lm2를 최선의 모형으로 선정하였다.

해당 모형에 따르면 여성의 gamble 액수가 남성에 비해 20% 수준에 불과하다. 또한 verbal이 1 증가할 수록 gamble은 약 33% 감소하고, status가 1 증가하면 gamble은 약 4% 증가한다. log(income)이 1 증가하면 log(gamble)은 약 1.12 증가하며, 이는 gamble이 income의 약 1.12 제곱 수준으로 지수적 형태의 증가를 보임을 보여준다.

summary(tg2\_lm2)

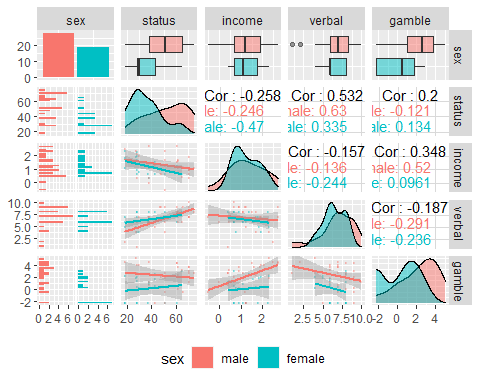
##   
## Call:  
## lm(formula = log(gamble) ~ sex + status + log(income) + verbal,   
## data = tg2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.8341 -1.0776 0.3515 1.1479 2.9882   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.41888 1.37844 1.029 0.30922   
## sexfemale -1.58871 0.63533 -2.501 0.01639 \*   
## status 0.03912 0.02189 1.787 0.08118 .   
## log(income) 1.12441 0.38220 2.942 0.00529 \*\*  
## verbal -0.39565 0.17306 -2.286 0.02735 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.808 on 42 degrees of freedom  
## Multiple R-squared: 0.412, Adjusted R-squared: 0.356   
## F-statistic: 7.357 on 4 and 42 DF, p-value: 0.0001386

exp(tg2\_lm2$coefficients)

## (Intercept) sexfemale status log(income) verbal   
## 4.1324693 0.2041891 1.0398982 3.0783954 0.6732416

tg2 %>%   
 mutate(gamble = log(gamble), income = log(income)) %>%   
 ggpairs(., legend = 1,   
 mapping = aes(colour = sex),  
 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`.



### Problem 2

*Plot the ToothGrowth data with len as the response. Fit a linear model to the data and check for possible simplification. Display the fitted regression lines on top of the data.*

data(ToothGrowth)  
head(ToothGrowth)

## len supp dose  
## 1 4.2 VC 0.5  
## 2 11.5 VC 0.5  
## 3 7.3 VC 0.5  
## 4 5.8 VC 0.5  
## 5 6.4 VC 0.5  
## 6 10.0 VC 0.5

str(ToothGrowth)

## 'data.frame': 60 obs. of 3 variables:  
## $ len : num 4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...  
## $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...  
## $ dose: num 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...

table(ToothGrowth$supp)

##   
## OJ VC   
## 30 30

table(ToothGrowth$dose)

##   
## 0.5 1 2   
## 20 20 20

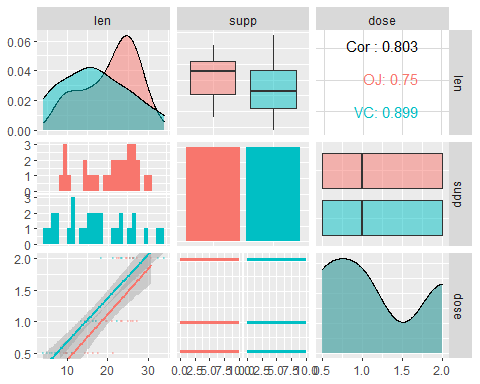
len를 종속변수로, 다른 변수를 설명변수로 하여 회귀모형 tg\_lm을 적합하였다.

pig\_lm <- lm(len ~ supp + dose, data = ToothGrowth)  
summary(pig\_lm)

##   
## Call:  
## lm(formula = len ~ supp + dose, data = ToothGrowth)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.600 -3.700 0.373 2.116 8.800   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.2725 1.2824 7.231 1.31e-09 \*\*\*  
## suppVC -3.7000 1.0936 -3.383 0.0013 \*\*   
## dose 9.7636 0.8768 11.135 6.31e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.236 on 57 degrees of freedom  
## Multiple R-squared: 0.7038, Adjusted R-squared: 0.6934   
## F-statistic: 67.72 on 2 and 57 DF, p-value: 8.716e-16

ggpairs(ToothGrowth, mapping = aes(colour = supp),  
 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")

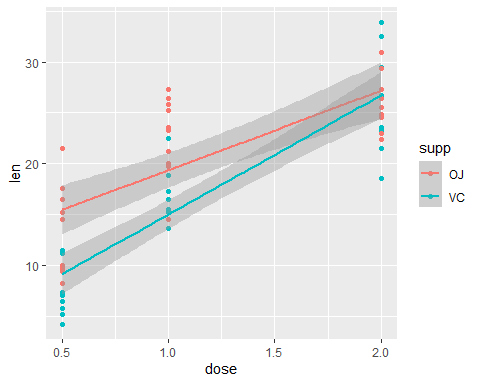
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



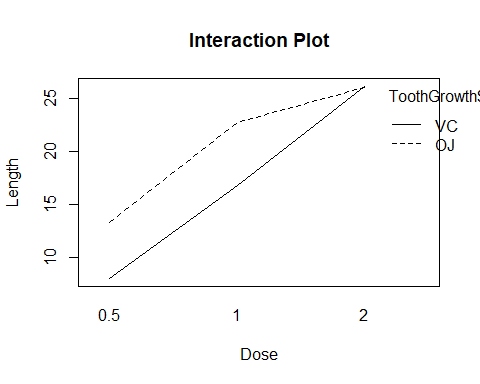
산점도에 ggplot2의 stat\_smooth를 사용하여 회귀직선을 적합한 결과 및 Interaction plot을 통한 검토 결과 supp에 따른 기울기 차이가 존재하여 두 설명변수 사이의 교호작용을 검토할 필요성이 있음을 확인할 수 있다.

ggplot(data = ToothGrowth, aes(x = dose, y = len, colour = supp))+  
 geom\_point() + stat\_smooth(method = lm, se = TRUE)

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



interaction.plot(ToothGrowth$dose, ToothGrowth$supp, ToothGrowth$len,  
 xlab = "Dose", ylab = "Length", main = "Interaction Plot")



교호작용을 고려한 회귀모형을 pig\_lm2로 적합하였다. simplication 가능성을 검토하기 위하여 stepAIC를 통해 변수선택을 실시한 결과, 원래 식이 final model로 제시되었다. 결정계수 및 p-value 모두 pig\_lm2에서 개선되었음을 확인할 수 있다. 모든 설명변수가 유의한 것으로 나타났는데, OJ보다 VC를 급여할 때 len이 8.3 가량 짧고, OJ 급여시 dose가 1 증가함에 따라 len도 7.8가량 늘어나며, VC 급여시 이 증가분이 3.9가량 더 늘어나는 것을 확인할 수 있다.

pig\_lm2 <- lm(len ~ supp\*dose, data = ToothGrowth)  
summary(pig\_lm2)

##   
## Call:  
## lm(formula = len ~ supp \* dose, data = ToothGrowth)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.2264 -2.8462 0.0504 2.2893 7.9386   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11.550 1.581 7.304 1.09e-09 \*\*\*  
## suppVC -8.255 2.236 -3.691 0.000507 \*\*\*  
## dose 7.811 1.195 6.534 2.03e-08 \*\*\*  
## suppVC:dose 3.904 1.691 2.309 0.024631 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.083 on 56 degrees of freedom  
## Multiple R-squared: 0.7296, Adjusted R-squared: 0.7151   
## F-statistic: 50.36 on 3 and 56 DF, p-value: 6.521e-16

pig\_lm2\_step <- stepAIC(pig\_lm2, direction = "back")

## Start: AIC=172.68  
## len ~ supp \* dose  
##   
## Df Sum of Sq RSS AIC  
## <none> 933.63 172.68  
## - supp:dose 1 88.92 1022.56 176.14

pig\_lm2\_step$anova

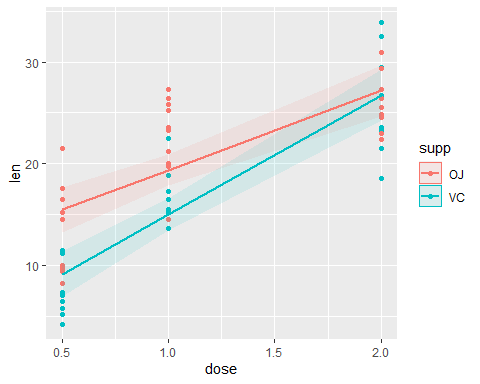
## Stepwise Model Path   
## Analysis of Deviance Table  
##   
## Initial Model:  
## len ~ supp \* dose  
##   
## Final Model:  
## len ~ supp \* dose  
##   
##   
## Step Df Deviance Resid. Df Resid. Dev AIC  
## 1 56 933.6349 172.6845

anova(pig\_lm, pig\_lm2\_step)

## Analysis of Variance Table  
##   
## Model 1: len ~ supp + dose  
## Model 2: len ~ supp \* dose  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 57 1022.56   
## 2 56 933.63 1 88.92 5.3335 0.02463 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

predict 함수를 활용하여 ToothGrowth 데이터에 pig\_lm2 모형을 통한 len 예측값 및 신뢰구간을 추가하여 ToothGrowth2 로 저장하였다. ToothGrowth3 데이터를 활용하여 산점도 위에 pig\_lm2 모형으로 적합한 회귀직선을 아래와 같이 표시하였다.

ToothGrowth2 <- ToothGrowth %>%   
 mutate(len\_pred = predict(pig\_lm2, data = ToothGrowth))  
  
pred\_conf <- predict(pig\_lm2, data = ToothGrowth, interval = "confidence")  
  
ToothGrowth2 <- cbind(ToothGrowth2, pred\_conf)  
  
ggplot(data = ToothGrowth2, aes(x = dose, y = len, colour = supp))+  
 geom\_point() + geom\_ribbon(aes(ymin = lwr, ymax = upr,   
 fill = supp, colour=NULL), alpha = 0.1) +  
 geom\_line(aes(x = dose, y = len\_pred, colour = supp), size = 1)



#### 추가 분석 실시

문제에서 주어진 내용 이외에 추가로 분석을 실시해보았다. 먼저 shapiro.test 결과 len은 정규성을 충족하고, var.test 결과 등분산 가정을 충족하는 바, two sample t-test 실시가 가능하다. t-test 실시 결과 p-value가 0.06으로, 미세한 차이로 유의수준 0.05에서 귀무가설을 기각할 수 없다. 그러므로 supp 종류(OJ or VC)가 len에 미치는 영향이 통계적으로 매우 유의하다고 할 수 없다.

shapiro.test(ToothGrowth$len)

##   
## Shapiro-Wilk normality test  
##   
## data: ToothGrowth$len  
## W = 0.96743, p-value = 0.1091

var.test(len ~ supp, data = ToothGrowth)

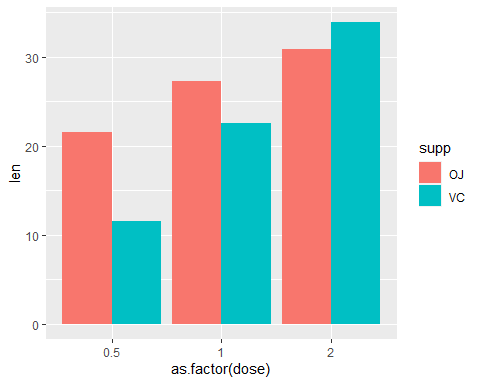
##   
## F test to compare two variances  
##   
## data: len by supp  
## F = 0.6386, num df = 29, denom df = 29, p-value = 0.2331  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.3039488 1.3416857  
## sample estimates:  
## ratio of variances   
## 0.6385951

t.test(len ~ supp, data = ToothGrowth, var.equal = TRUE)

##   
## Two Sample t-test  
##   
## data: len by supp  
## t = 1.9153, df = 58, p-value = 0.06039  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.1670064 7.5670064  
## sample estimates:  
## mean in group OJ mean in group VC   
## 20.66333 16.96333

barplot 상에서는 dose가 2일 때와 아닐 때 supp의 영향이 달라는 것이 보여, 이를 분석하고자 더미변수를 조정하여 회귀모형을 적합, 변수선택을 진행하였으며, pig\_lm4 형태의 회귀모형이 최종 선택되었다. 기존 모델보다 결정계수, p-value는 개선되었으나 산점도 그래프 상에 회귀직선을 표기하는 데에는 한계가 있었다.

ggplot(data = ToothGrowth, aes(x = as.factor(dose), y = len, fill = supp))+  
 geom\_bar(stat = "identity", position = "dodge")



ToothGrowth2 <- ToothGrowth %>%   
 mutate(dose1 = ifelse(dose == 1, 1, 0),  
 dose2 = ifelse(dose == 2, 1, 0))  
  
pig\_lm3 <- lm(len ~ supp + dose1 + dose2 + supp:dose1 + supp:dose2, data = ToothGrowth2)  
summary(pig\_lm3)

##   
## Call:  
## lm(formula = len ~ supp + dose1 + dose2 + supp:dose1 + supp:dose2,   
## data = ToothGrowth2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.20 -2.72 -0.27 2.65 8.27   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.230 1.148 11.521 3.60e-16 \*\*\*  
## suppVC -5.250 1.624 -3.233 0.00209 \*\*   
## dose1 9.470 1.624 5.831 3.18e-07 \*\*\*  
## dose2 12.830 1.624 7.900 1.43e-10 \*\*\*  
## suppVC:dose1 -0.680 2.297 -0.296 0.76831   
## suppVC:dose2 5.330 2.297 2.321 0.02411 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.631 on 54 degrees of freedom  
## Multiple R-squared: 0.7937, Adjusted R-squared: 0.7746   
## F-statistic: 41.56 on 5 and 54 DF, p-value: < 2.2e-16

pig\_lm3\_step <- stepAIC(pig\_lm3, direction = "both")

## Start: AIC=160.43  
## len ~ supp + dose1 + dose2 + supp:dose1 + supp:dose2  
##   
## Df Sum of Sq RSS AIC  
## - supp:dose1 1 1.156 713.26 158.53  
## <none> 712.11 160.43  
## - supp:dose2 1 71.022 783.13 164.14  
##   
## Step: AIC=158.53  
## len ~ supp + dose1 + dose2 + supp:dose2  
##   
## Df Sum of Sq RSS AIC  
## <none> 713.26 158.53  
## + supp:dose1 1 1.16 712.11 160.43  
## - supp:dose2 1 107.16 820.43 164.93  
## - dose1 1 833.57 1546.83 202.98

pig\_lm3\_step$anova

## Stepwise Model Path   
## Analysis of Deviance Table  
##   
## Initial Model:  
## len ~ supp + dose1 + dose2 + supp:dose1 + supp:dose2  
##   
## Final Model:  
## len ~ supp + dose1 + dose2 + supp:dose2  
##   
##   
## Step Df Deviance Resid. Df Resid. Dev AIC  
## 1 54 712.106 160.4329  
## 2 - supp:dose1 1 1.156 55 713.262 158.5303

pig\_lm4 <- lm(len ~ supp + dose1 + dose2 + supp:dose2, data = ToothGrowth2)  
summary(pig\_lm4)

##   
## Call:  
## lm(formula = len ~ supp + dose1 + dose2 + supp:dose2, data = ToothGrowth2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.030 -2.558 -0.440 2.777 8.100   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.4000 0.9862 13.587 < 2e-16 \*\*\*  
## suppVC -5.5900 1.1388 -4.909 8.56e-06 \*\*\*  
## dose1 9.1300 1.1388 8.017 8.17e-11 \*\*\*  
## dose2 12.6600 1.5065 8.404 1.93e-11 \*\*\*  
## suppVC:dose2 5.6700 1.9724 2.875 0.00574 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.601 on 55 degrees of freedom  
## Multiple R-squared: 0.7934, Adjusted R-squared: 0.7784   
## F-statistic: 52.8 on 4 and 55 DF, p-value: < 2.2e-16

### Problem 3

*Find a good model for the uswages data with wages as the response and all the other variables as predictors.*

<http://rstudio-pubs-static.s3.amazonaws.com/316501_10093f3aaff0490ca23c8bd3404185fc.html>

uswages 데이터를 로드하여 기술통계량을 확인한 결과, exper, 즉 경력이 음수로 들어간 이상데이터가 확인되어 처리가 필요한 것으로 판단된다.

data(uswages)  
head(uswages)

## wage educ exper race smsa ne mw so we pt  
## 6085 771.60 18 18 0 1 1 0 0 0 0  
## 23701 617.28 15 20 0 1 0 0 0 1 0  
## 16208 957.83 16 9 0 1 0 0 1 0 0  
## 2720 617.28 12 24 0 1 1 0 0 0 0  
## 9723 902.18 14 12 0 1 0 1 0 0 0  
## 22239 299.15 12 33 0 1 0 0 0 1 0

str(uswages)

## 'data.frame': 2000 obs. of 10 variables:  
## $ wage : num 772 617 958 617 902 ...  
## $ educ : int 18 15 16 12 14 12 16 16 12 12 ...  
## $ exper: int 18 20 9 24 12 33 42 0 36 37 ...  
## $ race : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ smsa : int 1 1 1 1 1 1 1 1 1 0 ...  
## $ ne : int 1 0 0 1 0 0 0 0 0 0 ...  
## $ mw : int 0 0 0 0 1 0 0 1 0 1 ...  
## $ so : int 0 0 1 0 0 0 1 0 0 0 ...  
## $ we : int 0 1 0 0 0 1 0 0 1 0 ...  
## $ pt : int 0 0 0 0 0 0 1 1 1 0 ...

summary(uswages)

## wage educ exper race   
## Min. : 50.39 Min. : 0.00 Min. :-2.00 Min. :0.000   
## 1st Qu.: 308.64 1st Qu.:12.00 1st Qu.: 8.00 1st Qu.:0.000   
## Median : 522.32 Median :12.00 Median :15.00 Median :0.000   
## Mean : 608.12 Mean :13.11 Mean :18.41 Mean :0.078   
## 3rd Qu.: 783.48 3rd Qu.:16.00 3rd Qu.:27.00 3rd Qu.:0.000   
## Max. :7716.05 Max. :18.00 Max. :59.00 Max. :1.000   
## smsa ne mw so   
## Min. :0.000 Min. :0.000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:1.000 1st Qu.:0.000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :1.000 Median :0.000 Median :0.0000 Median :0.0000   
## Mean :0.756 Mean :0.229 Mean :0.2485 Mean :0.3125   
## 3rd Qu.:1.000 3rd Qu.:0.000 3rd Qu.:0.0000 3rd Qu.:1.0000   
## Max. :1.000 Max. :1.000 Max. :1.0000 Max. :1.0000   
## we pt   
## Min. :0.00 Min. :0.0000   
## 1st Qu.:0.00 1st Qu.:0.0000   
## Median :0.00 Median :0.0000   
## Mean :0.21 Mean :0.0925   
## 3rd Qu.:0.00 3rd Qu.:0.0000   
## Max. :1.00 Max. :1.0000

dplyr 패키지가 row name을 없애는 문제를 해결하기 위하여 tibble 패키지를 활용하여 row name을 별도 열로 저장하였으며, exper가 0보다 작은 경우 결측치로 처리하였다.

uswages2 <- uswages %>%   
 rownames\_to\_column() %>%   
 mutate(exper = ifelse(exper < 0, NA, exper))

이후 분석의 효과성 제고를 위하여 해당 결측치 33개를 제외하였다.

sum(!complete.cases(uswages))

## [1] 0

sum(is.na(uswages2$exper))

## [1] 33

uswages2 <- na.omit(uswages2)  
nrow(uswages2)

## [1] 1967

dummy 변수로 입력되어 있는 ne, mw, we, so를 보다 용이하게 handling 하기 위하여 area 라는 열의 factor 데이터로 전환하였다. 또한 race, smsa, pt 등도 데이터의 성격에 맞게 factor 데이터로 전환하였다.

#지역 데이터 확인  
uswages2 %>%   
 mutate(area\_sum = ne + mw + we + so) %>%   
 dplyr::select(area\_sum) %>%   
 table()

## .  
## 1   
## 1967

#factor 변환  
uswages2 <- uswages2 %>%   
 mutate(race = factor(race, labels = c("White", "Black")),  
 smsa = factor(smsa, labels = c("No", "Yes")),  
 area\_temp = 1\*ne + 2\*mw + 3\*we + 4\*so,  
 area = factor(area\_temp, labels = c("ne", "mw", "we", "so")),  
 pt = factor(pt, labels = c("No", "Yes"))) %>%   
 dplyr::select(-c("ne", "mw", "we", "so","area\_temp"))

ggpairs 를 통해 데이터의 분포형태를 관찰한 결과, wage는 분포의 치우침 및 형태를 고려할 때 log 변환으로 처리할 필요가 보인다.

summary(uswages2)

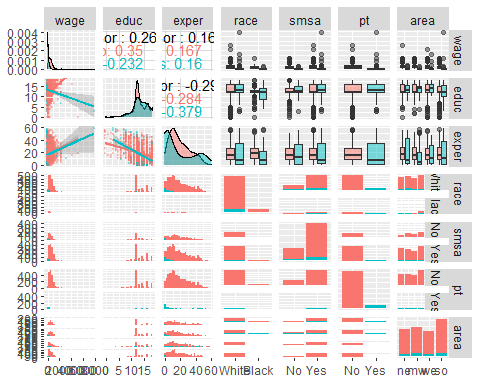
## rowname wage educ exper   
## Length:1967 Min. : 50.39 Min. : 0.00 Min. : 0.00   
## Class :character 1st Qu.: 314.69 1st Qu.:12.00 1st Qu.: 8.00   
## Mode :character Median : 522.32 Median :12.00 Median :16.00   
## Mean : 613.99 Mean :13.08 Mean :18.74   
## 3rd Qu.: 783.48 3rd Qu.:16.00 3rd Qu.:27.00   
## Max. :7716.05 Max. :18.00 Max. :59.00   
## race smsa pt area   
## White:1812 No : 483 No :1802 ne:448   
## Black: 155 Yes:1484 Yes: 165 mw:488   
## we:415   
## so:616   
##   
##

str(uswages2)

## 'data.frame': 1967 obs. of 8 variables:  
## $ rowname: chr "6085" "23701" "16208" "2720" ...  
## $ wage : num 772 617 958 617 902 ...  
## $ educ : int 18 15 16 12 14 12 16 16 12 12 ...  
## $ exper : int 18 20 9 24 12 33 42 0 36 37 ...  
## $ race : Factor w/ 2 levels "White","Black": 1 1 1 1 1 1 1 1 1 1 ...  
## $ smsa : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 1 ...  
## $ pt : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 2 2 1 ...  
## $ area : Factor w/ 4 levels "ne","mw","we",..: 1 3 4 1 2 3 4 2 3 2 ...

uswages2 %>%   
 dplyr::select(-"rowname") %>%   
 ggpairs(mapping = aes(colour = 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")

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

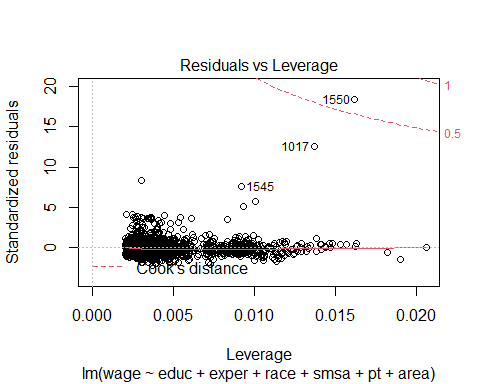
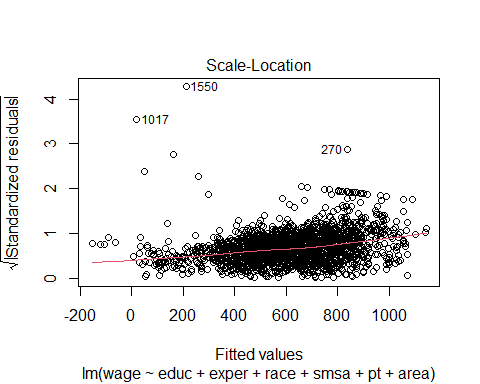
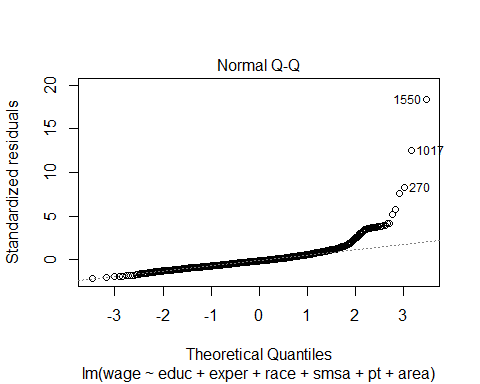
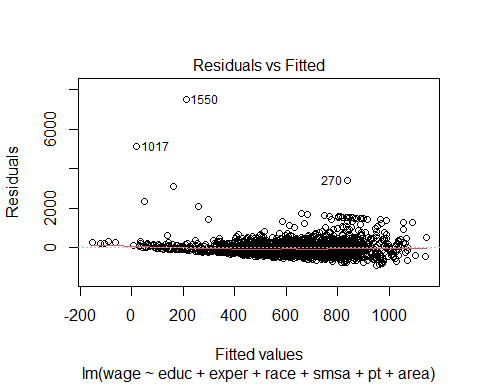


우선 로그변환 없이 wage를 종속변수로, 나머지 설명변수 모두를 포함하여 회귀모형을 적합하였다. p-value는 매우 유의하나 결정계수가 0.19로 낮고 diagnostic plot 상에서도 잔차의 정규성, 등분산성에 문제가 있는 것으로 판단된다. 또한, 해당 모형 상에서는 wage가 음수값이 도출될 수 있는 상황이나, 음수의 wage가 도출되는 것은 바람직하지 않다.

wage\_lm <- lm(wage ~ educ + exper + race + smsa + pt + area, data = uswages2)  
summary(wage\_lm)

##   
## Call:  
## lm(formula = wage ~ educ + exper + race + smsa + pt + area, data = uswages2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -877.5 -213.8 -55.4 129.8 7505.1   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -259.0701 55.1763 -4.695 2.85e-06 \*\*\*  
## educ 49.2987 3.2606 15.119 < 2e-16 \*\*\*  
## exper 8.9664 0.7369 12.167 < 2e-16 \*\*\*  
## raceBlack -121.8871 35.3490 -3.448 0.000576 \*\*\*  
## smsaYes 116.5704 21.8940 5.324 1.13e-07 \*\*\*  
## ptYes -326.2943 33.6231 -9.704 < 2e-16 \*\*\*  
## areamw -6.5410 27.1988 -0.240 0.809975   
## areawe 47.8021 28.2382 1.693 0.090650 .   
## areaso 2.8037 26.0787 0.108 0.914397   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 412.5 on 1958 degrees of freedom  
## Multiple R-squared: 0.1941, Adjusted R-squared: 0.1909   
## F-statistic: 58.96 on 8 and 1958 DF, p-value: < 2.2e-16

plot(wage\_lm)

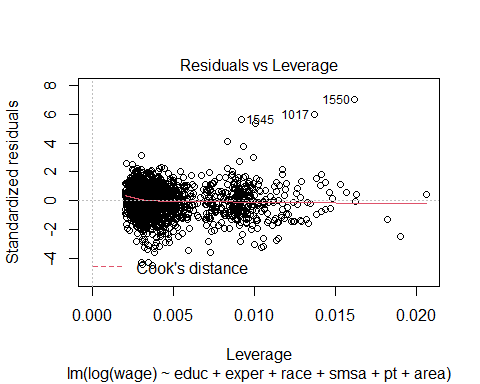
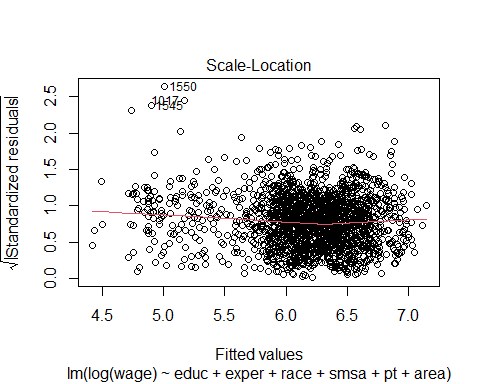
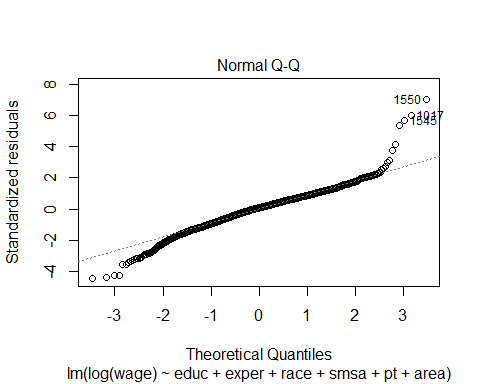
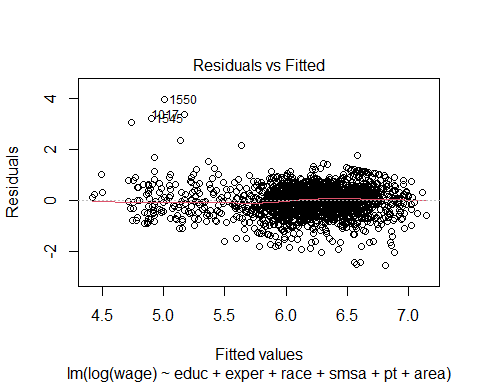


앞서 판단한 바와 같이 종속변수 wage를 로그 변환하여 회귀모형을 wage\_lm2로 다시 적합하였다. 음수의 wage 가 도출되는 문제가 해결되었을 뿐만 아니라 결정계수 또한 약 2배에 가까운 0.37로 증가하였음을 확인할 수 있다. 변수별 유의성에 있어서는 여전히 area를 제외한 모든 변수가 매우 유의하다. 한편 모형 진단을 위한 plot을 검토한 결과, 잔차의 등분산성 및 정규성이 개선되었으나 여전히 문제가 남아있는 것으로 보인다.

wage\_lm2 <- lm(log(wage) ~ educ + exper + race + smsa + pt + area, data = uswages2)  
summary(wage\_lm2)

##   
## Call:  
## lm(formula = log(wage) ~ educ + exper + race + smsa + pt + area,   
## data = uswages2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.5158 -0.3309 0.0504 0.3520 3.9446   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.710039 0.076052 61.932 < 2e-16 \*\*\*  
## educ 0.087515 0.004494 19.472 < 2e-16 \*\*\*  
## exper 0.015483 0.001016 15.243 < 2e-16 \*\*\*  
## raceBlack -0.215273 0.048723 -4.418 1.05e-05 \*\*\*  
## smsaYes 0.177741 0.030177 5.890 4.54e-09 \*\*\*  
## ptYes -1.067418 0.046344 -23.032 < 2e-16 \*\*\*  
## areamw 0.008320 0.037489 0.222 0.824   
## areawe 0.046846 0.038922 1.204 0.229   
## areaso 0.010080 0.035945 0.280 0.779   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5686 on 1958 degrees of freedom  
## Multiple R-squared: 0.368, Adjusted R-squared: 0.3654   
## F-statistic: 142.5 on 8 and 1958 DF, p-value: < 2.2e-16

plot(wage\_lm2)

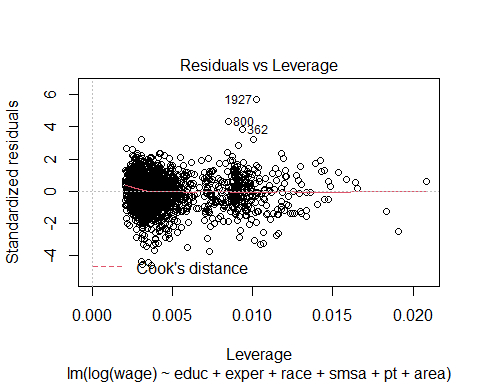
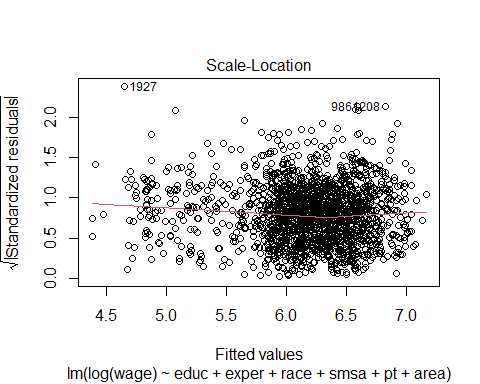
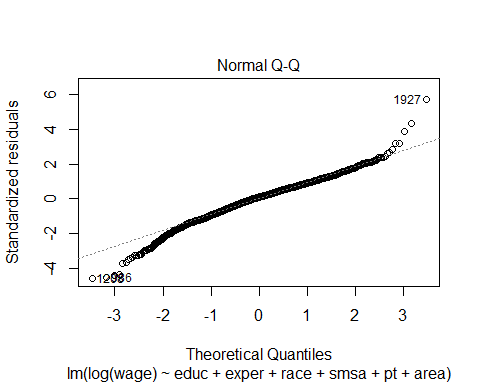
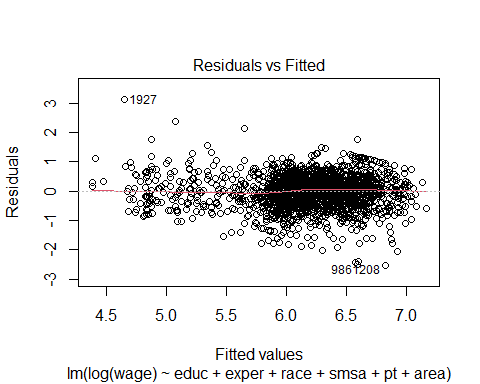


이를 개선하기 위하여 상기 diagostic 에서 지속적으로 outlier로 지목되는 3개의 데이터를 제외한 uswages3를 생성하여 다시 회귀모형을 적합해 보았다. p-value는 거의 유사하나 결정계수가 약 10%가량 개선되었음을 확인할 수 있다. diagnostic 상에서 잔차의 정규성 및 등분산성 또한 개선된 것이 확인된다. ggpairs에서도 관측할 수 있었던 여러 방면에서의 이상치를 모두 관리한다면 더욱 모델을 개선할 수 있을 것으로 보이나, 효율성의 측면에서 outlier 관리는 중단하고 이후 분석을 개시하였다.

uswages3 <- uswages2[-c(1545, 1017, 1550),]  
  
wage\_lm3 <- lm(log(wage) ~ educ + exper + race + smsa + pt + area, data = uswages3)  
summary(wage\_lm3)

##   
## Call:  
## lm(formula = log(wage) ~ educ + exper + race + smsa + pt + area,   
## data = uswages3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.52355 -0.32526 0.04878 0.35800 3.12782   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.6406720 0.0742235 62.523 < 2e-16 \*\*\*  
## educ 0.0930936 0.0043951 21.181 < 2e-16 \*\*\*  
## exper 0.0154416 0.0009874 15.638 < 2e-16 \*\*\*  
## raceBlack -0.2011520 0.0472874 -4.254 2.20e-05 \*\*\*  
## smsaYes 0.1670151 0.0292925 5.702 1.37e-08 \*\*\*  
## ptYes -1.1115009 0.0452277 -24.576 < 2e-16 \*\*\*  
## areamw 0.0158683 0.0363909 0.436 0.663   
## areawe 0.0471793 0.0378050 1.248 0.212   
## areaso 0.0114844 0.0349113 0.329 0.742   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5516 on 1955 degrees of freedom  
## Multiple R-squared: 0.3959, Adjusted R-squared: 0.3934   
## F-statistic: 160.1 on 8 and 1955 DF, p-value: < 2.2e-16

plot(wage\_lm3)



stepAIC 함수를 활용하여 변수선택을 진행한 결과, area 변수가 제거되었으며 이에 따른 최종 모형을 wage\_lm4로 저장하였다. 약 40%의 설명력을 가지며 모든 변수가 통계적으로 매우 유의하다. 교육수준 및 경력이 증가할 수록, 또한 표준대도시구역 거주 시 주급이 증가하며, 백인에 비해 흑인의 주급이 약 80% 수준이며, 파트타임 근무자의 주급은 일반 근무자의 30% 수준이었다.

변수 상 다중공선성 관련하여서는 vif가 1 수준으로 이슈가 없었으나, 잔차의 정규성, 등분산성 관련하여서는 모델을 개선할 여지가 있는 것으로 판단된다.

step <- stepAIC(wage\_lm3, direction = "both")

## Start: AIC=-2327.6  
## log(wage) ~ educ + exper + race + smsa + pt + area  
##   
## Df Sum of Sq RSS AIC  
## - area 3 0.520 595.45 -2331.9  
## <none> 594.93 -2327.6  
## - race 1 5.507 600.44 -2311.5  
## - smsa 1 9.893 604.82 -2297.2  
## - exper 1 74.419 669.35 -2098.1  
## - educ 1 136.528 731.46 -1923.8  
## - pt 1 183.793 778.72 -1800.9  
##   
## Step: AIC=-2331.88  
## log(wage) ~ educ + exper + race + smsa + pt  
##   
## Df Sum of Sq RSS AIC  
## <none> 595.45 -2331.9  
## + area 3 0.520 594.93 -2327.6  
## - race 1 5.943 601.39 -2314.4  
## - smsa 1 9.869 605.32 -2301.6  
## - exper 1 74.035 669.48 -2103.7  
## - educ 1 136.563 732.01 -1928.3  
## - pt 1 183.862 779.31 -1805.4

step$anova

## Stepwise Model Path   
## Analysis of Deviance Table  
##   
## Initial Model:  
## log(wage) ~ educ + exper + race + smsa + pt + area  
##   
## Final Model:  
## log(wage) ~ educ + exper + race + smsa + pt  
##   
##   
## Step Df Deviance Resid. Df Resid. Dev AIC  
## 1 1955 594.9289 -2327.598  
## 2 - area 3 0.5202788 1958 595.4492 -2331.882

wage\_lm4 <- lm(log(wage) ~ educ + exper + race + smsa + pt, data = uswages3)  
summary(wage\_lm4)

##   
## Call:  
## lm(formula = log(wage) ~ educ + exper + race + smsa + pt, data = uswages3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.53967 -0.32807 0.05106 0.35572 3.12072   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.6621211 0.0685692 67.991 < 2e-16 \*\*\*  
## educ 0.0929905 0.0043882 21.191 < 2e-16 \*\*\*  
## exper 0.0153779 0.0009856 15.603 < 2e-16 \*\*\*  
## raceBlack -0.2053326 0.0464498 -4.421 1.04e-05 \*\*\*  
## smsaYes 0.1655021 0.0290524 5.697 1.41e-08 \*\*\*  
## ptYes -1.1109243 0.0451809 -24.588 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5515 on 1958 degrees of freedom  
## Multiple R-squared: 0.3954, Adjusted R-squared: 0.3938   
## F-statistic: 256 on 5 and 1958 DF, p-value: < 2.2e-16

exp(wage\_lm4$coefficients)

## (Intercept) educ exper raceBlack smsaYes ptYes   
## 105.8603839 1.0974513 1.0154968 0.8143764 1.1799855 0.3292545

vif(wage\_lm4)

## educ exper race smsa pt   
## 1.112003 1.096309 1.012897 1.010867 1.003314

shapiro.test(wage\_lm4$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: wage\_lm4$residuals  
## W = 0.97755, p-value < 2.2e-16

ncvTest(wage\_lm4)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 8.845653, Df = 1, p = 0.0029379