This is an introducton to basic hypothesis testing in R. We have shown that, with a certain set of assumptions, the difference between the OLS estimator and the true parameter vector is distributed normally as shown in expression (2.63):

$$(\mathbf{b} - \beta)|\mathbf{X} \sim N(\mathbf{0},\ \sigma^2 \cdot (\mathbf{X}'\mathbf{X})^{-1})$$

We have also shown that $s^2 = \mathbf{e}'\mathbf{e}/(n-k)$ is an unbiased estimator of σ^2 in Section 2.3.4 of the lecture notes. The purpose of the section is not to rehash the lectures, but instead to use the results to practice indexing in R.

```
data <- read.csv("../data/auto.csv", header=TRUE)
names(data) <- c("price", "mpg", "weight")
y <- matrix(data$price)
X <- cbind(1, data$mpg, data$weight)
```

For reference, consider the regression output, using data we've seen before:

```
res <- lm(price ~ 1 + mpg + weight, data = data)
summary(res)


Call:
lm(formula = price ~ 1 + mpg + weight, data = data)

Residuals:
   Min     1Q Median     3Q    Max
 -3332  -1858   -504   1256   7507

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1946.0687  3597.0496   0.541  0.59019
mpg          -49.5122    86.1560  -0.575  0.56732
weight         1.7466     0.6414   2.723  0.00813 **
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 2514 on 71 degrees of freedom
Multiple R-squared:  0.2934,       Adjusted R-squared:  0.2735
F-statistic: 14.74 on 2 and 71 DF,  p-value: 4.425e-06
```

We print the full results to examine the F-statistic. Under normal circumstances, consider just printing `coef(summary(res))`. In order to perform individual t-tests, we will first have to identify the standard errors for each coefficient, noting the distribution in (2.63). The variance of the error, σ^2, can be numerically estimated, as shown below:

```
n <- nrow(X); k <- ncol(X)
P <- X %*% solve(t(X) %*% X) %*% t(X)
e <- (diag(n) - P) %*% y
s2 <- t(e) %*% e / (n - k)
print(s2)
```

```
        [,1]
[1,] 6320340
```

The vector of standard errors matches those reported from R's canned routine `lm()`, which is encouraging.

```
vcov.mat <- as.numeric(s2) * solve(t(X) %*% X)
se <- sqrt(diag(vcov.mat))
print(se)
```

```
[1] 3597.0495988   86.1560389    0.6413538
```

We can now use the vector of standard errors to perform the individual t-tests.

```
b <- solve(t(X) %*% X) %*% t(X) %*% y
apply(b / se, 1, function(t) {2 * pt(-abs(t), df = (n - k))})
```

```
[1] 0.590188628 0.567323727 0.008129813
```

Great! We have replicated the `Pr(>|t|)` column of the canned output. Now let's try to replicate the full regression F-statistic. This is a joint test of coefficient significance; are the coefficients jointly different from a zero vector? Max has a great description as to why this is different from three separate tests of significance. For now, note that we are testing joint significance by setting

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \Sigma\!\!\!\!= \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{1}$$

This is great. This simplifies the hell out of equation (2.81), which is fairly daunting at first:

$$F = \frac{(\mathbf{Rb} - \Sigma)'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \Sigma)/J}{s^2} = \frac{\mathbf{b}'(\mathbf{X}'\mathbf{X})\mathbf{b}/J}{s^2} \tag{2}$$

```
F <- t(b) %*% (t(X) %*% X) %*% b / (s2*3)
print(F)
```

```
         [,1]
[1,] 158.1714
```

Well shit. This is much larger than the reported F-statistic of 14.74. What happened? The problem is that we also included the intercept, whereas R assumes that this shouldn't be included in the joint test. Simplification failed. Let's try again.

```
R <- rbind(c(0, 1, 0), c(0, 0, 1)); J <- 2
select.var <- solve(R %*% solve(t(X) %*% X) %*% t(R))
F <- t(R %*% b) %*% select.var %*% (R %*% b) / (s2 * J)
print(c(F, pf(F, 2, 71, lower.tail=FALSE)))
```

```
[1] 1.473982e+01 4.424878e-06
```

It worked! And the probability of observing the F-statistic with degrees of freedom $J = 2$ and $n - k = 71$ is printed as well.

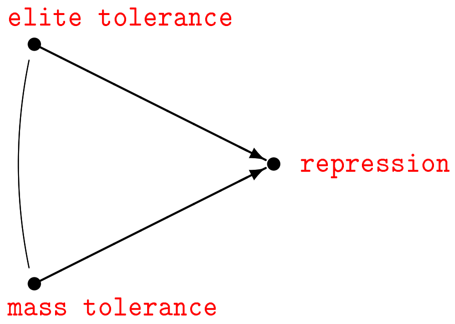

Figure 1: Path model, causes of McCarthyism, reproduced from Freedman (2009)

Puzzle (sort of)

We will replicate the results of a sort of silly study that examines the causes of McCarthyism, using a *path model*. The study was published in the *American Political Science Review* by Gibson (1988) and reprinted in Freedman (2009) to illustrate path models, which are essentially a simple graphical framework to keep track of direct and indirect causation. The path model is recreated in Figure (1), and shows that tolerance scores of both the masses and elites directly impact repression. This is a theoretical framework. The unlabeled connection between the tolerance nodes indicates association rather than causation. The repression and tolerance scores have been standardized, so that they have mean equal to zero and standard deviation equal to one.

The purpose of this exercise is to build up an intuition of the relationship between the OLS estimates and covariate correlations.
Gibson reports the correlation matrix for the path diagram:

	Masses	Elite	Repress
Masses	1.00	0.52	-0.26
Elite	0.52	1.00	-0.42
Repress	-0.26	-0.42	1.00

His model for political repression for $n = 36$ states is given by:

$$\texttt{repression} = \beta_1 \cdot \texttt{mass tolerance} + \beta_2 \cdot \texttt{elite tolerance} + \epsilon, \tag{3}$$

Denote `mass tolerance` as M, `elite tolerance` as E, and `repression` as R, such that Equation (3) becomes $R = \beta_1 M + \beta_2 E + \epsilon$. Finally, let $\mathbf{X} = [M \quad E]$, so that $R = \mathbf{X}\beta + \epsilon$.

Here is the kicker. Since, all variables were standardized, we know that

$$\frac{1}{n}\sum_{i=1}^{n} E_i = 0 \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^{n} E_i^2 = 1$$

This is true, also, for M and R. Being careful about matrix multiplication, we can compute the following:

$$\mathbf{X}'\mathbf{X} = \left[\begin{array}{cc} \sum_{i=1}^n M_i^2 & \sum_{i=1}^n M_i E_i \\ \sum_{i=1}^n M_i E_i & \sum_{i=1}^n E_i^2 \end{array}\right] = n \left[\begin{array}{cc} 1 & r_{ME} \\ r_{ME} & 1 \end{array}\right] = n \left[\begin{array}{cc} 1 & 0.52 \\ 0.52 & 1 \end{array}\right] \tag{4}$$

$$\mathbf{X}'R = \left[\begin{array}{c} \sum_{i=1}^{n} M_i R_i \\ \sum_{i=1}^{n} E_i R_i \end{array}\right] = n \left[\begin{array}{c} r_{MR} \\ r_{ER} \end{array}\right] = n \left[\begin{array}{c} -0.26 \\ -0.42 \end{array}\right] \tag{5}$$

We don't even need the actual data to compute the OLS coefficients! Specifically, $\beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'R$:

```
n <- 36
xtx <- n * matrix(c(1, 0.52, 0.52, 1), ncol = 2)
xtr <- n * matrix(c(-0.26, -0.42), ncol = 1)
(b <- solve(xtx) %*% xtr)
```

```
            [,1]
[1,] -0.05701754
[2,] -0.39035088
```

Given standardized tolerance and repression scores, we can use the following formula from page 85 in Freedman to calculate the model variance: $\hat{\sigma}^2 = 1 - \hat{\beta}_1^2 - \hat{\beta}_2^2 - 2\hat{\beta}_1\hat{\beta}_2 r_{ME}$

```
p <- 3
(sigma.hat.sq <- (n / (n - p)) * (1 - b[1]^2 - b[2]^2 - 2 * b[1] * b[2] * 0.52))
```

```
[1] 0.8958852
```

There is an implicit intercept, since the scores are standardized, so that $p = 3$. We compute the standard errors from the estimated covariance matrix, $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$. Note that $\mathbb{V}(\hat{\beta}_k|\mathbf{X}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}$:

```
vcov.mat <- sigma.hat.sq * solve(xtx)
se1 <- sqrt(vcov.mat[1,1])
se2 <- sqrt(vcov.mat[2,2])
pt(b[1]/se1, n - p)
pt(b[2]/se2, n - p)
```

```
[1] 0.3797346
[1] 0.02109715
```

The coefficient on `mass tolerance` is not significant, but the coefficient on `elite tolerance` is significant. But are the two coefficients significantly different from each other? Let $\mathbf{R} = [1 \quad -1]$ and $\sum[0]$. Then the following test statistic will test that the two coefficients are equal.

$$F = \frac{(\mathbf{R}\hat{\beta} - \sum)'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\beta} - \sum)}{\hat{\sigma}^2} \tag{6}$$

```
R <- t(matrix(c(-1, 1))); r <- 0
G <- R %*% b - r
(F <- (G %*% R %*% solve(xtx) %*% t(R) %*% t(G))/sigma.hat.sq)
```

```
           [,1]
[1,] 0.01435461
```

The test statistic follows the F-distribution, and is not significant at any reasonable p-value. So, while `elite tolerance` may be significant in the regression of repression on tolerance, it is not significantly different than the insignificant variable `mass tolerance`.