REGRESSION MODELS

One approach: Use theoretical considerations to develop a model for the mean function or other aspects of the conditional distribution.

The next two approaches require some terminology:

Error: e|x = Y|(X = x) - E(Y|X = x)
= Y|x - E(Y|x) for short

- So Y|x = E(Y|x) + e|x (Picture this …)
- E|x is a random variable
- E(e|x) = E(Y|x) - E(Y|x)) = E(Y|x) - E(Y|x) = 0
- Var(e|x) =
- The distribution of e|x is

Second approach:

Bivariate Normal Model: Suppose X and Y have a bivariate normal distribution.

Recall:
- Y|x is normal
- $E(Y|\,x) = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X)$ (linear mean function)
- $Var(Y|x) = \sigma_Y^2(1 - \rho^2)$ (constant variance)

Thus:

- E(Y|x) = a + bx
- $Var(Y|x) = \sigma^2$

where

b =

a =

σ^2 =

Implications for e|x:

- e|x ~

Third approach: Model the conditional distributions

"The" Simple Linear Regression Model

Version I:

Only one assumption: E(Y|x) is a linear function of x.

Typical notation: $E(Y|x) = \eta_0 + \eta_1 x$ (or $E(Y|x) = \beta_0 + \beta_1 x$)

Equivalent formulation: $Y|x = \eta_0 + \eta_1 x + e|x$

Interpretations of parameters: (Picture!)

η_1:

η_0 : (if …)

When model fits:

- X, Y bivariate normal
- Other situations
  Example: Blood lactic acid
    Why is this not bivariate normal?
- Model might also be used when mean function is not linear, but linear approximation is reasonable.

Version II: *Two assumptions*:

1. $E(Y|x) = \eta_0 + \eta_1 x$ (linear mean function)
2. $Var(Y|x) = \sigma^2$ (constant variance)

Equivalent formulation:

1'. $E(Y|x) = \eta_0 + \eta_1 x$ (linear mean function)

2': $Var(e|x) = \sigma^2$ (constant error variance)

[Draw a picture!]

When model fits:

- If X and Y have a bivariate normal distribution.
- Credible (at least approximately) in many other situations as well, for transformed variables if not for the original predictor. (i.e., it's often useful)

Until/unless otherwise stated, we will henceforth assume the Version II model -- i.e., we all assume conditions (1) and (2) (equivalently, (1') and (2').)

Thus we have *three parameters*:

$\boldsymbol{\eta_0}$, $\boldsymbol{\eta_1}$ (which determine E(Y|x) and $\boldsymbol{\sigma^2}$ (which determines Var(Y|x)

The goal: To estimate η_0 and η_1 (and later σ^2) from data.

Notation: The estimates of η_0 and η_1 will be called $\hat{\eta}_0$ and $\hat{\eta}_1$, respectively. From $\hat{\eta}_0$ and $\hat{\eta}_1$, we obtain an estimate

$$\hat{E}(Y|x) = \hat{\eta}_0 + \hat{\eta}_1 x$$

of E(Y|x).

Note: $\hat{E}(Y|x)$ is the same notation we used earlier for the lowess estimate of E(Y|x). Be sure to keep the two estimates straight.

More terminology:

- We label our data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
- $\hat{y}_i = \hat{\eta}_0 + \hat{\eta}_1 x_i$ is our resulting estimate $\hat{E}(Y|x_i)$ of $E(Y|x_i)$. It is called the i^{th} *fitted value* or i^{th} *fit*.
- $\hat{e}_i = y_i - \hat{y}_i$ is called the i^{th} *residual.*

Note: $\hat{e}_i$ (the residual) is analogous to but not the same as $e|x_i$ (the error). Indeed, $\hat{e}_i$ can be considered an estimate of the error $e_i = y_i - E(Y|x_i)$.

Picture:

Least Squares Regression

- Method of obtaining estimates $\hat{\eta}_0$ and $\hat{\eta}_1$ for η_0 and η_1

Consider lines $y = h_0 + h_1x$. We want the one that is "closest" to the data points (x_1, y_1), (x_2, y_2), … , (x_n, y_n) collectively.

What does "closest" mean? Various possibilities:

1. The usual math meaning: shortest perpendicular distance to point.
   Problems:
   - Gets unwieldy quickly.
   - We're really interested in getting close to y for a given x -- which suggests:

2. Minimize $\sum d_i$, where $d_i = y_i - (h_0 + h_1x_i)$ = vertical distance from point to candidate line. (Note: If the candidate line is the desired best fit then d_i = .)
   Problem: Some d_i's will be positive, some negative, so will cancel out in the sum.
   This suggests:

3. Minimize $\sum |d_i|$. This is feasible with modern computers, and is sometimes done.
   Problems:
   - This can be computationally difficult and lengthy.
   - The solution might not be unique.
     Example:
   - The method does not lend itself to inference about the fit.

4. Minimize $\sum d_i^2$
   This works!
   See demo.

Terminology:

- $\sum d_i^2$ is called the *residual sum of squares* (denoted *RSS*(h_0, h_1)) or the *objective function*.
- The values of h_0 and h_1 that minimize RSS(h_0, h_1) are denoted $\hat{\eta}_0$ and $\hat{\eta}_1$, respectively, and called the *ordinary least squares* (or *OLS*) *estimates* of η_0 and η_1