Regression Methods

ECON 5753 — University of Arkansas

Prof. Kyle Butts

February 2025

Conditional Expectation Function

Linear Models

Ordinary Least Squares

Statistical Properties / Inference

Bivariate Regression Example

Forecasting with Regression Model

Marginal Effects

Forecasting

We have an outcome variable Y and a set of p different predictor variables $X = (X_1, X_2, \dots, X_p)$.

The goal of forecasting is to take an observation's X values, $X = x$, and predict Y given that information.

- We want to know: *conditional* on $X = x$, what do we *expect* the value of Y to be.

f_0 as the Conditional Expectation Function

$$Y = f_0(X) + \varepsilon,$$

The last time, we called this best guess at y, $f_0(X)$. Today, we will call it the Conditional Expectation Function

Joint Distribution

For now, let's think of X as a single variable, e.g. someone's height. Let Y be someone's weight. To make notation easier, think of these as discrete variables (i.e. finite, but large, number of values)

Joint Distribution

For now, let's think of X as a single variable, e.g. someone's height. Let Y be someone's weight. To make notation easier, think of these as discrete variables (i.e. finite, but large, number of values)

For forecasting, it is not enough to know the *marginal* distributions of height and weight. We must know the *joint* distribution, i.e.:

$$\mathbb{P}(X = x, Y = y)$$

- The probability we sample a person with X equal to x ***and*** Y equal to y

Joint Distribution

$$\mathbb{P}(X = x, Y = y)$$

This is easy to estimate in our sample; we just count the number of times $X_i = x$ and $Y_i = y$ and divide by the total number of obseravtions

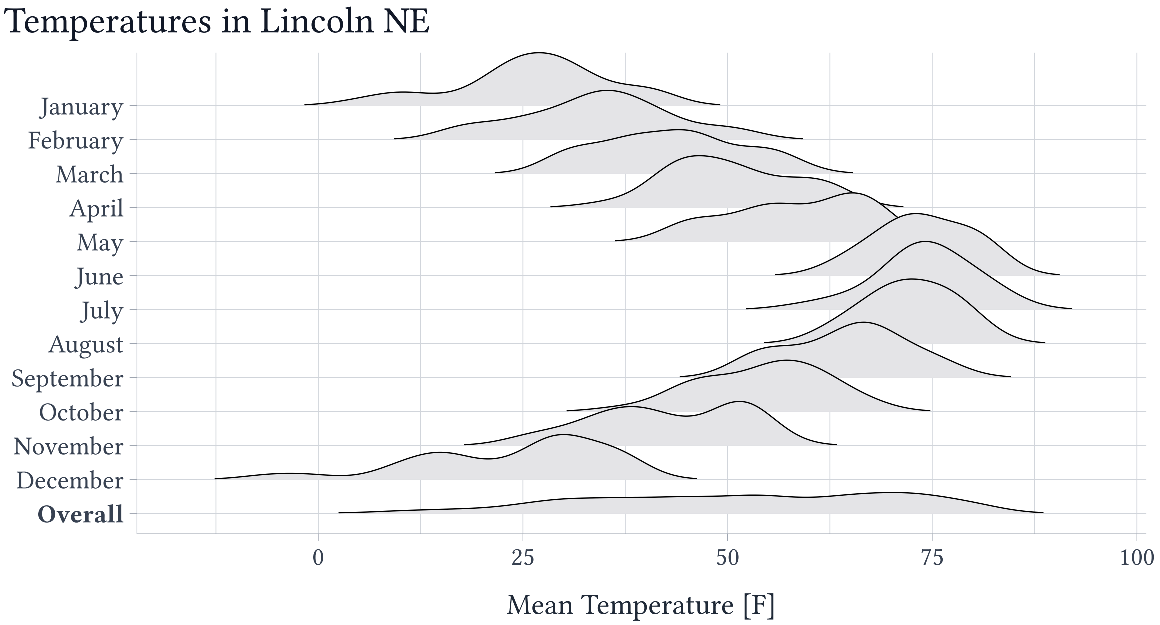
Conditional Probability

A related question we can ask is the conditional probability of $Y = y$ *given/conditional on* $X = x$:

$$\mathbb{P}(Y = y \mid X = x)$$

Think of it like this:

- You grab a random unit from your population and you observe that $X_i = x$
- Given that you know this information, you now have to take a guess at what the value of Y is.


Temperatures in Lincoln NE
January
February
March
April
May
June
July
August
September
October
November
December
Overall
0
25
50
75
100
Mean Temperature [F]

Conditional Probability

If knowing the value of X_i does not help you guess the value of Y_i, then

$$\mathbb{P}(Y = y \mid X = x) = \mathbb{P}(Y = y)$$

and we say X and Y are independent

- This means knowing X can not help you forecast Y!

Bayes' Rule

The joint-distribution and the conditional-distribution are connected via Bayes' Rule:

$$\mathbb{P}(Y = y \mid X = x) = \frac{\overbrace{\boxed{\mathbb{P}(Y = y, X = x)}}^{(\#\text{ of units with } X = x \textbf{ and } Y = y)\ /\ n}}{\underbrace{\boxed{\mathbb{P}(X = x)}}_{(\#\text{ of units with } X = x)\ /\ n}}$$

The intuition is:

- Count the number of people with $X = x$ *and* $Y = y$
- Divide by the number of people with $X = x$

Conditional Probability

Note that for all values of x, we have

$$\sum_{y} \mathbb{P}(Y = y \mid X = x) = 1$$

Intuition: "The conditional that Y equals something given $X = x$ is 1"

Note that for all values of x, we have

$$\sum_y \mathbb{P}(Y = y \mid X = x) = 1$$

Intuition: "The conditional that Y equals something given $X = x$ is 1"

You should think of the conditional probability as a new probability defined on the sub-population with $X_i = x$

Expectation

Remember the definition of the conditional expectation of a discrete varaible:

$$\mathbb{E}[Y] = \sum_{y} \mathbb{P}(Y = y)y$$

The average of the values Y can take, weighted by the probability they take those values

Expectation

$$\mathbb{E}[Y] = \sum_y \mathbb{P}(Y = y)y$$

If we observed everyone in the *population*, we could calculate this really easily:

- Take the average value of y in the population

Conditional Expectation

Similarly, the conditional expectation of Y given $X = x$ is:

$$\mathbb{E}[Y \mid X = x] = \sum_y \mathbb{P}(Y = y \mid X = x) y$$

The average of the values Y can take, weighted by the *conditional* probability they take those values

Conditional Expectation

$$\mathbb{E}[Y \mid X = x] = \sum_y \mathbb{P}(Y = y \mid X = x)y$$

If we observed everyone in the *population*, we could calculate this really easily:

- Subset to people with $X = x$
- Take the average value of y *within that subsample*

Conditional Expectation

In the previous lecture, we used the notation $f_0(x)$ to denote the conditional expectation function:

$$f_0(x) \equiv \mathbb{E}[Y \mid X = x]$$

This function takes x as an input and outputs the conditional expectation of Y given $X = x$

Estimating Conditional Expectation

In reality, we only observe a sample $(X_i, Y_i)_{i=1}^n$. We can estimate $f_0(x)$ at a point x in the same way:

- Subset to people with $X_i = x$
- Take the average value of Y_i *within that subsample.* Call this $\hat{f}(x)$

In math terms, this estimator is given by

$$\hat{f}(x) = \frac{1}{\underbrace{\sum_{i=1}^{n} \mathbb{1}[X_i = x]}_{\#\text{ of units with } X_i = x}} \overbrace{\sum_{i=1}^{n} Y_i \, \mathbb{1}[X_i = x]}^{\text{sum of } Y_i \text{ for units with } X_i = x}$$

Estimating Conditional Expectation

We can estimate $f_0(x)$ at a point x in the same way:

- Subset to people with $X_i = x$
- Take the average value of Y_i *within that subsample.* Call this $\hat{f}(x)$

When $n \to \infty$, we have $\hat{f}(x) \to f_0(x)$ for all values of x

- This estimator is consistent for the conditional expectation of Y given $X = x$

Difficulties with this estimator

This estimator is simple and works if we have *really large samples*. But what if we only have a few people with a value of $X_i = x$?

We are taking a sample mean with a few units; it will be very noisy

- The relative "n" in the law of large numbers is the number of units with $X_i = x$

Difficulties with this estimator

This estimator is simple and works if we have *really large samples*. But what if we only have a few people with a value of $X_i = x$?

We are taking a sample mean with a few units; it will be very noisy

- The relative "n" in the law of large numbers is the number of units with $X_i = x$

We do not use *any* of the data from nearby units, $X_i = x \pm$ a little

- Feels wasteful to throw out this information; do we really think Y changes dramatically as we move away from x a little?

Estimating Conditional Expectation

$$f_0(x) \equiv \mathbb{E}[Y \mid X = x]$$

There are two primary strategies we will discuss in this class:

1. Linear regression models [this topic]

   $\rightarrow$ Assume a functional form for $f_0(x)$

2. Non-parametric estimators [later]

   $\rightarrow$ The previous estimator or variants that pool over $(x - \delta, x + \delta)$

Conditional Expectation Function

Linear Models

Ordinary Least Squares

Statistical Properties / Inference

Bivariate Regression Example

Forecasting with Regression Model

Marginal Effects

Linear Model for the Conditioanl Expectation Function

$$f_0(x) \equiv \mathbb{E}[Y \mid X = x]$$

Let $X_i \equiv \begin{bmatrix} x_{i1} & \dots & x_{ip} \end{bmatrix}'$ be the vector of p explanatory variables.

Our first approach to estimating the conditional expectation function is to assume a linear model:

$$f_0(x) = x'\beta$$

Linear Model for the Conditioanl Expectation Function

Alternatively, you will see the model written out as

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

with the assumption $\mathbb{E}[u_i \mid X_i] = 0$.

The restriction ensures that $X_i'\beta$ *is* the CEF of Y_i:

$$\mathbb{E}[Y_i \mid X_i = x] = \mathbb{E}\big[X_i'\beta + u_i \mid X_i = x\big]$$

Linear Model for the Conditioanl Expectation Function

Alternatively, you will see the model written out as

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

with the assumption $\mathbb{E}[u_i \mid X_i] = 0$.

The restriction ensures that $X_i'\beta$ *is* the CEF of Y_i:

$$\begin{aligned} \mathbb{E}[Y_i \mid X_i = x] &= \mathbb{E}\left[X_i'\beta + u_i \mid X_i = x\right] \\ &= x'\beta + \mathbb{E}[u_i \mid X_i = x] \end{aligned}$$

Linear Model for the Conditioanl Expectation Function

Alternatively, you will see the model written out as

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

with the assumption $\mathbb{E}[u_i \mid X_i] = 0$.

The restriction ensures that $X_i'\beta$ *is* the CEF of Y_i:

$$\begin{aligned}
\mathbb{E}[Y_i \mid X_i = x] &= \mathbb{E}\left[X_i'\beta + u_i \mid X_i = x\right] \\
&= x'\beta + \mathbb{E}[u_i \mid X_i = x] \\
&= x'\beta
\end{aligned}$$

Note that there are many "linear" models for the CEF

$$f_0(x) = x_1\beta_1 + x_2\beta_2$$
$$f_0(x) = x_1\beta_1 + x_2\beta_2 + x_2^2\beta_3$$
$$f_0(x) = g_1(x_1)\beta_1 + g_2(x_1)\beta_2 + x_2\beta_3$$

where g_1 and g_2 are some known functions (polynomial term, indicator functions, etc.)

These are all *linear* models for the CEF, $\mathbb{E}[Y_i \mid X_i = x]$

- "linear model" = linear combinations of terms

Perhaps a better way to write this would be to define the control variables as

$$W_i = \begin{bmatrix} g_1(X_i) & \dots & g_K(X_i) \end{bmatrix}'$$

Then, we could write out model out as

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

with $\mathbb{E}[u_i \mid X_i] = 0$.

- This notation better distinguishes between covariates in model (e.g. polynomial of age) and variables you are conditioning on (e.g. age)

But, a lot of explanations of regression models do not make this difference very clear; instead just writing

$$Y_i = X_i\beta + u_i$$

where X_i really is W_i, i.e. can contain functions of the underlying covariates.

- I will try and make this distinction clear, but may fail at points

Error term restriction

The key assumption here is that in the model with

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

we have the conditiona mean-zero error term: $\mathbb{E}[u_i \mid X_i] = 0$.

This latter assumption depends on the terms included in W_i. Say the CEF of wages conditional on age is quadratic, but we only include the linear term

- Then the term $\text{age}^2 \beta_2$ will show up in the error term u_i. This will not be mean-zero given age!

Conditional Expectation Function

Linear Models

Ordinary Least Squares

Statistical Properties / Inference

Bivariate Regression Example

Forecasting with Regression Model

Marginal Effects

Fitting a regression model

$$Y_i = \underbrace{W_i'\beta}_{f_0(x)} + u_i$$

After that long diatribe on defining a linear model, we are now going to discuss estimation

Matrix Notation

Let Y be the $n \times 1$ vector of Y_i. Let $\boldsymbol{W}$ be the $n \times K$ matrix stacking W_i':

$$\boldsymbol{W} = \begin{bmatrix} W_1' \\ \vdots \\ W_n' \end{bmatrix}$$

- We generally *always* assume you have an intercept, i.e. $W_{i1} = 1$

Our model becomes

$$Y = \boldsymbol{W}\beta + u$$

Matrix Notation

Take a minute to verify that the following yields the regression model we think it does

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} W_{11} & \dots & W_{1K} \\ \vdots & \ddots & \vdots \\ W_{n1} & \dots & W_{nK} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_K \end{bmatrix} + \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}$$

Residuals

We can rearrange out model as $u = Y - \boldsymbol{W}\beta$. For a given guess at β, b, we have our regression residuals as

$$\hat{u}(b) = Y - \boldsymbol{W}b$$

- When evaulated at the OLS estimates $\hat{\beta}_{\text{OLS}}$, this is usually written just as $\hat{u}$

Residuals

$$\hat{u}(b) = Y - \boldsymbol{W}b$$

As we discussed in our previous topic, we can not just minimize the average residual, $\frac{1}{n}\iota'\hat{u}(b)$, because positive and negative errors "cancel out"

Residuals

$$\hat{u}(b) = Y - \boldsymbol{W} b$$

As we discussed in our previous topic, we can not just minimize the average residual, $\frac{1}{n}\iota' \hat{u}(b)$, because positive and negative errors "cancel out"

Instead, we will use the sum of squared residuals:

$$\hat{u}(b)'\hat{u}(b) = \left(Y - \boldsymbol{W} b\right)' \left(Y - \boldsymbol{W} b\right)$$

Sum of Squared Residuals

As a reminder, this matrix notation is indeed the "sum of squared residuals":

$$\begin{aligned} \hat{u}(b)'\hat{u}(b) &= \begin{bmatrix} \hat{u}_1(b) & \dots & \hat{u}_n(b) \end{bmatrix} \begin{bmatrix} \hat{u}_1(b) \\ \vdots \\ \hat{u}_n(b) \end{bmatrix} \\ &= \sum_i \hat{u}_i(b)^2 \end{aligned}$$

Ordinary Least Squares Problem

So, our estimation problem is to choose a b to minimize the sum of squared residuals:

$$
\begin{aligned}
\hat{\beta}_{\text{OLS}} &\equiv \underset{b}{\operatorname{argmin}}\, \hat{u}(b)'\hat{u}(b) \\
&= \underset{b}{\operatorname{argmin}}\, (Y - \boldsymbol{W}b)'\,(Y - \boldsymbol{W}b)
\end{aligned}
$$

Ordinary Least Squares Problem

$$\hat{\beta}_{\mathsf{OLS}} = \underset{b}{\operatorname{argmin}} \, (Y - \boldsymbol{W} b)' \, (Y - \boldsymbol{W} b)$$

Expanding out this product yields

$$(Y - \boldsymbol{W} b)' \, (Y - \boldsymbol{W} b) = Y'Y - b'\boldsymbol{W} Y - Y \boldsymbol{W} b + b' \boldsymbol{W}' \boldsymbol{W} b$$

It might not be immediately recognizable, but this is a *quadratic* function of b

First-order conditions

Taking the derivative and set $= 0$ will yield the minimum:

$$\frac{\partial}{\partial b}\left(Y'Y - b'\boldsymbol{W}'Y - Y\boldsymbol{W}'b + b'\boldsymbol{W}'\boldsymbol{W}b\right)$$

First-order conditions

Taking the derivative and set $= 0$ will yield the minimum:

$$\frac{\partial}{\partial b}\left(Y'Y - b'\boldsymbol{W}'Y - Y\boldsymbol{W}'b + b'\boldsymbol{W}'\boldsymbol{W}b\right)$$

Using our rules of matrix derivatives from Topic 1, this yields:

$$0 - \boldsymbol{W}'Y - \boldsymbol{W}'Y + 2\boldsymbol{W}'\boldsymbol{W}b$$

First-order conditions

Taking the derivative and set $= 0$ will yield the minimum:

$$\frac{\partial}{\partial b}\left(Y'Y - b'\boldsymbol{W}'Y - Y\boldsymbol{W}'b + b'\boldsymbol{W}'\boldsymbol{W}b\right)$$

Using our rules of matrix derivatives from Topic 1, this yields:

$$0 - \boldsymbol{W}'Y - \boldsymbol{W}'Y + 2\boldsymbol{W}'\boldsymbol{W}b$$

Setting this equal to 0, yields our first-order condition:

$$\left(\boldsymbol{W}'\boldsymbol{W}\right)\hat{\beta}_{\text{OLS}} = \boldsymbol{W}'Y$$

OLS Estimator

$$\hat{\beta}_{\text{OLS}} = \left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1}\boldsymbol{W}'Y$$

Recap: We have derived the OLS estimator from minimizing the sum of squared prediction errors (with the help of linear algebra)

Intuition of OLS Estimator

$$\hat{\beta}_{\text{OLS}} = \left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1}\boldsymbol{W}'Y$$

Say we have just an intercept ($W_{i1} = 1$), so that $\boldsymbol{W} = \iota$. In this case:

- $\boldsymbol{W}'\boldsymbol{W} = \iota'\iota = n$
- $\boldsymbol{W}'Y = \iota'Y = \sum_{i=1}^{n} Y_i$

Consequently $\hat{\beta}_{\text{OLS}} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is the sample mean

Intuition of OLS Estimator

$$\hat{\beta}_{\text{OLS}} = \left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1}\boldsymbol{W}'Y$$

Say we have an intercept ($W_{i1} = 1$) and a single explantory variable W_{i2}

It turns out (by the FWL theorem), that the regression of Y_i on $1, W_{i2}$ is equivalent to the regression of $Y_i - \bar{Y}$ on $W_{i2} - \bar{W}_2$.

Intuition of OLS Estimator

Thinking of the regression of $Y_i - \bar{Y}$ on $W_{i2} - \bar{W}_2$:

- $\boldsymbol{W}'\boldsymbol{W} = \sum_i (W_{i2} - \bar{W}_2)^2$ is $(n-1)$ times the sample variance of W_{i2}.
- $\boldsymbol{W}'Y = \sum_i (W_{i2} - \bar{W}_2)(Y_i - \bar{Y})$ is $(n-1)$ the sample covariance

Consequently, we have the bivariate regression formula: $\hat{\beta}_{\mathsf{OLS}} = \widehat{\mathsf{Cov}}(W_{i2}, Y_i) / \widehat{\mathsf{Var}}(W_{i2})$.

More generally, when we have $K-1$ covariates and an intercept, this is equivalent to the regression where Y and all the covariates are demeaned (without an intercept). Then,

$$\begin{aligned}\hat{\beta}_{\text{OLS}} &= \left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1}\boldsymbol{W}'Y \\ &= \left[\widehat{\text{Var}}(W_i)\right]^{-1}\widehat{\text{Cov}}(W_i, Y_i)\end{aligned}$$

- $\widehat{\text{Var}}(W_i)^{-1}$ is the covariance matrix of all the of variables
- $\widehat{\text{Cov}}(W_i, Y_i)$ is the $K-1$ vector of covariances between each W_{ik} and Y_i

Conditional Expectation Function

Linear Models

Ordinary Least Squares

Statistical Properties / Inference

Bivariate Regression Example

Forecasting with Regression Model

Marginal Effects

Sample distribution of $\hat{\beta}_{\text{OLS}}$

In repeated sampling, we will get different draws of u_i for each unit. This will create different estimates of $\hat{\beta}$.

Say the true model is $y_i = W_i' \beta_0 + u_i$. Assuming we did a good job modeling the conditional expectation function, then we can assume $\mathbb{E}[u_i \mid X_i] = 0$

- Remember that W_i are functions of X_i

Sample distribution of $\hat{\beta}_{\text{OLS}}$

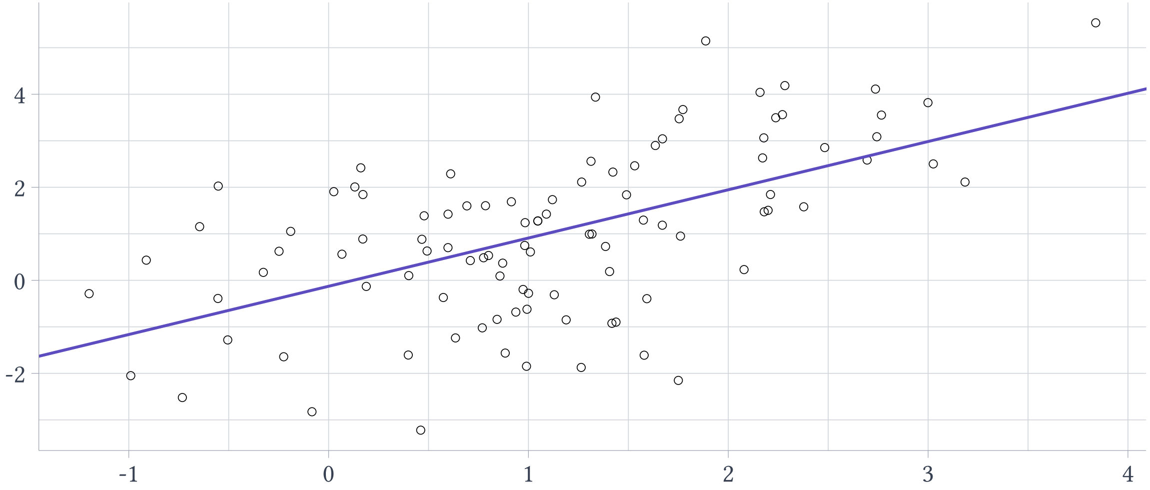
Simulation

As a simple example, do a Monte Carlo simulation:

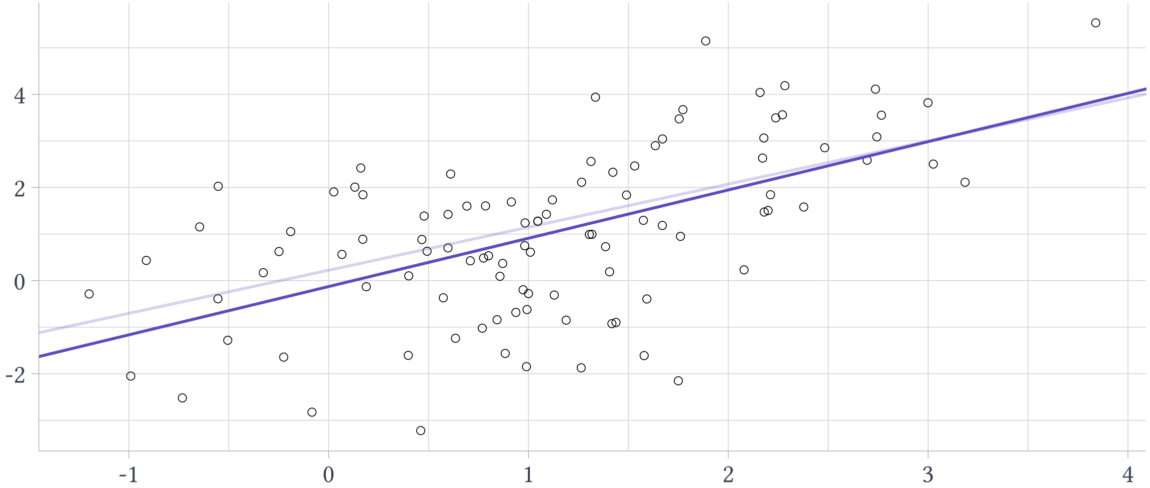
- $x_i \sim \mathcal{N}(1, 1)$
- $\varepsilon_i \sim \mathcal{N}(0, 1.5^2)$
- $y_i = x_i * 1 + \varepsilon_i$

Draw $B = 2500$ different samples each with $n = 100$ observations. Estimate regression of y_i on an intercept and x_i.

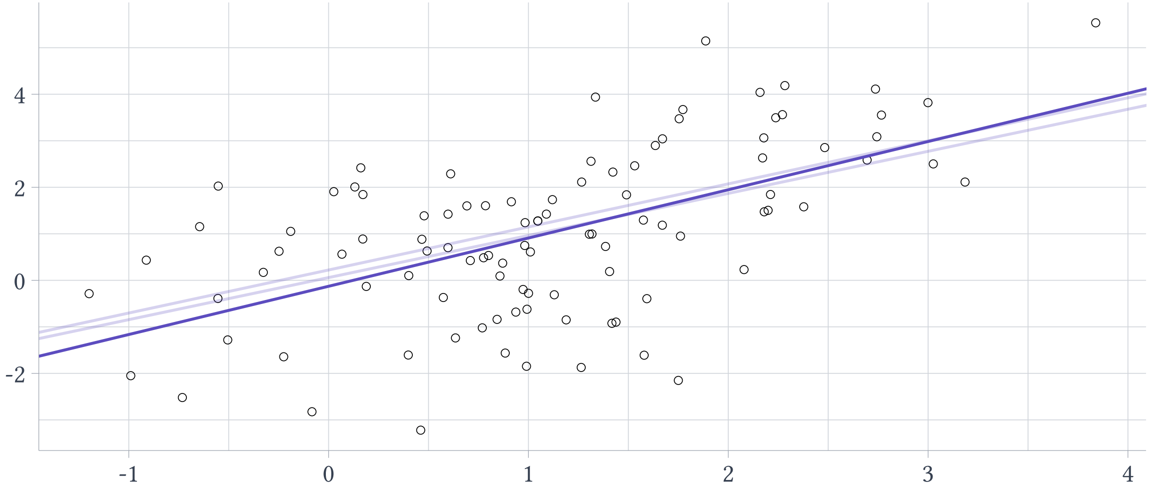
Original Sample



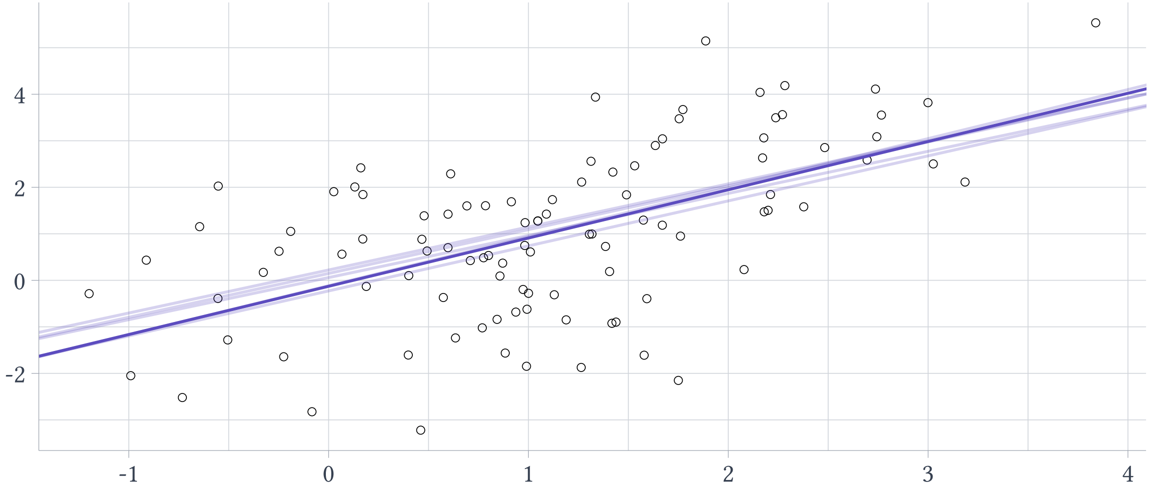
Original Sample + 1 Extra Sample



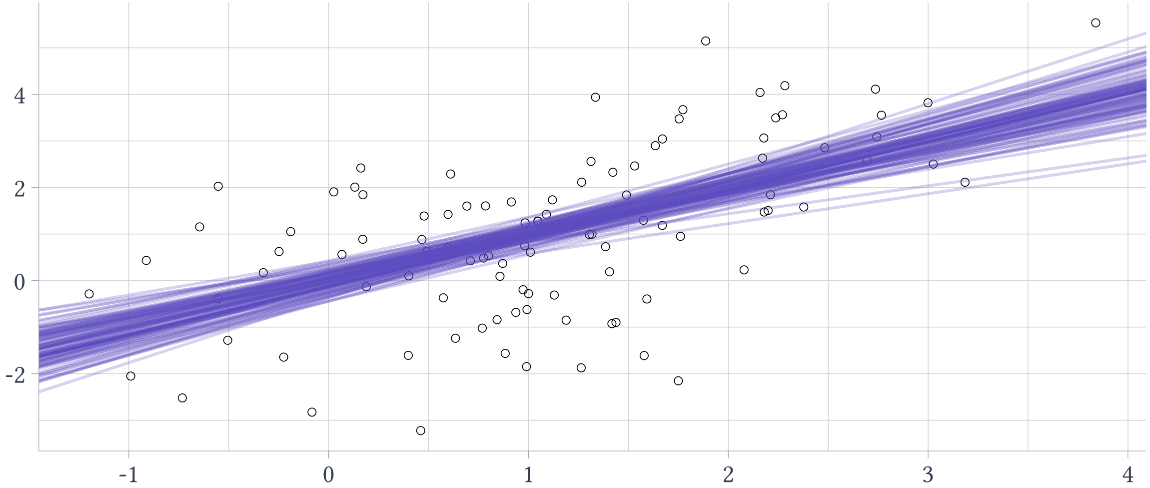
Original Sample + 2 Extra Samples

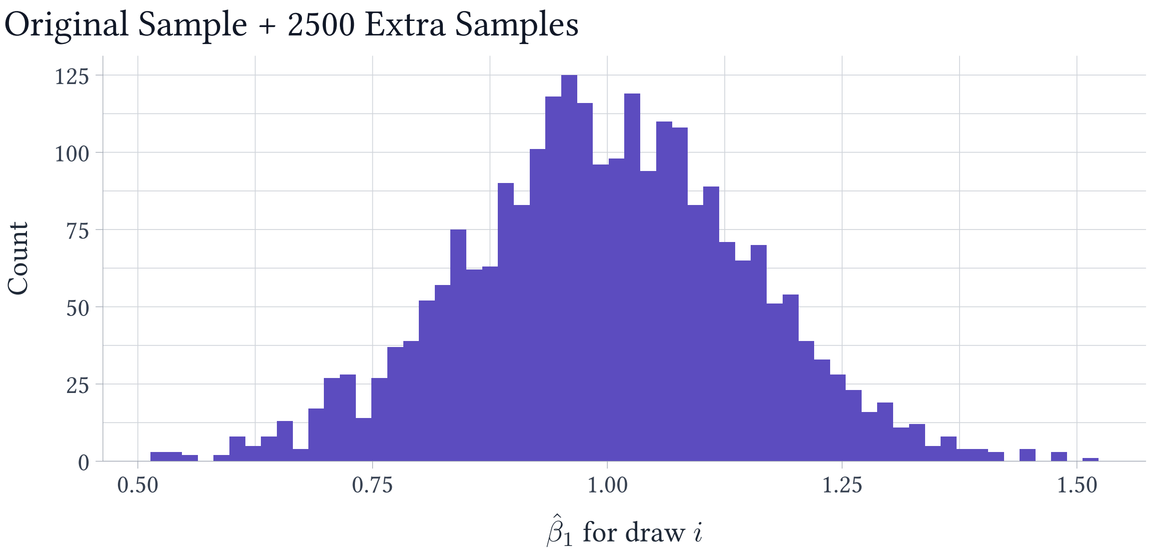


Original Sample + 5 Extra Samples



Original Sample + 100 Extra Samples




Original Sample + 2500 Extra Samples
Count
125
100
75
50
25
0
0.50
0.75
1.00
1.25
1.50
$\hat{\beta}_1$ for draw i

Statistical properties

The true model is $y_i = W_i' \beta_0 + u_i$ with $\mathbb{E}[u_i \mid X_i] = 0$.

Plugging this into our OLS estimator, we have:

$$
\begin{aligned}
\hat{\beta}_{\text{OLS}} &= \left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1} \boldsymbol{W}'Y \\
&= \left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1} \boldsymbol{W}'\left(\boldsymbol{W}\beta_0 + u\right)
\end{aligned}
$$

The true model is $y_i = W_i' \beta_0 + u_i$ with $\mathbb{E}[u_i \mid X_i] = 0$.

Plugging this into our OLS estimator, we have:

$$\begin{aligned}
\hat{\beta}_{\text{OLS}} &= \left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1} \boldsymbol{W}'Y \\
&= \left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1} \boldsymbol{W}'\left(\boldsymbol{W}\beta_0 + u\right) \\
&= \left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1} \boldsymbol{W}'\boldsymbol{W}\beta_0 + \left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1} \boldsymbol{W}'u \\
&= \beta_0 + \left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1} \boldsymbol{W}'u
\end{aligned}$$

Unbiasedness

Our previous slide shows

$$\hat{\beta}_{\text{OLS}} = \beta_0 + \left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1} \boldsymbol{W}'u$$

Using $\mathbb{E}[u_i \mid X_i] = 0$, we can show unbiasedness of our estimator:

$$\begin{aligned} \mathbb{E}\left[\hat{\beta}_{\text{OLS}}\right] &= \beta_0 + \mathbb{E}\left[\left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1} \boldsymbol{W}'u\right] \\ &= \beta_0 \end{aligned}$$

Error-term covariance

$$\hat{\beta}_{\text{OLS}} = \beta_0 + \left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1}\boldsymbol{W}'u$$

For the distribution of $\hat{\beta}_{\text{OLS}}$, we first need to discuss the covariance of the error term.

We write the variance as $\boldsymbol{\Sigma} = \mathbb{E}[uu']$ which has typical element $\sigma_{i,j} = \mathbb{E}[u_i u_j]$

Independent Errors

Our error term u has variance:

$$\mathbf{\Sigma} = \mathbb{E}\left[uu'\right]$$

If each unit is independent, we have $\sigma_{i,j} = 0$ whenever $i \neq j$. If this is true, we have

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

Independent Errors

We could estimate this matrix using the residuals $\hat{u}_i = y_i - W_i' \hat{\beta}_{\text{OLS}}$:

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} \hat{u}_1^2 & 0 & \dots & 0 \\ 0 & \hat{u}_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \hat{u}_n^2 \end{bmatrix}$$

This estimator is not consistent for $\boldsymbol{\Sigma}$ since $\hat{u}_i \neq u_i$, but this turns out to be okay when estimating the variance of $\hat{\beta}_{\text{OLS}}$

Inference on $\hat{\beta}_{\text{OLS}}$

$$\hat{\beta}_{\text{OLS}} = \beta_0 + \left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1} \boldsymbol{W}' u$$

If we write out the summation in the final term, you can see this is a weighted sum of idiosyncratic shocks, u_i

$$\hat{\beta}_{\text{OLS}} = \beta_0 + \left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1} \sum_{i=1}^{n} W_i u_i$$

Inference on $\hat{\beta}_{\text{OLS}}$

Subtracting β_0 and multiplying by $\sqrt{n}$, we have:

$$\sqrt{n}\left(\hat{\beta}_{\text{OLS}} - \beta_0\right) = \left(\frac{1}{n}\boldsymbol{W}'\boldsymbol{W}\right)^{-1} \boxed{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i u_i}$$

apply a central-limit theorem

The term $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i u_i$ has mean 0 (from unbiasedness) and has variance:

$$\mathbb{E}\left[\boldsymbol{W}'uu'\boldsymbol{W}\right] = \boldsymbol{W}'\boldsymbol{\Sigma}\boldsymbol{W}.$$

Inference on $\hat{\beta}_{\text{OLS}}$

$$\sqrt{n}\left(\hat{\beta}_{\text{OLS}} - \beta_0\right) = \left(\frac{1}{n}\boldsymbol{W}'\boldsymbol{W}\right)^{-1} \boxed{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i u_i}$$

Using the central limit theorem, we have that $\frac{1}{\sqrt{n}}\boldsymbol{W}'u$ is normally distributed, and we are multiplying it by a matrix $\left(\frac{1}{n}\boldsymbol{W}'\boldsymbol{W}\right)^{-1}$, so we have:

$$\hat{\beta}_{\text{OLS}} \sim \mathcal{N}\left(\beta_0, \left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1}\boldsymbol{W}'\boldsymbol{\Sigma}\boldsymbol{W}\left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1}\right)$$

Inference on $\hat{\beta}_{\text{OLS}}$

$$\hat{\beta}_{\text{OLS}} \sim \mathcal{N}\left(\beta_0, \left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1} \boldsymbol{W}'\boldsymbol{\Sigma}\boldsymbol{W} \left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1}\right)$$

The variance is a $K \times K$ matrix with diagonal elements $\text{Var}\left(\hat{\beta}_{\text{OLS},k}\right)$

- Take square-root of diagonal elements to get standard deviation of the estimators

Inference on $\hat{\beta}_{\text{OLS}}$

$$\hat{\beta}_{\text{OLS}} \sim \mathcal{N}\left(\beta_0, \left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1} \boldsymbol{W}'\boldsymbol{\Sigma}\boldsymbol{W} \left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1}\right)$$

The variance is a $K \times K$ matrix with diagonal elements $\text{Var}\left(\hat{\beta}_{\text{OLS},k}\right)$

- Take square-root of diagonal elements to get standard deviation of the estimators
- The off-diagonal elements tell us how slope coefficients might be correlated with one another in repeated samples

Inference on $\hat{\beta}_{\text{OLS}}$

Let $\text{Var}\left(\hat{\beta}_{\text{OLS},k}\right)$ be the k-th diagonal, then we have

$$\hat{\beta}_{\text{OLS},k} \sim \mathcal{N}\left(\beta_{0,k},\ \text{Var}\left(\hat{\beta}_{\text{OLS},k}\right)\right)$$

Since we have a statistic $\hat{\beta}_{\text{OLS},k}$ that has a sample distribution that is normally-distributed, we can do standard statistical techniques:

- Confidence intervals and hypothesis testing

Standard Errors

We can take our estimate $\hat{\Sigma}$ consisting of $\hat{u}_i^2$ on the diagonals and estimate the variance of $\hat{\beta}_{\text{OLS}}$:

$$\left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1}\boldsymbol{W}'\hat{\boldsymbol{\Sigma}}\boldsymbol{W}\left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1}$$

- This is called the 'HC1' estimator ('`, r`' in Stata)

Standard Errors

We can take our estimate $\hat{\Sigma}$ consisting of $\hat{u}_i^2$ on the diagonals and estimate the variance of $\hat{\beta}_{\text{OLS}}$:

$$\left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1}\boldsymbol{W}'\hat{\boldsymbol{\Sigma}}\boldsymbol{W}\left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1}$$

- This is called the 'HC1' estimator ('`, r`' in Stata)

For inference on a coefficient, take square-root of the k-th diagonal element

- This is called the standard error (our estimate for the standard deviation of $\hat{\beta}_{\text{OLS},k}$)

Conditional Expectation Function

Linear Models

Ordinary Least Squares

Statistical Properties / Inference

Bivariate Regression Example

Forecasting with Regression Model

Marginal Effects

Monte Carlo example

From our simulation, the true regression line is

$$y_i = 0 + x_i * 1 + \varepsilon_i$$

- ε is homoskedastic so that $\sigma_i^2 = 1.5$ for all i
- $x_i \sim \mathcal{N}(1,1)$

Our regression model was $y_i = \beta_0 + x_i\beta_1 + u_i$, i.e. $W_i = (1, x_i)'$.

In our simulation, we can derive the variance of $\hat{\beta}_{\text{OLS}}$:

$$\boldsymbol{W}'\boldsymbol{W} = \begin{bmatrix} \sum_i 1 & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix} \approx n \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$$

And

$$\boldsymbol{W}'\Sigma\boldsymbol{W} = \begin{bmatrix} \sum_i \sigma_i^2 & \sum_i x_i\sigma_i^2 \\ \sum_i x_i\sigma_i^2 & \sum_i x_i^2\sigma_i^2 \end{bmatrix} \approx n \begin{bmatrix} 1.5 & 1.5 \\ 1.5 & 3 \end{bmatrix}$$

Sample distribution

Taking $\left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1}\boldsymbol{W}'\Sigma\boldsymbol{W}\left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1}$ yields

$$\approx \frac{1}{n}\begin{bmatrix} 3 & -1.5 \\ -1.5 & 1.5 \end{bmatrix}$$

- Check my linear algebra for practice

Sample distribution

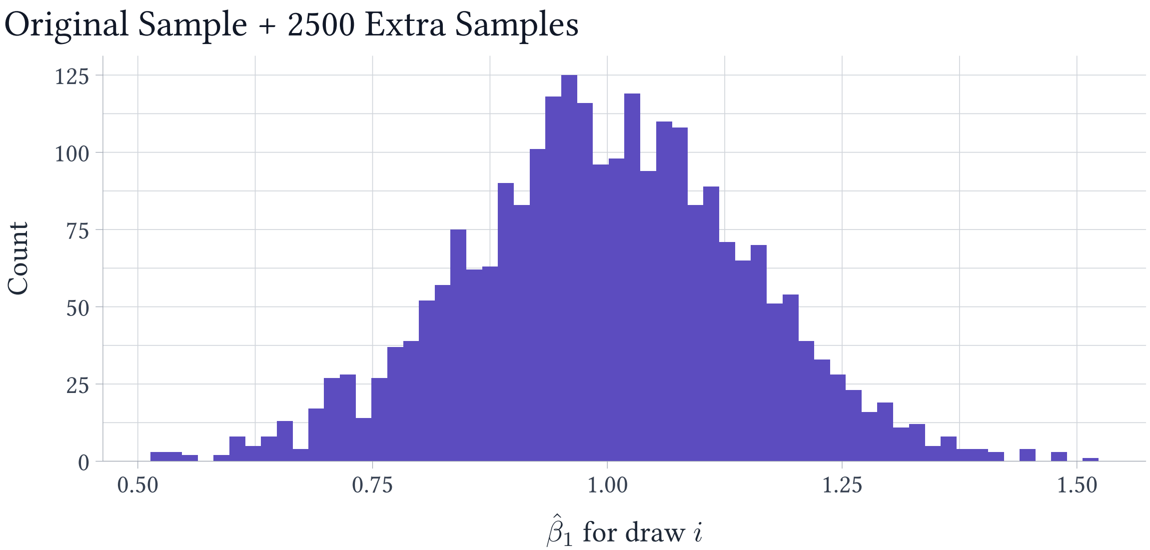
With our 100 observations, we have that

$$\mathsf{Var}\left(\hat{\beta}_{\mathsf{OLS}}\right) \approx \begin{bmatrix} 0.03 & -0.015 \\ -0.015 & 0.015 \end{bmatrix}$$

The standard deviation of $\hat{\beta}_1$ is $\sqrt{0.015} \approx 0.1225$.

- 95% of estimates should be 1 ± 0.245

Let's check that with our Monte Carlo simulations..



Original Sample + 2500 Extra Samples
Count
125
100
75
50
25
0
0.50
0.75
1.00
1.25
1.50
$\hat{\beta}_1$ for draw i

Sample Distribution of Regression Coefficients

In general, with homoskedastic errors, the slope coefficient has distribution:

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_{1,0}, \frac{1}{n}\frac{\mathsf{Var}(\varepsilon)}{\mathsf{Var}(X)}\right)$$

The standard error has the following properties:

- Shrinks with sample size
- Grows with the variance of the error term
- Shrinks with the variance of X

Standard Error

Our standard error estimator is given by

$$\mathsf{SE}(\hat{\beta}_1) = \sqrt{\frac{\mathsf{Var}(\hat{\varepsilon})/n}{\mathsf{Var}(X)}}$$

$\mathsf{Var}(\hat{\varepsilon})$ assumes homoskedasticity; otherwise we need to use the 'general' HC1 formula

Confidence intervals for $\hat{\beta}_1$

Since we have an approximately normally distributed random variable, we can form confidence intervals just like before:

$$\left[\hat{\beta}_1 - 1.96 * \mathsf{SE}(\hat{\beta}_1), \hat{\beta}_1 + 1.96 * \mathsf{SE}(\hat{\beta}_1)\right]$$

Confidence intervals for $\hat{\beta}_1$

Since we have an approximately normally distributed random variable, we can form confidence intervals just like before:

$$\left[\hat{\beta}_1 - 1.96 * \mathsf{SE}(\hat{\beta}_1), \hat{\beta}_1 + 1.96 * \mathsf{SE}(\hat{\beta}_1)\right]$$

The interpretation is as before: across repeated samples, 95% of samples' confidence intervals will contain the true value $\beta_{1,0}$

Conditional Expectation Function

Linear Models

Ordinary Least Squares

Statistical Properties / Inference

Bivariate Regression Example

Forecasting with Regression Model

Marginal Effects

Forecasting with our fitted model

We have a model

$$Y = \boldsymbol{W}\beta + u$$

that we fit using ordinary-least squares. From the previous section, we have

$$\hat{\beta}_{\text{OLS}} \sim \mathcal{N}\left(\beta_0, \boldsymbol{V}\right)$$

where $\boldsymbol{V} = \left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1}\boldsymbol{W}'\boldsymbol{\Sigma}\boldsymbol{W}\left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1}$

Forecasting with our fitted model

We want to evaluate this model at a particular value of W_i, we'll call it w. The forecasted value is given by

$$\hat{Y} = w'\hat{\beta}_{\textsf{OLS}} = \sum_{k=1}^{K} w_k \hat{\beta}_{\textsf{OLS},k}$$

Uncertainty from our regression coefficients will translate to uncertainty about our $\hat{Y}$

Forecasting with our fitted model

We have

$$\begin{aligned}\hat{Y} &= w'\hat{\beta}_{\text{OLS}} \\ &= \boxed{w'\beta_0} + w'\left(\hat{\beta}_{\text{OLS}} - \beta_0\right)\end{aligned}$$

$$= f_0(w) = \mathbb{E}[Y \mid X = x]$$

The forecasted value is the conditional expectation function (assuming our model is correct) plus noise

Inference on our Forecast

$$\hat{Y} = w'\hat{\beta}_{\text{OLS}}$$

Note that our forecast takes a normally distributed object $\hat{\beta}_{\text{OLS}}$, and mulitplies it by a row-vector, w. From topic 1, we have

$$\hat{Y} = w'\hat{\beta}_{\text{OLS}} \sim \mathcal{N}\left(w'\beta_0, w'\boldsymbol{V}w\right)$$

Let's illustrate this with our simulation. We will predict our regression model at $x = 1.5$.

Recall with $n = 100$, we had:

$$\mathsf{Var}\left(\hat{\beta}_{\mathsf{OLS}}\right) \approx \begin{bmatrix} 0.03 & -0.015 \\ -0.015 & 0.015 \end{bmatrix}$$

Our model has $\mathbb{E}[Y_i \mid X_i = 1.5] = 1 * 1.5 = 1.5$.

Monte Carlo simulation

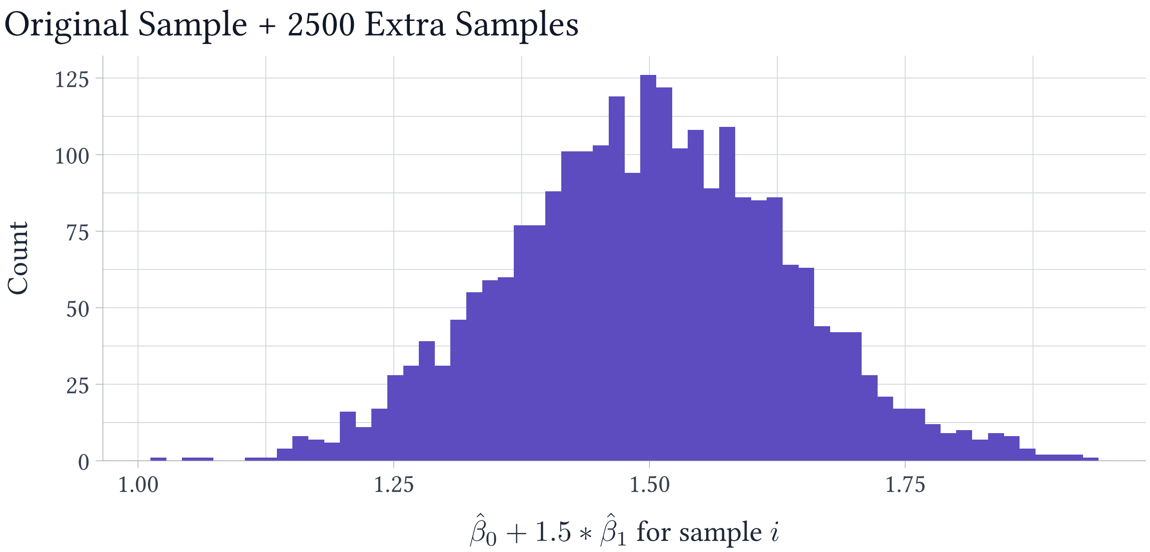
Our forecast, $\hat{Y}$ has variance

$$\begin{bmatrix} 1 & 1.5 \end{bmatrix} \begin{bmatrix} 0.03 & -0.015 \\ -0.015 & 0.015 \end{bmatrix} \begin{bmatrix} 1 \\ 1.5 \end{bmatrix} = 0.01875$$

The standard deviation of our forecast is $\sqrt{0.01875} \approx 0.137$.

- 95% of estimates should be 1.5 ± 0.274

Let's check that with our Monte Carlo simulations..


Original Sample + 2500 Extra Samples
Count
125
100
75
50
25
0
1.00
1.25
1.50
1.75
$\hat{\beta}_0 + 1.5 * \hat{\beta}_1$ for sample i

Often times, we want to compare forecasted values at two points: w_1 and w_2

> "Compare two individuals, one with value w_1 and one with value w_2. How do we predict $\hat{Y}_1$ and $\hat{Y}_2$ will differ?"

The simplest way to do this is to compare $w_1'\hat{\beta}_{\text{OLS}}$ and $w_2'\hat{\beta}_{\text{OLS}}$ directly

> "Compare two individuals, one with value w_1 and one with value w_2. How do we predict $\hat{Y}_1$ and $\hat{Y}_2$ will differ?"

It is important to remember that the goal of forecasting is to predict Y as well as possible

- When units have larger x, maybe they tend to have larger z_1 and smaller z_2. Regression will use that information when predicting $\hat{\beta}_{\text{OLS}}$

Correct regression interpretation

Avoid making causal claims that *changing* w_k *causes* a change in Y

✓ Our regression model predicts that a one unit increase in w_k is associated with a $\hat{\beta}_{\text{OLS},k}$ units increase/decrease in Y

✓ Compare two people, one with a value of $w_k = \tilde{w}$ and one with $w_k = \tilde{w} + 1$. Our regression model predicts the latter has $\hat{\beta}_{\text{OLS},k}$ increase/decrease in Y.

× Increase w_k by one unit increases/decreases Y by $\hat{\beta}_{\text{OLS},k}$ units

Correct regression interpretation

In general, you should use the following language:

Correct regression interpretation

Often we want to think about changing X_i instead of changing W_i;

- E.g. if we change age (X_i), we change age ($W_{2,i}$) and age^2 ($W_{3,i}$)

To make this more clear, we can write our model, noting the dependence of W on X:

$$f(X) = W(X)\beta = \sum_{k=1}^{K} g_k(X)\beta_k$$

Marginal (predictive) Effects

We can ask how $\hat{f}(X)$ changes when we change one element of X, x_ℓ (e.g. age).

To do so, we can take the derivative of $\hat{f}(X)$ with respect to x_ℓ and plug in a point X

$$\frac{\partial}{\partial x_\ell}\hat{f}(X) = \sum_{k=1}^{K} \frac{\partial}{\partial x_\ell} g_k(X)\hat{\beta}_{\text{OLS},k}$$

This is called the marginal (predictive) effect of x_ℓ

- I put predictive to emphasize this is not the *causal* effect of experimentally changing x_ℓ for a unit

Marginal (predictive) Effects

$$\frac{\partial}{\partial x_\ell}\hat{f}(X) = \sum_{k=1}^{K} \frac{\partial}{\partial x_\ell} g_k(X)\hat{\beta}_{\text{OLS},k}$$

In the case where we just include each variable linearly, i.e. $g_k(X) = X_k$, then this reduces to the standard $\hat{\beta}_{\text{OLS},k}$ being our estimated marginal effect.

Marginal (predictive) Effects

$$\frac{\partial}{\partial x_\ell}\hat{f}(X) = \sum_{k=1}^{K} \frac{\partial}{\partial x_\ell} g_k(X) \hat{\beta}_{\text{OLS},k}$$

In the case where we just include each variable linearly, i.e. $g_k(X) = X_k$, then this reduces to the standard $\hat{\beta}_{\text{OLS},k}$ being our estimated marginal effect.

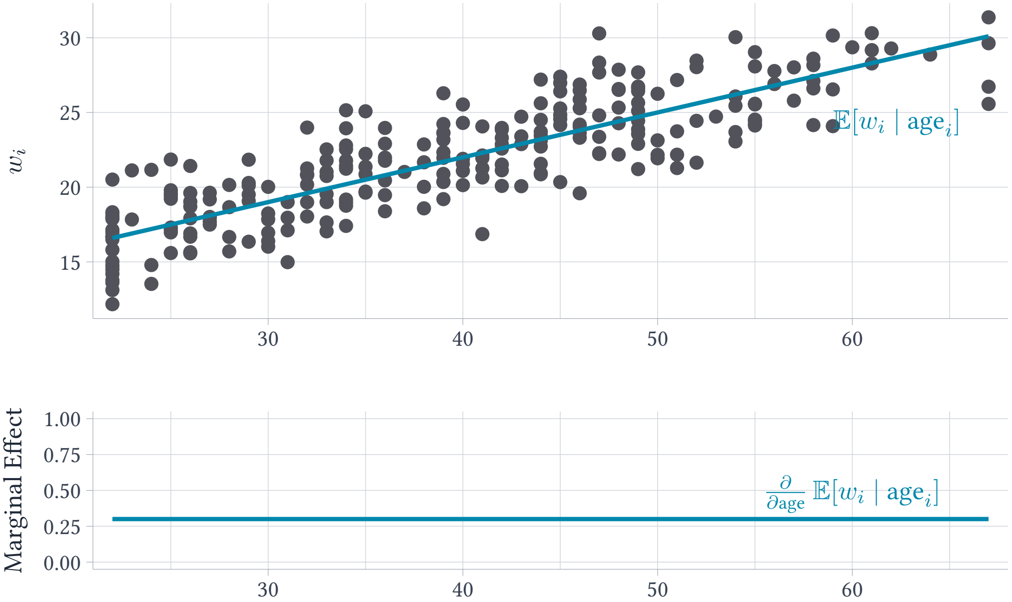
In the next topic, we will practice this when we have other functions of variables in g_k

Marginal Effects

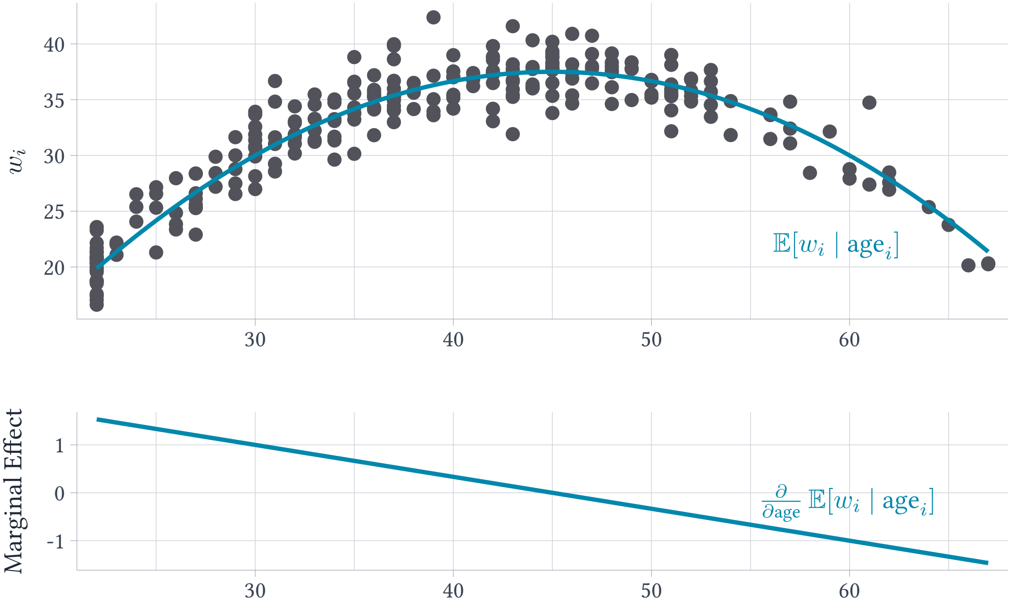
$$\frac{\partial}{\partial x_\ell}\hat{f}(X) = \sum_{k=1}^{K} \frac{\partial}{\partial x_\ell} g_k(X)\hat{\beta}_{\mathsf{OLS},k}$$

This is holding fixed all the other valuables at the original covariate values: $x_{1,i}, \ldots, x_{K,i}$ and only changing x_ℓ

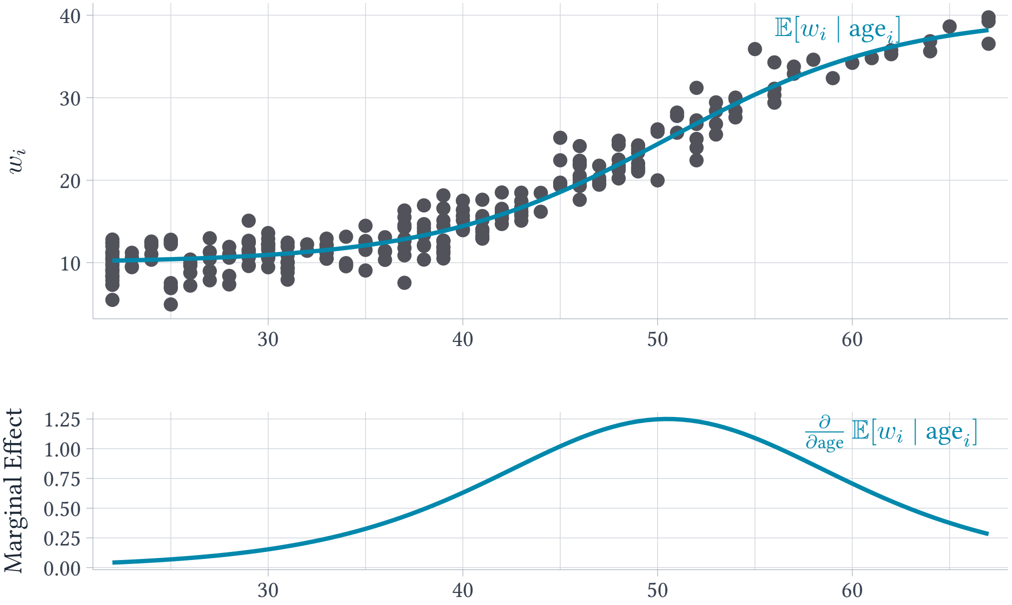
- multiple W_k can change from changing a particular x_ℓ



w_i
30
25
20
15
$\mathbb{E}[w_i \mid \text{age}_i]$
30
40
50
60
Marginal Effect
1.00
0.75
0.50
0.25
0.00
$\frac{\partial}{\partial \text{age}} \mathbb{E}[w_i \mid \text{age}_i]$
30
40
50
60



w_i
40
35
30
25
20
30
40
50
60
$\mathbb{E}[w_i \mid \text{age}_i]$
Marginal Effect
1
0
-1
30
40
50
60
$\frac{\partial}{\partial \text{age}} \mathbb{E}[w_i \mid \text{age}_i]$



w_i
40
30
20
10
30
40
50
60
$\mathbb{E}[w_i \mid \text{age}_i]$
Marginal Effect
1.25
1.00
0.75
0.50
0.25
0.00
30
40
50
60
$\frac{\partial}{\partial \text{age}} \mathbb{E}[w_i \mid \text{age}_i]$