

**Lecture 4: Hypothesis Testing**  
**in the Classical Linear Regression Model**

**(Hayashi, pp. 33-46)**

- 1. Some terminology**
- 2. The normality assumption**
- 3.  $t$  tests, confidence intervals, and  $p$ -values for individual regression coefficients**
- 4.  $F$  tests for multiple regression coefficients and for linear restrictions on coefficients**

# 1. Terminology

**All of econometrics (and pretty much all of empirical economics in general) is about “conditional inference”, in the following sense:**

**We start with a *maintained hypothesis*, or “model” which we assume is true, before we even touch any data.** This maintained hypothesis gives us enough structure to pose questions rigorously and formulate formal, statistical tests.

**Given this maintained hypothesis, we can then do formal statistical tests about whether a specific statement, called the *null hypothesis*, is consistent with the data we observe.** The technology of statistics allows us to make quite precise statements about the chances that the null hypothesis is true, given our data and model. If those chances are very low, (e.g. less than .1%, 1%, or 5%), we say the null hypothesis is *rejected* by the data. These are cases where the data “speaks decisively” about something.

**But note two things that are true in general:**

- 1. “Tests” or probability statements about the *null hypothesis* ARE NOT VALID if the (untested) maintained hypothesis is false.**
- 2. If the data does not reject a null hypothesis at, say, the 5% level, THAT DOES NOT MEAN THE NULL HYPOTHESIS IS TRUE. It only means that *we can't be more than 95% sure it is false.***

## **A Comment:**

It's probably fair to say that there are two competing/complementary 'styles' of research in economics today:

A)- the 'structural' approach:

- very detailed set of maintained assumptions (for example a dynamic equilibrium model of the entire economy).

- estimated parameters are often 'fundamental' economic ones, e.g. the level of risk aversion or intertemporal substitution in the U function.

- allows strong statements about policy effects on prices, quantities, utility and welfare IF the maintained assumptions are true.

B)- the 'Experimental' approach

-includes 'natural experiments'

-goal is to make credible inferences about the causal effects of X on y using a *minimum* number of (ideally zero) maintained assumptions.

-approaches include natural, lab and field experiments, IV and RD design, semi/nonparametric econometrics; most of the focus is on the exogeneity assumption.

-does not produce conclusions about utility, welfare.

-the 'gold standard' in this style of research is the *randomized clinical trial*.

Good reference: Angrist and Pischke, *Mostly Harmless Econometrics*

**Today's example :** Suppose the classical linear regression model (Assumptions 1-4) is true. Add one more assumption (normality of the error term; to be described). Then we can test a variety of null hypotheses about the true parameter vector,  $\beta$ . **BUT ALL OF THESE TESTS ARE VALID ONLY IF THE MAINTAINED ASSUMPTIONS OF THE CLASSICAL LINEAR REGRESSION MODEL (plus normality) ARE TRUE.**

## 2. The Normality Assumption

**Assumption 5** (*in addition to linearity, strict exogeneity, no multicollinearity, and homoskedasticity*):

The distribution of  $\boldsymbol{\varepsilon}$ , conditional on  $\mathbf{X}$ , is multivariate normal.

(see any statistics textbook for a description of the multivariate normal distribution).

## Useful Properties of the Normal Distribution:

1. The univariate normal has only two parameters, the mean and the variance. So, if a univariate normal distribution's mean and variance are independent of  $\mathbf{X}$ , then that distribution is identical for all  $\mathbf{X}$ .
2. If two variables are jointly normally distributed, then a lack of correlation between them implies they are statistically independent. (Recall that correlation measures the strength of *linear* relationships only).
3. A linear function of a normally distributed variable is also normally distributed.
4. If  $x$  and  $y$  are jointly normally distributed, then the distribution of  $x$  conditional on  $y$  is normal with a mean that is linear in  $y$  and a conditional variance that is independent of  $y$ .

**Now, if we add Assumption 5 (joint normality of  $\boldsymbol{\varepsilon}$ ) to Assumptions 2 and 4 (strict exogeneity plus homoskedasticity), it follows that:**

**$\boldsymbol{\varepsilon} \mid \mathbf{X}$  is distributed according to  $N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ .**

Further, recall once again that the sampling error of the OLS coefficient vector is given by:

$$\mathbf{b} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{A}\boldsymbol{\varepsilon}$$

So, since  $\boldsymbol{\varepsilon}$  is multivariate normal, so is  $\mathbf{b} - \boldsymbol{\beta}$ . Also, from previous results, we have already calculated the mean and variance of  $\mathbf{b} - \boldsymbol{\beta}$ . The mean is zero (by unbiasedness), and the variance is  $\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$ .

**Putting these facts together, it follows that:**

**$\mathbf{b} - \boldsymbol{\beta} \mid \mathbf{X}$  is distributed according to  $N(\mathbf{0}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1})$ .**

**The above property is the basis for most of the hypothesis testing in everyday economic research.**

### 3. Tests Concerning Individual Regression Coefficients

Consider the null hypothesis:  $H_0 : \beta_k = \bar{\beta}_k$

(often, but not always, we are interested in the case where  $\bar{\beta}_k = 0$ )

*If the null hypothesis is true* (i.e. “under the null”),

$(b - \bar{\beta}_k) | \mathbf{X}$  is distributed according to  $N(0, \sigma^2 (\mathbf{X}'\mathbf{X})_{kk}^{-1})$

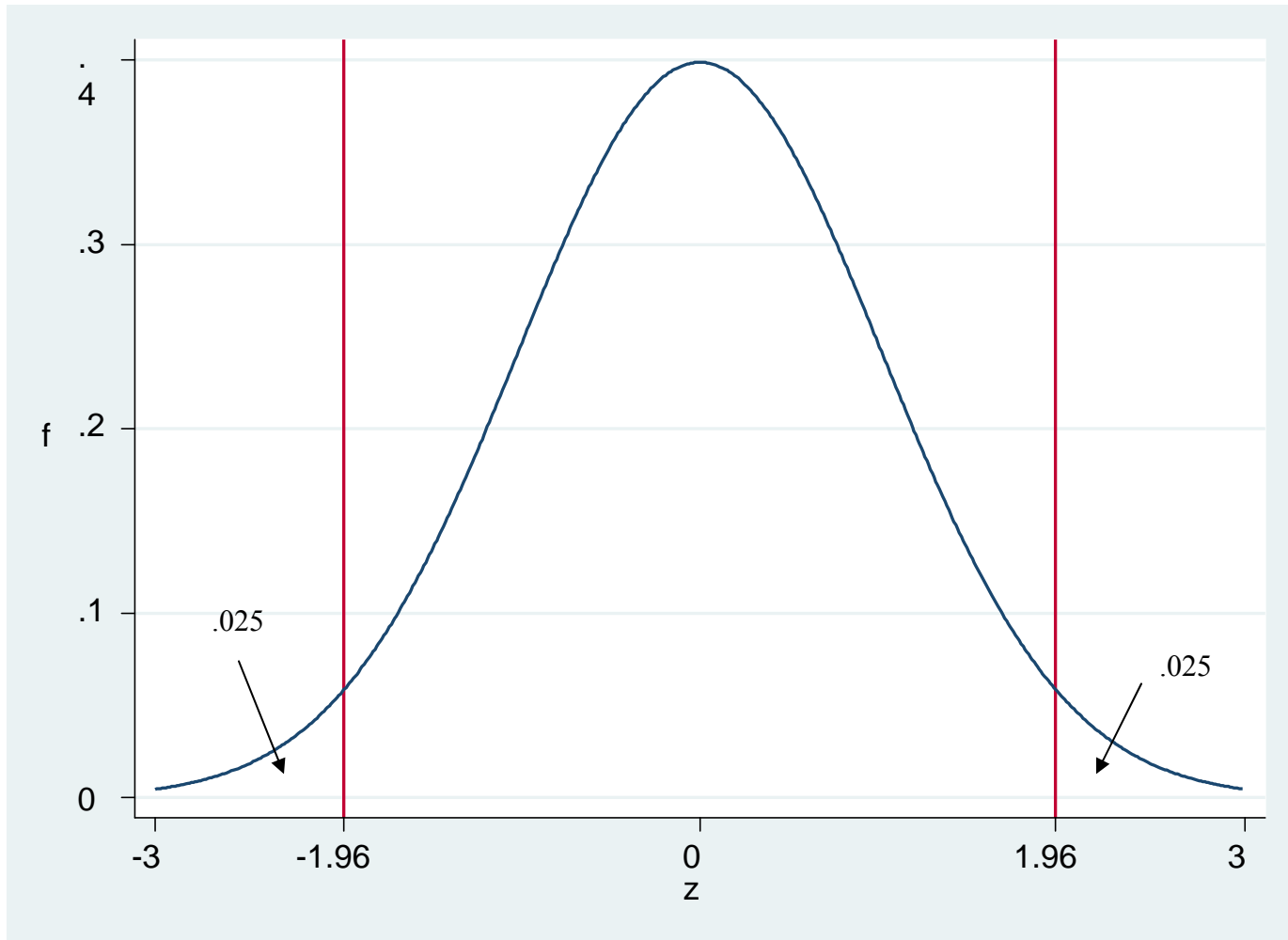
In words, if the true value of  $\beta_k$  is  $\bar{\beta}_k$ , and if each element of  $\boldsymbol{\varepsilon}$  is drawn independently from a normal distribution with mean 0 and variance  $\sigma^2$ , then, estimating this regression on a number of independently-drawn samples, the distribution of the difference between the estimated and true value of the  $k$ th coefficient is normal with mean zero and variance  $\sigma^2 (\mathbf{X}'\mathbf{X})_{kk}^{-1}$ .

Next, we use the fact that if  $u$  is  $N(0, \sigma^2)$ ,  $u/\sigma$  is  $N(0, 1)$ . So:

$z_k \equiv \frac{b_k - \bar{\beta}_k}{\sqrt{\sigma^2 (\mathbf{X}'\mathbf{X})_{kk}^{-1}}}$  follows a standard normal distribution.

Thus, suppose we calculated  $z_k$  and it was 3. That means that, under our null hypothesis, the chances of drawing the particular  $\varepsilon$  vector that yielded our estimate of  $b$  were very low. In such a case we would reject the null hypothesis.

More concretely, if  $z_k < -1.96$  or  $z_k > 1.96$ , we can reject the null hypothesis with 95% confidence (because  $\Phi(-1.96) = .025$  and  $\Phi(1.96) = .975$ , where  $\Phi$  is the standard normal cdf.):



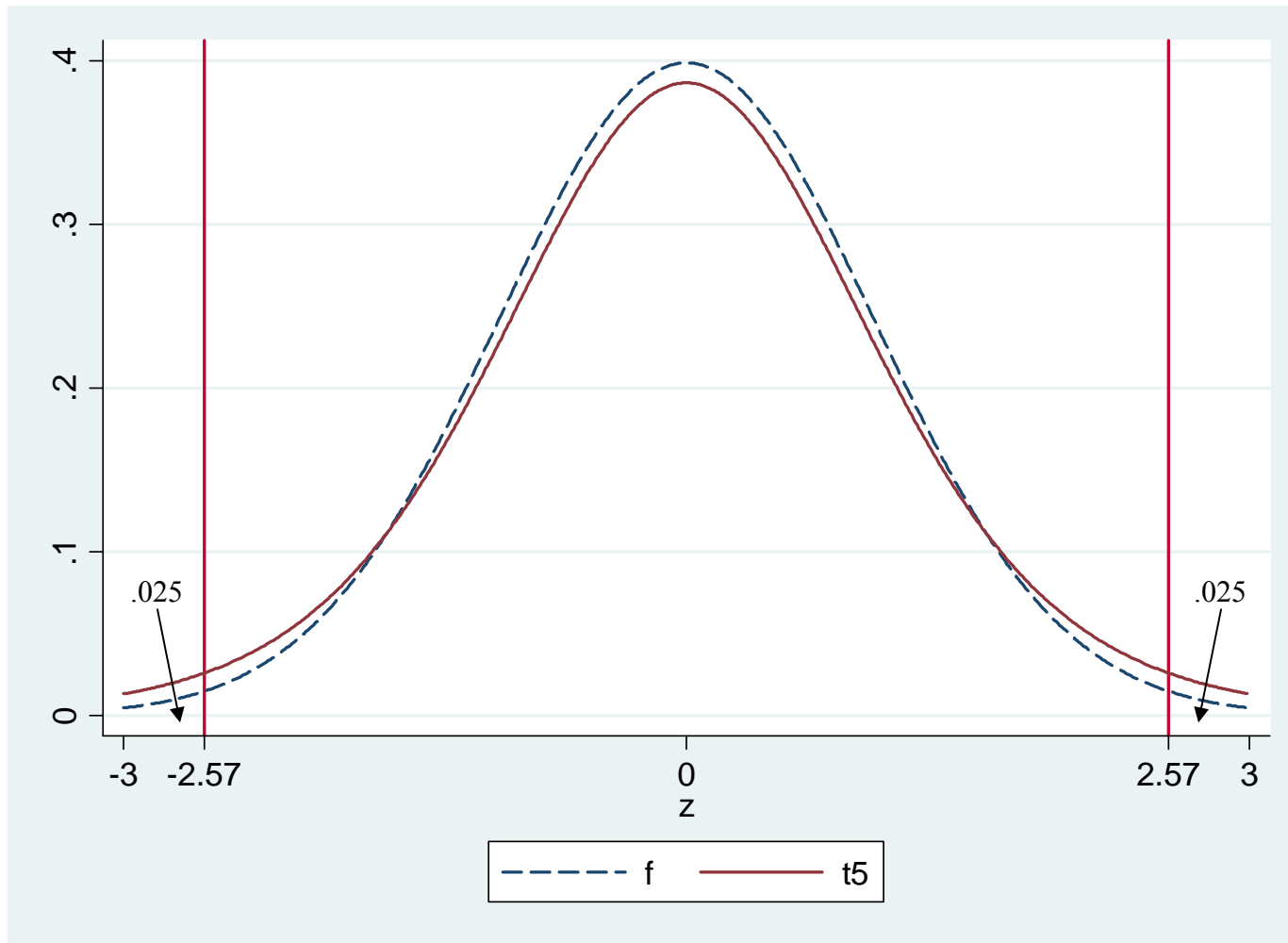
Of course, there's only one problem with the above procedure:

The calculations use  $\sigma^2$ , when in fact we don't know  $\sigma^2$ .

Solution: Use an unbiased estimate,  $s^2 = \frac{SSR}{n-K} = \frac{\mathbf{e}'\mathbf{e}}{n-K}$ , in place of  $\sigma^2$ .

Now,  $t_k \equiv \frac{b_k - \bar{\beta}_k}{\sqrt{s^2 (\mathbf{X}'\mathbf{X})_{kk}^{-1}}}$  follows a  $t$  distribution with  $n-K$  degrees of freedom. The  $t$  distribution looks like the normal, but with fatter tails (the tails get fatter as  $n-K$  shrinks):

$t$  distribution with 5 degrees of freedom:



Summary of  $t$  tests for individual regression coefficients:

1. Estimate the regression and calculate  $t_k \equiv \frac{b_k - \bar{\beta}_k}{\sqrt{s^2 (\mathbf{X}'\mathbf{X})_{kk}^{-1}}}$  for your null hypothesis (e.g.  $\bar{\beta}_k = 0$ ).

2. Choose a threshold level of significance, e.g. 1%. Look up the critical values of  $t$  for that significance level and for the degrees of freedom ( $n-K$ ) in your sample. (They will be equal and opposite in sign, e.g. -2.57 and 2.57) (If you were doing this by hand in Stata you would use the *invttail* function). Call these critical values  $-t^*$  and  $t^*$ .

3. Compare your  $t$  statistic to the critical values. If it lies outside the acceptance region, you reject the null at the level of significance you've selected.

Another approach: *Confidence intervals*

Note: We often refer to the denominator of  $t_k$ ,  $\sqrt{s^2 (\mathbf{X}'\mathbf{X})_{kk}^{-1}}$ , as the *standard error* of  $b_k$  ( $SE(b_k)$ ). So we accept  $H_0$  whenever:

$$-t^* < \frac{b_k - \bar{\beta}_k}{SE(b_k)} < t^*, \quad \text{or} \quad -t^* SE(b_k) < b_k - \bar{\beta}_k < t^* SE(b_k), \quad \text{or:}$$

$$b_k - t^* SE(b_k) < \bar{\beta}_k < b_k + t^* SE(b_k).$$

The interval  $[b_k - t^* SE(b_k), b_k + t^* SE(b_k)]$  is called the ( $x\%$ ) confidence interval for  $b_k$ . The nice thing about C.I.s is that they can be constructed without reference to a specific null hypothesis,  $\bar{\beta}_k$ . We can just construct the C.I.s for the estimated coefficients and say that we can reject, with ( $x\%$ ) confidence, all null hypotheses that lie outside this interval.

Rough rule: reject all parameter values that differ from the estimated coefficient by more than two standard errors.

Yet another approach: *p-values*.

Suppose we don't want to specify an arbitrary significance level in advance.

Then we can:

1) Compute the t-ratio  $t_k \equiv \frac{b_k - \bar{\beta}_k}{\sqrt{s^2 (\mathbf{X}'\mathbf{X})_{kk}^{-1}}}$ , as before.

2) Calculate the probability that a *t*-variate with  $n-K$  degrees of freedom exceeds  $t_k$  in absolute value. (In other words, what is the *highest level of significance* at which you could reject the null hypothesis that  $\beta_k = \bar{\beta}_k$ ?).

Formally,  $p$  is implicitly defined by:  $\text{Prob} (-|t_k| < t < |t_k|) = 1-p$

Note: when Stata and other statistical packages report  $p$ -values for estimated coefficients, these refer to the specific null hypothesis that  $\bar{\beta}_k = 0$ .

## 4. Tests Concerning Multiple Regression Coefficients

Now, consider the null hypothesis:

$$H_0 : \underbrace{\mathbf{R}}_{\#r \times K} \underbrace{\boldsymbol{\beta}}_{K \times 1} = \underbrace{\mathbf{r}}_{\#r \times 1}$$

Examples include:

$\beta_k = 0$  (One restriction,  $\#r=1$ ).

$\boldsymbol{\beta} = \mathbf{0}$  ( $K$  restrictions,  $\#r=K$ ). (All the coefficients are zero)

All the coefficients but the constant are zero ( $\#r=K-1$ )

$\beta_1 = -\beta_2$  (The coefficients on variables 1 and 2 are equal but opposite in sign). (One restriction)

... and many others. We assume that  $\mathbf{R}$  is of full row rank.

*Proposition:* Under assumptions 1-5, and under the null hypothesis  $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ , the  $F$ -ratio, defined as:

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

$$= (\mathbf{R}\mathbf{b} - \mathbf{r})' [\mathbf{R}(\mathit{Varhat}(\mathbf{b} | \mathbf{X}))\mathbf{R}']^{-1} (\mathbf{R}\mathbf{b} - \mathbf{r}) / \# \mathbf{r}$$

(where  $\mathit{Varhat}(\mathbf{b} | \mathbf{X}) = s^2 (\mathbf{X}'\mathbf{X})^{-1}$  is the *estimated* variance-covariance matrix of the coefficients),

follows the  $F$  distribution with degrees of freedom  $\#r$  and  $n-K$ .

So, to conduct an  $F$ -test of the restriction  $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ ,

Estimate the regression and calculate the  $F$  statistic for your null hypothesis, using the above formula.

2. Choose a threshold level of significance, e.g. 1%. Look up the critical value of  $F$  for that significance level (there will be only one; it will be positive), and for degrees of freedom  $\#r$  and  $n-K$ . Call this critical value  $F^*$ .

3. Compare your  $F$  statistic to the critical value. If  $F > F^*$ , you reject the null at the level of significance you've selected.

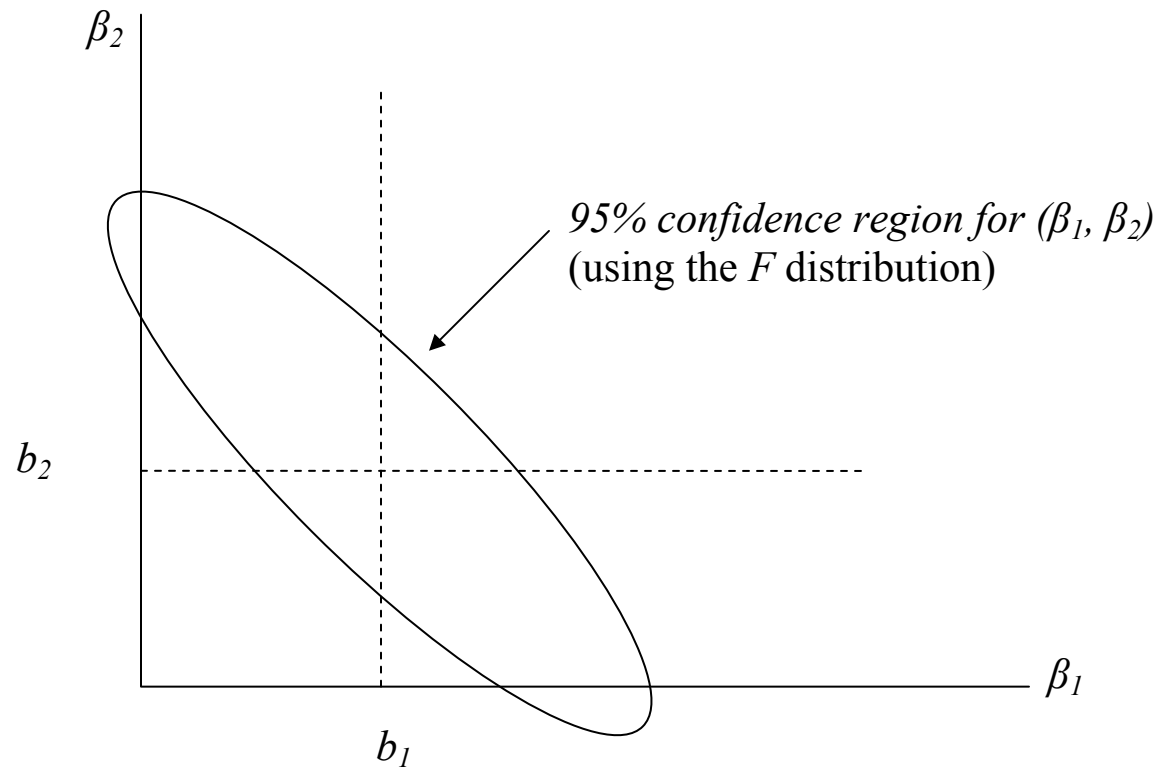
## Applications of $F$ tests:

-sometimes we're interested in the significance of a group of variables representing a specific type of factor: personality indicators, all the geographic indicators, etc.

-to allow for nonlinearity, often represent a single 'underlying' variable by a polynomial or a set of dummies (step function)

- $X$  variables are often correlated with each other. In consequence, the data might sometimes show quite conclusively that *either*  $x_1$  or  $x_2$  has a strong effect on  $y$ , but can't say which one it is. In such cases,  $t$ -tests show that both variables are insignificant, but an  $F$  test reveals the deeper, underlying reality:

A case where an  $F$ -test rejects the hypothesis  $\beta_1 = \beta_2 = 0$ , while individual  $t$ -tests could accept both  $\beta_1 = 0$  and  $\beta_2 = 0$ :



(this would occur if  $\text{Cov}(b_1, b_2) < 0$ )

A final note on  $F$  tests:

Suppose you were able to estimate the “restricted” model, i.e. minimize the sum of squared residuals while imposing the restriction  $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ . (This is often much easier to do than it appears).

Then there’s a much simpler formula for the  $F$  statistic, namely:

$$F = \frac{(SSR_R - SSR_U) / \# \mathbf{r}}{SSR_U / (n - K)}$$

Intuitively, this is a measure of how much your  $SSR$  (i.e. the regression’s ‘unexplained variance’) rises when you impose the restriction, adjusted for the number of restrictions imposed.

This is our first example of a *Likelihood Ratio (LR) Test*.