STOR 565 Spring 2019 Homework 2

Due on 02/05/2019 in Class

Coleman Breen

Remark. This homework aims to help you go through the necessary preliminary from linear regression. Credits for **Theoretical Part** and **Computational Part** are in total 100 pts. If you receive more points that 100 (say via attempting extra credit/optional questions) then your score will be rounded to 100. **If you are aiming to get full points, it is your duty to make sure you have attempted enough problems to get 100 pts**. For **Computational Part**, please complete your answer in the **RMarkdown** file and summit your printed PDF (or doc or html) homework created by it.

Computational Part

1. (*21 pt*) Consider the dataset "Boston" in predicting the crime rate at Boston with associated covariates.

```
head(Boston)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12
##   lstat medv
## 1  4.98 24.0
## 2  9.14 21.6
## 3  4.03 34.7
## 4  2.94 33.4
## 5  5.33 36.2
## 6  5.21 28.7
```

Suppose you would like to predict the crime rate with explantory variables

- `medv` - Median value of owner-occupied homes
- `dis` - Weighted mean of distances to employement centers
- `indus` - Proportion of non-retail business acres

Run the linear model using the code below. You can do so either by copying and pasting the code into the R console, or by clicking the green arrow in the code 'chunk' (grey box where the code is written).

```
mod1 <- lm(crim ~ medv + dis + indus, data = Boston)
summary(mod1)
```

```
##
## Call:
## lm(formula = crim ~ medv + dis + indus, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.625  -3.345  -1.242   1.608  78.994
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 11.67738    2.12190   5.503 5.95e-08 ***
## medv        -0.26061    0.04204  -6.199 1.19e-09 ***
## dis         -0.96320    0.22758  -4.232 2.75e-05 ***
## indus        0.13145    0.07728   1.701   0.0896 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.519 on 502 degrees of freedom
## Multiple R-squared:  0.2404, Adjusted R-squared:  0.2358
## F-statistic: 52.95 on 3 and 502 DF,  p-value: < 2.2e-16
```

Answer the following questions.

(i) What do the following quantities that appear in the above output mean in the linear model? Provide a breif description.

- `t value` and `Pr(>|t|)` of `medv`

Answer: `t value` – this is the number of standard deviations that our model's `Estimate` is away from zero (the null hypothesis). In this case the `t value` is about -6, so the `medv` estimate is ~6 standard deviations below zero.

`Pr(>|t|)` – this is the probability of getting a `t value` as (or more) extreme than the observed. So, the probability of getting a `t value` of -6 or less is very near zero, indicating that we can reject the null hypothesis (that the estimate for `medv` is zero).

- `Multiple R-squared`

Answer:

`Multiple R-squared` is the proportion of the variance in `crim` that is explained by `medv`, `dis`, and `indus`.

- `F-statistic`, `DF` and corresponding `p-value`

Answer:

`F-statistic` is the ratio of the variance explained by the model to that explained by error. In our case, the `F-statistic` is about 53. So the linear model accounts for more than 50 times that explained by error.

`DF` is the degrees of freedom needed to calculate the `F-statisitc` from an F-distribution. The `3` comes from 3 predictor variables + 1 predictor (that's all constants) - 1. The `502` comes from 506 observations - 4 predictor variables (including one that's all constants).

`p-value` is the probability that none of `medv`, `dis`, and `indus` bleong in the model (the null hypothesis). Because our `p-value` is close to zero, we can reject the null and conclude that at least one of those three variables belongs.

***

(ii) Are the following sentences True of False? Briefly justify your answer.

- `indus` is not a significant predictor of crim at the 0.1 level.

Answer: False. It is significant at the 0.001 level, which is smaller than 0.1. So it is also significant at the 0.1 level.

- `Multiple R-squared` is preferred to `Adjusted R-squared` as it takes into account all the variables.

Answer: True. It penalizes including an excessive number of variables.

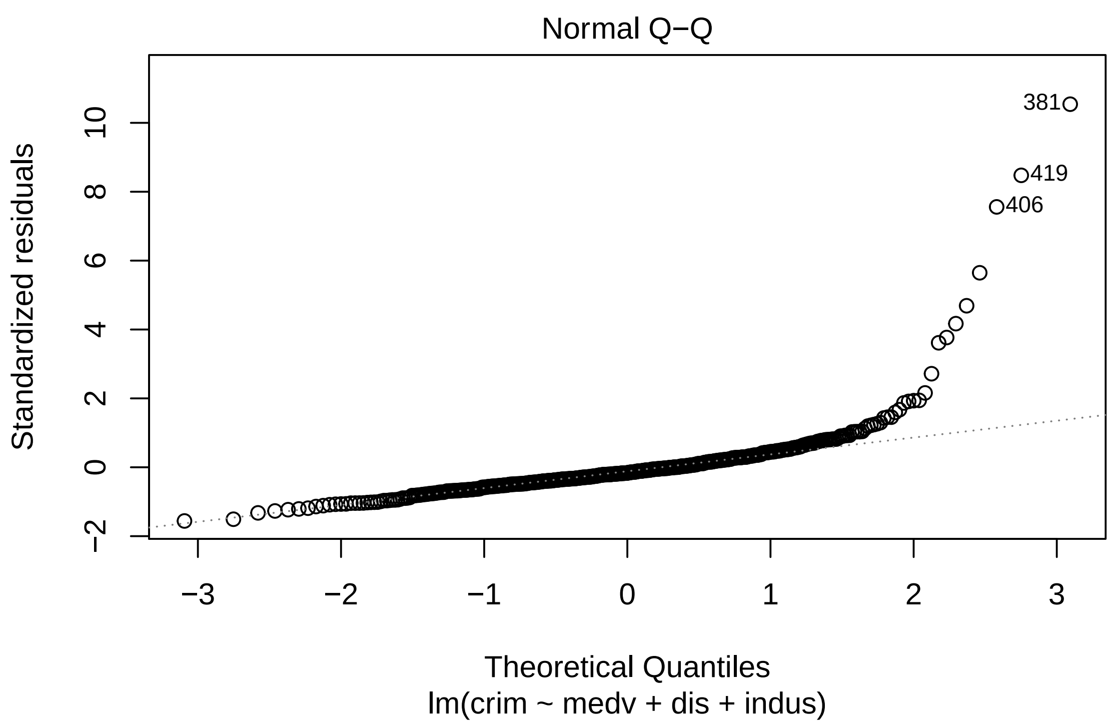
- `medv` has a negative effect on the response.

Answer: True. Becasue the estimate for `medv` is negative (-.26061).

- Our model residuals appear to be normally distributed.

Hint. You need to access to the model residuals in justifying the last sentence. The following commands might help.

```r
# Obtain the residuals
res1 <- residuals(mod1)

# Normal QQ-plot of residuals
plot(mod1, 2)
```




```r
# Conduct a Normality test via Shapiro-Wilk and Kolmogorov-Smirnov test
shapiro.test(res1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res1
## W = 0.59766, p-value < 2.2e-16
```

```r
ks.test(res1, "pnorm")
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  res1
## D = 0.39475, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Answer: False. These date are not normal. Both the Kolmogorov-Smirnov and Shapiro-Wilk tests indicate we can reject the null (that the data are normally distirbuted). Additionally, the QQ plot shows a rightward skew.

---

2. (*25 pt*) For this exercise, we will use a dataset with summary information about American colleges and universities in 2013. The following code chunk retrieves it directly from the website, saving you from having to download it. The data is saved in the object called `amcoll`.

```r
amcoll <- read.csv('http://www-bcf.usc.edu/~gareth/ISL/College.csv')
```

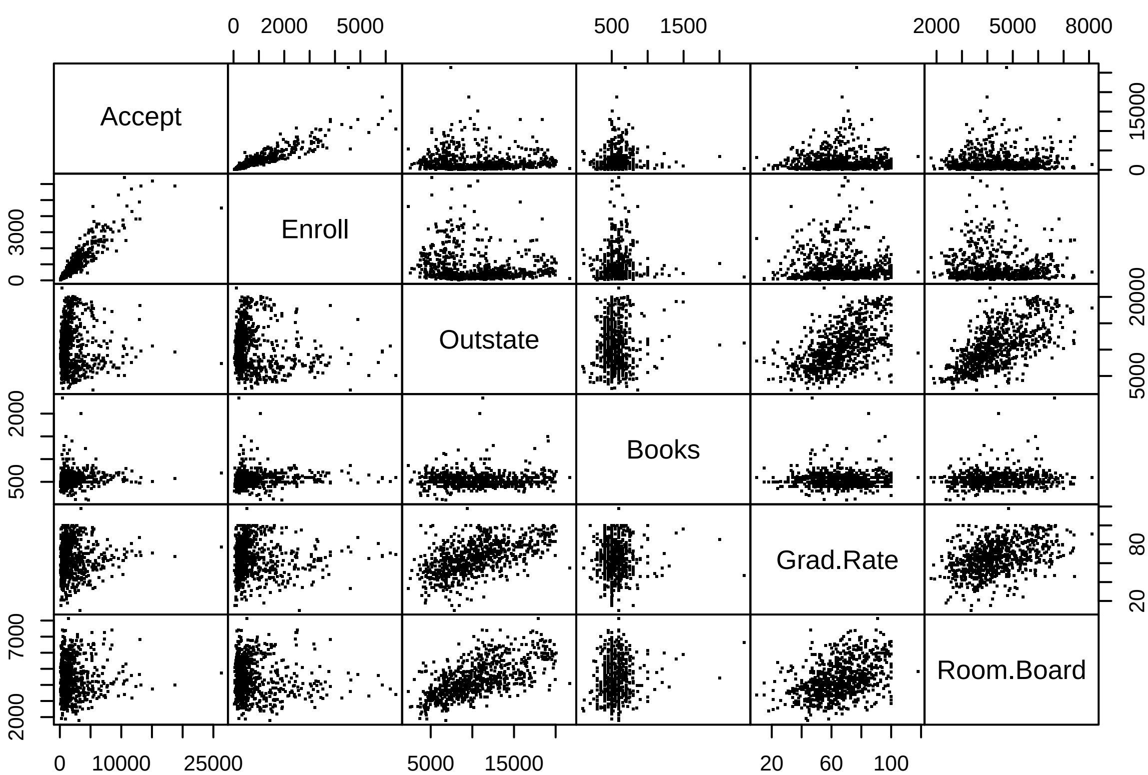
Suppose that we are curious about what factors at a university play an important role in the room and board each semester (column `Room.Board`). Answer the following questions.

(a) Based on some research into the area, you believe that the five most important predictors for the room and board amount are

- the number of students who accepted admission *Accept*
- the number of students who are currently enrolled *Enroll*
- the out of state tuition for a semester *Outstate*
- the average cost of books per year *Books*
- the graduation rate of the students *Grad.Rate*

Plot a pairwise scatterplot of these variables along with the room and board cost, and comment on any trends. If you don't know how to plot such a scatterplot, see for example Pairs and other computational notes from U Wisc. Include your pairwise scatter plot as part of what you turn in.

```r
#--> Prune dataset
library(tidyverse)
amcoll2 <- amcoll %>%
  select(Accept, Enroll, Outstate, Books, Grad.Rate, Room.Board)

#--> Pairwise scatter plot
pairs(amcoll2, gap = 0, pch = ".")
```



Because we are interested in predicting `Room.Board` I am focusing on that first. I see a pretty strong positive relationship with both `Grad.Rate` and `Outstate`. There may be a positive assosiation between out of state tuition rates and room and board costs. There may also be a positive assosiation between graduation rate and room and board costs.

Besides `Room.Board`, I also see positive assosiations between (1) `Accept` and `Enroll` and (2) `Room.Board` and `Outstate`.

(b) Run a linear model of Room.Board on the 5 features above. Suppose we decide that .01 is our level of significance (so p-values have to be above .01 to count as significant). Discuss the findings of your linear model. In particular you should find that one of the features is **not** significant.

```
#--> Compute model
mod2 <- lm(Room.Board ~ Accept+Enroll+Outstate+Books+Grad.Rate, data = amcoll2)
summary(mod2)
```

```
##
## Call:
## lm(formula = Room.Board ~ Accept + Enroll + Outstate + Books +
##     Grad.Rate, data = amcoll2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2329.8  -544.4  -100.3   496.7  2880.7
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.013e+03  1.532e+02  13.141  < 2e-16 ***
## Accept       1.409e-01  3.012e-02   4.677 3.43e-06 ***
## Enroll      -2.905e-01  8.033e-02  -3.616 0.000318 ***
## Outstate     1.590e-01  9.135e-03  17.404  < 2e-16 ***
## Books        6.458e-01  1.773e-01   3.642 0.000288 ***
## Grad.Rate    4.147e+00  2.071e+00   2.003 0.045544 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 808.4 on 771 degrees of freedom
## Multiple R-squared:  0.4601, Adjusted R-squared:  0.4566
## F-statistic: 131.4 on 5 and 771 DF,  p-value: < 2.2e-16
```

At the .01 level, `Grad.Rate` is not a significant predictor variable in our model. The other four variables (`Accept`, `Enroll`, `Outstate`, and `Books`) are all significant preidctors of `Room.Board`. `Enroll` is negative, so we would expect that a higher student enrollment predicts cheaper room and board costs. This makes sense because of the university is able to spend less per student on fixed costs. As the other three (`Accept`, `Outstate`, and `Books`) increase, so does `Room.Board`.

(c) Write a function `kfold.cv.lm()` which performs the following. You can either write this from scratch or use any standard package in R or see the book for example code etc.

Input Arguments:

```
- k: integer number of disjoint sets
- seed: numeric value to set random number generator seed for reproducability
- X: $n \times p$ design matrix
- y: $n \times 1$ numeric response
- which.betas: $p \times 1$ logical specifying which predictors to be included in a regression
```

Output:

Avg.MSPE (average training error over your folds = $\frac{1}{10}\sum_{\text{fold } i}$ prediction error using model obtained from remaining folds), *Avg.MSE* $\frac{1}{10}\sum_{\text{fold } i}$ average training error using model obtained from remaining folds)

Description: Function performs k-fold cross-validation on the linear regression model of y on X for predictors *which.betas*. Returns both the average MSE of the training data and the average MSPE of the test data.

```r
kfold.cv.lm <- function(k, seed, x, y, which.betas) {
  #--> Load library
  library(tidyverse)
  #--> Set the seed
  set.seed(seed)

  #--> Generate a sequence of k integers
  num_reps <- (nrow(x) / k) + 1 # overshoot a bit
  int_sequence <- rep(1:k, num_reps)
  int_sequence <- int_sequence[1:nrow(x)] # trim off any extra

  #--> Shuffle the integer sequence
  shuffled <- sample(int_sequence)

  #--> Add the shuffled sequence (k folds) as a pseudo-variable
  x <- cbind(x, shuffled)

  #--> Correct which.betas (since we added another column of integers)
  which.betas <- c(which.betas, FALSE)

  #--> Initialize vectors
  MSPE <- rep(0, k)
  MSE <- rep(0, k)

  #--> Loop and calculate
  for (i in 1:k) {
    #--> Pick out the folds and predictors we want
    y_test <- y[shuffled == k, ]
    y_train <- y[shuffled != k, ]

    test <- cbind(x[x[, ncol(x)] == k, which.betas], y_test)
    train <- cbind(x[x[, ncol(x)] != k, which.betas], y_train)

    #--> Fit a linear model
    model <- lm(y_train ~., data=train)

    #--> Compute errors
    MSPE[k] <- mean(model$residuals ^ 2)
    MSE[k] <- mean((train$y_train - predict.lm(model, train)) ^ 2)
  }

  #--> Sum and return
  Avg.MSPE <- sum(MSPE) / k
  Avg.MSE <- sum(MSE) / k

  return(data.frame("Avg.MSE" = Avg.MSE, "Avg.MSPE" = Avg.MSPE))
}
```

(d) Use your function `kfold.cv.lm()` to perform 10 folder cross validation on the college data for the following two models:

- the full model on the 5 features above;
- the model where you leave out the feature you found to be insgnificant in (b).

Which of the two is a "better" model and why?

```
#--> Put in x and y format
x <- select(amcoll2, -Room.Board)
y <- select(amcoll2, Room.Board)

#--> Model on 5 features
model1 <- kfold.cv.lm(10, 1729, x, y, rep(TRUE, 5))

#--> Model on 4 features
model2 <- kfold.cv.lm(10, 1729, x, y, c(rep(TRUE, 4), FALSE))

print(model1)
```

```
##    Avg.MSE Avg.MSPE
## 1 65341.71 65341.71
```

```
print(model2)
```

```
##    Avg.MSE Avg.MSPE
## 1 65583.15 65583.15
```

Answer: The first model–which includes all five variables–is the better model because it creates a lower MSPE than the model with four variables. While this could be because we are including more variables and merely giving our model more degrees of freedom, I believe that `Grad.Rate` is truly important. Even though it was not significant at the .001 level, it *is* significant at the .05 level. There is less than a five percent chance of observing a meaningful (non-zero) `Grad.Rate` beta-coefficient by chance alone. Therefore, I believe it contributes meaningfully to our model and it is consistent with the ouput of out 10 fold validation, which shows the first model performing better since it has a lower average MSPE.

---

3. (*25 pt*, Textbook Exercises 3.10) This question should be answered using the `Carseats` data set.

```
head(Carseats)
```

```
##   Sales CompPrice Income Advertising Population Price ShelveLoc Age
## 1  9.50       138     73          11        276   120       Bad  42
## 2 11.22       111     48          16        260    83      Good  65
## 3 10.06       113     35          10        269    80    Medium  59
## 4  7.40       117    100           4        466    97    Medium  55
## 5  4.15       141     64           3        340   128       Bad  38
## 6 10.81       124    113          13        501    72       Bad  78
##   Education Urban  US
## 1        17   Yes Yes
## 2        10   Yes Yes
## 3        12   Yes Yes
## 4        14   Yes Yes
## 5        13   Yes  No
## 6        16    No Yes
```

(a) Fit a multiple regression model to predict `Sales` using `Price`, `Urban`, and `US`. Then, display a summary of the linear model using the `summary` function.

```r
carsales_lm <- lm(Sales~Price+Urban+US, data=Carseats)

summary(carsales_lm)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

(b) Write a one- or two-sentence interpretation of each coefficient in the model. Be careful: some of the variables in the model are qualitative!

Answer: `Price` is a significant predictor of `Sales` at the alpha = almost 0 level. As `Price` increases, `Sales` decreases ever so slightly. Perhaps people with Ferraris don't have much need for car seats.
`USYes`, also is a significant predictor of `Sales` at the almost 0 level. A store located in the US is likely to sell more carseats. Perhaps this is because there are laws that require carseats (or at least the type that this company makes) for child passengers.
`UrbanYes` is not a significant predictor. I cannot think of a logical reason why someone living in a city would be more/less likely to buy a car seat than someone else. This is consistent with the fact that `UrbanYes` is not a significant predictor.
`Intercept` is significant as well, indicating that we have a `Sales` intercept that is well above zero. This indicates that there is a certain "automatic" sale volume if all other variables are zero.

(c) Based on the output in part (a): For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

Answer: `Intercept`, `Price`, and `USYes`.

(d) On the basis of your response to the previous question, a model with fewer predictors, using only the predictors for which there is evidence of association with the outcome. Display a summary of the linear model using the `summary` function.

```r
carsales_lm <- lm(Sales~Price+US, data=Carseats)

summary(carsales_lm)
```

```
##
## Call:
## lm(formula = Sales ~ Price + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9269 -1.6286 -0.0574  1.5766  7.0515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
## Price       -0.05448    0.00523 -10.416  < 2e-16 ***
## USYes        1.19964    0.25846   4.641 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.469 on 397 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2354
## F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

(e) In a few sentences: How well do the models in (a) and (d) fit the data? Justify your response with information from the outputs of part (a) and (d).

Answer: Neither model fits particularly well. I like the second model better because all of its predictors are significant at the alpha = (almost) 0 level. However, both of them fail to achieve a decent R-squared value. In practice, I would have a hard time selling this model to anyone because there is so much variance. I would try to grow a random forest (like Breiman discussed in his article) rather than assume a linear relationship.

---

4. (*14 pt Optional*) Note: this question is optinal and if you do want to do it, you will need to do the heavy lifting in terms of finding the data, cleaning the data etc. We will not be able to help you too much with respect to the above data "carpentry" issues.

Search online for a dataset that **you are interested in** where you think you can apply linear regression (i.e. your data has a continuous response and a bunch of real valued features). Data sets from the book (ISLR) website are not allowed and more importantly try to find something that makes you curious to find the answers.

(a) My sister Paige, a shining light in my life, is living in Belgium and playing carrilon. She even has a blog about it. While there, she has developed a taste for not only Belgian beer, but French wine. Her love of oenology is so deep that she is applying to several graduate programs in wine tourism. To make her life easier, I would like to look at the relationships of different measurements of wine on one another. Is there a way to predict, say, alcohol content if you know other measurements? I obtained this data from the University of California at Irvine Machine Learning Data Collection.

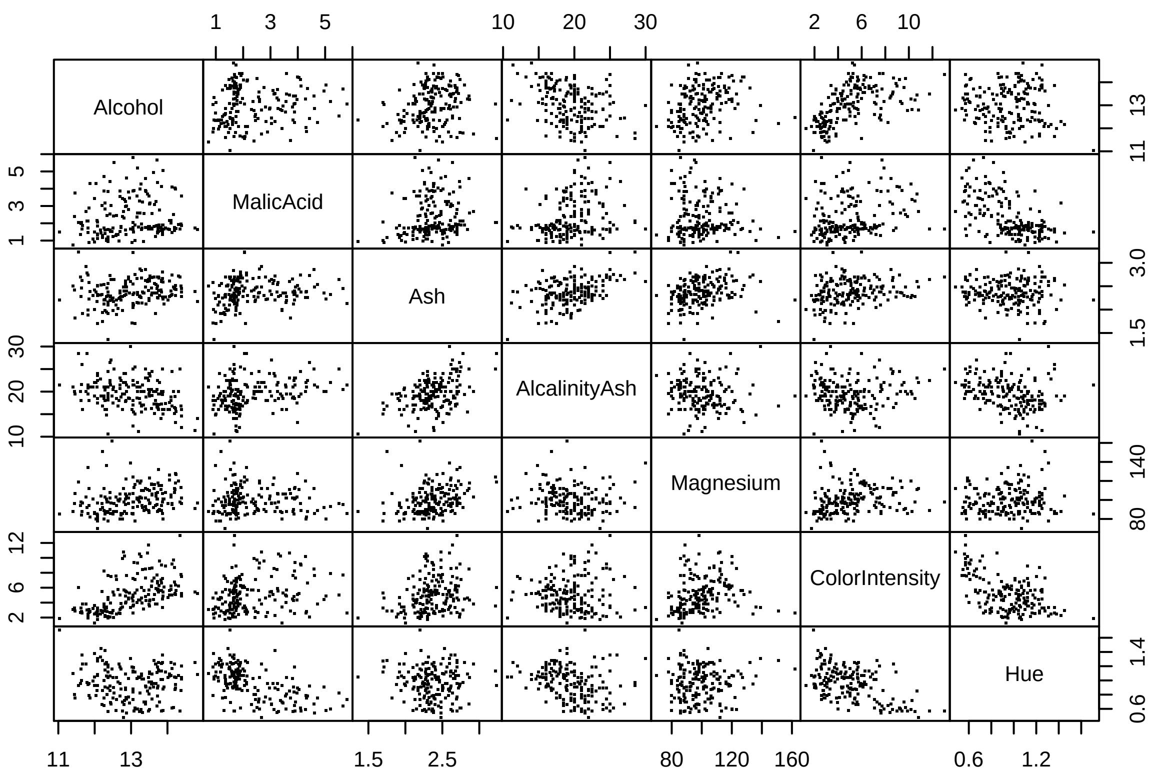
(b) Plot a pairwise scatter plot between the response and some (at least 2) of the features.

```r
#--> Read and add variable names
wine <- read.csv("wine.data", header=FALSE)
wine <- wine[, -1]
names(wine) <- c("Alcohol", "MalicAcid", "Ash", "AlcalinityAsh", "Magnesium", "TotalPhenols",
                 "Flavanoids", "NonflavPhenols", "Proanthocyanins", "ColorIntensity", "Hue",
                 "OD280", "Proline")

wine2 <- wine %>%
```

```r
  select(Alcohol, MalicAcid, Ash, AlcalinityAsh, Magnesium, ColorIntensity, Hue) # get rid of all that

head(wine2)
```

```
##   Alcohol MalicAcid  Ash AlcalinityAsh Magnesium ColorIntensity  Hue
## 1   14.23      1.71 2.43          15.6       127           5.64 1.04
## 2   13.20      1.78 2.14          11.2       100           4.38 1.05
## 3   13.16      2.36 2.67          18.6       101           5.68 1.03
## 4   14.37      1.95 2.50          16.8       113           7.80 0.86
## 5   13.24      2.59 2.87          21.0       118           4.32 1.04
## 6   14.20      1.76 2.45          15.2       112           6.75 1.05
```

```r
#--> Plot
pairs(wine2, gap = 0, pch = ".")
```



(c) Run a linear model to learn the relationship between the features and the response and extract information from the lm function (what variables seem significant and what do not)?

```r
winemodel <- lm(Alcohol~., data=wine2)
summary(winemodel)
```

```
##
## Call:
## lm(formula = Alcohol ~ ., data = wine2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.55592 -0.41207  0.05028  0.40221  1.39180
```

```
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    11.064362   0.579257  19.101  < 2e-16 ***
## MalicAcid       0.113801   0.048999   2.323  0.02138 *
## Ash             0.643251   0.204426   3.147  0.00195 **
## AlcalinityAsh  -0.097173   0.016484  -5.895 1.95e-08 ***
## Magnesium       0.003469   0.003383   1.025  0.30667
## ColorIntensity  0.194590   0.024710   7.875 3.77e-13 ***
## Hue             0.743922   0.282848   2.630  0.00931 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5899 on 171 degrees of freedom
## Multiple R-squared:  0.4899, Adjusted R-squared:  0.472
## F-statistic: 27.37 on 6 and 171 DF,  p-value: < 2.2e-16
```

Explain in words (e.g. to someone who has no math or stat background) your findings.

I want to predict alcohol content of the wines made in my vineyard without having to actually measure the alcohol content. When I run a linear model using different chemical and physical properties of the wine, I see that I can predict the alcohol content pretty well. In fact, nearly 50% of our variance in alcohol content can be explained by our model. The predictors that did a good job (p-value < 0.01) were `Ash`, `AlcalinityAsh`, `ColorIntensity`, and `Hue`. I've been told by my sister, the oenologist, that high ash and ash alcalinity are good things. I'll take her word for it. `Ash`, `ColorIntensity`, and `Hue` all positively correlate with `Alcohol`, so as one goes up, so do the others. `AshAlcalinity` on the other hand negatively correlates with `Alcohol`, so as one goes up, the other goes down. The next time I see my sister, I'll be sure to order the wine with the "most intense hue" and "high ash content." I'm sure the bartender will love that.