**Week 2 supervised learning regression; BIOS 7747; Jake Krol; 2023-09-20**

Build and evaluate a regression model that can predict the systolic blood pressure after 30 days of treatment. Provide:

**1: A mathematical equation that predicts the systolic blood pressure after 30 days of treatment.**

$$\hat{y} = -7.407 + 0.0583x_1 + 2.786x_2 + 1.022x_3 - 3.068x_4$$

Where $\hat{y}$ = predicted systolic blood pressure (BP); $x_1$ = Age; $x_2$ = Sex; $x_3$ = Initial blood pressure (mm Hg); $x_4$ = Drug dose

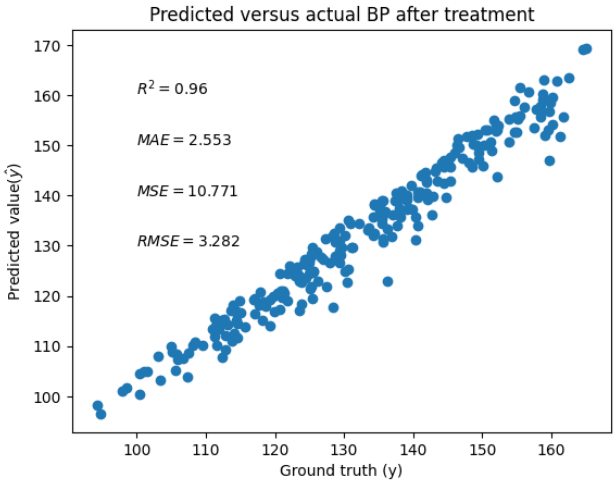


**Figure 1: Ground truth BP after treatment versus predicted BP after treatment**

**2. A performance evaluation of the predictive model in the training dataset. Please, use both quantitative metrics and graphic representations to explain your model's performance.**

To preface, we are fitting the model on the original training dataset, the model is likely overfit and may not truly represent the relationship between the predictors and outcome for new patients.

For predictive performance of the regression model. The model has **low error when fitted against the original training data (RMSE = 3.282 BP [mm Hg])** (**Fig 1**). The amount of variance explained in the outcome via the predictor variables is high with **R^2 ~ 0.96**. However, figure 1 only suggests overall performance and does not tell us about the relationship between individual predictors and the outcome. Further evaluation is needed for linear/non-linear relationships and significance.

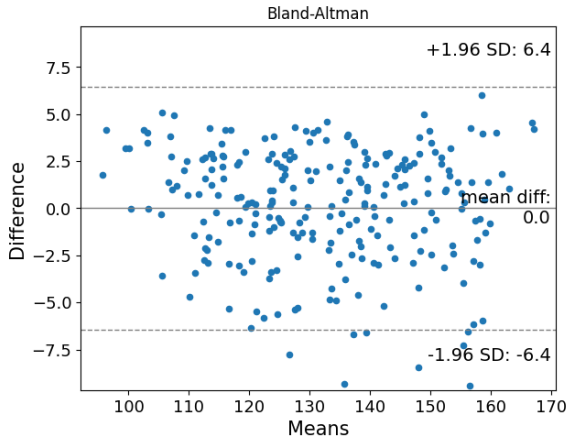


**Figure 2: Bland-Altman plot**

After visualizing the residual, $y - \hat{y}$, versus the paired means of ground truths and prediction, $mean(y_i, \hat{y}_i)$, in a Bland-Altman plot (**Fig 2**), it's shown that the **residuals are not centered around the mean difference of 0**. The increased amount of high magnitude of negative residuals compared to positive residuals indicates the regression model is over-estimating outcome BP. The model and ground truth BP do not agree.

**Week 2 supervised learning regression; BIOS 7747; Jake Krol; 2023-09-20**

Discuss the quantitative effect of every individual variable in the outcomes. For each available variable, would you say that:

1. **it has a linear effect in the treatment outcomes?**
2. **it has a non-linear effect in the treatment outcomes?**
3. **it has no significant effect in the treatment outcomes?**

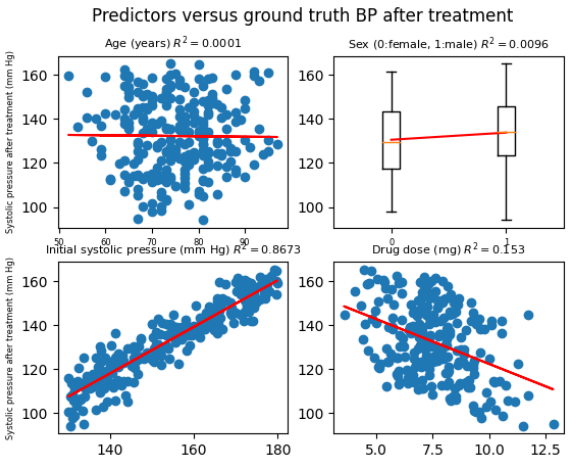


**Figure 3: Predictors v ground truth BP after treatment**

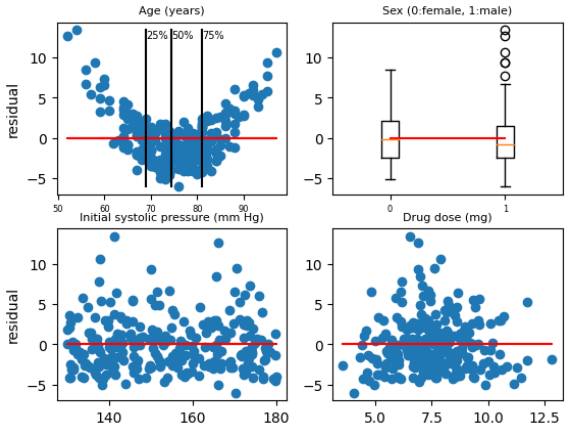


**Figure 4: Predictors and Residuals**

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -7.4072 | 3.216 | -2.303 | 0.022 | -13.741 | -1.073 |
| Age (years) | 0.0583 | 0.024 | 2.386 | 0.018 | 0.010 | 0.107 |
| Sex (0:female, 1:male) | 2.7864 | 0.422 | 6.610 | 0.000 | 1.956 | 3.617 |
| Initial systolic pressure (mm Hg) | 1.0224 | 0.015 | 70.299 | 0.000 | 0.994 | 1.051 |
| Drug dose (mg) | -3.0678 | 0.133 | -23.091 | 0.000 | -3.329 | -2.806 |

**Figure 5: p-values of coefficients**

**Figure 3** plots each predictor against the ground truth BP after treatment to get a glance at possible linear relationships. R squared values were calculated, too; however, we cannot have strong confidence in linearity until we also see the residual plots for each predictor, calculated as $y - \hat{y}$, (**Figure 4**) where we expect the residuals to be uniformly distributed across the predictor values in the case of linearity. **Figure 5** shows the p-values of the predictors to test against the null hypothesis that the predictor has no significant effect on the outcome.

The high R^2 (0.86) (**Fig 3**) and relatively* uniform residual plot (**Fig 4)** for **initial systolic BP** indicates this variable has a moderate* **linear** effect on the BP after treatment. Also, a p-value of less than 0.05 (**Fig 5**) rejects the null hypothesis of no significant correlation, instead suggesting **significant correlation** between initial systolic BP and BP after treatment.

**Age** has a low correlation with outcome BP R^2 (0.0001) (**Fig 3**) and a non-uniform residual plot (**Fig 4**) which suggests there is a **no linear** relationship between age and outcome BP; the model underpredicts for the lower and upper quartiles of age values. However, **Figure 5** shows a **significant correlation** between age and outcome BP (p-value < 0.05) that likely cannot be modeled accurately using linear methods.

Following a similar trend, **sex** has a low correlation with outcome BP (R^2 ~ 0.0096) (**Fig 3**) and the residual plot shows the model is slightly over-predicting, residuals tend to be negative, for male sex (**Fig 4**); the low correlation and non-uniform residual distribution across sex suggests a **non-linear** relationship. Again, despite non-linearity, the relationship between sex and outcome BP is **significant** ($p < 0.05$).

The remaining predictor variable, **drug dose,** also has a low correlation with outcome BP (R^2 ~ 0.153) (**Fig 3**), and the residual plot is non-uniform (**Fig 4**) with a greater frequency of underpredicted outcomes for values near the median of the drug dose value distribution. The lack of uniformly distributed residuals indicates **non-linearity** between drug dose and outcome BP. The correlation between drug dose and outcome BP is **significant** ($p < 0.05$) (**Figure 5**).