Introduction to Linear Regression

CS109A Introduction to Data Science

Pavlos Protopapas, Kevin Rader, and Chris Gumb

Chrissy Cadigan

Lecture Outline

Simple Linear Regression

Multi-linear Regression

Interpreting Model Parameters

Scaling

Collinearity

Qualitative Predictors

Lecture Outline

Simple Linear Regression

Multi-linear Regression

Interpreting Model Parameters

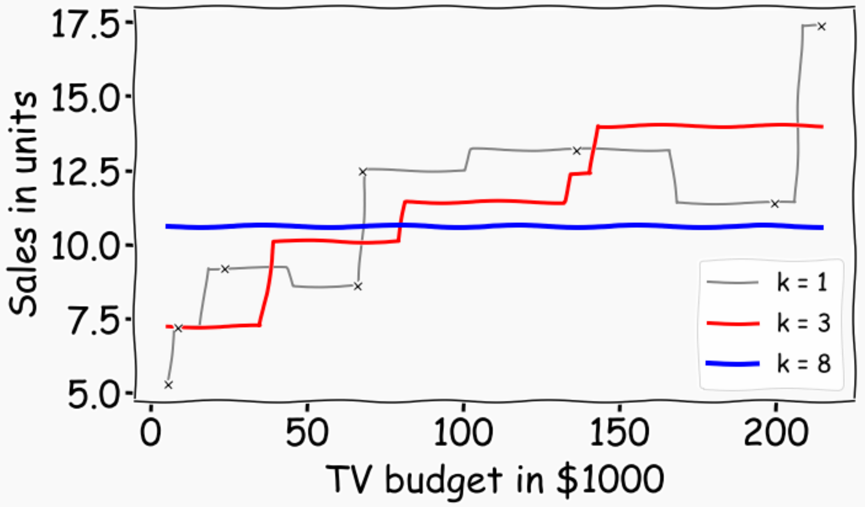
Scaling

Collinearity

Qualitative Predictors

Linear Models

kNN model



Note that when building our kNN model for prediction, which is non-parametric, we did not compute a closed-form solution for $\hat{f}$. So, what happens when we pose the question

How much more in sales can we expect if we double the TV advertising budget?

Linear Regression

Linear Models

We can build a model by first assuming a simple form of f:

$$\boxed{f(x) = \beta_0 + \beta_1 x}$$

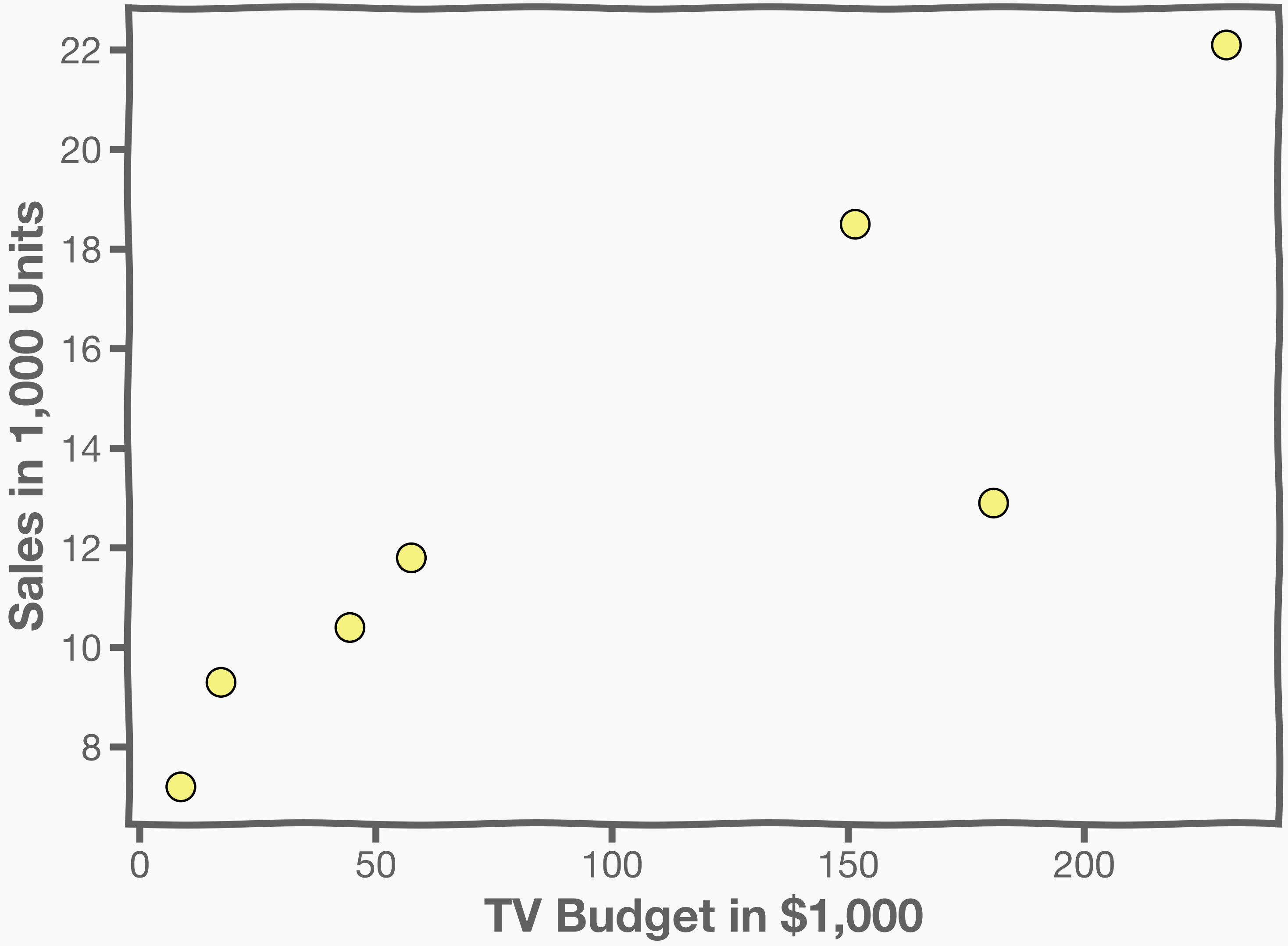
… then it follows that our estimate is:

$$\boxed{\hat{y} = \hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x}$$

where $\hat{\beta}_1$ and $\hat{\beta}_0$ are **estimates** of β_1 and β_0 respectively, that we compute using observations.

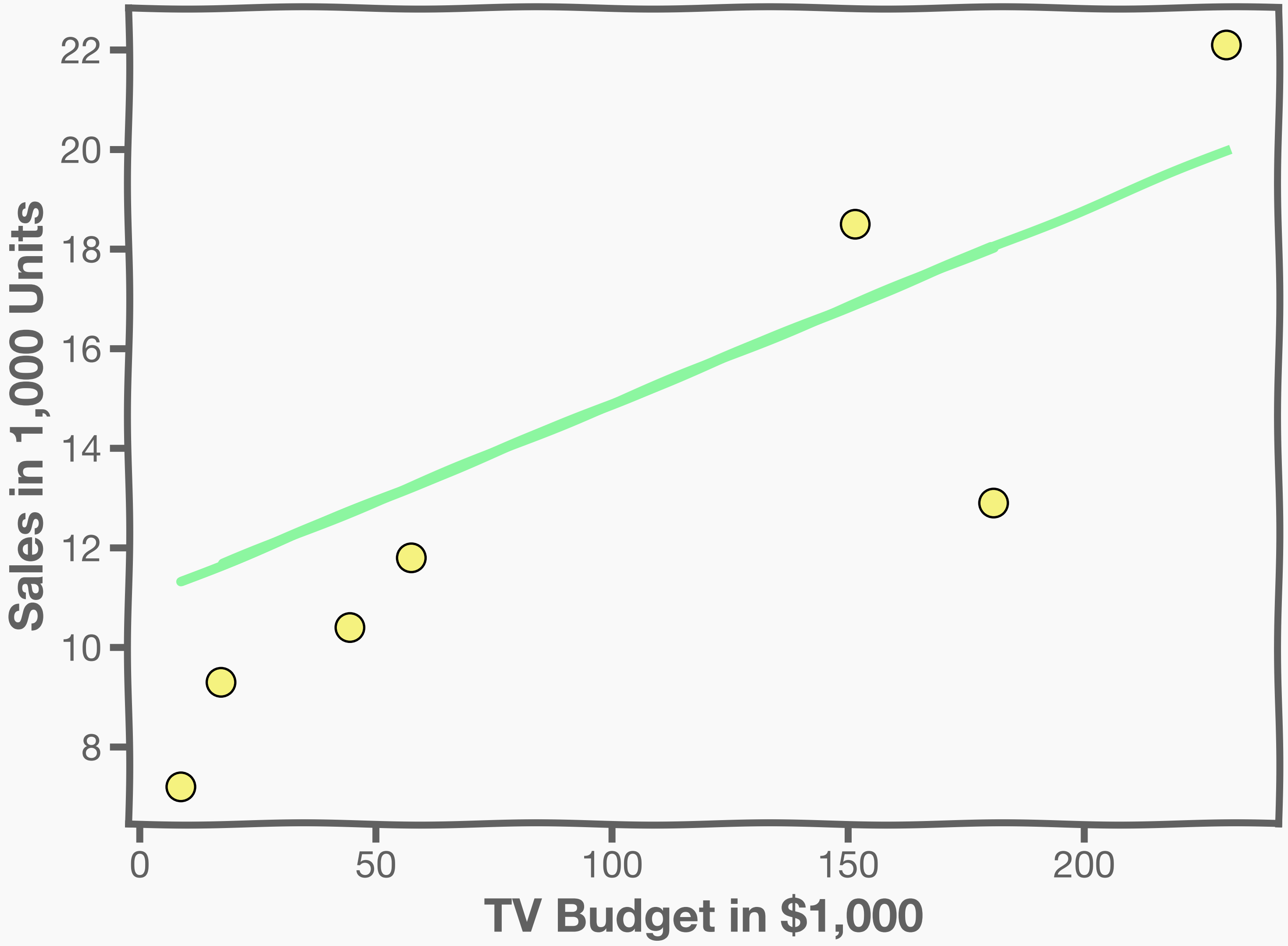
Estimate of the regression coefficients

For a given data set



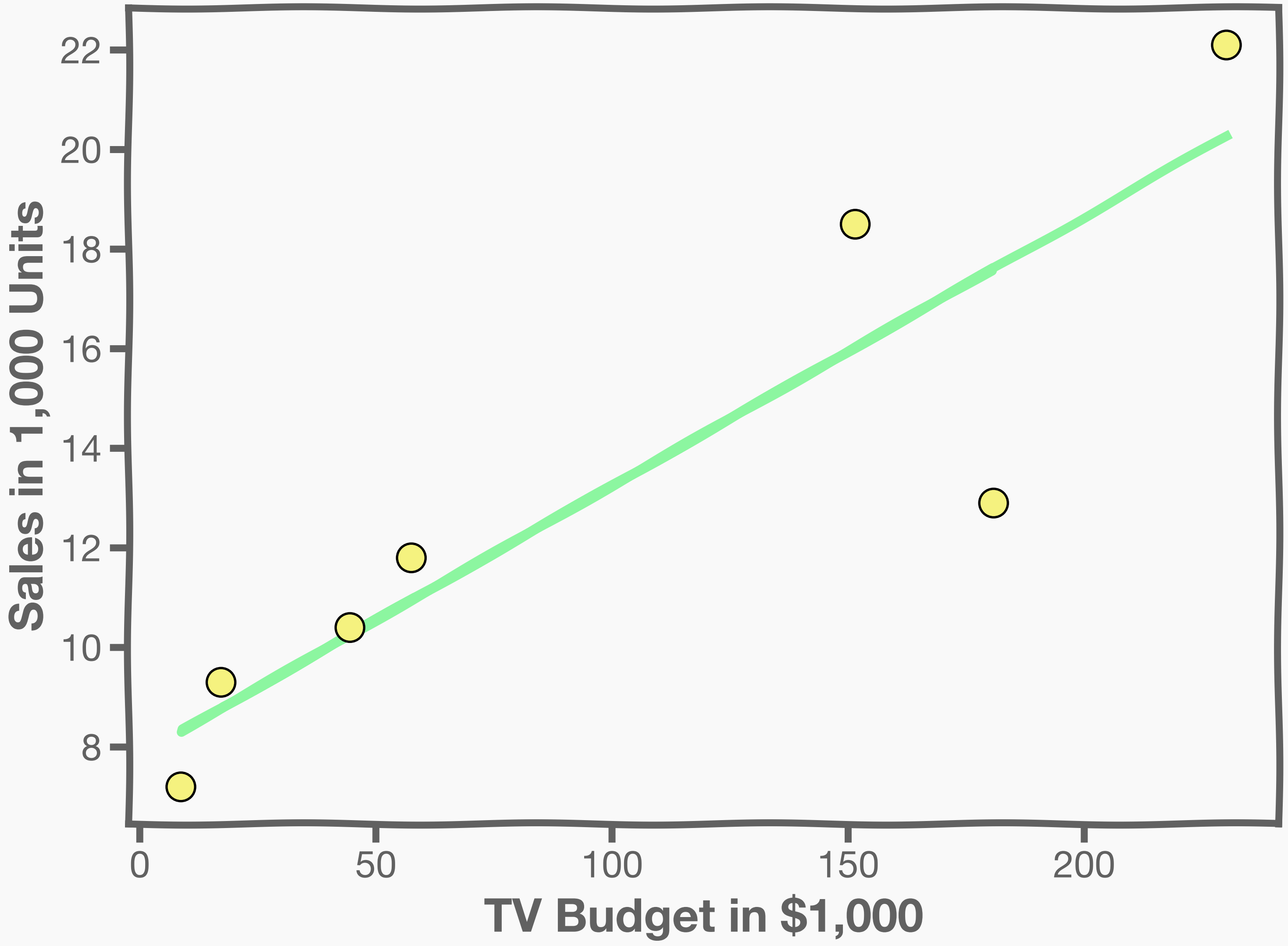
Estimate of the regression coefficients (cont)

Is this line good?



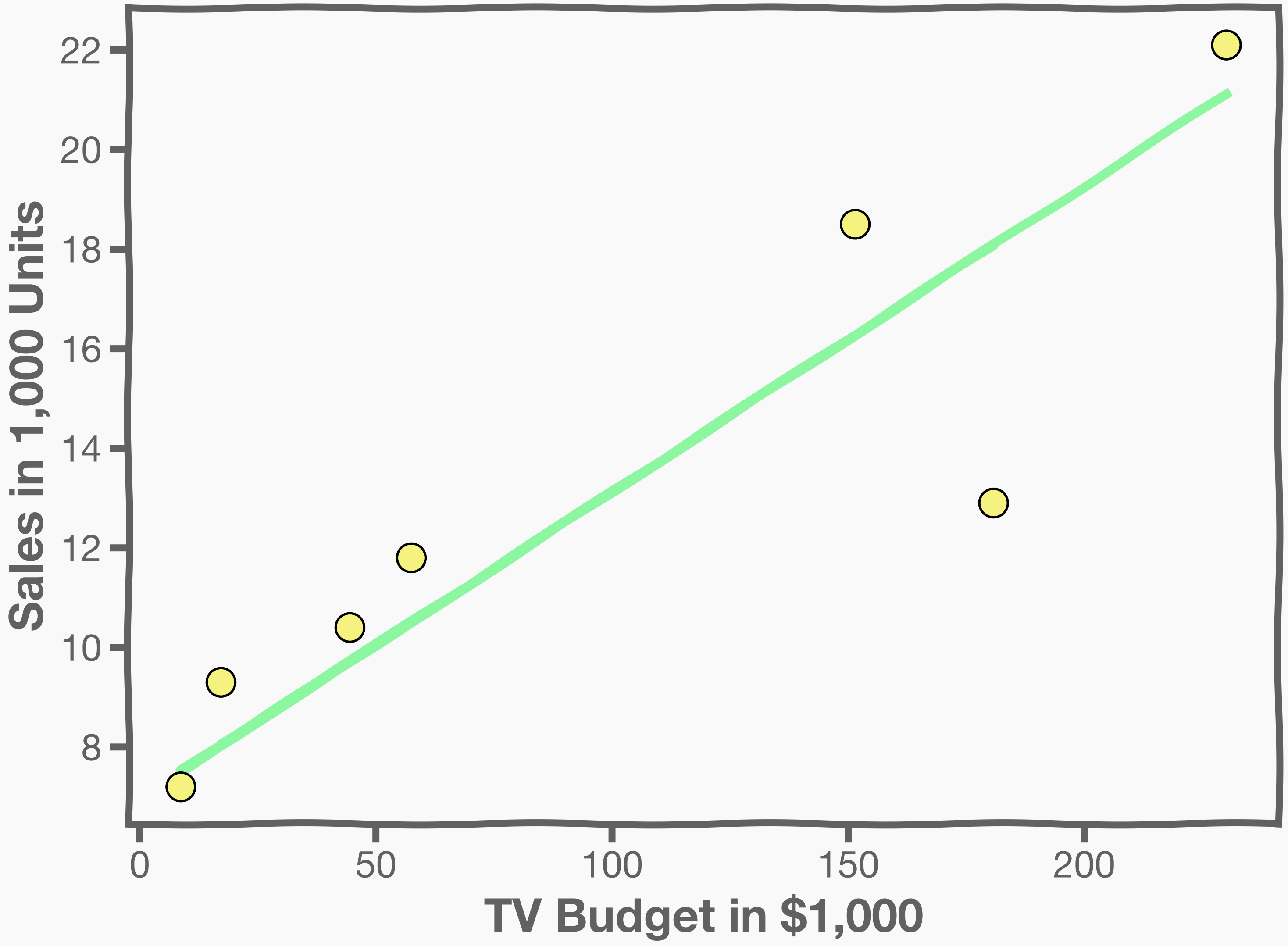
Estimate of the regression coefficients (cont)

Maybe this one?



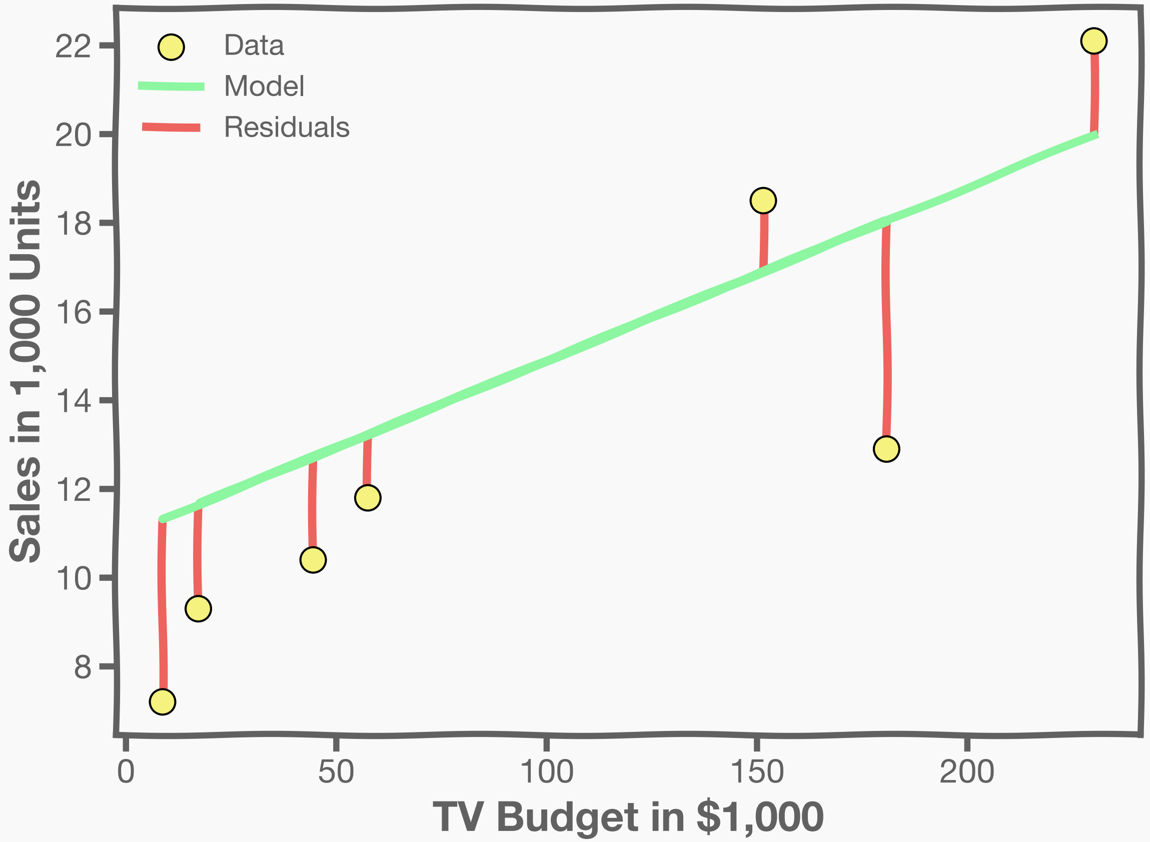
Estimate of the regression coefficients (cont)

Or this one?



Estimate of the regression coefficients (cont.)

Question: Which line is the best?



As before, for each observation (x_n, y_n), the absolute residuals, $r_i = |y_i - \hat{y}_i|$ quantify the error at each observation.

Estimate of the regression coefficients (cont.)

AGAIN, we use the MSE as our **loss function,**

$$L(\beta_0, \beta_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$$

We choose β_1 and β_0 that minimizes the predictive errors made by our model, i.e., minimize our loss function.

Then the optimal values, $\hat{\beta}_0$ and $\hat{\beta}_1$, should be:

$$\boxed{\widehat{\beta}_0, \widehat{\beta}_1 = \underset{\beta_0, \beta_1}{\operatorname{argmin}} L(\beta_0, \beta_1).}$$

Estimate of the regression coefficients (cont.)

AGAIN, we use the MSE as our **loss function,**

$$L(\beta_0, \beta_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$$

FIND THE VALUES OF β_0 AND β_1 THAT YIELD THE SMALLEST VALUE OF L

We choose β_1 and β_0 that minimizes the predictive errors made by our model, i.e., minimize our loss function.

Then the optimal values, $\hat{\beta}_0$ and $\hat{\beta}_1$, should be:

$$\widehat{\beta_0}, \widehat{\beta_1} = \underset{\beta_0, \beta_1}{\operatorname{argmin}} L(\beta_0, \beta_1).$$

WE CALL THIS **FITTING** OR **TRAINING** THE MODEL

Introducing...

SK-Learn

Import sklearn's linear model LinearRegression

```
>>> from sklearn.linear_model import LinearRegression
>>> df = pd.read_csv('Advertising.csv')
>>> X= df[['TV']].values
>>> y = df['Sales'].values
```

SK-Learn

```
>>> from sklearn.linear_model import LinearRegression
>>> df = pd.read_csv('Advertising.csv')
>>> X= df[['TV']].values
>>> y = df['Sales'].values
>>> reg = LinearRegression()
>>> reg.fit(X, y)
```

Instantiate the model

Use the method `fit()` from the model `LinearRegression`. This method finds the values of β_0 and β_1

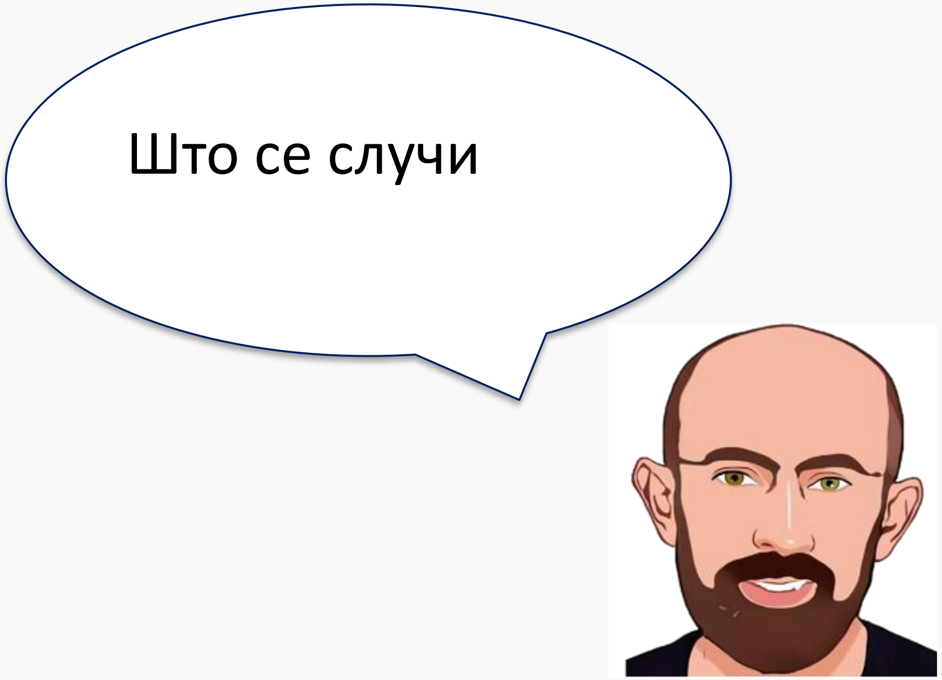
SK-Learn

```
>>> from sklearn.linear_model import LinearRegression
>>> df = pd.read_csv('Advertising.csv')
>>> X= df[['TV']].values
>>> y = df['Sales'].values
>>> reg = LinearRegression()
>>> reg.fit(X, y)
>>> reg.coef_
array([[0.04665056]])
>>> reg.intercept_
array([7.08543108])
>>> reg.predict(np.array([[100]]))
array([[11.75048733]])
```

Use the fitted model (i.e. uses the values of β_0 and β_1 found in the `.fit()` to predict y.

$$y = \beta_0 + \beta_1 x$$

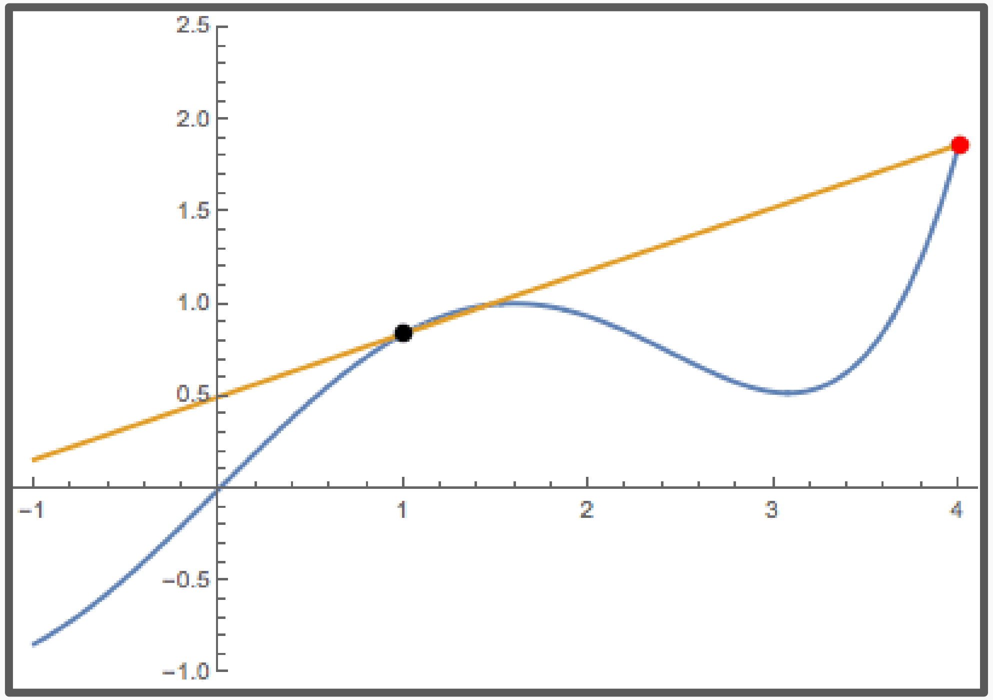
```
>>> reg.fit(X, y)
```



Derivative definition

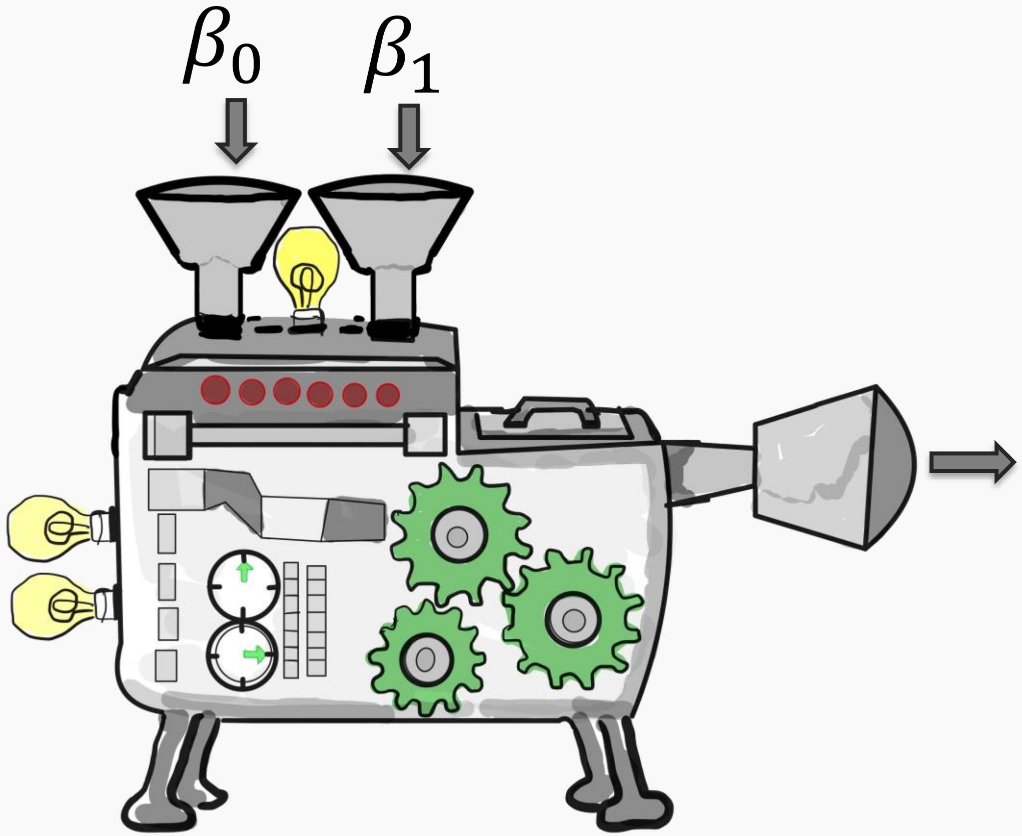
A derivative is the instantaneous rate of change of a single valued function. Given a function f(x) the derivative can be defined as:

$$f'(x) = \frac{df}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$



Partial derivatives

For a loss function L that depends on β_0, β_1 we need the partial derivatives, $\frac{\partial L}{\partial \beta_i}$. Partial derivatives indicate the rate of change of the function with respect to one variable while keeping the others fixed.



$L(\beta_0, \beta_1)$

$$L(\beta_0, \beta_1)$$

$$\frac{\partial L}{\partial \beta_0} \qquad \frac{\partial L}{\partial \beta_1}$$

Partial derivative example

If $L(\beta_0, \beta_1) = \big(y - (\beta_1 x + \beta_0)\big)^2$ then what is $\dfrac{\partial L}{\partial \beta_0}$?

Partial derivative example

If $L(\beta_0, \beta_1) = \left(y - (\beta_1 x + \beta_0)\right)^2$ then what is $\dfrac{\partial L}{\partial \beta_0}$?

$$\frac{\partial L(f(\beta_0))}{\partial \beta_0} = \frac{\partial L}{\partial f}\frac{\partial f}{\partial \beta_0}$$

Partial derivative $\dfrac{\partial L}{\partial \beta_0}$

If $L(\beta_0, \beta_1) = \left(y - (\beta_1 x + \beta_0)\right)^2$ then what is $\dfrac{\partial L}{\partial \beta_0}$?

$$L = \underbrace{(\underbrace{y - \beta_1 x - \beta_0}_{f(\beta_0)})^2}_{L = f^2}$$

$$\frac{\partial L}{\partial \beta_0} = \frac{\partial L}{\partial f}\frac{\partial f}{\partial \beta_0} \qquad L = f^2 \Rightarrow \frac{\partial L}{\partial f} = 2f \qquad f = y - \beta_1 x - \beta_0 \Rightarrow \frac{\partial f}{\partial \beta_0} = -1$$

$$\frac{\partial L}{\partial \beta_0} = \frac{\partial L}{\partial f}\frac{\partial f}{\partial \beta_0} = -2f = -2(y - \beta_1 x - \beta_0)$$

Partial derivative $\frac{\partial L}{\partial \beta_1}$

If $L(\beta_0, \beta_1) = \left(y - (\beta_1 x + \beta_0)\right)^2$ then what is $\frac{\partial L}{\partial \beta_1}$?

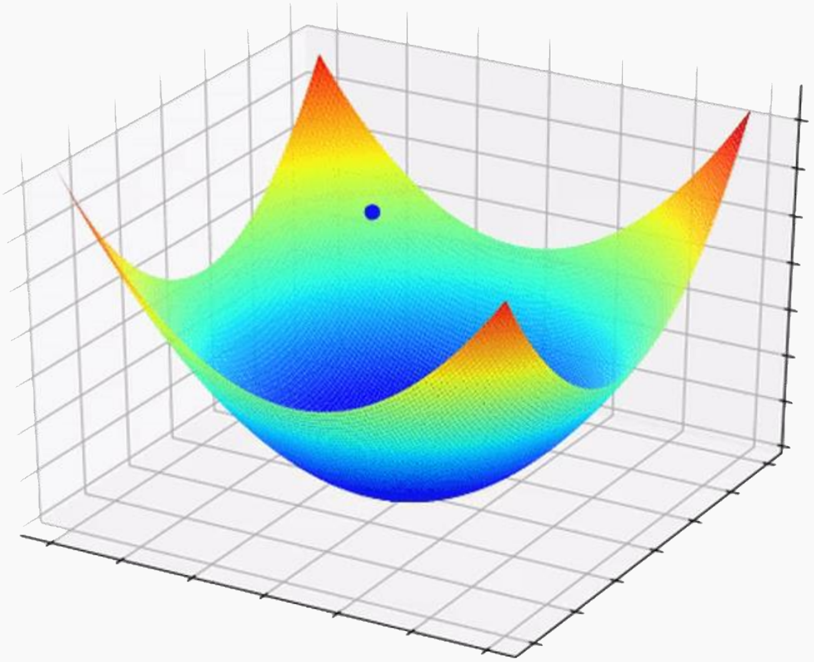
$$L = (y - \beta_1 x - \beta_0)^2$$

$$\frac{\partial L}{\partial \beta_1} = \frac{\partial L}{\partial f}\frac{\partial f}{\partial \beta_1} \qquad L = f^2 \Rightarrow \frac{\partial L}{\partial f} = 2f \qquad f = y - \beta_1 x - \beta_0 \Rightarrow \frac{\partial f}{\partial \beta_1} = -x$$

$$\frac{\partial L}{\partial \beta_1} = \frac{\partial L}{\partial f}\frac{\partial f}{\partial \beta_1} = -2xf = -2x(y - \beta_1 x - \beta_0)$$

Optimization

How does one minimize a loss function?



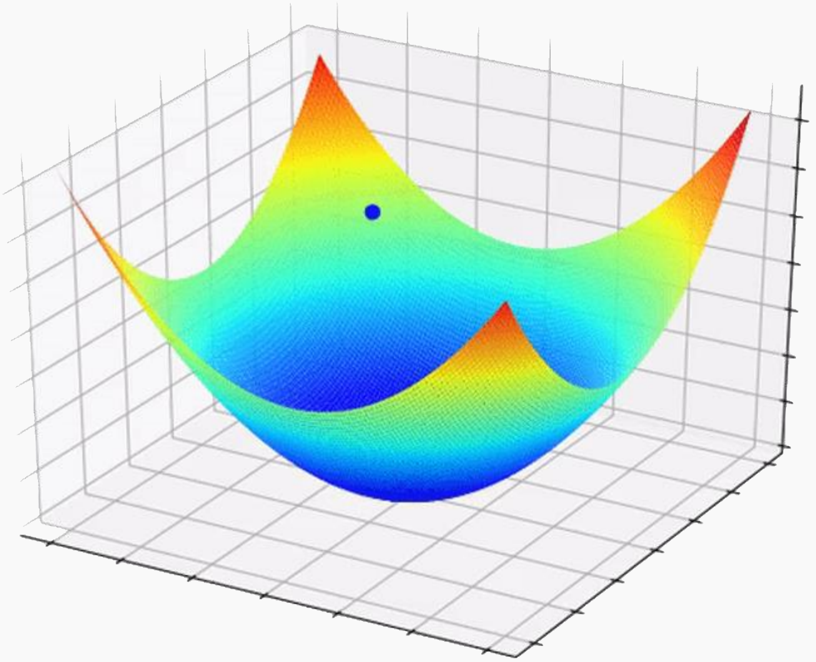
The global minima or maxima of $L(\beta_0, \beta_1)$ must occur at a point where the gradient (slope) is:

$$\nabla L = \left[\frac{\partial L}{\partial \beta_0}, \frac{\partial L}{\partial \beta_1}\right] = 0$$

- Brute Force: Try every combination
- Greedy Algorithm: Gradient Descent
- Closed-form Solution: Solve the above equation for β_0, β_1

Optimization

How does one minimize a loss function?



The global minima or maxima of $L(\beta_0, \beta_1)$ must occur at a point where the gradient (slope) is:

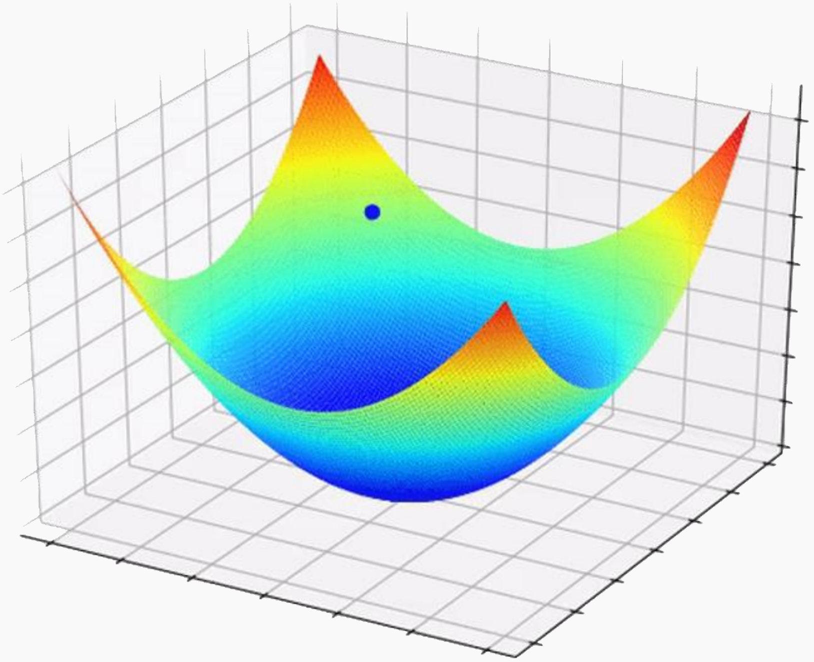
$$\nabla L = \left[\frac{\partial L}{\partial \beta_0}, \frac{\partial L}{\partial \beta_1}\right] = 0$$

- **Brute Force:** Try every combination
- **Greedy Algorithm:** Gradient Descent
- **Closed-form Solution: Solve the above equation for $\boldsymbol{\beta_0}, \boldsymbol{\beta_1}$**

Optimization

The gradient is a vector that contains all the partial derivatives of the function with respect to its variables. The nabla symbol (∇) is used to denote the gradient operation

How does one minimize a loss functi



The global minima axima of $L(\beta_0, \beta_1)$ must occur at a poi where the gradient (slope) is:

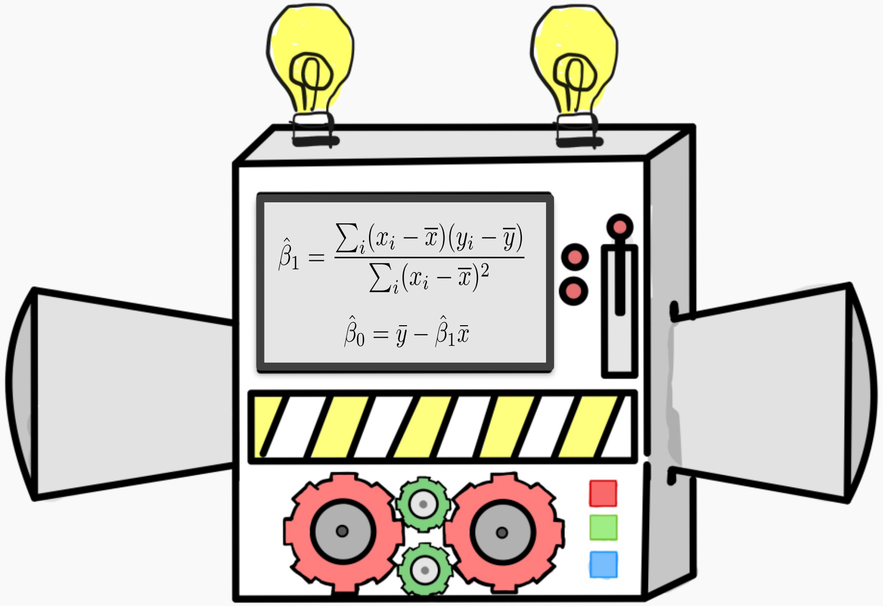
$$\nabla L = \left[\frac{\partial L}{\partial \beta_0}, \frac{\partial L}{\partial \beta_1}\right] = 0$$

- Brute Force: Try every combination
- Greedy Algorithm: Gradient Descent
- **Closed-form Solution: Solve the above equation for $\boldsymbol{\beta_0, \beta_1}$**

Optimization

$$\nabla L = \left[\frac{\partial L}{\partial \beta_0}, \frac{\partial L}{\partial \beta_1}\right] = 0$$

$$\frac{\partial L}{\partial \beta_0} = -2(y - \beta_1 x \ - \beta_0) = 0$$



$$\frac{\partial L}{\partial \beta_1} = -2x(y - \beta_1 x \ - \beta_0) = 0$$

Optimization

Sum over the data

Data: predictors values

Average value of x

Average value of y

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Summary: Estimate of the regression coefficients

We use MSE as our **loss function,**

$$L\left(\beta_0, \beta_1\right)=\frac{1}{n} \sum_{i=1}^n\left(y_i-\widehat{y_i}\right)^2=\frac{1}{n} \sum_{i=1}^n\left[y_i-\left(\beta_1 x_i+\beta_0\right)\right]^2$$

We choose $\hat{\beta}_1$and $\hat{\beta}_0$ in order to minimize the predictive errors made by our model, i.e. minimize our loss function.

Then the optimal values for $\hat{\beta}_0$ and $\hat{\beta}_1$ should be:

$$\widehat{\beta}_0, \widehat{\beta}_1=\underset{\beta_0, \beta_1}{\operatorname{argmin}} L(\beta_0, \beta_1).$$

Summary: Estimate of the regression coefficients

We use MSE as our **loss function,**

$$L(\beta_0, \beta_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}[$$

FIND THE VALUES OF β_0 AND β_1 THAT YIELD THE SMALLEST VALUE OF L

We choose $\hat{\beta}_1$and $\hat{\beta}_0$ in order to minimize the predictive errors made by our model, i.e. minimize our loss function.

Then the optimal values for $\hat{\beta}_0$ and $\hat{\beta}_1$ should be:

$$\widehat{\beta}_0, \widehat{\beta}_1 = \underset{\beta_0,\beta_1}{\operatorname{argmin}} L(\beta_0, \beta_1).$$

WE CALL THIS **FITTING** OR **TRAINING** THE MODEL

Estimate of the regression coefficients: analytical solution

Take the gradient of the loss function and find the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ where the gradient is zero: $\nabla L = \left[\frac{\partial L}{\partial \beta_0}, \frac{\partial L}{\partial \beta_1}\right]$=0

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where $\bar{y}$ and $\bar{x}$ are sample means.

The line:

is called the **regression line**.

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

Estimate of the regression coefficients: analytical solution

Take the gradient of the loss function and find t ere the gradient is zero: $\nabla L = \left[\frac{\partial L}{\partial \beta_0}, \frac{\partial L}{\partial \beta_1}\right]$=0

Finding the exact solution only works for rare cases. Linear regression is one of such rare cases.

$$\boxed{\begin{gathered}\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}\end{gathered}}$$

where $\bar{y}$ and $\bar{x}$ are sample means.

The line:

is called the **regression line**.

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