

# ARE212: Section 04

Dan Hammer

August 21, 2012

This is an introduction 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  $\sigma^2$  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

```

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,  $\sigma^2$ , 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

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{r} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad (1)$$

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

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

```
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.