Rayanna Harduarsingh

October 14th, 2020

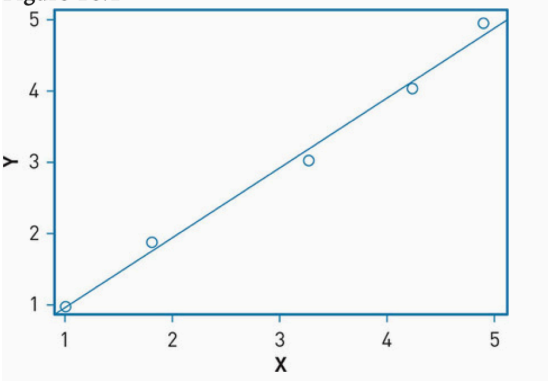
IST 687

Module 8 Notes

Chapter 16/Module 8: Linear Models

- key aims of data science is to find relationships between sets of data
- Prediction models are created from a statistical analysis process that analyze data in which the user supplies and then calculates a set of numerical coefficients that help us with prediction.
- Linear Modeling (or linear regression) used for prediction
    - A line displaying a set of data points that represents the connection between an independent variable and a dependent variable.

Figure 16.1



• This uses the best fitting line method.

• Shows how the relationship between an input (independent) variable on the horizontal X-axis relates to the output (dependent) values on the Y-axis.

• The output variable is dependent (is a function of) the independent variable."

- **"Correlation does not mean causation."**

TABLE 16.1

	oilChanges	repairs	miles
1	3	300	20100
2	5	300	23200
3	2	500	19200
4	3	400	22100
5	1	700	18400
6	4	420	23400
7	6	100	17900
8	4	290	19900
9	3	475	20100
10	2	620	24100
11	0	600	18200
12	10	0	19600
13	7	200	20800
14	8	50	19700

• **Independent: Miles and oilChange**s

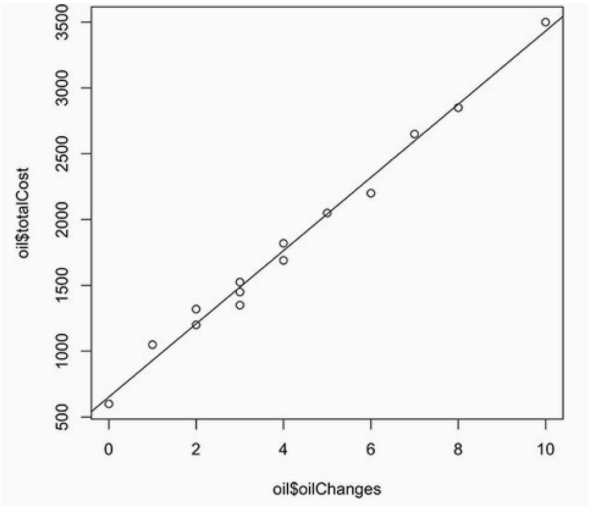
•These variables stand alone and isn't influenced by the other variables.

• **Dependent: Repairs**

•Repairs are measured & affected by the Miles and oil changes variables. Hence, this is the variable we're trying to predict.

R Code

- Showing the plot of points:
    - > plot(oil$oilChanges, oil$repairs)
- Building a linear model:
    - > model1 <- lm(formula=repairs ~ oilChanges, data=oil)
        - The lm() command places its output in a data structure.
        - The squiggly line [~], which is called a tilde character, is part of the syntax that tells lm() which independent and dependent variables to include in the model.
- Linear model using the number of oilChanges to predict totalCost and plots the results:
    - > oil$oilChangeCost <- oil$oilChanges * 350
    - > oil$totalCost <- oil$oilChangeCost + oil$repairs
    - > m <- lm(formula=totalCost ~ oilChanges, data=oil)
    - > plot(oil$oilChanges, oil$totalCost)
    - > abline(m)



• Analysis shows that we shouldn't do any oil changes.

R Functions Used in This Chapter

abline()	Plots a best-fitting line on top of a scatter plot.
lm()	Stands for linear models and, for this chapter, multiple regression.
predict()	Uses a model to predict a variable (output).
plot()	Is a general purpose graphing function that has many uses in R.
ggplot()	Uses geom_point and stat_smooth.
View()	Shows a dataframe in an easy-to-read format.

Question: How did you come up with the 350 number in this command below?

> oil$oilChangeCost <- oil$oilChanges * 350