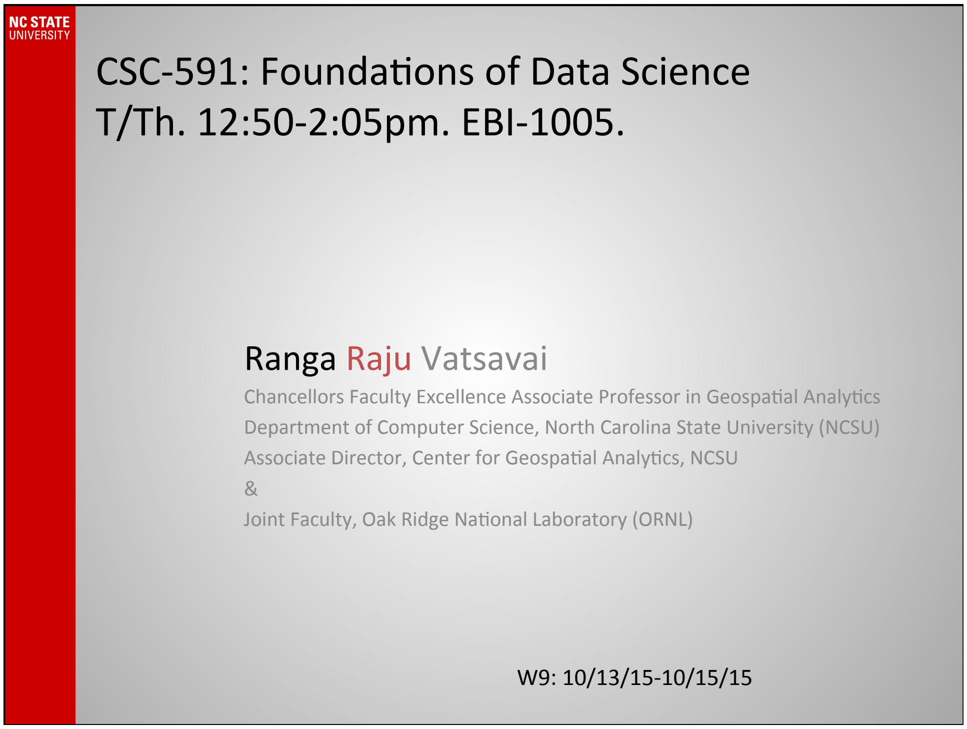
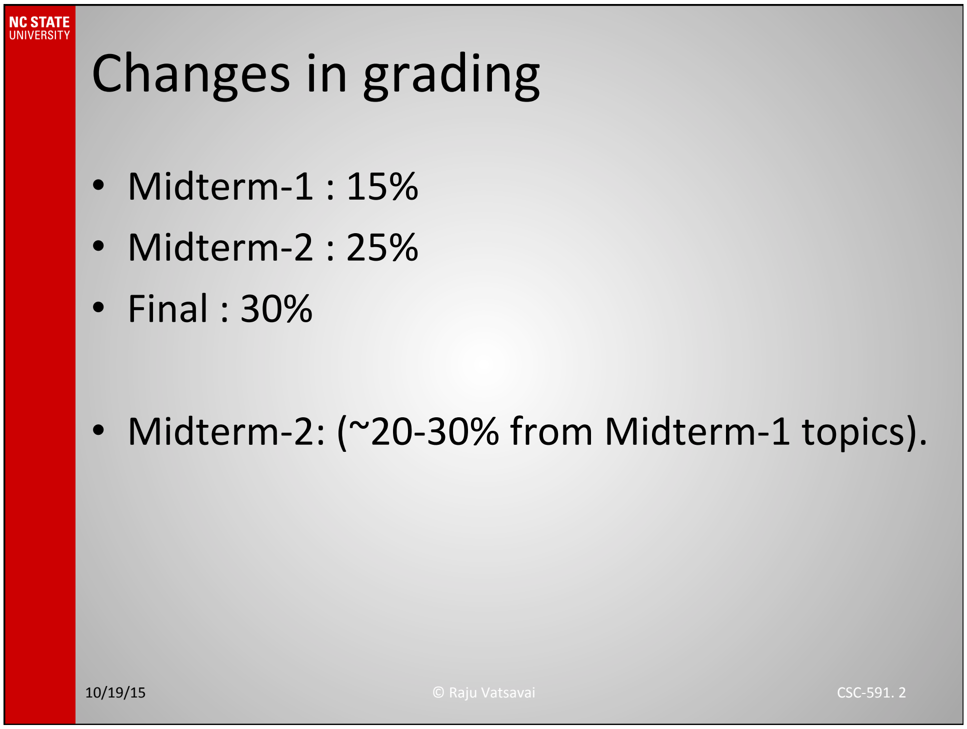


NC STATE UNIVERSITY

CSC-591: Foundations of Data Science
T/Th. 12:50-2:05pm. EBI-1005.

Ranga Raju Vatsavai
Chancellors Faculty Excellence Associate Professor in Geospatial Analytics
Department of Computer Science, North Carolina State University (NCSU)
Associate Director, Center for Geospatial Analytics, NCSU
&
Joint Faculty, Oak Ridge National Laboratory (ORNL)

W9: 10/13/15-10/15/15

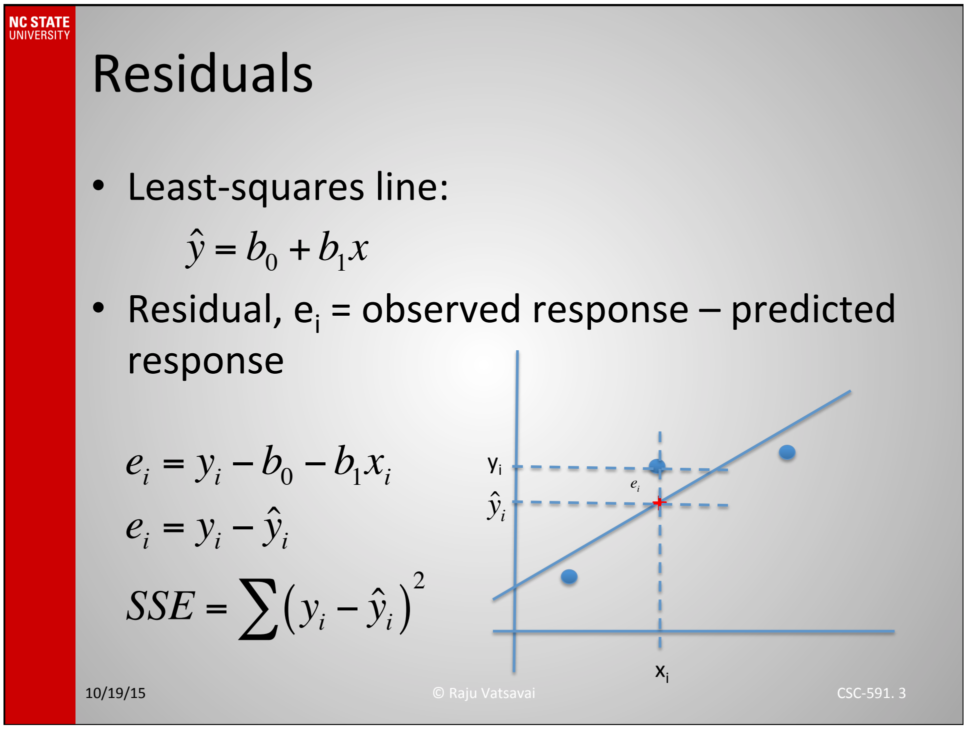


NC STATE UNIVERSITY

Changes in grading

- Midterm-1 : 15%
- Midterm-2 : 25%
- Final : 30%

- Midterm-2: (~20-30% from Midterm-1 topics).



NC STATE UNIVERSITY

Residuals

- Least-squares line:

$$\hat{y} = b_0 + b_1 x$$

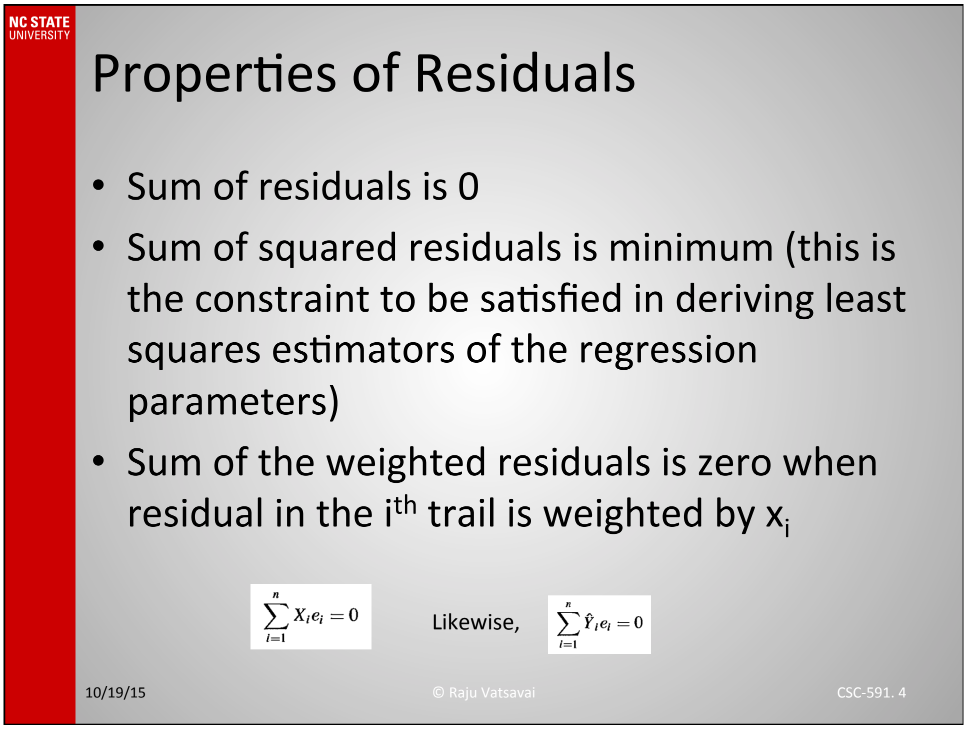
- Residual, e_i = observed response – predicted response

$$e_i = y_i - b_0 - b_1 x_i$$

$$e_i = y_i - \hat{y}_i$$

$$SSE = \sum (y_i - \hat{y}_i)^2$$

y_i
$\hat{y}_i$
e_i
x_i



NC STATE UNIVERSITY

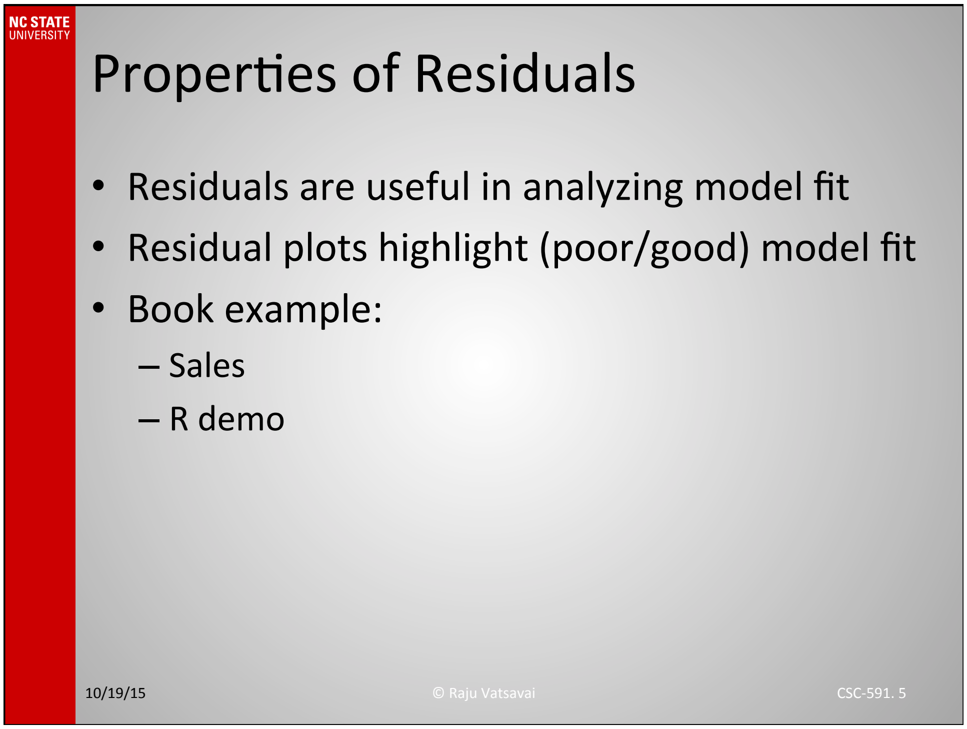
Properties of Residuals

- Sum of residuals is 0
- Sum of squared residuals is minimum (this is the constraint to be satisfied in deriving least squares estimators of the regression parameters)
- Sum of the weighted residuals is zero when residual in the i^{th} trail is weighted by x_i

$$\sum_{i=1}^{n} X_i e_i = 0$$

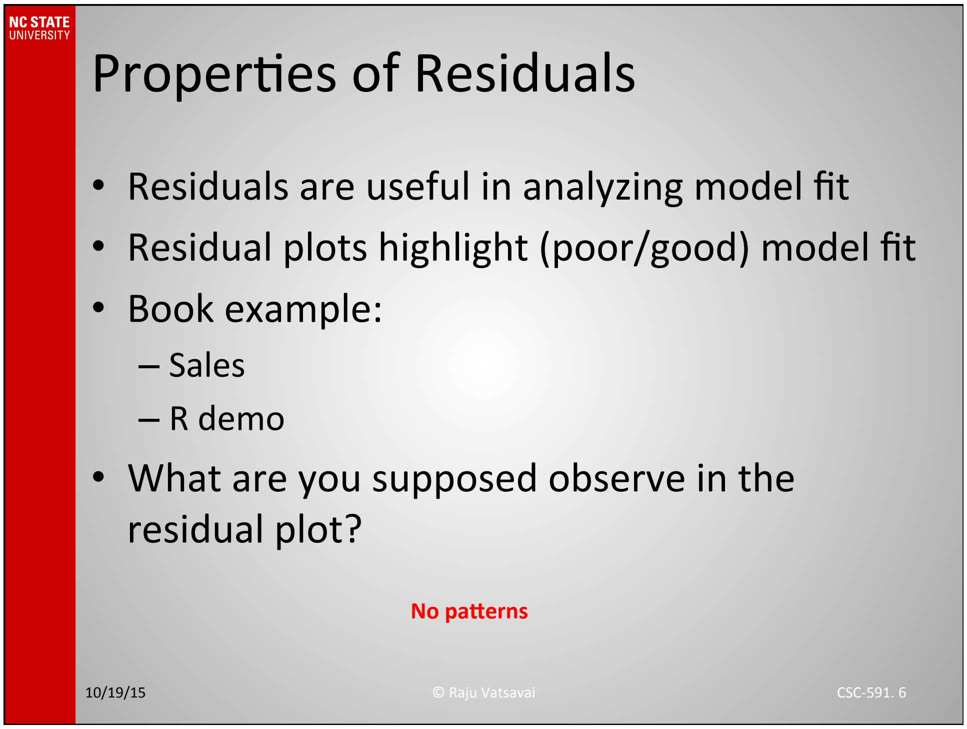
Likewise,

$$\sum_{i=1}^{n} \hat{Y}_i e_i = 0$$



Properties of Residuals

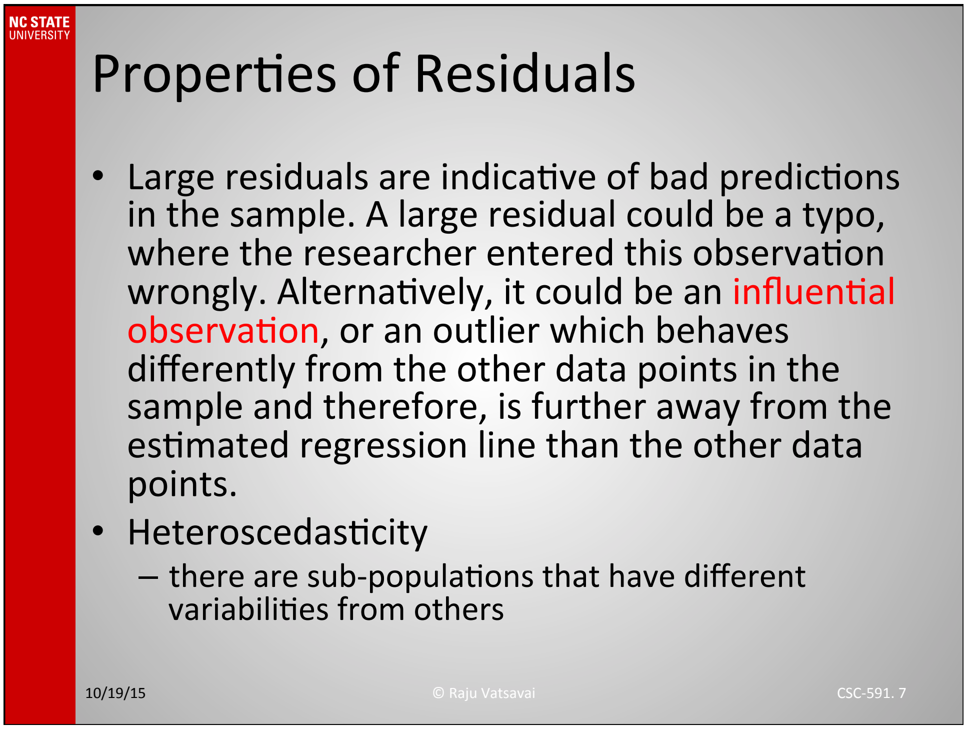
- Residuals are useful in analyzing model fit
- Residual plots highlight (poor/good) model fit
- Book example:
  – Sales
  – R demo



Properties of Residuals

- Residuals are useful in analyzing model fit
- Residual plots highlight (poor/good) model fit
- Book example:
  – Sales
  – R demo
- What are you supposed observe in the residual plot?

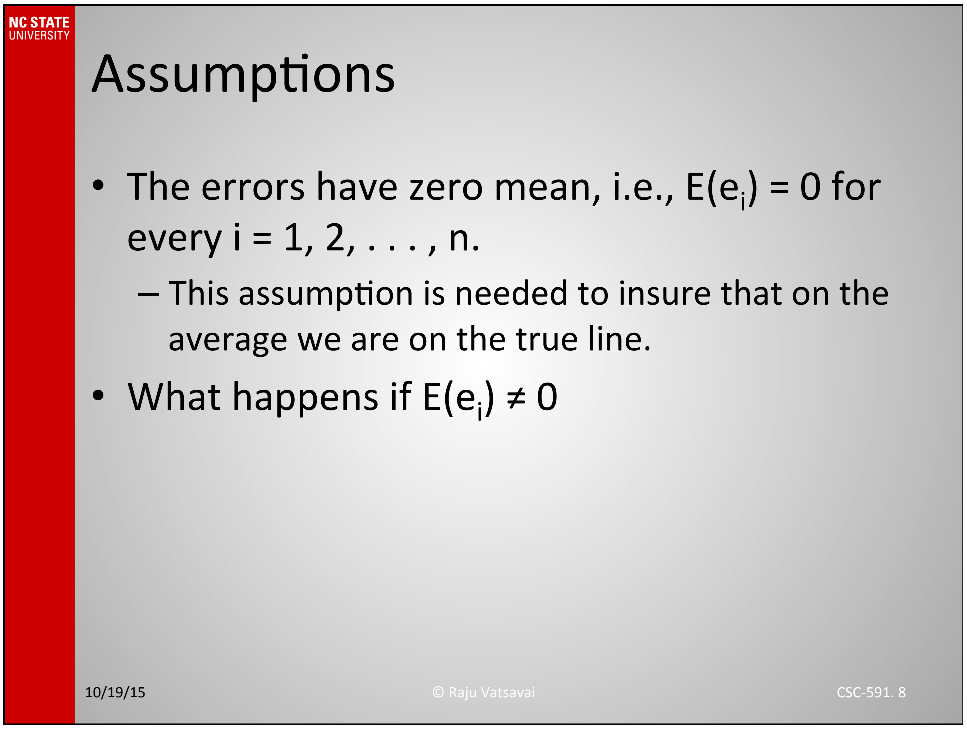
No patterns



NC STATE UNIVERSITY

Properties of Residuals

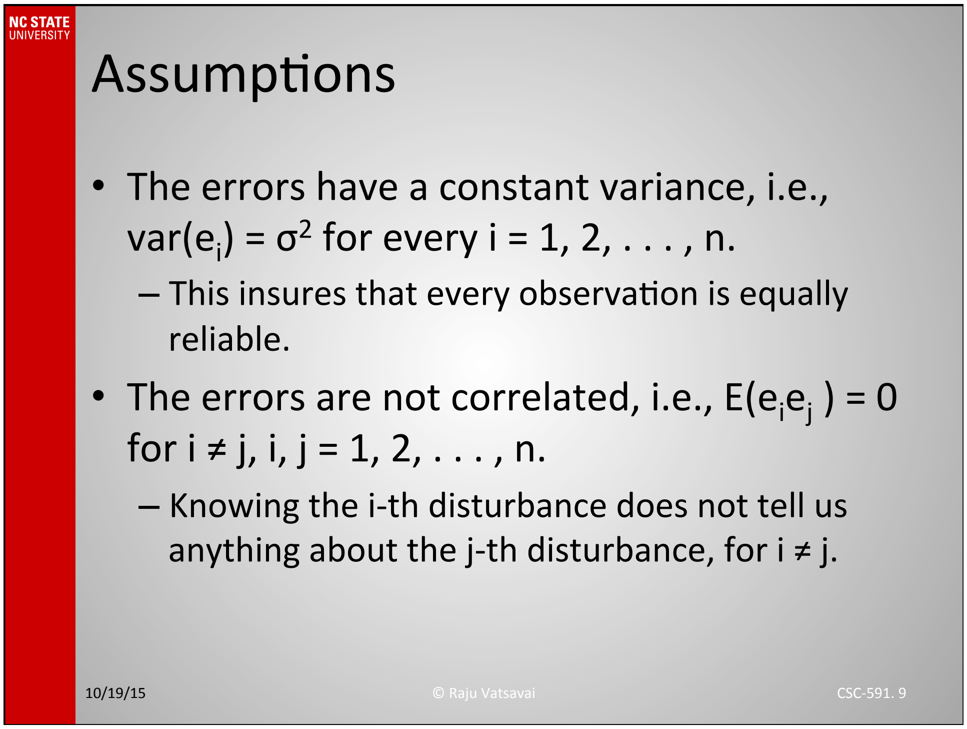
- Large residuals are indicative of bad predictions in the sample. A large residual could be a typo, where the researcher entered this observation wrongly. Alternatively, it could be an influential observation, or an outlier which behaves differently from the other data points in the sample and therefore, is further away from the estimated regression line than the other data points.
- Heteroscedasticity
  - there are sub-populations that have different variabilities from others



NC STATE UNIVERSITY

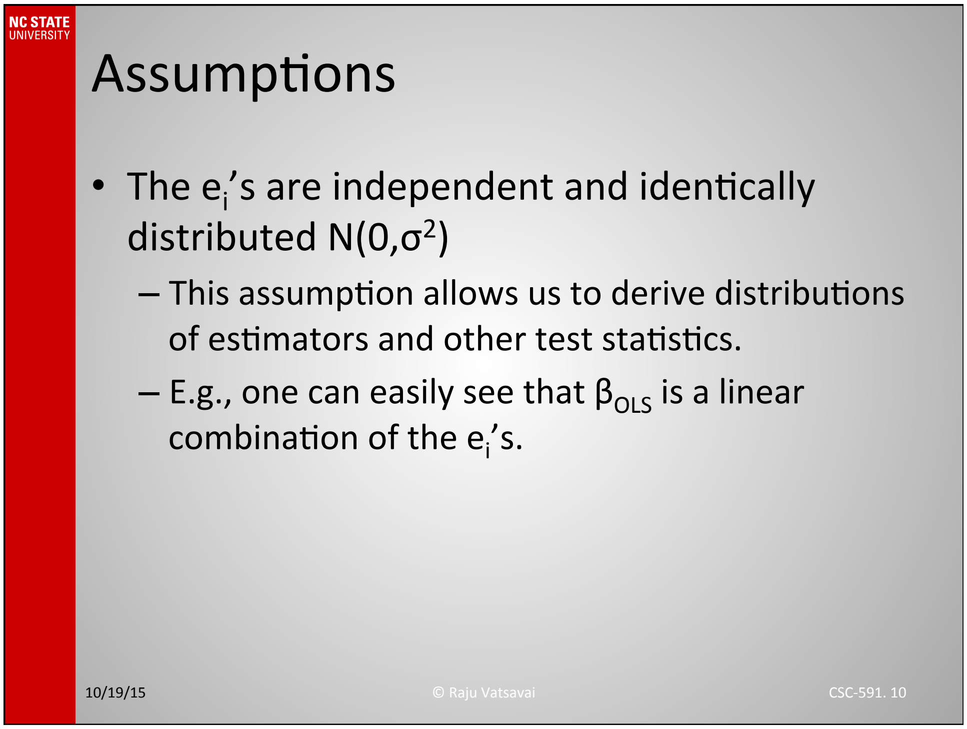
Assumptions

- The errors have zero mean, i.e., $E(e_i) = 0$ for every $i = 1, 2, . . . , n$.
  - This assumption is needed to insure that on the average we are on the true line.
- What happens if $E(e_i) \neq 0$



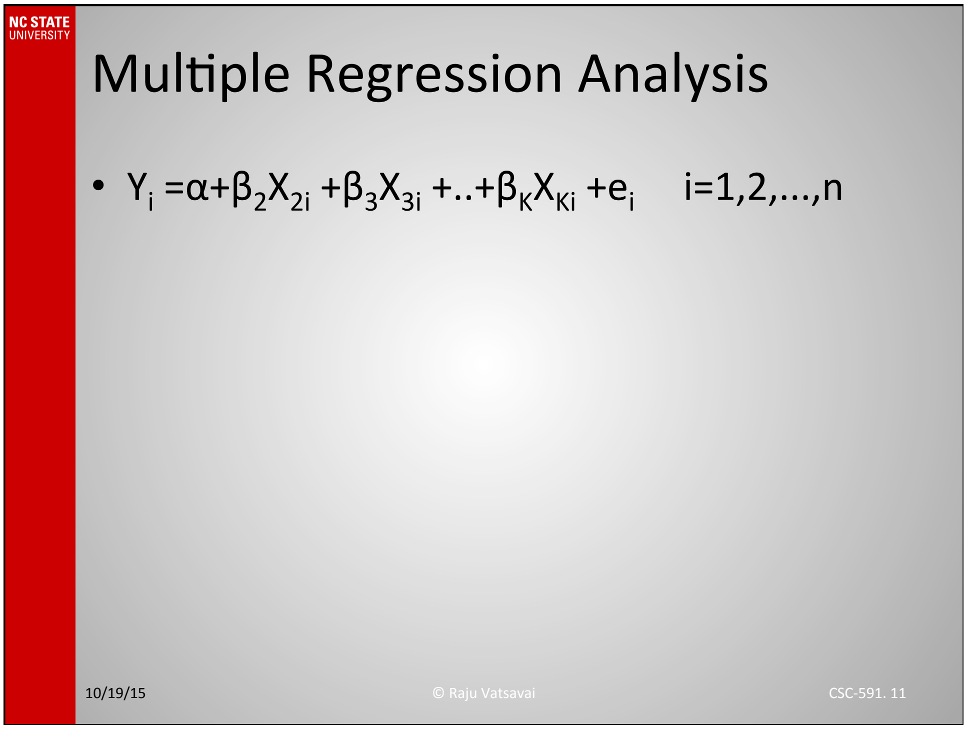
Assumptions

- The errors have a constant variance, i.e., var(e_i) = σ^2 for every i = 1, 2, . . . , n.
  - This insures that every observation is equally reliable.
- The errors are not correlated, i.e., E($e_i e_j$) = 0 for i ≠ j, i, j = 1, 2, . . . , n.
  - Knowing the i-th disturbance does not tell us anything about the j-th disturbance, for i ≠ j.



Assumptions

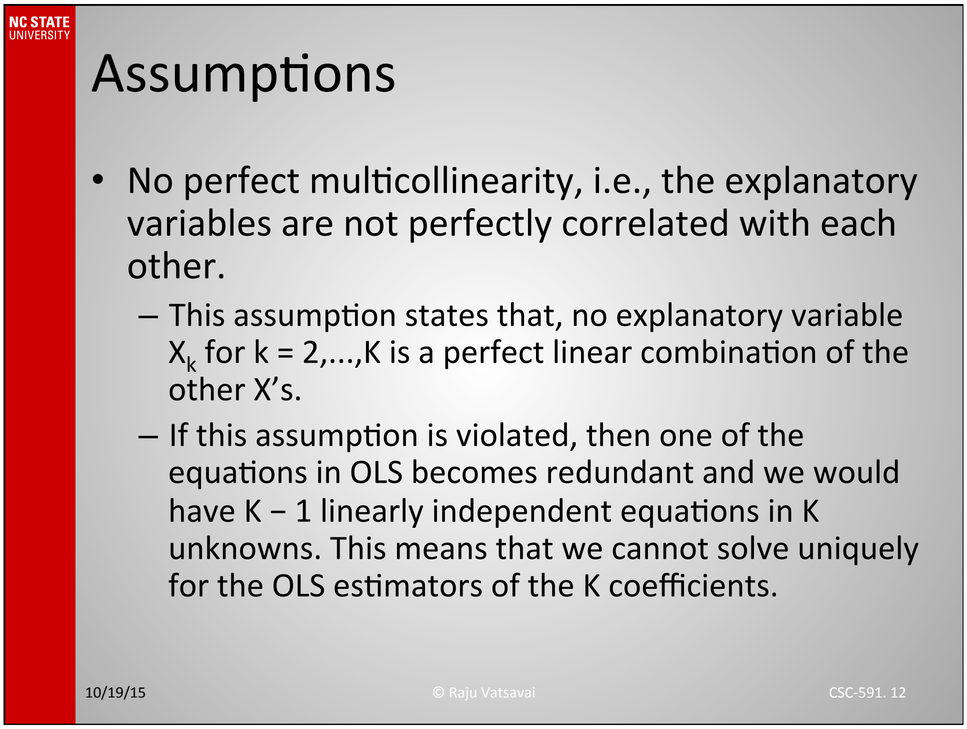
- The e_i's are independent and identically distributed N(0,σ^2)
  - This assumption allows us to derive distributions of estimators and other test statistics.
  - E.g., one can easily see that β_{OLS} is a linear combination of the e_i's.



NC STATE UNIVERSITY

Multiple Regression Analysis

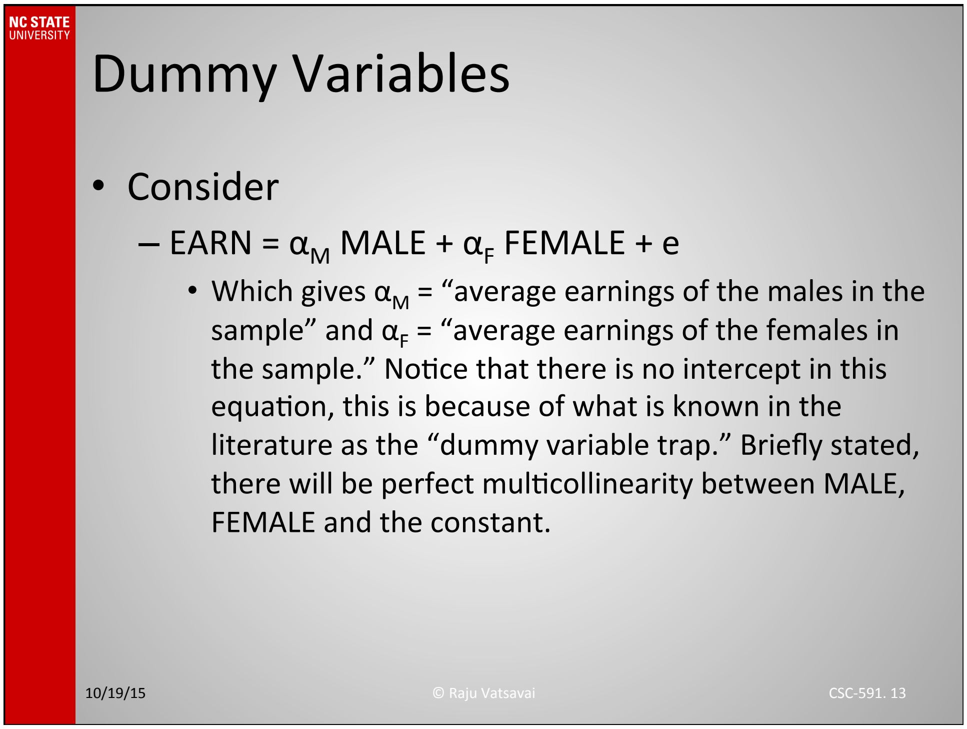
- $Y_i = \alpha + \beta_2 X_{2i} + \beta_3 X_{3i} + .. + \beta_K X_{Ki} + e_i \quad i=1,2,...,n$



NC STATE UNIVERSITY

Assumptions

- No perfect multicollinearity, i.e., the explanatory variables are not perfectly correlated with each other.
  - This assumption states that, no explanatory variable X_k for k = 2,...,K is a perfect linear combination of the other X's.
  - If this assumption is violated, then one of the equations in OLS becomes redundant and we would have K − 1 linearly independent equations in K unknowns. This means that we cannot solve uniquely for the OLS estimators of the K coefficients.



Dummy Variables

- Consider
  - EARN = α_M MALE + α_F FEMALE + e
    - Which gives α_M = “average earnings of the males in the sample” and α_F = “average earnings of the females in the sample.” Notice that there is no intercept in this equation, this is because of what is known in the literature as the “dummy variable trap.” Briefly stated, there will be perfect multicollinearity between MALE, FEMALE and the constant.