All Things Regression

Part I

Fernando Hoces la Guardia
07/19/2022

Regression Journey

- Regression as Matching on Groups. Ch2 of MM up to page 68 (not included).
- Regression as Line Fitting and Conditional Expectation. Ch2 of MM, Appendix.
- Multiple Regression and Omitted Variable Bias. Ch2 of MM pages 68-79 and Appendix.
- All Things Regression: Anatomy, Inference, Logarithms, Binary Outcomes, and R^2. Ch2 of MM, Appendix + others.

Regression Journey

- Regression as Matching on Groups. Ch2 of MM up to page 68 (not included).
- Regression as Line Fitting and Conditional Expectation. Ch2 of MM, Appendix.
- Multiple Regression and Omitted Variable Bias. Ch2 of MM pages 68-79 and Appendix.
- **All Things Regression: Anatomy, Inference, Logarithms, Binary Outcomes, and R^2. Ch2 of MM, Appendix + others.**

Today and Tomorrow's Lecture

- Regression Anatomy
- Regression Inference
- Non-linearities:
  - Logarithms
  - Others
- Binary Outcomes
- R^2

Regression Anatomy

Regression Anatomy

- In addition to the intuition of regression as matching in subgroups, here we will explore another interpretation of what does it mean to control for multiple variables (regressors)

- We started with our exploration to regression with just on regressor: $Y_i = \alpha + \beta P_i + e_i$

- We then added multiple regressors and interpreted the beta coefficient as a weighted average of difference within subgroups.

- The first resgression is sometimes called a bivariate regression (or bivariate analysis, aka univariate analysis, in the sense that there is only one independent variable).

"Controlling For" a Second Interpretation 1/2

- In a **multiple** regression like the following:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$$

- The coefficient of X_{1i} (β_1) is the same as the one obtained from a **bivariate** regression between the outcome variable (Y_i) and the residual term $\widetilde{X}_{1i}$, that corresponds to the following (auxiliary) regression:

$$X_{1i} = \pi_0 + \pi_1 X_{2i} + \widetilde{X}_{1i}$$

Meaning:

$$\beta_1 = \frac{\text{Cov}\left(\text{Y}_i, \widetilde{X}_{1i}\right)}{\text{Var}\left(\widetilde{X}_{1i}\right)}$$

"Controlling For" a Second Interpretation 2/2

$$X_{1i} = \pi_0 + \pi_1 X_{2i} + \widetilde{X}_{1i}$$

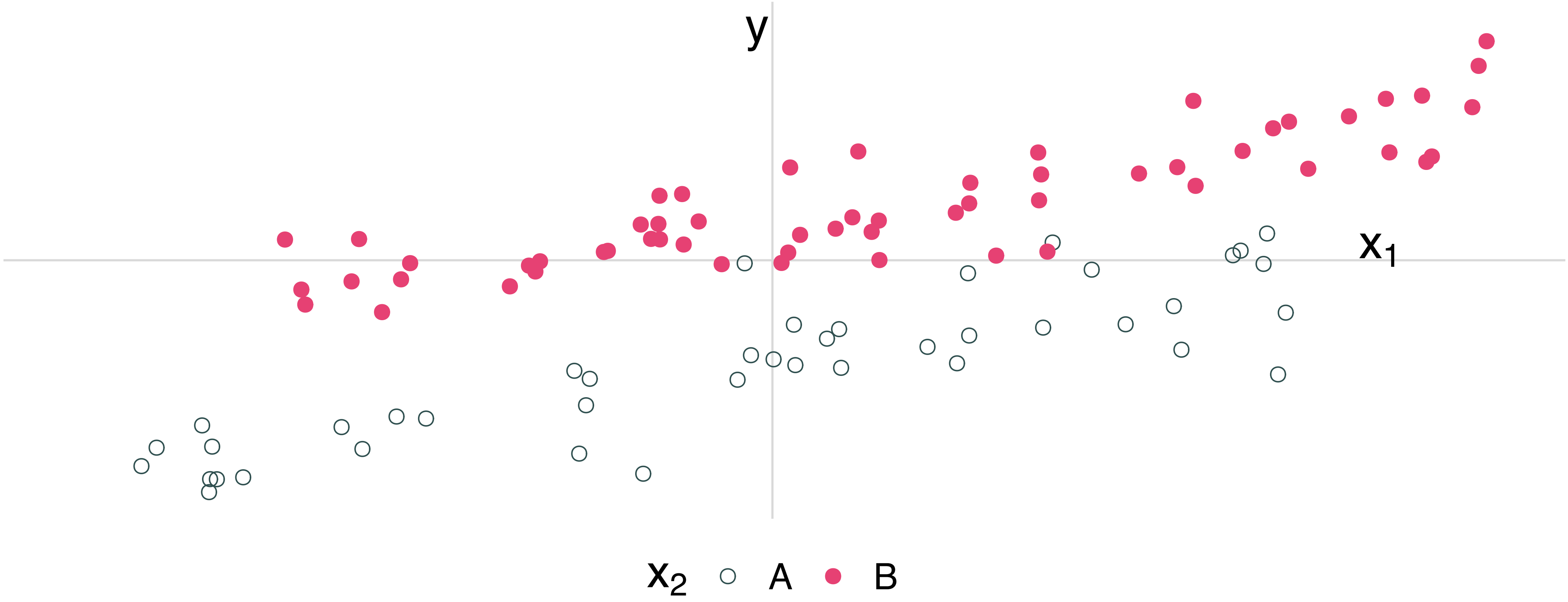
- Let's think about what this residual means:
  - All variation (information) in X_{1i} that cannot be explained by variation (information) in X_{2i}.
  - Then the bivariate regression (of Y_i and $\widetilde{X}_{1i}$) is basically regressing Y_i on "all of X_{1i} that is not explained by X_{2i}" or "all of X_{1i}, removing, or controlling for, the variation in X_{2i}"

Regression Anatomy: Visually

- This formula also applies for the residual after regression Y_i on X_{2i}, and this last one has a nice visual interpretation.
- (Regression Anatomy here is a simplified version of a more general idea called the Frisch-Waugh-Lovell theorem, it is outside of the scope of the course, but if you learn linear algebra, it has a really cool interpretation)
- Graphical example (Again from the great slides of Ed Rubin) for the case where X_2i is a binary variable

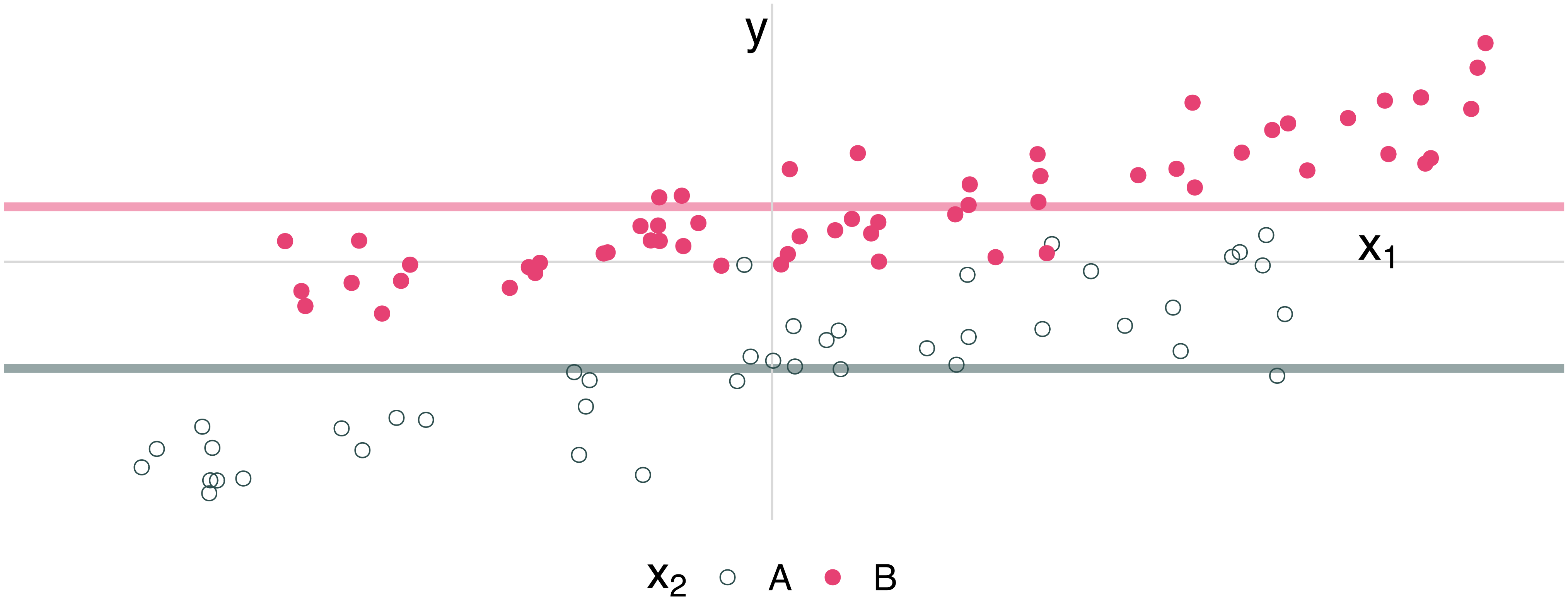
Regression Anatomy: Visually

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$$



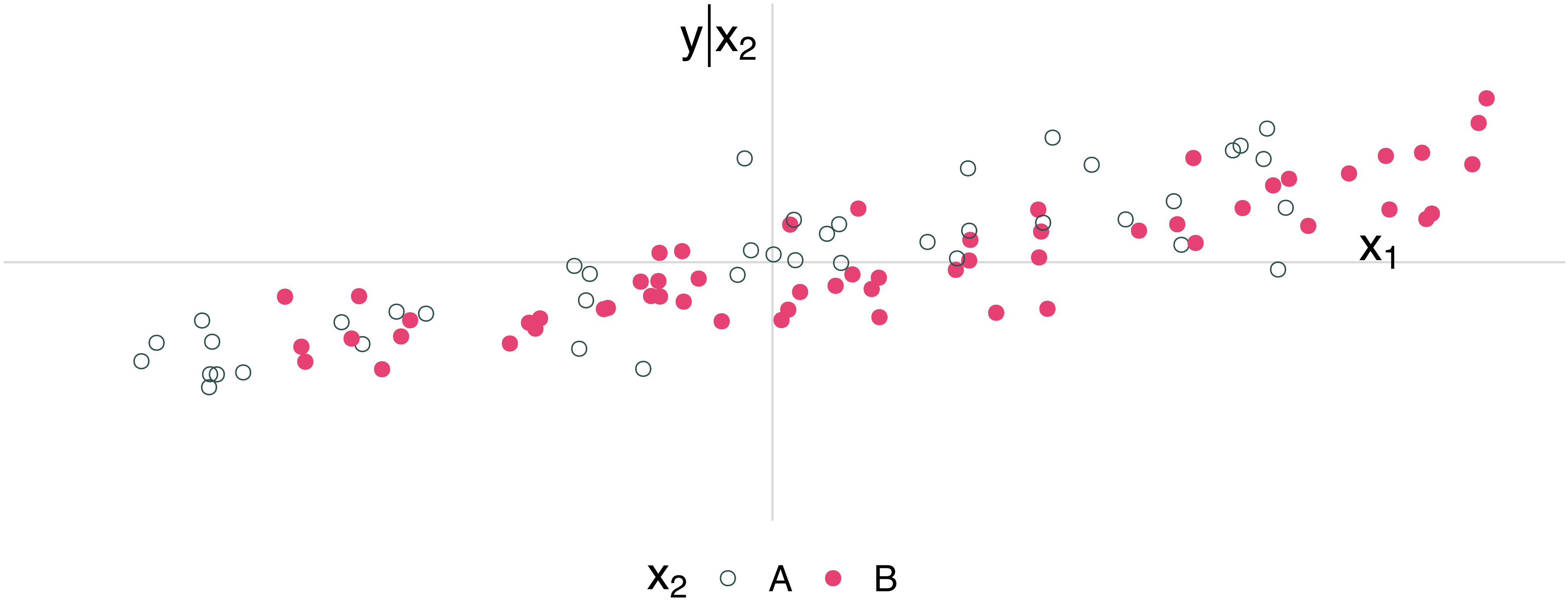
Regression Anatomy: Visually

β_1 gives the relationship between y and x_1 *after controlling for* x_2



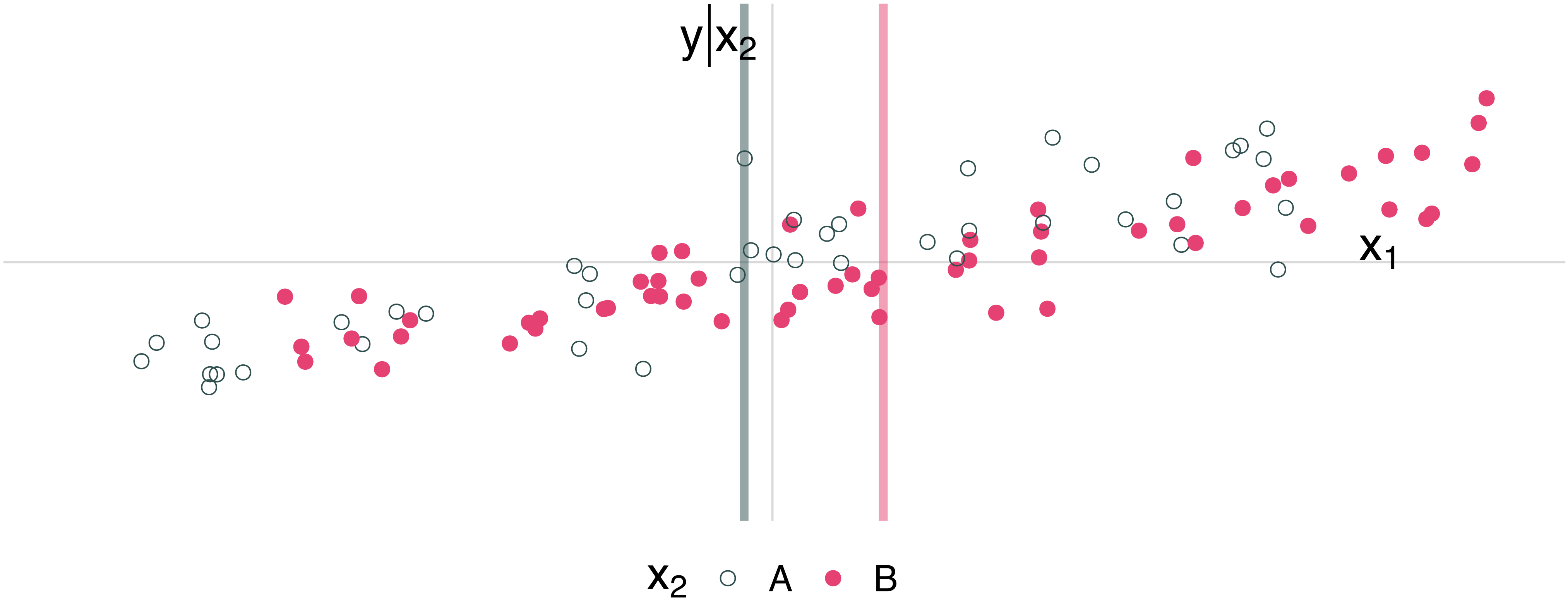
Regression Anatomy: Visually

β_1 gives the relationship between y and x_1 *after controlling for* x_2



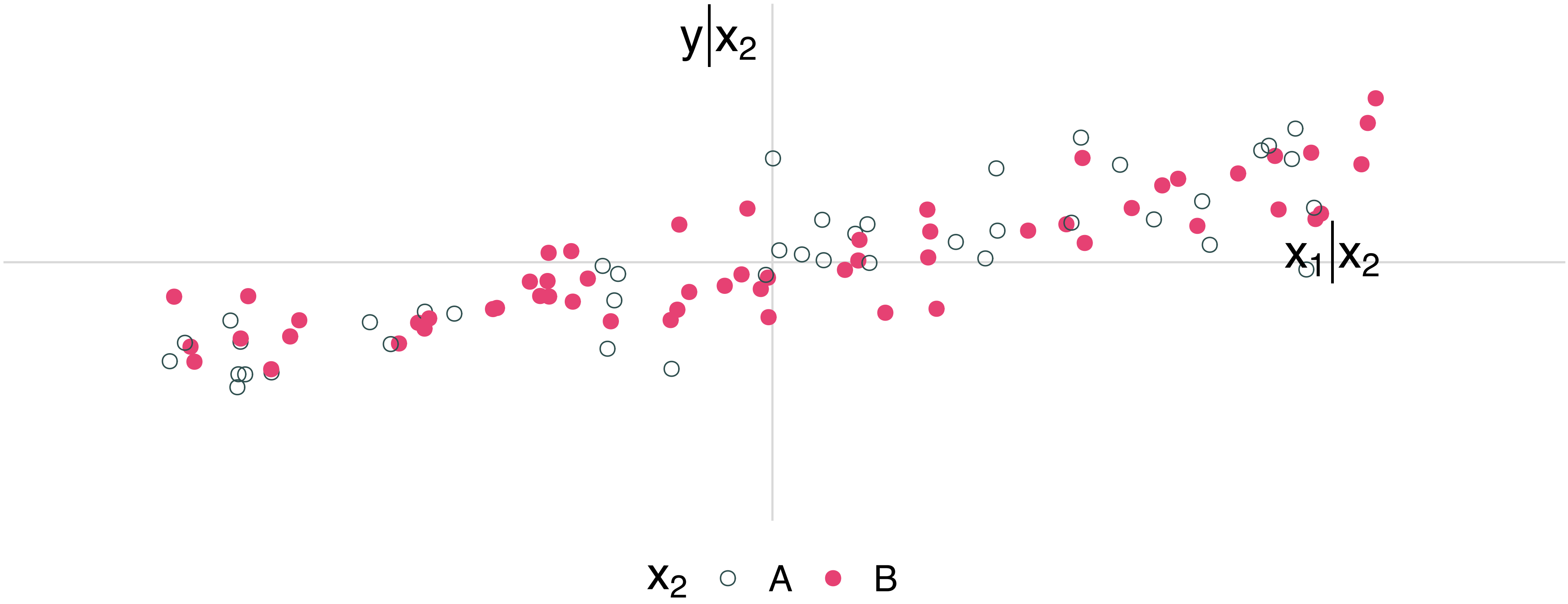
Regression Anatomy: Visually

β_1 gives the relationship between y and x_1 *after controlling for* x_2



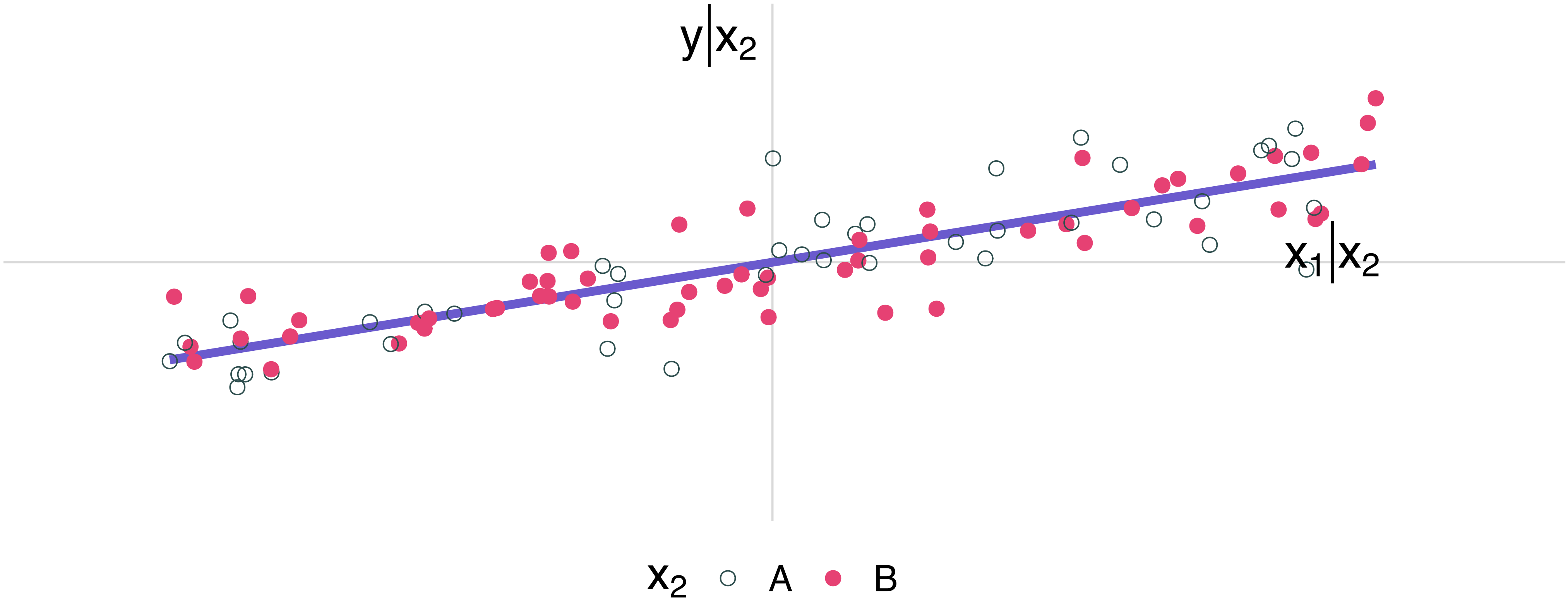
Regression Anatomy: Visually

β_1 gives the relationship between y and x_1 *after controlling for* x_2



Regression Anatomy: Visually

β_1 gives the relationship between y and x_1 *after controlling for* x_2



Regression Anatomy: Many Regressors 1/2

- This logic, of removing the variation explained by other regressors and turning a multivariate regression into a bivariate regression, applies to any number of regressors.

- Hence the multivariate regression:

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \ldots \beta_K X_{Ki} + e_i}$$

- The coefficient of $\boldsymbol{X_{ki}}$ $(\boldsymbol{\beta_k})$ is the same as the one obtained from a bivariate regression between the outcome variable $(\boldsymbol{Y_i})$ and the residual term $\widetilde{\boldsymbol{X}}_{\boldsymbol{ki}}$, that corresponds to the following (auxiliary) regression:

$$\boldsymbol{X_{ki} = \pi_0 + \pi_1 X_{1i} + \pi_1 X_{2i} + \ldots \pi_{k-1} X_{k-1,i} + \pi_{k+1} X_{k+1,i} + \ldots + \beta_K X_{Ki} + \widetilde{X}_{k1i}}$$

Regression Anatomy: Many Regressors 2/2

With:

$$\beta_k = \frac{\mathrm{Cov}\left(\mathrm{Y}_i,\, \widetilde{X}_{ki}\right)}{\mathrm{Var}\left(\widetilde{X}_{ki}\right)}$$

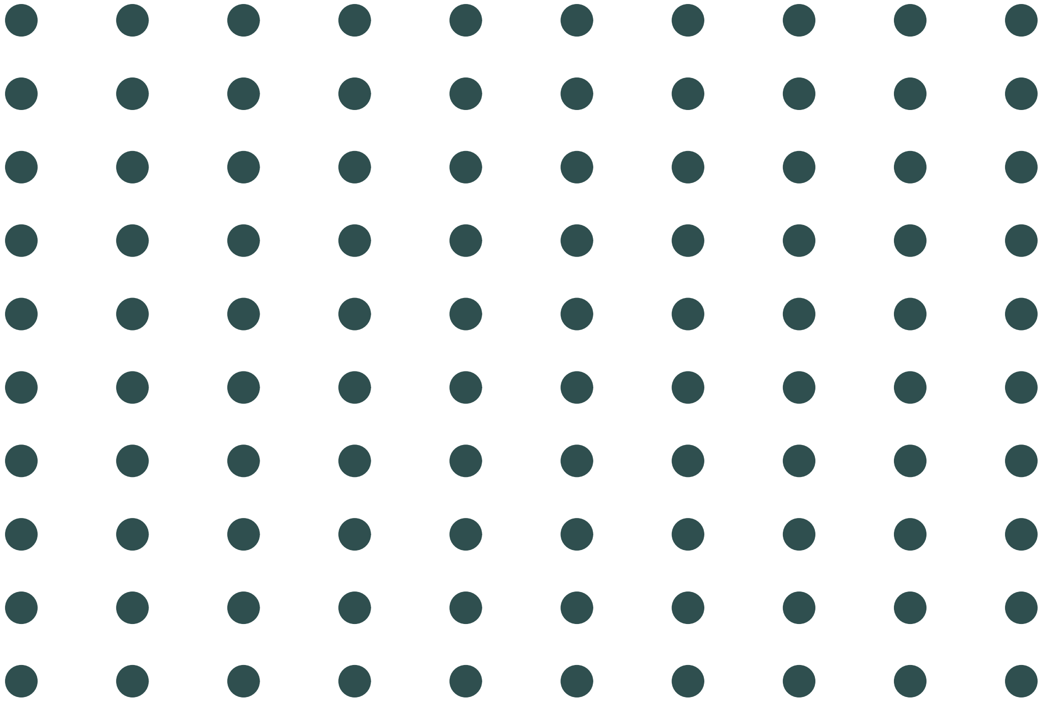
- With this approach, "controlling for" can be understood as "removing all the variation between the variable of interest (X_{ki}) and all the other controls"

Today and Tomorrow's Lecture

- Regression Anatomy
- **Regression Inference**
- Non-linearities:
  - Logarithms
  - Others
- Binary Outcomes
- R^2

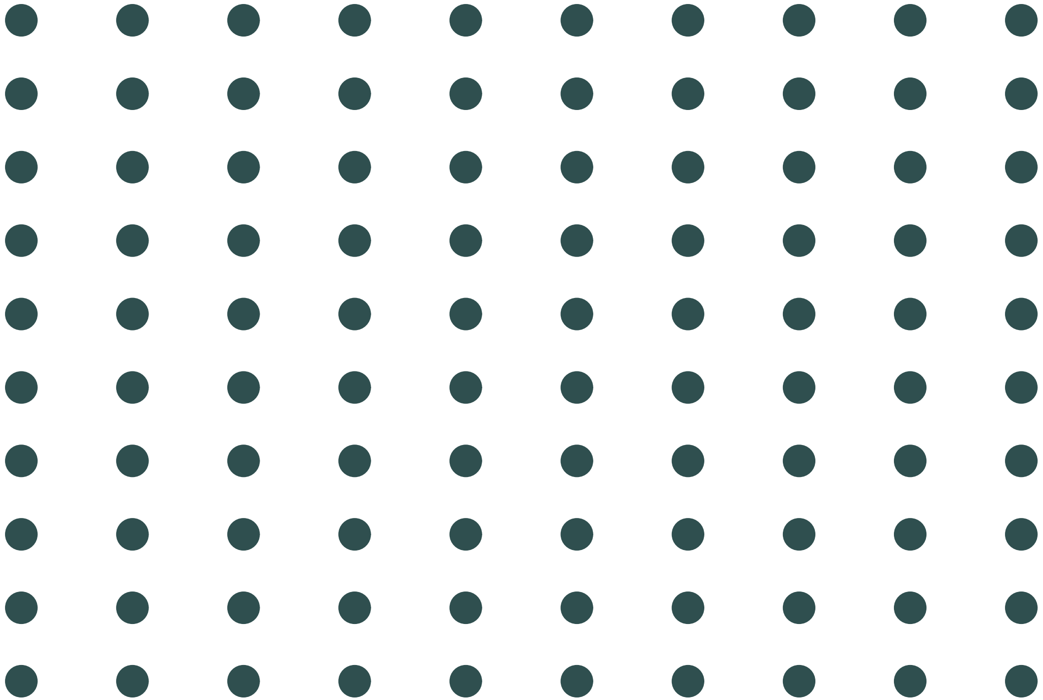
Regression Inference

Until Now We Have Focus On **The** Population Regression

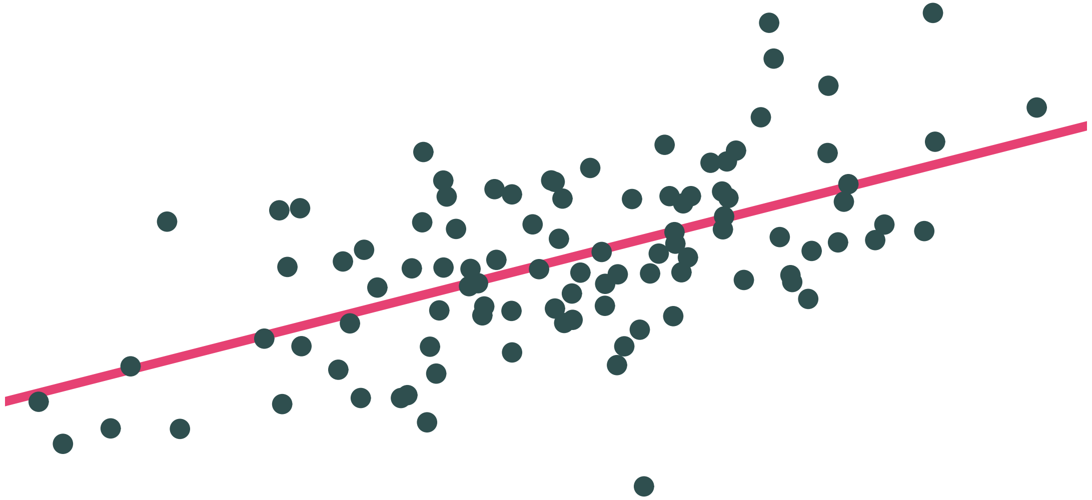


Population

Until Now We Have Focus On **The** Population Regression



Population

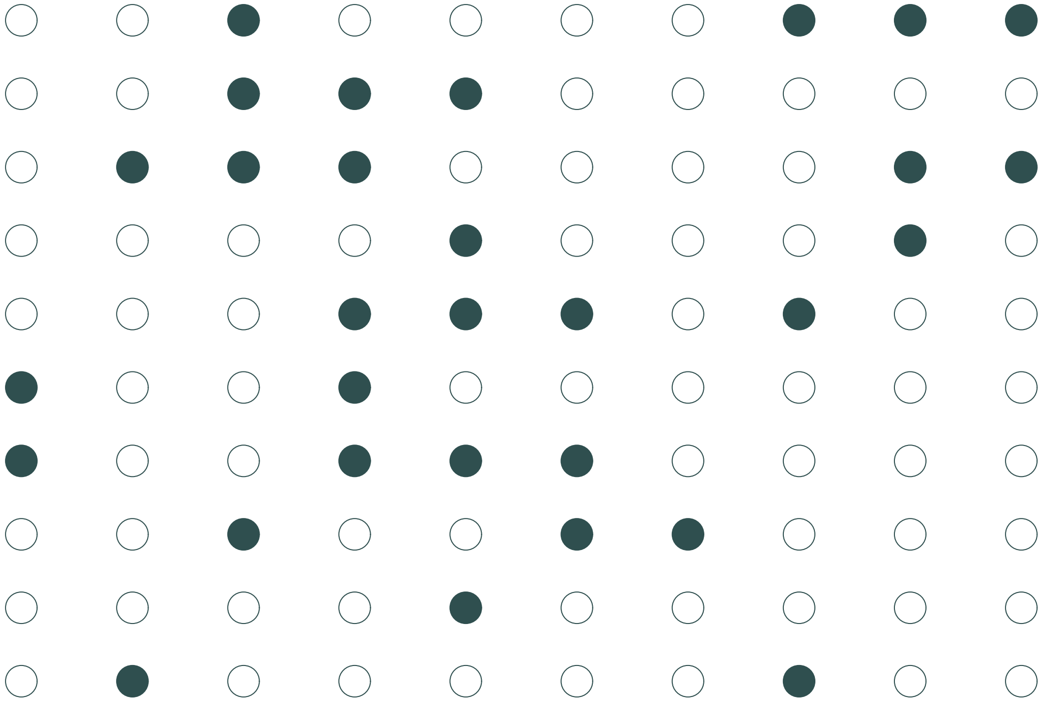


Population relationship

$$Y_i = 2.53 + 0.57X_i + e_i$$

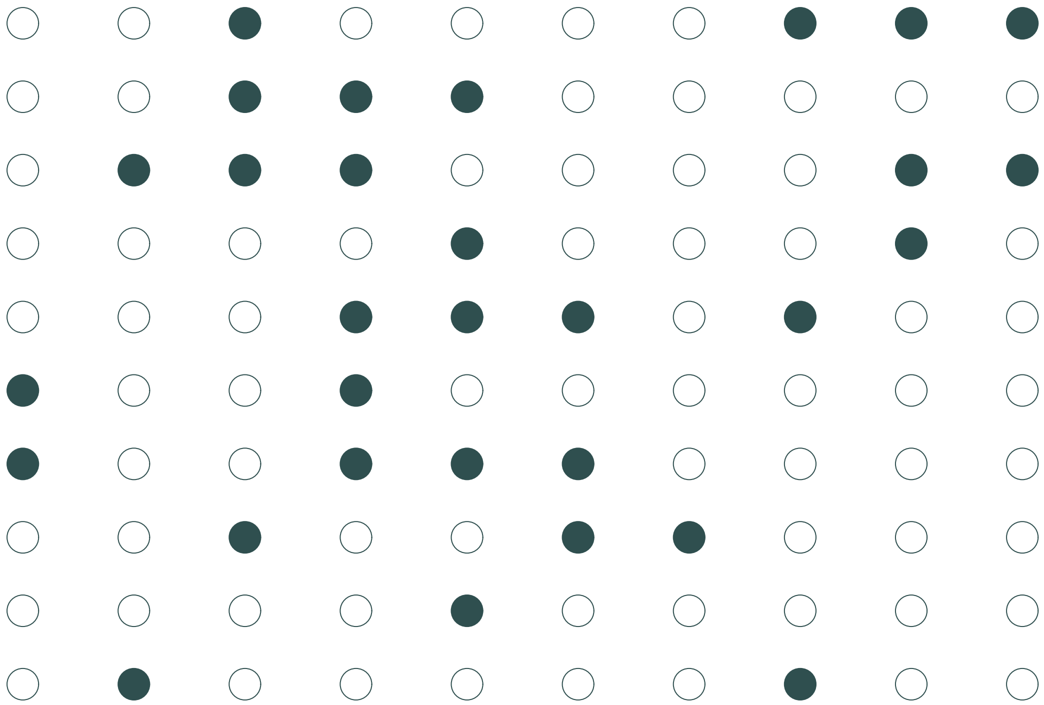
$$Y_i = \alpha + \beta X_i + e_i$$

Now We Focus On Regressions Based On A Sample

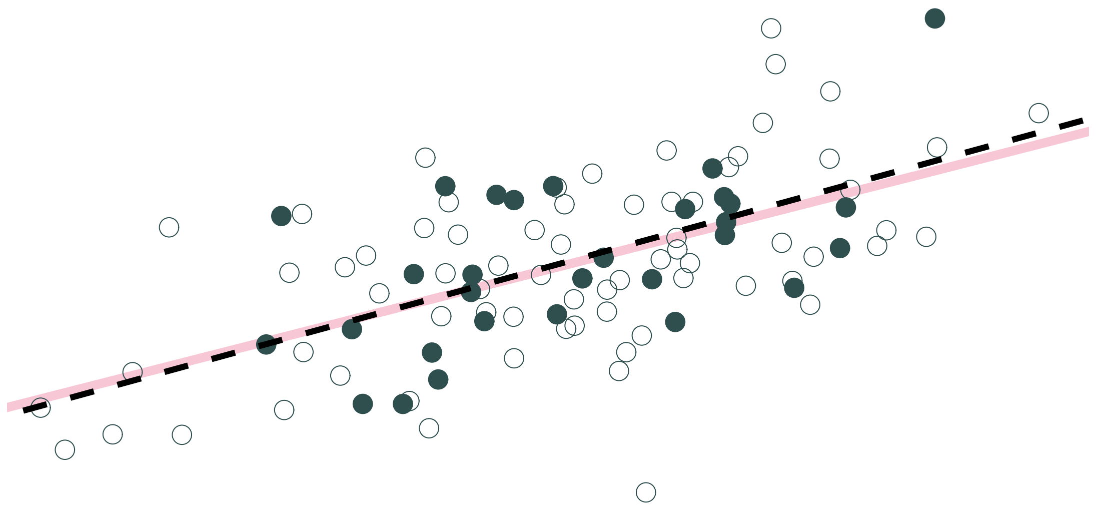


Sample 1: 30 random individuals

Now We Focus On Regressions Based On A Sample



Sample 1: 30 random individuals



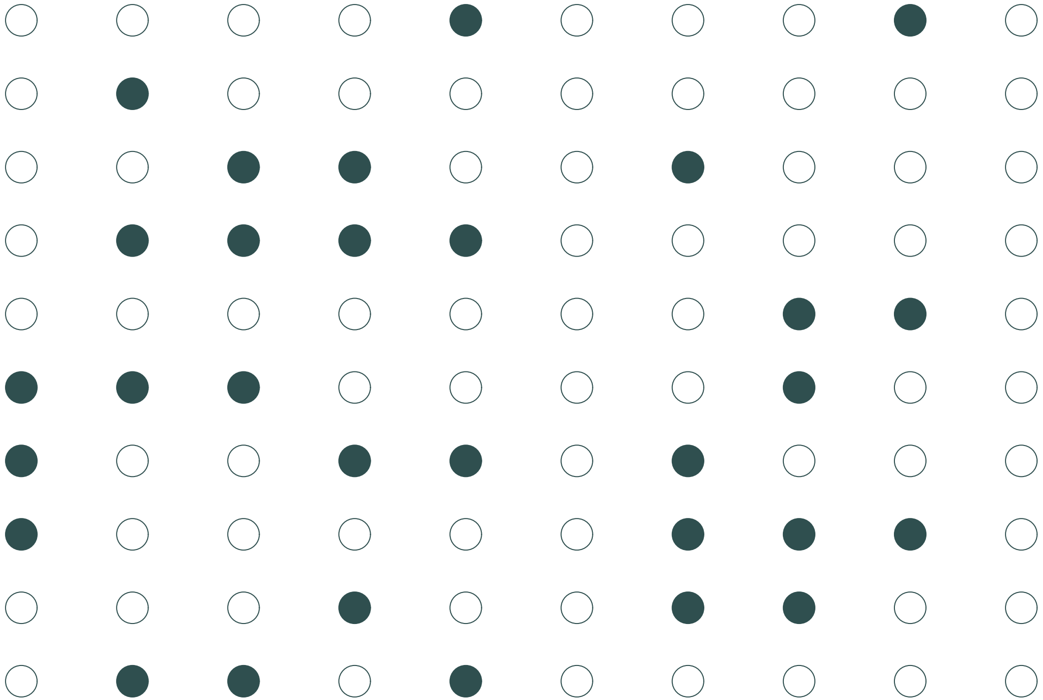
Population relationship

$$Y_i = 2.53 + 0.57X_i + u_i$$

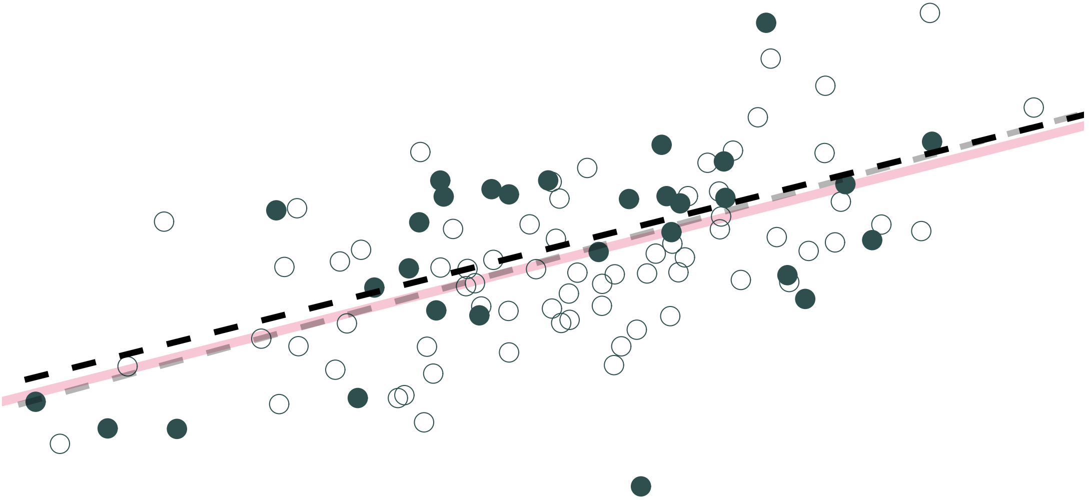
Sample relationship

$$\hat{Y}_i = 2.36 + 0.61X_i$$

Now We Focus On Regressions Based On A Sample



Sample 2: 30 random individuals



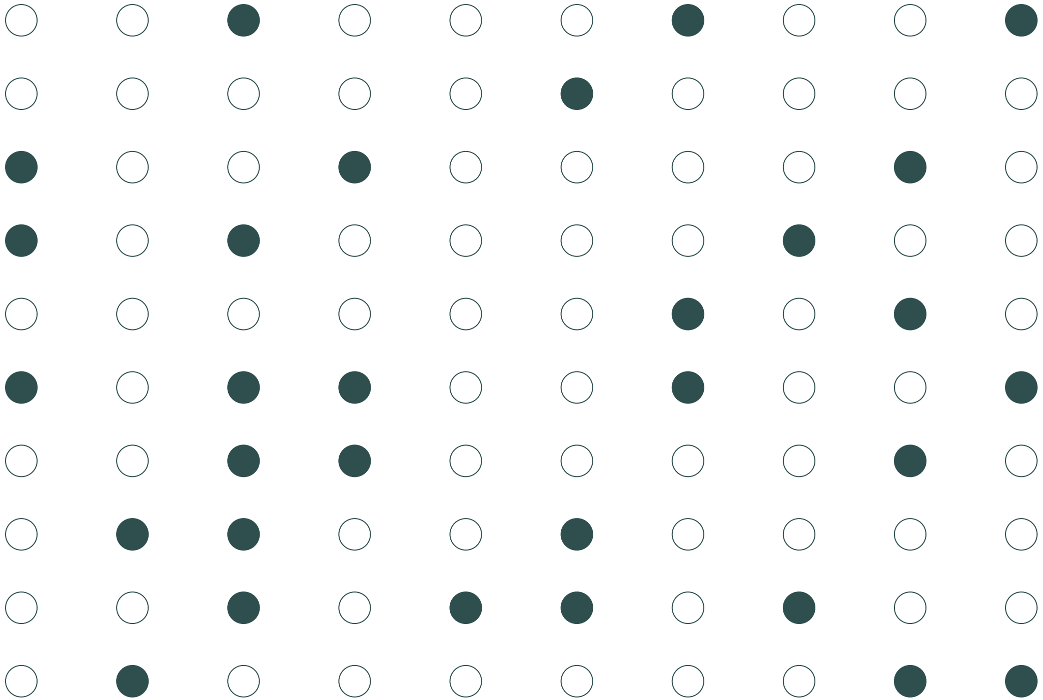
Population relationship

$$Y_i = 2.53 + 0.57X_i + u_i$$

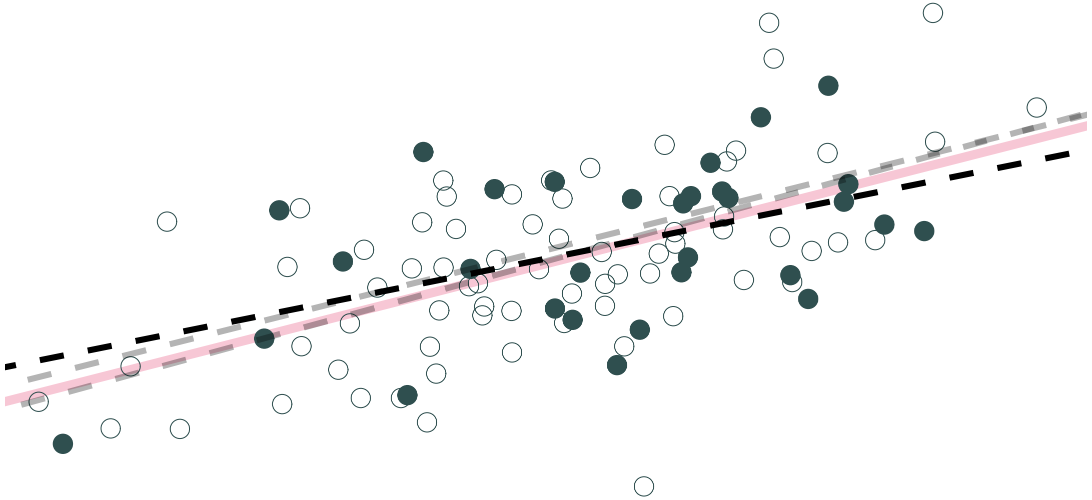
Sample relationship

$$\hat{Y}_i = 2.79 + 0.56X_i$$

Now We Focus On Regressions Based On A Sample



Sample 3: 30 random individuals



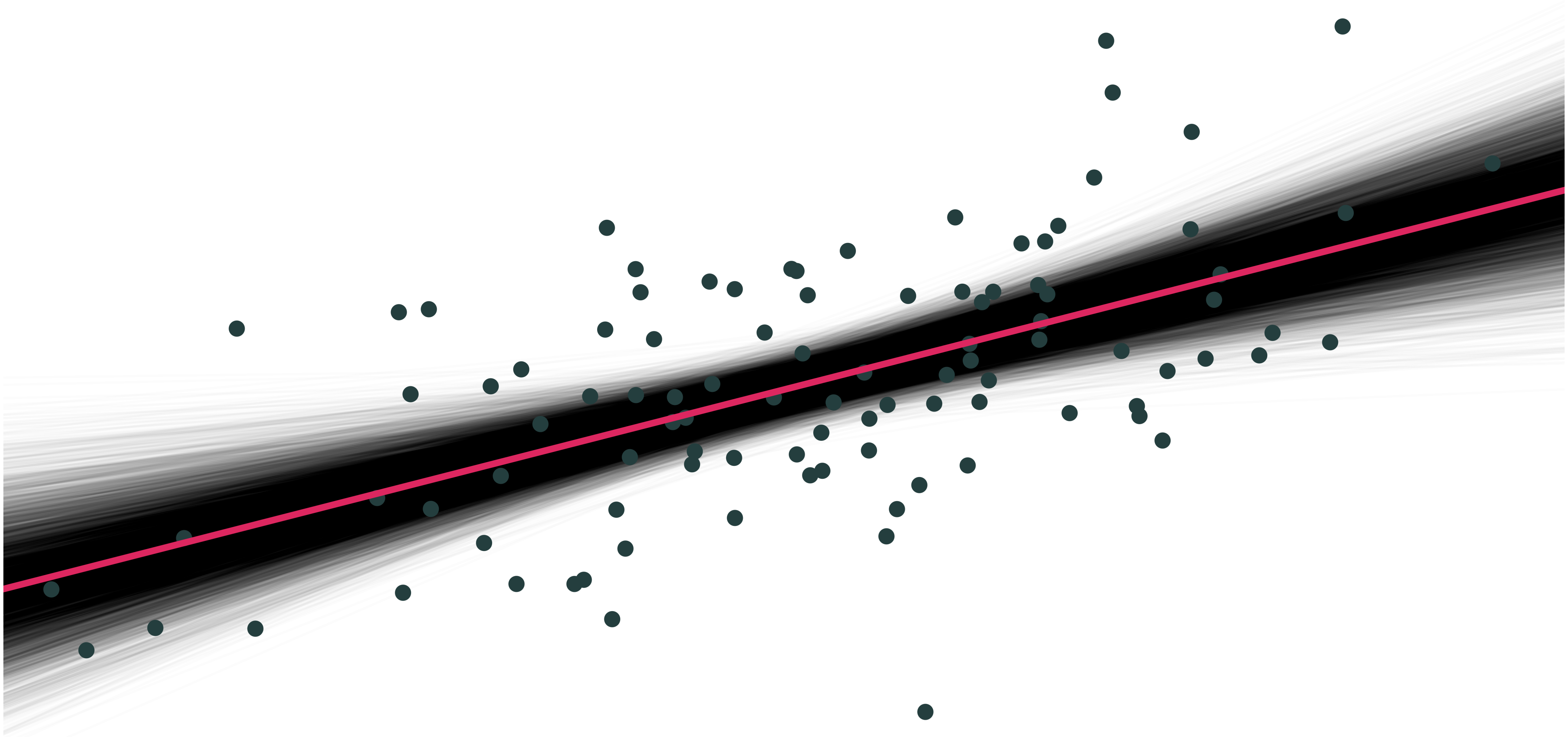
Population relationship

$$Y_i = 2.53 + 0.57X_i + u_i$$

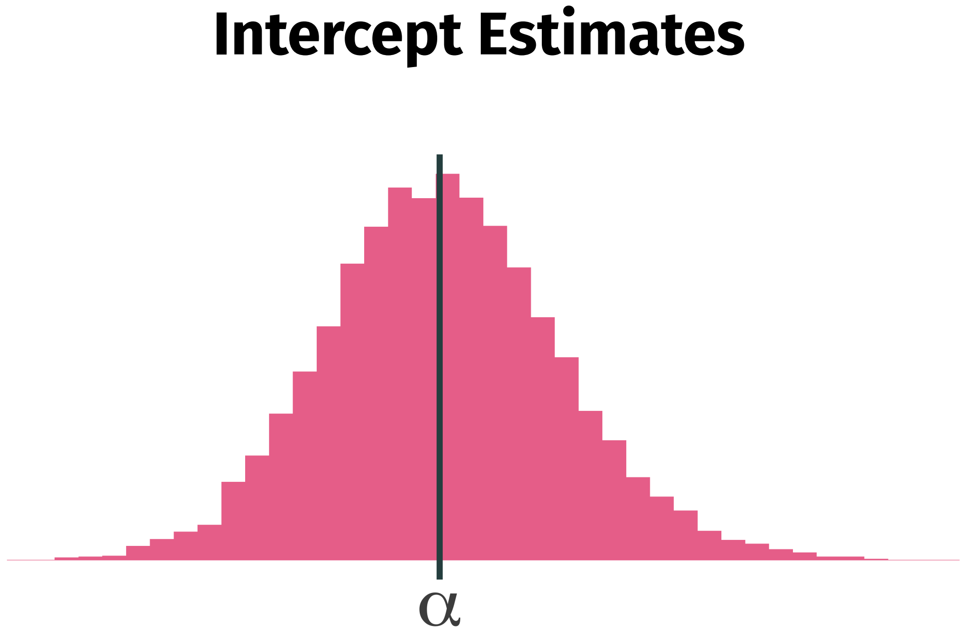
Sample relationship

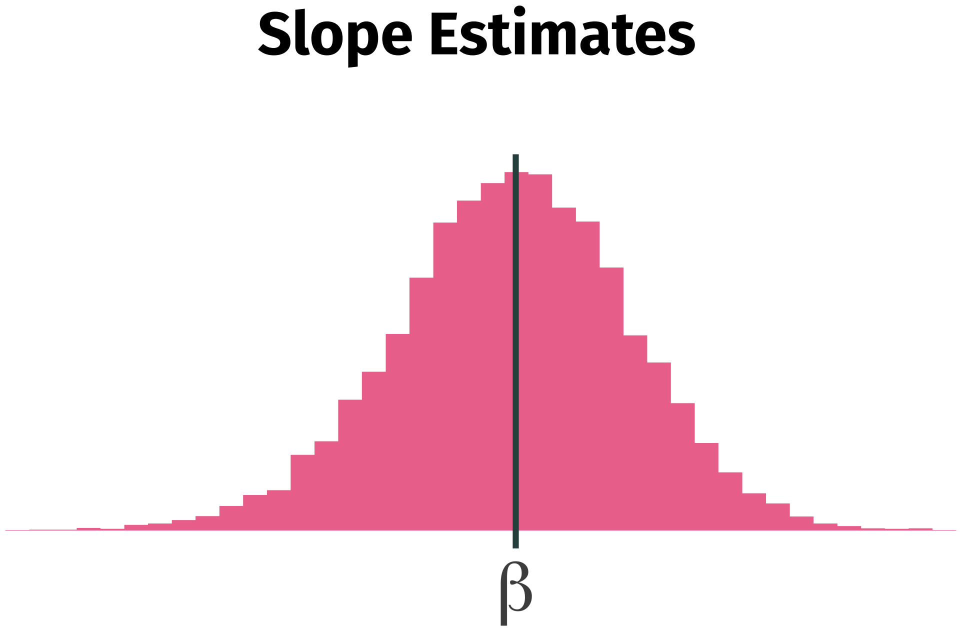
$$\hat{Y}_i = 3.21 + 0.45X_i$$

Repeat **10,000 times** (Monte Carlo simulation).



CLT in Action





- The estimated coefficients are a linear combination (similar to a summation) of independent random variables. Hence the CLT applies.

- Let $\widehat{\beta}$ be the estimated coefficient of the slope, CLT tells us: $\widehat{\beta} \sim \boldsymbol{N}(\beta, \boldsymbol{SE}(\widehat{\beta}))$

Standard Errors of Estimated Coefficients 1/3

- Remember that the standard deviation of the sample mean, what we called standard errors, is:

$$SE(\overline{Y}) = \frac{\sigma_Y}{\sqrt{n}}$$

- A similar formula applied also to the difference in means $\widehat{\mu} = \overline{Y}_1 - \overline{Y}_0$.

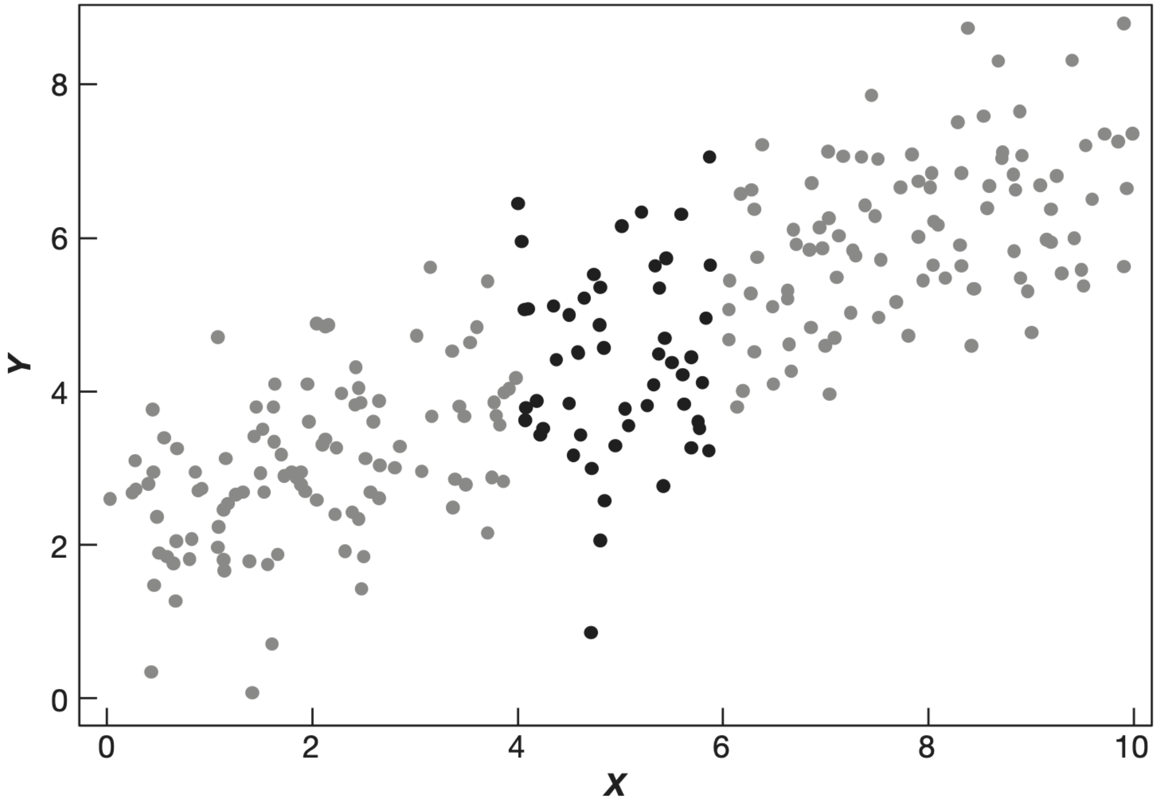
- Following a similar intuition, here we will state that the standard error of the estimated regression coefficient of interest is:

$$SE(\widehat{\beta}) = \frac{\sigma_e}{\sqrt{n}} \times \frac{1}{\sigma_X}$$

Standard Errors of Estimated Coefficients 2/3

- One regressor: $SE(\widehat{\beta}) = \frac{\sigma_e}{\sqrt{n}} \times \frac{1}{\sigma_X}$
- n plays a similar role as for the previous SEs.
- σ_e: is the standard deviation of the residual. As X explains (fits) more of Y this standard deviation gets smaller. As X explains more of Y, the precision of $\widehat{\beta}$ increases.
- σ_X: is the standard deviation of the variable X. As X varies more, the precision of $\widehat{\beta}$ increases.

FIGURE 2.2
Variance in X is good



Standard Errors of Estimated Coefficients 3/3

- The standard error of a coefficient $\widehat{\beta}_k$ in a multivariate regression is:

$$SE(\widehat{\beta}_k) = \frac{\sigma_e}{\sqrt{n}} \times \frac{1}{\sigma_{\widetilde{X}_k}}$$

- Where $\widehat{\beta}_k$ comes from a multivariate regression:

$$Y_i = \alpha + \sum_{k=1}^{K} \beta_k X_{ki} + e_i$$

- And $\widetilde{X}_{ki}$ is the residual from regression anatomy:

$$X_{ki} = \pi_0 + \sum_{j=1}^{k-1} \pi_j X_{ki} + \sum_{j=k+1}^{K} \pi_j X_{ki} + \widetilde{X}_{ki}$$

- $\sigma_{\widetilde{X}_k}$ is the standard deviation of the residual $\widetilde{X}_{ki}$. It represents all the variation that is left in X_k after controlling for all other regressors. By construction it will be less than σ_{X_k}. Notice the trade-off of adding more regressors.

Collinearity 1/2

- Collinearity is a problem of regression that happens when two or more regressors are closely correlated ("colinear").

- In the non-extreme case of perfect collinearity, regression will still work, but the resulting SE will be inflated. Let's look at the SE formula to see why:

$$SE(\widehat{\beta}_k) = \frac{\sigma_e}{\sqrt{n}} \times \frac{1}{\sigma_{\widetilde{X}_k}}$$

- If X_k is highly collinear, with one or more other regressors, it will render a very small residual in the auxiliary regression, resulting in turn in a very small $\sigma_{\widetilde{X}_k}$. Given that this last term is in the denominator, the SE will become very large, rendering any coefficient statistically insignificant.

Collinearity 2/2

$$SE(\widehat{\beta}_k) = \frac{\sigma_e}{\sqrt{n}} \times \frac{1}{\sigma_{\widetilde{X}_k}}$$

- The extreme version of this problem is when one regressor is perfectly correlated with one or more regressors (making it a linear combination of the regressor).

- In this case the residual is zero, and so is its variance in the auxiliary regression.

- Under perfect collinearity (aka multicollinearity) the software that is runnin the regression will do one of two things: (i) drop one or more of the regressor to avoid perfect collinearity, or (ii) don't run the regression (saying something like "cannot invert matrix").

- (Perfect collinearity is the reason why we don't include two binary variables two describe two groups, as they would be perfectly collinear with the intercept)

Robust Standard Errors

- One underlying assumption behind the SEs discussed so far is that the residual does not change in a systematic way across the Xs.

- For an example of how this assumption does not hold, look draw this pattern on the board.

- There is a modified version of the SEs that is robust to this problem. In the sense that when the problem is present, it solves it, and when its not, it doesn't do harm.

- This is the most common formula for standard errors that is reported in current research.

Repeat Hypothesis Testing and Confidence Intervals

- Now that we have our SEs, the procedure to conduct hypothesis tests, and build confidence intervals for estimated coefficients $(\widehat{\beta})$, is the same as discussed in the statistical inference lecture:

1. Define a null hypothesis β_0 (usually $\beta_0 = 0$)
2. Construct a t-statistic: $t(\beta_0)$ by subtracting the null and dividing by the SE.
3. Compute the p-value as probability that we observe a t-statistic as extreme as the obtained in the sample, if the null is true. You don't need to obtain the exact p-value, but you are asked to remember that the probability that this t-statistic is larger than 1 is about 30%, of being larger than 2 is about 5%, and of being larger than 3 is less than 1% (from the $N(0,1)$).

Today and Tomorrow's Lecture

- Regression Anatomy
- Regression Inference
- Non-linearities:
  - Logarithms
  - Others
- Binary Outcomes
- R^2

Non-linearities

Acknowledgments

- Ed Rubin's Graduate Econometrics
- Kyle Raze's Undergraduate Econometrics 1
- MM