Regression and the 2-Sample t

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

Regression and the 2-Sample t

Introduction

- In this section, we assume that we have actually successfully conducted a completely randomized, independent sample experiment resulting in two groups of scores.
- The null hypothesis is that, in the populations these two samples represent, the populations means are equal.
- The common statistical assumption is that the populations are also normally distributed and have equal variance.
- The null hypothesis is usually expressed as

$$H_0 : \quad \mu_1 = \mu_2 \tag{1}$$

- More generally, you might write it as

$$H_0 : \quad \mu_1 - \mu_2 = \kappa_0 \tag{2}$$

where κ_0 is usually zero, but may also represent some target value for the mean difference.

Introduction

- In this module, we review two classic approaches to testing this hypothesis.
  1. The 2-sample,independent sample t-test. This is the method you probably saw as an undergraduate.
  2. Fitting a regression model and performing an analysis of variance. You may have seen this method, but may have been taught that it is a special case of a statistical method called "Analysis of Variance," without being told that the analysis of variance is actually linear regression.
- We begin by reviewing the classic t-test, and then move on to discuss the regression approach.

The 2-Sample Student t Test

- The t-test is calculated as

$$t_{n_1+n_2-2} = \frac{\overline{Y}_{\bullet 1} - \overline{Y}_{\bullet 2} - \kappa_0}{\sqrt{w\hat{\sigma}^2}} \tag{3}$$

where κ_0 is the null-hypothesized value of $\mu_1 - \mu_2$, almost always zero and hence often omitted,

$$w = \frac{1}{n_1} + \frac{1}{n_2} = \frac{n_1 + n_2}{n_1 n_2} \tag{4}$$

and

$$\hat{\sigma}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{5}$$

- If the null hypothesis is true, under the assupmptions of the test, the statistic has a Student t distribution with $n_1 + n_2 - 2$ degrees of freedom.
- Let's look at a quick example, first using hand computation (with the assistance of R), then using a slightly more general approach commonly used in R.

The 2-Sample Student t Test

- Suppose we have a simple data set in which we have only 5 observations in Group 1 and 6 in Group 2.
- Here are the data.

```
> Group.1 <- c(102,131,119,109,111)
> Group.2 <- c(104,98,110,119,99,88)
```

- On the next slide we'll process the formula from Equation 3.

The 2-Sample Student t Test

```
> numerator <- mean(Group.1) - mean(Group.2)
> n.1 <- length(Group.1)
> n.2 <- length(Group.2)
> w <- (1/n.1 + 1/n.2)
> df <- n.1 + n.2 -2
> sigma.hat.squared <- ((n.1-1) * var(Group.1) +
+   (n.2-1)*var(Group.2) )/df
> t <- numerator/sqrt(w*sigma.hat.squared)
> t

[1] 1.73214
```

The value of 1.73 is well below the critical value required for rejection.

The 2-Sample Student t Test

A quicker way is to use a built-in function in R.

```
> t.test(Group.1, Group.2, var.equal=TRUE, paired=FALSE)

        Two Sample t-test

data:  Group.1 and Group.2
t = 1.7321, df = 9, p-value = 0.1173
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.488287 26.288287
sample estimates:
mean of x mean of y
    114.4     103.0
```

Reassuringly, we get the same result. Notice that this procedure also automatically returns a 95% confidence interval on the quantity $\mu_1 - \mu_2$, calculated as

$$\overline{Y}_{\bullet 1} - \overline{Y}_{\bullet 2} \pm t^*_{.975, n_1+n_2-2}\sqrt{w\hat{\sigma}^2}$$

A Regression Modeling Approach

Nested Models

- In the previous section, we showed how to compare two means with the classic t-statistic.
- Another way of thinking about the same statistical test is that comparing two means involves comparing two nested regression models.
  1. One model reproduces the data for both groups from a single mean.
  2. The second, more complex model, says that the two groups may have unequal means.
  3. Via regression analysis, we compare these two nested models.

A Regression Modeling Approach

Nested Models

But what exactly are nested models?

A Regression Modeling Approach

When are Models Nested?

- One model is nested within another if it can be expressed as a special case of the other in which the parameters of the second are constrained versions of the parameters of the first.
- For example, consider two regression models, $E(Y|X = x) = \beta_1 x + \beta_0$, and $E(Y|X = x) = \beta_0$.
- The second model is nested within the first because it is a special case of the first model in which β_1 is constrained to be zero.
- If a model stays the same except for some parameters that are dropped, the simpler model is generally nested within the more complex version.

A Regression Modeling Approach

When are Models Nested?

- It turns out, there is a simple, general procedure for comparing any two nested regression models with each other.
- As we begin this discussion, I want to draw your attention to something.
- When two models are nested, the more complex model always fits the data in the sample (or the variables in the population) at least as well as the less complex model, in the sense that the sum of squared errors will be at least as small and almost always smaller.

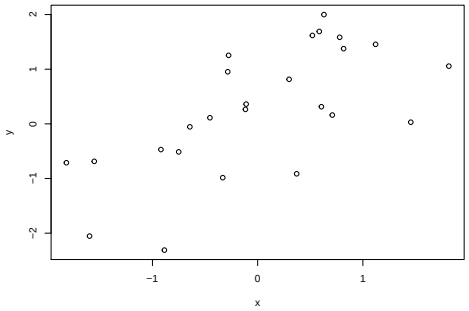
Why?

A Regression Modeling Approach

A Model Comparison F-Test for Nested Models

- Let's define and plot some artificial data on two variables.

```
> set.seed(12345)
> x <- rnorm(25)
> e <- rnorm(25,0,sqrt(1/2))
> y <- sqrt(1/2)*x + e
> plot(x,y)
```



A Regression Modeling Approach

A Model Comparison F-Test for Nested Models

- We want to predict y from x using least squares linear regression.
- We seek to fit a model of the form

$$y_i = \beta_0 + \beta_1 x_i + e_i = \hat{y}_i + e_i$$

  while minimizing the sum of squared errors in the "up-down" plot direction.
- As usual, we fit such a model in R by creating a "fit object" and examining its contents.
- We see that the formula for $\hat{y}_i$ is a straight line with slope β_1 and intercept β_0.

A Regression Modeling Approach

A Model Comparison F-Test for Nested Models

- We create a representation of the model with a model specification formula.
- As we noted before, the formula corresponds to the model stated on the previous slide in a specific way:
  1. Instead of an equal sign, a "$\sim$" is used.
  2. The coefficients themselves are not listed, only the predictor variables.
  3. The error term is not listed
  4. The intercept term generally does not need to be listed, but can be listed with a "1".
- So the model on the previous page is translated as `y ~ x`.

A Regression Modeling Approach

A Model Comparison F-Test for Nested Models

- We create the fit object as follows.

```
> fit.A <- lm(y ~ x)
```

- Once we have created the fit object, we can examine its contents.

A Regression Modeling Approach

A Model Comparison F-Test for Nested Models

```
> summary(fit.A)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8459 -0.6692  0.2133  0.5082  1.2330

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2549     0.1754   1.453 0.159709
x             0.8111     0.1894   4.282 0.000279 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8771 on 23 degrees of freedom
Multiple R-squared: 0.4435,        Adjusted R-squared: 0.4193
F-statistic: 18.33 on 1 and 23 DF,  p-value: 0.0002791
```

A Regression Modeling Approach

A Model Comparison F-Test for Nested Models

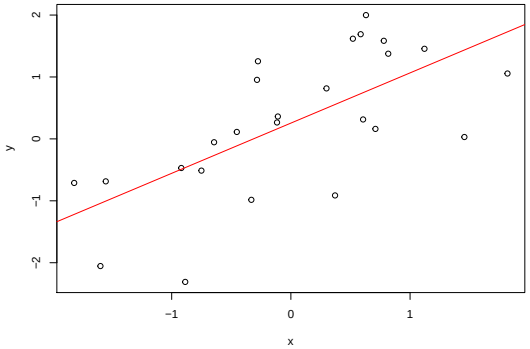
- As before, we see the printed coefficients for the intercept and for x.
- There are statistical t tests for each coefficient. These are tests of the null hypothesis that the coefficient is zero.
- There is also a test of the hypothesis that the squared multiple correlation (the square of the correlation between $\hat{y}$ and y) is zero.
- Standard errors are also printed, so you can compute confidence intervals. (How would you do that quickly "in your head?" (C.P.)
- The intercept is not significantly different from zero. Does that surprise you? (C.P.)
- The squared correlation is .4435. What is the squared correlation in the population? (C.P.)

A Regression Modeling Approach

A Model Comparison F-Test for Nested Models

- Let's add a red best-fitting straight line.

```
> plot(x,y)
> abline(fit.A,col='red')
```



A Regression Modeling Approach

A Model Comparison F-Test for Nested Models

- If we have more than one predictor, we have a multiple regression model.
- Suppose, for example, we add another predictor w to our artificial data set.
- We design this predictor to be completely uncorrelated with the other predictor and the criterion, so this predictor is, in the population, of no value.
- We do this by making this new predictor, w, a set of random numbers.

```
> w <- rnorm(25)
```

- Now our model becomes

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 w_i + e_i$$

A Regression Modeling Approach

A Model Comparison F-Test for Nested Models

- How do we compare this new model, with an added predictor variable w, with the old model?
- Notice that the old model, with predictor variable x, is nested within the new model, which has predictor variables x and w.
- In such a situation, we can perform a generalized hierarchical F test.
- To begin with, we can ask a simple question.
- How would we specify and fit the model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 w_i + e_i$$

in R?

A Regression Modeling Approach

A Model Comparison F-Test for Nested Models

That's right,

```
> fit.B <- lm(y ~ x + w)
> summary(fit.B)

Call:
lm(formula = y ~ x + w)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8475 -0.6693  0.2198  0.5108  1.2298

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.254043   0.181833   1.397 0.176312
x           0.812727   0.202128   4.021 0.000573 ***
w           0.004366   0.152239   0.029 0.977380
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8968 on 22 degrees of freedom
Multiple R-squared: 0.4435,        Adjusted R-squared: 0.393
F-statistic: 8.768 on 2 and 22 DF,  p-value: 0.001584
```

A Regression Modeling Approach

A Model Comparison F-Test for Nested Models

- We have fit this newer, more complex model and we can of course examine its coefficients and the associated statistical tests.
- We saw earlier that this new model *must* have a smaller sum of squared errors than Model A in the sample.
- But can we reject the null hypothesis that Model B is actually no better than Model A in the population?
- On the next slide, we describe a general procedure that has far-reaching implications.

A Regression Modeling Approach

Partial F-Tests: A General Approach

- Suppose Model B includes Model A as a special case. That is, Model A is a special case of Model B where some terms have coefficients of zero. Then Model A is nested within Model B.
- We define $\texttt{RSS}_b$ to be the sum of squared residuals for Model B, $\texttt{RSS}_a$ the sum of squared residuals for Model A.
- Since Model A is a special case of Model B, model B is more complex so $\texttt{RSS}_a$ will always be as least as large as $\texttt{RSS}_b$.
- We define df_b to be $n - p_b$, where p_b is the number of regressors in Model B *including the intercept*, and correspondingly $df_a = n - p_a$.
- Then, to compare Model B against Model A, we compute the partial $F-$statistic as follows.

$$F_{df_a - df_b, df_b} = \frac{MS_{comparison}}{MS_{res}} = \frac{(\texttt{RSS}_a - \texttt{RSS}_b)/(p_b - p_a)}{\texttt{RSS}_b/df_b} \tag{6}$$

A Regression Modeling Approach

Partial F-Tests: A General Approach

- That seems like a lot, but it is really rather simple.
- The F is the ratio of two mean squares.
- The numerator is the the difference between the sum of squares divided by the difference in degrees of freedom (or, alternatively, the difference in the number of regressor terms).
- The denominator is the sum of squares for the more complex (better fitting) model divided by its degrees of freedom.
- The calculations are not quite as hard as they seem at first glance.
- In any case, we can relax! R is going to do this for you automatically with its `anova` command.

A Regression Modeling Approach

Partial F-Tests: A General Approach

```
> anova(fit.A,fit.B)

Analysis of Variance Table

Model 1: y ~ x
Model 2: y ~ x + w
  Res.Df    RSS Df  Sum of Sq     F Pr(>F)
1     23 17.694
2     22 17.693  1 0.00066144 8e-04 0.9774
```

A Regression Modeling Approach

Partial F-Tests: A General Approach

- Line 2 of the output gives the F-test for comparing Model B with Model A.
- Notice that the p value is 0.974, the same as the p value for the regression coefficient attached to w in the previous output.
- We cannot reject the null hypothesis that the two models fit equally well, so we do not have sufficient information to declare that the regression coefficient for w, when it is added to the model, is non-zero.
- This might make it seem that the `anova` F statistic is superfluous, in that one can obtain similar information from the Wald test on the regression coefficient.
- In general, though, the `anova` test proves to be very valuable, because while the t test only works for a single parameter, the `anova` test can be used to compare models that differ by several parameters.

A Regression Modeling Approach

Partial F-Tests: A General Approach

- What happens if we apply the anova function to a single model with 2 predictors?

```
> anova(fit.B)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          1 14.1025 14.1025 17.5352 0.0003817 ***
w          1  0.0007  0.0007  0.0008 0.9773798
Residuals 22 17.6933  0.8042
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Automatically, the command runs a sequence of tests. It adds the terms in the regression equation one at a time, each time computing a difference test.

A Regression Modeling Approach

Partial F-Tests: A General Approach

- That's really cool — can save a huge amount of work.
- But what happens if we call the `anova` command with just a single model with a single predictor? Will the command choke? It seems like there is nothing to compare.

```
> anova(fit.A)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x          1 14.102 14.1025  18.331 0.0002791 ***
Residuals 23 17.694  0.7693
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The function did not choke. But what happened?
- Note that the p-value for this test is the same as the p-value for the overall test of zero squared multiple correlation shown in the output summary for `fit.A`.
- What is going on?

A Regression Modeling Approach

Partial F-Tests: A General Approach

```
> summary(fit.A)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8459 -0.6692  0.2133  0.5082  1.2330

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2549     0.1754   1.453 0.159709
x             0.8111     0.1894   4.282 0.000279 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8771 on 23 degrees of freedom
Multiple R-squared: 0.4435,        Adjusted R-squared: 0.4193
F-statistic: 18.33 on 1 and 23 DF,  p-value: 0.0002791
```

A Regression Modeling Approach

Partial F-Tests: A General Approach

- To demonstrate, let's fit a model with just an intercept.

```
> fit.0 <- lm(y ~ 1)
```

- Recall that the 1 in the model formula stands for the intercept.
- Now let's *explicitly* perform a partial F-test comparing `fit.0` with `fit.A`.

A Regression Modeling Approach

Partial F-Tests: A General Approach

Notice that we get essentially the same result as when we simply applied the `anova` function to the `fit.A` object.

```
> anova(fit.0,fit.A)

Analysis of Variance Table

Model 1: y ~ 1
Model 2: y ~ x
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     24 31.796
2     23 17.694  1    14.102 18.331 0.0002791 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A Regression Modeling Approach

Partial F-Tests: A General Approach

- In a roundabout way, we've shown that the standard F-test that $R^2 = 0$ (or that $\beta_1 = 0$) computed on a model with a single regressor variable is actually a test comparing two models.
- The simpler model (hidden "behind the scenes" with standard procedures) has just an intercept.
- The more complex model adds a single regressor.
- Now we show how this F-test is equivalent to the t-test of equal means.

A Regression Modeling Approach

Setting Up The Data

- Recall our earlier simple data sets

```
> Group.1 <- c(102,131,119,109,111)
> Group.2 <- c(104,98,110,119,99,88)
```

- Two set these data up for analysis using the regression approach, we need to do two things:
  1. Concatenate the data to produce one long vector of scores.

     ```
     > Score <- c(Group.1, Group.2)
     ```
  2. Create a binary "dummy variable" called `Group` that is coded 1 if the person is in Group 1 and 0 if the person is in Group 2.

     ```
     > Group <- c(rep(1,5),rep(0,6))
     ```

A Regression Modeling Approach

Setting Up The Data

Optionally, you can place the data into a dataframe. Here is what the data look like.

```
> t.data <- data.frame(Group,Score)
> t.data

   Group Score
1      1   102
2      1   131
3      1   119
4      1   109
5      1   111
6      0   104
7      0    98
8      0   110
9      0   119
10     0    99
11     0    88
```

A Regression Modeling Approach

Analyzing the Regression Model

- Now let's revisit the regression model.
- We're going to discover that when a binary dummy variable is used, a regression model takes on some interesting new characteristics.
- Model A in our case has only an intercept. And all the data are in just one column.
- The model can be expressed as

$$E(Y) = \beta_0 \tag{7}$$

and

$$Var(Y) = \sigma_2 \tag{8}$$

because there is no X!

- In other words, Model A says that all observations, regardless of their group, have a mean β_0 and a variance σ^2.
- Equal means, equal variances. Does that sound familiar?
- It should, because that is the model for the null hypothesis in the two sample, independent sample t test.

A Regression Modeling Approach

Analyzing the Regression Model

- Now we add the `Group` variable as a predictor.
- The model might be written

$$E(Y|Group = x) = \beta_0 + \beta_1 x \tag{9}$$

and

$$\text{Var}(Y|Group = x) = \sigma^2 \tag{10}$$

- However, because X takes on only the values 1 and 0, we can see that actually, the above model represents two "group-specific" models.
- As before, both groups have variances of σ^2.
- But consider the model for $E(Y|Group = x)$. Since *Group* takes on only two values, the model of Equation 9 can be written

$$\begin{aligned} E(Y|Group = 0) &= \beta_0 \\ E(Y|Group = 1) &= \beta_0 + \beta_1 \end{aligned}$$

A Regression Modeling Approach

Analyzing the Regression Model

- We see then, that Model B allows Group 2 ($x = 0$) to have a mean of β_0, and Group 1 ($x = 1$) to have a mean of $\beta_0 + \beta_1$.
- If we can reject Model A in favor of more complex Model B, this is the same as rejecting the null hypothesis of equal means.

A Regression Modeling Approach

Analyzing the Regression Model

- Let's compare the two models.

```
> fit.A <- lm(Score ~ 1)
> fit.B <- lm(Score ~ 1 + Group)
> anova(fit.A,fit.B)
Analysis of Variance Table

Model 1: Score ~ 1
Model 2: Score ~ 1 + Group
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     10 1417.6
2      9 1063.2  1    354.44 3.0003 0.1173
```

- The F statistic is 3.00 with 1 and 9 degrees of freedom.

A Regression Modeling Approach

Analyzing the Regression Model

- One fact discussed in introductory courses is that the square of a t statistic with ν degrees of freedom has an F distribution with 1 and ν degrees of freedom.
- Conversely, if we take the square root of the F statistic of 3.00, we obtain the absolute value of a t statistic.
- Notice that the square root of the F value of 3.00 obtained here is equal to of 1.73 we obtained earlier.
- If we look at the summary output for Model B, we see that the t statistic for the null hypothesis that $\beta_1 = 0$ is identical to the t statistic obtained earlier.
- That is, of course, because $\beta_1 = 0$ in our setup if and only if the two groups have equal means. The β_1 coefficient represents $\mu_1 - \mu_2$ in our setup.

A Regression Modeling Approach

Analyzing the Regression Model

```
> summary(fit.B)

Call:
lm(formula = Score ~ 1 + Group)

Residuals:
   Min     1Q Median     3Q    Max
 -15.0   -5.2   -3.4    5.8   16.6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  103.000      4.437  23.213 2.43e-09 ***
Group         11.400      6.581   1.732    0.117
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.87 on 9 degrees of freedom
Multiple R-squared:  0.25,        Adjusted R-squared: 0.1667
F-statistic:     3 on 1 and 9 DF,  p-value: 0.1173
```

A Regression Modeling Approach

An Easy Confidence Interval on the Mean Difference

- A common analysis done in conjunction with the t-test is to construct a confidence interval on the quantity $\mu_1 - \mu_2$.
- Do you see a simple way to obtain a confidence interval on one of the above quantity from the output from `summary(fit.B)`?

A Regression Modeling Approach

An Easy Confidence Interval on the Mean Difference

- If you look at β_1 carefully, you will see that it actually represents $\mu_1 - \mu_2$, because it represents the amount that must be added to μ_2 to model μ_1 correctly.
- Consequently, a confidence interval on β_1 is a confidence interval on $\mu_1 - \mu_2$.
- You can obtain such confidence intervals calculated precisely by using the `confint` function.

A Regression Modeling Approach

An Easy Confidence Interval on the Mean Difference

```
> confint(fit.B)

                2.5 %    97.5 %
(Intercept) 92.962319 113.03768
Group       -3.488287  26.28829
```