Session 5

OLS using Matrix Algebra

Anthony Tay

This Version: 17 Sep 2025

Session 5

- OLS Estimation of the Linear Regression Model
  - Theory expressed using matrix algebra
  - OLS properties
  - Applications

Setting Up The MLR Model

Population: $E(Y \mid X_1, X_2, \dots, X_{K-1}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{K-1} X_{K-1}$

Define $\epsilon = Y - \beta_0 - \beta_1 X_1 - \beta_2 X_2 - \cdots - \beta_{K-1} X_{K-1}$

$X_1, \dots, X_{K-1}$ refer to $K-1$ different regressors, not $K-1$ obs. of a variable X

We have

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{K-1} X_{K-1} + \epsilon\,,\ E(\epsilon \mid X_1, \dots, X_{K-1}) = 0$$

Setting Up The MLR Model

Representative iid sample from the population: $\{Y_i, X_{i1}, X_{i2}, \dots, X_{i,K-1}\}_{i=1}^{n}$

Sample satisfies

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{K-1} X_{i,K-1} + \epsilon_i$$

$$E(\epsilon_i \mid X_{11}, \dots, X_{n1}, X_{12}, \dots, X_{n2}, \dots, X_{1,K-1}, \dots, X_{n,K-1}) = 0$$

$$E(\epsilon_i^2 \mid X_{11}, \dots, X_{n1}, X_{12}, \dots, X_{n2}, \dots, X_{1,K-1}, \dots, X_{n,K-1}) = \sigma_i^2 \text{ or } \sigma^2$$

$$E(\epsilon_i \epsilon_j \mid X_{11}, \dots, X_{n1}, X_{12}, \dots, X_{n2}, \dots, X_{1,K-1}, \dots, X_{n,K-1}) = 0$$

for all $i, j = 1, \dots, n$, $i \neq j$

Setting Up The MLR Model

The equations

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{K-1} X_{i,K-1} + \epsilon_i \quad \text{for all} \quad i = 1, \ldots, n$$

can be written

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,K-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,K-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,K-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{K-1} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\text{i.e.,} \quad y = X\beta + \varepsilon$$

X_{ij} is ith observation of the jth regressor

Setting Up The MLR Model

We can partition the regressor matrix by observation:

$$X = \left[\begin{array}{ccccc} 1 & X_{11} & X_{12} & \cdots & X_{1,K-1} \\ \hline 1 & X_{21} & X_{22} & \cdots & X_{2,K-1} \\ \hline \vdots & \vdots & \vdots & \ddots & \vdots \\ \hline 1 & X_{n1} & X_{n2} & \cdots & X_{n,K-1} \end{array}\right] = \begin{bmatrix} X_{1*} \\ X_{2*} \\ \vdots \\ X_{n*} \end{bmatrix}$$

and write the model as

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_{1*} \\ X_{2*} \\ \vdots \\ X_{n*} \end{bmatrix} \beta + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \quad \text{or} \quad Y_i = X_{i*}\beta + \epsilon_i \,, \; i = 1, 2, ..., n$$

Setting Up The MLR Model

It is sometimes helpful to partition the regressor matrix by variable:

$$X = \left[\begin{array}{c|c|c|c|c} 1 & X_{11} & X_{12} & \cdots & X_{1,K-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,K-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,K-1} \end{array}\right] = \begin{bmatrix} i_n & X_{*1} & X_{*2} & \cdots X_{*,K-1} \end{bmatrix}$$

OLS estimation requires no perfect collinearity in the sample:

$$Xc = c_0 + c_1 X_{*1} + c_2 X_{*2} + \cdots + c_{K-1} X_{*,K-1} = 0_{n\times 1} \iff c = 0_K$$

where $c = \begin{bmatrix} c_0 & \cdots & c_{K-1} \end{bmatrix}^{\mathrm{T}}$. This assumption can also be written:

$$Xc \neq 0_n \quad \Leftrightarrow \quad c \neq 0_K$$

Setting Up The MLR Model

The assumption

$$E(\epsilon_i \mid X_{11}, \ldots, X_{n1}, X_{12}, \ldots, X_{n2}, \ldots, X_{1,K-1}, \ldots, X_{n,K-1}) = 0 \text{ for } i = 1, \ldots, n$$

can be written as

$$\begin{bmatrix} E(\epsilon_1 \mid X) \\ E(\epsilon_2 \mid X) \\ \vdots \\ E(\epsilon_n \mid X) \end{bmatrix} = 0_n \iff E(\varepsilon \mid X) = 0_n$$

Setting Up The MLR Model

The assumptions

$$E(\epsilon_i^2 \mid X_{11}, \ldots, X_{n1}, X_{12}, \ldots, X_{n2}, \ldots, X_{1,K-1}, \ldots, X_{n,K-1}) = \sigma_i^2 \quad \text{for} \quad i = 1, \ldots, n$$

$$E(\epsilon_i \epsilon_j \mid X_{11}, \ldots, X_{n1}, X_{12}, \ldots, X_{n2}, \ldots, X_{1,K-1}, \ldots, X_{n,K-1}) = 0 \quad \text{for} \quad i, j = 1, \ldots, n \, , i \neq j$$

can be written as

$$\begin{bmatrix} E(\epsilon_1^2 \mid X) & E(\epsilon_1 \epsilon_2 \mid X) & \cdots & E(\epsilon_1 \epsilon_n \mid X) \\ E(\epsilon_2 \epsilon_1 \mid X) & E(\epsilon_2^2 \mid X) & \cdots & E(\epsilon_2 \epsilon_n \mid X) \\ \vdots & \vdots & \ddots & \vdots \\ E(\epsilon_n \epsilon_1 \mid X) & E(\epsilon_n \epsilon_2 \mid X) & \cdots & E(\epsilon_n^2 \mid X) \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

or

$$Var(\varepsilon \mid X) = E(\varepsilon \varepsilon^{\mathrm{T}} \mid X) = \mathsf{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$$

Setting Up The MLR Model

If the noise terms are homoskedastic, i.e.,

$$E(\epsilon_i^2 \mid X_{11}, \ldots, X_{n1}, X_{12}, \ldots, X_{n2}, \ldots, X_{1,K-1}, \ldots, X_{n,K-1}) = \sigma^2 \quad \text{for} \quad i = 1, \ldots, n,$$

then we have

$$E(\varepsilon\varepsilon^{\mathrm{T}} \mid X) = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 I_n$$

which we will assume for now

OLS Estimation of MLR

In matrix form, the MLR is

$$Y = X\beta + \varepsilon\,,\ E(\varepsilon \mid X) = 0\,,\ E(\varepsilon\varepsilon^{\mathrm{T}} \mid X) = \sigma^2 I_n$$

For any potential estimator $\hat{\beta}$, define

- Fitted values: $\hat{y} = X\hat{\beta}$
- Residuals: $\hat{\varepsilon} = y - \hat{y} = y - X\hat{\beta}$
- Residual Sum of Squares $RSS(\hat{\beta})$: $\sum_{i=1}^{n} \hat{\epsilon}_i^2 = \hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon}$

OLS Estimation of MLR

We have

$$\begin{aligned} RSS(\hat{\beta}) = \hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon} &= (y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta}) \\ &= (y^{\mathrm{T}} - \hat{\beta}^{\mathrm{T}}X^{\mathrm{T}})(y - X\hat{\beta}) \\ &= y^{\mathrm{T}}y - \hat{\beta}^{\mathrm{T}}X^{\mathrm{T}}y - y^{\mathrm{T}}X\hat{\beta} + \hat{\beta}^{\mathrm{T}}X^{\mathrm{T}}X\hat{\beta} \\ &= y^{\mathrm{T}}y - 2\hat{\beta}^{\mathrm{T}}X^{\mathrm{T}}y + \hat{\beta}^{\mathrm{T}}X^{\mathrm{T}}X\hat{\beta} \end{aligned}$$

OLS:

$$\begin{aligned} \hat{\beta}^{ols} &= \arg\min_{\hat{\beta}} RSS(\hat{\beta}) \\ &= \arg\min_{\hat{\beta}} y^{\mathrm{T}}y - 2\hat{\beta}^{\mathrm{T}}X^{\mathrm{T}}y + \hat{\beta}^{\mathrm{T}}X^{\mathrm{T}}X\hat{\beta} \end{aligned}$$

OLS Estimation of MLR

First and second derivatives are

$$\frac{\partial RSS(\hat{\beta})}{\partial \beta} = -2X^{\mathrm{T}}y + 2X^{\mathrm{T}}X\hat{\beta} \quad \text{and} \quad \frac{\partial^2 RSS(\hat{\beta})}{\partial \beta \, \partial \beta^{\mathrm{T}}} = 2X^{\mathrm{T}}X$$

FOC: $\left.\frac{\partial SSR(\hat{\beta})}{\partial \beta}\right|_{\hat{\beta}^{ols}} = -2X^{\mathrm{T}}y + 2X^{\mathrm{T}}X\hat{\beta}^{ols} = 0_K$ which implies

$$\hat{\beta}^{ols} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$$

- $Xc \neq 0 \Leftrightarrow c \neq 0$ means that $X^{\mathrm{T}}X$ in non-singular, and
- $c^{\mathrm{T}}X^{\mathrm{T}}Xc = (Xc)^{\mathrm{T}}(Xc) > 0$ (Hessian is pos. def.) so $\hat{\beta}^{ols}$ minimizes $SSR(\hat{\beta})$

OLS Estimation of MLR

Another way of expressing $\widehat{\beta}^{ols}$

Writing

$$X = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,K-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,K-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,K-1} \end{bmatrix} = \begin{bmatrix} X_{1*} \\ X_{2*} \\ \vdots \\ X_{n*} \end{bmatrix}$$

which emphasizes observations.

Note X_{i*} is $1 \times K$ vector comprising ith observation of all variables (including '1' for the intercept term)

OLS Estimation of MLR

Then

$$\begin{aligned}
\widehat{\beta}^{ols} &= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y \\
&= \left\{ \begin{bmatrix} X_{1*}^{\mathrm{T}} & X_{2*}^{\mathrm{T}} & \cdots & X_{n*}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} X_{1*} \\ X_{2*} \\ \vdots \\ X_{n*} \end{bmatrix} \right\}^{-1} \begin{bmatrix} X_{1*}^{\mathrm{T}} & X_{2*}^{\mathrm{T}} & \cdots & X_{n*}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \\
&= \underbrace{\left(\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*} \right)^{-1}}_{\text{sum of } k \times k \text{ matrices}} \underbrace{\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} Y_i}_{\text{sum of } k \times 1 \text{ vectors}}
\end{aligned}$$

OLS Estimation of MLR (Quick Summary)

- MLR: $Y = X\beta + \epsilon,\ E(\epsilon \mid X) = 0,\ Var(\epsilon \mid X) = E(\epsilon\epsilon^{\mathrm{T}} \mid X) = \sigma^2 I_n$
- OLS Estimator for β: $\hat{\beta}^{ols} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$
- OLS Fitted values:

$$\hat{y}^{ols} = X\hat{\beta}^{ols} = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y = Py \ \text{ where } \ P = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$$

- OLS Residuals:

$$\begin{aligned}\hat{\epsilon}^{ols} &= y - \hat{y}^{ols} = y - X\hat{\beta}^{ols} = y - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y \\ &= (I_n - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})y = My \ \text{ where } \ M = I_n - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\end{aligned}$$

Where helpful to do so, I will place "ols" marker in subscript instead of superscript

OLS Estimation of MLR (More on the FOC)

The FOC can be written as: $X^{\mathrm{T}}\hat{\epsilon}^{ols} = 0_{K\times 1}$

$$\text{Since} \quad X = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,K-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,K-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,K-1} \end{bmatrix} = \begin{bmatrix} i_n & X_{*1} & X_{*2} & \ldots & X_{*,K-1} \end{bmatrix}$$

$$\text{the FOC says:} \quad X^{\mathrm{T}}\hat{\varepsilon}_{ols} = \begin{bmatrix} i_n^{\mathrm{T}} \\ X_{*1}^{\mathrm{T}} \\ X_{*2}^{\mathrm{T}} \\ \vdots \\ X_{*,K-1}^{\mathrm{T}} \end{bmatrix} \hat{\varepsilon}^{ols} = \begin{bmatrix} i_n^{\mathrm{T}}\hat{\varepsilon}^{ols} \\ X_{*1}^{\mathrm{T}}\hat{\varepsilon}^{ols} \\ X_{*2}^{\mathrm{T}}\hat{\varepsilon}^{ols} \\ \vdots \\ X_{*,K-1}^{\mathrm{T}}\hat{\varepsilon}^{ols} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} \hat{\epsilon}_i^{ols} \\ \sum_{i=1}^{n} X_{i1}\hat{\epsilon}_i^{ols} \\ \sum_{i=1}^{n} X_{i2}\hat{\epsilon}_i^{ols} \\ \vdots \\ \sum_{i=1}^{n} X_{i,K-1}\hat{\epsilon}_i^{ols} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

OLS Estimation of MLR (More on the FOC)

The first line $\sum_{i=1}^{n} \hat{\epsilon}_i^{ols} = 0$ means that $\overline{\hat{\epsilon}^{ols}} = 0$ so

$$\sum_{i=1}^{n} X_{ij}\hat{\epsilon}_i^{ols} = 0 \iff \text{smp. cov.}(X_{ij}, \hat{\epsilon}_i^{ols}) = 0$$

for each variable X_j, $j = 1, \ldots, K-1$

OLS estimator $\hat{\beta}^{ols}$ makes OLS residuals

- mean zero and
- uncorrelated with each regressor

Goodness of Fit

$TSS = FSS + RSS$ decomposition continues to hold in the general MLR case

Let $M_0 = I_n - i_n(i_n^{\mathrm{T}} i_n)^{-1} i_n^{\mathrm{T}} = I_n - (1/n) i_n i_n^{\mathrm{T}}$ which is symmetric and idempotent

- $M_0^{\mathrm{T}} = M_0$ and $M_0 M_0 = M_0$

We have

- $M_0 y = \begin{bmatrix} Y_1 - \overline{Y} \\ Y_2 - \overline{Y} \\ \vdots \\ Y_n - \overline{Y} \end{bmatrix}$ and $y^{\mathrm{T}} M_0^{\mathrm{T}} M_0 y = y^{\mathrm{T}} M_0 y = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$
- $M_0 \hat{\varepsilon}^{ols} = \hat{\varepsilon}^{ols}$

Goodness of Fit

Therefore

$$y = \hat{y}_{ols} + \hat{\varepsilon}_{ols}$$

$$M_0 y = M_0 \hat{y}_{ols} + \hat{\varepsilon}_{ols}$$

$$y^{\mathrm{T}} M_0^{\mathrm{T}} M_0 y = (M_0 \hat{y}_{ols} + \hat{\varepsilon}_{ols})^{\mathrm{T}} (M_0 \hat{y}_{ols} + \hat{\varepsilon}_{ols})$$

$$y^{\mathrm{T}} M_0 y = \hat{y}_{ols}^{\mathrm{T}} M_0 \hat{y}_{ols} + \hat{\varepsilon}_{ols}^{\mathrm{T}} \hat{y}_{ols} + \hat{y}_{ols}^{\mathrm{T}} \hat{\varepsilon}_{ols} + \hat{\varepsilon}_{ols}^{\mathrm{T}} \hat{\varepsilon}_{ols}$$

$$y^{\mathrm{T}} M_0 y = \hat{y}_{ols}^{\mathrm{T}} M_0 \hat{y}_{ols} + \hat{\varepsilon}_{ols}^{\mathrm{T}} \hat{\varepsilon}_{ols}$$

$$TSS = FSS + RSS$$

Goodness of Fit

We can use this to define the goodness-of-fit measure:

$$R^2 = 1 - \frac{RSS}{TSS}$$

and "Adjusted R^2"

$$\text{Adj.-}R^2 = 1 - \frac{\frac{1}{n-K}RSS}{\frac{1}{n-1}TSS} = 1 - \frac{RSS}{TSS}\frac{n-1}{n-K}$$

OLS Estimator Properties: Unbiasedness

- OLS estimator $\widehat{\beta}^{ols}$ is unbiased:

$$\widehat{\beta}^{ols} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}(X\beta + \varepsilon) = \beta + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\varepsilon$$

$$\begin{aligned} E(\widehat{\beta}^{ols} \mid X) &= \beta + E((X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\varepsilon \mid X) \\ &= \beta + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}E(\varepsilon \mid X) \\ &= \beta \end{aligned}$$

$$E(\widehat{\beta}^{ols}) = \beta$$

OLS Estimator Variance (Homoskedasticity Case)

Under homoskedasticity $Var(\varepsilon\varepsilon^{\mathrm{T}} \mid X) = \sigma^2 I_n$, we have

$$\begin{aligned} Var(\hat{\beta}^{ols} \mid X) &= Var(\beta + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\varepsilon \mid X) \\ &= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\, Var(\varepsilon \mid X) X (X^{\mathrm{T}}X)^{-1} \\ &= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}(\sigma^2 I_n) X (X^{\mathrm{T}}X)^{-1} \\ &= \sigma^2 (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}} I_n X (X^{\mathrm{T}}X)^{-1} \\ &= \sigma^2 (X^{\mathrm{T}}X)^{-1} \end{aligned}$$

but must estimate σ^2

OLS Estimator Variance (Homoskedasticity Case)

Unbiased estimator of σ^2 is $\widehat{\sigma^2} = \dfrac{1}{n-K}\sum_{i=1}^{n}\hat{\epsilon}_{i,ols}^2 = \dfrac{\hat{\varepsilon}_{ols}^{\mathrm{T}}\hat{\varepsilon}_{ols}}{n-K}$

Proof: $\hat{\varepsilon}_{ols} = My$ where $M = I_n - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$

note that

- $MX = (I_n - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})X = X - X = 0_{n\times K}$
- M is symmetric (Exercise!)
- M is idempotent, i.e., $MM = M$ (Exercise!)

Therefore $\hat{\varepsilon}_{ols} = My = M(X\beta + \varepsilon) = M\varepsilon$ and

$$\hat{\varepsilon}_{ols}^{\mathrm{T}}\hat{\varepsilon}_{ols} = (M\varepsilon)^{\mathrm{T}}M\varepsilon = \varepsilon^T M^{\mathrm{T}}M\varepsilon = \varepsilon^{\mathrm{T}}MM\varepsilon = \varepsilon^{\mathrm{T}}M\varepsilon$$

OLS Estimator Variance (Homoskedasticity Case)

$$
\begin{aligned}
E(\hat{\varepsilon}_{ols}^{\mathrm{T}}\hat{\varepsilon}_{ols} \mid X) &= E(\varepsilon^{\mathrm{T}} M \varepsilon \mid X) \\
&= E(\mathrm{trace}(\varepsilon^{\mathrm{T}} M \varepsilon) \mid X) \quad \text{because } \varepsilon^{\mathrm{T}} M \varepsilon \text{ is a scalar} \\
&= E(\mathrm{trace}(\varepsilon\varepsilon^{\mathrm{T}} M) \mid X) \\
&= \mathrm{trace}(E(\varepsilon\varepsilon^{\mathrm{T}} M) \mid X) \quad \text{since trace is a sum} \\
&= \mathrm{trace}(E(\varepsilon\varepsilon^{\mathrm{T}} \mid X) M) = \mathrm{trace}(\sigma^2 I_n M) = \sigma^2 \mathrm{trace}(M) \\
&= \sigma^2 \mathrm{trace}(I_n - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}) = \sigma^2(\mathrm{trace}(I_n) - \mathrm{trace}(X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})) \\
&= \sigma^2(n - \mathrm{trace}((X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X)) = \sigma^2(n - \mathrm{trace}(I_K)) \\
&= \sigma^2(n-K)
\end{aligned}
$$

Therefore $E\left(\dfrac{\hat{\varepsilon}_{ols}^{\mathrm{T}}\hat{\varepsilon}_{ols}}{n-K} \,\middle|\, X\right) = \sigma^2$, which implies $E\left(\dfrac{\hat{\varepsilon}_{ols}^{\mathrm{T}}\hat{\varepsilon}_{ols}}{n-K}\right) = \sigma^2$

OLS Estimator Variance (Heteroskedasticity Case)

$$Var(\hat{\beta}^{ols} \mid X) = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\,Var(\varepsilon \mid X)X(X^{\mathrm{T}}X)^{-1}$$

$$= \left(\sum_{i=1}^{n} X_{i*}^{\mathrm{T}}X_{i*}\right)^{-1} \begin{bmatrix} X_{1*}^{\mathrm{T}} & X_{2*}^{\mathrm{T}} & \cdots & X_{n*}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} \begin{bmatrix} X_{1*} \\ X_{2*} \\ \vdots \\ X_{n*} \end{bmatrix} \left(\sum_{i=1}^{n} X_{i*}^{\mathrm{T}}X_{i*}\right)^{-1}$$

$$= \left(\sum_{i=1}^{n} X_{i*}^{\mathrm{T}}X_{i*}\right)^{-1} \left(\sum_{i=1}^{n} \sigma_i^2 X_{i*}^{\mathrm{T}}X_{i*}\right) \left(\sum_{i=1}^{n} X_{i*}^{\mathrm{T}}X_{i*}\right)^{-1}$$

Heteroskedasticity-Robust Estimator Variance-Covariance Matrix: replace σ_i^2 with $\hat{\epsilon}_{i,ols}^2$

$$\widehat{Var_{HC0}}(\hat{\beta}^{ols}) = \left(\sum_{i=1}^{n} X_{i*}^{\mathrm{T}}X_{i*}\right)^{-1} \left(\sum_{i=1}^{n} \hat{\epsilon}_{i,ols}^2 X_{i*}^{\mathrm{T}}X_{i*}\right) \left(\sum_{i=1}^{n} X_{i*}^{\mathrm{T}}X_{i*}\right)^{-1}$$

OLS (Best Linear Unbiased, under Homoskedasticity)

We have shown that $\widehat{\beta}^{ols}$ is unbiased

It is also "linear"

- A linear estimator is one with the form $\widetilde{\beta} = Ay$
- Each $\widetilde{\beta}_j = \sum_{i=1}^{n} a_{ji} Y_i$
- OLS estimator is $\widehat{\beta}^{ols} = \underbrace{(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}}_{A} y$

Now we show: OLS estimators are BLU (they have the smallest variance among all linear unbiased estimators)

OLS (Best Linear Unbiased, under Homoskedasticity)

BLU in the sense that

$$Var(c^{\mathrm{T}}\hat{\beta}^{ols} \mid X) \leq Var(c^{\mathrm{T}}\tilde{\beta} \mid X)$$

for all $K \times 1$ vectors c, and for all unbiased estimators of the form $\tilde{\beta} = By$

- each individual $\hat{\beta}_k$ is BLU
- all linear combinations of $\hat{\beta}$ are BLU

Consider predicting Y at the new observation $X_{0*} = \begin{bmatrix} 1 & X_{01} & \dots & X_{0,K-1} \end{bmatrix}$. OLS predictor is

$$\hat{Y}(X_{0*}) = X_{0*}\,\hat{\beta}^{ols}$$

which is a linear combination of the parameter estimates in $\hat{\beta}^{ols}$, i.e., OLS prediction rule gives us the most precise linear unbiased prediction of Y at X_{0*}

OLS (Best Linear Unbiased, under Homoskedasticity)

Proof of Efficiency: let $\tilde{\beta} = By$ be an unbiased estimator where $B \neq (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$

Let D be such that $B = D + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$, so that

$$\begin{aligned}\tilde{\beta} = By &= (D + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})y \\ &= (D + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})(X\beta + \varepsilon) \\ &= DX\beta + D\varepsilon + \beta + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\varepsilon\end{aligned}$$

To ensure unbiasedness of $\tilde{\beta}$, we must assume $DX = 0$

OLS (Best Linear Unbiased, under Homoskedasticity)

Then

$$
\begin{aligned}
Var(\tilde{\beta} \mid X) &= E((\tilde{\beta}-\beta)(\tilde{\beta}-\beta)^{\mathrm{T}} \mid X) \\
&= E((D+(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})\varepsilon\varepsilon^{\mathrm{T}}(D+(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})^{\mathrm{T}} \mid X) \\
&= (D+(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})E(\varepsilon\varepsilon^{\mathrm{T}} \mid X)(D+(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})^{\mathrm{T}} \\
&= \sigma^2(D+(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})(D+(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})^{\mathrm{T}} \\
&= \sigma^2(D+(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})(D^{\mathrm{T}}+X(X^{\mathrm{T}}X)^{-1}) \\
&= \sigma^2[DD^{\mathrm{T}}+(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}D^{\mathrm{T}}+DX(X^{\mathrm{T}}X)^{-1}+(X^{\mathrm{T}}X)^{-1}] \\
&= \sigma^2[DD^{\mathrm{T}}+(X^{\mathrm{T}}X)^{-1}] = \sigma^2DD^{\mathrm{T}}+\sigma^2(X^{\mathrm{T}}X)^{-1} = \sigma^2DD^{\mathrm{T}}+Var(\hat{\beta} \mid X)
\end{aligned}
$$

OLS (Best Linear Unbiased, under Homoskedasticity)

Therefore

$$\begin{aligned} Var(c^{\mathrm{T}}\tilde{\beta} \mid X) &= c^{\mathrm{T}} Var(\tilde{\beta} \mid X)c \\ &= c^{\mathrm{T}}(\sigma^2 DD^{\mathrm{T}} + Var(\hat{\beta} \mid X))c \\ &= \sigma^2 c^{\mathrm{T}} DD^{\mathrm{T}} c + c^{\mathrm{T}} Var(\hat{\beta} \mid X))c \\ &= \sigma^2 (D^{\mathrm{T}}c)^{\mathrm{T}} D^{\mathrm{T}} c + Var(c^{\mathrm{T}}\hat{\beta} \mid X)) \geq Var(c^{\mathrm{T}}\hat{\beta} \mid X)) \end{aligned}$$

NB: $D^{\mathrm{T}}c$ is a vector, therefore $(D^{\mathrm{T}}c)^{\mathrm{T}} D^{\mathrm{T}} c$ is a sum of squares, which cannot be negative

Example

- Estimate Multiple Linear Regression using `earnings.csv` data

$$\ln(earn) = \beta_0 + \beta_1 educ + \beta_2 tenure + \beta_3 age + \beta_4 age^2 + \epsilon$$

- We'll compute estimates twice, for illustration
  - Using formulas derived in this session
  - Using existing R functions and packages

```
library(tidyverse)
dat1 <- read_csv("data\\earnings2019.csv", show_col_types=FALSE)
```

Example

```
mlr <- function(y, X){
  n <- dim(X)[1]
  K <- dim(X)[2]
  XTXinv <- solve(t(X)%*%X)
  betahat <- XTXinv %*% t(X)%*%y
  yhat <- X %*% betahat
  ehat <- y - yhat
  sigmasqhat <- sum(ehat^2)/(dim(X)[1]-dim(X)[2])
  betahatvar <- sigmasqhat * XTXinv
  betahatse  <- sqrt(diag(betahatvar))
  betahat_t  <- betahat/betahatse
  results <- cbind(
    betahat, betahatse, betahat_t, 2*pt(-abs(betahat_t), n-K)
  )
  colnames(results) <- c("coef.", "s.e.", "t-stat", "p-value")
  model_return <- list("results"=results, "sigmasqhat"=sigmasqhat, "betahatvar"=betahatvar)
    return(model_return)
  }
```

Example

```
y = log(dat1$earn)
X = cbind("intercept"=1, "educ"=dat1$educ, "tenure"=dat1$tenure,
          "age"=dat1$age, "agesq"=dat1$age^2)
model2 <- mlr(y,X)
model2$results %>% round(5)
```

```
             coef.    s.e.    t-stat p-value
intercept -0.14085 0.11333  -1.24284 0.21399
educ       0.12590 0.00379  33.21175 0.00000
tenure     0.01545 0.00108  14.26744 0.00000
age        0.06251 0.00478  13.08669 0.00000
agesq     -0.00067 0.00005 -12.76758 0.00000
```

Example

Estimated variance-covariance matrix (not usually reported)

```
model2$betahatvar
```

	intercept	educ	tenure	age	agesq
intercept	1.284416e-02	-1.880320e-04	1.206658e-05	-4.679158e-04	4.957007e-06
educ	-1.880320e-04	1.436923e-05	-2.315549e-08	-9.040899e-07	1.083268e-08
tenure	1.206658e-05	-2.315549e-08	1.172805e-06	-6.690121e-07	2.934586e-09
age	-4.679158e-04	-9.040899e-07	-6.690121e-07	2.281319e-05	-2.471022e-07
agesq	4.957007e-06	1.083268e-08	2.934586e-09	-2.471022e-07	2.747833e-09

Example

Using the `lm()` function

```
model2lm <- lm(log(earn)~educ + tenure + age + I(age^2), data=dat1)
model2lm %>% summary %>% coefficients  %>% round(5)
```

```
             Estimate Std. Error   t value Pr(>|t|)
(Intercept) -0.14085    0.11333  -1.24284  0.21399
educ         0.12590    0.00379  33.21175  0.00000
tenure       0.01545    0.00108  14.26744  0.00000
age          0.06251    0.00478  13.08669  0.00000
I(age^2)    -0.00067    0.00005 -12.76758  0.00000
```

Example

Estimated var-covariance matrix

```
vcov(model2lm)
```

```
             (Intercept)          educ        tenure           age
(Intercept)  1.284416e-02 -1.880320e-04  1.206658e-05 -4.679158e-04
educ        -1.880320e-04  1.436923e-05 -2.315549e-08 -9.040899e-07
tenure       1.206658e-05 -2.315549e-08  1.172805e-06 -6.690121e-07
age         -4.679158e-04 -9.040899e-07 -6.690121e-07  2.281319e-05
I(age^2)     4.957007e-06  1.083268e-08  2.934586e-09 -2.471022e-07
                I(age^2)
(Intercept)  4.957007e-06
educ         1.083268e-08
tenure       2.934586e-09
age         -2.471022e-07
I(age^2)     2.747833e-09
```

Hypothesis Testing

We develop matrix formulas for the t- and F-tests of linear hypotheses concerning elements of β

If we assume

$$y = X\beta + \varepsilon\,,\ \varepsilon \mid X \sim \text{Normal}_n(0, \sigma^2 I)\,,$$

then we have

$$\hat{\beta} \mid X \sim \text{Normal}_K(\beta, \sigma^2 (X^{\mathrm{T}} X)^{-1})$$

since $\hat{\beta} = \beta + (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} \varepsilon$ is a constant plus a linear combination of normally distributed noise terms

Hypothesis Testing

A general single linear hypothesis can be written as

$$H_0 : r^{\mathrm{T}}\beta = r_0 \quad \text{vs} \quad r^{\mathrm{T}}\beta \neq r_0.$$

Example: To test $\beta_1 + \beta_2 = 1$ in the regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i,$$

set $r^{\mathrm{T}} = \begin{bmatrix} 0 & 1 & 1 \end{bmatrix}$ and $r_0 = 1$.

Hypothesis Testing

We have

$$r^{\mathrm{T}}\hat{\beta} \mid X \sim \text{Normal}(r^{\mathrm{T}}\beta, r^{\mathrm{T}}(\sigma^2(X^{\mathrm{T}}X)^{-1})r).$$

If the null hypothesis $r^{\mathrm{T}}\beta = r_0$ holds, then

$$r^{\mathrm{T}}\hat{\beta} \mid X \sim \text{Normal}(r_0, r^{\mathrm{T}}(\sigma^2(X^{\mathrm{T}}X)^{-1})r)$$

and

$$\frac{r^{\mathrm{T}}\hat{\beta} - r_0}{\sqrt{r^{\mathrm{T}}(\sigma^2(X^{\mathrm{T}}X)^{-1})r}} \sim \text{Normal}(0, 1).$$

Hypothesis Testing

Furthermore, it can be shown that if we replace σ^2 with $\widehat{\sigma^2}$, then

$$\begin{aligned} t &= \frac{r^{\mathrm{T}}\hat{\beta} - r_0}{\sqrt{r^{\mathrm{T}}(\widehat{\sigma^2}(X^{\mathrm{T}}X)^{-1})r}} \\ &= \frac{r^{\mathrm{T}}\hat{\beta} - r_0}{\sqrt{r^{\mathrm{T}}\widehat{Var}(\hat{\beta} \mid X)r}} \sim \mathrm{t}(n-K). \end{aligned}$$

This can be used to test the hypothesis $H_0 : r^{\mathrm{T}}\beta = r_0$ in the usual way.

Hypothesis Testing

To test multiple hypotheses jointly, write the hypotheses as

$$H_0 : \mathcal{R}\beta = r_0 \quad \text{vs} \quad H_A : \mathcal{R}\beta \neq r_0$$

where now $\mathcal{R}$ is a $(J \times K)$ matrix, and r_0 is a $(J \times 1)$ vector.

To test the hypotheses

$$H_0 : \beta_1 + \beta_2 = 1 \text{ and } \beta_3 = 0 \quad \text{vs} \quad H_A : \beta_1 + \beta_2 \neq 1 \text{ or } \beta_3 \neq 0 \text{ (or both)},$$

in the regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

set the matrices $\mathcal{R}$ and r_0 to

$$\mathcal{R} = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad r_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Hypothesis Testing

To test multiple hypotheses jointly, we can again compare the residual sum of squares from the restricted and unrestricted regressions

In particular, it can be shown that if the null is true, then

$$F = \frac{(\hat{\varepsilon}_{rls}^{\mathrm{T}}\hat{\varepsilon}_{rls} - \hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon})/J}{\hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon}/(n-K)} \sim \mathrm{F}(J, n-K)$$

where J is # restrictions being tested and $\hat{\varepsilon}$ is the vector of unrestricted OLS residuals

Reject H_0 if the F-statistic is "improbably large" under null, meaning that $F > \mathrm{F}_{1-\alpha}(J, n-K)$ where $\mathrm{F}_{1-\alpha}(J, n-K)$ is the $1-\alpha$ percentile of the $\mathrm{F}(J, n-K)$ distribution. Typically $\alpha = 0.01$, 0.05 or 0.1.

Hypothesis Testing

It turns out that in practice, one does not actually have to compute the restricted regression. It can be shown that

$$\hat{\varepsilon}_{rls}^{\mathrm{T}}\hat{\varepsilon}_{rls} - \hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon} = (\mathcal{R}\hat{\beta} - r_0)^{\mathrm{T}}(\mathcal{R}(X^{\mathrm{T}}X)^{-1}\mathcal{R}^{\mathrm{T}})^{-1}(\mathcal{R}\hat{\beta} - r_0)$$

where $\hat{\beta}$ is the unrestricted OLS estimators (for proof, see Econometrics Notes)

Furthermore, denominator of F-statistic is $\widehat{\sigma^2}$, therefore

$$\begin{aligned} F &= (\mathcal{R}\hat{\beta} - r_0)^{\mathrm{T}}(\mathcal{R}(\widehat{\sigma^2}(X^{\mathrm{T}}X)^{-1})\mathcal{R}^{\mathrm{T}})^{-1}(\mathcal{R}\hat{\beta} - r_0)/J \\ &= (\mathcal{R}\hat{\beta} - r_0)^{\mathrm{T}}(\mathcal{R}\,\widehat{Var}(\hat{\beta} \mid X)\,\mathcal{R}^{\mathrm{T}})^{-1}(\mathcal{R}\hat{\beta} - r_0)/J \\ &\sim \mathsf{F}(J, n-K) \end{aligned}$$

Hypothesis Testing

If the error terms are not normally distributed, we can use the asymptotically valid chi-sq version of the test, namely

$$JF \overset{a}{\sim} \chi^2(J)$$

If errors are heteroskedastic, we can replace $\widehat{Var}(\hat{\beta} \mid X)$ with the (asymptotically valid) heteroskedasticity-robust versions

Hypothesis Testing: Example

We use the following example

$$\ln earn = \beta_0 + \beta_1 educ + \beta_2 black + \beta_3 female + \beta_4 black.female + \epsilon$$

Use t-test to test: $H_0 : \beta_2 = \beta_3$ vs $H_A : \beta_2 \neq \beta_3$

Use F-test to test: $H_0 : \beta_2 = \beta_3$ and $\beta_4 = 0$ versus $H_0 : \beta_2 \neq \beta_3$ or $\beta_4 \neq 0$

```
library(car)
dat2<-read_csv("data\\earnings2019.csv",show_col_types=FALSE) %>%
  mutate(female=1-male,
         white=if_else(race=="White", 1,0),
         black=if_else(race=="Black", 1,0),
         other=if_else(race=="Other", 1,0)) %>% select(-race)
mdl_unres <- lm(log(earn) ~ educ + black*female, data=dat2)
```

Hypothesis Testing: Example

```
mdl_unres %>% summary %>% coefficients %>% round(5)
```

```
             Estimate Std. Error   t value Pr(>|t|)
(Intercept)   1.53686    0.05709  26.92194  0.00000
educ          0.12786    0.00388  32.95769  0.00000
black        -0.26005    0.02696  -9.64677  0.00000
female       -0.28066    0.01959 -14.32381  0.00000
black:female  0.08730    0.03549   2.46010  0.01392
```

Hypothesis Testing: Example

```
linearHypothesis(mdl_unres, c("black=female"))
```

```
Linear hypothesis test:
black - female = 0

Model 1: restricted model
Model 2: log(earn) ~ educ + black * female

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   4942 1603.7
2   4941 1603.5  1   0.18778 0.5786 0.4469
```

Hypothesis Testing: Example

```
linearHypothesis(mdl_unres, c("black=female", "black:female=0"))
```

```
Linear hypothesis test:
black - female = 0
black:female = 0

Model 1: restricted model
Model 2: log(earn) ~ educ + black * female

  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1   4943 1606.5
2   4941 1603.5  2    2.9507 4.5461 0.01065 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis Testing: Example

For illustration, we compute the two tests using the formulas derived in this section

```
y = log(dat2$earn)
X = cbind("intercept"=1, "educ"=dat2$educ, "black"=dat2$black, "female"=dat2$female,
          "black.female" = dat2$black * dat2$female)
n = dim(X)[1]
K = dim(X)[2]
model1 <- mlr(y,X)
model1$results  %>% round(5)
```

```
                coef.    s.e.    t-stat p-value
intercept     1.53686 0.05709  26.92194 0.00000
educ          0.12786 0.00388  32.95769 0.00000
black        -0.26005 0.02696  -9.64677 0.00000
female       -0.28066 0.01959 -14.32381 0.00000
black.female  0.08730 0.03549   2.46010 0.01392
```

Hypothesis Testing: Example

```
# t-test of "black=female"
r = matrix(c(0,0,1,-1,0), ncol=1)
r0 = 0
betahat <- as.matrix(model1$results[,1])
tstat <- (t(r) %*% betahat - r0)/sqrt(t(r) %*% model1$betahatvar %*% r)
tstat_pval <- 2*pt(-abs(tstat), n-K)
cat("\n Test: b2=b3")
cat("\n tstat:", round(tstat,5), "   p-value:", round(tstat_pval,5), "\n")
```

```
Test: b2=b3
tstat: 0.76067    p-value: 0.44689
```

Hypothesis Testing: Example

```
# F-test of "black=female" and "black.female=0"
R <- matrix(c(0,0,1,-1,0,
              0,0,0, 0,1), nrow=2, byrow=TRUE)
J <- dim(R)[1]
r0 <- matrix(c(0,0), nrow = 2)
Fstat <- t(R%*%betahat - r0) %*%
              solve(R %*% model1$betahatvar %*% t(R)) %*%
                  (R %*% betahat - r0)/J
Fstat_pval <- 1-pf(Fstat, J, n-K)
cat("\n Test: b2=b3, b4=0")
cat("\n Fstat:", round(Fstat,5), "   p-value:", round(Fstat_pval,5),"\n")
```

```
 Test: b2=b3, b4=0
 Fstat: 4.54611    p-value: 0.01065
```

Asymptotic Properties of OLS Estimators

Multivariate versions of LLN and CLT:

If $\{Z_i\}_{i=1}^n$ iid vectors of random variables with $E(Z_i) = 0$ and $Var(Z_i) = \Omega$, for all i, then

$$\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i = \begin{bmatrix} (1/n)\sum_{i=1}^n Z_{1i} \\ (1/n)\sum_{i=1}^n Z_{2i} \\ \vdots \\ (1/n)\sum_{i=1}^n Z_{ki} \end{bmatrix} \xrightarrow{p} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix} = \mu$$

$$\sqrt{n}\,\overline{Z} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i = \begin{bmatrix} (1/\sqrt{n})\sum_{i=1}^n Z_{1i} \\ (1/\sqrt{n})\sum_{i=1}^n Z_{2i} \\ \vdots \\ (1/\sqrt{n})\sum_{i=1}^n Z_{ki} \end{bmatrix} \xrightarrow{d} \text{Normal}_k(0, \Omega)$$

Asymptotic Properties

To talk about limiting distributions, we have to scale $\widehat{\beta}$. Use

$$\sqrt{n}(\widehat{\beta}^{ols} - \beta) = \left(\frac{1}{n} \sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*} \right)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_{i*}^{\mathrm{T}} \epsilon_i \right)$$

Our assumptions and the CLT imply

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_{i*} \epsilon_i \xrightarrow{d} \mathsf{Normal}_{k+1}(0, S)$$

therefore

$$\sqrt{n}(\widehat{\beta}^{ols} - \beta) = \underbrace{\left(\frac{1}{n} \sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*} \right)^{-1}}_{\xrightarrow{p} \Sigma_{xx}^{-1}} \underbrace{\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_{i*}^{\mathrm{T}} \epsilon_i \right)}_{\xrightarrow{p} \mathsf{Normal}_{k+1}(0,S)} \xrightarrow{d} \mathsf{Normal}_{k+1}(0, \Sigma_{xx}^{-1} S \Sigma_{xx}^{-1})$$

Asymptotic Properties

That is, $\hat{\beta}^{ols}$ is consistent, with asymptotic variance $Avar(\hat{\beta}^{ols}) = \Sigma_{xx}^{-1} S \Sigma_{xx}^{-1}$. This result justifies the approximation

$$Var(\hat{\beta}^{ols}) \approx (1/n)\Sigma_{xx}^{-1} S \Sigma_{xx}^{-1}.$$

An obvious estimator for Σ_{xx} is

$$\hat{\Sigma}_{xx} = \frac{1}{n}\sum_{i=1}^{N} X_{i*}^{\mathrm{T}} X_{i*} = \frac{1}{n} X^{\mathrm{T}} X$$

Some additional assumptions (see advanced econometrics textbooks) guarantee

$$\hat{S} = \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^2 X_{i*}^{\mathrm{T}} X_{i*} \xrightarrow{p} S.$$

Asymptotic Properties

This allows us to consistently estimate the asymptotic variance of $\hat{\beta}$ by

$$\widehat{Avar}(\hat{\beta}^{ols}) = \left(\frac{1}{n}\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*}\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^2 X_{i*}^{\mathrm{T}} X_{i*}\right) \left(\frac{1}{n}\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*}\right)^{-1}$$

and justifies the use of

$$\begin{aligned}
\widehat{Var}_{HC0}(\hat{\beta}^{ols}) &= \frac{1}{n}\widehat{Avar}(\hat{\beta}^{ols}) \\
&= \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*}\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^2 X_{i*}^{\mathrm{T}} X_{i*}\right) \left(\frac{1}{n}\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*}\right)^{-1} \\
&= \left(\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*}\right)^{-1} \left(\sum_{i=1}^{n} \hat{\epsilon}_i^2 X_{i*}^{\mathrm{T}} X_{i*}\right) \left(\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*}\right)^{-1}
\end{aligned}$$

Example

We adapt previously written `mlr` function to give HC0 standard errors:

```
mlr_hc0 <- function(y, X){
  n <- dim(X)[1]; K <- dim(X)[2]
  XTXinv <- solve(t(X)%*%X)
  betahat <- XTXinv %*% t(X)%*%y; yhat <- X %*% betahat; ehat <- y - yhat
  sigmasqhat <- sum(ehat^2)/(dim(X)[1]-dim(X)[2])
  hatS = 0
  for (i in 1:n){
    xi = as.matrix(X[i,])
    hatS = hatS + ehat[i]^2*xi%*%t(xi)
  }
  betahatvar_hc0 <- XTXinv %*% hatS %*% XTXinv ## XTXinv computed earlier
  betahatse_hc0  <- sqrt(diag(betahatvar_hc0))
  betahat_t_hc0  <- betahat/betahatse_hc0
  results <- cbind(betahat, betahatse_hc0, betahat_t_hc0, 2*pnorm(-abs(betahat_t_hc0), 0, 1))
  colnames(results) <- c("coef.", "s.e.(hc0)", "t-stat", "p-value")
  model_return <- list("results"=results, "betahatvar"=betahatvar_hc0)
    return(model_return)
  }
```

Example

$$\ln earn = \beta_0 + \beta_1 educ + \beta_2 black + \beta_3 female + \beta_4 black.female + \epsilon$$

```
y = log(dat1$earn)
X = cbind("intercept"=1, "educ"=dat2$educ, "black"=dat2$black, "female"=dat2$female,
          "black.female"=dat2$black * dat2$female)
model2_hc0 <- mlr_hc0(y,X)
cat("Estimation results:\n"); model2_hc0$results %>% round(6)
```

```
Estimation results:
                coef. s.e.(hc0)     t-stat p-value
intercept    1.536856  0.056516  27.193109 0.00000
educ         0.127858  0.003927  32.555323 0.00000
black       -0.260050  0.026926  -9.657793 0.00000
female      -0.280661  0.020170 -13.915056 0.00000
black.female 0.087299  0.034640   2.520150 0.01173
```

Example

s.e. based on the following variance-covariance matrix

```
cat("\n variance-covariance matrix (hc0) of betahat:\n")
model2_hc0$betahatvar %>% round(6)
```

```
 variance-covariance matrix (hc0) of betahat:
             intercept      educ     black    female black.female
intercept     0.003194 -0.000214 -0.000290 -0.000070     0.000188
educ         -0.000214  0.000015  0.000005 -0.000011     0.000003
black        -0.000290  0.000005  0.000725  0.000224    -0.000724
female       -0.000070 -0.000011  0.000224  0.000407    -0.000400
black.female  0.000188  0.000003 -0.000724 -0.000400     0.001200
```

Example

Alternatives for calculating HC variance-covariance matrices

- `hccm()` from `car` package

```
model3 <- lm(log(earn) ~ educ + black*female, data=dat2)
betahatvar_HC0_car = hccm(model3,type="hc0")
round(betahatvar_HC0_car,6)
```

```
             (Intercept)      educ     black    female black:female
(Intercept)     0.003194 -0.000214 -0.000290 -0.000070     0.000188
educ           -0.000214  0.000015  0.000005 -0.000011     0.000003
black          -0.000290  0.000005  0.000725  0.000224    -0.000724
female         -0.000070 -0.000011  0.000224  0.000407    -0.000400
black:female    0.000188  0.000003 -0.000724 -0.000400     0.001200
```

Example

- vcovHC() function from the sandwich package.

```
model3 <- lm(log(earn) ~ educ + black*female, data=dat2)
betahatvar_HC0_sando = sandwich::vcovHC(model3,type="HC0")
round(betahatvar_HC0_sando,6)
```

```
             (Intercept)      educ     black    female black:female
(Intercept)     0.003194 -0.000214 -0.000290 -0.000070     0.000188
educ           -0.000214  0.000015  0.000005 -0.000011     0.000003
black          -0.000290  0.000005  0.000725  0.000224    -0.000724
female         -0.000070 -0.000011  0.000224  0.000407    -0.000400
black:female    0.000188  0.000003 -0.000724 -0.000400     0.001200
```

Example

Can use HC variance estimator to construct heteroskedasticity robust t and F statistics

- replace $\widehat{Var}(\hat{\beta} \mid X)$ with the heteroskedasticity robust variance estimator version
- this is done for you in `linearHypothesis()` by selecting "Chisq" for test option

Example

```
linearHypothesis(model3, c("black=female", "black:female=0"),
                 vcov=betahatvar_HC0_sando, test="Chisq")
```

```
Linear hypothesis test:
black - female = 0
black:female = 0

Model 1: restricted model
Model 2: log(earn) ~ educ + black * female

Note: Coefficient covariance matrix supplied.

  Res.Df Df  Chisq Pr(>Chisq)
1   4943
2   4941  2 9.6196    0.00815 **
```