Lecture 24

Part 5 Linear Regression

Multiple Linear Regression

Multiple Linear Regression

- Multiple linear regression uses a single model to investigate how two or more independent variables, denoted $X_1, X_2, \ldots, X_k$, are related to the dependent variable, denoted Y.
- Ideally, we should include as many independent variables into the regression model *as are believed to be related* to the dependent variable.

Advantages

- Can use a *single* model to determine:
  - Which independent variables might be truly related to the dependent variable.
  - The nature of these relationships.
- Compared to a simple linear regression model, a multiple linear regression model will generally:
  - *Fit* the data better.
  - Produce better *predictions*, provided all the independent variables are truly related to the dependent variable.

Caveats

- We should not include as many independent variables as we can into the regression model.
- Why?
  - *Model selection* is a very important aspect of multiple linear regression.
  - Danger of *over-fitting* our sample data (can degrade predictive performance of model).
  - Problem of *multicollinearity* (parameter estimates become unreliable).
  - We can't fit a multiple linear regression model with more independent variables than observations in our sample.

Multiple Linear Regression Model

- The **multiple linear regression model** assumes that the relationship between the dependent and independent variables is given by:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

- β_0 is the intercept parameter.
- $\beta_1, \ldots, \beta_k$ are the **coefficient parameters** for the independent variables.
- ϵ is the error variable.

Multiple Linear Regression Model

- The **multiple linear regression model** assumes that the relationship between the dependent and independent variables is given by:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

- β_0 is the intercept parameter.
- $\beta_1, \ldots, \beta_k$ are the **coefficient parameters** for the independent variables.
- ϵ is the error variable.

Linearity

- The "linear" in linear regression refers to linearity in the coefficient parameters.
- For example, the following is a multiple linear regression model:

$$Y = \beta_0 + \beta_1 X_1^2 + \beta_2 \log X_2 + \beta_3 \frac{1}{X_3} + \epsilon$$

- Because we can rewrite the model as:

$$Y = \beta_0 + \beta_1 W_1 + \beta_2 W_2 + \beta_3 W_3 + \epsilon$$

where $W_1 = X_1^2$, $W_2 = \log X_2$ and $W_3 = \frac{1}{X_3}$.

Linearity

- However, the following is not a multiple linear regression model:

$$Y = \beta_0 + \beta_1 \cos(X_1 + \beta_2) + X_2^{\beta_3} + \epsilon$$

- Scatter plots of Y against each independent variable can help to determine in what form they should appear in the model.

Sample Data

- Our sample data now consists of a set of $k+1$ values for each observation:

$$\{(Y_1, X_{11}, X_{21}, \ldots, X_{k1}), \ldots, (Y_n, X_{1n}, X_{2n}, \ldots, X_{kn})\}$$

- The multiple linear regression model states that the Y_i value for the ith observation can be expressed as:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \epsilon_i$$

Response Surface

- Rather than a straight line, the multiple linear regression model is a *response surface* described by the following equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

- Due to inherent variability in the population, observations will not lie *exactly* on the response surface.
- The error variable ϵ_i signifies how far the Y_i value for each observation is from the response surface.

Model Assumptions

- We have the same assumptions as we did with simple linear regression, which are again stated in terms of ϵ.
- Namely, that the errors:
  - Are normally distributed.
  - Have mean equal to 0.
  - Have constant variance denoted by σ_ϵ^2.
  - Are independent.
- That is, $\epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$.

Multiple Linear Regression Model

- Based on the model assumptions, another way to state the multiple linear regression model is that Y is normally distributed with mean equal to:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

and variance equal to:

$$V(Y) = \sigma_\epsilon^2$$

Interpreting the Coefficient Parameters

- Suppose we have the following multiple linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

- How does the response surface change when we increase the value of X_1 by one unit?
- Consider two observations with X values given by (x_1, x_2) and $(x_1 + 1, x_2)$.

Interpreting the Coefficient Parameters

- Original observation (x_1, x_2):

$$E(Y_{\text{orig}}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- New observation $(x_1 + 1, x_2)$:

$$\begin{aligned} E(Y_{\text{new}}) &= \beta_0 + \beta_1(x_1 + 1) + \beta_2 x_2 \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_1 \\ &= E(Y_{\text{orig}}) + \beta_1 \end{aligned}$$

Interpreting the Coefficient Parameters

- An increase in X_1 by one unit leads to a change in the response surface by the amount β_1.
- The response surface can equivalently be thought of as the expected value of Y, so β_1 is also the *expected change* in Y when X_1 increases by one unit.
- In general, a coefficient parameter β_j represents the expected change in Y when X_j is increased by one unit, with all other independent variables held fixed.

Parameter Estimation

- Suppose we have obtained estimates $\hat{\beta}_0$, $\hat{\beta}_1$, ..., $\hat{\beta}_k$, for the parameters β_0, β_1, ..., β_k.
- The estimated or **fitted regression model** is:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k$$

- For each observation in our sample, the **fitted value** is given by:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki}$$

  and the **residual** is given by:

$$e_i = Y_i - \hat{Y}_i$$

Method of Least Squares

- Parameter estimates are again chosen as the values that make the residuals as small as possible.
- That is, the parameters β_0, β_1, ..., β_k are estimated by minimizing the **sum of squared residuals**:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2$$

$$= \sum_{i=1}^{n} \left(Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki} \right) \right)^2$$

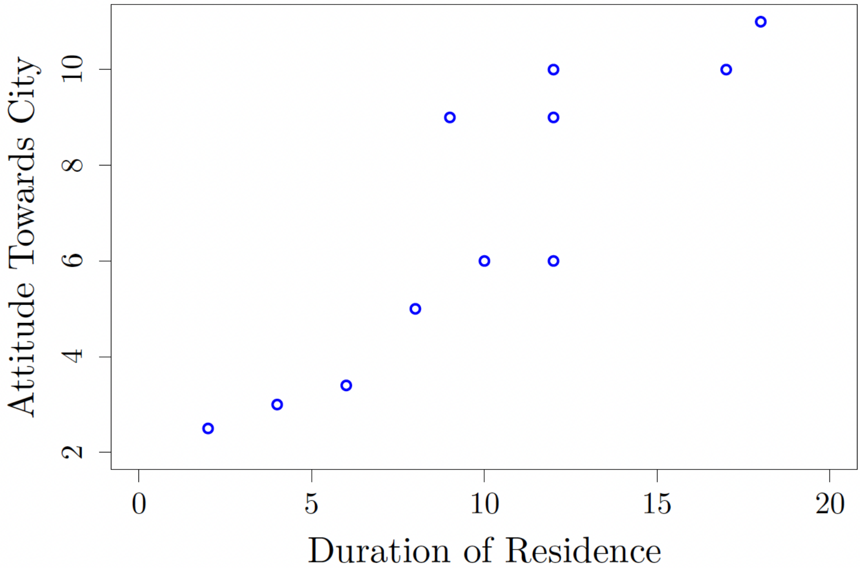
Parameter Estimation

- Unlike simple linear regression, we will not be calculating parameter estimates for multiple linear regression by hand.
- Instead, we will rely on software to fit the model, and the focus will be on *understanding* and *interpreting* the computer output.

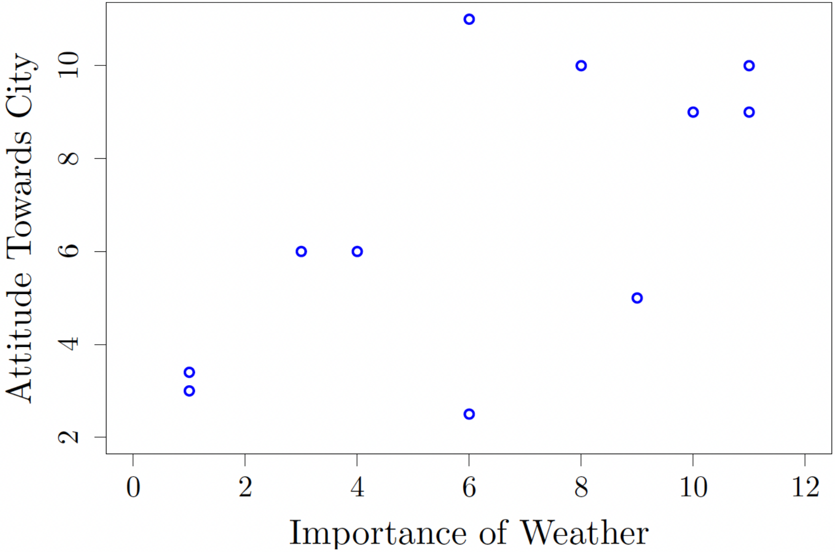
Attitude Towards the City Example

- Suppose we want to see whether people's attitudes towards the city they live in are linearly related to *two* variables:
  - Their duration of residence.
  - The importance they attach to weather.
- Last topic, the first step in the analysis was to construct a scatter plot to "eyeball" the data.
- Now that we have two independent variables, we should construct two scatter plots:
  - Attitude against duration of residence.
  - Attitude against importance attached to weather.

Attitude against Duration of Residence



Attitude against Importance of Weather



Model Specification

- Next, we need to specify our model.
- Let Y denote attitude towards the city, X_1 denote the duration of residence and X_2 denote the importance attached to weather.
- Since both X_1 and X_2 appear to be linearly related to Y, a reasonable model might be:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

R Output

```
Call:
lm(formula = attitude ~ duration + weather, data = city.dat)

Residuals:
Min        1Q   Median       3Q       Max
-1.56859 -0.79732  0.03449  0.47779  1.82480


Coefficients:
Estimate                Std. Error t stat. Pr(>|t|)
(Intercept)  0.45755    0.94094   0.486  0.639817
duration     0.46751    0.08907   5.249  0.000775
weather      0.26344    0.11784   2.236  0.055810

Residual standard error: 1.243 on 8 degrees of freedom
Multiple R-squared: 0.8724,  Adjusted R-squared: 0.8405
F-statistic: 27.35 on 2 and 8 DF,  p-value: 0.0002649
```