Cars II

This project deals with the linear regression model and how we fit this with the computer.

1 introduction

We have the following model that relates the y values to the x values:

$$y_i = b_0 + b_1 x_i + \text{error}_i.$$

That is, given an x value, we get the y value by finding the position on the line $b_0 + b_1 x$ and then add an "error" term.

We estimate the values of b_0 and b_1 from the data (x_i, y_i) using the formulas

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}, \quad b_0 = \bar{y} - b_1 \bar{x}.$$

That is, the regression line goes through $(\bar{x}, \bar{y})$ and has slope b_1.

The prediction line is $\hat{y} = b_0 + b_1 x$. We use this to predict a y value for a given x_i. Call this $\hat{y}_i = b_0 + b_1 x_i$. Then the residual error is

$$e_i = y_i - \hat{y}_i.$$

The residuals, are the collection of residual errors. They tell us about the aptness of the model.

1.1 Finding the estimates

With the computer we can find the estimated values and residuals with the aid of the `lm()` function. If `x` and `y` store the data then the command `lm(y ∼ x)` will find the coefficients. Usually you store this in a model as in this example.

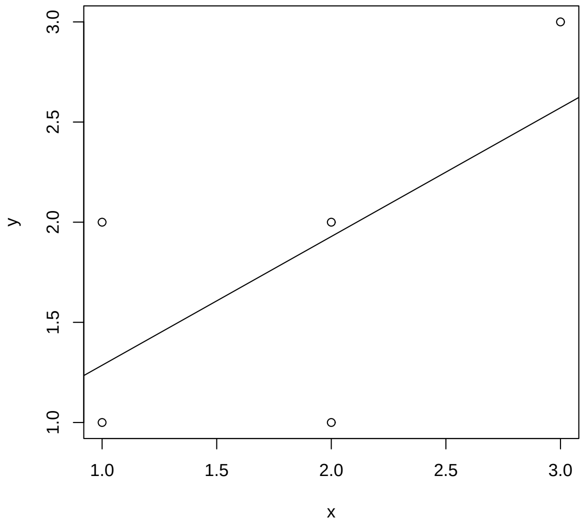
```
> x = c(1, 1, 2, 2, 3)
> y = c(1, 2, 1, 2, 3)
> plot(x, y)
> res = lm(y ~ x)
> res

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
      0.643        0.643
```

From the output we see that the prediction line is $\hat{y} = .063 + .643x$. How do we draw that on our plot? The function `abline` will do it for us with `abline(res)`.

The tilde (∼) is used for statistical models and the expression `y ∼ x` is a model formula. You can also use the model formula with plotting as we did with boxplots. For example both `plot(x,y)` and `plot(y ∼ x)` produce the same graphic. The latter makes it easier to reuse typing.

So the above could be done with

```
> plot(y ~ x)
> res = lm(y ~ x)
> abline(res)
```



2 The Cars data

We will use the `Cars93` data set again. This is a built in data set that accompanies the `MASS` package. To load it we type the following.

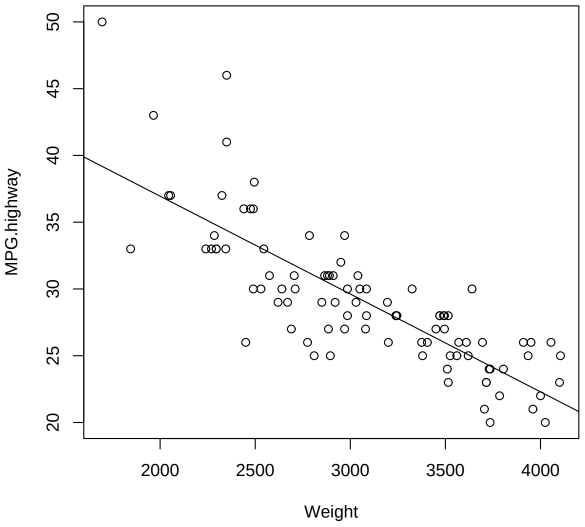
```
> library(MASS)
> data(Cars93)
> attach(Cars93)
```

Now all the variables are visible. Use `names(Cars93)` to see the names.

We will model city mileage by weight as an example.

To fit the model we replace `y` above by `MPG.city` and `x` by `Weight`.

```
> plot(MPG.highway ~ Weight)
> res = lm(MPG.highway ~ Weight)
> abline(res)
> res

Call:
lm(formula = MPG.highway ~ Weight)

Coefficients:
(Intercept)       Weight
   51.60137     -0.00733
```



What is the predicted city mileage of a 3000 pound car? We answer this with the prediction line.

```
> 51.6013 - 0.00733 * 3000
```

```
[1] 29.61
```

Question 1: Fit the linear model with `MPG.highway` modeled by `Weight`. Find the predicted highway mileage of a 6400 pound HUMMER H2 and a 2524 pound MINI Cooper.

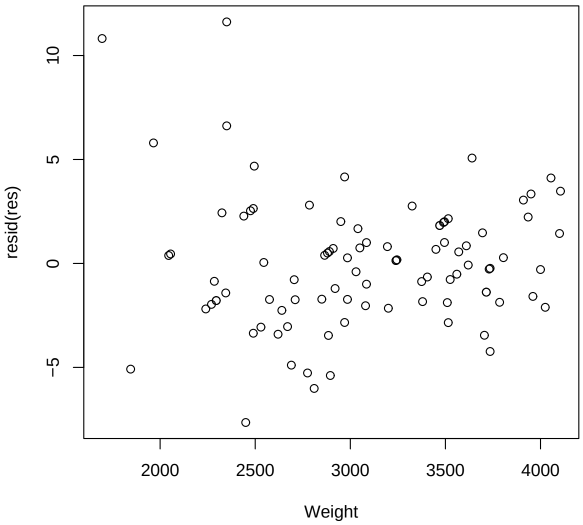
Question 2: Model `MPG.highway` by `EngineSize`. What is the predicted mileage of a 2.8 liter Audi wagon.

2.1 residual plots

A plot of residuals can inform us of the aptness of the linear model for a relationship. To make the plots, you need the residuals which are returned with the command `resid(res)` if `res` stores the output of `lm()`.

For example, the residuals for the model of `MPG.highway` modeled by `Weight` are plotted with

```
> res = lm(MPG.highway ~ Weight)
> plot(Weight, resid(res))
```



From the graph we see that the bulk of the data is scattered but has roughly the same spread. For the small weight cars, there is more spread. If there were lots of these or some really large residuals we would have to think twice about using a linear model.

Question 3: Make a residual plot for the model `MPG.highway` by `EngineSize`. Are they scattered with a common spread? Does the spread increase or decrease? Is there a trend in the residuals?

Question 4: Model `Weight` by `Length`. Make a residual plot. Does the linear model seem appropriate?

Question 5: Model `Turn.circle` by `Length`. Make a residual plot. Does the linear model seem appropriate?

2.2 The five number summary of regression

The five number summary for regression consists of the means, the standard deviations and the correlation coefficient. This is because, the prediction line formula can be rewritten as

$$\frac{\hat{y}_i - \overline{y}}{s_y} = r\frac{x_i - \overline{x}}{s_x}$$

This says, that when we find the regression line for the z-scores of a data set, the slope and correlation are the same. To see this we can use the `scale` function as follows. (We first store the formula so we don't have to retype.)

```
> f = formula(scale(MPG.highway) ~ scale(Weight))
> plot(f)
> res = lm(f)
> abline(res)
> res

Call:
lm(formula = f)
```

```
Coefficients:
  (Intercept)  scale(Weight)
    -1.08e-16      -8.11e-01

> cor(MPG.highway, Weight)

[1] -0.8107
```

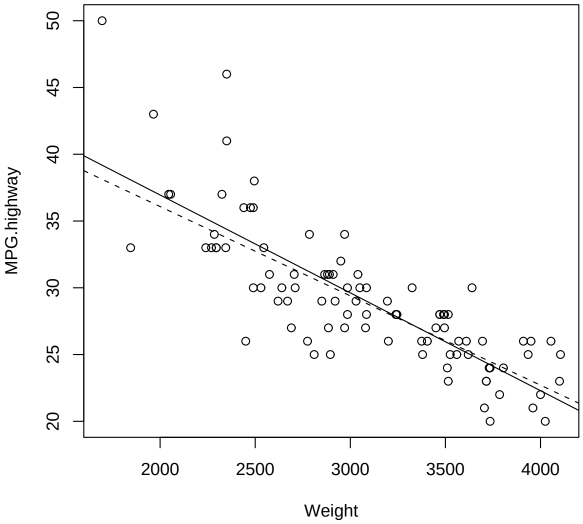
Question 6: Repeat the above for the model of `Turn.circle` by `Length`. Do the correlation and slope of regression line coincide?

2.3 resistant regression

When there are bivariate outliers in a data set, the regression coefficients can be affected so that the predictions are not very good. If the data has a linear trend, except for a few outliers, resistant regression methods should be used. In the MASS package the `rlm()` function performs resistant regression. You use it identically as `lm()`.

For example, to draw both a regression line and a resistant one for highway mileage modeled by weight do

```
> f = formula(MPG.highway ~ Weight)
> plot(f)
> abline(lm(f))
> abline(rlm(f), lty = 2)
```



The argument `lty=2` draws the line with dashes.

The graph shows a slightly flatter slope as the two high-mileage cars do not have the same influence.

Question 7: Model the `Length` of a car by the number of `Passengers`. Draw both the regression line and the resistant regression line. Is there much difference? Did you expect much?

2.3.1 Problems

💻 Question 8: Model `MPG.city` by `Price`. Use your model to predict the gas mileage of a $50,245 HUMMER H2.

What is the residual amount if the actual mileage is 10/13 mpg.

Look at the residual plot. Does it show the linear model as appropriate? Explain.

💻 Question 9: Model `Fuel.tank.capacity` by `MPG.highway`. Using residuals assess if a linear model is a good fit for the data.

One might think that manufactures wan't to have a certain range in mind, so that low mileage cars have large fuel tanks. This would suggest a model of `1/Fuel.tank.capacity` modeled by `MPG.highway`. Use residuals to assess this model.

Which model seems to be more accurate based on the residuals?

💻 Question 10: Model the `Max.Price` by `Min.Price`. Fit a linear model and look at the residuals. Are there any outliers? In what sense? If there are, also add a resistant regression line. Did the slope change much?

💻 Question 11: Model `Length` by `Width`. Guess the correlation and then find it using `cor`.

```
> detach(Cars93)
```