# P8130: Biostatistical Methods I

## Lecture 15: Multiple Linear Regression

Cody Chiuzan, PhD
Department of Biostatistics
Mailman School of Public Health (MSPH)

# Recap

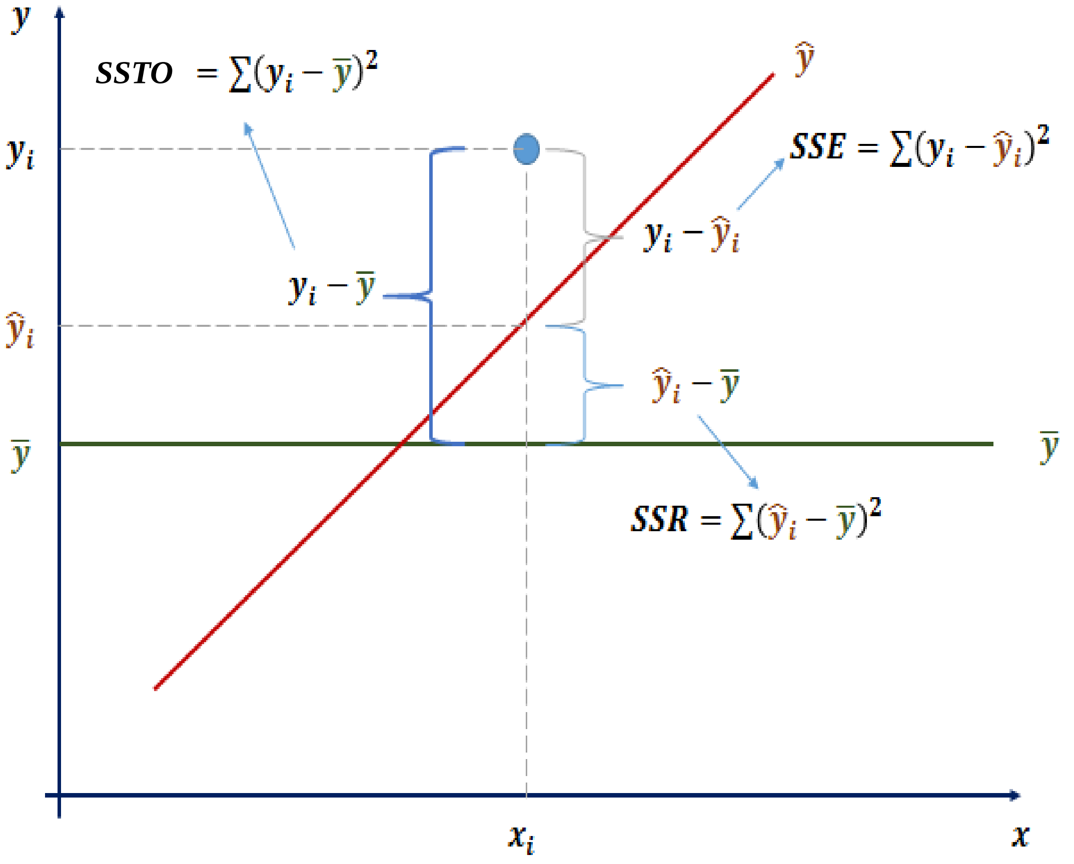
- Lecture 14 presented ‘Inferences in SLR’
  - Hypothesis testing for slope (and intercept)
  - Confidence and prediction intervals
    - Confidence interval (CI): inference on a parameter, range meant to cover the value of the parameter.
    - Prediction interval (PI): a range of values to be taken by a random variable (wider than CI)
  - Correlation (strength of the association) vs slope (rate of change)
  - Coefficient of determination: $r^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$

# Recap

- Partitioning the Sum of Squares:

***SSTO = SSR + SSE***

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$



# Partitioning the Sum of Squares

- Analysis of variance (ANOVA) vs Regression (practically the same)
- ANOVA results generalize immediately to multiple linear regression

ANOVA table for Simple Linear Regression

| Source of Variation | $SS$ | $df$ | $MS$ | $E\{MS\}$ |
|---|---|---|---|---|
| Regression | $SSR = \Sigma(\hat{Y}_i - \bar{Y})^2$ | 1 | $MSR = \frac{SSR}{1}$ | $\sigma^2 + \beta_1^2 \Sigma(X_i - \bar{X})^2$ |
| Error | $SSE = \Sigma(Y_i - \hat{Y}_i)^2$ | $n-2$ | $MSE = \frac{SSE}{n-2}$ | $\sigma^2$ |
| Total | $SSTO = \Sigma(Y_i - \bar{Y})^2$ | $n-1$ | | |

# Multiple Linear Regression (MLR)

- Allows inferences about the relationship between the outcome (Y) and a predictor (X), *<u>while adjusting for other predictors (covariates)</u>*

- Used to control/account for confounding and extremely useful in observational studies

- A consequence of not adjusting: the estimated coefficients corresponding to a certain predictor might be *<u>biased</u>*

# Estimation Bias

- Let us assume that you have two predictors $X_1$and $X_2$, but you are interested primarily in the <u>relationship between $Y$ and $X_1$</u>.

- Should you fit a regression model including only $X_1$ or $(X_1 and\ X_2)$?

- It depends:
  - The ‘bias’ is a function of correlation between the two covariates
  - If the correlation is high -> bias will be high
  - If the correlation is small -> bias will be small
    - If small or zero correlation, then no need to adjust for $X_2$

# MLR: Motivation

- Definitely more realistic than SLR
- Can fit multiple predictors of different types (continuous and categorical)
- Improvement in estimation
- Allows testing of multiple effects and their interaction(s)

# MLR: Formulation

- Data are observed from *n* subjects: $(Y_i, X_{i1}, X_{i2}, \dots, X_{ip})$ for $i = 1, 2, \dots, n$.

- The MLR model is given by:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots \beta_p X_{ip} + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2).$$

- Same assumptions as in SLR:
  - Uncorrelated error terms with mean 0 and constant variance (i.i.d.)
  - Linearity in parameters (p predictors, p+1 parameters – why?)

- How to we find the model estimates?
  - Least Squares and Maximum Likelihood Methods

# MLR: Matrix Notation

- The general form of a linear model is given by:

$$\tilde{Y} = \mathbf{X}\tilde{\beta} + \tilde{\varepsilon}$$

- Where $\tilde{Y}$ is the $N \times 1$ vector of observed responses

  $\mathbf{X}$ is the $N \times (p + 1)$ design matrix of fixed constants

  $\tilde{\beta}$ is the $(p + 1) \times 1$ vector of fixed, but unknown parameters

  $\tilde{\varepsilon}$ is the $N \times 1$ vector of (unobserved) errors

- In this formulation, $p$ denotes the number of predictors.

# MLR: Interpretation(s)

- In multiple linear regression, each model coefficient still shows the difference/change in the expected outcome, but …

- ‘*Adjusted for*’ or ‘*Controlling for*’ or ‘*Holding all other variables constant*’

- It is imperative that your interpretation includes one of the phrases above.

# MLR Interpretation: Example

MLR using two covariates: BEDS and INFRISK (data ‘Hospital.csv’)

lm(formula = data_hosp$LOS ~ data_hosp$BEDS + data_hosp$INFRISK)

Coefficients:

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 6.2703521 | 0.5038751 | 12.444 | < 2e-16 *** |
| data_hosp$BEDS | 0.0024747 | 0.0008236 | 3.005 | 0.00329 ** |
| data_hosp$INFRISK | 0.6323812 | 0.1184476 | 5.339 | 5.08e-07 *** |

---

Residual standard error: 1.568 on 110 degrees of freedom
Multiple R-squared: 0.3388, Adjusted R-squared: 0.3268
F-statistic: 28.19 on 2 and 110 DF, p-value: 1.31e-10

# MLR Interpretation: Example

- Write the fitted line equation:

- Interpret the coefficient corresponding to predictor BEDS

- Calculate and compare the LOS for two hospitals with 400 and 500 beds …

# MLR: Other Type of Predictors

- MLR accommodates not only continuous predictors but also:
  - Categorical (e.g., Gender, Treatment Arms)
  - Ordinal (e.g., Disease Severity/Status)
- If only one predictor and it is categorical, then we have a ONE-WAY ANOVA
- In regression, categorical variables are modeled using dummy or indicator variables.

# Indicator Variables

- The indicator or ‘dummy’ variable is defined as:

$$I(condition) = 1,\ if\ condition\ is\ TRUE$$

$$I(condition) = 0,\ if\ condition\ is\ FALSE$$

- For example, let us take variable GENDER:

$$I(gender) = 1,\ if\ male\ is\ TRUE$$

$$I(gender) = 0,\ if\ male\ is\ FALSE$$

- Notice that indicator variables only take values 0 and 1.
- The number of indicator variables is always $p$-$1$, where $p$ represents the number of levels for the qualitative predictor

# Indicator Variables

- In a clinical trial, let variable TREATMENT have 4 levels corresponding to the number of treatment arms:
  - ‘Placebo’, ‘Chemotherapy’, ‘Immuno Agent’, ‘Chemotherapy + Immuno Agent’
- Regression will use 3 dummy variables to model TREATMENT

| **Treatment** | **$I_1$** | **$I_2$** | **$I_3$** |
|---|---|---|---|
| Placebo | 0 | 0 | 0 |
| Immuno | 1 | 0 | 0 |
| Chemo | 0 | 1 | 0 |
| Immuno + Chemo | 0 | 0 | 1 |

- Placebo is considered the reference category (not accounted for?)

# Indicator Variables

- The regression model for the clinical trial example will change to:

$$E(Y) = \beta_0 + \beta_1 I(Trt = Immuno) + \beta_2 I(Trt = Chemo) + \beta_3\, I(Trt = Combo)$$

$$E(Y|Trt = Immuno) =$$

$$E(Y|Trt = Chemo) =$$

$$E(Y|Trt = Combo) =$$

$$E(Y|Trt = Placebo) =$$

- R will generate results for each level of the categorical predictor vs the reference category
- What if we want an overall (general) test for the Treatment variable?

# MLR Categorical Predictors: Matrix Notation

- Using the same clinical trial example with 4 treatment arms, write the matrix formulation for the regression model:

$$\tilde{Y} = \mathbf{X}\tilde{\beta} + \tilde{\varepsilon}$$

- What are the dimensions of each vector/matrix?

- What are the components of each vector/matrix?

# MLR: R Practice

- Fit MLR with continuous/categorical predictors
- Compare the ‘intercept’ vs ‘no intercept’ models with categorical predictors
- Change the reference category for categorical predictors
- ANOVA ‘general test’ for a categorical predictor

# MLR: Interactions

- Interactions can be formed between:
  - Two continuous predictors (difficult to interpret)
  - Continuous x categorical
  - Categorical x categorical
  - Etc.

- Most common are two-way interactions, but three-, four-way interactions are also possible (Don’t!)
- <u>Interaction</u>: the effect of one independent predictor on the dependent variable depends on the values of another independent predictor (two-way interaction).

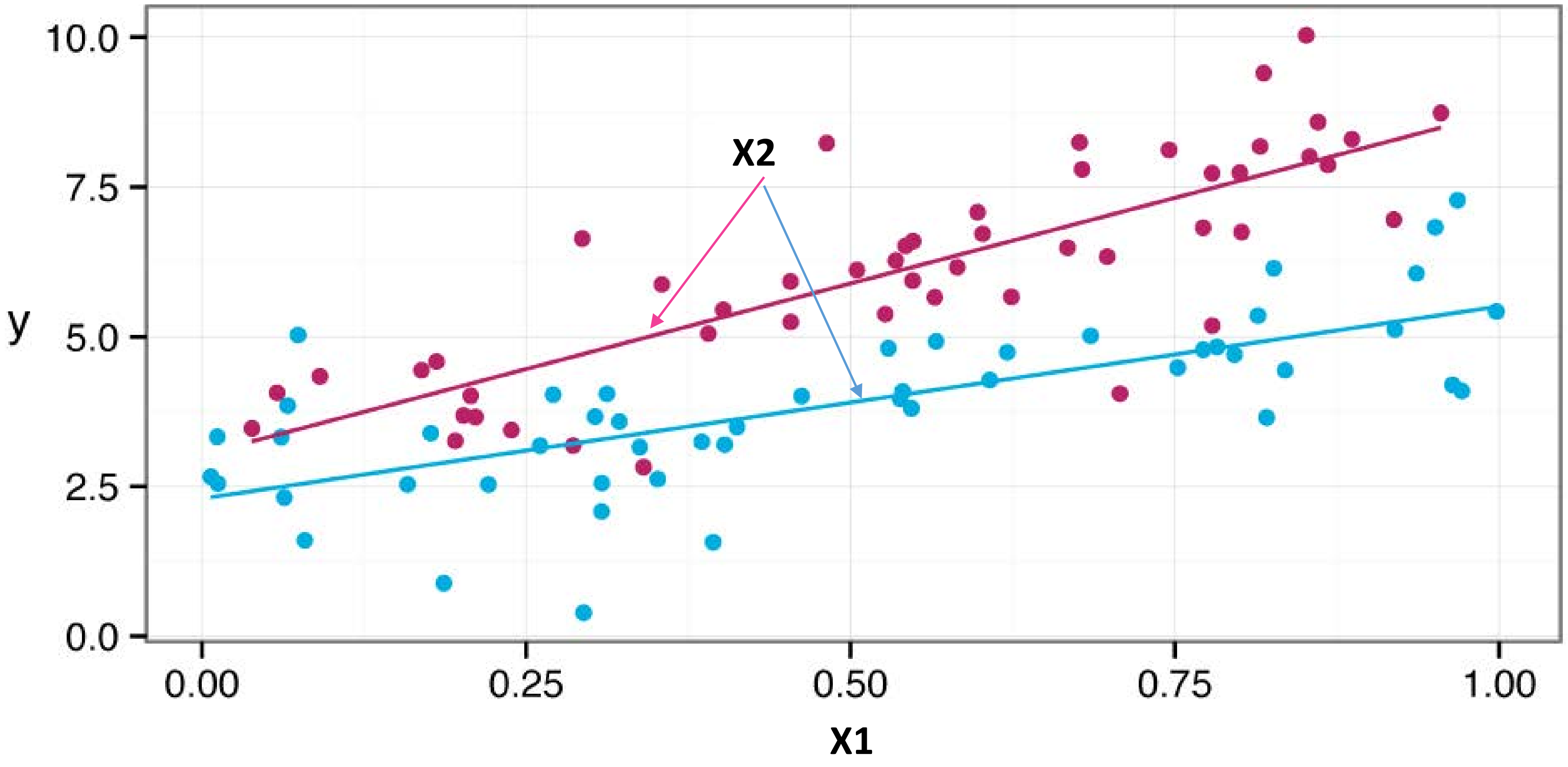
# MLR: Interactions

- Suppose that we want to test if the effect of $X_1$ on $Y$ is different with respect to the levels of $X_2$
- We can we fit separate regression models for each level category of $X_2$.

  OR
- Add interaction effects to the model.
- Careful: if the interaction term is significant, then you CANNOT assess the main effects separately
  - You can estimate the response for different predictor values, but the interaction term needs to be taken into consideration.

# MLR: Interactions

- Indication of interaction: Unparallel slopes



# MLR: Interactions

- Model interactions
  - If the interaction term is NOT significant, remove it and re-fit the model only with the main effects
  - If the interaction term is significant, you can only calculate the estimated Ys, taking into account the interaction term
- Again, let us consider the clinical trial example with TREATMENT, GENDER and their interaction, i.e., the *Saturated Model:*

$$
\begin{aligned}
E(Y) = \beta_0 + \beta_1 I(Trt = Immuno) + \beta_2 I(Trt = Chemo) + \beta_3\, I(Trt = Combo) + \\
\beta_4 I(Gender = Male) + \beta_5 I(Gender = Male) \cdot I(Trt = Immuno) + \\
\beta_6 I(Gender = Male) \cdot I(Trt = Chemo) + \beta_7 I(Gender = Male) \cdot I(Trt = Combo)
\end{aligned}
$$

- *What is the expected response for a male subject treated with chemo?*

# Readings

Kutner *et al.*, Applied Linear Statistical Models

- Chapter 6, Sections: 6.1 – 6.5
- Chapter 8, Sections: 8.3 – 8.5