Chapter 11: Theory of Linear Models

- In past chapters, we considered our data to be a single sample $Y_1, \ldots, Y_n$ which was iid.
- This implied that $E(Y_i)$ was constant
- In many situations, we measure or observe two or more variables, and $E(Y_i)$ depends on the value(s) of the other variable(s).

Examples: (1)

(2)

- In regression models, our main variable of interest, Y, is called the dependent variable (or response variable).
- Its expected value depends on an independent variable (denoted x) or several independent variables (denoted $x_1, x_2, \ldots, x_k$).

Note: "Independent" here has nothing to do with the notion of "independent random variables".

- A probabilistic model implies that the independent variable(s) cannot predict the response with certainty.
- There is random error in a probabilistic model (unlike a deterministic model).

Example: $Y = \beta_0 + \beta_1 x + \varepsilon$

where β_0 and β_1 are constants and the random component ε is a r.v. with $E(\varepsilon) = 0$.

- For a given value of x:

Picture:

Note: At each x value, Y follows a different distribution:

- We usually assume the mean of Y may _____ with x, but the shape of Y's distribution and the variance of Y are _____ at each x value.

Example of Sample Data Following Such a Model

11.2 Linear Statistical Models

- While we could choose to model $E(Y)$ as any function $f(x_1, \ldots, x_k)$ of our independent variable(s), in this chapter we focus on linear models.

- By this we mean models that are a linear function of the parameters $\beta_0, \beta_1, \ldots$ (not necessarily linear in $x_1, x_2, \ldots$).

Examples of Linear Models

Defn: A linear statistical model takes the form:

where $\beta_0, \beta_1, \ldots, \beta_k$ are unknown parameters, ε is a r.v. (we will assume $E(\varepsilon) = 0$) and $x_1, x_2, \ldots, x_k$ are variables assuming known values.

Potential Problems: (1) The form of our specified model may not match the reality of our data (an issue of applied data analysis)

(2) Even if the model is correct, we do not know $\beta_0, \beta_1, \ldots, \beta_k$ (we will have to estimate them)

- Once we estimate the β_i's, they often have meaningful interpretations in the context of the data set.

11.3 Least Squares Estimation

- Let us consider observing n paired observations on a dependent variable $y_1, \ldots, y_n$ and an independent variable $x_1, \ldots, x_n$ and fitting a simple linear regression model:

Example scatterplot of data:

- Which estimates of β_0 and β_1 yield the straight line that "best fits" these data?
- We want to choose a line that is "close" to the observed data points.

Least Squares Method: Choose the line so that the sum of the squared vertical deviations between the points and the line is as small as possible.

Picture:

- Denote the estimates producing this "least squares line" as
- Note: is an estimator of
- Then for $i=1, \ldots, n$,

is the i-th "fitted value".

Picture:

- The SSE (sum of squared errors) is thus:

- We minimize SSE with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ by _____:

- Solve these "least-squares equations" simultaneously for $\hat{\beta}_0$ and $\hat{\beta}_1$:

Example 1: Researchers measured skiers' time until exhaustion (running at a fixed pace on a treadmill) to relate this to the skiers' time in a 20-km ski race. Data for 5 skiers:

X (exhaustion time in min): 8.4 9.0 9.6 10.0 11.0

Y (20 km ski time in min): 71.4 68.7 69.4 63.0 62.6

Picture:

- Find the estimated least squares line and estimate the expected 20-km ski time for skiers with a 10½-minute treadmill exhaustion time.