SUPPLEMENT FOR CHAPTER 27: MORE REGRESSION TESTS -
F-TEST AND TEST FOR ZERO POPULATION CORRELATION

I. The F-Test for Regression

Often software will include something called an *Analysis of Variance Table* in regression output. For example, here is Minitab output for regression using the Nenana ice-breakup data (with the Analysis of Variance table highlighted in bold):

```
The regression equation is
BrkupDte = 128 - 0.0659 Year since 1900

Predictor            Coef  SE Coef      T      P
Constant          128.496    1.590  80.80  0.000
Year since 1900  -0.06587  0.02445  -2.69  0.009

S = 5.72706   R-Sq = 7.9%   R-Sq(adj) = 6.8%

Analysis of Variance
Source          DF       SS      MS     F      P
Regression       1   238.06  238.06  7.26  0.009
Residual Error  85  2787.93   32.80
Total           86  3025.99
```

Notice that:

1. The p-value listed in the Analysis of Variance table at the end of the line "Regression" is exactly the same as the p-value listed under the line "Year since 1900" in the first table of the output.
2. The number 7.26 listed under "F" in line "Regression" is (up to rounding error) the square of the number -2.69 listed under T in the first table of the output.
3. The number 7.26 listed under "F" in line "Regression" is also (up to rounding error) the quotient of the two numbers listed under "MS" in the Analysis of Variance table: 238.06/32.80 = 7.2579 …
4. The last line under SS in the Analysis of Variance table is the sum of the two numbers above it: 3025.99 = 238.06 + 2787.93
5. For each of the lines "Regression" and "Residual Error" in the Analysis of Variance table, the item in column MS equals the item in column SS divided by the item in column DF.
6. The last item in the column DF is the sum of the first two.

Here's a summary of what's going on:

First, recall the following from the handout *Supplement for Chapter 8: Why r^2 is the Fraction of Variation Accounted for by Regression*:

i) The *total sum of squares* is defined as $SS_{total} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ (p. 1)

ii) The *residual sum of squares* is defined as $SS_R = \sum_{i=1}^{n} e_i^2$ (p. 1)

iii) The *model sum of squares* (also called the *sum of squares for regression*) is defined as $SS_M = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ (p. 1)

iv) $SS_{total} = SS_R + SS_M$ (p. 2)

The column labeled SS in the Analysis of Variance table lists these three sums of squares; the line labels indicate which sum of squares is in which line. By (iv), the third sum of squares must indeed be the sum of the first two (explaining item (4) above).

To explain items (5) and (6):

- Each sum of squares listed in the SS column has an associated "degrees of freedom", listed in the DF column. The degrees of freedom for the total sum of squares is defined as n (the number of observations), the degrees of freedom for the residual sum of squares is defined as n-1, and the degrees of freedom of the regression (model) sum of squares is defined as 1. (This explains item (6).)
- The items in the MS column are defined as the quotient of the corresponding sum of squares by its degrees of freedom. They are called "mean squares:" The *mean square for regression* (also called the *model mean square*) and the *mean square error* (also called the *residual mean square*). (This explains item (5).)

Comments:

- The mean square for regression is called *MSR* for short, and the mean square error is called *MSE* for short.
- Note that MSE = $SS_R/(n-1) = s_e^2$.
- Much of this may seem like much ado about nothing. Indeed, for simple linear regression, the Analysis of Variance Table doesn't really give any new information. That is why the textbook doesn't discuss it in Chapter 27. However, the table does give something new in the context of multiple regression; see pp. 792 – 793. It also shows a connection between regression and Analysis of Variance (chapters 29 and 30), which can in fact be largely combined into one theory of Linear Models.
- However, some users of statistics do use parts of the Analysis of Variance Table in reporting their results even for simple linear regression, so it is worth being aware of in case you encounter it.

To explain item (3) above: The entry in the Regression row and F column is *defined* as F = MSR/MSE. (More explanation of F below.)

To explain item (2) above, first note that

$$MSR = SS_R = \Sigma (\hat{y}_i - \bar{y})$$

$$= \Sigma(b_0 + b_1 x_i - b_o - b_1 \bar{x})^2$$

$$= b_1^2 \, \Sigma(x_i - \bar{x})^2$$

Using a little algebra, we can use this to see that

$$\text{MSR/MSE} = [b_1^2\ \Sigma(x_i - \bar{x})^2]/\text{MSE} = b_1^2 \div \left[\frac{MSE}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

$$= b_1^2 / (SE(b_1))^2 \qquad \text{(since MSE} = s_e^2\text{)}$$

$$= t^2,$$

where $t = b_1 / SE(b_1)$ is the test statistic for testing H_0: $\beta_1 = 0$.

To explain item (1) (as well as F): We first need to introduce the *F-distribution*. It's a little more complicated than other distributions we've dealt with, because it has two associated degrees of freedom.

Definition: The *F distribution with ν_1 degrees of freedom in the numerator and ν_2 degrees of freedom in the denominator* is the distribution of a random variable of the form $\frac{U_1/\nu_1}{U_2/\nu_2}$, where

- U_1 and U_2 are independent,
- $U_1 \sim \chi^2(\nu_1)$, and
- $U_2 \sim \chi^2(\nu_2)$.

We will write $F(\nu_1, \nu_2)$ for this distribution.

Now recall that the t distribution t_k with k degrees of freedom is the distribution of a random variable which is of the form $\frac{Z}{\sqrt{U/k}}$ where

- $Z \sim N(0,1)$,
- $U \sim \chi^2(k)$, and
- Z and U are independent.

Comparing these two definitions (and remembering that the square of a standard normal random variable has a $\chi^2(1)$ distribution) shows *that the square of a t_k random variable has a F(1, k) distribution.*

In our situation, the t-statistic for the test with H_0: $\beta_1 = 0$ has a t_{n-2} distribution , so MSR/MSE (which was shown above to equal t^2) has an F distribution with 1 degree of freedom in the numerator and n-2 degrees of freedom in the denominator.

The upshot is that there is an alternate test for H_0: $\beta_1 = 0$ – an F test, using test-statistic F = MSR/MSE, and doing a (one-sided) test comparing with an F distribution.

Again, this may seem like much ado about nothing. However, in multiple regression the F-test has a different interpretation: The null hypothesis for the F-test is that *all* non-constant regression coefficients are zero, which is different from testing whether a *single* regression coefficient is zero.

II. The Test for Non-Zero Population Correlation

You might encounter another hypothesis test related to simple linear regression that is not included in the textbook: the t-test for zero population correlation.

The *population correlation*, denoted ρ, is the population analogue of the sample correlation r. So it is defined when we have two random variables, X and Y. You may have encountered the definition in probability:

$$\rho = \mathrm{Cov}(X,Y)/(\mathrm{Var}(X)\mathrm{Var}(Y))^{1/2},$$

where Var(X) and Var(Y) are the (population) variances of the random variables X and Y, and Cov(X,Y) is the *covariance*, defined by

$$\mathrm{Cov}(X,Y) = E([X-E(X)][Y-E(Y)]).$$

If you work out the formulas for a finite population, you will see that the formula for ρ looks just like the formula for the (sample correlation) r, *except* that there is an n in the denominator instead of n-1.

ρ is a *parameter*, since it refers to the entire population, not just a sample. If we have a sample, then the sample correlation r is an *estimate* of the population parameter ρ.

Recall (pp. 181, 187) that the least squares slope and the sample correlation r are related by

$$b_1 = r(s_y/s_x).$$

It can be shown in an analogous manner that (if the model assumptions are satisfied), the population slope and population correlation are related by an analogous formula

$$\beta_1 = \rho\frac{\sigma_Y}{\sigma_X},$$

where σ_X and σ_Y are the (population) variances of X and Y, respectively.

From this we see that $\beta_1 = 0$ if and only if $\rho = 0$.[1] So testing for $\rho = 0$ is the same as testing for $\beta_1 = 0$: use test statistic

$$t = \frac{b_1}{SE(b_1)},$$

which has the t(n-2) distribution under the null hypothesis H_0: $\rho = 0$ (i.e., $\beta_1 = 0$).

Using various identities and some algebra (details omitted), it is possible to show that

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

so it is possible to do the hypothesis test just knowing r.

This may be useful when you don't have access to the actual data but do know r.

---

[1]You may ask, "What if $\sigma_Y = 0$? What if $\sigma_X = 0$?" In either of these cases, ρ is undefined, so talking about ρ doesn't make sense.