Researching Participant Engagement and Success in Harvard University's edX Courses

CS555 Project

SiCheng Yi, HaoYu Gong, YiMeng Wang

Content

Research Scenario and Questions:

Scenario: The dataset provided comes from a study of online courses offered by Harvard and MIT through the edX platform since 2012. This dataset includes various attributes related to these courses, such as the number of participants, course content, and completion rates.

Research Questions:

Perform a simple linear regression with Participants and Certified of > 50% Course Content Accessed, and briefly summarize your conclusions?

Calculate the least squares regression equation that predictsParticipants (Course Content Accessed) from `Audited (> 50% Course Content Accessed)`, Certified, `% Audited`,`Total Course Hours (Thousands)`. Formally test (using the 5-step procedure) whether the set of these predictors is associated with a prestige score at the $\alpha = 0.05$ level?

Assess whether the model assumptions are met.

Are there significant outliers or influential points in the dataset that could skew the results of the analysis?

Description of the Data Set:

The dataset comprises information on 290 online courses from Harvard and MIT, including:

Institution: The provider of the course (MIT or Harvard).

Course Number: A unique identifier for each course.

Launch Date: The start date of the course.

Course Title: The name of the course.

Instructors: Names of the course instructors.

Course Subject: The subject area of the course.

Year: The last year the course was offered.

Honor Code Certificates: Indicates if the course offers honor code certificates (1 for yes, 0 for no).

Participants (Course Content Accessed): Number of participants who accessed the course.

Audited (> 50% Course Content Accessed): Number of participants who audited more than 50% of the course.

Data Cleaning Performed:

The dataset was filtered to remove courses provided by MIT.

The percentage of played video data was converted to numeric format and non-numeric characters were removed.

Missing values (NA) were omitted to clean the data.

Link: *https://www.kaggle.com/datasets/edx/course-study/*

Statistic Method

(1) Linear Regression Analysis:

Firstly, we applied simple linear regression analysis to investigate the relationship between "Participants" and "Certified" for courses with more than 50% of course content accessed. We used the lm function to fit the regression model.

Then, we performed multiple linear regression analysis to predict "Participants (Course Content Accessed)" based on several predictor variables, including "Audited (> 50% Course Content Accessed)", "Certified", "% Audited", and "Total Course Hours (Thousands)". We used the lm function to fit the multiple linear regression model.

(2) Outlier Detection:

In the regression analysis, we conducted outlier detection. We used methods such as Q-Q plots (qqplot) and Cook's distance (Cooks's Distance) to identify and detect outliers.

(3) Significance Testing:

We utilized ANOVA (Analysis of Variance) to perform variance analysis to determine if the various predictor variables in the multiple linear regression model are associated with the prestige score. We also computed confidence intervals using the confint function.

In summary, our analysis methods include linear regression, outlier detection, and significance testing to explore relationships within the data and determine if there are statistically significant findings.

Results Report

A. Establish model1

We initially attempted to construct our first model, where 'Participants (Course Content Accessed)' is the dependent variable, and all other variables are independent variables.

Below is the summary of the model1. (model1)

```
Call:
lm(formula = `Participants (Course Content Accessed)` ~ ., data = train)

Residuals:
   Min     1Q Median     3Q    Max
-56733  -4306    530   4234  41899

Coefficients: (1 not defined because of singularities)
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                            19556.6124 21324.9570   0.917 0.361609
`Audited (> 50% Course Content Accessed)`   10.0804     0.5988  16.834  < 2e-16 ***
Certified                                  -17.4582     2.3251  -7.509 4.62e-11 ***
`% Audited`                               -450.3326   120.8659  -3.726 0.000343 ***
`% Certified`                              295.9613   345.5626   0.856 0.394068
`% Certified of > 50% Course Content Accessed`   218.7686   148.0256   1.478 0.143002
`% Played Video`                            -8.5620   110.4550  -0.078 0.938389
`% Posted in Forum`                         22.8677   184.1100   0.124 0.901435
`% Grade Higher Than Zero`                 -69.5678   134.5855  -0.517 0.606520
`Total Course Hours (Thousands)`            53.3509    15.4875   3.445 0.000878 ***
`Median Hours for Certification`          -122.5625    90.7062  -1.351 0.180094
`Median Age`                               -10.0060   433.4703  -0.023 0.981636
`% Male`                                   118.9761   130.2097   0.914 0.363357
`% Female`                                      NA         NA      NA       NA
`% Bachelor's Degree or Higher`            -247.5471   188.2308  -1.315 0.191885
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10830 on 88 degrees of freedom
Multiple R-squared:  0.9404,    Adjusted R-squared:  0.9316
F-statistic: 106.8 on 13 and 88 DF,  p-value: < 2.2e-16
```

Interpretation of Coefficients:

Intercept: The estimated value for 'Participants' when all other independent variables are zero is approximately 19556.61.

Audited (> 50% Course Content Accessed): This coefficient represents the change in 'Participants' for each unit increase in 'Audited (> 50% Course Content Accessed)'. It is approximately 10.08 and is statistically significant (p-value < 2e-16), indicating a positive correlation.

Certified: This coefficient represents the change in 'Participants' for each unit increase in 'Certified'. It is approximately -17.46 and is statistically significant (p-value < 4.62e-11), indicating a negative correlation.

% Audited: This coefficient represents the change in 'Participants' for each unit increase in '% Audited'. It is approximately -450.33 and is statistically significant (p-value < 0.000343), indicating a negative correlation.

Total Course Hours (Thousands): This coefficient represents the change in 'Participants' for each unit increase in 'Total Course Hours (Thousands)'. It is approximately 53.3509 and is statistically significant (p-value < 0.000878), indicating a positive correlation.

Model Performance:

Multiple R-squared: The value of 0.9404 suggests that the model explains 94.04% of the variance in the dependent variable. This is a high value, indicating a good fit of the model to the data.

B. Establish model

Through the first model, we find that 'Audited (> 50% Course Content Accessed)' has a significant impact on the dependent variable. Therefore, we create a separate model using only these two variables. Below is the summary of the model. (model)

```
Call:
lm(formula = `Participants (Course Content Accessed)` ~ `Audited (> 50% Course Content Accessed)`,
    data = train)

Residuals:
   Min     1Q Median     3Q    Max
-93928  -4253   3213   6411  67007

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                          -8901.8313  2004.2864  -4.441 2.31e-05 ***
`Audited (> 50% Course Content Accessed)`     9.4776     0.4258  22.258  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17050 on 100 degrees of freedom
Multiple R-squared:  0.8321,    Adjusted R-squared:  0.8304
F-statistic: 495.4 on 1 and 100 DF,  p-value: < 2.2e-16
```

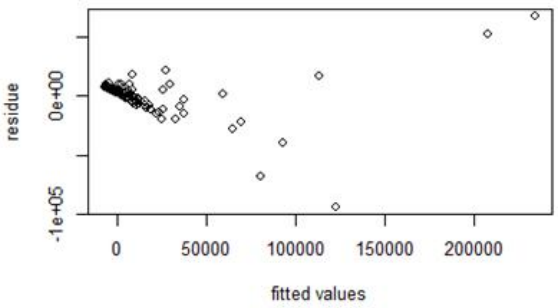
Interpretation of Coefficients:

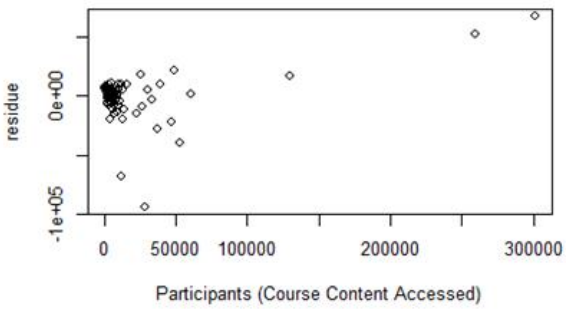
Intercept: The estimated intercept value is -8901.83. This value indicates that when the value of 'Audited (> 50% Course Content Accessed)' is zero, i.e., no one has audited more than 50% of the course content, the estimated value of 'Participants' is approximately -8901.83. In this context, the interpretation of the intercept does not make much sense.

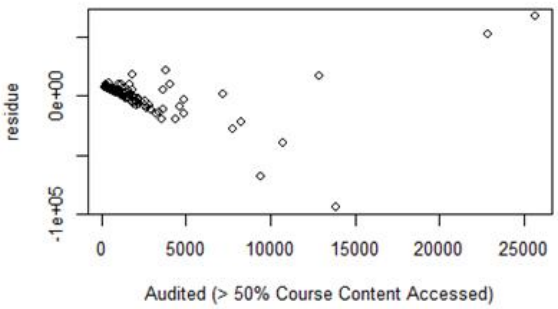
Audited (> 50% Course Content Accessed): The estimated coefficient value for this variable is approximately 9.48, and it is highly significant statistically (p-value < 2e-16). This implies that for every additional person who has audited more than 50% of the course content, the estimated value of 'Participants' increases by 9.48. In other words, there is a positive correlation between the number of people who have audited more than 50% of the course content and the number of participants.

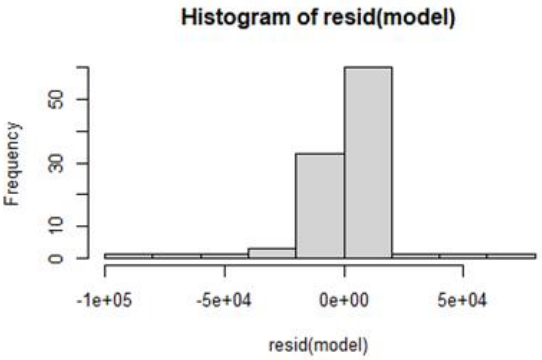
Model Performance:

Multiple R-squared: The value of the Multiple R-squared is 0.8321, indicating that the model explains 83.21% of the variance in the dependent variable 'Participants'. This is a relatively high value, suggesting that the independent variable 'Audited (> 50% Course Content Accessed)' plays a significant role in explaining the variance of the dependent variable.







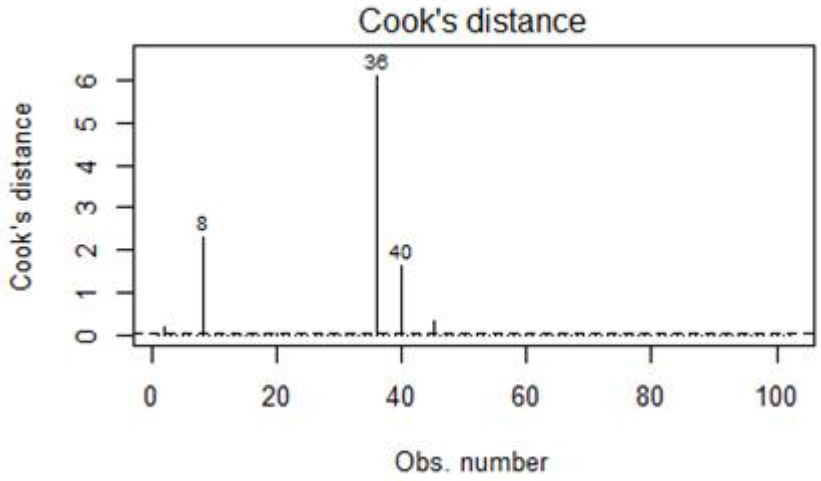


Outlier Test:

	rstudent	unadjusted p-value	Bonferroni p
40	-7.033638	2.6362e-10	2.6889e-08
36	5.511602	2.8283e-07	2.8848e-05
45	-4.479989	2.0054e-05	2.0455e-03
8	3.753970	2.9377e-04	2.9964e-02

The outlier test indicates that there may be four outliers in the model.

Influential Test:




```
# A tibble: 6 × 15
  Participants (Course Content Access…¹ Audited (> 50% Cours…² Certified `% Audited` `% Certified` % Certified of > 50%…³
                                  <dbl>                  <dbl>     <dbl>        <dbl>          <dbl>                  <dbl>
1                                129400                  12888      1439         9.96           1.11                   11.1
2                                 52521                  10729      5058         20.4           9.64                   47.1
3                                259577                  22894      1208         8.82           0.47                   5.24
4                                301082                  25637      1523         8.51           0.51                   5.45
5                                 28425                  13849      2685         48.7           9.45                   19.4
6                                 12007                   9446      1404         78.7           11.7                   14.9
# ℹ abbreviated names: ¹`Participants (Course Content Accessed)`, ²`Audited (> 50% Course Content Accessed)`,
#   ³`% Certified of > 50% Course Content Accessed`
# ℹ 9 more variables: `% Played Video` <dbl>, `% Posted in Forum` <dbl>, `% Grade Higher Than Zero` <dbl>,
#   `Total Course Hours (Thousands)` <dbl>, `Median Hours for Certification` <dbl>, `Median Age` <dbl>, `% Male` <dbl>,
#   `% Female` <dbl>, `% Bachelor's Degree or Higher` <dbl>
```

C. Establish model2

From the initial model (model1), we extracted the most significant independent variables and used them to construct a new model. Below is the summary of this new model. (model2)

```
Call:
lm(formula = `Participants (Course Content Accessed)` ~ `Audited (> 50% Course Content Accessed)` +
    Certified + `% Audited` + `Total Course Hours (Thousands)`,
    data = train)

Residuals:
   Min     1Q Median     3Q    Max
-66554  -3926    946   6409  48625

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                          11288.6880  3050.5268   3.701 0.000357 ***
`Audited (> 50% Course Content Accessed)`     9.9865     0.4668  21.392  < 2e-16 ***
Certified                            -14.5001     1.7662  -8.210 9.60e-13 ***
`% Audited`                          -406.8939    74.2659  -5.479 3.38e-07 ***
`Total Course Hours (Thousands)`       38.7400    15.6413   2.477 0.014988 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12190 on 97 degrees of freedom
Multiple R-squared:  0.9167,    Adjusted R-squared:  0.9133
F-statistic:   267 on 4 and 97 DF,  p-value: < 2.2e-16
```

the R-value is 0.9167.

D. Establish model3

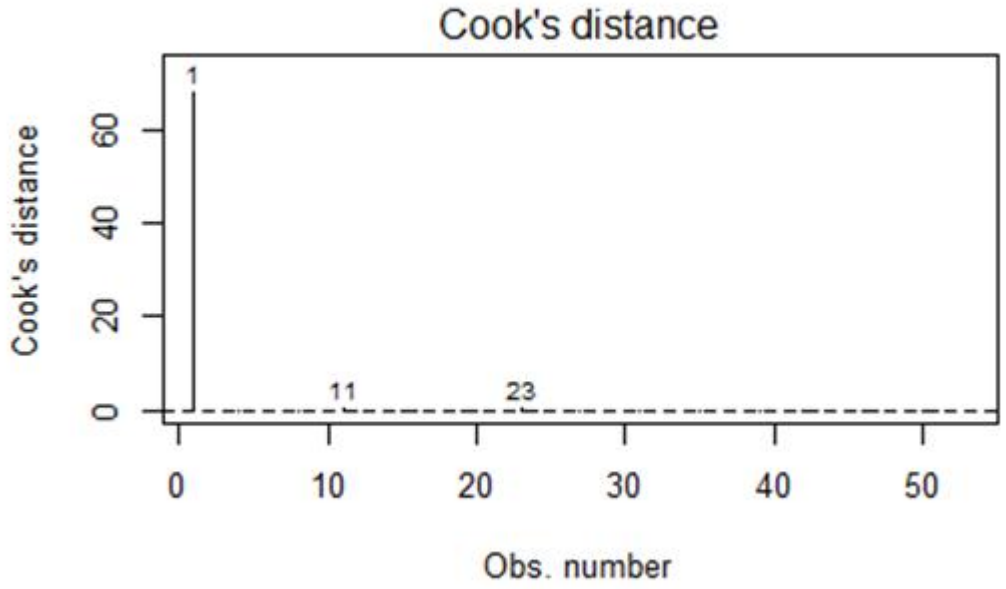
By transforming the data to exclude outliers and conducting another multiple linear regression (MLR) analysis, a new model, model3, was established. Below is the summary of this new model. (model3)

```
Call:
lm(formula = `Participants (Course Content Accessed)` ~ `Audited (> 50% Course Content Accessed)` +
    Certified + `% Audited` + `Total Course Hours (Thousands)`,
    data = data2)

Residuals:
    Min      1Q  Median      3Q     Max
-7689.1 -1669.6  -382.8  2343.5  6102.0

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                        7048.0730  1746.3088   4.036 0.000195 ***
`Audited (> 50% Course Content Accessed)`    8.3486     0.8977   9.300 2.58e-12 ***
Certified                           -14.6782     1.7114  -8.577 3.00e-11 ***
`% Audited`                        -251.9738    46.7609  -5.389 2.12e-06 ***
`Total Course Hours (Thousands)`    146.1536    49.9455   2.926 0.005228 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3108 on 48 degrees of freedom
Multiple R-squared:  0.9729,    Adjusted R-squared:  0.9707
F-statistic: 431.4 on 4 and 48 DF,  p-value: < 2.2e-16
```

the R-value is 0.9729.

This R-value is significantly higher than the previous model which did not account for the removal of outliers.

Influential Test:




```
# A tibble: 5 × 15
  Participants (Course Content Access…¹ Audited (> 50% Cours…² Certified `% Audited` `% Certified` % Certified of > 50%…³
                                  <dbl>                  <dbl>     <dbl>       <dbl>          <dbl>                  <dbl>
1                                129400                  12888      1439        9.96           1.11                   11.1
2                                  8493                   2137      1183        25.2           13.9                    55.4
3                                 26086                   4637      1583        17.8           6.07                    34.1
4                                  7008                   3248       961        46.4           13.7                    29.6
5                                  4262                   3518       377        82.6           8.85                    10.7
# ℹ abbreviated names: ¹`Participants (Course Content Accessed)`, ²`Audited (> 50% Course Content Accessed)`,
#   ³`% Certified of > 50% Course Content Accessed`
# ℹ 9 more variables: `% Played Video` <dbl>, `% Posted in Forum` <dbl>, `% Grade Higher Than Zero` <dbl>,
#   `Total Course Hours (Thousands)` <dbl>, `Median Hours for Certification` <dbl>, `Median Age` <dbl>, `% Male` <dbl>,
#   `% Female` <dbl>, `% Bachelor's Degree or Higher` <dbl>
```

E. Hypothesis Testing Using F-Test

(1) Hypotheses Formulation

Null Hypothesis (H0): All independent variables are ineffective, meaning all coefficients are zero.

Alternative Hypothesis (H1): At least one independent variable is effective, meaning not all coefficients are zero.

Alpha (α): 0.05

(2) Testing Approach

Conduct an F-test to compare the model with all independent variables against a model with no independent variables.

(3) Decision Criteria

If the computed F-statistic exceeds the critical value derived from the F-distribution, the null hypothesis is rejected, supporting the alternative hypothesis of at least one significant independent variable.

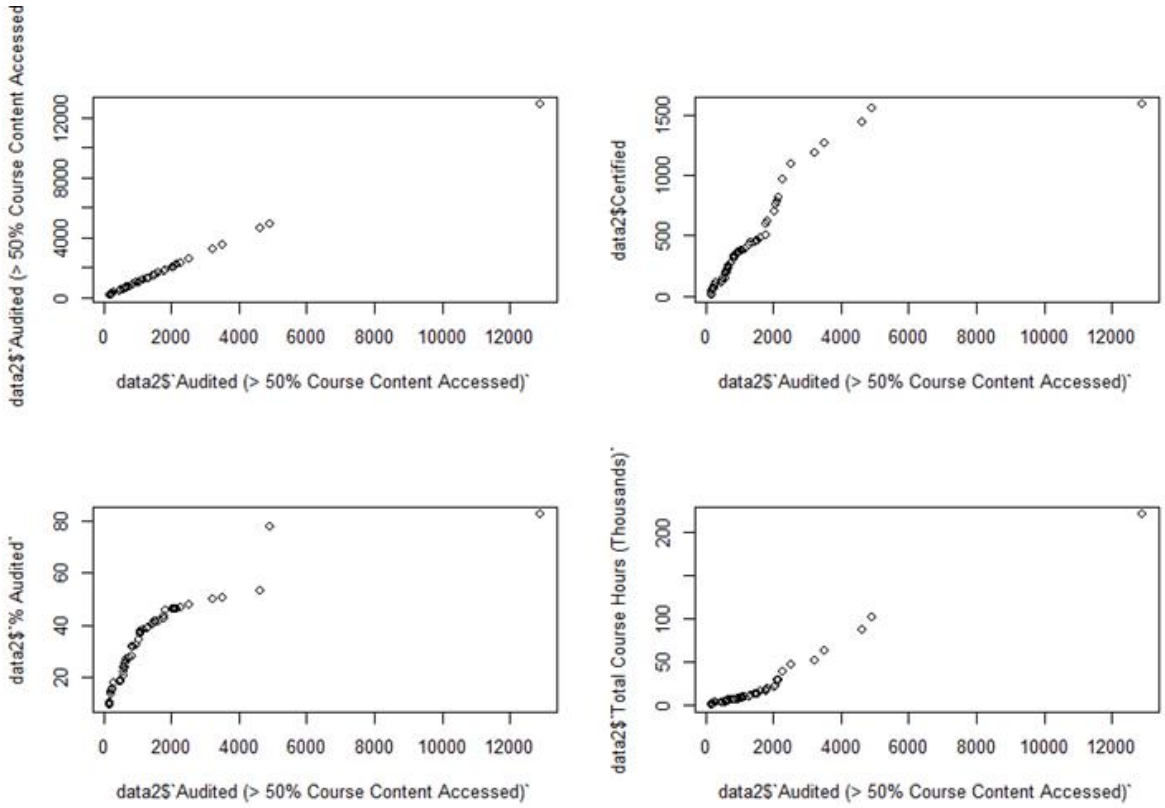
(4) Results

The F-statistic in the model is 431.4. Using the qf function, the critical value from the F-distribution is determined to be qf(0.95, df1=4, df2=97) = 2.46548.

(5) Conclusion

We can confidently reject the null hypothesis, suggesting that the model contains at least one significant predictor.

F.QQ-Plot



We use qqplots compare various variables from a dataset labeled data2. The variables compared are:

Audited(>50% Course Content Accessed)vs. Audited (> 50% Course Content Accessed)

Audited (> 50% Course Content Accessed) vs. Certified

Audited (> 50% Course Content Accessed) vs. % Audited

Audited (> 50% Course Content Accessed) vs. Total Course Hours (Thousands)

The plots indicate that there are distributional differences between 'Audited (> 50% Course Content Accessed)' and 'Certified', '% Audited', and 'Total Course Hours (Thousands)'.

G. ANOVA table

```
Analysis of Variance Table

Response: Participants (Course Content Accessed)
                                        Df     Sum Sq    Mean Sq  F value    Pr(>F)
`Audited (> 50% Course Content Accessed)`  1 1.5046e+10 1.5046e+10 1557.321 < 2.2e-16 ***
Certified                                  1 2.5525e+08 2.5525e+08   26.420 4.996e-06 ***
`% Audited`                                1 1.2875e+09 1.2875e+09  133.263 1.876e-15 ***
`Total Course Hours (Thousands)`           1 8.2730e+07 8.2730e+07    8.563  0.005228 **
Residuals                                 48 4.6375e+08 9.6614e+06
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In summary, the ANOVA table suggests that 'Audited (> 50% Course Content Accessed)', 'Certified', and '% Audited' are significant predictors of the number of participants accessing course content. However, 'Total Course Hours (Thousands)' may not be as strongly associated with the dependent variable as the other predictors in the model.

H. 95% Confidence Intervals

	2.5 %	97.5 %
(Intercept)	3536.883912	10559.26217
`Audited (> 50% Course Content Accessed)`	6.543575	10.15354
Certified	-18.119162	-11.23718
`% Audited`	-345.992813	-157.95483
`Total Course Hours (Thousands)`	45.731360	246.57578

This table displays the 95% confidence intervals for each coefficient in our linear regression model, model3. A confidence interval is a range of estimated values for the coefficient, and we can be quite certain (in this case, at a 95% confidence level) that this range includes the true value of the coefficient. Specifically, each row in the table represents a model parameter (coefficient) and provides the lower (2.5%) and upper (97.5%) limits of the confidence interval for that parameter at the 95% confidence level.

(Intercept): The 95% confidence interval for the intercept ranges from approximately 3536.88 to 10559.26. This means we can be quite confident that, if all other variables are zero, the true value of 'Participants (Course Content Accessed)' will fall within this range.

Audited (> 50% Course Content Accessed): The confidence interval for the coefficient of this independent variable ranges from approximately 6.54 to 10.15. Thus, we can be quite confident that for each additional participant who audited more than 50% of the content, the average increase in 'Participants (Course Content Accessed)' falls within this range.

Certified: The confidence interval for the coefficient of this independent variable ranges from approximately -18.12 to -11.24. The negative sign indicates that as 'Certified' increases, the expected value of 'Participants (Course Content Accessed)' decreases. We can be 95% certain that the true amount of decrease falls within this range.

% Audited: The confidence interval for the coefficient of this independent variable ranges from approximately -346.00 to -157.95. This is also a negative confidence interval, suggesting a negative correlation between '% Audited' and 'Participants (Course Content Accessed)'.

Total Course Hours (Thousands): The confidence interval for the coefficient of this independent variable ranges from approximately 45.73 to 246.58. This positive confidence interval indicates that as 'Total Course Hours (Thousands)' increases, the expected value of 'Participants (Course Content Accessed)' also increases, and we can be 95% certain that the true amount of increase lies within this range.

Overall Project Conclusion:

Our project involved a comprehensive analysis of course participation, focusing on the relationship between various variables and the number of participants accessing course content. Through linear regression analysis, outlier detection, and hypothesis testing, we gained valuable insights.

1.Key Findings:

- ✓ The initial model (model1) highlighted significant correlations between course participation and variables such as 'Audited (> 50% Course Content Accessed)' (positive correlation), 'Certified,' '% Audited,' and 'Total Course Hours (Thousands)' (negative correlations).
- ✓ The second model emphasized the positive impact of individuals auditing more than 50% of the course content on the number of participants.
- ✓ A refined model (model2) and an outlier-excluded model (model3) demonstrated improved performance, with model3 achieving a significantly higher R-value.

2.Statistical Significance:

- ✓ Hypothesis testing using an F-test confidently rejected the null hypothesis, indicating the presence of at least one significant predictor in the model.
- ✓ ANOVA results highlighted the significance of 'Audited (> 50% Course Content Accessed),' 'Certified,' and '% Audited' in predicting course participation.

3.Practical Implications:

- ✓ The identified predictors offer actionable insights for course organizers to enhance participant engagement.

- ✓ The positive impact of individuals auditing a substantial portion of the course content suggests the importance of content accessibility.

Limitations and Considerations:

Assumptions of linearity and potential collinearity among variables may influence the model's accuracy. The removal of outliers and the absence of a time factor are factors that may impact the generalizability of the findings.

Recommendations for Future Research:

Explore additional factors influencing course participation, such as demographic variables or course structure.

Consider dynamic models that incorporate the temporal aspect for a more comprehensive analysis.

In conclusion, our project provides valuable insights into the factors influencing course participation. The identified predictors can inform strategic decisions for course design and outreach efforts. However, careful consideration of model limitations and avenues for further research is essential for a nuanced understanding of the dynamics involved.

[1] Kaggle, “Online Courses from Harvard and MIT”, 2016, https://www.kaggle.com/datasets/edx/course-study/