R preparation: Linear Regression

Quantitative Reasoning

Preparation for class on 2020-10-01

In Chapter 7, we will be going through linear regression. We will be using data from the survey you completed prior to the mid-semester break in order to practice running linear regressions. As a preparation for our next lesson, let's use `shoe` to predict `height`.

Import `survey.csv` and Inspect

To begin, lets import `survey.csv` and inspect.

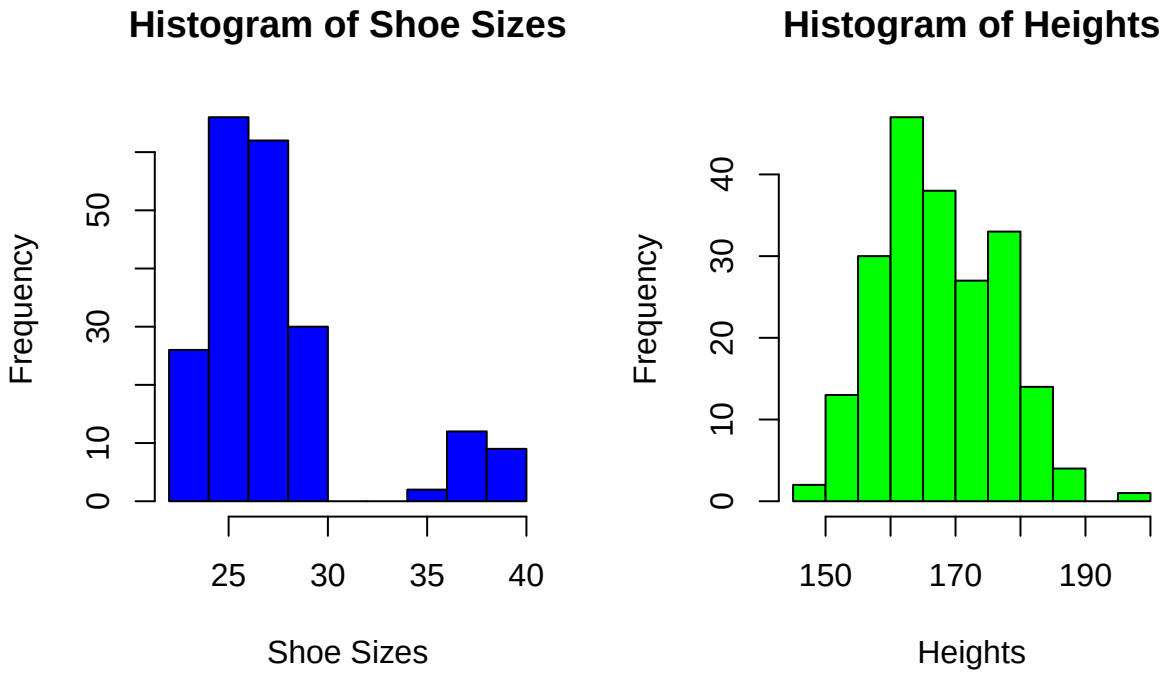
```
survey <- read.csv("survey.csv")
```

```
str(survey)
```

Histograms and Scatterplot of Height and Shoe Size

Before we use `shoe` to predict `height`, let's plot out histograms of the two variables to assess their shapes:

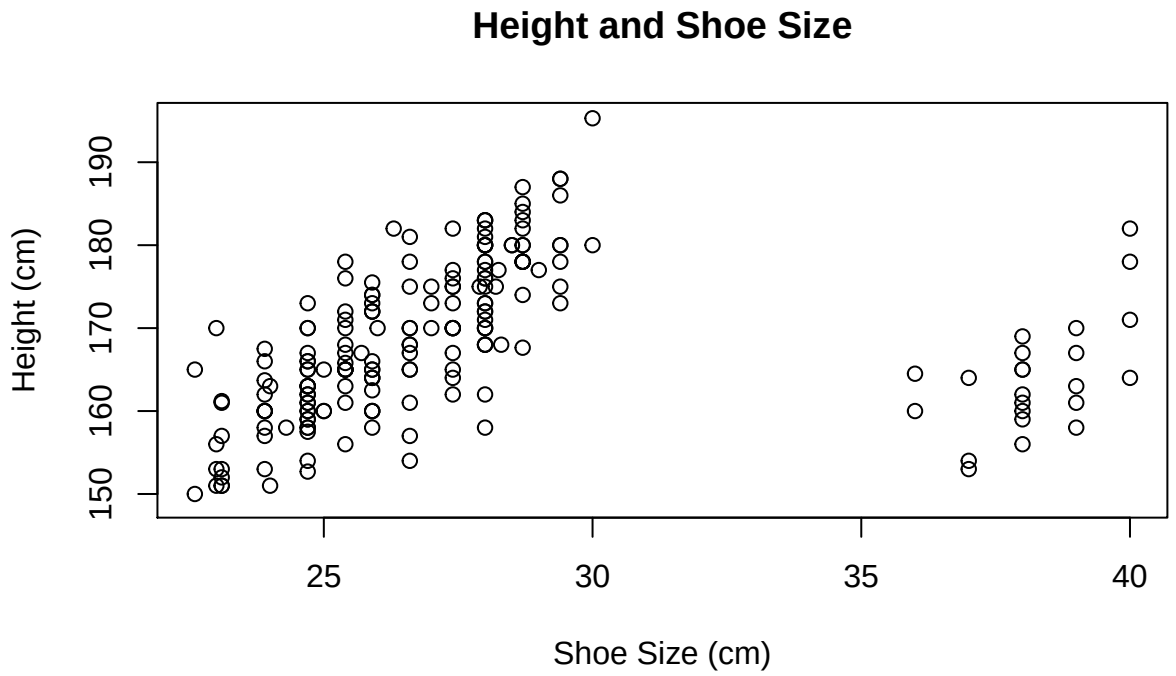
```
par(mfrow=c(1,2))
hist(survey$shoe,   main="Histogram of Shoe Sizes", xlab="Shoe Sizes", col="blue")
hist(survey$height, main="Histogram of Heights",    xlab="Heights",    col="green")
```




```
par(mfrow=c(1,1))
```

Now let's plot a scatter plot to visualise the relationship between the variables:

```
plot(height~shoe, data=survey,
     main="Height and Shoe Size",
     ylab="Height (cm)",
     xlab="Shoe Size (cm)")
```



Note the outliers off of the right tail of the histogram for `shoe` and in the lower right of the scatterplot. Recall the discussion from Monday's class that these outliers are probably errors: Respondents most likely misread the conversion table for shoe sizes. Rather than enter their shoe size in centimetres as requested, they presumably entered European size instead.

Though it is not usually a good idea to do so, let's ignore the outliers for a minute. Let's instead focus on the main body of the two histograms and the main cluster of points in the scatter plot. With respect to the histograms, the main body of both seems to be more or less normally distributed. With respect to the scatter plot, the points are arrayed in a fairly compact cloud that it seems like we might draw a straight line through, rising from the left to the right. It seems like a linear model might be appropriate.

Linear Regression

We can use `shoe` to predict `height` with a linear model using the `lm()` command in R. We will store the results of the linear model in a new object we will call `lm_hs`. We can then print the contents of `lm_hs` and see the estimated y-intercept as well as the slope of the line of best fit for the data.

```
lm_hs <- lm(height ~ shoe, data=survey)
lm_hs
```

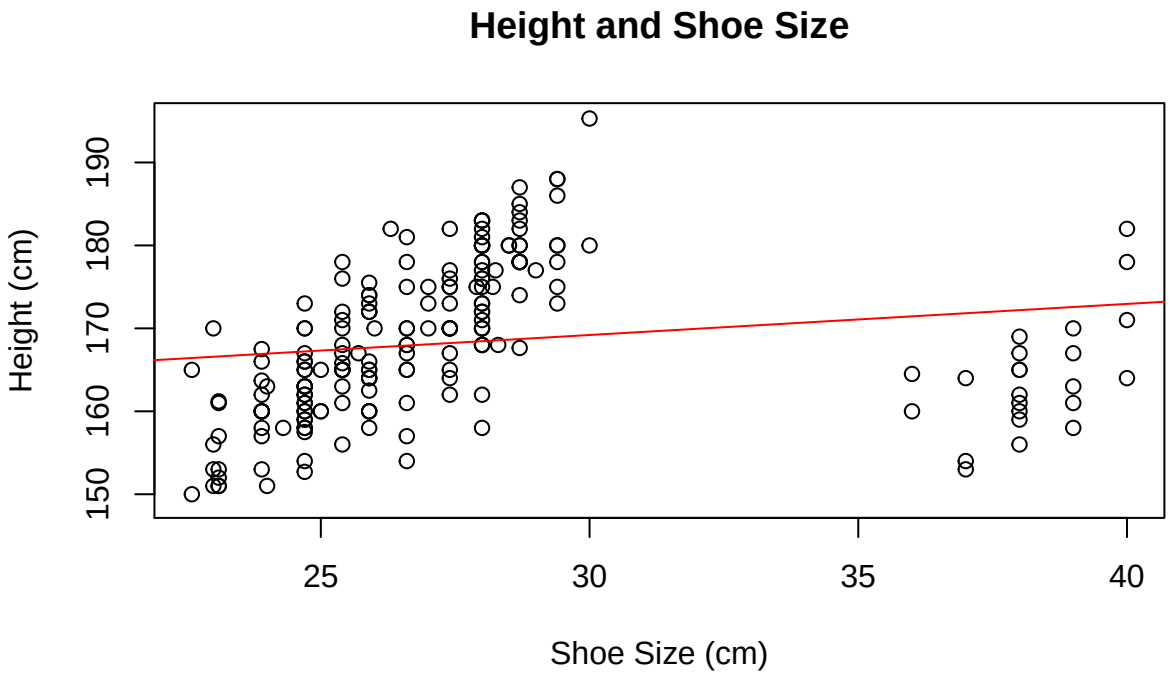
```
##
## Call:
## lm(formula = height ~ shoe, data = survey)
##
## Coefficients:
## (Intercept)         shoe
##    157.9563       0.3747
```

The estimated y-intercept produced by the linear model implies that the line of best fit intercepts with the y-axis at approximately 158.0 centimetres. The estimated slope implies that for each additional centimetre shoe size, the model predicts that an individual's height will be an additional 0.37 centimetres taller.

Admittedly, the estimated y-intercept and slope seem a bit odd when you take a moment to think about them: In particular, the slope would imply that the relationship between `shoe size` and `height` is such that for an additional centimetre increase in `shoe size`, the prediction for `height` increases by less than a centimetre. Since people's heights vary more than their shoe sizes, we would think that the slope should probably be steeper.

To get a better sense of what might be going on, lets plot out the scatter plot again but this time use `abline()` to add the line of best fit based on the linear model `lm_hs`.

```
plot(height~shoe, data=survey,
     main="Height and Shoe Size",
     ylab="Height (cm)",
     xlab="Shoe Size (cm)")
abline(lm_hs, col="red") # line of best fit
```



Ah! We forgot to exclude the outliers when we estimated the linear model! It appears that the line of best fit produced by the model thus took into account the outliers. The outliers appear to have "flattened" the slope such that it goes through both the main cluster of points in the scatter plot as well as the outliers. At the same time, the cluster of

Excluding the Outliers

The obvious solution here would seem to be that we should re-estimate the linear model, but exclude the outliers (`shoe` > 35) when doing so. Let's do so and see if the estimated y-intercept and slope seems more reasonable.

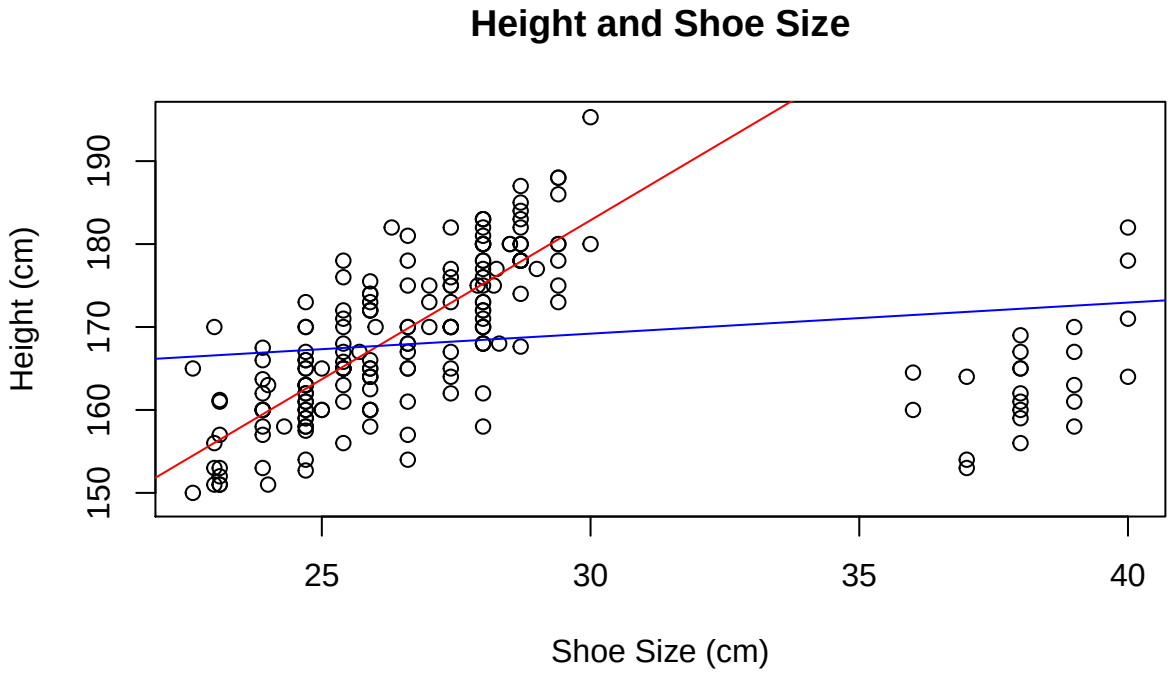
```
lm_hs_main <- lm( height ~ shoe, data=survey[survey$shoe<35,])
lm_hs_main
```

```
##
## Call:
## lm(formula = height ~ shoe, data = survey[survey$shoe < 35, ])
##
## Coefficients:
## (Intercept)         shoe
##      67.830        3.835
```

OK, so that seems a bit better: Excluding outliers, the y-intercept is 67.8 and the slope is 3.8. If we focus on the slope, then this seems more reasonable as the slope is now greater than 1. The slope now implies that for each additional centimetre of shoe size, the model predicts that height will be an additional 3.8 centimetres.

Let's also check out the scatter plot to see if the line of best fit from `lm_hs_main` now seems to fit the main cluster of points. Let's also add in the line of best fit from `lm_hs` for the sake of comparison.

```
plot(height~shoe, data=survey,
     main="Height and Shoe Size",
     ylab="Height (cm)",
     xlab="Shoe Size (cm)")
abline(lm_hs_main, col="red") # new line of best fit
abline(lm_hs, col="blue") # old line of best fit
```



Much better! The line of best fit now seems to pass directly through the centre of the main cluster of points. Moreover, the estimated slope seems much more realistic.