Multiple Regression Model

Thach Pham

15/11/2018

The experiment uses R Core Team (2018) and some packages developed by Wickham et al. (2018), Allaire et al. (2018), Schloerke et al. (2018), Hothorn et al. (2018), Fox, Weisberg, and Price (2018), Hlavac (2018), Wickham (2017), and Shea (2018).

```
library(wooldridge)
library(stargazer)
library(car)
library(lmtest)
library(tidyverse)
library(GGally)
# For ggiraphExtra require: devtools, libcairo2-dev and libudunits2-dev
# devtools::install_github("cardiomoon/ggiraphExtra")
# library(ggiraph)
# library(ggiraphExtra)
```

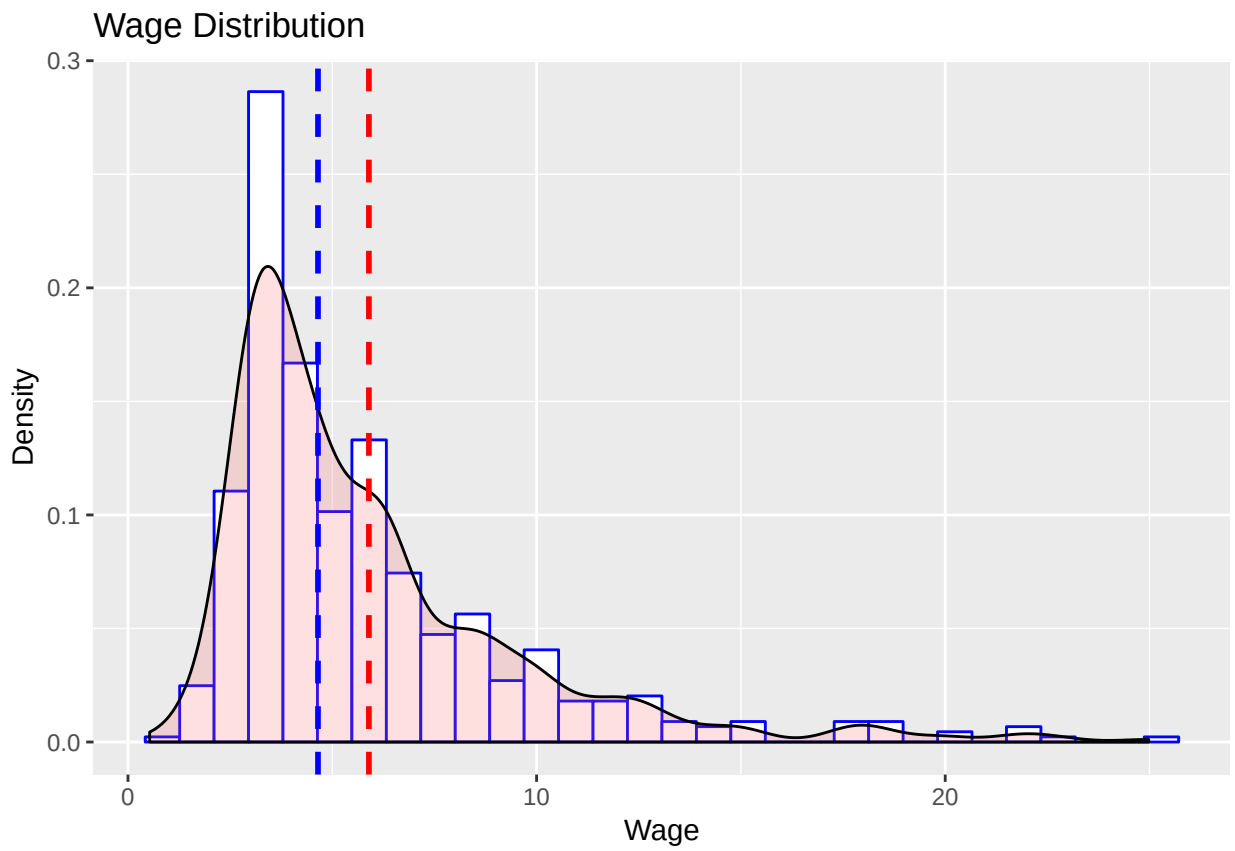
Descriptive Statistics of wage1 data

```
data(wage1)
desc.wage1 <- select(wage1, c(wage, educ, exper, tenure, lwage))
stargazer(desc.wage1, type = "text", style = "qje", title = "Descriptive Statistics",
          digits = 4, out = "MRO/Descriptive Statistics.doc")
```

```
##
## Descriptive Statistics
## ---------------------------------------------------------------
## Statistic  N   Mean   St. Dev.   Min   Pctl(25) Pctl(75)   Max
## ===============================================================
## wage      526 5.8961   3.6931  0.5300   3.3300   6.8800  24.9800
## educ      526 12.5627  2.7690     0       12       14      18
## exper     526 17.0171 13.5722     1       5        26      51
## tenure    526 5.1046   7.2245     0       0        7       44
## lwage     526 1.6233   0.5315  -0.6349  1.2030   1.9286  3.2181
## ---------------------------------------------------------------
```

Distribution of Wage

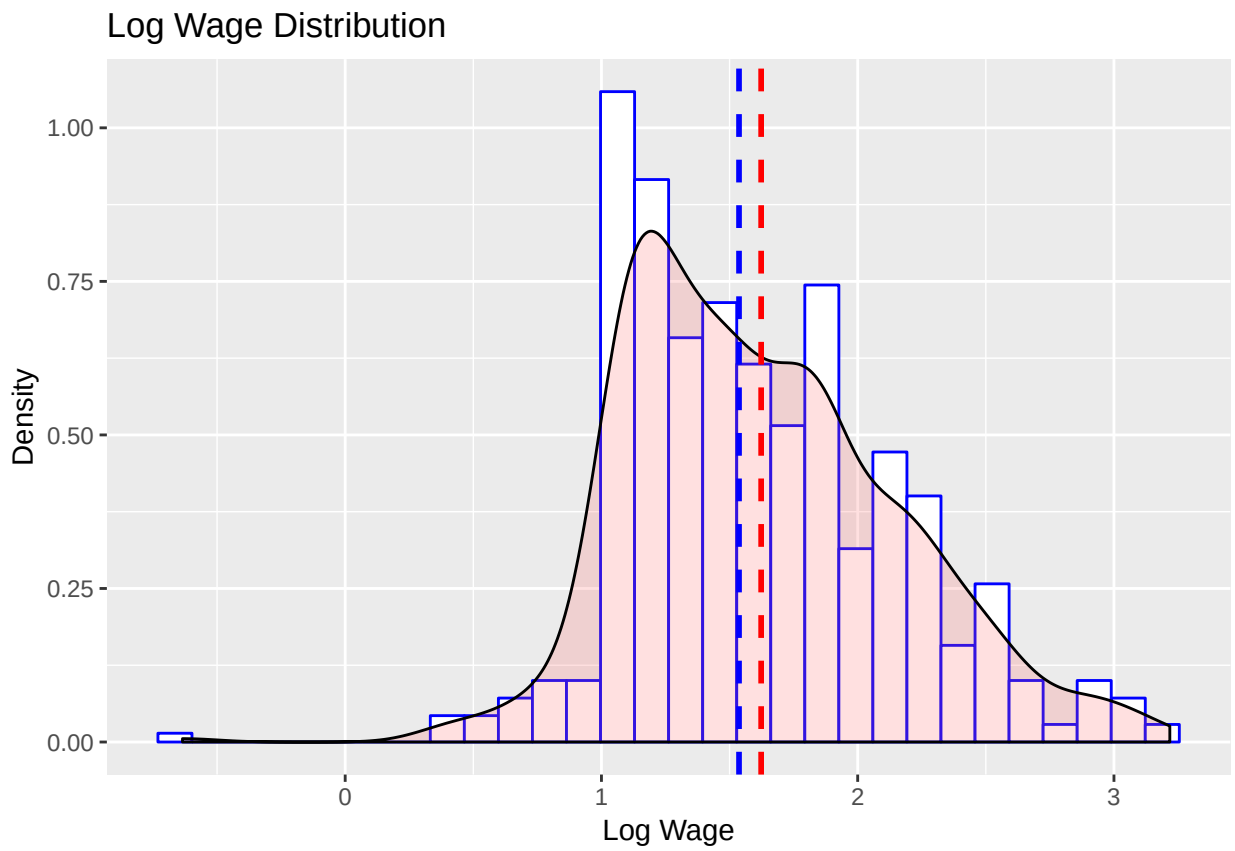
```
(wage.dist <- ggplot(desc.wage1, aes(wage)) +
  geom_histogram(aes(y = ..density..), bindwidth = 0.5, colour = "blue", fill = "white") +
  geom_density(alpha = 0.2, fill = "#FF6666") +
  labs(x = "Wage", y = "Density", title = "Wage Distribution") +
  geom_vline(aes(xintercept = mean(wage)), color = "red", linetype = "dashed", size = 1) +
  geom_vline(aes(xintercept = median(wage)), color = "blue", linetype ="dashed", size = 1))
```




```
ggsave("MRO/Wage Distribution.png", wage.dist)
```

Distribution of Log Wage

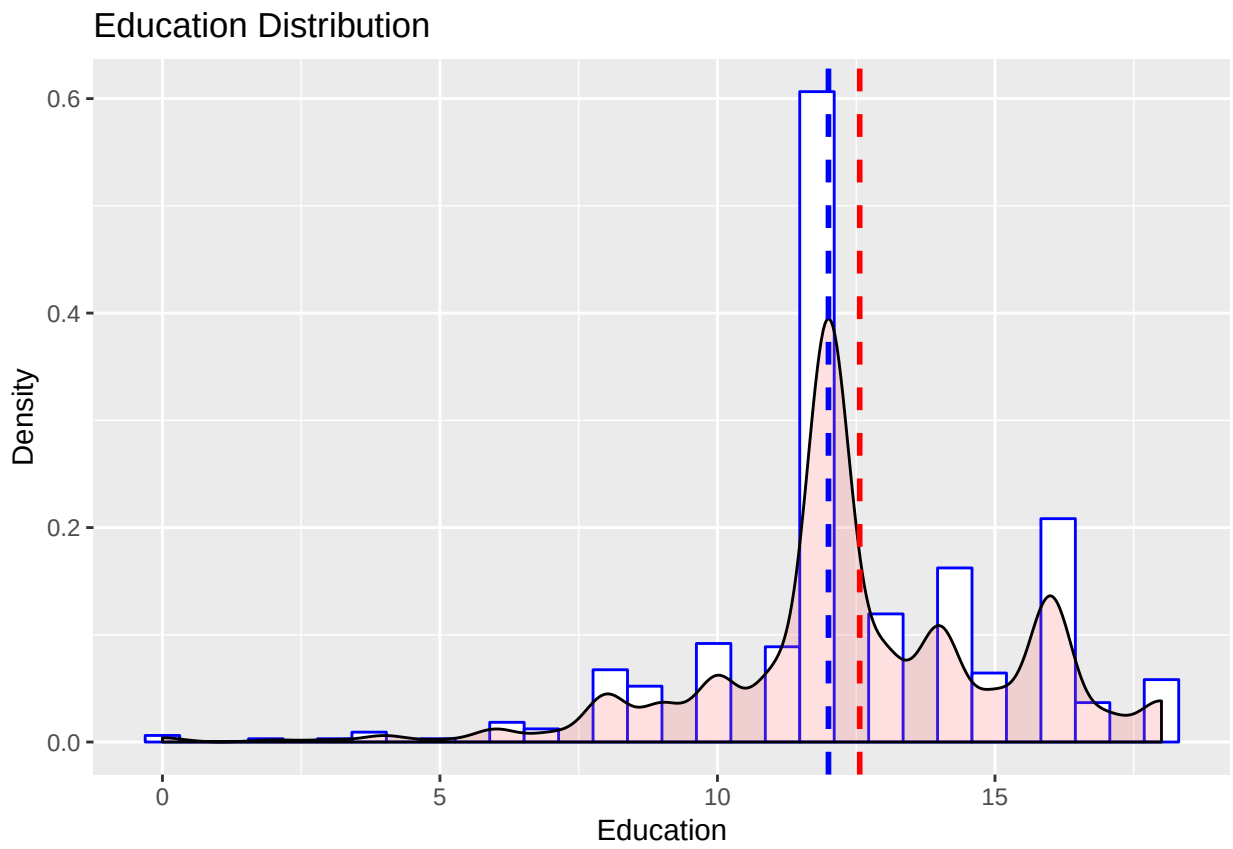
```
(lwage.dist <- ggplot(desc.wage1, aes(lwage)) +
  geom_histogram(aes(y = ..density..), bindwidth = 0.5, colour = "blue", fill = "white") +
  geom_density(alpha = 0.2, fill = "#FF6666") +
  labs(x = "Log Wage", y = "Density", title = "Log Wage Distribution") +
  geom_vline(aes(xintercept = mean(lwage)), color = "red", linetype = "dashed", size = 1) +
  geom_vline(aes(xintercept = median(lwage)), color = "blue", linetype ="dashed", size = 1))
```




```
ggsave("MRO/Log Wage Distribution.png", lwage.dist)
```

Distribution of Education

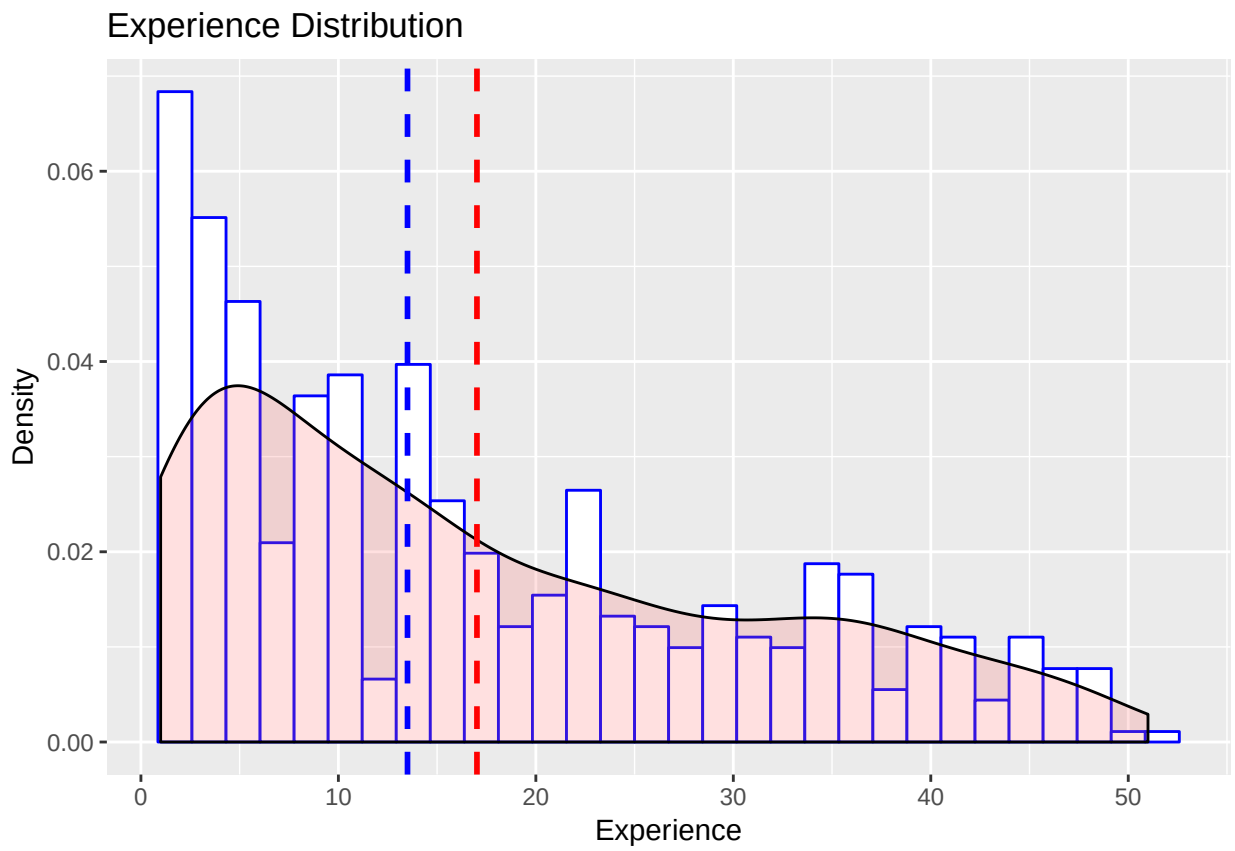
```
(educ.dist <- ggplot(desc.wage1, aes(educ)) +
  geom_histogram(aes(y = ..density..), bindwidth = 0.5, colour = "blue", fill = "white") +
  geom_density(alpha = 0.2, fill = "#FF6666") +
  labs(x = "Education", y = "Density", title = "Education Distribution") +
  geom_vline(aes(xintercept = mean(educ)), color = "red", linetype = "dashed", size = 1) +
  geom_vline(aes(xintercept = median(educ)), color = "blue", linetype ="dashed", size = 1))
```




```
ggsave("MRO/Education Distribution.png", educ.dist)
```

Distribution of Experience

```
(exper.dist <- ggplot(desc.wage1, aes(exper)) +
  geom_histogram(aes(y = ..density..), bindwidth = 0.5, colour = "blue", fill = "white") +
  geom_density(alpha = 0.2, fill = "#FF6666") +
  labs(x = "Experience", y = "Density", title = "Experience Distribution") +
  geom_vline(aes(xintercept = mean(exper)), color = "red", linetype = "dashed", size = 1) +
  geom_vline(aes(xintercept = median(exper)), color = "blue", linetype ="dashed", size = 1))
```



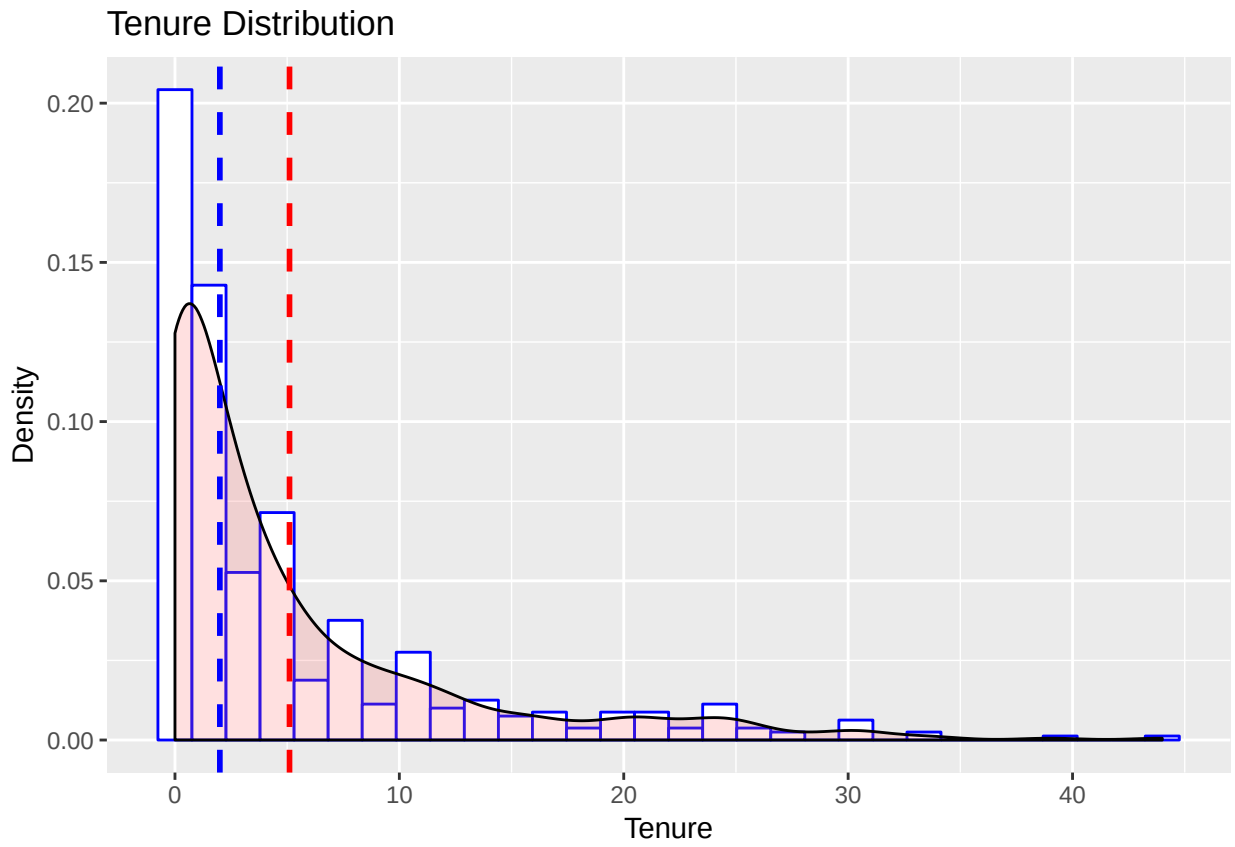

```
ggsave("MRO/Experience Distribution.png", exper.dist)
```

Distribution of Tenure

```
(tenure.dist <- ggplot(desc.wage1, aes(tenure)) +
  geom_histogram(aes(y = ..density..), bindwidth = 0.5, colour = "blue", fill = "white") +
  geom_density(alpha = 0.2, fill = "#FF6666") +
  labs(x = "Tenure", y = "Density", title = "Tenure Distribution") +
  geom_vline(aes(xintercept = mean(tenure)), color = "red", linetype = "dashed", size = 1) +
  geom_vline(aes(xintercept = median(tenure)), color = "blue", linetype ="dashed", size = 1))
```

```
## Warning: Ignoring unknown parameters: bindwidth
```

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




```
ggsave("MRO/Tenure Distribution.png", tenure.dist)
```

```
## Saving 6.5 x 4.5 in image
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Covariance and Correlation Matrix

```
cov(desc.wage1)
```

```
##              wage        educ       exper    tenure     lwage
## wage   13.638884   4.1508640   5.6590763  9.255208 1.8394674
## educ    4.150864   7.6674851 -11.2572660 -1.123715 0.6344412
## exper   5.659076 -11.2572660 184.2035162 48.956303 0.8034574
## tenure  9.255208  -1.1237154  48.9563027 52.192855 1.2500910
## lwage   1.839467   0.6344412   0.8034574  1.250091 0.2825329
```

Correlation Matrix

```
cor.graph <- ggpairs(desc.wage1, title = "Correlation Matrix")
ggsave("MRO/Correltaion Matrix.png", cor.graph)
```

Model 1

```
model1 <- lm(wage ~ educ + exper + tenure, data = desc.wage1)
stargazer(model1, type = "text", title = "lm(wage ~ educ + exper + tenure)",
          style = "qje", out = "MRO/model1.doc")
```

```
##
## lm(wage ~ educ + exper + tenure)
## ===============================================================
##                                         wage
## ---------------------------------------------------------------
## educ                                   0.599***
##                                        (0.051)
##
## exper                                   0.022*
##                                        (0.012)
##
## tenure                                 0.169***
##                                        (0.022)
##
## Constant                              -2.873***
##                                        (0.729)
##
## N                                        526
## R2                                       0.306
## Adjusted R2                              0.302
## Residual Std. Error                3.084 (df = 522)
## F Statistic                   76.873*** (df = 3; 522)
## ===============================================================
## Notes:                ***Significant at the 1 percent level.
##                        **Significant at the 5 percent level.
##                        *Significant at the 10 percent level.
```

```
#ggPredict(model1, se = T, interactive = T)
```

Multicollinearity of Model 1

```
vif(model1)
```

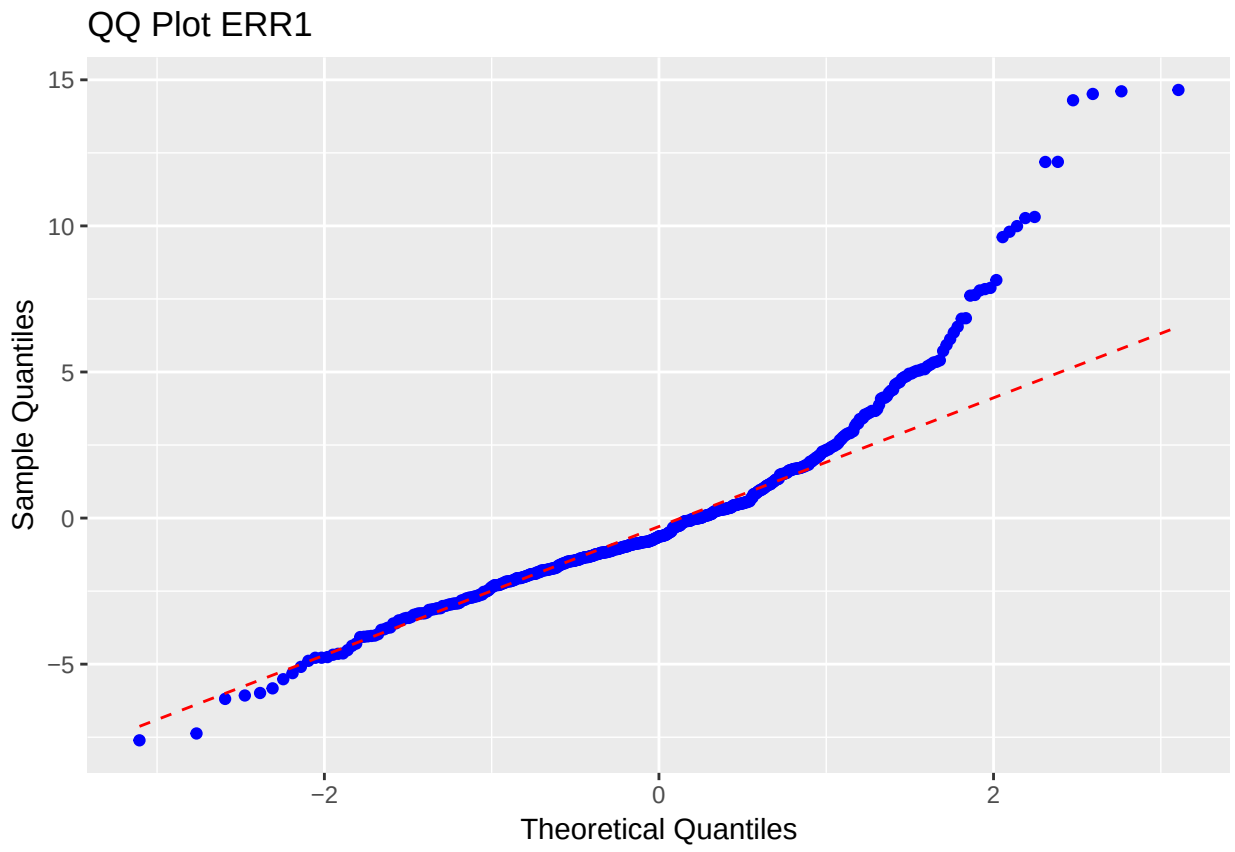
```
##     educ    exper   tenure
## 1.112771 1.477618 1.349296
```

Checking Normal Distribution of Error Term (Model 1)

```
shapiro.test(residuals(model1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model1)
## W = 0.89317, p-value < 2.2e-16
```

```
(error.model1 <- ggplot(model1, aes(sample = model1$residuals)) +
  stat_qq(col = "blue") +
  stat_qq_line(col = "red", lty = 2) +
  labs(x = "Theoretical Quantiles", y = "Sample Quantiles", title = "QQ Plot ERR1"))
```




```
ggsave("MRO/QQ Plot ERR1.png", error.model1)
```

```
## Saving 6.5 x 4.5 in image
```

Test of Heteroskedasticity of Model 1

```r
bptest(model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 43.096, df = 3, p-value = 2.349e-09
```

Robust coefficients using White's robust SE (Model1)

```r
coeftest(model1, vcov = hccm(model1, type = "hc0"))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -2.872735   0.804340 -3.5715 0.0003877 ***
## educ         0.598965   0.060781  9.8544 < 2.2e-16 ***
## exper        0.022340   0.010515  2.1246 0.0340881 *
## tenure       0.169269   0.029167  5.8035 1.128e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model 2

```r
model2 <- lm(lwage ~ educ + exper + tenure, data = desc.wage1)
stargazer(model2, type = "text", title = "lm(lwage ~ educ + exper + tenure)",
          style = "qje", out = "MRO/model2.doc")
```

```
##
## lm(lwage ~ educ + exper + tenure)
## ================================================
##                                  lwage
## ------------------------------------------------
## educ                            0.092***
##                                 (0.007)
##
## exper                           0.004**
##                                 (0.002)
##
## tenure                          0.022***
##                                 (0.003)
##
## Constant                        0.284***
##                                 (0.104)
##
```

```
## N                                   526
## R2                                0.316
## Adjusted R2                       0.312
## Residual Std. Error          0.441 (df = 522)
## F Statistic                80.391*** (df = 3; 522)
## ============================================================
## Notes:              ***Significant at the 1 percent level.
##                      **Significant at the 5 percent level.
##                      *Significant at the 10 percent level.
```

```
#ggPredict(model2, se = T, interactive = T)
```

Multicollinearity of Model 2

```
vif(model2)
```

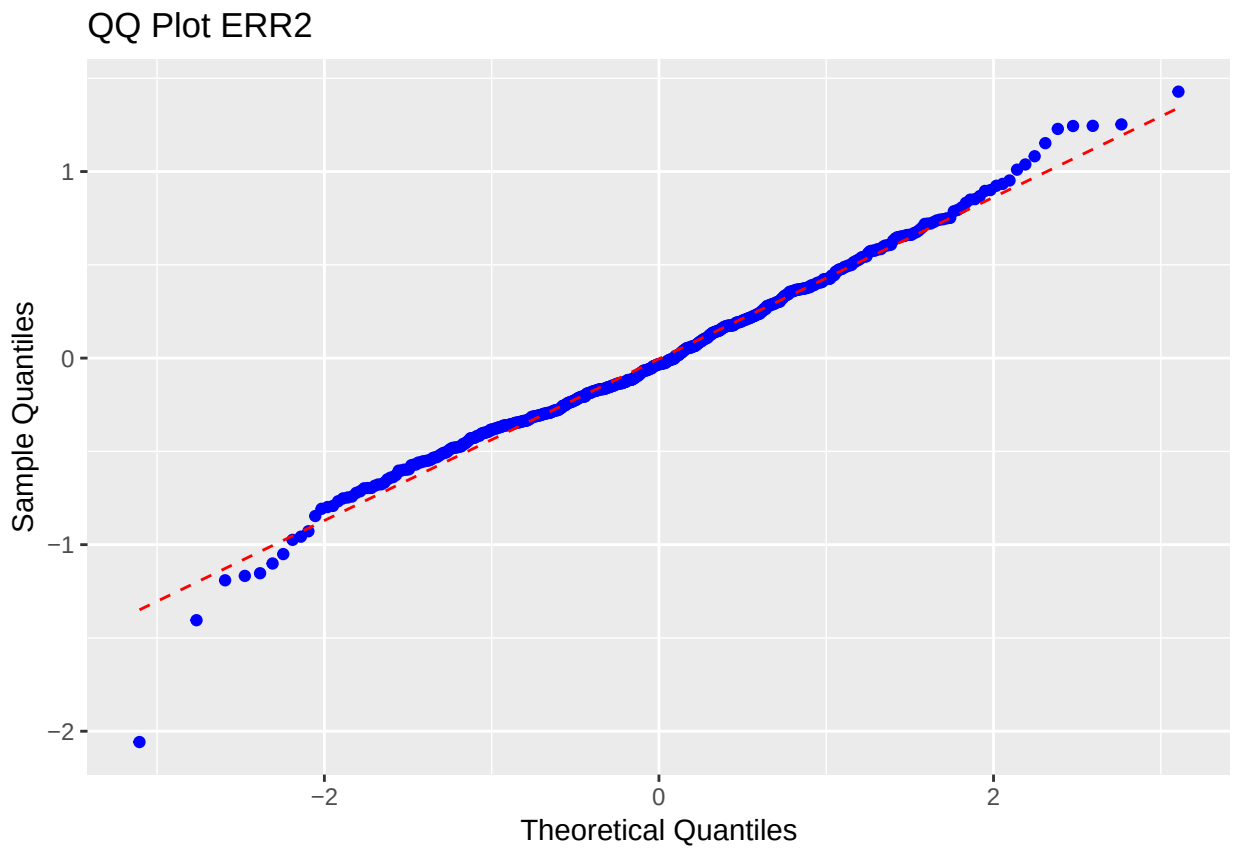
```
##     educ    exper   tenure
## 1.112771 1.477618 1.349296
```

Checking Normal Distribution of Error Term (Model 2)

```
shapiro.test(residuals(model2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model2)
## W = 0.98946, p-value = 0.000787
```

```
(error.model2 <- ggplot(model2, aes(sample = model2$residuals)) +
  stat_qq(col = "blue") +
  stat_qq_line(col = "red", lty = 2) +
  labs(x = "Theoretical Quantiles", y = "Sample Quantiles", title = "QQ Plot ERR2"))
```




```
ggsave("MRO/QQ Plot ERR2.png", error.model2)
```

```
## Saving 6.5 x 4.5 in image
```

Test of Heteroskedasticity of Model 2

```
bptest(model2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 10.761, df = 3, p-value = 0.01309
```

Robust coefficients using White's robust SE (Model 2)

```
coeftest(model2, vcov = hccm(model2, type = "hc0"))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 0.2843595  0.1112813  2.5553   0.01089 *
## educ        0.0920290  0.0078910 11.6625 < 2.2e-16 ***
## exper       0.0041211  0.0017392  2.3695   0.01817 *
## tenure      0.0220672  0.0037676  5.8571 8.343e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Reference

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, and Winston Chang. 2018. *Rmarkdown: Dynamic Documents for R*. https://CRAN.R-project.org/package=rmarkdown.

Fox, John, Sanford Weisberg, and Brad Price. 2018. *Car: Companion to Applied Regression*. https://CRAN.R-project.org/package=car.

Hlavac, Marek. 2018. *Stargazer: Well-Formatted Regression and Summary Statistics Tables*. https://CRAN.R-project.org/package=stargazer.

Hothorn, Torsten, Achim Zeileis, Richard W. Farebrother, and Clint Cummins. 2018. *Lmtest: Testing Linear Regression Models*. https://CRAN.R-project.org/package=lmtest.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Schloerke, Barret, Jason Crowley, Di Cook, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg, and Joseph Larmarange. 2018. *GGally: Extension to 'Ggplot2'*. https://CRAN.R-project.org/package=GGally.

Shea, Justin M. 2018. *Wooldridge: 111 Data Sets from "Introductory Econometrics: A Modern Approach, 6e" by Jeffrey M. Wooldridge*. https://CRAN.R-project.org/package=wooldridge.

Wickham, Hadley. 2017. *Tidyverse: Easily Install and Load the 'Tidyverse'*. https://CRAN.R-project.org/package=tidyverse.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, and Kara Woo. 2018. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.