STAT 308 – Chapter 8/Appendix B

Background Information

We have discussed simple linear regression techniques, where we have are using a singular explanatory (x) variable to make a linear prediction about a singular repsonse (y) variable. However, suppose that we have mulitple explanatory ($x_1, x_2, \ldots, x_n$) variables that we can use to make a better prediction of y (i.e. lower the variance of y given $x_1, x_2, \ldots, x_n$). This chapter will introduce the concept of multiple linear regression.

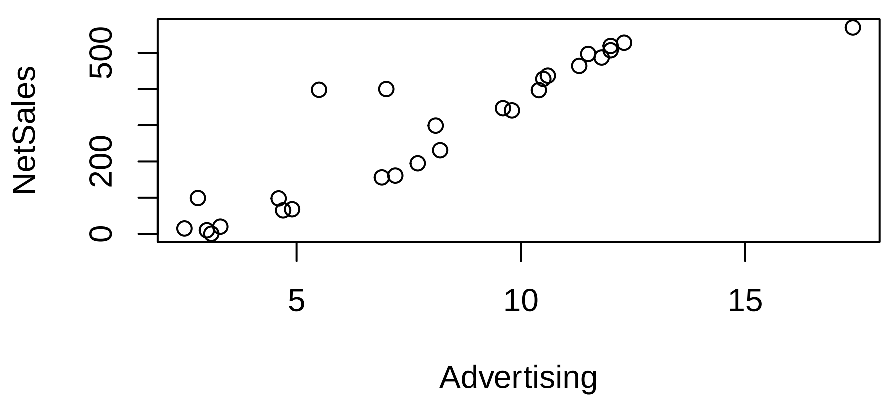
Motivating Example

Consider the following dataset on sales at All Green franchises with the following variables: - `NetSales`: Net Sales of the franchise (in 1000s of dollars) - `SqFt`: Square footage of the store (in 1000s) - `Inventory`: Amount of inventory (in 1000s of dollars) - `Advertising`: Amount spent on advertising (in 1000s of dollars) - `SizeofDist`: Size of district (per 1000 families) - `NoofStores`: Number of competing stores in the district

```
allgreen <- read.csv("../data/AllGreen.csv")
```

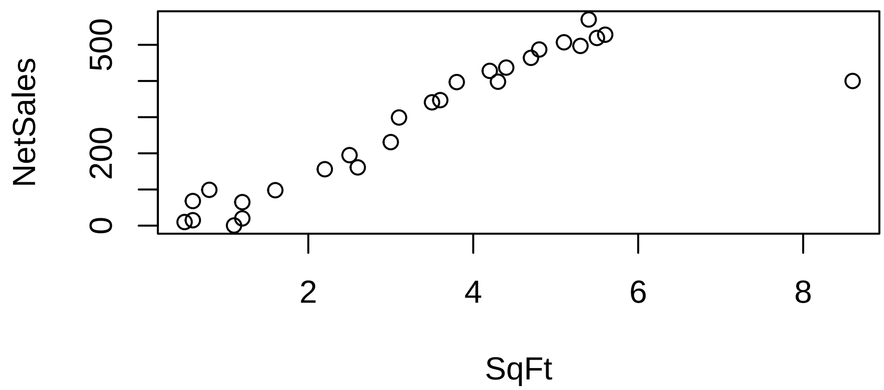
Previously, we have only discussed using one variable to make linear predictions about the response. For example, suppose we believe that the net sales of a franchise is linearly related to the amount of money they spent in advertising.

```
plot(NetSales ~ Advertising,allgreen)
```



Or, perhaps we can say the net sales is linearly related to the square footage of the store.

```r
plot(NetSales ~ SqFt,allgreen)
```



Now, what if we can take both of these explanatory variables, advertising and square footage, and use them **both** in conjunction to make a linear prediction of the net sales of the franchise.

```r
## You do NOT need to run this code.  This is used only for in class visualization.
# library(plotly)
# plot_ly(allgreen,
#         x=~Advertising, # First explanatory variable
#         y=~SqFt, # Second explanatory variable,
#         z=~NetSales) %>% # Response variable
#   add_markers()
```

Important Defintions

Multiple Linear Regression: An extension of simple linear regression into where we can estimate a response variable Y based on multiple explanatory variables $X_1, X_2, \ldots, X_p$. p is the number of variables in our linear structure.

Assumptions Needed for Multiple Linear Regression Models

(Note, these assumptions are the same for simple linear regression, just extended to multiple linear regression)

- Existence: For any given value of $X_1, \ldots, X_p$, Y is a random variable with a certain probability distribution with a finite mean and variance. Define:
  - $\mu_{Y|X_1,\ldots,X_p}$ - the population mean of Y for a fixed $X_1, \ldots, X_p$
  - $\sigma^2_{Y|X_1,\ldots,X_p}$ - the population variance of Y for a fixed $X_1, \ldots, X_p$
- Independence: The observed values of Y are statistically independent of one another given $X_1, \ldots, X_p$

- Linearity: $\mu_{Y|X} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$ or equivalently

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

where ϵ is the error of the multiple linear model.

The next two assumptions discuss the distribution of ϵ.

- Homoscedasticity: The variance of Y is the same for different given values of $X_1, \ldots, X_p$ (i.e. $\sigma^2_{Y|X_1,\ldots,X_p} = \sigma^2$)
- Normality: For any fixed value of $X_1, \ldots, X_p$, Y is normally distributed. This fact makes analysis of the data easier.

All of these assumptions put together lead us to our full mathematical model we will assume:

$$Y \sim \mathcal{N}(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p, \sigma^2)$$

.

"Least squares' ' regression estimate

Recall from Chapter 5 that, for a simple linear regression model, the estimates for the regression parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ can be found by minimizing the sum of squared errors

$$\text{SSE} = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2.$$

For multiple linear regression, this estimation procedure is the same, except now we have additional β parameters that we need to estimate. In other words, to find estimates for the regression parameters $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, we need to minimize

$$\text{SSE} = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_p x_{i,p}))^2.$$

where $x_{i,j}$ is the i^{th} observation from variable j for $j = 1, \ldots, p$.

Information about least squares regression

Regression variance: $s^2_{y|x_1,\ldots,x_p} = s^2 = \frac{1}{n-(p+1)} SSE$.

Why do we divide by $n - (p+1)$?

$r^2 = \frac{SST - SSE}{SST}$.

Full ANOVA table for multiple linear regression

	df	Sums of Squares	Mean Square	f Value	Pr(>f)
Model	p	SSM = SSY − SSE	$\text{MSM} = \frac{\text{SSM}}{p}$	$\frac{\text{MSM}}{\text{MSE}}$	$Pr(F_{p,n-(p+1)} > \frac{\text{MSM}}{\text{MSE}})$
Error	$n - (p+1)$	SSE	$\text{MSE} = \frac{\text{SSE}}{n-(p+1)}$		
Total	$n - 1$	$\text{SSY} = \sum_{i=1}^{n} (y_i - \bar{y})^2$			

Multiple Linear Regression in Matrix Form

There is not a nice scalar form to calculate the regression coefficients from a multiple linear regression model, but we can calculate this using matrix algebra.

Matrix Transpose:

Consider a matrix $\boldsymbol{A}_{2\times3}$. The transpose of the matrix $\boldsymbol{A}'$ is the same matrix with the rows and columns flipped (i.e. the 1st row of $\boldsymbol{A}$ is the 1st column of $\boldsymbol{A}'$)

Matrix Addition:

Consider matrices $\boldsymbol{A}_{2\times3} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$ and $\boldsymbol{B}_{2\times3} = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix}$, then

$$\boldsymbol{A} + \boldsymbol{B} = \begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} & a_{13}+b_{13} \\ a_{21}+b_{21} & a_{22}+b_{22} & a_{23}+b_{23} \end{bmatrix}$$

Matrix Multiplication:

Consider matrices $\boldsymbol{A}_{2\times3} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$ and $\boldsymbol{B}_{3\times1} = \begin{bmatrix} b_{11} \\ b_{21} \\ b_{31} \end{bmatrix}$, then

$$\boldsymbol{A} \times \boldsymbol{B} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} \end{bmatrix}$$

Diagonal Matrix:

A diagonal matrix is a square matrix (i.e. same number of rows and columns) where all of the off-diagonal values is 0 ($a_{ij} = 0$ when $i \neq j$).

Matrix Inverse:

Consider a square matrix $\boldsymbol{A}_{n\times n}$. The inverse of the matrix $\boldsymbol{A}^{-1}$ is the matrix where $\boldsymbol{A}\times\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\times\boldsymbol{A} = \boldsymbol{I}$ where $\boldsymbol{I}$ is the identity matrix – a diagonal matrix where the values in the diagonal are all 1$.

Connection to Multiple Linear Regression:

Now, consider the following matrices:

$$\boldsymbol{Y}_{n\times1} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$\boldsymbol{X}_{n\times(p+1)} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix}$$

$$\boldsymbol{\beta}_{(p+1)\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p+1} \end{bmatrix}$$

$$\boldsymbol{\epsilon}_{n\times 1} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Note that, we previously stated

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon.$$

Using matrix algebra, this is equivalent to saying

$$\boldsymbol{Y}_{n\times 1} = \boldsymbol{X}_{n\times(p+1)}\boldsymbol{\beta}_{(p+1)\times 1} + \boldsymbol{\epsilon}_{n\times 1}.$$

For ease of use, I will not include the subscript describing the matrix dimensions in the rest of these lecture notes.

The sum of squared errors used to find the least squares regression estimates can also be found using matrix notation:

$$SSE = \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_p x_{i,p}))^2 = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}).$$

Using this information, the least squares regression estimate for $\boldsymbol{\beta}$ is calculated by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}.$$

Other items you can get from matrix algebra for multiple linear regression:

Vector of Predicted values: $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$

Sum of Squared Errors: SSE $= \boldsymbol{Y}'\boldsymbol{Y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y}$

Model Sums of Squares: SSE $= \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y} - n\bar{Y}^2$

Variance of Regression Estimates: $s^2_{\hat{\beta}_k}$ for $k = 1, \ldots, p$ is k^th diagonal element of $s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$

Let's see this in action with the allgreen dataset.