1 Instructions

This homework is essentially a miniaturized Final Project. This is your chance to apply the concepts of the Final to a friendlier set of data in a collaborative forum. This is also your chance to get feedback on your writing, formatting, etc.

For this homework, we will use a rather popular dataset for regression — predicting the progression of diabetes. Ten baseline variables were obtained for each of $n = 442$ diabetes patients along with the response variable of interest, a quantitative measure of disease progression one year after baseline. The data file `diabetes_students.csv` has all of this data, except the response variable has been removed for the last 50 rows. Those 50 rows will serve as a test set. A description of the columns is in Table 1.

Variable	Description
AGE	Patient's age (years)
SEX	Patient's sex
BMI	Body-Mass Index
BP	Average Blood Pressure
S1 – S6	Blood Serum Measurements
Y	Response Variable (disease progression after one year)

Table 1: Variables used in `diabetes.csv`

1. **Data preprocessing**: Perform a train/validation split. How much of your data you use for validation and how you use it is up to you. I have already held out the test data.
2. Build a **regression model** predicting `Y` by using the predictor variables. You may transform any variables and include any interactions or higher order terms you deem necessary, but you do not have to. You may also build several candidate models, compare them, and choose one but you do not have to. Include in your write-up the following:
   - One-paragraph executive summary of results
   - Your final model as an equation
   - Steps you took to create the model(s) (i.e., feature selection)
   - Assessment of regression assumptions (plus tests and plots as needed)
   - Remedial steps (if any) taken to create a valid model that passes assumption tests, and follow-up reassessment of assumptions
   - Overall assessment of your final model (e.g., measures of usefulness, inference and interpretations of features, etc.), including comparison of RMSE for training and validation sets.
3. **Model test**: Generate predicted Y values (column name `Y`) and a 90% Prediction Interval (column names `Lower` and `Upper`) for each data point in `diabetes_test.csv`. Save this as `predictions.csv` and submit it with your assignment. I will evaluate the RMSE for your test set and Prediction Interval coverage of the actual data and give you feedback on how your model performed. The individual(s) with the lowest test RMSE will get 10 bonus points.

   Remember that any transformations must be reversed for the values in `predictions.csv` by applying inverse transformations. See lesson 39 and 40 for examples.

4. If you decide to eliminate the intercept from your model, you'll find that R^2 is not comparable and loses its meaning. Some languages (such as R) automatically alter the formulation to instead calculate the percentage of variance explained by the model. You can use the below code to calculate the percentage of variance explained for any model — including intercept or not.

```
from sklearn.metrics import explained_variance_score
explained_variance_score(y_train,model.predict())
```

2 Grading

I will be grading you on both the correct application of techniques and the quality of your writing. This includes the effectiveness of your communication of the steps you took for model building and assumption testings as well as your reporting on the usefulness of the model. Don't spend too much time building your model unless you feel very confident with the basics; find a decent model and move on. While model quality is important, your effectiveness at communicating it is far more important for this assignment.

3 Submission

To reiterate, I expect the following files to be submitted:

1. PDF report
   - No raw Python code.
   - All plots include elements needed to stand alone (titles, axis labels, etc.) and are clear (increase size/dpi as needed)
2. Jupyter notebook(s)
3. `predictions.csv`
   - Must include the index column plus columns `Y`, `Lower`, and `Upper` at a minimum.
   - Please do not sort or re-order these predictions.

4 Hints

- This is real data. There is no perfect model. There is no perfect transformation. Do your best.
- Use your validation set to compare between candidate models.
- If you transform your response, don't forget to back-transform your predictions and intervals.
- You could spend a few hours on this. You could spend 80 hours on this. You could probably explore this data for months.
- Write an outline before you start writing. This is a paper, not a traditional math assignment.
- RMSE will likely be in the region of 50. If you're getting numbers significantly different from that, something is going wrong.