Using F tests instead of t-tests

We can also test the hypothesis $H_0{:}\beta_1 = 0$ versus $H_1{:}\beta_1 \neq 0$ with an F test.

$$F = \frac{\text{MSRegression}}{\text{MSError}}$$

This test is mathematically identical to the previous test of $H_0{:}\beta_1=0$ done with the t-test (see demonstration in text, which only demonstrates that $F=t^2$ for $H_0{:}\beta_1=0$). The probabilities are identical.

More generally, the first column in the F tables ($F_{1,den\ df}$) is equivalent to the square of t.

ie.

$$t^2_{\gamma\ df\ ;\ \frac{\alpha}{2}} = F_{1,\gamma\ df\ ;\alpha}$$

Text mentions that t-test has the advantage that it can test one-tailed hypotheses, while the F cannot.

Also, the t readily tests hypotheses other than $\beta_1 = 0$. This can also be done with a non-central F test, but this is more difficult.

SAS has a test statement in PROC REG which produces an F value for testing values other than 0, but you should know that:

1) the t-test you would do is the same, and

2) The SAS test is a two tailed test. The t-test can be either one or two-tailed

3) The P value (P>F) given by SAS for the F value from the “TEST" statement is exactly the same as it would be for the t-test.

Expected Mean Squares for Regression

Recall from ANOVA (simple CRD, balanced) that

$\mathrm{E(MSE)} = \sigma^2$

$\mathrm{E(MSTreatments)} = \sigma^2 + \mathrm{n}\sigma_\tau^2$

where σ^2 is the residual variance, σ_τ^2 is the treatment variance and n is the number of replicates in each treatment

and the quantity that we wanted to test was σ_τ^2

The test used was $\mathrm{F} = \frac{\sigma^2 + \mathrm{n}\sigma_\tau^2}{\sigma^2}$

We can see that

1) F will be 1 if $\sigma_\tau^2 = 0$. This would be the null hypothesis

2) Power (the ability to detect a difference which exists) increases as we increase n (sample size) or σ_τ^2 (the treatment differences) or as we reduce σ^2 (the random error term).

Likewise for regression,

$\mathrm{E(MSE)} = \sigma^2$ (this is deviations from regression)

$\mathrm{E(MSRegression)} = \sigma^2 + \beta_1^2\Sigma(\mathrm{X}_i - \bar{\mathrm{X}})^2$

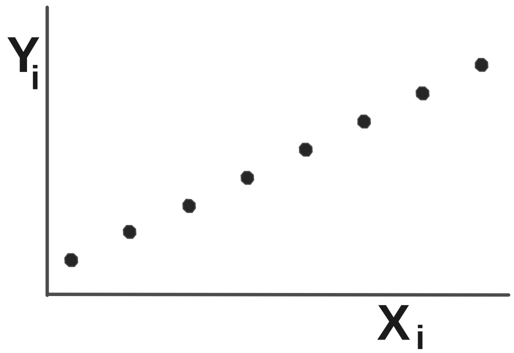
The test used was $\mathrm{F} = \frac{\sigma^2 + \beta_1^2\Sigma(\mathrm{X}_i - \bar{\mathrm{X}})^2}{\sigma^2}$

1) F will be 1 if $\beta_1 = 0$. This would be the null hypothesis. Also, since β_i is squared, this will be a two tailed test. For one tailed tests use the t-test.

2) Power (the ability to detect a difference which exists) increases as we increase β_i (regression coefficient) or $\Sigma(\mathrm{X}_i - \bar{\mathrm{X}})^2$ (the corrected SS of X_i) or as we reduce σ^2 (the random error term).

Note that power increases as $\Sigma(\mathrm{X}_i - \bar{\mathrm{X}})^2$ increases, this occurs as we

a) increase the distance from X_i to $\bar{\mathrm{X}}$. (Were is best place to put X_i?)



but only if we know that it is a straight line

b) increase n, since more squared differences are added $\Sigma(\mathrm{X}_i - \bar{\mathrm{X}})^2$

Also note that the term $\beta_1^2\Sigma(\mathrm{X}_i - \bar{\mathrm{X}})^2$ will be positive since β_i is squared and the SSX_i will be positive. Therefore, this test is one tailed.

EXAMPLE: Using SAS to test hypotheses about β_0 and β_1

EXST7034 - EXAMPLE 1

Program Statements

```
*************************************************************;
*** EXST7034 Example 1 using PC-SAS                        ***;
*** Problem from Neter, Wasserman & Kuttner 1989, 22.19   ***;
*************************************************************;
OPTIONS LS=80 PS=61 NOCENTER NODATE NONUMBER;
DATA ONE; INFILE CARDS MISSOVER;
          TITLE1 'EXST7034 - EXAMPLE 1';
   INPUT X Y;
CARDS;
raw data here
;
PROC SORT; BY X Y;
PROC PRINT;  TITLE2 'Raw Data Listing';
PROC REG;    TITLE2 'Regression Models done with SAS REG
procedure';
   MODEL  Y = X / XPX I P CLM;   TEST X = 5; RUN;
```

```
Model: MODEL1
Model Crossproducts X'X X'Y Y'Y
X'X                 INTERCEP                 X                 Y
INTERCEP                  10                10               142
X                         10                20               182
Y                        142               182              2194

X'X Inverse, Parameter Estimates, and SSE
                    INTERCEP                 X                 Y
INTERCEP                 0.2              -0.1              10.2
X                       -0.1               0.1                 4
Y                       10.2                 4              17.6

EXST7034 - EXAMPLE 1
Regression Models done with SAS REG procedure
Dependent Variable: Y
Analysis of Variance
                        Sum of          Mean
Source      DF         Squares        Square       F Value       Prob>F
Model        1       160.00000     160.00000        72.727       0.0001
Error        8        17.60000       2.20000
C Total      9       177.60000

    Root MSE         1.48324     R-square         0.9009
    Dep Mean        14.20000     Adj R-sq         0.8885
    C.V.            10.44535
```

```
Parameter Estimates
                     Parameter      Standard    T for H0:
Variable   DF     Estimate         Error   Parameter=0    Prob > |T|
INTERCEP    1   10.200000    0.66332496        15.377        0.0001
X           1    4.000000    0.46904158         8.528        0.0001
```

Note: $8.528^2 = 72.72678$

Output from the PROC REG "TEST" option for "TEST X = 5;"

```
Dependent Variable: Y
Numerator:      10.0000  DF:    1   F value:   4.5455
Denominator:        2.2  DF:    8   Prob>F:   0.0656
```

Notes:

1) t test of parameter estimate (= 8.528) is equal to the square root of the F test of the model. $F = 72.727$; $\sqrt{F} = \sqrt{72.727} = 8.529$. These are the same test.

2) The value for the standard error of b_1 is

$$\mathrm{Var}(b_1) = \frac{n\hat{\sigma}^2}{n\left[\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}\right]} = \frac{MSE}{\Sigma(X_i - \bar{X})^2} = \frac{2.2}{20 - \frac{10^2}{10}} = 0.22 = s_{b_1}^2$$

$$s_{b_1} = \sqrt{0.22} = 0.46904158$$

Which is also equal to the square root of MSE*c_{ii} from the $(X'X)^{-1}$ matrix, where MSE = 2.2 and $c_{11} = 0.1$.

3) The value for the standard error of b_0 is

$$\mathrm{Var}(b_0) = \frac{\Sigma X_i^2 \sigma^2}{n\left[\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}\right]} = \frac{\Sigma X_i^2 MSE}{n\Sigma(X_i - \bar{X})^2} = \frac{20*2.2}{10*10} = 0.44$$

$$s_{b_0} = \sqrt{0.44} = 0.66332496$$

4) The TEST option was used to test the hypothesis that H_0: β_1=5. The alternative would be the two tailed alternative that H_1:$\beta_1 \neq 5$.

The option produced the results: F = 4.5455, P(>F) = 0.0656
Which should be the square of t, or $t = \sqrt{F} = 2.132$.

$$t = \frac{(b_1 - \beta_{1_0})}{s_{b_1}} = \frac{b_1 - 5}{s_{b_1}} = \frac{4.0 - 5}{0.46904158} = \frac{1}{0.46904158} = 2.132$$

EXST7034 - EXAMPLE 1 : Vial breakage regressed on number of airline transfers.

Example of confidence limits for the regression line at various values of X_i. A missing value was included with an X value of 4.
Regression Models done with SAS REG procedure

Obs	Dep Var Y	Predict Value	Std Err Predict	Lower95% Mean	Upper95% Mean	Residual
1	8.0000	10.2000	0.663	8.6704	11.7296	-2.2000
2	9.0000	10.2000	0.663	8.6704	11.7296	-1.2000
3	11.0000	10.2000	0.663	8.6704	11.7296	0.8000
4	12.0000	10.2000	0.663	8.6704	11.7296	1.8000
5	13.0000	14.2000	0.469	13.1184	15.2816	-1.2000
6	15.0000	14.2000	0.469	13.1184	15.2816	0.8000
7	16.0000	14.2000	0.469	13.1184	15.2816	1.8000
8	17.0000	18.2000	0.663	16.6704	19.7296	-1.2000
9	19.0000	18.2000	0.663	16.6704	19.7296	0.8000
10	22.0000	22.2000	1.049	19.7814	24.6186	-0.2000
11	.	26.2000	1.483	22.7796	29.6204	.

```
Sum of Residuals                -1.59872E-14
Sum of Squared Residuals             17.6000
Predicted Resid SS (Press)           25.8529
```

Example of confidence limits for a new point at various values of X_i. A missing value was included with an X value of 4.

Regression Models done with SAS REG procedure

Obs	Dep Var Y	Predict Value	Std Err Predict	Lower95% Predict	Upper95% Predict	Residual
1	8.0000	10.2000	0.663	6.4532	13.9468	-2.2000
2	9.0000	10.2000	0.663	6.4532	13.9468	-1.2000
3	11.0000	10.2000	0.663	6.4532	13.9468	0.8000
4	12.0000	10.2000	0.663	6.4532	13.9468	1.8000
5	13.0000	14.2000	0.469	10.6127	17.7873	-1.2000
6	15.0000	14.2000	0.469	10.6127	17.7873	0.8000
7	16.0000	14.2000	0.469	10.6127	17.7873	1.8000
8	17.0000	18.2000	0.663	14.4532	21.9468	-1.2000
9	19.0000	18.2000	0.663	14.4532	21.9468	0.8000
10	22.0000	22.2000	1.049	18.0109	26.3891	-0.2000
11	.	26.2000	1.483	21.3628	31.0372	.

```
Sum of Residuals                -1.59872E-14
Sum of Squared Residuals             17.6000
Predicted Resid SS (Press)           25.8529
```

Summary of the results due to the assumptions made

(a) $S^2 = MSE$ then $E(S^2) = \sigma^2$

(b) Distributions

(1) b_i is distributed $N[\beta, \sigma^2(X'X)^{-1}]$

We do not assume $Cov(\beta_i,\beta_j) = 0$ as with the Y's. More later.

(2) $\frac{MSReg}{MSE}$ is distributed $F_{(df\ MSReg,\ df\ MSE)}$

For multiple regression this is a joint test, so the distribution has a noncentrality parameter which is zero when $\beta_1, \beta_2, \ldots \beta_k$ equals zero. (When H_o is true)

(3) In particular

$$\frac{b_i - \beta_i}{\sqrt{S^2 c_{ii}}} \text{ is distributed } t_{(df\ Error)}$$

where the c_{ii} is the Gaussian multiplier from $(X'X)^{-1}$

(c) What if the distribution of Y_i is not normal?

1) If the departure is small, the distribution is still reasonably symmetric, then the regression coefficients will be approximately normal and the effect on confidence intervals and tests of hypothesis will be small.

2) Even if the departure from normality is great, the regression coefficients have a property called asymptotic normality, such that under most conditions the the distribution approaches normality as the sample size increases.

Later we will also discuss transformations which will “normalize" the data, aiding in meeting this assumption.

Variance of $E(Y_i)$ for the simple linear model

$$\hat{Y}_i = b_0 + b_1X_i$$

Sampling Distribution of $\hat{Y}_i$

as with the variances of β_i's, $\hat{Y}_i$ is a linear combination of the Y_i and is normal

$$E(\hat{Y}_i) = E(Y_i)$$

$$\mathrm{Var}(\hat{Y}_i) = \sigma^2 \left(\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\Sigma(X_i - X)^2} \right)$$

In practice σ^2 would be estimated by MSE.

Note that the variance for $\hat{Y}_i$ is very similar to the variance of b_0. This is because b_0 is a special case of $\hat{Y}_i$ where X_i=0.

Also note that the value of the numerator of the second term will increase as the distance between $\bar{X}$ and X_i increases. This is because the regression line is most stable at $\bar{X}$, and uncertainty increases as we get farther from $\bar{X}$.

Sampling Distribution of $\frac{\hat{Y}_i - E(Y_i)}{s_{\hat{Y}_i}}$

as with the other normally distributed statistics examined, this will follow students t distribution with n-2 degrees of freedom.

The t distribution can be used either for testing an hypothesis about $\hat{Y}_i$ or for placing a confidence interval on $\hat{Y}_i$.

Example : From *vial breakage regressed on number of airline transfers* example

Place a confidence interval on the regression line for the amount of breakage for 3 transfers.

$$s^2_{\hat{Y}_i} = \text{MSE}\left(\frac{1}{n} + \frac{(X_i-\bar{X})^2}{\Sigma(X_i-\bar{X})^2}\right) = 2.2\left(\frac{1}{10} + \frac{(3-1)^2}{20-\frac{10^2}{10}}\right) =$$

$$2.2\left(\frac{1}{10} + \frac{4}{10}\right) = 2.2*\frac{5}{10} = 1.1$$

$$s_{\hat{Y}_i} = \sqrt{1.1} = 1.0488$$

since $t_{\frac{\alpha}{2},\, 8\, df} = 2.306$, then

$$P(\hat{Y}_{X=3} - t_{1-\frac{\alpha}{2},n-2}\, s_{\hat{Y}_i} \le E(\hat{Y}) \le \hat{Y}_{X=3} + t_{1-\frac{\alpha}{2},n-2}\, s_{\hat{Y}_i}) = 1\text{-}\alpha$$

$$P(22.2 - 2.306*1.0488 \le E(\hat{Y}) \le 22.2 + 2.306*1.0488) = 1\text{-}\alpha$$

$$P(19.781 \le E(\hat{Y}) \le 24.619) = 0.95$$

Check this against the SAS output