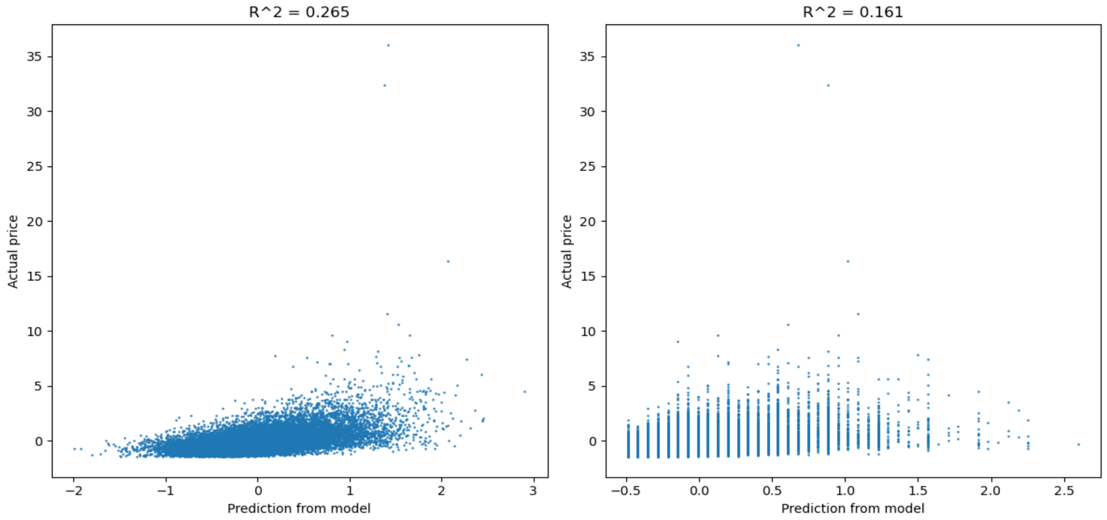
1. I first cleaned the original data by dropping all rows with NAs in them. I also dropped the rows with “other” as the gender and replaced all “male” tags with number 0 and “female” tags with number 1. After that I normalized all columns with quantitative data (including the dummy variables which I am not sure if is appropriate or not). I then built the multiple linear regression model using the total income as the outcome variable and all other quantitative data columns (excluding income details and two dummy variables) as the predictor variables.

The rows with NAs and “other” gender are not very useful for building the linear regression model if we want to include the dummy variables and given my current level of knowledge I do not know how to extract values from them. I normalized the data because to find out which predictor is the most useful we need to compare their normalized regression coefficients.

The R-square of the multiple linear regression model is 0.265. Years of experience is the predictor with the largest coefficient. In a single linear regression model using YOE as the predictor, R-square value is 0.161.

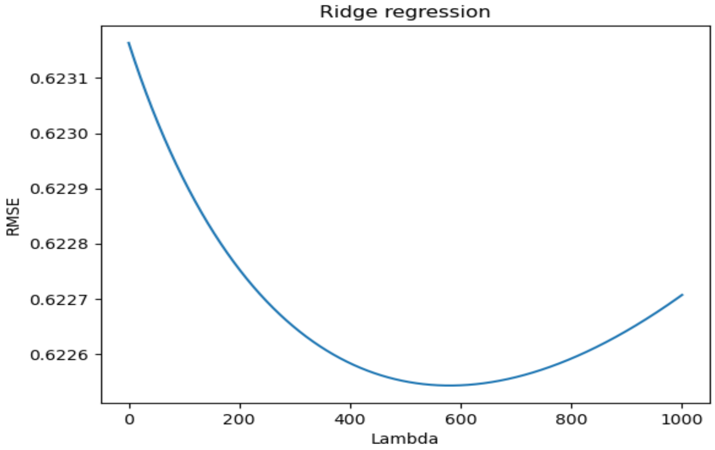


YOE is the best predictor. It explains about 60% of the variance explained by the multiple linear regression model.

2. I divided the predictors’ data in Q1 into training and testing sets and plotted the RMSE of the Ridge models trained by the training set using different lambda values when they are applied to the testing set. The Lambda value corresponding to the lowest RMSE was found.

To determine the best choice of Lambda we need to use hyperparameter tuning in which we try to find the lambda that leads to the Ridge model with the lowest RMSE on the testing set. A train-test split can prevent overfitting or leakage.

The Ridge model achieves an R-square value of 0.285 using the optimal lambda 579.9.

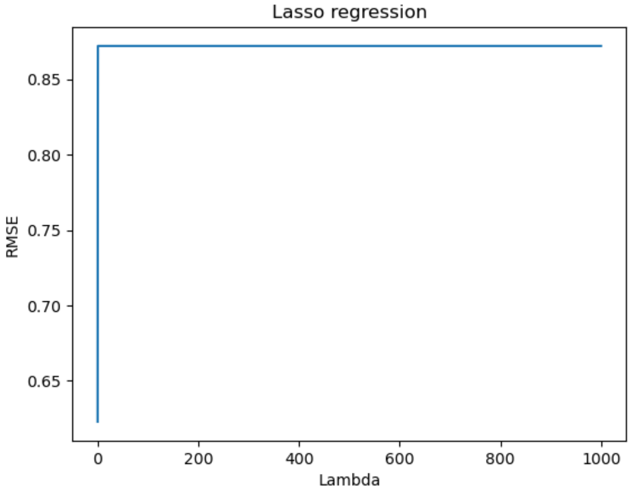


Ridge regression is slightly better than OLS, but the Lambda is suspiciously large for unknown reasons.

3. I divided the predictors' data in Q1 into training and testing sets and plotted the RMSE of the Lasso models trained by the training set using different Lambda values when they are applied to the testing set. The Lambda value corresponding to the lowest RMSE was found.

To determine the best choice of Lambda we need to use hyperparameter tuning in which we try to find the Lambda that leads to the Lasso model with the lowest RMSE on the testing set. A train-test split can prevent overfitting or leakage.

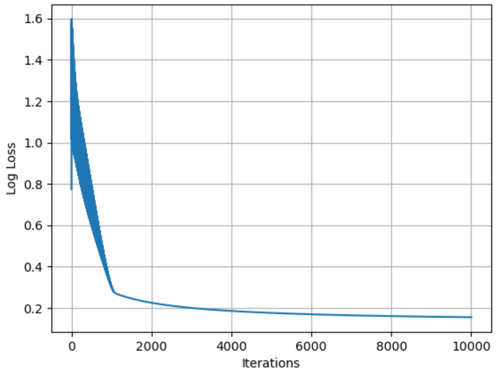
Lasso regression achieves an R-square value of 0.290 using the optimal lambda 0.001.



Lasso regression demonstrated a slight advantage over the other two regression methods. The predictor Race=white's coefficient was shrinked to 0.

4. A logistic regression model is built using the gradient descent method to determine the best weight and bias vectors w and b that results in the smallest cost.

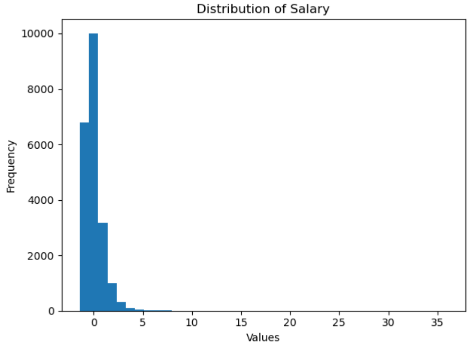
Due to large amounts of data the iteration is very slow. The number of iterations is set as 10000, and the learning rate is 0.0001. The weight w is finalized as 0.0588.

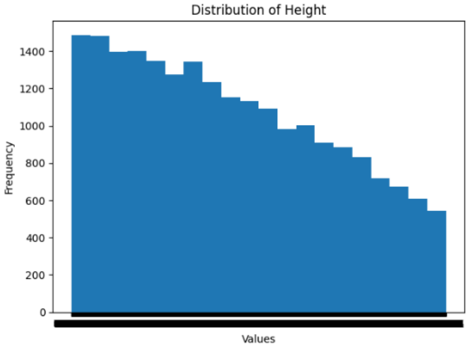


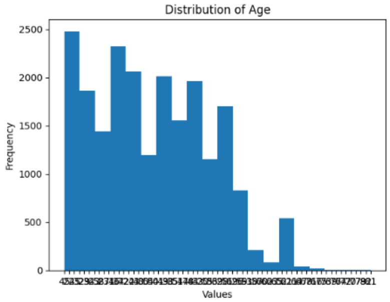
The beta seems to be very small and thus probably insignificant.

5. The median value of total compensation is calculated and data smaller than this median are replaced by 0 while numbers bigger than the median are changed into 1. Four logistic regression models with the same output variable (compensation=high/low) but different predictors (years of relevant experience, age, height, SAT score and GPA) are trained using the same process as that of Q4.

Extra Credit







The three predictors don't seem to be normally distributed. This is not too surprising because the data points (people) included in this data frame are all tech industry employees, who share some intrinsic similarities with each other and do not correctly reflect the qualities of the whole population.