Modeling Earnings with Regression

COMP 4441
Vanessa Cox & Gayla Hess
Autumn 2023

Research Question

How are earnings affected by alma mater?

Background

Our analysis will determine which variables most influenced a student's future earnings. The data set obtained from DASL provides information from 706 private and public institutions. In addition to school name and location, standardized test scores (ACT and SAT), price, price with aid, percent need, merit aided and category of school were included.

Importance of data and context within society

Future earning potential is important to students as they consider educational opportunities. The earnings outcome based on institution type is relevant to society for studies of economics (classes and how each relates monetarily), trends (e.g., voting statistics), and relates to recent changes related to college admission standards.

Approach

Data Cleaning

We removed the variables school and place during the beginning data cleaning steps of our project. We got rid of the state variable as well and replaced it with 1 of 4 U.S. geographic regions: West, South, Midwest, or Northeast. The variables school, place, and even state, created too many unique factors to statistically assess for earning disparities. SAT and ACT scores were found to be collinear. ACT was kept while SAT and SAT.ACT were not needed in the analysis.

Comparison of private vs public institutions and explanation of statistical test chosen

After running multiple visualizations and tests on the private and public earnings data, it seems neither are normally distributed nor their variances equal. However, we have 268 observations of public school earnings and 438 observations of private school earnings, so running a Welch's t-test is acceptable. The Welch Two Sample t-test showed a p-value of 0.7174 and we failed to reject the H_0 of $\mu_1 = \mu_2$.

Power Analysis

We performed a power analysis of the statistical method we used to compare public/private earnings. The values for n_1 and n_2 are the sizes of the two populations, private (n_1=438) and public (n_2=268). The effect size was calculated using a pooled standard deviation and was double checked using the cohensD() function (d=0.02671961) and significance level set at p=0.05. We then calculated the power to be 0.06366854. This interprets as having a 6% chance of rejecting the null hypothesis when the null hypothesis is false or a 94% chance of a Type II error. After reviewing the density plot of earnings and the public variable, we have concluded the low power calculation is due to the distributions of both populations being overlaid.

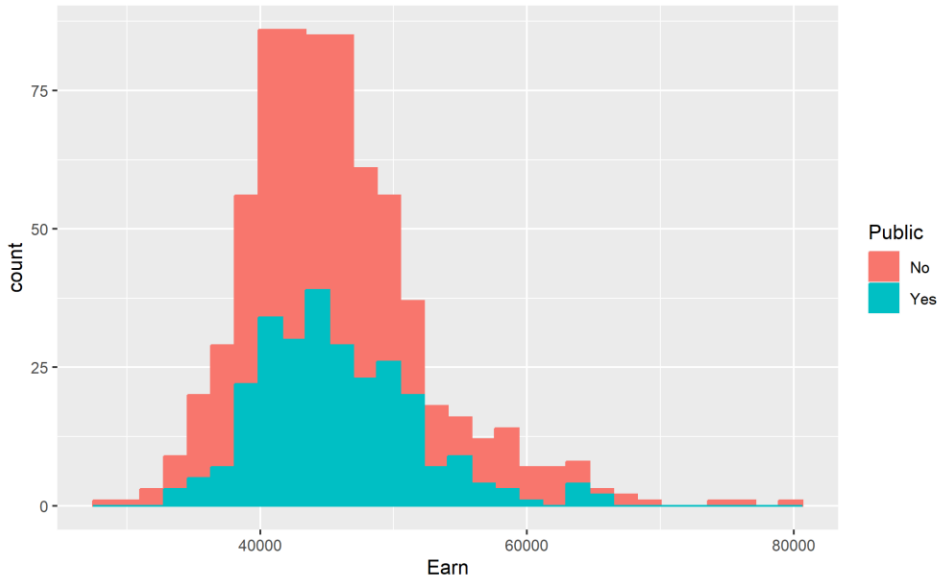

Figure 1: Histogram of count vs earnings of Private and Public Institutions

Why regression is more powerful than simply comparing two populations

Regression models are more powerful than simply comparing two populations in that they create predictive models, control for confounding variables, and can describe complex relationships.

Exploring the data with Exploratory Graphs

Scatter plots (with and without a regression line), box plots, histograms, density plots and qq plots were created to view the data. Numeric independent variables appear linearly related to earnings. Relationships between both numerical and non-numerical data were also examined. Looking at the box plot of Regions vs. Earnings the means seem about the same, however the variances differ by region with the most variation appearing in the West.

Explanation of a few Model Building Methods for Regression

Stepwise Regression:

Stepwise regression is a method of fitting regression models by considering each variable to be added or removed from the model based on a modeler-chosen metric. It does not consider all possible models, and it produces a single regression model when the algorithm ends. Within stepwise regression, you can specifically do forward, backward, or both-direction stepwise selection.

Forward Stepwise Selection:

1. Begins with a model that contains no variables (null model)

2. Then starts adding most significant variables one after the other.

3. Stops once pre-specified stopping rule is reached or until all the variables under consideration are included in the model

Backward Stepwise Selection:

1. Begins with a model that contains all variables under consideration (full model)

2. Then starts removing the least significant variables one after the other

3. Stops once pre-specified stopping rule is reached or until no variable is left in the model

Both-Direction Stepwise Regression

1. Begins with null model.

2. Add predictors to model sequentially (same as forward stepwise selection). However, after adding each predictor, predictors are removed if they are no longer significant (like backward stepwise selection).

3. Stops once pre-specified stopping rule is reached or until no variable is left in the model

Best Subsets Regression

Best subsets regression is also known as “all possible regressions” and “all possible models.” The best subsets regression fits all models based on the independent variables specified. Specifically, it fits all models with every combination and number of independent variables possible. After fitting all the models, the best subsets regression displays the best fitting models for every number of independent variables. For example, we have 6 (if we exclude “Public”) independent variables, or predictors, in our dataset, so it would display the best fitting 1-predictor, 2-predictor, 3-predictor, 4-predictor, 5-predictor, and 6-predictor models. Usually, either adjusted R-squared or Mallows’ Cp is the criterion for picking the best fitting models for this process, however we will be using AIC.

Two-way interaction terms

Regression models take independent variables and establish a relationship with dependent variables which helps create predictive models. A simple linear regression can be seen as $y = b_0 + b_1x_1$ where "b_1" represents how much of an effect the independent variable "x_1" has on the outcome and in what direction (+/-). Having multiple independent variables would look like $y = b_0 + b_1x_1 + b_2x_2 + b_3x_3$ and so on. However, independent variables do not always act independently of each other, and that interaction can create a significant impact on the outcome. This is modeled as $y = b_0 + b_1x_1 + b_2x_2 + b_3 (x_1*x_2)$ where the multiplicative variable is called the interaction term. Specifically, in this case, it is a two-way interaction term because there are two independent variables, but there are higher ordered interaction terms with more than two.

Addressing transformed variables

After log, root, and inverse transforming the independent and dependent variables separately and assessing for a significant linear relationship, we found none of the transformations contributed to a more significant linear relationship compared to the un-transformed variables.

Building the best model to predict earnings (without the variables school, place, or public)

Forward Stepwise Regression Using P-Values

The regression method we used was with the forward stepwise regression, where independent variables were selected for the model if they had a $p \leq 0.05$. The significant variables selected by this model were: ACT, Pct.need, and Region. This following is the summary table from the forward stepwise regression:

Selection Summary

Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	ACT	0.3035	0.3025	182.4163	14197.7617	5615.5940
2	Pct.need	0.3602	0.3584	110.4029	13760.2325	5389.2015
3	Region	0.4038	0.3994	58.4966	13717.7676	5213.8688

Backward Stepwise Regression Using P-Values

We next created a model using backward stepwise regression, where independent variables were removed from the model based on $p > 0.05$. The following is a summary of the eliminated variables:

Elimination Summary

Step	Variable Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	Price.with.aid	0.3613	0.3542	4.1280	12750.8086	5010.6905
2	merit.aided	0.4052	0.3999	58.8013	13718.1865	5211.7006
3	Price	0.4038	0.3994	58.4966	13717.7676	5213.8688

Note the three variables left after they were removed were the same variables produced as the forward stepwise regression.

Stepwise Regression Using P-Values

Next, we created a linear model using both-direction stepwise regression where variables were both added and eliminated based on $p = 0.05$. Variables were added if $p \leq 0.05$ and removed if $p > 0.05$. This accounts for the changes in p-values that occur as the model is being built. For example, an independent variable may have a significantly low p-value to start, but as the model changes it may become probabilistically insignificant. The following is the both-direction stepwise regression summary:

Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	ACT	addition	0.303	0.302	182.4160	14197.7617	5615.5940
2	Pct.need	addition	0.360	0.358	110.4030	13760.2325	5389.2015
3	Region	addition	0.404	0.399	58.4970	13717.7676	5213.8688

Stepwise (Forward, Backward, Both-Direction) Model Analysis

The variables selected by all three methods, forward, backward, and both-direction, based on $p = 0.05$, were the same. This resulted in the same linear model for all three methods. The linear model is:

$$Earn = 27756.61 + 887.67(ACT) - 94.47(Pct.need) + 2174.85(RegionNortheast) - 742.74(RegionSouth) + 2910.92(RegionWest)$$

Best Subsets Regression

For our final model building method, we used best subsets regression. This model produces the best models for all independent variables based on the metric chosen by the modeler, in this case we chose AIC. The Akaike Information Criterion (AIC) is a mathematical method for evaluating how well a model fits the data it was generated from. If a model is more than 2 AIC units lower than another, then it is considered significantly better than that model. The lowest AIC in the best subsets regression for our model is the 6-predictor model ($AIC = 12751.6655$) . However, this is a model where we keep all independent variables and is overfitted. The next two best models are the 3-predictor (Pct.need, ACT, Region) and 5-predictor (Price, Price.with.aid, Pct.need, ACT, Region) models. The AIC's for both these models are close enough to be considered about the same in terms of being the best model. We will talk more about this when the final model is discussed. The following is our best subsets regression summary table:

Best Subsets Regression

Model Index	Predictors
1	ACT
2	Pct.need ACT
3	Pct.need ACT Region
4	Price Pct.need ACT Region
5	Price Price.with.aid Pct.need ACT Region
6	Price Price.with.aid Pct.need merit.aided ACT Region

Subsets Regression Summary

Model	R-Square	Adj. R-Square	Pred R-Square	C(p)	AIC	SBIC	SBC	MSEP	FPE	HSP	APC
1	0.3035	0.3025	0.2985	182.4163	14197.7617	NA	14211.4406	22263636917.6925	31624230.1074	44857.6047	0.7005
2	0.3602	0.3584	0.3536	110.4029	13760.2325	NA	13778.3619	19952922707.2545	29170320.6579	42523.4161	0.6454
3	0.4038	0.3994	0.3923	58.4966	13717.7676	NA	13749.4939	18621094821.5497	27342707.2102	39859.8653	0.6032
4	0.4052	0.3999	0.3918	58.8013	13718.1865	NA	13754.4452	18605570597.1528	27359507.6333	39885.2031	0.6035
5	0.4087	0.4026	0.3933	56.4188	13716.0825	NA	13756.8735	18521952860.4053	27275965.8400	39764.4269	0.6017
6	0.3624	0.3544	0.3423	5.0000	12751.6655	NA	12796.2958	16039784512.0352	25376065.1593	39655.5150	0.6517

How many different models (without interaction and transformation) can be produced?

$$\sum_{i=1}^{7} \binom{n}{k_i} = \frac{7!}{1!(7-1)!} + \frac{7!}{2!(7-2)!} + \frac{7!}{3!(7-3)!} + \frac{7!}{4!(7-4)!} + \frac{7!}{5!(7-5)!} + \frac{7!}{6!(7-6)!} + \frac{7!}{7!(7-7)!} = 127$$

Based on the 7 variables (we are including the variable “Public” here), there are 127 possible models.

Probability at least one variable not important for regression has been included in the model:

$$(1 - 0.7012) \cdot (1 - 2e^{-16}) \cdot (1 - 3.5e^{-5}) \cdot (1 - 9.38e^{-5}) \cdot (1 - 0.0254) \cdot (1 - 2.69e^{-5}) \cdot (1 - 2.07e^{-7}) = 0.26105$$

$$1 - 0.26105 = 0.73895$$

There is a 73.895% chance at least one variable included in the regression is not important.

Select a final model and assess the model assumptions

Based on the best subsets regression, we have two models to choose from. We have the best 3-predictor model that contains the independent variables Pct.need, ACT, and Region, and the best 5-predictor model that contains the independent variables Price, Price.with.aid, Pct.need, ACT, Region. After cross referencing the two models given by the best subsets regression, and the three models given by the forward stepwise, backward stepwise, and combined stepwise regression, we have concluded the best

model for our data is the 3-predictor model with the variables Pct.need, ACT, and Region. The model has an $AIC = 13717.77$ and is:

$$Earn = 27756.61 + 887.67(ACT) - 94.47(Pct.need) + 2174.85(RegionNortheast) - 742.74(RegionSouth) + 2910.92(RegionWest)$$

Furthermore, after assessing the significance of potential two-way interaction terms, we added Pct.need*ACT to our model. This model has an $AIC = 13692.5$ and is:

$$Earn = 27756.61 + 887.67(ACT) - 94.47(Pct.need) + 2174.85(RegionNortheast) - 742.74(RegionSouth) + 2910.92(RegionWest) - 19.83(ACT * Pct.need)$$

The model assumptions for linear regression are linearity, independence, homoscedasticity, and normality. In our exploratory period and experimentation with transformations, we deemed our independent and dependent variables to be linearly related. The assumption of independence comes from an understanding of our data, and after we removed the variable SAT from our data as it was collinear with ACT, we agree the assumption of independence is met. Furthermore, an analysis of the residuals concluded the assumption of homoscedasticity is met. However, the analysis of residuals concluded the assumption of normality is not met; and additionally, transformations did not aid in meeting the assumption of normality. Considering the number of observations in our data set, we are confident the legitimacy of our model is not compromised by the assumption of normality being unmet.

Adding the variable public & assessing its impact and importance to the final regression model

The following is our final model including the public variable:

```
Call:
lm(formula = Earn ~ +ACT + Pct.need + Region + Pct.need:ACT +
    Public, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-12566.2  -3239.6   -399.1   2772.8  23052.1

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -3663.71    5732.41  -0.639   0.5230
ACT              2100.97     219.24   9.583  < 2e-16 ***
Pct.need          428.48      93.10   4.602 4.99e-06 ***
RegionNortheast  1992.05     496.44   4.013 6.67e-05 ***
RegionSouth     -1366.11     575.79  -2.373   0.0179 *
RegionWest       2629.94     627.30   4.192 3.12e-05 ***
PublicYes        1382.28     545.63   2.533   0.0115 *
ACT:Pct.need      -20.69       3.78  -5.474 6.20e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5095 on 679 degrees of freedom
  (19 observations deleted due to missingness)
Multiple R-squared:  0.4324,    Adjusted R-squared:  0.4265
F-statistic: 73.89 on 7 and 679 DF,  p-value: < 2.2e-16
```

Here, we can see the public variable has a statistically significant influence on earnings based on the p-value ($p = 0.0115$). Furthermore, this model has an $AIC = 13688.04$ which is significantly lower than the previous model.

Explain the logic of adding the variable public to the regression model as the last step

Adding a variable last ensures it is statistically the hardest for the variable to be significant. We first wanted to assess all other variables to explain the difference in earnings, and then reassess with the public variable included in the regression.

Conclusion

Our model found earnings can be modeled with the following equation:

$$\begin{aligned} Earnings = & -3663.71 + 2100.97(ACT) + 428.48(Pct.need) + 1992.05(RegionNortheast) \\ & - 1366.11(RegionSouth) + 2629.94(RegionWest) + 1382.28(PublicYes) \\ & - 20.69(ACT * Pct.need) \end{aligned}$$

Circling back to the t-test we performed at the beginning of our analysis, we first deemed the difference in earnings between public and private school students to be statistically insignificant. However, whether a university is private or public plays a significant role in determining earnings according to our linear model. Specifically, attending a private school corresponds to an average of $1,382.28 (95% CI: $310.94-2453.61) increase in earnings compared to private school attendees. This exemplifies the statistical power of regression analyses and the importance of understanding your data and the corresponding statistical tests you run on it.

The location of the school also plays a role in earnings with the Midwest region as the default variable in this model. Schools in the South will see graduates have an increased income of $1,366.11 on average while those in the West will earn $2,629.94 more on average. Additionally, ACT scores, identified as a significant variable early on in analysis, have the greatest impact. For every point scored on the ACT, graduates earn an average of $2,100.97 (95% CI $1670.50-2531.43) more.

The adjusted R^2 of our final model is 0.4265 indicating 42.65% of the variability in earnings is explained by our model. A more complete analysis could be done if the dataset included lurking variables such as degree obtained, GPA, internship experience, legacy status, work experience, etc. Even though the R^2 indicates a low explanation of variability, we believe we have the strongest possible model given the data at hand.