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

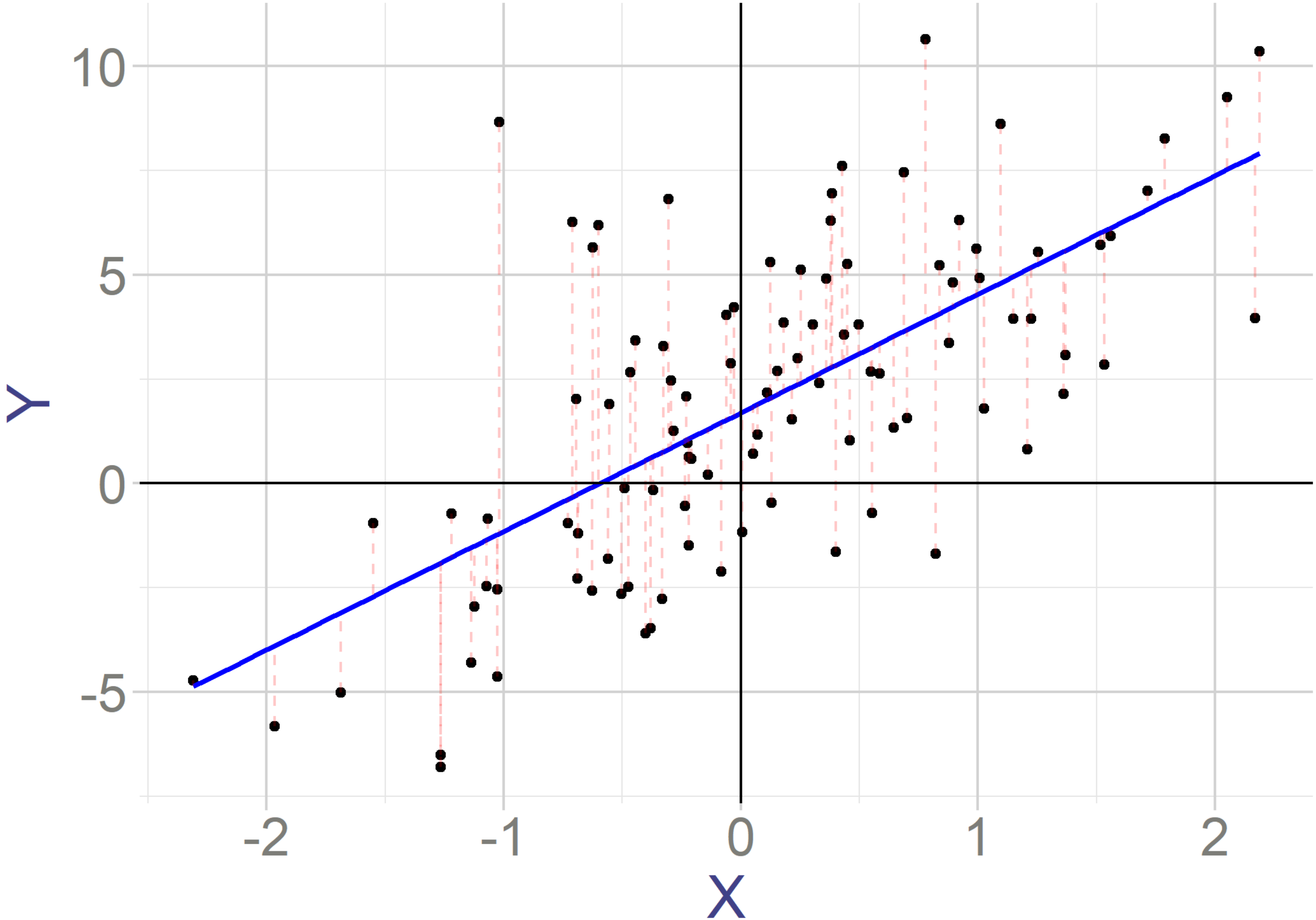
1. Suppose you have paired data: $\{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\}$

2. 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)



10
5
0
-5
-2
-1
0
1
2
Y
X

Assumptions

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

Assumptions:

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. $cov(u_i, u_j) = 0$

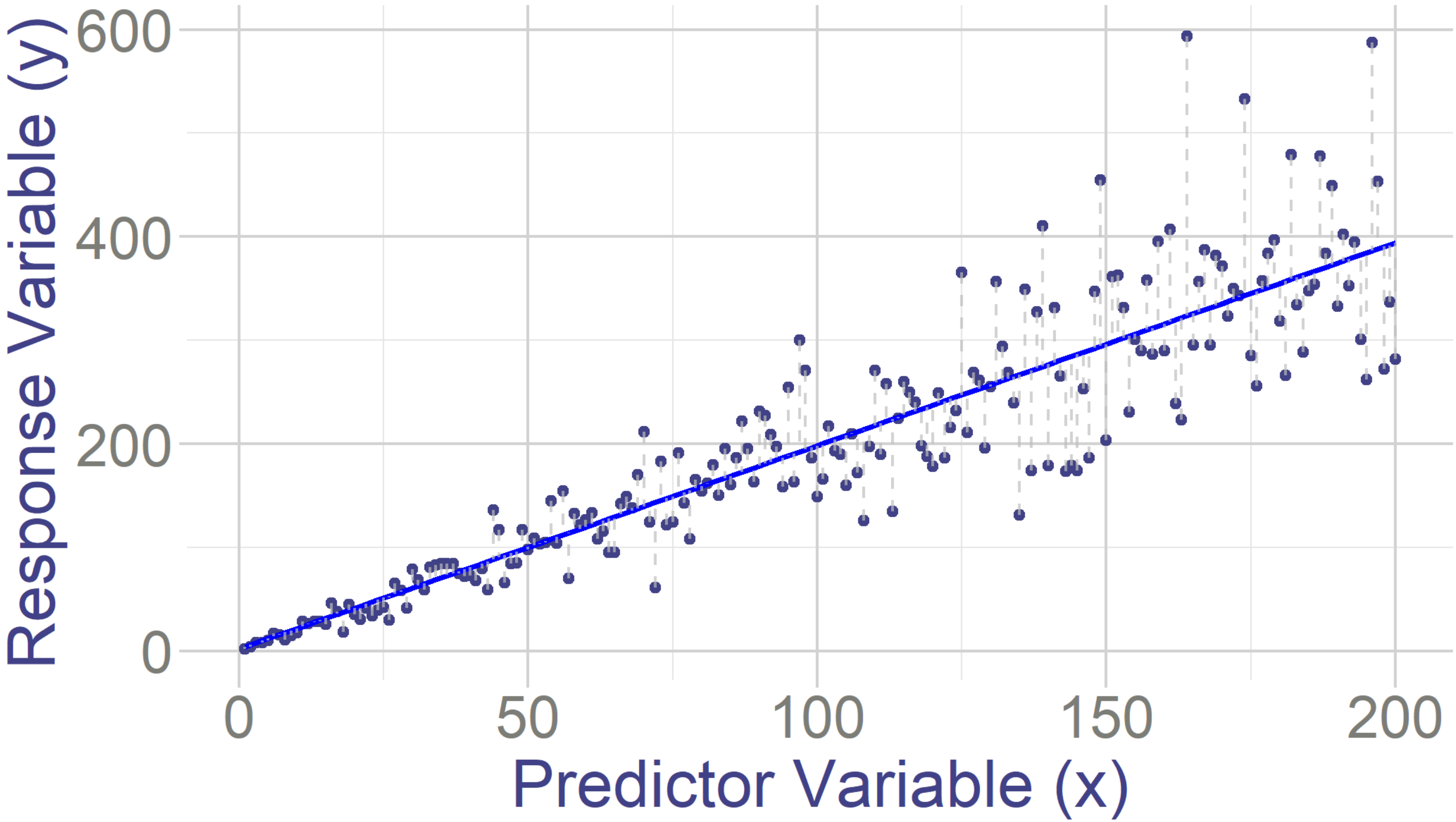
Model is linear in the parameter and with additive error term

- 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
  - $y_i = (\beta_0 + \beta_1 x_i) * e_i$
  - $y_i = \beta_0 + x_i^{\beta_1} + e_i$
  - $y_i = log(\beta_0 + \beta_1 x_i + e_i)$
  - $y_i = \beta_0 + (\beta_1 x_i + e_i)^2$

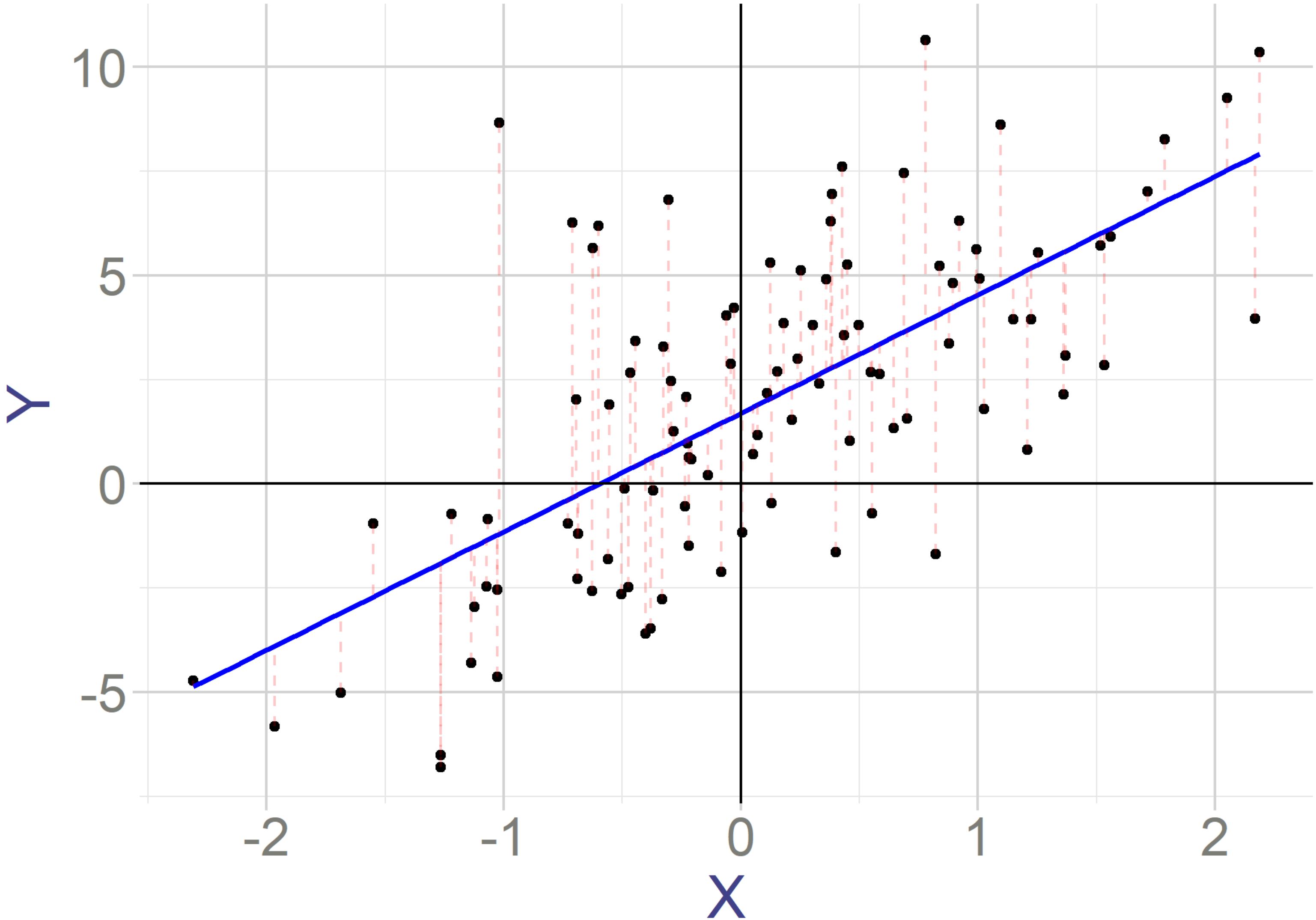
2 is in the app

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

What happens if this is not true?



Let's go back to our regression line



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

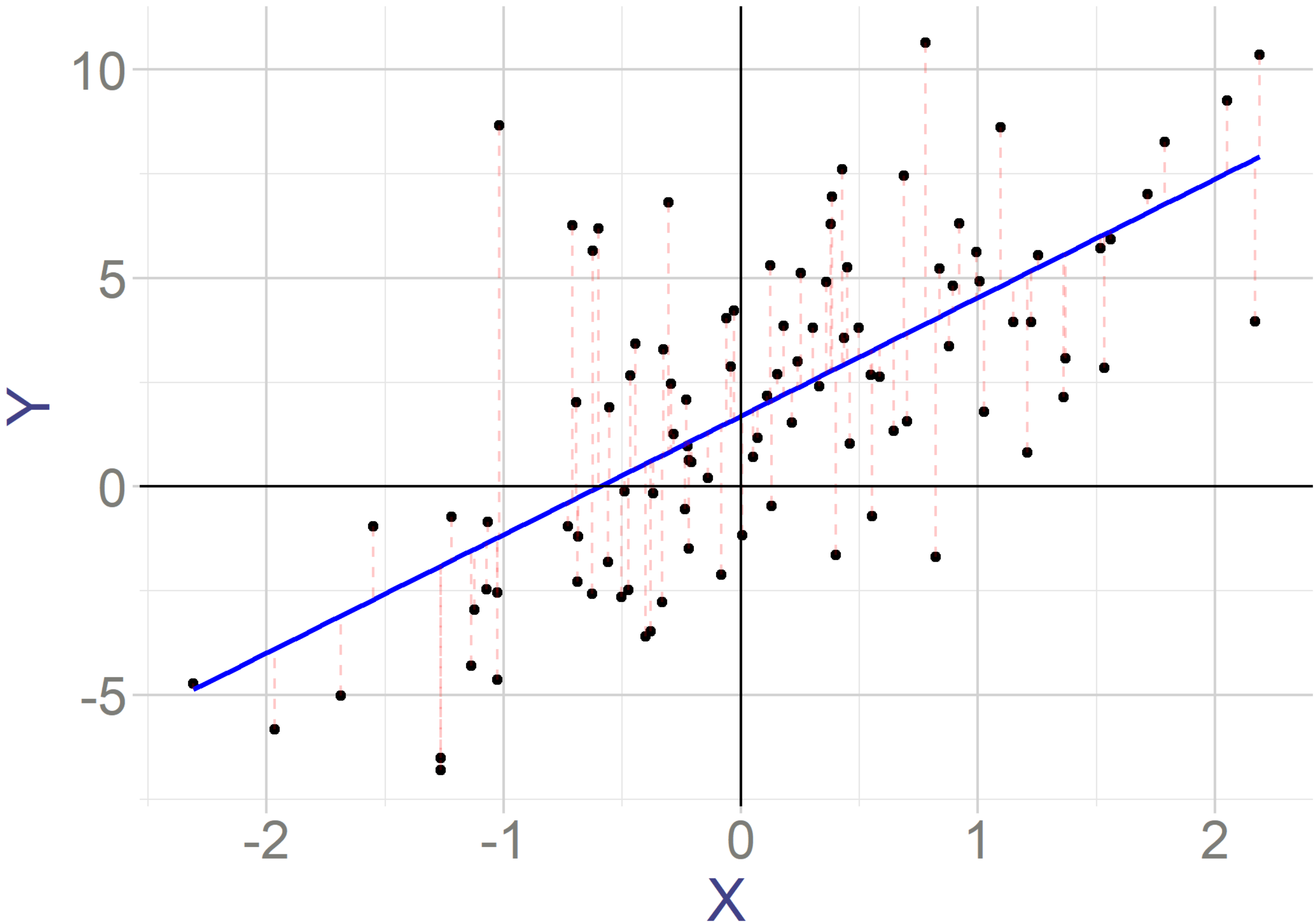
Once estimated, we can write y_i as

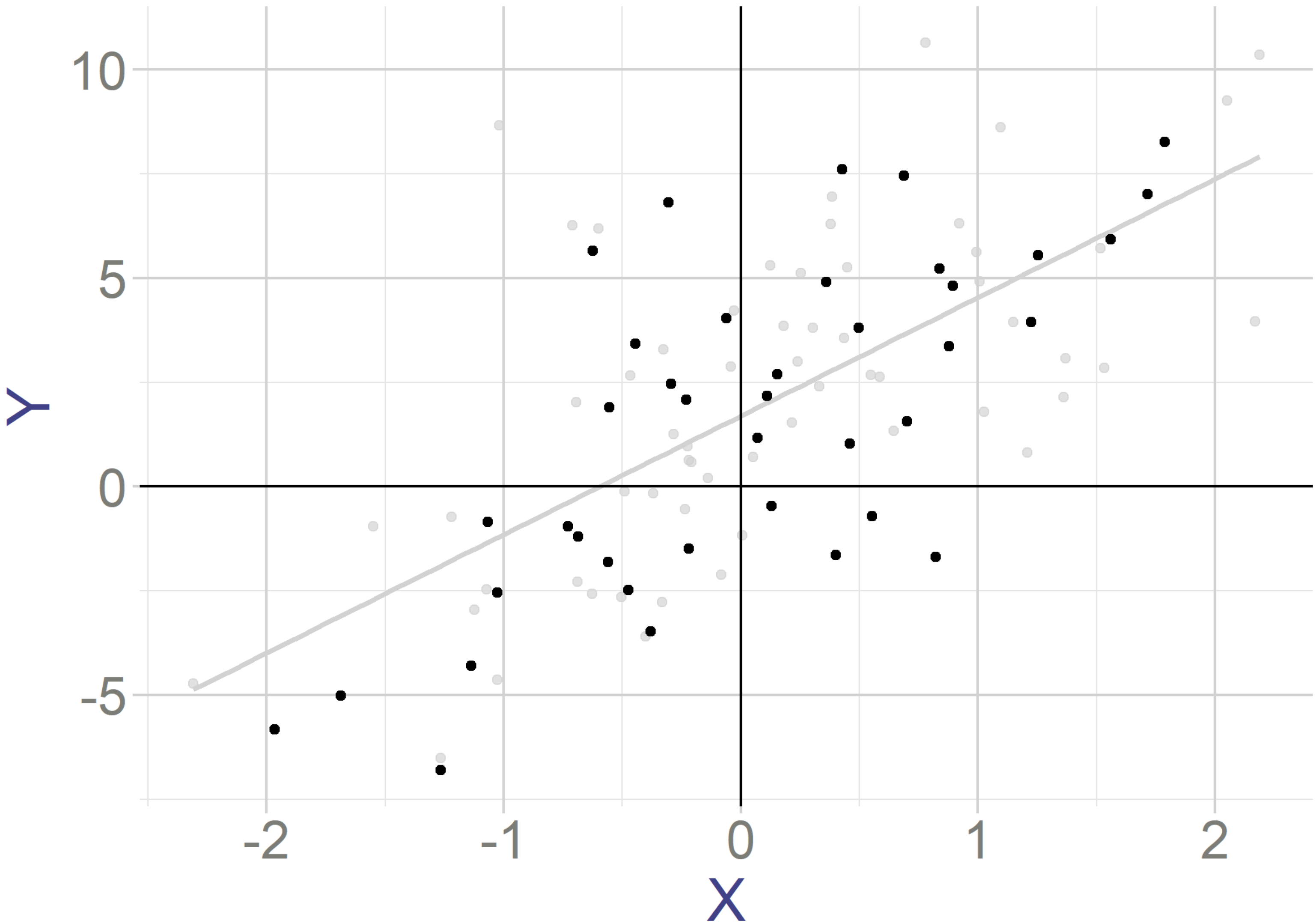
$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$$

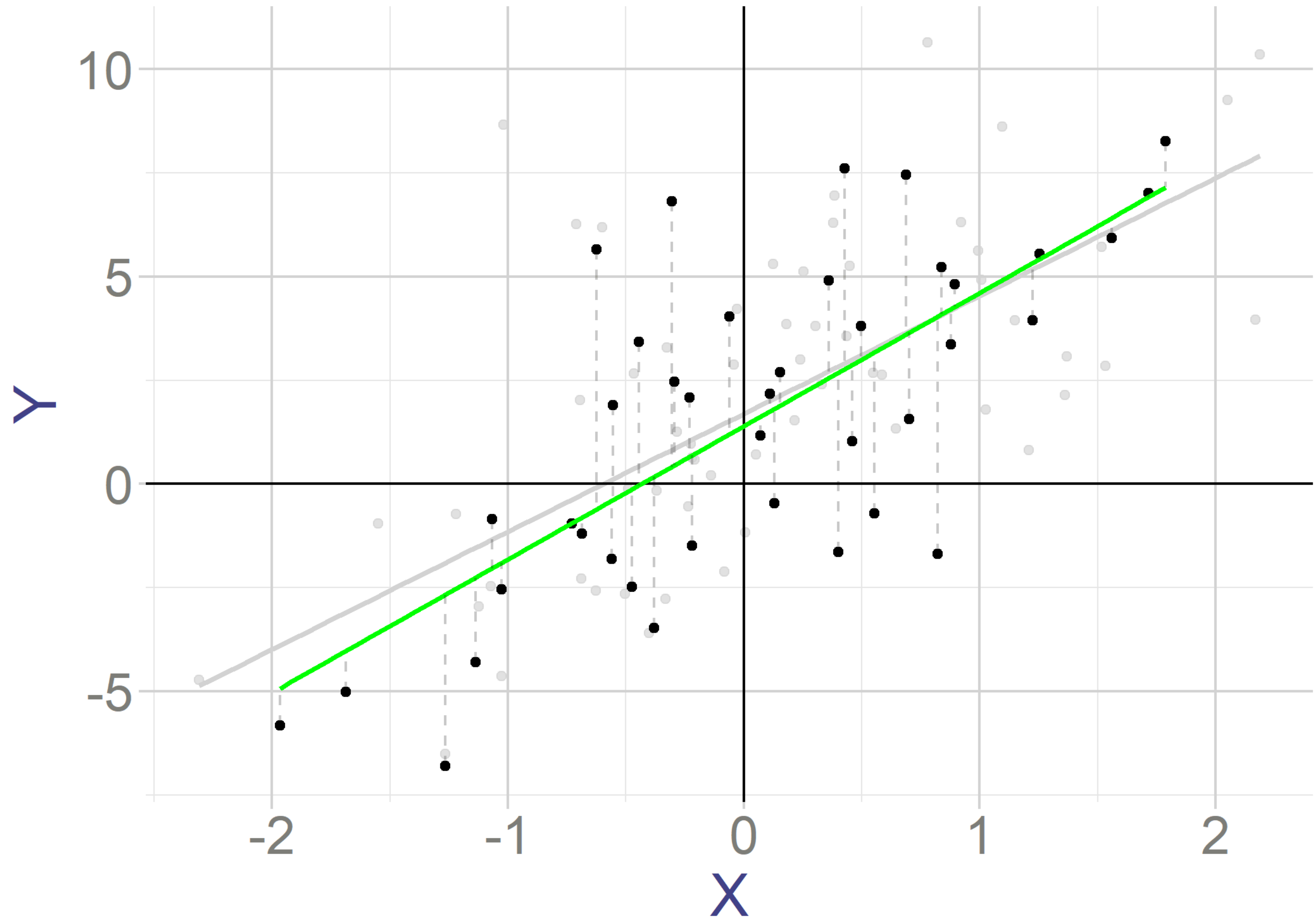
Error term here reflects both uncertainly about parameters and the random part present in population model

We can predict y_i for any x_i using our estimates

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


10
5
0
-5
Y
-2
-1
0
1
2
X


10
5
0
-5
Y
-2
-1
0
1
2
X



But 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 $SSE = \sum_i^n e_i^2$

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

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

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

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

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:

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

Setting this derivative to zero:

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

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

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:

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

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

Back to Motivating example

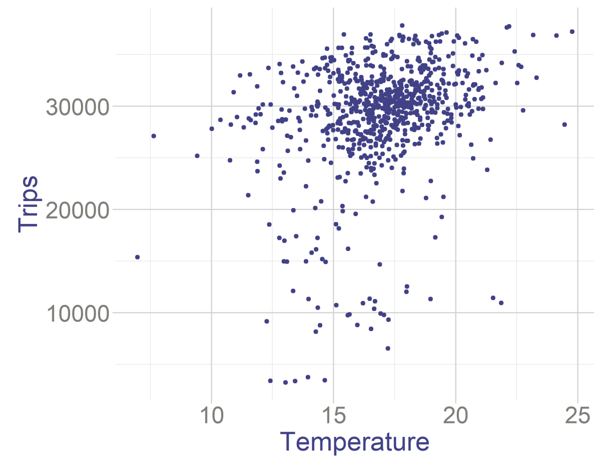
Show 7 entries

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

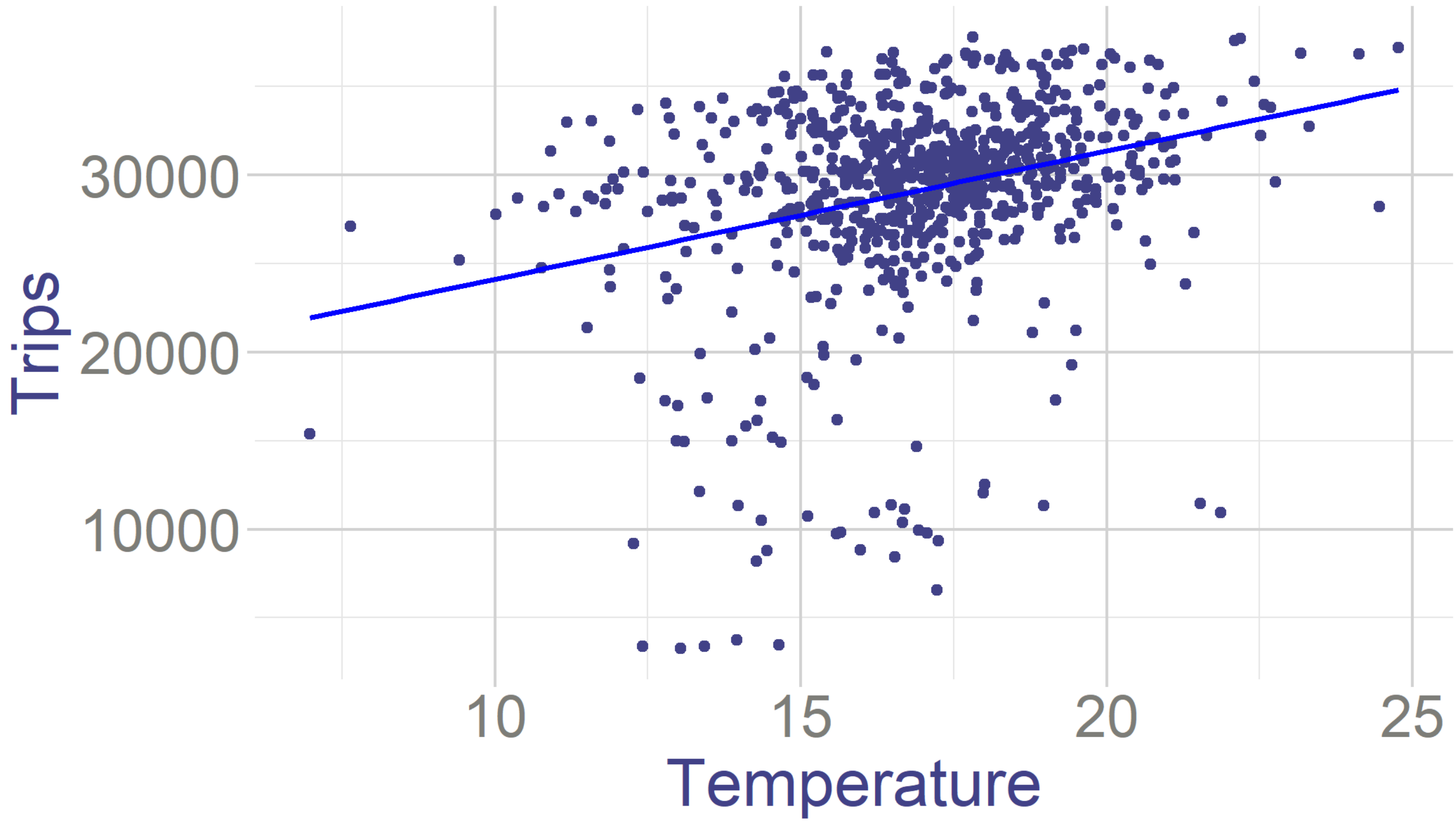
Next



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)
```

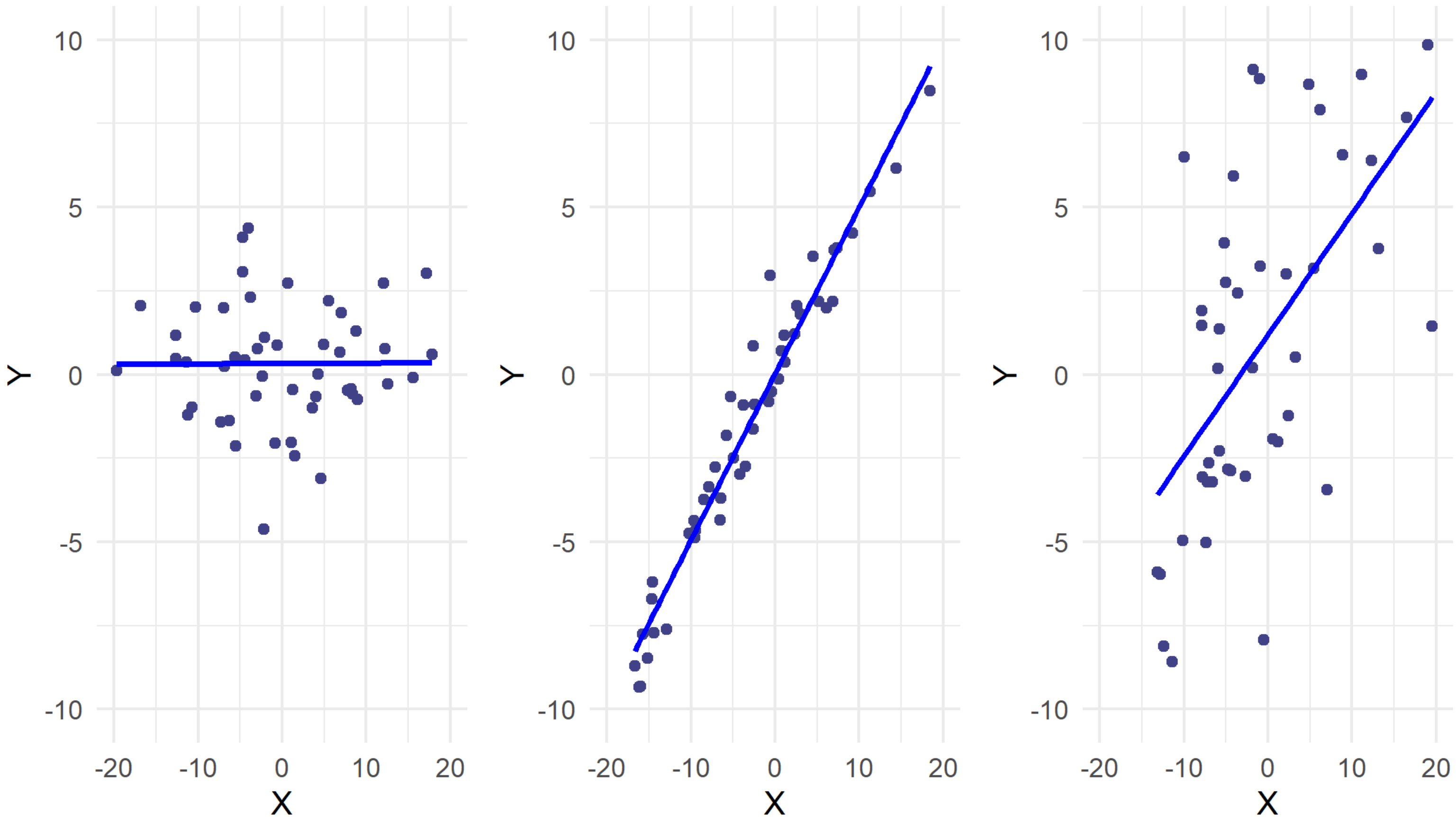
```
##
## Call:
## lm(formula = Trips ~ TMP, data = Data_BP)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24010.5  -1508.4    774.5   2920.5   8900.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16892.66    1427.32  11.835   <2e-16 ***
## 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
```

Properties of this estimator

Here is a couple of cool useful properties of OLS. Let's derive them:

- $\sum e_i = \sum(y_i - \hat{y}_i) = 0$
- $\sum y_i = \sum \hat{y}_i$
- $\hat{y}_i|(x_i = 0) = 0 * \hat{\beta}_1 + \hat{\beta}_0 = \hat{\beta}_0$
- $\sum x_i e_i = 0$
- $\sum \hat{y}_i e_i = 0$
- $var(e_i) = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-2} = \frac{SSE}{n-2}$

Fit of linear regression



Measure of fit - R squared

How much we managed to explain with our regression?

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

Measure of fit is:

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

Intuition:

- How much more 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.

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

Moreover, we can show that:

$$R^2 = (\rho(x,y))^2 = \hat{\beta}_1^2 \frac{S_{xx}}{S_{yy}} = \hat{\beta}_1^2 \frac{\sum(x_i - \bar{x})^2}{\sum(y_i - \bar{y})^2}$$

How much of bike usage does the temperature explains?

- $\beta_1 = 723.55$
- $S_{xx} = var(x) * (n-1) = 4043.965$
- $S_{yy} = var(y) * (n-1) = 24012556582$

```
##
## 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:

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

Regression through the origin

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 throuth the origin?

Regression with a categorical variable

- What if x_i is a categorical variable?
- Example: $x_i = 1$ if female, $x_i = 0$ if male
- We called it a binary variable, or a dummy variable

$$\hat{\beta}_0 = \bar{y}_{x_i=0}$$

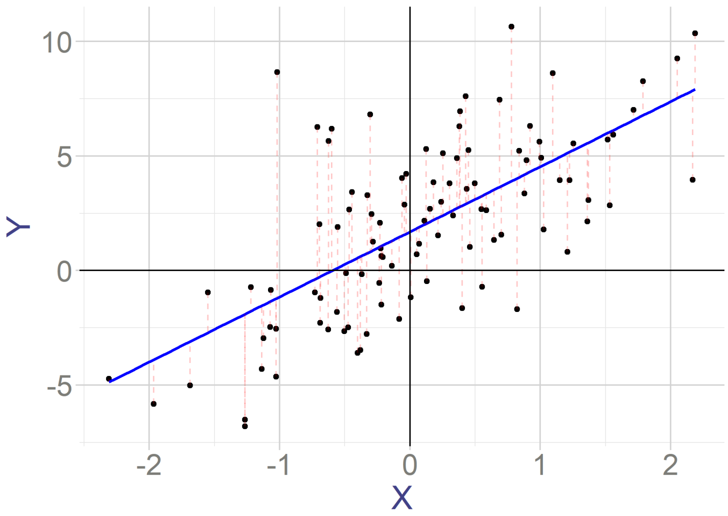
and

$$\hat{\beta}_1 = \bar{y}_{x_i=1} - \bar{y}_{x_i=0}$$

Statistical Properties of OLS

Uncertainty in the Estimate

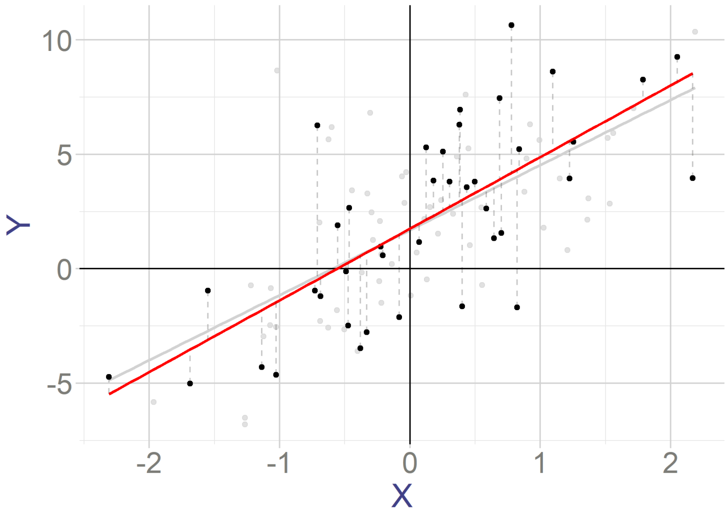
We only have samples, and yet we want to learn something about the population parameters



Population Regression

$$y_i = 1.69 + 2.84x_i + u_i$$

Uncertainty in the Estimate



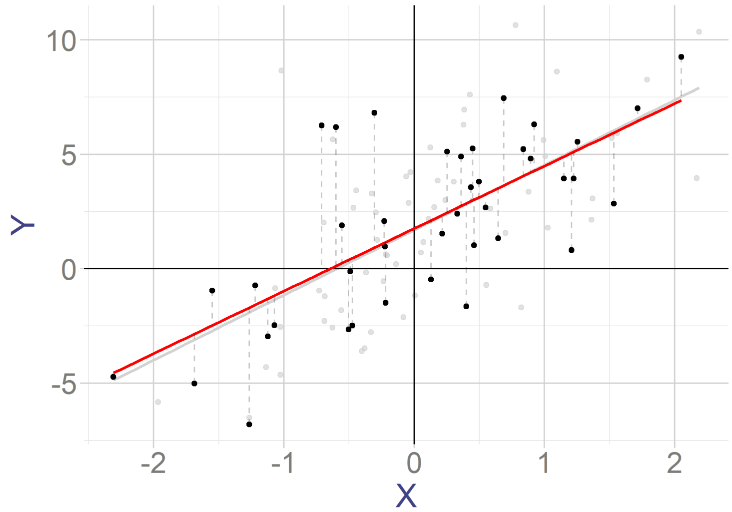
Population Regression

$$y_i = 1.69 + 2.84x_i + u_i$$

Sample Estimate

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

Uncertainty in the Estimate



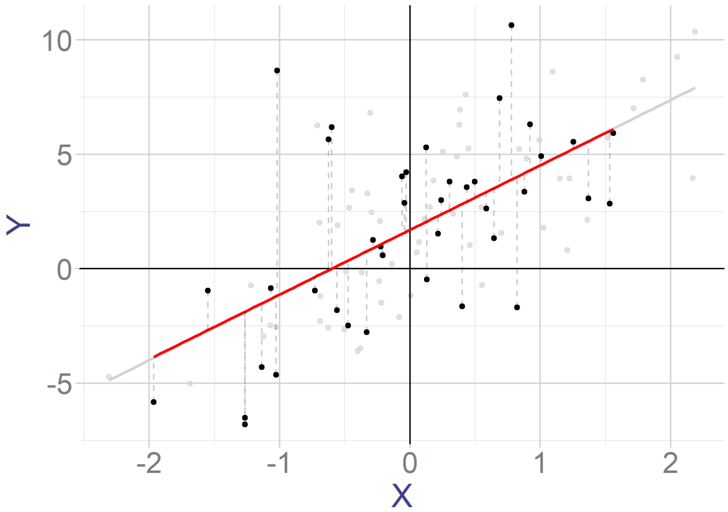
Population Regression

$$y_i = 1.69 + 2.84x_i + u_i$$

Sample Estimate

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

Uncertainty in the Estimate



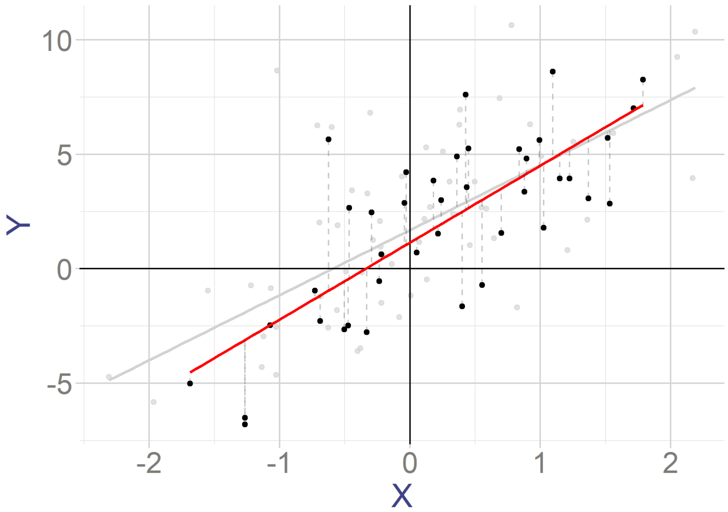
Population Regression

$$y_i = 1.69 + 2.84x_i + u_i$$

Sample Estimate

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

Uncertainty in the Estimate



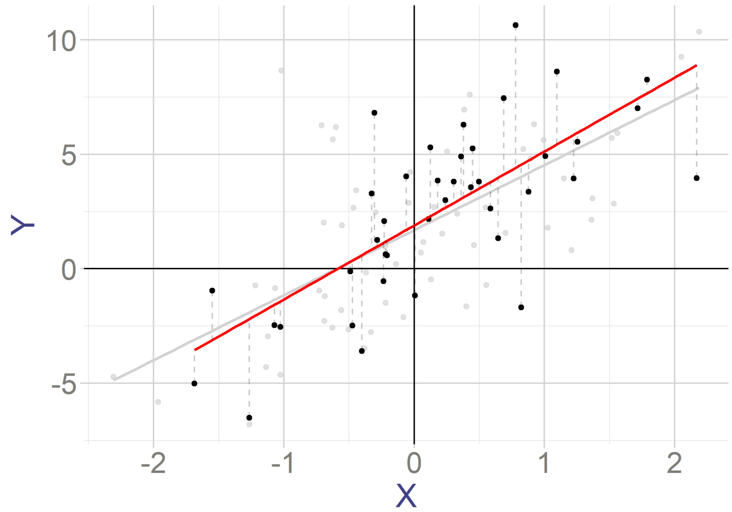
Population Regression

$$y_i = 1.69 + 2.84x_i + u_i$$

Sample Estimate

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

Uncertainty in the Estimate



Population Regression

$$y_i = 1.69 + 2.84x_i + u_i$$

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: $F = 1.8C + \mathbf{32}$

- 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:

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

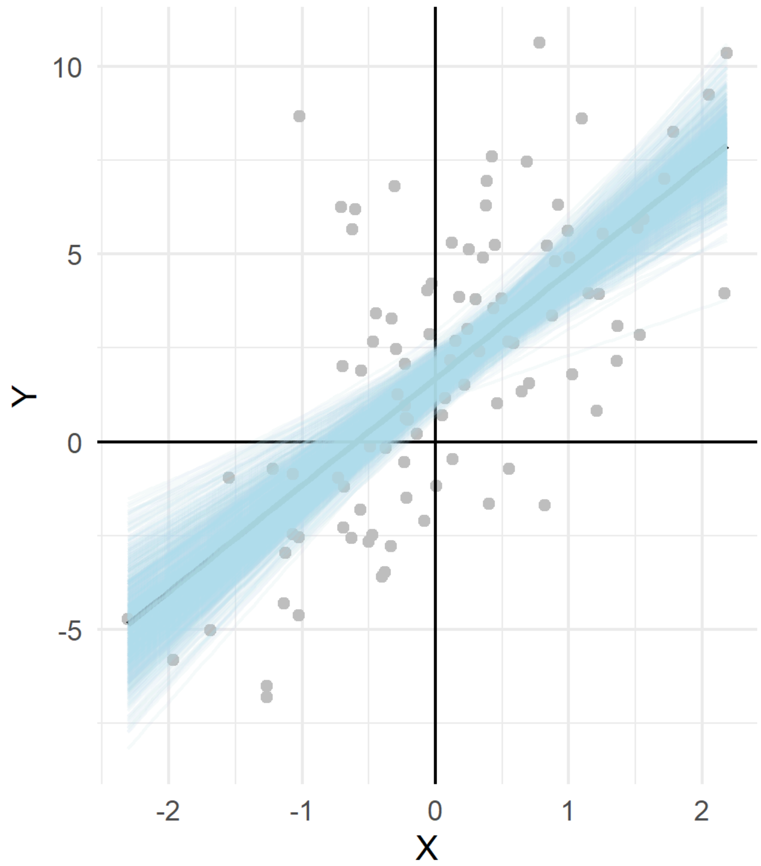
Inference

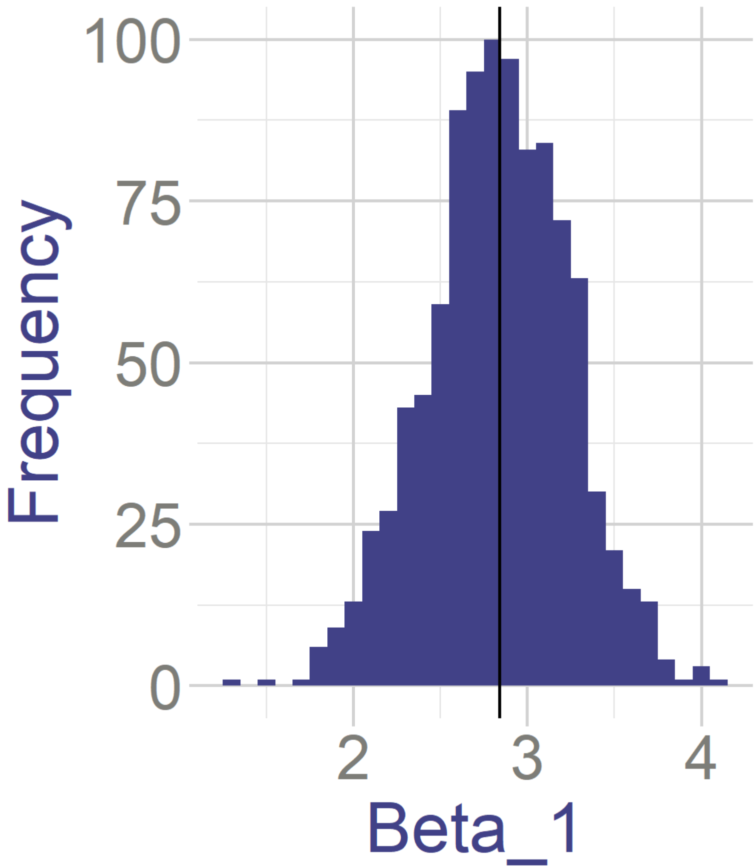
- Until now, we haven't made any assumptions about the **distributions** of the underlying data or β
  - We don't need it for calculating coefficients β_0 or β_1
  - We don't need it for making predictions $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
  - We don't need it to calculate variance or expectation 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 β
  - In particular, we will assume that population errors are normally distributed: $u_i \sim N(0, \sigma)$
  - This will help us to determine the distribution of β
  - 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: $y_i | x_i \sim N(\hat{\beta}_0 + \hat{\beta}_1 x_i, \sigma)$

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

$$y_i = 1.69 + 2.84x_i + u_i$$

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:

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

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

Hypothesis Testing

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

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

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

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

With that, we can use usual hypothesis testing procedures

Example:

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

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

GRAPHS: slope - show with graph the alternative and the null

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

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.

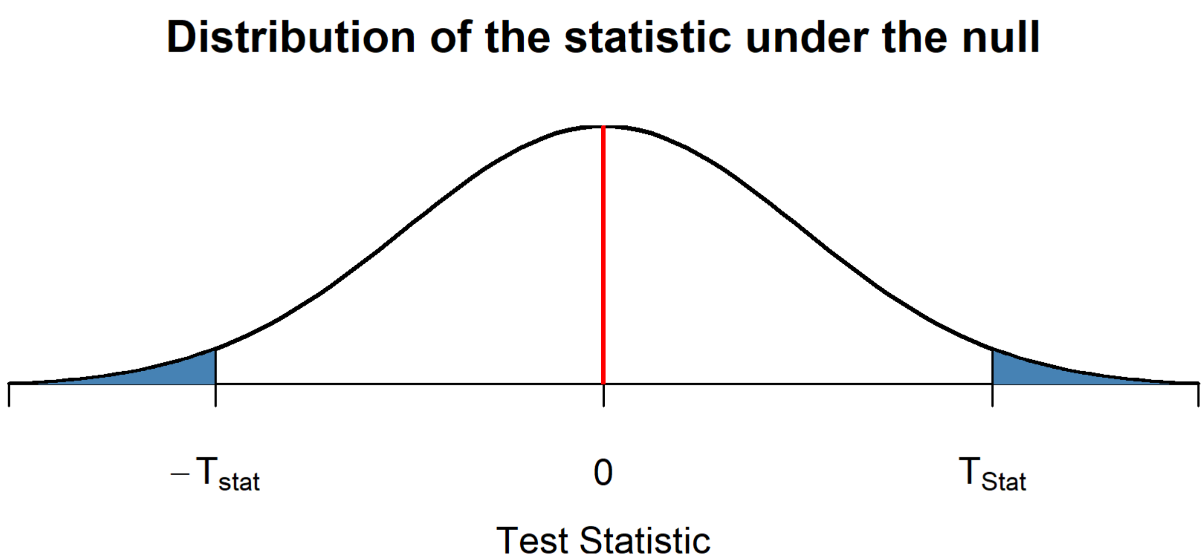
P-Value

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:

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

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

And Similarly for β_0

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

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:

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

Confidence Intervals

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

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

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

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

- Interpretation?
  - If we take a lot of samples, and calculate confidence interval using data

Confidence Intervals

R code

```
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, $var(\text{new response}) = var(\text{mean response}) + var(u_i)$

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?

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

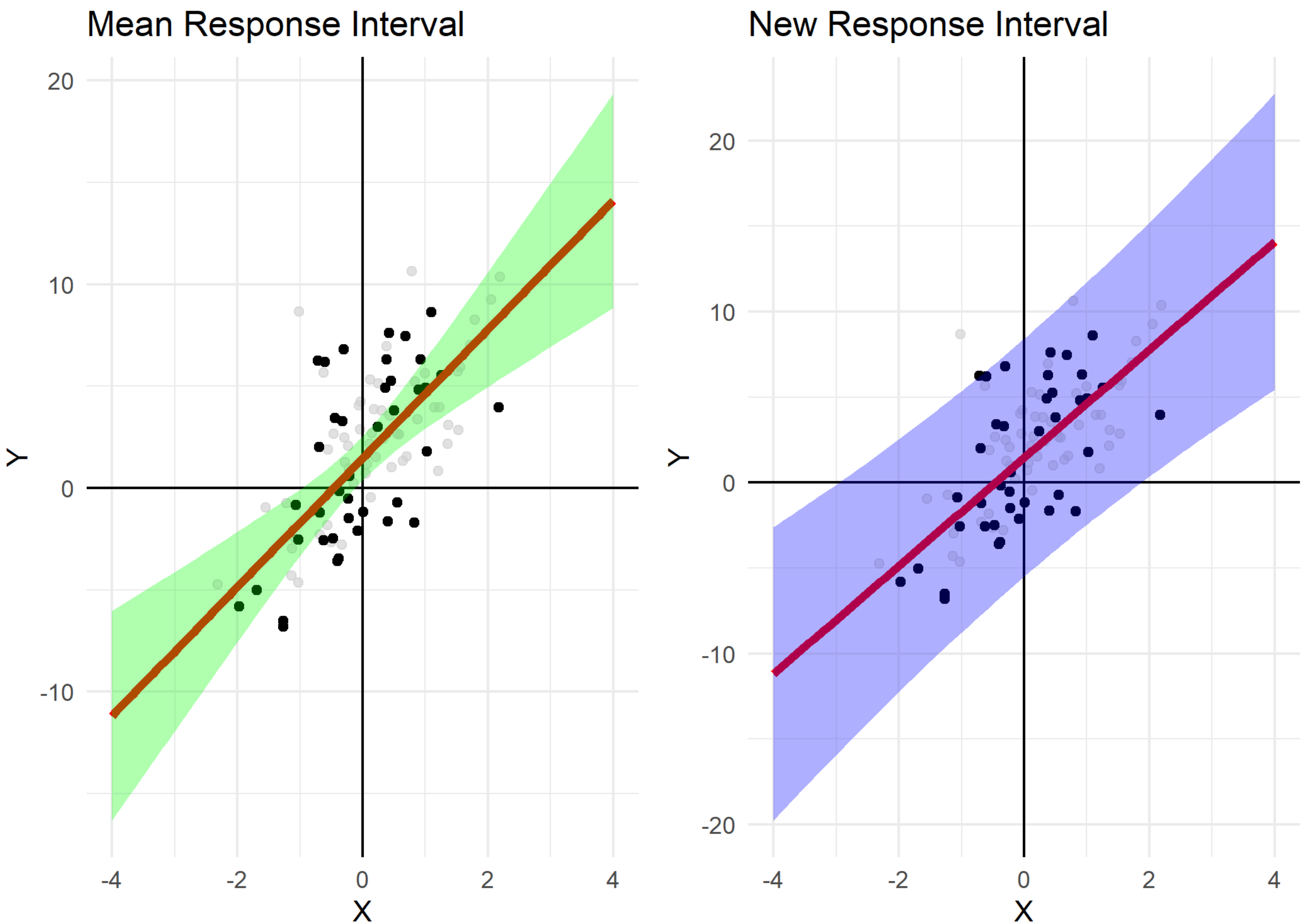
What's the distribution:

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

Confidence Intervals

We can build the confidence intervals as before:

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


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

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?