Student Test Score Prediction w/ Linear Regression

Giorgos Tzimas

June 2023

1 Dataset Description

1.1 Background

The panel dataset contains cross-section data from the High School and Beyond survey conducted by the Department of Education in 1980. The data was collected in order to study the relationship between early high school experiences and the students' educational experiences in high school and after, as well as the effects of how family, community, school and classroom factors affect student performance.

For further information: **https://nces.ed.gov/surveys/hsb/**

1.2 Observations and Attributes

The dataset comprises of **4,739** entries with **15** distinct features. These features contain a range of information pertaining to both the student, the school they attend and their parents.

1.3 Data Types

Out of the 15 features in the dataset:

- 8 are of type "object" or text
- 2 are of type "integer"
- 5 are of type "float"

2 Data Cleaning

None of our records or features in the dataset contain any missing values. Therefore, we will not have to use any imputation methods in this case.

Feature	Description
gender	Student's gender
ethnicity	Student's ethnicity (African-American, Hispanic or other)
score	Base year composite test score
fcollege	Is the father a college graduate?
mcollege	Is the mother a college graduate?
home	Does the family own their home?
urban	Is the school in an urban area?
unemp	County unemployment rate in 1980.
wage	State hourly wage in manufacturing in 1980.
distance	Distance from 4-year college (in 10 miles).
tuition	Average state 4-year college tuition (in 1000 USD).
education	Number of years of education.
income	Is the family income above USD 25,000 per year?
region	School region (West or other).

Table 1: Dataset features and their descriptions

Feature	Data Type
ID	int64
gender	object
ethnicity	object
score	float64
fcollege	object
mcollege	object
home	object
urban	object

Feature	Data Type
unemp	float64
wage	float64
distance	float64
tuition	float64
education	int64
income	object
region	object

Table 2: Dataset attributes and their types

3 Data Transformation

All of the categorical variables in the dataset will be one-hot encoded (dummy-coded) into numeric variables with n-1 columns (n=number of levels in the feature).

- n_gender: 1 if "male", 0 if "female
- n_ethnicity_hispanic: 1 if "hispanic", 0 if otherwise
- n_ethnicity_afam: 1 if "afam", 0 if otherwise
- n_fcollege: 1 if "yes", 0 if "no"
- n_mcollege: 1 if "yes", 0 if "no"

Feature	Missing
ID	0
gender	0
ethnicity	0
score	0
fcollege	0
mcollege	0
home	0
urban	0

Feature	Missing
unemp	0
wage	0
distance	0
tuition	0
education	0
income	0
region	0

Table 3: Features with number of missing values

- n_home: 1 if "yes", 0 if "no"
- n_urban: 1 if "yes", 0 if "no"
- n_region: 1 if "west", 0 if "other"

	ID	score	unemp	wage	distance	tuition	education	n_gender	n_ethnicity_afam	n_ethnicity_hispanic	n_fcollege	n_mcollege	home_yes	urban_yes	n_income	n_region
0	1	39.150002	6.200000	8.090000	0.200000	0.889150	12	1	0	0	1	0	1	1	1	0
1	2	48.869999	6.200000	8.090000	0.200000	0.889150	12	0	0	0	0	0	1	1	0	0
2	3	48.740002	6.200000	8.090000	0.200000	0.889150	12	1	0	0	0	0	1	1	0	0
3	4	40.400002	6.200000	8.090000	0.200000	0.889150	12	1	1	0	0	0	1	1	0	0
4	5	40.480000	5.600000	8.090000	0.400000	0.889150	13	0	0	0	0	0	0	1	0	0

Table 4: Dataset after dummy-coding

4 Exploratory Data Analysis

In this section, we will visually explore our dataset to gain some insight into the relationships between different features.

4.1 Correlation Matrix

	score	unemp	wage	distance	tuition	education	n_gender	n_ethnicity_afam	n_ethnicity_hispanic	n_fcollege	n_mcollege	home_yes	urban_yes	n_income	n_region
score	1.000000	-0.025309	0.116627	-0.067979	0.129858	0.465187	0.080169	-0.284252	-0.161187	0.250970	0.189563	0.125874	-0.085139	0.178368	-0.026034
unemp	-0.025309	1.000000	0.266771	0.293036	0.184027	-0.014746	-0.028380	-0.059376	0.088610	-0.100055	-0.086024	0.005036	-0.052140	-0.078126	-0.041865
wage	0.116627	0.266771	1.000000	-0.000390	0.317727	0.023858	0.027211	-0.133696	-0.099519	0.030929	0.016939	0.067455	-0.032566	0.071773	-0.083654
distance	-0.067979	0.293036	-0.000390	1.000000	-0.100981	-0.093183	-0.003441	-0.100022	0.058226	-0.105474	-0.081523	0.019605	-0.289175	-0.080422	0.068089
tuition	0.129858	0.184027	0.317727	-0.100981	1.000000	0.039534	0.009025	0.059007	-0.304375	0.028796	0.036125	-0.000320	-0.016181	0.053930	-0.582348
education	0.465187	-0.014746	0.023858	-0.093183	0.039534	1.000000	0.009764	-0.090034	-0.043374	0.284356	0.225177	0.096519	-0.015005	0.219166	-0.024145
n_gender	0.080169	-0.028380	0.027211	-0.003441	0.009025	0.009764	1.000000	-0.039639	0.018810	0.040697	0.019911	0.038162	0.009728	0.058891	-0.013418
n_ethnicity_afam	-0.284252	-0.059376	-0.133696	-0.100022	0.059007	-0.090034	-0.039639	1.000000	-0.216348	-0.101376	-0.018085	-0.126593	0.183720	-0.096960	-0.144095
n_ethnicity_hispanic	-0.161187	0.088610	-0.099519	0.058226	-0.304375	-0.043374	0.018810	-0.216348	1.000000	-0.112351	-0.106208	-0.063879	0.077073	-0.128255	0.207679
n_fcollege	0.250970	-0.100055	0.030929	-0.105474	0.028796	0.284356	0.040697	-0.101376	-0.112351	1.000000	0.429710	0.077525	-0.052510	0.353563	0.029686
n_mcollege	0.189563	-0.086024	0.016939	-0.081523	0.036125	0.225177	0.019911	-0.018085	-0.106208	0.429710	1.000000	0.063916	-0.034306	0.245670	-0.011578
home_yes	0.125874	0.005036	0.067455	0.019605	-0.000320	0.096519	0.038162	-0.126593	-0.063879	0.077525	0.063916	1.000000	-0.096868	0.138826	0.004868
urban_yes	-0.085139	-0.052140	-0.032566	-0.289175	-0.016181	-0.015005	0.009728	0.183720	0.077073	-0.052510	-0.034306	-0.096868	1.000000	-0.071642	-0.052117
n_income	0.178368	-0.078126	0.071773	-0.080422	0.053930	0.219166	0.058891	-0.096960	-0.128255	0.353563	0.245670	0.138826	-0.071642	1.000000	0.007449
n_region	-0.026034	-0.041865	-0.083654	0.068089	-0.582348	-0.024145	-0.013418	-0.144095	0.207679	0.029686	-0.011578	0.004868	-0.052117	0.007449	1.000000

Table 5: Correlation matrix for all features

- None of our independent variables are strongly correlated with each other, which suggests multicollinearity might not be an issue when building the model
- The feature with the strongest correlation to our dependent variable (score) is education (0.46)
- The second strongest correlated variable with score is n_ethnicity_afam (-0.28)

4.2 Distribution of "score"

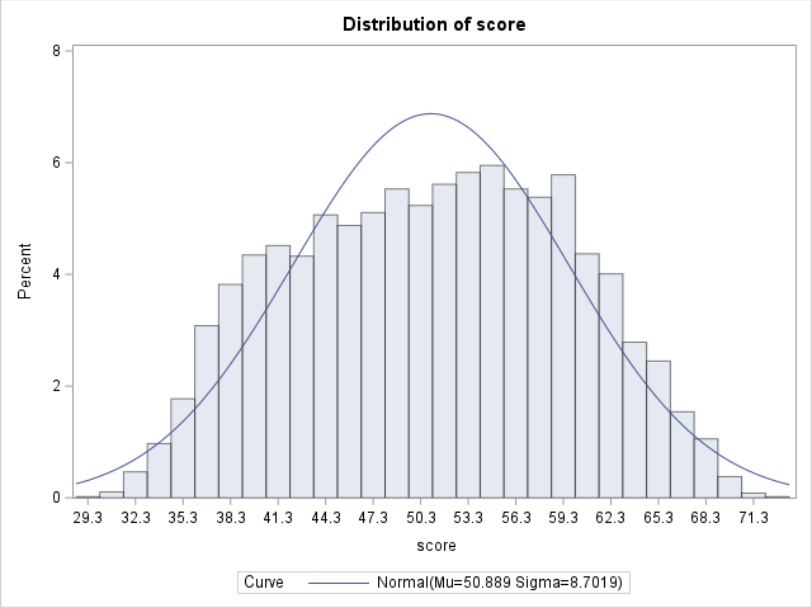

Figure 1: Distribution of variable "score"

Measure	Values
Mean	50.889029
Median	51.189999
Mode	56.020000
St. Dev.	8.700991
Variance	75.707252
Min	28.950001
Max	72.809998
Range	43.859997
IQR	13.844999
Skew	-0.032655

Figure 2: Location and variability metrics

- Our dependent variable is normally distributed with a negligible skewness of -0.03

- The mean score for all students is 50.89
- The standard deviation for student score is 8.7
- The lowest score is 28.95 and the largest is 72.81

4.3 Distribution of "unemp"

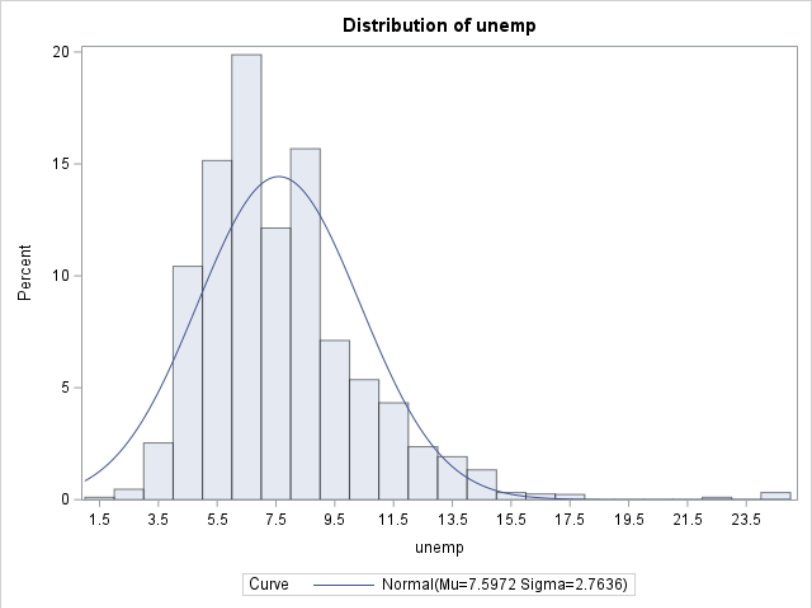

Figure 3: Distribution of variable "unemp"

Measure	Values
Mean	7.597215
Median	7.100000
Mode	8.000000
St. Dev.	2.763289
Variance	7.635768
Min	1.400000
Max	24.900000
Range	23.500000
IQR	3.000000
Skew	1.558486

Figure 4: Location and variability metrics

- The county unemployment rate is right-tailed with a skewness of 1.56
- The mean unemployment rate is 7.60 and the median is 7.1
- The standard deviation is 2.76
- The lowest unemployment rate is 1.40 and the largest is 24.90
- This distribution's right-skewness is due to potentially outlier values in the [21.5, 24.9] range
- We will take the log of "unemp" and compare its distribution to the original

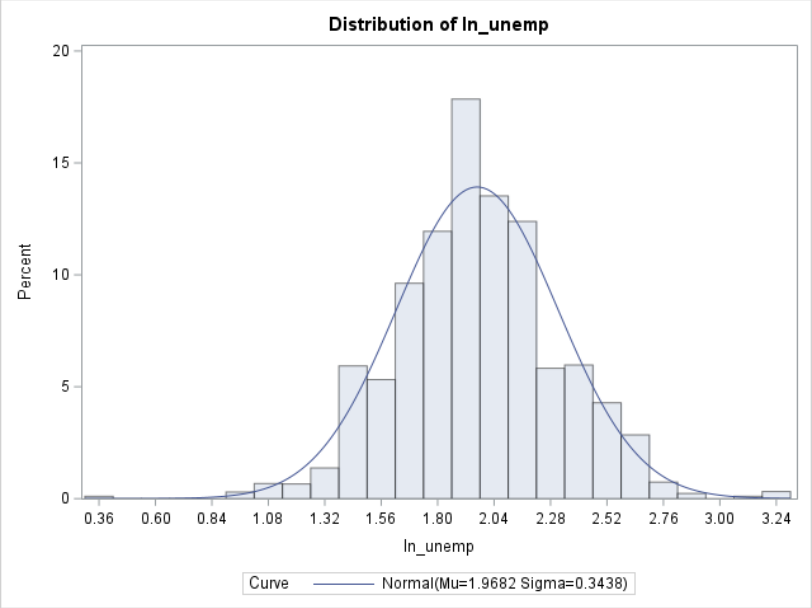

Figure 5: Distribution of variable "ln_unemp"

Measure	Values
Mean	1.968177
Median	1.960095
Mode	2.079442
St. Dev.	0.343759
Variance	0.118170
Min	0.336472
Max	3.214868
Range	2.878396
IQR	0.411099
Skew	0.017166

Figure 6: Location and variability metrics

- The right-tailedness of the original variable is mitigated by taking its natural logarithm with a new skew of 0.02

4.4 Distribution of "wage"

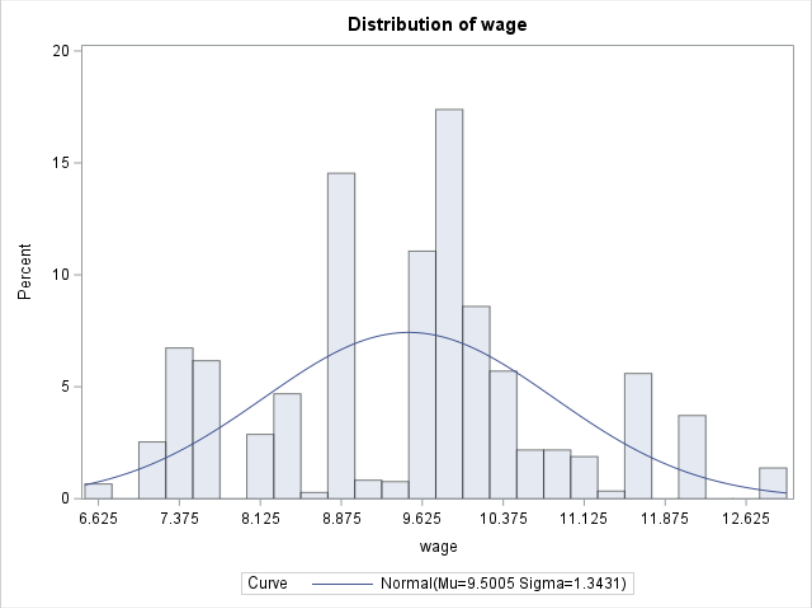

Figure 7: Distribution of variable "wage"

Measure	Values
Mean	9.500506
Median	9.680000
Mode	8.890000
St. Dev.	1.342925
Variance	1.803449
Min	6.590000
Max	12.960000
Range	6.370000
IQR	1.299999
Skew	0.093066

Figure 8: Location and variability metrics

- The hourly manufacturing wage is approximately normally distributed with a skewness of 0.09
- The mean wage is 9.50 and the median is 9.68
- The standard deviation is
- The tails of the distribution are "fatter" than a standard normal distribution
- The lowest wage is 6.59 and the largest is 12.96

4.5 Distribution of "distance"

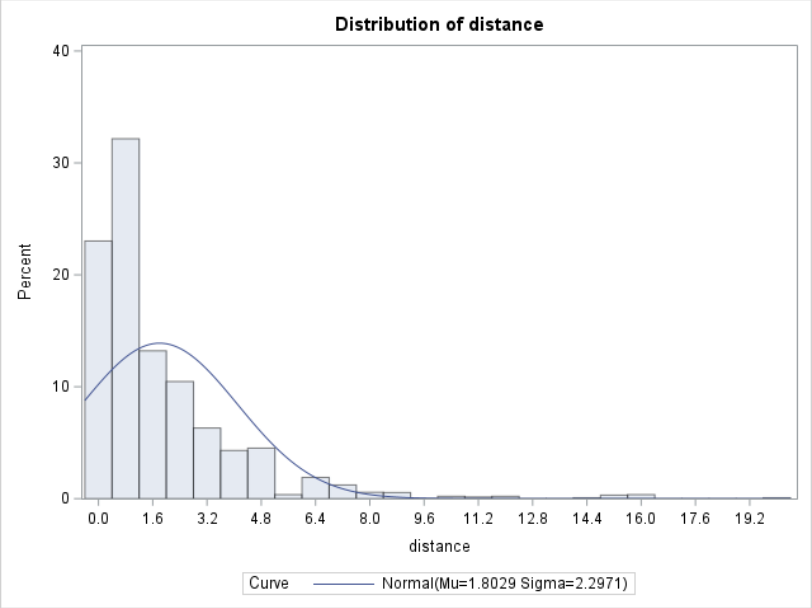

Figure 9: Distribution of variable "distance"

Measure	Values
Mean	1.802870
Median	1.000000
Mode	0.500000
St. Dev.	2.296885
Variance	5.275683
Min	0.000000
Max	20.000000
Range	20.000000
IQR	2.100000
Skew	2.999513

Figure 10: Location and variability metrics

- Distance from college in 10 miles is heavily right-tailed with a skewness of 3.00
- The mean distance is 1.8 (18 miles) and the median is 1.0 (10 miles)
- The standard deviation is 2.3 (23 miles)
- The lowest distance is 0.26 (2.6 miles) and the highest is 1.40 (14 miles)
- This distribution's right-skewness is due to potentially outlier values in the [8.5, 20] range
- Same as with "unemp", we will take the natural log of this feature

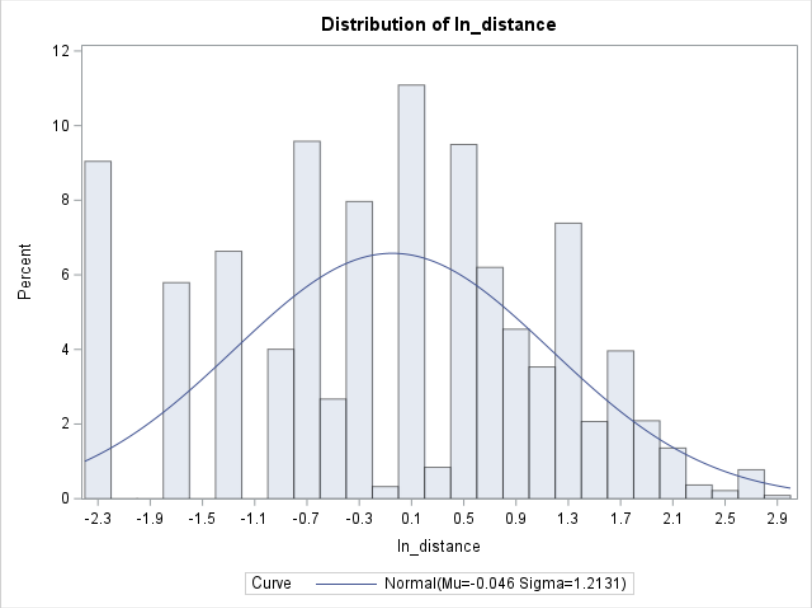

Figure 11: Distribution of variable "distance"

Measure	Values
Mean	-0.045369
Median	0.000000
Mode	0.000000
St. Dev.	1.200915
Variance	1.442198
Min	-2.302585
Max	2.995732
Range	5.298317
IQR	1.609438
Skew	-0.143513

Figure 12: Location and variability metrics

- The right-tailedness of the original variable is mitigated by taking its natural logarithm with a new skew of -0.14

4.6 Distribution of "tuition"

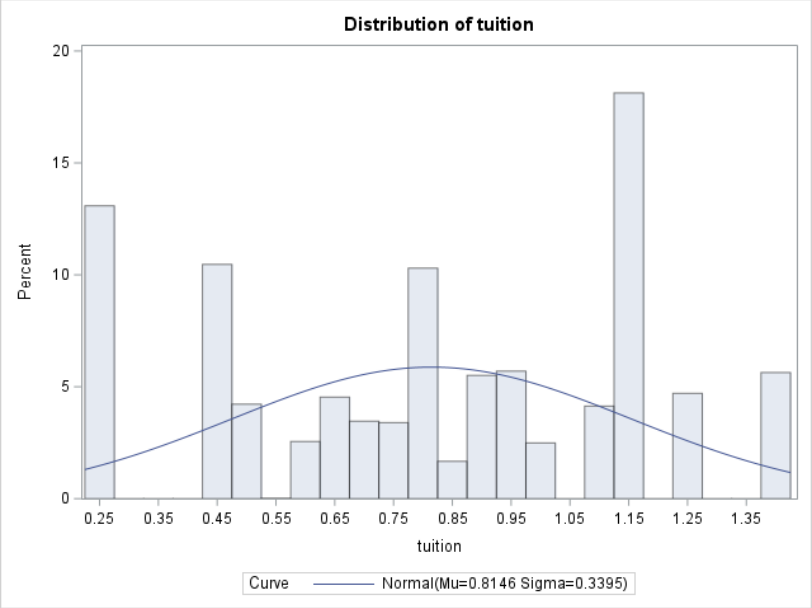

Figure 13: Distribution of variable "distance"

Measure	Values
Mean	0.814608
Median	0.824480
Mode	0.257510
St. Dev.	0.339468
Variance	0.115239
Min	0.257510
Max	1.404160
Range	1.146650
IQR	0.642030
Skew	-0.151913

Figure 14: Location and variability metrics

- Four year college tuition in $1000 is approximately normally distributed with a skewness of -0.15
- The mean tuition amount is 0.81 ($810) and the median is 0.82 ($820)
- The standard deviation is 0.34 ($340)
- The lowest tuition amount is 0.26 ($260) and the largest is 1.40 ($1,400)
- The tails of the distribution are "fatter" than a standard normal distribution, suggesting many records present towards the tails and potential outliers

4.7 Distribution of "score" by Gender

Scores by Gender

Overall Statistics					
Min	28.95	Mean	50.88903	Max	72.81
Pooled Std Dev	8.674816				

Statistics by Gender		
Min	30.91	28.95
Mean	50.25633	51.65808
Q1	43.385	44.77
Q2	50.235	52.02
Q3	56.87	58.82
Max	72.81	71.36
Range	41.9	42.41
Std Dev	8.482948	8.902488

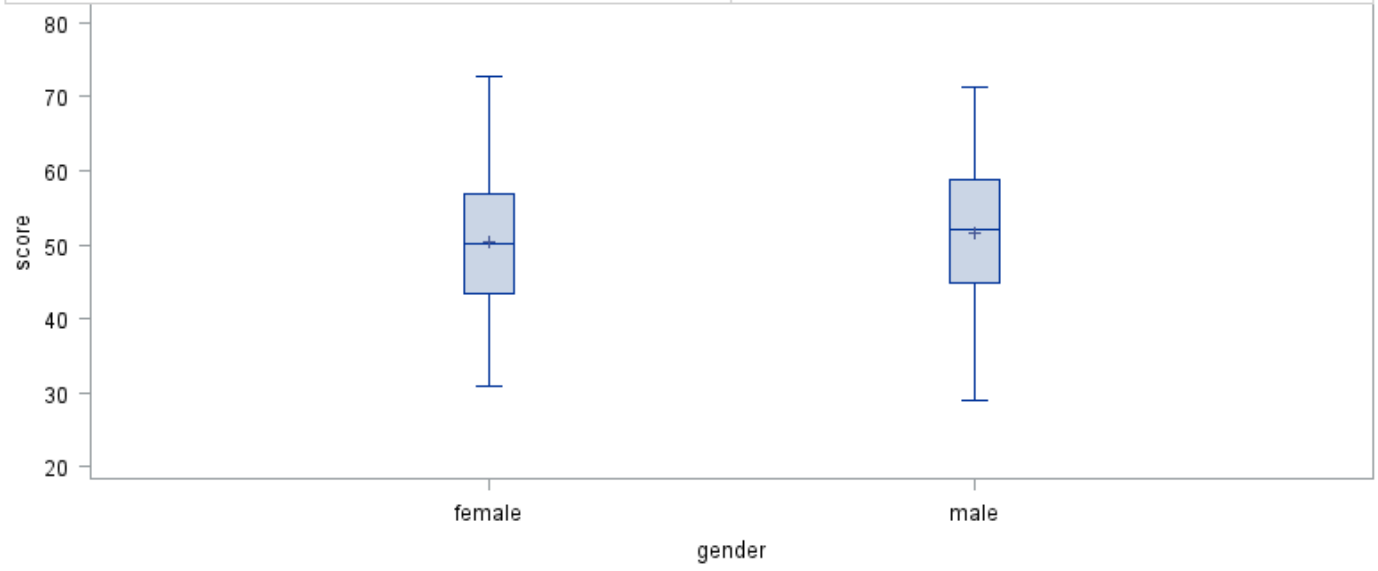

Figure 15: Distribution of student scores by gender

- The mean score is not significantly different based on gender, with males at 50.26 and females at 51.66
- The range of scores is also approximately equal, with males at 41.9 and females at 42.41
- Same with standard deviations, with males at 8.48 and females at 8.90

4.8 Distribution of "score" by Ethnicity

Scores by Ethnicity

Overall Statistics					
Min	28.95	Mean	50.88903	Max	72.81
Pooled Std Dev	8.105135				

Statistics by Ethnicity			
Min	28.95	32.1	31.05
Mean	45.34246	47.99839	53.17423
Q1	38.96	41.36	47.02
Q2	43.785	47.89	53.99
Q3	51.13	53.41	59.44
Max	69.96	68.86	72.81
Range	41.01	36.76	41.76
Std Dev	8.031027	7.811662	8.208713

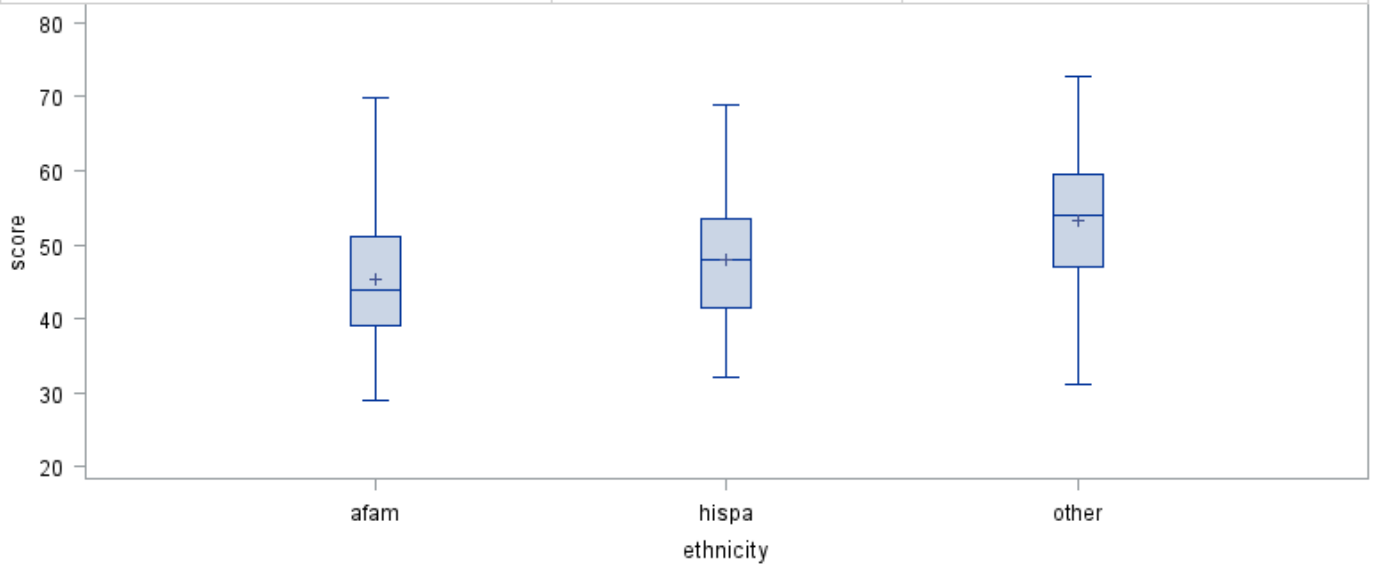

Figure 16: Distribution of student scores by ethnicity

- Average score is the highest for other ethnicities (53.17), followed by Hispanics (48) and African-Americans (45.34)
- The top 25% of students of other ethnicities have a score of 59.44, followed by Hispanics (53.41) and African-Americans (51.13)
- The standard deviation is largest for students of other ethnicities (8.21), followed by African-Americans (8.03) and Hispanics (7.81)

4.9 Distribution of "score" by Father's Graduation Status

Scores by Father Graduation Status

Overall Statistics					
Min	28.95	Mean	50.88903	Max	72.81
Pooled Std Dev	8.424293				

Statistics by Father Graduation Status		
Min	28.95	32.52
Mean	49.76975	55.14934
Q1	42.84	49.27
Q2	49.81	56.085
Q3	56.4	61.51
Max	72.81	71.36
Range	43.86	38.84
Std Dev	8.512329	8.08017

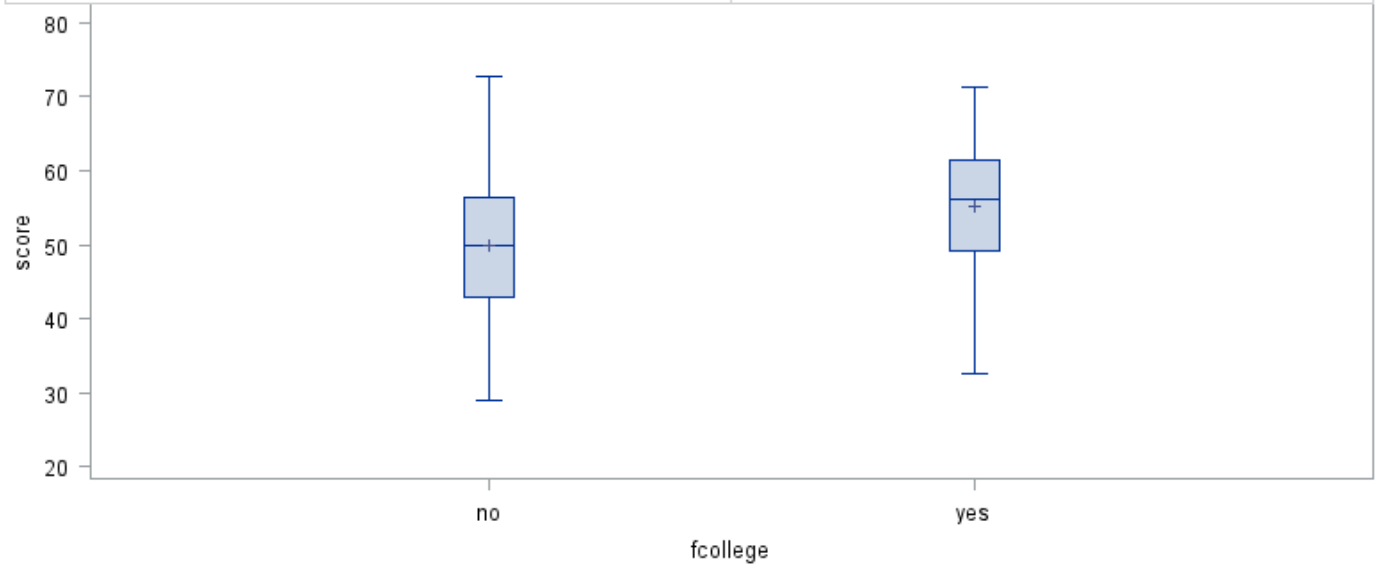

Figure 17: Distribution of student scores by father's graduation status

- Students have a higher average score when the father has graduated college (55.14) compared to not (49.77)
- Father's graduation status seems to have an effect on student performance
- Student scores tend to vary more for students with non-graduate parents (8.51) compared to graduates (8.08)

4.10 Distribution of "score" by Mother's Graduation Status

Scores by Mother Graduation Status

Overall Statistics					
Min	28.95	Mean	50.88903	Max	72.81
Pooled Std Dev	8.545033				

Statistics by Mother Graduation Status		
Min	28.95	32.14
Mean	50.23083	55.02224
Q1	43.35	49.02
Q2	50.32	55.98
Q3	56.97	61.51
Max	70.7	72.81
Range	41.75	40.67
Std Dev	8.549997	8.513754

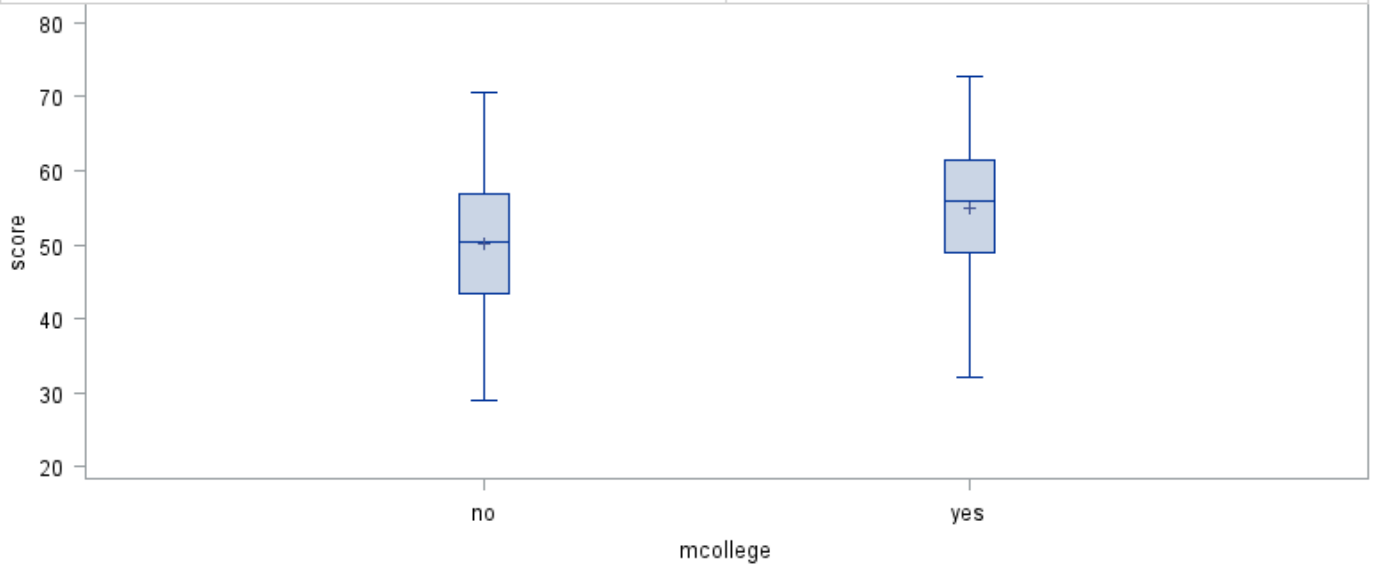

Figure 18: Distribution of student scores by mother's graduation status

- A similar pattern is present when it comes to the mother's graduation status
- Students have a higher average score when the mother has graduated college (55.02) compared to not (50.23)
- Mother's graduation status seems to have an effect on student performance

4.11 Distribution of "score" by Home Ownership

Scores by Home Ownership

Overall Statistics					
Min	28.95	Mean	50.88903	Max	72.81
Pooled Std Dev	8.633608				

Statistics by Home Ownership		
Min	30.78	28.95
Mean	48.54969	51.40179
Q1	40.805	44.78
Q2	48.115	51.66
Q3	55.735	58.15
Max	68.96	72.81
Range	38.18	43.86
Std Dev	8.99037	8.553494

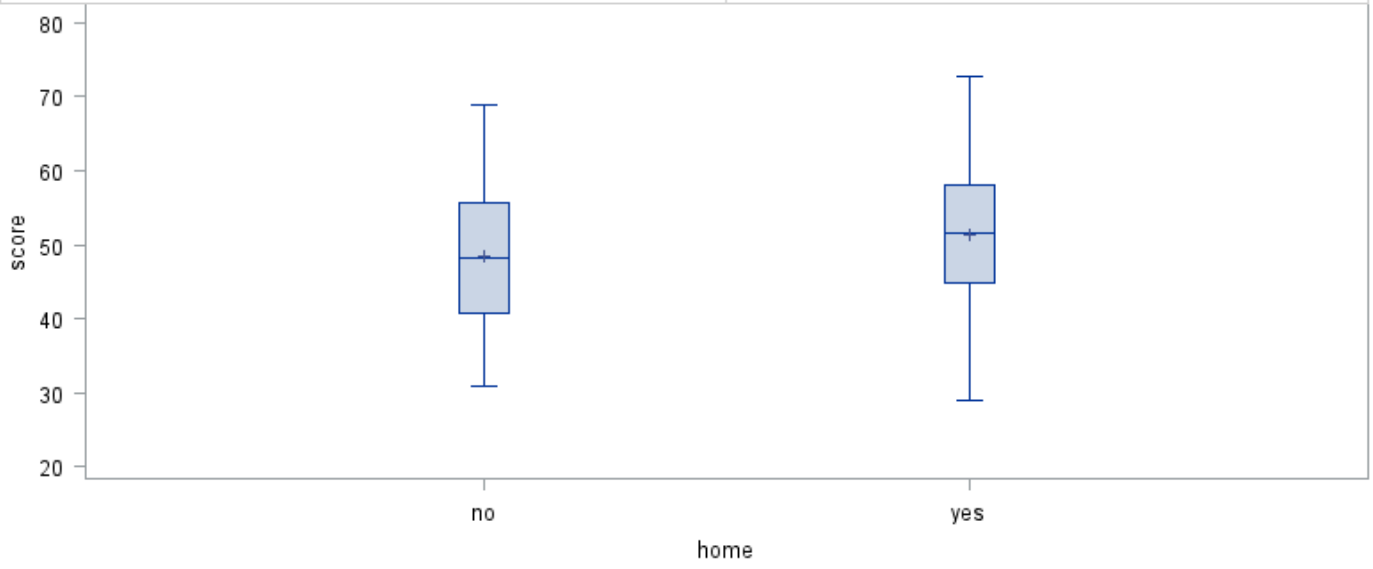

Figure 19: Distribution of student scores by home ownership

- Student's average score is slightly higher for students that own their homes (51.40) compared to students that do not (48.55)
- The range of scores is larger for students that own their homes (43.86) compared to students that do not (38.18)
- Student scores are more dispersed for students that do not own their homes (8.99) compared to students that do (8.55)

4.12 Distribution of "score" by School Location

Scores by Urban School Area

Overall Statistics					
Min	28.95	Mean	50.88903	Max	72.81
Pooled Std Dev	8.671228				

Statistics by Urban School Area		
Min	30.78	28.95
Mean	51.29728	49.54482
Q1	44.5	42.44
Q2	51.65	49.29
Q3	58.02	56.505
Max	72.81	71.36
Range	42.03	42.41
Std Dev	8.606787	8.880234

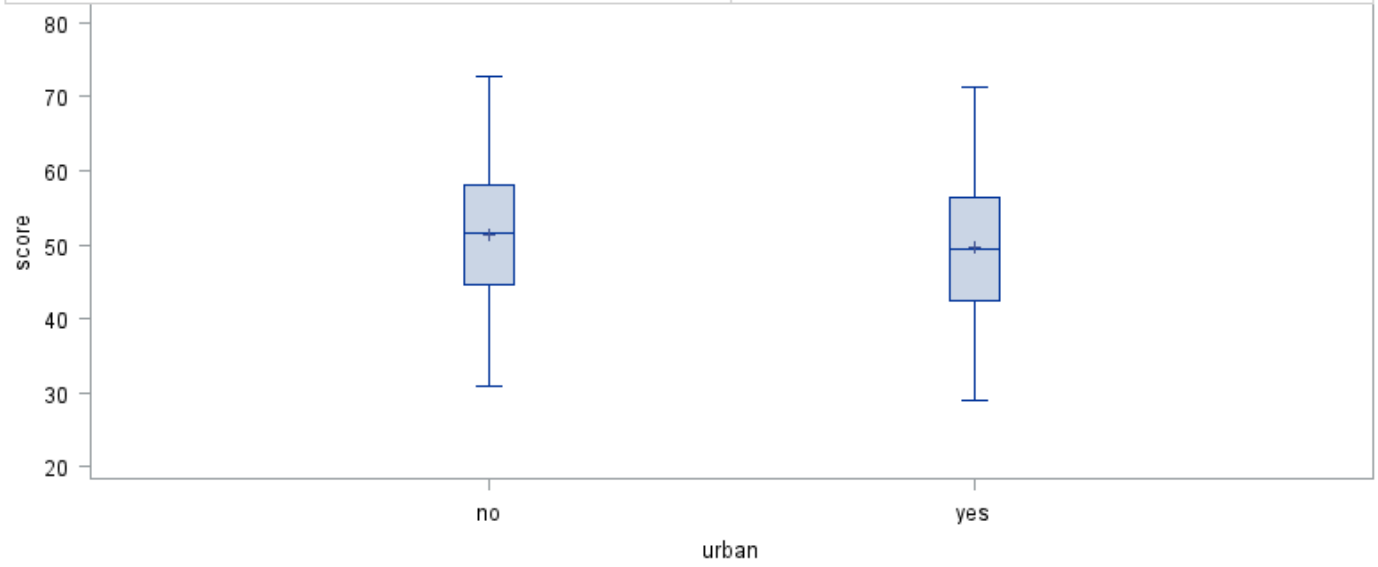

Figure 20: Distribution of student scores based on whether the school is classified as urban

- Students that attend non-urban schools tend to have slightly higher scores on average (51.30) compared to urban (49.54)
- Range of scores is similar for both non-urban (42.03) and urban (42.41) school students
- Standard deviations are also similar for non-urban (8.61) and urban (8.88) school students

4.13 Distribution of "score" by Income Category

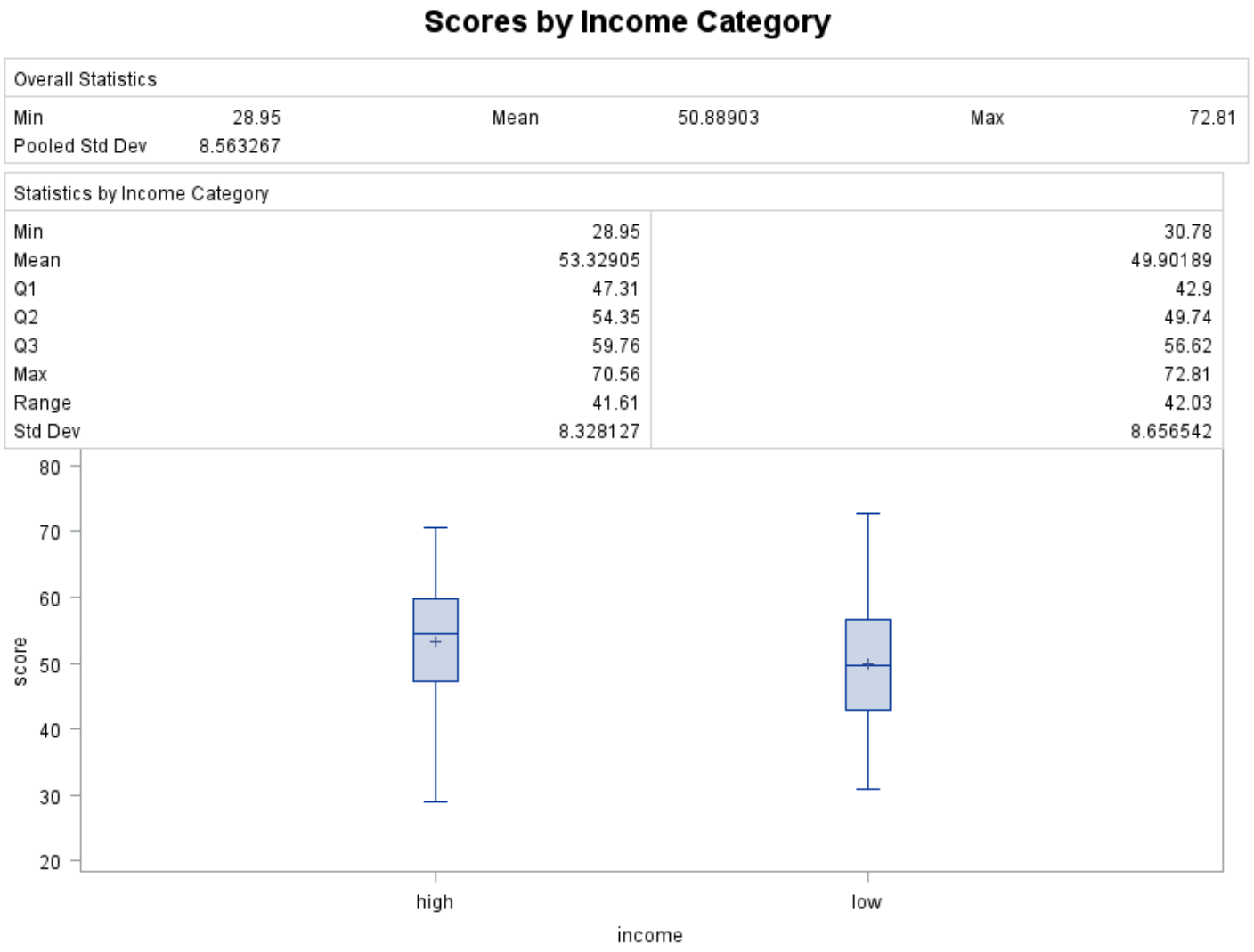

Figure 21: Distribution of student scores based on income category

- Students that belong in the high-income category tend to score slightly higher (53.33) on average than students in the low-income category (49.90)
- Student scores are less dispersed for high-income students (8.33) compared to low-income (8.66)

5 Building the Linear Regression Model

In this section, we will be creating two Linear Regression models using the adjusted R-Squared method. One will contain the previously discussed features in the dataset and the other will include interaction terms created from our features.

5.1 Model 1: Adj. R-Squared Method

The adjusted R-squared selection method is a technique used to assess and compare the performance of regression models. It is an extension of the traditional R-squared metric that takes into account the number of predictors in a model and adjusts for the degrees of freedom.

Adj- R-Squared Model

The REG Procedure
Model: MODEL1
Dependent Variable: score

Number of Observations Read	4739
Number of Observations Used	4645
Number of Observations with Missing Values	94

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	118299	9099.95629	181.58	<.0001
Error	4631	232090	50.11654		
Corrected Total	4644	350389			

Root MSE	7.07930	R-Square	0.3376
Dependent Mean	50.86258	Adj R-Sq	0.3358
Coeff Var	13.91849		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	22.61142	1.22402	18.47	<.0001	0
n_gender	1	1.12011	0.20942	5.35	<.0001	1.00692
n_ethnicity_hispanic	1	-3.64288	0.29287	-12.44	<.0001	1.23237
n_ethnicity_afam	1	-6.40597	0.30505	-21.00	<.0001	1.18153
n_fcollege	1	1.17342	0.29634	3.96	<.0001	1.32933
n_mcollege	1	1.03377	0.33904	3.05	0.0023	1.25398
n_home	1	0.79219	0.27677	2.86	0.0042	1.04094
n_urban	1	-0.65691	0.27076	-2.43	0.0153	1.20937
n_region	1	0.66473	0.32788	2.03	0.0427	1.58651
ln_unemp	1	-0.69385	0.32711	-2.12	0.0340	1.17796
ln_distance	1	-0.39033	0.09513	-4.10	<.0001	1.23422
wage	1	0.18382	0.08493	2.16	0.0305	1.21762
tuition	1	2.10226	0.42218	4.98	<.0001	1.89290
education	1	1.91055	0.06154	31.05	<.0001	1.12246

Figure 22: Linear Regression model with highest adj. R^2 value

$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$
H_a : At least one coefficient $\beta_j \neq 0$

- This model has a total of 13 features (not including score).
- All of our features are statistically significant (p-value <.05)
- Our F-Statistic is 181.58, meaning our overall model as a whole is statistically significant in explaining the variation in the response variable (p-value <.0001)
- The model's R^2 value is 0.3376 and adjusted R^2 is 0.3358
- Root Mean Square Error is 7.07930
- The feature with the strongest predicting power is education (t-value=31.05, p-value<.0001)

- There is no presence of strong feature collinearity, as all of the VIF scores are below 10

5.2 Model 2: Adj. R-Squared Method w/ Interaction Terms

After adding multiple interaction terms to the model, the only added feature with a statistically significant effect was gender and education (p-value <.05). The addition of this interaction term made n_region non-statistically significant, so it was removed from this model.

Model w/ Interaction Terms

The REG Procedure
Model: MODEL1
Dependent Variable: score

Number of Observations Read	4739
Number of Observations Used	4645
Number of Observations with Missing Values	94

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	118712	9131.72642	182.53	<.0001
Error	4631	231677	50.02736		
Corrected Total	4644	350389			

Root MSE	7.07300	R-Square	0.3388
Dependent Mean	50.86258	Adj R-Sq	0.3369
Coeff Var	13.90610		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	25.62291	1.43708	17.83	<.0001	0
n_gender	1	-4.54801	1.62344	-2.80	0.0051	60.61602
n_ethnicity_hispanic	1	-3.64071	0.29260	-12.44	<.0001	1.23237
n_ethnicity_afam	1	-6.47665	0.30326	-21.36	<.0001	1.16978
n_fcollege	1	1.17725	0.29572	3.98	<.0001	1.32618
n_mcollege	1	1.06954	0.33888	3.16	0.0016	1.25507
n_home	1	0.77973	0.27647	2.82	0.0048	1.04057
n_urban	1	-0.70337	0.26999	-2.61	0.0092	1.20465
ln_unemp	1	-0.64048	0.32497	-1.97	0.0488	1.16462
ln_distance	1	-0.39310	0.09500	-4.14	<.0001	1.23283
wage	1	0.19465	0.08449	2.30	0.0213	1.20708
tuition	1	1.59480	0.34614	4.61	<.0001	1.27472
education	1	1.71854	0.08154	21.08	<.0001	1.97412
n_gender_education	1	0.41019	0.11661	3.52	0.0004	61.66924

Figure 23: Linear Regression model with highest adj. R^2 value and added interaction terms

$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$
H_a : At least one coefficient $\beta_j \neq 0$

- This model has a total of 13 features (not including score)
- All of our features are statistically significant (p-value <.05)
- Our F-Statistic is 182.53 with a p-value <.0001, meaning our overall model as a whole is statistically significant in explaining the variation in the response variable (greater than previous model)
- The model's R^2 value is 0.3388 and adjusted R^2 is 0.3369 (greater than previous model)
- Root Mean Square Error is 7.0730 (less than previous model)
- The feature with the strongest predicting power is n_ethnicity_afam (t-value=-21.36, p-value <.0001)

- Although the interaction term has a VIF greater than 10, this is to be expected when including an interaction term in the model. In addition, it has statistical significance (p-value <0.05)

6 Final Model

After comparing both of our models on different metrics, our second model seems to be performing slightly better, with a larger F-Statistic, greater R^2 and adjusted R^2 and lesser Root Mean Square Error. We will further explore the performance of the model by analyzing its residuals.

6.1 Residual Analysis

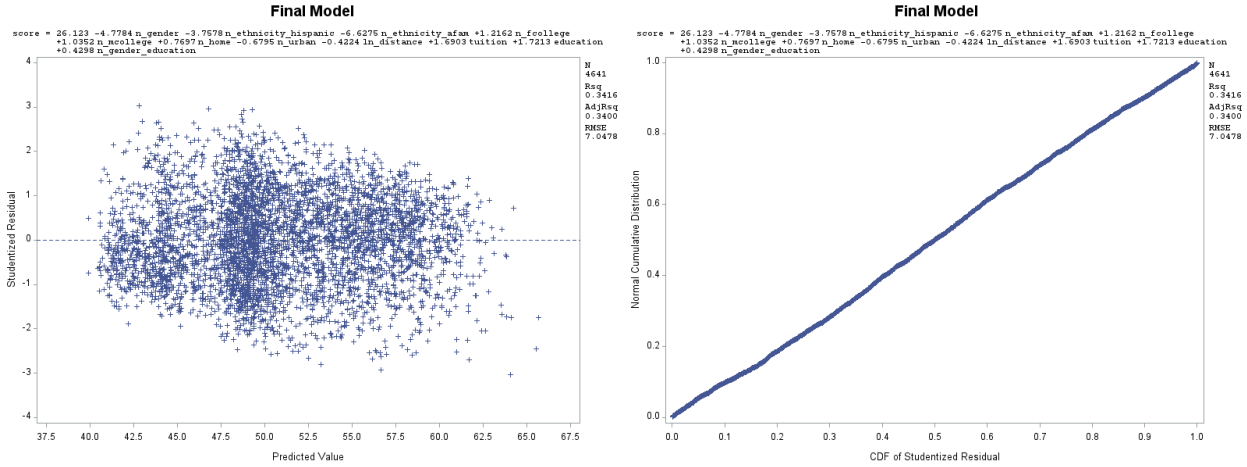

Figure 24: Studentized residuals for the final model

The studentized residual of the model is normally distributed when plotted against the Normal Cumulative Distribution. Also, the studentized residual for the predicted variable does not have any strong patterns present and is randomly distributed. However, some residuals are greater than 3 standard deviations away from the center. These points could be outliers that are influencing our model output. We will check for outliers and influential points in the dataset.

6.2 Outliers and Influential Points

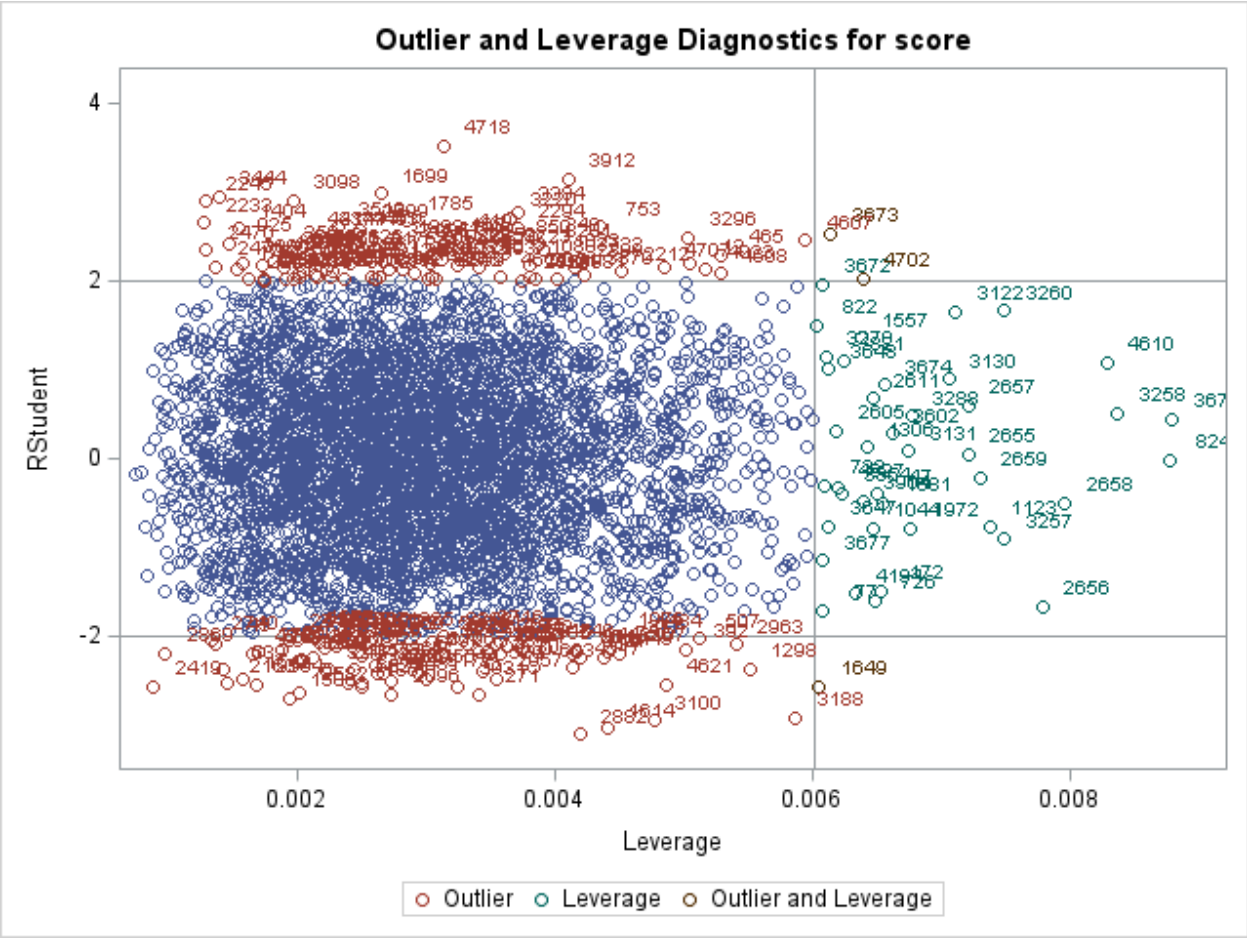

Figure 25: Outlier and leverage graph based on Student's R and Leverage

Our dataset contains quite a few records that have a studentized residual greater or lesser than 3 standard deviations from the center. In addition, a few data points have leverage greater than 0.006 (2p/n, p=14, n=4,739). These datapoints will be removed, leaving us with a total of 4,641 observations.

6.3 Final Model

Upon removing all outliers and influential points from the dataset, it was observed that the variables "wage" and "ln_unemp" exhibited no statistically significant relationship (p-value >.05). After removing these variables, we are left with 11 predictors.
$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$
H_a : At least one coefficient $\beta_j \neq 0$

- This model has a total of 11 features (not including score)
- All of our features are statistically significant (p-value <.05)

The REG Procedure
Model: MODEL1
Dependent Variable: score

Number of Observations Read	4641
Number of Observations Used	4641

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	119301	10846	218.34	<.0001
Error	4629	229931	49.67188		
Corrected Total	4640	349232			

Root MSE	7.04783	R-Square	0.3416
Dependent Mean	50.86080	Adj R-Sq	0.3400
Coeff Var	13.85709		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	26.12307	1.17467	22.24	<.0001	0
n_gender	1	-4.77839	1.61827	-2.95	0.0032	60.60326
n_ethnicity_hispanic	1	-3.75783	0.28808	-13.04	<.0001	1.20288
n_ethnicity_afam	1	-6.62754	0.29866	-22.19	<.0001	1.14129
n_fcollege	1	1.21624	0.29410	4.14	<.0001	1.32073
n_mcollege	1	1.03523	0.33773	3.07	0.0022	1.25360
n_home	1	0.76973	0.27530	2.80	0.0052	1.03795
n_urban	1	-0.67949	0.26904	-2.53	0.0116	1.20361
ln_distance	1	-0.42244	0.09363	-4.51	<.0001	1.20433
tuition	1	1.69031	0.32288	5.24	<.0001	1.11606
education	1	1.72125	0.08125	21.19	<.0001	1.97139
n_gender_education	1	0.42977	0.11625	3.70	0.0002	61.64743

Figure 26: Linear Regression model with highest adj. R^2 value and added interaction terms

- Our F-Statistic is 218.34 with a p-value <.0001, much greater than any previous model
- The model's R^2 value is 0.3416 and adjusted R^2 is 0.34, which is also greater than the previous models
- Root Mean Square Error is 7.04783
- The feature with the strongest predicting power is n_ethnicity_afam (t-value=-22.19, p-value <.0001)

6.4 Parameter Estimates

- n_gender: Student score decreases by 4.77% when the student is male
- n_ethnicity_hispanic: Student score decreases by 3.76% when the student is Hispanic compared to other students (control group=n_ethnicity_other)
- n_ethnicity_afam: Student score decreases by 6.63% when the student is African American compared to other students
- n_fcollege: Student scores increases by 1.22% when the father is a graduate
- n_mcollege: Student score increases by 1.04% when the mother is a graduate
- n_home: Student score increases by 0.77% when their family owns their home
- n_urban: Student score decreases by 0.68% when the school is in an urban area
- ln_distance: Student score decreases by (exp(-0.42)-1)*100 = 34% for each additional 10 miles in distance (1 unit=10 miles)

- tuition: Student score increases by 1.69% for every $1000 in tuition
- education: Student score increases by 0.43% for each additional year of education when the gender is female (n_gender=0)
- n_gender_education: Student score increases by (1.72+0.43 =) 2.5% for each additional year of education when the gender is male (n_gender=1)

7 Training and Testing

Our dataset will be split into 80% training and 20% testing randomly.

7.1 Training Results

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Number of Observations Read	4641
Number of Observations Used	3713
Number of Observations with Missing Values	928

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	96093	8735.74566	177.87	<.0001
Error	3701	181766	49.11278		
Corrected Total	3712	277860			

Root MSE	7.00805	R-Square	0.3458
Dependent Mean	50.93479	Adj R-Sq	0.3439
Coeff Var	13.75887		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	26.91073	1.30709	20.59	<.0001
n_gender	1	-5.49904	1.80205	-3.05	0.0023
n_ethnicity_hispanic	1	-3.93086	0.32012	-12.28	<.0001
n_ethnicity_afam	1	-6.68088	0.33371	-20.02	<.0001
n_fcollege	1	1.32230	0.32958	4.01	<.0001
n_mcollege	1	1.02021	0.37783	2.70	0.0070
n_home	1	0.77519	0.30642	2.53	0.0115
n_urban	1	-0.60327	0.29764	-2.03	0.0428
ln_distance	1	-0.39660	0.10398	-3.81	0.0001
tuition	1	2.04264	0.35933	5.68	<.0001
education	1	1.64287	0.09051	18.15	<.0001
n_gender_education	1	0.49191	0.12938	3.80	0.0001

Figure 27: Linear Regression model model results on training set

- Our F-Statistic for the training data is 177.87 with a p-value <.0001
- The model's R^2 value is 0.3458 and adjusted R^2 is 0.3439
- Root Mean Squared Error is 7.00805

7.2 Testing Results

Model Summary

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	928	7.21779	5.95339

Model Summary

The CORR Procedure

2 Variables:	score phat

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
score	928	50.56473	8.76830	46924	31.70000	69.32000	
phat	928	50.83265	5.05600	47173	40.67425	64.07653	Predicted Value of new_y

Pearson Correlation Coefficients, N = 928
Prob > |r| under H0: Rho=0

	score	phat
score	1.00000	0.56786 <.0001
phat Predicted Value of new_y	0.56786 <.0001	1.00000

Figure 28: Linear Regression model model results on testing set

- Root Mean Squared Error= 7.21779
- $R^2 = \hat{p}^2 = 0.32246498$
- CV $R^2 = 0.3458 - 0.32246498 = 0.02333502$

Our model is performing equally well on the testing set as on the training set. RMSE and R^2 do not wildly differ from each other in both cases, meaning the model is not overfitting the training data.

7.3 Testing Model on New Records

After training and testing the model on the previous dataset, we will create some new records and see how the model performs on previously unseen data.

Obs	n_gender	n_ethnicity_hispanic	n_ethnicity_afam	n_fcollege	n_mcollege	n_home	n_urban	ln_distance	tuition	education	n_gender_education
1	1	1	0	1	0	1	0	-0.92000	0.86	11	12
2	0	0	1	0	0	1	1	1.02000	0.72	10	0
3	1	0	1	0	0	0	0	1.30000	0.68	12	12

Figure 29: New records manually created to test the model

Testing Model

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	45.6744	0.5128	44.6690	46.6797	31.8976	59.4511	.
2	.	37.8967	0.5139	36.8891	38.9042	24.1197	51.6736	.
3	.	41.2216	0.4571	40.3253	42.1179	27.4524	54.9908	.

Figure 30: Testing model performance on new unseen records

- Obs. 1: Predicted= 45.6744, CI= [31.8976, 59.4511]
- Obs. 2: Predicted= 37.8967, CI= [24.1197, 51.6736]
- Obs. 3: Predicted= 41.2216, CI= [27.4524, 54.9908]