Exam 1 Evan Cranmer

Evan Cranmer

2024-10-01

1.1 Interpreting the estimates for each coefficient

Intercept: The expected mean value of the outcome, y, is 2.4 when x1, x2, and x3 are equal to 0.

x1: If the beta was statistically significant, the change in expected mean value of y is 0.02 for every one-unit increase in x1, holding other variables constant. However, since our p-value is greater than our alpha value of .05, we *fail to reject the null*, which states that x1 = 0. Simply put, there is not enough evidence to conclude that x2 has a statistically significant effect on y, holding other variables constant. Since it is close to .05 however, it could possibly hold substantive significance. Still, it is not statistically significant.

x2: If the beta was statistically significant, the change in expected mean value of y is -0.12 for every one-unit increase in x2, holding other variables constant. However, since our p-value is greater than our alpha value of .05, we *fail to reject the null*, which states that x2 = 0. There is not enough evidence to conclude that x2 has a statistically significant effect on y, holding other variables constant. The p-value for this coefficient is pretty far from .05, at .18.

x3: The change in expected mean value of y is 0.15 for every one-unit increase in x3, holding other variables constant. Since our p-value is less than .001, we reject the null hypothesis that states x3 holds a value of zero.

1.2 Constructing confidence intervals for each

```r
#for each regression coefficient... formula is CI = estimte +/- (t-value)*(se)

#x1
x1lowCI <- 0.02292-(1.802*0.01272)
x1upCI <- 0.02292+(1.802*0.01272)

x1CI <- print(c(x1lowCI,x1upCI))
```

```
## [1] -0.00000144  0.04584144
```

```
#x2

x2lowCI <- -0.11583-(-1.380*0.08391)
x2upCI <- -0.11583+(-1.380*0.08391)

x2CI <- print(c(x2lowCI,x2upCI))
```

```
## [1] -0.0000342 -0.2316258
```

```
#x3 CIs

x3lowCI <- 0.15450-(6.192*0.02495)
x3upCI <- 0.15450+(6.192*0.02495)

x3CI <- print(c(x3lowCI,x3upCI))
```

```
## [1] 0.0000096 0.3089904
```

Interpretation of each confidence interval

x1: We are 95% confident that the true population beta is between -0.00000144 and 0.04584144. Since this interval includes zero there is this beta may not be statistically different from 0.

x2: We are 95% confident that the true population beta is between -0.0000342 and -0.2316258. Since this interval includes zero there is a chance this beta may not be statistically different from 0.

x3: We are 95% confident that the true population beta is between 0.0000096 and 0.3089904. This interval is close to zero, but does not include zero. Therefore, we can reject the null hypothesis that states x3 is equal to 0/has a non-effect on y.

1.3 Interpreting t-value of -1.380

This t-value corresponds to x2… which is our second beta coefficient. Our hypotheses are questioning whether x2 is different than 0.

Our null hypothesis is H_0: Beta2 = 0 Our alternative hypothesis is H_a: Beta2 does not equal 0

Since the p-value for our t-value, which is equal to .18, is much higher than our significance level at .05, we fail to reject the null hypothesis. We can conclude there is not enough statistical evidence to suggest that x2 has a non-zero effect on y.

1.4 Calculating missing values

```
#(A) Calculating sum of squares. If mean square = sum of squares/ degrees of freedom..
#Sum of squares = mean of squares * df

A_1.4 <- 0.122 * 1

B_1.4 <- 0.451 * 3

C_1.4 <- 57.478/32

#F= MS/MSE
D_1.4 <- (64.447/1.796)

print(c(A_1.4, B_1.4, C_1.4, D_1.4))
```

```
## [1]  0.122000  1.353000  1.796188 35.883630
```

Based on our results, our missing values are as follows:

A = 0.122 B = 1.353 C = 1.796 D = 35.884

1.5 Comparing our another model

Adjusted R-squared: First I'll be comparing the models on adjusted R-squared becasue the models have a differing number of parameters that could inflate the r-squared value. Using adjusted R-squared will account for this difference. Our first model in 1.1 has an adjusted r-squared value of 0.57, which is greater than our r-squared value of 0.54 in our model for 1.5. This means that approximately 57.0% of the variance in the dependent variable is explained by the independent variables, and the value was adjusted for the number of parameters in our model. Since more variance is explained by the first model than the second, model 1.1 seems to be a better fit based on adjusted R-squared.

F-test: Model 1.1 has a larger F-statistic, meaning it explains a larger portion of variation in the dependent variable compared to an intercept-only model. The p-value for the f-test in model 1.1 is <.001, meaning we can reject the null which says an intercept-only model explains the variance just as well as our model. Even though both models are statistically significant with a p value <.001, our p-value for model 1.1 is < our p-value for model 1.5. So, I would still choose model 1 because I'd be more confident in our model based on the p-values for the f statistic.

1.6 Null and Alternative Hypothesis for F-statistic in model 1

Our null hypothesis is H_0: Beta1 = Beta2 = Beta3 = 0 Our alternative hypothesis is H_a: Beta1 or Beta2 or Beta 3 does not equal 0.

I would reject the null hypothesis for model 1.1 because the p-value is < .001 which is less than our critical value of .05. This signifies that at least one of the independent variables in the model is has a non-zero effect on our dependent variable.

2.1 Loading CAschools CSV

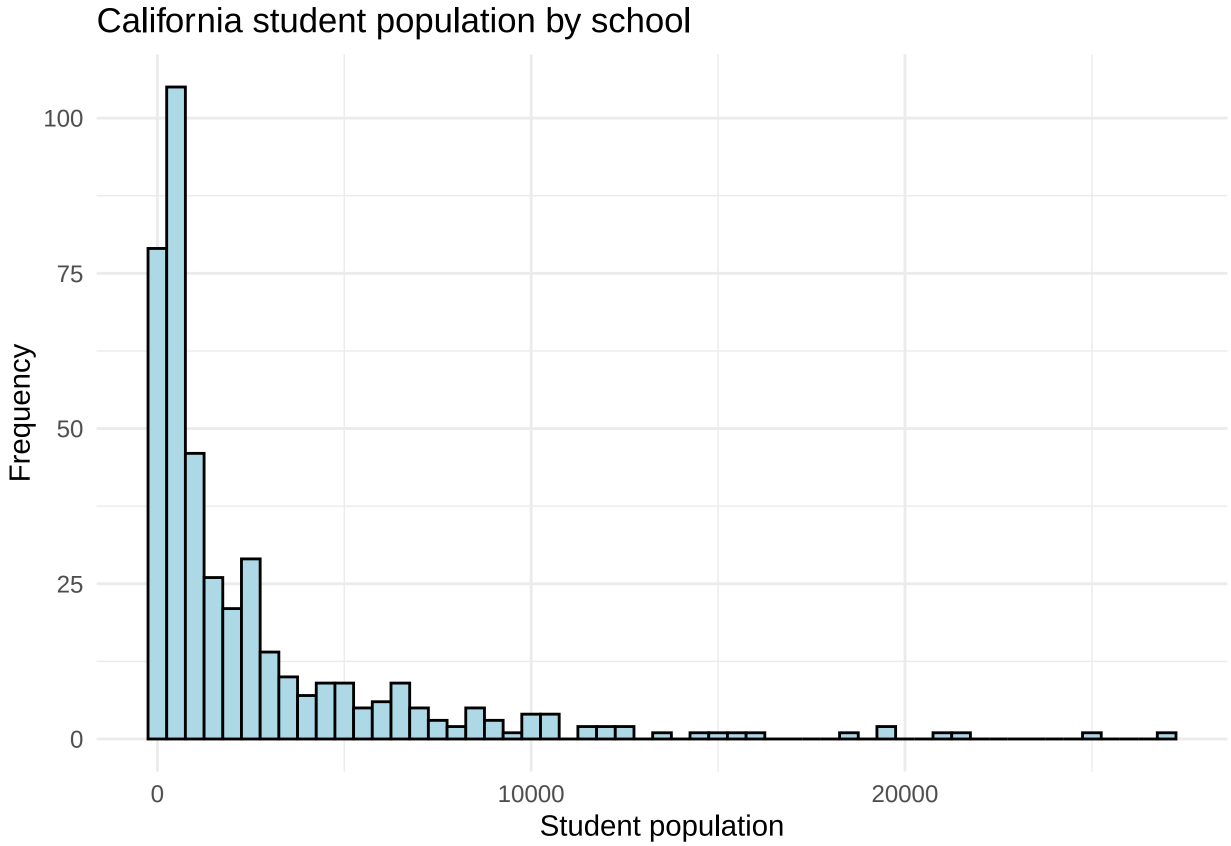
```
#importing data
CAschools <- read.csv("/Users/bcranmer9/Documents/SURV615/CASchools.csv")

#looking at summary statistics for all variables in our dataset
summary(CAschools)
```

```
##     rownames         district         school             county
##  Min.   :  1.0   Min.   :61382   Length:420         Length:420
##  1st Qu.:105.8   1st Qu.:64308   Class :character   Class :character
##  Median :210.5   Median :67760   Mode  :character   Mode  :character
##  Mean   :210.5   Mean   :67473
##  3rd Qu.:315.2   3rd Qu.:70419
##  Max.   :420.0   Max.   :75440
##     grades            students          teachers          calworks
##  Length:420         Min.   :   81.0   Min.   :   4.85   Min.   : 0.000
##  Class :character   1st Qu.:  379.0   1st Qu.:  19.66   1st Qu.: 4.395
##  Mode  :character   Median :  950.5   Median :  48.56   Median :10.520
##                     Mean   : 2628.8   Mean   : 129.07   Mean   :13.246
##                     3rd Qu.: 3008.0   3rd Qu.: 146.35   3rd Qu.:18.981
##                     Max.   :27176.0   Max.   :1429.00   Max.   :78.994
##      lunch           computer        expenditure        income
##  Min.   :  0.00   Min.   :   0.0   Min.   :3926   Min.   : 5.335
##  1st Qu.: 23.28   1st Qu.:  46.0   1st Qu.:4906   1st Qu.:10.639
##  Median : 41.75   Median : 117.5   Median :5215   Median :13.728
##  Mean   : 44.71   Mean   : 303.4   Mean   :5312   Mean   :15.317
##  3rd Qu.: 66.86   3rd Qu.: 375.2   3rd Qu.:5601   3rd Qu.:17.629
##  Max.   :100.00   Max.   :3324.0   Max.   :7712   Max.   :55.328
##     english            read             math
##  Min.   : 0.000   Min.   :604.5   Min.   :605.4
##  1st Qu.: 1.941   1st Qu.:640.4   1st Qu.:639.4
##  Median : 8.778   Median :655.8   Median :652.5
##  Mean   :15.768   Mean   :655.0   Mean   :653.3
##  3rd Qu.:22.970   3rd Qu.:668.7   3rd Qu.:665.9
##  Max.   :85.540   Max.   :704.0   Max.   :709.5
```

```
#I'm going to pick out three to examine visually

#total student enrollment
students_summary <-summary(CAschools$students)
students_summary
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    81.0   379.0   950.5  2628.8  3008.0 27176.0
```

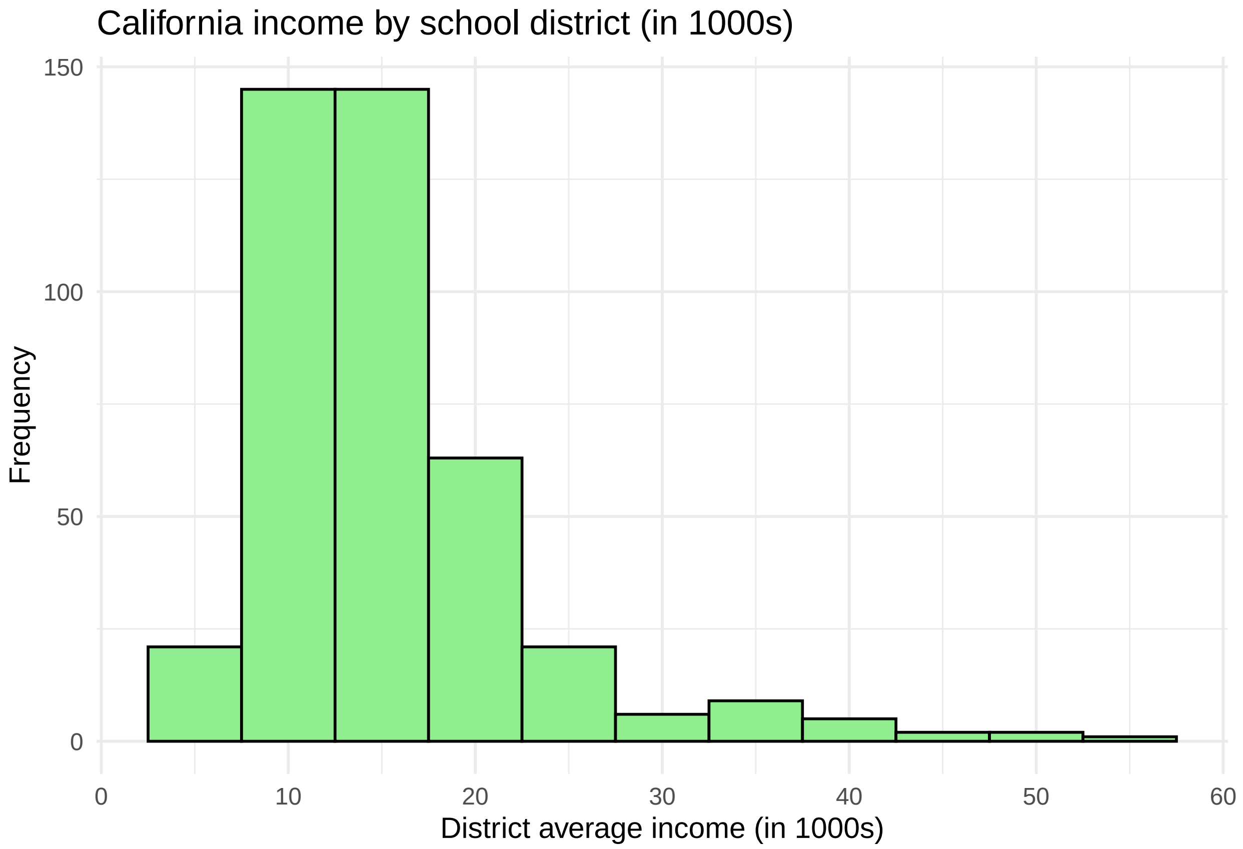
```
students_plot <- ggplot(CAschools, aes(x = students)) +
  geom_histogram(binwidth = 500, fill = "lightblue", color = "black") +
  labs(title = "California student population by school", x = "Student population", y =
  theme_minimal()
students_plot
```



```r
#income by district (in 1000s)
income_summary <- summary(CAschools$income)
income_summary
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.335  10.639  13.728  15.317  17.629  55.328
```

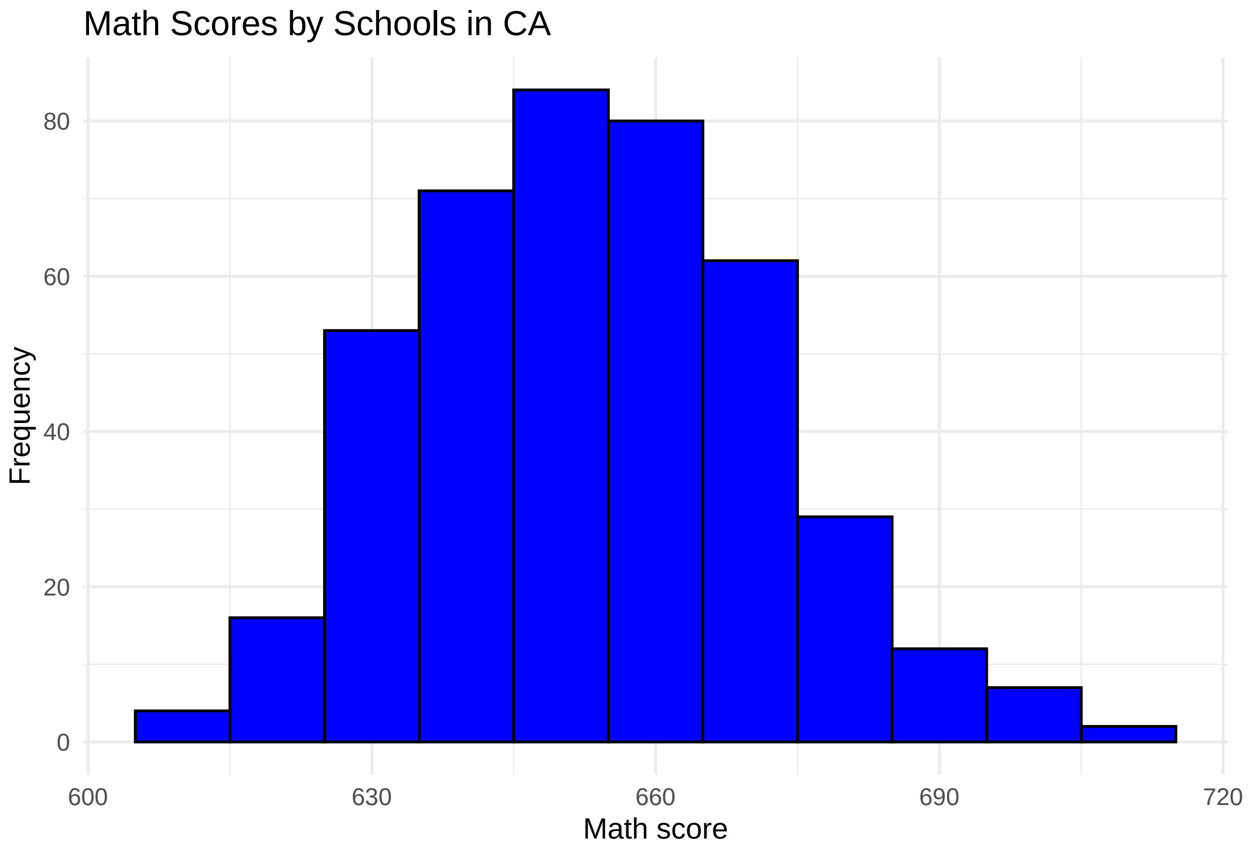
```r
income_plot <- ggplot(CAschools, aes(x = income)) +
  geom_histogram(binwidth = 5, fill = "lightgreen", color = "black") +
  labs(title = "California income by school district (in 1000s)", x = "District average
  theme_minimal()
income_plot
```



```r
#average math test score
math_summary <- summary(CAschools$math)
math_summary
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   605.4   639.4   652.5   653.3   665.9   709.5
```

```
math_plot <- ggplot(CAschools, aes(x = math)) +
  geom_histogram(binwidth = 10, fill = "blue", color = "black") +
  labs(title = "Math Scores by Schools in CA", x = "Math score", y = "Frequency") +
  theme_minimal()
math_plot
```



After running the summary() function on CAschools dataset, I find that the dataset contains 14 variables and 1 column that repeats the rowname of each case. We can, at a glance, see the distribution by looking at the median, mean, and quartiles for each variable. Instead of going through all variables in depth, lets look at 3 variables that are interesting to me: students, income, and math.

Distribution of “students”: This variable gathers the total enrollment for each school in California. The student population has a median is 950.5 students in each school, but a mean of 2628.8 because of the large difference in value between median and mean (median>mean). This suggests that the distribution of students is right-skewed. If we plot this graphically, we can visually confirm that the data is right-skewed.

Distribution of “income”: This variable has data on the district average income by the 1000s. We find that the median district average income is $13,728, with 25% quartile at 10,639 USD and 75% quartile at 17,629 USD. Graphically, this also looks right-skewed.

Distribution of “math”: This variable has data on the average math score by school. This distribution looks normal (more evenly distributed). The median math score is 652.5, with 25 and 75% quartiles being 639.4 and 665.9, respectively.

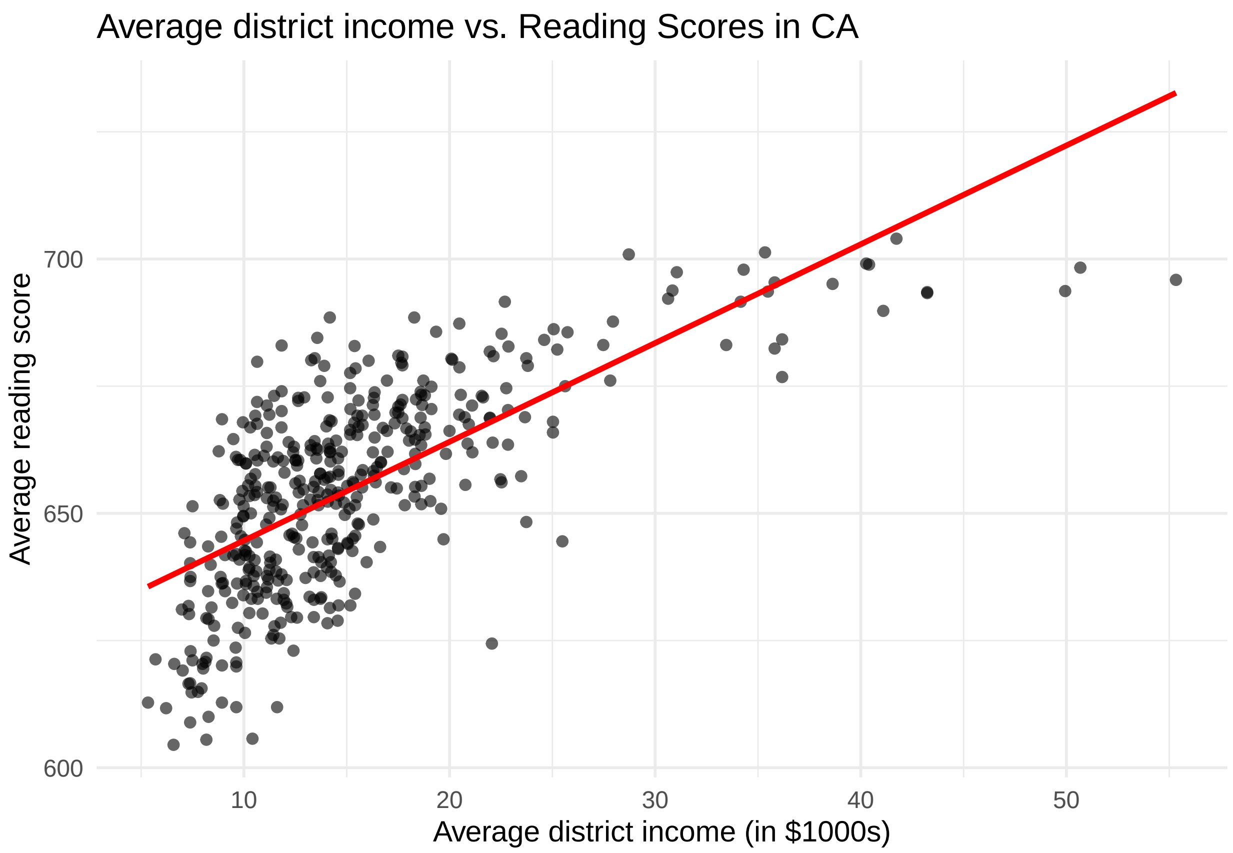
2.3 Creating a regression model

```
#
schoolmodel <- lm(read ~ income, data = CAschools)

summary(schoolmodel)
```

```
##
## Call:
## lm(formula = read ~ income, data = CAschools)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.665 -10.113   0.998  10.675  35.742
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 625.22768    1.65072  378.76   <2e-16 ***
## income        1.94187    0.09749   19.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.42 on 418 degrees of freedom
## Multiple R-squared:  0.487,  Adjusted R-squared:  0.4857
## F-statistic: 396.7 on 1 and 418 DF,  p-value: < 2.2e-16
```

```
#making our scatterplot
school_plot <- ggplot(CAschools, aes(x = income, y = read)) +
  geom_point(alpha = 0.6) +
  labs(
    title = "Average district income vs. Reading Scores in CA",
    x = "Average district income (in $1000s)",
    y = "Average reading score",
  ) +
  geom_smooth(method='lm', color="red", se=FALSE)+
  theme_minimal()

school_plot
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



For every one-unit increase in average district income, the expected change in mean value of reading test scores increases by approximately 1.94.

2.4 Creating a multiple regression model

```
schoolmodel2 <- lm(read ~ income + students, data = CAschools)
summary(schoolmodel2)
```

```
##
## Call:
## lm(formula = read ~ income + students, data = CAschools)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.340  -9.817   1.606   9.459  33.080
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.278e+02  1.634e+00 384.137  < 2e-16 ***
## income       1.958e+00  9.343e-02  20.961  < 2e-16 ***
## students    -1.071e-03  1.725e-04  -6.207 1.31e-09 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 13.81 on 417 degrees of freedom
## Multiple R-squared:  0.5303, Adjusted R-squared:  0.5281
## F-statistic: 235.4 on 2 and 417 DF,  p-value: < 2.2e-16
```

Our intercept does not make intuitive sense, because it is saying that a school with no students in a district with no income would have an expected reading score of 627.8. To make the intercept more interpretable, I would center both of the variables around their means. To do this, we would have to create two new variables in the dataset, taking the observed values for each and subtracting the mean of that variable. After centering the variable, we could interpret the intercept as the expected reading score for those with a mean level of income and mean level of students.

Another issue related to our model, but not our intercept, is the scale of students. Right now, the effect of students is so small because one-unit in “students” is literally one student. We could use a scaling techniques to divide the number of students by say, 1000 (or another number), so the effect would be more profound/interpretable.

2.5 Comparing our models using R-squared, Adjusted R-squared, and ANOVA

```
#global f test to compare our nested models
anova(schoolmodel, schoolmodel2)
```

```
## Analysis of Variance Table
##
## Model 1: read ~ income
## Model 2: read ~ income + students
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    418 86918
## 2    417 79568  1    7350.4 38.522 1.308e-09 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Model 1 R squared and Adjusted R squared:

Our R-squared value for this model is approximately 0.487. Which means that approximately 48.7% of the variance in the dependent variable is explained by the independent variable.

Our adjusted R-squared value for this model is approximately .4857. Which means that approximately 48.57% of the variance in the dependent variable is explained by the independent variable, and the value was adjusted for the number of parameters in our model

Model 2 R squared and Adjusted R squared: Our R-squared value for this model is approximately 0.530. Which means that approximately 53.0% of the variance in the dependent variable is explained by the independent variables.

Our adjusted R-squared value for this model is approximately .528. Which means that approximately 52.8% of the variance in the dependent variable is explained by the independent variables, and the value was adjusted for the number of parameters in our model

Analysis of Variance Global F test

After performing a global F test comparing both of the models, we find that the p-value for testing H0 : RSSreduced = RSS Full, is $< .001$. We can reject H0, stating that there is no difference in fit between the models, and conclude that the second model is a better fit. Aside from the p-value, we can see that the second model with income and student variables has a lower RSS, meaning the model has less unexplained variance in the dependent variable.

2.6 Why are r-squared and adjusted r-squared different for the same models

The adjusted r-squared value is different than the r-squared value because it considers parsimony by accounting for sample size and number of parameters in the equation. For example, in model 2 since we have 2 variables this is treated differently than a model with only 1 independent variable (like model 1). We account for the difference in parameters by putting the number of parameters in the denominator of our equation (n-p). This is because we don't want to throw in any random variable into our model and inflate our R-squared, if it doesn't improve the fit to a significant degree.