Raquel Prado
Department of Statistics
Fall 2018 Fall 2018

Name: _____

Midterm Exam

The midterm is closed-book, you are only allowed to use one page of notes and a calculator.
Please attach your formula sheet.

Problem	Possible Points	Your points
1	20	
2	15	
3	30	
4	8	
5	27	
Total	100	

Statistics
Midterm Exam AMS-204.

In the HELP (Health Evaluation and Linkage to Primary Care) study, investigators were interested in determining predictors of severe depressive symptoms (measured by the Center for Epidemiologic Studies–Depression Scale, `cesd`) amongst a cohort enrolled at a substance abuse treatment facility. The `HELPrct` dataset is a subset from the HELP study data restricted to 453 subjects.

The following sets of commands summarize information about a number of linear models with response variable `cesd` where:

- `cesd`: depression score with high values indicating more depressive symptoms;
- `substance`: factor that indicates the type of substance abuse with 3 categories, namely, alcohol, cocaine, or heroine;
- `mcs`: mental component score (continuous measure of mental well-being) with lower scores indicating worse status;
- `homeless`: factor indicating housing status with two categories, homeless or housed (note that **R** codes housed as `homelesshoused`);
- `sex`: factor with levels male or female.

```
> library(mosaicData)
> attach(HELPrct)
> My_HELPrct=data.frame(cesd=cesd,substance=substance,mcs=mcs,
+                       homeless=homeless,sex=sex)
> detach(HELPrct)
> head(My_HELPrct,4)

  cesd substance       mcs homeless    sex
1   49   cocaine 25.111990   housed   male
2   30   alcohol 26.670307 homeless   male
3   39    heroin  6.762923   housed   male
4   15    heroin 43.967880   housed female

> tail(My_HELPrct,4)

    cesd substance      mcs homeless    sex
450   37   alcohol 62.17550   housed   male
451   28    heroin 33.43454 homeless female
452   11   cocaine 54.42482 homeless   male
453   35   alcohol 30.21223 homeless   male

> summary(My_HELPrct)
```

```
      cesd         substance        mcs             homeless       sex
 Min.   : 1.00   alcohol:177   Min.   : 6.763   homeless:209   female:107
 1st Qu.:25.00   cocaine:152   1st Qu.:21.676   housed  :244   male  :346
 Median :34.00   heroin :124   Median :28.602
 Mean   :32.85                 Mean   :31.677
 3rd Qu.:41.00                 3rd Qu.:40.941
 Max.   :60.00                 Max.   :62.175
```

```
> attach(My_HELPrct)
```

1. (20 points total) Consider the following model, referred to as `M1`:

```
> M1=lm(cesd~mcs)
> summary(M1)

Call:
lm(formula = cesd ~ mcs)

Residuals:
     Min       1Q   Median       3Q      Max
-27.3593  -6.7277  -0.0024   6.2374  24.4239

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 53.90219    1.14723   46.98   <2e-16 ***
mcs         -0.66467    0.03357  -19.80   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.164 on 451 degrees of freedom
Multiple R-squared:  0.465,        Adjusted R-squared:  0.4638
F-statistic:   392 on 1 and 451 DF,  p-value: < 2.2e-16
```

   (a) (6 points) Write down model **M1** in matrix form, i.e., specify $\mathbf{y}$, $\mathbf{X}$, $\boldsymbol{\beta}$, and $\boldsymbol{\epsilon}$ in

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

(b) (5 points) Based on the `R` output above, what are the least squares estimates (LSE) of the components of the $\boldsymbol{\beta}$ vector? How do you interpret the LSE of second component of this vector? What does it say about the relationship between `cesd` and `mcs`?

(c) (5 points) The `t-value -19.80` and its corresponding `p-value` can be used for a particular `t-test`. Write down the null and alternative hypotheses in this test as well as your conclusion (i.e., reject the null or fail to reject) in the context of this example. Use a significance level of 0.05.

(d) (4 points) The output above also provides an `F-statistic` of `392` and a corresponding `p-value`. Write down the null and alternative hypothesis of this `F-test` Is this test equivalent to the `t-test` above? Can we think of this `F-test` as a way to compare two models? If so, what are these models?

2. (15 points total) Now consider the following model `M2`:

```
> M2=lm(cesd~substance*sex)
> summary(M2)

Call:
lm(formula = cesd ~ substance * sex)

Residuals:
     Min       1Q   Median       3Q      Max
-31.1667  -8.8191   0.8919   8.1348  27.0244

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)                40.278      2.015  19.984  < 2e-16 ***
substancecocaine           -7.302      2.762  -2.644  0.00849 **
substanceheroin            -2.111      2.989  -0.706  0.48044
sexmale                    -7.413      2.258  -3.283  0.00111 **
substancecocaine:sexmale    2.545      3.160   0.805  0.42098
substanceheroin:sexmale     3.065      3.396   0.903  0.36720
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.09 on 447 degrees of freedom
Multiple R-squared:  0.07655,        Adjusted R-squared:  0.06622
F-statistic: 7.411 on 5 and 447 DF,  p-value: 1.081e-06
```

   (a) (5 points) Specify the ANOVA model considered here, i.e., provide the equation that specifies the model for any given response variable $y_{i,j,k}$ where i indexes the level of the factor substance ($i = 1, 2, 3$), j indexes the level of the factor sex ($j = 1, 2$), and k indexes the individual in the group that corresponds to substance level i and sex level j. Assume that there are $n_{i,j}$ individuals for a given combination of factor levels, i.e., $k = 1, \ldots, n_{i,j}$.

(b) (2 points) Based on the LSEs of the model parameters provided by the **R** output, what is the fitted value obtained from this model for the `cesd` of a female who consumed alcohol?

(c) (3 points) Based on the LSEs of the model parameters provided by the **R** output, what is the fitted value obtained from this model for the `cesd` of a male who consumed heroine?

(d) (5 points) The following ANOVA table provides an `F-value` of `0.4989` and a corresponding `p-value` of `0.6075227`. These values correspond to a particular test. Write down the null and alternative hypotheses for this test and provide your conclusions in the context of the example and model considered here. Use a significance level of 0.05. Would you suggest fitting a different model based on this test? If so, what model would you consider?

```
> anova(M2)
Analysis of Variance Table

Response: cesd
               Df Sum Sq Mean Sq F value    Pr(>F)
```

```
substance       2   2704 1352.06  9.2454 0.0001163 ***
sex             1   2569 2569.02 17.5671 3.344e-05 ***
substance:sex   2    146   72.96  0.4989 0.6075227
Residuals     447  65369  146.24
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

3. (30 points total) Now consider the following model M3:

```
> M3=lm(cesd~sex+substance+mcs)
> summary(M3)

Call:
lm(formula = cesd ~ sex + substance + mcs)

Residuals:
     Min       1Q   Median       3Q      Max
-24.4790  -6.4370   0.2378   6.3425  24.6366

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      57.44472    1.42410  40.338  < 2e-16 ***
sexmale          -3.22924    1.00255  -3.221 0.001370 **
substancecocaine -3.68922    0.99807  -3.696 0.000246 ***
substanceheroin  -1.85677    1.05737  -1.756 0.079768 .
mcs              -0.64351    0.03365 -19.126  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.972 on 448 degrees of freedom
Multiple R-squared:  0.4905,        Adjusted R-squared:  0.486
F-statistic: 107.8 on 4 and 448 DF,  p-value: < 2.2e-16

> anova(M3)

Analysis of Variance Table

Response: cesd
           Df Sum Sq Mean Sq F value    Pr(>F)
sex         1   2287  2286.7  28.404 1.563e-07 ***
substance   2   2986  1493.2  18.548 1.823e-08 ***
mcs         1  29449 29449.0 365.801 < 2.2e-16 ***
Residuals 448  36066    80.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(a) (5 points) Write down the equation(s) that define the linear model M3.

(b) (15 points) The ANOVA table provides 3 `p-values` that correspond to 3 specific `F-tests`. Provide the null and alternative hypotheses for each of these tests and your conclusions in the context of this example. Use a significance level of 0.05.

Statistics
Midterm Exam AMS-204.

(c) (5 points) Based on the LSEs obtained for this model, what is the fitted value of `cesd` for a female who consumed heroine and has a `mcs` score of 33.43?

(d) (5 points) The `t-value -1.756` and corresponding `p-value` 0.079768 on the summary table for model `M3` above corresponds to a particular test. Write down the null and alternative hypotheses in this case and provide your conclusions in the context of this example. Use a significance level of 0.05. Based on your conclusions, what are the implications in terms of predictive values of `cesd` for subjects that consumed heroine vs those subjects who consumed alcohol?

4. (8 points) Which of the 3 models considered above would you choose? Justify your reasoning.

5. (27 points total) Consider the following two models, `M4` and `M5`:

```
> M4=lm(cesd~mcs+sex)
> summary(M4)

Call:
lm(formula = cesd ~ mcs + sex)

Residuals:
     Min       1Q   Median       3Q      Max
-26.4794  -6.3204   0.4515   6.3082  24.7669

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 55.78720    1.30904  42.617  < 2e-16 ***
mcs         -0.65308    0.03353 -19.476  < 2e-16 ***
sexmale     -2.94873    1.01249  -2.912  0.00377 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.088 on 450 degrees of freedom
Multiple R-squared:  0.4749,        Adjusted R-squared:  0.4726
F-statistic: 203.5 on 2 and 450 DF,  p-value: < 2.2e-16
```

and

```
> M5=lm(cesd~mcs*sex)
> summary(M5)

Call:
lm(formula = cesd ~ mcs * sex)

Residuals:
     Min       1Q   Median       3Q      Max
-26.2228  -6.4094   0.3737   6.3473  24.3220

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 57.34852    2.23939  25.609   <2e-16 ***
mcs         -0.70703    0.07117  -9.934   <2e-16 ***
sexmale     -5.01100    2.60451  -1.924    0.055 .
mcs:sexmale  0.06935    0.08070   0.859    0.391
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.091 on 449 degrees of freedom
Multiple R-squared:  0.4758,        Adjusted R-squared:  0.4723
F-statistic: 135.8 on 3 and 449 DF,  p-value: < 2.2e-16
```

(a) (5 points) Write down the equation(s) that define model `M4`.

(b) (7 points) Write down the equation(s) that define model `M5`.

(c) (7 points) What is the fitted value for `cesd` obtained from model `M4` for a male with `mcs` score of 33? What is the fitted value obtained from model `M5` for the same individual?

(d) (6 points) Based on the information provided by the summary tables of models `M4` and `M5`, which of these two models would you choose? Justify your answer.

(e) (3 points) Based on the information provided above would you eliminate the factor `sex` from the model? Justify your answer.