CST 383 - Intro to Data Science

Dr. Glenn Bruns

# Lab: Linear regression, part 2

In this lab we will continue to perform linear regression using the CPU data set.

1. Read the CPU data:

df = pd.read\_csv("https://raw.githubusercontent.com/grbruns/cst383/master/machine.csv")

df.index = df['vendor']+' '+df['model']

df.drop(['vendor', 'model'], axis=1, inplace=True)

df['cs'] = np.round(1e3/df['myct'], 2) # clock speed in MHz

1. Create NumPy arrays X and y from the data. The target variable y should come from column 'prp' of df. For X, choose two other columns of df.
2. Create training and test sets using train\_test\_split(). Use Scikit-Learn's LinearRegression class to create a linear model from X\_train and y\_train.
3. What are the coefficients of your model?
4. Are both the predictor variables of your model important in predicting ‘prp’?
5. What is the R-squared value for your model? (Use X\_train and y\_train to get the R-squared value.) Did you get a good R-squared value? What's the best possible R-squared value?
6. Produce a scatterplot in which the predicted prp values are shown on the x axis and the actual prp values are shown on the y axis.
7. Repeat steps 2-7 using a different pair of predictor variables.
8. Repeat steps 2-7 using all predictor variables. Which predictors appear to be most important?