565_HW2

Matthew Stoebe

2025-04-04

#Question1

a. The response variable is the student's reading scores. The predictor of interest is the amount of weekly average screen time. The hypothesis is that students who spend too much time on screens struggle to read.

```r
tvread <- read.table("Data/tvread.txt", header=TRUE)
model1 <- lm(score ~ hours, data = tvread)
summary(model1)
```

```
##
## Call:
## lm(formula = score ~ hours, data = tvread)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.449 -18.515   3.951  18.617  46.284
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  66.9813    10.4736   6.395 9.72e-08 ***
## hours         1.6667     0.5297   3.147    0.003 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.51 on 43 degrees of freedom
## Multiple R-squared:  0.1872, Adjusted R-squared:  0.1683
## F-statistic: 9.901 on 1 and 43 DF,  p-value: 0.002997
```

b. There is a significant relationship between hours spent on screens and the student's reading scores. We see a coefficient of 1.667 which means that for each hour of screen time, th reading score increasees by 1.66. This is suprising

c.

```r
model2 <- lm(score ~ hours + grade, data = tvread)
summary(model2)
```

```
##
## Call:
```

```
## lm(formula = score ~ hours + grade, data = tvread)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.1755  -5.7624  -0.0756   6.4046  17.1842
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  45.9214     4.0579  11.317 2.46e-14 ***
## hours        -0.8229     0.2460  -3.346  0.00174 **
## grade        21.2100     1.2779  16.598  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.387 on 42 degrees of freedom
## Multiple R-squared:  0.8925, Adjusted R-squared:  0.8874
## F-statistic: 174.3 on 2 and 42 DF,  p-value: < 2.2e-16
```

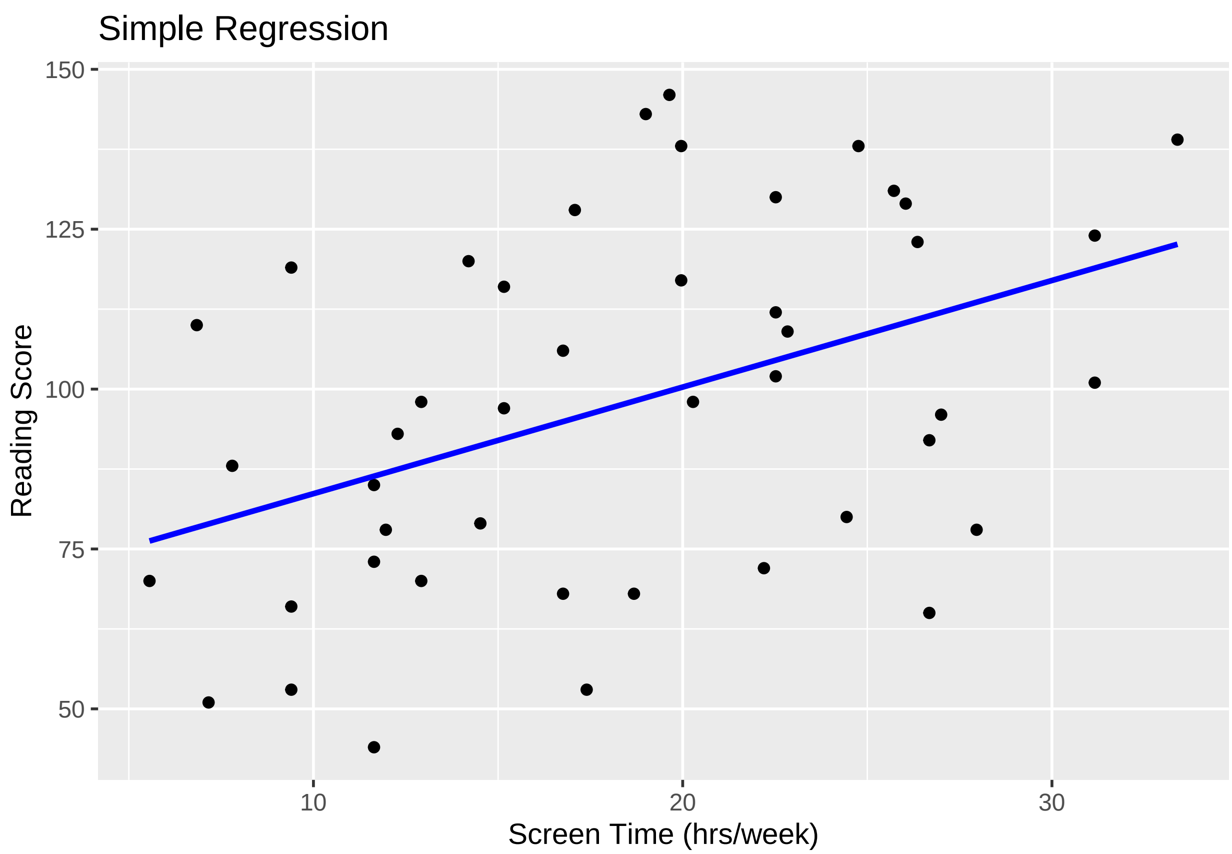
c. When we include grade of child, we see that the relationship between reading score and hours of screen time flipps. More specifically, for each hour spent on a screen, holding the student's grade constant, the students test score decreases by .8229 points. This aligns with the expected hypothesis. and makes sense as students in higher grades may have both better reading scores, and more screentime leading to confounding.

```r
library(ggplot2)
library(gridExtra)

p1 <- ggplot(tvread, aes(x = hours, y = score)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Simple Regression",
       x = "Screen Time (hrs/week)",
       y = "Reading Score")

p1
```

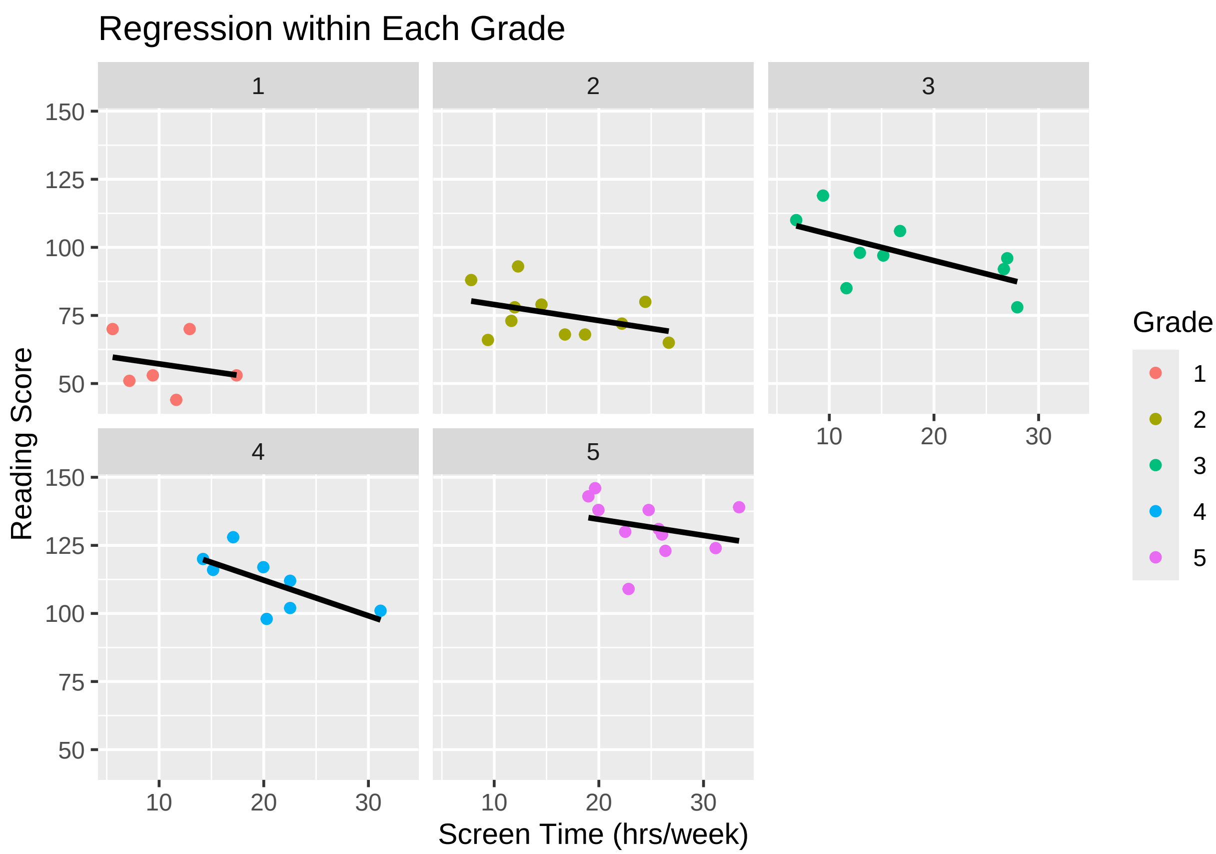
```
## `geom_smooth()` using formula = 'y ~ x'
```




```
p2 <- ggplot(tvread, aes(x = hours, y = score)) +
  geom_point(aes(color = factor(grade))) +
  geom_smooth(method = "lm", se = FALSE, color = "black") +
  facet_wrap(~ grade) +
  labs(title = "Regression within Each Grade",
       x = "Screen Time (hrs/week)",
       y = "Reading Score",
       color = "Grade")

p2
```

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



d. While it appears that at an aggregate level that more screen time leads to higher levels, this does not account for the fact that students in higher grade levels both spend more time on their devices and are better readers than students at lower grade levels. This confounds ouroutput. However, after controlling for the student's grade level, we see that increased screen time does actually lead to worse test scores.

```r
studdybuddy <- read.table("Data/studdybuddy.txt", header=TRUE)

t.test(sat ~ sb, data = studdybuddy)
```

```
##
##  Welch Two Sample t-test
##
## data:  sat by sb
## t = -2.9606, df = 92.627, p-value = 0.0039
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -226.4401  -44.6176
## sample estimates:
## mean in group 0 mean in group 1
##        1000.846        1136.375
```

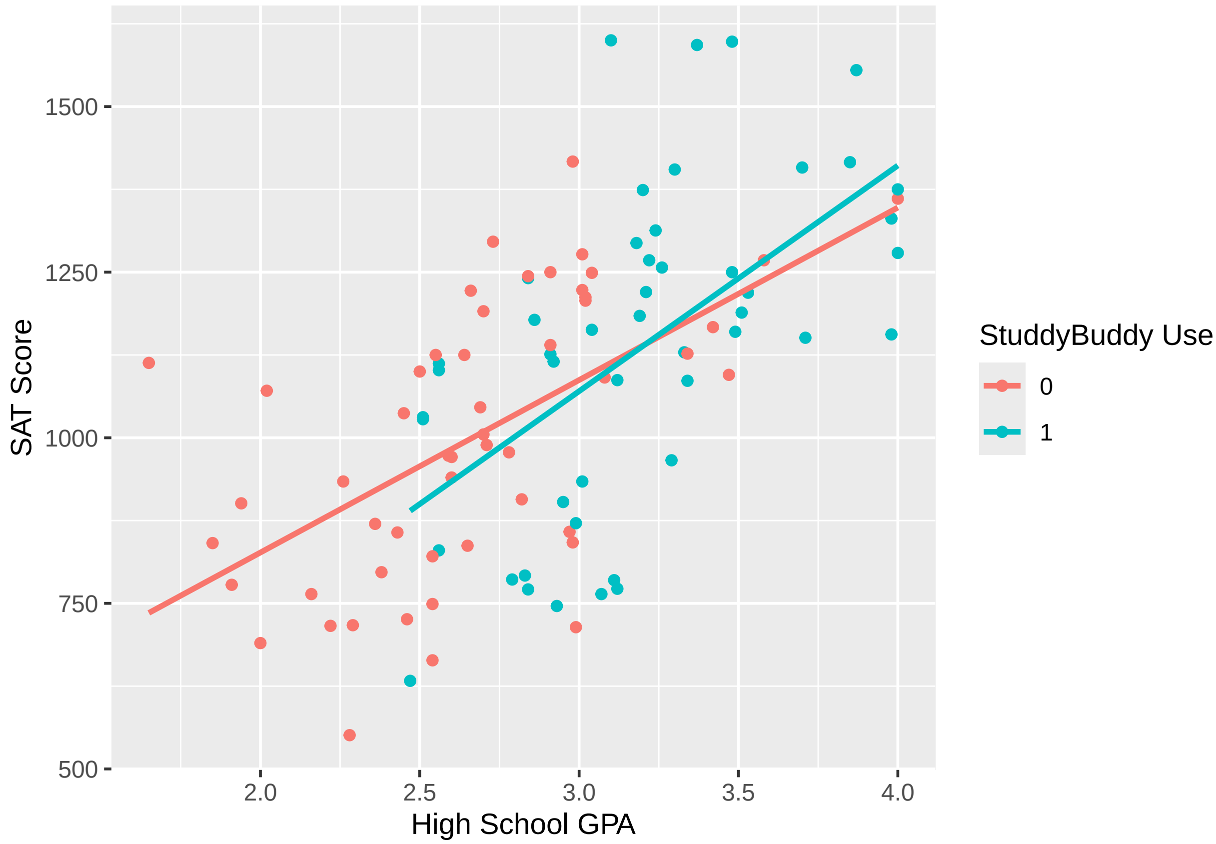
a. There is a significant difference.

```r
model_sat <- lm(sat ~ sb + gpa, data = studdybuddy)
summary(model_sat)
```

```
##
## Call:
## lm(formula = sat ~ sb + gpa, data = studdybuddy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -381.79 -139.38   -4.86  126.10  491.39
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   212.37     117.90   1.801   0.0748 .
## sb            -19.68      43.82  -0.449   0.6544
## gpa           295.46      43.09   6.856 6.58e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 187.4 on 97 degrees of freedom
## Multiple R-squared:  0.3824, Adjusted R-squared:  0.3697
## F-statistic: 30.03 on 2 and 97 DF,  p-value: 7.066e-11
```

b. There is no signifficant effect of study-budy use on sat score after controlling for student GPA.

```r
ggplot(studdybuddy, aes(x=gpa, y=sat, color=factor(sb))) +
  geom_point() +
  geom_smooth(method="lm", formula = y ~ x, se=FALSE) +
  labs(color = "StuddyBuddy Use", x="High School GPA", y="SAT Score")
```



c. The first model does not account for existing GPA. Students with higher GPAs are more likely to use study buddy and also have higher SAT scores. As such, this is a confoudder and creates the illusion that study buddy use increases sat scores. When we control for study buddy usage, we see that the regression lines are almost parallel implying that study-buddy users do not perform better than non study-buddy users.