Multiple Regression Analysis

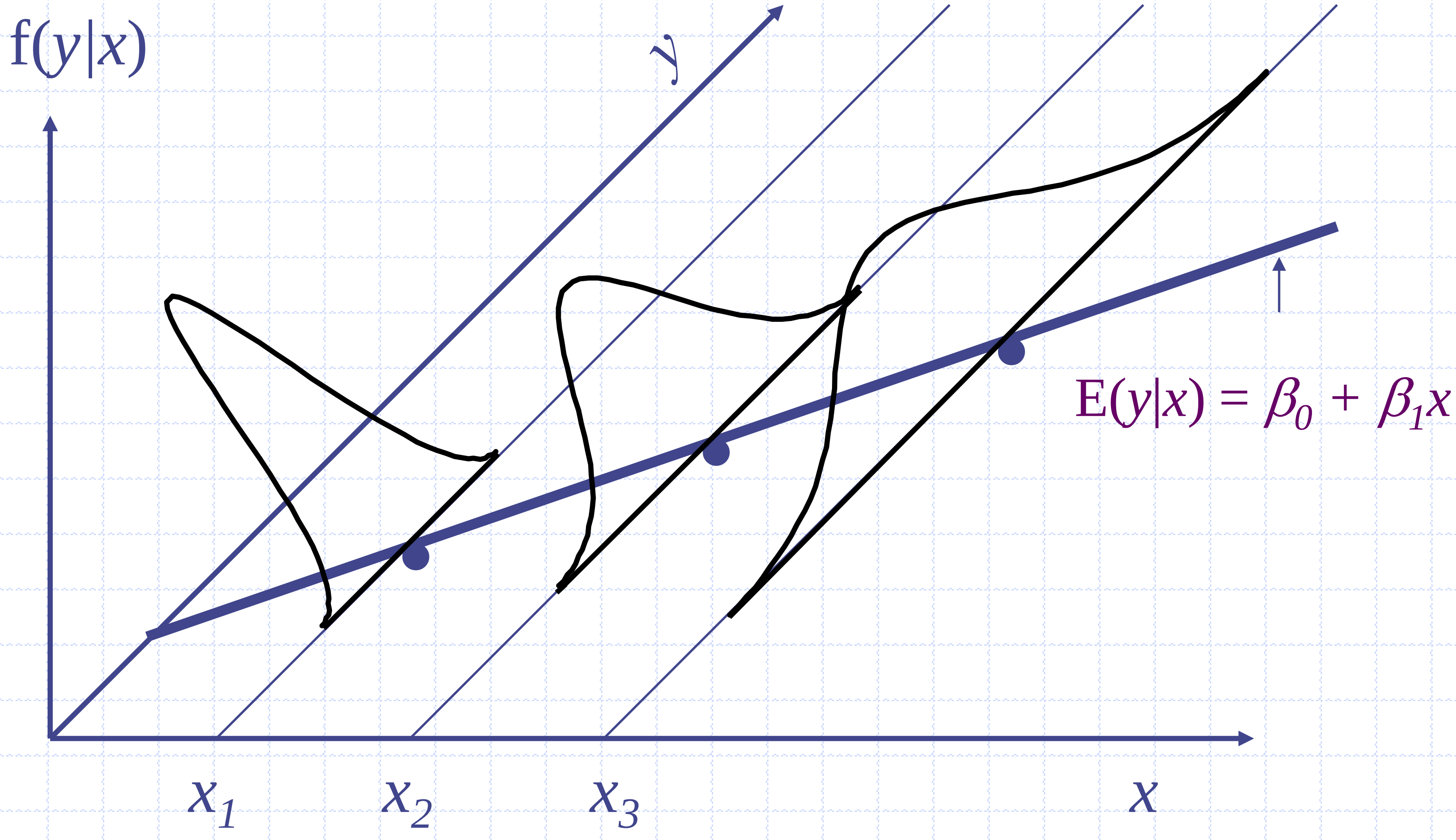
- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$

- 6. Heteroskedasticity

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



Why Worry About Heteroskedasticity?

- OLS is still unbiased and consistent, even if we do not assume homoskedasticity
- The standard errors of the estimates <u>are</u> 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

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

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

Variance with Heteroskedasticity

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

$$\hat{Var}\left(\hat{\beta}_j\right) = \frac{\sum \hat{r}_{ij}\hat{u}_i^2}{SST_j^2}$$, 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

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 $n \longrightarrow \infty$ it's all the same, though

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
- In Stata, robust standard errors are easily obtained using the robust option of reg

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 $\breve{r}_1 \breve{u}, \breve{r}_2 \breve{u}, \ldots, \breve{r}_q \breve{u}$, with <u>no</u> intercept
- The LM statistic is $n - \mathrm{SSR}_1$, where SSR_1 is the sum of squared residuals from this final regression

Testing for Heteroskedasticity

- Essentially want to test H_0: $\text{Var}(u|x_1, x_2, \ldots, x_k) = \sigma^2$, 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: $\delta_1 = \delta_2 = \ldots = \delta_k = 0$

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

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 quickly

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

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

Case of form being known up to a multiplicative constant

- Suppose the heteroskedasticity can be modeled as $\mathrm{Var}(u|\boldsymbol{x}) = \sigma^2 h(\boldsymbol{x})$, 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 $\mathrm{Var}(u_i/\sqrt{h_i}|\boldsymbol{x}) = \sigma^2$, because we know $\mathrm{Var}(u|\boldsymbol{x}) = \sigma^2 h_i$
- So, if we divided our whole equation by $\sqrt{h_i}$ we would have a model where the error is homoskedastic

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 $\mathrm{Var}(u_i|\boldsymbol{x}_i)$

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 $1/h_i$)

More on WLS

- WLS is great if we know what Var($u_i|\mathbf{x}_i$) 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

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
- $\mathrm{Var}(u|\boldsymbol{x}) = \sigma^2\exp(\delta_0 + \delta_1 x_1 + \ldots + \delta_k x_k)$
- Since we don't know the δ, must estimate

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

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

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