exam1

Sunny Lee

2/26/2021

1)

```
grade=c(seq(40,60,5),seq(50,70,5),seq(60,80,5),seq(70,90,5))
studytime=c(rep(5,5),rep(10,5),rep(15,5),rep(20,5))
sleeptime=5+sample.int(5, 20, replace = TRUE)
newdf=data.frame(studytime,sleeptime,grade)
```

a)

```
cor(grade, studytime)
```

```
## [1] 0.8451543
```

```
cor(grade, sleeptime)
```

```
## [1] 0.3498243
```

Using the cor() command, we find that grade is very heavily correlated to studytime and not so much with sleeptime.

b)

```
library("ggpubr")
```

```
## Loading required package: ggplot2
```

```
modelstudy <- lm(grade~studytime)
modelsleep <- lm(grade~sleeptime)
summary(modelstudy)
```

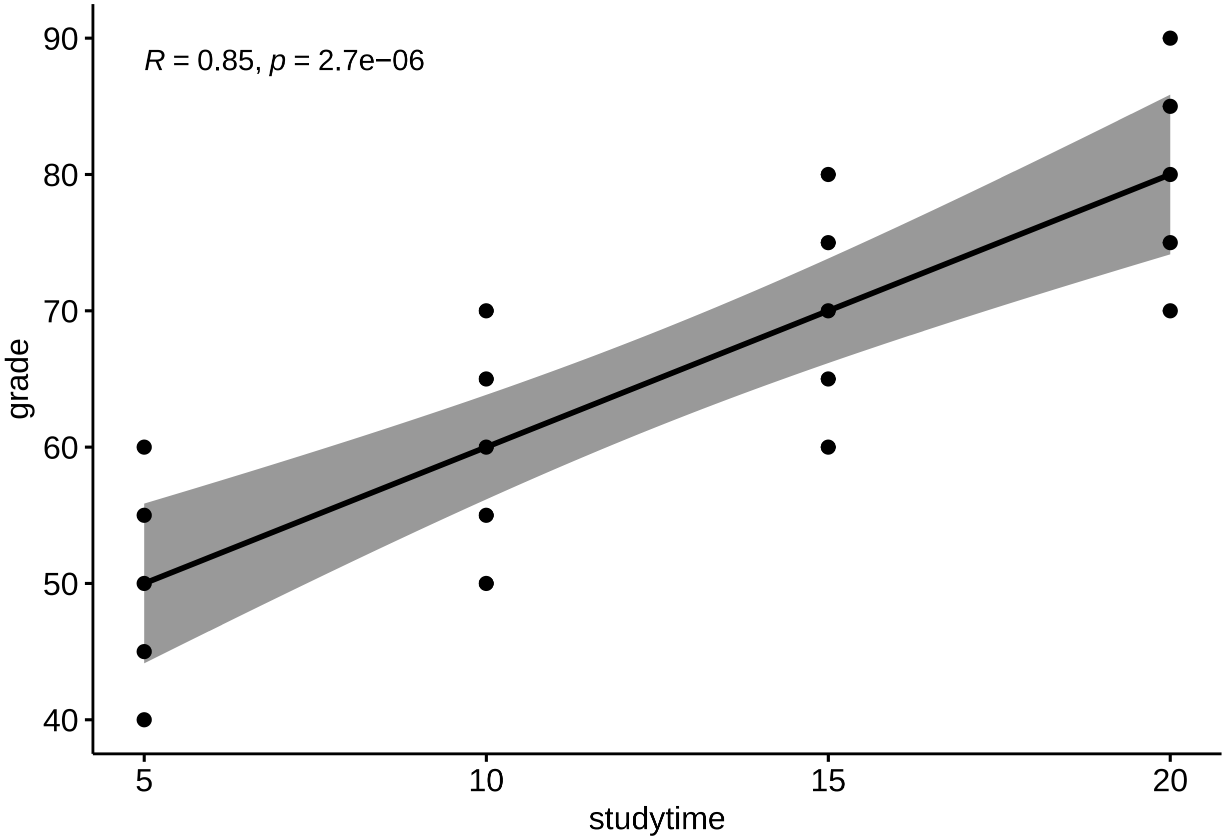
```
##
## Call:
## lm(formula = grade ~ studytime)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##    -10     -5      0      5     10
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  40.0000     4.0825   9.798 1.22e-08 ***
## studytime     2.0000     0.2981   6.708 2.73e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.454 on 18 degrees of freedom
## Multiple R-squared:  0.7143, Adjusted R-squared:  0.6984
## F-statistic:    45 on 1 and 18 DF,  p-value: 2.731e-06
```

```
summary(modelsleep)
```

```
##
## Call:
## lm(formula = grade ~ sleeptime)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.1105  -7.0241   0.4759   8.2168  20.7168
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   40.729     15.596   2.611   0.0177 *
## sleeptime      3.173      2.003   1.584   0.1305
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.06 on 18 degrees of freedom
## Multiple R-squared:  0.1224, Adjusted R-squared:  0.07362
## F-statistic:  2.51 on 1 and 18 DF,  p-value: 0.1305
```

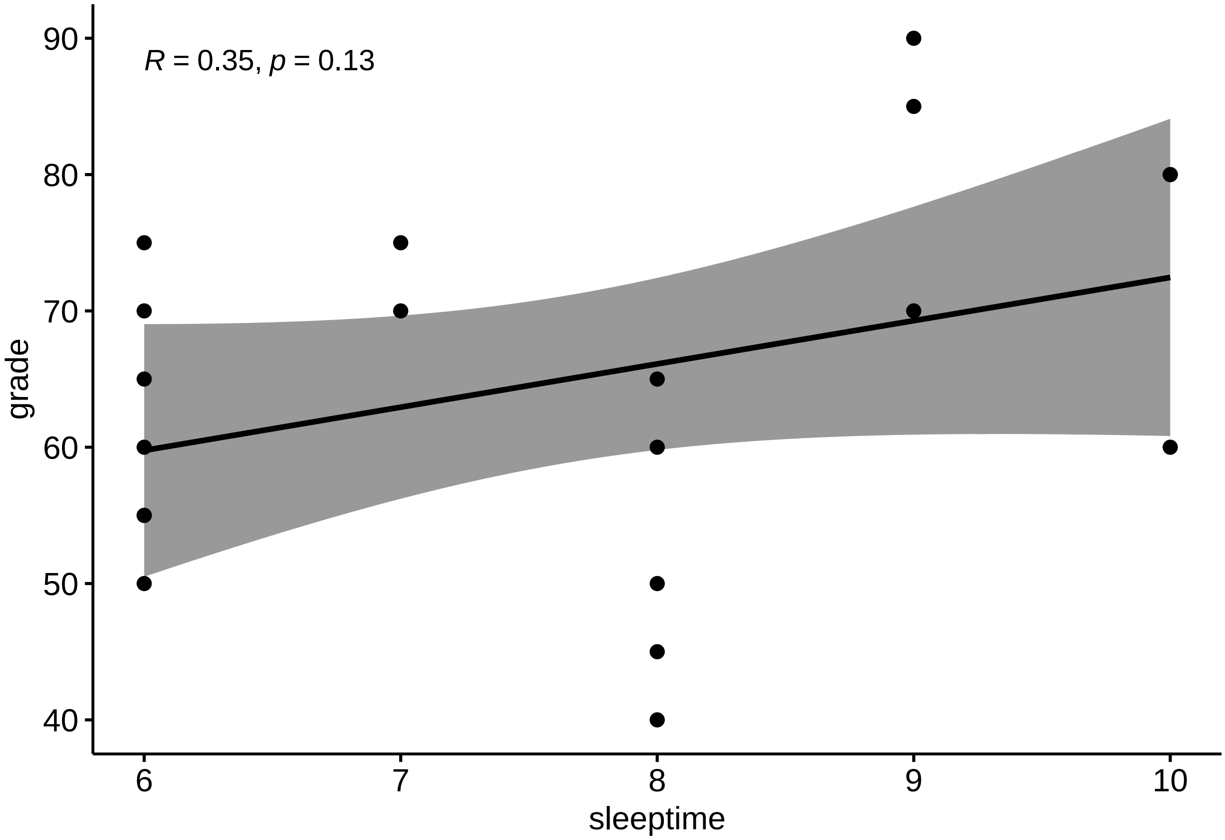
```
ggscatter(newdf,x="studytime",y="grade",add="reg.line",conf.int=TRUE,cor.coef=TRUE)
```

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



```
ggscatter(newdf,x="sleeptime",y="grade",add="reg.line",conf.int=TRUE,cor.coef=TRUE)
```

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



Looking at the summary for modelstudy, we find that our R^2 value is quite high and our p value for our coefficient is quite low. Thus, for modelstudy, we can reject the null hypothesis and conclude that studytime is a good predictor for grade. As for our modelsleep, we find that R^2 is quite low and our p value is quite high. Thus, for modelsleep, we cannnot reject the null hypothesis, thus sleeptime is not a good predictor of grade. We can also see this relation in the ggscatter plots, as the gray area for sleeptime is much more varied than the studytime plot.

c)

```
modelstudy$coefficients
```

```
## (Intercept)   studytime 
##          40           2
```

From the coefficients of our modelstudy, we find that our regression line equation is: $40 + 2(studytime)$. Thus, to find the grade if $studytime = 12$, we simply plug in 12 for studytime: $40 + 2(12) = 64$. Thus, we would expect if $studytime = 12$, the predicted grade would be 64.

d) If we look at the scatter plot of the grade vs studytime data, we see that there is a range of grade values for each studytime value. Since we are only applying a linear model, we cannot account for every single variance in grade for each studytime. Thus, while our model might guess some of the grades perfectly, there will inevitably be some error for each studytime.

2)

a)

```
cvmodel <- lm(grade[-1]~studytime[-1])
cvmodel$coefficients
```

```
##   (Intercept) studytime[-1] 
##     42.325581      1.860465
```

```
studytime[1]
```

```
## [1] 5
```

```
grade[1]
```

```
## [1] 40
```

```
pred <- 42.325581 + 1.860465*(studytime[1])
(grade[1] - pred)^2
```

```
## [1] 135.2082
```

b) Using the coefficients and the studytime[1] $= 5$ value, we can estimate our grade and compare it to the actual value. Thus, using our linear model, our estimate for grade comes out to be $42.325581 + 1.860465(5) = 51.62791$.

c) In order to calculate our MSE, we only need to calculate the RSS, as in our LOOCV, $n = 1$. Thus, by subtracting the actual and the predicted and squaring the value, we get: 135.2082 as our MSE.

3)

a) $\lim_{x\to\infty} \frac{e^{-x}}{1+e^{-x}}$. As this function goes to infinity, e^{-x} will go to zero as $e^{-x} = \frac{1}{e^x}$ and as e^x gets larger, e^{-x} gets smaller. Thus, our numerator will go to 0 as our denominator goes to 1. Thus, the limit of our function as $x \to \infty$ is 0.

b) Here, if we "plug in" our $-\infty$, we will get $\frac{\infty}{\infty}$. Since this approach will not work, we can use L'hopital's rule and take the derivative of the numerator and denominator which will converge to the same value as the original. Thus, taking the derivative of the numerator and denominator, we get $\frac{e^x}{e^x} = 1$. Thus, as $x \to -\infty$, our function will approach 1.

c) Another way to write our function is: $e^{-x}(1 + e^{-x})^{-1}$. By using this function and the product rule: $\frac{e^{-2x}}{(1+e^{-x})^2} - \frac{e^{-x}}{1+e^{-x}} = \frac{e^{-2x}-e^{-x}-e^{-2x}}{(1+e^{-x})^2} = -\frac{e^{-x}}{(1+e^{-x})^2}$. Thus, since our numerator and denominator are both always positive, and a negative times a positive is negative, we conclude our function is always decreasing.

4)

a) This is classification, as we are classifying our data into discrete groups, not predicting continuous values.

b) Since we have a total of 100 data points in total, our test error rate would be $\frac{16+3}{100} = .19$, thus our test error rate is 19%.