Crash Course:
Bases de Estadística y Econometría

Paúl J. Corcuera

Informática par Economistas
Universidad de Piura
Ciclo 2024-II

**Nota basada en material del curso de Jonathan Roth y Peter Hull en Brown University.*

¿Así que quieres saber de machine learning? Primero las bases

¿Qué es Econometría?

$\rightarrow$ Es la caja de herramientas estadísticas que usamos los economistas para responder preguntas economicas usando data

Algunas preguntas que nos interesan:

¿Qué es Econometría?

$\rightarrow$ Es la caja de herramientas estadísticas que usamos los economistas para responder preguntas economicas usando data

Algunas preguntas que nos interesan:

- ¿Ha aumentado la desigualdad económica desde los años 60?
- ¿Cómo afecta el aumento de salario mínimo en el empleo?
- ¿Cuál va a ser la tasa de desempleo el otro año?

¿Qué es Econometría?

$\rightarrow$ Es la caja de herramientas estadísticas que usamos los economistas para responder preguntas economicas usando data

Algunas preguntas que nos interesan:

- ¿Ha aumentado la desigualdad económica desde los años 60?
  - **Preg. Descriptiva:** pregunta cómo las cosas son (o fueron) en la realidad
- ¿Cómo afecta el aumento de salario mínimo en el empleo?
- ¿Cuál va a ser la tasa de desempleo el otro año?

¿Qué es Econometría?

$\rightarrow$ Es la caja de herramientas estadísticas que usamos los economistas para responder preguntas economicas usando data

Algunas preguntas que nos interesan:

- ¿Ha aumentado la desigualdad económica desde los años 60?
    - **Preg. Descriptiva:** pregunta cómo las cosas son (o fueron) en la realidad
- ¿Cómo afecta el aumento de salario mínimo en el empleo?
    - **Preg. Causal:** ¿Qué hubiera pasado en un mundo contrafactual?
- ¿Cuál va a ser la tasa de desempleo el otro año?

¿Qué es Econometría?

$\rightarrow$ Es la caja de herramientas estadísticas que usamos los economistas para responder preguntas economicas usando data

Algunas preguntas que nos interesan:

- ¿Ha aumentado la desigualdad económica desde los años 60?
  - **Preg. Descriptiva:** pregunta cómo las cosas son (o fueron) en la realidad
- ¿Cómo afecta el aumento de salario mínimo en el empleo?
  - **Preg. Causal:** ¿Qué hubiera pasado en un mundo contrafactual?
- ¿Cuál va a ser la tasa de desempleo el otro año?
  - **Preg. Predicción:** ¿Qué pasará el otro año?

¿Qué es Econometría?

→ Es la caja de herramientas estadísticas que usamos los economistas para responder preguntas economicas usando data

Algunas preguntas que nos interesan:

- ¿Ha aumentado la desigualdad económica desde los años 60?
  - **Preg. Descriptiva:** pregunta cómo las cosas son (o fueron) en la realidad
- ¿Cómo afecta el aumento de salario mínimo en el empleo?
  - **Preg. Causal:** ¿Qué hubiera pasado en un mundo contrafactual?
- ¿Cuál va a ser la tasa de desempleo el otro año?
  - **Preg. Predicción:** ¿Qué pasará el otro año?

Por lo general, los economistas nos concentramos en las primeras dos, con un énfasis en preguntas causales

Why is answering these questions hard?

- For descriptive Qs: we only observe data for a **sample** of individuals, not for the full **population**
    - Example: we want to know how the distribution of income in the US has changed. But we only observe income for a survey of workers

Why is answering these questions hard?

- For descriptive Qs: we only observe data for a **sample** of individuals, not for the full **population**
  - Example: we want to know how the distribution of income in the US has changed. But we only observe income for a survey of workers

- Best case scenario:
  Our sample is **randomly** selected from the population
  - E.g., the workers in our survey were drawn out of hat with names of all possible workers
  - If so, need to account for the fact that by chance the sample might have different characteristics from the population

Why is answering these questions hard?

- For descriptive Qs: we only observe data for a **sample** of individuals, not for the full **population**
  - Example: we want to know how the distribution of income in the US has changed. But we only observe income for a survey of workers

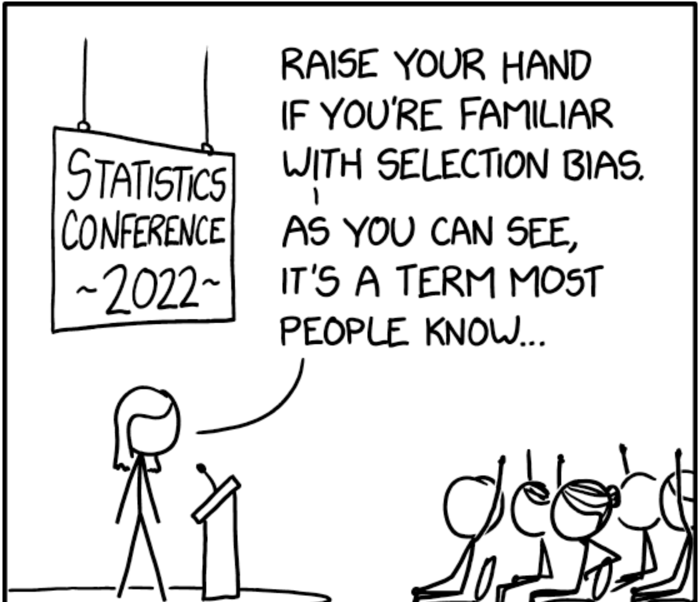
- Best case scenario:
  Our sample is **randomly** selected from the population
  - E.g., the workers in our survey were drawn out of hat with names of all possible workers
  - If so, need to account for the fact that by chance the sample might have different characteristics from the population

- Worst case scenario: our sample is *not representative* of the population that we care about
  - E.g., workers with certain characteristics were more likely to respond to the survey

- In 1948, Chicago Tribune writes that Thomas Dewey defeats Harry Truman in the 1948

- In 1948, Chicago Tribune writes that Thomas Dewey defeats Harry Truman in the 1948

Selection bias referes to settings like Dewey-Truman where the sample is not drawn randomly from the population of interest



Why is answering these questions hard? (Part II)

- Answering causal questions is often *even harder* than descriptive ones. Why?

Why is answering these questions hard? (Part II)

- Answering causal questions is often *even harder* than descriptive ones. Why?
- Causal Qs involve both a descriptive component (what are outcomes in reality?) and a *counterfactual* component (how would things have been under a different treatment?)

Why is answering these questions hard? (Part II)

- Answering causal questions is often *even harder* than descriptive ones. Why?
- Causal Qs involve both a descriptive component (what are outcomes in reality?) and a *counterfactual* component (how would things have been under a different treatment?)
- Example: what is the causal effect on your earnings of going to Brown instead of URI?
  - Descriptive Q: how much do Brown students earn after graduation?
  - Counterfactual Q: how much would Brown students have earned if they went to URI?

Why is answering these questions hard? (Part II)

- Answering causal questions is often *even harder* than descriptive ones. Why?
- Causal Qs involve both a descriptive component (what are outcomes in reality?) and a *counterfactual* component (how would things have been under a different treatment?)
- Example: what is the causal effect on your earnings of going to Brown instead of URI?
  - Descriptive Q: how much do Brown students earn after graduation?
  - Counterfactual Q: how much would Brown students have earned if they went to URI?
- Counterfactual Qs can't ever be answered with data alone. Need additional assumptions to learn about them!

Splitting up the problem

- When thinking about causal Qs, it's often easier to split the problem in two
- **Identification:** what could we learn about the parameters we care about (causal effects) if we had the observable data for the entire population
  - Need to make assumptions about how observed outcomes relate to outcomes that would have been realized under different treatments
- **Statistics**: what can we learn about the full population that we care about from the finite sample that we have?
  - Need to understand the process by which our data is generated from the full population

Framework for thinking about these steps

- **Sample:** the data that you actually observe
  - A survey of students from Brown and URI graduates about their earnings

Framework for thinking about these steps

- **Sample:** the data that you actually observe
  - A survey of students from Brown and URI graduates about their earnings
- **Estimator:** a function of the data in the sample
  - Difference in earnings between Brown and URI students in survey

Framework for thinking about these steps

- **Sample:** the data that you actually observe
  - A survey of students from Brown and URI graduates about their earnings
- **Estimator:** a function of the data in the sample
  - Difference in earnings between Brown and URI students in survey
- **Estimand:** a function of the observable data for the *population*
  - Difference in earnings between all Brown and URI students

Framework for thinking about these steps

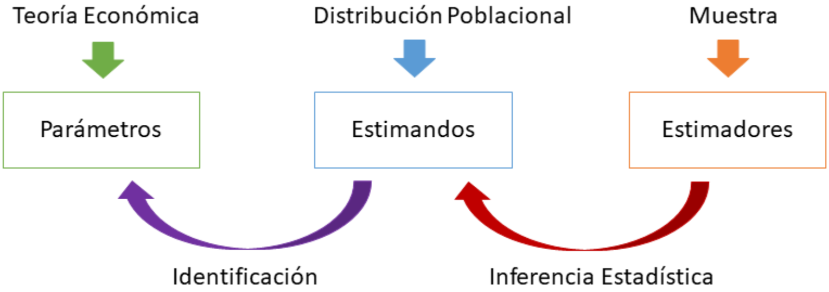
- **Sample:** the data that you actually observe
  - A survey of students from Brown and URI graduates about their earnings
- **Estimator:** a function of the data in the sample
  - Difference in earnings between Brown and URI students in survey
- **Estimand:** a function of the observable data for the *population*
  - Difference in earnings between all Brown and URI students
- **Target (aka structural) parameter:** what we actually care about
  - Causal effect on earnings of going to Brown relative to URI

Framework for thinking about these steps

- **Sample:** the data that you actually observe
  - A survey of students from Brown and URI graduates about their earnings
- **Estimator:** a function of the data in the sample
  - Difference in earnings between Brown and URI students in survey
- **Estimand:** a function of the observable data for the *population*
  - Difference in earnings between all Brown and URI students
- **Target (aka structural) parameter:** what we actually care about
  - Causal effect on earnings of going to Brown relative to URI
- The process of learning about the *estimand* from the estimator constructed with your *sample* is called **statistical estimation/inference**.

Framework for thinking about these steps

- **Sample:** the data that you actually observe
  - A survey of students from Brown and URI graduates about their earnings
- **Estimator:** a function of the data in the sample
  - Difference in earnings between Brown and URI students in survey
- **Estimand:** a function of the observable data for the *population*
  - Difference in earnings between all Brown and URI students
- **Target (aka structural) parameter:** what we actually care about
  - Causal effect on earnings of going to Brown relative to URI
- The process of learning about the *estimand* from the estimator constructed with your *sample* is called **statistical estimation/inference**.
- The process of learning about the *parameter* from the *estimand* is called **identification**.



$\theta = E[X_i]/E[Y_i]$

$E[X_i]$

$E[Y_i]$

$\sum X_i/n$

$\sum Y_i/n$

Let's add some math...

- Introduce **potential outcomes** notation
  - Super useful framework for thinking about causality!
    See the 2021 Nobel Prize writeup on Canvas!

Let's add some math...

- Introduce **potential outcomes** notation
  - Super useful framework for thinking about causality!
    See the 2021 Nobel Prize writeup on Canvas!

- $D_i =$ indicator if get treatment (1 if Brown, 0 if URI)

Let's add some math...

- Introduce **potential outcomes** notation
  - Super useful framework for thinking about causality!
    See the 2021 Nobel Prize writeup on Canvas!
- D_i = indicator if get treatment (1 if Brown, 0 if URI)
- $Y_i(1)$ = outcome under treament = earnings at Brown
- $Y_i(0)$ = outcome under control = earnings at URI

Let's add some math...

- Introduce **potential outcomes** notation
  - Super useful framework for thinking about causality!
    See the 2021 Nobel Prize writeup on Canvas!
- D_i = indicator if get treatment (1 if Brown, 0 if URI)
- $Y_i(1)$ = outcome under treament = earnings at Brown
- $Y_i(0)$ = outcome under control = earnings at URI
- Observed outcome Y_i is $Y_i(1)$ if $D_i = 1$ and $Y_i(0)$ if $D_i = 0$. (Y_i is your actual earnings)

Let's add some math...

- Introduce **potential outcomes** notation
  - Super useful framework for thinking about causality!
    See the 2021 Nobel Prize writeup on Canvas!
- D_i = indicator if get treatment (1 if Brown, 0 if URI)
- $Y_i(1)$ = outcome under treament = earnings at Brown
- $Y_i(0)$ = outcome under control = earnings at URI
- Observed outcome Y_i is $Y_i(1)$ if $D_i = 1$ and $Y_i(0)$ if $D_i = 0$. (Y_i is your actual earnings)
- We can write the observed outcome as $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$

- Example sample: (Y_i, D_i) for $i = 1, ...N$. Data with earnings and where you went to school

- Example sample: (Y_i, D_i) for $i = 1, ...N$. Data with earnings and where you went to school
- Example estimator:
    - Difference in sample mean of earnings for people who went to Brown and people who went to URI:

$$\underbrace{\frac{1}{N_1}\sum_{i:D_i=1} Y_i}_{\text{Avg earnings at Brown in sample}} \quad - \quad \underbrace{\frac{1}{N_0}\sum_{i:D_i=0} Y_i}_{\text{Avg earnings at URI in sample}}$$

- Example sample: (Y_i, D_i) for $i = 1, \ldots N$. Data with earnings and where you went to school
- Example estimator:
  - Difference in sample mean of earnings for people who went to Brown and people who went to URI:

$$\underbrace{\frac{1}{N_1}\sum_{i:D_i=1} Y_i}_{\text{Avg earnings at Brown in sample}} \quad - \quad \underbrace{\frac{1}{N_0}\sum_{i:D_i=0} Y_i}_{\text{Avg earnings at URI in sample}}$$

- Example estimand:
  - Difference in population mean of earnings for people went to Brown and people who went to URI:

$$\underbrace{E[Y_i|D_i = 1]}_{\text{Avg earnings at Brown in population}} \quad - \quad \underbrace{E[Y_i|D_i = 0]}_{\text{Avg earnings at URI in population}}$$

- Example sample: (Y_i, D_i) for $i = 1, ...N$. Data with earnings and where you went to school
- Example estimator:
  - Difference in sample mean of earnings for people who went to Brown and people who went to URI:

$$\underbrace{\frac{1}{N_1}\sum_{i:D_i=1} Y_i}_{\text{Avg earnings at Brown in sample}} \quad - \quad \underbrace{\frac{1}{N_0}\sum_{i:D_i=0} Y_i}_{\text{Avg earnings at URI in sample}}$$

- Example estimand:
  - Difference in population mean of earnings for people went to Brown and people who went to URI:

$$\underbrace{E[Y_i|D_i = 1]}_{\text{Avg earnings at Brown in population}} \quad - \quad \underbrace{E[Y_i|D_i = 0]}_{\text{Avg earnings at URI in population}}$$

- Example target parameter:
  - Causal effect of Brown for Brown students:

$$\underbrace{E[Y_i(1)|D_i = 1]}_{\text{Earnings at Brown for Brown students in pop}} \quad - \quad \underbrace{E[Y_i(0)|D_i = 1]}_{\text{Earnings at URI for Brown students in pop}} \quad .$$

Why is causal identification hard?

- Thought experiment: suppose we had data on earnings for *every* Brown and URI graduate
- We can learn from the data:

$$\underbrace{E[Y_i(1)|D_i = 1]}_{\text{Earnings at Brown for Brown Students}} \quad \text{and} \quad \underbrace{E[Y_i(0)|D_i = 0]}_{\text{Earnings at URI for URI students}}$$

Why is causal identification hard?

- Thought experiment: suppose we had data on earnings for *every* Brown and URI graduate
- We can learn from the data:

$$\underbrace{E[Y_i(1)|D_i = 1]}_{\text{Earnings at Brown for Brown Students}} \quad \text{and} \quad \underbrace{E[Y_i(0)|D_i = 0]}_{\text{Earnings at URI for URI students}}$$

- The causal effect of Brown for Brown students is

$$\underbrace{E[Y_i(1)|D_i = 1]}_{\text{Earnings at Brown for Brown Students}} - \underbrace{E[Y_i(0)|D_i = 1]}_{\text{Earnings at URI for Brown Students}}$$

Why is causal identification hard?

- Thought experiment: suppose we had data on earnings for *every* Brown and URI graduate
- We can learn from the data:

$$\underbrace{E[Y_i(1)|D_i = 1]}_{\text{Earnings at Brown for Brown Students}} \quad \text{and} \quad \underbrace{E[Y_i(0)|D_i = 0]}_{\text{Earnings at URI for URI students}}$$

- The causal effect of Brown for Brown students is

$$\underbrace{E[Y_i(1)|D_i = 1]}_{\text{Earnings at Brown for Brown Students}} - \underbrace{E[Y_i(0)|D_i = 1]}_{\text{Earnings at URI for Brown Students}}$$

- The data doesn't tell us $\underbrace{E[Y_i(0)|D_i = 1]}_{\text{Earnings at URI for Brown Students}}$. Why not?

Why is causal identification hard?

- Thought experiment: suppose we had data on earnings for *every* Brown and URI graduate
- We can learn from the data:

$$\underbrace{E[Y_i(1)|D_i = 1]}_{\text{Earnings at Brown for Brown Students}} \quad \text{and} \quad \underbrace{E[Y_i(0)|D_i = 0]}_{\text{Earnings at URI for URI students}}$$

- The causal effect of Brown for Brown students is

$$\underbrace{E[Y_i(1)|D_i = 1]}_{\text{Earnings at Brown for Brown Students}} - \underbrace{E[Y_i(0)|D_i = 1]}_{\text{Earnings at URI for Brown Students}}$$

- The data doesn't tell us $\underbrace{E[Y_i(0)|D_i = 1]}_{\text{Earnings at URI for Brown Students}}$. Why not?
  - Because we never see Brown students going to URI!

- One idea to solve this problem would be to assume that:

$$\underbrace{E[Y_i(0)|D_i = 1]}_{\text{Earnings at URI for Brown Students}} = \underbrace{E[Y_i(0)|D_i = 0]}_{\text{Earnings at URI for URI Students}}$$

- Why might this give us the wrong answer?

- One idea to solve this problem would be to assume that:

$$\underbrace{E[Y_i(0)|D_i = 1]}_{\text{Earnings at URI for Brown Students}} = \underbrace{E[Y_i(0)|D_i = 0]}_{\text{Earnings at URI for URI Students}}$$

- Why might this give us the wrong answer?
- Because Brown students may be different from URI students in other ways that would affect their earnings (regardless of where they went to college)
  - Academic ability, family background, career goals, etc.
- These differences are referred to as *omitted variables* or *confounding factors*

What about experiments?

- The gold standard for learning about causal effects is a randomized controlled trial (RCT), aka experiment
- Suppose that the Brown and URI administration randomized who got into which college (assume these are the only 2 colleges for simplicity)
- Since college is randomly assigned, the only thing that differs between Brown and URI students is the college they went to
- Hence,

$$\underbrace{E[Y_i(0)|D_i = 1]}_{\text{Earnings at URI for Brown Students}} = \underbrace{E[Y_i(0)|D_i = 0]}_{\text{Earnings at URI for URI Students}}$$

since we've eliminated any confounding factors

But running experiments is often hard/impossible

- Unfortunately, Brown/URI have not let us randomize who gets into which college
  - At least not yet! If you could convince them to do this, it'd make for a cool senior thesis!
- Likewise, it is difficult to convince states to randomize their minimum wages, or other policies
- In some cases, randomization is not just difficult but would be immoral
  - "What is the causal effect of spousal death on labor supply?"

But running experiments is often hard/impossible

- Unfortunately, Brown/URI have not let us randomize who gets into which college
    - At least not yet! If you could convince them to do this, it'd make for a cool senior thesis!
- Likewise, it is difficult to convince states to randomize their minimum wages, or other policies
- In some cases, randomization is not just difficult but would be immoral
    - "What is the causal effect of spousal death on labor supply?"
- In this course, we'll discuss tools economists try to use when running experiments is not possible.

Course Roadmap – Where we're going

- **Part I ($\sim$ 7 lectures): Review of probability/statistics**. This will give us a mathematical language to talk about:
  1. *Statistical estimation/inference:* how does the sample we observe relate to the population of interest
  2. *Identification:* how do observable features of the population relate to (causal) parameters we care about

Course Roadmap – Where we're going

- **Part I ($\sim$ 7 lectures): Review of probability/statistics**. This will give us a mathematical language to talk about:
  1. *Statistical estimation/inference:* how does the sample we observe relate to the population of interest
  2. *Identification:* how do observable features of the population relate to (causal) parameters we care about

- **Part II ($\sim$ 9 lectures): Linear regression:** We'll discuss ordinary least squares (OLS), the workhorse model for estimation in econometrics. When does it work, and when will it fail?

Course Roadmap – Where we're going

- **Part I ($\sim$ 7 lectures): Review of probability/statistics**. This will give us a mathematical language to talk about:
  1. *Statistical estimation/inference:* how does the sample we observe relate to the population of interest
  2. *Identification:* how do observable features of the population relate to (causal) parameters we care about

- **Part II ($\sim$ 9 lectures): Linear regression:** We'll discuss ordinary least squares (OLS), the workhorse model for estimation in econometrics. When does it work, and when will it fail?

- **Part III ($\sim$ 7 lectures:) Other "quasi-experimental" strategies**: We'll discuss other strategies for "mimicking" an experiment when it's not available, including instrumental variables (IV) and regression discontinuity (RD)

Outline

1. Deriving Multivariate Regression and OLS
2. Regression and Causality
3. Regression Odds and Ends

Moving Beyond One "Regressor"

- So far we've talked about regression as a way of approximating the CEF $E[Y_i|X_i = x] \approx \alpha + x\beta$ for a single scalar X_i
  - We then showed how the estimand (α, β) can be estimated by OLS

Moving Beyond One "Regressor"

- So far we've talked about regression as a way of approximating the CEF $E[Y_i|X_i = x] \approx \alpha + x\beta$ for a single scalar X_i
  - We then showed how the estimand (α,β) can be estimated by OLS
- Next we'll see how this can be generalized to approximate/estimate $E[Y_i|\mathbf{X}_i = \mathbf{x}] \approx \mathbf{x}'\boldsymbol{\beta}$ for a vector $\mathbf{X_i} = (1, X_{i1}, ..., X_{iK})'$
  - Note: As usual, I'll be putting vectors/matrices in bold type-face

Moving Beyond One "Regressor"

- So far we've talked about regression as a way of approximating the CEF $E[Y_i|X_i = x] \approx \alpha + x\beta$ for a single scalar X_i
  - We then showed how the estimand (α,β) can be estimated by OLS
- Next we'll see how this can be generalized to approximate/estimate $E[Y_i|\mathbf{X}_i = \mathbf{x}] \approx \mathbf{x}'\boldsymbol{\beta}$ for a vector $\mathbf{X_i} = (1, X_{i1}, ..., X_{iK})'$
  - Note: As usual, I'll be putting vectors/matrices in bold type-face
- Two main motivations for this:

Moving Beyond One "Regressor"

- So far we've talked about regression as a way of approximating the CEF $E[Y_i|X_i = x] \approx \alpha + x\beta$ for a single scalar X_i
  - We then showed how the estimand (α,β) can be estimated by OLS

- Next we'll see how this can be generalized to approximate/estimate $E[Y_i|\mathbf{X}_i = \mathbf{x}] \approx \mathbf{x}'\boldsymbol{\beta}$ for a vector $\mathbf{X_i} = (1, X_{i1}, ..., X_{iK})'$
  - Note: As usual, I'll be putting vectors/matrices in bold type-face

- Two main motivations for this:

1. We want to use regression to identify causal effects, but conditional unconfoundedness is only plausible with multiple controls
   - In the Brown/URI example, we may want to control for high school GPA, family income, SAT, race ...

Moving Beyond One "Regressor"

- So far we've talked about regression as a way of approximating the CEF $E[Y_i|X_i = x] \approx \alpha + x\beta$ for a single scalar X_i
  - We then showed how the estimand (α,β) can be estimated by OLS
- Next we'll see how this can be generalized to approximate/estimate $E[Y_i|\mathbf{X}_i = \mathbf{x}] \approx \mathbf{x}'\boldsymbol{\beta}$ for a vector $\mathbf{X_i} = (1, X_{i1}, ..., X_{iK})'$
  - Note: As usual, I'll be putting vectors/matrices in bold type-face
- Two main motivations for this:

1. We want to use regression to identify causal effects, but conditional unconfoundedness is only plausible with multiple controls
   - In the Brown/URI example, we may want to control for high school GPA, family income, SAT, race ...
2. We want a *nonlinear* CEF approx.: e.g. $E[Y_i \mid X_i] \approx \alpha + X_i\beta + X_i^2\gamma$
   - We can "trick" regression into doing this by setting $\mathbf{X_i} = (1, X_i, X_i^2)'$