XIAO_Lin_Solutions_lab6

Lin Xiao

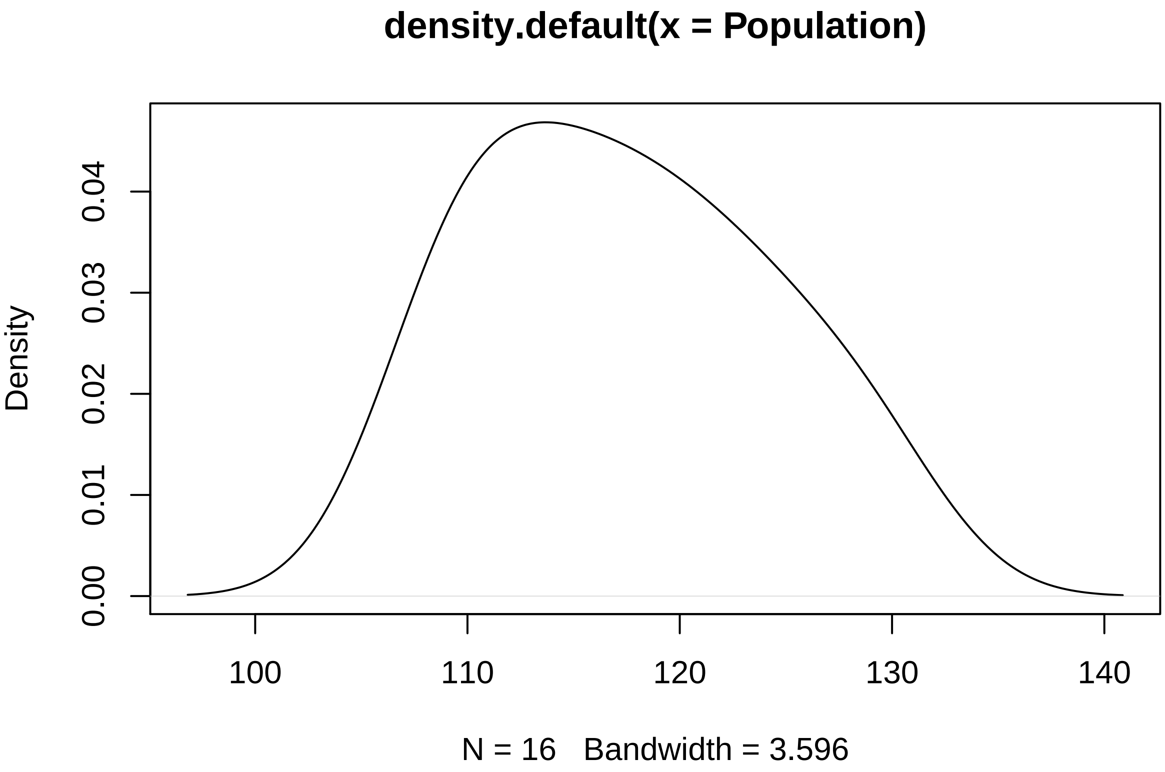
October 19, 2015

1. Your goal is to build a regression model for Gross National Product (GNP) based on two input variables: number of people employed and the total population using the longley data in R. Start by performing graphical exploratory data analysis: create univariate density estimates and scatterplots to understand the bivariate features of the data. Do you see anything interesting?

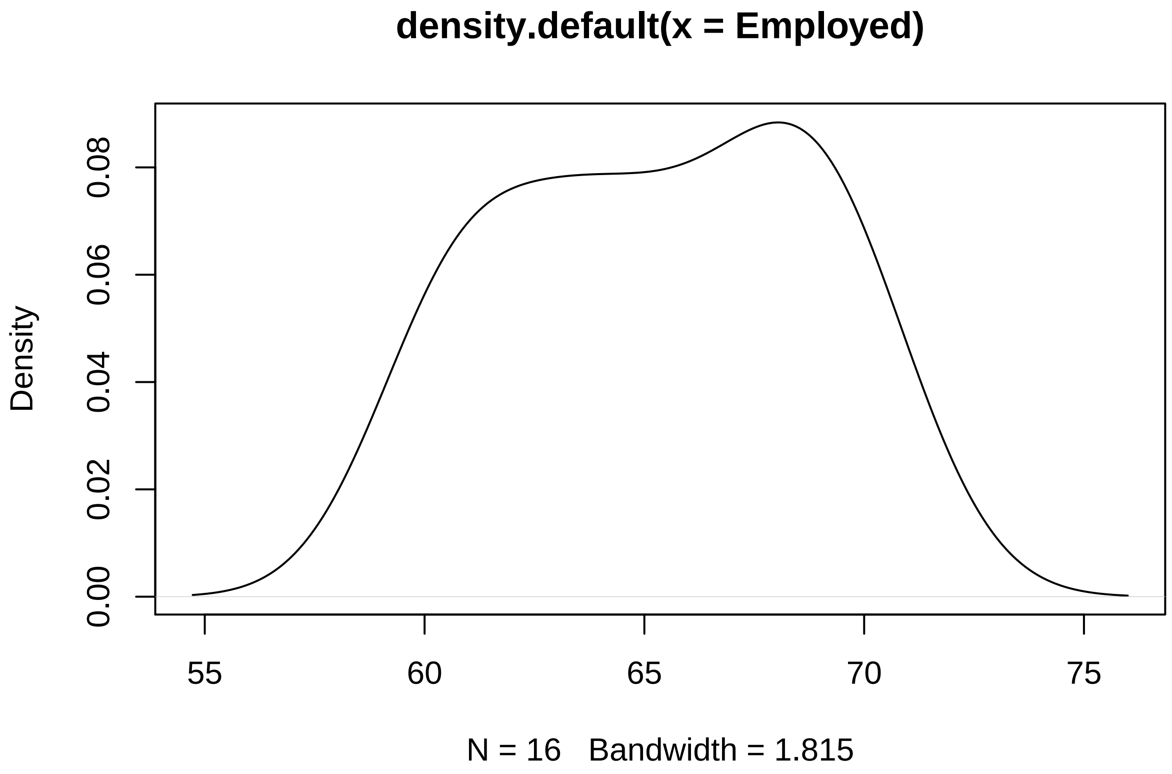
```
require(xtable)
```

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

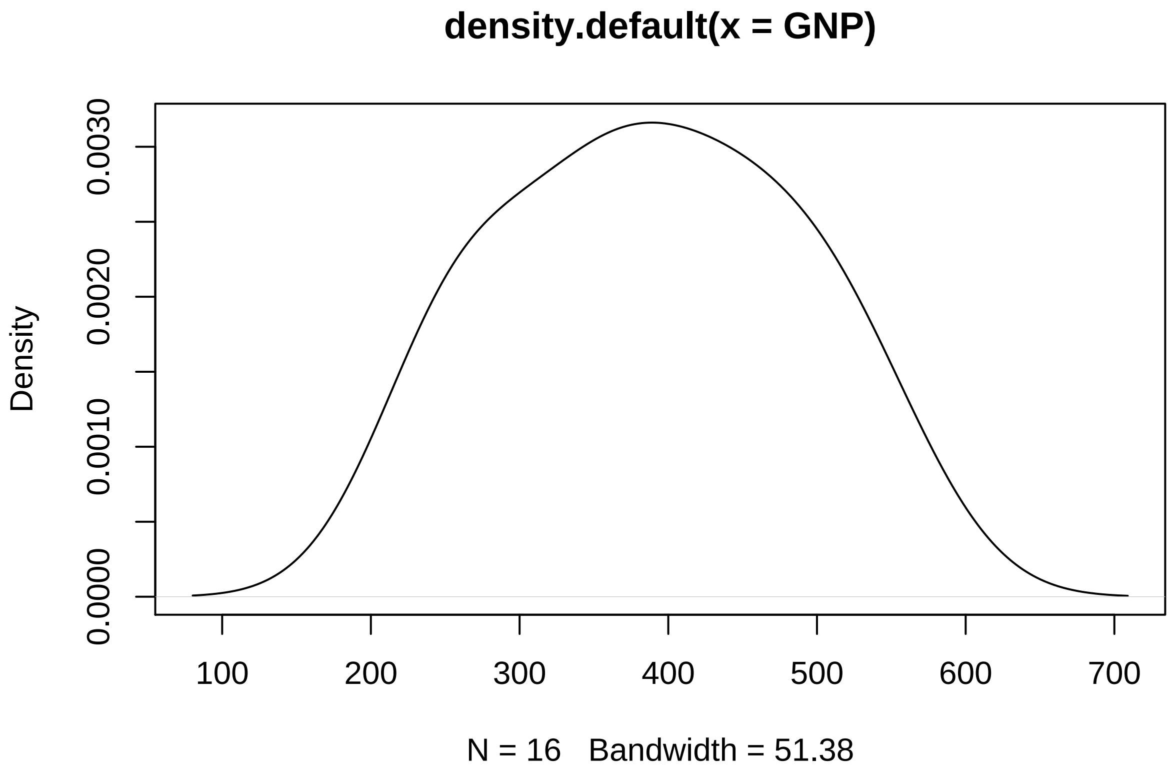
```
data("longley")
attach(longley)
plot(density(Population))
```



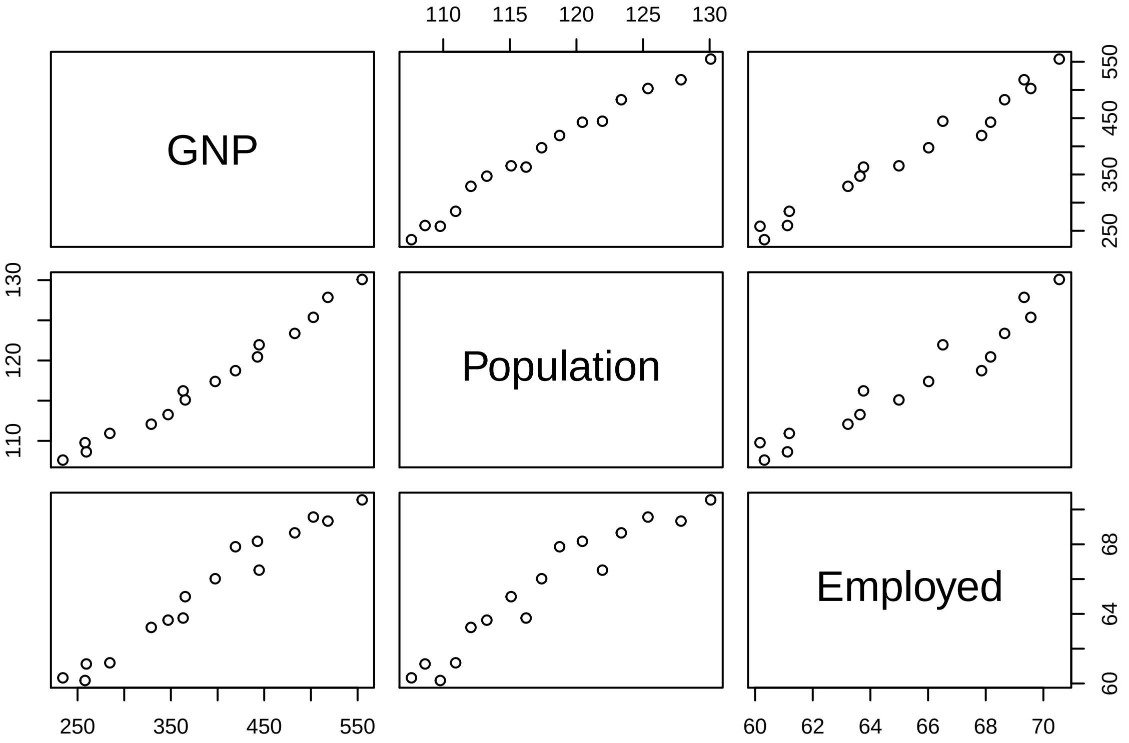
```
plot(density(Employed))
```



```r
plot(density(GNP))
```



```
# Plot the scatter plot matrix
pairs(cbind(GNP,Population,Employed))
```



We can see from the univariate density plots that they are almost normally distributed, except for tiny skewness for the Employment. Also from the scatter plot matrix we can see that not only the GNP is highly correlated to the Employment and the Population, but there is also multicollinearity between Population and the Employment. That is something we should be worried about.

2. Build the regression model, and provide the coefficients of the model and details on their significance using xtable. Interpret your coefficients. Is the intercept meaningful? What can you do to make the intercept more meaningful?

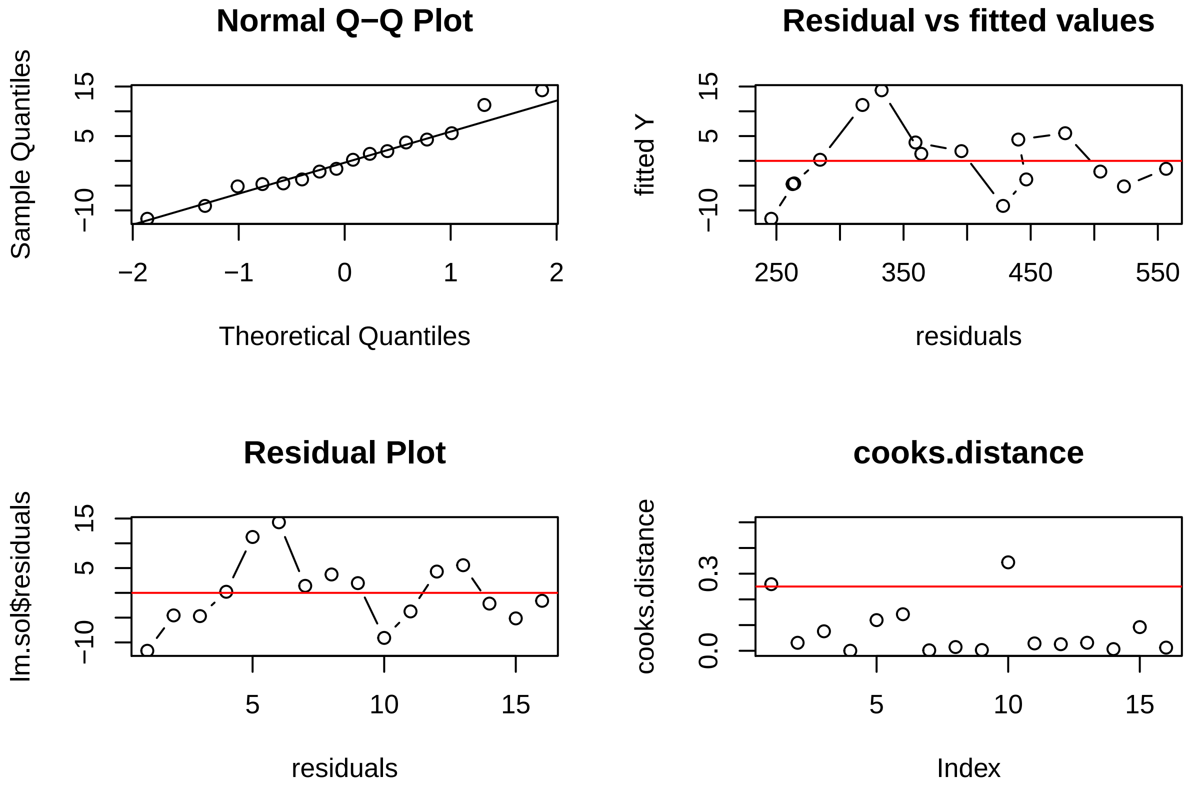
	Estimates	Significance
(Intercept)	-1372.0954	0.000000
Population	8.5561	0.000001
Employed	11.5606	0.000050

For the intercept, it means if we set the Population and the Employed both equal to 0, GNP has the value of -1372.0954. And if we hold Employed constant, a unit increase in Population will result in 8.5561 increase in GNP. If we hold Population constant, a unit increase in Employed will result in 11.5606 increase in GNP. The intercept is not meaningful since the Population and the Employed value will never be zero in reality. I would rescale X by subtracting medroid{X} from X. And then the intercept is simply the GNP value when X takes its medroid value.

3. Perform regression diagnostics via graphical methods: Assess normality of your residuals, constant variance, independence as well any potentially influential points using Cook's Distance with a threshold

value of 4. Be sure to detail what you see. Do you need to transform your data? If you were to transform your data, how would it impact the interpretation of your model?

```r
par(mfrow=c(2,2))

# Exam normality of the residuals
qqnorm(lm.sol$residuals)
qqline(lm.sol$residuals)

# Exam the variance constancy
plot(fitted(lm.sol), lm.sol$residuals, type="b", xlab = "residuals",
     ylab = "fitted Y", main = "Residual vs fitted values")
abline(a=0, b=0, col="red")

# Exam the independence of residuals
plot(lm.sol$residuals,type="b", xlab = "residuals", main = "Residual Plot")
abline(a=0, b=0, col="red")

# Using cooks distance to find out the influential points
plot(cooks.distance(lm.sol), ylab = "cooks.distance",
     main = "cooks.distance", ylim = c(0,0.5))
abline(a=4/nrow(longley), b=0, col="red")
```



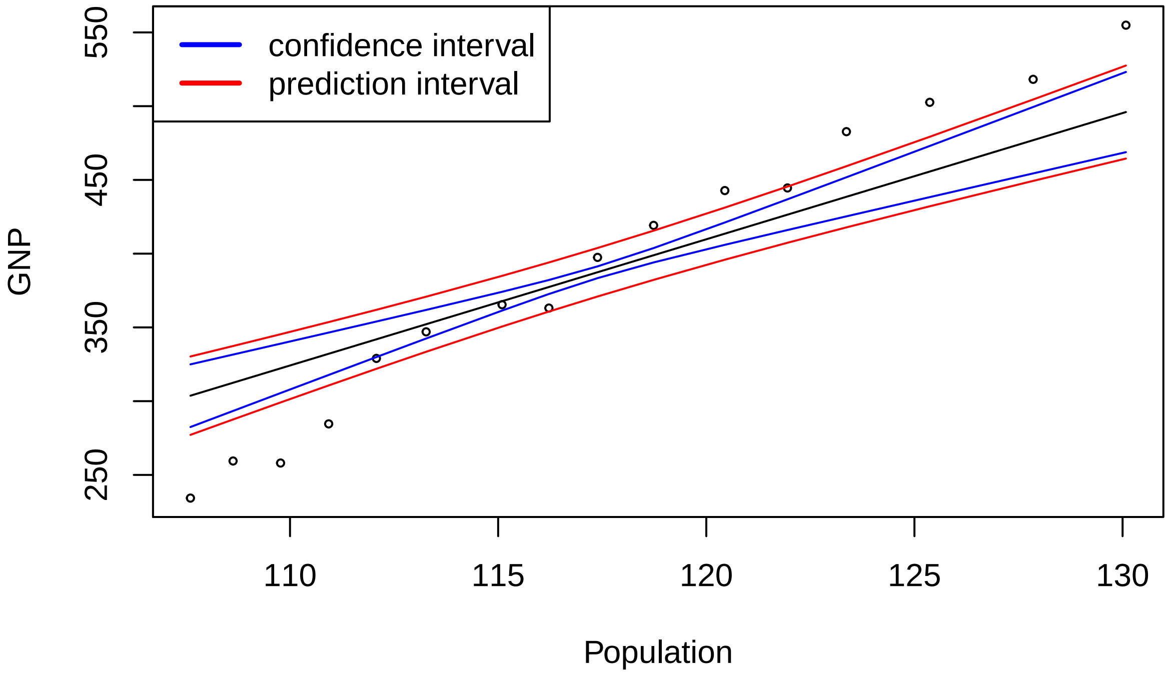
```r
# Return the positions of influential points
which(as.vector(cooks.distance(lm.sol)) > 4/nrow(longley))
```

```
## [1]  1 10
```

We can see from the first plot that the residuals are almost normaly distributed, but for the residual plots(the second and the third) we can tell the variance is not constant and the residuals are not independent, because they are not scattered randomly around 0. Also the 1st and 10th points are influential points that needs to be taken out.
So we have to transform the data. One way to do it is to get rid of the influential points and refit the model, it would not make huge difference in interpreting the model. Another way to transform is to take logs(or reciprocal, exponetial and other functions) on either Y or X(or part of regressors), refit the model and exam the constancy of variance, and independence of residuals.
For example, if we take log on GNP data, a unit increase in Population holding Employment constant will result in GNP with increase of $e^{\beta_{Population}}$.

4. Create a plot of Population against GNP that shows the fitted regression line holding Employment at its mean value. Add prediction and confidence intervals to your plot based on the same assumption in different colors. Where are the intervals narrowest? What do you expect will happened to the intervals as $n \to \infty$.

```r
par(mfrow=c(1,1))
# Reconstruct the data frame for X
newdata <- data.frame("x0" = rep(1,length(Employed)), "Population" = Population,
                      "Employed" = rep(mean(Employed),length(Employed)))
fitted <- predict(lm.sol, newdata)

# Plot GNP against Population
plot(Population, GNP, cex=0.5)

# Show the fitted line
lines(Population, fitted)

# Get the confidence intervals
CI <- predict(lm.sol, newdata= newdata, interval = 'confidence')

# Get the prediction intervals
PI <- predict(lm.sol, newdata= newdata, interval = 'predict')

# Draw the interval lines
lines(Population,PI[,3], col="red")
lines(Population,PI[,2], col="red")
lines(Population,CI[,3], col="blue")
lines(Population,CI[,2], col="blue")
legend("topleft", c("confidence interval","prediction interval"),lty=c(1,1),
       lwd=c(2.5,2.5 ), col=c("blue", "red"))
```



```r
# Find out the narrowest place(year)
unique(which(PI[,3]-PI[,2]==min(PI[,3]-PI[,2])))
```

```
## [1] 9
```

```r
unique(which(CI[,3]-CI[,2]==min(CI[,3]-CI[,2])))
```

```
## [1] 9
```

We can see from above that the intervals reach the narrowest in the 9th row of longley data, since the population was increasing during the years, the narrowest happened in 1955.
Since a confidence interval for y for a given x^* is

$$\widehat{y} \pm t^*_{n-2} s_y \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{(n-1)s_x^2}}$$

And a prediction interval for a given x^* is

$$\widehat{y} \pm t^*_{n-2} s_y \sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{(n-1)s_x^2}}$$

As $n \to \infty$, the confidence intervals shrink into set of points, while prediction intervals become

$$\widehat{y} \pm t^*_{n-2} s_y$$