Data 8 Spring 2024

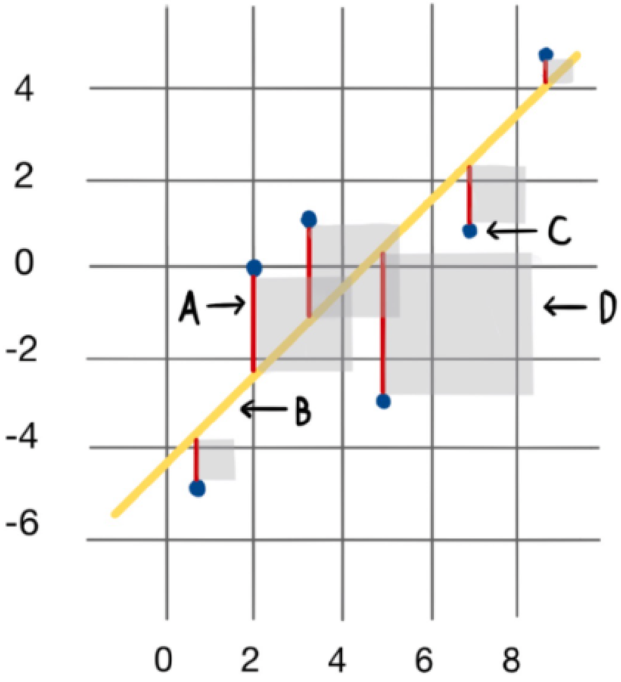
RMSE, Prediction Intervals, and k-Nearest Neighbors

Project 3

April 2024

1. RMSE

Identify each of the elements in the plot below as one of the following: regression line, error, squared error, data point.

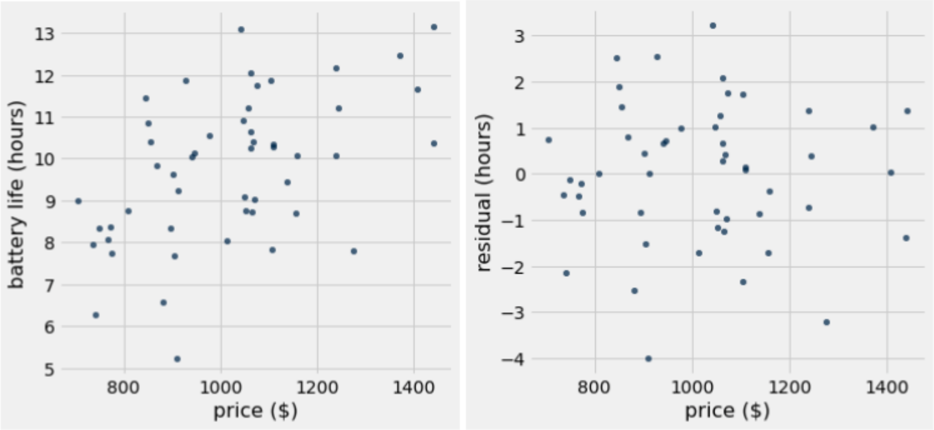


2. Prediction Intervals

Jessica is looking to buy a new laptop for her birthday. She has a table `laptops` with information on different laptops with two columns:

- **price** (float): the price of the laptop in dollars
- **battery life** (float): the battery life of the laptop in hours

a. Inspect the following scatter plot and residual plot of Jessica's data. Would using linear regression be reasonable for this dataset?



b. Jessica wants to use a regression line to predict the battery life of a laptop given the price. Define the `fitted_value` function below which takes in the following arguments:

- **table** (Table): a table with the data points used to generate the regression line
- **x** (string): the column name for the x variable
- **y** (string): the column name for the y variable
- **given_x** (float): the x value we want to make a prediction at

The function should return a float by using a regression line to predict a y-value for the given x-value. Assume the `slope(tbl, x, y)` and `intercept(tbl, x, y)` functions are defined as in lecture.

```
def fitted_value(table, x, y, given_x):

    m = ________________________________

    b = ________________________________

    return ________________________________
```

c. Assume the average price of a laptop in Jessica's dataset is $1000. Jessica generates a 90% confidence interval for the battery life of laptops priced at $1100 and $700.

i. Which one of these two intervals would be wider? Why?

ii. Does the answer to the previous part change if we used a different confidence level? Why or why not?

d. Jessica believes that a laptop with a price of $1300 should have a battery life of 14 hours. Complete the following code to test his hypothesis with a 4% p-value cutoff. Assume Jessica has properly simulated 1000 predicted battery lives for a laptop with price $1300 and stored them in the array `predictions`

```
left = ________________________________

right = ________________________________

if ________________________________:

    print("Fail to reject the null hypothesis")

else:

    print("Reject the null hypothesis")
```

3. Prediction Intervals T/F

For this question, assume we are using the same dataset from Question 2 and the data is well suited for linear regression. You find the correlation between price and battery life to be 0.8.

a. **True or False**: A 90% prediction interval for a laptop with price $0 will have nearly the same lower and upper bounds as a 90% confidence interval for the true intercept of the regression line in original units.

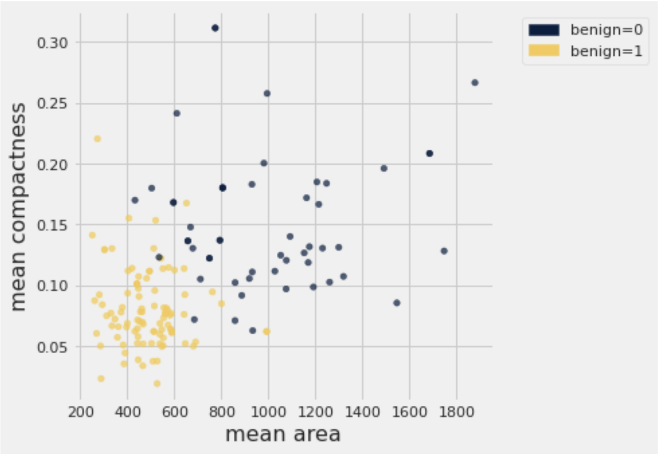
b. **True or False**: A 90% prediction interval for a laptop with price 1 in standard units will have nearly the same lower and upper bounds as a 90% confidence interval for the true correlation of the regression line.

c. **True or False**: If we constructed 100 90% confidence intervals and 100 95% confidence intervals for the battery life of a laptop with price $950, less of the 95% CIs will contain the true battery life of a laptop with price $950.

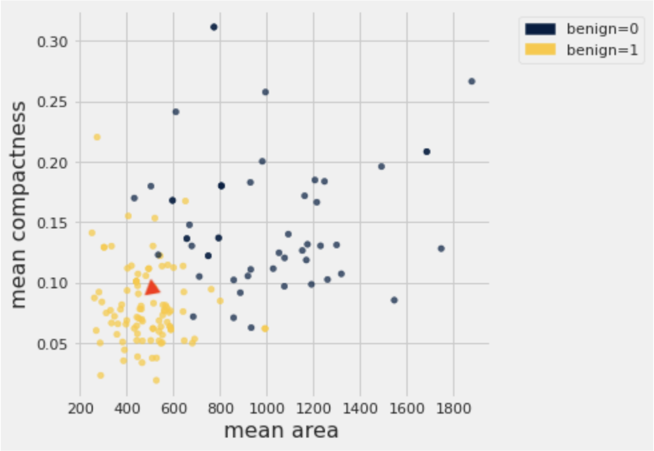
4. kNN Classifier

Significant research has been done to understand whether a breast tumor is benign or malignant. Sarah wants to create a classifier that predicts whether a tumor is benign or not.

a. Sarah begins by attempting to classify a new tumor based on the average compactness and average area of the tumor. Draw the decision boundary that the k nearest neighbors algorithm (with k = 3) would generate for this problem.



b. Now Sarah wants to classify a new tumor (represented as a triangle in the scatter plot on the next page). Describe the steps she would take to classify this new point based on a k nearest neighbors classifier with k=3.



c. Zach suggests that Sarah should use a different k for her classifier like k=4 or k=8. Is Zach's suggestion reasonable?

d. When trying to develop a classifier, we split our original dataset into a training and a test set. We don't look at or use the test set until we have finished training. Why is that a good idea in general? What might happen if we didn't?

e. Suppose in our breast tumor training dataset we have 30 benign=0 data points and 45 benign=1 data points. What k values are too large?

f. Dana suggests that we use a constant classifier which will always predict the class that is most common in the training set. In our test set, there are 12 benign=0 data points and 34 benign=1 data points. What will the accuracy of the constant classifier be on our test set?

g. Aside from the proportion of correct classifications, what are some other metrics we might consider to measure the quality of our predictions?