2/1/23 Day 3

Ch. 3 Linear Regression

1) Is there a relationship between advertising budget and sales?
2) How strong is the relationship between advertising budget and sales?
3) Which media contribute to sales?
4) How accurately can we estimate the effect of each medium on sales?
5) How accurately can we predict future sales?
6) Is the relationship linear?
7) Is there interaction among the advertising media?

3.1: Simple Linear Regression

(3.1) $Y \approx \beta_0 + \beta_1 X$

β_0, β_1 are coefficients or parameters, estimate by
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

3.1.1: Estimating the Coefficients

Consider a dataset $\{(x_i, y_i)\}_{i=1}^n$. We fit $\hat{\beta}_0, \hat{\beta}_1$ to this dataset using least squares; i.e., we minimize the residual sum of squares

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

One can show that the parameters are

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

3.1.2: Assessing the Accuracy of the Coefficient Estimates

We are assuming that $Y = \beta_0 + \beta_1 X + \varepsilon$ (3.5)

(3.5) is the population regression line, whereas estimation gives the least squares line.

• The least-squares are unbiased

For the population mean μ and the sample mean $\hat{\mu}$

$$Var(\hat{\mu}) = [SE(\hat{\mu})]^2 = \frac{\sigma^2}{n}$$

where $SE(\hat{\mu})$ is the standard error of $\hat{\mu}$ and σ is the standard deviation of Y (provided that the errors are uncorrelated).

$$SE(\hat{\beta}_0)^2 = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right], \quad SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

where $\sigma^2 = Var(\varepsilon)$

• $SE(\hat{\beta}_1)^2$ is smaller when the x_i are more spread out

• We have to estimate σ, which we get by the residual standard error

$$RSE = \sqrt{RSS/(n-2)}$$

Standard errors give <u>confidence intervals</u> in the usual way.

Hypothesis testing:

H_0: There is no relationship between X and Y.
H_a: There is some relationship between X and Y.

$H_0: \beta_1 = 0$
$H_a: \beta_1 \neq 0$

Compute a t statistic

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

which measures how far $\hat{\beta}_1$ is from 0 in units of standard errors

3.1.3: Assessing the Accuracy of the Model

Typically, the quality of fit of a linear regression is assessed by the residual standard error (RSE) and the R^2 statistic.

RSE

- RSE is the average amount that the response will deviate from the true regression line.
- Measures lack of fit; the smaller the better.

R^2

- RSE is measured in the units of Y, so it can be difficult to interpret.
- R^2 gives a proportion of variance explained

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

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

~~Note that~~

$$\cancel{\frac{RSS}{TSS} = \frac{RSS/n}{TSS/n} = \frac{MSE}{Var}}$$

~~so we're talking about variance explained.~~

TSS: variation in sample before regression

RSS: variation in sample unexplained by regression

TSS-RSS = variance explained

- R^2 is a measure of the linear relationship between X and Y.

Note that in the simple linear regression case, R^2 is identical to the correlation