GENERAL ASSEMBLY

DATA SCIENCE

CLASS 6: LINEAR REGRESSION

I. BASIC FORM
II. ESTIMATING COEFFICIENTS
III. DETERMINING OVERALL MODEL RELEVANCE
IV. UNDERSTANDING MODEL COEFFICIENTS
V. GOTCHAS
VI. CATEGORICAL VARIABLES

I. BASIC FORM

BASIC FORM

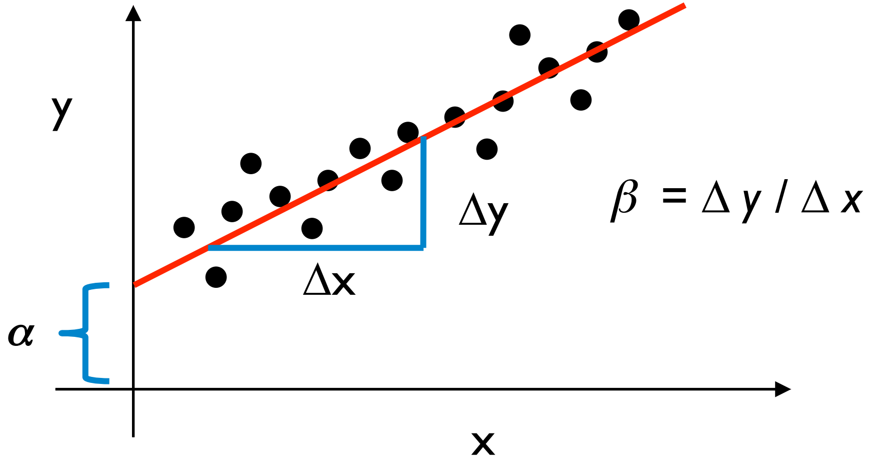
- A regression model, in its most basic sense, is a **functional relationship between your explanatory variables and your response.**

- A simple linear regression model captures a linear relationship (β) between your response (y) and one covariate (x), plus a constant (α) and random error (ε):

$$y = \alpha + \beta x + \varepsilon$$

HOW DO WE INTEPRET THE MODEL?

- Look at the chart to your right. It's an interpretation of a simple linear model when y and x are both continuous.
- Here's how to interpret $y = \alpha + \beta x + \varepsilon$:
  - ***y*** is your **response feature** (the feature we want to predict)
  - ***x*** is your **explanatory feature** (the feature we use to train the model)
  - $\boldsymbol{\alpha}$ is your **intercept** (where the regression model crosses the y-axis)
  - β is your **regression coefficient** (the model parameter)
  - ε is your **residual** (the model's error)



We can extend our model to several explanatory features, giving us a multiple linear regression model:

$$y = \alpha + \beta_1 x_1 + \ldots + \beta_n x_n + \varepsilon$$

WHY DO WE CARE ABOUT LINEAR REGRESSION?

- Linear regression, and its cousin, logistic regression, are primarily used to extract universal inference of the effects of explanatory features on the response feature.

- Linear regression lets you understand the effect of one feature, controlling for all the other identified explanatory features.
    - For a nominal (non-transformed) linear relationship, you can say: “controlling for all other factors, eating a healthy diet will increase your age by 2 years.”
    - For a log-transformed linear relationship, you can say, “controlling for all other factors, eating a healthy diet increases age by 20%”.

- Linear regression is usually not the most predictive technique, however, with non-generalizable machine learning techniques (like KNN), it is hard to make the sort of inference that I mentioned above!

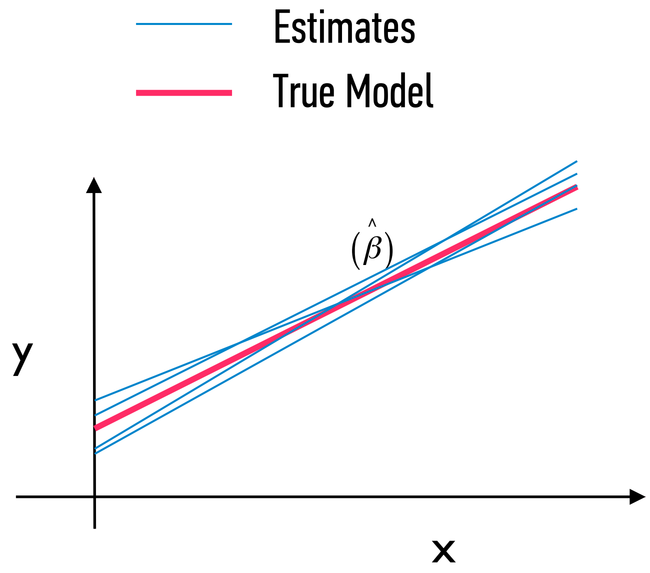
II. ESTIMATING COEFFICIENTS

ESTIMATING COEFFICIENTS

- We measure the relationship between a covariate and our response feature by estimating the covariate's regression coefficient, $(\hat{\beta})$.

- β-hat is different from β as b-hat is an **estimate** of the general model, based on the sample of data we are observing.

- As we are estimating, we need to quantify our confidence that the model's estimates are reflective of truth.

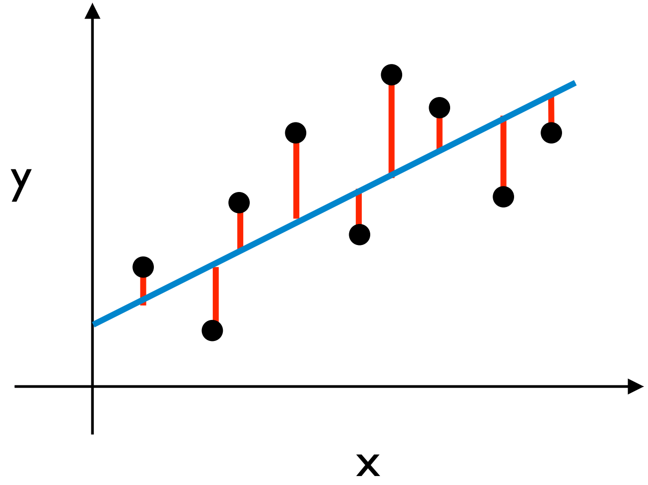


- We estimate the coefficients of a linear model by finding the values of β and α that minimize the **sum of the square of the model error** (residuals) in the sample data.

Model Prediction

$$SS_{residuals} = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

Observed Result



- We minimize the **sum of the square of the model error** via the following equation:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

- Now, let's all pull out our (sometimes painful memories of vector calculus.

- Say you have the following two vectors to the right: an **x vector** for your intercept and covariate, and a **y vector** for your response.

$$X = \begin{pmatrix} 1, & 3.385 \\ 1, & 0.48 \\ 1, & 1.35 \\ 1, & 465 \\ 1, & 36.33 \end{pmatrix} \qquad Y = \begin{pmatrix} 44.5 \\ 15.5 \\ 8.1 \\ 423 \\ 119.5 \end{pmatrix}$$

$$\hat{\beta} = \boxed{(X^T X)}^{-1} X^T Y$$

$$X = \begin{pmatrix} 1, & 3.385 \\ 1, & 0.48 \\ 1, & 1.35 \\ 1, & 465 \\ 1, & 36.33 \end{pmatrix} \qquad Y = \begin{pmatrix} 44.5 \\ 15.5 \\ 8.1 \\ 423 \\ 119.5 \end{pmatrix}$$

Transposing simply means flipping the columns and rows

The transposition multiplied and summed up in the final result.

$$X^T X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 3.385 & 0.48 & 1.35 & 465 & 36.33 \end{pmatrix} \begin{pmatrix} 1, & 3.385 \\ 1, & 0.48 \\ 1, & 1.35 \\ 1, & 465 \\ 1, & 36.33 \end{pmatrix} = \begin{pmatrix} 5 & 506.54 \\ 506.54 & 217558.38 \end{pmatrix}$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Only square matrices can be inverted

$$(X^T X)^{-1} = \begin{pmatrix} 5 & 506.54 \\ 506.54 & 217558.38 \end{pmatrix}^{-1} = \begin{pmatrix} 0.26 & -6.1 \cdot 10^{-4} \\ -6.1 \cdot 10^{-4} & 6.0 \cdot 10^{-6} \end{pmatrix}$$

Taking the inverse of a 2x2 matrix simply means swapping across diagonals, and dividing each value by the subtracted cross products of the original matrix (i.e. the determinant).

$$\frac{217558.38}{5 \times 217558.38 - 506.54 \times 506.54}$$

$$X = \begin{pmatrix} 1, & 3.385 \\ 1, & 0.48 \\ 1, & 1.35 \\ 1, & 465 \\ 1, & 36.33 \end{pmatrix} \qquad Y = \begin{pmatrix} 44.5 \\ 15.5 \\ 8.1 \\ 423 \\ 119.5 \end{pmatrix}$$

$$\hat{\beta} = (X^T X)^{-1} \boxed{X^T Y}$$

$$X^T Y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 3.385 & 0.48 & 1.35 & 465 & 36.33 \end{pmatrix} \begin{pmatrix} 44.5 \\ 15.5 \\ 8.1 \\ 423 \\ 119.5 \end{pmatrix} = \begin{pmatrix} 610.6 \\ 201205.4 \end{pmatrix}$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} 0.26 & -6.1 \cdot 10^{-4} \\ -6.1 \cdot 10^{-4} & 6.0 \cdot 10^{-6} \end{pmatrix} \begin{pmatrix} 610.6 \\ 201205.4 \end{pmatrix} = \begin{pmatrix} 37.201 \\ 0.838 \end{pmatrix}$$

III. DETERMINING OVERALL MODEL RELEVANCE

INTERPRETING THE OUPUT

- A data scientist interprets the quality of the model and its coefficients based on the following measures:

    - **Overall Model Relevance:** Root mean squared error and R^2

    - **Coefficient estimates:** Confidence intervals and p-values.

- Overall model relevance is primarily assessed through root mean squared error (RMSE) and r-squared.

- **Root mean squared error** in effect shows the average 'deviance' of your predicted values from actuals. It is calculated via:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2}$$

- RMSE can be used to **compare relevance** of different regression models, as lower MSEs mean lower actual model error in the data.

- R-squared is a measure of **goodness of fit.** It calculates the proportion of variance of the data explained by the model.
    - R-squared ranges from 0 (no variance explained) to 1 (all variance explained).
    - It's calculated by dividing the sum of the squares of the residuals in the regression model with the total sum of the squared difference between the data and its mean.

$$R^2 \equiv 1 - \frac{SS_{\rm res}}{SS_{\rm tot}}.$$

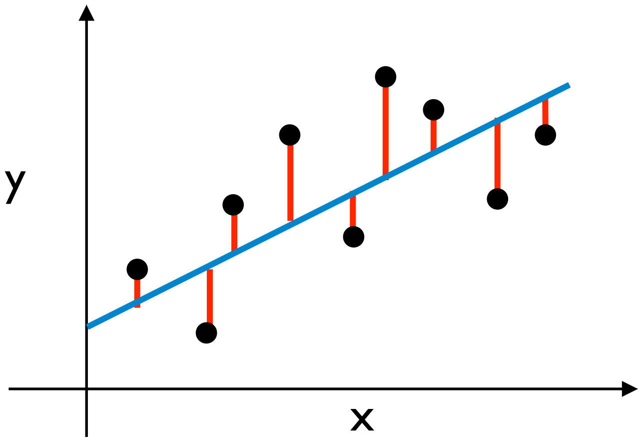
where

$$SS_{\rm res} = \sum_i (y_i - f_i)^2$$

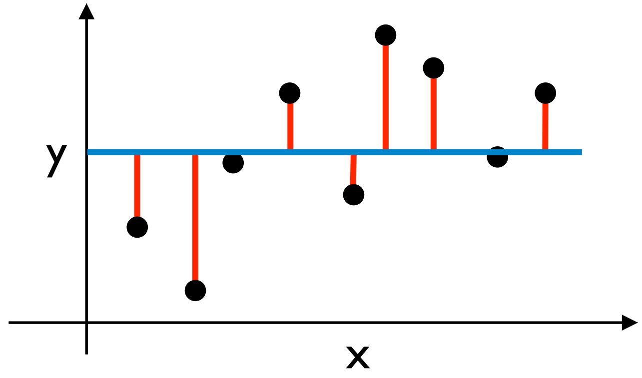
and

$$SS_{\rm tot} = \sum_i (y_i - \bar{y})^2,$$

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2$$



$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2,$$



- A ‘good’ r-squared value depends on the domain.
- However, as a benchmark, most models should not have a r-squared under 0.05—this typically shows that your model is not explaining the data well.
- But, watch out! R-squared has a few problems with it:
  - **Goodness of fit does not equal accuracy!**
  - By definition, adding more covariates to the model improves r-squared, even though they may do nothing to improve model accuracy or quality.
    - Adjusted R^2, exists to compensate this, as it takes into account the model complexity.

$$Adjusted\ \ R^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

$p = Input\ Variables \quad n = Samples$

As p increases:
- Denominator decreases
- Fraction increases
- Adjusted R^2 decreases

IV. UNDERSTANDING MODEL COEFFICIENTS

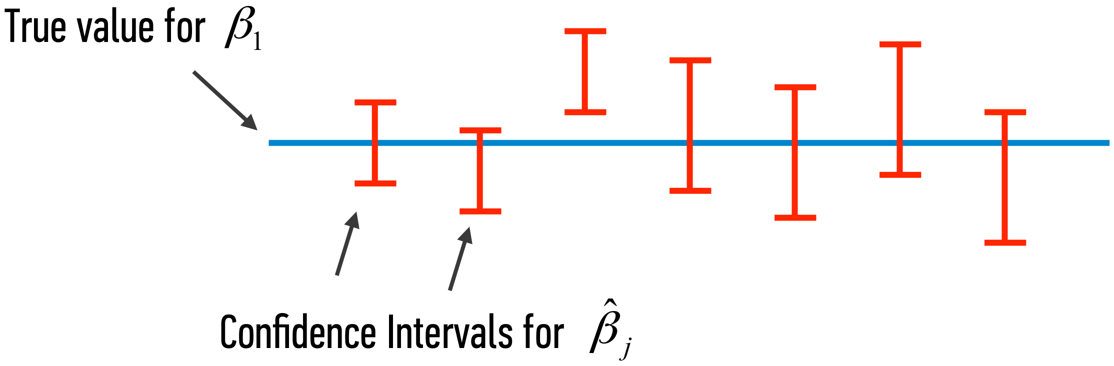
- Recall our model equation for multiple linear regression:
    - $y = \alpha + \beta_I x_I + \ldots + \beta_n x_n + \varepsilon$
- Also recall the meaning of β:
    - $\beta = \Delta y / \Delta x$
- How do we know that a covariant β is meaningful in the model?
    - We look at the p-value associated with the coefficient t-value.
- What is a p-value?
    - A p-value is the probability of observing the outcome (e.g., the coefficient estimate) if the null hypothesis for linear regression coefficients is true ($p < 0.05$ is typically considered significant).

- What is the null hypothesis for linear regression coefficients?
  - That there is no relationship between X and Y.
- In al cases, the p-value is greater than 0.05 when 0 falls within the 95% confidence interval of the regression coefficient.

$$H_0: \beta_j = 0$$

$$H_a: \beta_j \neq 0$$

- **How do we interpret a 95% confidence interval?**
  - 95% of the time, the true coefficients will be within the interval range.



V. GOTCHAS

COMMON PROBLEMS

- Linear modeling is a parametric technique, meaning that it relies on specific assumptions about the underlying data:
    - Independence, linearity, and additivity of the relationship between explanatory and response features
    - Homoscedasticity of the errors
    - Normality of the Error Distribution
    - Statistical independence of the errors

- These often don’t at the onset hold true!

Source: http://people.duke.edu/~rnau/testing.htm

- Collinearity exists whenever there is a correlation between two or more explanatory features.
  - When this occurs, you can no longer vary your covariates independently to extract their effect on the response feature.
  - **Practically, collinearity reduces confidence in your coefficient estimates.**

- You can identify collinear variables via a correlation matrix.
  - The most popular way to measure correlation is via the Pearson product-moment coefficient (a.k.a., correlation coefficient).

$$r = \frac{Cov(x, y)}{s_y s_x}$$

Covariance of x and y

Sample standard deviation

- Here's the full equation:

Observed x Average x

Covariance of x and y

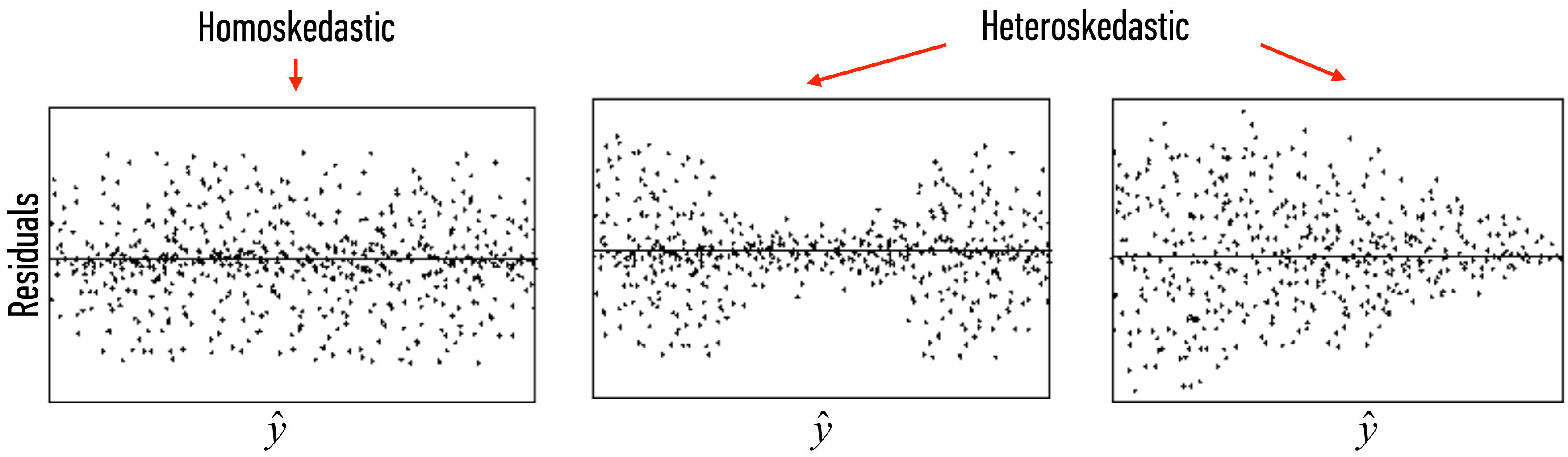
$$r = \frac{Cov(x, y)}{s_y s_x} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})}}$$

Sample standard deviation

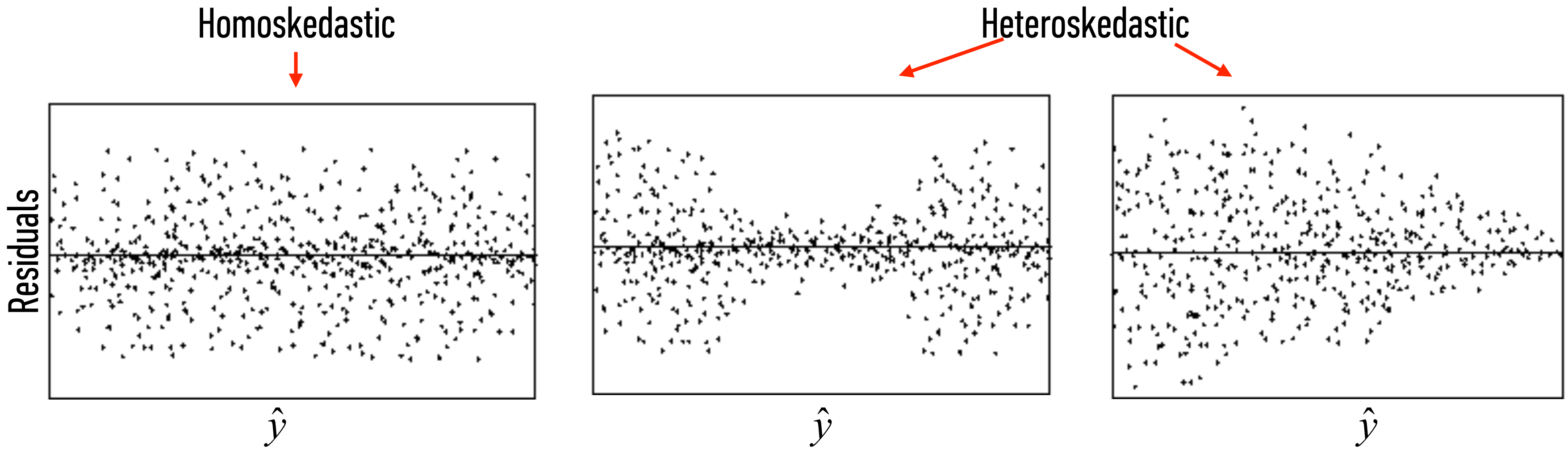
- Once you've identified perfectly correlated covariates in your correlation matrix, eliminate all but one of the covariates, or combine them into an interaction term.

PROBLEM #2: HETEROSKEDASTICITY

- Heteroskedasticity occurs when there is non-constant variance in the error terms (residuals) of your model with respect to the magnitude of your response feature.
  - (literally: hetero=different, skedasis=dispersion).



- How do you identify heteroskedasticity?
  - Plot the residuals against the predicted response variable.



- How does heteroskedasticity affect my model?
  - It will distort and therefore decrease confidence in coefficient and prediction estimates.

- Why does heteroskedasticity reduce confidence in the model?
  - Because standard errors, confidence intervals, and hypothesis tests all rely on constant error variance.

- You can correct for heteroskedasticity in two ways:

1. **Log-transform the response variable.**
    - Coefficients now correspond to percentage change in response variable, rather than unit change.

2. **Use Weighted Least Squares, i.e. a 'robust' regression.**
    - Weights themselves are an input to the model. This typically means observations with greater deviation contribute less to estimates associated with the coefficients.

VI. CATEGORICAL FEATURES

- Linear regression (like most algorithms we will learn in this class) can only accept numerical input.
- We incorporate categorical features into the algorithm by creating k-1 binary ("dummy") features.

Major (k=4)
Computer Science
Engineering
Business
Literature
Business
Engineering

Engineering	Business	Literature
0	0	0
1	0	0
0	1	0
0	0	1
0	1	0
1	0	0

Computer Science is the reference

- Why do we have k-1 and not k dummy features?
    - Because k features would create collinearity!
    - The absence of K (i.e. all K's at zero) represents the 'reference' feature that the regression estimates without K.
        - So, choose your omitted K wisely!

- If the categorical data has a clear rank or order, you can represent the data with integers.
    - For example, [strongly disagree, disagree, neutral, agree, strongly agree] can be represented as [1, 2, 3, 4, 5].

IN-CLASS EXERCISE: LINEAR REGRESSION

1. Aggregate our dataset's Batting table data on the yearly level before 2005.
2. Run an OLS regression where hits is your explanatory feature and runs scored per year is your response.
   - Interpret its results, calculate R-squared and RMSE, and examine the residuals for heteroskedasticity.
3. Run an OLS on stolen bases and runs scored per year.
   - Compare its coefficients, R-squared, and RMSE to the previous example.
4. Create dummy features representing time and re-run the regression.
5. Create a multivariate regression including hits, stolen bases, and our dummy variables and compare its output to #2-4.
6. Compare the predictive accuracy of the multivariate regression and the regression from step #2 using new data.

QUESTIONS?