3.2 Multiple Linear Regression

Prof. Lauren Perry

Multiple Linear Regression

In practice, X is usually composed of more than one predictor variable.

- Multiple linear regression will allow us to deal with multiple inputs.
  - Want to put all useful inputs into the model at once.
- It also allows us to better model the case where the relationship between X and Y is not linear.

Multiple Linear Regression

For p distinct predictors, the linear regression model takes the form

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

where X_j is the jth predictor and β_j quantifies the association between that variable and the response.

- We say that β_j is the average change in Y for a one unit increase in X_j, holding all other predictors fixed.

Estimating the Regression Coefficients

We make predictions using the formula

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

and we again estimate our parameters by minimizing the sum of squared residuals

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

the solutions to which are most easily represented using matrix algebra.

Estimating the Regression Coefficients

We will find these coefficients using R:

```
ads <- read.csv("~/Courses/STAT 196M/datasets/Advertising.csv")
mod1 <- lm(sales ~ TV + radio + newspaper, data=ads)
round(mod1$coefficients,3)
```

```
## (Intercept)          TV       radio   newspaper
##       2.939       0.046       0.189      -0.001
```

(Recall that the advertising data is in *thousands*.)

Estimating the Regression Coefficients

```r
summary(mod1)
```

produces the following:

```
Coefficient   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
```

Overall Model Fit

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

- This is a little more complex than in the simple linear regression setting, where we could just examine β_1.
- Here, we test

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

versus

$$H_a : \text{at least one } \beta_j \text{ is non-zero}$$

Overall Model Fit

This test uses an F-statistic:

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

where again $\text{TSS} = \sum(y_i - \bar{y})^2$ and $\text{RSS} = \sum(y_i - \hat{y}_i)^2$.

When there is no relationship between the predictors, we expect the F ratio to be close to 1.

Overall Model Fit

In R, the command `summary(mod1)` also produces the following:

```
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
```

Examining A Subset of Coefficients

Sometimes, we have reason to test whether a particular subset of q of the p coefficients are zero:

$$H_0 : \text{all of the } q \text{ coefficients are zero}$$

Here, we fit a second model that uses all of the variables *except* for the q variables of interest.

- We call this model's residual sum of squares RSS_0. Then

$$F = \frac{(\text{RSS}_0 - \text{RSS})/q}{\text{RSS}/(n-p-1)}$$

- The p-values provided earlier in the coefficient output correspond to the setting where the single corresponding variable is omitted.
  - i.e., the partial effect of adding that variable to the model.

Overall Model Fit

If at least one coefficient has a small p-value, why do we still need to look at the overall F-statistic?

- About 5% of the p-values associated with each variable will be below 0.05 *just by chance*.
- So, with a lot of predictors, it's relatively likely that we would see small p-values even if there is no association between the predictors and the response.
  - The F-statistic adjusts for number of predictors, so it doesn't have this problem.
- Thus, we want to examine overall model fit as well as the significance of each coefficient.

Deciding on Important Variables

Once we've decided the model is useful overall, we want to figure out *which* predictors are useful.

- We *could* just look at the p-values for each coefficient, but this can lead to some issues.
  - Ex: if p is large, we may make some false discoveries.
- Instead, we use *variable selection methods*.

Variable Selection Methods

The ideal approach is to examine models for all possible subsets of the predictors.

We can then compare these models using statistics like

1. Mallow's C_p
2. Akaike information criterion (AIC)
3. Bayesian information criterion (BIC)
4. Adjusted R^2

These are studied more extensively in Chapter 6.

Variable Selection Methods

Unfortunately, examining all possible subsets isn't always feasible.

- For $p = 2$ predictors, there are four possible models:
  - $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$
  - $Y = \beta_0 + \beta_1 x_1 + \epsilon$
  - $Y = \beta_0 + \beta_2 x_2 + \epsilon$
  - $Y = \beta_0 + \epsilon$
- But the number of possible models grows quickly!
- For p input variables, there are a total of 2^p possible subsets.

Variable Selection Methods

We need a way to automate variable selection that doesn't require us to examine all possible subsets.

Unfortunately, the methods discussed at this point in the textbook tend to lead to a variety of problems, so we will hold off on other options until Chapter 6.

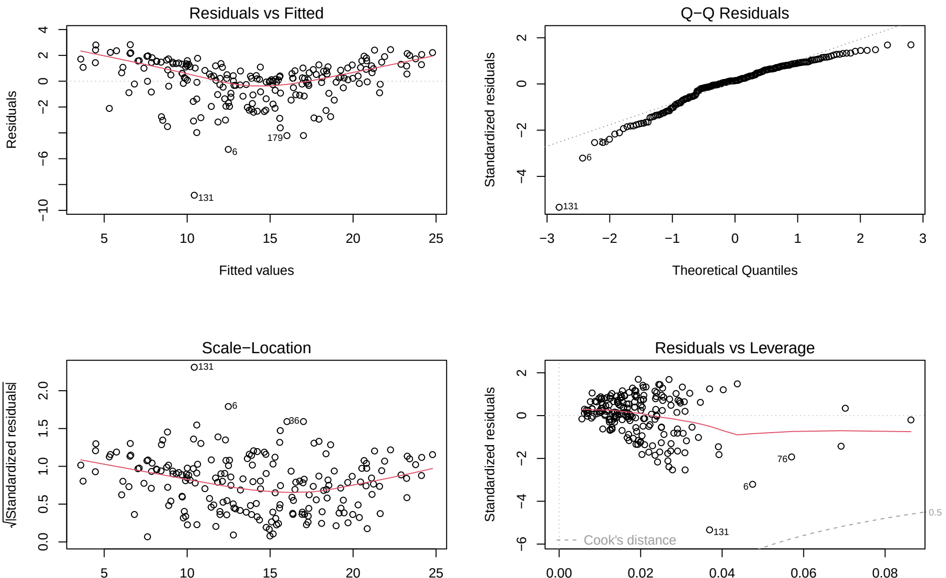
Model Fit

- Correlation R and the coefficient of determination R^2 are conceptually the same for multiple linear regression.
- However, R^2 will *always* increase as more variables are added to the model.
- Instead, we will use an *adjusted* R^2 value that takes into account the number of input variables p.

$$R^2_{adj} = 1 - \left[\left(\frac{n-1}{n-p-1} \right) (1 - R^2) \right]$$

  - We interpret R^2_{adj} the same way as R^2.
  - This value is shown in the regression model summary output in `R`.

We can also examine graphical summaries for model fit.



Predictions

It's straightforward to plug in values of X to the estimated regression line.

Sources of error/uncertainty:

1. The coefficients are estimates, so $\hat{f}(X)$ is only an estimate for $f(X)$.
    - A source of reducible error.
    - We can calculate confidence intervals for $\hat{Y}$.
2. In practice, assuming linearity is probably only an approximation.
    - Another source of reducible error, *model bias*.
    - We generally ignore this if the model is "good enough".
3. Random error ϵ.
    - Irreducible error.
    - We can also calculate *prediction intervals* for $\hat{Y}$.

Prediction Intervals

- Confidence intervals quantify uncertainty for a *mean*.
  - For a 95% CI, we say that 95% of intervals of that form will contain the true value of $f(X)$.
  - I.e., the *average* outcome y for a point x.
- Prediction intervals quantify uncertainty for a *single point*.
  - For a 95% PI, we say that 95% of intervals of that form will contain the true value of Y for a specific point.
  - I.e., the *specific* outcome y for a point x.