Multiple Regression Analysis

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + u$

- Inference

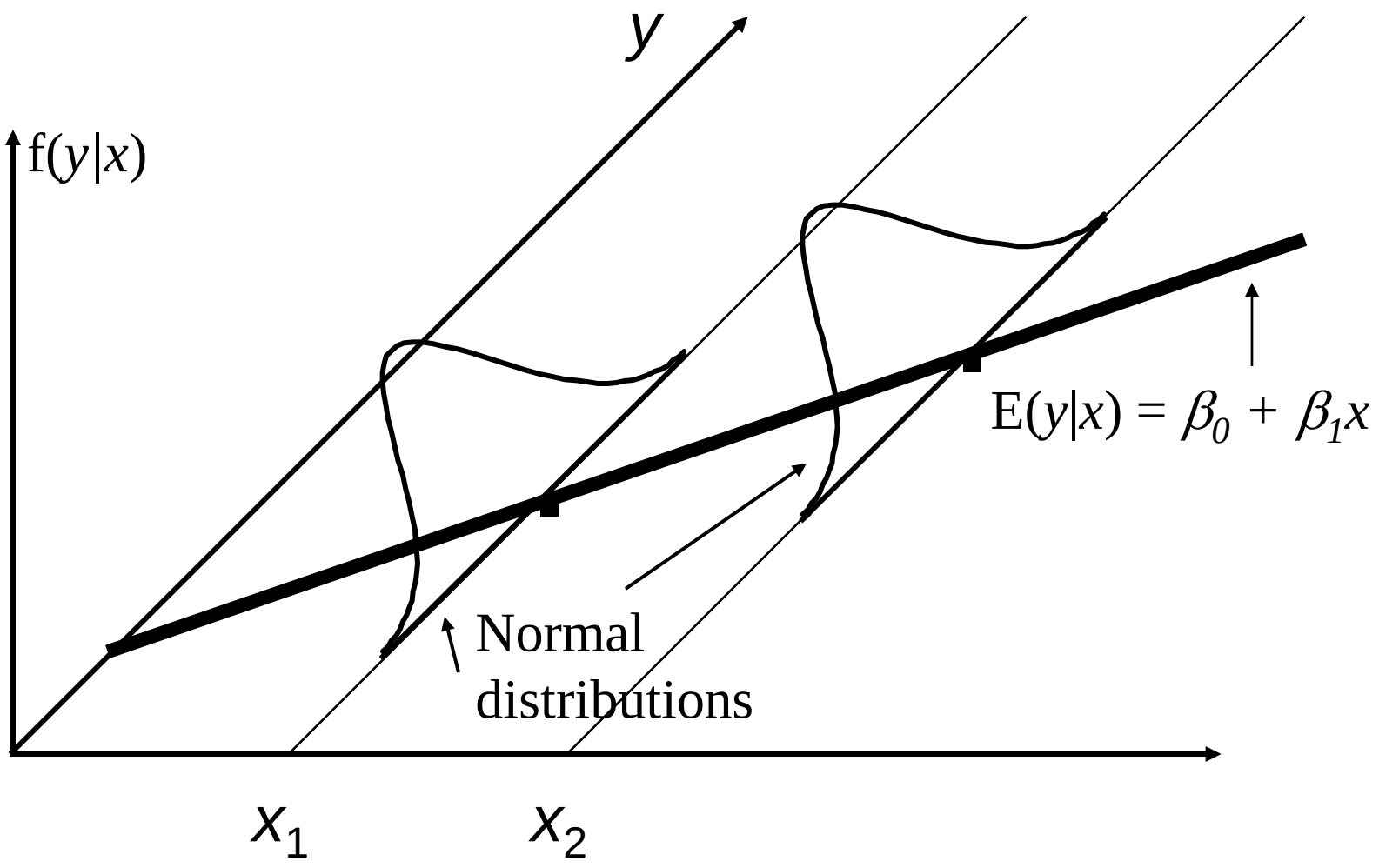
Assumptions of the Classical Linear Model (CLM)

- So far, we know that given the Gauss-Markov assumptions, OLS is BLUE,
- In order to do classical hypothesis testing, we need to add another assumption (beyond the Gauss-Markov assumptions)
- Assume that u is independent of $x_1, x_2, \ldots, x_k$ and u is normally distributed with zero mean and variance σ^2: $u \sim \text{Normal}(0, \sigma^2)$

CLM Assumptions (cont)

- Under CLM, OLS is not only BLUE, but is the minimum variance unbiased estimator
- We can summarize the population assumptions of CLM as follows
- $y|\boldsymbol{x} \sim \text{Normal}(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k, \sigma^2)$
- While for now we just assume normality, clear that sometimes not the case
- Large samples will let us drop normality

The homoskedastic normal distribution with a single explanatory variable



Normal Sampling Distributions

Under the CLM assumptions, conditional on the sample values of the independent variables

$\hat{\beta}_j \sim \text{Normal}\left[\beta_j, Var\left(\hat{\beta}_j\right)\right]$, so that

$$\left(\hat{\beta}_j - \beta_j\right) \Big/ sd\left(\hat{\beta}_j\right) \sim \text{Normal}(0,1)$$

$\hat{\beta}_j$ is distributed normally because it is a linear combination of the errors

The *t* Test

Under the CLM assumptions

$$\left(\hat{\beta}_j - \beta_j\right) \Big/ se\left(\hat{\beta}_j\right) \sim t_{n-k-1}$$

Note this is a *t* distribution (vs normal)

because we have to estimate σ^2 by $\hat{\sigma}^2$

Note the degrees of freedom : $n-k-1$

The *t* Test (cont)

- Knowing the sampling distribution for the standardized estimator allows us to carry out hypothesis tests
- Start with a null hypothesis
- For example, H_0: $\beta_j=0$
- If accept null, then accept that x_j has no effect on y, controlling for other x's

The *t* Test (cont)

To perform our test we first need to form "the" t statistic for $\hat{\beta}_j$: $t_{\hat{\beta}_j} \equiv \hat{\beta}_j / se(\hat{\beta}_j)$

We will then use our t statistic along with a rejection rule to determine whether to accept the null hypothesis, H_0

t Test: One-Sided Alternatives

- Besides our null, H_0, we need an alternative hypothesis, H_1, and a significance level
- H_1 may be one-sided, or two-sided
- H_1: $\beta_j > 0$ and H_1: $\beta_j < 0$ are one-sided
- H_1: $\beta_j \neq 0$ is a two-sided alternative
- If we want to have only a 5% probability of rejecting H_0 if it is really true, then we say our significance level is 5%

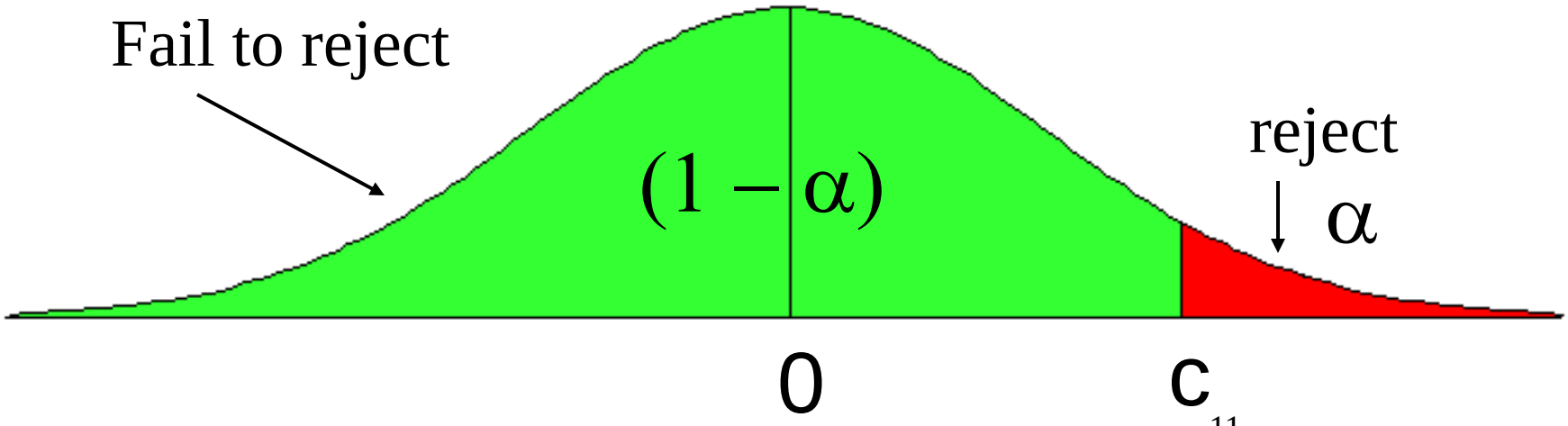
One-Sided Alternatives (cont)

- Having picked a significance level, α, we look up the $(1 - \alpha)^{th}$ percentile in a *t* distribution with $n - k - 1$ df and call this *c*, the critical value
- We can reject the null hypothesis if the *t* statistic is greater than the critical value
- If the *t* statistic is less than the critical value then we fail to reject the null

One-Sided Alternatives (cont)

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i$$

$H_0: \beta_j = 0$ $\quad$ $H_1: \beta_j > 0$



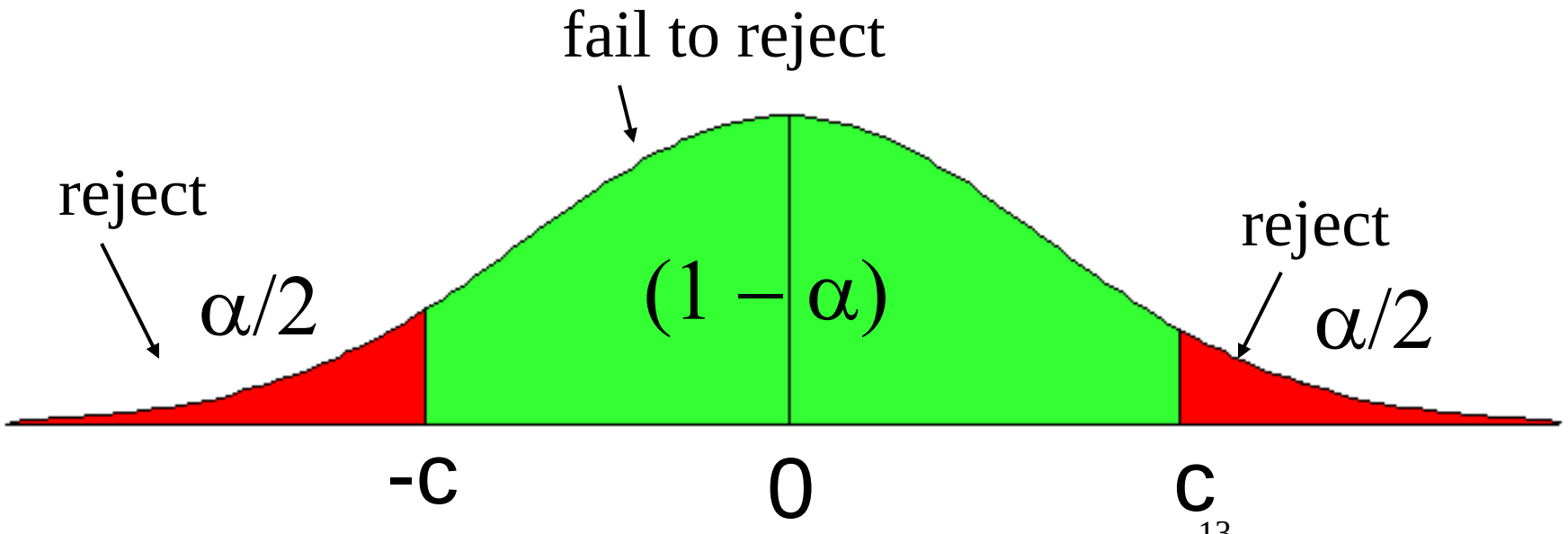
One-sided vs Two-sided

- Because the *t* distribution is symmetric, testing $H_1: \beta_j < 0$ is straightforward. The critical value is just the negative of before
- We can reject the null if the *t* statistic < –*c*, and if the *t* statistic > than –*c* then we fail to reject the null
- For a two-sided test, we set the critical value based on $\alpha/2$ and reject $H_1: \beta_j \neq 0$ if the absolute value of the *t* statistic > *c*

Two-Sided Alternatives

$$y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik} + u_i$$

$H_0: \beta_j = 0$ $\qquad$ $H_1: \beta_j \neq 0$



Summary for H_0: $\beta_j = 0$

- Unless otherwise stated, the alternative is assumed to be two-sided
- If we reject the null, we typically say "x_j is statistically significant at the α % level"
- If we fail to reject the null, we typically say "x_j is statistically insignificant at the α % level"

Testing other hypotheses

- A more general form of the *t* statistic recognizes that we may want to test something like H_0: $\beta_j = a_j$
- In this case, the appropriate *t* statistic is

$$t = \left(\hat{\beta}_j - a_j\right) \Big/ se\left(\hat{\beta}_j\right), \text{ where}$$

$$a_j = 0 \text{ for the standard test}$$

Confidence Intervals

- Another way to use classical statistical testing is to construct a confidence interval using the same critical value as was used for a two-sided test
- A (1 - α) % confidence interval is defined as

$$\hat{\beta}_j \pm c \bullet se\left(\hat{\beta}_j\right), \text{ where c is the } \left(1 - \frac{\alpha}{2}\right) \text{percentile}$$

$$\text{in a } t_{n-k-1} \text{ distribution}$$

Computing *p*-values for *t* tests

- An alternative to the classical approach is to ask, “what is the smallest significance level at which the null would be rejected?”
- So, compute the *t* statistic, and then look up what percentile it is in the appropriate *t* distribution – this is the *p*-value
- *p*-value is the probability we would observe the *t* statistic we did, if the null were true

Stata and *p*-values, *t* tests, etc.

- Most computer packages will compute the *p*-value for you, assuming a two-sided test
- If you really want a one-sided alternative, just divide the two-sided *p*-value by 2
- GRETL provides the *t* statistic, *p*-value, and 95% confidence interval for H_0: $\beta_j = 0$ for you, labelled “t”, “P > |t|” and “[95% Conf. Interval]”, respectively

Testing a Linear Combination

- Suppose instead of testing whether β_1 is equal to a constant, you want to test if it is equal to another parameter, that is $H_0 : \beta_1 = \beta_2$
- Use same basic procedure for forming a t statistic

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{se\left(\hat{\beta}_1 - \hat{\beta}_2\right)}$$

Testing Linear Combo (cont)

Since

$$se\left(\hat{\beta}_1 - \hat{\beta}_2\right) = \sqrt{Var\left(\hat{\beta}_1 - \hat{\beta}_2\right)}, \text{ then}$$

$$Var\left(\hat{\beta}_1 - \hat{\beta}_2\right) = Var\left(\hat{\beta}_1\right) + Var\left(\hat{\beta}_2\right) - 2Cov\left(\hat{\beta}_1, \hat{\beta}_2\right)$$

$$se\left(\hat{\beta}_1 - \hat{\beta}_2\right) = \left\{\left[se\left(\hat{\beta}_1\right)\right]^2 + \left[se\left(\hat{\beta}_2\right)\right]^2 - 2s_{12}\right\}^{1/2}$$

where s_{12} is an estimate of $Cov\left(\hat{\beta}_1, \hat{\beta}_2\right)$

Testing a Linear Combo (cont)

- So, to use formula, need s_{12}, which standard output does not have
- Many packages will have an option to get it, or will just perform the test for you
- More generally, you can always restate the problem to get the test you want

Example:

- Suppose you are interested in the effect of campaign expenditures on outcomes
- Model is $voteA = \beta_0 + \beta_1 \log(expendA) + \beta_2 \log(expendB) + \beta_3 prtystrA + u$
- H_0: $\beta_1 = -\beta_2$, or H_0: $\theta_1 = \beta_1 + \beta_2 = 0$
- $\beta_1 = \theta_1 - \beta_2$, so substitute in and rearrange $\Rightarrow$ $voteA = \beta_0 + \theta_1 \log(expendA) + \beta_2 \log(expendB - expendA) + \beta_3 prtystrA + u$

Example (cont):

- This is the same model as originally, but now you get a standard error for $\beta_1 - \beta_2 = \theta_1$ directly from the basic regression
- Any linear combination of parameters could be tested in a similar manner
- Other examples of hypotheses about a single linear combination of parameters:
  - $\beta_1 = 1 + \beta_2$; $\beta_1 = 5\beta_2$; $\beta_1 = -1/2\beta_2$; etc

Multiple Linear Restrictions

- Everything we've done so far has involved testing a single linear restriction, (e.g. $\beta_1 = 0$ or $\beta_1 = \beta_2$)
- However, we may want to jointly test multiple hypotheses about our parameters
- A typical example is testing "exclusion restrictions" – we want to know if a group of parameters are all equal to zero

Testing Exclusion Restrictions

- Now the null hypothesis might be something like H_0: $\beta_{k-q+1} = 0, \ldots, \beta_k = 0$
- The alternative is just H_1: H_0 is not true
- Can't just check each *t* statistic separately, because we want to know if the *q* parameters are jointly significant at a given level – it is possible for none to be individually significant at that level

Exclusion Restrictions (cont)

- To do the test we need to estimate the "restricted model" without $x_{k-q+1}, \ldots, x_k$ included, as well as the "unrestricted model" with all *x*'s included
- Intuitively, we want to know if the change in SSR is big enough to warrant inclusion of $x_{k-q+1}, \ldots, x_k$

$$F \equiv \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}, \text{ where}$$

r is restricted and ur is unrestricted

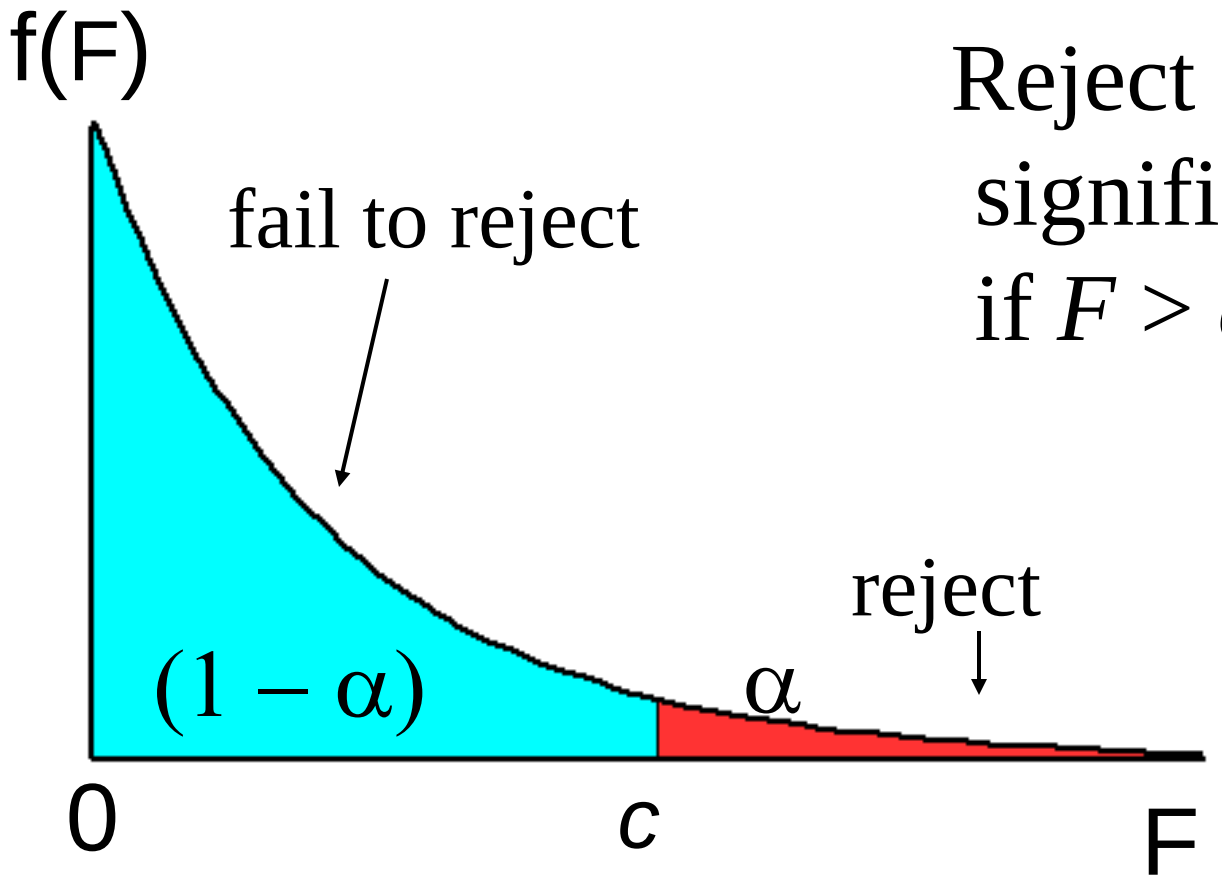
The *F* statistic

- The *F* statistic is always positive, since the SSR from the restricted model can't be less than the SSR from the unrestricted
- Essentially the *F* statistic is measuring the relative increase in SSR when moving from the unrestricted to restricted model
- q = number of restrictions, or $df_r - df_{ur}$
- $n - k - 1 = df_{ur}$

The *F* statistic (cont)

- To decide if the increase in SSR when we move to a restricted model is "big enough" to reject the exclusions, we need to know about the sampling distribution of our *F* stat
- Not surprisingly, $F \sim F_{q,n-k-1}$, where q is referred to as the numerator degrees of freedom and $n - k - 1$ as the denominator degrees of freedom

The *F* statistic (cont)



Reject H_0 at α significance level if $F > c$

The R^2 form of the F statistic

- Because the SSR's may be large and unwieldy, an alternative form of the formula is useful
- We use the fact that SSR = SST(1 – R^2) for any regression, so can substitute in for SSR_u and SSR_{ur}

$$F \equiv \frac{\left(R_{ur}^2 - R_r^2\right)/q}{\left(1 - R_{ur}^2\right)/(n-k-1)}, \text{ where again}$$

r is restricted and ur is unrestricted

Overall Significance

- A special case of exclusion restrictions is to test $H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0$
- Since the R^2 from a model with only an intercept will be zero, the *F* statistic is simply

$$F = \frac{R^2 / k}{\left(1 - R^2\right) / \left(n - k - 1\right)}$$

General Linear Restrictions

- The basic form of the *F* statistic will work for any set of linear restrictions
- First estimate the unrestricted model and then estimate the restricted model
- In each case, make note of the SSR
- Imposing the restrictions can be tricky – will likely have to redefine variables again

Example:

- Use same voting model as before
- Model is $voteA = \beta_0 + \beta_1 \log(expendA) + \beta_2 \log(expendB) + \beta_3 prtystrA + u$
- now null is H_0: $\beta_1 = 1$, $\beta_3 = 0$
- Substituting in the restrictions: $voteA = \beta_0 + \log(expendA) + \beta_2 \log(expendB) + u$, so
- Use $voteA - \log(expendA) = \beta_0 + \beta_2 \log(expendB) + u$ as restricted model

F Statistic Summary

- Just as with *t* statistics, p-values can be calculated by looking up the percentile in the appropriate *F* distribution
- Stata will do this by entering: display fprob(*q, n – k* – 1, *F*), where the appropriate values of *F, q,*and *n – k* – 1 are used
- If only one exclusion is being tested, then *F* = t^2, and the *p*-values will be the same

Multiple Regression Analysis

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + u$
- Dummy Variables

Econometrics V Lecture IV

Dummy Variables

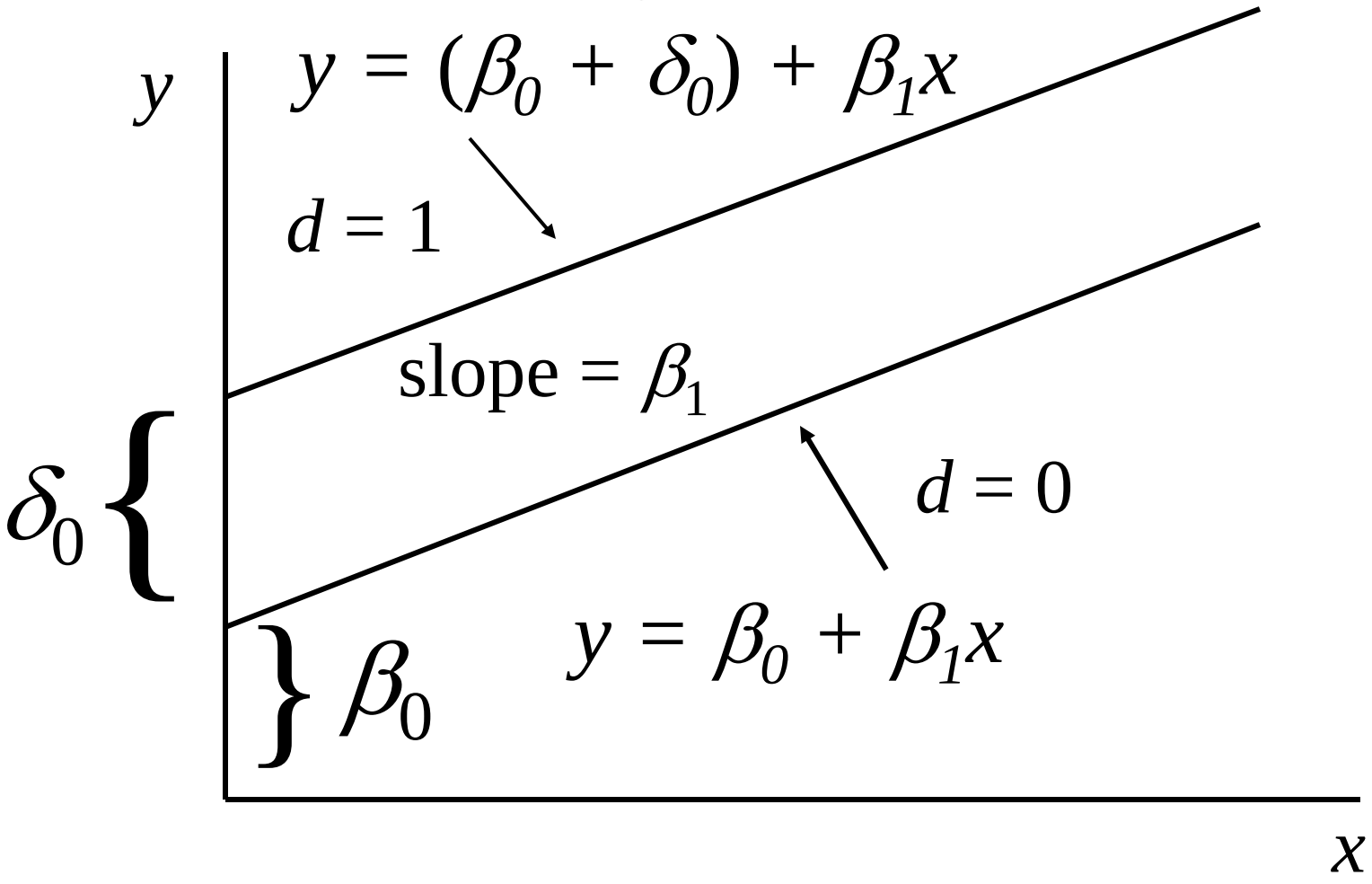
- A dummy variable is a variable that takes on the value 1 or 0
- Examples: male (= 1 if are male, 0 otherwise), south (= 1 if in the south, 0 otherwise), etc.
- Dummy variables are also called binary variables, for obvious reasons

Econometrics V Lecture IV

A Dummy Independent Variable

- Consider a simple model with one continuous variable (x) and one dummy (d)
- $y = \beta_0 + \delta_0 d + \beta_1 x + u$
- This can be interpreted as an intercept shift
- If d = 0, then $y = \beta_0 + \beta_1 x + u$
- If d = 1, then $y = (\beta_0 + \delta_0) + \beta_1 x + u$
- The case of d = 0 is the base group

Example of $\delta_0 > 0$



Econometrics V Lecture IV

Dummies for Multiple Categories

- We can use dummy variables to control for something with multiple categories
- Suppose everyone in your data is either a HS dropout, HS grad only, or college grad
- To compare HS and college grads to HS dropouts, include 2 dummy variables
- hsgrad = 1 if HS grad only, 0 otherwise; and colgrad = 1 if college grad, 0 otherwise

Econometrics V Lecture IV

Multiple Categories (cont)

- Any categorical variable can be turned into a set of dummy variables
- Because the base group is represented by the intercept, if there are n categories there should be n – 1 dummy variables
- If there are a lot of categories, it may make sense to group some together
- Example: top 10 ranking, 11 – 25, etc.

Econometrics V Lecture IV
Interactions Among Dummies

- Interacting dummy variables is like subdividing the group
- Example: have dummies for male, as well as hsgrad and colgrad
- Add male*hsgrad and male*colgrad, for a total of 5 dummy variables –> 6 categories
- Base group is female HS dropouts
- hsgrad is for female HS grads, colgrad is for female college grads
- The interactions reflect male HS grad 41 s and male college grads

Econometrics V Lecture IV

More on Dummy Interactions

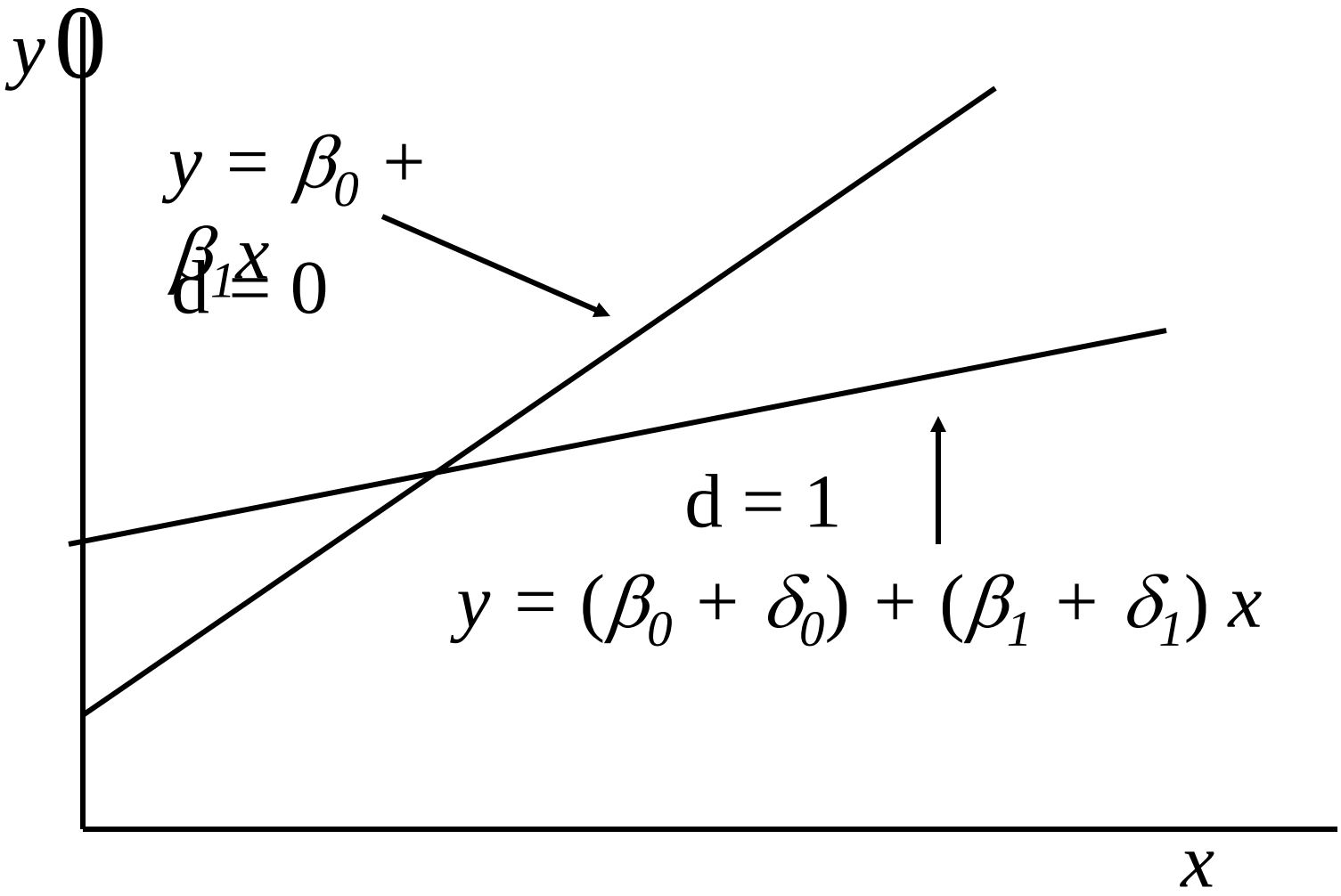
- Formally, the model is $y = \beta_0 + \delta_1 male + \delta_2 hsgrad + \delta_3 colgrad + \delta_4 male*hsgrad + \delta_5 male*colgrad + \beta_1 x + u$, then, for example:
- If male = 0 and hsgrad = 0 and colgrad = 0
- $y = \beta_0 + \beta_1 x + u$
- If male = 0 and hsgrad = 1 and colgrad = 0
- $y = \beta_0 + \delta_2 hsgrad + \beta_1 x + u$
- If male = 1 and hsgrad = 0 and colgrad = 1
- $y = \beta_0 + \delta_1 male + \delta_3 colgrad + \delta_5 male*colgrad + \beta_1 x + u$

Econometrics V Lecture IV

Other Interactions with Dummies

- Can also consider interacting a dummy variable, *d*, with a continuous variable, *x*
- $y = \beta_0 + \delta_1 d + \beta_1 x + \delta_2 d^* x + u$
- If $d = 0$, then $y = \beta_0 + \beta_1 x + u$
- If $d = 1$, then $y = (\beta_0 + \delta_1) + (\beta_1 + \delta_2) x + u$
- This is interpreted as a change in the slope

Example of $\delta_0 > 0$ and $\delta_1 <$



Econometrics V Lecture IV

Testing for Differences Across Groups

- Testing whether a regression function is different for one group versus another can be thought of as simply testing for the joint significance of the dummy and its interactions with all other *x* variables
- So, you can estimate the model with all the interactions and without and form an *F* statistic, but this could be unwieldy

Econometrics V Lecture IV
The Chow Test

- Turns out you can compute the proper F statistic without running the unrestricted model with interactions with all *k* continuous variables
- If run the restricted model for group one and get SSR_1, then for group two and get SSR_2
- Run the restricted model for all to get SSR, then

$$F = \frac{\left[SSR - \left(SSR_1 + SSR_2\right)\right]}{SSR_1 + SSR_2} \bullet \frac{\left[n - 2\left(k+1\right)\right]}{k+1}$$

Econometrics V Lecture IV

The Chow Test (continued)

- The Chow test is really just a simple *F* test for exclusion restrictions, but we've realized that $SSR_{ur} = SSR_1 + SSR_2$
- Note, we have $k + 1$ restrictions (each of the slope coefficients and the intercept)
- Note the unrestricted model would estimate 2 different intercepts and 2 different slope coefficients, so the df is $n - 2k - 2$

Econometrics V Lecture IV

Linear Probability Model

- $P(y = 1|x) = E(y|x)$, when y is a binary variable, so we can write our model as
- $P(y = 1|x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$
- So, the interpretation of β_j is the change in the probability of success when x_j changes
- The predicted y is the predicted probability of success
- Potential problem that can be outside [0,1]

Econometrics V Lecture IV

Linear Probability Model (cont)

- Even without predictions outside of [0,1], we may estimate effects that imply a change in *x* changes the probability by more than +1 or –1, so best to use changes near mean
- This model will violate assumption of homoskedasticity, so will affect inference
- Despite drawbacks, it's usually a good place to start when *y* is binary

Econometrics V Lecture IV

Caveats on Program Evaluation

- A typical use of a dummy variable is when we are looking for a program effect
- For example, we may have individuals that received job training, or welfare, etc
- We need to remember that usually individuals choose whether to participate in a program, which may lead to a self-selection problem

Econometrics V Lecture IV

Self-selection Problems

- If we can control for everything that is correlated with both participation and the outcome of interest then it's not a problem
- Often, though, there are unobservables that are correlated with participation
- In this case, the estimate of the program effect is biased, and we don't want to set policy based on it!

Multiple Regression Analysis

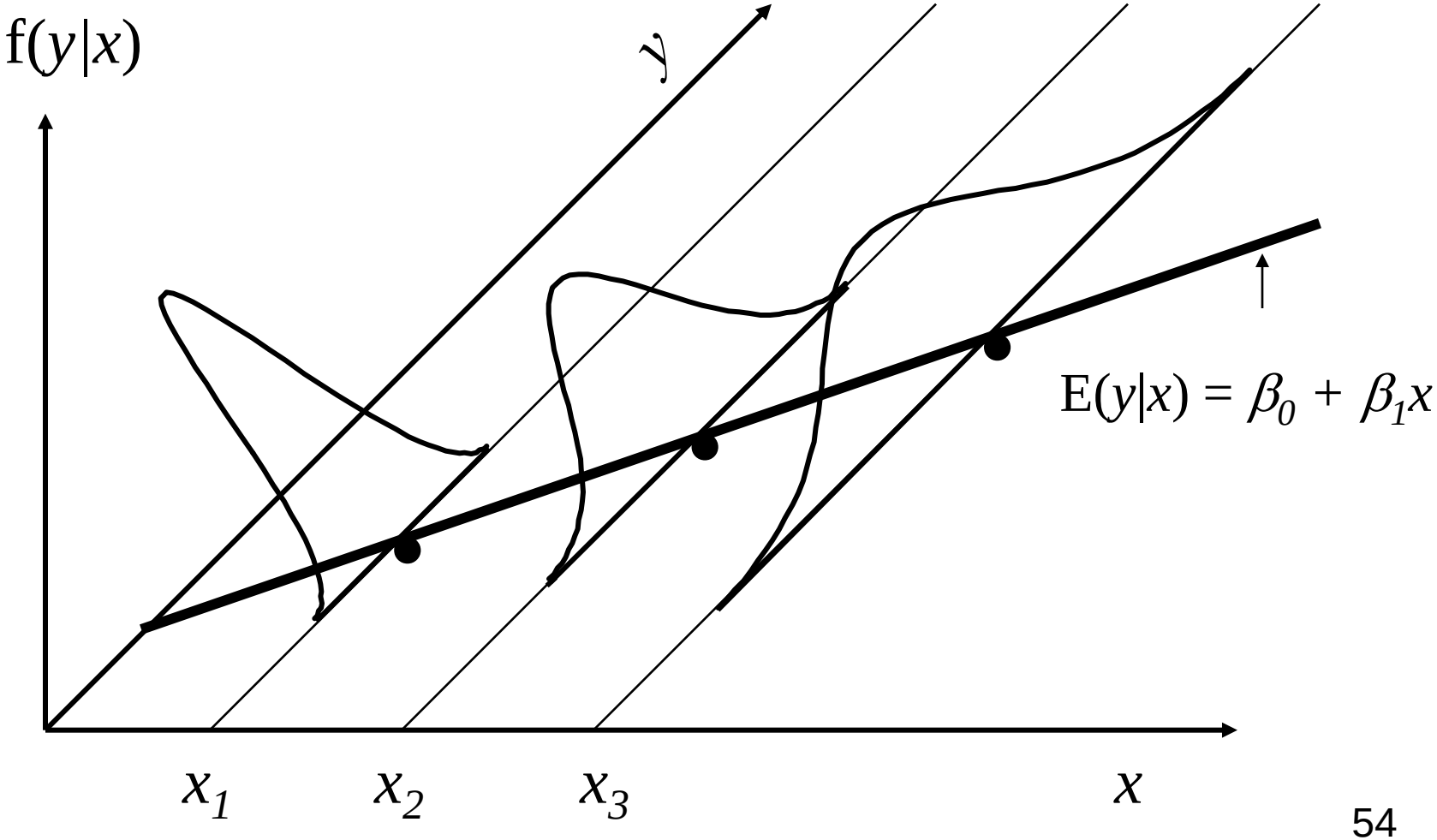
- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + u$
- Heteroskedasticity

Econometrics V Lecture IV

What is Heteroskedasticity

- Recall the assumption of homoskedasticity implied that conditional on the explanatory variables, the variance of the unobserved error, *u*, was constant
- If this is not true, that is if the variance of *u* is different for different values of the *x*'s, then the errors are heteroskedastic
- Example: estimating returns to education and ability is unobservable, and think the variance in ability differs by educational attainment

Example of Heteroskedasticity



Econometrics V Lecture IV
Why Worry About Heteroskedasticity?

- OLS is still unbiased and consistent, even if we do not assume homoskedasticity
- The standard errors of the estimates are biased if we have heteroskedasticity
- If the standard errors are biased, we can not use the usual *t* statistics or *F* statistics or *LM* statistics for drawing inferences

Econometrics V Lecture IV
Variance with Heteroskedasticity

For the simple case, $\hat{\beta}_1 = \beta_1 + \frac{\sum (x_i - \bar{x}) u_i}{\sum (x_i - \bar{x})^2}$, so

$$Var(\hat{\beta}_1) = \frac{\sum (x_i - \bar{x})^2 \sigma_i^2}{SST_x^2} \quad (1),$$

where $SST_x = \sum (x_i - \bar{x})^2$

A valid estimator for this when $\sigma_i^2 \neq \sigma^2$is

$\frac{\sum (x_i - \bar{x})^2 \hat{u}_i^2}{SST_x^2}$, where $\hat{u}_i$ are are the OLS residuals

Econometrics V Lecture IV

- Since the standard error of $\hat{\beta}_1$ is based
- Directly on estimating the variance $\hat{\beta}_1$
- We need to be able to estimate the previous eqn on the middle of the previous slide.
- When heteroscedasticity is present. White (1980) showed this can be done. Let $\hat{U}_i$ denote the OLS residuals from the initial regression of y on x

Econometrics V Lecture IV

- Then a valid estimator of Var $\hat{\beta}_1$ for heteroscedasticity of any form is:

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \hat{u}_i^2}{SST_x^2}$$

- Why? It can be shown that when the equation above is multiplied by n the sample size, its converges in probability to:

Econometrics V Lecture IV

$$E\frac{\left[\left(x_i - u_x\right)^2 u_1^{\,i}\right]}{\left(\sigma_x{}^2\right)^2} \quad which\,is\,the\,probability\lim it\,of\;n\,times\,(1)$$

This is what is necessary for justifying the use of standard errors to construct confidence intervals and t statistics

Econometrics V Lecture IV
Variance with Heteroskedasticity

For the general multiple regression model, a valid estimator of $Var\left(\hat{\beta}_j\right)$ with heteroskedasticity is

$\hat{Var}\left(\hat{\beta}_j\right) = \dfrac{\sum \hat{r}_{ij}^2 \hat{u}_i^2}{SST_j^2}$, where $\hat{r}_{ij}$ is the i^{th} residual from regressing x_j on all other independent variables, and SST_j is the sum of squared residuals from this regression

Econometrics V Lecture IV

Robust Standard Errors

- Now that we have a consistent estimate of the variance, the square root can be used as a standard error for inference
- Typically call these robust standard errors
- Sometimes the estimated variance is corrected for degrees of freedom by multiplying by $n/(n - k - 1)$
- As $n \rightarrow \infty$ it's all the same, though

Econometrics V Lecture IV

Robust Standard Errors (cont)

- Important to remember that these robust standard errors only have asymptotic justification – with small sample sizes *t* statistics formed with robust standard errors will not have a distribution close to the *t*, and inferences will not be correct

Econometrics V Lecture IV

A Robust *LM* Statistic

- Run OLS on the restricted model and save the residuals $\breve{u}$
- Regress each of the excluded variables on all of the included variables (q different regressions) and save each set of residuals $\breve{r}_1, \breve{r}_2, \ldots, \breve{r}_q$
- Regress a variable defined to be = 1 on $\breve{r}_1 \breve{u}$, $\breve{r}_2 \breve{u}$, …, $\breve{r}_q \breve{u}$, with no intercept
- The LM statistic is $n - SSR_1$, where SSR_1 is the sum of squared residuals from this final regression

Econometrics V Lecture IV

Testing for Heteroskedasticity

- Essentially want to test H_0: $Var(u|x_1, x_2, \ldots, x_k) = \sigma^2$, which is equivalent to H_0: $E(u^2|x_1, x_2, \ldots, x_k) = E(u^2) = \sigma^2$
- If assume the relationship between u^2 and x_j will be linear, can test as a linear restriction
- So, for $u^2 = \delta_0 + \delta_1 x_1 + \ldots + \delta_k x_k + v$) this means testing H_0: $\delta_1 = \delta_2 = \ldots = \delta_k = 0$

Econometrics V Lecture IV
The Breusch-Pagan Test

- Don't observe the error, but can estimate it with the residuals from the OLS regression
- After regressing the residuals squared on all of the *x*'s, can use the R^2 to form an *F* or *LM* test
- The *F* statistic is just the reported *F* statistic for overall significance of the regression, $F = [R^2/k]/[(1 - R^2)/(n - k - 1)]$, which is distributed $F_{k, n-k-1}$
- The *LM* statistic is $LM = nR^2$, which is distributed χ^2_k

Econometrics V Lecture IV
The White Test

- The Breusch-Pagan test will detect any linear forms of heteroskedasticity
- The White test allows for nonlinearities by using squares and crossproducts of all the *x*'s
- Still just using an F or LM to test whether all the x_j, x_j^2, and $x_j x_h$ are jointly significant
- This can get to be unwieldy pretty

Econometrics V Lecture IV

Alternate form of the White test

- Consider that the fitted values from OLS, $\hat{y}$, are a function of all the x's
- Thus, $\hat{y}^2$ will be a function of the squares and crossproducts and $\hat{y}$ and $\hat{y}^2$ can proxy for all of the x_j, x_j^2, and $x_j x_h$, so
- Regress the residuals squared on $\hat{y}$ and $\hat{y}^2$ and use the R^2 to form an F or LM statistic
- Note only testing for 2 restrictions now

Econometrics V Lecture IV

Weighted Least Squares

- While it's always possible to estimate robust standard errors for OLS estimates, if we know something about the specific form of the heteroskedasticity, we can obtain more efficient estimates than OLS
- The basic idea is going to be to transform the model into one that has homoskedastic errors – called weighted least squares

Econometrics V Lecture IV

Case of form being known up to a multiplicative constant

- Suppose the heteroskedasticity can be modeled as $\text{Var}(u|\boldsymbol{x}) = \sigma^2 h(\boldsymbol{x})$, where the trick is to figure out what $h(\boldsymbol{x}) \equiv h_i$ looks like
- $E(u_i/\sqrt{h_i}|\boldsymbol{x}) = 0$, because h_i is only a function of $\boldsymbol{x}$, and $\text{Var}(u_i/\sqrt{h_i}|\boldsymbol{x}) = \sigma^2$, because we know $\text{Var}(u|\boldsymbol{x}) = \sigma^2 h_i$
- So, if we divided our whole equation by $\sqrt{h_i}$ we would have a model where the error is homoskedastic

Econometrics V Lecture IV
Generalized Least Squares

- Estimating the transformed equation by OLS is an example of generalized least squares (GLS)
- GLS will be BLUE in this case
- GLS is a weighted least squares (WLS) procedure where each squared residual is weighted by the inverse of $\text{Var}(u_i|\boldsymbol{x}_i)$

Econometrics V Lecture IV

Weighted Least Squares

- While it is intuitive to see why performing OLS on a transformed equation is appropriate, it can be tedious to do the transformation
- Weighted least squares is a way of getting the same thing, without the transformation
- Idea is to minimize the weighted sum of squares (weighted by $1/h_i$)

Econometrics V Lecture IV

More on WLS

- WLS is great if we know what Var($u_i|\boldsymbol{x}_i$) looks like
- In most cases, won't know form of heteroskedasticity
- Example where do is if data is aggregated, but model is individual level
- Want to weight each aggregate observation by the inverse of the number of individuals

Econometrics V Lecture IV

Feasible GLS

- More typical is the case where you don't know the form of the heteroskedasticity
- In this case, you need to estimate $h(\boldsymbol{x}_i)$
- Typically, we start with the assumption of a fairly flexible model, such as
- $\text{Var}(u|\boldsymbol{x}) = \sigma^2\exp(\delta_0 + \delta_1 x_1 + \ldots + \delta_k x_k)$
- Since we don't know the δ, must estimate

Econometrics V Lecture IV

Feasible GLS (continued)

- Our assumption implies that $u^2 = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + \ldots + \delta_k x_k) v$
- Where $E(v|\boldsymbol{x}) = 1$, then if $E(v) = 1$
- $\ln(u^2) = \alpha_0 + \delta_1 x_1 + \ldots + \delta_k x_k + e$
- Where $E(e) = 1$ and e is independent of $\boldsymbol{x}$
- Now, we know that $\hat{u}$ is an estimate of u, so we can estimate this by OLS

Econometrics V Lecture IV
Feasible GLS (continued)

- Now, an estimate of h is obtained as $\hat{h} = \exp(\hat{g})$, and the inverse of this is our weight
- So, what did we do?
- Run the original OLS model, save the residuals, $\hat{u}$, square them and take the log
- Regress $\ln(\hat{u}^2)$ on all of the independent variables and get the fitted values, $\hat{g}$
- Do WLS using $1/\exp(\hat{g})$ as the weight

Econometrics V Lecture IV
WLS Wrapup

- When doing F tests with WLS, form the weights from the unrestricted model and use those weights to do WLS on the restricted model as well as the unrestricted model
- Remember we are using WLS just for efficiency – OLS is still unbiased & consistent
- Estimates will still be different due to sampling error, but if they are very different then it's likely that some other Gauss-Markov assumption is false

Multiple Regression Analysis

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + u$
- Specification and Data Problems

Econometrics V Lecture IV

Functional Form

- We've seen that a linear regression can really fit nonlinear relationships
- Can use logs on RHS, LHS or both
- Can use quadratic forms of *x*'s
- Can use interactions of *x*'s
- How do we know if we've gotten the right functional form for our model?

Econometrics V Lecture IV

Functional Form (continued)

- First, use economic theory to guide you
- Think about the interpretation
- Does it make more sense for x to affect y in percentage (use logs) or absolute terms?
- Does it make more sense for the derivative of x_1 to vary with x_1 (quadratic) or with x_2 (interactions) or to be fixed?

Econometrics V Lecture IV

Functional Form (continued)

- We already know how to test joint exclusion restrictions to see if higher order terms or interactions belong in the model
- It can be tedious to add and test extra terms, plus may find a square term matters when really using logs would be even better
- A test of functional form is Ramsey's regression specification error test (RESET)

Econometrics V Lecture IV
Ramsey's RESET

- RESET relies on a trick similar to the special form of the White test
- Instead of adding functions of the *x*'s directly, we add and test functions of $\hat{y}$
- So, estimate $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_1 \hat{y}^3 + error$ and test
- H_0: $\delta_1 = 0$, $\delta_2 = 0$ using $F \sim F_{2,n-k-3}$ or $LM \sim \chi^2_2$

Econometrics V Lecture IV

No nested Alternative Tests

- If the models have the same dependent variables, but nonnested *x*'s could still just make a giant model with the *x*'s from both and test joint exclusion restrictions that lead to one model or the other
- An alternative, the Davidson-MacKinnon test, uses $\hat{y}$ from one model as regressor in the second model and tests for significance

Econometrics V Lecture IV
Nonnested Alternatives (cont)

- More difficult if one model uses *y* and the other uses ln(*y*)
- Can follow same basic logic and transform predicted ln(*y*) to get $\hat{y}$ for the second step
- In any case, Davidson-MacKinnon test may reject neither or both models rather than clearly preferring one specification

Econometrics V Lecture IV

Proxy Variables

- What if model is misspecified because no data is available on an important *x* variable?
- It may be possible to avoid omitted variable bias by using a proxy variable
- A proxy variable must be related to the unobservable variable – for example: $x_3^* = \delta_0 + \delta_3 x_3 + v_3$, where * implies unobserved
- Now suppose we just substitute x_3 for x_3^*

Econometrics V Lecture IV

Proxy Variables (continued)

- What do we need for for this solution to give us consistent estimates of β_1 and β_2?
- $E(x_3^* \mid x_1, x_2, x_3) = E(x_3^* \mid x_3) = \delta_0 + \delta_3 x_3$
- That is, u is uncorrelated with x_1, x_2 and x_3^* and v_3 is uncorrelated with x_1, x_2 and x_3
- So really running $y = (\beta_0 + \beta_3\delta_0) + \beta_1 x_1 + \beta_2 x_2 + \beta_3\delta_3 x_3 + (u + \beta_3 v_3)$ and have just redefined intercept, error term x_3 coefficient

Econometrics V Lecture IV

Proxy Variables (continued)

- Without out assumptions, can end up with biased estimates
- Say $x_3^* = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + v_3$
- Then really running $y = (\beta_0 + \beta_3\delta_0) + (\beta_1 + \beta_3\delta_1)\, x_1 + (\beta_2 + \beta_3\delta_2)\, x_2 + \beta_3\delta_3 x_3 + (u + \beta_3 v_3)$
- Bias will depend on signs of β_3 and δ_j
- This bias may still be smaller than omitted variable bias, though

Econometrics V Lecture IV

Lagged Dependent Variables

- What if there are unobserved variables, and you can't find reasonable proxy variables?
- May be possible to include a lagged dependent variable to account for omitted variables that contribute to both past and current levels of *y*
- Obviously, you must think past and current *y* are related for this to make sense

Econometrics V Lecture IV

Measurement Error

- Sometimes we have the variable we want, but we think it is measured with error
- Examples: A survey asks how many hours did you work over the last year, or how many weeks you used child care when your child was young
- Measurement error in *y* different from measurement error in *x*

Econometrics V Lecture IV

Measurement Error in a Dependent Variable

- Define measurement error as $e_0 = y - y^*$
- Thus, really estimating $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u + e_0$
- When will OLS produce unbiased results?
- If e_0 and x_j, u are uncorrelated is unbiased
- If $E(e_0) \neq 0$ then β_0 will be biased, though
- While unbiased, larger variances than with no measurement error

Econometrics V Lecture IV

Measurement Error in an Explanatory Variable

- Define measurement error as $e_1 = x_1 - x_1^*$
- Assume $E(e_1) = 0$, $E(y| x_1^*, x_1) = E(y| x_1^*)$
- Really estimating $y = \beta_0 + \beta_1 x_1 + (u - \beta_1 e_1)$
- The effect of measurement error on OLS estimates depends on our assumption about the correlation between e_1 and x_1
- Suppose $Cov(x_1, e_1) = 0$
- OLS remains unbiased, variances larger

Econometrics V Lecture IV

Measurement Error in an Explanatory Variable (cont)

- Suppose Cov(x_1^*, e_1) = 0, known as the classical errors-in-variables assumption, then
- Cov(x_1, e_1) = E($x_1 e_1$) = E($x_1^* e_1$) + E(e_1^2) = 0 + σ_e^2
- x_1 is correlated with the error so estimate is biased

$$\text{plim}\left(\hat{\beta}_1\right) = \beta_1 + \frac{Cov\left(x_1, u - \beta_1 e_1\right)}{Var\left(x_1\right)} = \beta_1 - \frac{\beta_1 \sigma_e^2}{\sigma_{x^*}^2 + \sigma_e^2}$$

$$= \beta_1\left(1 - \frac{\sigma_e^2}{\sigma_{x^*}^2 + \sigma_e^2}\right) = \beta_1\left(\frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_e^2}\right)$$

Econometrics V Lecture IV

Measurement Error in an Explanatory Variable (cont)

- Notice that the multiplicative error is just $\text{Var}(x_1^*)/\text{Var}(x_1)$
- Since $\text{Var}(x_1^*)/\text{Var}(x_1) < 1$, the estimate is biased toward zero – called attenuation bias
- It's more complicated with a multiple regression, but can still expect attenuation bias with classical errors in variables

Econometrics V Lecture IV

Missing Data – Is it a Problem?

- If any observation is missing data on one of the variables in the model, it can't be used
- If data is missing at random, using a sample restricted to observations with no missing values will be fine
- A problem can arise if the data is missing systematically – say high income individuals refuse to provide income data

Econometrics V Lecture IV

Nonrandom Samples

- If the sample is chosen on the basis of an *x* variable, then estimates are unbiased
- If the sample is chosen on the basis of the *y* variable, then we have sample selection bias
- Sample selection can be more subtle
- Say looking at wages for workers – since people choose to work this isn't the same as wage offers

Econometrics V Lecture IV

Outliers

- Sometimes an individual observation can be very different from the others, and can have a large effect on the outcome
- Sometimes this outlier will simply be do to errors in data entry – one reason why looking at summary statistics is important
- Sometimes the observation will just truly be very different from the others

Econometrics V Lecture IV

Outliers (continued)

- Not unreasonable to fix observations where it's clear there was just an extra zero entered or left off, etc.
- Not unreasonable to drop observations that appear to be extreme outliers, although readers may prefer to see estimates with and without the outliers
- Can use GRETL to investigate outliers