STAT 425: Midterm 1

Instructor: A. Chronopoulou

Thursday, October 7, 2021 **5pm**

Instructions:

- This is a closed-notes, closed-books exam. You are allowed to have a cheat sheet.
- You can use your calculator, but cell-phones are not allowed to be used as calculators.
- The duration of the exam is 1h 20min.
- Read the questions carefully!
- Please mark your answers clearly, reference the R output you use and make sure that you *show your work!*
- Be precise with the use of appropriate notation.
- In the last page of the exam (i.e. pg. 8), you will find a list of T distribution and F distribution values.

In the end of the exam, you have 15 minutes to scan **your exam** and upload it on *Gradescope*.

GOOD LUCK!

1. **Theoretical Question** *[30 points]*
   Let $\hat{\beta}_{GLS}$ be the Generalized Least Squares estimator of β in

   $$\mathbf{y} = \mathbf{X}\beta + \varepsilon,$$

   where $\mathbf{y}$ is an $n \times 1$ random vector, $\mathbf{X}$ is an $n \times p$ design matrix with full rank $p < n$, and ε is an $n \times 1$ random vector with mean 0 and variance-covariance $Cov(\varepsilon) = \Sigma$.

   (a) *[10 points]* Show that $E(\mathbf{y}) = \mathbf{X}\beta$ and $Cov(\mathbf{y}) = \Sigma$.

   (b) *[20 points]* Show that $E(\hat{\beta}_{GLS}) = \beta$ and $Cov(\hat{\beta}_{GLS}) = (\mathbf{X}^T\Sigma\mathbf{X})^{-1}$.

2. **Simple Linear Regression: Airfreight Containers** *[70 points]*
   A substance used in biological and medical research is shipped by airfreight to users in cartons of 1,000 ampules[1]. Data involving 10 shipments were collected on the number of times the carton was transferred from one address to another over the shipment route, X, and the number of samples found to be broken upon arrival, Y.

   Based on the `R` output provided to you in page 3, answer the following questions:

   (a) *[5 points]* State the fitted regression line.

   (b) *[5 points]* Estimate the expected number of broken ampules when *one* transfer is made.

   (c) *[5 points]* Estimate the increase in the expected number of ampules broken when there are *two* transfers as compared to *one* transfer.

   (d) *[10 points]* Conduct a t-test to decide whether or not there is a linear association between number of times a carton is transferred and number of broken ampules. Use level of significance 0.05. State the alternatives, decision rule and conclusion.

   (e) *[10 points]* Interpret the intercept of the regression line in the context of the problem, and obtain a 95% confidence interval for it.

   (f) *[10 points]* Because of changes in airline routines, shipments may have to be transferred more frequently than in the past. Obtain a 99% interval estimate for the expected number of ampules broken when three transfers are made.

   (g) *[10 points]* The next shipment will entail three transfers. Obtain a 99% interval estimate for the number of broken ampules for this upcoming shipment.

   (h) *[5 points]* Will the 99% confidence band for the regression line, when three transfers are made, be wider or narrower compared to the interval in (f)? Explain why.

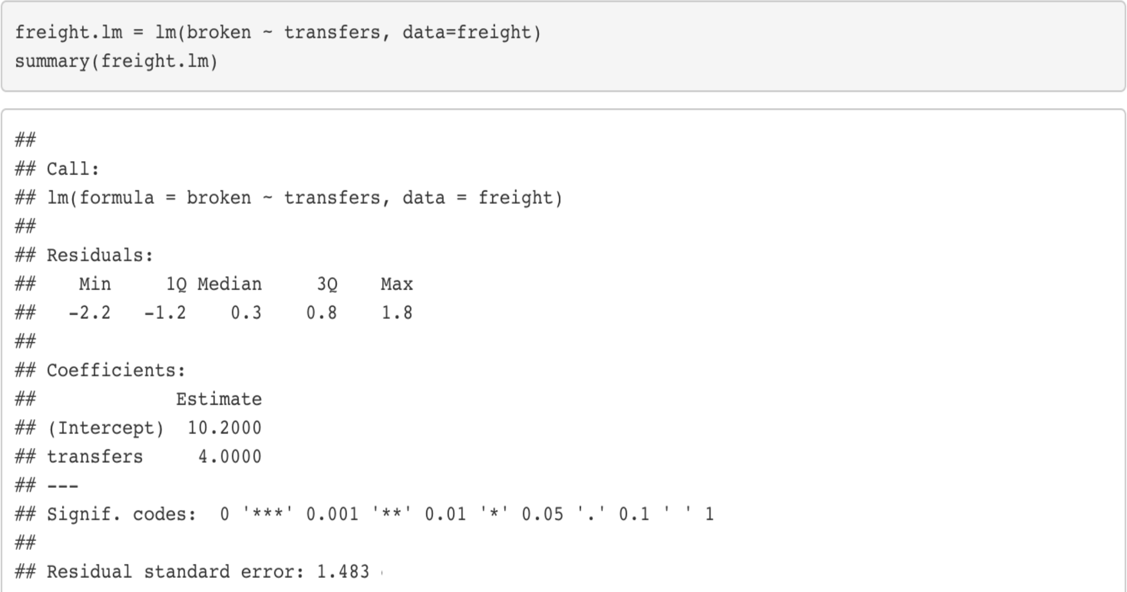
   (i) *[5 points]* Compute the boundary values of the 99% confidence band for the regression line when three transfers are made.

   (j) *[5 points]* What proportion of variation in Y is explained by the regression model?

[1] An ampule is a sealed glass capsule containing a liquid.

R Output for the Freight SLR Problem #2.

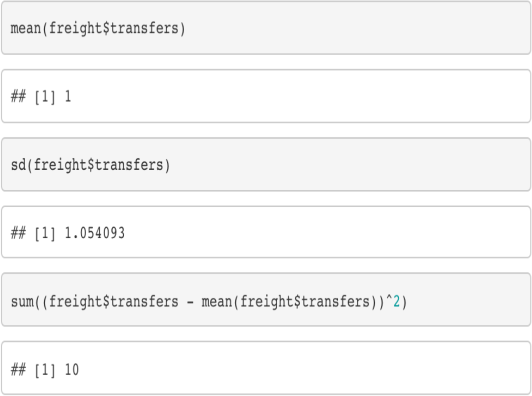
In the code below, the data frame that contains the data is called `freight`, the response Y is called `broken`, and the predictor X is called `transfers`.



```
freight.lm = lm(broken ~ transfers, data=freight)
summary(freight.lm)
```

```
##
## Call:
## lm(formula = broken ~ transfers, data = freight)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -2.2   -1.2    0.3    0.8    1.8
##
## Coefficients:
##             Estimate
## (Intercept)  10.2000
## transfers     4.0000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.483
```

Figure 1: `summary(freight)` partial output.



```
mean(freight$transfers)
```

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

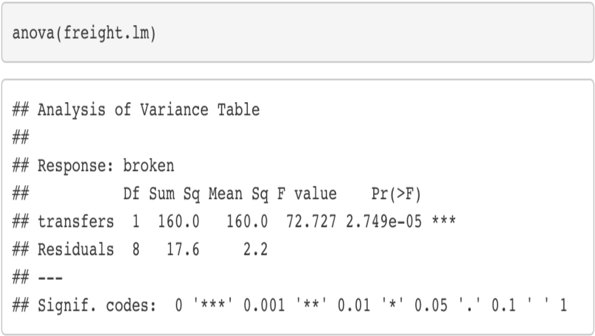
```
sd(freight$transfers)
```

```
## [1] 1.054093
```

```
sum((freight$transfers - mean(freight$transfers))^2)
```

```
## [1] 10
```

Figure 2: Summary Statistics for the `freight` data set.



```
anova(freight.lm)
```

```
## Analysis of Variance Table
##
## Response: broken
##           Df Sum Sq Mean Sq F value    Pr(>F)
## transfers  1  160.0   160.0  72.727 2.749e-05 ***
## Residuals  8   17.6     2.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 3: ANOVA Table for the fitted `freight.lm` regression model.

3. **Multiple Linear Regression: Tobacco Leaves** *[40 points]*
A researcher interested in the burning time of of tobacco leaves, collected a random sample of 30 packets of leaves and recorded the nitrogen content by percentage weight, X_1, the chlorine content by percentage weight, X_2, the potassium content by percentage weight, X_3, and the burn time in seconds, Y.

   In all the following questions (a)-(d), use significance level $\alpha = 0.05$. Based on the `R` output provided to you in pages 5, 6, and 7, answer the following questions:

   (a) *[10 points]* Based on the diagnostic plots and tests provided, what are the conclusions you can draw about the normality, constant variance and linearity assumptions? Clearly reference the plots/statistics that you use to draw conclusions.

   (b) *[10 points]* The Box-Cox method suggests transforming the response by taking a logarithm. Using the `log` transformation of the response, the researcher refitted the model. Based on the diagnostic plots and tests provided for the refitted model, what are the conclusions you can draw about the normality, constant variance and linearity assumptions? Clearly reference the plots/statistics that you use to draw conclusions.

   (c) *[10 points]* Test whether X_3 can be dropped from the regression model given that X_1 and X_2 are retained. State the null/alternative hypotheses, decision rule and conclusion.

   (d) *[10 points]* Test whether X_1 and X_2 can be dropped from the regression model given that X_3 is retained. State the null/alternative hypotheses, decision rule and conclusion.

R Output for the Tobacco Leaves MLR Problem #3.

In the code below, the data frame that contains the data is called `leafburn`, the response Y is called `burntime`, and the predictors are `nitrogen` (X_1), `chlorine` (X_2), and `potassium` (X_3).

```
leafburn.mlr = lm(burntime ~ ., data=leafburn)
par(mfrow=c(1,2))
plot(leafburn.mlr, which=1)
plot(leafburn.mlr, which=2)
```

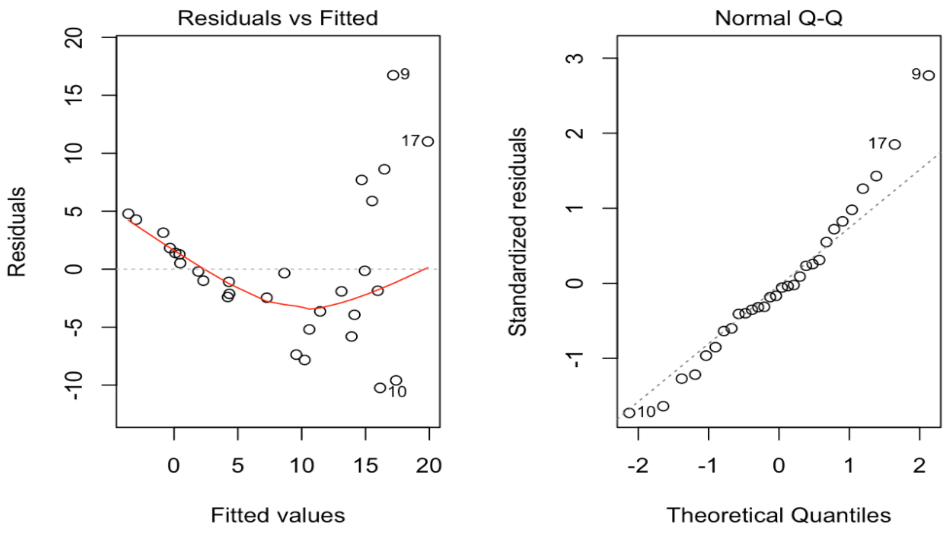
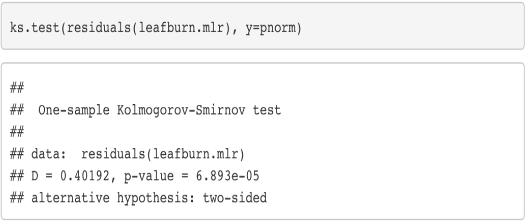


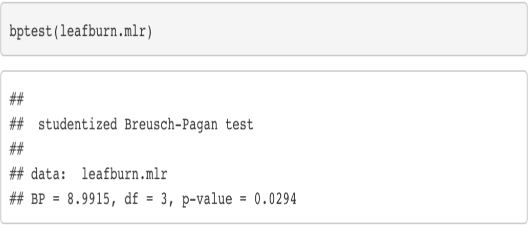
Figure 4:
(i) Fitted vs. Residuals Plot (ii) Residuals QQ Plot



```
ks.test(residuals(leafburn.mlr), y=pnorm)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  residuals(leafburn.mlr)
## D = 0.40192, p-value = 6.893e-05
## alternative hypothesis: two-sided
```

Figure 5: Kolmogorov-Smirnov Test



```
bptest(leafburn.mlr)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  leafburn.mlr
## BP = 8.9915, df = 3, p-value = 0.0294
```

Figure 6: Breusch-Pagan Test

```
leafburn.mlr.transform = lm(log(burntime) ~ ., data=leafburn)
par(mfrow=c(1,2))
plot(leafburn.mlr.transform, which=1)
plot(leafburn.mlr.transform, which=2)
```

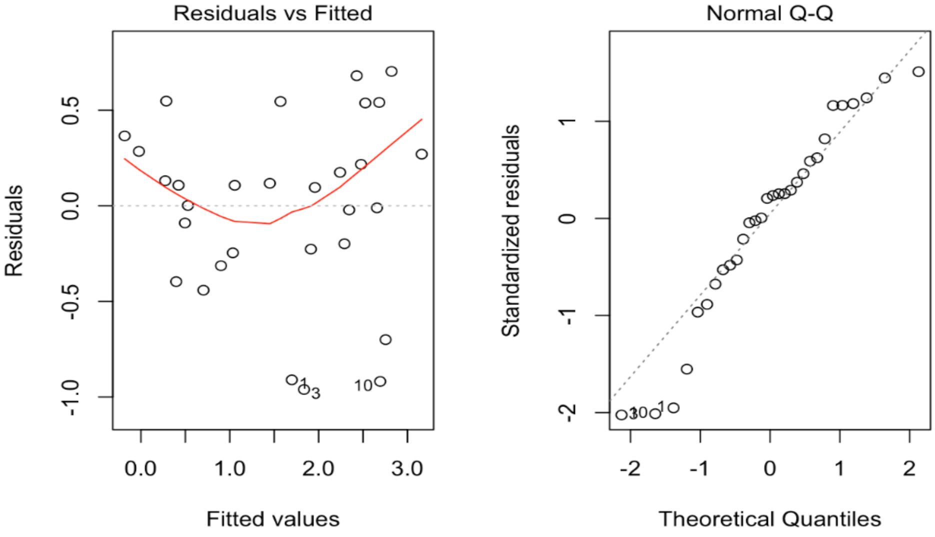
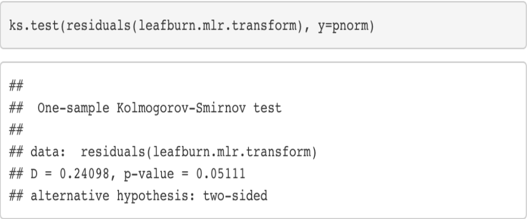


Figure 7:

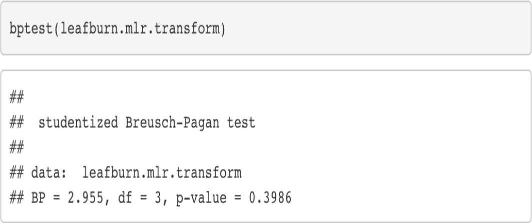
(i) Fitted vs. Residuals Plot (ii) Residuals QQ Plot



```
ks.test(residuals(leafburn.mlr.transform), y=pnorm)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  residuals(leafburn.mlr.transform)
## D = 0.24098, p-value = 0.05111
## alternative hypothesis: two-sided
```

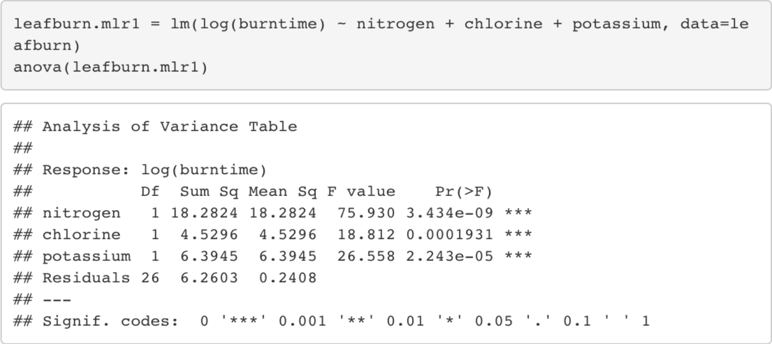
Figure 8: Kolmogorov-Smirnov Test



```
bptest(leafburn.mlr.transform)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  leafburn.mlr.transform
## BP = 2.955, df = 3, p-value = 0.3986
```

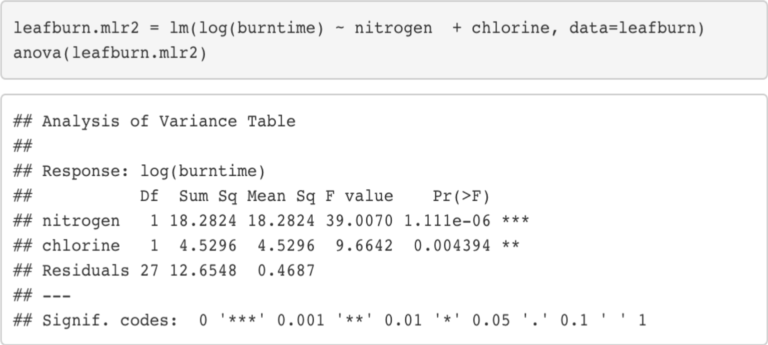
Figure 9: Breusch-Pagan Test



```
leafburn.mlr1 = lm(log(burntime) ~ nitrogen + chlorine + potassium, data=le
afburn)
anova(leafburn.mlr1)
```

```
## Analysis of Variance Table
##
## Response: log(burntime)
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## nitrogen   1 18.2824 18.2824  75.930 3.434e-09 ***
## chlorine   1  4.5296  4.5296  18.812 0.0001931 ***
## potassium  1  6.3945  6.3945  26.558 2.243e-05 ***
## Residuals 26  6.2603  0.2408
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

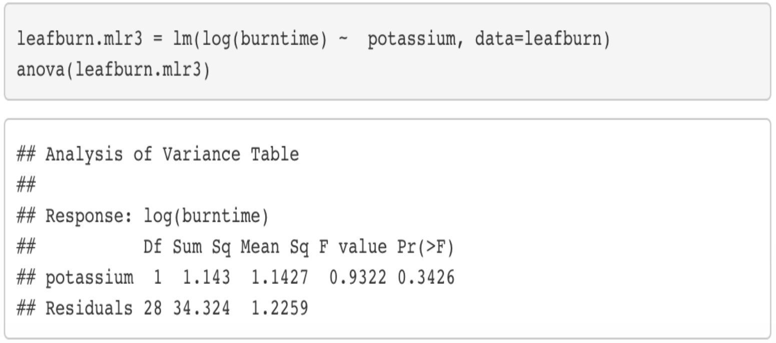
Figure 10: ANOVA Table for Model 1



```
leafburn.mlr2 = lm(log(burntime) ~ nitrogen  + chlorine, data=leafburn)
anova(leafburn.mlr2)
```

```
## Analysis of Variance Table
##
## Response: log(burntime)
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## nitrogen   1 18.2824 18.2824 39.0070 1.111e-06 ***
## chlorine   1  4.5296  4.5296  9.6642  0.004394 **
## Residuals 27 12.6548  0.4687
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 11: ANOVA Table for Model 2

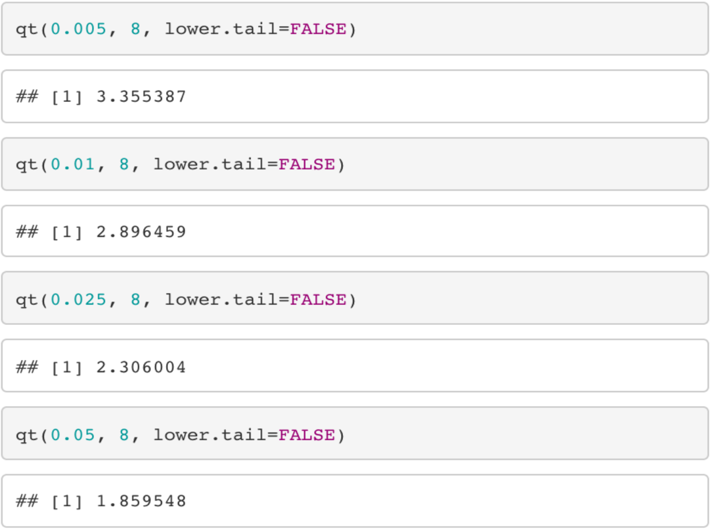


```
leafburn.mlr3 = lm(log(burntime) ~  potassium, data=leafburn)
anova(leafburn.mlr3)
```

```
## Analysis of Variance Table
##
## Response: log(burntime)
##           Df Sum Sq Mean Sq F value Pr(>F)
## potassium  1  1.143  1.1427  0.9322 0.3426
## Residuals 28 34.324  1.2259
```

Figure 12: ANOVA Table for Model 3

F and *T* distribution Critical Values.



```
qt(0.005, 8, lower.tail=FALSE)
```

```
## [1] 3.355387
```

```
qt(0.01, 8, lower.tail=FALSE)
```

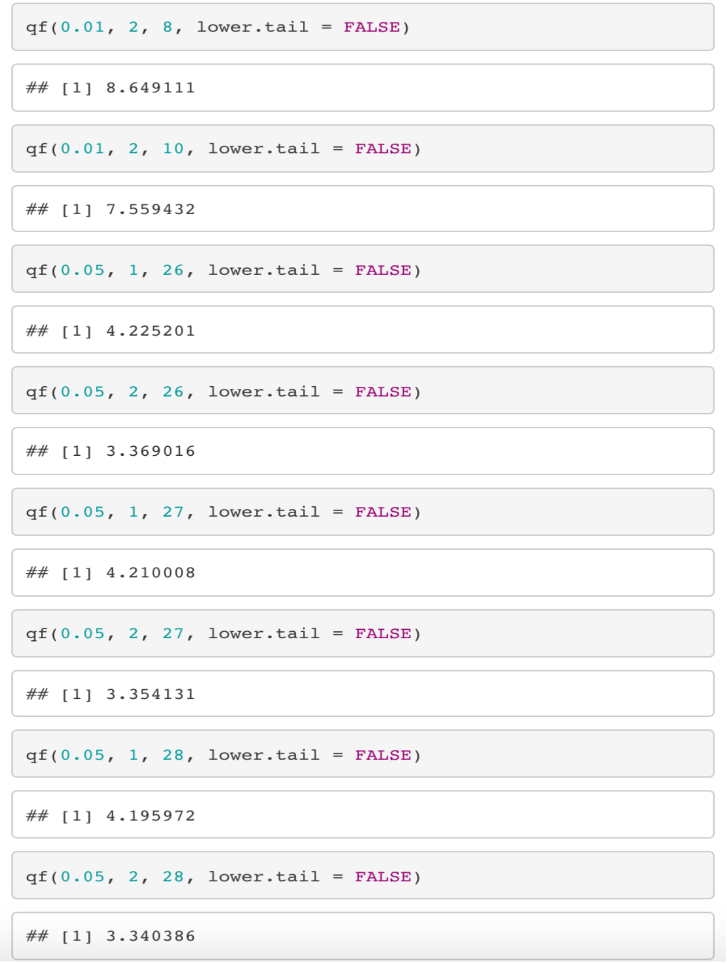
```
## [1] 2.896459
```

```
qt(0.025, 8, lower.tail=FALSE)
```

```
## [1] 2.306004
```

```
qt(0.05, 8, lower.tail=FALSE)
```

```
## [1] 1.859548
```



```
qf(0.01, 2, 8, lower.tail = FALSE)
```

```
## [1] 8.649111
```

```
qf(0.01, 2, 10, lower.tail = FALSE)
```

```
## [1] 7.559432
```

```
qf(0.05, 1, 26, lower.tail = FALSE)
```

```
## [1] 4.225201
```

```
qf(0.05, 2, 26, lower.tail = FALSE)
```

```
## [1] 3.369016
```

```
qf(0.05, 1, 27, lower.tail = FALSE)
```

```
## [1] 4.210008
```

```
qf(0.05, 2, 27, lower.tail = FALSE)
```

```
## [1] 3.354131
```

```
qf(0.05, 1, 28, lower.tail = FALSE)
```

```
## [1] 4.195972
```

```
qf(0.05, 2, 28, lower.tail = FALSE)
```

```
## [1] 3.340386
```