STAT 452 – Statistical Modeling I

Case Study 1

Paolo Furlanetto Ferrari

03/28/203

Summary

I. Introduction

In this Case Study, we analyzed the grades of multiple assignments in four semesters of the discipline Stat 452. The goal is to model the grades of the second exam, Exam 2, as a function of the grades of other assignments and also to the semester in which the data was taken. Throughout this Case Study, we chose is a linear regression as the model. In the next sessions, we provide an overview of the data as well a comparison between multiple linear regression models with different numbers of predictors. Next, we describe the methodology of ANOVA that we use to test the models and select the best model [1]. Finally, we conclude by analyzing the performance of the selected model against the real data.

II. Data Overview

We start by looking qualitatively at the dependence of the grades of Exam 2 versus the other grades in the dataset. Figure 1 shows 5 scatterplots of Exam 2 plotted in the y-axis versus the other variables in the x-axis, representing, from left to right, the grades of Exam 1, the Project, mean Case Study, mean Homework grade, and Participation level. We also plot the regression line (shown in blue) for each of the variables taken separately, using distinct simple linear regression models. As expected, all variables have a positive correlation with Exam 2.

A clear limitation of the linear models happens because of the truncation of the grades lower than 0 and higher than 100. This is specially evident in the Project and Case Study scatterplots because of the high occurrence of 0s and 100s. These deviations already indicate that these two variables may not be good predictors of Exam 2, as will be seen later in Section 3. Notably, the lines with largest slopes seem to be Exam 1 and Homework, indicating a higher correlation between these variables and the Exam 2 score.

Next, we analyzed the dependence of the Exam 2 scores with respect to the different semesters. Figure 3 shows a box-plot of the Exam 2 grades for the four semesters, with the average of each semester annotated in red. The right plot of Figure 3 shows a scatterplot of the raw data for the 4 semesters, along with a regression line in blue, fitted using a simple linear regression model. While there's not a clear trend of Exam 2 with Semester, the linear fit indicates a slight decrease on the mean scores.

To get a more quantitative sense of the data, Table 1 shows the correlation coefficients of Exam 2 with respect to each of the variables shown on Figures 1 and 3. The single linear regression coefficients are also shown. We conclude from this early analysis that the variables Exam 1, Homework and Semester Number are the ones with highest correlation with Exam 2. This observation will be corroborated when the multiple linear regression models are compared.

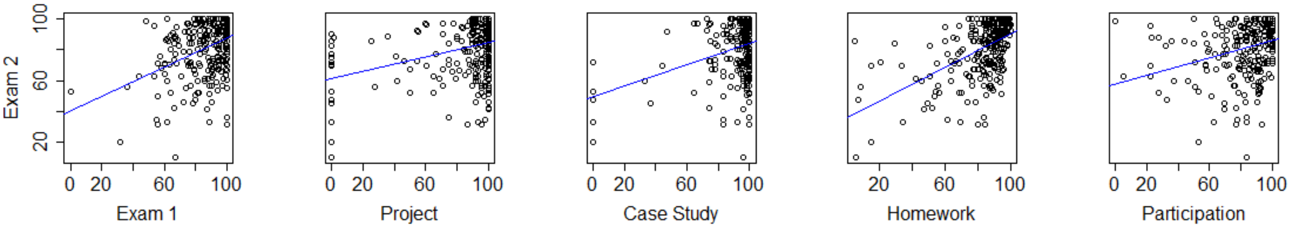

Figure 1: Scatterplots of Exam 2 (y-axis) vs. other scores in the dataset. Regression lines are shown in blue.

III. Model Comparison

Throughout this section, we use a significance level of 0.05. The first attempted multiple linear regression model takes into account all variables as predictors. Table 2 shows the obtained coefficients and their corresponding p-values. We conclude that the variables Project, Case Study and Participation are candidates for being dropped out from the model since the p-values of their coefficients are larger than 0.05. We also note that the values of their coefficients are the much closer to 0 compared to the variables Exam 1, Homework and Semester.

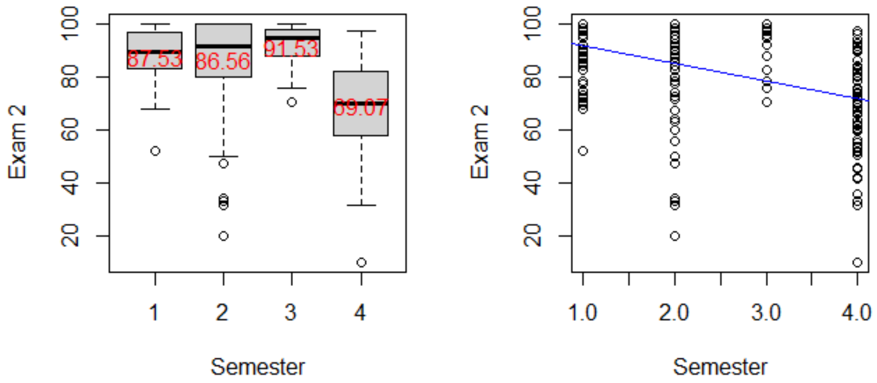

Figure 2: Box-plot and scatterplots of Exam 2 (y-axis) vs. semester number. The average for each semester is annotated in red for the left plot. In the right plot, the regression line is shown in blue.

From this observation, we next evaluate 5 different models in Table 3. These models correspond, respectively, to the full model (FM), FM with Project (P) and Case Study (C) removed FM – (P + C), with Participation (Pa) and C removed FM – (C + Pa), with P and Pa removed FM – (P + Pa), and with the three variables removed at the same time FM – (P+C+Pa). For each model, Table 3 displays the R^2 coefficient, residual sums of squares (RSS), and the AIC statistic [1]. We see that the model with highest R^2 and lowest AIC is the model with the three predictors removed, thus indicating that it is the most correct one (Model 5). While not shown here, the summary of each of the Models 2-4 indicated that the variable which was kept relative to Model 5 should be removed according to the p-value associated to its coefficient.

Table 1: Correlation coefficient and single linear regression (SLR) coefficients of Exam 2 scores versus the different variables.

	Exam 1	Project	Case Study	Homework	Participation	Semester
Correlation Coefficient	0.35	0.33	0.31	0.54	0.29	-0.42
SLR Coefficient	0.47	0.23	0.35	0.55	0.28	-6.63

Table 2: Regression coefficients and p-values of the full model.

	Exam 1	Project	Case Study	Homework	Participation	Semester
Coefficient	0.35	0.01	0.02	0.40	-0.03	-7.51
p-value	9 e-9	0.69	0.69	5 e-10	0.60	2 e-16

Table 3: Metrics for comparison of the five models tested in this work.

	R^2	RSS	AIC
Model 1: FM	0.5405	38962.8	2158.6
Model 2: FM – (P+C)	0.5396	39035.9	2155.2
Model 3: FM – (C+Pa)	0.5397	39028.7	2155.1
Model 4: FM – (P+Pa)	0.5398	39017.5	2155.4
Model 5: FM – (P + C + Pa)	0.5393	39065.4	2153.4

From these findings, we conclude that the best model is the one where Exam 2 depends on the variables Exam 1, Homework and Semester only. Table 4 shows the obtained coefficients along with the confidence intervals of each one. Here, we used a Bonferroni correction to calculate the Cis of each coefficient, by dividing the original confidence level by 4.

Table 4: Coefficients and p-values of final model.

	Exam 1	Homework	Semester	Intercept
Coefficient	0.35 ± 0.057	0.41 ± 0.047	-7.54 ± 0.664	36.02 ± 4.748
p-value	4e-9	8e-16	<2e-16	4e-13

IV. Model Evaluation

To evaluate the quality of the fit of the final model, we qualitatively looked both at the predictions and residuals of the model versus the actual Exam 2 scores, shown in the left panel of Figure 3. As expected, the residuals are approximately uncorrelated to the Exam 2 scores and with a mean 0 (the line y = 0 is shown in red for comparison). In addition, the right plot shows that the predicted and actual values have a correlation near 1, as indicated by the y = x line shown in green.

Finally, the left panel of Figure 4 shows a histogram of the residuals of the model. The near-normal distribution of the residuals around 0 indicates that the assumptions behind the linear regression model are reasonable. Corroborating this, the right panel in Figure 4 shows a quantile-quantile (QQ) plot of the residuals, together with a blue line indicating the approximately linear dependence of the theoretical quantiles given by a standard normal distribution and the sample residual quantiles. As shown in the QQ plot, deviation from the normality of the residuals is more pronounced near the edges of the data, where the residuals are larger than 20. This is most likely a result of the truncation of the Exam 2 scores at 100.

In conclusion, the analysis in this work shows that the expected grade of a student in their second exam can be predicted by knowing their grade in exam 1, their homework grade and in what semester they are taking the course. Students which perform well on exam 1 and their homework tend to perform better in exam 2. In addition, the model shows a small tendency for students to perform worse in the exam versus students in the previous semester.

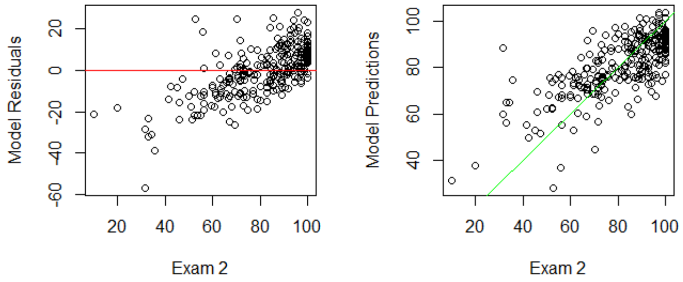

Figure 3: Scatterplot of the model residuals (left) and predicted scores (right) vs the real Exam 2 scores. The lines y = 0 and y = x are added in red and green respectively as visual guides.

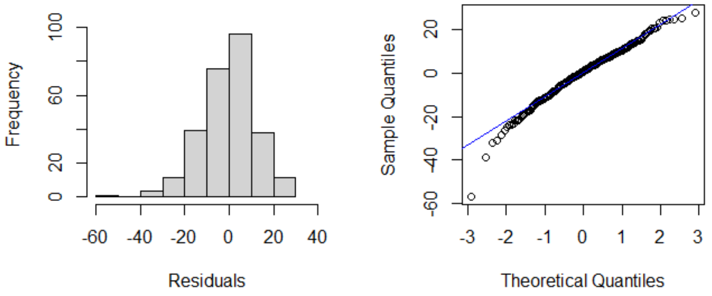

Figure 4: Histogram of the model residuals (left) and QQ plot of residuals vs a standard normal distribution (right).

References

[1] Faraway, J. Practical Regression and Anova using R (2002).