st538 Project 1 Week 3

Philip Ourso

4/13/2022

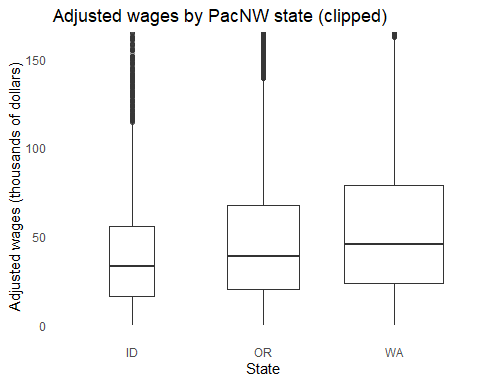
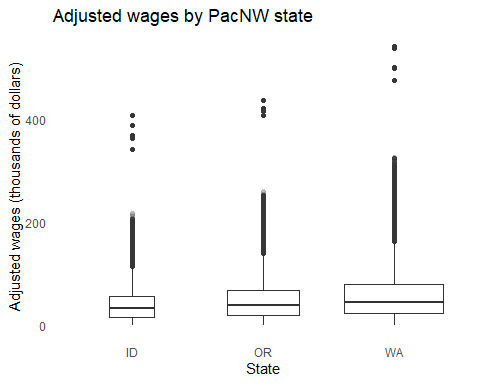
## Project #1 - Interim Report

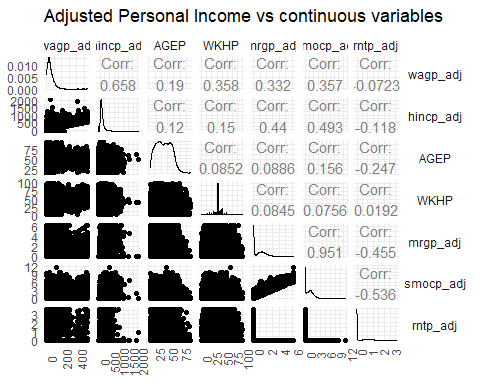
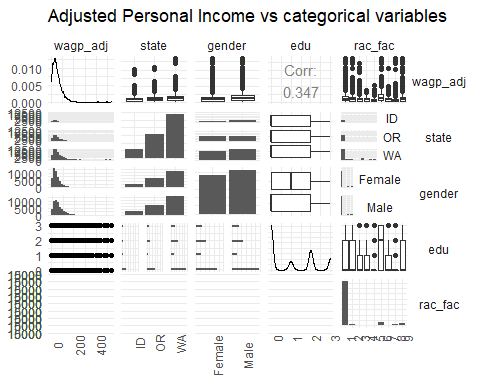
After accounting for gender, race, educational attainment and typical hours worked, is there an effect of state residency on total person’s income in the Pacific Northwest states of Washington, Oregon and Idaho?

The dataset used included 5-year PUMS data for the states of Washington, Oregon and Idaho, for years 2015-2019. Income and housing cost variables were adjusted and string data were recoded in some cases for convenience.  
Specific variables include:  
- ESR: employment status  
- SEX: gender  
- MRGP: mortgage payment amount  
- COW: class of worker  
- AGEP: age  
- RAC1P: race  
- SCHL: educational attainment  
- WKHP: usual hours worked per week  
- SMOCP: selected monthly owner costs  
- WAGP: wages or salary income  
- RNTP: monthly rent

### Exploratory data analysis

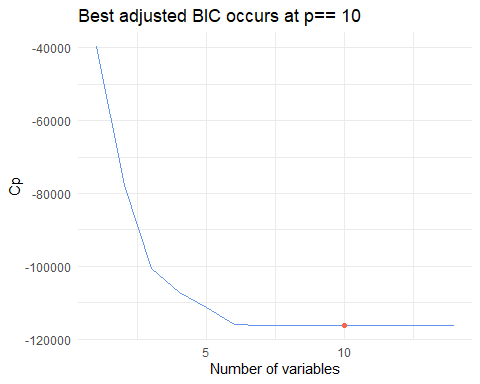
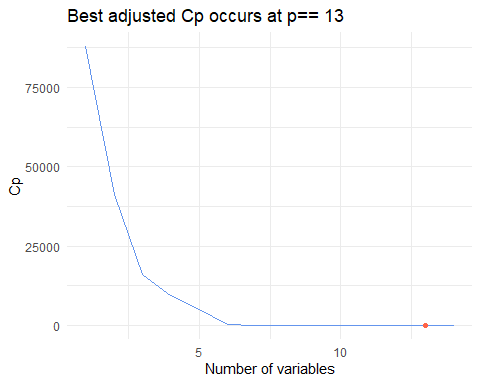
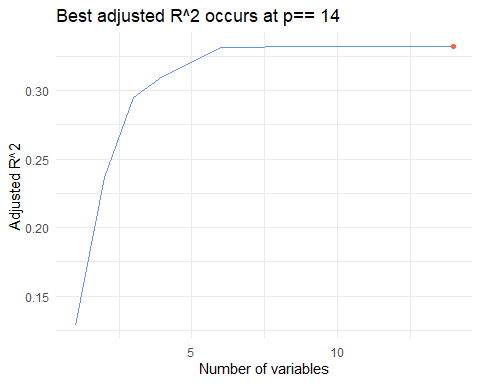
Variables expressing amounts of money were scaled to thousands of dollars and employment status was used to filter individuals not actively employed. Educational attainment was binned to reduce the number of levels.  
Boxplots and density plots of adjusted income by state are difficult to read in the presence of outliers, so the scale of observations is restricted to get a better look at the bulk of the data. Idaho exibits a slightly lower median, while Washington has a slightly higher interquartile range.  
Pairs plots were created to get a quick snapshot of any potential correlations. Due to the size of the dataset however, a subset of 100,000 observations was considered. Furthermore, categorical variables were considered separately to continuous variables. The log transformation of continuous variables was also considered separately.  
Of the categorical variables, educational attainment exhibited the most obvious trend, with advanced degree holders exhibiting higher personal income. Gender, class of worker and the aforementioned state also showed variation. Reviewing continuous variables revealed the high likelihood of muticollinearity between mortgage payment (MRGP) and selected monthly owner costs (SMOCP), not surprising considering the presumed relationship between those pairs. Mortgage payment was dropped from consideration.  
Log transformations of continuous variable revealed no obvious signals.





### Linear models

Withholding the state variable, for which inference is desired, best subset selection is used to determine which variables to include in the resulting model. Both adjusted R^2 and Cp prefer nearly all variables. Considering BIC reduces this to 10 variables.  
Taking the BIC-preferred variables as a starting point, a rich model is fit first. A number of variables are subsequently withheld, one at a time, and F-tests performed to compare reduced models to the full model. In all cases we fail to reject the null hypothesis and the full model is retained.

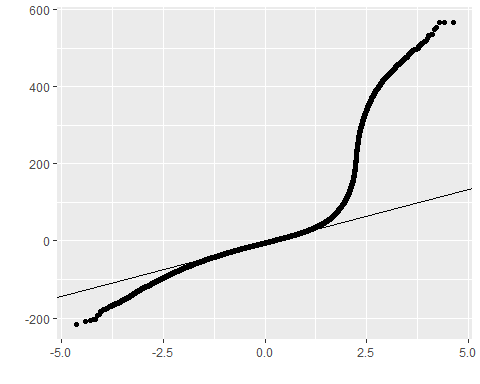
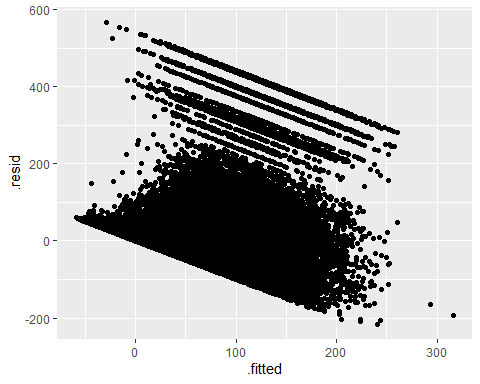


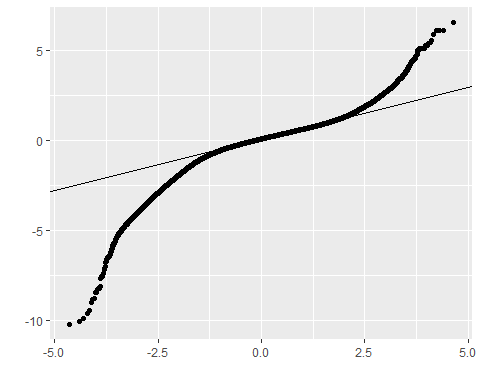
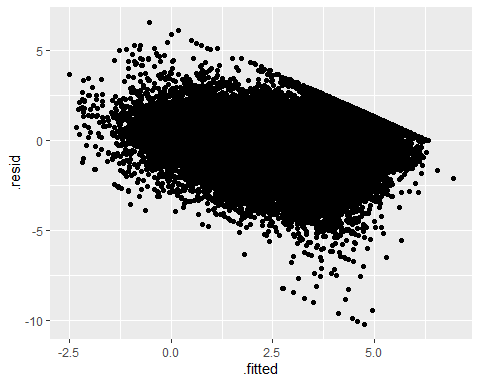
## (Intercept) genderMale edu rac\_fac2 rac\_fac3 rac\_fac6   
## -66.3221230 16.7907332 12.9619489 -6.4351123 1.6881011 2.7002311   
## rac\_fac7 rac\_fac8 rac\_fac9 AGEP WKHP smocp\_adj   
## -5.5963456 -6.4711698 -1.5462539 0.5520107 1.3899524 17.1980075   
## rntp\_adj   
## 13.2174033

### Model diagnostics

Proceeding with the full model identified by best subset selection, model diagnostics reveal large residual errors with extreme deviation from normality, indicating a violation of the assumptions of linear regression. The plot of residual error against fitted value shows several concerning issues: a large percentage of the population has residual errors of hundreds of thousands of dollars, predicted wages are negative for a subset of the population and there is a downward trend indicating that the residual error becomes negative as fitted values become larger.  
Plots of residuals against explanatory variables reveal further concerning trends, as the constant variance assumption appears to be violated, particularly at the lower and upper values of the predictors. Lastly, the QQ-plot exhibits an extreme departure from normality at upper quantiles.

Returning to model definition, response is instead modeled as the logarithm of wages and the logarithm of continuous predictors is include in best subset selection. Using the logarithm of wages earned results in a higher adjusted R^2, although logarithms of predictors aren’t included in the optimal subset. Repeating model diagnostics continues to reveal model issues: while deviation from normality assumption is somewhat reduced, there is clearly non-constant variance at higher values of explanatory variables and residuals continue to be large.





### Conclusion

Clearly the model, as identified by best subset selection and fitted by multiple linear regression, exhibits several serious violations of assumptions. For completeness sake, in order to answer the question posited at the outset, a linear regression model is fit with all previously included explanatory variables and Pacific Northwest state, as a factor. Reviewing the results from this new model indicates that, with respect to Idaho residency as a baseline, only Washington residency is found to be significant at a p-value < 0.0001. The coefficient for the Washington indicator is estimated to be 5.45, indicating a non-trivial impact upon wages (in thousands of dollars).  
Even with the obvious improvement from modeling the logarithm of wages earned, serious issues remain. Including state information in the model with the logarithm response now results in the statistical significance of both Oregon and Washington residency with respect to Idaho residency. Oregon residency is estimated to correspond to roughly 9% higher wages, while Washington residency approaches 20% higher wages.

##   
## Call:  
## lm(formula = log(wagp\_adj) ~ gender + edu + rac\_fac + AGEP +   
## WKHP + smocp\_adj + rntp\_adj + state, data = df.red.state.tmp)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.7032 -0.3338 0.1188 0.4941 4.6966   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.3915104 0.0080723 48.501 < 2e-16 \*\*\*  
## genderMale 0.1907854 0.0031462 60.640 < 2e-16 \*\*\*  
## edu 0.2122204 0.0014211 149.334 < 2e-16 \*\*\*  
## rac\_fac2 -0.1663948 0.0107475 -15.482 < 2e-16 \*\*\*  
## rac\_fac3 -0.0151995 0.0137405 -1.106 0.269   
## rac\_fac4 -0.0516999 0.0589283 -0.877 0.380   
## rac\_fac5 -0.0941130 0.0401287 -2.345 0.019 \*   
## rac\_fac6 -0.0010948 0.0062946 -0.174 0.862   
## rac\_fac7 -0.0968923 0.0229425 -4.223 2.41e-05 \*\*\*  
## rac\_fac8 -0.1448172 0.0091138 -15.890 < 2e-16 \*\*\*  
## rac\_fac9 -0.0740751 0.0081558 -9.082 < 2e-16 \*\*\*  
## AGEP 0.0147437 0.0001109 132.939 < 2e-16 \*\*\*  
## WKHP 0.0475760 0.0001306 364.180 < 2e-16 \*\*\*  
## smocp\_adj 0.1786372 0.0015918 112.224 < 2e-16 \*\*\*  
## rntp\_adj 0.1659216 0.0030425 54.535 < 2e-16 \*\*\*  
## stateOR 0.0866406 0.0053144 16.303 < 2e-16 \*\*\*  
## stateWA 0.1714405 0.0050524 33.932 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8204 on 288392 degrees of freedom  
## Multiple R-squared: 0.4864, Adjusted R-squared: 0.4863   
## F-statistic: 1.707e+04 on 16 and 288392 DF, p-value: < 2.2e-16

### Appendix: R code

knitr::opts\_chunk$set(echo = FALSE)  
library(tidyverse)  
library(GGally)  
library(leaps)  
#library(graphics)  
library(Rmisc)  
  
  
# copied from: Week2\_LR2\_code.R  
if(! (packageVersion("ggplot2") >= "2.0.0")) {  
 stop("This version of stat\_qqline require ggplot2 version 2.0.0 or greater")  
}  
  
StatQQLine <- ggproto("StatQQLine", Stat,  
 compute\_group = function(data, scales, distribution = qnorm, dparams = list()) {  
 data <- remove\_missing(data, na.rm = TRUE, "sample", name = "stat\_qqline")  
 y <- quantile(data$sample, c(0.25, 0.75))  
 x <- do.call(distribution, c(list(p = c(0.25, 0.75)), dparams))  
 slope <- diff(y)/diff(x)  
 int <- y[1L] - slope \* x[1L]  
 data.frame(slope = slope, intercept = int)  
 },  
   
 required\_aes = c("sample")  
)  
  
stat\_qqline <- function(mapping = NULL, data = NULL, geom = "abline",  
 position = "identity", na.rm = FALSE, show.legend = NA,  
 distribution = qnorm, dparams = list(),  
 inherit.aes = TRUE, ...) {  
 layer(  
 stat = StatQQLine, data = data, mapping = mapping, geom = geom,  
 position = position, show.legend = show.legend, inherit.aes = inherit.aes,  
 params = list(na.rm = na.rm, distribution = distribution, dparams = dparams, ...)  
 )  
}  
# read data  
df = read.csv(unz("ACSPUMS5Y2019\_2022-04-16T191614.zip",  
 "ACSPUMS5Y2019\_2022-04-16T165733.csv"  
 ), header = T)  
# dim(df)  
# head(df,20)  
  
# char's adding col to end?  
df = subset(df, select=-c(X))  
#colnames(df)  
  
# check for missing  
#df[is.na(df),]  
  
# recode state, gender as string  
df$state = ifelse(df$ST == 16, "ID",  
 ifelse(df$ST == 41, "OR",  
 "WA"  
 )  
 )  
df$gender = ifelse(df$SEX==1, "Male", "Female")  
  
# adjust for inflation  
df = df %>%   
 mutate(pincp\_adj = (PINCP\*ADJINC)/1000,  
 hincp\_adj = (HINCP\*ADJINC)/1000,  
 mrgp\_adj = (MRGP\*ADJHSG)/1000,  
 smocp\_adj = (SMOCP\*ADJHSG)/1000,  
 rntp\_adj = (RNTP\*ADJHSG)/1000,  
 wagp\_adj = (WAGP\*ADJINC)/1000  
 )  
  
# convert school and class of worker to factor  
df = df %>%  
 mutate(  
 schl\_fac = factor(SCHL),  
 cow\_fac = factor(COW),  
 rac\_fac = factor(RAC1P)  
 )  
  
# retain only records in labor force  
df = df %>%  
 filter(  
 ESR != 0,  
 ESR != 3,  
 ESR != 6,  
 WKHP > 0,  
 WKHP < 99,  
 wagp\_adj > 0  
 )  
  
# recode education to reduce levels  
df = df %>%  
 mutate(  
 edu = case\_when(  
 SCHL <= 19 ~ 0, # high school or less  
 SCHL == 20 ~ 1, # associates  
 SCHL == 21 ~ 2, # bachelors  
 SCHL > 21 ~ 3 # master's or doctoral  
 ),  
 edu = factor(edu)  
 )  
  
  
  
# univariate distributions  
  
# unadjusted box plots  
#ggplot(df, aes(x=factor(ST), y=PINCP)) +  
# geom\_boxplot(aes(group=ST)) +  
# theme\_minimal()  
  
# adjusted income by state  
p0 = ggplot(df, aes(x=factor(state), y=wagp\_adj)) +  
 geom\_boxplot(aes(group=state), varwidth=T, outlier.alpha=0.1) +  
 theme\_minimal() +   
 theme(  
 panel.grid = element\_blank()  
 ) +  
 labs(  
 x = "State",  
 y = "Adjusted wages (thousands of dollars)",  
 title = "Adjusted wages by PacNW state"  
 )  
p0  
  
# view only up to whiskers  
ylim = boxplot.stats(df$wagp\_adj)$stats[c(1,5)]  
p0 +   
 coord\_cartesian(ylim=ylim\*1.05) +  
 labs(  
 title = "Adjusted wages by PacNW state (clipped)"  
 )  
  
## adjusted income by state  
# ggplot(df, aes(x=wagp\_adj)) +  
# geom\_density(aes(color=factor(state))  
# ) +  
# theme\_minimal() +   
# theme(  
# panel.grid = element\_blank()  
# ) +  
# labs(  
# y = "Density",  
# x = "Adjusted wages (thousands of dollars)",  
# title = "Adjusted wages by PacNW state"  
# )  
  
# # adjusted income by gender  
# ggplot(df, aes(x=gender, y=wagp\_adj)) +  
# geom\_boxplot(aes(group=gender), varwidth=T, outlier.alpha=0.1) +  
# theme\_minimal() +   
# theme(  
# panel.grid = element\_blank()  
# ) +  
# labs(  
# x = "State",  
# y = "Adjusted wages (thousands of dollars)",  
# title = "Adjusted income by gender"  
# )  
  
# # adjusted income by school  
# ggplot(df, aes(x=edu, y=wagp\_adj)) +  
# geom\_boxplot(aes(group=edu), varwidth=T, outlier.alpha=0.1) +  
# theme\_minimal() +   
# theme(  
# panel.grid = element\_blank()  
# ) +  
# labs(  
# x = "Education",  
# y = "Adjusted wages (thousands of dollars)",  
# title = "Adjusted wages by education"  
# )  
#   
# # adjusted income by work  
# ggplot(df, aes(x=cow\_fac, y=wagp\_adj)) +  
# geom\_boxplot(aes(group=cow\_fac), varwidth=T, outlier.alpha=0.1) +  
# theme\_minimal() +   
# theme(  
# panel.grid = element\_blank()  
# ) +  
# labs(  
# x = "Class of worker",  
# y = "Adjusted wages (thousands of dollars)",  
# title = "Adjusted wages by class of worker"  
# )  
#   
# # adjusted income by age  
# ggplot(df, aes(x=AGEP, y=wagp\_adj)) +  
# geom\_hex() +  
# theme\_minimal() +   
# theme(  
# panel.grid = element\_blank()  
# ) +  
# labs(  
# x = "Age",  
# y = "Adjusted wages (thousands of dollars)",  
# title = "Adjusted wages by age"  
# )  
#   
# # adjusted income by age  
# ggplot(df, aes(x=AGEP, y=wagp\_adj)) +  
# geom\_point(alpha=0.1) +  
# theme\_minimal() +   
# theme(  
# panel.grid = element\_blank()  
# ) +  
# labs(  
# x = "Age",  
# y = "Adjusted wages (thousands of dollars)",  
# title = "Adjusted wages by age"  
# )  
#   
# # adjusted income by hours/ week  
# ggplot(df, aes(x=WKHP, y=wagp\_adj)) +  
# geom\_point(alpha=0.1) +  
# theme\_minimal() +   
# theme(  
# panel.grid = element\_blank()  
# ) +  
# labs(  
# x = "Age",  
# y = "Adjusted wages (thousands of dollars)",  
# title = "Adjusted wages by age"  
# )  
# ggpairs categorical vars  
ggpairs(subset(df[sample(nrow(df),100000),],   
 select = c(wagp\_adj, state, gender, edu, rac\_fac ))) +  
 theme\_minimal() +  
 theme(  
 axis.text.x = element\_text(angle=90),  
 strip.text.y.right = element\_text(angle=0)  
 ) +   
 labs(  
 title = "Adjusted Personal Income vs categorical variables"  
 )  
  
# ggpairs continuous vars  
ggpairs(subset(df[sample(nrow(df),100000),],   
 select = c(wagp\_adj, hincp\_adj, AGEP, WKHP, mrgp\_adj, smocp\_adj, rntp\_adj))) +  
 theme\_minimal() +  
 theme(  
 axis.text.x = element\_text(angle=90),  
 strip.text.y.right = element\_text(angle=0)  
 ) +  
 labs(  
 title = "Adjusted Personal Income vs continuous variables"  
 )  
# consider log transformations of continuous vars  
df.log = subset(df, select=c(wagp\_adj, hincp\_adj, AGEP, WKHP, mrgp\_adj, smocp\_adj, rntp\_adj))  
df.log = df.log %>%  
 mutate(wagp\_adj.log = log(wagp\_adj),  
 hincp\_adj.log = log(hincp\_adj),  
 AGEP.log = log(AGEP),  
 WKHP.log = log(WKHP),  
 mrgp\_adj.log = log(mrgp\_adj),  
 smocp\_adj.log = log(smocp\_adj),  
 rntp\_adj.log = log(rntp\_adj)  
 )  
  
ggpairs(subset(df.log[sample(nrow(df),100000),],   
 select = c(wagp\_adj, wagp\_adj.log, hincp\_adj.log, AGEP.log,   
 WKHP.log, mrgp\_adj.log, smocp\_adj.log, rntp\_adj.log))) +  
 theme\_minimal() +  
 theme(  
 axis.text.x = element\_text(angle=90),  
 strip.text.y.right = element\_text(angle=0)  
 ) +  
 labs(  
 title = "Adj wages vs cont. variables log transforms"  
 )  
  
# best subset selection, leave state out  
df.red = subset(df,   
 select = c(wagp\_adj, gender, edu, rac\_fac, AGEP, WKHP, smocp\_adj, rntp\_adj) )  
  
regfit.full = regsubsets(wagp\_adj ~ ., df.red, nvmax=20)  
reg.summary = summary(regfit.full)  
  
tb = tibble(  
 p = 1:length(reg.summary$rsq),  
 rsq = reg.summary$rsq,  
 rss = reg.summary$rss,  
 adjr2 = reg.summary$adjr2,  
 cp = reg.summary$cp,  
 bic = reg.summary$bic,  
)  
  
# adjusted R^2  
ggplot(tb, aes(x = p,y = adjr2)) +  
 geom\_line(color = 'cornflowerblue') +  
 geom\_point(data=tb[tb$p==which.max( reg.summary$adjr2 ),],   
 mapping=aes(x = p,y = adjr2),color='tomato') +  
 theme\_minimal() +  
 labs(  
 x = "Number of variables",  
 y = "Adjusted R^2",  
 title = paste("Best adjusted R^2 occurs at p==", toString(which.max( reg.summary$adjr2 )))  
 )  
  
# adjusted Cp  
ggplot(tb, aes(x = p,y = cp)) +  
 geom\_line(color = 'cornflowerblue') +  
 geom\_point(data=tb[tb$p==which.min( reg.summary$cp ),],   
 mapping=aes(x = p,y = cp),color='tomato') +  
 theme\_minimal() +  
 labs(  
 x = "Number of variables",  
 y = "Cp",  
 title = paste("Best adjusted Cp occurs at p==", toString(which.min( reg.summary$cp )))  
 )  
  
# adjusted Cp  
ggplot(tb, aes(x = p,y = bic)) +  
 geom\_line(color = 'cornflowerblue') +  
 geom\_point(data=tb[tb$p==which.min( reg.summary$bic ),],   
 mapping=aes(x = p,y = bic),color='tomato') +  
 theme\_minimal() +  
 labs(  
 x = "Number of variables",  
 y = "Cp",  
 title = paste("Best adjusted BIC occurs at p==", toString(which.min( reg.summary$bic )))  
 )  
  
coef(regfit.full, 12)  
  
lm.full = lm(wagp\_adj ~ gender + edu + rac\_fac + AGEP + WKHP + smocp\_adj + rntp\_adj, df.red)  
summary(lm.full)  
  
# consider reduced models  
lm.red1 = lm(wagp\_adj ~ gender + edu + AGEP + WKHP + smocp\_adj + rntp\_adj, df.red)  
anova(lm.red1, lm.full)  
  
lm.red2 = lm(wagp\_adj ~ gender + edu + rac\_fac + AGEP + WKHP + smocp\_adj, df.red)  
anova(lm.red2, lm.full)  
  
lm.red3 = lm(wagp\_adj ~ edu + rac\_fac + AGEP + WKHP + smocp\_adj + rntp\_adj, df.red)  
anova(lm.red3, lm.full)  
  
lm.red4 = lm(wagp\_adj ~ gender + rac\_fac + AGEP + WKHP + smocp\_adj + rntp\_adj, df.red)  
anova(lm.red4, lm.full)  
  
lm.full.log = lm(log(wagp\_adj) ~ gender + edu + rac\_fac + AGEP + WKHP + smocp\_adj + rntp\_adj, df.red)  
summary(lm.full.log)  
# log model  
  
df.tmp = df.red %>%  
 mutate(  
 WKHP.log = log(WKHP),  
 AGEP.log = log(AGEP),  
 WAGP.log = log(wagp\_adj)  
 )   
  
  
regfit.test = regsubsets(WAGP.log ~ gender + edu + rac\_fac + AGEP + WKHP + smocp\_adj + rntp\_adj + WKHP.log + AGEP.log, df.tmp, nvmax=20)  
reg.summary.tmp = summary(regfit.test)  
  
  
tb.tmp = tibble(  
 p = 1:length(reg.summary.tmp$rsq),  
 rsq = reg.summary.tmp$rsq,  
 rss = reg.summary.tmp$rss,  
 adjr2 = reg.summary.tmp$adjr2,  
 cp = reg.summary.tmp$cp,  
 bic = reg.summary.tmp$bic,  
)  
  
# adjusted R^2  
ggplot(tb.tmp, aes(x = p,y = adjr2)) +  
 geom\_line(color = 'cornflowerblue') +  
 geom\_point(data=tb.tmp[tb.tmp$p==which.max( reg.summary.tmp$adjr2 ),],   
 mapping=aes(x = p,y = adjr2),color='tomato') +  
 theme\_minimal() +  
 labs(  
 x = "Number of variables",  
 y = "Adjusted R^2",  
 title = paste("Best adjusted R^2 occurs at p==", toString(which.max( reg.summary.tmp$adjr2 )))  
 )  
  
# adjusted Cp  
ggplot(tb.tmp, aes(x = p,y = cp)) +  
 geom\_line(color = 'cornflowerblue') +  
 geom\_point(data=tb.tmp[tb.tmp$p==which.min( reg.summary.tmp$cp ),],   
 mapping=aes(x = p,y = cp),color='tomato') +  
 theme\_minimal() +  
 labs(  
 x = "Number of variables",  
 y = "Cp",  
 title = paste("Best adjusted Cp occurs at p==", toString(which.min( reg.summary.tmp$cp )))  
 )  
  
# adjusted Cp  
ggplot(tb.tmp, aes(x = p,y = bic)) +  
 geom\_line(color = 'cornflowerblue') +  
 geom\_point(data=tb.tmp[tb.tmp$p==which.min( reg.summary.tmp$bic ),],   
 mapping=aes(x = p,y = bic),color='tomato') +  
 theme\_minimal() +  
 labs(  
 x = "Number of variables",  
 y = "Cp",  
 title = paste("Best adjusted BIC occurs at p==", toString(which.min( reg.summary.tmp$bic )))  
 )  
  
coef(regfit.full, 13)  
  
# residual vs fitted  
df.red.fort = fortify(lm.full, df.red)  
qplot(.fitted, .resid, data=df.red.fort)  
  
# # residual vs explanatory  
# qplot(AGEP, .resid, data = df.red.fort)   
# qplot(WKHP, .resid, data = df.red.fort)   
# #qplot(WKWN, .resid, data = df.red.fort)   
# qplot(smocp\_adj, .resid, data = df.red.fort)   
# qplot(rntp\_adj, .resid, data = df.red.fort)   
# #multiplot(p1, p2, p3, p4, p5, cols=5)  
  
# qqplot  
qplot(sample = df.red.fort$.resid) + stat\_qqline()   
  
  
lm.test = lm(WAGP.log ~ gender + edu + rac\_fac + AGEP + WKHP + smocp\_adj + rntp\_adj + WKHP.log + AGEP.log, df.tmp)  
#summary(lm.test)  
  
# check reduced model  
  
# residual vs fitted  
df.tmp.fort = fortify(lm.test, df.tmp)  
qplot(.fitted, .resid, data=df.tmp.fort)  
  
# # residual vs explanatory  
# qplot(AGEP, .resid, data = df.tmp.fort)   
# qplot(WKHP, .resid, data = df.tmp.fort)   
# qplot(smocp\_adj, .resid, data = df.tmp.fort)   
# qplot(rntp\_adj, .resid, data = df.tmp.fort)   
# #multiplot(p1, p2, p3, p4, p5, cols=5)  
  
# qqplot  
qplot(sample = df.tmp.fort$.resid) + stat\_qqline()   
  
df.red.state = subset(df,   
 select = c(wagp\_adj, gender, edu, rac\_fac, AGEP, WKHP, smocp\_adj, rntp\_adj, state) )  
lm.full.state = lm(wagp\_adj ~ gender + edu + rac\_fac + AGEP + WKHP + smocp\_adj + rntp\_adj + state, df.red.state)  
summary(lm.full.state)   
  
df.red.state.tmp = subset(df,   
 select = c(wagp\_adj, gender, edu, rac\_fac, AGEP, WKHP, smocp\_adj, rntp\_adj, state) )  
lm.full.state.tmp = lm(log(wagp\_adj) ~ gender + edu + rac\_fac + AGEP + WKHP + smocp\_adj + rntp\_adj + state, df.red.state.tmp)  
summary(lm.full.state.tmp)