STAT7203: Applied Probability and Statistics
Week 12 Exercises

1. According to Hubble's law, relative velocity v (km/s) of any two galaxies separated by a distance D (Mega parsec – 1 parsec is 3.09×10^{13} km) is given by

$$v = H_0 D,$$

where H_0 is Hubble's constant. If the expansion of the universe was linear, then $1/H_0$ (Hubble Time) would give the age of the universe. The velocities and distances of 24 galaxies containing Cepheid stars is given in the file hubble.xlsx.

Some useful R commands for this problem:

- We can read in a csv file using the function `read.csv`. If you have a file `myDataFile.csv` in your working directory, then you can read in the data using

```
myData = read.csv('myDataFile.csv', header=TRUE)
```

The argument `header=TRUE` indicates that the first line of the file has the names of the variables.

- `lm` is the built in function to fit the linear regression model. If the data is stored in a dataframe called `myData`, then model is fitted using

```
myModel = lm('Response ~ Explanatory', data=myData)
```

- A summary of the fitted model can be obtained using the function `summary`.

```
summary(myModel)
```

- Some diagnostic plots can be generated using

```
plot(myModel)
```

The first plot is a plot of residuals versus fitted values. The second plot has the residuals against the theoretical quantiles of the standard normal distribution. We do not cover the other plots generated in this course.

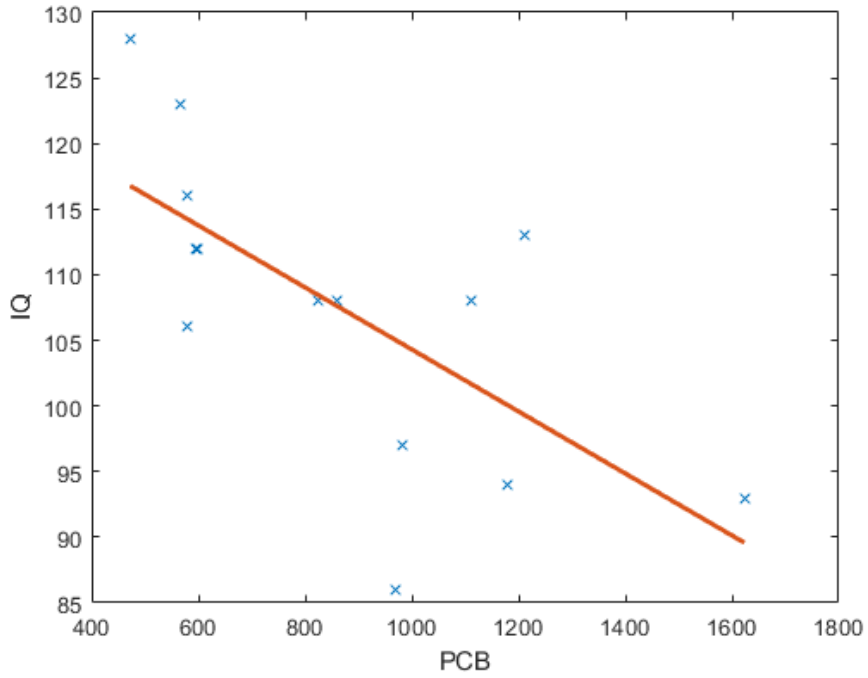
(a) Fit the linear regression model $V = \beta_0 + \beta_1 D + \varepsilon$, where $\varepsilon \sim \mathsf{N}(0, \sigma^2)$ to the hubble data. Assess the suitability of the linear regression model with diagnostic plots.

(b) Assuming the linear regression model is appropriate, is the data consistent with $\beta_0 = 0$.

(c) In the linear regression model, Hubble's constant is β_1. Construct a 99% confidence interval for the Hubble constant.

(d) Construct a 95% confidence interval for the mean relative velocity of two galaxies separated by 10 Mega pasecs.

2. Polychlorinated biphenyls (PCBs) were once used in industry but were banned in the 1970s because of concerns about their toxicity. Despite the ban, PCBs can still be detected in most people because they are persistent in the environment. A team of researchers recorded the amount of PCBs detected in maternal milk from mothers who had eaten fish from a particular lake considered to be contaminated with PCBs. They subsequently administered an IQ test to the children when they were 11 years old. The data from 14 mothers and their eldest child are shown in the following scatter plot along with the least-squares line fitting a linear relationship between the two variables:



A regression analysis produced the following edited summary:

```
Coefficients:
             Estimate Std. Error
(Intercept) 127.937156   6.962961
PCB          -0.023631   0.007529
```

(a) Briefly interpret the value -0.023631.

(b) Carry out a t-test to assess whether there is evidence of a association between maternal milk PCB levels and IQ outcome. Show your working and state your conclusion.

(c) Construct a 95% confidence interval for the intercept of the regression line.

(d) Based on the coefficients for the least-squares line provided by MATLAB, estimate the mean IQ of children if their mothers had a maternal milk PCB measurement of 1400 ng/g. If the child has an IQ 102, what is the residual?

(e) State the assumptions underlying linear regression.

(f) The plot below plots the residuals of the linear regression against the theoretical quantiles of the standard normal distribution. Comment on the validity of the assumptions of the linear regression model.