Student Name: Eldor Fozilov

Student Number: 20192032

# Small Project 2

## Problem 1

(i) The **minimum** values of the variables **atndrte**, **priGPA**, and **ACT** are 6.25, 0.857 and 13, respectively.
The **maximum** values of the variables **atndrte**, **priGPA**, and **ACT** are 100, 2.93 and 32, respectively.
The **average** values of the variables **atndrte**, **priGPA**, and **ACT** are approximately 81.71, 2.59, and 22.51, respectively.

(ii) $atendrte\ (hat) = 75.7 + 17.261priGPA - 1.717ACT$

N = 680, R-squared = 0.2906

The fact that the intercept is equal to 75.7 means that when prior GPA and ACT score are both equal to 0, the predicted student's attendance rate is 75.7 (in percentage). However, since prior GPA and ACT score are very unlikely to be zero, the intercept does not have a useful meaning.

(iii) The fact that the estimated coefficient of priGPA is equal to 17.261 means that holding other things constant, if prior GPA increases by one point, then the attendance rate is predicted to increase by 17.261 (in percentage). This does not contradict common sense as we would expect that students with higher prior GPA would have higher attendance rate.

The estimated coefficient of ACT is surprisingly equal to – 1.717, which means that holding other things constant, an increase in ACT by one point results in 1.717 (in percentage) decrease in attendance rate. We would make a hypothesis that as student's ACT increase, he or she will have higher attendance rate ($B_2 >= 0$). However, the data does not support our hypothesis: the t-statistic of the variable ACT's estimated coefficient is very low (-10.16). This might mean that 1) we need to reject our hypothesis, 2) there were mistakes in collecting the sample, 3) there might be many other variables that are highly correlated with ACT and attendance rate, but they were not included in the model.

(iv) If priGPA = 3.65 and ACT = 20, the predicted **atendrte** is approximately equal to 104.36. The result means that those students who have priGPA=3.65 and ACT = 20 have very high predicted attendance rate, although predicted attendance rate of 104.36% is not actually possible since the maximum can only be 100%. There is only one student in the sample

who has priGPA = 3.65 and ACT = 20 (observation number 569), and he has high attendance rate (87.5%) as expected.

(v) If Student A has priGPA = 3.1 and ACT = 21 and Student B has priGPA = 2.1 and ACT = 26, the predicted difference in their attendance rates is equal to 25.846 (in percentage)

Problem 2

(i) $math10\ (hat) = -20.36 + 6.23\log(expend) - 0.30lnchprg$

N = 408, R-squared = 0.18

The fact that the estimated coefficient of the variable **log(expend)** is equal to 6.23 means that, holding other things constant, if expenditure per student increase by 1%, then the expected increase in **math10** (the percentage of tenth graders at a high school receiving a passing score on a standardized mathematics exam) will be 6.23 percentage points. We expected that as expenditure per student increases, the percentage of students passing the math exam will also increase. The data supports our expectations.

However, the estimated coefficient of **lnchprg** (the percentage of students who are eligible for the school lunch program subsidized by the government) is equal to a negative number, -0.30. This means that, holding other things constant, the expected decrease in math10 as a result of one point increase in **lnchprg** is equal to 0.30. This number might be surprising or not. On the one hand, we would expect that all other factors being equal, if a student who is too poor to eat regular meals becomes eligible for the lunch program, his or her performance should improve. On the other hand, the existence of the lunch program subsidized by the government might reflect the poverty rate of students, the lack of school quality and resources, and thus an increase in the number of students in the lunch program might actually have relationships with the decrease in exam performance. Since the sample size is pretty large and as the p-value is very low (0.0367), we might have to believe the results and consider the latter view as being closer to the truth. ($B_2 < 0$)

(ii) The estimated intercept is equal to -20.36. However, it does not provide any useful information because variable **math10** cannot be a negative number even when log(expend) and lnchprg are both equal to 0. Also, although it makes sense to set **lnchprg** equal zero since it is possible that no student is participating in the existing lunch program, setting log(expend) to 0 (in other words, setting expend to 1) does not make sense since the expenditure per student is much higher than 1$ in reality according to the provided data.

(iii) $math10\ (hat) = -69.34 + 11.16\log\ (expend)$

$N = 408, R - squared = 0.03$

When we run the simple regression of math10 on log(expend), the estimated slope coefficient is approximately equal to 11.16. When compared to the estimated coefficient in part (i), the estimated spending effect is much larger in this case.

(iv) The correlation between log(expend) and lnchprg is approximately equal to -0.19. This makes sense because if schools have high expenditure per student, they don't have to rely on government-funded programs like the lunch program. On the other hand, if schools don't have enough resources and high expenditure per student, they might have to rely more on the help of the government-funded programs.

(v)