STAT X12 HW 3

Due Sept 20 at 11:59 pm

HW 3 Instructions:

This is a self-organized group homework (groups of size 2, 3, or 4). You must find a partner for this assignment for it to be graded. If you are having trouble finding a group, post in the Water Cooler channel in MS Teams and maybe you can find a group to join or someone else who isn't in a group. You will get a bonus (TBD but between 1 and 2%) for combining with someone new for the homework (one new pairing within the group is enough). You will need to note the new combination in the final question to get the bonus.

At the end of the assignment you will document any discussions that have a substantive impact on your results.

Make sure you check the knitted version of your answers for errors. We will only grade what we can see.

Back pain and sitting or sitting and back pain? (continued)

- Gupta, N., Christiansen, C., Hallman, D., Korshoj, M., Carneiro, I., and Holtermann, A. (2015) Is Objectively Measured Sitting Time Associated with Low Back Pain? A Cross-Sectional Investigation in the NOMAD study. *PLoS ONE* 10(3): e0121159. doi:10.1371/journal.pone.0121159

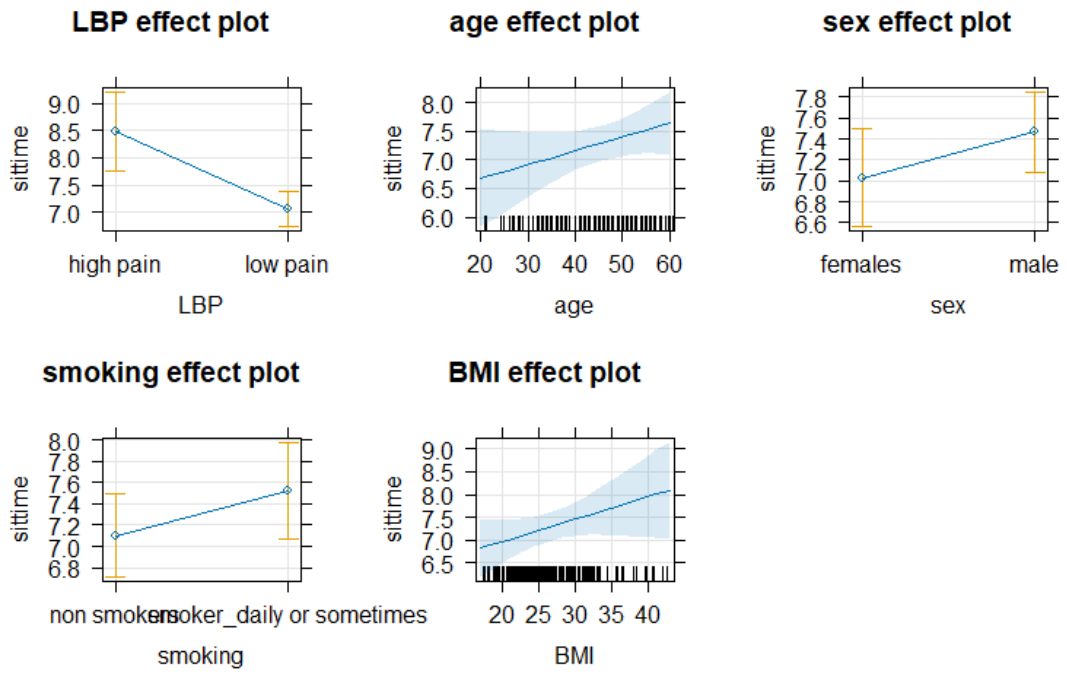
```
gupta_2015 <- read_csv("gupta_2015_sheet1.csv")
g2 <- gupta_2015 %>% dplyr::rename_all(list(~make.names(., unique=T)))
g2 <- g2 %>% dplyr::rename(age = 'age.in.years',
                           seniority = 'job.seniority.in.months',
                           BMI = 'body.mass.index.in.kg.m2',
                           influence = 'Influence.at.work.in.0.100.',
                           LBP = 'low.back.pain.categories',
                           smoking = 'smoking.status.categories',
                           sittime = 'total.sitting.time.in.hours',
                           occlift =
'occupational.lifting.carrying.time.categories',
                          sittimepercent
='total.sitting.time.as.percentage.of.the.mean.measured.time.per.day',
                          sittimepercent_cat =
'categories.of.the.percent.total.sitting.time',
                          sittimehours_cat =
'categories.of.the.total.sitting.time.in.hours') %>%
                    mutate(sex = factor(sex),
                           LBP = factor(LBP),
                           smoking = factor(smoking),
```

```
                           occlift = factor(occlift),
                           sittimepercent_cat = factor(sittimepercent_cat),
                           sittimehours_cat = factor(sittimehours_cat),
                           occlift = fct_recode(occlift,
                                                high = 'high all the time 3/4
time, and \xbd of the time',
                                                moderate = 'moderate 1/4
lift',
                                                rarely = 'rarely, very little
and never lift')
                    )

g2R <- g2 %>% drop_na(sittime, LBP, age, sex, smoking, BMI)

lm1A <- lm(sittime ~ LBP + age + sex + smoking + BMI, data = g2R)
summary(lm1A)

##
## Call:
## lm(formula = sittime ~ LBP + age + sex + smoking + BMI, data = g2R)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3864 -1.4416 -0.1314  1.6021  5.2548
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        5.66056    1.14029   4.964 1.57e-06
## LBPlow pain                       -1.42988    0.40666  -3.516 0.000551
## age                                0.02415    0.01611   1.499 0.135531
## sexmale                            0.43643    0.31464   1.387 0.167091
## smokingsmoker_daily or sometimes   0.42856    0.30635   1.399 0.163510
## BMI                                0.04875    0.03094   1.576 0.116822
##
## Residual standard error: 2.065 on 184 degrees of freedom
## Multiple R-squared:  0.09737,    Adjusted R-squared:  0.07284
## F-statistic:  3.97 on 5 and 184 DF,  p-value: 0.001922

lm1A %>% tbl_regression(intercept = T)
```

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	5.7	3.4, 7.9	<0.001
LBP			
high pain	—	—	
low pain	-1.4	-2.2, -0.63	<0.001
age	0.02	-0.01, 0.06	0.14
sex			
females	—	—	
male	0.44	-0.18, 1.1	0.2
smoking			
non smokers	—	—	
smoker_daily or sometimes	0.43	-0.18, 1.0	0.2
BMI	0.05	-0.01, 0.11	0.12

[1]CI = Confidence Interval

```
plot(allEffects(lm1A), grid = T)
```



1) Suppose that the researchers run the following code. What are the null and alternative hypotheses being assessed in the terms of the slope coefficients from the theoretical model: $\mu\{sittime|LBP, age, sex, smoking, BMI\} = \beta_0 + \beta_1 I_{LBP=Low} + \beta_2 age + \beta_3 I_{sex=male} + \beta_4 I_{smoking=daily} + \beta_5 BMI$?

- $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$
- $H_A: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 \neq 0$

```
lmR <- lm(sittime ~ LBP, data = g2R)
anova(lmR, lm1A)

## Analysis of Variance Table
##
## Model 1: sittime ~ LBP
## Model 2: sittime ~ LBP + age + sex + smoking + BMI
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1    188 817.76
## 2    184 784.49  4    33.273 1.951 0.1038
```

2) Write an evidence sentence for the previous result, making sure you include information on what you would do based on this result - would you go with the full or reduced model based on this result?

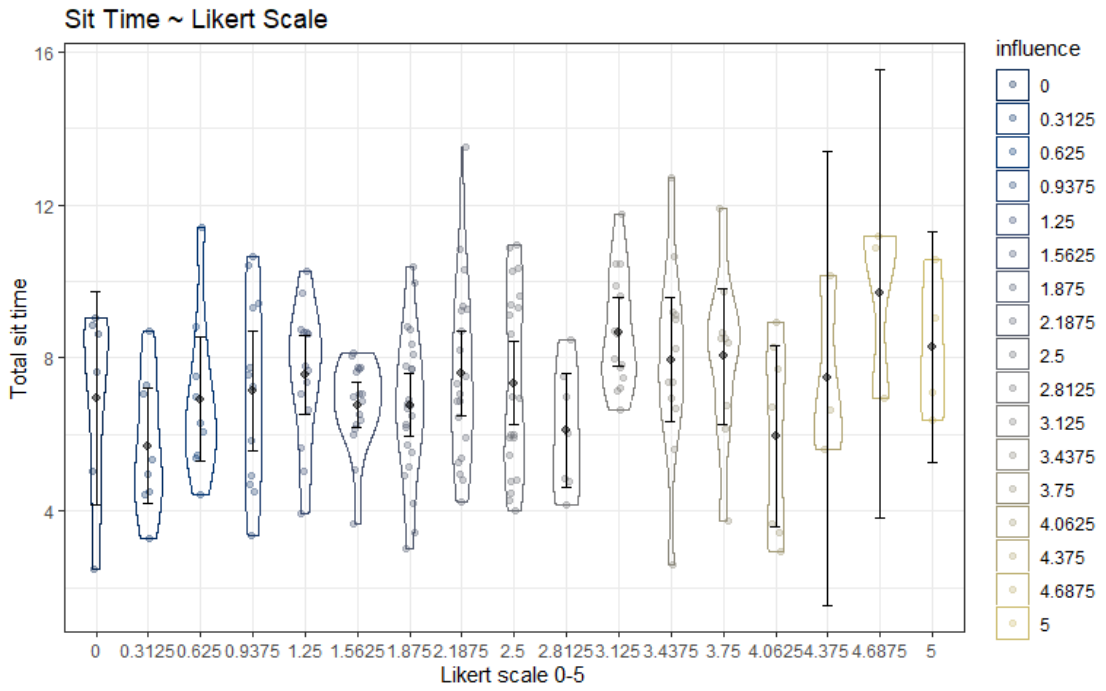
There is no evidence to reject the null hypothesis that we need to include the main effects of age, sex, smoking, and BMI for their impacts on the sittime~LBP model F(4, 184) = 1.951, P-value = 0.104.

Exploring occupational lifting and influence:

3) The `influence` variable was described as "The influence at work (decision authority/latitude) of the workers was determined by the 4-item scale from the Copenhagen Psychosocial Questionnaire with Cronbach's Alpha of 0.77. A sample item is "Do you have a large degree of influence concerning your work?". The responses were scored on a Likert scale with response categories ranging from 0 (never) to 5 (always). A composite scale measuring influence at work was constructed by calculating the mean rating of all four items. For the analysis, this scale was recorded to 0–100 scale, whereby a larger score represented a higher degree of influence at work." It seems like it would be more meaningful to interpret this variable on the original 0 to 5 point scale than as a percentage (so its units would be "points"). Use `mutate` to create a new version of the variable in the data frame that rescales the variable to have a minimum of 0 and maximum of 5 in the `g3` data set created below. Then make a plot of the sitting time (it is still the response) versus the rescaled influence variable you created. Remember to label the x and y-axes in the plot. Discuss the strength and direction of the relationship based on the plot.

```
g3 <- g2R %>% drop_na(occlift, influence) #Use for questions 3 to 8

g3 <- g3 %>%
  mutate(influence = influence/20)

print(g3$likert_scale)

## NULL

?enhanced_stripchart
enhanced_stripchart(g3, sittime~influence)+
                      labs(title = 'Sit Time ~ Likert Scale',
```

```
                    x = 'Likert scale 0-5',
                    y = 'Total sit time')
```



Based on the enhanced strip chart, it appears that the the average total sit time is relatively similar among all likert scale values. There are many more observations between likert scale 1-3 than the <1 and >3. It is apparent that there are fewer observations of around 0 and 4 to 5, which can be seen by the fewer number of dots and the spread of the error. When investigating likert scale values of 4 to 5, many of these higher ranking employees may be considered white-collar, and thus did not report sitting time.

4) The occupational lift variable (`occlift`) is discussed as "The self-reported information on total time spent on occupational carrying and lifting was collected using a question: How much of your working time do you spend carrying or lifting things? with six response categories summarized into three groups; high lifting/carrying time (almost all the time, approximately 3/4 of the time, and 1/2 of the time), moderate lifting/carrying time (1/4 of the time) and low lifting/carrying time (rarely/very little and never)." Plot sitting time versus lifting categories (improve x and y-axis labels in the plot) and generate an F-test for the null hypothesis of no difference in the true mean sitting time across the lift categories. Report the test statistic, distribution under the null hypothesis, and p-value. No "evidence" sentence or discussion needed, just those test details.

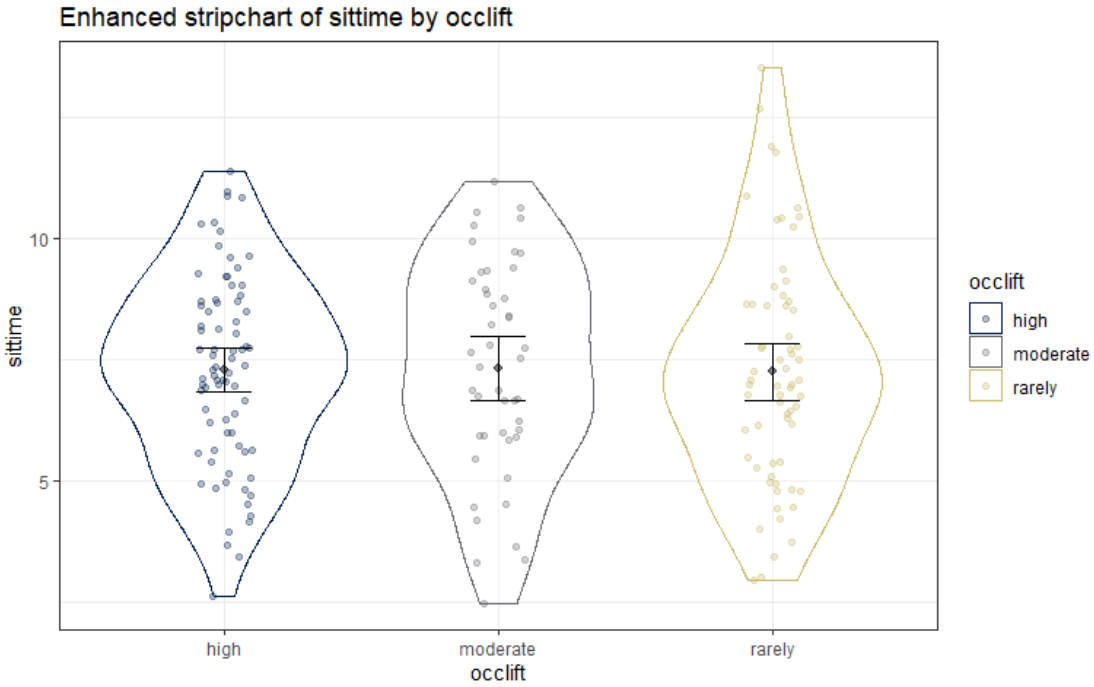
```
is.factor(g3$occlift)
```

```
## [1] TRUE
```

```
unique(g3$occlift)
```

```
## [1] moderate high     rarely
## Levels: high moderate rarely
```

```
enhanced_stripchart(g3, sittime~occlift)
```



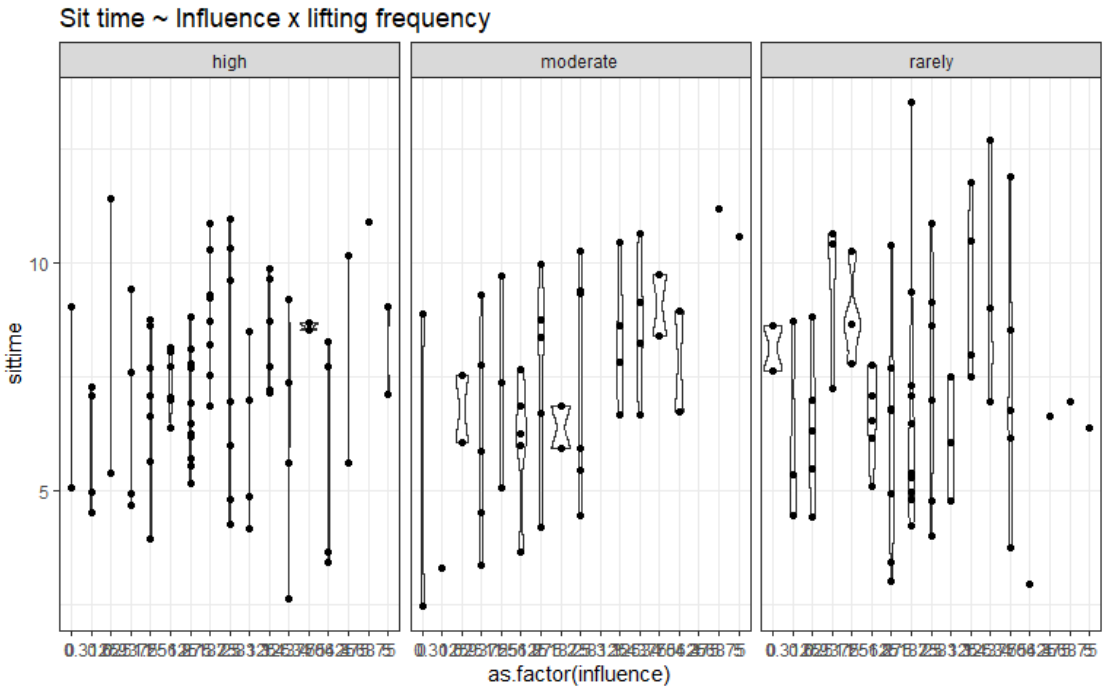

```
lm1 <- lm(sittime ~ occlift, data = g3)
summary(lm1)

##
## Call:
## lm(formula = sittime ~ occlift, data = g3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8435 -1.4635 -0.0735  1.4365  6.2674
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.29346    0.24472  29.803   <2e-16
## occliftmoderate  0.03002    0.40179   0.075    0.941
## occliftrarely   -0.04085    0.36298  -0.113    0.911
##
## Residual standard error: 2.161 on 186 degrees of freedom
## Multiple R-squared:  0.0001626,  Adjusted R-squared:  -0.01059
## F-statistic: 0.01512 on 2 and 186 DF,  p-value: 0.985

Anova(lm1)

## Anova Table (Type II tests)
##
```

```
## Response: sittime
##           Sum Sq  Df F value Pr(>F)
## occlift     0.14   2  0.0151  0.985
## Residuals 868.85 186
```

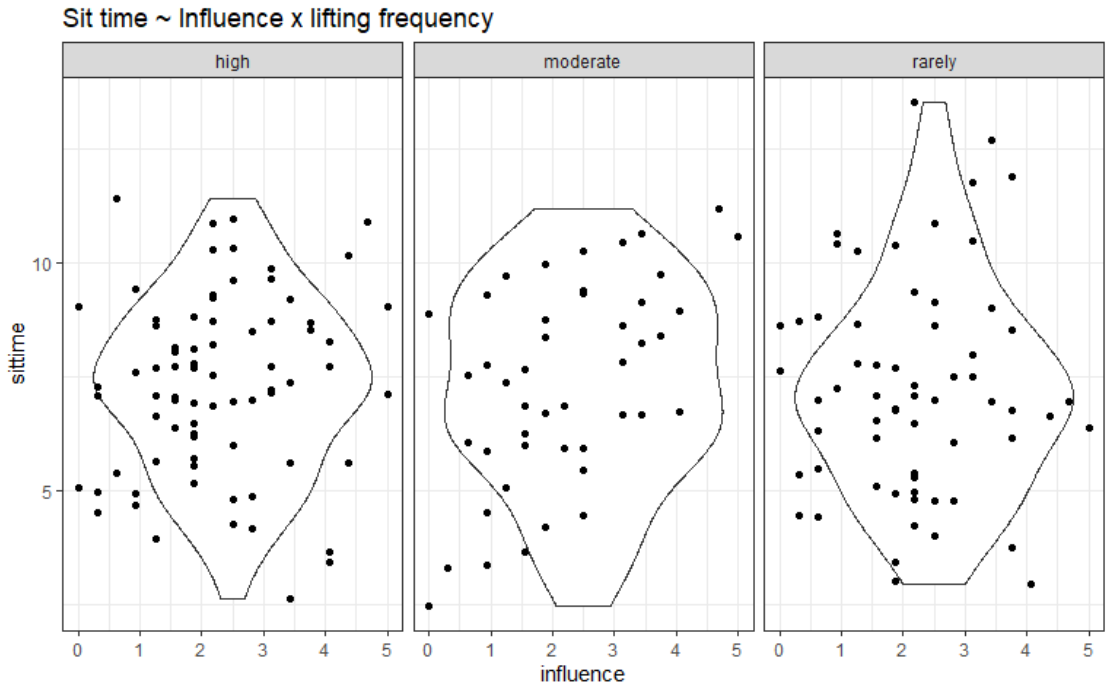
F(2, 186) = 0.015, P-value = 0.985

5) Facet the previous plot of sitting time versus your new version of influence based on the lift categories. Discuss the relationships displayed and the potential for an interaction between influence and lift categories *on the sitting time*. How could you explain the result based on the variables being measured?

```
ggplot(g3, aes(y = sittime, x = as.factor(influence)))+
  geom_violin()+
  geom_point()+
  facet_wrap(~occlift)+
  labs(title = 'Sit time ~ Influence x lifting frequency')
```




```
ggplot(g3, aes(y = sittime, x = influence))+
  geom_violin()+
  geom_point()+
  facet_wrap(~occlift)+
  labs(title = 'Sit time ~ Influence x lifting frequency')
```



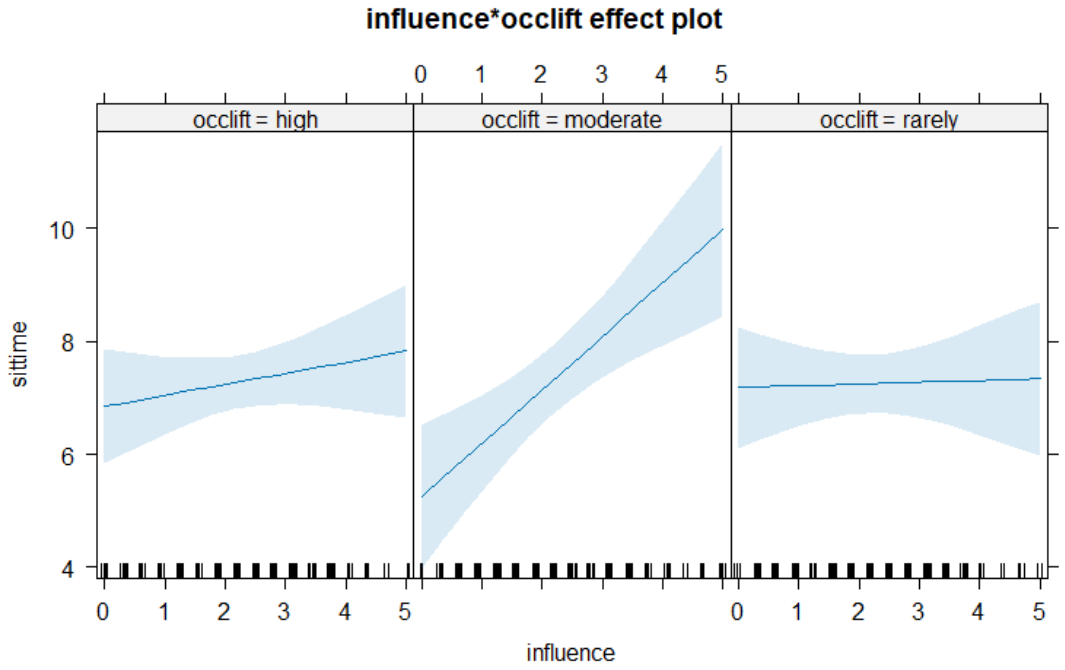
It is hard to tease out a trend in the potential interaction based on this figure. It looks like the spread of observations is greater in the rarely lifting category than the other two. However, for all three lifting categories, it appears that the likert scale-2 value has the most observations while 0 has the fewest. The high group appears to have no trend, the moderate group appears to have a positive relationship between sit time and likert scale, and the rarely group appears to have a negative relationship, but this is hard to confirm by the spread of these data.

6) Fit the interaction model with your new version of the influence variable and occupation lifting, run a model summary, generate two versions of the effects plot (regular version with "faceted" panels and a multiline version that overlays the lines with CI bands and different line types), and write out the estimated model (define any indicators used).

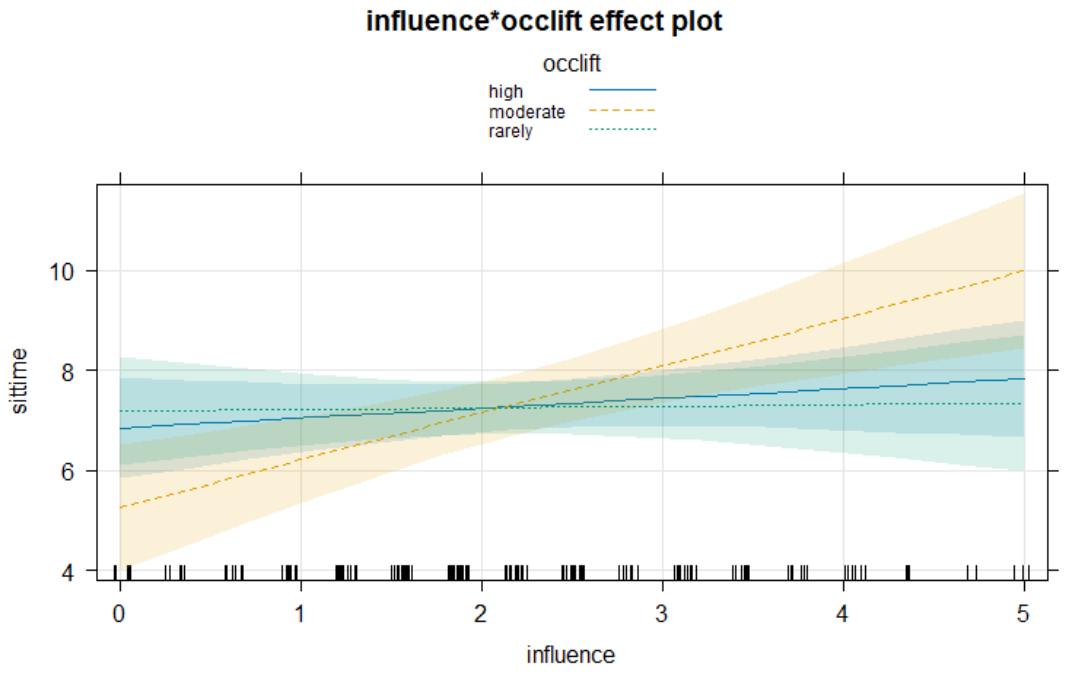
```
lm2 <- lm(sittime ~ influence*occlift, data = g3)
summary(lm2)

##
## Call:
## lm(formula = sittime ~ influence * occlift, data = g3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9026 -1.5817 -0.1537  1.4422  6.2660
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  6.8492     0.5120  13.377   <2e-16
## influence                    0.1959     0.2000   0.979   0.3288
## occliftmoderate             -1.5911     0.8185  -1.944   0.0534
```

```
## occliftrarely                0.3360     0.7479   0.449   0.6538
## influence:occliftmoderate    0.7482     0.3244   2.306   0.0222
## influence:occliftrarely     -0.1645     0.2999  -0.548   0.5841
##
## Residual standard error: 2.097 on 183 degrees of freedom
## Multiple R-squared:  0.07423,    Adjusted R-squared:  0.04894
## F-statistic: 2.935 on 5 and 183 DF,  p-value: 0.01421
```

```
plot(allEffects(lm2))
```




```
plot(allEffects(lm2), multiline = T, grid = T, confint = list(style =
'auto'), lty = 1:3)
```



Predicated

- $\hat{\mu}\{sittime|influence * occlift\} = \beta_0 + \beta_1(influence) + \beta_2 I_{Occlift=moderate} + \beta_3 I_{Occlift=rarely} + \beta_4(influence \cdot I_{Occlift=moderate} + \beta_5(influence \cdot I_{Occlift=rarely})$

Estimated

- $\hat{\mu}\{sittime|influence * occlift\} = 6.8492 + 0.1959(influence) - 1.5911 \cdot I_{Occlift=moderate} + 0.336 \cdot I_{Occlift=rarely} + 0.7482(influence) \cdot I_{Occlift=moderate} - 0.1645(influence) \cdot I_{Occlift=rarely})$
- Where $I_{Occlift=moderate}$ is 1 for moderate lift frequency, and 0 otherwise
- Where $I_{Occlift=rarely}$ is 1 for rarely lift frequency, and 0 otherwise

7) Report simplified versions of the model for each of the three levels of occupational lift, showing your work.

For high:

$$\hat{\mu}\{sittime|influence * (occlift = high)\} = 6.8492 + 0.1959(influence) - 1.5911 \cdot 0 + 0.336 \cdot 0 + 0.7482(influence) \cdot 0 - 0.1645(influence) \cdot 0)$$

$$\hat{\mu}\{sittime|influence * (occlift = high)\} = 6.8492 + 0.1959(influence)$$

For moderate:

$$\hat{\mu}\{sittime|influence * (occlift = moderate)\}$$
$$= 6.8492 + 0.1959(influence) - 1.5911 \cdot 1 + 0.336 \cdot 0 + 0.7482(influence) \cdot 1 - 0.1645(influence) \cdot 0)$$

$$\hat{\mu}\{sittime|influence * (occlift = moderate)\} = 5.2581 + 0.8868(influence)$$

For rarely:

$$\hat{\mu}\{sittime|influence * (occlift = rarely)\}$$
$$= 6.8492 + 0.1959(influence) - 1.5911 \cdot 0 + 0.5800 \cdot 1 + 0.7482(influence) \cdot 0 - 0.1645(influence) \cdot 1)$$

$$\hat{\mu}\{sittime|influence * (occlift = rarely)\} = 7.2374 - 0.0314(influence)$$

8) Then report size (slope) interpretation**s** for `influence` for the `high` and `moderate` groups. You won't be able to add a CI to the interpretation for the `moderate` group (yet), but you can get one from the model for the `high` level.

```
# lm(sittime ~ influence*occlift)
summary(lm2)

##
## Call:
## lm(formula = sittime ~ influence * occlift, data = g3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9026 -1.5817 -0.1537  1.4422  6.2660
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  6.8492     0.5120  13.377   <2e-16
## influence                    0.1959     0.2000   0.979   0.3288
## occliftmoderate             -1.5911     0.8185  -1.944   0.0534
## occliftrarely                0.3360     0.7479   0.449   0.6538
## influence:occliftmoderate    0.7482     0.3244   2.306   0.0222
## influence:occliftrarely     -0.1645     0.2999  -0.548   0.5841
##
## Residual standard error: 2.097 on 183 degrees of freedom
## Multiple R-squared:  0.07423,    Adjusted R-squared:  0.04894
## F-statistic: 2.935 on 5 and 183 DF,  p-value: 0.01421

confint(lm2)

##                                2.5 %     97.5 %
## (Intercept)                5.8390715 7.85942678
## influence                 -0.1988016 0.59058754
## occliftmoderate           -3.2059278 0.02377735
## occliftrarely             -1.1395372 1.81153465
## influence:occliftmoderate  0.1081288 1.38836352
## influence:occliftrarely   -0.7561781 0.42722828
```

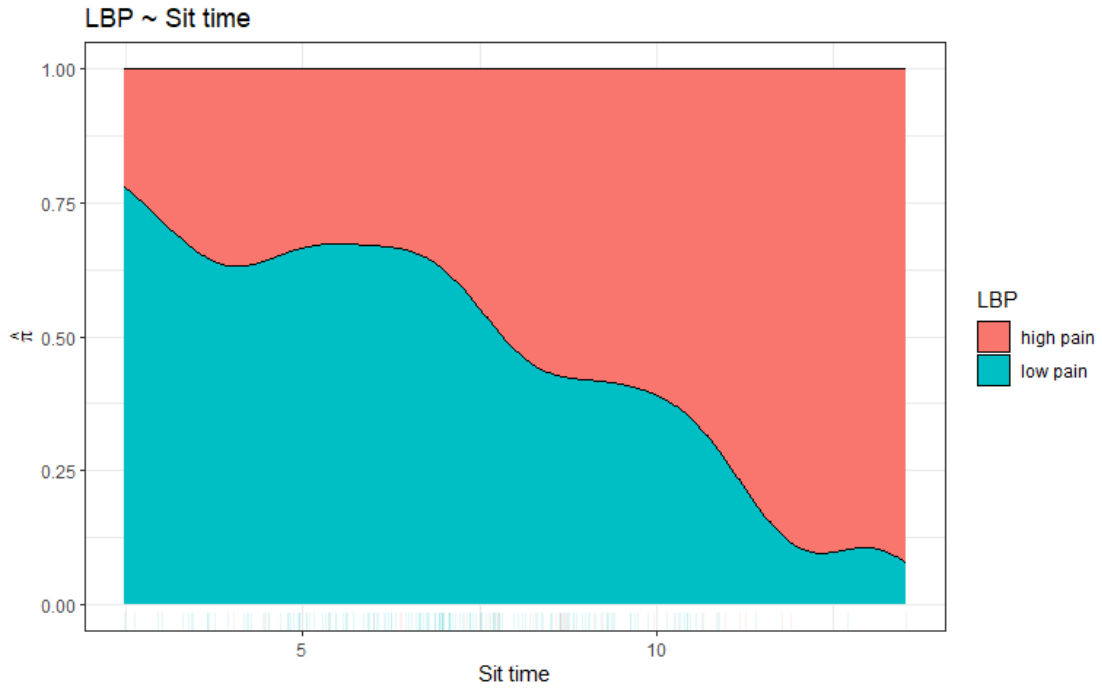
For the high frequency lifting group the slope for influence is 0.1959. This tells us that for every 1-unit increase of the influence variable, the sit time will increase by 0.1959. hours. Thus showing there is a slight positive relationship between the work influence and amount of time spent sitting for the individuals that do high levels of lifting (95% CI: -0.1988016 to 0.59058754).

For the moderate frequency lifting group the slope for influence is 0.8868. This tells us that every 1-unit increase of the influence variable, the sit time will decrease by 0.8868 hours. Thus showing there is a positive relationship between the work influence and amount of time spent sitting for the individuals that do moderate levels of lifting.

Switching gears… LBP as a response instead of explanatory

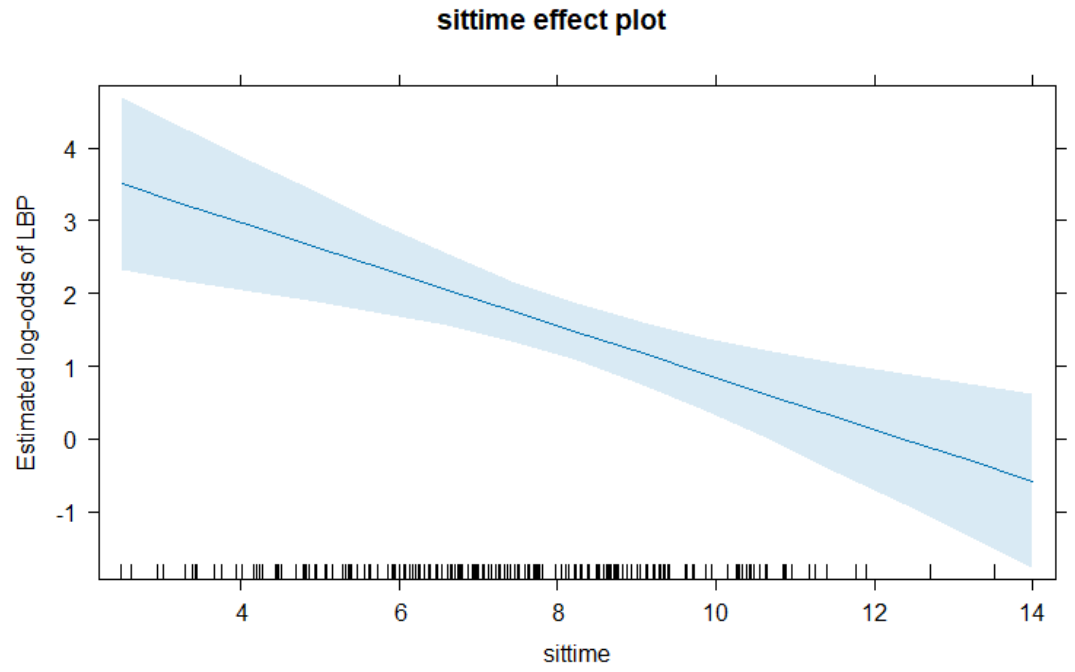
9) Make a plot of low back pain (LBP) as the response variable versus sitting time. Discuss the pattern in the relationship. You can use the entire data set (g2) of 201 observations for the data for this plot and related work that follows (this will match the "Crude" row in their Table 3 for the total sitting time). Also use the default choice for the bw value (remove the specification to get this to happen). Remember to improve the x and y axes in the plot.

```
g2 %>%
  ggplot(aes(x = sittime, fill = LBP))+
  geom_density(position = 'fill')+
  geom_rug(aes(col = LBP), alpha = 0.1)+
  labs(title = 'LBP ~ Sit time',
       x = 'Sit time',
       y = expression(hat(pi)))
```

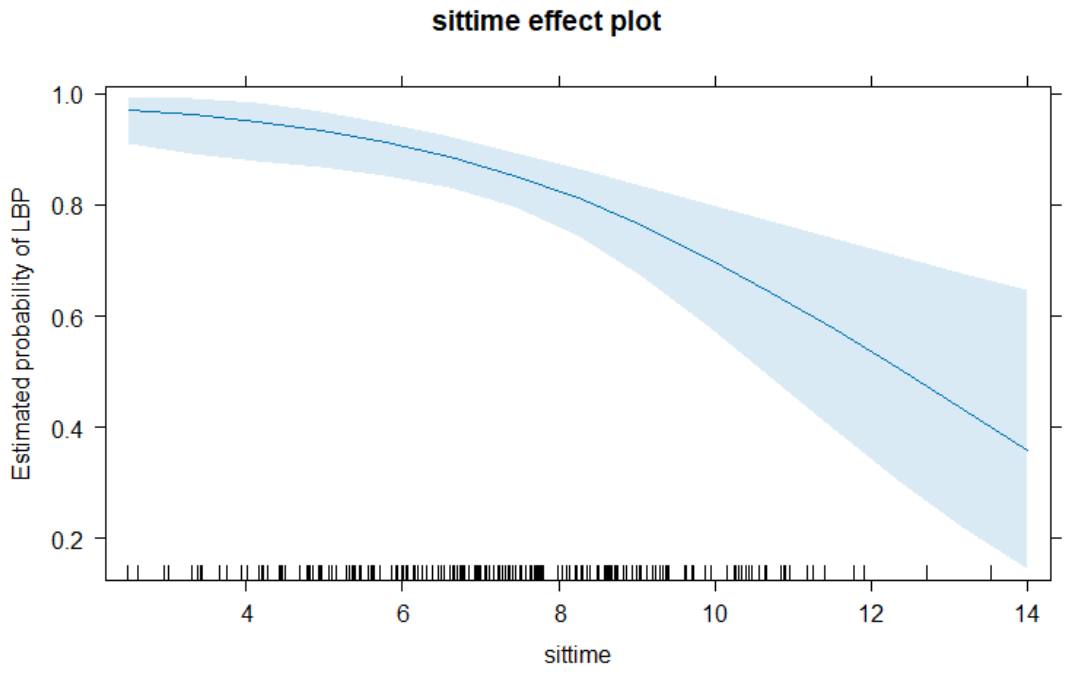


10) Fit an appropriate model with LBP as the response and total sitting time as the predictor. Make effects plots on the link and response scales. Compare the previous plot to the effects plot on the response scale to determine which of the two levels of LBP the function is treating as a "success".

```
glm1 <- glm(LBP ~ sittime, family = binomial, data = g2)
summary(glm1)

##
## Call:
## glm(formula = LBP ~ sittime, family = binomial, data = g2)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.39402    0.84191   5.219  1.8e-07
## sittime     -0.35509    0.09932  -3.575  0.00035
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 179.51  on 200  degrees of freedom
## Residual deviance: 165.17  on 199  degrees of freedom
## AIC: 169.17
##
## Number of Fisher Scoring iterations: 5

plot(allEffects(glm1), type = 'link', ylab = 'Estimated log-odds of LBP')
```




```
plot(allEffects(glm1), type = 'response', ylab = 'Estimated probability of
LBP')
```



It appears that the success group is *high pain*.

11) Sometimes we would like to have the other level as the success, so we can use `relevel` to switch the baseline level. The following code switches that to create an alternate version of `LBP` called `LBP2`. Refit the previous model with this new response and remake the effects plot on the response scale. No discussion.

```
g2 <- g2 %>% mutate(LBP2 = relevel(LBP, "low pain"))
glm2 <- glm(LBP2 ~ sittime, family = binomial, data = g2)
summary(glm2)

##
## Call:
## glm(formula = LBP2 ~ sittime, family = binomial, data = g2)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.39402    0.84191  -5.219  1.8e-07
## sittime      0.35509    0.09932   3.575  0.00035
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 179.51  on 200  degrees of freedom
## Residual deviance: 165.17  on 199  degrees of freedom
## AIC: 169.17
##
## Number of Fisher Scoring iterations: 5
```

12) For both of your models, generate nice looking model summaries using `tbl_regression(exponentiate = T)`. This table provides $exp(\hat{\beta}_1)$ and its

associated 95% CI. One of these results will closely match the result in Table 3 for the "Crude" result. What does that tell you about the "success" used in their model?

```
glm1 %>% tbl_regression(exponentiate = T)
```

Characteristic	OR[1]	95% CI[1]	p-value
sittime	0.70	0.57, 0.85	<0.001

[1]OR = Odds Ratio, CI = Confidence Interval

```
glm2 %>% tbl_regression(exponentiate = T)
```

Characteristic	OR[1]	95% CI[1]	p-value
sittime	1.43	1.18, 1.75	<0.001

[1]OR = Odds Ratio, CI = Confidence Interval

The success in their model is low pain. Labeled *glm2* in my code, this exponentiated table matches that of the crude model line in table 3.

13) Document any resources you used outside of those provided in this class. This includes, but is not limited to, students in other groups and generative AI. If the resource is not static, discuss how you used it and which questions you used it for. Report "NONE" if you did not use any.

Chat GPT for some stuff

14) Document if there is a new partnering in this homework submission for a bonus. If none is noted, we will assume this is a pre-existing group.

This is a new group. All three members are new