Problem Set 1b

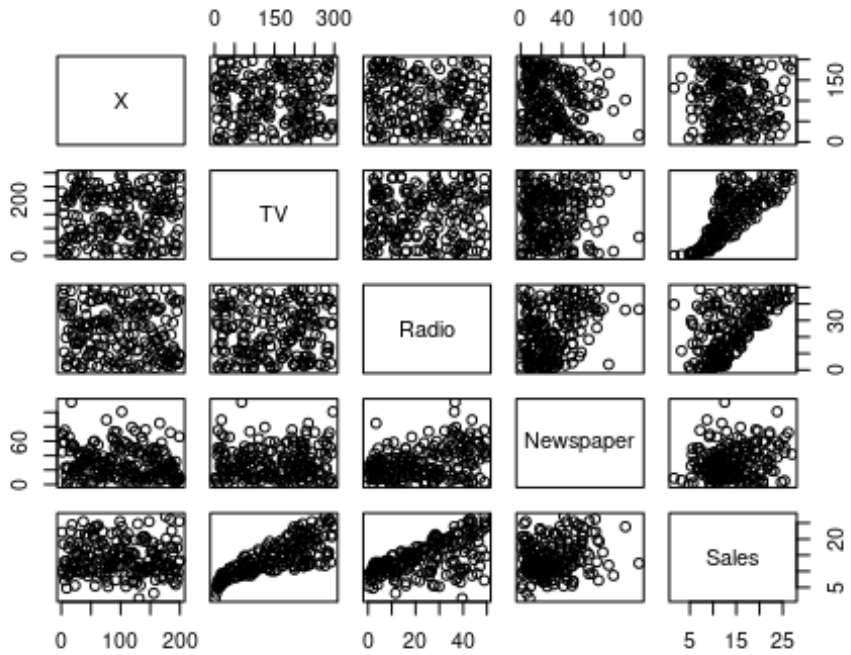
Anirudh Penmatcha

2023-09-12

```
library(car)

##################################

# Part 1: R Questions

##################################

# Question 1: Loading, summarizing and plotting the dataset

dataframe <- read.csv("Advertising.csv")
summary(dataframe)
```

```
##        X                TV              Radio          Newspaper     
##  Min.   :  1.00   Min.   :  0.70   Min.   : 0.000   Min.   :  0.30  
##  1st Qu.: 50.75   1st Qu.: 74.38   1st Qu.: 9.975   1st Qu.: 12.75  
##  Median :100.50   Median :149.75   Median :22.900   Median : 25.75  
##  Mean   :100.50   Mean   :147.04   Mean   :23.264   Mean   : 30.55  
##  3rd Qu.:150.25   3rd Qu.:218.82   3rd Qu.:36.525   3rd Qu.: 45.10  
##  Max.   :200.00   Max.   :296.40   Max.   :49.600   Max.   :114.00  
##      Sales      
##  Min.   : 1.60  
##  1st Qu.:10.38  
##  Median :12.90  
##  Mean   :14.02  
##  3rd Qu.:17.40  
##  Max.   :27.00
```

```
plot(dataframe)
```




```
###################################

# Question 2: Simple Linear Regression

# Yes, there is a relationship between sales and the mediums of
advertisement.
# TV and Sales have a clear linear relationship. With more advertisements on
TV, the sales are almost proportionally high.
# Radio and Sales also share somewhat of a relationship, however, it isn't as
linear as with TV and Sales.
# Newspaper and Sales don't show much of a relationship. Which means
investing much of the advertising budget in Newspapers will not be
worthwhile.

print(dataframe[5])

##     Sales
## 1    22.1
## 2    10.4
## 3     9.3
## 4    18.5
## 5    12.9
## 6     7.2
## 7    11.8
## 8    13.2
## 9     4.8
```

```
## 10   10.6
## 11    8.6
## 12   17.4
## 13    9.2
## 14    9.7
## 15   19.0
## 16   22.4
## 17   12.5
## 18   24.4
## 19   11.3
## 20   14.6
## 21   18.0
## 22   12.5
## 23    5.6
## 24   15.5
## 25    9.7
## 26   12.0
## 27   15.0
## 28   15.9
## 29   18.9
## 30   10.5
## 31   21.4
## 32   11.9
## 33    9.6
## 34   17.4
## 35    9.5
## 36   12.8
## 37   25.4
## 38   14.7
## 39   10.1
## 40   21.5
## 41   16.6
## 42   17.1
## 43   20.7
## 44   12.9
## 45    8.5
## 46   14.9
## 47   10.6
## 48   23.2
## 49   14.8
## 50    9.7
## 51   11.4
## 52   10.7
## 53   22.6
## 54   21.2
## 55   20.2
## 56   23.7
## 57    5.5
## 58   13.2
## 59   23.8
```

```
## 60   18.4
## 61    8.1
## 62   24.2
## 63   15.7
## 64   14.0
## 65   18.0
## 66    9.3
## 67    9.5
## 68   13.4
## 69   18.9
## 70   22.3
## 71   18.3
## 72   12.4
## 73    8.8
## 74   11.0
## 75   17.0
## 76    8.7
## 77    6.9
## 78   14.2
## 79    5.3
## 80   11.0
## 81   11.8
## 82   12.3
## 83   11.3
## 84   13.6
## 85   21.7
## 86   15.2
## 87   12.0
## 88   16.0
## 89   12.9
## 90   16.7
## 91   11.2
## 92    7.3
## 93   19.4
## 94   22.2
## 95   11.5
## 96   16.9
## 97   11.7
## 98   15.5
## 99   25.4
## 100  17.2
## 101  11.7
## 102  23.8
## 103  14.8
## 104  14.7
## 105  20.7
## 106  19.2
## 107   7.2
## 108   8.7
## 109   5.3
```

```
## 110  19.8
## 111  13.4
## 112  21.8
## 113  14.1
## 114  15.9
## 115  14.6
## 116  12.6
## 117  12.2
## 118   9.4
## 119  15.9
## 120   6.6
## 121  15.5
## 122   7.0
## 123  11.6
## 124  15.2
## 125  19.7
## 126  10.6
## 127   6.6
## 128   8.8
## 129  24.7
## 130   9.7
## 131   1.6
## 132  12.7
## 133   5.7
## 134  19.6
## 135  10.8
## 136  11.6
## 137   9.5
## 138  20.8
## 139   9.6
## 140  20.7
## 141  10.9
## 142  19.2
## 143  20.1
## 144  10.4
## 145  11.4
## 146  10.3
## 147  13.2
## 148  25.4
## 149  10.9
## 150  10.1
## 151  16.1
## 152  11.6
## 153  16.6
## 154  19.0
## 155  15.6
## 156   3.2
## 157  15.3
## 158  10.1
## 159   7.3
```

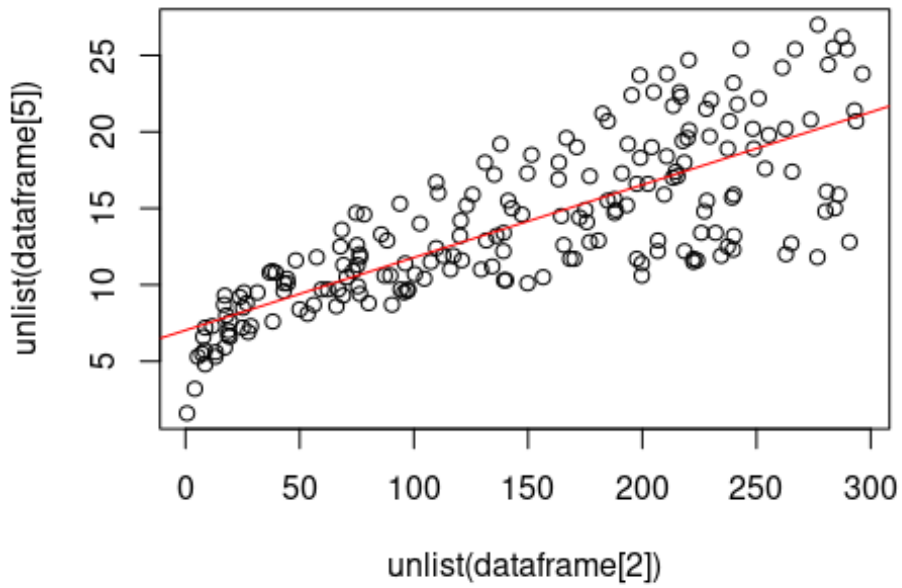
```
## 160  12.9
## 161  14.4
## 162  13.3
## 163  14.9
## 164  18.0
## 165  11.9
## 166  11.9
## 167   8.0
## 168  12.2
## 169  17.1
## 170  15.0
## 171   8.4
## 172  14.5
## 173   7.6
## 174  11.7
## 175  11.5
## 176  27.0
## 177  20.2
## 178  11.7
## 179  11.8
## 180  12.6
## 181  10.5
## 182  12.2
## 183   8.7
## 184  26.2
## 185  17.6
## 186  22.6
## 187  10.3
## 188  17.3
## 189  15.9
## 190   6.7
## 191  10.8
## 192   9.9
## 193   5.9
## 194  19.6
## 195  17.3
## 196   7.6
## 197   9.7
## 198  12.8
## 199  25.5
## 200  13.4
```

```r
# Running a simple regression over each of the variables
lm_model_TV <- lm(unlist(dataframe[5]) ~ unlist(dataframe[2]), data =
dataframe)
summary(lm_model_TV)
```

```
##
## Call:
## lm(formula = unlist(dataframe[5]) ~ unlist(dataframe[2]), data =
```

```
dataframe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           7.032594   0.457843   15.36   <2e-16 ***
## unlist(dataframe[2]) 0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

```
plot(unlist(dataframe[2]), unlist(dataframe[5]))
abline(lm_model_TV, col = "red")
```

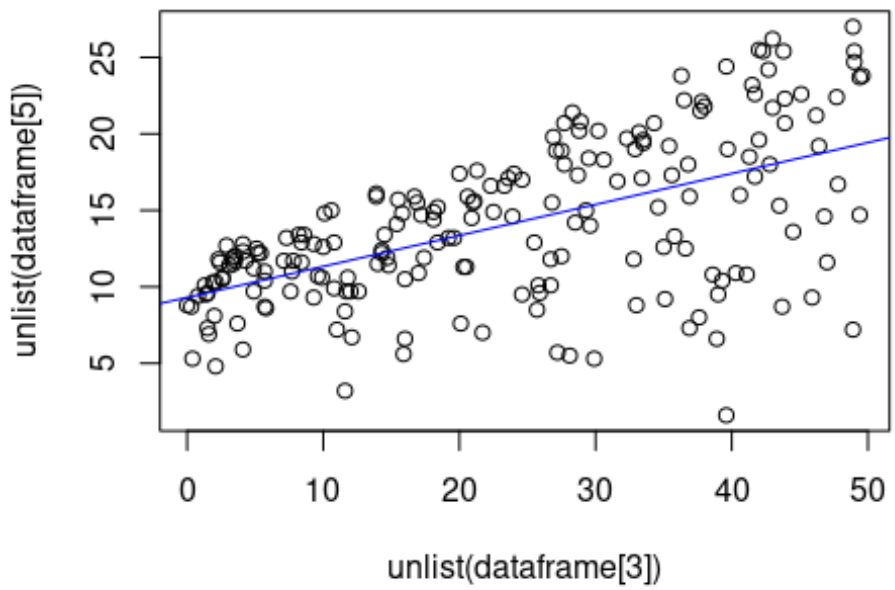



```
lm_model_RADIO <- lm(unlist(dataframe[5]) ~ unlist(dataframe[3]), data =
dataframe)
summary(lm_model_RADIO)
```

```
##
## Call:
## lm(formula = unlist(dataframe[5]) ~ unlist(dataframe[3]), data =
```

```
dataframe)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7305  -2.1324   0.7707   2.7775   8.1810
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           9.31164    0.56290  16.542   <2e-16 ***
## unlist(dataframe[3])  0.20250    0.02041   9.921   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.275 on 198 degrees of freedom
## Multiple R-squared:  0.332,  Adjusted R-squared:  0.3287
## F-statistic: 98.42 on 1 and 198 DF,  p-value: < 2.2e-16
```

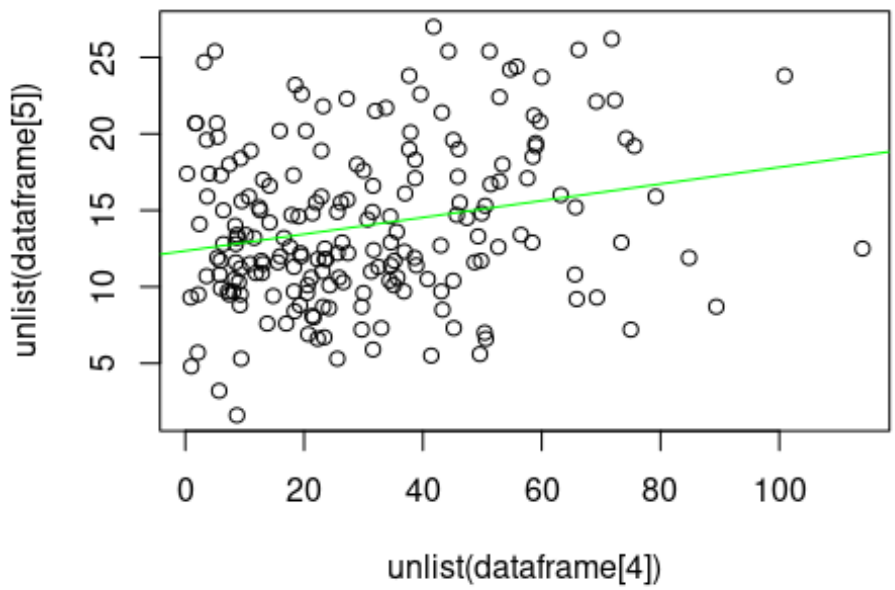
```
plot(unlist(dataframe[3]), unlist(dataframe[5]))
abline(lm_model_RADIO, col = "blue")
```




```
lm_model_NEWSPAPER <- lm(unlist(dataframe[5]) ~ unlist(dataframe[4]), data =
dataframe)
summary(lm_model_NEWSPAPER)
```

```
##
## Call:
## lm(formula = unlist(dataframe[5]) ~ unlist(dataframe[4]), data =
```

```
dataframe)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2272  -3.3873  -0.8392   3.5059  12.7751
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          12.35141    0.62142   19.88  < 2e-16 ***
## unlist(dataframe[4])  0.05469    0.01658    3.30  0.00115 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.092 on 198 degrees of freedom
## Multiple R-squared:  0.05212,    Adjusted R-squared:  0.04733
## F-statistic: 10.89 on 1 and 198 DF,  p-value: 0.001148

plot(unlist(dataframe[4]), unlist(dataframe[5]))
abline(lm_model_NEWSPAPER, col = "green")
```




```
# We see from the graphs that the coefficients of TV and Sales model have a
good fit. Radio and Sales have an average fit.
# And Newspaper and Sales has the worst fit. As for each medium's
contribution to sales, TV and Radio definitely contribute, but
# Newspaper doesn't seem to.

###################################
```

```r
# Question 3: Multiple Linear Regression

mult_lm_model <- lm(unlist(dataframe[5]) ~ unlist(dataframe[4]) +
unlist(dataframe[3]) + unlist(dataframe[2]), data = dataframe)
summary(mult_lm_model)

##
## Call:
## lm(formula = unlist(dataframe[5]) ~ unlist(dataframe[4]) +
unlist(dataframe[3]) +
##     unlist(dataframe[2]), data = dataframe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.938889   0.311908   9.422   <2e-16 ***
## unlist(dataframe[4]) -0.001037   0.005871  -0.177     0.86
## unlist(dataframe[3])  0.188530   0.008611  21.893   <2e-16 ***
## unlist(dataframe[2])  0.045765   0.001395  32.809   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16

mult_lm_model$coefficients

##          (Intercept) unlist(dataframe[4]) unlist(dataframe[3])
##          2.938889369         -0.001037493          0.188530017
## unlist(dataframe[2])
##          0.045764645

# The coefficient of newspaper is negative while TV and Radio are positive.
We also see the p-value given in the summary as
# less than 2.2e-16 which means that coefficients are statistically
significant because typically a p-value < 0.05 is considered
# statistically significant.

# Do they all contribute to sales?
# Newspaper definitely doesn't because of the negative relationship. But TV
and Radio do due to the positive coefficients.

# Reconciling results of multiple and simple regressions for newspaper
# If we look at the coefficients of the simple Linear Regression's model and
compare it with the respective coefficients of the
```

```r
# Multiple Linear Regression models, they aren't too far apart. It won't be
the exact same but will be close to each other
# because in multiple Linear Regression model, it's trying to fit it for all
the three advertising mediums.

# How strong is the relationship between advertising and sales?
# It's mostly okay because it's not the strongest with Radio and Newspaper
but if a business had to invest their budget
# into advertisements for increasing their sales, then they should do it only
in TV and Radio because they have good relationship
# with sales.

# Discussing R-squared results
# The R-Squared value is computed to be 0.8972 or 89.72% which is very good.
It means that we got a good fit and the model is
# able to accurately predict the output for 90% of the data. However, it is
also important to keep in mind to use other
# metrics

# Plotting a 3d graph of Sales, TV and Radio.

#scatter3d(Sales~TV+Radio)

##################################
# Question 4: Models with interaction terms

lm_model_TV_Radio <- lm(unlist(dataframe[5]) ~ unlist(dataframe[3]) *
unlist(dataframe[2]), data = dataframe)
summary(lm_model_TV_Radio)

##
## Call:
## lm(formula = unlist(dataframe[5]) ~ unlist(dataframe[3]) *
unlist(dataframe[2]),
##     data = dataframe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3366 -0.4028  0.1831  0.5948  1.5246
##
## Coefficients:
##                                          Estimate Std. Error t value
Pr(>|t|)
## (Intercept)                             6.750e+00  2.479e-01  27.233
<2e-16
## unlist(dataframe[3])                    2.886e-02  8.905e-03   3.241
0.0014
## unlist(dataframe[2])                    1.910e-02  1.504e-03  12.699
<2e-16
```

```
## unlist(dataframe[3]):unlist(dataframe[2]) 1.086e-03  5.242e-05  20.727
<2e-16
##
## (Intercept)                                ***
## unlist(dataframe[3])                       **
## unlist(dataframe[2])                       ***
## unlist(dataframe[3]):unlist(dataframe[2]) ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9435 on 196 degrees of freedom
## Multiple R-squared:  0.9678, Adjusted R-squared:  0.9673
## F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16

# R-squared = 0.9678 (or) 96.78; F-statistic = 1963
# It seems like the R-Squared has gone up by a lot more. And the F-statistic
is much higher which means it is
# statistically significant and does a much better job of explaining the
variation in the dependent variable, which means it estimates the output
# quite precisely. So yes, there is a lot of synergy between TV and Radio due
to the improved performance that we've observed.

# Experimenting with variations in interaction terms
lm_model_TV_Newspaper <- lm(unlist(dataframe[5]) ~ unlist(dataframe[4]) *
unlist(dataframe[2]), data = dataframe)
summary(lm_model_TV_Newspaper)

##
## Call:
## lm(formula = unlist(dataframe[5]) ~ unlist(dataframe[4]) *
unlist(dataframe[2]),
##     data = dataframe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1860 -1.5521 -0.0648  1.8062  8.7276
##
## Coefficients:
##                                             Estimate Std. Error t value
Pr(>|t|)
## (Intercept)                                6.4042175  0.7333818   8.732
1.1e-15
## unlist(dataframe[4])                       0.0241103  0.0192716   1.251
0.212
## unlist(dataframe[2])                       0.0426585  0.0043105   9.896  <
2e-16
## unlist(dataframe[4]):unlist(dataframe[2]) 0.0001324  0.0001079   1.228
0.221
##
## (Intercept)                                ***
```

```
## unlist(dataframe[4])
## unlist(dataframe[2])                        ***
## unlist(dataframe[4]):unlist(dataframe[2])
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.117 on 196 degrees of freedom
## Multiple R-squared:  0.6485, Adjusted R-squared:  0.6432
## F-statistic: 120.6 on 3 and 196 DF,  p-value: < 2.2e-16

# R-squared = 0.6458 (or) 64.58%; F-statistic = 120.6

lm_model_Radio_Newspaper <- lm(unlist(dataframe[5]) ~ unlist(dataframe[4]) *
unlist(dataframe[3]), data = dataframe)
summary(lm_model_Radio_Newspaper)

##
## Call:
## lm(formula = unlist(dataframe[5]) ~ unlist(dataframe[4]) *
unlist(dataframe[3]),
##     data = dataframe)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.6981  -2.1955   0.7567   2.7191   8.2228
##
## Coefficients:
##                                             Estimate Std. Error t value
## (Intercept)                                 8.7904734  1.0224848   8.597
## unlist(dataframe[4])                        0.0220611  0.0345866   0.638
## unlist(dataframe[3])                        0.2145684  0.0382985   5.603
## unlist(dataframe[4]):unlist(dataframe[3]) -0.0005259  0.0010642  -0.494
##                                            Pr(>|t|)
## (Intercept)                                2.58e-15 ***
## unlist(dataframe[4])                          0.524
## unlist(dataframe[3])                       7.08e-08 ***
## unlist(dataframe[4]):unlist(dataframe[3])     0.622
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.292 on 196 degrees of freedom
## Multiple R-squared:  0.3335, Adjusted R-squared:  0.3233
## F-statistic:  32.7 on 3 and 196 DF,  p-value: < 2.2e-16

# R-squared = 0.3335 (or) 33.35%; F-statistic = 32.7

##################################

# Question 5: Optimize sales
```

```r
# How should the budget be divided between TV & Radio?
budget_TV_Radio <- lm(unlist(dataframe[5]) ~ unlist(dataframe[3]) *
unlist(dataframe[2]), data = dataframe)
summary(budget_TV_Radio)

## *I'm not sure how to answer Question 5* ##


###################################

# Part 2: Reading

###################################

# What is the goal of Machine Learning?
# To develop high performance models that give useful predictions under
computing restraints

# What does Varian mean by "good out of sample predictions"?
# It means to get good estimates or predictions on data that the model hasn't
seen yet. Sample here is the data with which
# the model was estimated. So out of sample would mean data points outside
this sample.

# What is overfitting?
# How Varian explains this is when a model fits linear independent variables
perfectly with the training data, but don't predict
# well with data outside the training set, then the model is considered to be
overfitting the training set.

# What is model complexity?
# If we visualize a model and observe one that has overfit, it will have a
lot of depressions and curves so it touches all the
# points. However, one that is not overfit or underfit, will look less
twisted and bent with a fit which can be considered a good
# one. So these are different complexities in models.

# What is the training data?
# The training data is the dataset with which we estimate our model
```