REGRESSION MODELS

One approach: Use theoretical considerations specific to the situation to develop a specific model for the mean function or other aspects of the conditional distribution. Possibly use data to estimate coefficients or other parameters.

The next two approaches (which have broader applicability) make model assumptions about joint or conditional distributions. They 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
- $E(e|x) = E(Y|x - E(Y|x)) = E(Y|x) - E(Y|x) = 0$
- $\mathrm{Var}(e|x) =$
- The distribution of $e|x$ is

Second approach: Start with assumptions about the joint distributions of the variables.

Example: **The Bivariate Normal Model**. Suppose X and Y have a bivariate normal distribution. (Of course, we need to have evidence that this model assumption is reasonable or approximately true before we are justified in using this model.)

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

Thus:

- $E(Y|x) = \alpha + \beta x$
- $\mathrm{Var}(Y|x) = \sigma^2$

where

$\beta =$

$\alpha =$

$\sigma^2 =$

Implications for $e|x$:

- $e|x \sim$

Third approach: Model the conditional distributions

Most widely used/basic example: **"The" Simple Linear Regression Model**

Only one explanatory variable is involved.

Version 1:

Only one assumption: 1. 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 …)

Some cases where this 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.

Note: In this model, Y is a random variable, but X need not be.

We have *two parameters*: $\boldsymbol{\eta_0}, \boldsymbol{\eta_1}$ These determine E(Y|x)

We need to estimate η_0 and η_1 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 <u>not</u> the same as $e|x_i$ (the error). Indeed, $\hat{e}_i$ can be considered an estimate of the error $e_i = y_i - E(Y|x_i)$.

Draw a picture:

Idea behind estimation methods: Consider lines $y = h_0 + h_1x$. We want the one that is "closest" to the data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ 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 method is called "Minimum Absolute Deviation," (MAD) or "Least Absolute Deviation" (LAD). It is feasible with modern computers, and increasingly available. (e.g., Stata and R's quantreg package)
   Problems:
   - It can be computationally difficult and lengthy.
   - The solution might not be unique.
     Example:
   - The method does not lend itself as readily to inference for the estimates.

   Strong points: It may be preferable to method 4 (below) in some situations; e.g.:
   - There is concern that outliers might be too influential.
   - The conditional distributions Y|X are not symmetric and the goal is to estimate the conditional median rather than the conditional mean.
   - The conditional distributions have heavy tails.

4. Minimize $\sum d_i^2$ ("Method of Least Squares")
   This works well!
   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*