Chapter 1
Linear models

Simon Fraser University
ECON 483
Summer 2023

Disclaimer

Outline

- Multivariate linear models: Definition, assumptions
- The Ordinary Least Squares estimator
  - Principle
  - Goodness of fit
  - Properties
  - Inference
- Illustration in **RStudio**
- Violation of model assumptions
- Assumption 1: Linearity
- Assumption 2: i.i.d sample
- Assumption 3: Exogeneity
- Assumption 4: No collinearity
- Assumption 5: Homoskedasticity

Multivariate linear model

- Consider the following model:

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + ... + \beta_K X_{i,K} + u_i}$$

- There are $\boldsymbol{K+1}$ different variables (the $\boldsymbol{+1}$ is the intercept), with their associated coefficients
- In vector notation:

$$\boldsymbol{Y_i = (1\ X_{i,1}\ X_{i,2}\ ...\ X_{i,K}) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix} + u_i}$$

$$\boldsymbol{Y_i = X_i' \beta + u_i}$$

Multivariate linear model

- $\boldsymbol{X}_i'$ could be $(\mathbf{1}\ \boldsymbol{Age}_i\ \boldsymbol{Education}_i\ \ldots\ \boldsymbol{Income}_i)$
- We can then stack these rows on top of each other to get the covariates matrix

$$\boldsymbol{X} \equiv \begin{pmatrix} \mathbf{1} & \boldsymbol{Age_1} & \boldsymbol{Education_1} & \ldots & \boldsymbol{Income_1} \\ \mathbf{1} & \boldsymbol{Age_2} & \boldsymbol{Education_2} & \ldots & \boldsymbol{Income_2} \\ \vdots & \vdots & \vdots & \vdots & \\ \mathbf{1} & \boldsymbol{Age_n} & \boldsymbol{Education_n} & \ldots & \boldsymbol{Income_n} \end{pmatrix}$$

Linear models: Assumptions

Assumption 1 : Linearity

Consider the following model:

$$Y_i = X_i'\beta + u_i$$

where:

- Y_i is the **dependent variable**
- X_i is the vector of $K+1$ **independent variables** or **explanatory variables**
- β is the $K+1$ vector **parameters** of interest: β_0, β_1,..., β_K
- u_i is the **error term**

- Note: The model is **linear in the parameters**

Linear model (cont'd)

Assumption 2 : i.i.d. sample

An Independent and Identically Distributed sample is available:
$\{x_{i,1}, x_{i,2}, ..., x_{i,K}, y_i\}, \ i = 1, ..., n$

- Note: For Assumption 2 we use lower case letters are they are observations of the random variables $X_{i,1}, X_{i,2}, ..., X_{i,K}$ and Y_i

Linear model (cont'd)

Assumption 3 : Exogeneity

The error term has an expectation of 0 conditional on X_i:

$$\mathbb{E}[\boldsymbol{u_i}|\boldsymbol{X_i}] = \mathbf{0} \quad \forall \boldsymbol{i} = \mathbf{1}, ..., \boldsymbol{n}$$

- In words: When fixing $\boldsymbol{X_i}$ to a value $\boldsymbol{x}$, the error term $\boldsymbol{u_i}$ is on average 0. I.e. when looking at individuals for which $\boldsymbol{X} = \boldsymbol{x}$, their error term is equal to 0 on average
- Note: $\boldsymbol{X_i}$ is now a vector, so $\boldsymbol{x}$ is a vector of numbers too
- Note: By the law of iterated expectation it means $\mathbb{E}[\boldsymbol{u_i}] = \mathbb{E}[\mathbb{E}(\boldsymbol{u_i}|\boldsymbol{X_i})] = \mathbb{E}[\mathbf{0}] = \mathbf{0}$
- Interpretation: $\mathbb{E}[\boldsymbol{Y_i}|\boldsymbol{X_i} = \boldsymbol{x_i}] = \mathbb{E}[\boldsymbol{X_i'\beta} + \boldsymbol{u_i}|\boldsymbol{X_i} = \boldsymbol{x_i}] = \boldsymbol{x_i'\beta}$

Linear model (cont'd)

Assumption 4 : No collinearity

The $[n \times (K+1)]$ matrix X gathering the observations x_i has rank $K+1$, i.e. it is full column rank.

- In words: No covariate can be expressed as a linear combination of the other covariates. For instance, do not include X_i and a covariate $Z_i = aX_i$ where a is a constant
- Interpretation: Using X_i or Z_i is equivalent to using the same information, so using both is redundant
- Numerically speaking, that implies a problem when computing the inverse of a matrix. It is the equivalent of dividing by 0 when we divide by a number

Linear model (cont'd)

Assumption 5 : Homoskedasticity

The conditional variance of the error term does not change with i:

$$\mathbb{V}[u_i|\boldsymbol{X}_i] = \sigma^2(x_i) = \sigma^2$$

- Note: Assumption 5 states that the variance of the error term does not depend on the value taken by $\boldsymbol{X}_i$

Assumptions: Remarks

- Assumption 1 assumes the form of the relationship between $\boldsymbol{Y_i}$ and $\boldsymbol{X_i}$. A linear form makes the coefficients interpretable: A coefficient $\boldsymbol{\beta_k}$ associated to a variable $\boldsymbol{X_{i,k}}$ represents the marginal effect of $\boldsymbol{X_{i,k}}$ on $\boldsymbol{Y_i}$: $\boldsymbol{\beta_k = \frac{\partial Y_i}{\partial X_{i,k}}}$, **keeping all the other X's constant**
- Assumption 2 is **needed** to show the consistency and asymptotic normality of the OLS estimator
- Assumption 3 is **needed** to show the consistency and asymptotic normality of the OLS estimator
- Assumption 4 is **needed** to show the consistency and asymptotic normality of the OLS estimator
- Assumption 5 allows the OLS estimator to have an appealing property for inference (BLUE). It is not needed to get **consistency**, **unbiasedness** or **asymptotic normality**

Ordinary Least Squares

The Ordinary Least Squares (OLS) estimator

- As in the univariate case, we want to minimize the distance between the predictions and the actual data:

$$\hat{\beta}_{OLS} \equiv \underset{\{\beta_0,\ldots,\beta_K\}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i'\beta)^2$$

- We now have $K+1$ first order conditions. But they look all the same (same as the univariate case)
- Matrix notation can help us solve the system of equations more easily
- The objective function can be rewritten in matrix terms
- Let $Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$ be $[n \times 1]$ the vector of data about the dependent variable

The Ordinary Least Squares (OLS) estimator

- Let

$$\boldsymbol{X} \equiv \begin{pmatrix} \mathbf{1} & \boldsymbol{x_{1,1}} & \dots & \boldsymbol{x_{1,K}} \\ \mathbf{1} & \boldsymbol{x_{2,1}} & \dots & \boldsymbol{x_{2,K}} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{1} & \boldsymbol{x_{n,1}} & \dots & \boldsymbol{x_{n,K}} \end{pmatrix}$$

  be the $[\boldsymbol{n} \times (\boldsymbol{K}+\mathbf{1})]$ matrix of data containing the observations for the covariates
- The first column is full of ones: It is the intercept!

The Ordinary Least Squares (OLS) estimator: FOC

- Time to take the first order conditions!
- There will be $\boldsymbol{K+1}$ of them, and hence a system of $\boldsymbol{K+1}$ equations with $\boldsymbol{K+1}$ unknowns
- That is where linear algebra meets calculus. It allows us to take a derivative with respect to a vector, i.e. with respect to each variable in a vector at once, and the system of equations will be easier to solve that the traditional substitution method

$$\frac{\partial Q}{\partial \beta}(\hat{\beta}_{OLS}) = \underset{[(K+1)\times n]}{-2\boldsymbol{X}'}(\underset{[n\times 1]}{\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}_{OLS}}) = \underset{[(K+1)\times 1]}{\boldsymbol{0}}$$

$$\Leftrightarrow \underset{[(K+1)\times(K+1)]}{\boldsymbol{X}'\boldsymbol{X}}\hat{\beta}_{OLS} = \underset{[(K+1)\times 1]}{\boldsymbol{X}'\boldsymbol{Y}}$$

$$\Leftrightarrow \underset{[(K+1)\times 1]}{\hat{\beta}_{OLS}} = \underset{[(K+1)\times(K+1)]}{(\boldsymbol{X}'\boldsymbol{X})^{-1}}\underset{[(K+1)\times 1]}{\boldsymbol{X}'\boldsymbol{Y}}$$

The Ordinary Least Squares (OLS) estimator

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$$

- The $(X'X)^{-1}$ term is the equivalent of $1/\widehat{Var}(X_i)$ in the univariate case

$$X'X = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{1,1} & x_{2,1} & \dots & x_{n,1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1,K} & x_{2,K} & \dots & x_{n,K} \end{pmatrix} \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,K} \\ 1 & x_{2,1} & \dots & x_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,K} \end{pmatrix}$$

$$X'X = \begin{pmatrix} n & \sum_{i=1}^n x_{i,1} & \dots & \sum_{i=1}^n x_{i,K} \\ \sum_{i=1}^n x_{i,1} & \sum_{i=1}^n x_{i,1}^2 & \dots & \sum_{i=1}^n x_{i,1}x_{i,K} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{i,K} & \sum_{i=1}^n x_{i,1}x_{i,K} & \dots & \sum_{i=1}^n x_{i,K}^2 \end{pmatrix}$$

The OLS estimator formula

$$\boldsymbol{X'X} = \begin{pmatrix} \boldsymbol{n} & \sum_{i=1}^{n} \boldsymbol{x_{i,1}} & \cdots & \sum_{i=1}^{n} \boldsymbol{x_{i,K}} \\ \sum_{i=1}^{n} \boldsymbol{x_{i,1}} & \sum_{i=1}^{n} \boldsymbol{x_{i,1}^2} & \cdots & \sum_{i=1}^{n} \boldsymbol{x_{i,1}x_{i,K}} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} \boldsymbol{x_{i,K}} & \sum_{i=1}^{n} \boldsymbol{x_{i,1}x_{i,K}} & \cdots & \sum_{i=1}^{n} \boldsymbol{x_{i,K}^2} \end{pmatrix}$$

- We can see portions of sample variances and covariances
- The inverse of that matrix will contain elements of sample variances and covariances too
- But **careful**: In the multivariate case, $\hat{\boldsymbol{\beta}}_{\boldsymbol{1}} \neq \frac{\widehat{Cov}(\boldsymbol{X_{i,1}}, \boldsymbol{Y_i})}{\widehat{Var}(\boldsymbol{X_{i,1}})}$!
- It is because the covariance between all the different $\boldsymbol{X}$'s gets in the way

The OLS estimator formula

- The $\boldsymbol{X'Y}$ term is the equivalent of $\widehat{\boldsymbol{Cov}}(\boldsymbol{X_i}, \boldsymbol{Y_i})$ in the univariate case

$$\boldsymbol{X'X} = \begin{pmatrix} \mathbf{1} & \mathbf{1} & \dots & \mathbf{1} \\ \boldsymbol{x_{1,1}} & \boldsymbol{x_{2,1}} & \dots & \boldsymbol{x_{n,1}} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x_{1,K}} & \boldsymbol{x_{2,K}} & \dots & \boldsymbol{x_{n,K}} \end{pmatrix} \begin{pmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \\ \vdots \\ \boldsymbol{y_n} \end{pmatrix}$$

$$\boldsymbol{X'Y} = \begin{pmatrix} \sum_{i=1}^{n} \boldsymbol{y_i} \\ \sum_{i=1}^{n} \boldsymbol{y_i x_{i,1}} \\ \vdots \\ \sum_{i=1}^{n} \boldsymbol{y_i x_{i,K}} \end{pmatrix}$$

OLS as a linear estimator

- The OLS estimator is part of the class of **linear estimators**

Definition: Linear estimators

A linear estimator for a dependent variable $\boldsymbol{Y_i}$ is of the form

$$\hat{\boldsymbol{Y}} \equiv \boldsymbol{LY}$$

where $\boldsymbol{L}$ is a $[\boldsymbol{n} \times \boldsymbol{n}]$ matrix. Equivalently:

$$\hat{\boldsymbol{y}}_i = \sum_{\boldsymbol{j=1}}^{\boldsymbol{n}} \boldsymbol{L}_{i,j} \boldsymbol{y}_j$$

In other words, the predictions are made by taking a linear combination of the observations of the dependent variable

The OLS estimator: Estimate and predictions

- We now have an estimate of the **marginal effect** of **each** $\boldsymbol{X_k}$ on $\boldsymbol{Y_i}$: $\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}$
- We can also build predictions of $\boldsymbol{y_i}$ for any value of $\boldsymbol{x_i}$:

$$\hat{\boldsymbol{y}}_{\boldsymbol{i}} = \hat{\boldsymbol{\beta}}_{\boldsymbol{0}} + \hat{\boldsymbol{\beta}}_{\boldsymbol{1}}\boldsymbol{x}_{\boldsymbol{i},\boldsymbol{1}} + \hat{\boldsymbol{\beta}}_{\boldsymbol{2}}\boldsymbol{x}_{\boldsymbol{i},\boldsymbol{2}} + \ldots + \hat{\boldsymbol{\beta}}_{\boldsymbol{K}}\boldsymbol{x}_{\boldsymbol{i},\boldsymbol{K}}$$

- Note: If $\boldsymbol{x_{i,k}} = \boldsymbol{0}$ for each $\boldsymbol{k}$, then $\hat{\boldsymbol{y}}_{\boldsymbol{i}} = \hat{\boldsymbol{\beta}}_{\boldsymbol{0}}$, the intercept of the estimated regression line
- It would make less sense to plot $\boldsymbol{Y_i}$ against $\boldsymbol{X_{i,k}}$ since there are many other $\boldsymbol{X}$'s. Or we would need to fix them first
- A plot of $\boldsymbol{Y_i}$ against $\hat{\boldsymbol{Y}}_{\boldsymbol{i}}$ always makes sense: The closer the points to the 45 degrees line, the closer the predictions to the actual data

The OLS estimator: Interpretation

- Say we have a regression with 2 variables, $\boldsymbol{X_{i,1}}$ and $\boldsymbol{X_{i,2}}$
- We find and $\hat{\boldsymbol{\beta}}_{\mathbf{0}} = \mathbf{2}$, $\hat{\boldsymbol{\beta}}_{\mathbf{1}} = \mathbf{3.5}$ and $\hat{\boldsymbol{\beta}}_{\mathbf{2}} = \mathbf{5}$
- It means that if $\boldsymbol{x_{i,1}}$ increases by **one** unit, $\boldsymbol{y_i}$ increases by $\frac{\partial \boldsymbol{y_i}}{\partial \boldsymbol{x_{i,1}}} = \hat{\boldsymbol{\beta}}_{\mathbf{1}} = \mathbf{3.5}$ units
- And if $\boldsymbol{x_{i,2}}$ increases by **one** unit, $\boldsymbol{y_i}$ increases by $\frac{\partial \boldsymbol{y_i}}{\partial \boldsymbol{x_{i,2}}} = \hat{\boldsymbol{\beta}}_{\mathbf{2}} = \mathbf{5}$ units
- Typically, we look at the change in one covariate at a time. All the other ones are held fixed ("ceteris paribus": Everything else being equal)
- $\hat{\boldsymbol{\beta}}_{\mathbf{1}}$ is the estimate of the marginal effect of $\boldsymbol{x_{i,1}}$ on $\boldsymbol{y_i}$ (analogously for $\hat{\boldsymbol{\beta}}_{\mathbf{2}}$)
- For an observation where $\boldsymbol{x_{i,1}} = \mathbf{50}$ and $\boldsymbol{x_{i,2}} = \mathbf{12}$, we predict that observation will have $\hat{\boldsymbol{y}}_{\boldsymbol{i}} = \hat{\boldsymbol{\beta}}_{\mathbf{0}} + \hat{\boldsymbol{\beta}}_{\mathbf{1}} \times \mathbf{50} + \hat{\boldsymbol{\beta}}_{\mathbf{2}} \times \mathbf{12} = \mathbf{2} + \mathbf{3.5} \times \mathbf{50} + \mathbf{5} \times \mathbf{12} = \mathbf{237}$

Fitted values and residuals

- $\hat{\boldsymbol{y}}_{\boldsymbol{i}}$ are the predictions, i.e. the points on the regression "line" (a line with one covariate, a surface with two, a volume with three, etc. . .)
- The residuals are defined as before: $\hat{\boldsymbol{u}}_{\boldsymbol{i}} = \boldsymbol{y}_{\boldsymbol{i}} - \hat{\boldsymbol{y}}_{\boldsymbol{i}}$
- $\hat{\boldsymbol{u}}_{\boldsymbol{i}}$ is **not** an estimate of the error term $\boldsymbol{u}_{\boldsymbol{i}}$ (which is a random variable, not a parameter to estimate). It is just what is left by the regression
- The average value of the residuals is then 0: $\bar{\hat{\boldsymbol{u}}} = \bar{\boldsymbol{y}} - \bar{\hat{\boldsymbol{y}}} = \bar{\boldsymbol{y}} - \bar{\boldsymbol{y}} = \mathbf{0}$
- So the average data point is **perfectly** predicted by the OLS estimator

Goodness-of-fit

- The **R-squared** is the proportion of the variance in $\boldsymbol{Y_i}$ that is explained by the model, i.e. by our $\hat{\boldsymbol{Y}}_{\boldsymbol{i}}$. It can be expressed in terms of the $\boldsymbol{ESS}$ but we can also use the **Sum of squared residuals**, or $\boldsymbol{SSR}$:

$$\boldsymbol{R^2} \equiv \frac{\boldsymbol{ESS}}{\boldsymbol{TSS}} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} = 1 - \frac{\boldsymbol{SSR}}{\boldsymbol{TSS}} = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2}{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

  where $\bar{\boldsymbol{y}}$ is the sample mean of $\boldsymbol{y_i}$

- The **adjusted R-squared**, denoted $\bar{\boldsymbol{R}}^{\boldsymbol{2}}$, is a modified version of the $\boldsymbol{R^2}$:

$$\bar{\boldsymbol{R}}^{\boldsymbol{2}} = 1 - \frac{\boldsymbol{SSR}/(\boldsymbol{n}-\boldsymbol{K}-1)}{\boldsymbol{TSS}/(\boldsymbol{n}-1)}$$

- Idea: Adding more variables for a given $\boldsymbol{n}$ increases $\boldsymbol{R^2}$, although too many variables is not good either. $\bar{\boldsymbol{R}}^{\boldsymbol{2}}$ might increase when $\boldsymbol{K}$ increases, but **not** if the extra variables are irrelevant

Properties of the OLS estimator

Properties of the OLS estimator

- By computing the $\hat{\beta}$ coefficients, we hope to learn some thing about the true β
- Like other estimators (sample average/proportion, corrected sample variance), it is relevant to ask whether the OLS estimator is consistent, biased, and asymptotically normal
- The assumptions made previously will be crucial in determining these properties
- Let $\hat{\beta}_{OLS} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_K)$ and let $\beta = (\beta_0, \beta_1, ..., \beta_K)$

OLS estimator properties: Consistency

Theorem 1 : Consistency of the OLS estimator

If Assumptions **1**, **2**, **3** and **4** are satisfied, then:

$$\hat{\beta}_{OLS} \xrightarrow{\mathbb{P}} \beta$$

where $\xrightarrow{\mathbb{P}}$ denotes convergence in probability, i.e. $\forall \varepsilon > 0$, $\mathbb{P}\left(|\hat{\beta}_{OLS} - \beta| > \varepsilon\right) \to 0$ as $n \to \infty$

OLS estimator properties: Unbiasedness

Theorem 2 : Unbiasedness of the OLS estimator

If Assumptions 1, 2, 3 and 4 are satisfied, then the OLS estimator is **unbiased**, i.e.:

$$\mathbb{E}[\hat{\boldsymbol{\beta}}_{\boldsymbol{OLS}}] = \boldsymbol{\beta}$$

- Assumption 3 is particularly important for that property
- Idea: Compute $\mathbb{E}[\hat{\boldsymbol{\beta}}_{\boldsymbol{OLS}}|\boldsymbol{X} = \boldsymbol{x}]$, then take the expectation of it again to get $\mathbb{E}[\hat{\boldsymbol{\beta}}_{\boldsymbol{OLS}}]$ by the law of iterated expectations
- We will later cover the case where Assumption 3 is not satisfied, in which case $\hat{\boldsymbol{\beta}}_{\boldsymbol{OLS}}$ is **biased** and **inconsistent**

OLS estimator properties: Asymptotic normality

- Recall that we have

$$\hat{\beta}_{OLS} - \beta = (X'X)^{-1}X'U$$

Theorem 3 : Asymptotic normality of the OLS estimator

If Assumptions 1, 2, 3, 4 and 5 are satisfied, then:

$$\sqrt{n}\left(\hat{\beta}_{OLS} - \beta\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2(\mathbb{E}[X_i X_i'])^{-1}\right)$$

where $\xrightarrow{d}$ refers to convergence in distribution.

- Theorem 3 is an application of the **Central Limit Theorem** on the right hand side of the expression above
- $(\mathbb{E}[X_i X_i'])^{-1}$ is a $[(K+1) \times (K+1)]$ matrix

Best Linear Unbiased Estimator

Theorem 4 : B.L.U.E.

If Assumptions If Assumptions **1**, **2**, **3**, **4** and **5** are satisfied, then the OLS estimator is the **Best Linear Unbiased Estimator (BLUE)**, in the sense that its variance is the smallest possible variance a linear unbiased estimator can achieve, i.e.:

$$\mathbb{V}[\hat{\boldsymbol{\beta}}_{\boldsymbol{OLS}}] \leq \mathbb{V}[\tilde{\boldsymbol{\beta}}]$$

where $\tilde{\boldsymbol{\beta}}$ is any other linear unbiased estimator.

Inference and hypothesis testing

Single hypothesis testing

- We can test hypotheses about the value of $\beta_0, \beta_1, ..., \beta_K$ individually using the same procedure as before since we have the asymptotic distribution from the CLT: $\mathcal{N}\left(0, \sigma^2(\mathbb{E}[X_i X_i'])^{-1}\right)$
- Say, we want to test the hypothesis $\mathcal{H}_0$: $\beta_1 = \beta_{\mathcal{H}_0}$ vs $\mathcal{H}_1$: $\beta_1 \neq \beta_{\mathcal{H}_0}$
- But we first need to estimate σ^2
- It can be shown that the sum of squared residuals $\hat{\sigma}^2 = \frac{SSR}{n-K-1} = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-K-1}$ is a **consistent estimator** of σ^2
- The $n - K - 1$ is coming from the fact that there are $K + 1$ estimated parameters and we need to divide by $n - (K + 1)$ to get the Student distribution we get for sample means

Single hypothesis testing (cont'd)

- For $(\mathbb{E}[\boldsymbol{X_i X_i'}])^{-1}$, sample estimates for each element in the matrix will do (using $\boldsymbol{n-1}$ is not incorrect, but it makes little difference when $\boldsymbol{n}$ is large, and both estimators are consistent anyway)
- For the variance of $\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}$, we need to look into the $\boldsymbol{k^{th}}$ diagonal element of $\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x_i x_i'}\right)^{-1}$
- So under $\boldsymbol{\mathcal{H}_0}$, $\boldsymbol{t = \sqrt{n}\frac{\hat{\beta}_k - \beta_{\mathcal{H}_0}}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}_{k,k}}} \sim t_{n-K-1}}$
- If the sample size is big enough (typically $\boldsymbol{n > 100}$), then $\boldsymbol{t_{n-K-1}}$ can be replaced by $\boldsymbol{\mathcal{N}(0,1)}$
- As fancy as it looks, even if the bottom term has a matrix in it, we only look at the $\boldsymbol{k^{th}}$ diagonal element to compute the standard error of $\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}$

Single hypothesis testing (cont'd)

- Typically, regressions results look like this

$$\hat{Y}_i = \underset{(se(\hat{\beta}_0))}{\hat{\beta}_0} + \underset{(se(\hat{\beta}_1))}{\hat{\beta}_1} X_{i,1} + \underset{(se(\hat{\beta}_2))}{\hat{\beta}_2} X_{i,2}$$

- The number in parentheses are the **standard errors**, i.e. the estimates of the standard deviations of the $\hat{\beta}$
- They correspond to the square root of the diagonal elements of $\hat{\sigma}^2(\frac{1}{n}\sum_{i=1}^{n} x_i x_i')^{-1}$
- For instance:

$$\widehat{Wage}_i = \underset{(0.1)}{0.67} + \underset{(0.3)}{2.1}\, Age_i - \underset{(0.2)}{3.7}\, nkids_i$$

Single hypothesis testing

- Let us test $\mathcal{H}_0$: $\beta_1 = b$ vs $\mathcal{H}_1$: $\beta_1 \neq b$
- The test statistic would be

$$t = \frac{\hat{\beta}_1 - b}{0.3}$$

- If $b = 0$, the test statistic would equal 7
- The critical value will come from the Normal distribution is the sample size is large enough ($n > 100$) or the Student distribution otherwise (with $n - K - 1$ degrees of freedom)
- For a 5% level test, the critical value from the normal distribution is 1.96
- The p-value will follow from looking at the probability of observing a higher number than the test statistic using the appropriate distribution

Joint hypothesis testing: The F-test

- What if we want to test 2 or more hypotheses at the same time?
- Example: $\boldsymbol{\beta_0 = 5}$ **and** $\boldsymbol{\beta_3 = 12}$
- We need a **joint hypotheses** test
- The **F-test** is used to test several hypotheses at the same time
- It consists in comparing two regressions: One with all the variables of interest (the **unrestricted model**), and one where plug the $\boldsymbol{\mathcal{H}_0}$ values and estimate the betas that don't appear in $\boldsymbol{\mathcal{H}_0}$ (the **restricted model**)
- Not covered in this course

Illustration in **RStudio**

Do Fast-Food Chains Price Discriminate on the Race and Income Characteristics of an Area? (1997)

- Kathryn Graddy studied the impact of the proportion of black people in an area on the price of fast food items
- Idea: KFC restaurants serve the same items everywhere, but not always at the same price. Why?
  - Is it a reflection of the costs associated to serving in one area vs another?
  - Is it due to the composition of the demand? (Rich vs poor households? Black vs white customers)
- In order to answer the question, she used data on item prices from fast food stores across different areas, and regressed these prices on population related variables as well as other economic indicators: Proportion of black people in the area, average income in the area, housing prices in the area, etc

Graddy (1997)

```
model <- lm(pfries ~ prpblck + lincome + prppov + BK + KFC, data = discrim)
summary(model)
```

```
##
## Call:
## lm(formula = pfries ~ prpblck + lincome + prppov + BK + KFC,
##     data = discrim)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.26395 -0.06476 -0.00111  0.05536  0.40641
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.94705    0.35349  -2.679  0.00770 **
## prpblck      0.08368    0.03682   2.272  0.02362 *
## lincome      0.17349    0.03219   5.390 1.23e-07 ***
## prppov       0.43048    0.15857   2.715  0.00693 **
## BK          -0.03199    0.01097  -2.917  0.00374 **
## KFC         -0.08714    0.01368  -6.371 5.35e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09667 on 386 degrees of freedom
##   (18 observations deleted due to missingness)
## Multiple R-squared:  0.1792, Adjusted R-squared:  0.1685
## F-statistic: 16.85 on 5 and 386 DF,  p-value: 4.52e-15
```

Graddy (1997)

- The results above suggests that an increase in the black population in the area of the restaurant leads to an increase in the price of fries in a fast food!
- The proportion of black people is a number between 0 and 1, so looking at a 1 unit increase does not make sense
- A **0.5** increase in the proportion of black people in the area leads to a $\mathbf{0.5 \times 0.08368 = \$0.04184}$ increase in the price of fries)
- Interesting. And results are similar for the price of sodas and of a meal overall. What could explain that?

Graddy (1997)

- We can easily make predictions by extracting the $\hat{\beta}$'s

```
beta_hats <- model$coefficients   # Extracts the beta hats in a vector
# Set some values for the X's
p_black <- 0.1
linc <- 10
prppov <- 0.05
BK <- 1
KFC <- 0
x <- c(1, p_black, linc, prppov, BK, KFC)
pred <- sum(x*beta_hats)
pred <-   beta_hats[1] + beta_hats[2]*p_black+
          beta_hats[3]*linc + beta_hats[4]*prppov+
          beta_hats[5]*BK + beta_hats[6]*KFC
```

Graddy (1997)

- We can easily make predictions by extracting the $\hat{\beta}$'s

```
pred
```

```
## (Intercept)
##    0.785728
```

Graddy (1997)

- It could be pure discrimination, yes. Some stores would want to charge black people more for that reason
- Do black people have a more inelastic demand for fast foods than other people?
- What are the characteristics of these neighborhoods vs other neighborhoods?
- It could also be due to factors that were not measured and correlated with the proportion of black people
- These factors are inside the error term, and will result in a violation of Assumption **3**

Violation of assumptions

Linear model assumption 1: Linearity

- Assumption 1 imposes linearity of the model: $\boldsymbol{Y_i = X_i'\beta + u_i}$
- Linearity in the parameters $\boldsymbol{\beta_0}$, not in $\boldsymbol{X_i}$
- Example: $\boldsymbol{Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,1}^2 + u_i}$ is a linear model
- Example: $\boldsymbol{Y_i = (X_i'\beta_0)^2 + u_i}$ is **not** a linear model as it is not linear in $\boldsymbol{\beta_0}$
- Exact linearity never really exists, but approximate linearity is enough. Visualize data to justify this assumption

Illustration in R: nonlinearities

```
n <- 500                # sample size
x1 <- rnorm(n)            # "draw" n numbers from a normal distribution
x2 <- log(x1^2)
y <-  x2 + rnorm(n)
data <- data.frame(y, x1, x2)
```

Illustration in R: nonlinearities

- If we regress $\boldsymbol{Y_i}$ on $\boldsymbol{X_{i,1}}$:

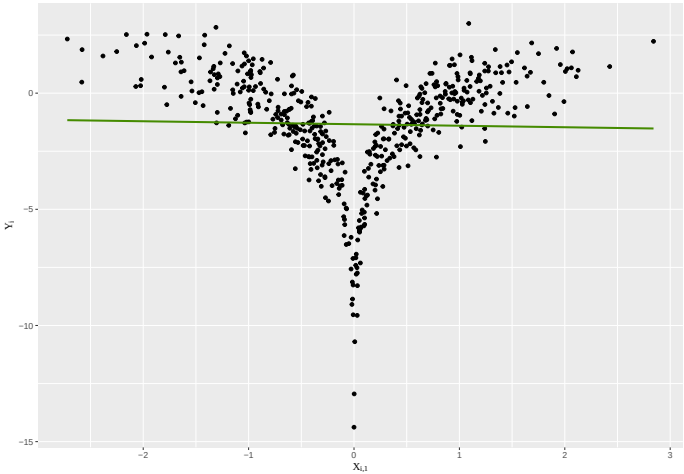


Illustration in R: nonlinearities

- Instead, regress $\boldsymbol{Y_i}$ on $\boldsymbol{\ln(X_{i,1})}$:

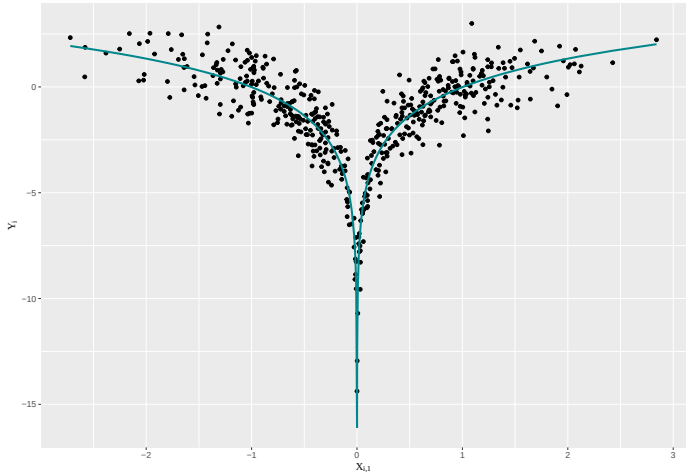


Illustration in R: nonlinearities

- Instead, regress $\boldsymbol{Y_i}$ on polynomials of $\boldsymbol{X_{i,1}}$:

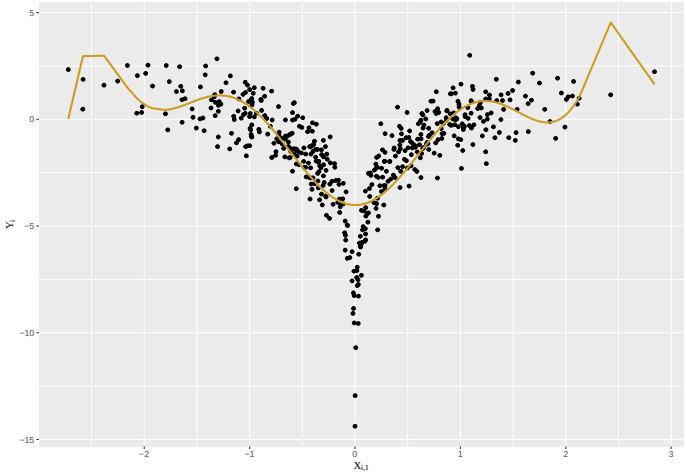
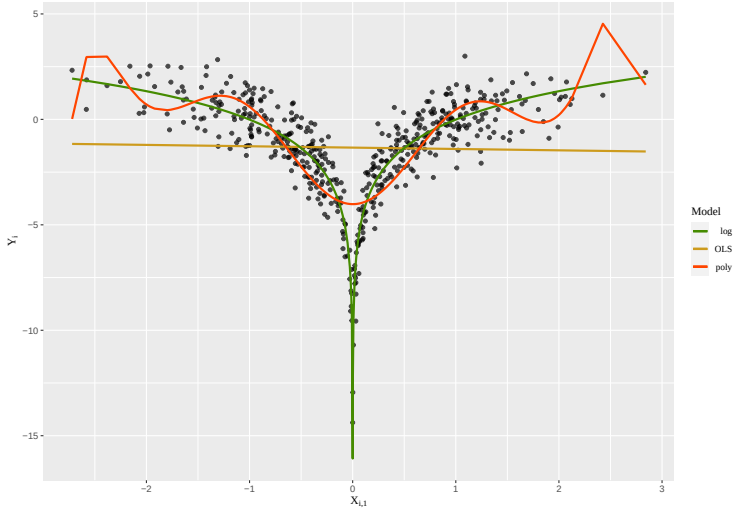


Illustration in R: nonlinearities



Linear model assumption 2: i.i.d sample

- Assumption **2** says the sample is composed of Independent and Identically Distributed (aka i.i.d.) observations $\{x_i, y_i\}, \ i = 1, ..., n$ is realistic if one knows the conditions under which the sample was obtained
- Without it, the OLS estimator might still be **consistent** and **unbiased**, but it requires more technical assumptions for laws of large numbers to apply

Linear model assumption 3: Exogeneity

- Assumption **3** is about the conditional expectation of the error term: $\mathbb{E}[\boldsymbol{u_i}|\boldsymbol{X_i}] = \mathbf{0} \quad \forall \boldsymbol{i} = \mathbf{1}, ..., \boldsymbol{n}$
- If not satisfied, it roughly means some $\boldsymbol{X_{i,j}}$ are somehow correlated with $\boldsymbol{u_i}$: These $\boldsymbol{X_{i,j}}$ are said to be **endogenous**, as opposed to **exogenous**
- It could happen for various reasons, among which **omitted variables** which are "buried" in the error term and correlated with $\boldsymbol{X_i}$
- Consequence: The OLS estimator is not **unbiased** anymore, nor it is **consistent**
- How do we estimate $\boldsymbol{\beta}$ in a reliable way then?

The Two Stage Least Squares (2SLS) estimator

- Consider the following model:

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_{i,1} + ... + \beta_K X_{i,K} + u_i}$$

- Say $\boldsymbol{X_{i,1}}$ is **endogenous**, and all the other covariates are **exogenous**
- Estimating $\boldsymbol{\beta_0, \beta_2, ..., \beta_K}$ consistently is doable via OLS
- To get reliable estimates for $\boldsymbol{\beta_1}$, we need to satisfy Assumption **3** somehow
- Consider a variable $\boldsymbol{Z_i}$ with the following two features:
  - $\mathbb{E}[\boldsymbol{u_i}|\boldsymbol{Z_i}] = \boldsymbol{0}$: $\boldsymbol{Z_i}$ is **exogenous**
  - $\boldsymbol{Z_i}$ is correlated with $\boldsymbol{X_{i,k}}$ but does not directly affect $\boldsymbol{Y_i}$: $\boldsymbol{Z_i}$ is **relevant**
- We can use $\boldsymbol{Z_i}$ to "represent" $\boldsymbol{X_{i,k}}$ while being exogenous. $\boldsymbol{Z_i}$ is called an **instrument**

The 2SLS estimator: First stage

- In order to rid $\boldsymbol{X_{i,1}}$ of its endogeneity, consider the following regression:

$$\boldsymbol{x_{i,1} = \pi_0 + \pi_1 z_i + \pi_2 x_{i,2} + ... + \pi_K x_{i,K} + v_i}$$

- This model is called the **first stage equation**, or **reduced form equation**
- Note: $\boldsymbol{X_{i,1}}$ is expressed as a combination of **exogenous** variables, so the source of the endogeneity of $\boldsymbol{X_{i,1}}$ comes from the correlation between the error terms $\boldsymbol{u_i}$ and $\boldsymbol{v_i}$
- Thus $\mathbb{E}[\boldsymbol{v_i | Z_i, X_{i,2}, ..., X_{i,K}}] = \boldsymbol{0}$ is naturally satisfied (they are exogenous for the first stage as well)
- Estimating this model via OLS yields the predictions $\boldsymbol{\hat{x}_{i,1}}$ which are exogenous!

The 2SLS estimator: Second stage

- Consider now the following regression equation:

$$\boldsymbol{Y_i = \beta_0 + \beta_1 \hat{X}_{i,1} + ... + \beta_K X_{i,K} + u_i}$$

- This model is called the **second stage equation**, or **structural form equation**
- Estimate this model via OLS, and the estimates are **consistent** and **asymptotically normal**
- However, they are **biased**. But for a big enough sample, the bias is negligible
- The final estimates were obtained from estimating two stages, hence the name **Two Stage Least Squares (2SLS or TSLS)**

The 2SLS estimator: Procedure

Algorithm: The 2SLS estimator

- Regress the endogenous variable on all the exogenous variables + instrument(s) via OLS. Get the predictions of the endogenous variable
- Regress the dependent variable $\boldsymbol{Y_i}$ on the predictions from the first stage + all the exogenous variables via OLS to obtain $\hat{\boldsymbol{\beta}}_{\boldsymbol{2SLS}}$

Remarks:

- The 2SLS estimates will have higher standard errors than the OLS ones. Intuitively, it comes from adding extra variation due the inclusion of the first stage predictions
- The $\boldsymbol{R^2}$ from a 2SLS estimation can be **negative** and thus difficult to interpret/use
- The OLS estimate for the endogenous variable is typically biased in the direction of the correlation between the endogenous variable and the error term

Endogeneity: Illustration in R

```r
n <- 100   # sample size
z <- rnorm(n)
u <- rnorm(n, mean = 0, sd = 1)  # generating the error term. Normal distribution here
v <- rnorm(n, mean = 0, sd = 1)  # generating the error term. Normal distribution here
x2 <- rnorm(n)    # An exogenous variable correlated with x1 but not the error term
x1 <- z + u + 0.5*x2 + v    # X depends on z and x2 but also epsilon
x2 <- rnorm(n)    # An exogenous variable correlated with x1 but not the error term
beta_0 <- c(1, 2, 3)   # an intercept and a coefficient on x.
y <- cbind(1, x1, x2)%*%beta_0 + u  # here we are creating the dependent variable Y
data <- data.frame(y, x1, x2, z)
```

Endogeneity: Illustration in R (OLS in small sample)

```r
ols <- lm(y ~ x1 + x2) # Regression with an intercept
summary(ols)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.84377 -0.60845  0.09258  0.61241  1.62297
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.98551    0.08386   11.75   <2e-16 ***
## x1           2.33034    0.04906   47.50   <2e-16 ***
## x2           3.08082    0.09300   33.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8383 on 97 degrees of freedom
## Multiple R-squared:  0.9674, Adjusted R-squared:  0.9667
## F-statistic:  1438 on 2 and 97 DF,  p-value: < 2.2e-16
```

Endogeneity: Illustration in R (OLS in big sample)

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1437 -0.5601  0.0127  0.5629  3.4046
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.003923   0.008455   118.7   <2e-16 ***
## x1          2.308181   0.004721   488.9   <2e-16 ***
## x2          3.005893   0.008573   350.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8454 on 9997 degrees of freedom
## Multiple R-squared:  0.9731, Adjusted R-squared:  0.9731
## F-statistic: 1.806e+05 on 2 and 9997 DF,  p-value: < 2.2e-16
```

Endogeneity: Illustration in R (2SLS in 2 steps)

```r
stage_1 <- lm(x1 ~ x2 + z, data = data)
x1_hat <- stage_1$fitted
stage_2 <- lm(y ~ x1_hat + x2, data = data)
summary(stage_2)
```

```
##
## Call:
## lm(formula = y ~ x1_hat + x2, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.6195  -2.5287   0.0633   2.4979  13.3412
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.99713    0.03732   26.72   <2e-16 ***
## x1_hat       1.99385    0.03775   52.82   <2e-16 ***
## x2           3.00458    0.03783   79.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.731 on 9997 degrees of freedom
## Multiple R-squared:  0.4755, Adjusted R-squared:  0.4754
## F-statistic:  4531 on 2 and 9997 DF,  p-value: < 2.2e-16
```

Endogeneity: Illustration in R (2SLS in one step with ivreg)

```
library(AER)
```

```
tsls <- ivreg(y ~ x1 + x2 | z + x2, data = data)
summary(tsls)
```

```
##
## Call:
## ivreg(formula = y ~ x1 + x2 | z + x2, data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -4.087534 -0.684820  0.005748  0.681170  3.957158
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.99713    0.01016   98.14   <2e-16 ***
## x1           1.99385    0.01028  194.03   <2e-16 ***
## x2           3.00458    0.01030  291.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.016 on 9997 degrees of freedom
## Multiple R-Squared: 0.9611,  Adjusted R-squared: 0.9611
```

Linear model assumption 4: No collinearity

- Assumption 4 says that no $\boldsymbol{X_{i,j}}$ can be expressed as a linear combination of the other $\boldsymbol{X_i}$'s
- If unsatisfied, it means two explanatory variables (or more) are linearly related: They are said to be **collinear**
- Example: $\boldsymbol{X_{i,1} = aX_{i,2} + bX_{i,3}}$
- Including $\boldsymbol{X_{i,1}}$, $\boldsymbol{X_{i,2}}$ and $\boldsymbol{X_{i,3}}$ in the same regression is redundant as $\boldsymbol{X_{i,2}}$ and $\boldsymbol{X_{i,3}}$ contain enough information for $\boldsymbol{X_{i,1}}$
- In practice: The software returns a numerical error (**perfect collinearity**), or standard errors are very big (**imperfect collinearity**)
- With many variables, collinearity is more likely. Be careful!

Multicollinearity: The dummy variable trap

- Qualitative variables sometimes have more than 2 possible values: red/green/blue, white/black/asian,...
- One might want to create a binary variable for each possible value and include them in the regression
- Problem: The sum of these binary variables is always equal to 1! They are collinear
- So if there are $\boldsymbol{p}$ different categories, include $\boldsymbol{p-1}$ binary variables!
- The one not included in the default category, i.e the reference to which all the other binary variables compare

Multicollinearity: Polynomial regression

- We saw how adding polynomial terms can help capture non linearities and improve predictions
- One might want to add many polynomial terms to make sure they don't miss any curvature
- Problem: As one adds polynomial terms, each extra term is close to being explained by a linear combination of the previous terms
- So do not include too many terms! Use t-tests to check if the extra ones are relevant

Collinearity: Illustration in R

```
n <- 100                # sample size
x1 <- rnorm(n)            # "draw" n numbers from a normal distribution
x2 <- rnorm(n)            # "draw" n numbers from a normal distribution
x3 <- 0.5*x1 - 3*x2
y <-  x1 + x2 + rnorm(n)
data <- data.frame(y, x1, x2, x3)
```

Collinearity: Illustration in R

```
ols_123 <- lm(y ~ x1 + x2 + x3, data = data)
summary(ols_123)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.77902 -0.66465 -0.08205  0.61677  2.37013
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.02211    0.09455  -0.234    0.816
## x1           1.04622    0.09391  11.140   <2e-16 ***
## x2           0.98733    0.09218  10.711   <2e-16 ***
## x3                NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9392 on 97 degrees of freedom
## Multiple R-squared:  0.7085, Adjusted R-squared:  0.7025
```

Collinearity: Illustration in R

```
ols_23 <- lm(y ~ x2 + x3, data = data)
summary(ols_23)
```

```
##
## Call:
## lm(formula = y ~ x2 + x3, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.77902 -0.66465 -0.08205  0.61677  2.37013
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.02211    0.09455  -0.234    0.816
## x2           7.26466    0.57216  12.697   <2e-16 ***
## x3           2.09244    0.18783  11.140   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9392 on 97 degrees of freedom
## Multiple R-squared:  0.7085, Adjusted R-squared:  0.7025
## F-statistic: 117.9 on 2 and 97 DF,  p-value: < 2.2e-16
```

Collinearity: Illustration in R

```
ols_13 <- lm(y ~ x1 + x3, data = data)
summary(ols_13)
```

```
##
## Call:
## lm(formula = y ~ x1 + x3, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.77902 -0.66465 -0.08205  0.61677  2.37013
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.02211    0.09455  -0.234    0.816
## x1           1.21078    0.09536  12.697   <2e-16 ***
## x3          -0.32911    0.03073 -10.711   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9392 on 97 degrees of freedom
## Multiple R-squared:  0.7085, Adjusted R-squared:  0.7025
## F-statistic: 117.9 on 2 and 97 DF,  p-value: < 2.2e-16
```

Collinearity: Illustration in R

```
ols_12 <- lm(y ~ x1 + x2, data = data)
summary(ols_12)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.77902 -0.66465 -0.08205  0.61677  2.37013
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.02211    0.09455  -0.234    0.816
## x1           1.04622    0.09391  11.140   <2e-16 ***
## x2           0.98733    0.09218  10.711   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9392 on 97 degrees of freedom
## Multiple R-squared:  0.7085, Adjusted R-squared:  0.7025
## F-statistic: 117.9 on 2 and 97 DF,  p-value: < 2.2e-16
```

Detecting collinearity and fixing it

There are many ways to detect multicollinearity. Some of them include:

- Large changes in the estimated coefficients when adding or removing a covariate
- Insignificant variables in a multivariate regression, but significant variables when estimating univariate regressions
- Look at $\boldsymbol{R_k^2}$, the $\boldsymbol{R^2}$ from regressing $\boldsymbol{X_{i,k}}$ on all the other covariates and compute the **Variance Inflation Factor** $\boldsymbol{VIF_k = \frac{1}{1-R_k^2}}$. If $\boldsymbol{VIF_k > 10}$ for some $\boldsymbol{k}$, there is a multicollinearity issue

How to fix the problem?

- Drop some variables (careful which, as you might introduce an omitted variable bias)
- Get more data: Collinearity is more likely is small samples, and bigger samples
- Use alternative estimation methods: **Ridge regression**, **partial least squares**, **Principal Component Analysis** (more on them later)

Linear model assumption 5: Homoskedasticity

- Assumption **5** imposes the same variance on the error term, i.e. the variance does not change with the values of the $\boldsymbol{X}_{i,j}$
- If unsatisfied, we then say that errors are **heteroskedastic**, as opposed to **homoskedastic**
- If the other assumptions are satisfied, the OLS estimator is still **consistent** and **unbiased**
- But the OLS estimator loses an appealing property: its variance is not the smallest possible one: It is not **BLUE** anymore
- Consequence: Confidence intervals produced by the OLS estimator are narrower than the true ones, leading to potentially incorrect conclusions
- What can we do?

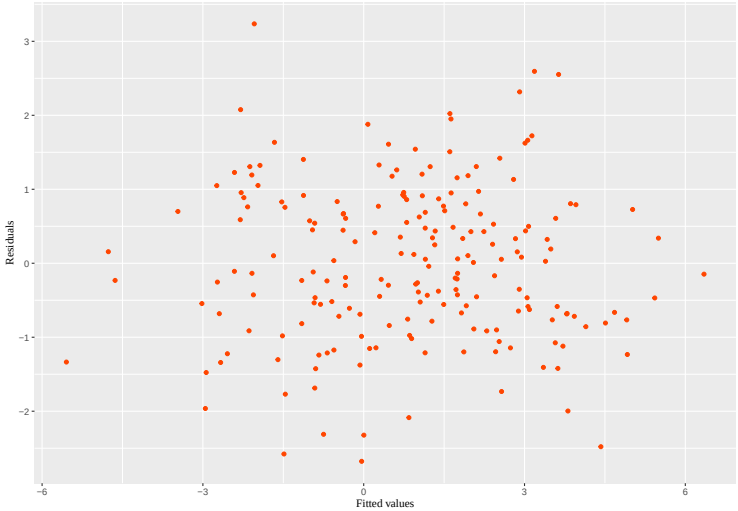
Detecting heteroskedasticity

- One can detect heteroskedasticity by plotting the residuals of an OLS estimation on the fitted values
- If there is no pattern in the variance of the residuals, then there is no evidence of heteroskedasticity
- If residuals display more variance for different values of the fitted values, chances are there is heteroskedasticity
- Two hypotheses tests can be performed as well: Breusch-Pagan and White test

Homoskedasticity: Illustration in R

```
n <- 200  # sample size
x <- rnorm(n)  # generating the explanatory variable. Normal distribution here
u <- rnorm(n)
beta_0 <- c(1, 2)   # an intercept and a coefficient on x.
y <- cbind(1, x)%*%beta_0 + u  # here we are creating the dependent variable Y
ols <- lm(y ~ x)
fit <- ols$fitted
resi <- ols$residuals
data <- data.frame(y, x, fit, resi)
```

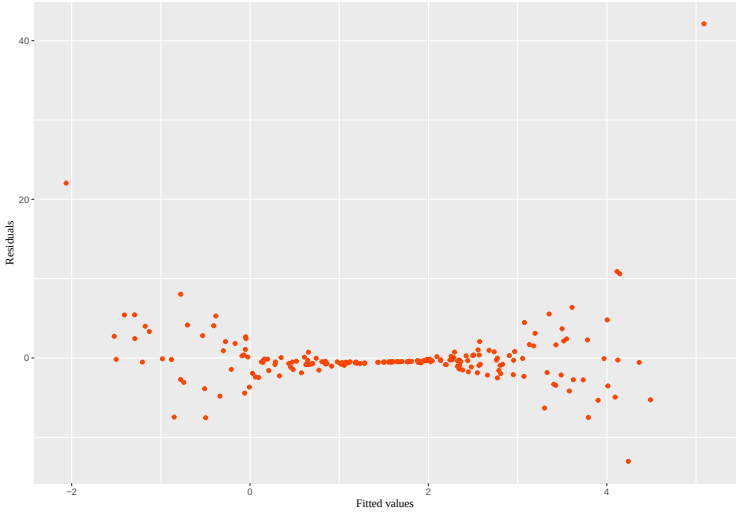
Homoskedasticity: Illustration in R



Detecting heteroskedasticity: Illustration in R

```r
n <- 200  # sample size
x <- rnorm(n)  # generating the explanatory variable. Normal distribution here
sigma <- 2
u <- c()
for(i in 1:n)
{
  u[i] <- rnorm(1, mean = 0, sd = sigma*x[i]^2)
}
beta_0 <- c(1, 2)   # an intercept and a coefficient on x.
y <- cbind(1, x)%*%beta_0 + u  # here we are creating the dependent variable Y
ols <- lm(y ~ x)
fit <- ols$fitted
resi <- ols$residuals
data <- data.frame(y, x, fit, resi)
```

Detecting heteroskedasticity: Illustration in R



Detecting heteroskedasticity with test statistics

- Assumption 5 says that

$$\mathbb{V}[\boldsymbol{u}_i|\boldsymbol{X}_i] = \boldsymbol{\sigma^2}$$

  and remember that

$$\mathbb{V}[\boldsymbol{u}_i|\boldsymbol{X}_i] = \mathbb{E}[\boldsymbol{u}_i^2|\boldsymbol{X}_i] - (\mathbb{E}[\boldsymbol{u}_i|\boldsymbol{X}_i])^2 = \mathbb{E}[\boldsymbol{u}_i^2|\boldsymbol{X}_i]$$

- We would like to test the null hypothesis $\boldsymbol{\mathcal{H}_0 : \mathbb{V}[u_i|X_i] = \sigma^2}$
- And the point of a regression is to estimate $\mathbb{E}[\boldsymbol{Y}_i|\boldsymbol{X}_i = \boldsymbol{x}_i] = \boldsymbol{\beta_0 + \beta_1 x_i}$
- It suggests the following regression:

$$\boldsymbol{u_i^2 = \delta_0 + \delta_1 X_{i,1} + ... + \delta_K X_{i,K} + e_i}$$

- We don't know $\boldsymbol{u_i^2}$... but we can get $\boldsymbol{\hat{u}_i^2}$, a kind of estimate!
- That is the gist of the next two tests

Detecting heteroskedasticity: The Breusch-Pagan test

- If there is heteroskedasticity, the variance of the error term is a function of the covariates
- We do not know the variance, but we have used residuals $\hat{u}_i^2$ in the past to estimate it
- The idea of the Breusch-Pagan test is to regress $\hat{u}_i^2$ on the covariates, and test the joint significance of the regression
- If we reject the null hypothesis, we have heteroskedasticity

Detecting heteroskedasticity: The Breusch-Pagan test (cont'd)

Algorithm: Breusch-Pagan test (1970)

- Estimate the model via OLS, get the residuals $\hat{u}_i$
- Regress $\hat{u}_i^2$ on $X_{i,1}$, ..., $X_{i,K}$
- Perform a F-test on all the covariates (5% is a good default significance level)
- If the null is rejected, there is evidence of heteroskedasticity

- Note: If we suspect the variance of u_i only depends on a subset of covariates, it is possible to only include the suspects in the regression

Detecting heteroskedasticity: The White test

- White suggested using the squares and the interactions of the covariates
- If there are two covariates, include $\boldsymbol{X_{i,1}}$, $\boldsymbol{X_{i,2}}$ but also $\boldsymbol{X_{i,1}^2}$, $\boldsymbol{X_{i,2}^2}$ and $\boldsymbol{X_{i,1}X_{i,2}}$
- If there are three covariates, there would be 9 terms!
- It gets quickly out of hands, and eats degrees of freedom away...
- But $\boldsymbol{\hat{y}_i}$ is itself a function of the covariates!

Detecting heteroskedasticity: The White test (cont'd)

- It suggests we can use $\hat{y}_i$ and $\hat{y}_i^2$ and check their significance:

$$\hat{u}_i^2 = \alpha_0 + \alpha_1 \hat{y}_i + \alpha_2 \hat{y}_i^2 + \varepsilon_i$$

- Way more tractable than using all these interaction terms!
- It is a modified version of the test with interactions and imposes more restrictions, but is still useful

Detecting heteroskedasticity: The White test (cont'd)

Algorithm: White test (1980)

- Estimate the model via OLS, get the residuals $\hat{u}_i$ and the fitted values $\hat{y}_i$
- Regress $\hat{u}_i^2$ on $\hat{y}_i$ and $\hat{y}_i^2$
- Use a F-test to test the joint significance of $\hat{y}_i$ and $\hat{y}_i^2$ (5% is a good default significance level)
- If the null is rejected, there is evidence of heteroskedasticity

Dealing with known heteroskedasticity

- To see how to fix the heteroskedasticity problem, consider the following univariate model:

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_i + u_i}$$

  where Assumptions **1**, **2**, **3** and **4** are satisfied

- Assume $\mathbb{V}[\boldsymbol{u_i}|\boldsymbol{x_i}] = \boldsymbol{\sigma^2 f(x_i)}$

- Note: This assumption means that the error terms are **independent** but **not identically distributed** as variances differ across $\boldsymbol{i}$

- Say, we know what $\boldsymbol{f(X_i)}$ is. Divide both sides of the regression equation by $\sqrt{\boldsymbol{f(X_i)}}$ and define

$$\boldsymbol{Y_i^* = \frac{Y_i}{\sqrt{f(X_i)}},\ X_i^* = \frac{X_i}{\sqrt{f(X_i)}},\ W_i^* = \frac{1}{\sqrt{f(X_i)}},\ u_i^* = \frac{u_i}{\sqrt{f(X_i)}}}$$

- The last variable $\boldsymbol{W_i^*}$ is the intercept, $\mathbf{1}$, divided by $\sqrt{\boldsymbol{f(X_i)}}$

Dealing with known heteroskedasticity (cont'd)

- The transformed model is:

$$\boldsymbol{Y_i^* = \beta_0 W_i^* + \beta_1 X_i^* + u_i^*}$$

- Now: $\mathbb{V}[\boldsymbol{u_i^*}|\boldsymbol{X_i}] = \mathbb{V}\left[\frac{\boldsymbol{u_i}}{\sqrt{\boldsymbol{f(X_i)}}}|\boldsymbol{X_i}\right] = \frac{\boldsymbol{1}}{\boldsymbol{f(x_i)}}\mathbb{V}[\boldsymbol{u_i}|\boldsymbol{X_i}] = \frac{\boldsymbol{1}}{\boldsymbol{f(x_i)}}\boldsymbol{\sigma^2 f(x_i) = \sigma^2}$
- Back to homoskedasticity!!
- Regress $\boldsymbol{Y_i^*}$ on $\boldsymbol{W_i^*}$ and $\boldsymbol{X_i^*}$ via OLS to get **consistent**, **unbiased** and **BLUE** estimates of $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$
- Note: You can add an intercept, but it will not be significant anyway
- The resulting estimator is called the **Generalized Least Squares (GLS)** estimator

Dealing with known heteroskedasticity (cont'd)

- In a multivariate context, the principle is the same: Divide everything by the standard deviation of $\boldsymbol{u_i}$
- If error terms are correlated (for instance, in time series regressions where data are not i.i.d.), we need to take into account covariance terms
- The variance of the error term vector is a matrix containing variances in the diagonal, and covariances off the diagonal (not covered in this course)
- Because at the end of the day, it is an OLS on variables that were re-weighted, that particular GLS estimator is called **Weighted Least Squares (WLS)**

Dealing with unknown heteroskedasticity

- What if we know there is heteroskedasticity, but we don't know the variance of the error term?
- It is a more realistic case unfortunately...
- Remember that

$$\mathbb{V}[\boldsymbol{u}_i|\boldsymbol{X}_i] = \mathbb{E}[\boldsymbol{u}_i^2|\boldsymbol{X}_i] - (\mathbb{E}[\boldsymbol{u}_i|\boldsymbol{X}_i])^2 = \mathbb{E}[\boldsymbol{u}_i^2|\boldsymbol{X}_i] = \boldsymbol{\sigma}^2(\boldsymbol{x}_i)$$

- And the point of a regression is to estimate $\mathbb{E}[\boldsymbol{Y}_i|\boldsymbol{X}_i = \boldsymbol{x}_i] = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\boldsymbol{x}_i$
- Using $\hat{\boldsymbol{u}}_i^2$ as the dependent variable (an "estimate" of $\boldsymbol{u}_i^2$), we can estimate $\mathbb{E}[\boldsymbol{u}_i^2|\boldsymbol{X}_i]$
- Thus, consider the following model:

$$\hat{\boldsymbol{u}}_i^2 = \boldsymbol{\alpha_0} + \boldsymbol{\alpha_1}\boldsymbol{X}_{i,1} + ... + \boldsymbol{\alpha_K}\boldsymbol{X}_{i,K} + \boldsymbol{\varepsilon}_i$$

Dealing with unknown heteroskedasticity

$$\hat{u}_i^2 = \alpha_0 + \alpha_1 X_{i,1} + ... + \alpha_K X_{i,K} + \varepsilon_i$$

- Note: we care about good prediction here, not about the interpretation of the estimated coefficients, so we don't have to worry about endogeneity issues
- We can use the fitted values of this regression as an estimate of $\mathbb{E}[u_i|X_i]$
- Problem: The fitted values might be negative, so dividing all variables by its square root would not work...

Dealing with unknown heteroskedasticity (cont'd)

- Instead, consider:

$$\ln(\hat{u}_i^2) = \delta_0 + \delta_1 X_{i,1} + ... + \delta_K X_{i,K} + e_i$$

- From that regression, we can recover the fitted values for $\hat{u}_i^2$ by taking the exponential, and an exponential is never negative!!
- What if we pick the wrong functional form?
- We want to predict well, so we want a **flexible** function, i.e. one that can approximate potential non linearities
- So we can add polynomial terms! (as long as we don't have collinearity)

Dealing with unknown heteroskedasticity

Algorithm: Feasible Generalized Least Squares (FGLS)

- Estimate the model via OLS, get the residuals $\hat{u}_i$
- Regress $\ln(\hat{u}_i^2)$ on $X_{i,1}, ..., X_{i,K}$ (and polynomial terms if desired), and get the fitted values $\widehat{\ln(\hat{u}_i^2)}$
- Get $\hat{\hat{u}}_i^2 = \exp(\widehat{\ln(\hat{u}_i^2)})$
- Get $Y_i^* = Y_i / \sqrt{\hat{\hat{u}}_i^2}$, $X_{i,1}^* = X_{i,1} / \sqrt{\hat{\hat{u}}_i^2}$ etc
- Regress Y_i^* on the X_i^*'s via OLS

- Note 1: Careful with F-tests after FGLS! Use the same weights in both the unrestricted and restricted regressions
- Note 2: If the functional form of the variance is not as the regression involving $\ln(\hat{u}_i^2)$, the standard errors obtained at the end are no longer valid, but we can always use the heteroskedasticity-robust standard error (see next slides)

Feasible GLS (FGSL): Illustration in R

```r
# Step 1
ols <- lm(y ~ x)
resi <- ols$residuals
resi2 <- resi^2
l_resi2 <- log(resi2)
# Adding With polynomial terms
x2 <- x^2
x3 <- x^3
# Step 2
res_model <- lm(l_resi2 ~ x + x2 + x3)
# step 3
res_fit <- exp(res_model$fitted)
```

Feasible GLS (FGSL): Illustration in R

```r
# Step 4
ystar <- y/sqrt(res_fit)
xstar <- x/sqrt(res_fit)
int <- 1/sqrt(res_fit)  # gotta transform the intercept too!
# Step 5
fgls_model <- lm(ystar ~  int + xstar)
summary(fgls_model)
```

```
##
## Call:
## lm(formula = ystar ~ int + xstar)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1485 -0.6724  0.0749  0.6675  4.7651
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.4788     0.2789   1.717 0.087575 .
## int           0.7319     0.1904   3.843 0.000164 ***
## xstar         1.6202     0.1372  11.809  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.611 on 197 degrees of freedom
## Multiple R-squared:  0.4552, Adjusted R-squared:  0.4496
## F-statistic: 82.29 on 2 and 197 DF,  p-value: < 2.2e-16
```

Feasible GLS (FGSL): Illustration in R

```r
# Step 1
ols <- lm(y ~ x)
resi <- ols$residuals
resi2 <- resi^2
l_resi2 <- log(resi2)
# FGLS With other flexible terms
x2 <- sin(x)
x3 <- cos(x)
# Step 2
res_model <- lm(l_resi2 ~ x + x2 + x3)
# Step 3
res_fit <- exp(res_model$fitted)
```

Feasible GLS (FGSL): Illustration in R

```r
# Step 4
ystar <- y/sqrt(res_fit)
xstar <- x/sqrt(res_fit)
int <- 1/sqrt(res_fit)  # gotta transform the intercept too!
# Step 5
fgls_model <- lm(ystar ~  int + xstar)
summary(fgls_model)
```

```
##
## Call:
## lm(formula = ystar ~ int + xstar)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6091 -0.7102  0.0836  0.6584  5.1679
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.4880     0.2447   1.995   0.0475 *
## int           0.7408     0.1610   4.600 7.54e-06 ***
## xstar         1.6154     0.1350  11.964  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.532 on 197 degrees of freedom
## Multiple R-squared:  0.4774, Adjusted R-squared:  0.4721
## F-statistic: 89.97 on 2 and 197 DF,  p-value: < 2.2e-16
```

Dealing with unknown heteroskedasticity: Robust standard errors

- Several methods can be used to deal with heteroskedasticity
- One consists in doing a first regression to obtain weights used on the observations in a second OLS regression: It is called Generalized Least Squares, or GLS
- Another alternative is to run OLS as usual (it is still unbiased and consistent after all), but to modify the formula for the variance of $\hat{\boldsymbol{\beta}}$ so that it takes heteroskedasticity into account
- That formula is valid **whether** there is heteroskedasticity or not
- This is why it is called the **heteroskedasticity-robust standard error**
- Proposed by Eicker (1967), Huber (1967) and White (1980) separately, they also go by the names **Eicker–Huber–White** or **White** standard errors

Robust standard errors: Illustration in R

- The ***sandwich*** package provides several formulas for different types of heteroskedasticity
- The name of the package comes from the fact that in multivariate regressions, the variance-covariance matrix has a sandwich aspect: $\mathbf{\Gamma\Omega\Gamma'}$ where each term is a matrix
- Here, we need to use ***vcovHC()*** (in the brackets, one must put the estimated model. I called mine "ols")
- Results are slightly different as there are additional adjustments with the package

Robust standard errors: Illustration in R

```
# Homoskedastic case
sqrt( sandwich::vcovHC(ols, type = "const") )
```

```
## Warning in sqrt(sandwich::vcovHC(ols, type = "const")): NaNs produced
```

```
##             (Intercept)         x
## (Intercept)   0.3035608       NaN
## x                 NaN 0.305091
```

```
# Heteroskedastic case
# Heteroskedasticity-robust variance-covariance matrix
sqrt( vcovHC(ols) )
```

```
##             (Intercept)         x
## (Intercept)   0.3011662 0.2904117
## x             0.2904117 0.6687603
```

Summary

- We learnt how linear regression works when there are multiple variables
- As long as the assumptions are satisfied, we can make reliable estimates of the coefficients of interest $\boldsymbol{\beta}$
- What about prediction?
- Do we actually need consistent and unbiased estimates $\hat{\boldsymbol{\beta}}$?
- In fact, it is possible to improve prediction **by introducing bias**
- So linear models are not so much used for prediction, although they are capable of producing some
- They impose a rigid functional form for $\boldsymbol{f(X_{i,1}, ..., X_{i,K})}$ that will limit predictive power