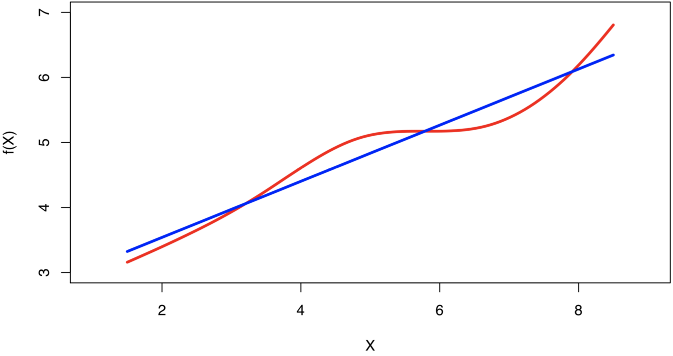
STARTED on 1/25/22
(week 2, lecture 3)

Linear Regression[1]

[1] Based on materials in ISLR Ch 3

- Linear regression is a simple approach to supervised learning. It assumes that the dependence of Y on $X_1, X_2, \cdots, X_p$ is linear.
- True regression functions are never linear!



- Although it may seem overly simplistic, linear regression is extremely useful both conceptually and practically.

Taking a step back

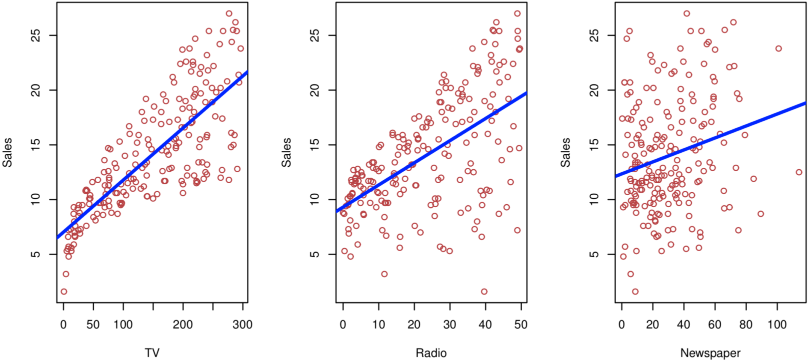
p is # of EV's (covariates)

primary focus in this topic.

- Simple linear regression: Y is univariate, $p = 1$
- Multiple linear regression: Y is univariate, $p > 1$ $--$ $>$ we discuss this as a way of introducing supervised learning
- Multivariate multiple linear regression: Y is a vector, $p > 1$

Consider the advertising data shown on the next slide.
Questions we might ask:

- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media contribute to sales?
- How accurately can we predict future sales?
- Is the relationship linear?
- Is there synergy among the advertising media?



Sales
25
20
15
10
5
0
50
100
200
300
TV
Sales
25
20
15
10
5
0
10
20
30
40
50
Radio
Sales
25
20
15
10
5
0
20
40
60
80
100
Newspaper

Simple linear regression using a single predictor X

- We assume a model
$$Y = \beta_0 + \beta_1 X + \epsilon,$$
where β_0 and β_1 are two unknown constants that represent the intercept and slope, also known as coefficients or parameters, and ϵ is the error term.
- Given some estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, we predict future sales using
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$
where $\hat{y}$ indicates a prediction of Y on the basis of $X = x$. The hat symbol denotes an estimated value.

Estimation of the parameters by least squares

- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X. Then $e_i = y_i - \hat{y}_i$ represents the ith residual.
- We define the residual sum of squares (RSS) as

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

or equivalently as

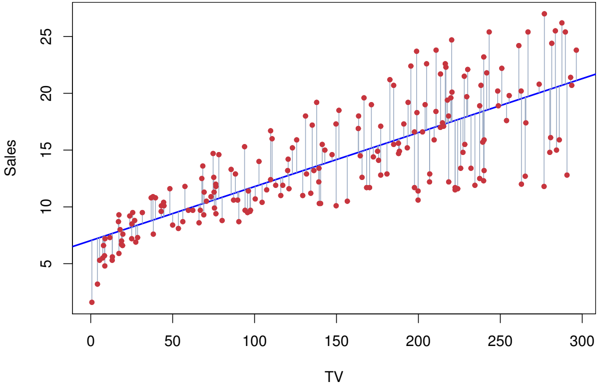
$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

- The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. The minimizing values can be shown to be

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ are the sample means.

Example: advertising data



The least squares fit for the regression of sales onto TV. In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.

- We compute the Residual Standard Error

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

where the residual sum-of-squares is $\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

- R-squared or fraction of variance explained is

$$R^2 = \frac{\text{TSS-RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}},$$

where $\mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the local total sum of squares.

(If we have >2 (i.e. the relationship this is not true.

- It can be shown that in this simple linear regression setting that $R^2 = r^2$, where r is the correlation between X and Y:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}. = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y}$$

Advertising data results

Quantity	Value
Residual Standard Error	3.26
R^2	0.612

- Here our model is

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

- We interpret β_j as the average effect on Y of a one unit increase in X_j, holding all other predictors fixed. In the advertising example, the model becomes

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{newspaper} + \epsilon.$$

- The ideal scenario is when the predictors are uncorrelated – a balanced design:
  - - Each coefficient can be estimated and tested separately.
  - - Interpretations such as "a unit change in X_j is associated with a β_j change in Y , while all the other variables stay fixed", are possible.
- Correlations amongst predictors cause problems:
  - - The variance of all coefficients tends to increase, sometimes dramatically
  - - Interpretations become hazardous — when X_j changes, everything else changes.
- Claims of causality should be avoided for observational data. Regression model finds association, not causation.
  - - Tall people have higher income

Estimation and Prediction for Multiple Regression

- Given estimates $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$, we can make predictions using the formula

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p.$$

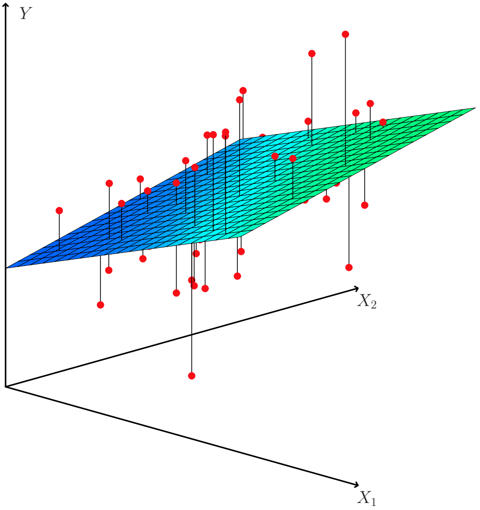
- We estimate $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$ as the values that minimize the sum of squared residuals

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2.$$

  The values of $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p)^T$ that minimize RSS are the multiple least squares regression coefficient estimates

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

  where X is the $n \times (p+1)$ design matrix and $y = (y_1, \ldots, y_n)^T$.


Y
X_2
X_1

Results for advertising data

	Coefficient
Intercept	2.939
TV	0.046
radio	0.189
newspaper	-0.001

Correlations:

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

magnitude v sign is different known as simpson paradox.

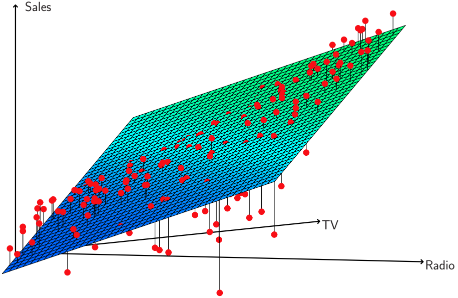
Interaction

- In our previous analysis of the *Advertising* data, we assumed that the effect on *sales* of increasing one advertising medium is independent of the amount spent on the other media.
- For example, the linear model

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \textit{newspaper} + \epsilon$$

  states that the average effect on *sales* of a one-unit increase in *TV* is always β_1, regardless of the amount spent on *radio*.

- But suppose that spending money on radio advertising actually increases the effectiveness of *TV* advertising, so that the slope term for TV should increase as *radio* increases.
- In this situation, given a fixed budget of \$100, 000, spending half on *radio* and half on *TV* may increase *sales* more than allocating the entire amount to either *TV* or to *radio*.
- In marketing, this is known as a synergy effect, and in statistics it is referred to as an interaction effect.



When levels of either *TV* or *radio* are low, then the true *sales* are lower than predicted by the linear model. But when advertising is split between the two media, then the model tends to underestimate *sales*.

Model takes the form

$$\begin{aligned} \text{sales} &= \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \textit{radio} \times \textit{TV} + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \text{radio}) \times \text{TV} + \beta_2 \times \text{radio} + \epsilon. \end{aligned}$$

Results:

	Coefficient
Intercept	6.7502
TV	0.0191
radio	0.0289
TV*radio	0.0011

Interpretation

- The R^2 for the interaction model is 96.8%, compared to only 89.7% for the model that predicts *sales* using *TV* and *radio* without an interaction term.
- This means that (96.8 - 89.7)/(100 - 89.7) = 69% of the variability in *sales* that remains after fitting the additive model has been explained by the interaction term.
- The coefficient estimates in the table suggest that an increase in TV advertising of \$1, 000 is associated with increased sales of

$$(\hat{\beta}_1 + \hat{\beta}_3 \times \text{radio}) \times 1000 = 19 + 1.1 \times \text{radio units.}$$

- An increase in radio advertising of \$1, 000 will be associated with an increase in sales of

$$(\hat{\beta}_2 + \hat{\beta}_3 \times \text{TV}) \times 1000 = 29 + 1.1 \times \text{TV units.}$$

STOP TUES 1/25/22 (Week 2; Lecture 3)