Notes

Questions to Answer

Recall the `Advertising` data from **Chapter 2**. Here are a few important questions that we might seek to address:

1. **Is there a relationship between advertising budget and sales?**
2. **How strong is the relationship between advertising budget and sales?** Does knowledge of the advertising budget provide a lot of information about product sales?
3. **Which media are associated with sales?**
4. **How large is the association between each medium and sales?** For every dollar spent on advertising in a particular medium, by what amount will sales increase?
5. **How accurately can we predict future sales?**
6. **Is the relationship linear?** If there is approximately a straight-line relationship between advertising expenditure in the various media and sales, then linear regression is an appropriate tool. If not, then it may still be possible to transform the predictor or the response so that linear regression can be used.
7. **Is there synergy among the advertising media?** Or, in stats terms, is there an interaction effect?

Simple Linear Regression: Definition

Simple linear regression: Very straightforward approach to predicting response Y on predictor X.

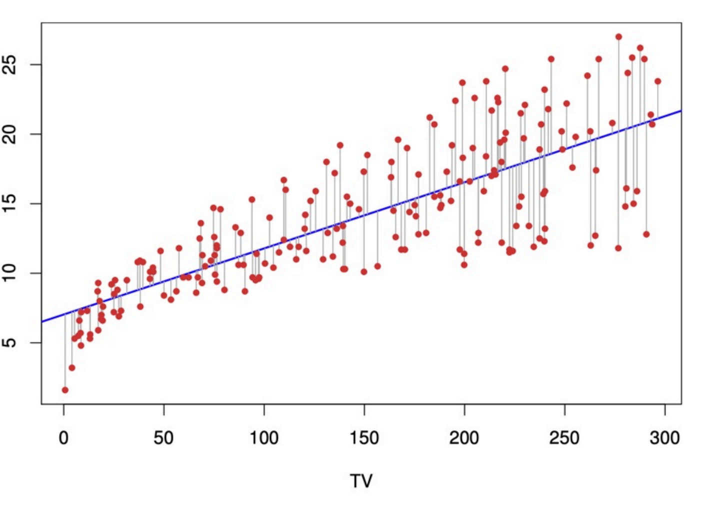
$$Y \approx \beta_0 + \beta_1 X$$

- Read "$\approx$" as *"is approximately modeled by."*
- β_0 = intercept
- β_1 = slope

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- $\hat{\beta}_0$ = our approximation of intercept
- $\hat{\beta}_1$ = our approximation of slope
- x = sample of X
- $\hat{y}$ = our prediction of Y from x
- hat symbol denotes "estimated value"
- Linear regression is a simple approach to supervised learning

Simple Linear Regression: Visualization



For the `Advertising` data, the least squares fit for the regression of `sales` onto `TV` is shown. The fit is found by minimizing the residual sum of squares. Each grey line segment represents a residual. In this case a linear fit captures the essence of the relationship, although it overestimates the trend in the left of the plot.

Simple Linear Regression: Math

- **RSS** = *residual sum of squares*

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

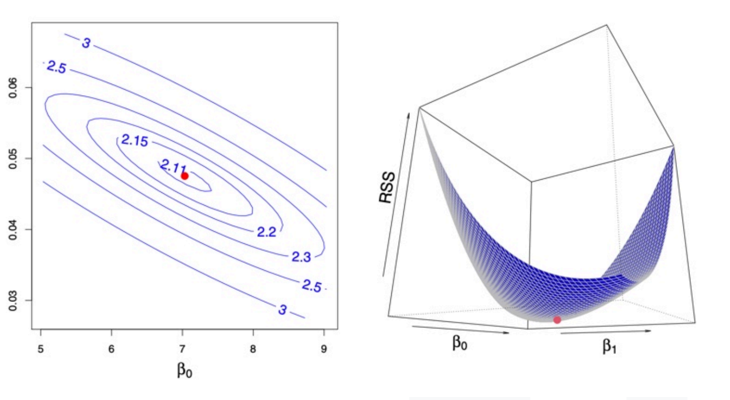
$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- $\bar{x}, \bar{y}$ = sample means of x and y

Visualization of Fit



Contour and three-dimensional plots of the RSS on the `Advertising` data, using `sales` as the response and `TV` as the predictor. The red dots correspond to the least squares estimates $hatbeta_0$ and $hatbeta_1$, given by (3.4).

Learning Objectives:

- Perform linear regression with a **single predictor variable.**

Assessing Accuracy of Coefficient Estimates

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- **RSE** = *residual standard error*
- Estimate of σ

$$\mathrm{RSE} = \sqrt{\frac{\mathrm{RSS}}{n-2}}$$

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

- **95% confidence interval:** a range of values such that with 95% probability, the range will contain the true unknown value of the parameter
  - If we take repeated samples and construct the confidence interval for each sample, 95% of the intervals will contain the true unknown value of the parameter

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1)$$

$$\hat{\beta}_0 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_0)$$

Learning Objectives:

- Estimate the **standard error** of regression coefficients.

Assessing the Accuracy of the Model

- **RSE** can be considered a measure of the *lack of fit* of the model. a
- R^2 statistic (also called coefficient of determination) provides an alternative that is in the form of a *proportion of the variance explained*, ranges from 0 to 1, a *good value* depends on the application.

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

where TSS is the *total sum of squarse*:

$$TSS = \Sigma(y_i - \bar{y})^2$$

Quiz: Can R^2 be negative?

Answer

Multiple Linear Regression

Multiple linear regression extends simple linear regression for *p* predictors:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon_i$$

- β_j is the *average* effect on Y from X_j holding all other predictors fixed.
- Fit is once again choosing the β_j that minimizes the RSS.
- Example in book shows that although fitting *sales* against *newspaper* alone indicated a significant slope (0.055 +- 0.017), when you include *radio* in a multiple regression, *newspaper* no longer has any significant effect. (-0.001 +- 0.006)

Important Questions

1. *Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?*

   F statistic close to 1 when there is no relationship, otherwise greater then 1.

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

2. *Do all the predictors help to explain Y, or is only a subset of the predictors useful?*

   p-values can help identify important predictors, but it is possible to be mislead by this especially with large number of predictors. Variable selection methods include Forward selection, backward selection and mixed. Topic is continued in Chapter 6.

3. *How well does the model fit the data?*

   R^2 still gives *proportion of the variance explained*, so look for values "close" to 1. Can also look at **RSE** which is generalized for multiple regression as:

$$RSE = \sqrt{\frac{1}{n-p-1} RSS}$$

4. *Given a set of predictor values, what response value should we predict, and how accurate is our prediction?*

   Three sets of uncertainty in predictions:

   - Uncertainty in the estimates of β_i
   - Model bias
   - Irreducible error ϵ

Qualitative Predictors

- Dummy variables: if there are k levels, introduce $k-1$ dummy variables which are equal to one ("one hot") when the underlying qualitative predictor takes that value. For example if there are 3 levels, introduce two new dummy variables and fit the model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

Qualitative Predicitor	x_{i1}	x_{i2}
level 0 (baseline)	0	0
level 1	1	0
level 2	0	1

- Coefficients are interpreted the average effect relative to the baseline.
- Alternative is to use index variables, a different coefficient for each level:

$$y_i = \beta_{01} + \beta_{02} + \beta_{03} + \epsilon_i$$

Extensions

- Interaction / Synergy effects

  Include a product term to account for synergy where one changes in one variable changes the association of the Y with another:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon_i$$

- Non-linear relationships (e.g. polynomial fits)

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots \beta_n X^n + \epsilon_i$$

Potential Problems

1. *Non-linear relationships*

   Residual plots are useful tool to see if any remaining trends exist. If so consider fitting transformation of the data.

2. *Correlation of Error Terms*

   Linear regression assumes that the error terms ϵ_i are uncorrelated. Residuals may indicate that this is not correct (obvious *tracking* in the data). One could also look at the autocorrelation of the residuals. What to do about it?

3. *Non-constant variance of error terms*

   Again this can be revealed by examining the residuals. Consider transformation of the predictors to remove non-constant variance. The figure below shows residuals demonstrating non-constant variance, and shows this being mitigated to a great extent by log transforming the data.

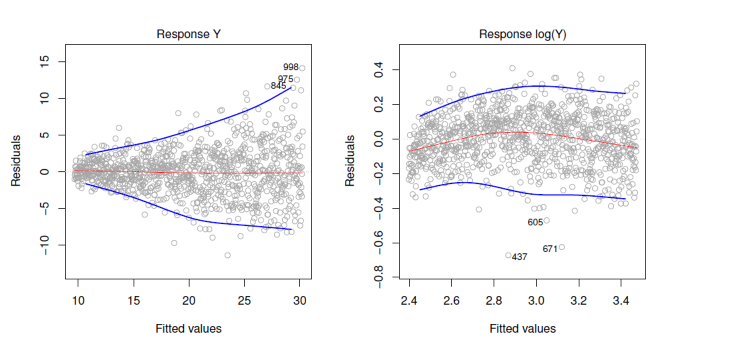


Figure 3.11

4. *Outliers*

   - Outliers are points with for which y_i is far from value predicted by the model (including irreducible error). See point labeled '20' in figure 3.13.
   - Detect outliers by plotting studentized residuals (residual e_i divided by the estimated error) and look for residuals larger then 3 standard deviations in absolute value.
   - An outlier may not effect the fit much but can have dramatic effect on the **RSE**.
   - Often outliers are mistakes in data collection and can be removed, but could also be an indicator of a deficient model.

5. *High Leverage Points*

   - These are points with unusual values of x_i. Examples is point labeled '41' in figure 3.13.
   - These points can have large impact on the fit, as in the example, including point 41 pulls slope up significantly.
   - Use *leverage statistic* to identify high leverage points, which can be hard to identify in multiple regression.

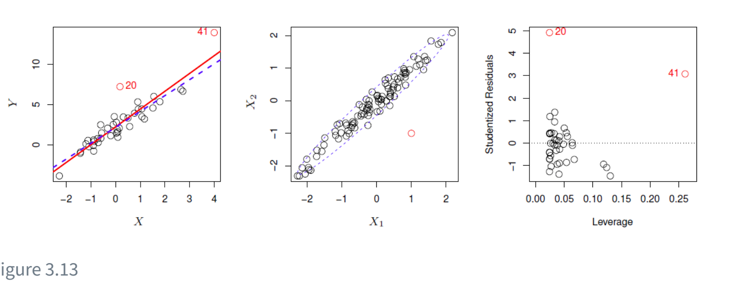


Figure 3.13

6. *Collinearity*

   - Two or more predictor variables are closely related to one another.
   - Simple collinearity can be identified by looking at correlations between predictors.
   - Causes the standard error to grow (and p-values to grow)
   - Often can be dealt with by removing one of the highly correlated predictors or combining them.
   - *Multicollinearity* (involving 3 or more predictors) is not so easy to identify. Use *Variance inflation factor*, which is the ratio of the variance of $\hat{\beta}_j$ when fitting the full model to fitting the parameter on its own. Can be computed using the formula:

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

where $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all the other predictors.

Answers to the Marketing Plan questions

1. **Is there a relationship between advertising budget and sales?**

   Tool: Multiple regression, look at F-statistic.

2. **How strong is the relationship between advertising budget and sales?**

   Tool: R^2 and **RSE**

3. **Which media are associated with sales?**

   Tool: p-values for each predictor's *t-statistic*. Explored further in chapter 6.

4. **How large is the association between each medium and sales?**

   Tool: Confidence intervals on $\hat{\beta}_j$

5. **How accurately can we predict future sales?**

   Tool:: Prediction intervals for individual response, confidence intervals for average response.

6. **Is the relationship linear?**

   Tool: Residual Plots

7. **Is there synergy among the advertising media?**

   Tool: Interaction terms and associated p-vales.

Comparison of Linear Regression with K-Nearest Neighbors

- This section examines the K-nearest neighbor (KNN) method (a non-parameteric method).
- This is essentially a k-point moving average.
- This serves to illustrate the Bias-Variance trade-off nicely.