R Functions for Exam 3

You will not need to generate any R commands for this exam. You will be shown one or more commands and asked to write what the result(s) will be. You will also be shown commands together with their output and asked to explain what the output means.

`cor()`
You should know that this function computes the correlation of two variables. Assuming you understand the concept of correlation, there's nothing else to know about `cor()`.

`pf()`
You should know that this function computes cumulative probabilities from F distributions. For example, `pf(F,df1,df2,lower.tail=FALSE)` gives you the probability of an F value greater than or equal to the value of `F` that you input, according to an F distribution with `df1` and `df2` degrees of freedom.

You should also know that `pf()` outputs a probability, meaning the result is always between 0 and 1. Higher values of `F` always give smaller probabilities (e.g., the probability of an F value greater than `2` must be less than the probability of an F value greater than `1`). The entire F distribution is above zero, so if you input a negative `F`, `pf()` will give a result of `1`.

`qf()`
You should know that this function computes quantiles for F distributions. The input you give it is a probability. For example, `qf(alpha,df1,df2,lower.tail=FALSE)` gives you the value of *F* that has a probability `alpha` of being exceeded.

`lm()`
You should know how `lm()` is used to get regression coefficients. If you see a command like `lm(variable1 ~ variable2 + variable3)`, you should know that `variable1` is the outcome, and `variable2` and `variable3` are the predictors. The output of `lm()` displays the regression coefficients, including the intercept. An example is below. You should be able to understand what all of these numbers mean.

```
Coefficients:
(Intercept)    variable2    variable3
   -0.10294      0.08038     -0.16039
```

`summary()` of a regression
If you input the results of a regression—from `lm()`—to the `summary()` function, you get a set of statistics for hypothesis tests for that regression. First, you get everything related to testing the reliability of each regression coefficient, including the name of the predictor that coefficient is for, the value of the coefficient in the sample, the coefficient's standard error, the t value (which is the coefficient divided by its standard error), and the two-tailed

p-value. An example set of output is below. You should be able to understand everything shown here.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.10294    0.08908  -1.156   0.2507
variable2    0.08038    0.09193   0.874   0.3841
variable3   -0.16039    0.08402  -1.909   0.0592 .
```

The other thing you get from the summary of a regression is information for testing the reliability of the regression as a whole, meaning the question of whether the regression explains anything meaningful about the outcome. First, you get the residual standard error, which is the square root of $MS_{residual}$. Since $MS_{residual}$ is an estimate of the population variance, the residual standard error is an estimate of the population standard deviation. The degrees of freedom for the residual standard error is also shown; this is the same thing as $df_{residual}$. The exam might ask you to write down the value of $MS_{residual}$ or $SS_{residual}$, based on the output of `summary(regression)`. The next thing the output gives you is R-squared, which R calls `Multiple R-squared` to distinguish it from r^2 from correlation (when there's only a single predictor). The Adjusted R-squared is something that we haven't discussed and that you won't be tested on (it estimates the variability explained just by chance and subtracts that from R-squared). Finally, the output gives you `F`, the degrees of freedom for the regression and the residual (in that order; notice that $df_{residual}$ is repeated from above), and the p-value. You should be able to understand all of these numbers—the residual standard error, R-squared, F, both degrees of freedom, and p—as well as how these numbers relate to each other.

```
Residual standard error: 0.8901 on 97 degrees of freedom
Multiple R-squared: 0.04312,  Adjusted R-squared: 0.02339
F-statistic: 2.186 on 2 and 97 DF,  p-value: 0.1179
```

`anova()`

If you input the model for an ANOVA—from `lm()`—to the `anova()` function, you get a set of statistics for hypothesis tests for the ANOVA. For example, you should be able to understand all the numbers in the output below. They show you *df*, *SS*, *MS*, *F*, and *p* for testing the main effects of both factors and their interaction. The last line shows *df*, *SS*, and *MS* for the residual.

```
> anova(lm(X ~ factor1*factor2))

Analysis of Variance Table

Response: X
                     Df Sum Sq Mean Sq F value Pr(>F)
factor1               2 10.405   5.203  1.9407 0.1697
factor2               1  3.360   3.360  1.2535 0.2761
factor1:factor2       2  2.502   1.251  0.4667 0.6337
Residuals            20 53.617   2.681
```