An Introduction to Statistical Learning with Applications in R

—

Second Edition

—

Gareth James, Daniela Witten, Trevor Hastie, & Robert Tibshirani

Graham Strickland

June 9, 2025

2 Statistical Learning

2.1

(a) Given a very large sample size n and a small number of predictors p, an inflexible method would be better than a flexible one, since the risk of overfitting is less.

(b) For the same reasons as (a), a flexible method would yield better results for small n and large p.

(c) A flexible method would yield better results, since non-linear functions cannot be accurately modelled by linear functions.

(d) If there is high variance in the error terms, an inflexible method would be better, since a flexible method would introduce even more variance in the values of $\hat{f}$.

2.2

(a) This is a classification problem, since we are trying to identify a qualitative trend in the data. It is an inference problem, since we are not trying to estimate future values of f. In this case, we have $n = 500$ and $p = 4$.

(b) This is a classification problem, since we are trying to classify the product as either a success or a failure. It is also a prediction problem, since we are looking to estimate a future output. We have $n = 20$ and $p = 14$.

(c) This is a regression problem since we have quantitative data and assume that it fits some function f, which we are attempting to estimate. Since this is a future estimate, it is a prediction problem. We have $n = 52$ and $p = 4$.

2.3

(a) We have an R plot of various error curves in Figure 1: Arbitrary polyno-

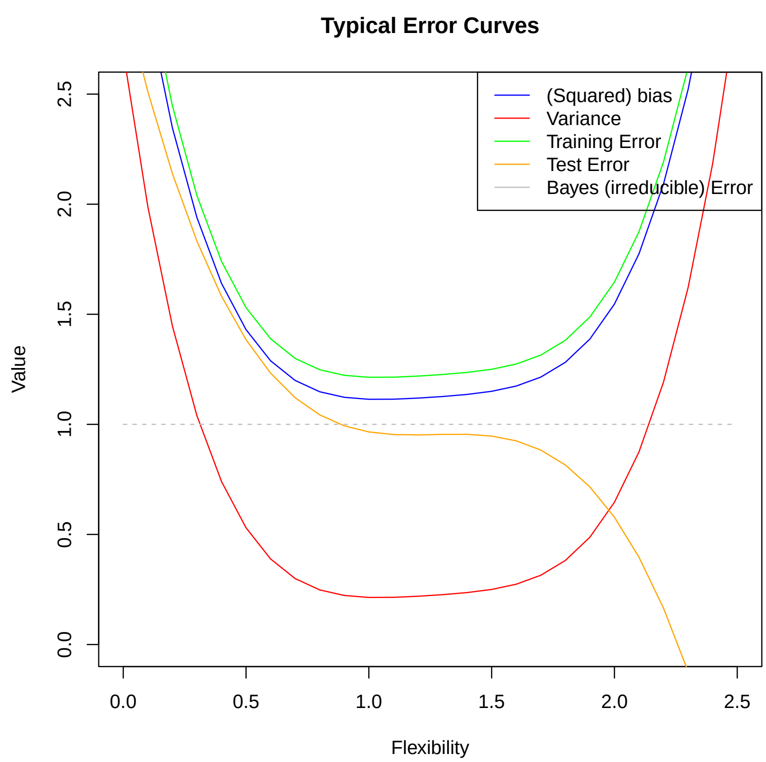

Figure 1: Typical (squared) bias, variance, training error, test error, and Bayes (irreducible) error

mials of the appropriate degree were used to generate these plots, since they are approximations, using the following script:

```
# Exercise 2.3(a) - Plots of error curves

squared_bias <- function(x) {
```

```
  return((x - 1.25)^4 + 0.01 * x^3 + 0.05 * x^2 - 0.1 * x + 1.15)
}

variance <- function(x) {
  return((x - 1.25)^4 + 0.01 * x^3 + 0.05 * x^2 - 0.1 * x + 0.25)
}

training_err <- function(x) {
  return((x - 1.25)^4 + 0.01 * x^3 + 0.05 * x^2 - 0.1 * x + 1.25)
}

test_err <- function(x) {
  return(-(x - 1.25)^3 + 0.05 * x^2 - 0.1 * x + 1.0)
}

bayes_err <- function(x) {
  return(rep(1.0, length(x)))
}

x <- seq(0, 2.5, by = 0.1)

y1 <- squared_bias(x)
y2 <- variance(x)
y3 <- training_err(x)
y4 <- test_err(x)
y5 <- bayes_err(x)

pdf("ex2_3_a.pdf")

plot(x, y1,
  type = "l", col = "blue", lwd = 1, ylim = c(0, 2.5), xlab = "Flexibility",
  ylab = "Value", main = "Typical Error Curves"
)

lines(x, y2, col = "red", lwd = 1)
lines(x, y3, col = "green", lwd = 1)
lines(x, y4, col = "orange", lwd = 1)
lines(x, y5, col = "gray", lwd = 1, lty = 2)

legend("topright",
  legend = c(
    "(Squared) bias", "Variance", "Training Error", "Test Error",
    "Bayes (irreducible) Error"
  ),
  col = c("blue", "red", "green", "orange", "gray"), lwd = 1
)

dev.off()
```

(b) We are given that the training error will always be greater than the (squared) bias, and that both will be u-shaped. The variance will always be less than the (squared) bias, while the test error will decrease as the model becomes more flexible. The Bayes error is irreducible, so it remains constant.

2.4

(a) Three real-life applications in which classification might be useful include:

(i) Determining a voter's party preference based upon the results of a census. The response would be a classification of a voter as being likely to vote for one of the possible political parties and the predictors would be features such as age, demographic, or location, which could be quantitative or qualitative. The goal in this case would be prediction, since we would most likely want to determine their vote in the next election.

(ii) Classification could also be used to determine the medical condition causing certain symptoms via medical imaging analysis. The response would be any of a number of medical conditions and the predictors would be features of the image, like abnormal textures or shapes within images provided by medical scanning. The goal would be inference, given that the condition already exists and we would like to determine its nature.

(iii) Another application would be determining a credit score for an individual. The response would be a natural number within a predetermined range and the predictors would be factors such as time taken to repay debts, number of credit lines awarded to the individual, and total accumulated debt. The goal would be prediction, since the credit score is used to determine the likelihood that the individual will repay their debts on time.

(b) Three real-life applications in which regression might be useful include:

(i) Attempting to predict the performance of a stock, given its past history would be a suitable application for regression. The response would be a numerical value indicating the expected price at a certain time and the predictors would be past values on a set of times. The goal would be prediction, given that we are estimating future performance.

(ii) Regression could also be used to determine the probability that a person will purchase a certain item from a retailer, given their past purchase history. The response would be a probability ($\Pr(X) \in [0, 1]$) and the predictors would be factors such as number of past purchases from that same retailer, spending history, credit rating, etc. The goal would also be prediction.

(iii) Another application would be determining the levels of a contaminant in a water supply, based upon readings from sources which are not directly drawn from the water supply itself, e.g., taps and waste water. The response would be a numeric value (say in mg/L) and the predictors would be the equivalent values in the other sources. The goal would be inference, since we are attempting to determine a current value.

(c) Three real-life applications in which cluster analysis might be useful include:

(i) A useful application of cluster analysis would be attempting to determine the species of certain related organisms given the degree to which they exhibit certain features.

(ii) Cluster analysis could also be used to determine whether certain geological samples fit within distinct groups based upon their chemical composition.

(iii) Another application of cluster analysis could be using segmenting consumers into certain target markets based upon their response to marketing surveys.

2.5

The advantages of a very flexible approach for regression or classification include accuracy if over-fitting has not been exhibited, while the disadvantages include interpretability, large variance in errors, and tendency for over-fitting to occur. A more flexible approach might be preferred when a set of data contains a large variation in the response and a less flexible approach would be preferable if the data tends to contain few outliers.

2.6

When utilising a parametric statistical learning approach, we attempt to estimate a countable number of parameters β_i for $i \in \{1, \ldots, p\}$ s.t. the observation (X, Y) has Y s.t.

$$Y \approx f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

This means we need to select the number of parameters used in order to minimize the error. In doing so, we risk overfitting the model to the data, so that the approximation does not closely match the form of f.

In contrast, a non-parametric method makes no assumption about the parameters used, with the advantage that we do not need to concern ourselves with the number of parameters used, but with the disadvantage that we must select a level of smoothness which makes the approximation to f easy to calculate without introducing unnecessary variation in the shape of the approximation relative to f.

2.7

(a) If we denote the number of observations by n and the number of variables by p, then we let x_{ij} denote the ith observation of the jth variable. We calculate the Euclidian distances using the formula

$$\begin{aligned} &d\big((X_1 = 0, X_2 = 0, X_3 = 0), (x_{i1}, x_{i2}, x_{i3})\big) \\ &= \sqrt{(0 - x_{i1})^2 + (0 - x_{i2})^2 + (0 - x_{i3})^2} \\ &= \sqrt{(-x_{i1})^2 + (-x_{i2})^2 + (-x_{i3})^2}, \end{aligned}$$

for $i \in \{1, \ldots, 6\}$.

Thus we have

$$\begin{aligned}
&d\big((X_1 = 0, X_2 = 0, X_3 = 0), (x_{11}, x_{12}, x_{13})\big)\\
&= \sqrt{9^2}\\
&= 9,\\
&d\big((X_1 = 0, X_2 = 0, X_3 = 0), (x_{21}, x_{22}, x_{23})\big)\\
&= \sqrt{2^2}\\
&= 2,\\
&d\big((X_1 = 0, X_2 = 0, X_3 = 0), (x_{31}, x_{32}, x_{33})\big)\\
&= \sqrt{1^2 + 3^2}\\
&= \sqrt{10}\\
&\approx 3.162278,\\
&d\big((X_1 = 0, X_2 = 0, X_3 = 0), (x_{41}, x_{42}, x_{43})\big)\\
&= \sqrt{1^2 + 2^2}\\
&= \sqrt{5}\\
&\approx 2.236068,\\
&d\big((X_1 = 0, X_2 = 0, X_3 = 0), (x_{51}, x_{52}, x_{53})\big)\\
&= \sqrt{(-1)^2 + 1^2}\\
&= \sqrt{2}\\
&\approx 1.414214,
\end{aligned}$$

and

$$\begin{aligned}
&d\big((X_1 = 0, X_2 = 0, X_3 = 0), (x_{61}, x_{62}, x_{63})\big)\\
&= \sqrt{1^2 + 1^2 + 1^2}\\
&= \sqrt{3}\\
&\approx 1.732051.
\end{aligned}$$

(b) Our prediction with $K = 1$ is that $Y =$ Green since the nearest neighbour to the point $(X_1 = 0, X_2 = 0, X_3 = 0)$ is the point given by (x_{51}, x_{52}, x_{53}) corresponding to observation 5, which yields the value $y =$ Green.

(c) With $K = 3$, we have $Y =$ Red, since of the 3 nearest neighbours calculated in (a), we have two with value $y =$ Red and only one with value $y =$ Green.

(d) We would expect the best value to be small, since this would yield a decision boundary which is highly flexible and more easily approximates the non-linear nature of the Bayes decision boundary.

2.8

(c) i. We have the following output from the `summary` function:

```
> summary(college)
 Private        Apps           Accept          Enroll       Top10perc
 No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
 Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
           Median : 1558   Median : 1110   Median : 434   Median :23.00
           Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56
           3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00
           Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00
   Top25perc      F.Undergrad     P.Undergrad         Outstate
 Min.   :  9.0   Min.   :  139   Min.   :    1.0   Min.   : 2340
 1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320
 Median : 54.0   Median : 1707   Median :  353.0   Median : 9990
 Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441
 3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925
 Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700
   Room.Board       Books           Personal         PhD
 Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00
 1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00
 Median :4200   Median : 500.0   Median :1200   Median : 75.00
 Mean   :4358   Mean   : 549.4   Mean   :1341   Mean   : 72.66
 3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00
 Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00
    Terminal       S.F.Ratio      perc.alumni        Expend
 Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186
 1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751
 Median : 82.0   Median :13.60   Median :21.00   Median : 8377
 Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660
 3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830
 Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233
   Grad.Rate
 Min.   : 10.00
 1st Qu.: 53.00
 Median : 65.00
 Mean   : 65.46
 3rd Qu.: 78.00
 Max.   :118.00
```

ii. In Figure 2, we have the scatterplot matrix of the first ten columns of `College` data.

iii. In Figure 3, we have the plots of the `Outstate` versus `Private` `College` data.

iv. We have the following output from the `summary(Elite)` function call:

```
> summary(Elite)
 No Yes
699  78
```

In Figure 4, we have the plots of the `Outstate` versus `Elite` `College` data.

v. In Figure 5, we have the histogram plots for four of the variables.

vi. We can say broadly that the elite colleges have higher out-of-state tuition costs and that the number of applications received is strongly correlated with the number of applicants accepted (obviously this is to be expected), but not as strongly correlated with the number of new students enrolled.

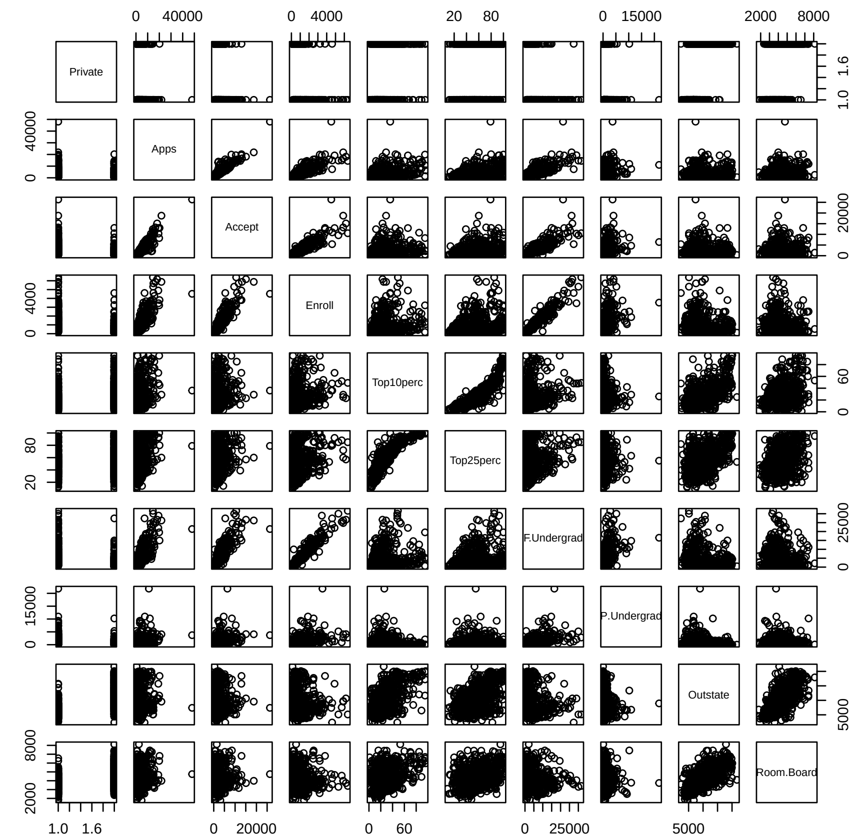

Figure 2: Scatterplot matrix of the first ten columns of College data

2.9

(a) The quantitative predictors are `mpg`, `cylinders`, `displacement`, `horsepower`, `weight`, `acceleration`, `year`, and `origin`. The only qualitative predictor is `name`.

(b) The range of each quantitative predictor is as follows:

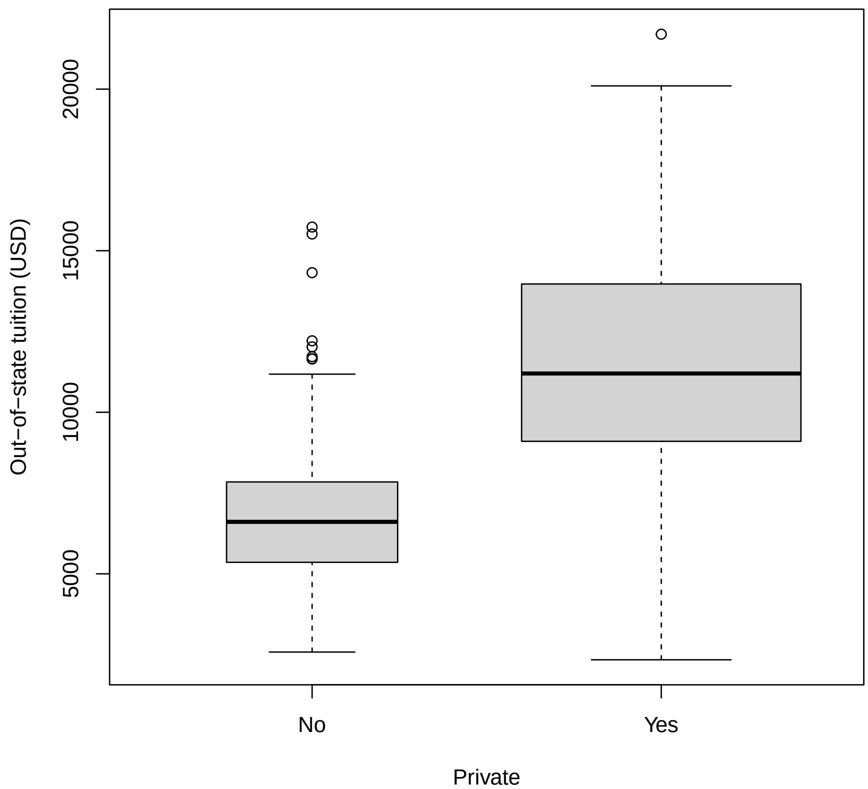

Figure 3: Outstate versus Private College data

Predictor	Minimum	Maximum
`mpg`	9.0	46.6
`cylinders`	3	8
`displacement`	68	455
`horsepower`	46	230
`weight`	1649	4997
`acceleration`	8.0	24.8
`year`	70	82
`origin`	1	3

(c) We have the following values for the mean and standard deviation:

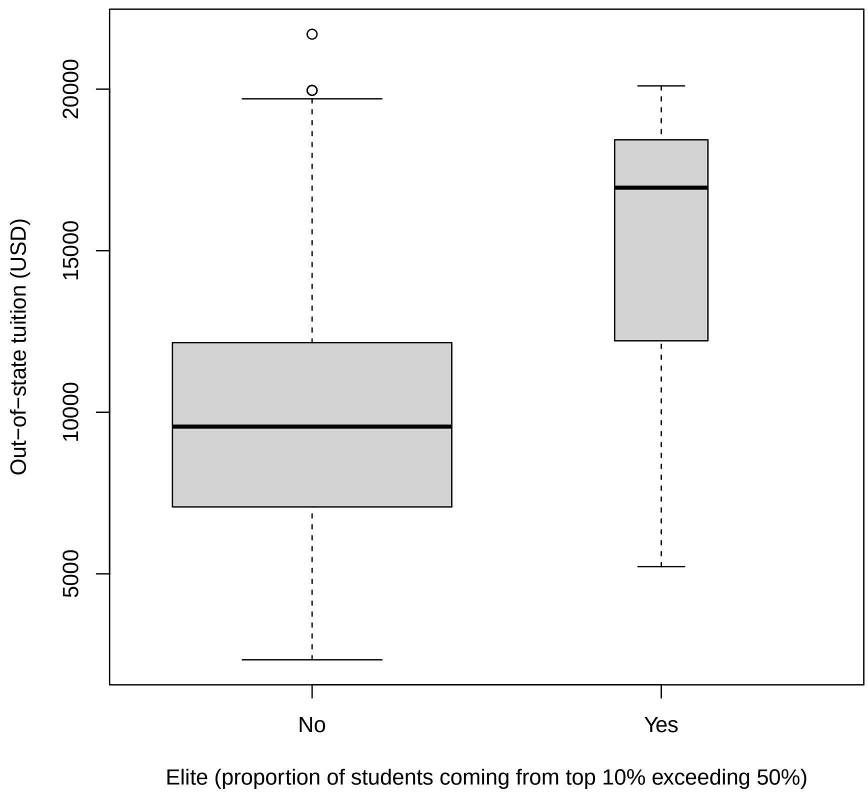

Figure 4: Outstate versus Elite College data

Predictor	Mean	Standard Deviation
`mpg`	23.44592	7.805007
`cylinders`	5.471939	1.705783
`displacement`	194.412	104.644
`horsepower`	104.4694	38.49116
`weight`	2977.584	849.4026
`acceleration`	15.54133	2.758864
`year`	75.97959	3.683737
`origin`	1.576531	0.8055182

(d) We have the following adjusted values for the range, mean, and standard deviation:

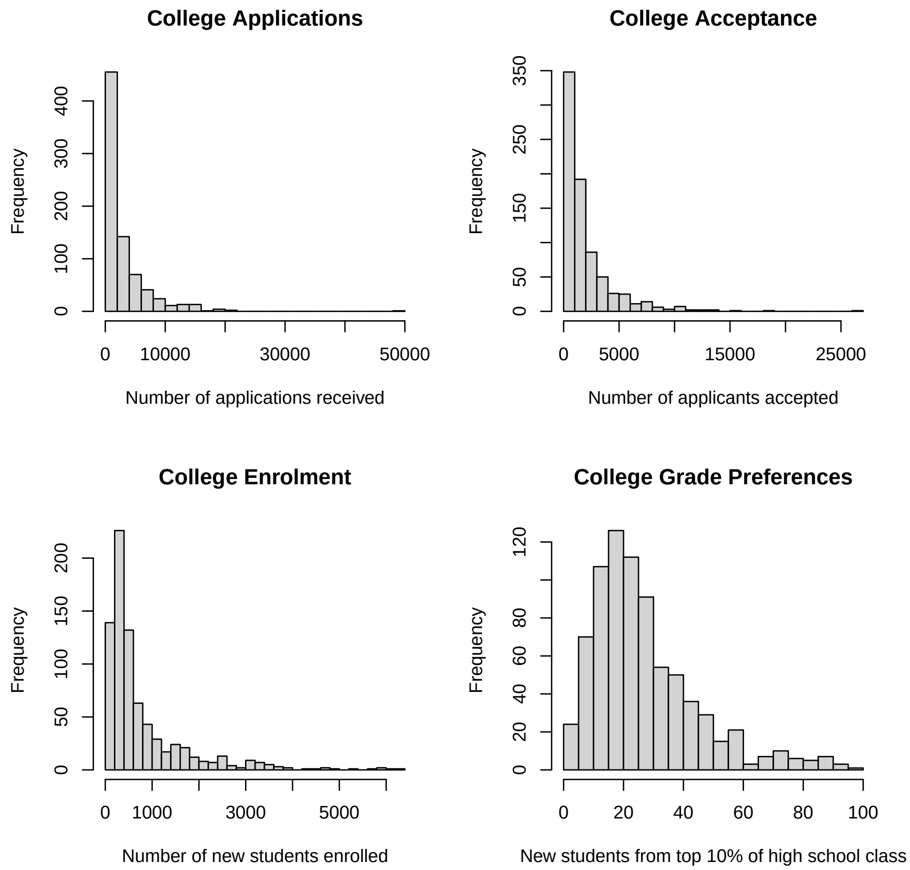

Figure 5: Plots of College data

Predictor	Min.	Max.	Mean	S.D.
`mpg`	11.0	46.6	24.40443	7.867283
`cylinders`	3	8	5.373418	1.654179
`displacement`	68	455	187.2405	99.67837
`horsepower`	46	230	100.7215	35.70885
`weight`	1649	4997	2935.972	811.3002
`acceleration`	8.5	24.8	15.7269	2.693721
`year`	70	82	77.14557	3.106217
`origin`	1	3	1.601266	0.81991

(e) In Figure 6, we have plots of the `Auto` data highlighting the relationships between some of the variables. As can be seen by these plots, there is a strong correlation between the number of cylinders and the gas mileage (`mpg`)of the vehicle in question. This relationship is somewhat the inverse

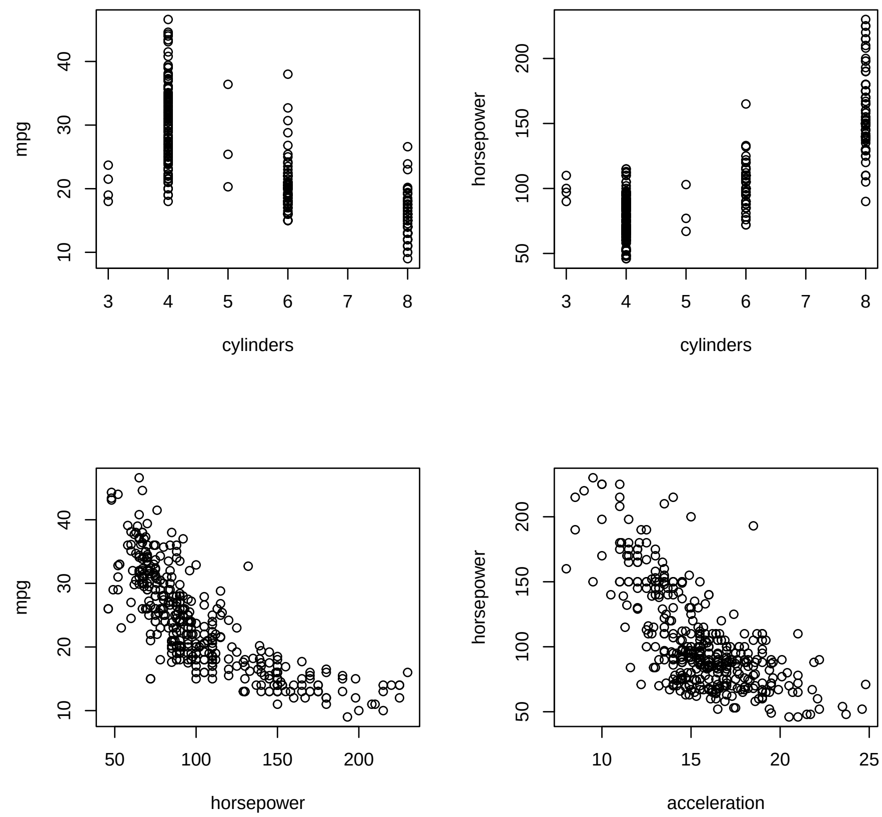

Figure 6: Scatterplots of Auto data

of that between cylinders and horsepower. We can also see that mpg and horsepower as well and, to a lesser extent, acceleration and horsepower are correlated.

(f) Yes, we can use both the number of cylinders and the horsepower to estimate the gas mileage of the vehicle, since these two variables have a strongly correlated relationship to gas mileage.

2.10

(a) We have the following output from calling `?Boston`:

```
Boston                  package:ISLR2                 R Documentation

Boston Data
```

```
Description:

     A data set containing housing values in 506 suburbs of Boston.

Usage:

     Boston

Format:

     A data frame with 506 rows and 13 variables.

     ‘crim’ per capita crime rate by town.

     ‘zn’ proportion of residential land zoned for lots over 25,000
          sq.ft.

     ‘indus’ proportion of non-retail business acres per town.

     ‘chas’ Charles River dummy variable (= 1 if tract bounds river; 0
          otherwise).

     ‘nox’ nitrogen oxides concentration (parts per 10 million).

     ‘rm’ average number of rooms per dwelling.

     ‘age’ proportion of owner-occupied units built prior to 1940.

     ‘dis’ weighted mean of distances to five Boston employment
          centres.

     ‘rad’ index of accessibility to radial highways.

     ‘tax’ full-value property-tax rate per $10,000.

     ‘ptratio’ pupil-teacher ratio by town.

     ‘lstat’ lower status of the population (percent).

     ‘medv’ median value of owner-occupied homes in $1000s.
```

(b) In Figure 7, we have some of the data from the `Boston` data set displayed in scatterplots.

From the scatterplots in this figure, we may see that as the distance to employment centres increases, the nitrogen oxide concentration decreases, perhaps indicating higher levels of nitrogen oxide in the inner city due to vehicles and industrial pollution, which coincides with where employment centres are located. Likewise, we see that owner-occupied units built prior to the year 1940 are concentrated in the city centre.

We also see that a higher number of rooms and value of owner-occupied homes tend to be concentrated amongst the lower status of the population.

(c) We can see that per capita crime rates are correlated with the lower status of the population by looking at Figure 8. It appears that high per capita crime rates are associated with a larger population in the lower status category.

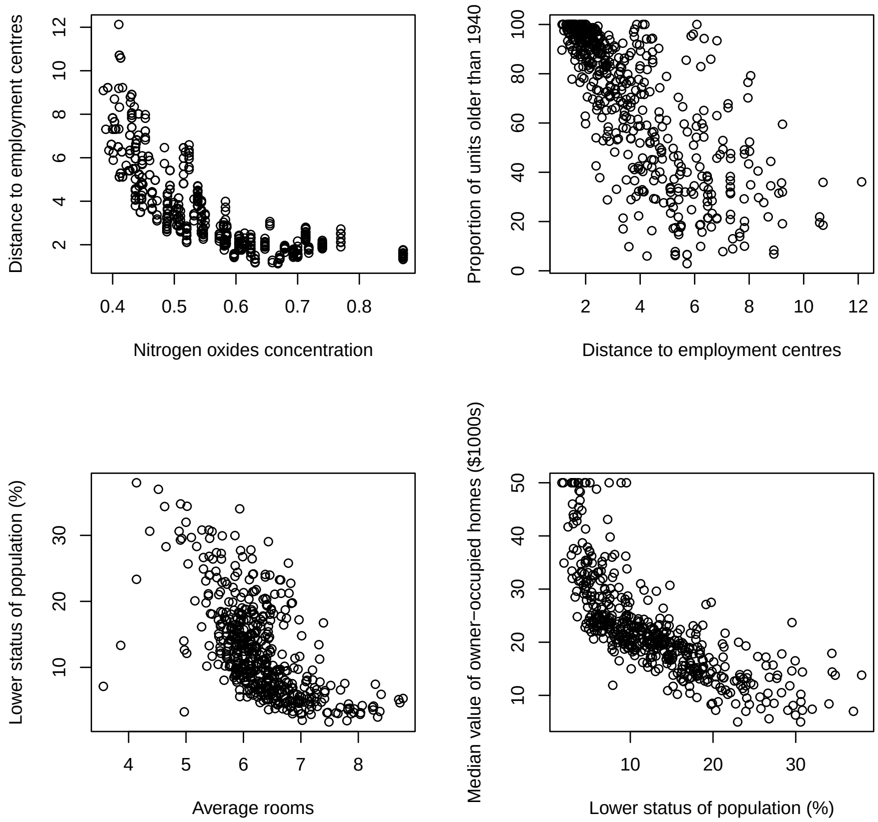

Figure 7: Scatterplots of Boston data

(d) Yes, we have the following summary:

```
> summary(Boston$crim)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
 0.00632  0.08204  0.25651  3.61352  3.67708 88.97620
```

While the median and mean for the `crim` data are 0.25651 and 3.61352, respectively, the range is $[0.00632, 88.97620]$, so that some census tracts must have particularly high crime rates. Tax rates appear to be more evenly distributed, as can be seen from the following:

```
> summary(Boston$tax)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  187.0   279.0   330.0   408.2   666.0   711.0
```

The same applies to pupil-teacher ratios:

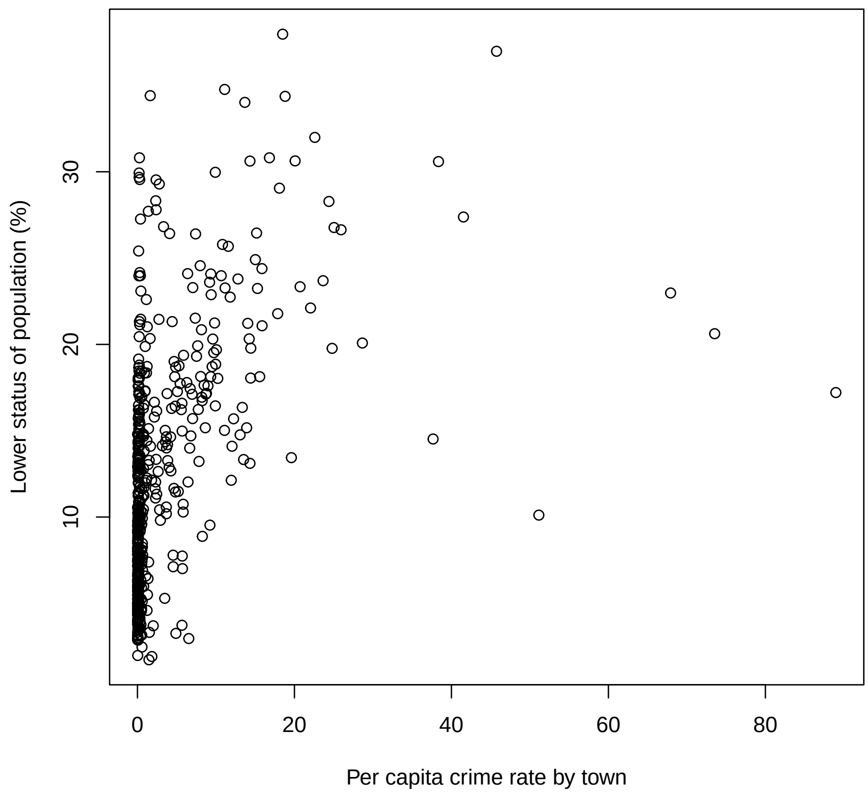

Figure 8: Scatterplot of crim vs lstat data

```
> summary(Boston$ptratio)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  12.60   17.40   19.05   18.46   20.20   22.00
```

(e) We have the following:

```
> sum(Boston$chas)
[1] 35
```

(f) We have the following:

```
> median(Boston$ptratio)
[1] 19.05
```

(g) The tract corresponding to observation $i = 399$, for which we have

```
> median(Boston$ptratio)
[1] 19.05
```

For that census tract, we have the following data:

```
> Boston[399,]
       crim zn indus chas   nox    rm age    dis rad tax ptratio lstat medv
399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2 30.59    5
```

We can see that the `crim` variable is far higher than the median value of 3.61352, and is also in the higher range for the variables `tax` and `ptratio` for which the range was determined in (d). This would likely indicate a correlation between these variables.

(h) We have the following code to find the number of census tracts which average more than 7 and 8 rooms per dwelling:

```
> sum(Boston$rm > 7)
[1] 64
> sum(Boston$rm > 8)
[1] 13
```

We have the following summary of all census tracts averaging more than 8 rooms:

```
> summary(Boston[Boston$rm > 8, ])
      crim               zn            indus             chas       
 Min.   :0.02009   Min.   : 0.00   Min.   : 2.680   Min.   :0.0000  
 1st Qu.:0.33147   1st Qu.: 0.00   1st Qu.: 3.970   1st Qu.:0.0000  
 Median :0.52014   Median : 0.00   Median : 6.200   Median :0.0000  
 Mean   :0.71879   Mean   :13.62   Mean   : 7.078   Mean   :0.1538  
 3rd Qu.:0.57834   3rd Qu.:20.00   3rd Qu.: 6.200   3rd Qu.:0.0000  
 Max.   :3.47428   Max.   :95.00   Max.   :19.580   Max.   :1.0000  
      nox               rm             age             dis       
 Min.   :0.4161   Min.   :8.034   Min.   : 8.40   Min.   :1.801  
 1st Qu.:0.5040   1st Qu.:8.247   1st Qu.:70.40   1st Qu.:2.288  
 Median :0.5070   Median :8.297   Median :78.30   Median :2.894  
 Mean   :0.5392   Mean   :8.349   Mean   :71.54   Mean   :3.430  
 3rd Qu.:0.6050   3rd Qu.:8.398   3rd Qu.:86.50   3rd Qu.:3.652  
 Max.   :0.7180   Max.   :8.780   Max.   :93.90   Max.   :8.907  
      rad              tax           ptratio          lstat           medv     
 Min.   : 2.000   Min.   :224.0   Min.   :13.00   Min.   :2.47   Min.   :21.9  
 1st Qu.: 5.000   1st Qu.:264.0   1st Qu.:14.70   1st Qu.:3.32   1st Qu.:41.7  
 Median : 7.000   Median :307.0   Median :17.40   Median :4.14   Median :48.3  
 Mean   : 7.462   Mean   :325.1   Mean   :16.36   Mean   :4.31   Mean   :44.2  
 3rd Qu.: 8.000   3rd Qu.:307.0   3rd Qu.:17.40   3rd Qu.:5.12   3rd Qu.:50.0  
 Max.   :24.000   Max.   :666.0   Max.   :20.20   Max.   :7.44   Max.   :50.0  
```

From the above summary, it becomes clear that census tracts with a large average number of rooms per dwelling have correspondingly low crime rates, relatively high tax rates, and relatively high pupil-teacher ratios.

3 Linear Regression

3.1

The four null hypotheses to which the p-values in Table 3.4 correspond are the following:

H_{0I}: There is no relationship between `sales` and β_0.

H_{0T}: There is no relationship between `sales` and `TV`.

H_{0r}: There is no relationship between `sales` and `radio`.

H_{0n}: There is no relationship between `sales` and `newspaper`.

The p-values for H_{0I}, H_{0T}, and H_{0r} are all below 0.0001, so that we can reject the null hypothesis in these four cases, and we can conclude that there do exist relationships between `sales`, β_0, `TV`, and `radio`. Since the p-value of a regression of `sales` on `newspaper` is relatively large, $p = 0.8599$, we can conclude that it is unlikely that `newspaper` alone will have an effect upon `sales`.

3.2

For the *K-nearest neighbours regression* (KNN regession), we are given a value for K and a prediction point x_0 and we first identify the K training observations that are closest to x_0. Then we estimate $f(x_0)$ using the average of all training responses in this set ($\mathcal{N}_0$), using

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in \mathcal{N}_0} y_i.$$

The *K-nearest neighbours* (KNN) classifier follows the same method, but instead of estimating $f(x_0)$ using an average, we estimate the conditional probability for class j as the fraction of points in $\mathcal{N}_0$ whose response values equal j:

$$\Pr(Y = j | X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j).$$

Then we classify the test observation x_0 to the class with the largest such probability.

Thus KNN regression estimates a function for prediction, while the KNN classifier method classifies an observation.

3.3

(a) Since $\hat{\beta}_1 = 20$, $\hat{\beta}_2 = 0.07$, and $\hat{\beta}_3 = 35$, we can conclude that IQ has little effect on starting salary by itself, while GPA and level have significant effects. The interaction term between GPA and level is given by $\hat{\beta}_5 = -10$, so that we can see a significant negative correlation between GPA and level, while $\hat{\beta}_4 = 0.01$ shows that the interaction between IP and GPA does not have a significant effect. Thus answer iv. is correct, since level will be a strong predictor, but GPA interacting with level will bring down the response if GPA is not high enough.

(b) For a college graduate with IQ of 110 and a GPA of 4.0, we have the following

$$\begin{aligned}\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5 \\ &= 50 + 20x_1 + 0.07x_2 + 35x_3 + 0.01x_4 - 10x_5 \\ &= 50 + 20(4.0) + 0.07(110) + 35(1) + 0.01(4.0 \cdot 110) - 10(4.0) \\ &\approx 137.1,\end{aligned}$$

so that the student is expected to earn a salary of \$137 100 per year.

(c) False, we would need to perform statistical tests in order to evaluate that the interaction terms have little effect, since we would then be able to compare p-values and standard errors. Thus our answer in (a) is only a guess using relative sizes, but we cannot be entirely sure that this provides the full picture without the required data.

3.4

(a) In both cases, the training RSS will be the same, since the least squares model uses the best linear approximation to the true relationship between X and Y.

(b) The test RSS will be greater for the cubic regression, since it does not model the real relationship between X and Y as well as the linear regression.

(c) The answer is the same as (a), since the training RSS will be minimised by the choice of any linear regression approach.

(d) There is not enough information, since we do not know which model more closely matches the non-linear relationship between X and Y.

3.5

Proof. We have

$$\begin{aligned}\hat{y}_i &= x_i\hat{\beta}\\ &= \frac{x_i\sum_{i'=1}^n x_{i'}y_{i'}}{\sum_{i'=1}^n x_{i'}^2}\\ &= \sum_{i'=1}^n\left(\frac{x_i x_{i'} y_{i'}}{\sum_{i'=1}^n x_{i'}^2}\right)\\ &= \sum_{i'=1}^n\left(\frac{x_i x_{i'}}{\sum_{i'=1}^n x_{i'}^2} y_{i'}\right)\\ &= \sum_{i'=1}^n a_{i'}y_{i'},\end{aligned}$$

where

$$a_{i'} = \frac{x_i x_{i'}}{\sum_{i'=1}^n x_{i'}^2},$$

for fixed i and variable i'. □

3.6

From (3.4), we have

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x},$$

so that we have

$$\begin{aligned}\hat{y}_i(\bar{x}) &= \hat{\beta}_0 + \hat{\beta}_1\bar{x}\\ &= \bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1\bar{x}\\ &= \bar{y},\end{aligned}$$

and we conclude that the least squares line passes through the point $(\bar{x}, \bar{y})$.

3.7

Proof. If we assume $\bar{x} = \bar{y} = 0$, then the simple linear regression of Y onto X is given by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Then we have

$$
\begin{aligned}
R^2 &= \frac{\text{TSS} - \text{RSS}}{\text{TSS}} \\
&= \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\
&= \frac{\sum_{i=1}^n [(y_i - \bar{y})^2 - (y_i - \hat{y}_i)^2]}{\sum_{i=1}^n (y_i - \bar{y})^2} \\
&= \frac{\sum_{i=1}^n [(y_i - \bar{y})^2 - (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\
&= \frac{\sum_{i=1}^n [(y_i - \bar{y})^2 - (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i)^2]}{\sum_{i=1}^n (y_i - \bar{y})^2} \\
&= \frac{\sum_{i=1}^n [(y_i - \bar{y})^2 - (y_i - \bar{y} + \hat{\beta}_1 [\bar{x} - x_i])^2]}{\sum_{i=1}^n (y_i - \bar{y})^2} \\
&= \frac{\sum_{i=1}^n [-2\hat{\beta}_1 (y_i - \bar{y})(\bar{x} - x_i) - \hat{\beta}_1^2 (\bar{x} - x_i)^2]}{\sum_{i=1}^n (y_i - \bar{y})^2} \\
&= \frac{\sum_{i=1}^n \hat{\beta}_1 (x_i - \bar{x})[2(y_i - \bar{y}) + \hat{\beta}_1 (\bar{x} - x_i)]}{\sum_{i=1}^n (y_i - \bar{y})^2} \\
&= \frac{\hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x})[2(y_i - \bar{y}) + \hat{\beta}_1 (\bar{x} - x_i)]}{\sum_{i=1}^n (y_i - \bar{y})^2} \\
&= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \sum_{i=1}^n (x_i - \bar{x})[2(y_i - \bar{y}) + \hat{\beta}_1 (\bar{x} - x_i)]}{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2} \\
&= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \sum_{i=1}^n (x_i - \bar{x})[2(y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})]}{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}.
\end{aligned}
$$

Now,

$$
\begin{aligned}
&2(y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})] \\
=\,&2(y_i - \bar{y}) - (x_i - \bar{x}) \frac{\sum_{j=1}^n (x_j - \bar{x})(y_j - \bar{y})}{\sum_{j=1}^n (x_j - \bar{x})^2} \\
=\,&\frac{2(y_i - \bar{y}) \sum_{j=1}^n (x_j - \bar{x})^2 - (x_i - \bar{x}) \sum_{j=1}^n (x_j - \bar{x})(y_j - \bar{y})}{\sum_{j=1}^n (x_j - \bar{x})^2} \\
=\,&\frac{2(y_i - \bar{y}) \sum_{j=1}^n (x_j - \bar{x})^2 - (y_i - \bar{y}) \sum_{j=1}^n (x_j - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2} \\
=\,&\frac{(y_i - \bar{y}) \sum_{j=1}^n (x_j - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2} \\
=\,&(y_i - \bar{y}),
\end{aligned}
$$

so that

$$\left[\mathrm{Cor}(X,Y)\right]^2 = \frac{\left[\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})\right]^2}{\sum_{i=1}^n (x_i-\bar{x})^2 \sum_{i=1}^n (y_i-\bar{y})^2}$$
$$= R^2.$$

□

3.8

(a) We have the following output:

```
> summary(lm.fit)

Call:
lm(formula = mpg ~ horsepower)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5710  -3.2592  -0.3435   2.7630  16.9240

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.935861   0.717499   55.66   <2e-16 ***
horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.906 on 390 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.6059,    Adjusted R-squared:  0.6049
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

i. Yes, it is safe to conclude that there is a relationship between the predictor and the response. We can see that from the p-values for `(Intercept)` and `horsepower`, which are very small, i.e., $p < 2.2 \times 10^{16}$, we may reject the null hypothesis.

ii. Since we have multiple R^2 and adjusted R^2 statistics of 0.6059 and 0.6049, respectively, we conclude that close to 2/3 of the variability in the response can be explained by the regression. From this, we conclude that the linear regression is a relatively good fit, and the relationship between the predictor and response is strong.

iii. Since `mpg` decreases as `horsepower` increases along the regression line, we see that the relationship between the predictor and response is negative.

iv. We have the following output:

```
> predict(lm.fit, data.frame(horsepower = 98), interval = "confidence")
       fit      lwr      upr
1 24.46708 23.97308 24.96108
```

from which we conclude that the associated 95% confidence interval is [23.97308, 24.96108] for a predicted value of 24.46708. For the prediction interval, we have

```
> predict(lm.fit, data.frame(horsepower = 98), interval = "prediction")
       fit     lwr      upr
1 24.46708 14.8094 34.12476
```

so that the interval is [14.8094, 34.12476] for the same predicted value.

(b) In Figure 9, we have a plot of the response and predictor, as well as the regression line using the `abline()` function.

(c)

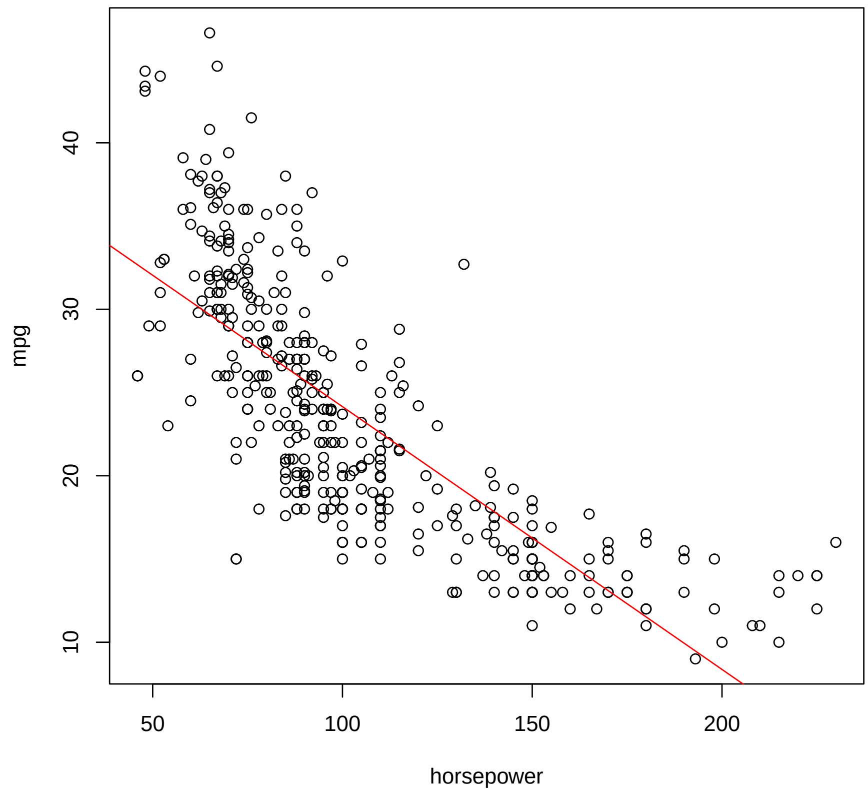

Figure 9: Response, predictor, and regression line for a linear regression of `mpg` on `horsepower`