Multivariate Regression: Interpretation and inference

EC420 MSU

Justin Kirkpatrick

Last updated February 14, 2021

This lecture

Goal:

1. Review multiple regression and "partialling out"
2. Review single variable inference
   - $Var(\hat{\beta}_1)$
   - SLR.1-SLR.4 + SLR.5
3. Extend to $Var(\hat{\beta}_j)$
   - MLR.1-MLR.4
   - MLR.5
   - **MLR.6**
4. Multicolinearity
5. OLS is **B.L.U.E.**
6. Heteroskedasticity

Multiple regression was estimating a PRF:

$$E[Y|X_1, X_2, \cdots, X_k] = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

And a SRF:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_k x_{i,k}$$

Last time...

Extending the single variable regression, we discussed "partialling out":

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{u} \text{ (the tilde here means "bad")}$$

Where the correct β_1 could be obtained if we "partialled out":

$$x_1 = \delta_0 + \delta_1 x_2 + v$$

$$\hat{v} = x_1 - \hat{\delta}_0 - \hat{\delta}_1 x_2$$

and regressing on the residual:

$$y = \beta_0 + \beta_1 \hat{v} + u$$

Yields an unbiased estimate $\hat{\beta}_1$ of the effect of a change in x_1 on the expectation of Y, *ceteris paribus*.

This tells us that OLS on two variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

is equivalent to "partialing out" each one:

$$y = \beta_0 + \beta_1 \hat{v} + \beta_2 \hat{w} + u$$

Where:

$$x_1 = \delta_0 + \delta_1 x_2 + v$$

and

$$\hat{v} = x_1 - \hat{\delta}_0 - \hat{\delta}_1 x_2$$

$\hat{v}$ is correlated with x_1, but is not correlated with x_2

$$x_2 = \gamma_0 + \gamma_1 x_1 + w$$

and

$$\hat{w} = x_2 - \hat{\gamma}_0 - \hat{\gamma}_1 x_1$$

$\hat{w}$ is correlated with x_2, but is not correlated with x_1

Using our def. of $\hat{\beta} = \frac{\widehat{Cov}(X,Y)}{\widehat{Var}(X)}$

The previous slide means:

$$\hat{\beta}_1 = \frac{\widehat{Cov}(X_1, Y)}{\widehat{Var}(X_1)} = \frac{\widehat{Cov}(\hat{v}, Y)}{\widehat{Var}(\hat{v})}$$

and naturally:

$$\hat{\beta}_2 = \frac{\widehat{Cov}(X_2, Y)}{\widehat{Var}(X_2)} = \frac{\widehat{Cov}(\hat{w}, Y)}{\widehat{Var}(\hat{w})}$$

- The residuals of a first-stage regression of x_1 on x_2 and vice versa.

Last time...

MICHIGAN STATE UNIVERSITY

Our R^2 measure still worked:

$$R^2 = \frac{SSE}{SST} \in [0,1]$$

Where:

- SSE = Sum of Squared Explained = $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- SST = Sum of Squared Total = $\sum_{i=1}^{n}(y_i - \bar{y})^2$
- SSR = Sum of Squared Residual = $\sum_{i=1}^{n}(\hat{u}_i - \bar{u})^2 = \sum_{i=1}^{n}\hat{u}_i^2$

And since $SSE + SSR = SST$, then $R^2 = \frac{SSE}{SST} = \left(1 - \frac{SSR}{SST}\right)$

R^2 gives us the fraction of total variance *explained by the model*

Gauss-Markov Assumptions for **single** variable x

SLR.1: In the population, y is a linear function of the parameters, x, and u:
$y = \beta_0 + \beta_1 x + u$

SLR.2: the sample $(y_i, x_i) : i = 1, 2, \cdots, n$ follows the population model and are independent.

SLR.3: "Sample Variation in the Explanatory (X) Variable". That is, x_i is not the same for all i's.

SLR.4: "Zero conditional mean". $E[u|x] = 0$

SLR.1-SLR.4 $\Rightarrow$ $\hat{\beta}$ is unbiased estimate of β.

SLR.5: $Var[u|x] = \sigma_u^2$ for all x. (*homoskedasticity*)

SLR.5 $\Rightarrow$ $Var(\hat{\beta}) = \frac{\sigma_u^2}{SST_x}$

MICHIGAN STATE UNIVERSITY

$Var(\hat{\beta})$ for single-variable regression:

We formulated β up to this point earlier:

$$\hat{\beta}_1 = \beta_1 + \frac{\sum((x_i - \bar{x})u_i)}{SST_x}$$

And then took the variance of this, noting that $Var(\beta_1) = 0$ because it is a (constant) population parameter.

$$Var(\hat{\beta}_1) = \frac{1}{SST_x^2} \times Var\left[\sum(x_i - \bar{x})u_i\right] = \frac{SST_x}{SST_x^2}\sigma_u^2 = \frac{1}{SST_x}\sigma_u^2$$

MICHIGAN STATE UNIVERSITY

And we could estimate σ_u^2:

$$\hat{\sigma}_u^2 = \frac{1}{(N-2)} \sum_{i=1}^{N} \hat{u}_i^2 = \frac{SSR}{N-2}$$

The $N-2$ is because we lost two *degrees of freedom* due to the two restrictions:

- $\sum \hat{u} = 0$
- $\sum \hat{u}_i x_i = 0$

Variance of estimators in multiple regression

Let's start by looking at σ_u^2

In multiple regression, we have more restrictions:

- $\sum \hat{u} = 0$
- $\sum \hat{u}_i x_{i,1} = 0$
- $\sum \hat{u}_i x_{i,2} = 0$
- $\sum \hat{u}_i x_{i,\ldots} = 0$

One for each β. So, when we have $\{\beta_0, \beta_1, \beta_2\}$, we lose 3 degrees of freedom:

$$\hat{\sigma}_u^2 = \frac{1}{(N-3)} \sum_{i=1}^{N} \hat{u}_i^2 = \frac{SSR}{N-3}$$

Generalizing to K variables

- We always count 1 for β_0
- We call the number of X's K
- We would use $\frac{1}{N-K-1}$ if there are K x's.
- N is the number of observations in our regression.

If $\hat{\beta} = \{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\}$

- Then we have $N - 2 - 1 = N - 3$ degrees of freedom

If $\hat{\beta} = \{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3\}$

- Then we have $N - 3 - 1 = N - 4$ degrees of freedom

Since

$$\widehat{Var}(\hat{\beta}) = \frac{\hat{\sigma}_u^2}{SST_x}$$

and

$$\hat{\sigma}_u^2 = \frac{SSR}{N-K-1}$$

More X's (more regressors) means the denominator on $\hat{\sigma}_u^2$ gets smaller

$\Rightarrow$ $\hat{\sigma}_u^2$ gets **larger**

$\Rightarrow$ $\widehat{Var}(\hat{\beta})$ gets **larger**

And thus our confidence intervals get larger, rejection region gets smaller, and we lose *precision*

So we know how to calculate our multivariate $\hat{\sigma}_u^2$.

What about the rest of $\frac{\hat{\sigma}_u^2}{SST_x}$? What is SST_x?

- Before, we had one x, so SST_x was straightforward.
- Now, we have 2 or more x's.

Each x has its own SST_x:

$$SST_{x_k} = \sum_{i=1}^{n} (x_{i,k} - \bar{x}_k)^2$$

Note that we are summing the x_k over all i.

SST_x example

The data:

x1	x2
1	-2
5	0
6	2

$$\bar{x}_1 = 4$$
$$\bar{x}_2 = 0$$

Which results in SST_x's of:

SST_{x_1} = 14
SST_{x_2} = 8

We need to make one more adjustment

We need to account for how unique the variance in each of the x_k's is.

- Imagine if we had two x's: x_j, x_k, but they were *almost* always the same number.
  - Think: temperature and rainfall.
- The estimates of the corresponding β's: β_j and β_k, should have a lot of variance to them because we aren't sure which is *actually* explaining the variation in y.

So, we are going to weight each SST_{x_j} by $(1 - R_j^2)$

Where R_j^2 is the R^2 of the regression $x_j = \delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k + v$.

That is, we will weight it by "how well is this variable, x_j, explained by all the other variables"

So with that weighted SST_{x_j}, $Var(\beta_j)$ is:

$$Var(\hat{\beta}_j) = \frac{\sigma_u^2}{SST_{x_j}(1 - R_j^2)}$$

And we can estimate this easily:

$$\widehat{Var}(\hat{\beta}_j) = \frac{\hat{\sigma}_u^2}{SST_{x_j}(1 - \hat{R}_j^2)}$$

When R_j^2 is high:

- Then x_j is explained almost completely by $x_1, \cdots, x_k$ (the other x's)
- R_j^2 is very high
- $(1 - R_j^2)$ is very small, close to 0
- $SST_{x_j}(1 - R_j^2)$ is very small, close to 0
- And thus, $Var(\hat{\beta}_j) = \frac{\hat{\sigma}_u^2}{SST_{x_j}(1-R_j^2)}$ is **very high** when R_j^2 is very high.
    - It is division by a small number near 0

What if $x_j = x_k$?

- What is the R_j^2?
  - What is the R^2 of the regression: $x_j = \beta_0 + \beta_1 x_k$?

If $R_j^2 = 1$, what is $Var(\hat{\beta}_j)$?

A problem, that's what it is.

When two x's are perfectly correlated, you have *multicolinearity*

- **Perfect** correlation occurs when $x_j = c + bx_k$, an *affine* transformation
- Degrees farenheit and degrees celsius are a perfect example

Variance of the estimator

X degrees Farenheit to Y degrees Celsisus conversion:

$$(X^{\circ}F - 32) \times \frac{5}{9} = Y^{\circ}C$$

```
##
## Call:
## lm(formula = C ~ F, data = df)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -4.210e-14 -7.180e-16  7.020e-16  1.401e-15  6.060e-15
##
## Coefficients:
##              Estimate Std. Error    t value Pr(>|t|)
## (Intercept) -1.778e+01  8.305e-16 -2.141e+16   <2e-16 ***
## F            5.556e-01  1.196e-17  4.644e+16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.596e-15 on 119 degrees of freedom
## Multiple R-squared:      1,    Adjusted R-squared:      1
## F-statistic: 2.157e+33 on 1 and 119 DF,  p-value: < 2.2e-16
```

Regression of C on F (perfect fit, **note the R^2**)

Variance of the estimator

MICHIGAN STATE UNIVERSITY

y	degC	degF
2	24	75.2
1	35	95.0
3	33	91.4
4	30	86.0
1	30	86.0

This matrix is not *full rank*

So the regression doesn't go so well

```r
summary(lm(y ~ degC + degF, degdf))
```

```
##
## Call:
## lm(formula = y ~ degC + degF, data = degdf)
##
## Residuals:
##       1       2       3       4       5
## -0.4220 -1.0405  0.8902  1.7861 -1.2139
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.25434    5.50894   0.591    0.596
## degC        -0.03468    0.17987  -0.193    0.859
## degF              NA         NA      NA       NA
##
## Residual standard error: 1.496 on 3 degrees of freedom
## Multiple R-squared:  0.01224,    Adjusted R-squared:  -0.317
## F-statistic: 0.03718 on 1 and 3 DF,  p-value: 0.8594
```

We have a bit of a problem when two of our x's are perfectly correlated.

What do we do about this?

In practice, we omit one of the x's

Won't this bias the result?

- Yes and no. If they are perfectly correlated, then *one* of the $\boldsymbol{x}$'s explains exactly what *both* $\boldsymbol{x}$'s could explain.

- But we will never, ever be able to tell *which* one is causal.
  - Think of the temperature example.
  - Can we tell if degrees F has an effect while degrees C doesn't?
  - Of course not!

Gauss-Markov Regression Assumptions:

MLR.1	The population, y is a linear function of the parameters x and u: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$
MLR.2	The sample $(y_i, x_i) : i = 1, 2, \cdots, n$ follows the population model and are independent
MLR.3	No multicolinearity / "full rank": x_j is not a linear transformation of x_k for all j, k.
MLR.4	Zero conditional mean: $E[u\|x_1, x_2, \cdots, x_k] = 0$ for all x.
MLR.5	$Var[u\|x_1, \cdots, x_k] = \sigma_u^2$ for all x.

Multiple regression assumptions

A neat thing happens when assumptions 1-5 hold

OLS is B.L.U.E.

- **B**est
  - Has the lowest variance
- **L**inear
  - β is a linear function of the data (e.g. it uses $Cov(Y, X)$)
- **U**nbiased
  - Is unbiased (showed for single; holds for multiple)
- **E**stimator

Of all linear, unbiased estimators, OLS is the most efficient

MICHIGAN STATE UNIVERSITY

Remember what we needed for inference

- $E[\hat{\beta}] = \beta$
- $Var(\hat{\beta})$
- That $\hat{\beta} \sim N(\beta, Var(\hat{\beta}))$

How do we know it's Normal?

- We will need more assumptions
  - Chapter 5 has weaker assumptions with a similar result

Assumption MLR.6: Normality of u

We can assume a normal distribution for the OLS estimator, $\hat{\beta}$, by assuming that the errors, u, are normally distributed in the population.

Assume:

$$u|x_1, x_2, \cdots, x_k \sim N(0, \sigma_u^2)$$

Then:

$$y|x_1, \cdots, x_k \sim N(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k, \sigma_u^2)$$

Note that this is the distribution of y conditional on the x's. All of the random variation comes from u. x's conditionally shift the mean deterministically.

I'm adding the subscript u to σ_u^2 for emphasis, but since u is the only source of random variation once we condition on x's, it is implied to be the only σ^2.

How do we get from normal u's to normal β?

Define $\hat{v}_j$ to be the residual of a regression of x_j on all other x's. In a two variable (x_j, x_k) example for observation i:

$$x_{i,j} = \hat{\delta}_0 + \hat{\delta}_1 x_{i,k} + \hat{v}_{i,j}$$

Then $\hat{\beta}_j$ is:

$$\hat{\beta}_j = \frac{\widehat{Cov}(\hat{v}, y)}{\widehat{Var}(\hat{v})} = \frac{\sum_{i=1}^{n} \hat{v}_{ij} y_i}{\sum_{s=1}^{n} \hat{v}_{sj}^2} = \sum_{i=1}^{n} w_{ij} u_i$$

Where:

$$w_{ij} = \frac{\hat{v}_{ij}}{\sum_{s=1}^{n} \hat{v}_{sj}^2}$$

MICHIGAN STATE UNIVERSITY

And a linear combination of normals is....normal!

(See Stats Review notes. Told you that property would come in handy).

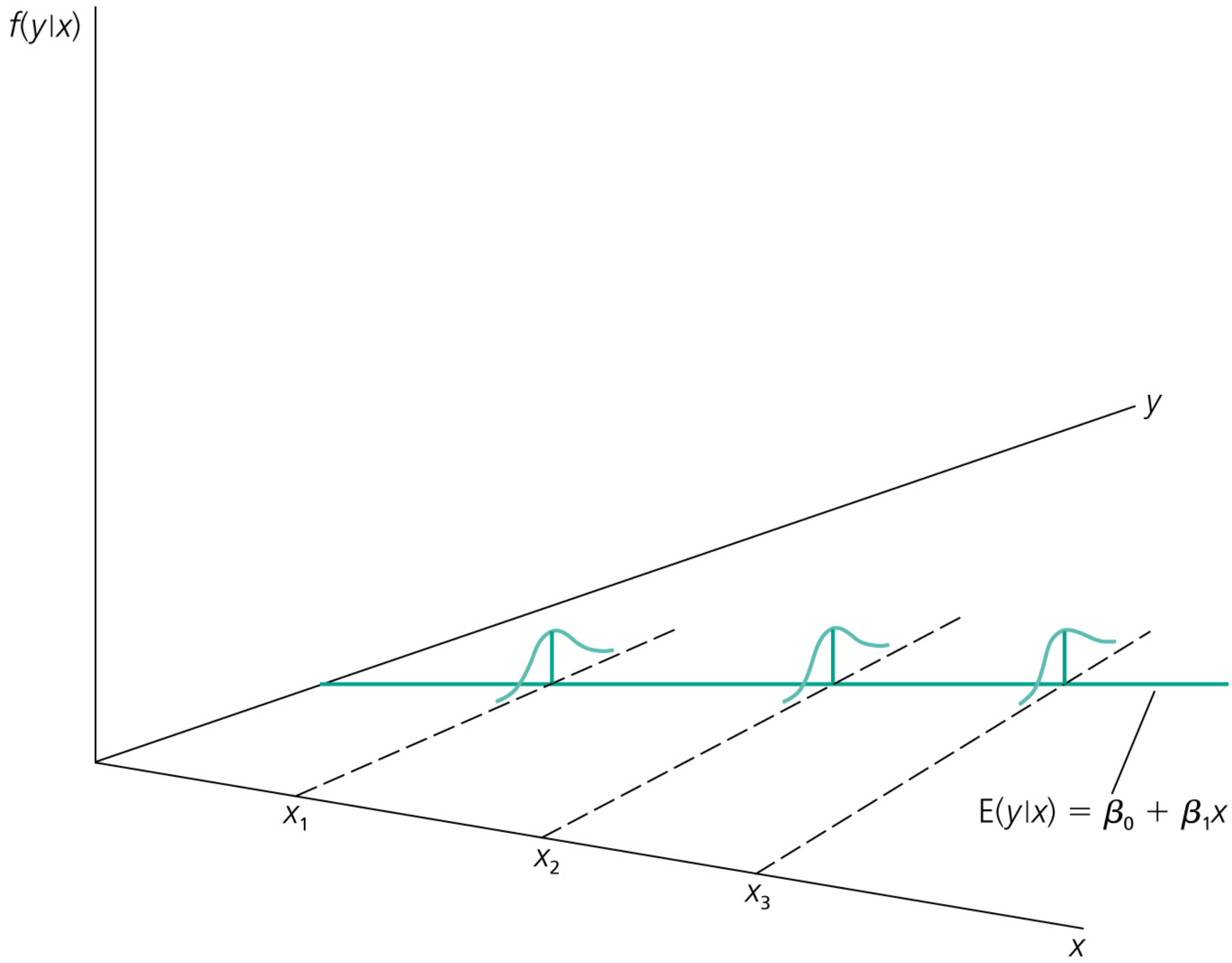
So now we know :

- $E[\hat{\beta}]$
- $Var(\hat{\beta})$
- That $\hat{\beta}$ really is normally distributed

That's what we need to start testing things!

$$\frac{\hat{\beta}-\beta}{se(\hat{\beta})} \sim N(0,1) \quad \text{and} \quad \frac{\hat{\beta}-\beta}{\hat{se}(\hat{\beta})} \sim t_{N-K-1}$$

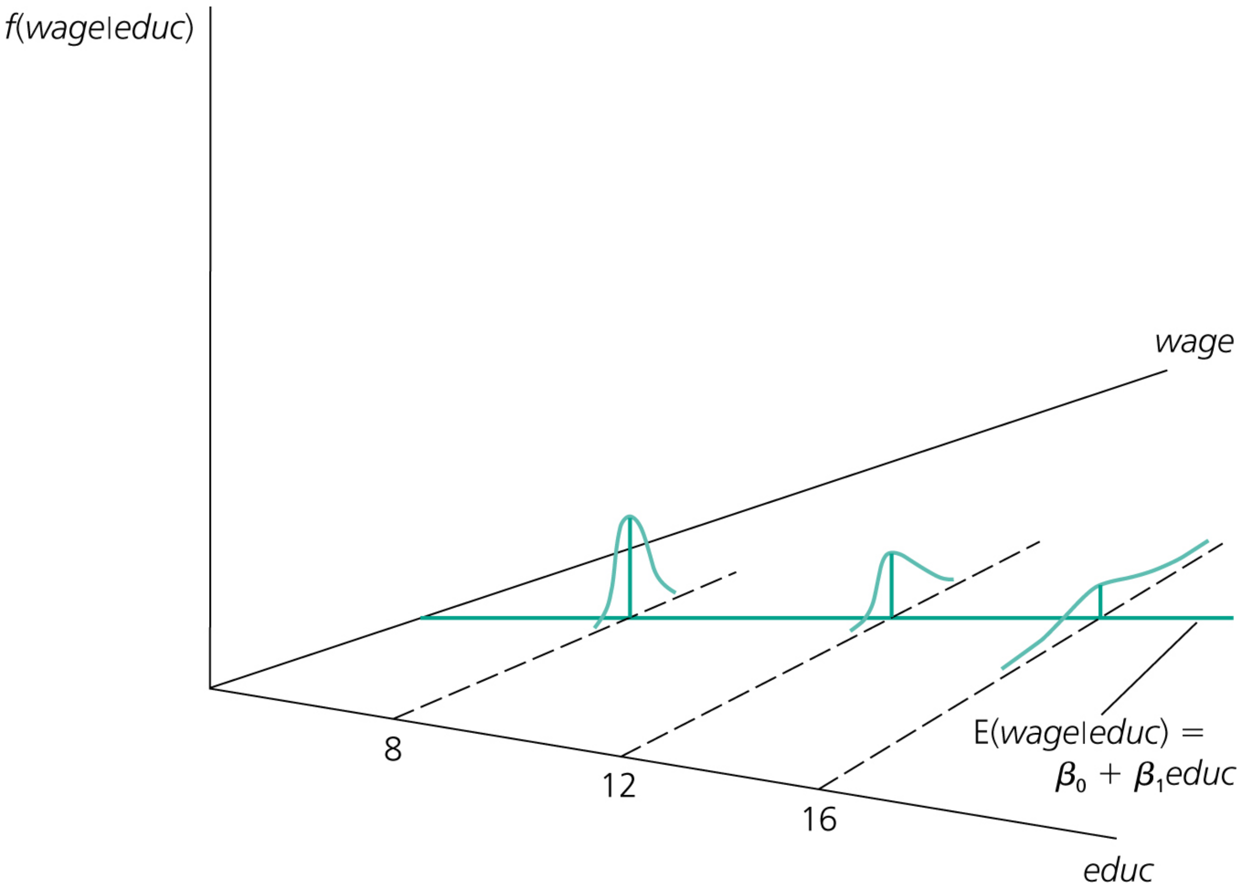
Now is a good time to revisit MLR.5, homoskedasticity:



Homoskedasticity (from Wooldridge)

Homoskedasticity

And what to do about heteroskedasticity:



Heteroskedasticity (from Wooldridge)

In practice, we have a very useful method of "correcting" for heteroskedasticity called "robust standard errors"

- Eicker-Huber-White Heterskedasticity-Consistent (HC) errors (1980)

In R, we can compute these errors fairly easily

We'll see in a few slides.

It comes at a cost, though: it inflates errors (make larger)

- Less likely to be "significant" (reject H_0) even if there is evidence to reject H_0.
- That's what it's supposed to do **if** there is heteroskedasticity
- But if there **isn't** heteroskedasticity, you are wasting some power.

Heteroskedasticity-robust standard errors: how do they work?

The problem is that x_j may be correlated with u and thus $\sigma_m^2 \neq \sigma_n^2$ - there is no common, single σ^2.

In the **single variable regression case**, we would account for this:

$$\widehat{Var}(\hat{\beta}_1) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2 \hat{u}_i^2}{(SST_x)^2}$$

Note that we have squared the sum-of-squares total in the denominator. The numerator looks a little like covariance, but it's more like the covariance of squared terms.

In multiple regression, things get complicated:

$$\hat{Var}(\hat{\beta}_j) = \frac{\sum_{i=1}^{N} \hat{v}_{ij}^2 \hat{u}_i^2}{(SSR_j)^2}$$

Which looks like the multivariate variance error, but with the extra SSR_j in the denominator, and $\hat{v}$, the residual from x_j on x_i

Homoskedasticity

If we adjust with heteroskedasticity-consistent errors (HC)

then we can relax MLR.5 and still have a *valid* estimate of the variance of $\hat{\beta}$.

Note that **heteroskedasticity-consistent errors** do not **ever** affect the point estimate of $\hat{\beta}$.

- The point estimates remain the same, but the error (and thus the significance) changes.
    - "Point estimate" refers to the value of $\hat{\beta}$, regardless of the variance.

Heteroskedasticity-consistent errors in R

- `install.packages(c('sandwich','lmtest'))`
- `require(sandwich)`
- `require(lmtest)`

- `myOLS = lm(Y ~ X1 + X2, df)`

- `coeftest(myOLS, vcov = vcovHC(myOLS, 'HC1'))`
  - `myOLS` is your linear regression object
  - `vcov` stands for "variance-covariance"
  - The `HC1` gives a specific type of HC errors
  - It is identical to the `, robust` errors in Stata.

If you do not adjust your standard errors, you must justify exactly why you are assuming homoskedasticity.

Homoskedasticity

```
wage2 = wooldridge::wage2
myOLS = lm(wage ~ educ + exper, wage2)
summary(myOLS)
```

```
##
## Call:
## lm(formula = wage ~ educ + exper, data = wage2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -924.38 -252.74  -40.88  198.16 2165.70
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -272.528    107.263  -2.541   0.0112 *
## educ          76.216      6.297  12.104  < 2e-16 ***
## exper         17.638      3.162   5.578 3.18e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 376.3 on 932 degrees of freedom
## Multiple R-squared:  0.1359,    Adjusted R-squared:  0.134
## F-statistic: 73.26 on 2 and 932 DF,  p-value: < 2.2e-16
```

Using HC errors

```
coeftest(myOLS, vcov = vcovHC(myOLS, 'HC1'))
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -272.5279   109.8965 -2.4799   0.01332 *
## educ          76.2164     6.7468 11.2966 < 2.2e-16 ***
## exper         17.6378     3.1126  5.6666 1.941e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- 'HC1' yields Stata-type robust errors
  - If you are planning on taking EC422 with Prof. Imberman, use 'HC1'.

MICHIGAN STATE UNIVERSITY

There are two meanings of the word "robust" in econometrics

- Robust standard errors, which is what we are discussing here
- A "robust" regression is one that is not affected by a particular specification issue
    - When we saw that we could include unrelated x's and not worry about getting bias, our regression was "robust"

Using the `fixest` package

```r
library(fixest)
myFEOLS = feols(wage ~ educ + exper, wage2)
# feols is fixed-effect OLS. We will get to fixed effects soon
summary(myFEOLS, se = 'hetero') # se = 'standard'
```

```
## OLS estimation, Dep. Var.: wage
## Observations: 935
## Standard-errors: Heteroskedasticity-robust
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -272.530   109.9000 -2.4799  0.013319 *
## educ          76.216     6.7468 11.2970 < 2.2e-16 ***
## exper         17.638     3.1126  5.6666  1.94e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 375.69   Adj. R2: 0.134
```

`fixest` package's `feols` lets you list the std. error correction in the `summary(...)` call, which is handy. It does a lot more as well.