Introduction to predictive modeling

Poojitha Gowthaman

22 July 2021

This is my baby step into predictive modeling and my very first introduction to how a simple equation of a line

$$Y = mx + c$$

can pretty much rule the world of Supervised Learning. This file is on a Q&A format. This helped me approach data science and introductory statistically modeling with an outlook - What is the problem that we are looking to solve?

QUESTION 1

Provide an example where classification would be used as the analysis. Describe the predictor(s), the response variable, and explain whether prediction or inference would be of primary interest.

Answer

We are looking for models that have a qualitative response for Classification. We make a decision based on a set of parameters that help us arrive at it. Lets take a data that holds weather information for 5 years - Temperature, Humidity and Wind. Based on a combination of temperature, humidity, the wind and if it's is sunny, cloudy or raining, we devise rules if we can play golf the next day or not.

The response variable here is Y - Decision on whether I play Golf or Not and/or if I need a jacket depending on the temperature.

The predictor variable here are the inputs - Temp, humidity or wind

Prediction is of primary interest here, because I would want to know if tomorrow is a good day to play Golf.

QUESTION 2

Provide an example (not used in class or the textbook) where regression would be used as the analysis. Describe the predictor(s), the response variable, and explain whether prediction or inference would be of primary interest.

Answer

We are looking for models that have a quantitative response for Regression. Agricultural scientists using regression to measure the effect of fertilizer and water on crop yields.

Fertilizer and water will be the **predictor variables** and crop yields will be the **response variable**. This model would be better used to infer how much water and fertilizer must be used to maximize crop yield.

QUESTION 3

Provide an example (not used in class or the textbook) where unsupervised learning (categorical response) would be used as the analysis. Describe the predictor(s), the response variable, and explain whether prediction or inference would be of primary interest.

Answer Unsupervised Learning is used in many sectors like Retail, Medical Imaging and Recommended Systems.

We can relate this to our everyday peeves on Amazon and Netflix. They gives us suggestions based our watch history. The shows are categorized into genres that we most watch and based on those clusters they would suggest the next possible series which could be of interest to us. These kind of systems would fall into both prediction and inference. From the supplier's point of view (Amazon or Netflix) they will have to make a fair amount of inference from the data to choose certain suggestions. And from the end user's perspective, it will be as though the model is predicting the next nest shows for me to watch. The predictors in this case would be all the genre of shows/sitcoms or movies that have a cluster. The response would be my probability of choosing the show that was suggested to watch next! A combination of both the cases gives a quality feedback for an unsupervised model.

QUESTION 4

Suppose we knew that the Bayes' classifier for a bivariate predictor space was a boundary that was extremely linear. If we were using K-nearest neighbors as a classifier, would you expect a larger or smaller value of K to provide a better approximation of the boundary? Explain.

Answer A larger K value would provide a better approximation of the boundary if it was extremely linear. We hope that our model gives a decision boundary that looks exactly like a Bayes classifier. We now have a low-variance and high-bias classifier. As K increases, the KNN classifier tends to form many clusters along the boundary to better approximate the Bayes decision line. However we should stop at the "best" value of K rather increasing the value too high to avoid overfitting. Our aim is to reduce the test error rate of the model and get the best accuracy.

As k increases for k-nearest neighbours, we will see increasingly simple boundaries. Therefore, if we have a simple (extremely linear) boundary, then we would expect a relatively large value of k to perform better as we require a less flexible version of the kNN model. Note: as always, there are caveats here depending on the interpretation of the question and things like sample size, etc.

QUESTION 5

My first dataset, I guess its been the same for generations - CarPrices.csv.

```r
setwd("/Volumes/Learnings/workspace/projects/my_modeling_learnings/")
CarPrices <- read.csv("CarPrice.csv")
head(CarPrices)
```

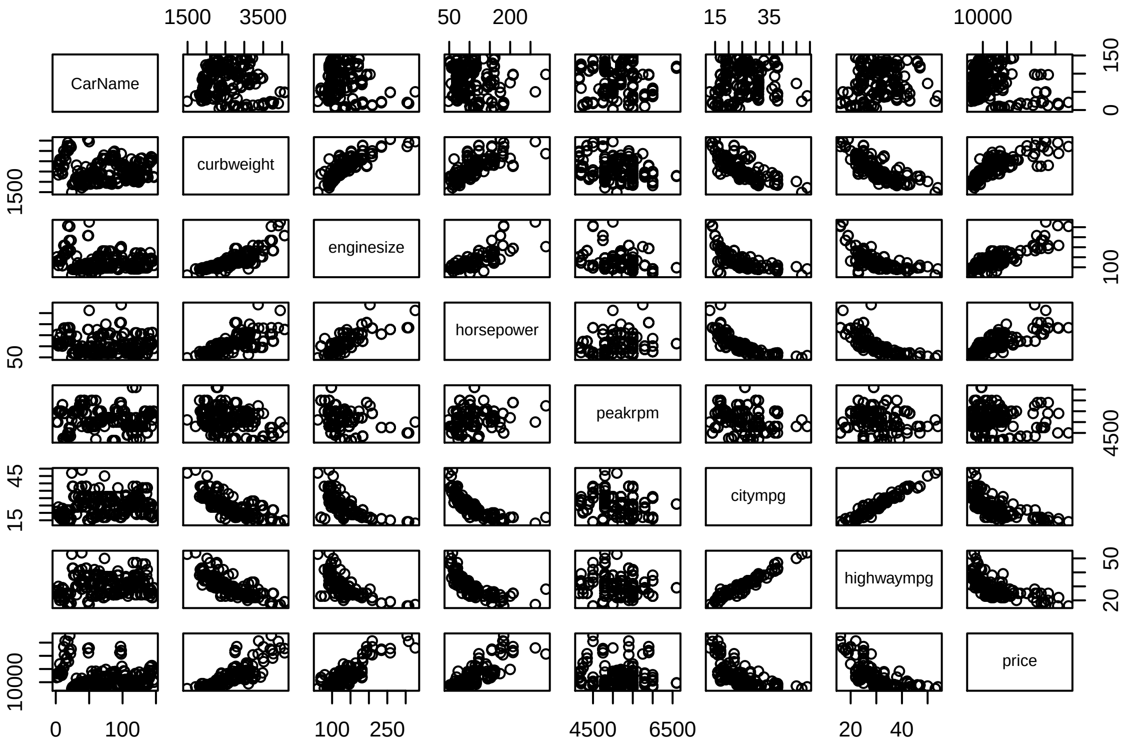
```
##                     CarName curbweight enginesize horsepower peakrpm citympg
## 1        alfa-romero giulia       2548        130        111    5000      21
## 2       alfa-romero stelvio       2548        130        111    5000      21
## 3 alfa-romero Quadrifoglio       2823        152        154    5000      19
## 4               audi 100 ls       2337        109        102    5500      24
## 5               audi 100ls       2824        136        115    5500      18
## 6                 audi fox       2507        136        110    5500      19
##   highwaympg price
## 1         27 13495
## 2         27 16500
## 3         26 16500
## 4         30 13950
## 5         22 17450
## 6         25 15250
```

#Question 5(a) Fit a simple linear model in R that attempts to predict the price of a car using only horsepower. Provide a summary of the linear model generated by R.

```
data(CarPrices)
```

```
## Warning in data(CarPrices): data set 'CarPrices' not found
```

```
plot(CarPrices)
```



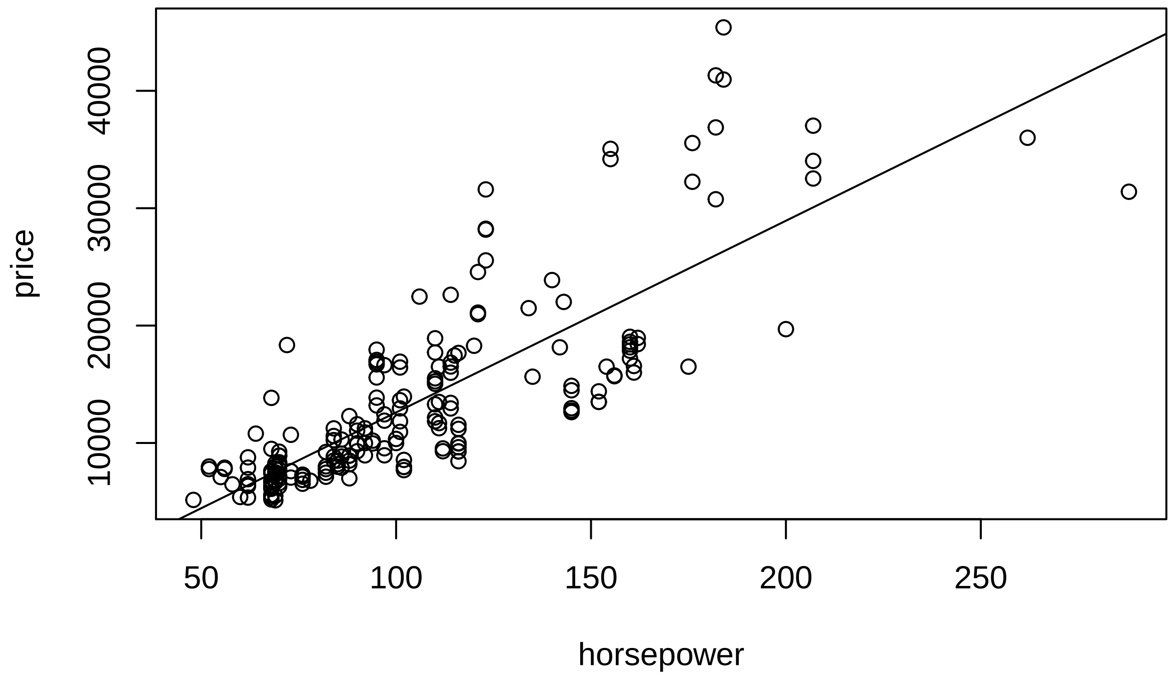
```
attach(CarPrices)
carlm <- lm(price~horsepower)
summary(carlm)
```

```
##
## Call:
## lm(formula = price ~ horsepower)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11897.5  -2350.4   -711.1   1644.6  19081.4
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3721.761    929.849  -4.003 8.78e-05 ***
## horsepower    163.263      8.351  19.549  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4717 on 203 degrees of freedom
## Multiple R-squared:  0.6531, Adjusted R-squared:  0.6514
## F-statistic: 382.2 on 1 and 203 DF,  p-value: < 2.2e-16
```

Question 5(b)

Plot the linear model. Do you have any concerns regarding the inferential assumptions?

```r
plot(price ~ horsepower, data=CarPrices)
abline(lm(price ~ horsepower, data=CarPrices,))
```



Concerns regarding Inferential Assumptions

- We would generally assume that the distance of the error ϵ_i has a constant variance. However looking at the plot above, there are some concerns with this assumption for the data distributed. It deviates outwards for higher values of horsepower and price. It definitely some noise in the higher values. They do not lie parallel to the regression line.
- The next assumption of ϵ_i being normally distributed needs more evidence to be verified. The error is

approximately normal for the model.

Question 5(c)

Write the fitted equation of the line. **Fitted Equation**

price = -3721.761 + 163.263 * Horsepower

- Intercept: When the horsepower is zero, the value of price of the car stands at -3721.761, which makes no sense absolutely. This may not be a very useful information for this case.
- Slope : 163.263. This is the coefficient estimate for the horsepower added. This tells us the change in prices caused by one unit increase in horsepower.
- The SE of the coefficient of horsepower estimate is observed to be 8.351,, which gives the idea of the scale of the variability of the estimate $\beta_1 cap$, which is 163.263 here but will vary with a standard deviation of approximately 8.351 around the true, unknown value of β_1 if we repeat the whole experiment many times. This level of variability accounts for a high bias in multiple samples.

Question 5(d)

Perform a complete hypothesis test for slope with alpha = 0.05 using the inferential assumptions referred to in lecture. If you had concerns in part (b), refer to them in the "Interpretation" part of the test.

1. **Hypothesis at 95% CI**

- Null Hypothesis

$$H_0 : \beta_1 = 0$$

*Alternate Hypothesis

$$H_1 : \beta_1 \neq 0$$

(2).**Assumptions**

The general assumptions associated with LM Models are:

- There exists some linear relationship between price and horsepower and not any other relation. In Caprices, we do observe a positive linear relationship.
- We would generally assume that the distance of the error ϵ_i has a constant variance.
- The error ϵ_i is normally distributed along the regression
- The error ϵ_i is independent of each other and the horsepower.

(3) **Test Statistic**

```
t = (163.263 - 0)/8.351
t
```

```
## [1] 19.55011
```

The value above is close to what is observed in the Summary of the Caprices, where t value for horsepower = 19.549. The P Value for the CarPrices Model is <2e-16. This value is extremely low when compared with the significance level of $\alpha = 0.05$ on a 95% confidence interval.

(4) **Decision**

Since the p value is extremely low and less than 0.05, we will reject the null hypothesis. There is enough evidence that price of cars is affected by changes in horsepower.

(5) **Interpretation**

- Since our decision is to reject the Null Hypothesis, we can strongly say that changes in horsepower are linearly associated with changes in prices.

- The intercept is a -negative value. We cannot have a negative price value for a car with zero horsepower. And We wouldn't interpret a lot around β_0 since our values are not clustered around zero in the CarPrices.
- we do observe the data to be scattered, we still have some unmeasured factors that influence the price of the cars.

Question 5(e)

Provide the R2 value and interpret it.

The R^2 value is 0.6531. This tells us how well we have fit the model. We are explaining majority of the variation of the price using just horsepower in the model.

65.31% is the statistical measure that represents the proportion of the variance for a price variable that's explained by an horsepower variable.

Question 6(f)

Is this model useful? Explain your reasoning.

The model is not entirely useful. We do see a lot of noise along the abline. There could be other factors that could suggest a strong relationship with the price of the car. We could use multiple linear regression to check this further. After viewing the plot in part (b) and finding the R2 value in part (e) is quite high, I can certainly say that the model is useful (in the sense that it appears this model would aid our prediction for price). That said, given the discussion of potential violations of the inferential assumptions, I would also say that we could improve on the model (for example, some transformation on our X variable might bring things closer in line to the assumptions).