DATA1001 Topic 5 Flashcards

Design of Experiments

Abyan Majid

Question	My answer
What are the 6 key steps in the linear regression framework?	1. Produce a scatter plot 2. Derive a regression line 3. Compute correlation coefficient r 4. Produce residual plot 5. Check assumptions 6. Make predictions
What type of data are regressions usually suitable for? And what is the most appropriate graphical summary for this type of data?	Regressions are suitable for bivariate data; we are specifically interested in whether some independent variable X can be used to predict some dependent variable Y. The most appropriate graphical summary for bivariate data is a scatter plot.
What does a linear association tell you? And then, what can you conclude from a (1) positive linear correlation, (2) negative linear correlation?	A linear association tells you how tightly clustered around a line the points in the scatter plot are. (1) A positive linear correlation tells you that y tends grows linearly with x (2) A negative linear correlation tells you that y does NOT tend to grow linearly with x
Identify numerical summaries (or pairs of numerical summaries) that can summarize a scatter plot?	(1) $(\bar{x}, \sigma_x)$ (2) $(\bar{y}, \sigma_y)$ (3) Correlation coefficient r
(1) What represents the center of a scatter plot? (2) What measures the horizontal spread of a scatter plot? (3) What measures the vertical spread of a scatter plot?	(1) Point of averages $(\bar{x}, \bar{y})$ (2) σ_x (3) σ_y
(1) What is a correlation coefficient r conceptually and mathematically? (2) Identify the range of numbers r can take, and then interpret what different values suggest!	(1) the correlation coefficient r measures the clustering of data around a given line. Mathematically, it is the mean product of the variables in standard units. (2) r takes values from -1 to 1. Common descriptors of r is as follows: - Very strong negative correlation (-0.8 to -1) - Strong negative correlation (-0.6 to -0.799) - Moderate negative correlation (-0.4 to -0.599) - Weak negative correlation (-0.2 to -0.399) - Very weak negative correlation (-0.001 to -0.199)

(3) Identify one way for computing the correlation coefficient given variables x and y (4) Recite the mathematical formula for r_{pop} and r_{sample}	- No correlation (0) - Very weak positive correlation (0.001 to 0.199) - Weak positive correlation (0.2 to 0.399) - Moderate positive correlation (0.4 to 0.599) - Strong positive correlation (0.6 to 0.799) - Very strong positive correlation (0.8 to 1) (3) `cor(x, y)` in R (4) Recall that r is the mean of the product of the variables in standard units! $r_{pop} = \frac{1}{n}\sum_{i=1}^{n}(\frac{y-\bar{y}}{SD_y})(\frac{x-\bar{x}}{SD_x})$ $r_{sample} = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{y-\bar{y}}{SD_y})(\frac{x-\bar{x}}{SD_x})$
If you interchange the variables, does the correlation coefficient stay the same?	Yes
Is correlation coefficient shift and scale invariant?	Yes
What does the SD line connect $(\bar{x}, \bar{y})$ to? Is it optimal?	$(\bar{x} + SD_x, \bar{y} + SD_y)$ It is less optimal than the regression line.
Identify 5 numerical summaries to fully describe a scatter plot?	(1) $\bar{x}$ (2) $\bar{y}$ (3) SD_x (4) SD_y (5) r
What does the regression line connect $(\bar{x}, \bar{y})$ to?	$(\bar{x} + SD_x, \bar{y} + rSD_y)$
Recite the formula for getting slope of a SD line and linear regression in terms of standard deviations!	Slope (SD line) = $\frac{SD_y}{SD_x}$ for $r \geq 0$, $\frac{-SD_y}{SD_x}$ for $r < 0$, Slope (Linear regression) = $r\frac{SD_y}{SD_x}$
Recite the equation that gives you the intercept?	By $\bar{y} = b\bar{x} + a$, intercept a is given by $a = \bar{y} - b\bar{x}$
What does the graph of averages do?	It shows you all y averages $\bar{y}$ for each unique x
What is a residual? What does the residual plot do? How do you tell from the residual plot if the data is linear?	A residual refers to the gap/difference between a point and the value of y the model predicts. We can conclude that the data is linear if the points in the residual plot are random/do not show any particular pattern.
$r = 0.8$ means 80% of the points are tightly clustered around the line	False

True or False?	
$r = 0.8$ means the points are twice as tightly clustered as $r = 0.4$ True or False?	False
What does the RMS error represent? Recite the quick formula for population RMS error in terms of y-th standard deviation!	Mean residual $RMS = \sqrt{mean\ of\ (gaps)^2}$
What is the RMS error when there is a perfect correlation (ie. $r = \pm 1$)?	0
What is the RMS error when there is no correlation (ie. $r = 0$)?	SD_y
(1) How does vertical strips tell if the data is homoscedastic or heteroscedastic? (2) Can the RMS error be used as a measure of spread for homoscedastic, heteroscedastic, or both types of data? (3) Can normal approximation be used within vertical strips given homoscedastic, heteroscedastic, or both types of data?	(1) Data is homoscedastic if the gaps from the points and the line are approximately equal. Otherwise, it is heteroscedastic. (2) RMS error can only be used as a measure of spread for homoscedastic data. (3) Normal approximation can only be used as a measure of spread for homoscedastic data
Identify 3 ways in which you can make a prediction with data! Then identify which method is most optimal!	(1) Get average of y for all x (2) Get average of y for all x within vertical strips (3) Use the regression line