ST518 - The independent samples t-test as SLR

Jason A. Osborne

N. C. State Univ.

T-test with SLR

Consider the good, old **equal variances t-test** of the hypothesis of equality of two population means, **sampled independently** from two populations with the same variance. Let Y_{ij} denote the j^{th} observation from the i^{th} sample, with n_i observations from population $i = 1, 2$. Let the population means be denoted μ_i. Then under $H_0 : \mu_1 = \mu_2$, the t-statistic below follows a t-distribution with $df = n_1 + n_2 = 2$:

$$t = \frac{\overline{Y}_{1\cdot} - \overline{Y}_{2\cdot}}{\sqrt{S_p^2(\frac{1}{n_1} + \frac{1}{n_2})}}$$

with S_p^2 denoted the weighted average of the two sample variances from the two samples, S_1^2 and S_2^2:

$$S_p^2 = \frac{n_1 - 1}{n_1 + n_2 - 2} S_1^2 + \frac{n_2 - 1}{n_1 + n_2 - 2} S_2^2$$

We will show that this t-test and the F-test from simple linear regression (SLR) are equivalent.

Before establishing equivalence, let's simulate some data and observe that $T^2 = F$

```
data one;
   do i=1 to 16;
      supp=(i>10);  *sample sizes are 10 and 6;

      y=10 + 2*supp + rannor(234)*2;
      *y normally dist'd with mean = 10 or 12, variance=4;

      output;
   end;
run;
proc ttest;
   class supp;
   var y;
run;
proc reg;
   model y=supp;
run;
```

```
The SAS System
The TTEST Procedure

supp           Method              N        Mean      Std Dev      Std Err
0                                 10      9.5930       2.0517       0.6488
1                                  6     11.2189       2.1512       0.8782
Diff (1-2)     Pooled                    -1.6260       2.0878       1.0781
Diff (1-2)     Satterthwaite             -1.6260                    1.0919

Method            Variances          DF     t Value     Pr > |t|
Pooled            Equal              14       -1.51       0.1538
Satterthwaite     Unequal        10.251       -1.49       0.1666

The REG Procedure
Model: MODEL1
                          Analysis of Variance

                                     Sum of           Mean
Source                    DF        Squares         Square    F Value    Pr > F
Model                      1        9.91396        9.91396       2.27    0.1538
Error                     14       61.02346        4.35882
Corrected Total           15       70.93742

Root MSE               2.08778    R-Square     0.1398
Dependent Mean        10.20272    Adj R-Sq     0.0783
Coeff Var             20.46296

                       Parameter       Standard
Variable       DF       Estimate          Error     t Value     Pr > |t|
Intercept       1        9.59299        0.66021       14.53       <.0001
supp            1        1.62595        1.07812        1.51       0.1538
```

In a regression formulation of this problem,

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} E_1 \\ \vdots \\ \\ \\ \vdots \\ E_n \end{pmatrix}$$

The least squares estimator of β is given by $\widehat{\beta} = (X'X)^{-1}X'Y$. Here,

$$\begin{aligned} X'X &= \begin{pmatrix} n_1 + n_2 & n_2 \\ n_2 & n_2 \end{pmatrix} \\ (X'X)^{-1} &= \frac{1}{n_1 n_2} \begin{pmatrix} n_2 & -n_2 \\ -n_2 & n_1 + n_2 \end{pmatrix} = \begin{pmatrix} 1/n_1 & -1/n_1 \\ -1/n_1 & (n_1+n_2)/(n_1 n_2) \end{pmatrix} \\ (X'Y)' &= \begin{pmatrix} \sum y_i, & n_2 \overline{y}_2 \end{pmatrix} \end{aligned}$$

And the product, $\widehat{\beta}$ is

$$(X'X)^{-1}X'Y = \begin{pmatrix} n_1+n_2 & n_2 \\ n_2 & n_2 \end{pmatrix} \begin{pmatrix} \sum y_i \\ n_2\overline{y}_2 \end{pmatrix} = \begin{pmatrix} \frac{\sum y_i}{n_1} - \frac{n_2\overline{y}_2}{n_1} \\ \frac{-\sum y_i}{n_1} + \frac{n_1+n_2}{n_1}\overline{y}_2 \end{pmatrix}$$

Note that $\overline{y} = \frac{n_1\overline{y}_1}{n_1+n+2} + \frac{n_2\overline{y}_2}{n_1+n_2}$ is a weighted average of $\overline{y}_1$ and $\overline{y}_2$ and $(n_1+n_2)\overline{y} = \sum y_i = n_1\overline{y}_1 + n_2\overline{y}_2$. The intecept and slope can then be simplified

$$\begin{aligned} \widehat{\beta}_0 &= \frac{n_1\overline{y}_1+n_2\overline{y}_2}{n_1} - \frac{n_2}{n_1}\overline{y}_2 &= \overline{y}_1 \\ \widehat{\beta}_1 &= \frac{-n_1\overline{y}_1-n_2\overline{y}_2}{n_1} + \overline{y}_2 + \frac{n_2}{n_1}\overline{y}_2 &= \overline{y}_2 - \overline{y}_1 \end{aligned}$$

To show that $MS(E) = S_p^2$ is a little more involved, but not too bad.

$$
\begin{aligned}
SSE &= \sum (Y_i - \widehat{Y}_i)^2 \\
&= (Y - \widehat{Y})'(Y - \widehat{Y}) \\
&= Y'(I - H)Y \\
H &= X(X'X)^{-1}X' \\
&= \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1/n_1 & -1/n_1 \\ -1/n_1 & \frac{n_1+n_2}{n_1 n_2} \end{pmatrix} \begin{pmatrix} 1 & \dots & 1 \\ 0 & \dots & 1 \end{pmatrix} \\
&= \dots \\
&= \left(\begin{array}{c|c} 1/n_1 * I_{n_1} & \mathbf{0} \\ \mathbf{0} & 1/n_2 * I_{n_2} \end{array} \right),
\end{aligned}
$$

an $(n_1 + n_2) \times (n_1 + n_2)$ matrix, with submatrices which are multiples of identity matrices I with n_1 and n_2 rows, respectively, and $\mathbf{0}$ denotes a submatrix of zeros.

$$
\begin{aligned}
HY &= \begin{pmatrix} \overline{Y}_1 \\ \vdots \\ \overline{Y}_1 \\ \overline{Y}_2 \\ \vdots \\ \overline{Y}_2 \end{pmatrix} = \widehat{Y} \\
e &= Y - \widehat{Y} \\
e'e &= SS(E) = \sum (Y_{1i} - \overline{Y}_1)^2 + \sum (Y_{2i} - \overline{Y}_2)^2 \\
&= (n_1 - 1) * S_1^2 + (n_2 - 1) S_2^2 \\
&= (n_1 + n_2 - 2) S_p^2 \\
MS(E) &= SS(E)/(n-2) \\
&= S_p^2
\end{aligned}
$$

To review, we have shown that $\widehat{\beta}_1 = \overline{Y}_2 - \overline{Y}_1$ and $MS(E) = S_p^2$.
It remains to show that $T^2 = MS(R)/MS(E)$.
Recall that $MS(R) = \sum(\widehat{Y}_i - \overline{Y})^2$, quantifying variation between least squares regression line, $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$ and horizontal line, $\overline{Y}$. Note that

$$
\begin{aligned}
MS(R) &= \sum(\widehat{Y}_i - \overline{Y})^2 \\
&= \sum\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i - (\widehat{\beta}_0 + \widehat{\beta}_1 \overline{x})\right)^2 \\
&= \widehat{\beta}_1^2 \sum(x_i - \overline{x})^2 \\
F &= MS(R)/MS(E) = \widehat{\beta}_1^2 \sum(x_i - \overline{x})^2 / MS(E) \\
&= \widehat{\beta}_1^2 / (MSE / \sum(x_i - \overline{x})^2)
\end{aligned}
$$

After some algebra, the reciprocal of
$\sum(x_i - \overline{x})^2) = (n_1 n_2^2 + n_1^2 n_2)/(n_1 + n_2)^2$ can be shown to be $1/n_1 + 1/n_2$
and

$$
F = MS(R)/MS(E) = \left(\frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{MS(E)(1/n_1 + 1/n_2)}}\right)^2
$$