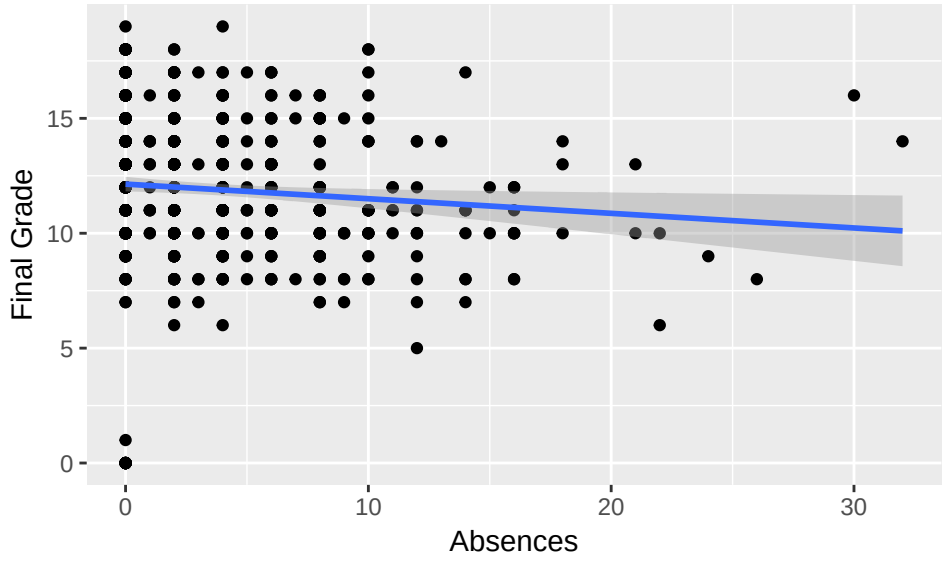
Report 2

Vanessa Machuca and Luis Espino

2/12/2018

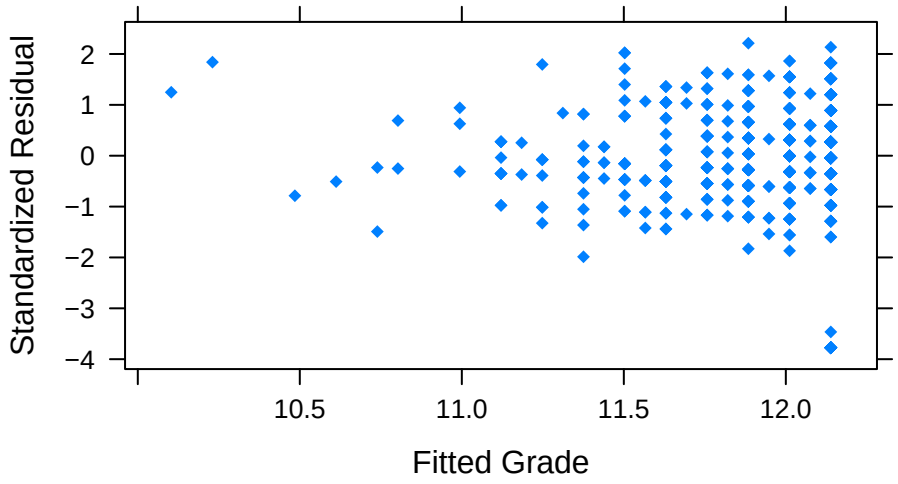
The data we are working with looks at student achievement a portuguese course at two secondary education Portuguese schools. We found the data through the UCI Machine Learning Repository. The observational units for the dataset are students. Our hypothesis is that a linear relationship exists between number of absences (absences) and final grade (G3). More specifically, we'd predict that a negative relationship exists between the two.

We will start out by looking at a plot of the explanatory variable (absences) vs. the reponse variable (G3)



and standardized residual vs. fitted grade.

```
G3_lm <- lm(G3 ~ absences, data=student_por)
xyplot(rstandard(G3_lm) ~ fitted(G3_lm), pch=18,xlab="Fitted Grade",ylab="Standardized Residual")
```



There seems to be a negative linear relationship between number of absences and final grade. It is slight, though. To further investigate this possible relationship, we compute a 99% CI for slope parameter β_1.

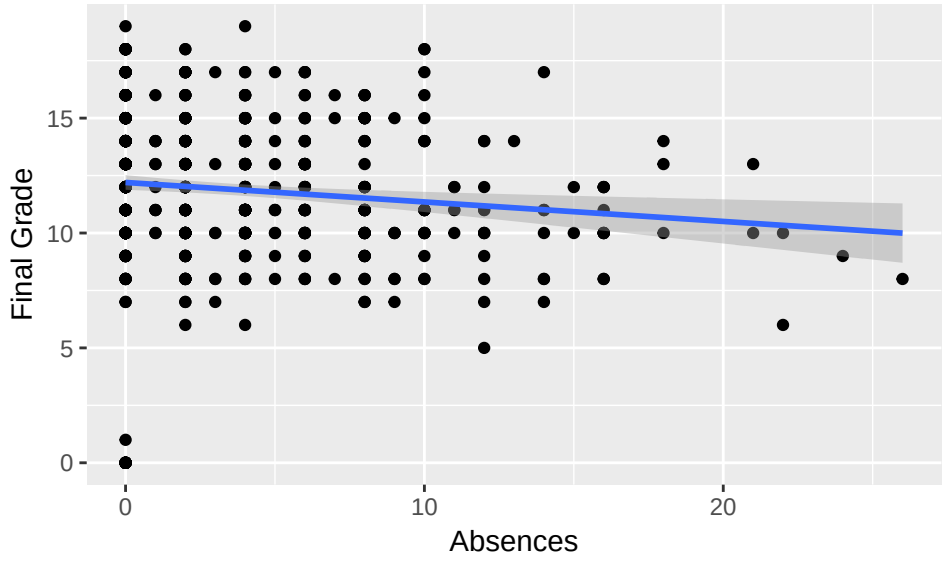
```
tidy(G3_lm, conf.int = TRUE, conf.level = 0.99)
```

```
##          term    estimate  std.error statistic       p.value   conf.low
## 1 (Intercept) 12.13880086 0.16099478 75.398725 1.086944e-322 11.7228790
## 2    absences -0.06361337 0.02725391 -2.334101  1.989562e-02 -0.1340225
##      conf.high
## 1 12.554722726
## 2  0.006795722
```

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.

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

```
student_porfiltered<- filter(student_por, absences < 30)
ggplot(student_porfiltered, aes(absences, G3)) + geom_point() + geom_smooth(method='lm') + labs(x="Abse
```




```
filteredG3_lm <- lm(G3 ~ absences, data=student_porfiltered)
tidy(filteredG3_lm, conf.int = TRUE, conf.level = 0.99)
```

```
##          term    estimate  std.error statistic       p.value   conf.low
## 1 (Intercept) 12.19942064 0.16286732  74.90404 2.025669e-320 11.7786573
## 2    absences -0.08474931 0.02877794  -2.94494  3.346507e-03 -0.1590964
##     conf.high
## 1 12.62018398
## 2 -0.01040226
```

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.

The data seem to be fairly symmetric about the model line, but lacks constant variability. We will now compute a prediction interval for an individual response at an interesting x value. Let's look at x=15.

```
newdata = data.frame(absences=15)
predict(filteredG3_lm, newdata, interval="predict")
```

```
##        fit      lwr      upr
```

```
## 1 10.92818 4.583994 17.27237
```

For a student with 15 absences, we get a predicted final grade of 10.9 and prediction interval of (4.58,17.3). This is a very wide interval, as can be expected both of a prediction interval (relative to a CI) and from looking at a relatively high x-value. Now to find the R^2.

```
summary(filteredG3_lm)$r.squared
```

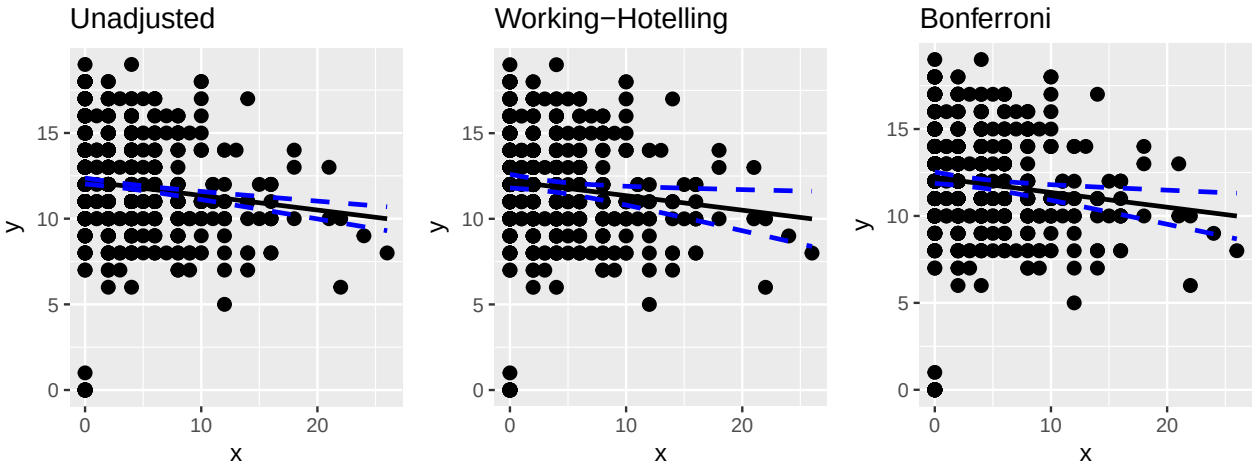
```
## [1] 0.0132676
```

We get an R^2 value of 0.013. This tells us that number of absences explains 1.3% of variability in final grades - quite a small percentage.

We expected a stronger linear relationship between number of absences and final grade. It's safe to say that number absences is likely related to other variables in the data set. In the future, then, we'd like to explore colinearity between absences and other variables like travel time, number of past class failures, and quality of family relationships. We are also interested in exploring the relationship between parent education and student school performance. The former may indicate whether or not a student who grew up in a family environment that promotes education, which might affect that student's performance.

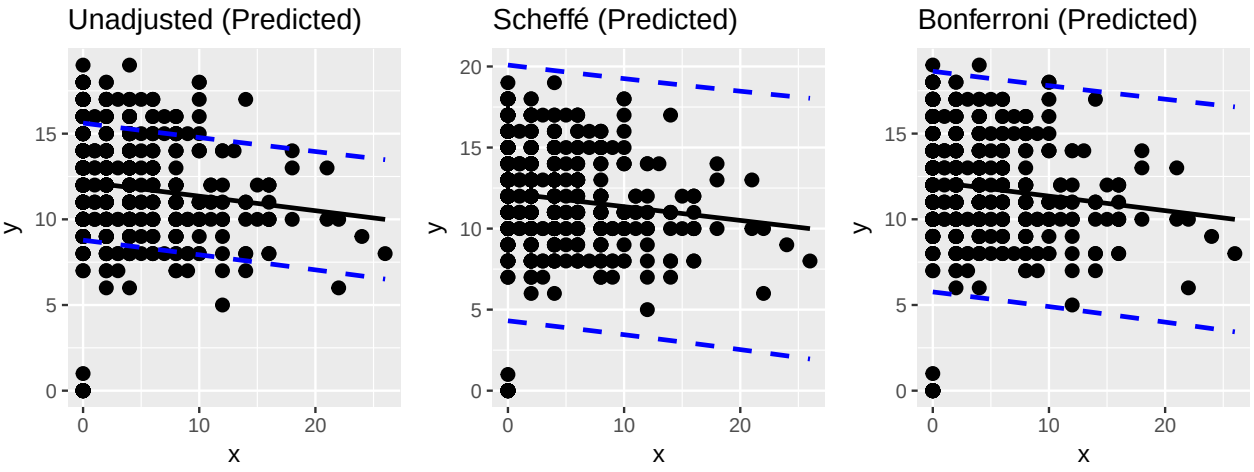
Simultaneous Inference

We will now carry out simultaneous inference on β_0 and β_1, utilizing a function from https://rpubs.com/aaronsc32/simultaneous-confidence-intervals. The plots below include bands for mean intervals using the Bonferroni method, Working-Hotelling method, and unadjusted CIs.




```
##            N     W     B
## Result 1.063 2.453 2.002
```

The plots below include bands for prediction intervals using the Bonferroni method, Scheffé method, and unadjusted PIs.




```
##            N     W     B
## Result 1.063 2.453 2.002
```

The level of significance, α, and power for a single test differs from the level of significance and power for a family of tests. For example, t-statistics based on the same sample data and MSE will be dependent on each other. Adjusting for multiple comparisons allows us to more accurately determine the actual level of significance and power for a family of tests and thus decreases the likelihood of type I errors. Thus, simultaneous inference procedues like Bonferroni are preferable when considering multiple comparisons. The Bonferroni bands are the narrower than the Scheffé and Working-Hotelling bands and, by the Bonferroni inequality, is 90% confident. We might, then, prefer the Bonferroni bands over the others.