BIOS 663 Homework 4

Marissa Ashner

4/8/2019

Problem 1

The following questions are on the data and model described in Q3 of HW3.

part i:

Examine the tolerances and variance inflation factors from this model. Do you think any collinearity is present based on tolerance and VIF? Why or why not?

Define tolerance as follows:

$$T_j = 1 - R_j^2$$

where $R_j^2 = R^2(X_j, \{X_1, \ldots X_{j-1}, X_{j+1}, \ldots X_{p-1}\})$ is the squared multiple correlation. Tolerances close to 1 are good, where tolerances close to 0 show worse collinearity.

Define the variance inflation factor as follows:

$$VIF_j = \frac{1}{1 - R_j^2} = \frac{1}{T_j}$$

A VIF close to 1 is good, where a VIF implies worse collinearity as it approaches infinity. The R-Code below calculates the Tolerance and VIF values for the model.

```
fit = lm(AVFVC ~ HEIGHT + WEIGHT + BMI + AREA + AGE + AVTREL + AVTRSP + AVTREL*AVTRSP + TEMP + BAROM + HUMID, data = dat2)
VIF = vif(fit)
Tol = 1/VIF
df = (rbind(VIF, Tol))
df %>% knitr::kable(align = c("c", "c"))
```

	HEIGHT	WEIGHT	BMI	AREA	AGE	AVTREL	AVTRSP	TEMP	BAROM	HUMID	AVTREL:AVTRSP
VIF	458.0476405	703.4462861	177.4503720	1364.8975242	1.0833448	580.389689	78.3930237	29.0137857	1.0590953	29.1669187	795.4489506
Tol	0.0021832	0.0014216	0.0056354	0.0007327	0.9230672	0.001723	0.0127562	0.0344664	0.9442021	0.0342854	0.0012572

Based on these values, it appears that there is a lot of collinearity present. This is because very few VIF values are close to 1, and many of the tolerance values are close to 0.

part ii:

Conduct an eigenanalysis of the scaled SSCP and correlation matrices, presenting a table formatted like Table 8.6.2 in Muller and Fetterman.

The Scaled SSCP Matrix can be defined as follows:

$$SSCP_s = D_s^{-0.5}(X'X)D_s^{-0.5}$$

where X is the design matrix includign the intercept, and D_s is a diagonal matrix with elements extracted from the diagonal of $X'X$.

and the correlation matrix as follows:

$$R = D_c^{-0.5} C D_c^{-0.5}$$

where C is the covariance matrix of the centered design matrix excluding the intercept, and D_c is a diagonal matrix of the extracted diagonal values of C.

The following R code calculates these two matrices for the model above, and performs an eigenanalysis on them both.

```
cov_int <- dat2 %>% mutate(INT = 1, AVTRELTRSP = AVTREL*AVTRSP) %>% select(INT, HEIGHT, WEIGHT, BMI, AREA, AGE, AVTREL, AVTRSP, AVTRELTRSP, T
EMP, BAROM, HUMID) %>% as.matrix()
xtx = t(cov_int)%*%cov_int
Ds_half <- diag(diag(xtx)^-0.5)
sscp <- Ds_half %*% xtx %*% Ds_half
eig_sscp <- eigen(sscp)$values
PCs_sscp <- prcomp(sscp)[2]
CI_sscp <- sqrt(eig_sscp[1]/eig_sscp)

covariates <- dat2 %>% mutate(AVTRELTRSP = AVTREL*AVTRSP) %>% select(HEIGHT, WEIGHT, BMI, AREA, AGE, AVTREL, AVTRSP, AVTRELTRSP, TEMP, BAROM,
HUMID)
cov_center <- apply(covariates, 2, function(y) y - mean(y))
C <- (t(cov_center)%*%cov_center)/dim(cov_center)[1]
Dc_half <- diag(diag(C)^-0.5)
R <- Dc_half %*% C %*% Dc_half
eig_corr <- eigen(R)$values
CI_corr <- sqrt(eig_corr[1]/eig_corr)
PCs_corr <- prcomp(R)[2]

df <- data.frame("Eigenvalue" = c("Correlation Matrix", eig_corr), "Condition Index" = c("Correlation Matrix", CI_corr))
df2 <- data.frame("Eigenvalue" = c("Scaled SSCP", eig_sscp), "Condition Index" = c("Scaled SSCP", CI_sscp))

df %>% knitr::kable(align = c("c", "c"))
```

Eigenvalue	Condition.Index
Correlation Matrix	Correlation Matrix
3.00984215183995	1
2.44782688677429	1.10887223583127
2.02476480717731	1.21922699035498
1.11320126901436	1.64431498130338
1.01325012874562	1.72350888345776
0.809279431894358	1.92851316812643
0.561095110548582	2.31608032840747
0.0177075849003461	13.0374361244415
0.00187373597726271	40.0790724581539
0.000705172710081123	65.3317229298841
0.000453720417828643	81.4474887486485

```
df2 %>% knitr::kable(align = c("c", "c"))
```

Eigenvalue	Condition.Index
Scaled SSCP	Scaled SSCP
11.9049479582918	1
0.0360382910658672	18.1753028593325
0.0293708868488598	20.132848271369
0.0161788245373072	27.1262817394066
0.00669911241875652	42.1555847347362
0.0049048528907778	49.2663919299516
0.0015549706468733	87.4989120284385
0.000251548277639727	217.546988935216
3.7467685203547e-05	563.683495875739
9.67075357565232e-06	1109.51605795048
4.86755151319751e-06	1563.89817359537

1.54903175502382e-06 2772.25764760603

(a):

Does there appear to be any collinearity between the intercept and the covariates? Why or why not? If so, list the variables.

Since the eigenvalues from the Scaled SSCP (which includes the intercept) show several eigenvalues near 0 and condition indices above 30 (namely the last 8), we know that there does appear to be collinearity issues. To identify which covariates this collinearity is between, we take a look at the last few PCs below.

```
PCs_sscp$rotation[,9:12]
```

```
##                PC9          PC10          PC11          PC12
##  [1,] -0.362557672  0.2106087110  0.6871965147  0.1473741435
##  [2,] -0.302287651 -0.3282233486 -0.4589345265  0.4834046812
##  [3,]  0.228929531  0.4172069832  0.1173983196  0.3239142829
##  [4,] -0.186467777 -0.2773877633 -0.1708233087 -0.0167347071
##  [5,] -0.097183258 -0.3259677478  0.1402711189 -0.7151909522
##  [6,]  0.008240427  0.0007674032  0.0006411418 -0.0004944857
##  [7,] -0.093561530  0.3902109785 -0.2888650047 -0.2038318770
##  [8,] -0.094260468  0.4147408015 -0.2908961834 -0.2059574809
##  [9,]  0.100295075 -0.3935164095  0.2925559731  0.2064610370
## [10,] -0.025805362  0.0012700485 -0.0127654815 -0.0270116541
## [11,]  0.806990526 -0.1072759346 -0.0237080809 -0.0096174923
## [12,]  0.029751201 -0.0035722176  0.0110939585  0.0220843112
```

From the PCA analysis, we can see that the covariates with the largest departures from 0 in the last four PCs are covariates 1 (i.e. the intercept), 2 (height), 3 (weight), 5 (area), 7 (avtrel), 8 (avtrsp), and 9 (avtrel*avtrsp).

(b):

Does there appear to be any collinearity among the covariates? Why or why not? If so, list the variables.

Since the eigenvalues from the Correlation Matrix (which does NOT includes the intercept) show several eigenvalues near 0 and condition indices above 30 (namely the last three), we know that there does appear to be collinearity issues. To identify which covariates this collinearity is between, we take a look at the last few PCs below.

```
PCs_corr$rotation[,10:11]
```

```
##               PC10         PC11
##  [1,]  0.108339574 -0.507595144
##  [2,]  0.544305474 -0.014314017
##  [3,] -0.130059319 -0.310181321
##  [4,] -0.544577040  0.582884584
##  [5,] -0.004893314 -0.001127961
##  [6,] -0.390534996 -0.350817108
##  [7,] -0.138224165 -0.124569696
##  [8,]  0.454476015  0.409062361
##  [9,] -0.012113720  0.009805947
## [10,] -0.002270905 -0.002276515
## [11,]  0.012911758 -0.012268259
```

From the PCA analysis, we can see that the covariates with the largest departures from 0 in the last four PCs are covariates 1 (height), 3 (BMI), 4 (area), 6 (avtrel), and 8 (avtrel*avtrsp).

Problem 2

Find the Box-Cox Transformation of the simulated data (BoxCox.dat) and compare the residual plots of the raw and transformed data.

The Box-Cox Transformations are a family of transformations of the response varaibles defined as:

$$Y_i(\pi) = \{ \frac{Y_i^{\pi} - 1}{\pi Y^{*(\pi-1)}} \quad \pi \neq 0 \quad \& \quad Y^{*} ln(Y_i) \quad \pi = 0 \}$$

where $Y^* = (\Pi_{i=1}^N Y_i)^{1/N}$. This corresponds to a transformation that is y^{π} for $\pi \neq 0$ and $log(y)$ otherwise. The transformation above puts the SSE of these on the same scale for the purpose of comparison and choosing the best π. We try the values of π between -1 and 1 incremented every 0.25 to compare the likelihoods and find the best transformation.

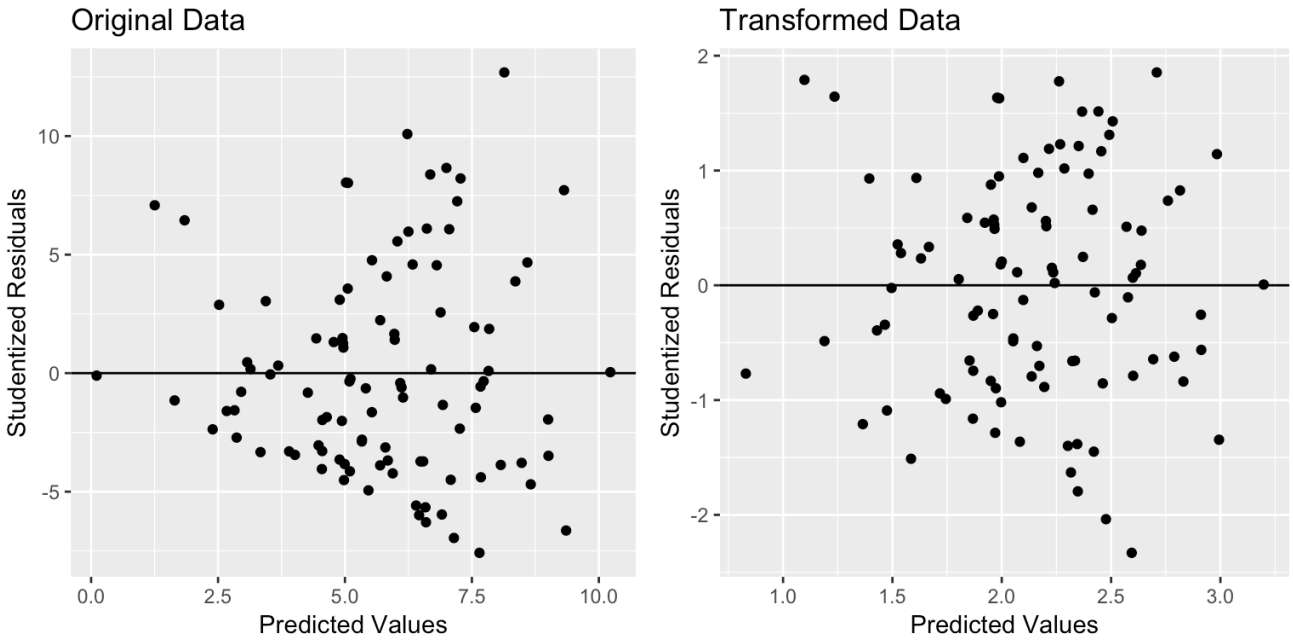
```
fit_bc <- lm(bc$V2 ~ bc$V1)
# Transformation Analysis
cols <- MASS::boxcox(fit_bc, seq(-1,1,1/4), plotit = FALSE)$x
like <- MASS::boxcox(fit_bc, seq(-1,1,1/4), plotit = FALSE)$y %>% as.matrix() %>% t()
knitr::kable(like, col.names = cols)
```

-1	-0.75	-0.5	-0.25	0	0.25	0.5	0.75	1
-685.0735	-549.0958	-425.2876	-325.0825	-264.4143	-241.4822	-239.6977	-248.8666	-264.7483

From the values above, we can see that the choice of π with the smallest likelihood is $\pi = 0.5$. Below, we transform the data using this result, and compare the residual plots of the two datasets.

```
lambda = 0.5
bc$V3 <- (bc$V2^lambda)
fit_bc_2 <- lm(bc$V3 ~ bc$V1)

plot1 <- ggplot() + geom_point(aes(fit_bc$fitted.values, fit_bc$residuals)) + geom_hline(aes(yintercept = 0)) + labs(x = "Predicted Values",
y = "Studentized Residuals", title = "Original Data")
plot2 <- ggplot() + geom_point(aes(fit_bc_2$fitted.values, fit_bc_2$residuals)) + geom_hline(aes(yintercept = 0)) + labs(x = "Predicted Value
s", y = "Studentized Residuals", title = "Transformed Data")
cowplot::plot_grid(plot1, plot2, nrow = 1)
```



From the plots above, it is clear that the transformation of the data yields better assumption validations than the original data. In particular, the graph on the left of the original data seems to fan out (i.e. more extreme residuals) as the predicted values are increased. The residuals are more randomly distributed around the x-axis in the transformed data.

Problem 3

part i: One-Way ANOVA:

For these questions, use the log of salivary cotinine as the response and task as the only predictor.

```
tobacco1 <- tobacco %>% mutate(LOGCOT = log(COTININE),
                               TASK1 = case_when(TASK == 1 ~ 1, TASK != 1 ~ 0),
                               TASK2 = case_when(TASK == 2 ~ 1, TASK != 2 ~ 0),
                               TASK3 = case_when(TASK == 3 ~ 1, TASK != 3 ~ 0),
                               TASK4 = case_when(TASK == 4 ~ 1, TASK != 4 ~ 0))
```

(a):

Report a test of whether all cell means are equal.

Consider the following model using the cell mean coding scheme:

$$y = \beta_1 I_{T1} + \beta_2 I_{T2} + \beta_3 I_{T3} + \beta_4 I_{T4}$$

where y is the log cotinine, and I_{Ti} is the indicator function associated with the ith task. In order to test whether all cell means are equal, we want to test the following set of hypotheses:

$$H_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4$$

which is equivalent to:

$$H_0 = \beta_1 - \beta_2 = 0, \beta_1 - \beta_3 = 0, \beta_1 - \beta_4 = 0$$

In order to test these hypotheses, we can use the overall F test, where:

$$F = (SSH/a)/\hat{\sigma}^2 \sim F_{G-1,n-G}$$

where $SSH = (\hat{\theta} - \theta_0)' M^{-1} (\hat{\theta} - \theta_0)$, $G = 4, n = 694$, and $\hat{\sigma}^2$ = $. It should also be noted that $M = C(X'X)^{-1}C'$.

```
X = tobacco1 %>% select(TASK1, TASK2, TASK3, TASK4) %>% as.matrix()
fit = lm(LOGCOT ~ -1 + TASK1 + TASK2 + TASK3 + TASK4, data = tobacco1)
thetahat = c((fit %>% summary)$coefficients[1,1] - (fit %>% summary)$coefficients[2,1],
             (fit %>% summary)$coefficients[1,1] - (fit %>% summary)$coefficients[3,1],
             (fit %>% summary)$coefficients[1,1] - (fit %>% summary)$coefficients[4,1])
mse = sum(fit$residuals^2)/(694-4)
C = matrix(c(1, 1, 1, -1, 0, 0, 0, -1, 0, 0, 0, -1), nrow = 3)
M = C %*% solve(t(X) %*% X) %*% t(C)
ssh = t(thetahat) %*% solve(M) %*% thetahat
f_obs = (ssh/3)/mse
p = 1-pf(f_obs, 4-1, 694-4)
## CAN ALSO USE linearHypothesis(fit, C)
```

From the code above, the test statistic $F = 116.2032527$ and the p-value is approximately 0. This means that we can reject the null hypothesis that all four cell means are equal. In other words, there is evidence to reject the fact that the four types of tobacco work have the same mean log cotinine level.

(b):

If your overall test of the task effect was significant, examine all pairwise comparisons using the Scheffe correction. Summarize your findings in a table including columns for the estimated mean difference, degrees of freedom, F statistic, p-value, and a Scheffe confidence interval for the mean difference. Explain your findings in language the investigator can understand.

Since the overall test of the task effect in (a) was significant, we will go forward with the pairwise comparisons. Since TASK has four levels, there will be $4 * (4 - 1)/2 = 6$ pairwise comparisons. Scheffe's correction provides a general technique for account for the fact that we are performing 6 tests.

To find the F statistic, we can take the square of the t statistic as follows:

$$F = t^2 = (\frac{\hat{\beta_i} - \hat{\beta_j}}{\sqrt{MSE(1/n_i + 1/n_j)}})^2 \sim F_{G-1,n-G}$$

where β_i and n_i are the mean log cotinine level and sample size for the ith task level. MSE is the mean squared error as calculated in the previous test. The critical region for this F test can be calculate by multiplying the F statistic by $G - 1 = 4 - 1 = 3$ to account for multiplicity in testing.

```
scheffe <- ScheffeTest(aov(LOGCOT ~ factor(TASK), data = tobacco1))
f1 <- ((summary(fit)$coefficients[1,1] - summary(fit)$coefficients[2,1]) /sqrt(mse*(1/sum(tobacco1$TASK1) + 1/sum(tobacco1$TASK2))))^2
f2 <- ((summary(fit)$coefficients[1,1] - summary(fit)$coefficients[3,1]) /sqrt(mse*(1/sum(tobacco1$TASK1) + 1/sum(tobacco1$TASK3))))^2
f3 <- ((summary(fit)$coefficients[1,1] - summary(fit)$coefficients[4,1]) /sqrt(mse*(1/sum(tobacco1$TASK1) + 1/sum(tobacco1$TASK4))))^2
f4 <- ((summary(fit)$coefficients[2,1] - summary(fit)$coefficients[3,1]) /sqrt(mse*(1/sum(tobacco1$TASK2) + 1/sum(tobacco1$TASK3))))^2
f5 <- ((summary(fit)$coefficients[2,1] - summary(fit)$coefficients[4,1]) /sqrt(mse*(1/sum(tobacco1$TASK2) + 1/sum(tobacco1$TASK4))))^2
f6 <- ((summary(fit)$coefficients[3,1] - summary(fit)$coefficients[4,1]) /sqrt(mse*(1/sum(tobacco1$TASK3) + 1/sum(tobacco1$TASK4))))^2
df <- data.frame(Diff = scheffe$`factor(TASK)`[,1], DF = "(3,690)", f = c(f1, f2, f3, f4, f5, f6), pval = scheffe$`factor(TASK)`[,4], CI_L =
scheffe$`factor(TASK)`[,2], CI_U = scheffe$`factor(TASK)`[,3])
df %>% knitr::kable(align = c("c", "c"))
```

	Diff	DF	f	pval	CI_L	CI_U
2-1	-0.9207815	(3,690)	19.57598	0.0002350	-1.503991	-0.3375720
3-1	-1.6738481	(3,690)	131.87148	0.0000000	-2.082328	-1.2653684
4-1	-2.6992523	(3,690)	332.68364	0.0000000	-3.113975	-2.2845298
3-2	-0.7530666	(3,690)	12.98968	0.0049075	-1.338617	-0.1675167
4-2	-1.7784708	(3,690)	71.37797	0.0000000	-2.368393	-1.1885490
4-3	-1.0254042	(3,690)	47.25875	0.0000000	-1.443412	-0.6073970

From the table above, it appears that all pairwise null hypotheses can be rejected. This means there is evidence to suggest that every mean log cotinine level for a certain task level is different than the mean log cotinine level for any other task level.

(c):

Provide a table of parameter estimates and standard errors using (a) cell mean coding and (b) reference cell coding, and give the interpretations of pa- rameters in both coding schemes. In addition, provide the C and θ_0 matrices used to test the hypothesis that average cotinine levels for workers involved in priming are greater than the average continine levels for all other workers.

For the cell mean coding, we use the model proposed in (a).

```
summary(fit)$coefficients
```

```
##       Estimate Std. Error  t value      Pr(>|t|)
## TASK1 4.508557  0.1022201 44.10636 5.762099e-203
## TASK2 3.587775  0.1812765 19.79172  2.168761e-69
## TASK3 2.834709  0.1039098 27.28047 9.869409e-112
## TASK4 1.809304  0.1070123 16.90745  6.478023e-54
```

The parameter estimates and standard errors are given in the code summary above. The interpretations are as follows: $beta_1$ is the mean log cotinine level for priming, $beta_2$ is the mean log cotinine level for barning, $beta_3$ is the mean log cotinine level for topping, and $beta_4$ is the mean log cotinine level for work not involving tobacco contact.

For the reference cell coding, we consider the following model:

$$y = \beta_1 + \beta_2 I_{T2} + \beta_3 I_{T3} + \beta_4 I_{T4}$$

where y is the log cotinine, and I_{Ti} is the indicator function associated with the ith task. It should be noted that TASK1 is the reference.

```
fit_ref = lm(LOGCOT ~ TASK2 + TASK3 + TASK4, data = tobacco1)
summary(fit_ref)$coefficients
```

```
##               Estimate Std. Error    t value      Pr(>|t|)
## (Intercept)  4.5085566  0.1022201  44.106361 5.762099e-203
## TASK2       -0.9207815  0.2081109  -4.424476  1.123234e-05
## TASK3       -1.6738481  0.1457607 -11.483531  4.699458e-28
## TASK4       -2.6992523  0.1479884 -18.239617  5.849323e-61
```

Again, the parameter estimates and standard errors are given in the code summary above. The interpretations are different than for the cell mean coding and are as follows: $beta_1$ is the intercept which again is the mean log cotinine level for priming, which is the reference level. $beta_2$ is the difference between the mean log cotinine level for barning and the mean log cotinine level for priming. Similarly, $beta_3$ is now the difference btween the mean log cotinine level for topping and the mean log cotinine level for priming, and $beta_4$ is the difference between the mean log cotinine level for work not involving tobacco contact and the mean log cotinine level for priming.

The TASK value corresponding with priming is 1, and so we want to test that the mean cotinine level for TASK1 is greater than the mean cotonine level for all others. We can test the following hypothesis using the cell mean coding:

$$H_0 = \beta_1 = (\beta_2 + \beta_3 + \beta_4)/3 \quad vs. \quad H_A = \beta_1 > (\beta_2 + \beta_3 + \beta_4)/3$$

This null hypothesis corresponds to:

$$H_0 = \beta_1 - \frac{1}{3}\beta_2 - \frac{1}{3}\beta_3 - \frac{1}{3}\beta_4 = 0$$

so $\theta_{\mathbf{0}} = [0]$ and $\mathbf{C} = [1 \quad -1/3 \quad -1/3 \quad -1/3]$

part ii: Two-Way ANOVA:

For these questions, use the log of salivary cotinine as the response and task and wet as predictors.

```
tobacco$WET <- tobacco$WET %>% as.factor()
tobacco$TASK <- tobacco$TASK %>% as.factor()
tobacco2 = tobacco %>% mutate(LOGCOT = log(COTININE),
                              WET0TASK1 = case_when(WET == 0 & TASK == 1 ~ 1,
                                                   WET != 0 | TASK != 1 ~ 0),
                              WET1TASK1 = case_when(WET == 1 & TASK == 1 ~ 1,
                                                   WET != 1 | TASK != 1 ~ 0),
                              WET0TASK2 = case_when(WET == 0 & TASK == 2 ~ 1,
                                                   WET != 0 | TASK != 2 ~ 0),
                              WET1TASK2 = case_when(WET == 1 & TASK == 2 ~ 1,
                                                   WET != 1 | TASK != 2 ~ 0),
                              WET0TASK3 = case_when(WET == 0 & TASK == 3 ~ 1,
                                                   WET != 0 | TASK != 3 ~ 0),
                              WET1TASK3 = case_when(WET == 1 & TASK == 3 ~ 1,
                                                   WET != 1 | TASK != 3 ~ 0),
                              WET0TASK4 = case_when(WET == 0 & TASK == 4 ~ 1,
                                                   WET != 0 | TASK != 4 ~ 0),
                              WET1TASK4 = case_when(WET == 1 & TASK == 4 ~ 1,
                                                   WET != 1 | TASK != 4 ~ 0))
```

(a):

Fit the two-way ANOVA model with full interaction, and interpret all parameter estimates in your model, clearly stating which coding scheme you used. Discuss the validity of the HILE Gauss assumptions for this model.

Consider the following model using the cell mean coding scheme:

$$y = \beta_1 I_{W0,T1} + \beta_2 I_{W1,T1} + \beta_3 I_{W0,T2} + \beta_4 I_{W1,T2} + \beta_5 I_{W0,T3} + \beta_6 I_{W1,T3} + \beta_7 I_{W0,T4} + \beta_8 I_{W1,T4}$$

where y is the log cotinine, and $I_{Ti,Wj}$ is the indicator function associated with the ith task and jth wet categories.

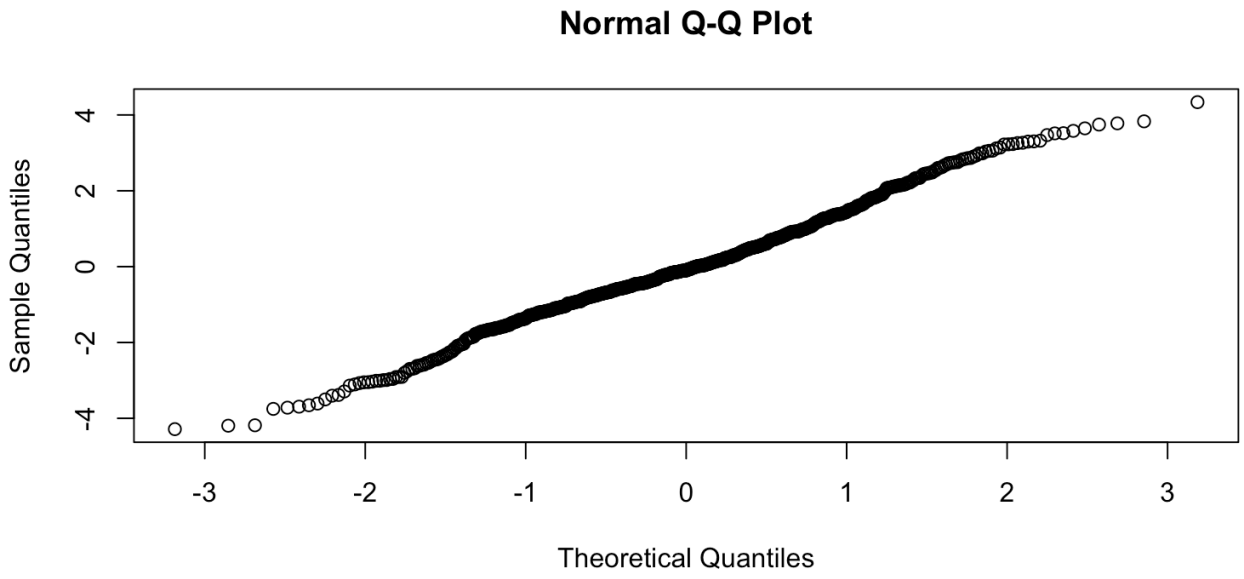
```
fit <- lm(LOGCOT ~ -1 + WET0TASK1 + WET1TASK1 + WET0TASK2 + WET1TASK2 + WET0TASK3 + WET1TASK3 + WET0TASK4 + WET1TASK4, data = tobacco2)
summary(fit)$coefficients
```

```
##           Estimate Std. Error   t value      Pr(>|t|)
## WET0TASK1 4.269337  0.1854712 23.018868  2.548339e-87
## WET1TASK1 4.613116  0.1226196 37.621351 6.459772e-169
## WET0TASK2 3.542748  0.2271549 15.596174  3.665240e-47
## WET1TASK2 3.667024  0.3013550 12.168449  5.528611e-31
## WET0TASK3 2.688185  0.2152536 12.488456  2.142659e-32
## WET1TASK3 2.879303  0.1187505 24.246652  2.808906e-94
## WET0TASK4 1.808889  0.1238562 14.604754  2.911782e-42
## WET1TASK4 1.810534  0.2130902  8.496563  1.211849e-16
```

The estimates for each of the parameters are shown in the summary statistics above. Since the cell mean coding is used, the interpretations are clear; each parameter represents the mean log cotinine level for the combination of WET and TASK levels listed. For example, β_1 is estimated by 4.269337 noted by the WET0TASK1 indicator variable.

In terms of HILE Gauss assumptions for this model, the only assumptions that are generally checked for ANOVA are H, I, and Gauss. The independence assumption is dependent on the design and the sampling scheme, and from the description of the design, I do not see any issues that would question the validity of the independence assumption. In terms of the homogeneity and gaussian errors assumptions, we can perform tests as done below to check these assumptions. We also note that the design is unbalanced (i.e. the sample size per cell ranges from 25 to 161), which is something to consider when using this model.

```
qqnorm(fit$residuals)
```



The linearity of the QQ-Plot above verifies the gaussian errors assumption.

```
leveneTest(LOGCOT ~ TASK*WET, data = tobacco2, center=median)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   7  18.635 < 2.2e-16 ***
##       686
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the Levene Test above, the p-value is < 2.2e-16, which means we reject the hypothesis that the homoegeneity of variance asssumption is satisfied. We should proceed with caution when using this model.

(b):

Based on this model, create a table of the estimated mean log cotinine levels, associated standard errors, and how each estimated mean is obtained from the model parameters (e.g., $\hat{\beta}_0 + \hat{\beta}_1$) for each task-wet combination.

```
beta <- intToUtf8(946)
params <- summary(fit)$coefficients[1:8,1:2] %>% as.data.frame()
params <- data.frame(params, "Parameter" = paste0(beta, c(1:8)))
params %>% knitr::kable(align = c("c", "c"))
```

	Estimate	Std..Error	Parameter
WET0TASK1	4.269337	0.1854712	β1
WET1TASK1	4.613116	0.1226196	β2
WET0TASK2	3.542748	0.2271549	β3
WET1TASK2	3.667024	0.3013550	β4
WET0TASK3	2.688185	0.2152536	β5
WET1TASK3	2.879303	0.1187505	β6
WET0TASK4	1.808889	0.1238562	β7
WET1TASK4	1.810534	0.2130902	β8

The estimates for mean log cotinine levels, their standard errors, and their relationship to the parameters are summarized in the table above. With cell mean coding, the parameter interpretations are clear; they each simply represent the mean log cotinine level for one task-wet combination.

part iii: The Full Model in Every Cell:

For these questions, use the log of salivary cotinine as the response and task and lnnsmoke as predictors.

```
tobacco3 <- tobacco %>% mutate(LOGCOT = log(COTININE),
                               TASK1 = case_when(TASK == 1 ~ 1, TASK != 1 ~ 0),
                               TASK2 = case_when(TASK == 2 ~ 1, TASK != 2 ~ 0),
                               TASK3 = case_when(TASK == 3 ~ 1, TASK != 3 ~ 0),
                               TASK4 = case_when(TASK == 4 ~ 1, TASK != 4 ~ 0))
```

(a):

Fit the full model in every cell. Provide and interpret estimates of all parameters for this model.

Consider the following model using the reference cell coding scheme:

$$y = \beta_1 + \beta_2 I_{T2} + \beta_3 I_{T3} + \beta_4 I_{T4} + \beta_5 X + \beta_6 I_{T2} X + \beta_7 I_{T3} X + \beta_8 I_{T4} X$$

where y is the log cotinine, X is the lnnsmoke varaible, and I_{Ti} is the indicator function associated with the ith task level. Note that TASK1 is the reference.

```
fit = lm(LOGCOT ~ factor(TASK) + factor(TASK)*LNNSMOKE, data = tobacco3)
summary(fit)$coefficients
```

```
##                          Estimate Std. Error    t value      Pr(>|t|)
## (Intercept)             4.3344142 0.09319895  46.507113 2.802358e-214
## factor(TASK)2          -1.2210290 0.19086147  -6.397462  2.924652e-10
## factor(TASK)3          -2.3221216 0.13561455 -17.122953  5.432182e-55
## factor(TASK)4          -3.4207011 0.13356892 -25.610007 4.927483e-102
## LNNSMOKE                0.2945994 0.08869134   3.321625  9.424000e-04
## factor(TASK)2:LNNSMOKE  0.4274696 0.16993370   2.515508  1.211344e-02
## factor(TASK)3:LNNSMOKE  0.9359153 0.12592833   7.432126  3.191248e-13
## factor(TASK)4:LNNSMOKE  1.4843094 0.13531844  10.969010  6.613806e-26
```

The estimates for each parameter can be seen by the summary above. The intercept, β_1, is the mean log continine level for priming when lnnsmoke is 0. The factor(TASK)2 estimate, for β_2, is the difference between the mean log continine level for barning and for priming when lnnsmoke is 0. Similarly, The factor(TASK)2 estimate, for β_3, is the difference between the mean log continine level for topping and for priming when lnnsmoke is 0 and the factor(TASK)4 estimate, for β_4, is the difference between the mean log continine level for work not involving tobacco and for priming when lnnsmoke is 0. The LNNSMOKE estimate, for β_5, is the

mean increase log continine level for a one unit increase in lnnsmoke (the natural log of 1 + number of cigarettes smoked a day) in those whose task is priming. Similiarly, factor(TASK)i:LNNSMOKE variables, estimating $\beta_6, \beta_7, beta_8$, are the mean increase log continine level for a one unit increase in lnnsmoke for those whose task is barning, topping, and no tobacco involvement, respectively.

(b):

Report an appropriate test of whether task is related to cotinine levels. If this test is significant, report step-down tests to determine exactly where differences lie. For all tests reported, be sure to state H_0 clearly and give explicit justification for which tests were used and why.

In order to test whether task is realted to cotinine levels, we want to test the following hypotheses:

$$H_0 = 0 = \beta_2 = \beta_3 = \beta_4 \quad \& \quad 0 = \beta_6 = \beta_7 = \beta_8$$

which is equivalent to:

$$H_0 = \beta_2 = 0, \beta_3 = 0, \beta_4 = 0, \beta_6 = 0, \beta_7 = 0, \beta_8 = 0,$$

These will test whether the intercepts for each task and the slopes for each task are equivalent to each other. In order to test these hypotheses, we can use the overall F test, where:

$$F = (SSH/a)/\hat{\sigma}^2 \sim F_{G-2,n-G}$$

where $SSH = (\hat{\theta} - \theta_0)' M^{-1} (\hat{\theta} - \theta_0)$, $G = 8, n = 694$, and $^2 = $. It should also be noted that $M = C(X'X)^{-1}C'$.

```
X = tobacco3 %>% mutate(INT = 1, LNNTASK2 = LNNSMOKE*TASK2, LNNTASK3 = LNNSMOKE*TASK3, LNNTASK4 = LNNSMOKE*TASK4) %>% select(INT, TASK2, TASK
3, TASK4, LNNSMOKE, LNNTASK2, LNNTASK3, LNNTASK4) %>% as.matrix()
C = matrix(c(0, 0, 0, 0, 0, 0,
             1, 0, 0, 0, 0, 0,
             0, 1, 0, 0, 0, 0,
             0, 0, 1, 0, 0, 0,
             0, 0, 0, 0, 0, 0,
             0, 0, 0, 1, 0, 0,
             0, 0, 0, 0, 1, 0,
             0, 0, 0, 0, 0, 1), nrow = 6)

linearHypothesis(fit, C)
```

```
## Linear hypothesis test
##
## Hypothesis:
## factor(TASK)2 = 0
## factor(TASK)3 = 0
## factor(TASK)4 = 0
## factor(TASK)2:LNNSMOKE = 0
## factor(TASK)3:LNNSMOKE = 0
## factor(TASK)4:LNNSMOKE = 0
##
## Model 1: restricted model
## Model 2: LOGCOT ~ factor(TASK) + factor(TASK) * LNNSMOKE
##
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    692 1806.03
## 2    686  883.86  6    922.17 119.29 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the code summary above, the F statistic is 119.29 and the p-value is < 2.2e-16, so we can reject the null hypothesis. This means there is evidence to reject the fact the the slopes for all four task levels are equal and the intercepts for all four task levels are equal.

Next, we perform a step down test to test whether the intercepts are equal. The null hypothesis for this test will be the first set of hypotheses previously listed above.

```
C = matrix(c(0, 0, 0,
             1, 0, 0,
             0, 1, 0,
             0, 0, 1,
             0, 0, 0,
             0, 0, 0,
             0, 0, 0,
             0, 0, 0), nrow = 3)
linearHypothesis(fit, C)
```

```
## Linear hypothesis test
##
## Hypothesis:
## factor(TASK)2 = 0
## factor(TASK)3 = 0
## factor(TASK)4 = 0
##
## Model 1: restricted model
## Model 2: LOGCOT ~ factor(TASK) + factor(TASK) * LNNSMOKE
##
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    689 1785.88
## 2    686  883.86  3    902.01 233.36 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the code summary above, the F statistic is 233.36 and the p-value is < 2.2e-16, so we can reject the null hypothesis. This means there is evidence to reject the fact the the intercepts for all four task levels are equal.

We will now follow the same process with the slopes for each task level. The null hypothesis for this test will be the second set of hypotheses previously listed above at the first test for this problem part.

```
C = matrix(c(0, 0, 0,
             0, 0, 0,
             0, 0, 0,
             0, 0, 0,
             0, 0, 0,
             1, 0, 0,
             0, 1, 0,
             0, 0, 1), nrow = 3)

linearHypothesis(fit, C)
```

```
## Linear hypothesis test
##
## Hypothesis:
## factor(TASK)2:LNNSMOKE = 0
## factor(TASK)3:LNNSMOKE = 0
## factor(TASK)4:LNNSMOKE = 0
##
## Model 1: restricted model
## Model 2: LOGCOT ~ factor(TASK) + factor(TASK) * LNNSMOKE
##
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    689 1053.51
## 2    686  883.86  3    169.64 43.889 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the code summary above, the F statistic is 43.889 and the p-value is < 2.2e-16, so we can reject the null hypothesis. This means there is evidence to reject the fact the the slopes for all four task levels are equal.

Next, we would want to step down even further to test pairwise comparisons of the intercepts and of the slopes. For the intercepts, to test whether the intercepts for TASK1 and TASK2 are equivalent, we would want to test $H_0 : \beta_1 = \beta_1 + \beta_2 \rightarrow \beta_2 = 0$. To test the equivalence of TASK1 and TASK3, as well as TASK1 and TASK4, we can follow the same process. For TASK2 and TASK3, we would want to test $H_0 : \beta_2 = \beta_3 \rightarrow \beta_2 - \beta_3 = 0$. Similar hypotheses would be tested for TASK2 and TASK4, as well as TASK3 and TASK4. We will perform a normal F test, but will look to reject the null as p-values smaller than $\alpha = 0.05/6 = .008$ using the Bonferroni correction, since we are running 6 tests. All F-statistics have degrees of freedom (1, 686), since we are performing one test with a size $n - G$ fitted model.

```
### 1 and 2
C = matrix(c(0, 1, 0, 0, 0, 0, 0, 0), nrow = 1)
test<-linearHypothesis(fit, C)
f12 <- test$F[2]
p12 <- test$`Pr(>F)`[2]

### 1 and 3
C = matrix(c(0, 0, 1, 0, 0, 0, 0, 0), nrow = 1)
test<-linearHypothesis(fit, C)
f13 <- test$F[2]
p13 <- test$`Pr(>F)`[2]

### 1 and 4
C = matrix(c(0, 0, 0, 1, 0, 0, 0, 0), nrow = 1)
test<-linearHypothesis(fit, C)
f14 <- test$F[2]
p14 <- test$`Pr(>F)`[2]

### 2 and 3
C = matrix(c(0, 1, -1, 0, 0, 0, 0, 0), nrow = 1)
test<-linearHypothesis(fit, C)
f23 <- test$F[2]
p23 <- test$`Pr(>F)`[2]

### 2 and 4
C = matrix(c(0, 1, 0, -1, 0, 0, 0, 0), nrow = 1)
test<-linearHypothesis(fit, C)
f24 <- test$F[2]
p24 <- test$`Pr(>F)`[2]

### 3 and 4
C = matrix(c(0, 0, 1, -1, 0, 0, 0, 0), nrow = 1)
test<-linearHypothesis(fit, C)
f34 <- test$F[2]
p34 <- test$`Pr(>F)`[2]

df <- data.frame(Test = c("TASK1:TASK2", "TASK1:TASK3", "TASK1:TASK4", "TASK2:TASK3", "TASK2:TASK4", "TASK3:TASK4"), Fvalue = c(f12, f13, f14
, f23, f24, f34), Pvalue = c(p12, p13, p14, p23, p24, p34))
df %>% knitr::kable(align = c("c", "c"), digits = c(10,  10, 20))
```

Test	Fvalue	Pvalue
TASK1:TASK2	40.92752	2.924652e-10
TASK1:TASK3	293.19551	0.000000e+00
TASK1:TASK4	655.87245	0.000000e+00
TASK2:TASK3	32.37628	1.882119e-08
TASK2:TASK4	131.13804	0.000000e+00
TASK3:TASK4	63.99177	5.325440e-15

From the summary table above, we can reject the null hypothesis in every case. That is, there is evidence to support the fact that all combinations of TASK intercepts are significnantly different than one another. We will perform an analogous analysis for the pairwise testing of the equivalence of slopes of the TASK levels.

```
### 1 and 2
C = matrix(c(0, 0, 0, 0, 0, 1, 0, 0), nrow = 1)
test<-linearHypothesis(fit, C)
f12 <- test$F[2]
p12 <- test$`Pr(>F)`[2]

### 1 and 3
C = matrix(c(0, 0, 0, 0, 0, 0, 1, 0), nrow = 1)
test<-linearHypothesis(fit, C)
f13 <- test$F[2]
p13 <- test$`Pr(>F)`[2]

### 1 and 4
C = matrix(c(0, 0, 0, 0, 0, 0, 0, 1), nrow = 1)
test<-linearHypothesis(fit, C)
f14 <- test$F[2]
p14 <- test$`Pr(>F)`[2]

### 2 and 3
C = matrix(c(0, 0, 0, 0, 0, 1, -1, 0), nrow = 1)
test<-linearHypothesis(fit, C)
f23 <- test$F[2]
p23 <- test$`Pr(>F)`[2]

### 2 and 4
C = matrix(c(0, 0, 0, 0, 0, 1, 0, -1), nrow = 1)
test<-linearHypothesis(fit, C)
f24 <- test$F[2]
p24 <- test$`Pr(>F)`[2]

### 3 and 4
C = matrix(c(0, 0, 0, 0, 0, 0, 1, -1), nrow = 1)
test<-linearHypothesis(fit, C)
f34 <- test$F[2]
p34 <- test$`Pr(>F)`[2]

df <- data.frame(Test = c("TASK1:TASK2", "TASK1:TASK3", "TASK1:TASK4", "TASK2:TASK3", "TASK2:TASK4", "TASK3:TASK4"), Fvalue = c(f12, f13, f14
, f23, f24, f34), Pvalue = c(p12, p13, p14, p23, p24, p34))
df %>% knitr::kable(align = c("c", "c"), digits = c(10,  10, 20))
```

Test	Fvalue	Pvalue
TASK1:TASK2	6.327780	1.211344e-02
TASK1:TASK3	55.236500	3.191248e-13
TASK1:TASK4	120.319184	0.000000e+00
TASK2:TASK3	8.913427	2.931579e-03
TASK2:TASK4	35.506803	4.063003e-09
TASK3:TASK4	16.311803	5.979385e-05

From the summary table above, we can reject the null hypothesis in almost every case. That is, there is evidence to support the fact that all combinations of TASK slopes (except TASK1:TASK2) are significnantly different than one another. the p-value for TASK1:TASK2 is greater than the Bonferroni corrected $\alpha = 0.008$ so there is not evidence to reject the fact that the mean increase in log cotinine levels when lnnsmoke is increased by one is different between priming and barning tasks.