A Regression Example: Relationship Between Advertising and sales

November 20, 2017

The Advertising Data

The Advertising dataset gives the sales of a product (in thousands) and the amount spent on different forms of advertising (TV, radio, newspaper) in thousands of dollars for 200 different regions. We want to investigate the relationship between amount spent on advertising and sales

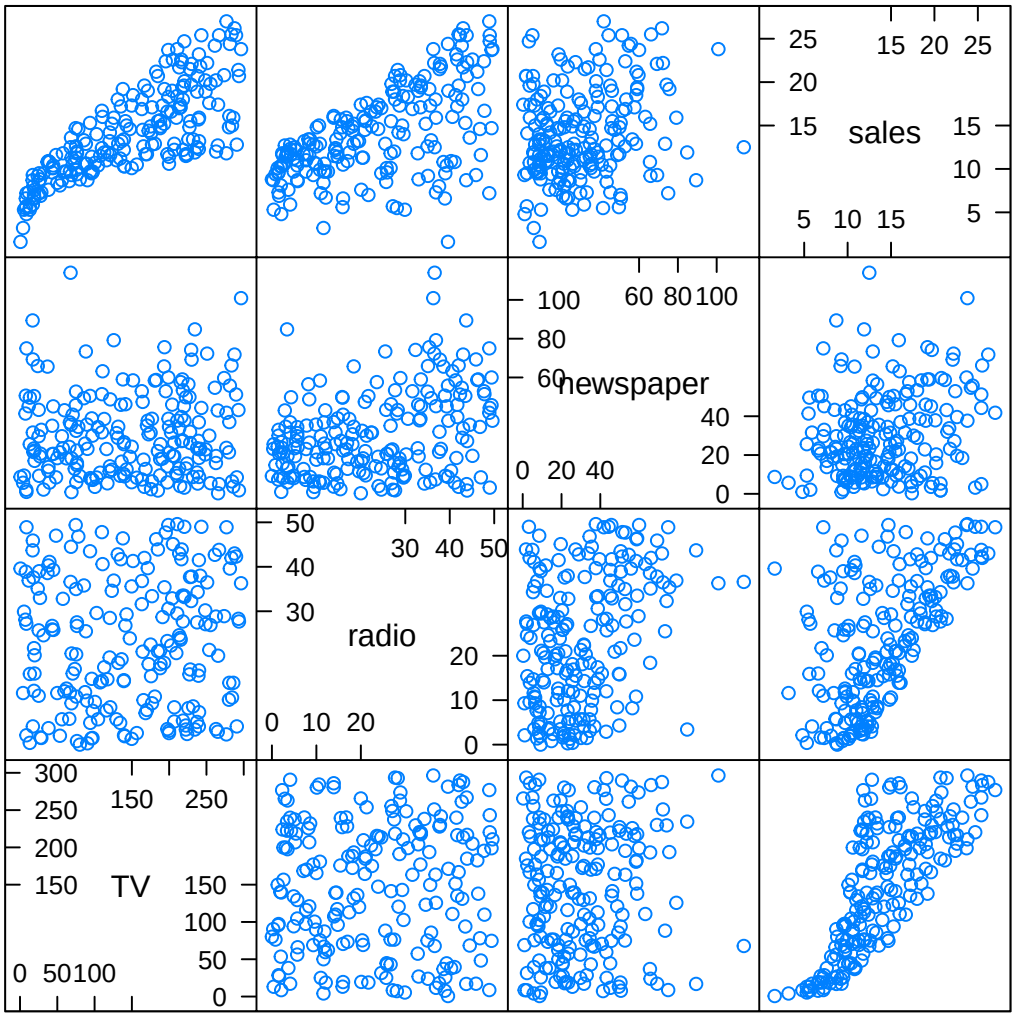
We read the data into R from the website for the ISLR text.

```
> adver <- read.csv("http://www-bcf.usc.edu/~gareth/ISL/Advertising.csv")
> adver <- adver[,-1]  #remove row numbers
> head(adver)

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

Exploratory Data Analysis

It is always a good idea to look at some plots to see if variables are linearly related. We first attach the dataframe to make extracting columns easier.

```
> search()  #lists all loaded packages and attached objects

[1] ".GlobalEnv"        "package:stats"     "package:graphics"
[4] "package:grDevices" "package:utils"     "package:datasets"
[7] "package:methods"   "Autoloads"         "package:base"

> attach(adver)  #now you can reference cols in adver as, for example, TV instead
> #of adver$TV
> search()  #see that adver is now in R's search path

 [1] ".GlobalEnv"        "adver"             "package:stats"
 [4] "package:graphics"  "package:grDevices" "package:utils"
 [7] "package:datasets"  "package:methods"   "Autoloads"
[10] "package:base"

> library(lattice)
> splom(adver)  #which predictors are highly correlated with sales?
```



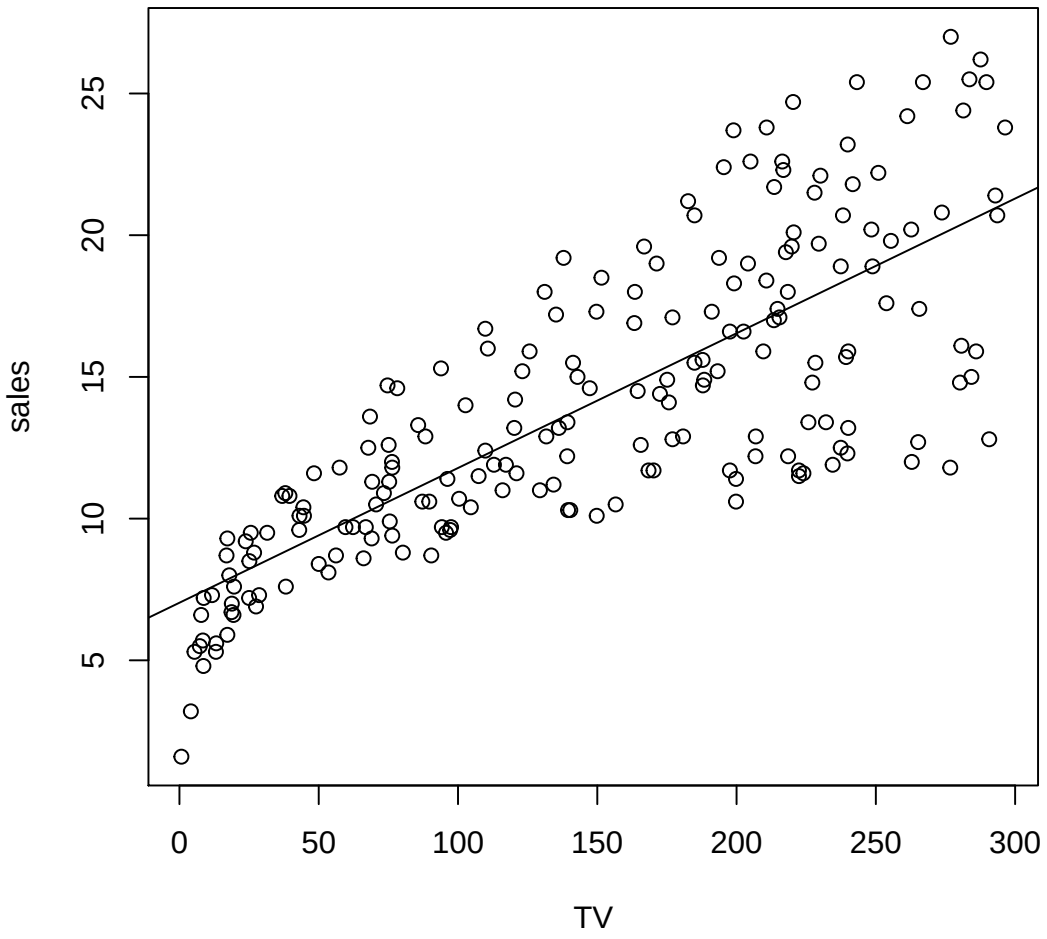
Scatter Plot Matrix

A Simple Linear Regression

Next, we run a simple regression to quantify the relationship between TV advertising and sales.

```
> tvreg <- lm(sales~TV,data=adver)
> summary(tvreg)

Call:
lm(formula = sales ~ TV, data = adver)

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 ***
TV          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(TV,sales)
> abline(tvreg)
```



Note: There is one small problem that shows up in the plot. Recall the regression model assumes that $\epsilon_i \sim N(0, \sigma^2)$, thus the errors should have the same variance regardless of the value of the x-variable, TV. However, the variance is more for large values of TV. We won't worry to much about that now; however, a log transform on sales helps with this problem.

Questions:

- How good is the fit of the line to the scatterplot? In other words, how strong is the predictive relationship between TV advertising and sales?

- Interpret the slope/coefficient of TV in the context of these data.

- Obtain a 95% confidence interval for the true slope and interpret it.

- Can we conclude the slope of the regression line is significantly different from 0? Why or why not?

- Interpret the intercept of the regression line in the context of these data.

Multiple Regression

We might also wonder about the relationship of radio and newspaper advertising with sales. We could run three separate simple linear regressions: TV vs. sales, radio vs. sales and newspaper vs. sales. However, it is better to run a multiple regression with TV, radio and newspaper as predictors and sales as the response variable.

In general, the form of the multiple regression model with 3 predictors is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$.

For the Advertising example, the specific form is:

$$sales = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times newspaper + \epsilon$$

We interpret the coefficient of a single predictor, β_2 for instance, as the average amount of change in Y corresponding to a one unit change in X_2 **when all other variables are held fixed**.

The fitted model is:

```
> multreg <- lm(sales ~ TV+radio+newspaper,data=adver)
> summary(multreg)

Call:
lm(formula = sales ~ TV + radio + newspaper, data = adver)

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

> cor(adver)

                  TV      radio  newspaper     sales
TV        1.00000000 0.05480866 0.05664787 0.7822244
radio     0.05480866 1.00000000 0.35410375 0.5762226
newspaper 0.05664787 0.35410375 1.00000000 0.2282990
sales     0.78222442 0.57622257 0.22829903 1.0000000
```

$$sales = 2.94 + 0.046 \times TV + 0.189 \times radio + -0.001 \times newspaper + \epsilon$$

Questions:

- Predict the sales if $1000 is spent on each of TV, radio and newspaper advertising.

- Interpret the coefficient of radio.

- What is strange about the coefficient of newspaper?

- How good is the fit of the regression line to the data?

- Are all coefficients significantly different from 0?

Variable Selection

In a regression, including more predictors always increases the r^2, even if the "predictor" is random noise. However, this does not always result in a model that has better predictive performance. Just as a decision tree can overfit to the training data and do poorly on test data, a regression model with too many predictors can overfit the training data and do poorly predicting test data. Variable selection in regression is like pruning a tree model. We want a model that fits the data well, but does not contain extra, useless variables. We use `stepAIC()` in the MASS library to pare the variables down to the most important ones. I won't explain its inner workings entirely. Howeve, the aim is to choose a set of variables which minimize the AIC, or Akaike Information Criterion, which balances good fit against not having too many variables.

```
> multreg <- lm(sales ~ TV+radio+newspaper,data=adver)
> library(MASS)
> stepAIC(object=multreg,scope=list(upper= ~., lower= ~1))

Start:  AIC=212.79
sales ~ TV + radio + newspaper

            Df Sum of Sq    RSS    AIC
- newspaper  1      0.09  556.9 210.82
<none>                    556.8 212.79
- radio      1   1361.74 1918.6 458.20
- TV         1   3058.01 3614.8 584.90

Step:  AIC=210.82
sales ~ TV + radio

            Df Sum of Sq    RSS    AIC
<none>                    556.9 210.82
+ newspaper  1      0.09  556.8 212.79
- radio      1   1545.62 2102.5 474.52
- TV         1   3061.57 3618.5 583.10

Call:
lm(formula = sales ~ TV + radio, data = adver)

Coefficients:
(Intercept)           TV        radio
    2.92110      0.04575      0.18799
```

After variables selection, we find the best model only contains the predictors TV and radio. The **adjusted** r^2 also balances fit with a penalty as more variables are included in the model. How has the adjusted r^2 changed after variable selection?

```
> bestreg <- lm(sales ~ TV+radio,data=adver)
> summary(bestreg)

Call:
lm(formula = sales ~ TV + radio, data = adver)

Residuals:
    Min      1Q  Median      3Q     Max
-8.7977 -0.8752  0.2422  1.1708  2.8328

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.92110    0.29449   9.919   <2e-16 ***
TV           0.04575    0.00139  32.909   <2e-16 ***
radio        0.18799    0.00804  23.382   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.681 on 197 degrees of freedom
Multiple R-squared:  0.8972,        Adjusted R-squared:  0.8962
F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16
```