

CS 457 Homework 10: Linear Regression

for Mon Sep 28 2026

Let's do some more linear regression! You will turn in your work at the end.

1. We'll start with a refresh on some basics. Load Anscombe's Quartet:

```
import pandas as pd
a = pd.read_csv("https://www.cs.unh.edu/~rum1/cs457/anscombe.csv")
```

This is actually four datasets in one. Check `a.columns` to see what's in the dataframe. The x values for sets 1–3 are in column `x123`, the fourth is at `x4`, and the y values are `y1` etc. Check out the four datasets visually — I recommend a scatterplot, as in:

```
import seaborn as sns
sns.scatterplot(data=a, x='x123', y='y1')
```

but of course you'll need to do four. They look pretty different, right?

Now get a sense for how misleading statistics can be by either using individual methods (like `a["y1"].mean()`, `min`, `max`, `std`) or an all-in-one command like `a.describe()` to get summary statistics. Now you know why plotting should always be the first step in understanding a dataset! Write up your findings, either in a separate document or as a 'text cell' right there in your colab notebook (hover over the bottom of your code cell to add a cell below it).

2. Now for some linear regression. Fit a linear model to each of the four data sets, as in:

```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import accuracy_score
# features always assumed to be 2D so X is a dataframe
X = a[["x123"]]
y = a["y1"]
model1 = LinearRegression().fit(X, y)
```

Print `model1.coef_` and `model1.intercept_` to see the regression coefficients for each model (don't forget the underscores!). Plot the models against the data they were trained on:

```
import seaborn as sns
import matplotlib.pyplot as plt
sns.scatterplot(data=a, x="x123", y="y1")
Xsorted = X.sort_values("x123")
plt.plot(Xsorted, model1.predict(Xsorted), color="red")
plt.show()
```

Again, write up your findings.

3. Now, write a Python function that takes a dataframe and two column labels, fits a linear model, prints the model parameters and shows a plot. Demonstrate this on at least two of the data sets.
4. Please email your write up, including your final function to Jack, CCing Wheeler. If you used colab, you should be able to 'Print to PDF' to generate your writeup. (In general that's not the best way, but for a simple notebook like this assignment it should suffice.)