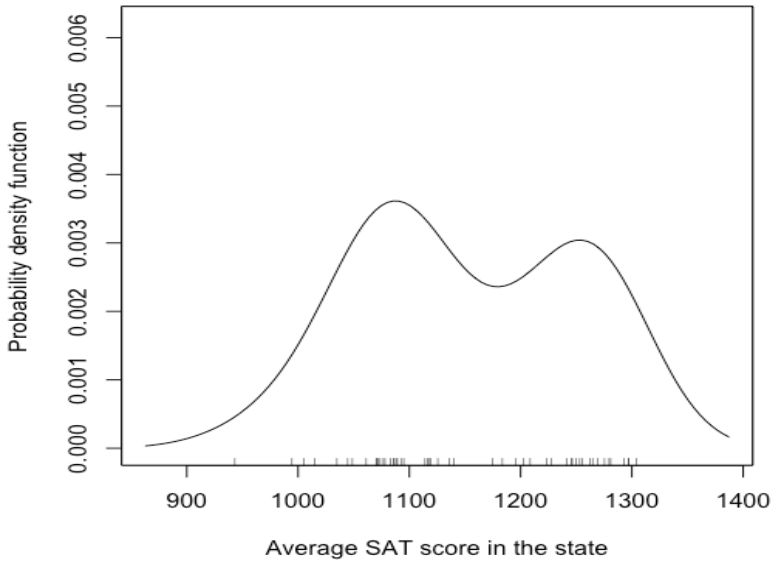
Assignment two (Ahmad Saquib Sina) #5304680

Part one

1. > sm. density (state_education $sat, xlab = "Average SAT score in the state")

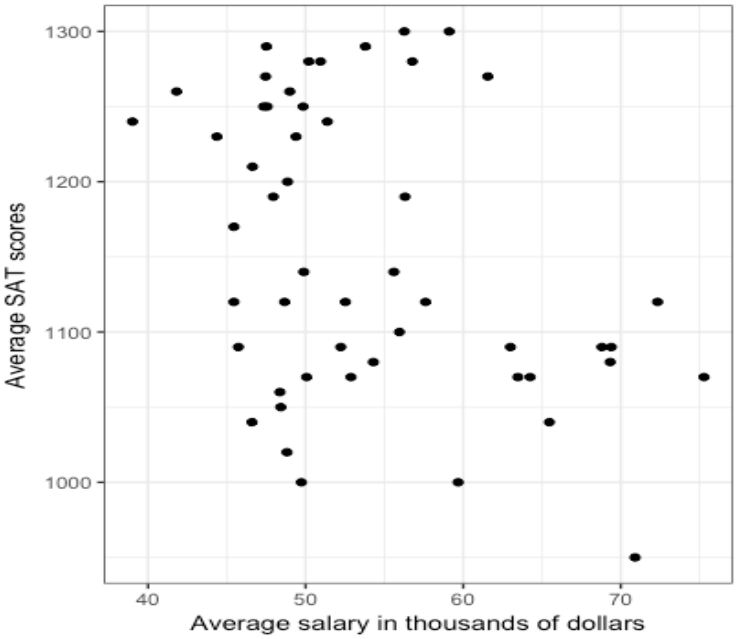


2. > educationtow = state_education %>% mutate (salary_thousand = salary/1000)
   > summary(educationtow)
   > educationtow %>% summarise (M =mean(sat), SD = sd(sat))
   > educationtow %>% summarise(M=mean(salary_thousand), SD = sd(salary_thousand))

   Descriptive statistics for SAT scores and salaries variables

	Mean	SD	Min.	Max.
Explanatory variable				
Average SAT scores	1153.529	96.66	950	1300
Response variable				
Salary_thousand	53.987	8.5063	39.02	75.28

3. Plot of the distribution of SAT score conditioned on teacher salary
   > ggplot (educationtow, aes (x= salary_thousand, y= sat)) + geom_point () + theme_bw () + xlab ("Average salary in thousands of dollars") + ylab ("Average SAT scores")



4. The relationship between the SAT scores and teacher salaries are weak, negative, and linear.
   Here, there are some outliers which are clustered, which makes the weak linear relation between SAT scores and teacher salaries.

5. Pearson correlation coefficient

   > educationtow %>% select (sat, salary_thousand) %>%
   + correlate ()
   # A tibble: 2 x 3

	rowname	sat	salary_thousand
	<chr>	<dbl>	<dbl>
1	sat	NA	-0.3863918
2	salary_thousand	-0.3863918	NA

   Therefore, the Pearson correlation coefficient is
   $r_{salary\_thousand,\ sat}$ = - 0.3863918

6. Yes, the Pearson correlation coefficient is an appropriate summary measure of the relationship. The value is low and that is negative. In the question 4, I have noticed that the relationship is linear but it is weak also. Therefore, Pearson correlation coefficient is an appropriate summary measure of the relationship.

7. > lm.1 = lm (sat ~ 1 + salary_thousand, educationtow)
   > lm.1

   Call:
   lm (formula = sat ~ 1 + salary_thousand, data = educationtow)

   Coefficients:

Intercept	salary_thousand
1390.570	-4.391

   sat = 1390.570 – 4.391 (salary_thousand)

8. Interpreting the value of the intercept and the slope from the regression equation using the context of the data

   The estimate for the intercept was 1390.570. Graphically, this value indicates the y-value where the line passes through the y-axis (i.e., y-intercept). As such, it gives the predicted value of Y when X = 0. Algebraically we get the same thing if we substitute 0 in for X_i in the estimated regression equation.

   $$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1(0)$$

   $$\hat{Y}_i = \hat{\beta}_0$$

   The predicted average SAT scores for all students is 1390.570 when the average salary in thousands of dollars is Zero (0).

   Slope Interpretation

   Recall from algebra that the slope of a line describes the change in Y versus the change in X. In regression, the slope describes the *predicted* change in $\hat{Y}$ for a one-unit difference in X. Therefore, each one unit difference in SAT scores is associated with a $ 4.391 predicted difference in salary_thousand and that is negative

9. Computing, reporting, and interpreting the value for R square based on values from the ANOVA decomposition
   Call:
   lm (formula = sat ~ 1 + salary_thousand, data = educationtow)

   Coefficients:

Intercept	salary_thousand
1390.570	-4.391

> sse.1 = sum ((educationtow$sat - (1390.570 - 4.391 * educationtow$salary_thousand)) ^2)
> sse.1
[1] 397417.7
> lm.0
Call:
lm (formula = sat ~ 1, data = educationtow)
Coefficients:
(Intercept)
1154
> sse.0 = sum ((educationtow$sat - 1154) ^ 2)
> sse.0
[1] 467176
> sse.0 - sse.1
[1] 69758.31
> (sse.0 -sse.1)/sse.0
[1] 0.1493191

So, PRE = 0.1493191
So, R^2 = 0.1493191

10. > View (educationtow)
> MEAN = mean(educationtow$salary_thousand)
> M9 = educationtow %>% mutate (center_salary_thousand = (salary_thousand - MEAN))
> MCENTEREDSALARY = M9 %>% filter (state == "Minnesota")
> MCENTEREDSALARY
The value of Minnesota's centered salary is $2.280961.

11. > mean(M9$center_salary_thousand)
[1] 1.531982e-15
> sd(M9$center_salary_thousand)
[1] 8.50639
> M9 %>% select (sat, center_salary_thousand) %>%
+ correlate ()
# A tibble: 2 x 3

	rowname	sat	center_salary_thousand
	<chr>	<dbl>	<dbl>
1	sat	NA	-0.3863918
2	center_salary_thousand	-0.3863918	NA

Here, we have found that the SD and Pearson correlation coefficient are same for centered and un-centered teacher salary. However, the mean is difference for each other.

12) > sd(educationtow$sat)
[1] 96.66072

> 96.66072 / 8.50639

[1] 11.36331

> 11.36331 * (-0.3863918)
[1] -4.39069

here, we have found that the slope of the regression I have found in #Question 7 is similar of the regression if we regress SAT scores on the centered teacher salaries by making reference to the values in the mathematical formula for slope, which is:

$r \times (SD_{outcome} / SD_{predictor})$

13. > lm.5 = lm (sat ~ 1 + center_salary_thousand, data = M9)
> lm.5
Call:
lm (formula = sat ~ 1 + center_salary_thousand, data = M9)

Coefficients:

Intercept	center_salary_thousand
1153.529	-4.391

sat = 1153.529 – 4.391 (center_salary_thousand)

14. Interpreting the value of the intercept and the slope from the regression equation using the context of the data
The estimate for the intercept was 1153.529. Graphically, this value indicates the y-value where the line passes through the y-axis (i.e., y-intercept). As such, it gives the predicted value of Y when X = 0. Algebraically we get the same thing if we substitute 0 in for X_i in the estimated regression equation.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1(0)$$

$$\hat{Y}_i = \hat{\beta}_0$$

The predicted average SAT scores for all students is 1153.529 when the center_salary_thousand is Zero (0).

Slope Interpretation Recall from algebra that the slope of a line describes the change in Y versus the change in X. In regression, the slope describes the *predicted* change in $\hat{Y}$ for a one-unit difference in X.

Therefore, each one unit difference in SAT scores is associated with a $ 4.391 predicted difference in salary_thousand and that is negative

15.

> mean(educationtow$salary_thousand)

[1] 53.98704

> sd(educationtow$salary_thousand)

[1] 8.50639

> mean(educationtow$sat)

[1] 1153.529

> sd(educationtow$sat)

[1] 96.66072

> M16 = mean(educationtow$salary_thousand)

> SD16 = sd(educationtow$salary_thousand)

> M17 = mean(educationtow$sat)

> SD17 = sd(educationtow$sat)

> educationseventeen = educationfifteen %>% mutate (z_sat = (sat - M17)/SD17, z_salary = (salay_thousand – M16))

> lm.8 = lm (z_sat ~ 1 + z_salary, educationseventeen)

> lm.8

Call:

lm (formula = z_sat ~ 1 + z_salary, data = educationseventeen)

Coefficients:

Intercept z_salary

3.501e-16 -3.864e-01

The regression equation is,

z_sat = 3.501e-16 - 3.864e-01 (z_salary)

= 3.501e-16 – 0.3864 (z_salary)

= 0 - 0.3864 (z_salary)

The estimate for the intercept was 0 (Zero). Graphically, this value indicates the y-value where the line passes through the y-axis (i.e., y-intercept). As such, it gives the predicted value of Y when X = 0. Algebraically we get the same thing if we substitute 0 in for X_i in the estimated regression equation.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1(0)$$

$$\hat{Y}_i = \hat{\beta}_0$$

The predicted z_sat scores for all students is 0 when the z_salary is Zero (0).

Slope Interpretation

Recall from algebra that the slope of a line describes the change in Y versus the change in X. In regression, the slope describes the *predicted* change in $\hat{Y}$ for a one-unit difference in X.

Therefore, each one unit difference in z_sat scores is associated with a 0.3864 predicted difference in z_salary and that is negative.

16. > sd(educationseventeen$z_salary)

[1] 1

> sd(educationseventeen$z_sat)

[1] 1

Using the mathematical formula for slope from Question # 12, we have found that

- 0.3864 x (1/1)

= - 0.3864

That is why, the slope from regressing $z_{outcome}$ or $z_{predictor}$ will be the correlation coefficient between the predictor and otucome