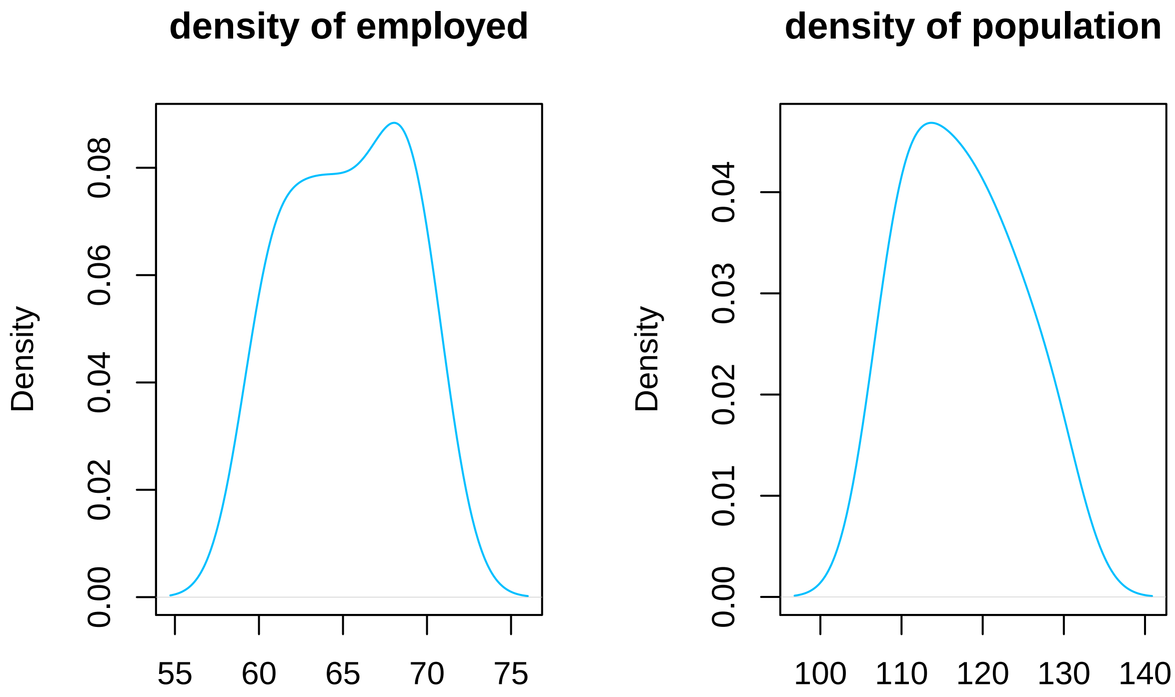
Lab 6

Hong Xu

October 19, 2015

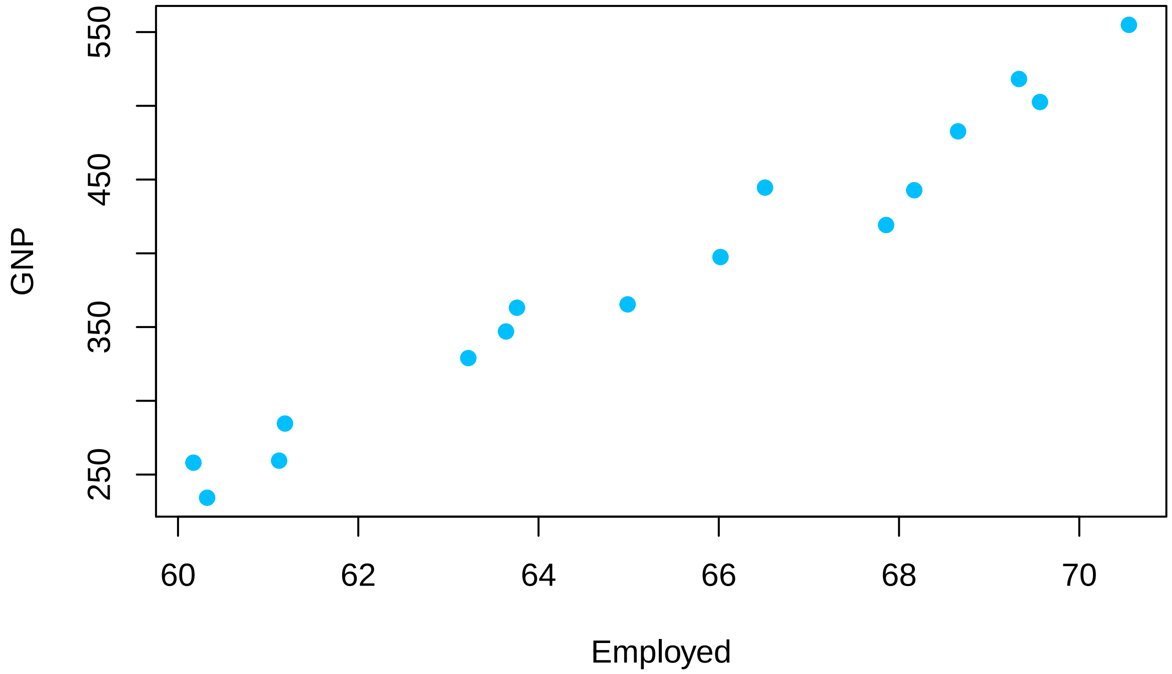
Task 1

```
## Longley Data ##
# univarite density estimate
old.par <- par(mfrow = c(1,2))
plot(density((longley$Employed)),
     col = "deepskyblue",
     xlab = "", main = "density of employed")
plot(density(longley$Population),
     col = "deepskyblue",
     xlab = "", main = "density of population")
```

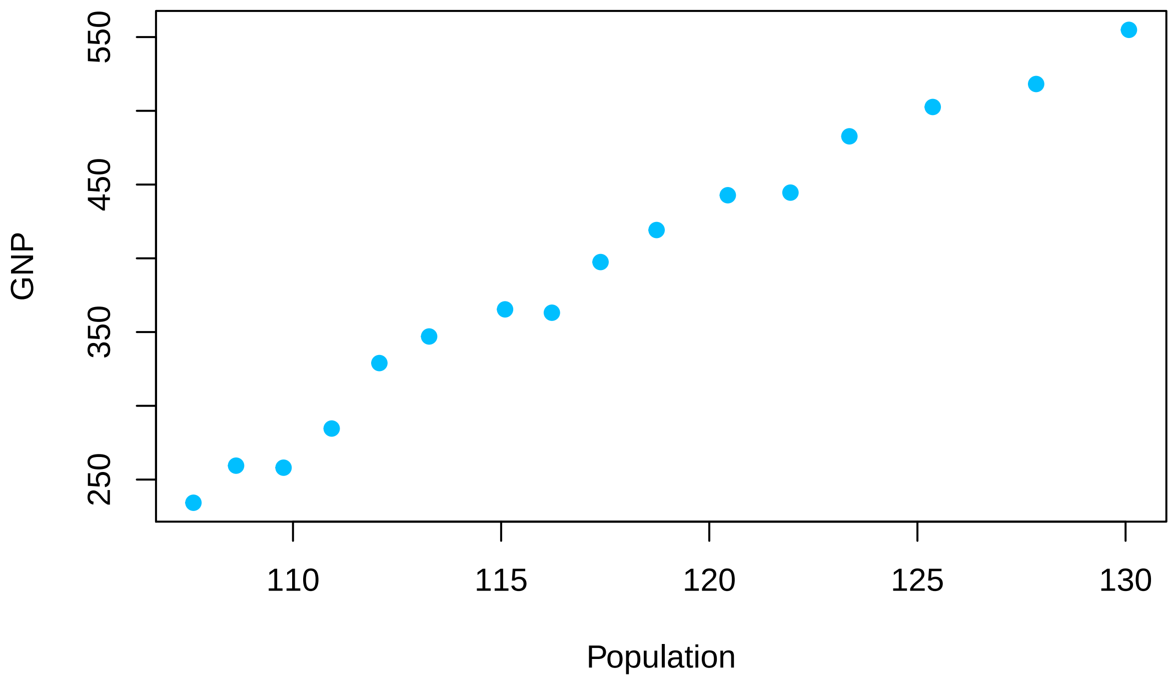


I started to see the distribution of both variables resemble the normal distribution.

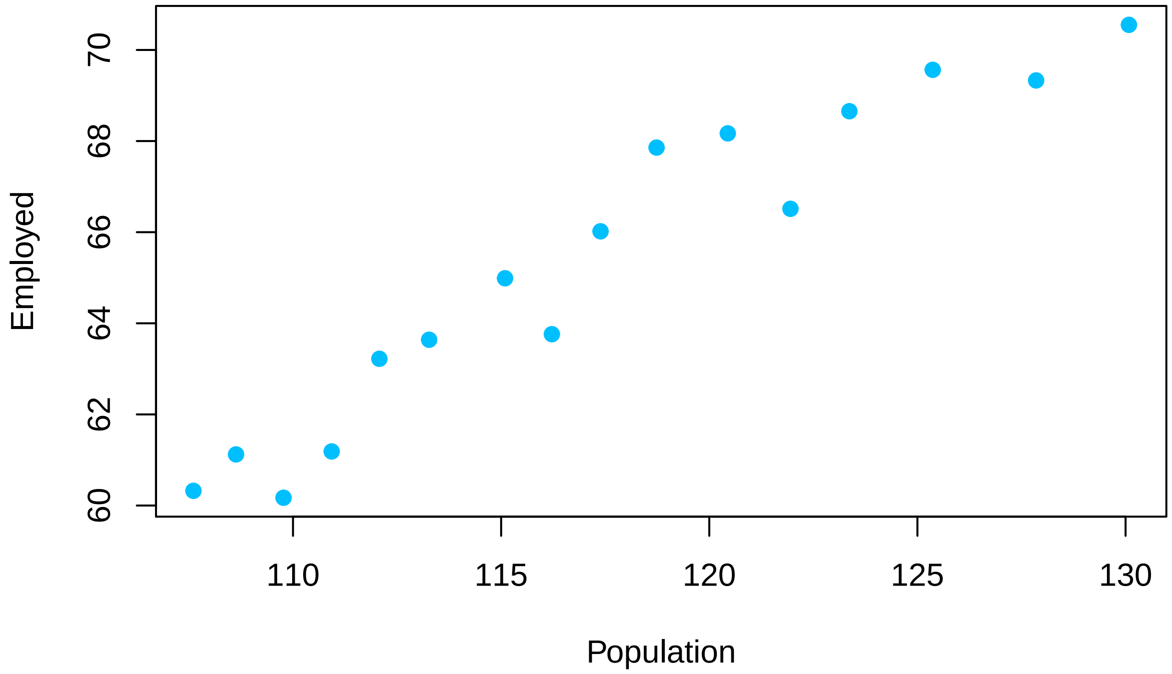
```
par(old.par)
# scatterplot of bivariate features
plot(longley$Employed, longley$GNP,
     type = "p", pch = 19, col = "deepskyblue",
     xlab = "Employed", ylab = "GNP")
```



```
plot(longley$Population, longley$GNP,
     type = "p", pch = 19, col = "deepskyblue",
     xlab = "Population", ylab = "GNP")
```



```r
plot(longley$Population, longley$Employed,
     type = "p", pch = 19, col = "deepskyblue",
     xlab = "Population", ylab = "Employed")
```



Both Employed and Population are highly positively correlated with GNP. However, there is also a correlation between Population and Employed.

Task 2

```
require(xtable)
```

```
## Loading required package: xtable
```

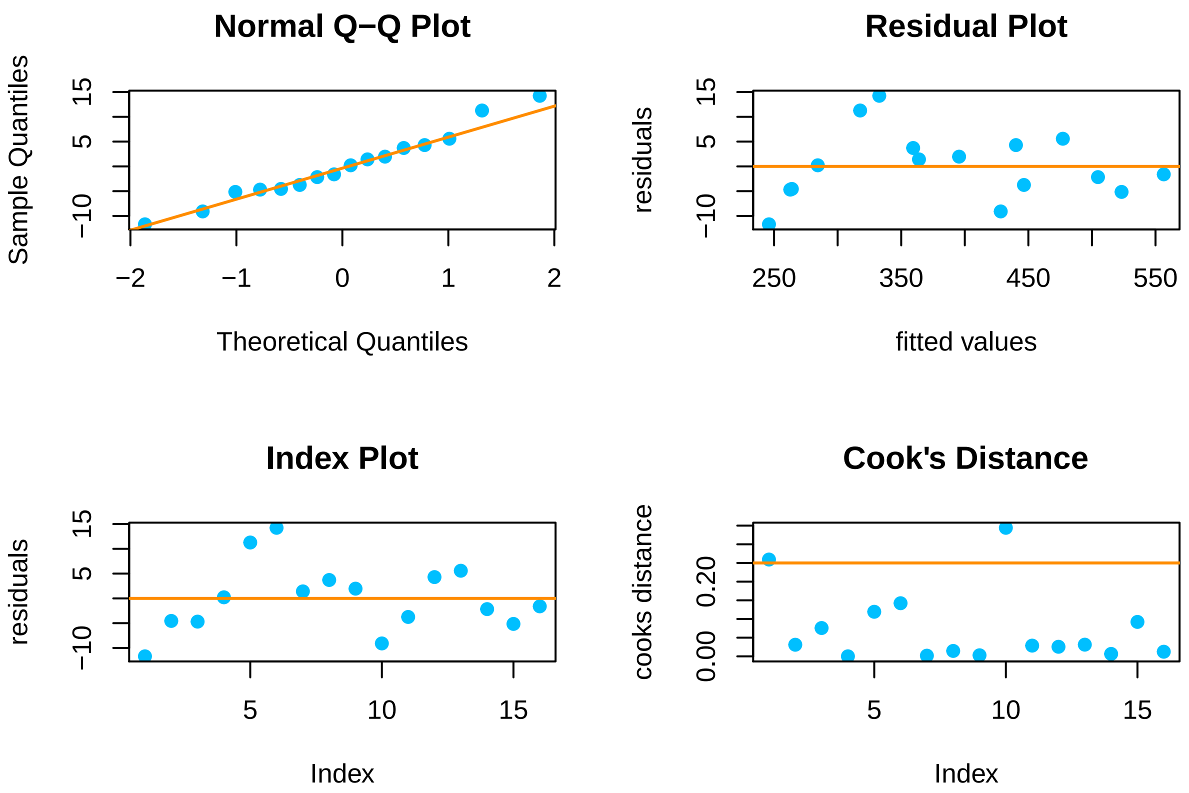
```
lm1 <- lm(GNP ~ Employed + Population, data = longley)
lm1.table <- xtable(summary(lm1))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1372.0954	36.1406	-37.97	0.0000
Employed	11.5606	1.9484	5.93	0.0000
Population	8.5561	0.9837	8.70	0.0000

The coefficients mean that, on average, one unit increase in Employed will result in 11.5606 increase in GNP; one unit increase in Population will increase 8.5561 GNP.
The intercept mean the level of GNP when both Population and Employed are at zero. It carries no meaning.

Task 3

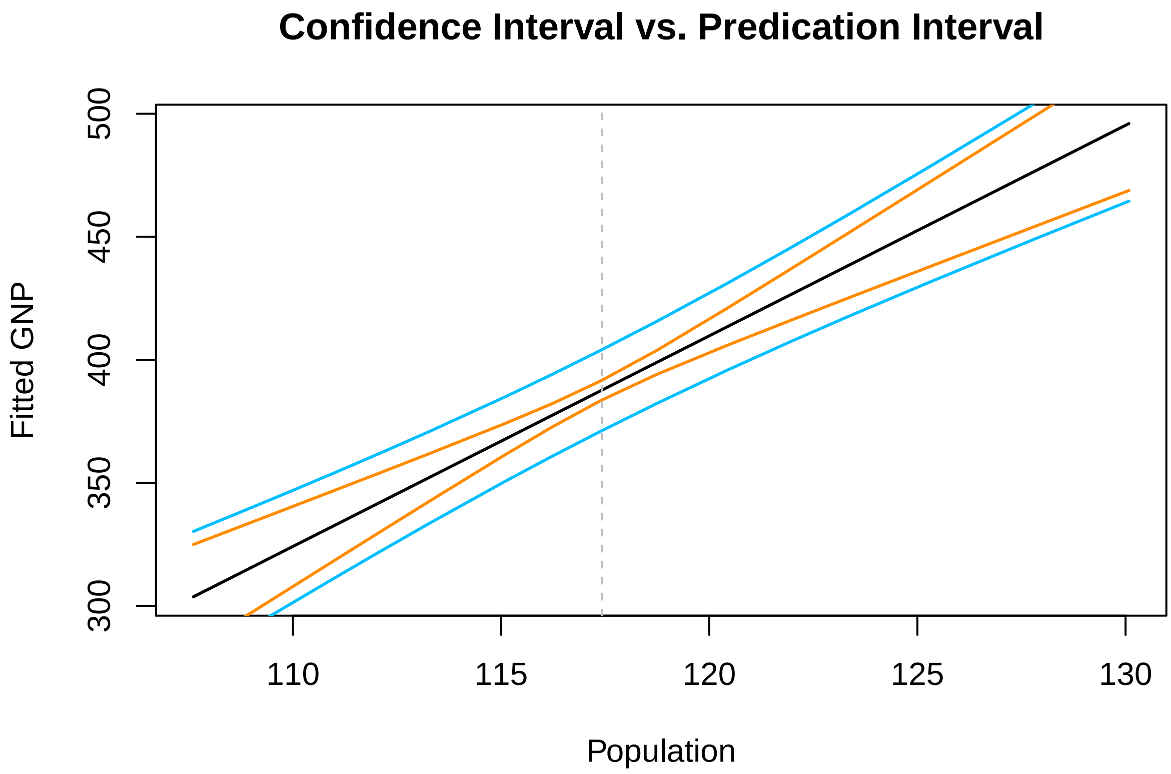
```r
## Run Diagnosis ##
old.par <- par(mfrow = c(2,2))
# normality of the residuals, using QQ plot #
qqnorm(lm1$residuals, pch = 19, col = "deepskyblue")
qqline(lm1$residuals, lwd = 1.5, col = "darkorange")
# homoscedasticity #
plot(lm1$fitted.values, lm1$residuals,
     type = "p", pch = 19, col = "deepskyblue",
     xlab = "fitted values",
     ylab = "residuals",
     main = "Residual Plot")
abline(h = 0, lwd = 1.5, col = "darkorange")
# independence, index plot #
plot(lm1$residuals,
     type = "p", pch = 19, col = "deepskyblue",
     ylab = "residuals",
     main = "Index Plot")
abline(h = 0, lwd = 1.5, col = "darkorange")
# Cook's D
plot(cooks.distance(lm1),
     type = "p", pch = 19, col = "deepskyblue",
     ylab = "cooks distance",
     main = "Cook's Distance")
abline(h = 4/nrow(longley), lwd = 1.5, col = "darkorange")
```



We observe that the normal assumption holds true except for the tail of the residual is a little bit flat (according to QQ plot); the mean of residuals is at zero and the constant variance holds true as well (According to Residual plot and Index plot).
According to Cook's Distance, only two of the points are outliers, which should be fine.
I don't think the data needs transformation here. If I were to transform, the interpretability of the model will be compromised and we cannot use coefficients directly as a measure of impact of the predictors.

Task 4

```r
par(old.par)
# Create a new data frame with Population values and Employed value held at mean value
new_df <- data.frame(Population = longley$Population, Employed = mean(longley$Employed))
# Make prediction
GNP_pred_ci <- predict(lm1, newdata = new_df, interval = "confidence")
GNP_pred_pi <- predict(lm1, newdata = new_df, interval = "prediction")
# Make plots
plot(new_df$Population, GNP_pred_ci[, "fit"],
     type = "l", lwd = 1.5,
     xlab = "Population",
     ylab = "Fitted GNP",
     main = "Confidence Interval vs. Predication Interval")
lines(new_df$Population, GNP_pred_ci[, "lwr"], col = "darkorange", lwd = 1.5)
lines(new_df$Population, GNP_pred_ci[, "upr"], col = "darkorange", lwd = 1.5)
lines(new_df$Population, GNP_pred_pi[, "lwr"], col = "deepskyblue", lwd = 1.5)
lines(new_df$Population, GNP_pred_pi[, "upr"], col = "deepskyblue", lwd = 1.5)
abline(v = mean(new_df$Population), col = "grey", lty = 2)
```



The prediction intervals are always wider than confidence intervals. The narrowest gets at the mean value of Population.
When n approximates infinite, the confidence interval will approach to zero, while prediction interval will become narrower but never reach to zero.