Berkeley Engineering | BerkeleyHaas

# PROFESSIONAL CERTIFICATE IN MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

## Office Hour with Mani K

**email:** mani.k@berkeley.edu

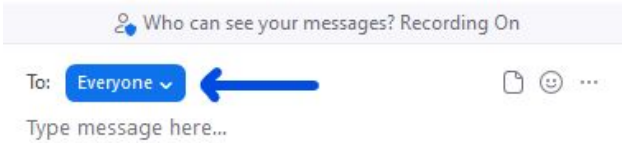
Oct 29, 2024 at 4pm PST

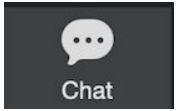
# GUIDELINES ON ZOOM

## Audio & Video

- Keep your microphone muted while presenters are speaking.
- Keep your camera on, preferably.


## Chat or use Q&A



# Agenda

- Linear Regression

  https://drive.google.com/file/d/11f_K7I-Jg4f5mF_kfmPQtSvhJpefAzHo/view?usp=sharing

  https://drive.google.com/file/d/1wcs_EBCcGBCVjh1XaZAeOCS345aat83O/view?usp=sharing

- Transformation

  https://drive.google.com/file/d/1Dk8yYKpKc2heDHy8p6XKnCzkVRT4hi-B/view?usp=sharing

  https://drive.google.com/file/d/1CL1bwIAgS_g1xB3uroiTxfABLV1ztq4-/view?usp=sharing

- Multiple Regression - Covariance

  https://drive.google.com/file/d/1N-r0TJ_QuqgzLT6A7Wt-ZtI_PyZU-z8z/view?usp=sharing

  https://drive.google.com/file/d/1UiUS85AVb4TyGz-7lnzwWoi5l9s8FcJ3/view?usp=sharing

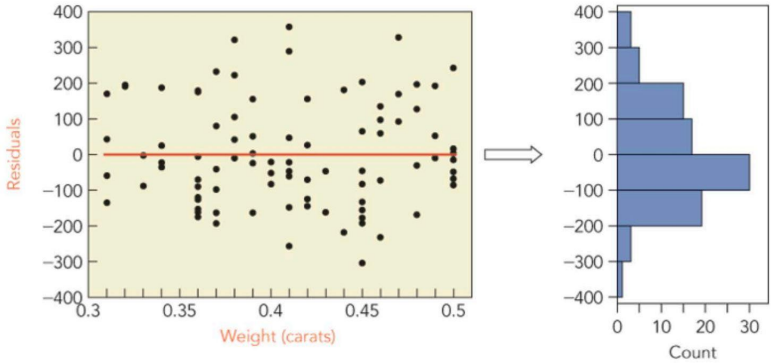
  https://drive.google.com/file/d/1i-r5-9X3ltVpwoZWjzmIpsJ-07W5dNJ6/view?usp=sharing

# Regression

- It’s about fitting a simple line with 2 variables
  - Independent/Explanatory variable - x axis
  - Dependent/Predicted/Response/Actuals variable - y axis
- If the association is somewhat linear then a line can be fit
  - $\hat{y} = b0 + b1*\dot{x}$ b0:Intercept b1: Slope $\hat{y}$: predicted response
- Residuals = $y - \hat{y}$ (this can also be negative, difference between actual & predicted)
- The best fit line is when squared of the residuals is kept to a minimum - Least Squared Regression
- If the residuals are higher and a lot of variance - the response may also depend on another variable we are not accounting for - lurking variable
- Perform multiple linear regression by adding more explanatory variables

# Regression

- Residuals capture variations
- Best way to identify good fit is using residual scatter plots - plot between residuals and the independent variable (x - axis)
- Divide the data into different scatter plots and check if the scatter plots looks similar
- Mean of residuals will usually be zero

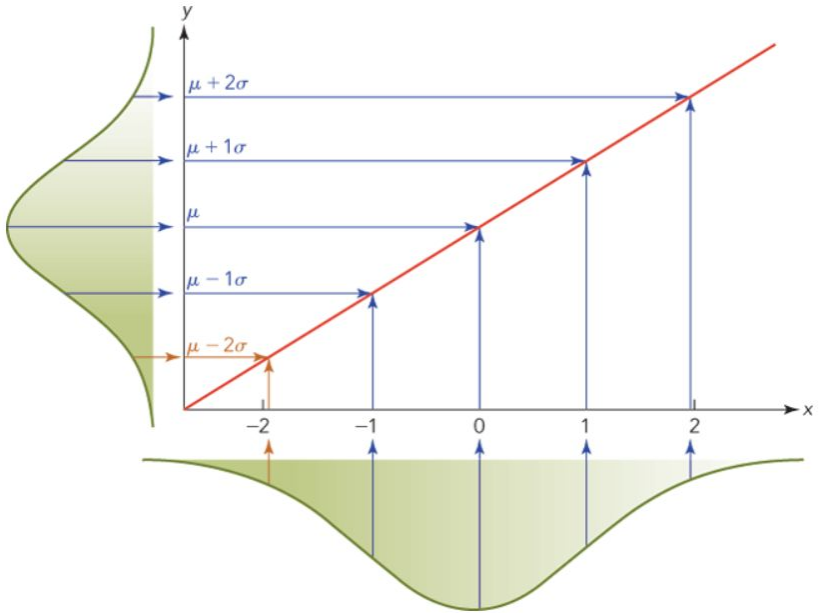


# Regression

- Check if residuals are normally distributed using a histogram of the residuals
- Check if the residuals are homoscedastic (like flat pipe) and not heteroscedastic using  a residual plot
- Check the quantile-quantile (QQ) plot if the residuals follows along the normal distribution line

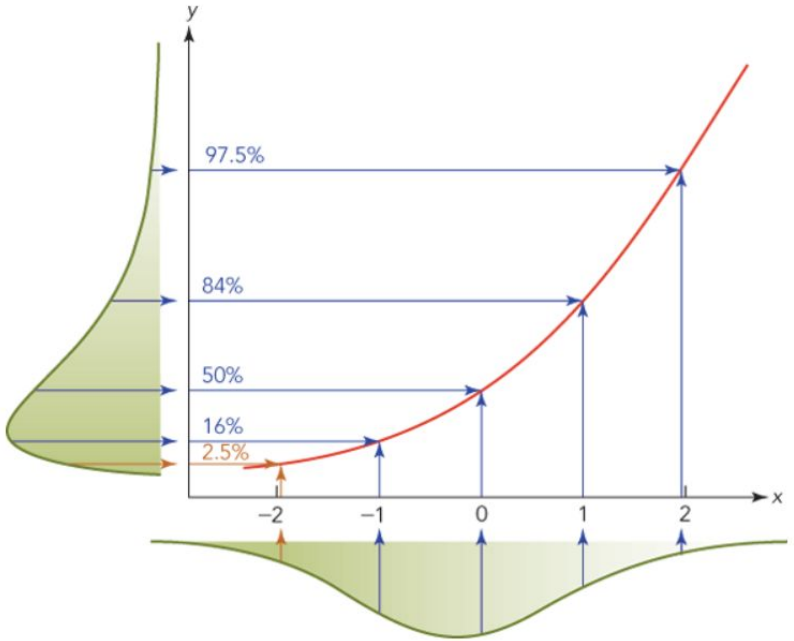
# Regression

- Normal Quantile Plot
  - x - axis (normal gaussian distribution)
  - y - axis ( if the distribution is also normal with a mean - μ and SD - $\sigma$ )
- Get a straight line by plotting these points



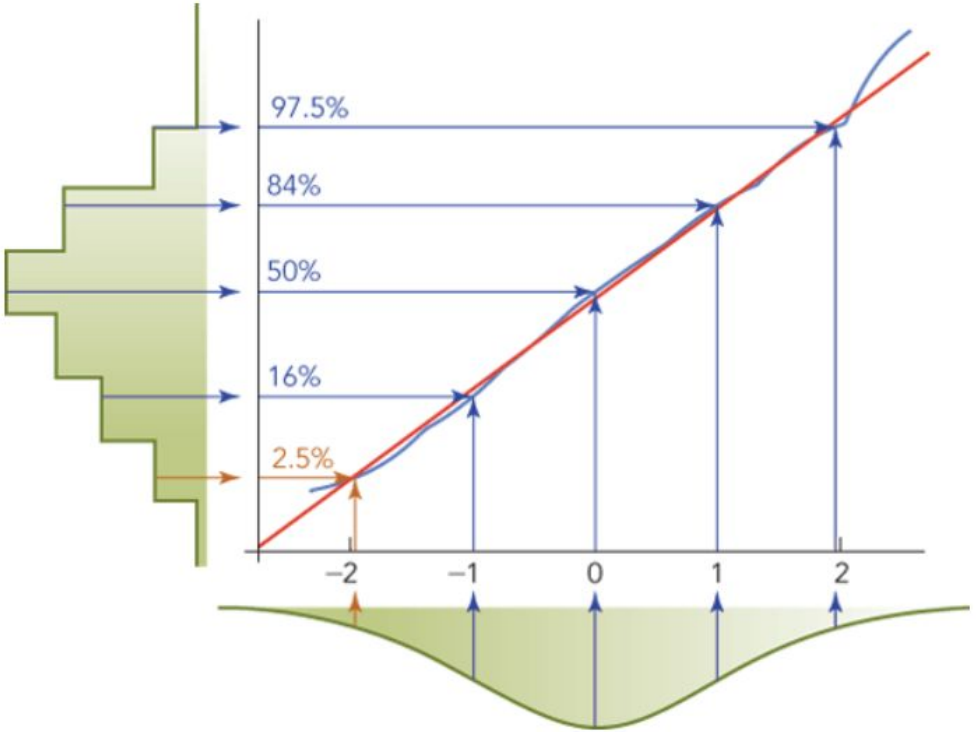
# Regression

- For a skewed distribution the quantile plot will look like this



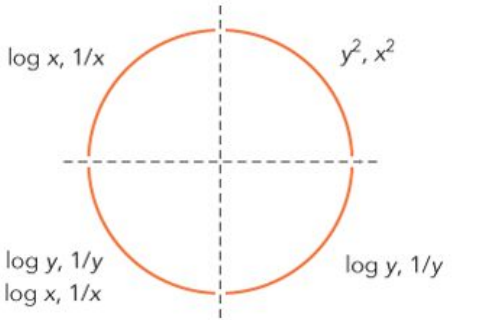
# Regression

- Residuals are part of a sample set ( not a population distribution)
- The histogram of the residuals is used to plot the quantile plot
- Showed in relation to the normal quantile plot



# Regression - Transformation

- Picking a transformation is a iterative process
- TUKEY AND MOSTELLER'S BULGING RULE - can be a good starting point

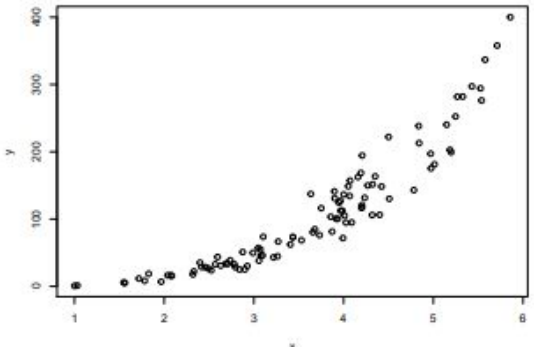



- The shape of the bulge (scatterplot) indicates the transformation of Y and/or X to straighten the relationship between them.

- For example if we have a scatter plot like this,

  -we are in the log y, 1/y quadrant

  - taking log y or 1/y is a good starting point



# Multiple Regression - Covariance

- Collinearity
  - If 2 or more explanatory variable are linearly correlated
  - Problems with p-values and statistical significance of the independent variables (this is our assumption that they are indeed independent!)
  - Small variations can cause huge changes to the output

- VIF - Variance Inflation Factor
  - Amount of unique variation in each explanatory variable and measures the effect of collinearity
  - VIF = 1/ (1-r2)
  - If 2 explanatory variables are totally uncorrelated then the r2 = 0 and so VIF = 1
  - Higher VIF denotes collinearity (general rule of thumb something > 5)

# Multiple Regression - Covariance

- For example

$$\begin{aligned} \textit{Estimated}\ \ \text{Sony}\ \ \text{Change} &= -0.4 + 0.9\ \ \text{Market}\ \ \%\ \ \text{Change} + 0.3\ \ \text{Dow}\ \ \%\ \ \text{Change} \\ &\quad + 0.7\ \ \text{Small-Big} - 0.1\ \ \text{High-Low} \end{aligned}$$

- To check collinearity we need to regress each explanatory variable to rest of the explanatory variables -
  - Step 1:
    - regress Market % Change over Dow, Small-Big & High-Low
    - regress Dow over Market % change, Small-Big & High-Low
    - regress Small-big over Market % change, Dow & High-Low
    - regress High-Low over Market % change, Dow & Small-Big
  - Step 2: Calculate VIF from the R2 from each of the regression
  - Step 3: Check for High Values of VIF
  - Step 4:
    - How to avoid collinearity ?
    - Remove one variable at a time and check
    - Combine several variables into one (average or difference)
    - This is where having domain expertise also helps!!

# Multiple Regression - Covariance

- Summary of VIF for Sony stock prediction

**Table 24.4 Summary of multiple regression showing variance inflation factors.**

| Term | Estimate | Std Error | *t*-Statistic | *p*-Value | VIF |
|---|---|---|---|---|---|
| Intercept | − 0.4340 | 0.5822 | − 0.75 | 0.4567 | — |
| Market % Change | 0.9040 | 0.3994 | 2.26 | 0.0245 | 9.83 |
| Dow % Change | 0.3191 | 0.4057 | 0.79 | 0.4322 | 9.02 |
| Small-Big | 0.6983 | 0.2065 | 3.38 | 0.0008 | 1.56 |
| High-Low | − 0.1450 | 0.1982 | − 0.73 | 0.4653 | 1.25 |