USWages-Dataset-Analysis.R

Shraddha Somani

```r
require(faraway)

## Loading required package: faraway

require(ggplot2)

## Loading required package: ggplot2

require(GGally)

## Loading required package: GGally

##
## Attaching package: 'GGally'

## The following object is masked from 'package:faraway':
##
##     happy

require(gridExtra)

## Loading required package: gridExtra

require(e1071)

## Loading required package: e1071

head(uswages,10)

##         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
## 14379 541.31   16    42    0    1  0  0  1  0  1
## 12878 148.39   16     0    0    1  0  1  0  0  1
## 23121 273.19   12    36    0    1  0  0  0  1  1
## 13086 666.67   12    37    0    0  0  1  0  0  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
```

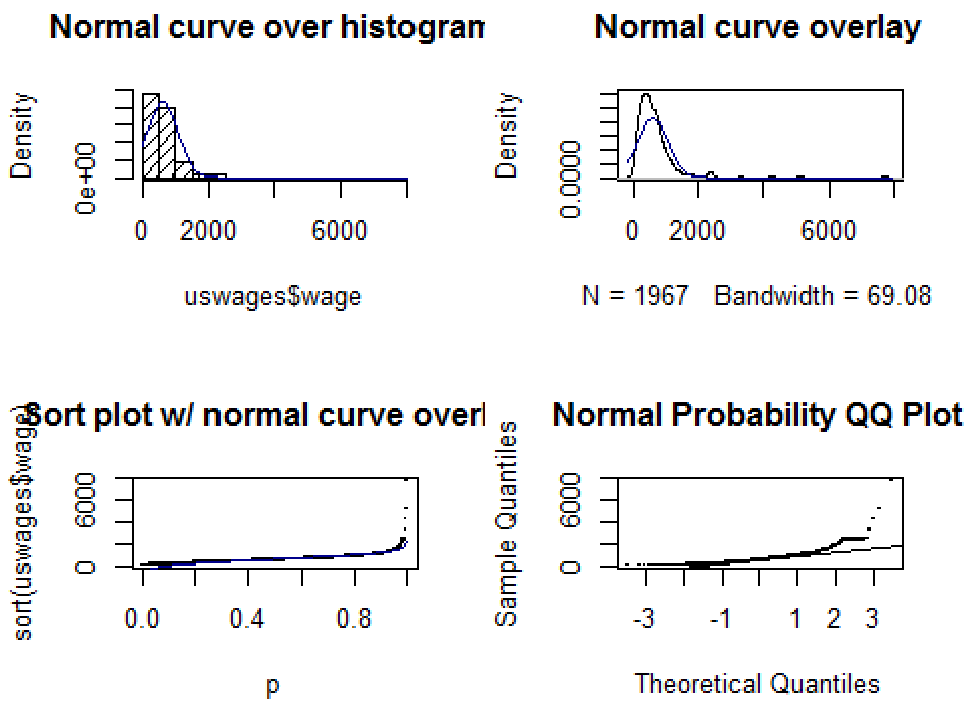
```r
# We see that exper has negative values.
uswages$exper[uswages$exper < 0] = NA
summary(uswages$exper)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00    8.00   16.00   18.74   27.00   59.00      33
```

```r
# Convert categorical variables to factors
uswages$race = factor(uswages$race)
levels(uswages$race) = c("White", "Black")
uswages$smsa = factor(uswages$smsa)
levels(uswages$smsa) = c("No", "Yes")
uswages$pt = factor(uswages$pt)
levels(uswages$pt) = c("No", "Yes")


# Convert set of dummy variables to one variable
uswages = data.frame(uswages, region = 1*uswages$ne + 2*uswages$mw +
3*uswages$so + 4*uswages$we)
uswages$region = factor(uswages$region)
levels(uswages$region) = c("ne", "mw", "so", "we")


# Deleting four regions ne, mw, so and we
uswages = subset(uswages, select = -c(ne:we))


# Take care of NA's
uswages = na.omit(uswages)

# 5 - Number Summary
summary(uswages)
```

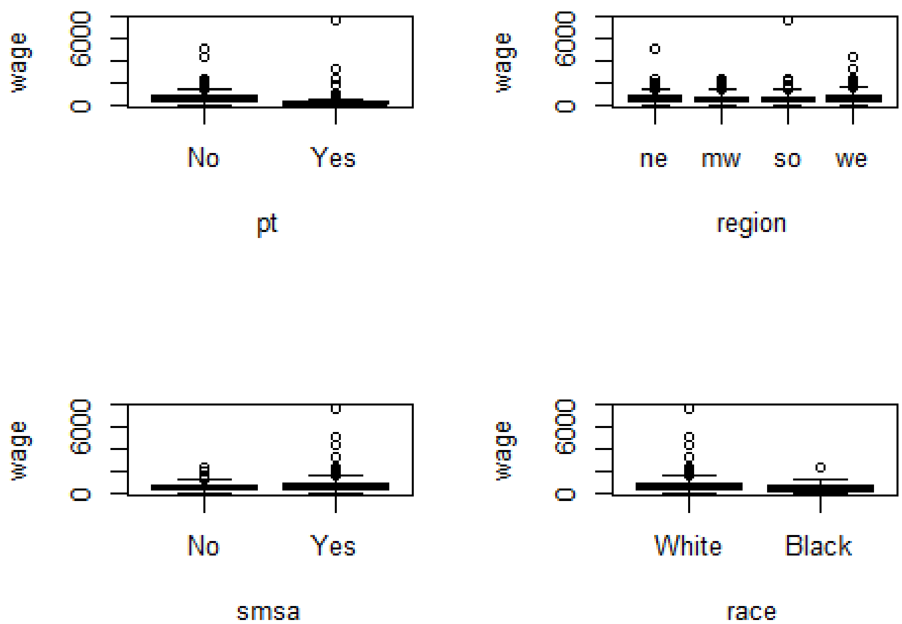
```
##       wage               educ            exper           race       smsa
##  Min.   :  50.39   Min.   : 0.00   Min.   : 0.00   White:1812   No : 483
##  1st Qu.: 314.69   1st Qu.:12.00   1st Qu.: 8.00   Black: 155   Yes:1484
##  Median : 522.32   Median :12.00   Median :16.00
##  Mean   : 613.99   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
##    pt        region
##  No :1802   ne:448
##  Yes: 165   mw:488
##             so:616
##             we:415
##
##

# Analysis:
# - If Mean and Median are unequal, skewness is present.
# - Skewed - Wage
# - Not Skwed - Educ and Exper
# - In Race, smsa, pt and region there is no concept of skewness because it
has binary values.
# - There is an unbalanced counts for race, smsa and pt.
# - This would tend to weaken the strength of a factor to predict the wages.


# Correlation
cor(uswages$wage, uswages$educ)

## [1] 0.2616368

# - There is a weak positive correlation between wage and educ.
cor(uswages$wage, uswages$exper)

## [1] 0.1694355

# - There is a weak positive correlation between wage and exper.
cor(uswages$educ,uswages$exper)

## [1] -0.2934846

# - There is a weak negative correlation between educ and exper.


# Distribution of wages
m = mean(uswages$wage, na.rm = TRUE)
std = sd(uswages$wage, na.rm = TRUE)
n = length(uswages$wage)
p = 1:n/(n+1)
oldpar = par(mfrow = c(2,2))
hist(
  uswages$wage,
```

```
  density = 20,
  breaks = 20,
  freq = FALSE,
  prob=TRUE,
  xlab = "uswages$wage",
  main = "Normal curve over histogram")
curve(
  dnorm(x, mean = m, sd = std),
  col = "darkblue",
  lwd = 0.25,
  add=TRUE)
plot(
  density(uswages$wage),
  main = "Normal curve overlay")
curve(
  dnorm(x, mean = m, sd = std),
  col = "darkblue",
  lwd = 0.25,
  add = TRUE)
plot(
  p,
  sort(uswages$wage),
  pch = ".",
  cex = 2,
  main = "Sort plot w/ normal curve overlay")
curve(
  qnorm(x, mean = m, sd = std),
  col = "darkblue",
  lwd = 0.25,
  add = TRUE)
qqnorm(
  uswages$wage,
  pch = ".",
  cex = 2,
  main = "Normal Probability QQ Plot")
qqline(uswages$wage)
```

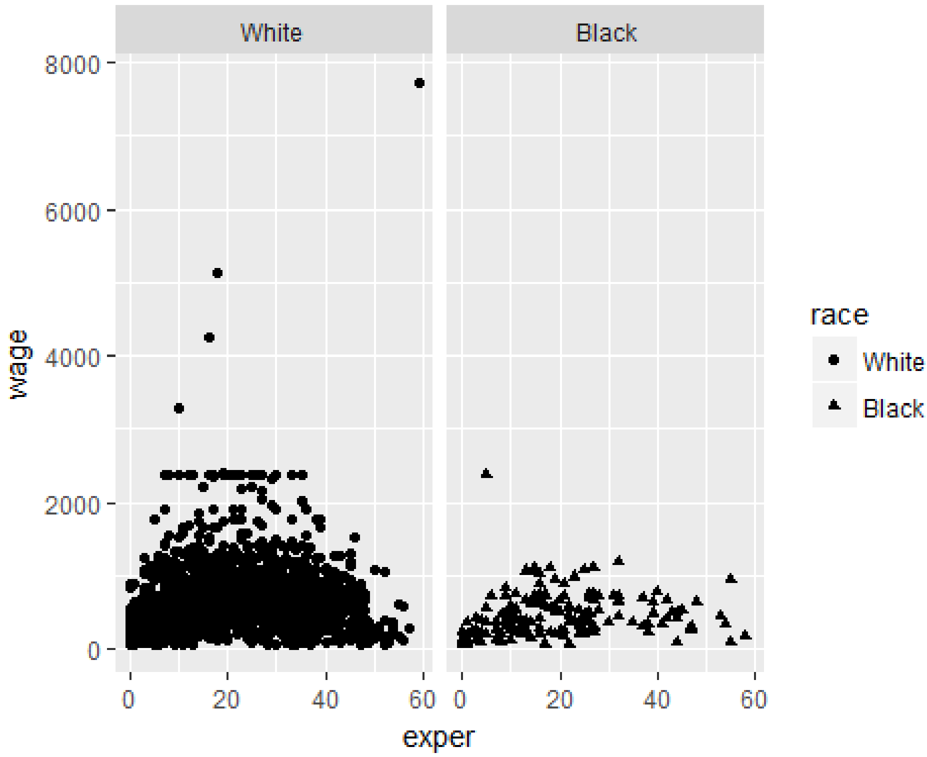



```
# Analysis:
# - There are outliers present.
# - From the density graph, it is clear that wage is positively skewed.
# - Majority of the wage lies between 50 to 2000.


# Boxplots
plot(wage ~ pt, data = uswages)
# Analysis:
# - The wages for part time are not that spread out compared to the people
who are working full time.
# - People doing full time have more wages compared to people working part
time.

plot(wage ~ region, data = uswages)
# Analysis:
# - The max wage value & the interquartile range for the people living in
we(West) is slightly more compared to the rest of the regions.
# - There are outliers present in all the regions except for mw(Middle West).

plot(wage ~ smsa, data = uswages)
# Analysis:
# - The max wage value & the interquartile range for the people living in
smsa(Standard Metropolitan Statistical Area) is slightly more compared to the
people not living in smsa.
# - There are outliers present in boxplot for the people living in
```

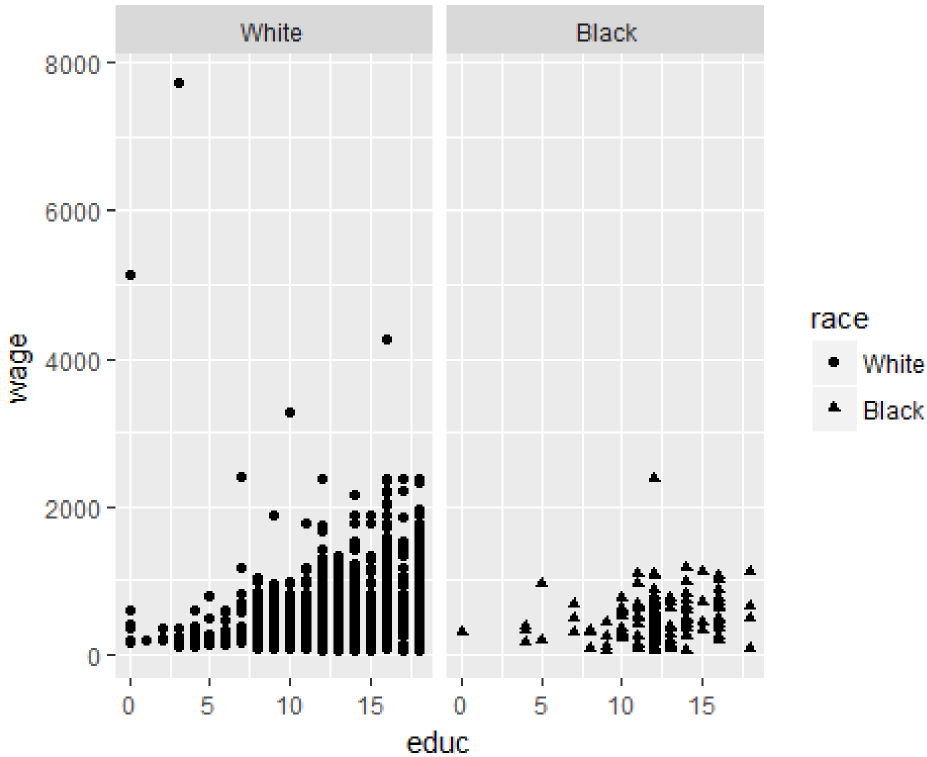
```r
smsa(Standard Metropolitan Statistical Area)

plot(wage ~ race, data = uswages)
```



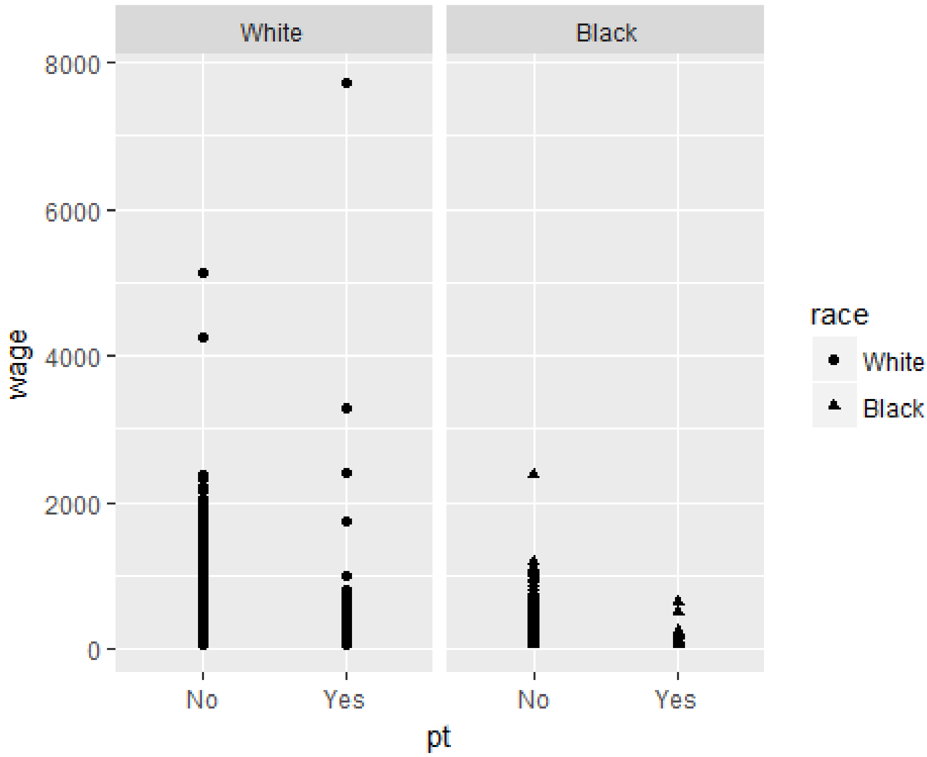

```r
# Analysis:
# - Whites earn more compared to blacks.
# - There are outliers present in whites.


# Experience vs Wage with respect to Race
ggplot(uswages,aes(x = exper, y = wage, shape = race, na.rm ='TRUE'))
+geom_point() +facet_grid(~ race)
```



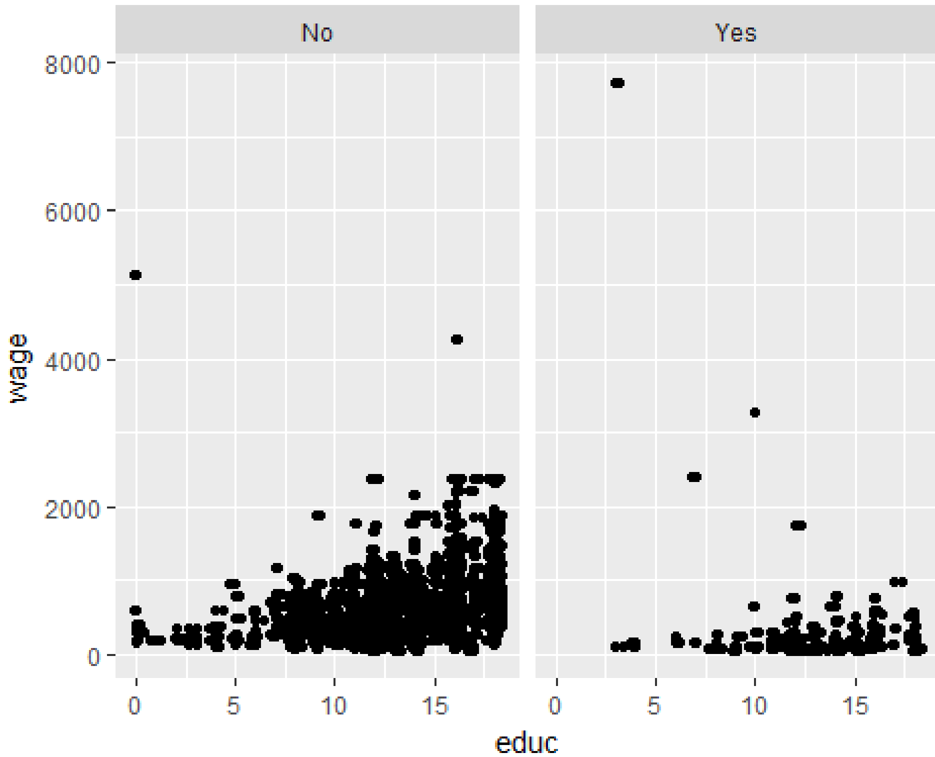

```r
# - Analysis:
# - We observe that compared to blacks there are more no of whites. Apart
from that what we observe that the wages of whites is also higher compared to
black.


# Education vs Wage with respect to Race
ggplot(uswages, aes(educ, wage, shape = race, na.rm = 'TRUE')) +geom_point()
+facet_grid(~race)
```



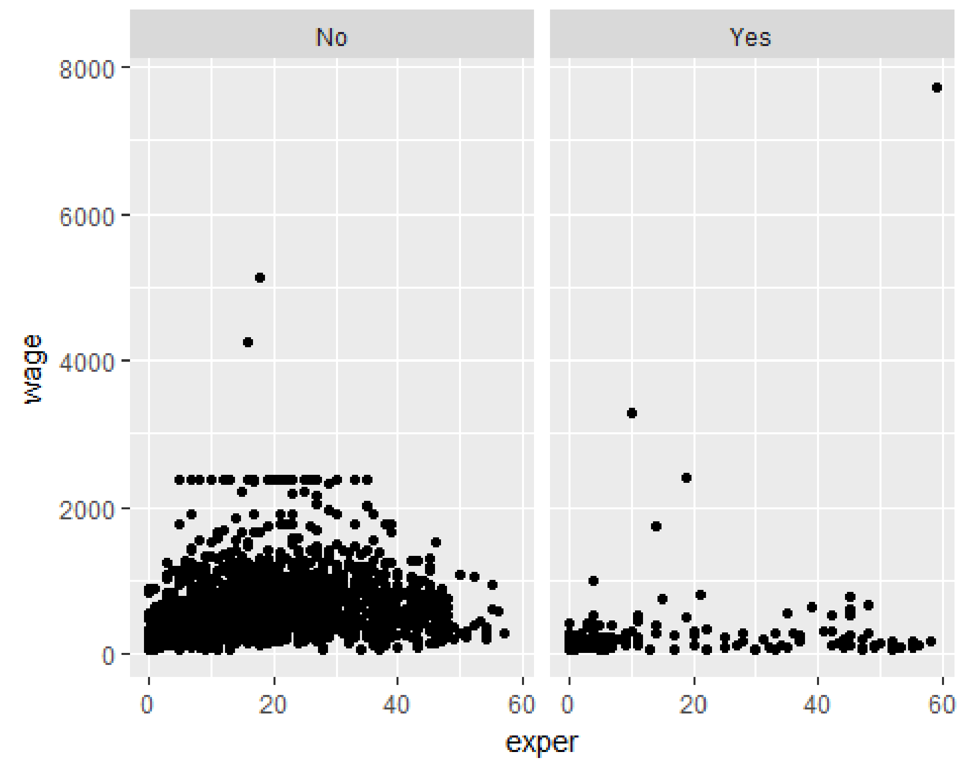

```
# Analysis:
# - The distribution of education in whites is spread out compared to blacks
and whites receive more wages.


# Part time vs Wage with respect to Race
ggplot(uswages,aes(x=pt,y= wage, shape= race,
na.rm='TRUE'))+geom_point()+facet_grid(~ race)
```



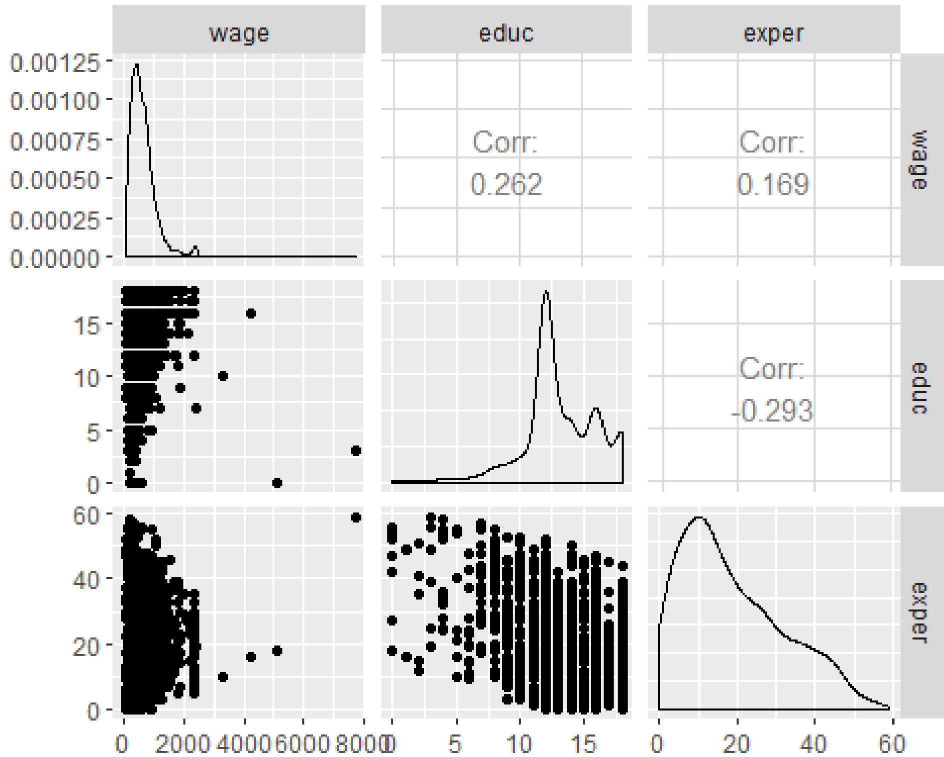

```
# Analysis:
# - Whites who dont work in Part Time are more in numbers compared to blacks
and earn more wages.
# - This statement holds good even for part time


# Education vs Wage with respect to Part time
ggplot(uswages, aes(educ, wage)) +facet_grid(~pt) +geom_point()
+geom_jitter()
```



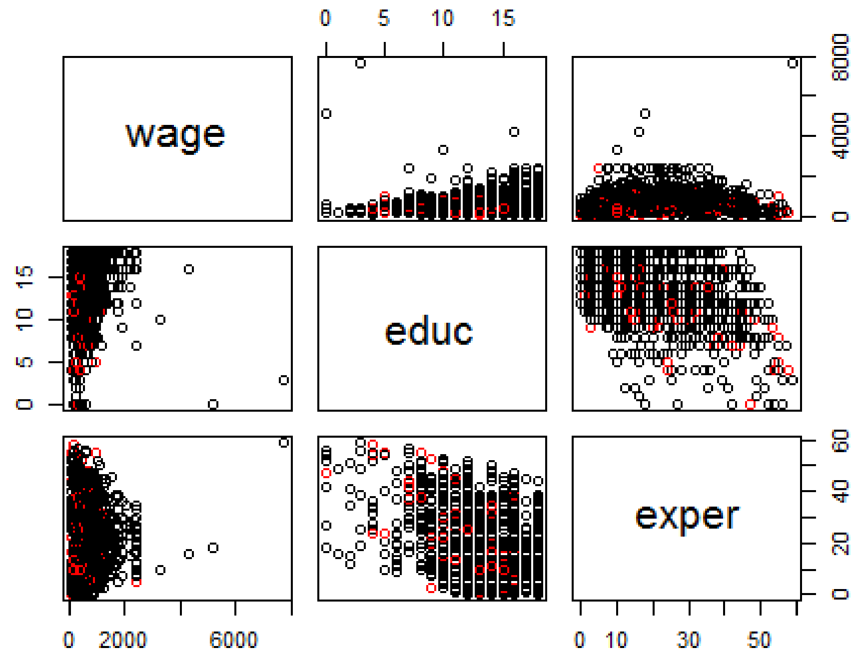

```
# Analysis:
# - People who are not working as part time are more and their wages are more
spread out compared to people who are working as part time.
# - People with no part time are distributed from 0 to 18 years of education.


# Experience vs Wage with respect to Part time
ggplot(uswages, aes(exper, wage)) +geom_point() +facet_grid(~pt)
```



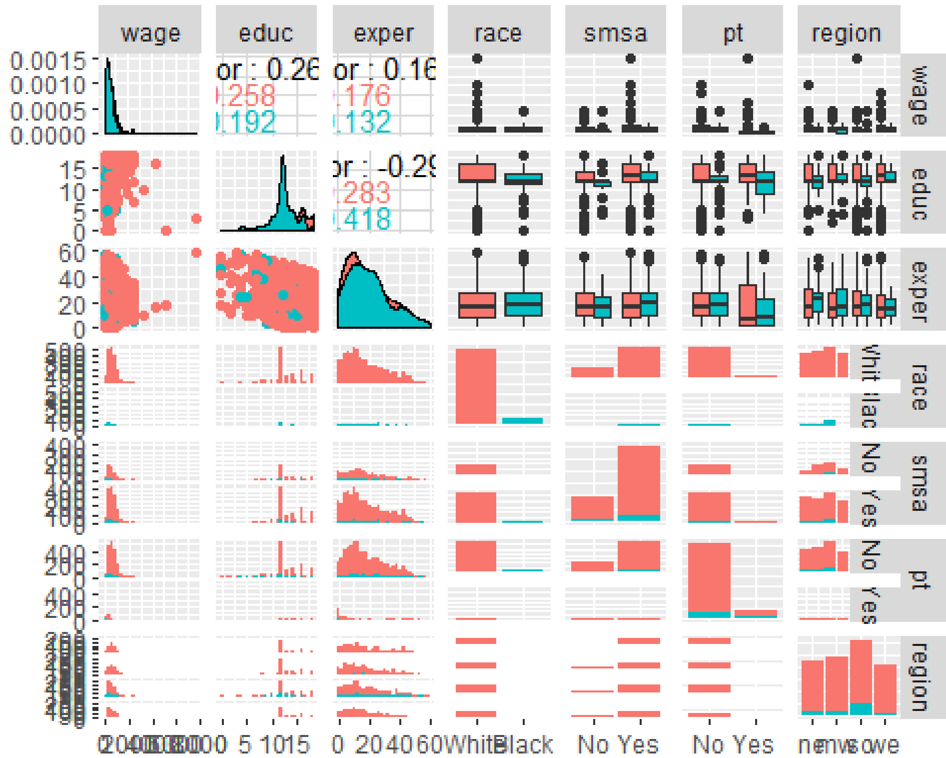

```
# - Analysis:
# - People who are not working as part time are more and their wages are more spread out compared to people who are working as part time.
# - People with no part time are distributed from 0 to 50 years of experience.


# Scatter plots for all the attributes
ggpairs(uswages, columns = c(1:3))
```




```
# Analysis:
# - Correlation between wage and educ is 0.26 which shows a weak positive linear relationship.
# - Correlation between wage and exper is 0.16 which shows a weak positive linear relationship.
# - Correlation between educ adn exper is -0.29 which shows a weak negative linear relationsip.

pairs(~ wage + educ + exper, data = uswages, col = uswages$race)
```

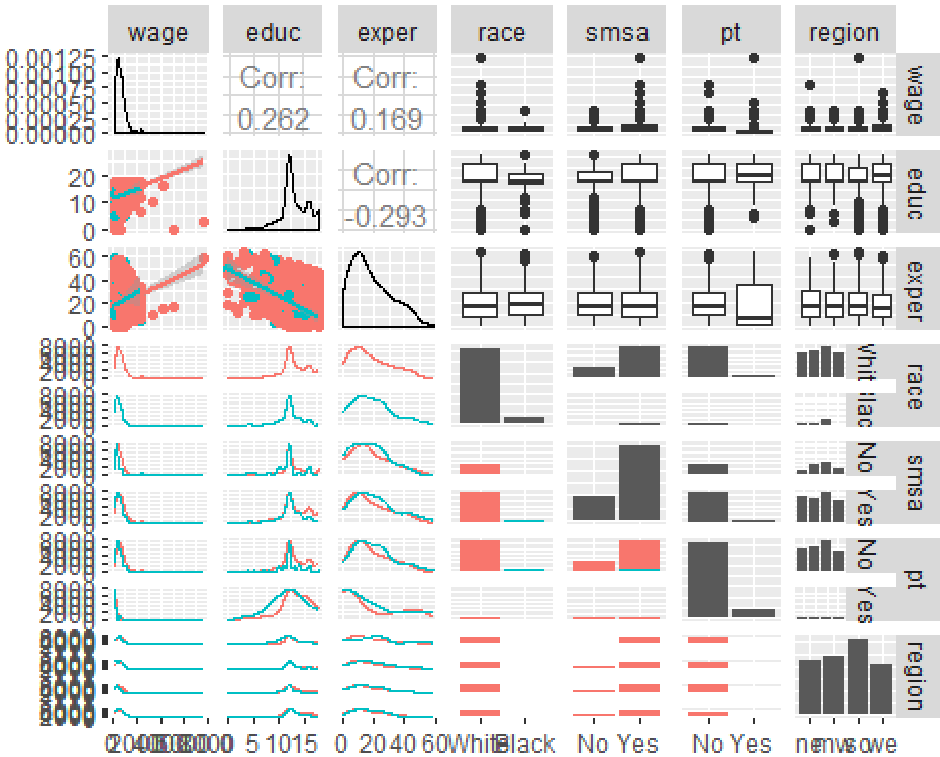



```
ggpairs(uswages, mapping = aes(colour = race))

## `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`.
```



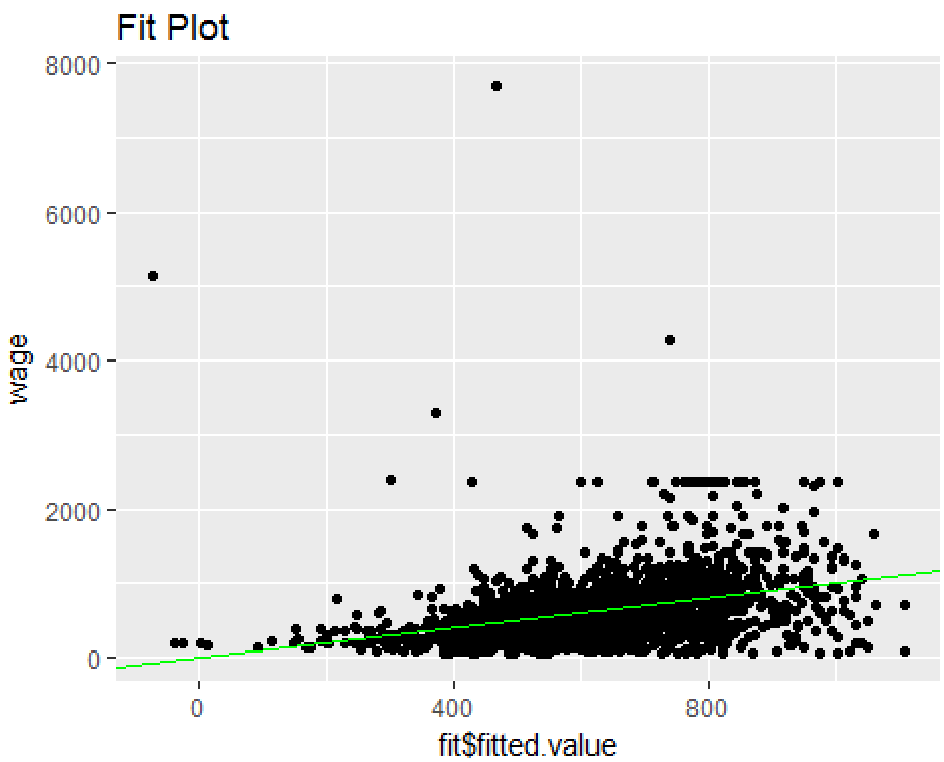

```r
# Analysis 1: Skewness
skewness(uswages$wage)

## [1] 3.833774

skewness(uswages$educ)

## [1] -0.6824895

skewness(uswages$exper)

## [1] 0.6671678

# - For wages, the graph is highly distributed and the data is positively
skewed i.e towards the right.
# - For educ, the graph is moderately distributed and the data is negatively
skewed i.e towards the left.
# - For exper, the graph is moderately distributed and the data is positively
skewed i.e towards the right.

# Analysis 2: Wages vs Factor variables
# - wage vs race - Whites earn more than the black do and the spread of wages
for whites is more.
# - wage vs smsa - The count and the wages of whites in smsa is very high
compared the blacks. The count of whites not living in smsa is high compared
to blacks but the wages are almost similar.
# - wage vs pt - The count and the wages of whites who work part-time is very
high compared the blacks.
```

```
# - wages vs region - The wages for all the people living in different
regions is almost the same.

# Analysis 3: Educ vs Factor variables
# - educ vs race - Whites are more educated than blacks.
# - educ vs smsa - Whites staying in smsa are more educated than blacks.
Whites not staying in smsa are more educated than blacks.
# - educ vs pt - There are more number of blacks who are educated and are
working as part time compared to whites. Whites not working as part time are
more educated than blacks.
# - educ vs region - In all the region Whites are more educated than blacks

# Analysis 4: Exper vs Factor variables
# - exper vs race - There are almost equal number of whites and blacks who
have same years of experience.
# - exper vs smsa - Whites living in smsa and not living in smsa an are
slightly more experienced than blacks.
# - exper vs pt - Whites working as part time have much more experience than
blacks. Whites and blacks have almost the same experience for those who are
not working as part time.
# - exper vs region - In all the region Whites are more experienced than
blacks apart from the ones staying in Middle West where the whites and
blacks have almost the same experience.


ggpairs(uswages, lower = list(continuous = "smooth", combo = "facetdensity",
mapping = aes(color = race)))
```

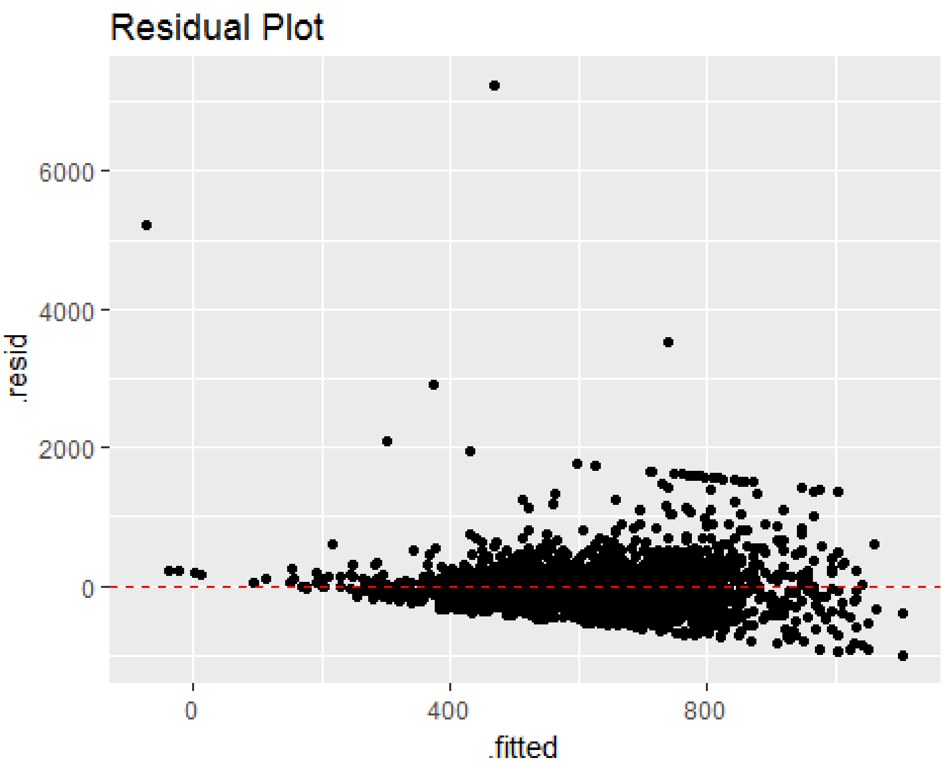



```
# Analysis:
# - The slope coefficient is based on two predictors - educ and exper.


# Fit Model
fit = lm(wage ~ educ + exper, uswages)
summary(fit)

##
## Call:
## lm(formula = wage ~ educ + exper, data = uswages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1014.7  -235.2   -52.1   150.1  7249.2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -239.1146    50.7111  -4.715 2.58e-06 ***
## educ          51.8654     3.3423  15.518  < 2e-16 ***
## exper          9.3287     0.7602  12.271  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 426.8 on 1964 degrees of freedom
## Multiple R-squared:  0.1348, Adjusted R-squared:  0.1339
## F-statistic:   153 on 2 and 1964 DF,  p-value: < 2.2e-16
```

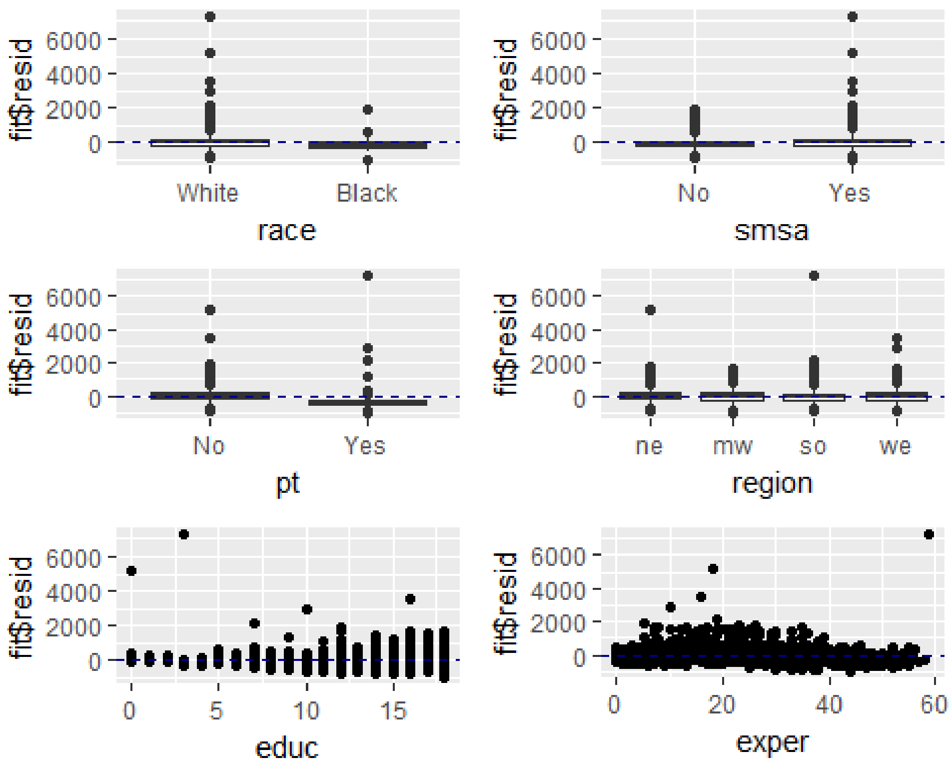
```r
# Analysis:
# - The slope coefficient are positive. So with increase in education and
experience there will be increase in wages.

deviance = deviance(fit)
deviance

## [1] 357763806

y = uswages$wage
totalss = sum((y-mean(y))^2)
totalss

## [1] 413500833

1 - deviance/totalss

## [1] 0.134793

summary(fit)$r.square

## [1] 0.134793

# Analysis:
# - The model is not a good fit because the value of R^2 is 0.13 which is
much less than 1.
# - 13% of the variance in the response variable can be explained by the
explanatory variables. The remaining 87% can be attributed to unknown,
lurking variables or inherent variability

c(summary(fit)$r.square, cor(fitted.values(fit), uswages$wage)^2)

## [1] 0.134793 0.134793

# Analysis:
# - The pearson correlation is equal to model summary.


# Fit Plot
qplot(fit$fitted.value, wage, data = uswages) +geom_abline(intercept = 0,
slope = 1, color="green") +ggtitle("Fit Plot")
```



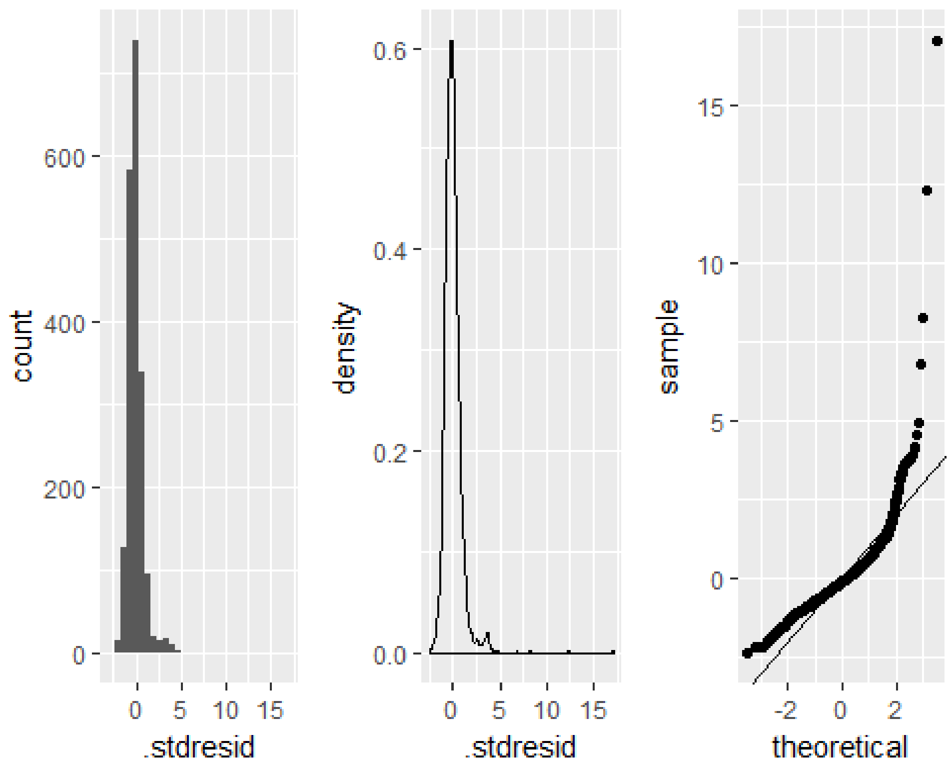

```
# Analysis:
# - It is not a good fit plot

# Residual Plot
ggplot(fit, aes(.fitted, .resid)) +geom_point() +geom_hline(yintercept = 0,
color = "red", linetype = "dashed") + ggtitle("Residual Plot")
```




```r
# Analysis:
# - The points in a residual plot are randomly dispersed around the
horizontal axis, a linear regression model is appropriate for the data.


# Exploring model structure
cor(fit$resid, uswages$wage)

## [1] 0.930165

# Analysis:
# - As the correlation is high, it indicates that there is some trouble with
our model.

plot1 = qplot(race, fit$resid, geom = "boxplot", data = uswages)
+geom_hline(yintercept = 0, color = "dark blue", linetype = "dashed")
plot2 = qplot(smsa, fit$resid, geom = "boxplot", data = uswages)
+geom_hline(yintercept = 0, color = "dark blue", linetype = "dashed")
plot3 = qplot(pt, fit$resid, geom = "boxplot", data = uswages)
+geom_hline(yintercept = 0, color = "dark blue", linetype = "dashed")
plot4 = qplot(region, fit$resid, geom = "boxplot", data = uswages)
+geom_hline(yintercept = 0, color = "dark blue", linetype = "dashed")
plot5 = qplot(educ, fit$resid, data = uswages) +geom_hline(yintercept = 0,
color = "dark blue", linetype = "dashed")
plot6 = qplot(exper, fit$resid, data = uswages) +geom_hline(yintercept = 0,
color = "dark blue", linetype = "dashed")
grid.arrange(plot1, plot2, plot3, plot4, plot5, plot6, nrow = 3)
```




```
# Analysis:
# - We see prounanced patters indicating we do not need to include square of the predictors or other transforms of the predictors.


# Normailty of the Residual
mod = fortify(fit)
plot7 = qplot(.stdresid, data = mod, geom = "histogram")
plot8 = qplot(.stdresid, data = mod, geom = "density")
plot9 = qplot(sample = .stdresid, data = mod, geom = "qq") +geom_abline()
grid.arrange(plot7, plot8, plot9, nrow = 1)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```r
# Analysis:
# - We see that the residual do not look as though they come from a normal
distribution.

g = lm(log(wage) ~ educ + exper + race + smsa + pt + region, data = uswages)
confint(g, level = 0.95)
```

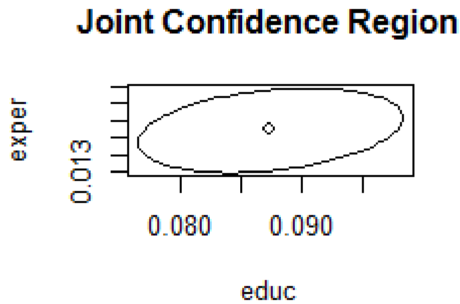
```
##                   2.5 %      97.5 %
## (Intercept)  4.56088834  4.85919020
## educ         0.07870075  0.09632893
## exper        0.01349117  0.01747522
## raceBlack   -0.31082788 -0.11971899
## smsaYes      0.11855745  0.23692397
## ptYes       -1.15830653 -0.97652860
## regionmw    -0.06520300  0.08184280
## regionso    -0.06041461  0.08057549
## regionwe    -0.02948699  0.12317864
```

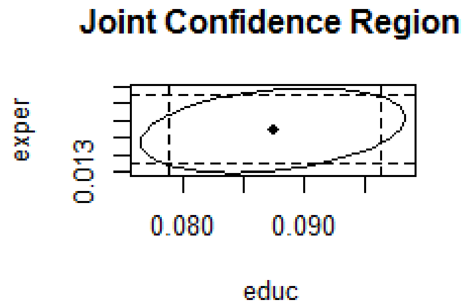
```r
# Compare Model
g1 = lm(log(wage) ~ educ + exper + race + smsa + pt, data = uswages)
confint(g1)
```

```
##                   2.5 %      97.5 %
## (Intercept)  4.59090124  4.86663286
## educ         0.07860326  0.09620455
## exper        0.01343477  0.01741146
## raceBlack   -0.31310266 -0.12537317
```

```
## smsaYes      0.11803316  0.23542167
## ptYes       -1.15785402 -0.97626013

tab1 = anova(g1,g)
tab1

## Analysis of Variance Table
##
## Model 1: log(wage) ~ educ + exper + race + smsa + pt
## Model 2: log(wage) ~ educ + exper + race + smsa + pt + region
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   1961 633.62
## 2   1958 633.06  3   0.55445 0.5716 0.6337

# P-value and F-value Calculation
g.bg = lm(log(wage) ~ educ + exper + race + smsa + pt + region, data =
uswages)
g.sm = lm(log(wage) ~ educ + exper + race + smsa + pt, data = uswages)
sse.sm = deviance(g.sm)
df.sm = df.residual(g.sm)
sse.bg = deviance(g.bg)
df.bg = df.residual(g.bg)
mse.prt = (sse.sm-sse.bg)/(df.sm-df.bg)
mse.bg = sse.bg/df.bg
f.ratio = mse.prt/mse.bg
f.ratio

## [1] 0.5716147

p.value = pf(f.ratio, df.sm-df.bg, df.bg, lower.tail=FALSE)
p.value

## [1] 0.6337081

# Analysis:
# - The P-value is 0.634 which is greater than 0.05.
# - Therefore, model 2 is better.


# Joint Confidence Region
install.packages("ellipse", repos = "http://cran.us.r-project.org", type =
"source")

## Installing package into 'C:/Users/Shraddha Somani/Documents/R/win-
library/3.3'
## (as 'lib' is unspecified)

library(ellipse)
plot(ellipse(g1, c("educ", "exper")), type = "l", main = "Joint Confidence
Region")
points(0,0)
points(coef(g1)["educ"], coef(g)["exper"])
```

```
g2 <- lm(log(wage) ~ race + smsa + pt, data = uswages)
plot(ellipse(g1, c("educ", "exper")), type = "l", main = "Joint Confidence
Region")
points(0,0)
points(coef(g)["educ"], coef(g)["exper"], pch=18)
abline(v=confint(g)["educ",], lty=2)
abline(h=confint(g)["exper",], lty=2)
compareg2g1 <- anova(g2, g1)
compareg2g1

## Analysis of Variance Table
##
## Model 1: log(wage) ~ race + smsa + pt
## Model 2: log(wage) ~ educ + exper + race + smsa + pt
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1   1963 788.00
## 2   1961 633.62  2    154.38 238.9 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Analysis:
# - If the p-value is less than or equal to the alpha (p < .05), then we
reject the null hypothesis, and we say the result is statistically
significant.
# - If the p-value is greater than alpha (p > .05), then we fail to reject
the null hypothesis, and we say that the result is statistically
nonsignificant (n.s.).
# - The F-Ratio 238.9 is big and since the p-value 2.2e-16 is much less than
0.05, we reject the null hypothesis H0 : ??educ = ??exper = 0.


# Prediction
g1 = lm(log(wage) ~ educ + exper + race + smsa + pt, data = uswages)
x0 = data.frame(educ = 12, exper = 5, race = "White", smsa = "Yes", pt =
"No", stringsAsFactors = FALSE)
predict(g1, x0, level = 0.95, interval = "confidence")

##        fit      lwr      upr
## 1 6.031457 5.986455 6.076459

x0 <- rbind(x0, data.frame(educ = 12, exper = 5, race = "Black", smsa =
"Yes", pt = "No"))
predict(g1, x0, level = 0.95, interval = "confidence")

##        fit      lwr      upr
## 1 6.031457 5.986455 6.076459
## 2 5.812219 5.716405 5.908033
```


Joint Confidence Region
exper
0.013
0.080
0.090
educ


Joint Confidence Region
exper
0.013
0.080
0.090
educ