Class 4a: Simple Linear Regression

Business Forecasting

Roadmap

This set of classes

- What is a simple linear regression?
- How to estimate it?
- How to test hypothesis in the regression?

Motivation

1. Suppose you are a consultant working for Ecobici
2. Your boss is worried about the impact of global warming on bike use
3. She wants to know: how the bike use will change when the temperature increases by 1 degreee
4. This is exactly what the linear regression will tell us!

Simple linear regression

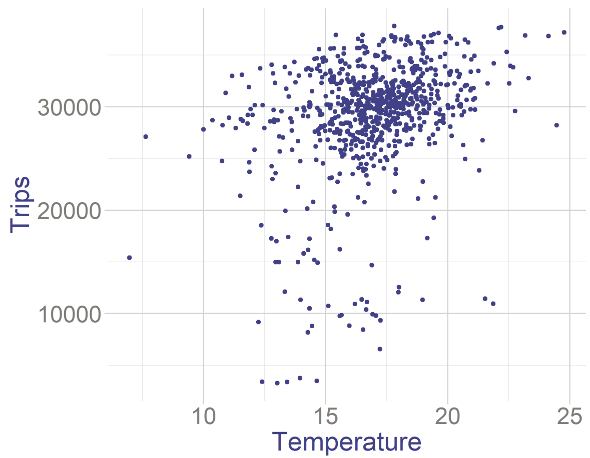
1. Suppose you have paired data: $\{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\}$

Show 7 entries

fecha_retiro	Trips	TMP	PM2.5
2017-01-02	20797	14.49	23.03
2017-01-03	26040	15.22	31.5
2017-01-04	27551	16.89	26.61
2017-01-05	28444	15.99	35.02
2017-01-06	26191	17.85	47.21
2017-01-09	31350	10.91	42.24
2017-01-10	33228	12.85	29.42

Showing 1 to 7 of 781 entries

Previous 1 2 3 4 5 … 112 Next



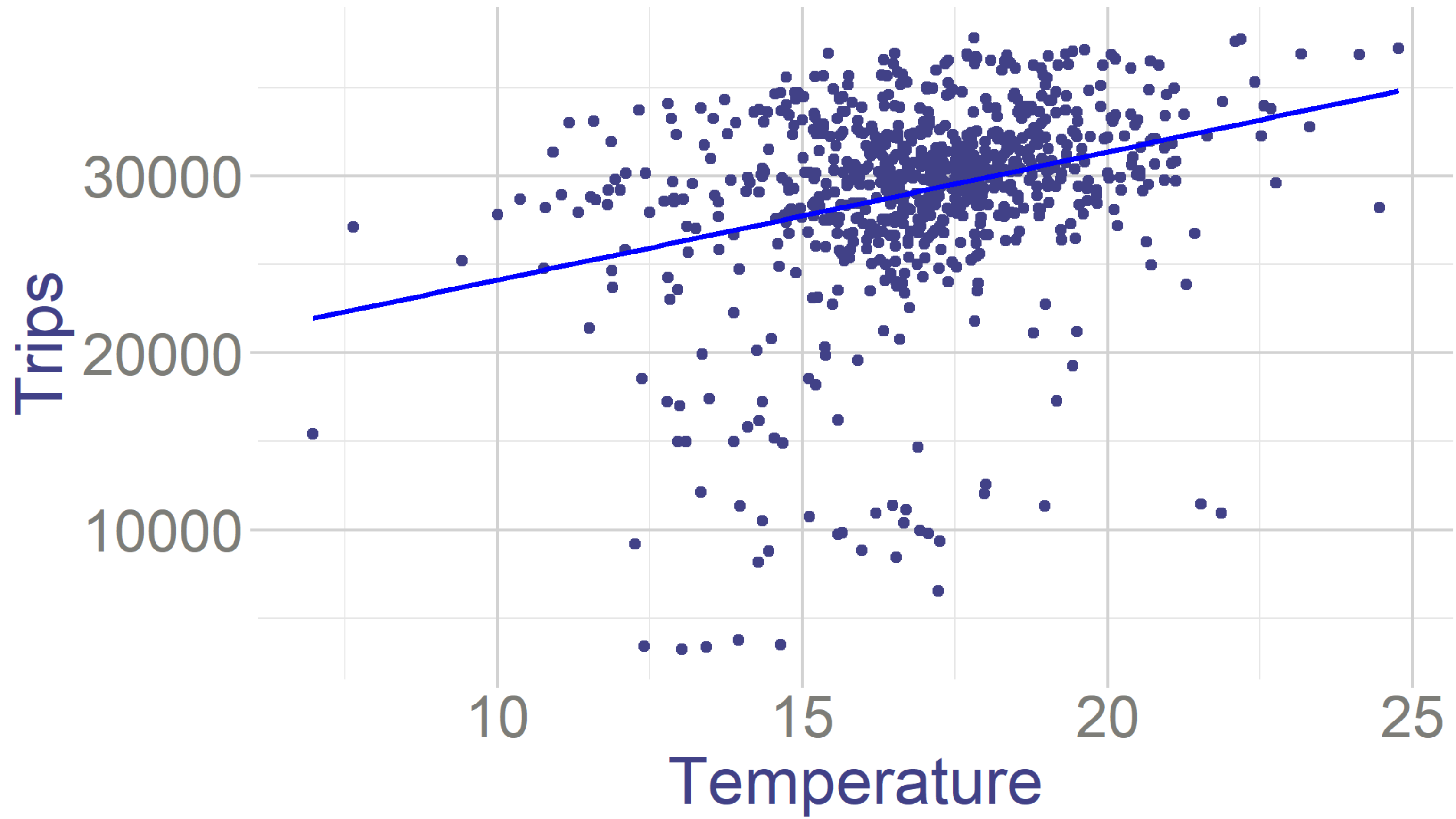
Simple linear regression

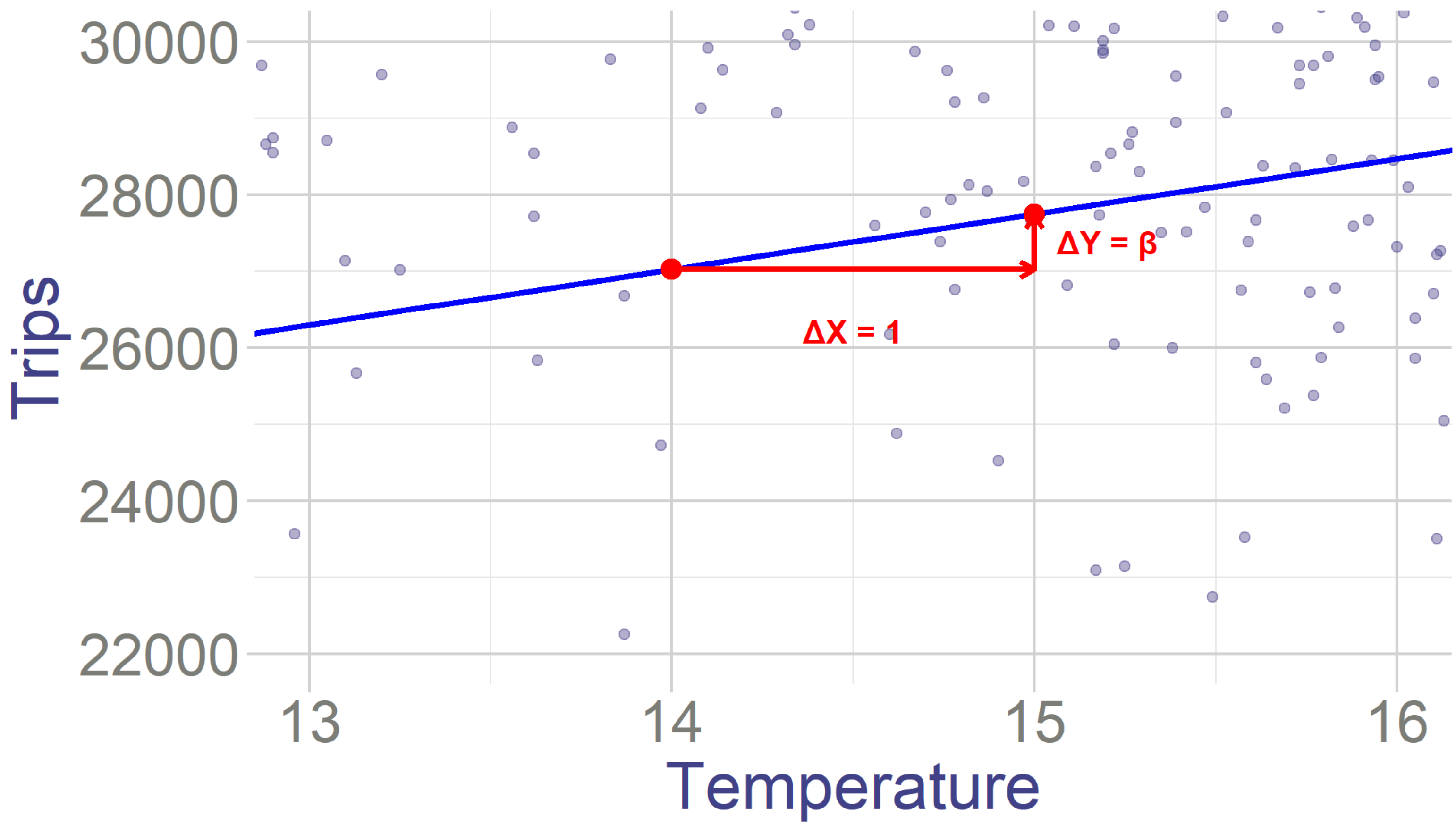
1. In the population, there exists a linear relationship between x_i and y_i of the form:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

Where:

- y_i is a dependent variable
- x_i is a independent variable, or regressor, or predictor
  - (suppose non-random)
- β_0 and β_1 are parameters
- β_1 tells you how y_i changes (on average) when we change x_i by one unit
- β_0 is intercept, where the line cuts y axis
- u_i is a random error term (unknown)


30000
20000
10000
Trips
10
15
20
25
Temperature


30000
28000
26000
24000
22000
Trips
13
14
15
16
Temperature
ΔY = β
ΔX = 1

Returns to Education

Card (Angrist and Krueger, 1991)

Context: How additional years of schooling affect workers' earnings in the labor market?

Finding: Each additional year of schooling increases wages by ~6–10%.

Question: "Suppose we see an estimated effect of $100 increase in monthly wages with each additional year of education. What's the regression equation behind this? What is Y, what is X, and what does β_1 mean in plain business terms?"

Forecasting Demand

Dinerstein et al (AER, 2018)

Context: How sensitive online consumer demand is to small price change?

Finding: Small price changes generate large changes in the demand in online shopping (eBay).

Question: "Suppose that if we increase a price of a keyboard by $1, the demand decreases on average by 200 units. Which variable is Y? Which is X?

Real Estate & Amenities

Glaeser & Kahn (Journal of Transport Geography, 2019)

Context: How proximity to amenities such as schools or subway stations influences housing prices?

Finding: Properties prices increase significantly near subways.

Question: “Suppose you’re told: each kilometer closer to a subway increases house price by $1205. What regression line must be behind that claim? What is Y, what is X, and what does β_1 mean?”

Advertising

Alpert et al. (Journal of Public Economics, 2023)

Context: Does advertising for medication actually increase doctor visits and prescriptions?

Finding: A 10% increase in views of ads for medication increased prescriptions by ~1.7%.

Question: “Suppose we see a finding: one more million views of ads increases monthly medication sales by 500. What regression did they run? What’s Y? What’s X? How would we interpret β_1?”

Gifts to Physicians

Newham & Valente (Journal of Health Economics, 2024),

Context: How payments from pharmaceutical companies to doctors affect prescription drug cost?

Finding: Each dollar of gift/payment to doctors leads to approximately $23 in increased prescription drug costs.

Question: "Here, β_1 from regression indicates that every \$1 in gifts yields an extra \$23 in drug costs. What model structure would get you this? Who might be Y, who is X, and what does β_1 mean for policy?"

Assumptions

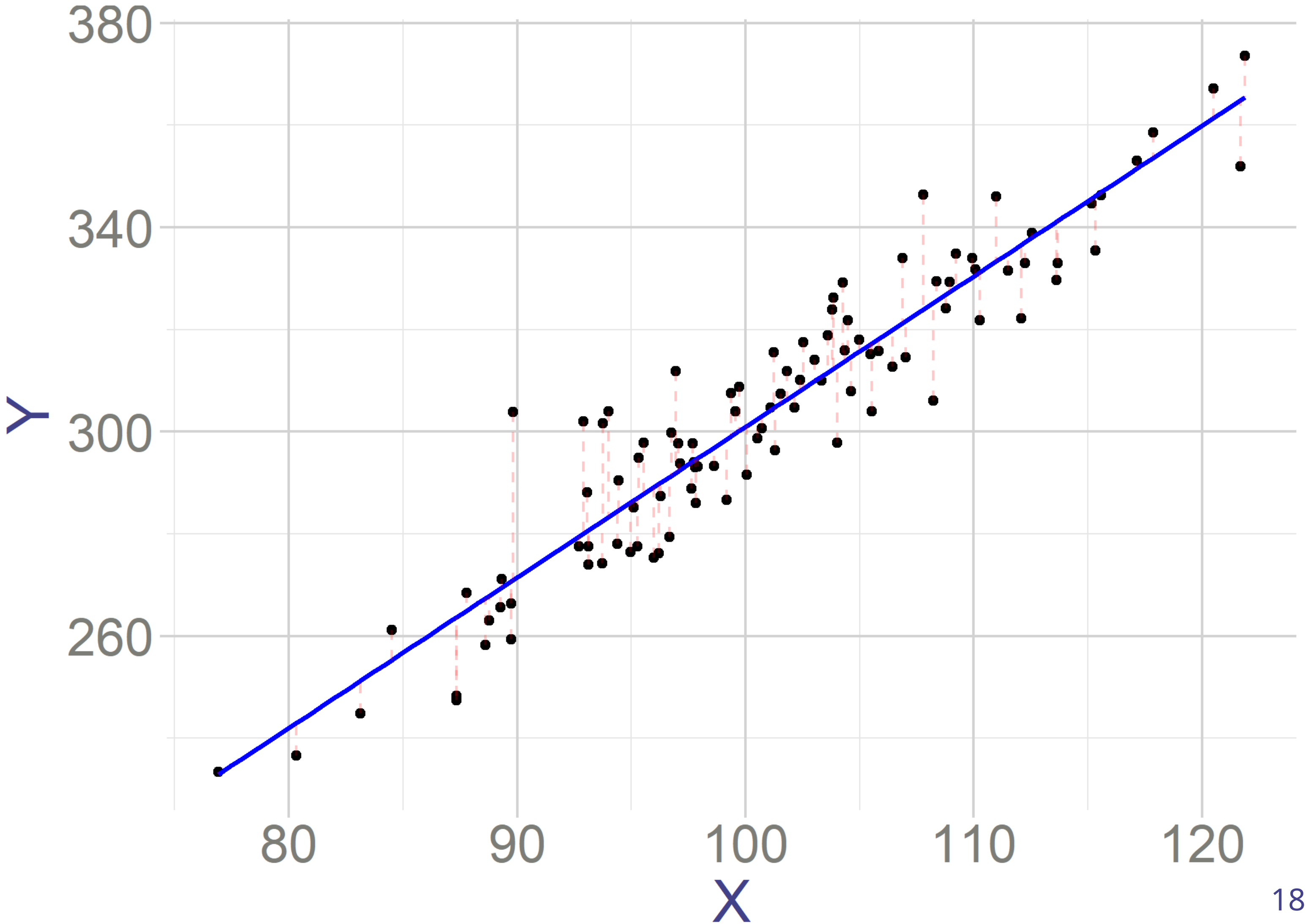
We can estimate β under some assumptions.

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

Here they are:

1. Model is linear in the parameter and with additive error term
2. $E(u_i) = 0 \rightarrow E(y_i|x = x_0) = \beta_0 + \beta_1 x_0$
3. $Var(u_i) = \sigma^2 \rightarrow var(y_i|x = x_0) = \sigma^2$
4. $cov(u_i, u_j) = 0$

General Example



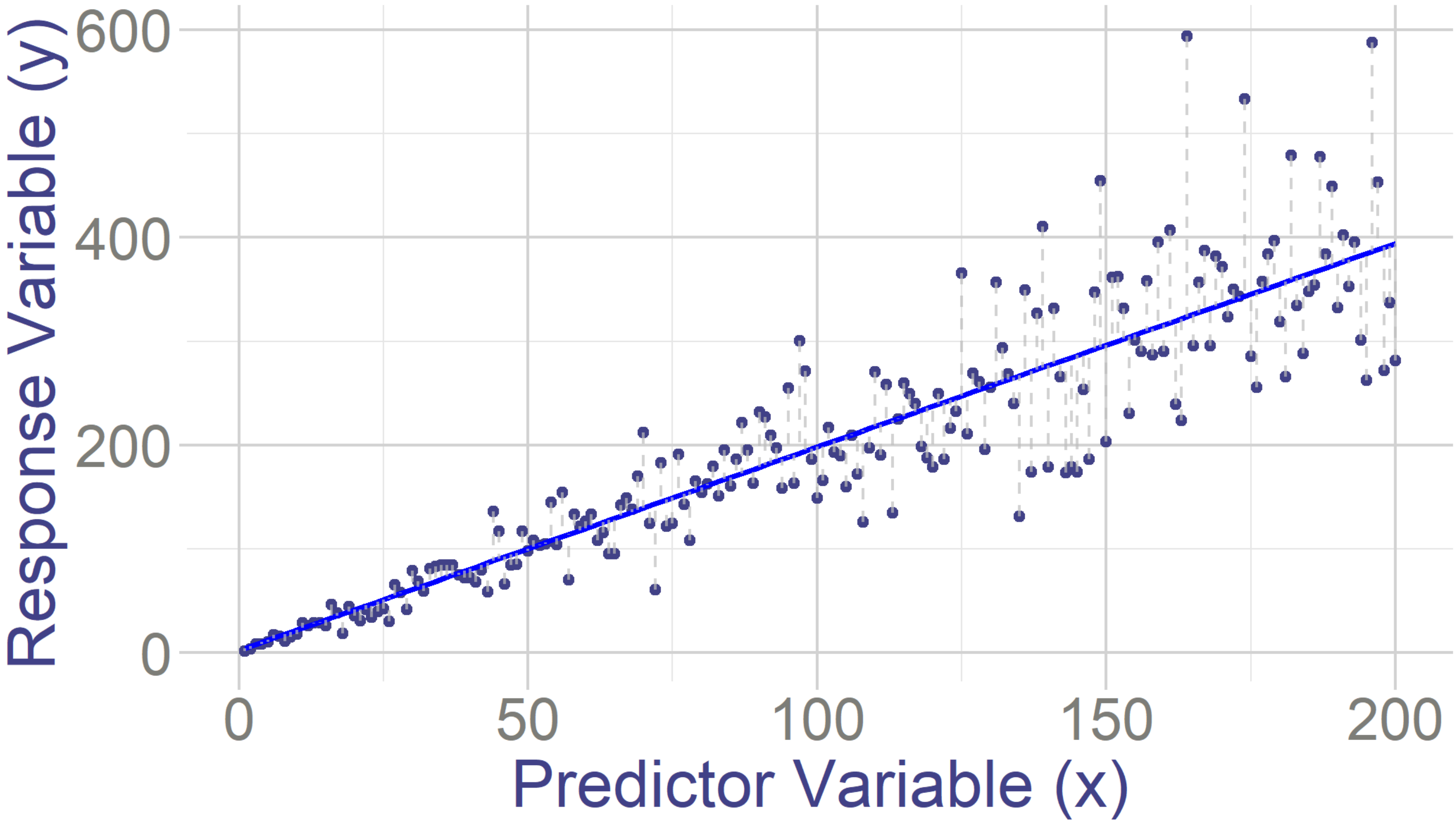
Model is linear in the parameter and with additive error term

- Linear models
  - $y_i = \beta_0 + \beta_1 x_i + e_i$
  - $y_i = \beta_0 + \beta_1 x_i^2 + e_i$
  - $y_i = \beta_0 + \beta_1 log(x)_i + e_i$
  - $y_i = \beta_0 + \beta_1 c^{x_i} + e_i$
- Not linear models
  - $y_i = (\beta_0 + \beta_1 x_i) * e_i$
  - $y_i = \beta_0 + x_i^{\beta_1} + e_i$
  - $y_i = log(\beta_0 + \beta_1 x_i + e_i)$
  - $y_i = \beta_0 + (\beta_1 x_i + e_i)^2$

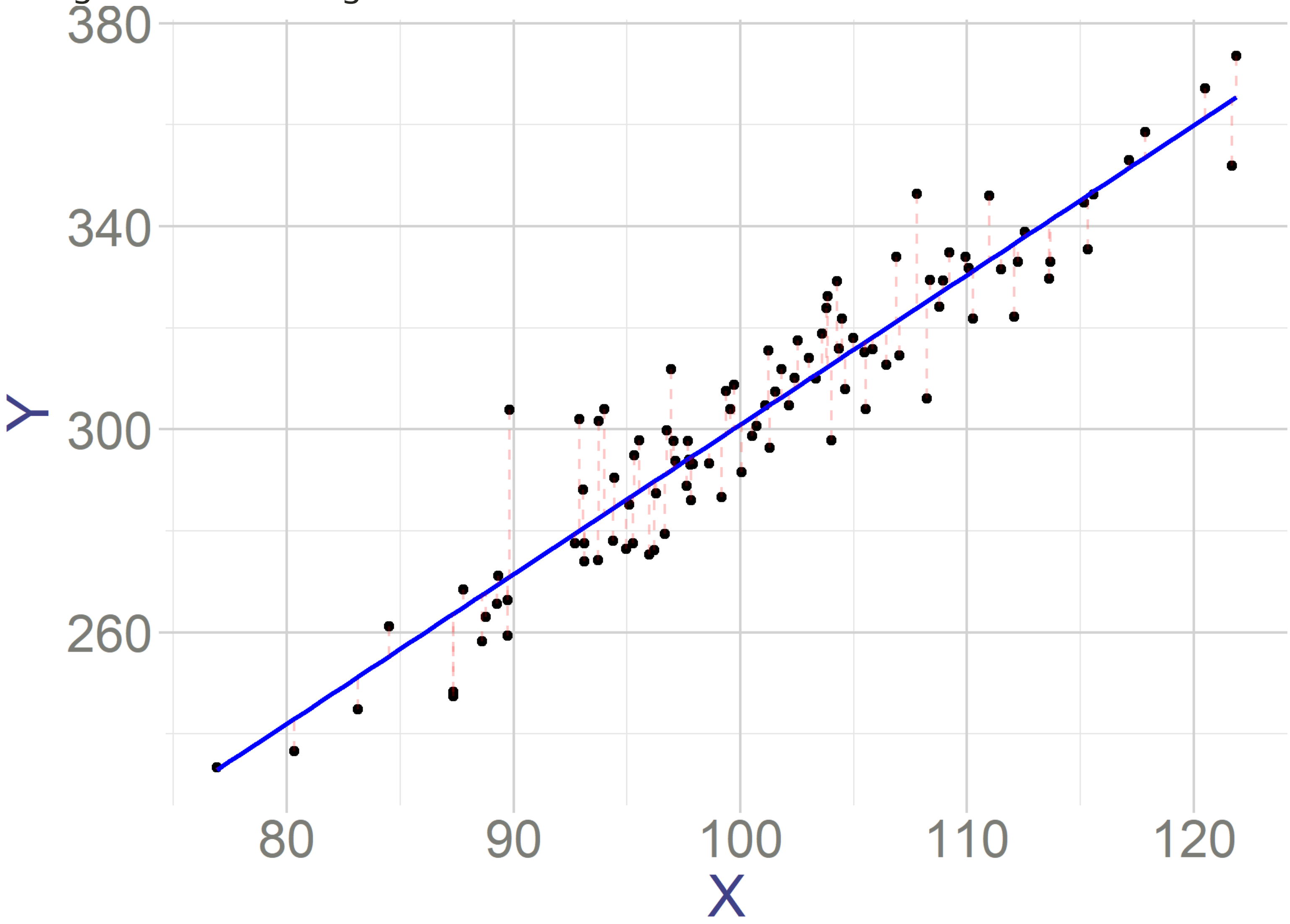
2 is in the app

$$Var(u_i) = \sigma^2$$

What happens if this is not true?



Let's go back to our regression line



Estimation of the parameters

We want to estimate the parameters in this linear relationship based on our **sample**.

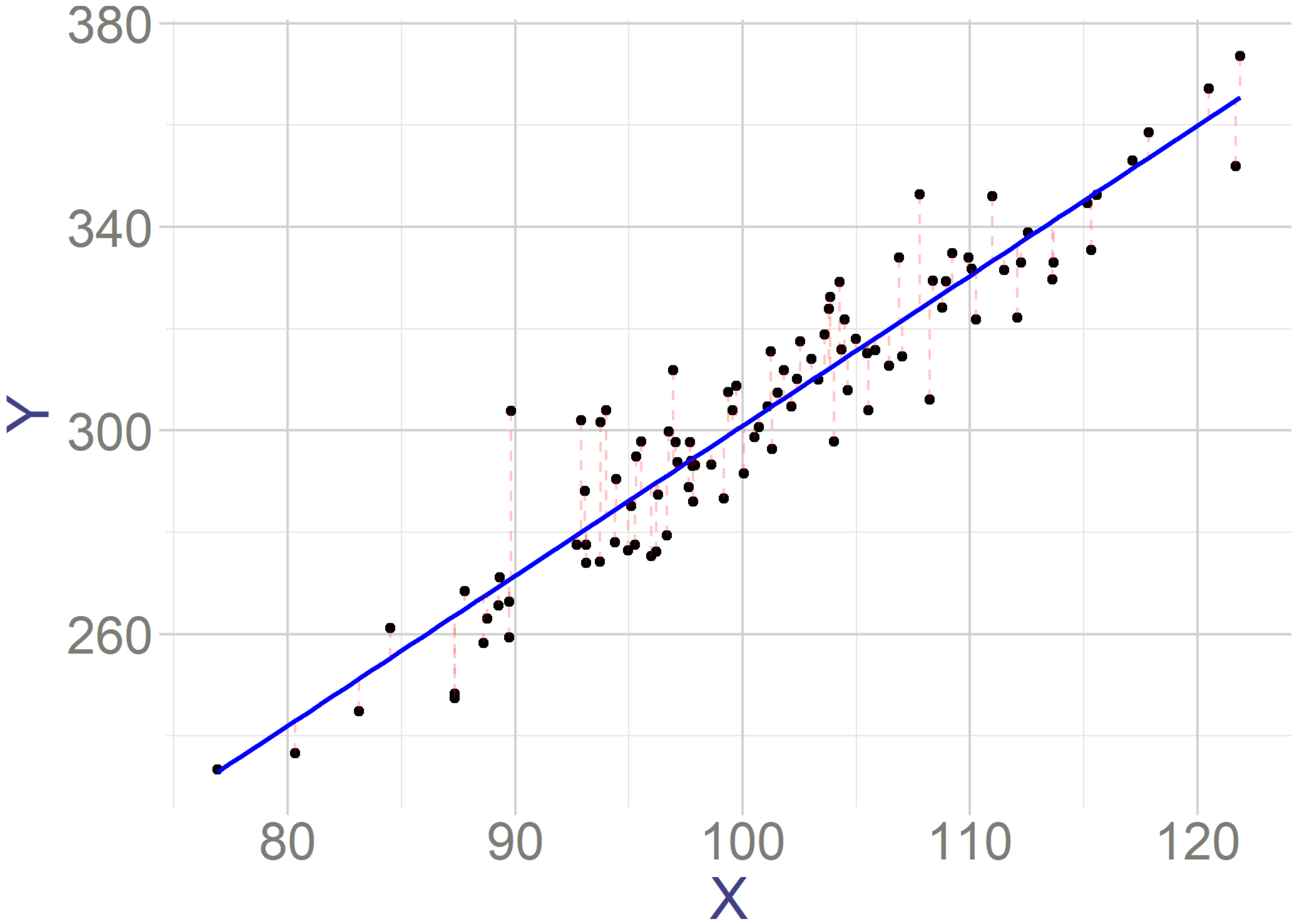
- Once estimated, we can write y_i as

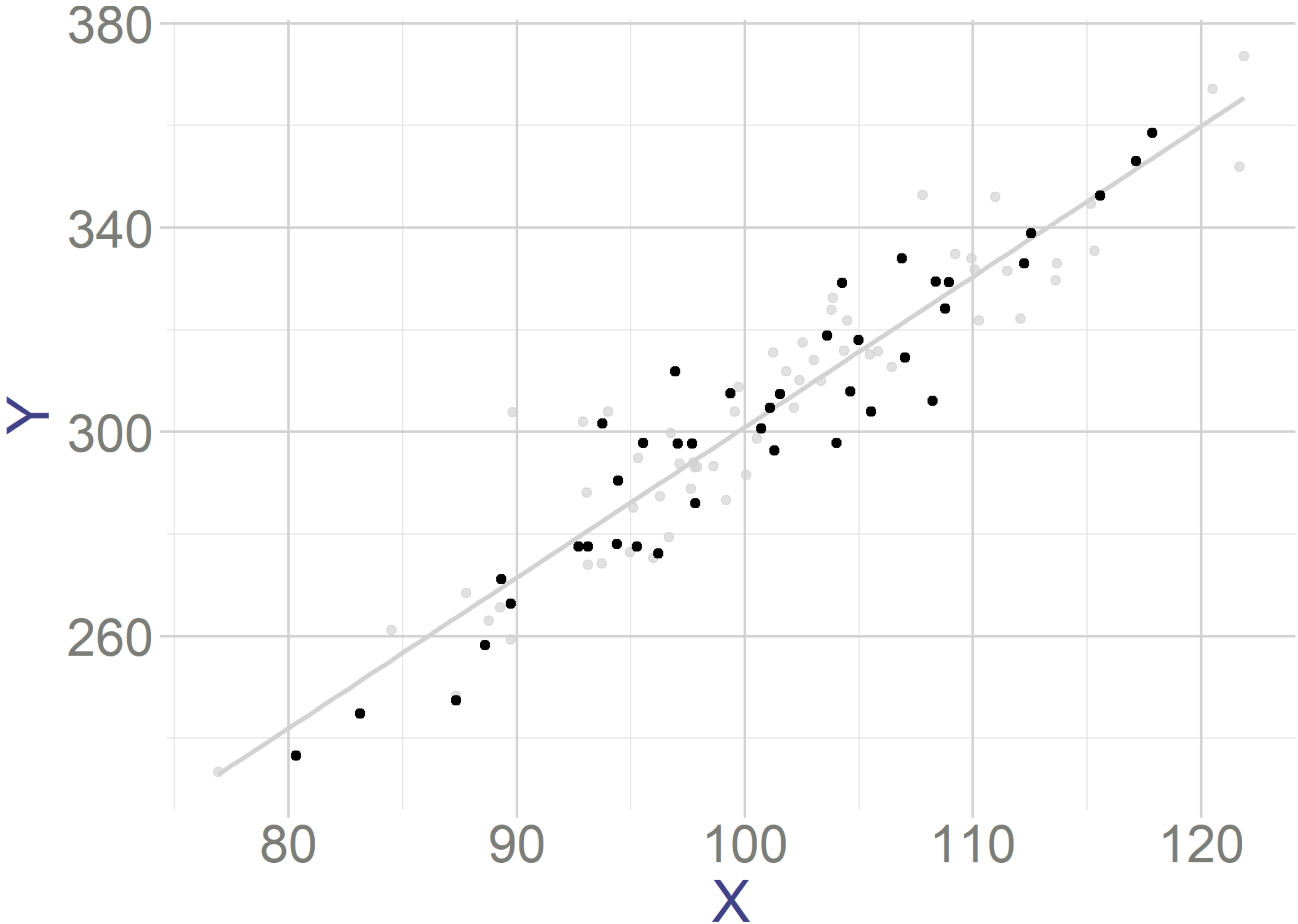
$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$$

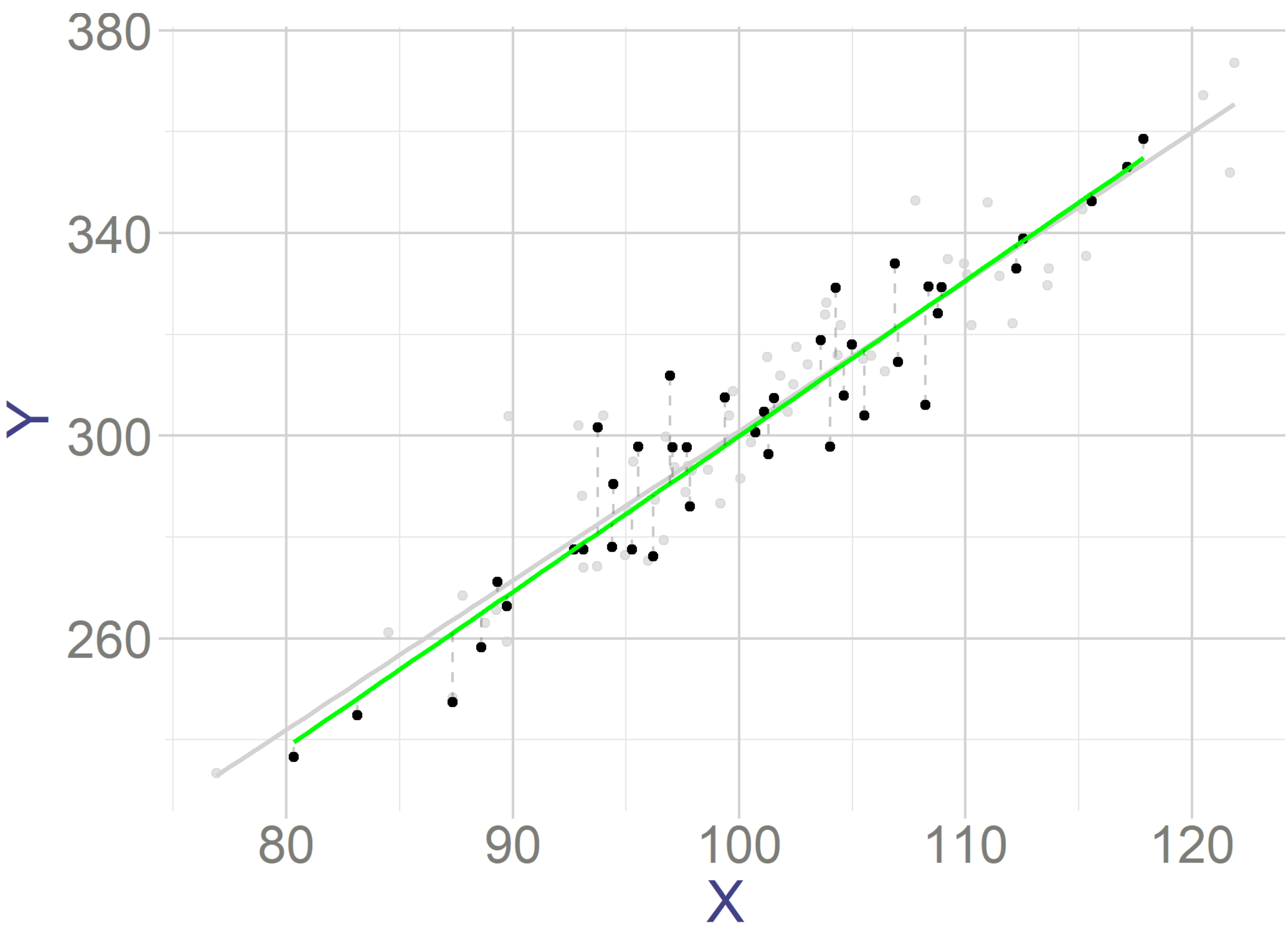
- Residua term (e) here reflects both uncertainly about parameters and the random part present in population model

- We can predict y_i for any x_i using our estimates

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$


380
340
300
260
80
90
100
110
120
X
Y


380
340
300
260
Y
80
90
100
110
120
X



- How do we find $\hat{\beta}_0$ and $\hat{\beta}_1$?

Best fit line

The best fitting line will minimize the sum of squared residuals $SSE = \sum_{i=1}^{n} e_i^2$

$$(\hat{\beta}_0, \hat{\beta}_1) = argmin_{b_0,b_1} SSE = argmin_{b_0,b_1} \sum_{i=1}^{n} e_i^2$$

$$\begin{aligned} SSE &= \sum_{i=1}^{n} e_i^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} \left(y_i - (b_0 + b_1 x_i)\right)^2 \end{aligned}$$

So effectively we are minimizing:

$$(\hat{\beta}_0, \hat{\beta}_1) = argmin_{b_0,b_1} SSE = argmin_{b_0,b_1} \sum_{i}^{n} \left(y_i - (b_0 + b_1 x_i)\right)^2$$

OLS

We called this estimator **OLS** - ordinary least squares

$$(\hat{\beta}_0, \hat{\beta}_1) = argmin_{b_0,b_1} SSE = argmin_{b_0,b_1} \sum_i^n (y_i - (b_0 + b_1 x_i))^2$$

✏️ Sidenote on Derivatives

To solve OLS, we'll need **derivatives**. A key tool is the **chain rule**: if

$$h(x) = f(g(x)),$$

then

$$h'(x) = f'(g(x)) \cdot g'(x).$$

Example (square function):

$$h(x) = (3x+1)^2 = f(g(x)),$$

Then

- $f(u) = u^2 \quad \Rightarrow \quad f'(u) = 2u$
- $g(x) = 3x+1 \quad \Rightarrow \quad g'(x) = 3$

and

$$h'(x) = f'(g(x)) \cdot g'(x) = 2(3x+1) \cdot 3.$$

Best fit line 1

To find the minimum of SSE, we take partial derivatives with respect to β_0 and β_1 and set them equal to zero:

Partial derivative with respect to β_0:

$$\frac{\partial SSE}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} \Big(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \Big)$$

Setting this derivative to zero:

$$-2 \sum_{i=1}^{n} \Big(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \Big) = 0$$

$$\hat{\beta}_0 n + \hat{\beta}_1 \sum x_i = \sum y_i$$

Best fit line 2

Partial derivative with respect to $\hat{\beta}_1$:

$$\frac{\partial SSE}{\partial \hat{\beta}_1} = 2\sum_{i=1}^{n} x_i \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right)$$

Setting this derivative to zero:

$$2\sum_{i=1}^{n} x_i \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right) = 0$$

$$\hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 = \sum x_i y_i$$

Best fit line

Putting it all together:

$$\hat{\beta}_0 n + \hat{\beta}_1 \sum x_i = \sum y_i$$

$$\hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x}{n} = \bar{y} - \hat{\beta}_1 \bar{x}$$

And plugging this here:

$$\hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 = \sum x_i y_i$$

We get:

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\widehat{cov(x_i, y_i)}}{\widehat{var(x_i)}}$$

Or

$$\hat{\beta}_1 = \frac{\widehat{cov(x_i, y_i)}}{\widehat{var(x_i)}} = \frac{\widehat{cov(x_i, y_i)}}{\sqrt{\widehat{var(x_i)}}\sqrt{\widehat{var(x_i)}}} \frac{\sqrt{\widehat{var(y_i)}}}{\sqrt{\widehat{var(y_i)}}} = \widehat{\rho(x, y)} \frac{\sqrt{\widehat{var(y_i)}}}{\sqrt{\widehat{var(x_i)}}}$$

Source: [https://observablehq.com/@yizhe-ang/interactive-visualization-of-linear-regression)

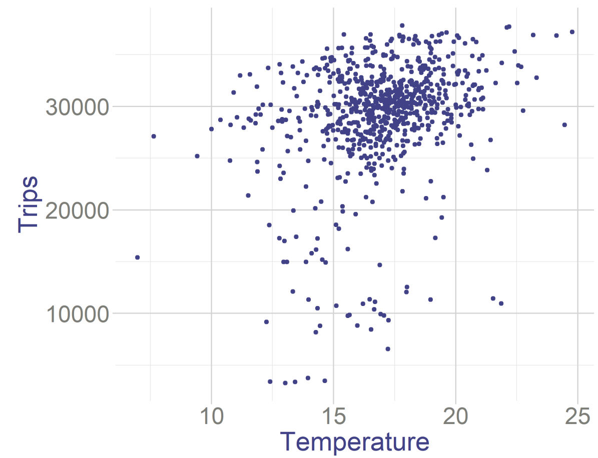
Back to Motivating example

Show 7 entries

fecha_retiro	Trips	TMP	PM2.5
2017-01-02	20797	14.49	23.03
2017-01-03	26040	15.22	31.5
2017-01-04	27551	16.89	26.61
2017-01-05	28444	15.99	35.02
2017-01-06	26191	17.85	47.21
2017-01-09	31350	10.91	42.24
2017-01-10	33228	12.85	29.42

Showing 1 to 7 of 781 entries

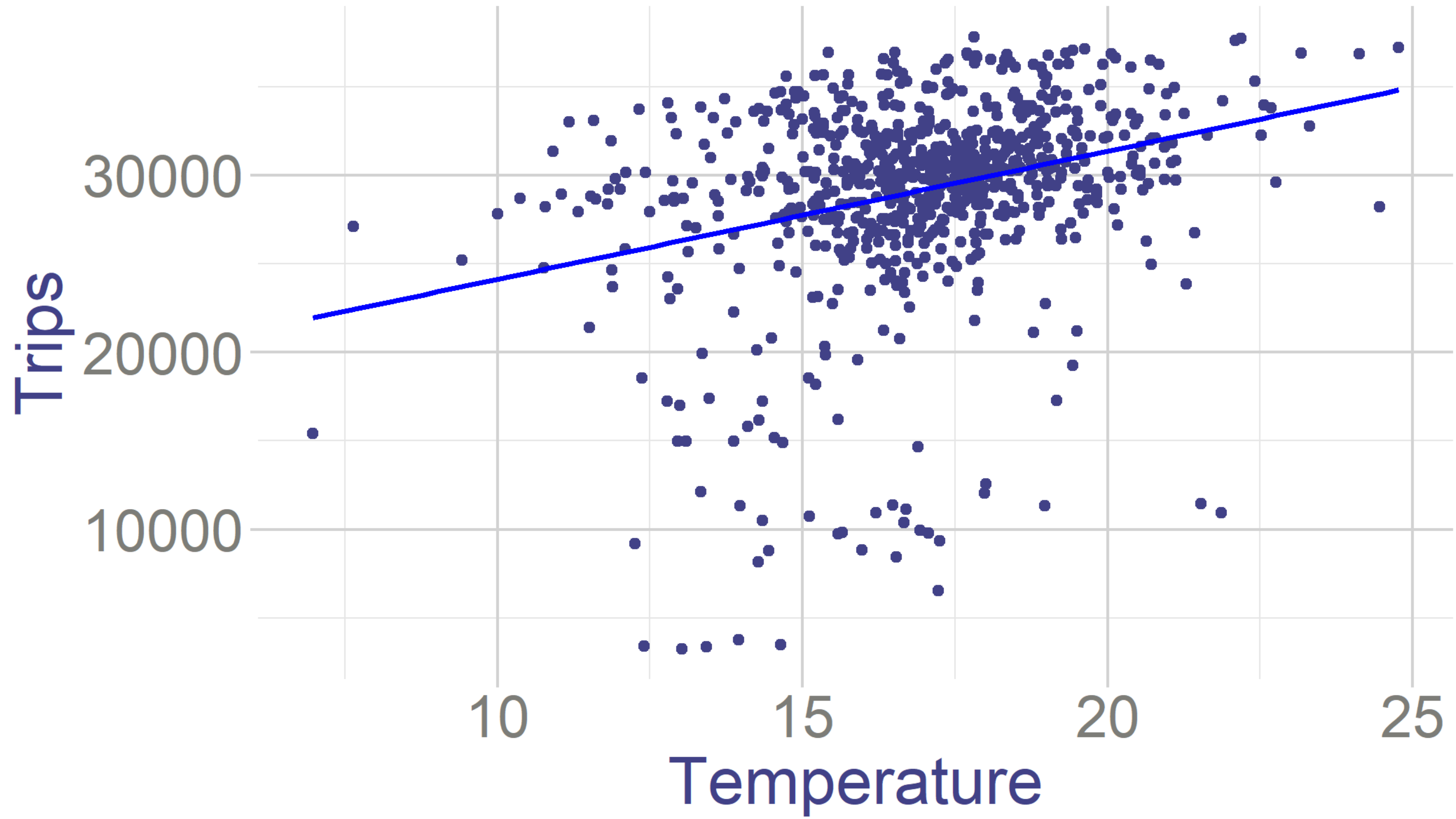
Previous 1 2 3 4 5 … 112 Next



We want to estimate the following relationship:

$$Trips_i = \beta_0 + \beta_1 Temperature_i + u_i$$

Best Fit Line



Regression output in R

```r
# Fit a linear regression model
lm_model <- lm(Trips ~ TMP, data = Data_BP)
# Display the summary of the linear regression model
summary(lm_model)
```

```
##
## Call:
## lm(formula = Trips ~ TMP, data = Data_BP)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24010.5  -1508.4    774.5   2920.5   8900.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16892.66    1427.32  11.835   <2e-16 ***
## TMP           723.55      83.37   8.679   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5302 on 779 degrees of freedom
## Multiple R-squared:  0.08817,    Adjusted R-squared:  0.087
## F-statistic: 75.32 on 1 and 779 DF,  p-value: < 2.2e-16
```

2. [**34 puntos**] You have been hired to analyse the relationship between campaign spending and vote share for the forthcoming presidential elections using a simple linear regression model in the form:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad ; \quad i = 1, ..., 150.$$

where

- y_i represents the share of votes received by the incumbent in the i^{th} election, that is, the candidate who has run for another charge in past elections (could be a mayor or another position that is elected by popular vote). Note that this is operationalised as a proportion of total votes obtained that it may take values between 0 and 1.
- x_i represents the share of total campaign spending by the incumbent in the i^{th} election who has been elected for a political position before. Note that this is operationalised as a proportion of total spending by all candidates and that it may take values between 0 and 1.
- ϵ_i is the i^{th} random error which satisfies Gauss– Markov's assumptions.

You have been provided with some statistics for data from 150 past elections such as

$$\bar{x} = 0.40 \quad ; \quad \bar{y} = 0.50 \quad ; \quad s_X = 0.20 \quad ; \quad s_Y = 0.15 \quad ; \quad r_{XY} = 0.60$$

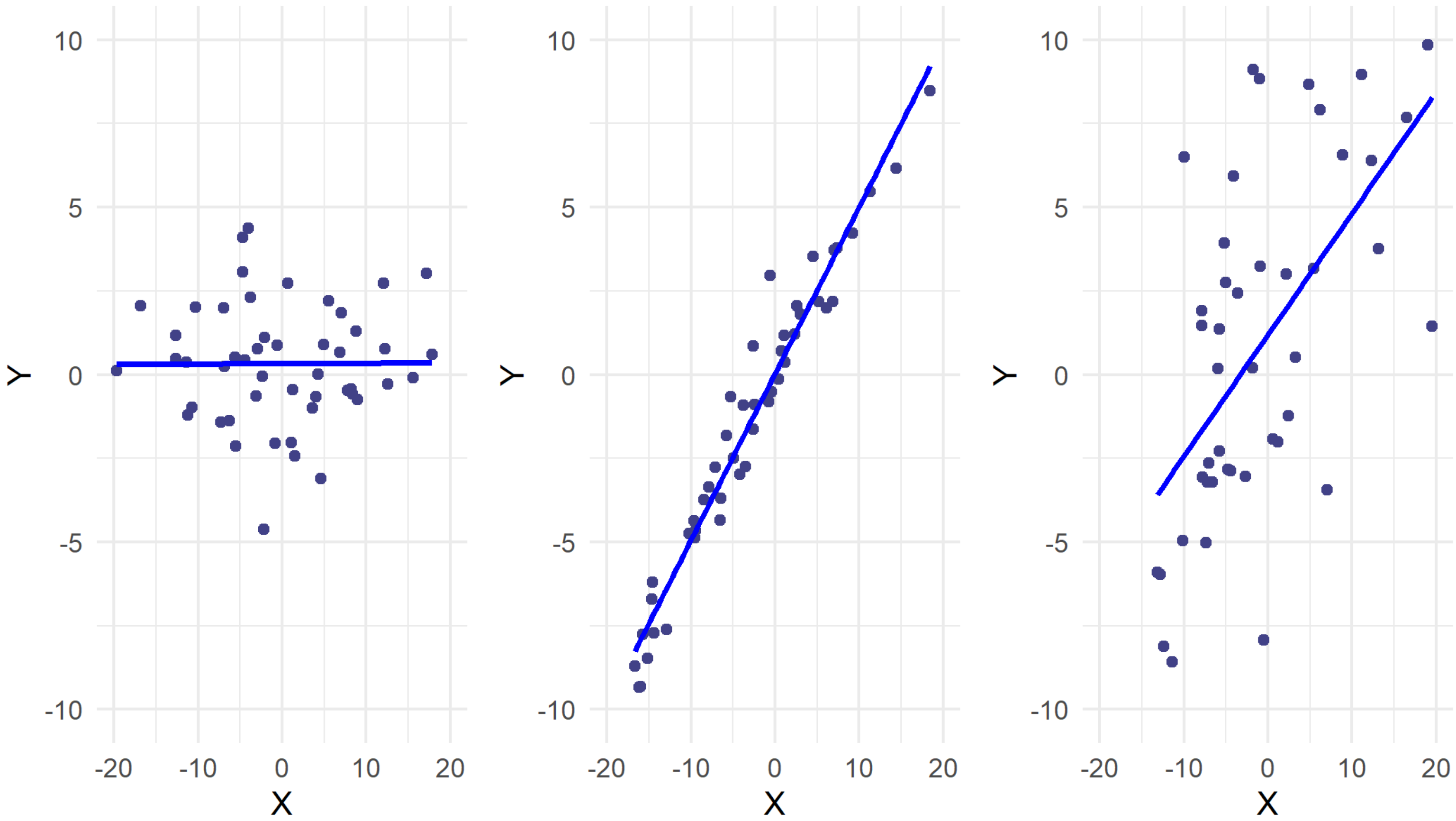
Answer the following questions with the infromation provided:

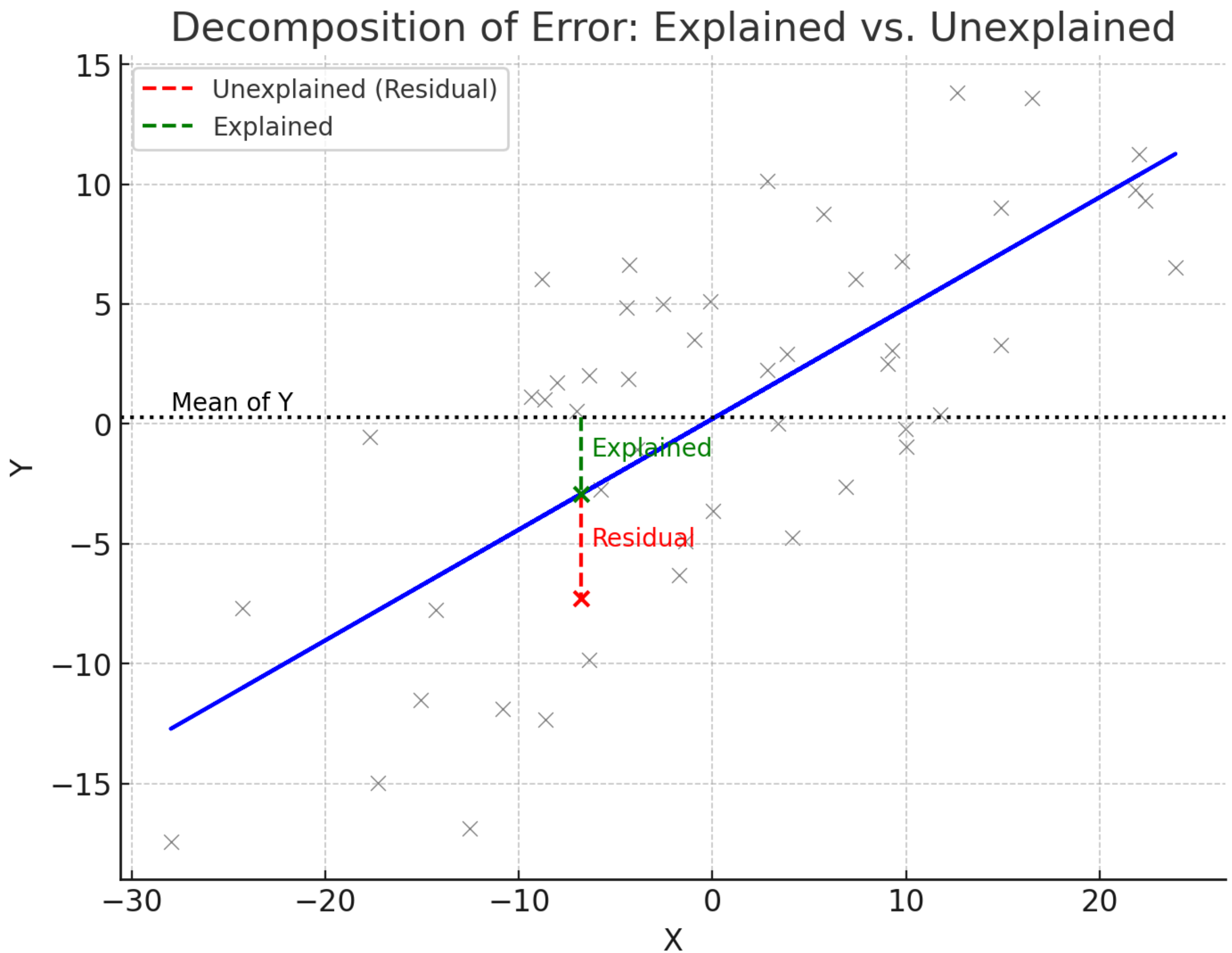
a) [6 puntos] Calculate the estimates for the model's parameters.

b) [6 puntos] Without making any formal inferential process, interpret the coefficients estimated.

c) [5 puntos] Determine how much campaign spending is needed to obtain at least 40 % of the total vote share.

Fit of linear regression




Decomposition of Error: Explained vs. Unexplained
Unexplained (Residual)
Explained
Mean of Y
Explained
Residual
Y
X
15
10
5
0
−5
−10
−15
−30
−20
−10
0
10
20

Measure of fit - R squared

How much we managed to explain with our regression?

- SST= total sum of squares = $S_{yy} = \sum(y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2$
- SSR= regression sum of squares = $\sum(\hat{y}_i - \bar{y})^2 = \sum \hat{y}_i^2 - n\bar{y}^2$

Measure of fit is:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Intuition:

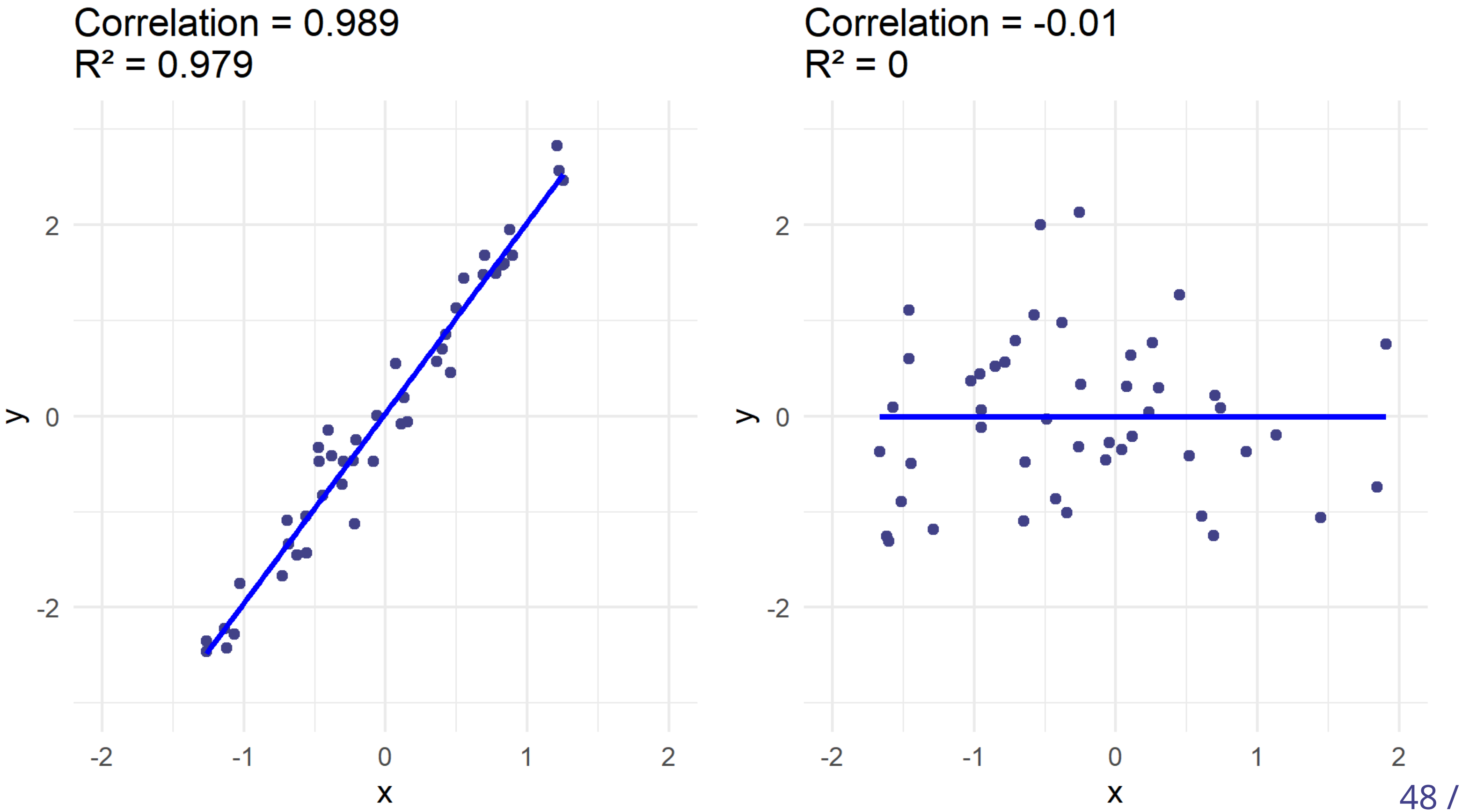
- How much variation in y can we explain with our model
- It is always between 0 and 1
  - In fact $SST = SSR + SSE = \sum(\hat{y}_i - \bar{y})^2 + \sum(\hat{y}_i - y_i)^2$
- SSE/SST is proportion that cannot be explained with the model
- so 1-SSE/SST is the variation that we can explain with the model

Illustration in the app

Measure of fit: R squared

If we have just one regressor, the R^2 is related to correlation between x and y.

$$R^2 = (\rho(x, y))^2$$



How much of bike usage does the temperature explains?

- Total Variation in y: $S_{yy} = \sum(y_i - \bar{y})^2 = 24012556582$
- Explained Variation in y: $SSR = \sum(\hat{y}_i - \bar{y})^2 = 2117129482$
- Unexplained Variation in y: $SSE = \sum \hat{e}^2 = 21895427100$

```
##
## Call:
## lm(formula = Trips ~ TMP, data = Data_BP)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24010.5  -1508.4    774.5   2920.5   8900.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16892.66    1427.32  11.835   <2e-16 ***
## TMP           723.55      83.37   8.679   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5302 on 779 degrees of freedom
## Multiple R-squared:  0.08817,    Adjusted R-squared:  0.087
## F-statistic: 75.32 on 1 and 779 DF,  p-value: < 2.2e-16
```

Scaling of variables:

- You built a linear regression explaining how one more peso spent on training improves the performance of the employee.
- You will present this regression to a client from US, who has no idea what a peso is.
- You need to translate it to dollars

Suppose that we used x and y in our sample to estimate $\hat{\beta}_1$ and $\hat{\beta}_0$.

- Let's say that the scale of x changed. New $z = ax + c$.
  - How do $\hat{\beta}_1$ and $\hat{\beta}_0$ change?
- Let's say that the scale of y changed. New $y' = by + d$.
  - How do $\hat{\beta}_1$ and $\hat{\beta}_0$ change?
- Suppose that $\bar{y} = 0$ and $\bar{x} = 0$. What is $\hat{\beta}_0$?

Scaling of variables:

Effect on slope is easiest derived using the definition with correlation:

$$
\begin{aligned}
\hat{\beta}_1' &= \text{cor}(z, y') \cdot \frac{\text{sd}(y')}{\text{sd}(z)} \\
&= \text{cor}(ax + c, by + d) \cdot \frac{\text{sd}(by + d)}{\text{sd}(ax + c)} \\
&= \text{cor}(x, y) \cdot \frac{b \cdot \text{sd}(y)}{a \cdot \text{sd}(x)} \\
&= \frac{b}{a}\hat{\beta}_1
\end{aligned}
$$

- correlation does not change when we scale variables
- adding constants does not matter for the slope
- multiplication of y or x changes the slope

Scaling of variables:

Effect on the intercept is easiest seen through its formula:

$$\hat{\beta}_0' = \bar{y}' - \hat{\beta}_1' \bar{z} \qquad = (b\bar{y} + d) - \left(\frac{b}{a}\hat{\beta}_1\right)(a\bar{x} + c) \;= b\bar{y} + d - b\hat{\beta}_1\bar{x} - \frac{b}{a}\hat{\beta}_1 c$$

- multiplying y changes the intercept
- adding a constant to y changes the intercept
- adding a constant to x changes the intercept
- multiplying x only changes the intercept if we also add a constant to x

6. [**5 puntos**]A group of experts used data relating weekly spending on food delivery through an *app* (Y) and reported monthly income (X), both measured in dollars, obtaining estimates in a regression:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

with $\hat{\beta}_j$ being least squares estimators for $j = 0, 1$. The analysis revealed that even when the reported income is zero, there was on average positive spending in the app. Additionally, it was found that income had a positive impact on spending in the app. Now, suppose you want to perform the same analysis but with both variables measured in pesos at an exchange rate of \$17.93 pesos per dollar, and you obtain new least squares estimations $\hat{\beta}_0^\star$ and $\hat{\beta}_1^\star$. Then, it is true that:

a) $\hat{\beta}_1^\star > \hat{\beta}_1$; *b*) $\hat{\beta}_1^\star < \hat{\beta}_1$; *c*) $\hat{\beta}_0^\star \geq \hat{\beta}_0$; *d*) $\hat{\beta}_0^\star < \hat{\beta}_0$

Regression through the origin (HOMEWORK)

Suppose the following model:

$$y_i = \beta_1 x_i + u_i$$

- What is the least square estimator for β_1?
- What happens if we use this estimator when it's not going through the origin?

Regression with a Categorical Variable

- Very often in data, we work with **binary (dummy) variables**.
- A **binary variable** takes the value:
  - 1 if the condition is true,
  - 0 otherwise.

Example 1:

$$x_i = \begin{cases} 1 & \text{if individual(i)is female} \\ 0 & \text{if individual(i)is male} \end{cases}$$

Example 2:

$$x_i = \begin{cases} 1 & \text{if transaction(i)is fraudulent} \\ 0 & \text{if transaction(i)is not fraudulent} \end{cases}$$

Example 3:

$$x_i = \begin{cases} 1 & \text{if client(i)made a purchase} \\ 0 & \text{if client(i)didn't make a purchase} \end{cases}$$

Regression with a categorical variable

- Suppose we regress y_i on a dummy x_i:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

- The OLS estimates have a simple interpretation:
  - $\hat{\beta}_0 = \bar{y}_{x_i=0}$: the **mean of y** for the group with $x = 0$
  - $\hat{\beta}_1 = \bar{y}_{x_i=1} - \bar{y}_{x_i=0}$: the **difference in group means** (change in y when x changes by 1)

Example:

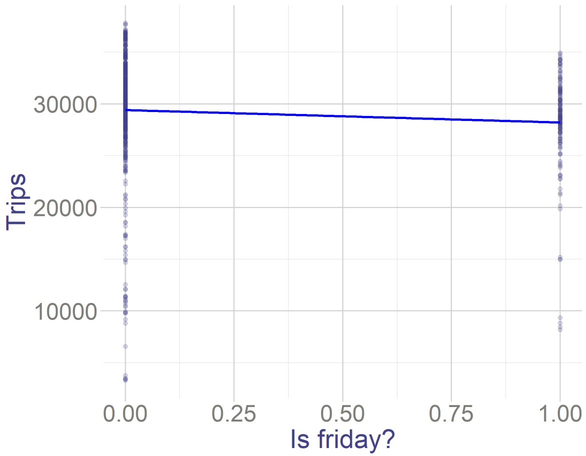
- Let $x_i = 1$ if female, 0 if male
- Then:
- $\hat{\beta}_0 = \bar{y}_{x_i=0} \quad \text{(mean outcome for males)}$
- $\hat{\beta}_1 = \bar{y}_{x_i=1} - \bar{y}_{x_i=0} \quad \text{(difference in means)}$

Show 7 entries

fecha_retiro	day_of_week	is_friday	Trips
2017-01-02	Mon	0	20797
2017-01-03	Tue	0	26040
2017-01-04	Wed	0	27551
2017-01-05	Thu	0	28444
2017-01-06	Fri	1	26191
2017-01-09	Mon	0	31350
2017-01-10	Tue	0	33228

Showing 1 to 7 of 781 entries

Previous 1 2 3 4 5 … 112 Next



- By how much trips change when I move from 0 (Not-friday) to 1 (Friday)?
- x changes by 1, y changes by β

```
##
## Call:
## lm(formula = Trips ~ is_friday, data = Data_BP)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26152.2  -1287.2    868.8   3016.8   8397.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  29410.2      221.2 132.929   <2e-16 ***
## is_friday    -1200.6      495.0  -2.425   0.0155 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5531 on 779 degrees of freedom
## Multiple R-squared:  0.007495,    Adjusted R-squared:  0.00622
## F-statistic: 5.882 on 1 and 779 DF,  p-value: 0.01552
```

Regression with a Categorical Outcome

- Suppose instead that y_i takes only two values: 0 or 1.

Example:
Suppose you work at Amazon and want to predict if a customer will return a product based on its rating.

$$y_i = \begin{cases} 1 & \text{if customer returns product } i \\ 0 & \text{if customer keeps product } i \end{cases}$$

We run the regression:

$$y_i = \beta_0 + \beta_1 x_i + u_i,$$

where x_i is the product rating (between 0 and 5).

Example Data

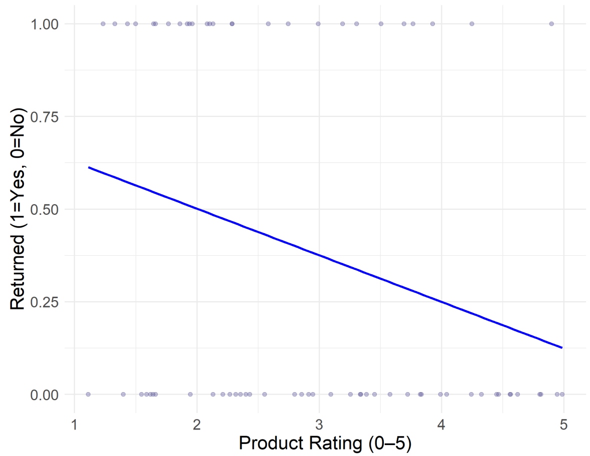
HEre provide an equivalent example of data like this, but for a made up data set on returns and ratings

Show 7 entries

rating	returned
4.944	0
1.548	0
4.621	0
3.305	1
2.582	1
2.799	0
3.826	0

Showing 1 to 7 of 70 entries

Previous 1 2 3 4 5 … 10 Next



Interpretation

- In a regression, we predict the **expected value** of y for a given x:

$$E(y|x) = \beta_0 + \beta_1 x$$

- The expected value of a variable is just its **mean**.
  - For a binary variable, the mean is simply the proportion of 1s.
- Therefore, $E(y|x)$ is the **proportion of returns** among products with rating x.
- Hence, the regression prediction is the **probability of return**.
- If $\hat{y} = 0.15$ for a 5-star product, we interpret it as a 30% chance of being returned
- The slope β_1 tells us how the probability changes when rating increases by one point:
  - If the probability of return is 27% at 4 stars and 30% at 5 stars, then

$$\beta_1 = 0.15 - 0.27 = -0.12$$

meaning each additional point reduces the probability of return by 12p.p

```
##
## Call:
## lm(formula = returned ~ rating, data = amazon_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6130 -0.3918 -0.1921  0.4904  0.8633
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.75285    0.15932   4.725  1.2e-05 ***
## rating      -0.12578    0.05103  -2.465   0.0162 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4732 on 68 degrees of freedom
## Multiple R-squared:  0.08202,    Adjusted R-squared:  0.06852
## F-statistic: 6.076 on 1 and 68 DF,  p-value: 0.01624
```

- What is the probability of return when rating is 3?

Interpretation

- β_1 describes change in probability of $y = 1$ when x changes by 1

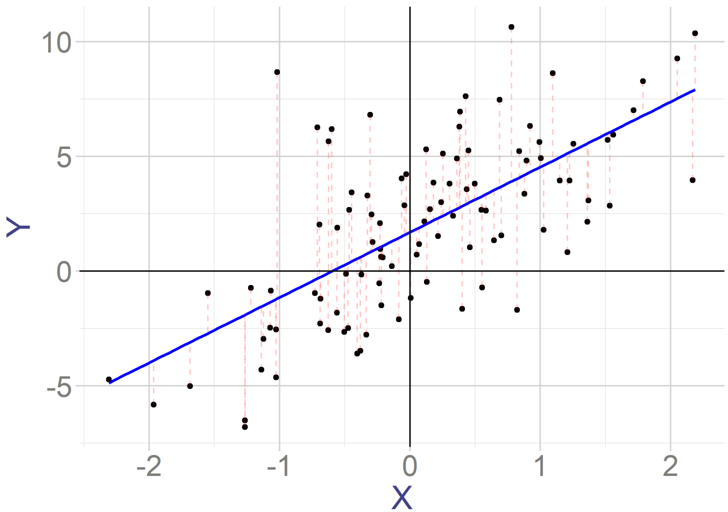
Limitations

- OLS can predict values **outside [0,1]**, which doesn't make sense for probabilities.
- That's why in practice we often move to **Logit/Probit models** — non linear models
- But OLS is still a useful for simplicity and interpretation.

Statistical Properties of OLS

Uncertainty in the Estimate

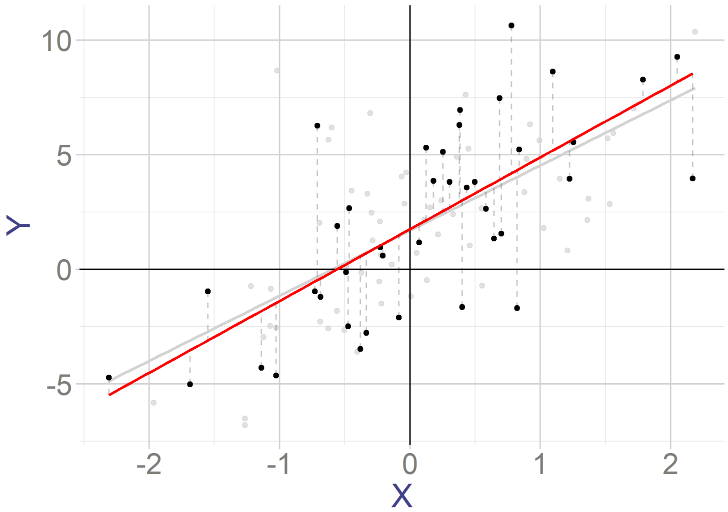
We only have samples, and yet we want to learn something about the population parameters



Population Regression

$$y_i = 6.22 + 2.95x_i + u_i$$

Uncertainty in the Estimate



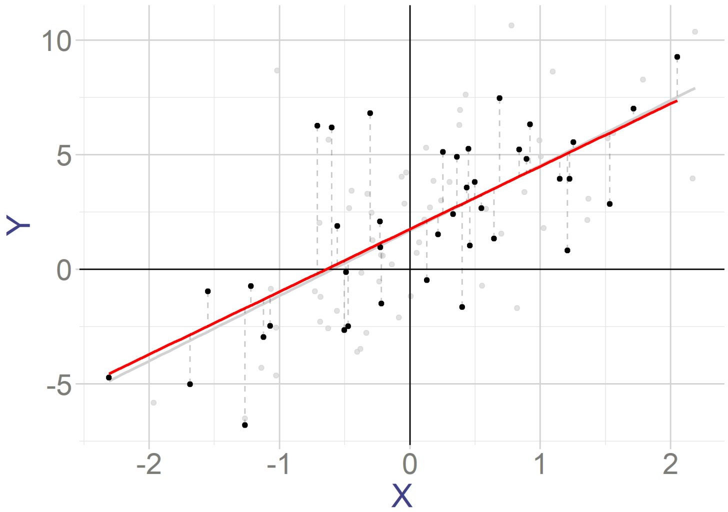
Population Regression

$$y_i = 6.22 + 2.95x_i + u_i$$

Sample Estimate

$$\hat{y}_i = 1.75 + 3.13x_i$$

Uncertainty in the Estimate



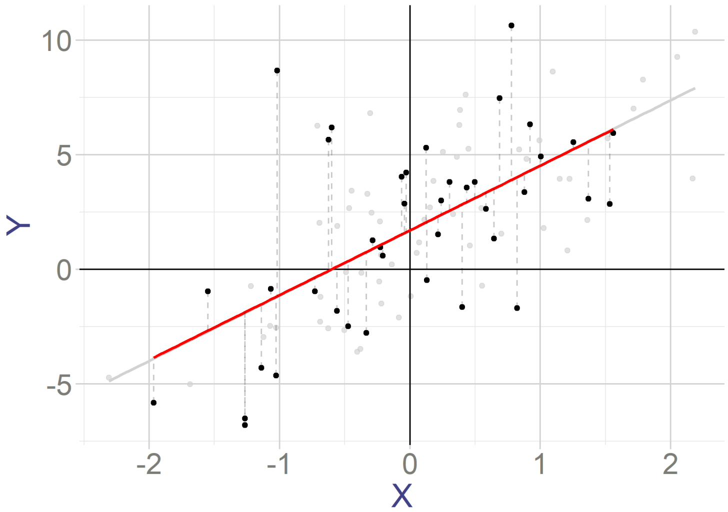
Population Regression

$$y_i = 6.22 + 2.95x_i + u_i$$

Sample Estimate

$$\hat{y}_i = 1.76 + 2.73x_i$$

Uncertainty in the Estimate



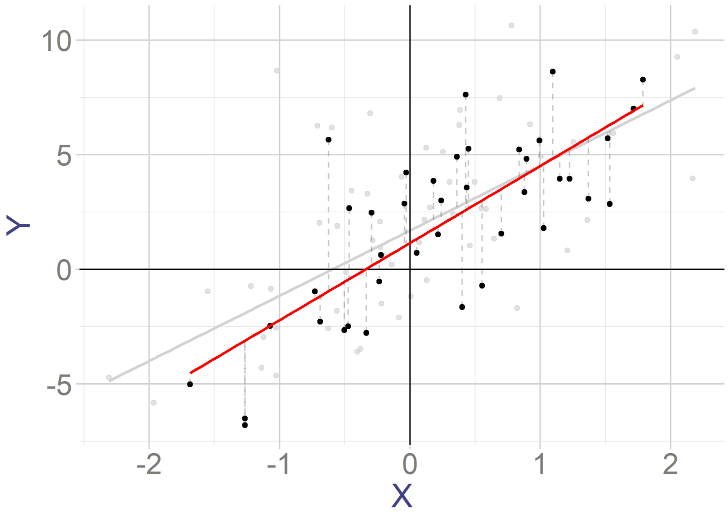
Population Regression

$$y_i = 6.22 + 2.95x_i + u_i$$

Sample Estimate

$$\hat{y}_i = 1.7 + 2.82x_i$$

Uncertainty in the Estimate



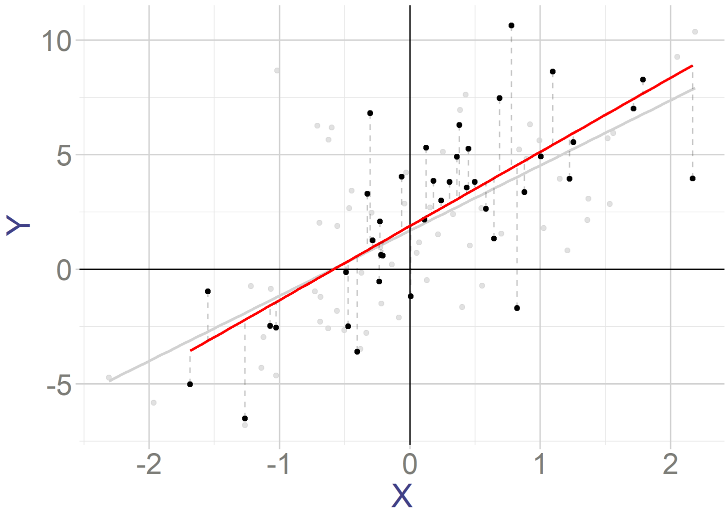
Population Regression

$$y_i = 6.22 + 2.95x_i + u_i$$

Sample Estimate

$$\hat{y}_i = 1.15 + 3.36x_i$$

Uncertainty in the Estimate



Population Regression

$$y_i = 6.22 + 2.95x_i + u_i$$

Sample Estimate

$$\hat{y}_i = 1.89 + 3.23x_i$$

Uncertainty in the Estimate

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimators
- And they are random variables
  - Because their values depend on the random samples
- Are they good estimators?
  - Are they unbiased?
  - Do they have small variance?

Uncertainty in the Estimate

Under these assumptions:

1. Relationship is linear in parameters with linear disturbance
2. $E(u_i) = 0$
3. $Var(u_i) = \sigma^2$
4. $cov(u_i, u_j) = 0$

- OLS is unbiased

$$E(\hat{\beta}_1) = E\left(\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}\right) = \beta_1 \qquad and \qquad E(\hat{\beta}_0) = \beta_0$$

- Assumption 1 is enough for being unbiased $E(u_i) = 0$

Uncertainty in the Estimate

- What is the variance of $\hat{\beta}_1$ and $\hat{\beta}_0$?

$$
\begin{aligned}
\mathrm{Var}(\hat{\beta}_1) &= \mathrm{Var}\left(\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}\right) \\
&= \mathrm{Var}\left(\sum_i \frac{(x_i - \bar{x}) y_i}{\sum_i (x_i - \bar{x})^2}\right) = \sum_i \left(\frac{(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}\right)^2 \mathrm{Var}(y_i) \\
&= \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} = \frac{\sigma^2}{S_{xx}}
\end{aligned}
$$

Because x_i don't change: $var(y_i) = var(\beta_0 + \beta_1 x_i + u_i) = var(u_i) = \sigma^2$

$$
\begin{aligned}
\mathrm{Var}(\hat{\beta}_0) &= \mathrm{Var}(\bar{y} - \hat{\beta}_1 \bar{x}) = \mathrm{Var}(\bar{y}) + \bar{x}^2 \mathrm{Var}(\hat{\beta}_1) - 2\bar{x}\underbrace{cov(\bar{y}, \hat{\beta}_1)}_{0} \\
&= \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{S_{xx}} = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)
\end{aligned}
$$

Standard error is standard deviation of the estimator: $SE(\hat{\beta}) = \sqrt{Var(\hat{\beta})}$

Uncertainty in the Estimate

- How to estimate the σ^2?

$$\hat{\sigma}^2 = \frac{\sum_i e_i^2}{n-2}$$

- Is unbiased for σ^2:

$$E(\hat{\sigma}^2) = E\left(\frac{\sum_i e_i^2}{n-2}\right) = \sigma^2$$

Regression Output

```
##
## Call:
## lm(formula = Trips ~ TMP, data = Data_BP)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24010.5  -1508.4    774.5   2920.5   8900.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16892.66    1427.32  11.835   <2e-16 ***
## TMP           723.55      83.37   8.679   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5302 on 779 degrees of freedom
## Multiple R-squared:  0.08817,    Adjusted R-squared:  0.087
## F-statistic: 75.32 on 1 and 779 DF,  p-value: < 2.2e-16
```

Problem:

Suppose that instead of measuring TMP in celcius, we measure it in $Farenheits$
Practically: $F = 1.8C + 32$

- How would β_1 and $SE(\hat{\beta}_1)$ change?

Gauss Markov Theorem

Under assumptions 1-4, among all linear and unbiased estimators, OLS has the smallest variance.

$$var(\hat{\beta}_1) \leq var(\hat{\beta}_1^{'}) \qquad and \qquad var(\hat{\beta}_0) \leq var(\hat{\beta}_0^{'})$$

Where $\hat{\beta}_1^{'}$ $\hat{\beta}_0^{'}$ are any linear and unbiased estimators of β_1 and β_0 respectively.

It's BLUE - Best, Linear, Unbiased Estimator

Linear estimator basically means it's a weighted sum of y_is:

$$\hat{\beta}_1^{'} = \sum_i c_i y_i$$

where c_i are some weights, usually function of x_i

In OLS:

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{\sum_i (x_i - \bar{x}) y_i}{\sum_i (x_i - \bar{x})^2} \qquad so \qquad c_i^{OLS} = \frac{(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

UPDATE on Gauss Markov

- Science is in progress
- A new paper in 2022 by Hansen shows linearity is not needed
- OLS, under our assumptions, is BUE (Best Unbiased Estimator)

Question 6 [5 points]:

Consider the linear model of the form:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

with $E[\epsilon_i] = 0$; $var(\epsilon_i) = \sigma_i^2 \neq \sigma^2$; $cov(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$, and the estimation of the model by Least Squares. Now consider the following statements:

A: The Least Squares estimators will no longer be unbiased.

B: The Least Squares estimators will no longer have minimum variance.

Then:

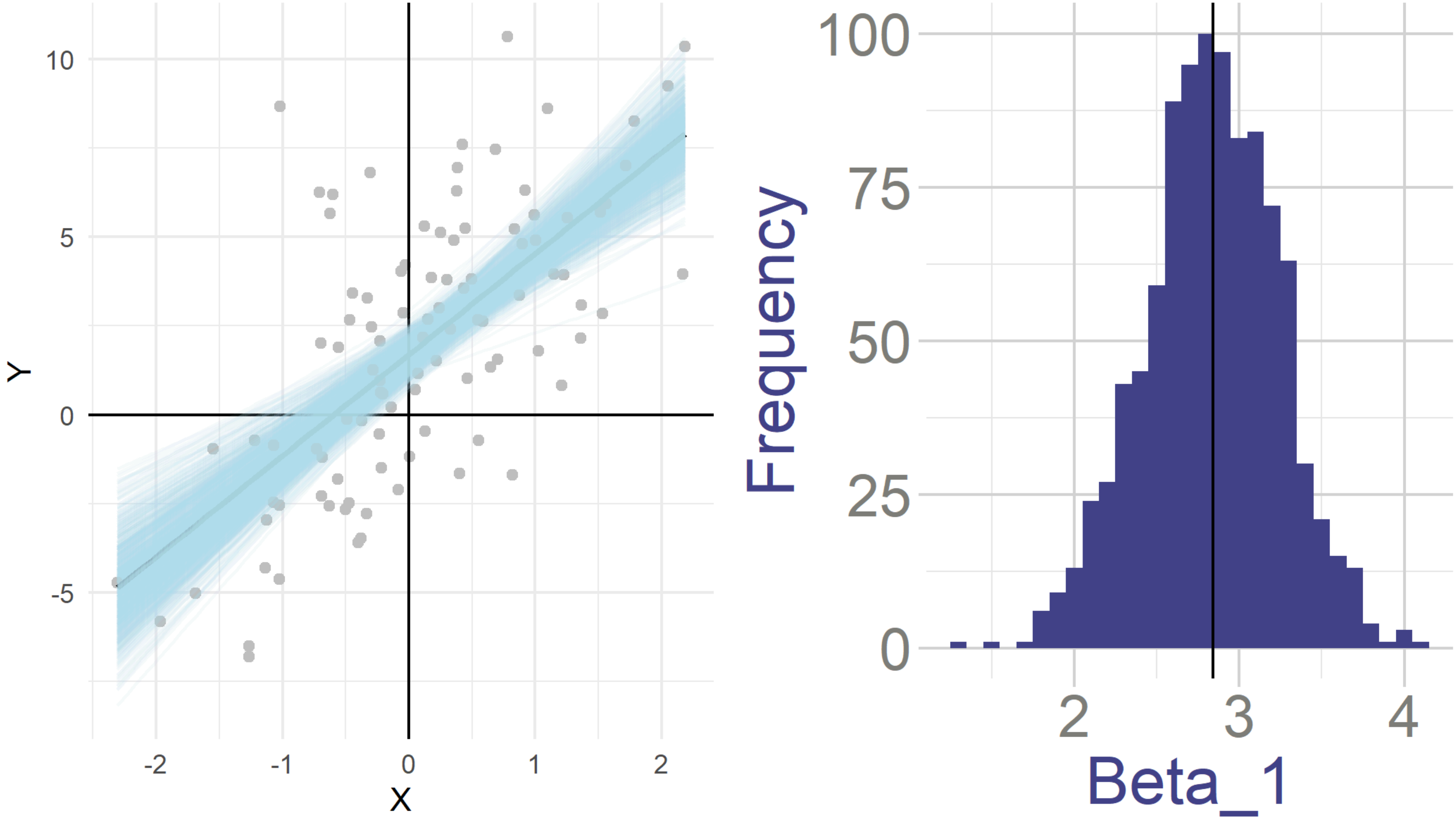
Inference

- Until now, we haven't made any assumptions about the **distributions** of the underlying data or β
    - We don't need it for calculating coefficients $\hat{\beta}_0$ or $\hat{\beta}_1$
    - We don't need it for making predictions $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
    - We don't need it to calculate variance or expectation of coefficients
    - We don't need it for Gauss-Markov Theorem
- However, to make **inference** (confidence intervals, hypothesis testing), we need to know something about distribution of $\hat{\beta}$
    - We will need to assume that population errors are normally distributed: $u_i \sim N(0, \sigma)$
    - For some results this can be relaxed with large samples (CLT)
    - y_i or x_i does not need to be normally distributed
    - But if $u_i \sim N(0, \sigma)$, then conditional on x_i: $y_i | x_i \sim N(\beta_0 + \beta_1 x_i, \sigma)$

Suppose I take 1000 samples of size 40 from the population where $u_i \sim N(0, 2)$:

$$y_i = 6.22 + 2.95x_i + u_i$$

And I estimate the β_1 and β_0 for each sample.



Distributions

Given that

- $u_i \sim N(0, \sigma)$
- linear combination of normal variables is normal

We can derive the following distributions:

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma}{\sqrt{S_{xx}}}\right) \qquad and \qquad \hat{\beta}_0 \sim N\left(\beta_0, \sigma\sqrt{(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}\right)$$

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$$

CLT For Regression

- In large samples, we can relax the normality assumption on u_i and use CLT:

$$\hat{\beta}_1 \quad \xrightarrow{d} \quad N\left(\beta_1, \frac{\sigma}{\sqrt{S_{xx}}}\right) and \qquad \hat{\beta}_0 \quad \xrightarrow{d} \quad N\left(\beta_0, \sigma\sqrt{(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}\right)$$

Why does this happen?

- The OLS slope is a **linear combination of the errors**:

$$\hat{\beta}1 = \beta_1 + \sum_i \underbrace{\tfrac{(x_i - \bar{x})}{S{xx}}}_{\text{weight}} u_i $$

- Define each weighted error as $Z_i = \frac{(x_i - \bar{x})}{S_{xx}} u_i$,so that

$$\hat{\beta}_1 - \beta_1 = \sum_i Z_i.$$

- By the **Central Limit Theorem**, the sum of many independent mean-zero variables is approximately normal:

$$\sum_i Z_i \xrightarrow{d} N\big(0,\ \mathrm{Var}(\sum_i Z_i)\big).$$

- Since $\mathrm{Var}(\sum_i Z_i) = \sum_i var(Z_i) = \frac{\sum_i (x_i - \bar{x})^2}{S_{xx}^2}\sigma^2 = \frac{S_{xx}}{S_{xx}^2}\sigma^2 = \frac{\sigma^2}{S_{xx}}$, we get

$$\hat{\beta}_1 - \beta_1 \xrightarrow{d} N\left(0,\ \frac{\sigma^2}{S_{xx}}\right).$$

Hypothesis Testing

Our **test statistic** for β_1 and it's distribution under the null hypothesis: $H_0 : \beta_1 = b_1$

$$T = \frac{\hat{\beta}_1 - b_1}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - b_1}{\frac{\hat{\sigma}}{\sqrt{S_{xx}}}} \sim t_{n-2}$$

Similarly, for β_0 the null hypothesis: $H_0 : \beta_0 = b_0$

$$T = \frac{\hat{\beta}_0 - b_0}{SE(\hat{\beta}_0)} = \frac{\hat{\beta}_0 - b_0}{\hat{\sigma}\sqrt{(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}} \sim t_{n-2}$$

With that, we can use usual hypothesis testing procedures

Example:

Does temperature predicts bike rides? Let's test it at $\alpha = 0.05$

$H_0 : \beta_1 = 0 \; H_A : \beta_1 \neq 0$

$$T_{test} = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{723.55}{83.37} = 8.679$$

We can compare it to critical value (n=781):

$$t_{779,\frac{\alpha}{2}} \approx z_{\frac{\alpha}{2}} = 1.96 < 8.679 = T_{test}$$

We confidently reject the the null that the temperature does not predict bike rides.

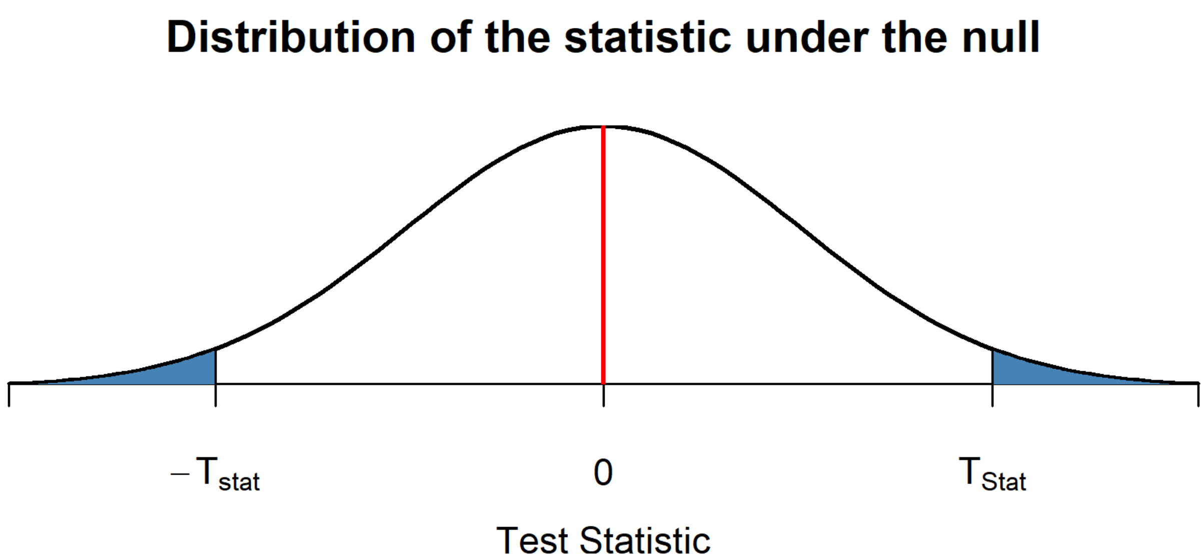
P-Value

Alternatively, calculate **p-value**: the probability of seeing our test statistic or a more extreme test statistic if the null hypothesis were true.

In regressions we usually use two-sided tests. Hence the p-value is:

$$p - value = 2 * P(t_{n-2,\frac{\alpha}{2}} > T_{test})$$

Small p-values mean that it would be unlikely to see our results if the null hypothesis were really true.



Regression Output

```
##
## Call:
## lm(formula = Trips ~ TMP, data = Data_BP)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24010.5  -1508.4    774.5   2920.5   8900.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16892.66    1427.32  11.835   <2e-16 ***
## TMP           723.55      83.37   8.679   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5302 on 779 degrees of freedom
## Multiple R-squared:  0.08817,    Adjusted R-squared:  0.087
## F-statistic: 75.32 on 1 and 779 DF,  p-value: < 2.2e-16
```

Confidence Intervals

Using the distributions, we can figure out confidence intervals for our estimates:

$$P(-t_{n-2,\frac{\alpha}{2}} < \frac{\hat{\beta}_1 - \beta}{SE(\hat{\beta}_1)} < t_{n-2,\frac{\alpha}{2}}) = 1 - \alpha$$

$$CI_{\beta_1} = \left(\hat{\beta}_1 - t_{n-2,\frac{\alpha}{2}} \underbrace{\frac{\hat{\sigma}}{\sqrt{S_{xx}}}}_{SE(\hat{\beta}_1)}, \hat{\beta}_1 + t_{n-2,\frac{\alpha}{2}} \underbrace{\frac{\hat{\sigma}}{\sqrt{S_{xx}}}}_{SE(\hat{\beta}_1)} \right)$$

And Similarly for β_0

$$CI_{\beta_0} = \left(\hat{\beta}_0 - t_{n-2,\frac{\alpha}{2}} \underbrace{\hat{\sigma}\sqrt{(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}}_{SE(\hat{\beta}_0)}, \hat{\beta}_0 + t_{n-2,\frac{\alpha}{2}} \underbrace{\hat{\sigma}\sqrt{(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}}_{SE(\hat{\beta}_0)} \right)$$

Confidence Intervals

What's the confidence 95% interval for the effect on temperature?

$$CI_{\beta_1} = (723.55 - 1.96 * 83.37, 723.55 + 1.96 * 83.37)$$

$$CI_{\beta_1} = (560.87, 886.23)$$

Confidence Intervals

Suppose we instead want to estimate the impact of pollution (PM10) on bike trips.

```
##
## Call:
## lm(formula = Trips ~ PM10, data = Data_BP)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27079.4  -1298.2    947.1   3155.8   8938.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28382.98     576.49  49.235   <2e-16 ***
## PM10           16.99      11.68   1.455    0.146
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5544 on 779 degrees of freedom
## Multiple R-squared:  0.002709,    Adjusted R-squared:  0.001429
## F-statistic: 2.116 on 1 and 779 DF,  p-value: 0.1462
```

- Can we reject null of no impact at 10%?
- What's the 90% confidence interval?

Confidence Intervals

Average response: What would be average number of rides on days with temperature of 30C?

$$(\bar{y}|x = x_0) = \hat{\beta}_0 + \hat{\beta}_1 x$$

What's the expectation?

$$E(\bar{y}|x = x_0) = E(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \beta_0 + \beta_1 x_0$$

What's the variance?

$$var(\bar{y}|x = x_0) = Var(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \sigma^2(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}})$$

What's the distribution:

$$(\bar{y}|x = x_0) \sim N\left(\beta_0 + \beta_1 x_0, \sigma\sqrt{(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}})}\right)$$

Confidence Intervals

We can build the confidence intervals as before:

$$CI_{(\bar{y}|x=x_0)} = \hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{n-2,\frac{\alpha}{2}} \underbrace{\hat{\sigma}\sqrt{(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}})}}_{SE}$$

Confidence Intervals

What would be 95% CI for average number of rides if temperature is 30C?

- $\hat{\beta}_0 = 16892.66$ and $\hat{\beta}_1 = 723.55$
- n=781
- $\bar{x} = 16.96$
- $S_{xx} = 4044$
- $\sum_i e^2 = 21895427100$
- $\hat{\sigma} = \sqrt{\frac{\sum_i e^2}{n-2}} = 5301.613$

$$CI_{(\bar{y}|x=x_0)} = 16892.66 + 723.55 * 30 \pm 1.96 * \underbrace{5301.613\sqrt{(\frac{1}{781} + \frac{(30-16.96)^2}{4044})}}_{SE}$$

$$CI_{(\bar{y}|x=x_0)} = 38599.16 \pm 2161.588$$

- Interpretation?
  - If we take a lot of samples, and calculate confidence interval using data from each, 95% of them would contain the true value
  - We are 95% confident, true value is in the interval

Confidence Intervals

R code

```r
lm_model <- lm(Trips ~ TMP, data = Data_BP)
new_data<- data.frame(TMP= c(30))
predict(lm_model, newdata = new_data, interval = "confidence", level = (
```

```
## $fit
##        fit      lwr      upr
## 1 38599.23 36434.32 40764.14
##
## $se.fit
## [1] 1102.851
##
## $df
## [1] 779
##
## $residual.scale
## [1] 5301.613
```

Mean response vs New response

- Suppose you are checking how people react to a new drug for balding. You estimated the following regressions:

$$\text{Number of hairs}/cm^2 = \hat{\beta}_0 + \hat{\beta}_1 \text{Amount of drug in mg}$$

- For now, you were only giving doses between 1-25mg. You want to increase dosage to 30mg.
- You can have two types of confidence intervals

- For **Mean Response**
    - Suppose you give 30mg to many, many people, and you are interested in average $\text{Number of hairs}/cm^2$ among those who got 30mg
    - Since you average among many people, the u_i individual error terms does not play a role ($E(u_i) = 0$)
    - The uncertainty comes from whether you did a good job estimating βs
- For **New Response**
    - Suppose you give 30mg to one person, and you are interested in their outcome.
    - Since there is only one person, u_i will play a role
    - Maybe you picked someone who naturally has a lot of hair, or who will be on other medication which makes him lose hair
    - Those factors avarage out in mean response, so don't play a role
    - There will be more uncertainty about this new response, hence wider CI
    - In particular, $var(\text{new response}) = var(\text{mean response}) + var(u_i)$
    - For this we need to make an assumption that errors are normal, CLT is not enough!
    - Because it's not only about the distribution of β but also error term

Confidence Intervals

New response: What would be the number of rides on some day with temperature 30C?

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

What's the expectation?

$$E(\hat{y}|x = x_0) = E(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \beta_0 + \beta_1 x_0$$

How much true value varies around this prediction?

$$var(y_0 - \hat{y}|x = x_0) = Var(\hat{\beta}_0 + \hat{\beta}_1 x_0) + Var(u_i) = \sigma^2(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}})$$

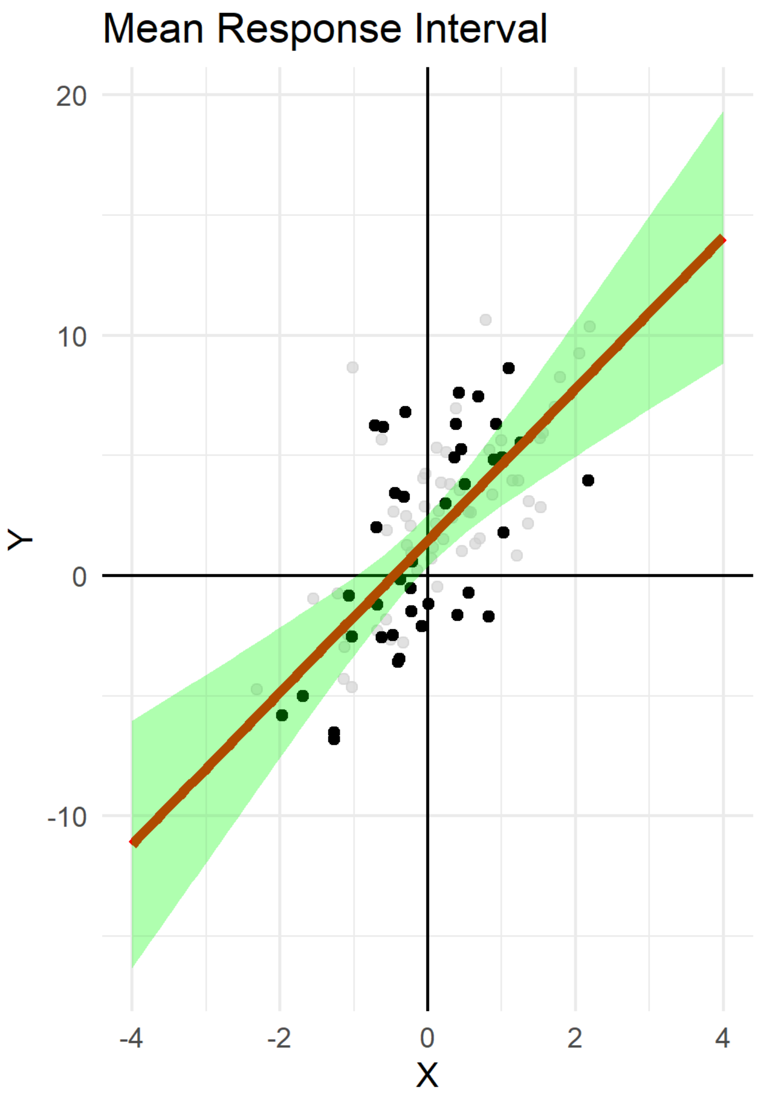
What's the distribution:

$$(\bar{y}|x = x_0) \sim N\left(\beta_0 + \beta_1 x_0, \sigma\sqrt{(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}})}\right)$$

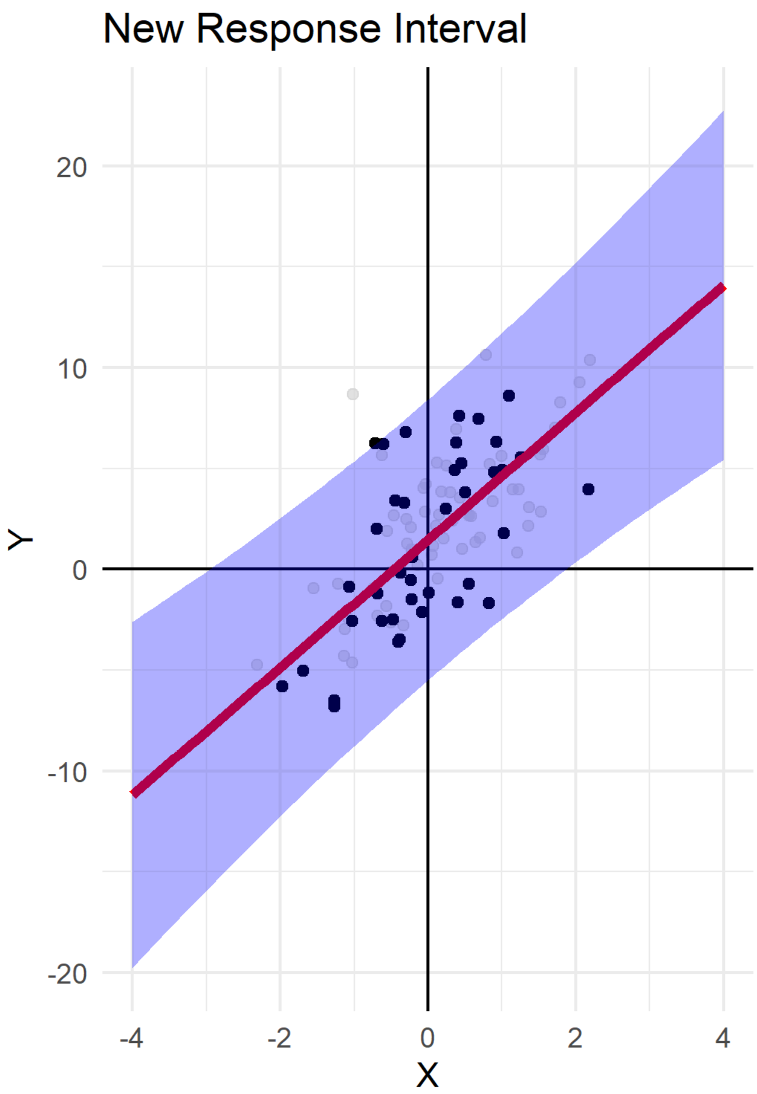
Confidence Intervals

We can build the confidence intervals as before:

$$CI_{(\bar{y}|x=x_0)} = \hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{n-2,\frac{\alpha}{2}} \underbrace{\hat{\sigma}\sqrt{(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}})}}_{SE}$$



Mean Response Interval
Y
X
20
10
0
-10
-4
-2
0
2
4



New Response Interval
Y
X
20
10
0
-10
-20
-4
-2
0
2
4

Confidence Intervals

What would be 95% CI for number of rides on some day with 30C?

R code

```
lm_model <- lm(Trips ~ TMP, data = Data_BP)
new_data<- data.frame(TMP= c(30))
predict(lm_model, newdata = new_data, interval = "predict", level = 0.95
```

```
## $fit
##        fit     lwr      upr
## 1 38599.23 27969.3 49229.16
##
## $se.fit
## [1] 1102.851
##
## $df
## [1] 779
##
## $residual.scale
## [1] 5301.613
```

Question

Suppose a model where we have employee's salary and their years of education. Predictor variable is education, response variable is salary. We try to establish the relationship between education and salary.

- What type of factors may affect the stochastic error u_i?
- Are they correlated with education?
- Would the estimator be unbiased?