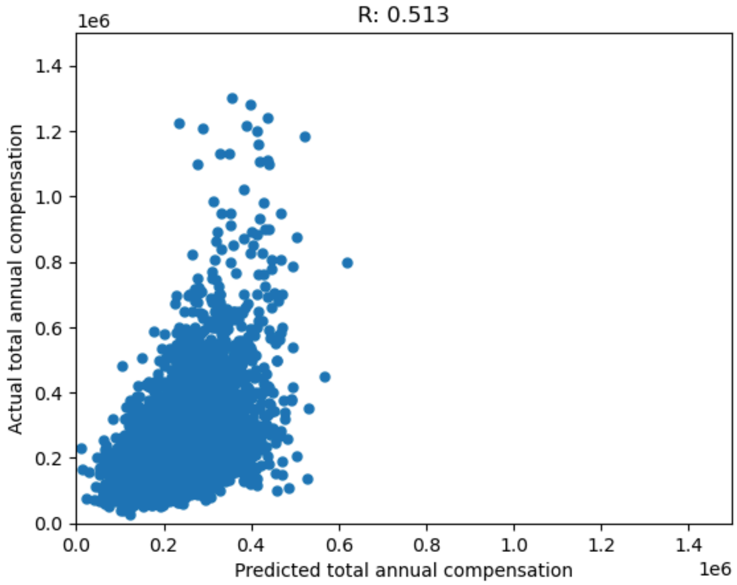
1. Using multiple linear regression: What is the best predictor of total annual compensation, how much variance is explained by this predictor vs. the full multiple regression model?

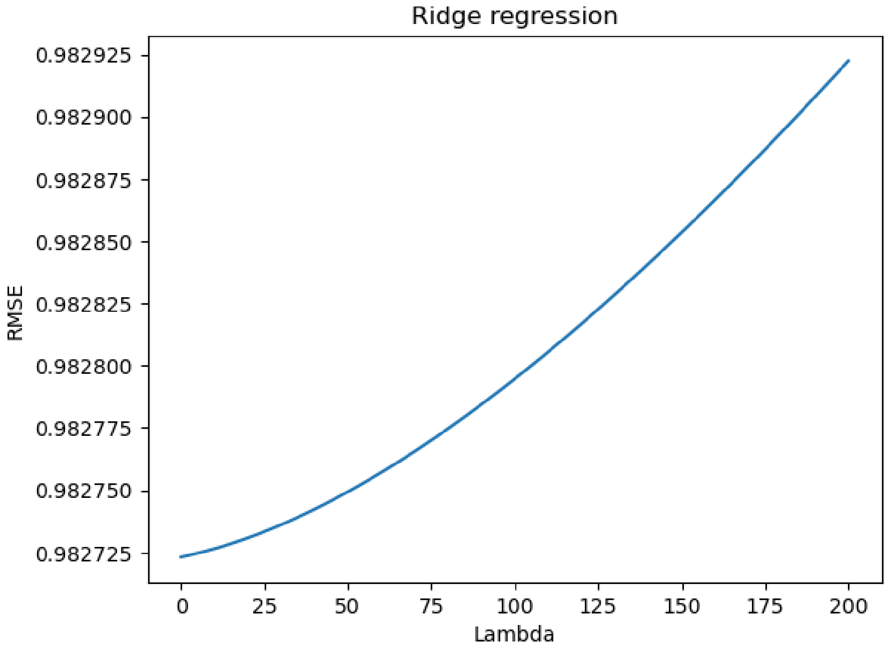
First, I removed the rows with no value for variables 21 or 22, because it makes it easier to remove columns later. Although this significantly reduces the size of the dataset, there are still an amount of entries which I believe is sufficiently large. I then split the zodiac variable into 11 dummy variables, leaving 1 out to avoid overdetermination. They are represented in columns 28-38. I also did this with gender, which is now represented in column 39, splitting by male and then all other genders. I removed the qualitative data and variables 5-7 from the dataset due to the reasoning given in the instruction document. I also removed variables 21 and 22 as they are categorical and do not make sense to interpret as they are, and the last variables from their respective sets of dummy variables (15 and 20), as they are already represented by the entry being in the model and not being represented by the other dummy variables.

I then split the data into a 60/40 training/test split and fit a multiple linear regression model. This gave an R^2 of 0.26257196482060896, which is not very high and indicated that total annual compensation is not very well predicted by all of the predictors together. The residual plot for this model is shown below. This plot shows us that the data is heteroscedastic, violating one of the assumptions of linear regression, and meaning predictions will be less accurate for larger total annual compensations. I then used the same split to fit a single linear regression for each of the predictors to find the largest R^2 value. The largest was 0.16481925763631178, which although not very high, was much larger than the others, which were mostly negative. The variable the model came from was years of relevant experience, meaning that of the predictors we have, years of relevant experience is the best predictor of total annual compensation. About 16.4819% of the variance was explained by this predictor versus the 26.2572% explained by the full multiple regression model, showing that the full multiple regression model improved on the predictions made by just years of relevant experience by a fair amount.



2. Using ridge regression to do the same as in 1): How does the model change or improve compared to OLS? What is the optimal lambda?

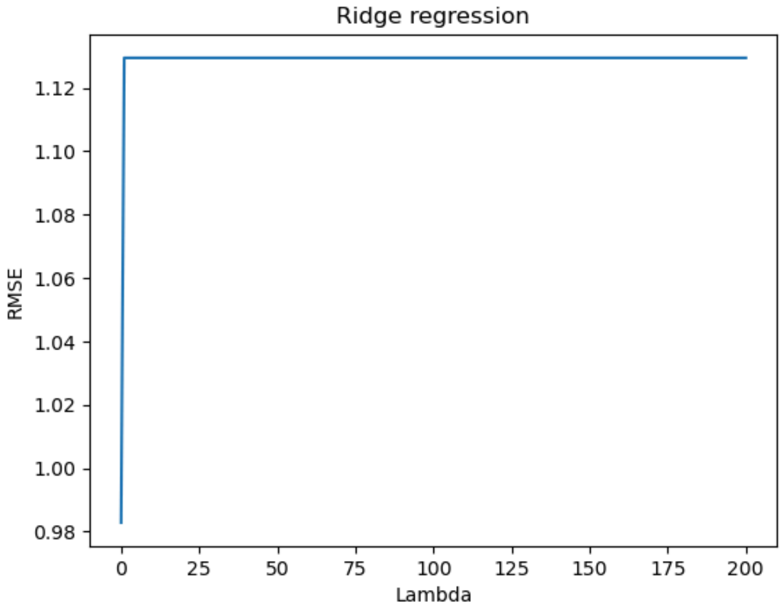
First, I standardized the data by converting each value in each column to its z-score. I did this because ridge regression penalizes features on a higher scale. I split the data using a 80/20 train/test split. I then graphed 200 lambda values between 0.001 and 200 against the root mean squared errors generated from the ridge regression models using those lambdas. I did this in order to find the optimal lambda, minimizing the root mean squared error, giving this graph below, which indicated that 0.001 was the optimal lambda.



Fitting a ridge regression using this optimal lambda gives a new R^2 of 0.241705308579764 02, which is slightly worse than the original model, but this is expected as we are reducing the relevancy of some variables. Splitting the normalized dataset in half and doing cross-validation using the ridge regression yields slightly better scores (0.30115607, 0.32283085, 0.33407484, 0.29783283, 0.20334034) than OLS (0.30115606, 0.32283084, 0.33407484, 0.29783282, 0.20334034), meaning that the ridge regression model is slightly better than the OLS one.

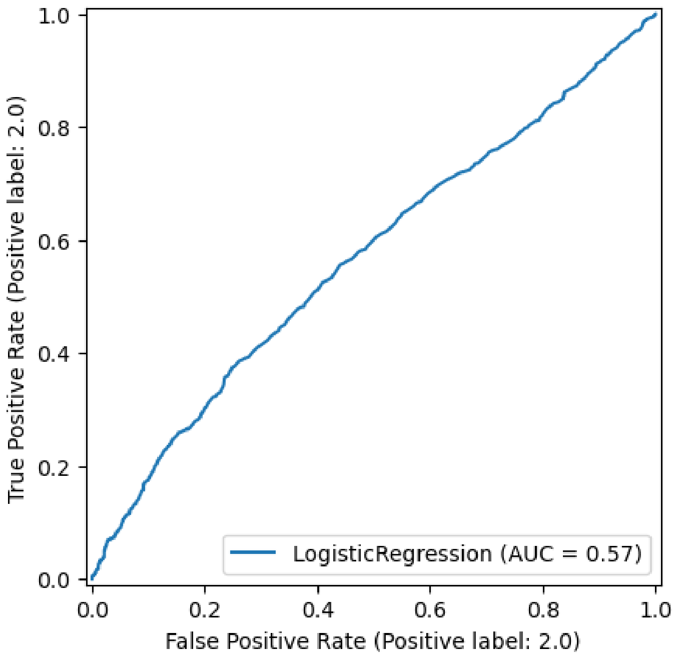
3. Using Lasso regression to do the same as in 1): How does the model change now? How many of the predictor betas are shrunk to exactly 0? What is the optimal lambda now?

I used the same process as the previous step, except fitting a Lasso regression instead of ridge. I got this graph for lambdas, which indicates that the optimal lambda is 0.001. Cross validation scores (0.30027829, 0.3233864 , 0.3331636 , 0.29826042, 0.20302568) indicated that this model was largely the same as the OLS one, although slightly worse and better at different times. Confusingly, none of the predictor betas were shrunk to exactly 0, although many were close.

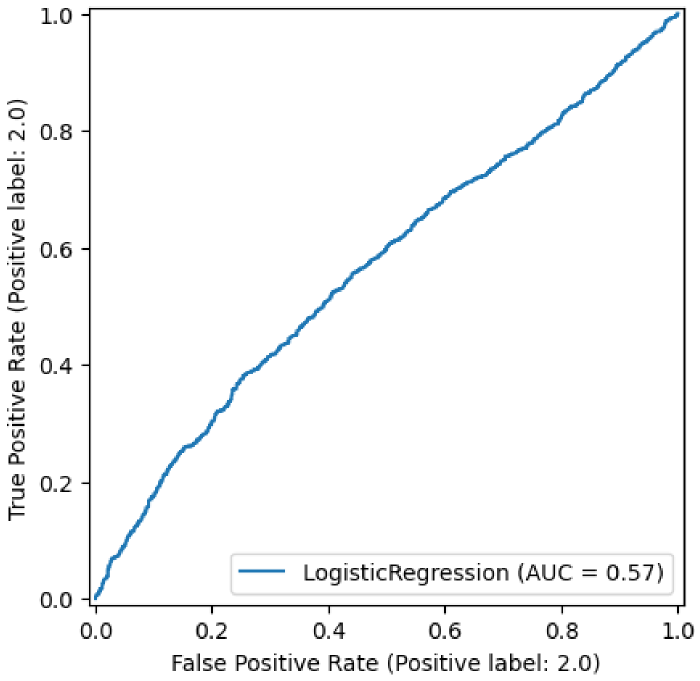


4. There is controversy as to the existence of a male/female gender pay gap in tech job compensation. Build a logistic regression model (with gender as the outcome variable) to see if there is an appreciable beta associated with total annual compensation with and without controlling for other factors.

I used the dummy variable for gender created in the first question. I grouped all genders aside from male with female, as they represent a very small part of this population and are likely to be affected by similar pay discrimination. I normalized the total annual compensation in order to get more meaningful betas. I split the data using an 80/20 training/test split. This model gave the ROC below. The model has an AUROC of 0.57, which is better than a random classifier. The beta associated with total annual compensation is 0.28952961, which is moderately appreciable.

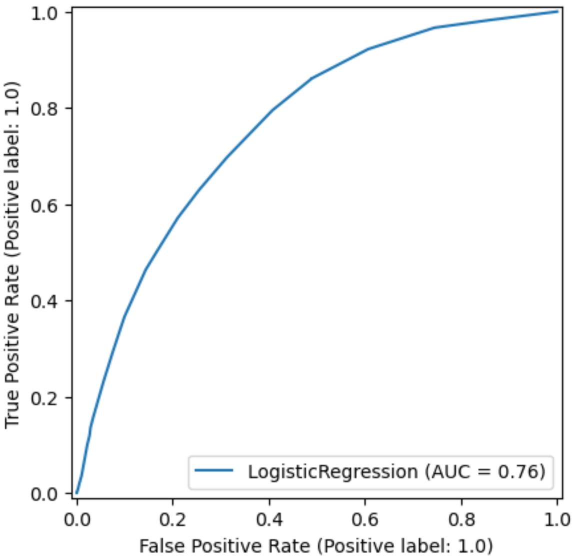


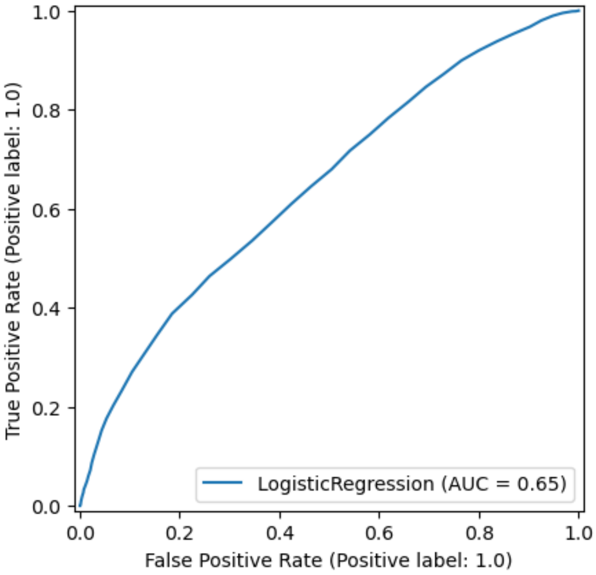
I then built another logistic regression model using all other features from the previous questions (excluding gender) with the addition of total annual income in order to control for other features. This gives a very similar ROC and AUROC. The beta for total annual compensation is 0.26847917, which is smaller than before, but not by much.

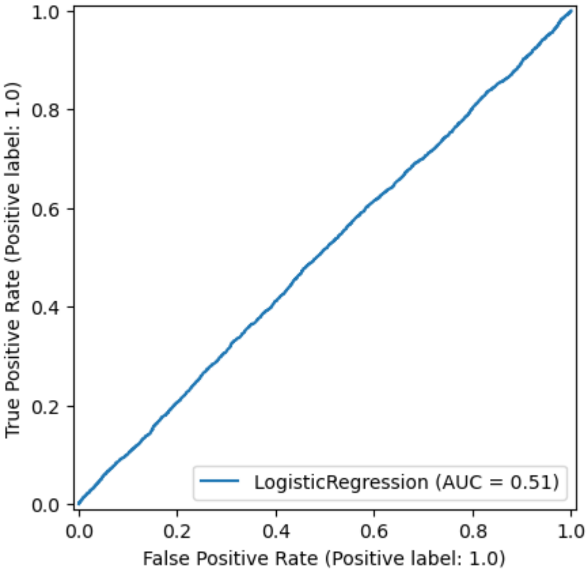


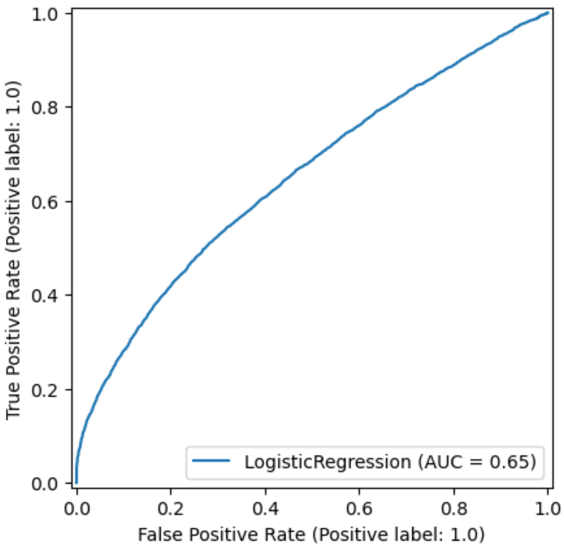
5. Build a logistic regression model to see if you can predict high and low pay from years of relevant experience, age, height, SAT score and GPA, respectively.

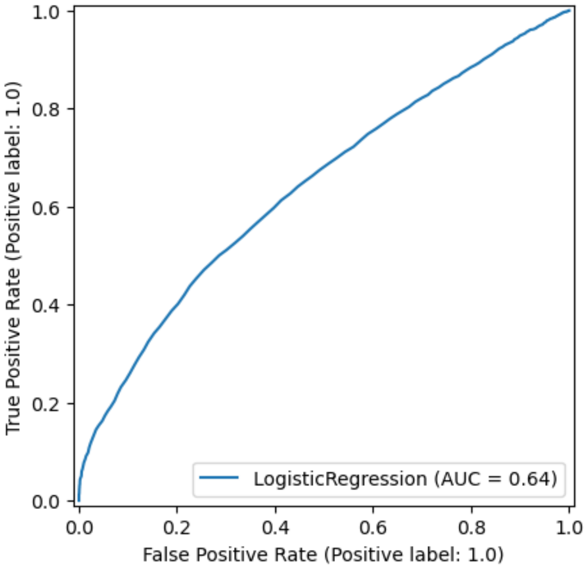
First, I created a dummy variable for high/low pay, considering all total annual compensation above the median to be high and all total annual compensation below or equal to the median to be low pay, represented by 1 and 0 respectively. Using a 80/20 training/test split and building logistic regression models for each variable gives the following ROCs in the order years of relevant experience, age, height, SAT score, and finally GPA going from left to right and then up to down.









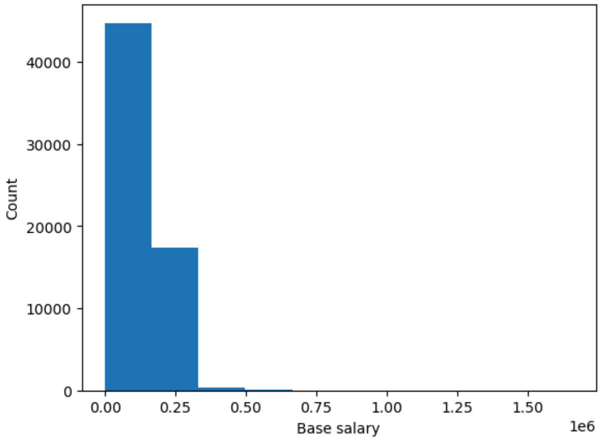


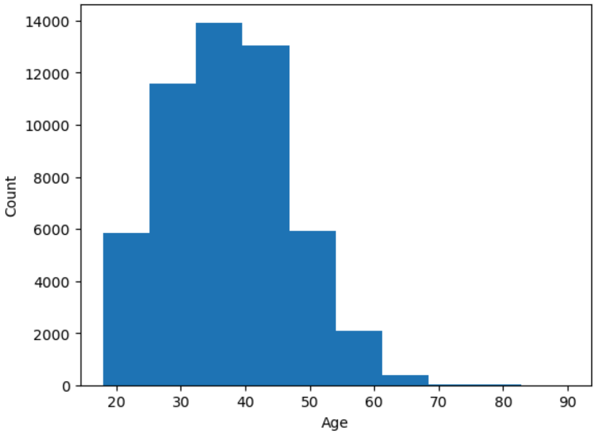
From the AUROCs, you can predict total annual compensation from years of relevant experience pretty well. The worst predictor is height, which is basically the same as predicting randomly. All of the other predictors are similarly slightly better than chance.

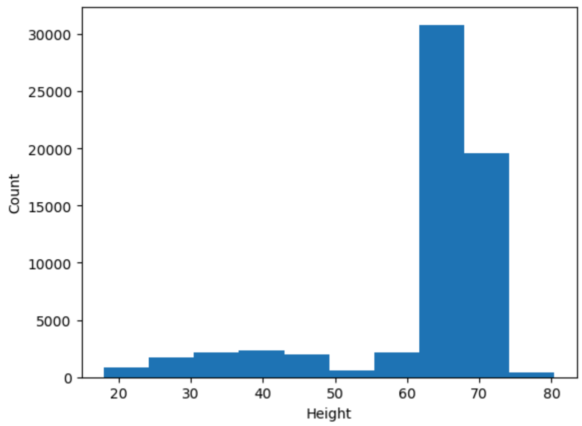
Extra credit:

a) Is salary, height or age normally distributed? Does this surprise you? Why or why not?

Plotting histograms gives the following charts.







Age and height are relatively normally distributed but not salary, which is not surprising. Normal distributions come from many independent factors combining, which is true of height and age for people in a company. However, salary is based on the structure of the company. It makes sense that there are exponentially less people as the base salary increases given that the amount of people on each level of a company typically also decreases as you go upwards, with salary increasing. It is also worth noting that some of the height data appears to be incorrect, as there is a fair amount of extremely short people.

b) Tell us something interesting about this dataset that is not already covered by the questions above and that is not obvious.