HW10

Grace Okamoto

11/3/2021

Problem 1

(A) Describe in your own words the difference between linear and logistic regression.

Welll, I suppose it's not a terribly good definition, but a linear regression is when a certain event's predicted values follows that of a linear equation, whereas a logistic regression follows that of a logistic equation. In linear regressions, the values can be any number, whereas in logistic regressions, the output is in probability.

(B) Given an example of a dataset that would be appropriate to analyze with multiple linear regression but not with logistic regression.

Housing prices could be measured with multiple linear regression, as their prices will not taper off as materials and labor increase.

(C) Given an example of a dataset that would be appropriate to analyze with logistic regression but not with linear regression.

A common example of logistic regressions is the odds of a particular candidate being elected.

(D) Given a model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

what is the interpretation of the coefficient β_2?

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

The coefficient of beta_2 is positive, meaning that there is a positive correlation between the variables.

(E) Given a model

$$\ln\left(\frac{p}{1-p}\right) = -5 + 3x_1$$

how much do the odds increase if x_1 increases by 1?

$$\ln\left(\frac{p}{1-p}\right) = -5 + 3x_1$$

$$\ln\left(\frac{p}{1-p}\right) = -5 + 4x_1$$

Problem 2

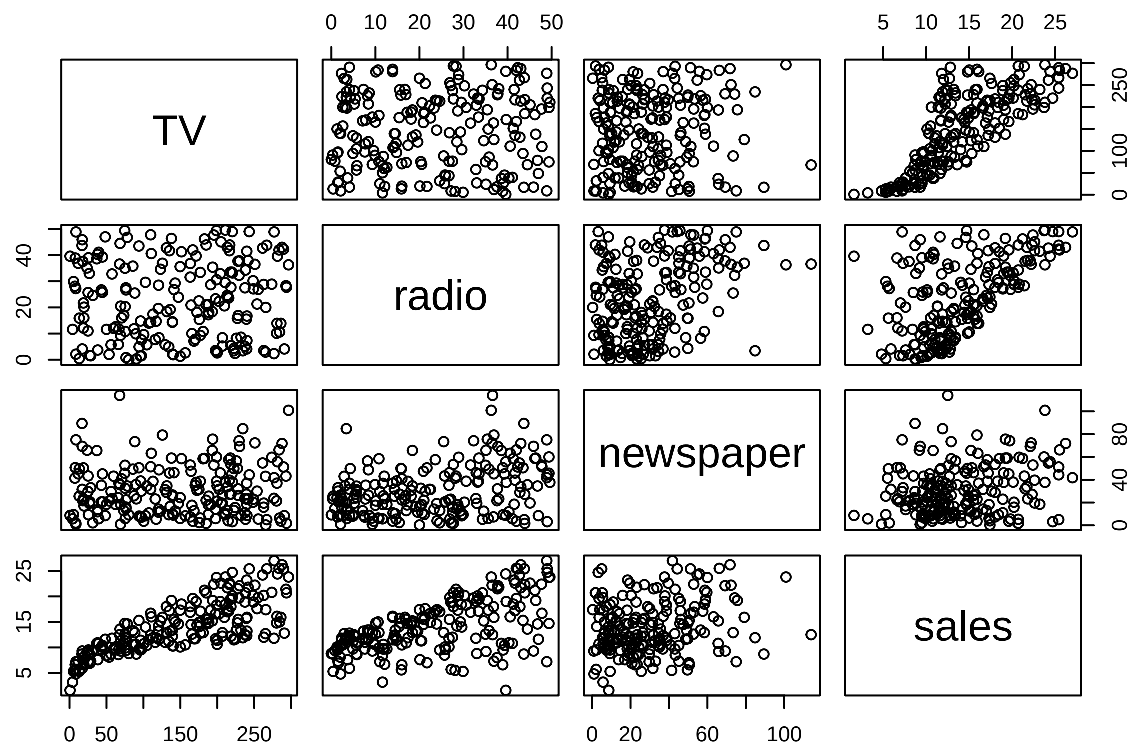
This problem checks your understanding of multiple linear regression and diagnosis of these fits. Start by loading in the Advertising.csv file as a dataframe.

```r
A <- read.csv("Advertising.csv")
head(A)
```

```
##   X    TV radio newspaper sales
## 1 1 230.1  37.8      69.2  22.1
## 2 2  44.5  39.3      45.1  10.4
## 3 3  17.2  45.9      69.3   9.3
## 4 4 151.5  41.3      58.5  18.5
## 5 5 180.8  10.8      58.4  12.9
## 6 6   8.7  48.9      75.0   7.2
```

(A) Make a scatterplot matrix of the columns of the dataframe. Each row of this data set represents a single media market and the TV, Newspaper, and Radio columns contain spending amounts related to each media type while the sales value is the number of units sold (in thousands) in that market. Based on this plot, do you think multiple linear regression is appropriate to attempt?

```r
pairs(A[,c(2:5)])
```



Multiple linear regression seems like a reasonable step to take, given that there seems to be a decent amount of correlation between some of the variables.

(B) Fit a multiple linear regression model using all three media columns as predictors with the sales column as the dependent variable.

```
mlrA <- lm(sales ~ TV+radio+newspaper, A)
summary(mlrA)
```

```
##
## Call:
## lm(formula = sales ~ TV + radio + newspaper, data = A)
##
## 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 ***
## TV           0.045765   0.001395  32.809   <2e-16 ***
## radio        0.188530   0.008611  21.893   <2e-16 ***
## newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## 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
```

(C) Write the linear equation estimated by your fit.

$$y = 2.938889 + 0.045765x_1 + 0.188530x_2 - 0.001037x_3$$

where y is sales, x_1 is TV, x_2 is radio, and x_3 is newspaper.

(D) Write the coefficient of determination for your fit.

The coefficient of determination is 0.8972.

(E) How many sales would your model predict for a market that spent 200 on TV, 50 on radio, and 100 on newspaper?

```
2.938889 + 0.045765*200 + 0.188530*50 - 0.001037*100
```

```
## [1] 21.41469
```

Problem 3

This problem checks your understanding of implementing logistic regression in R. Start by loading in the default_ISLR.csv file as a dataframe.

```
ISLR = read.csv("default_ISLR.csv")
head(ISLR)
```

```
##   default student   balance    income
## 1      No      No  729.5265 44361.625
## 2      No     Yes  817.1804 12106.135
## 3      No      No 1073.5492 31767.139
## 4      No      No  529.2506 35704.494
## 5      No      No  785.6559 38463.496
## 6      No     Yes  919.5885  7491.559
```

(A) Fit a logistic regression model to this data with response variable y being the default column, p being the probability of default, and x being the balance column.

```
ISLR_log <- glm (as.factor(default) ~ balance, data = ISLR,family = binomial )
summary(ISLR_log)
```

```
##
## Call:
## glm(formula = as.factor(default) ~ balance, family = binomial,
##     data = ISLR)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2697  -0.1465  -0.0589  -0.0221   3.7589
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.065e+01  3.612e-01  -29.49   <2e-16 ***
## balance      5.499e-03  2.204e-04   24.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1596.5  on 9998  degrees of freedom
## AIC: 1600.5
##
## Number of Fisher Scoring iterations: 8
```

(B) Compute the coefficients of the fitted model and write the corresponding equation for the log odds.

$$\ln\left(\frac{p}{1-p}\right) = -10.65 + 0.00599X$$

(C) What percentage of the provided data does your model correctly classify?

```
ISLR$temp <- ISLR_log$fitted.values
ISLR$predict <- ifelse(ISLR$temp<.5,"No","Yes")
head(ISLR)
```

```
##   default student   balance    income         temp predict
## 1      No      No  729.5265 44361.625 0.0013056797      No
## 2      No     Yes  817.1804 12106.135 0.0021125949      No
## 3      No      No 1073.5492 31767.139 0.0085947405      No
## 4      No      No  529.2506 35704.494 0.0004344368      No
## 5      No      No  785.6559 38463.496 0.0017769574      No
## 6      No     Yes  919.5885  7491.559 0.0037041528      No
```

```
mean(ISLR$predict == ISLR$default)
```

```
## [1] 0.9725
```

(D) What is the probability that someone with a balance of 1950 will default, according to your model?

```
exp(-10.65 + 0.00599 * 1950)/(1+exp(-10.65 + 0.00599 * 1950))
```

```
## [1] 0.7370128
```