Exam 1

Kevin Linares

2024-10-07

Table of contents

1. A researcher has obtained data with 39 observations. 3
    - 1.1 Interpret the estimate for each coefficient in the output below, including the intercept.[10pt] 3
    - 1.2 Calculate a 95% confidence interval for each of the regression coefficients from 1.1 (except the intercept). Interpret each. [15pt] 4
    - 1.3 What are the null and alternative hypotheses being tested that correspond to the the t value $= -1.380$? You can write this out mathematically or in plain language. What do you conclude? [5pt] 5
    - 1.4 You decide to add a fourth predictor, x4, to the model. The ANOVA table for this model is as follows. Fill in the blanks.[5pt] 5
    - 1.5 Using the output from 1.1 and that provided below, compare the models. State which model you would prefer and your justification. [5pt] 6
    - 1.6 The overal F statistic for the model from 1.1 is 18.06 giving a p-value of $2.981 \times 10 - 7$. Provide the null hypothesis being tested (and the alternative) and state whether you would reject or fail to reject the null hypothesis. [5pt] 7
2. Use CAschools.csv from Canvas. 8
    - 2.1 Load the CAschools dataset into R (show your code) and explore what variables the dataset contains and how they are distributed (do not print the entire dataset!). [5pt] 8
    - 2.3 Fit a simple linear regression model of read as a function of income and interpret the regression coefficient for the predictor. Create a scatter plot showing the relationship with a regression line (your outcome should be on the y axis). [15pt] 11
    - 2.4 Fit a multiple linear regression model with read as a function of income students. Does the intercept make sense? If not, how might you make it more interpretable? [10pt] 13

2.5 Compare the models from 2.2 and 2.3 using R2, R2 adj , and one other appropriate method we've discussed in class. Which model do you prefer and why? [20pt] 15
2.6 Why do R2 and R2 adj have different values for the same model? [5pt] . . . 17

```r
pacman::p_load(knitr, broom, scatterplot3d, ggthemes, tidyverse)

options(scipen = 999)

# set a standard graphic theme for plots from ggthemes package
theme_set(theme_hc())
```

Notes

- Label each answer with the appropriate question number in your R Markdown document (e.g., 1.1, 1.2, etc.).
- Clearly demonstrate your work. Where applicable, include any R code pertinent to your answer.
- Submit a single pdf file via Canvas by the deadline (930am October 8).
- You may consult any reference materials except tools that utilize AI (e.g., Chat GPT, Github, Copilot, etc.)
- This is not a group assignment so do not consult with your classmates. Your answers should be based on your own, individual work.
- The point value for each question is given in square brackets.

1. A researcher has obtained data with 39 observations.

- The dataset includes 5 variables: y, x1, x2, x3, and x4.

1.1 Interpret the estimate for each coefficient in the output below, including the intercept.[10pt]

- β_0: When $x_1, ..., x_4$ equal 0, the expected mean value of outcome Y is 2.40
- β_1: For every one unit increase in x_1, the expected mean value increase of outcome Y is .02, while holding other predictors constant.
- β_2: For every one unit increase in x_2, the expected mean value decrease of outcome Y is $-.12$, while holding other predictors constant.
- β_3: For every one unit increase in x_3, the expected mean value increase of outcome Y is .15, while holding other predictors constant.

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = dat)
##
## Residuals:
## Min 1Q Median 3Q Max
## -1.9121 -0.8780 -0.1565 0.8194 3.4597
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.40123 0.73995 3.245 0.00259 **
## x1 0.02292 0.01272 1.802 0.08010 .
## x2 -0.11583 0.08391 -1.380 0.17623
## x3 0.15450 0.02495 6.192 4.31e-07 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.296 on 35 degrees of freedom
## Multiple R-squared: 0.6075, Adjusted R-squared: 0.5739
## F-statistic: 18.06 on 3 and 35 DF, p-value: 2.981e-07
```

1.2 Calculate a 95% confidence interval for each of the regression coefficients from 1.1 (except the intercept). Interpret each. [15pt]

- We construct 95% confidence intervals using the following expression:

$$\hat{\beta}_i \pm t_{n-p}^{\alpha/2} se(\hat{\beta}_i)$$

- β_1: If we were to take repeated samples from the population of the same sample size, fit the same model on each sample, and estimate 95% CI we would expect the true population β_1 to fall within -0.003 and 0.049. However, this interval contains 0, and we can conclude that there is no significant evidence of a linear relationship between predictor x_1 and y.
- β_2: If we were to take repeated samples from the population of the same sample size, fit the same model on each sample, and estimate 95% CI we would expect the true population β_2 to fall within -0.286 and 0.055. However, this interval contains 0, and we can conclude that there is no significant evidence of a linear relationship between predictor x_2 and y.
- β_3: If we were to take repeated samples from the population of the same sample size, fit the same model on each sample, and estimate 95% CI we would expect the true population β_3 to fall within 0.104 and 0.205. This interval does not contain 0, and we can conclude that there is significant evidence of a linear relationship between predictor x_3 and y.

```r
deg_fre <-  39 - 4 # sample size - parameters
t_score <- qt(p=.05/2, df=deg_fre, lower.tail=F)

q_1_mod <- tibble(predictors = c("x1", "x2", "x3"),
                  est =  c(0.02292, -0.11583, 0.15450),
                  stand_error = c(0.01272, 0.08391 , 0.02495) ) |>
  mutate(lower = est - (t_score * stand_error),
         upper = est + (t_score * stand_error)) |>
  mutate_at(2:5, round, 4)

q_1_mod |> kable()
```

predictors	est	stand_error	lower	upper
x1	0.0229	0.0127	-0.0029	0.0487
x2	-0.1158	0.0839	-0.2862	0.0545
x3	0.1545	0.0250	0.1038	0.2052

1.3 What are the null and alternative hypotheses being tested that correspond to the the t value = −1.380? You can write this out mathematically or in plain language. What do you conclude? [5pt]

- For $\hat{\beta}_2$, the Null hypothesis ($H_0 : \beta_2 = 0$) is that there is no relationship between x_2 and y, versus the alternative hypothesis ($H_a : \beta_2 \neq 0$) that there is some relationship between x_2 and y. To test the Null hypothesis we want to determine if the $\hat{\beta}_2$ coefficient is sufficiently far from 0, and to determine this we need the standard error to compute the $t - statistic$ (expression given below). Our $t - statistic = -1.380$, which measures the number of standard deviations that $\hat{\beta}_2$ is away from 0. We interpret this as it is unlikely that we would observe such a substantial association or more extreme between the predictor and the response due to chance; therefore, we reject the Null hypothesis and conclude that there is an association between x_2 and y.

$$t = \frac{\hat{\beta}_i - 0}{se\hat{\beta}_i}$$

1.4 You decide to add a fourth predictor, x4, to the model. The ANOVA table for this model is as follows. Fill in the blanks.[5pt]

- Below in the code chunk are the filled in answers for the full Anova table, and we used the MSE to compute missing F values.

```
# Response: Y      df      sum squares      mean square      F value      pr
# x1               1     0.122             0.122            0.0681    0.7957775
# x2               1     26.487          26.487           14.7463     0.0005477
# x3               1     64.447          64.447           35.8836     1.12E-06
# x4               3     1.353             0.451            0.2509    0.8601274
# residual       32    RSS = 57.478  MSE = 1.7962
```

1.5 Using the output from 1.1 and that provided below, compare the models. State which model you would prefer and your justification. [5pt]

- We compare the adjusted R^2 estimates for models 1.1 and 1.5, and find that the parsimonious model with less parameters being estimated produces a higher estimate of .574 compared to the more complicated model estimate of .545. Moreover, the parsimonious model 1.1 explains 57% of the variability in the Y outcome.
- We can also use the information from the summary outputs for both models to calculate an F-test statistic for comparison. We see that the F-ratio is less than 1 which would result in a $p - value > .05$, this means that adding the three additional categorical predictors does not improve model fit.
- After interrogating both model's adjusted R^2 and using the F statistic **we prefer the more parsimonious model** with no categorical predictors seeing as how adding more predictors only complicates the model with out much in return in terms of minimizing the least squares.

```
## Call:
## lm(formula = y ~ x1 + x2 + x3 + factor(x4), data = dat)
##
## Residuals:
## Min 1Q Median 3Q Max
## -1.7384 -0.8434 -0.0912 0.9116 3.5110
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.70560 1.72414 0.989 0.330
## x1 0.03231 0.02242 1.441 0.159
## x2 -0.12456 0.09965 -1.250 0.220
## x3 0.15297 0.02981 5.131 1.36e-05 ***
## factor(x4)2 0.56608 1.59327 0.355 0.725
## factor(x4)3 1.01602 1.76798 0.575 0.570
## factor(x4)4 0.71668 1.12949 0.635 0.530
## ---
## Signif. codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.34 on 32 degrees of freedom
## Multiple R-squared: 0.6165, Adjusted R-squared: 0.5446
## F-statistic: 8.575 on 6 and 32 DF, p-value: 1.368e-05
```

```
anova_table <- tibble(
  mod = c("mod_reduced", "mod_full"),
  RSE = c(1.296, 1.340),
  df = c(35, 32)) |>
  mutate(RSS = (RSE^2) * df)

anova_table |> kable()
```

mod	RSE	df	RSS
mod_reduced	1.296	35	58.78656
mod_full	1.340	32	57.45920

```
print(str_c( "Our F-statistic = ", round(
  ( (anova_table[[1,4]] - anova_table[2, 4]) /
      (anova_table[1, 3] - anova_table[2, 3]) ) /
    (anova_table[2, 4] / anova_table[2, 3]),
  3) |> pull(), " With a difference in DF of ",
  (anova_table[1, 3] - anova_table[2, 3]) |> pull()) )
```

```
[1] "Our F-statistic = 0.246 With a difference in DF of 3"
```

1.6 The overal F statistic for the model from 1.1 is 18.06 giving a p-value of $2.981 \times 10 - 7$. Provide the null hypothesis being tested (and the alternative) and state whether you would reject or fail to reject the null hypothesis. [5pt]

- The F statistic from model 1.1 is testing the Null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_j = 0$, while the alternative hypothesis $H_a : \beta_j \neq 0$. This means that we are testing the relationship between the Y outcome and the predictors $(\beta_1, ..., \beta_j)$, and in a multiple regression setting, we use the F-test to examine whether all model predictors are equal to 0, meaning that they to no have a relationship with the outcome variable. When there is no relationship between the outcome and predictors in the model, the F-statistic takes on a value close to 1. However, if there is a relationship, or at least one predictor is not equal to 0, than we expect the F-statistic to be > 1 and far away. In our example, an F-statistic of 18.06 is far away from 1, which is why our p-value was $< .05$, and provides compelling evidence against the Null hypothesis, therefore **we reject the Null**, and state that at least one predictor in our model is not equal to 0, and therefore related to outcome variable Y.

2. Use CAschools.csv from Canvas.

- The dataset contains data on test performance, school characteristics and student demographic back- grounds for school districts in California.
- Information on the specific variables can be found in R_California Test Score Data.pdf from Canvas

2.1 Load the CAschools dataset into R (show your code) and explore what variables the dataset contains and how they are distributed (do not print the entire dataset!). [5pt]

```
ca_schools <- read_csv(
  "~/UMD/classes/stat_mod_ML_1_SURV615/exams/CASchools.csv")

dim(ca_schools)
```

```
[1] 420  15
```

```
#head(ca_schools) # don't print
```

- The California schools data set has 45 counties represented, and 85% of schools have grades from Kinder-to-8th grade (15% from kinder-to-6th). We explore the distributions of the quantitative variables in the data set using boxplots as it allows us to visualize the ranges, outliers, median, and IQR for each variable. From the plot we can see that variables expenditure per student (mean = $5,312), percent qualifying for reduced-price lunch (mean = 45%), average math scores (mean = 653), and average reading scores (mean = 655) which all seem to be normally distributed. We can also see that percent qualifying for CalWorks, number of computers, English scores, student total enrollment, number of teachers, and district average income are left-tailed skewed, and therefore we report medians here (calworks = 10.5%; computer = 118; english = 8.8; income = $13.700; students = 951; teachers = 49). Furthermore, we see that variables students and teachers have some extreme outliers, for students we see schools with as high as 27,176 enrolled students, and for teachers a max of 1,429.

```
# how many districts
ca_schools |> distinct(county) |> count() |> kable()
```

n
45

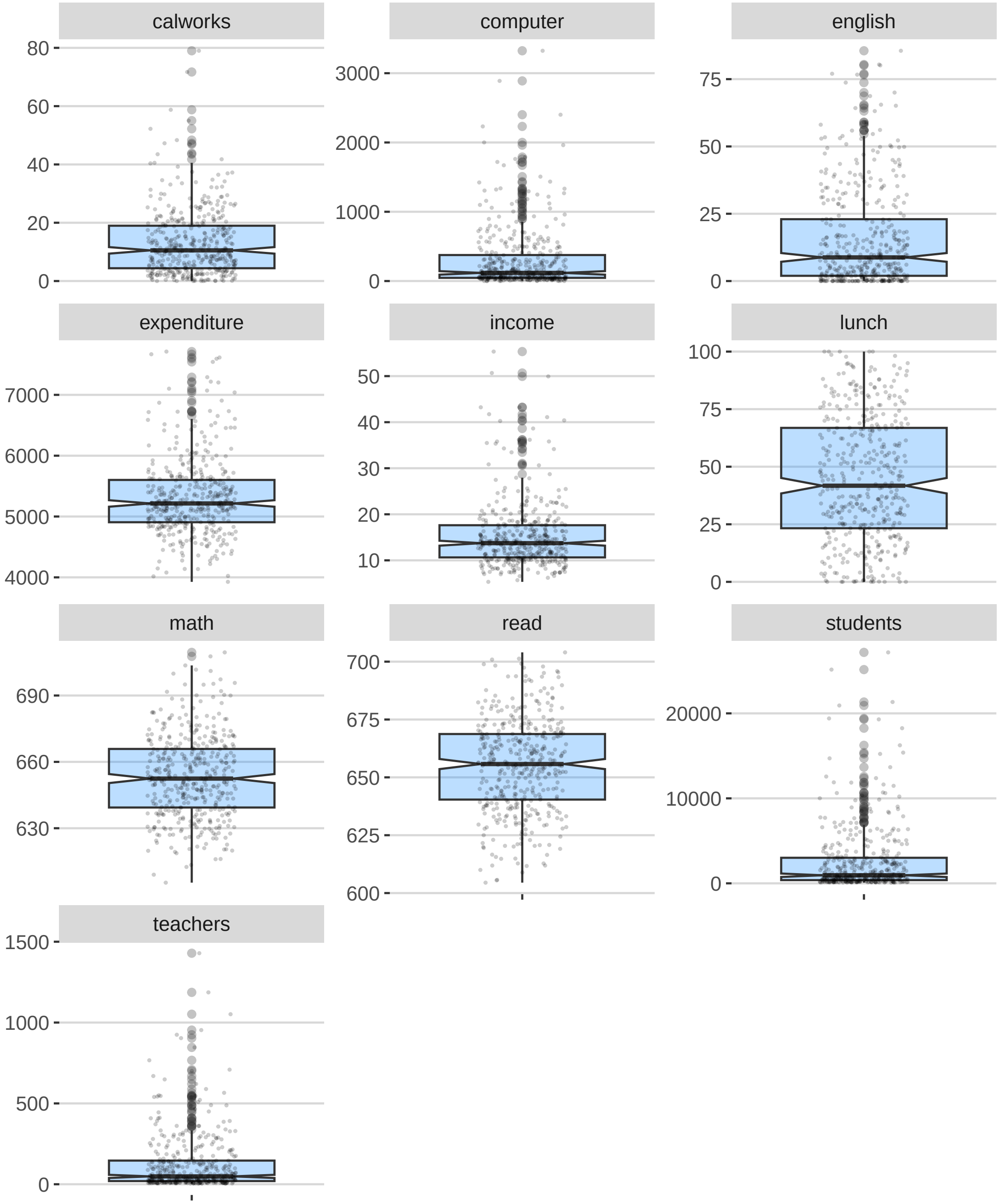
```
# school grade span?
ca_schools |>
  count(grades) |>
  mutate(percent = round(n/sum(n), 2) *100) |>
  kable()
```

grades	n	percent
KK-06	61	15
KK-08	359	85

```
# summary statistics (do not print)
# ca_schools |>
#   select(students:math) |> # select quantitative variables
#   summary()
```

```
# visualize distribution of quantitative variables
ca_schools |>
  select(rownames, students:math) |> # select quantitative variables
  pivot_longer(-rownames, names_to = "indicator", values_to="output") |>
  ggplot(aes(x="", y=output)) +
  geom_boxplot(notch=TRUE, fill="dodgerblue", alpha=.3) +
  geom_jitter(alpha=.2, size=.3, position=position_jitter(0.2)) +
  facet_wrap(~indicator, scales = "free_y", ncol=3) +
  theme(text=element_text(size=12)) +
  xlab("") + ylab("") +
  ggtitle(
    "Distribution of quantitative variables in the California schools data")
```

Distribution of quantitative variables in the California schools data



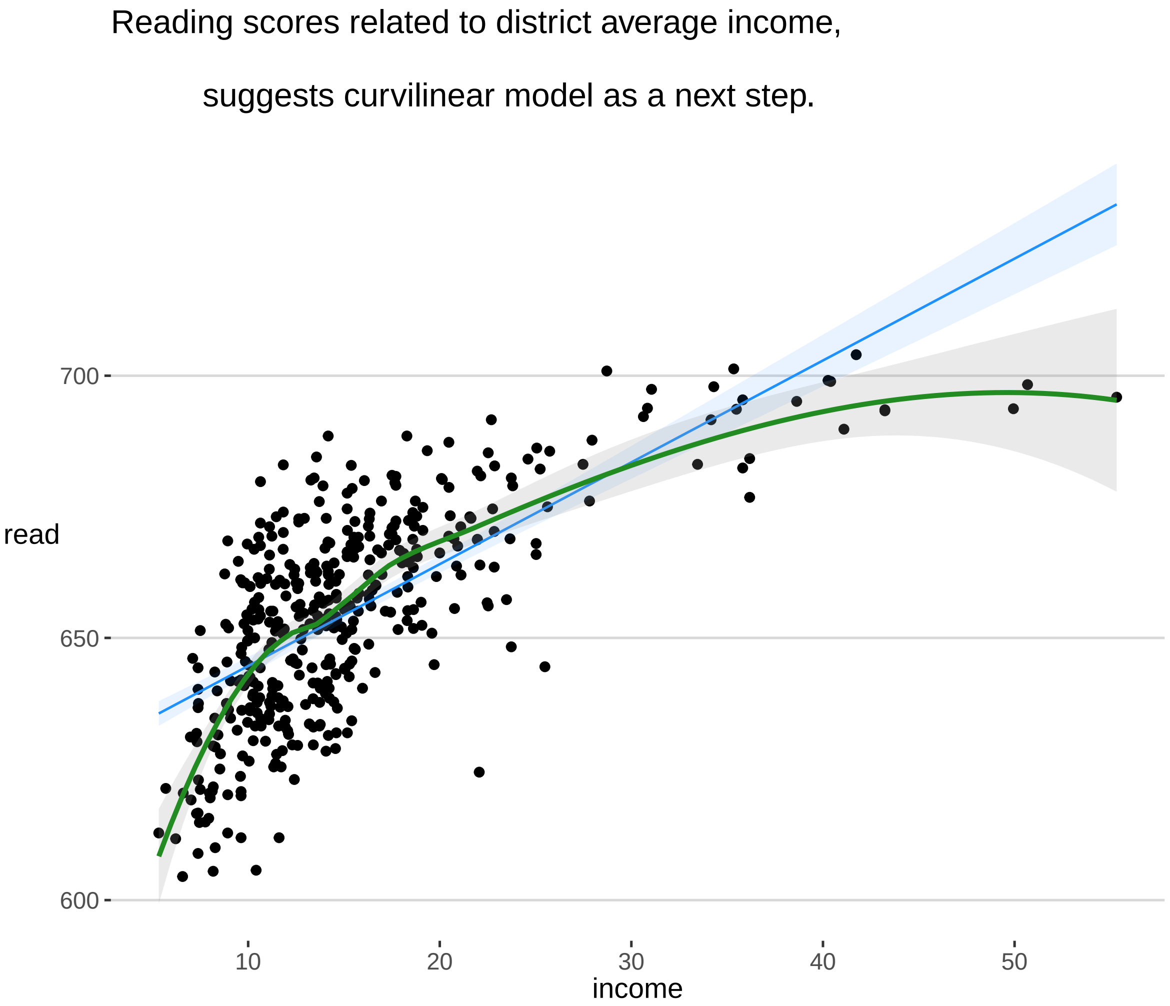
2.3 Fit a simple linear regression model of read as a function of income and interpret the regression coefficient for the predictor. Create a scatter plot showing the relationship with a regression line (your outcome should be on the y axis). [15pt]

- For every one unit increase in district average income per $\$1,000USD$, the expected mean value of school average reading test scores increases by 1.94 points. Our t-value of 19.92 suggests that β_1 is different from 0, and we reject the Null hypothesis ($\beta_1 = 0$). Furthermore, district average income explains 49% ($R^2_a = .486$) of the variation in school average reading scores. When we plot income on the x axis and read on the y axis alongside our line of best fit we see a positive linear relationship, as average income increases so do reading scores. However, we also plot 95% prediction confidence intervals and find that residuals are high in this model, meaning that our prediction error is high, and perhaps a linear regression model does not represent the observed data well. Alternatively, the forest green line in the scatterplot below posits that perhaps a curvilinear model may fit the observed data better than our current linear model.

```r
read_inc_mod  <- lm(read ~ income, ca_schools)
summary(read_inc_mod)
```

```
Call:
lm(formula = read ~ income, data = ca_schools)

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 <0.0000000000000002 ***
income        1.94187    0.09749   19.92 <0.0000000000000002 ***
---
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: < 0.00000000000000022
```

```
augment(read_inc_mod, interval = "confidence",
        conf.level = 0.95) |>
  ggplot() +
  geom_point( aes(x=income, y=read) ) +
  geom_line( aes(x=income, y=.fitted), color="dodgerblue"  ) +
  geom_ribbon(aes(x=income, ymin = .lower, ymax=.upper),
              alpha=.1, fill="dodgerblue") +
  geom_smooth(aes(x=income, y=read), method="loess",
              color="forestgreen", alpha=.2) +
  ggtitle("Reading scores related to district average income,\n
          suggests curvilinear model as a next step.")
```



Reading scores related to district average income,

suggests curvilinear model as a next step.

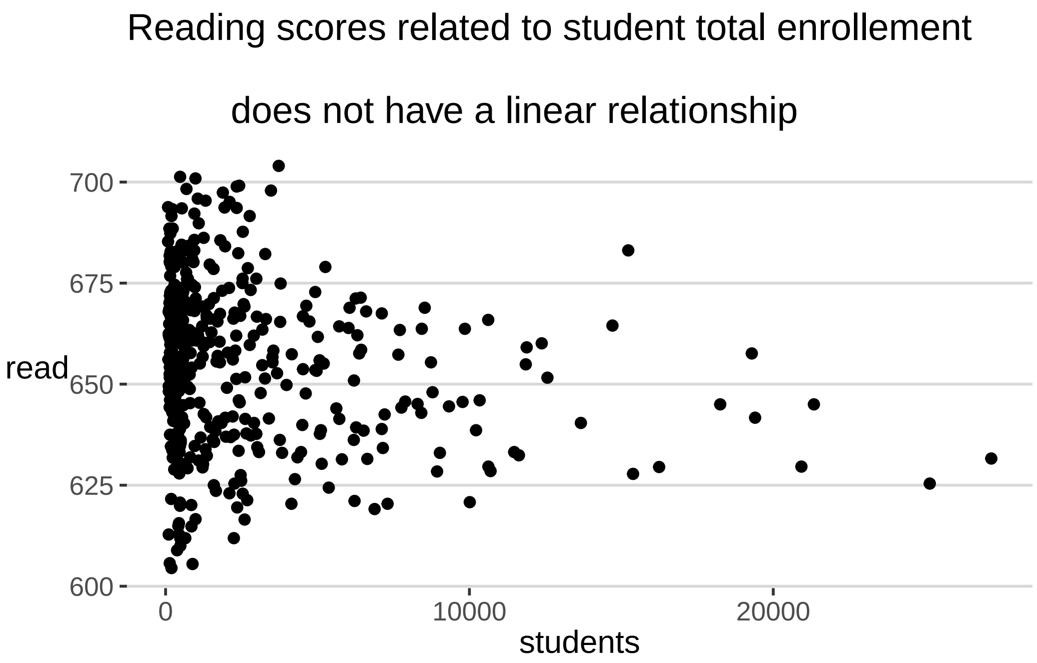
2.4 Fit a multiple linear regression model with read as a function of income students. Does the intercept make sense? If not, how might you make it more interpretable? [10pt]

- In a model with reading scores as a function of average district income and total student enrollment, the intercept β_0 is not interpretable because realistically there is no school with a district of $0 income or 0 enrolled students. We could mean center both of these variables to make the intercept interpretable, but we learned in 2.3 that these two variables are skewed. In particular, students and reading scores as can be seen in the scatterplot below do not appear to have a linear relationship. This is problematic, and we can also see it in the residuals versus fitted plot below, as we see a pattern in the residuals, which might be an indication of heteroskedasticity. What we may consider doing instead of centering students is binning them into categories and using this new categorical variable as a predictor in the model.

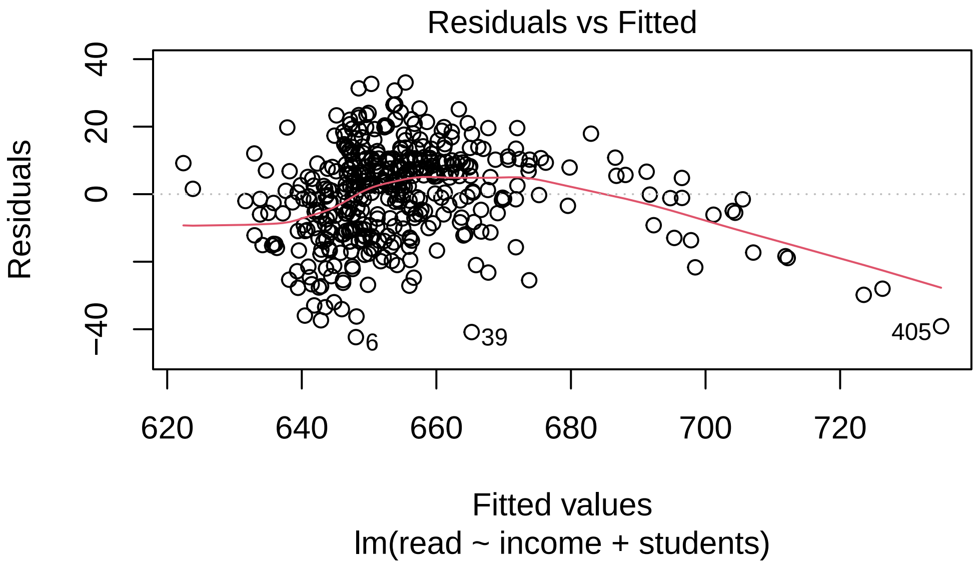
```
read_inc_stu_mod  <- lm(read ~ income + students, ca_schools)
summary(read_inc_stu_mod)
```

```
Call:
lm(formula = read ~ income + students, data = ca_schools)

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) 627.7903785   1.6342876 384.137 < 0.0000000000000002 ***
income        1.9583323   0.0934279  20.961 < 0.0000000000000002 ***
students     -0.0010708   0.0001725  -6.207        0.00000000131 ***
---
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: < 0.00000000000000022
```

```
ca_schools |> ggplot() +
  geom_point( aes(x=students, y=read) ) +
  ggtitle("Reading scores related to student total enrollement\n
          does not have a linear relationship")
```



```r
plot(read_inc_stu_mod, which=c(1))
```



2.5 Compare the models from 2.2 and 2.3 using R2, R2 adj , and one other appropriate method we've discussed in class. Which model do you prefer and why? [20pt]

- When comparing the R^2 and R_a^2 for both of these models, we see that the modes with only income explains about 49% of variation in reading scores versus the model with income and students 53%. These are not too distinct so we will use an F statistic to provide us with more direction as to which model to choose since the simple model is nested within the multiple regression model. These models differ in the use of **students**, meaning that one model includes this variable and the other does not. Therefore, the general F statistic will test whether this variable leads to a significant improvement. We reject the Null hypothesis, $H_0 : RSS_{reduced} = RSS_{full}$, $(F - statistic = 38.5, \Delta(df) = 1, p < .05)$, the model that includes students reduces the residual sum of squares more so than the model with just income as a predictor. Finally, we plot a 3d scatterplot and we see most of the observations clustering together with lots of variability around the predicted plane, or line of best fit. It is unclear how the outliers in students is affecting the prediction line as we do not see huge errors associated with them. We prefer the more parsimonious **simple linear regression model of reading scores as a function of income** given that the students variable poses modeling challenges that must be further evaluated, and we argue here that the multiple regression model is not explaining much more variability in outcome Y by including the additional variable as we saw with the R_a^2 estimates.

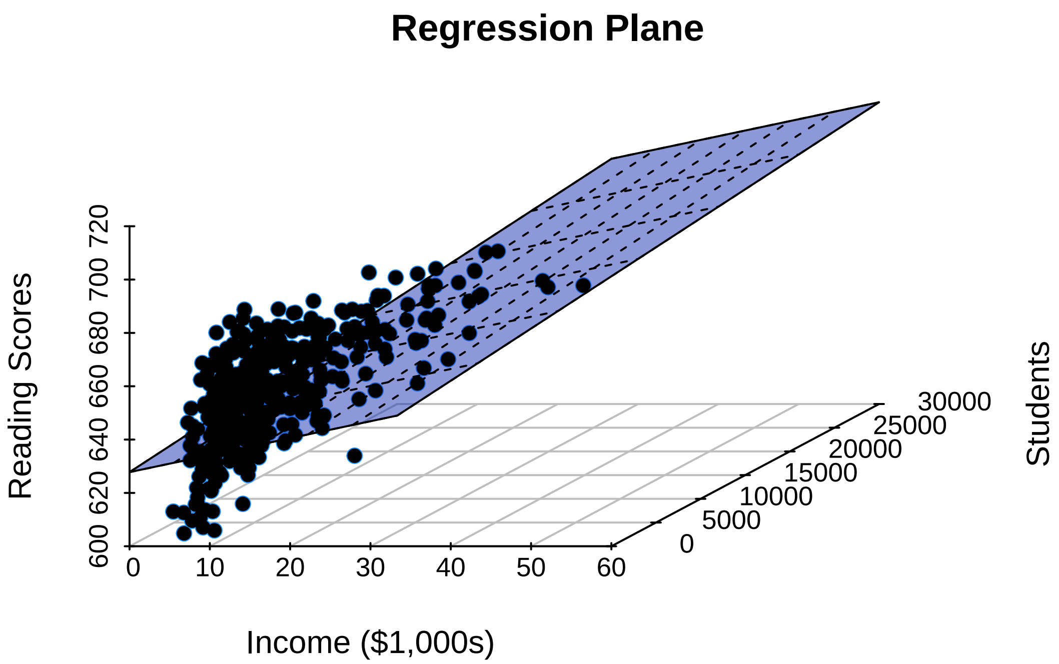
```r
glance(read_inc_mod)[c(1,2)] |>
  bind_rows(
    glance(read_inc_stu_mod)[c(1,2)]) |>
  mutate(models = c("income", "income_students")) |>
  mutate_at(1:2, round, 3) |>
  relocate(models) |>
  kable()
```

models	r.squared	adj.r.squared
income	0.487	0.486
income_students	0.530	0.528

```
anova(read_inc_mod, read_inc_stu_mod)
```

```
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 0.000000001308 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit_3_sp <- scatterplot3d(
  ca_schools$income, ca_schools$students, ca_schools$read,
  color = "dodgerblue", main = "Regression Plane", grid = TRUE, box = FALSE,
  mar = c(2.5, 2.5, 2, 1.5), angle = 40, ylab = "Students",
  xlab = "Income ($1,000s)", zlab = "Reading Scores")

fit_3_sp$plane3d(read_inc_stu_mod, draw_polygon = TRUE, draw_lines = TRUE,
                 polygon_args = list(col = rgb(.1, .2, .7, .5)))

fit_3_sp$points3d(ca_schools$income, ca_schools$students, ca_schools$read, pch=16)
```



2.6 Why do R2 and R2 adj have different values for the same model? [5pt]

- R^2 as a goodness of fit measure how well the model fits the data and it is expressed as $R^2 = 1 - RSS/TSS$. The implications for using R^2 is that by adding more variables to the model, the RSS decreases, and thus increases R^2. Therefore, adjusted R_a^2 is expressed as:

$$R_a^2 = 1 - \frac{RSS/n - p}{TSS/n - 1}$$

- where n is the sample size and p is the number of predictors being estimated. This means that R_a^2 adjusts for degrees of freedom, and by adding predictors, it will only increase R_a^2 if the additional predictor(s) has some predictive value to offer in the model. Therefore, these two measures are different for multiple regression models.