STAT 516 --- STATISTICAL METHODS II

STAT 516 is primarily about <u>linear models</u>.

<u>Model</u>: A mathematical equation describing (approximating) the relationship between two (or more) variables.

● Any assumptions we make about the variables are also part of the model.

<u>Simple Linear Regression (SLR) Modeling</u>

● Analyzes the relationship between <u>two quantitative</u> variables.

● We have a sample, and for each observation, we have data observed for two variables:

<u>Dependent (Response) Variable</u>: Measures major outcome of interest in study (often denoted Y)

<u>Independent (Predictor) Variable</u>: Another variable whose value may explain, predict or affect the value of the dependent variable (often denoted X)

Example:

● In SLR, we assume the relationship between *Y* and *X* can be mathematically approximated by a straight-line equation.

● We assume this is a <u>statistical</u> relationship: not a perfect linear relationship, but an <u>approximately</u> linear one.

Example: Consider the relationship between

X =

Y =

We might expect that gas spending changes with distance traveled – maybe nearly linearly.

● If we took a sample of trips and measured *X* and *Y* for each, would the data fall exactly along a line?

Picture:

● Our goal is often to predict *Y* (or to estimate the mean of *Y*) based on a given value of *X*.

Examples:

<u>Simple Linear Regression Model</u>: (expressed mathematically)

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Deterministic Component:

Random Component:

Regression Coefficients:

β_0 =

β_1 =

ε =

We assume ε has a

Since ε has mean 0, the mean (expected value) of *Y*, for a given *X*-value, is:

- **This is called the conditional mean of *Y*.**

● **The deterministic part of the SLR model is simply the mean of *Y* <u>for any value</u> of *X*:**

Example: Suppose $\beta_0 = 2$, $\beta_1 = 1$.

Picture:

●**When $X = 1$, E(Y) =**

● **When $X = 2$, E(Y) =**

● **The actual *Y* values we observe for these *X* values are a little different – they vary along with the random error component ε.**

Assumptions for the SLR model:

- **The linear model is correctly specified**
- **The error terms are independent across observations**
- **The error terms are normally distributed**
- **The error terms have the same variance, σ^2, across observations**

<u>Notes</u>:

- **Even if *Y* is linearly related to *X*, we rarely conclude that *X* <u>causes</u> *Y*.**

  -- This would require eliminating all unobserved factors as possible causes for *Y*.

- **We should not use the regression line for extrapolation: that is, predicting *Y* for any *X* values outside the range of our observed *X* values.**

  -- We have no evidence that a linear relationship is appropriate outside the observed range.

Picture:

Example: **Data gathered on 58 houses (Table 7.2, p. 328)**

X = size of house (in thousands of square feet)
Y = selling price of house (in thousands of dollars)

● **Is a linear relationship between X and Y appropriate?**

On computer, examine a scatter plot of the sample data.

● **How to choose the "best" slope and intercept for these data?**

Estimating Parameters

● **β_0 and β_1 are unknown parameters.**
● **We use the sample data to find estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.**
● **Typically done by choosing $\hat{\beta}_0$ and $\hat{\beta}_1$ to produce the least-squares regression line:**

Picture:

For each data point, <u>predicted</u> Y-value is denoted $\hat{Y}$.

Picture:

- **Residual (or error) = $Y - \hat{Y}$ for each data point.**

- **We want our line to make these residuals as small as possible.**

<u>Least-squares line</u>: The line chosen so that the <u>sum of squared residuals</u> (SSE) is minimized.

- **Choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize:**

__Example__: **(House Price data):**
The following can be calculated from the sample:

So the estimates are:

Our estimated regression line is:

● Typically, we calculate the least-squares estimates on the computer.

__Interpretations__ **of estimated slope and intercept:**