Report 2

Vanessa Machuca and Luis Espino

2/12/2018

Our hypothesis is that a linear relationship exists between number of absences (absences) and final grade (G3). More specifically, we'd predict that a negative relationship exists between the two.

UNTRANSFORMED VARIABLES

absences vs. final grade plot using ggplot

```
ggplot(student_por, aes(absences, G3)) + geom_point()
```

absences vs. final grade plot and residual vs. predicted using xyplot

```
G3_lm <- lm(G3 ~ absences, data=student_por)
xyplot(G3 ~ absences, data=student_por, type=c("p", "r"), pch=18)
xyplot(rstandard(G3_lm) ~ fitted(G3_lm), pch=18)
```

Strange. The second plot seems to be the mirror immage of the first. Am I doing something wrong? Anyway, there seems to be a negative linear relatinoshios between number of absences and final grade. It is slight, though. ***continue checking assumptions. figure out residual xyplot.***

99% CI for slope parameter β_1

```
tidy(G3_lm, conf.int = TRUE, conf.level = 0.99)
```

The CI is (-0.134, 0.0067). It contains zero, so we cannot be confident that the population slope is not zero. That is, there may not be a linear relationship between number of absences and final grade. Let's try transforming the variables, and then removing some outliers.

TRANSFORMED VARIABLES

```
student_por2<-filter(student_por, absences > 0 & G3 > 0)
ggplot(student_por2, aes(log(absences), G3)) + geom_point()

G3_lm2 <- lm(G3 ~ log(absences), data=student_por2)
tidy(G3_lm2, conf.int = TRUE, conf.level = 0.99)
```

```
ggplot(student_por2, aes(absences, log(G3))) + geom_point()

G3_lm3 <- lm(log(G3) ~ absences, data=student_por2)
tidy(G3_lm3, conf.int = TRUE, conf.level = 0.99)
```

Mmmmm, I don't like this. Let's try filtering out outliers!

REMOVING OUTLIERS

Let's remove the final grades for student who were absent 30 or more times - 2 outliers in total.

```r
student_porfiltered<- filter(student_por, absences < 30)
ggplot(student_porfiltered, aes(absences, G3)) + geom_point()
filteredG3_lm <- lm(G3 ~ absences, data=student_por_28)
tidy(filteredG3_lm, conf.int = TRUE, conf.level = 0.99)
```

The 99% CI for the population slope is now (-0.1591,-0.0104). This interval does not contain 0, so we can be confident that the population slope is not zero - there is a linear relationship between number of absences and final grade. Because we looked at a subset of the explanatory variables, though, the model is only appropriate for fewer than 30 absences.

ASSESS FIT OF MODEL W/O OUTLIERS