

Introduction to *Urban* Data Science

Machine Learning for Everyone

(EPA1316)

Lecture 9

Trivik Verma

THIS IS YOUR MACHINE LEARNING SYSTEM?

YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.

WHAT IF THE ANSWERS ARE WRONG?

JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.



Feedback

1. There has been no introduction to Brightspace for non TU students so they were a bit lost in the beginning.
2. Most students spend most time on Introduction to Data Science and the least on Understanding International Grand Challenges.
3. Online lectures: Students are happy with the quality of online lectures. Python course for TPM students has a really good methods for engagement during lectures using polls. Students encourage other professors to use these polls too, or to find other ways to interact with students.
4. Sometimes students find the lecture a bit too easy, examples are nice but are sometimes a bit too superficial, examples could go deeper.
5. Engagement: polls in python course by Julia work good, suggestion to do that. These are anonymous and low key. Questions like: 'did you understand this?', 'do you want another example?'

Feedback – Paper Discussions

1. There is a lot of literature to read for something that is not graded, so not much motivation and thus priority is given to this.
2. Friday paper discussions: Sometimes people end in breakout room with people who didn't read papers. It is not clear how much it will help with your final assignment. People have stopped reading papers more and more over the weeks due to the overall workload of the course/quarter.
3. Emphasize the need of the paper discussion. The physical group (who is following oral presentation at Wijnhaven right afterwards) is really enjoying it, you see the reason of the discussion. Difference between online and offline experience is very large. Online paper discussion is more awkward and people keep more quiet.
4. The discussion now feels like a side option next to the course.

Feedback – Labs+Assignments

Final assignment

1. Some students would like to be able to know already what the final assignments will look like, they would like being able to look ahead. Some students also don't really like this idea since it would be overwhelming. Mixed opinions.

Alignment

1. Labs are not very aligned with theory from that week sometimes. Students say they are learning theory from the labs instead of from the lectures.
2. Monday and Friday for data science instead of Wednesday and Friday would give more time for deadlines.

Homework + labs is a lot. Sometimes redundant. Homework (even if voluntary) gives some students pressure since they want to do all.

Last Time

- Machine Learning
- Predicting a Variable
- Error evaluation
- Model comparison
- Fitness of models

Today

- Linear models
- Estimate of the regression coefficients
- Model evaluation
- Interpretation

Linear Models

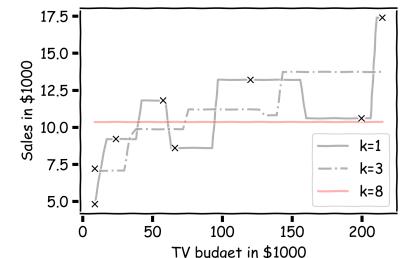
Note that in building our kNN model for prediction, we did not compute a closed form for \hat{f} .

What if we ask the question:

“how much more sales do we expect if we double the TV advertising budget?”

Alternatively, we can build a model by first assuming a simple form of f :

$$f(x) = \beta_0 + \beta_1 X$$



Linear Regression

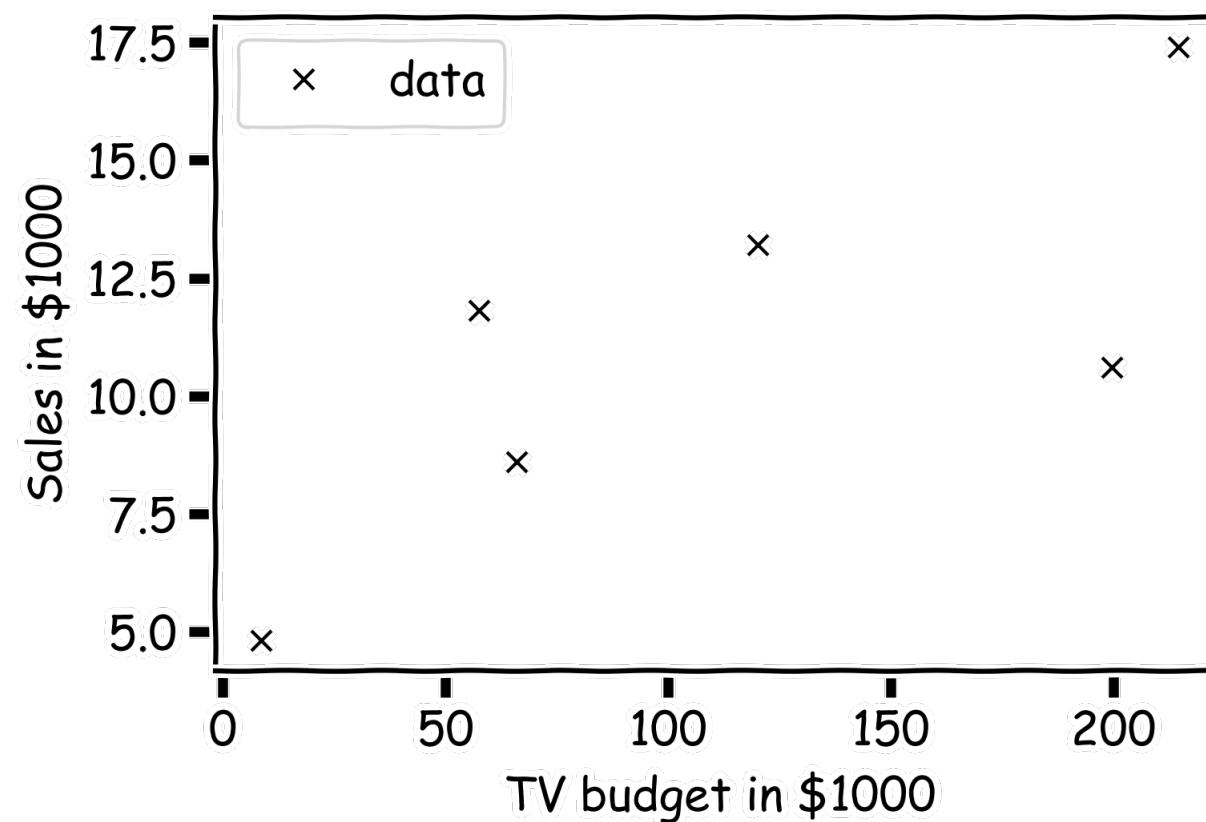
... then it follows that our estimate is:

$$\hat{Y} = \hat{f}(X) = \hat{\beta}_1 X + \hat{\beta}_0$$

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

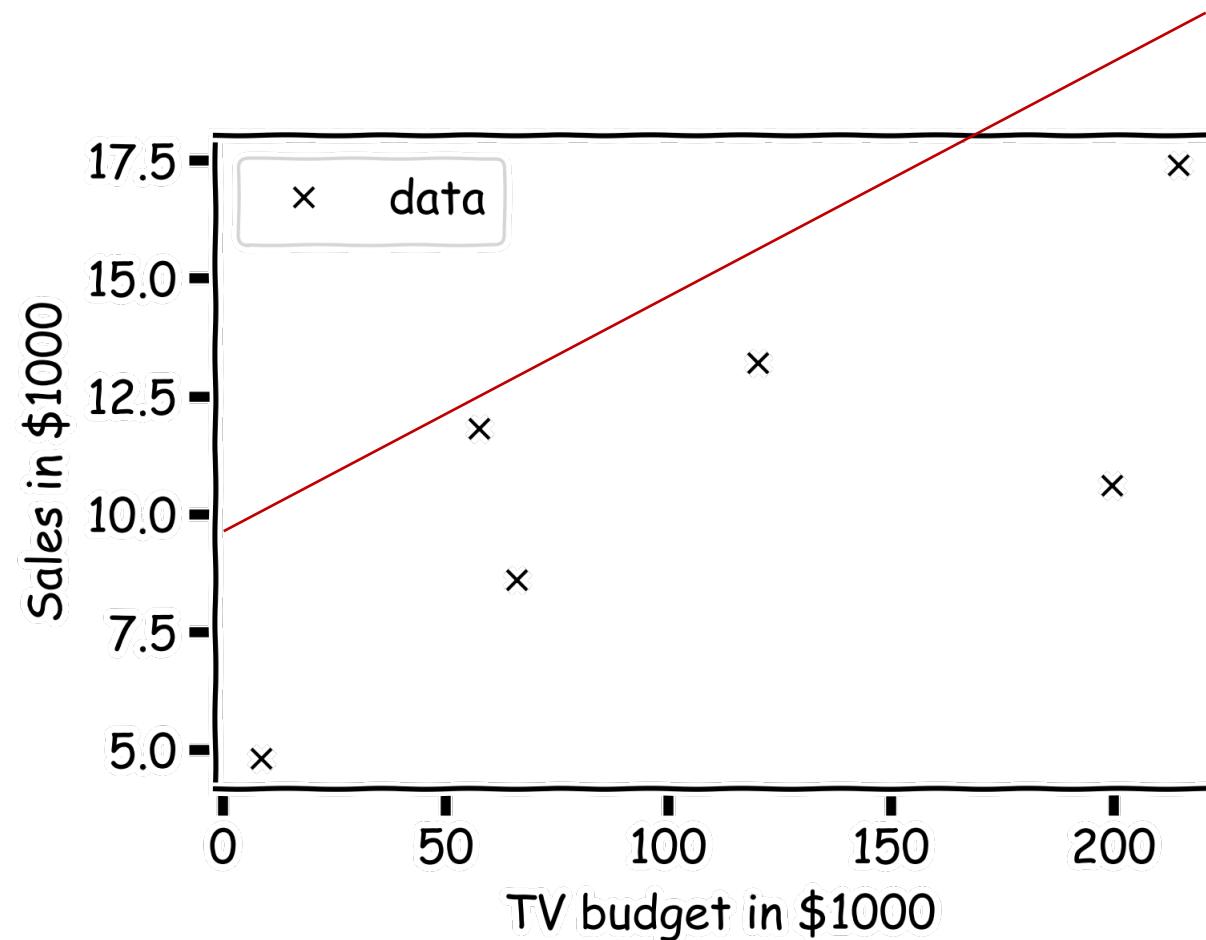
Estimate of the regression coefficients

For a given data set



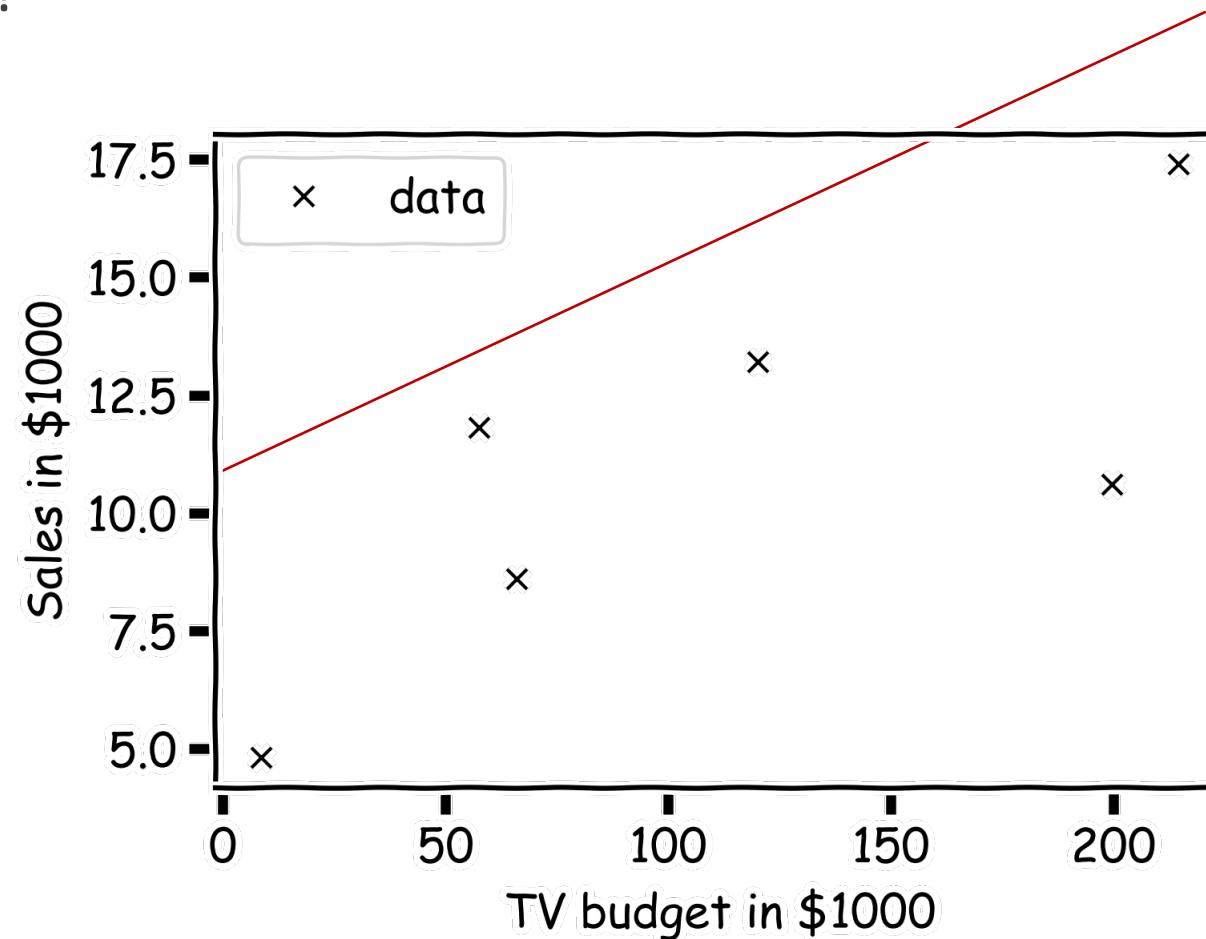
Estimate of the regression coefficients (cont)

Is this line good?



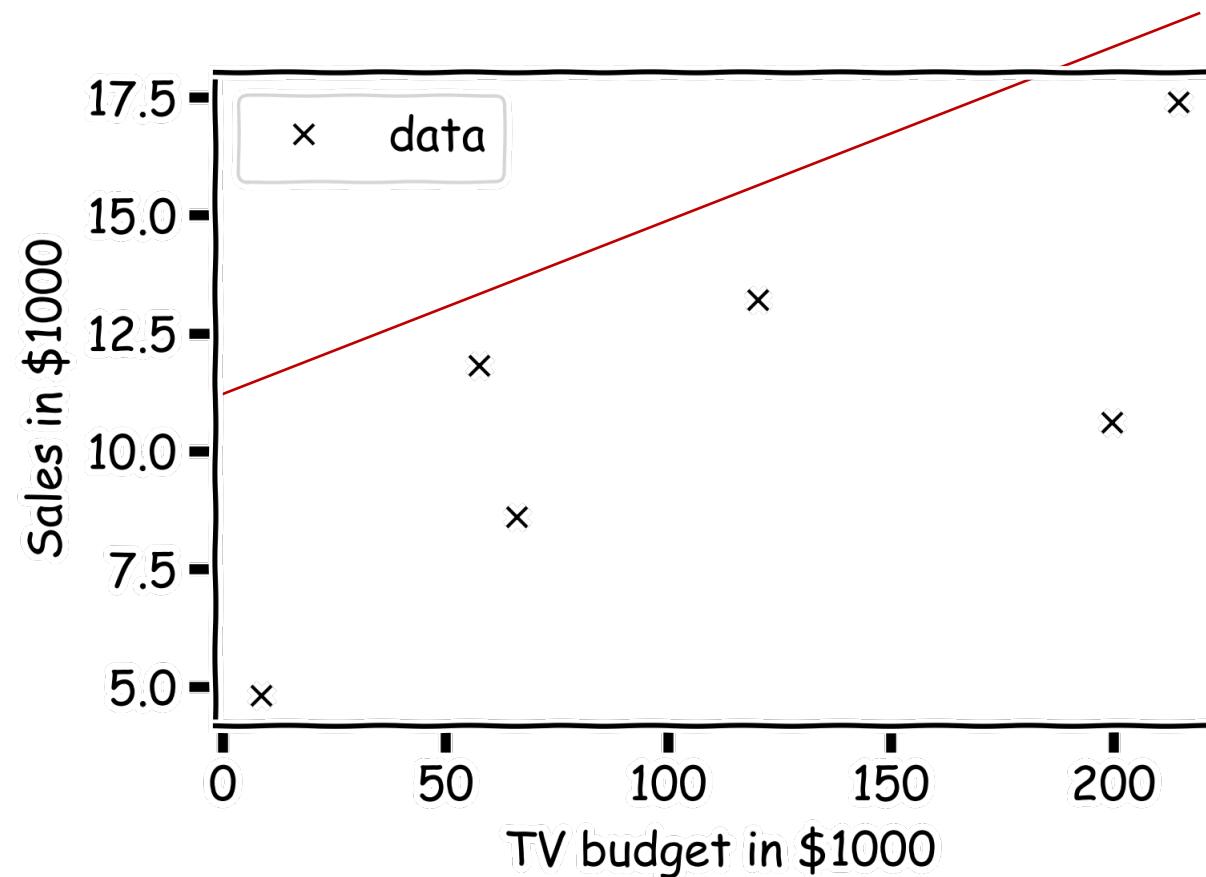
Estimate of the regression coefficients (cont)

Maybe this one?



Estimate of the regression coefficients (cont)

Or this one?

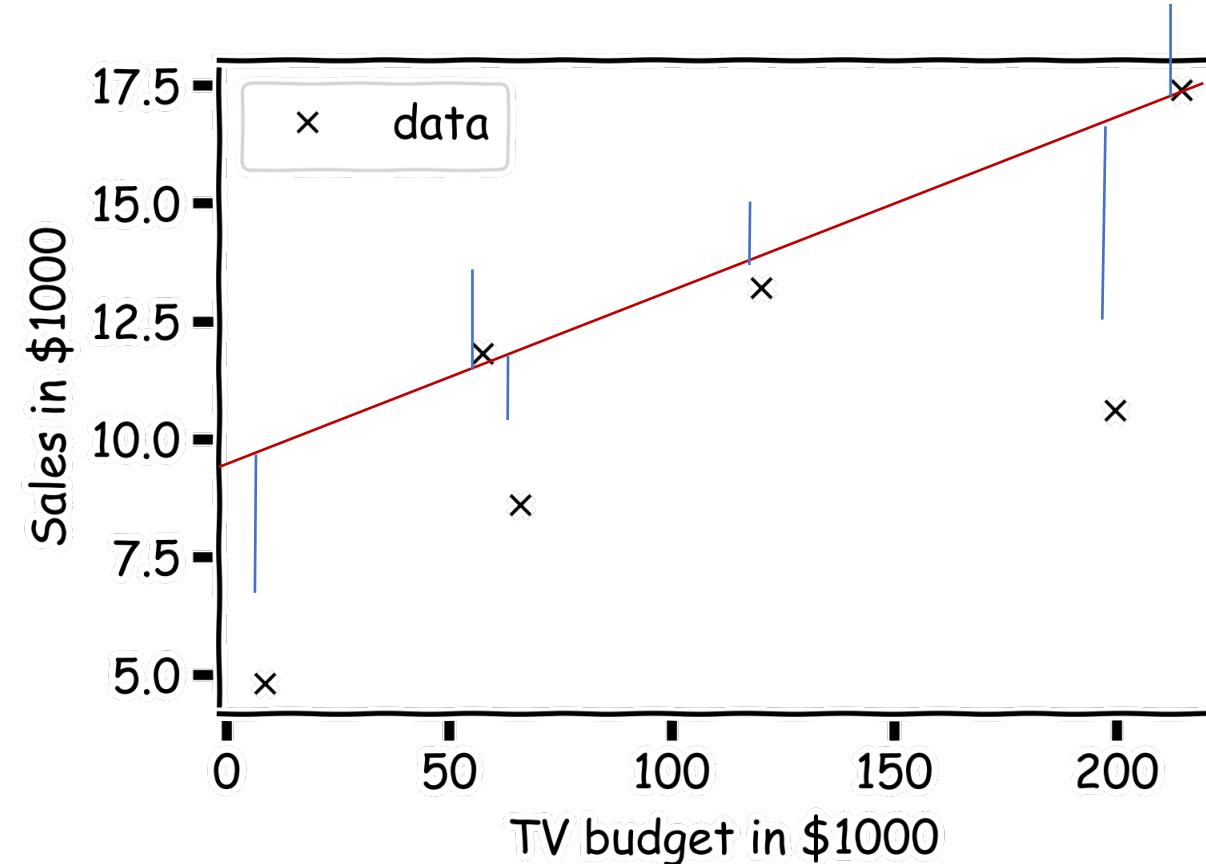


Estimate of the regression coefficients (cont)

Question: Which line is the best?

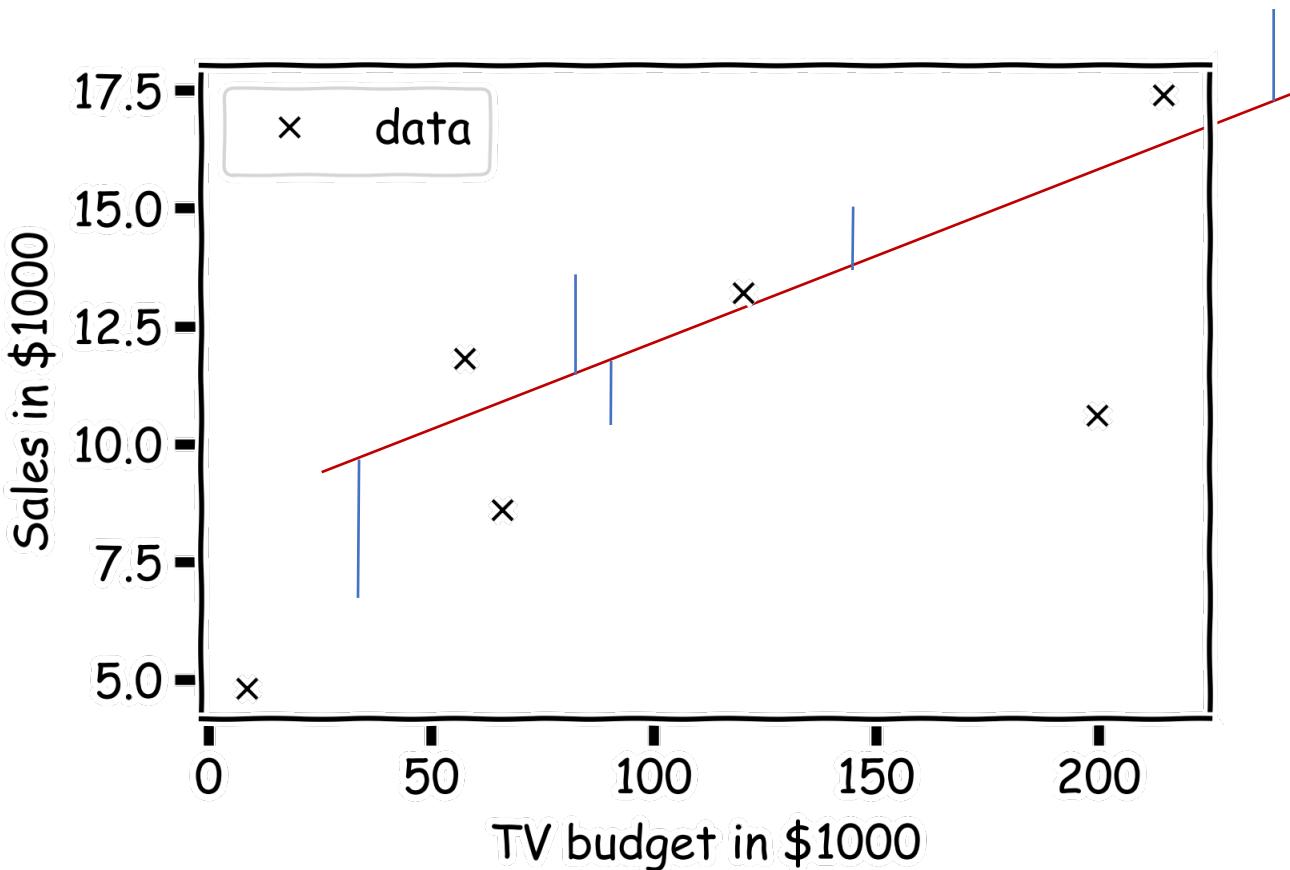
For each observation (x_n, y_n) , the **absolute residual** is calculating the residuals

$$r_i = |y_i - \hat{y}_i|.$$



Loss Function: Aggregate Residuals

How do we aggregate residuals across the entire dataset?



1. Max Absolute Error
2. Mean Absolute Error
3. Mean Squared Error

Estimate of the regression coefficients (cont)

- Again we use MSE as our **loss function**,

$$L(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^n [y_i - (\beta_1 X + \beta_0)]^2.$$

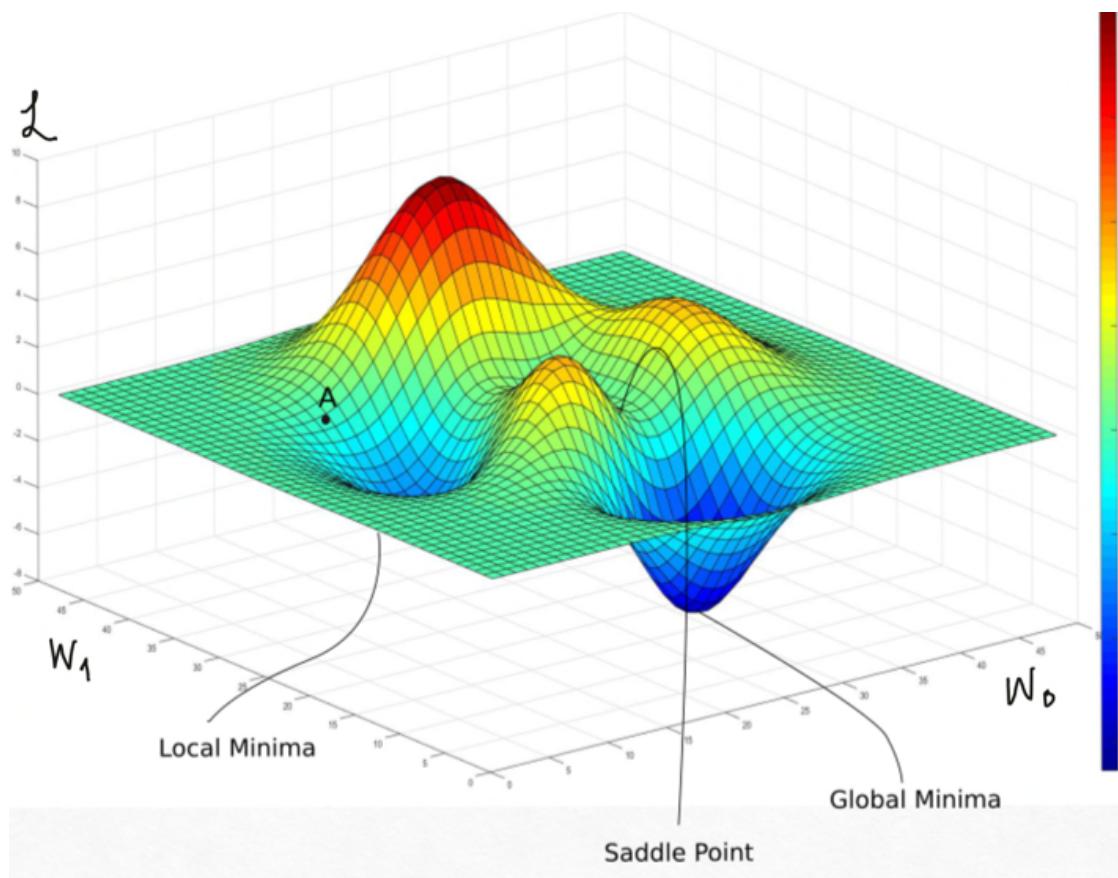
- We choose $\hat{\beta}_1$ and $\hat{\beta}_0$ in order to minimize the predictive errors made by our model, i.e. minimize our loss function.
- Then the optimal values for $\hat{\beta}_0$ and $\hat{\beta}_1$ should be:

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

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

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)

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

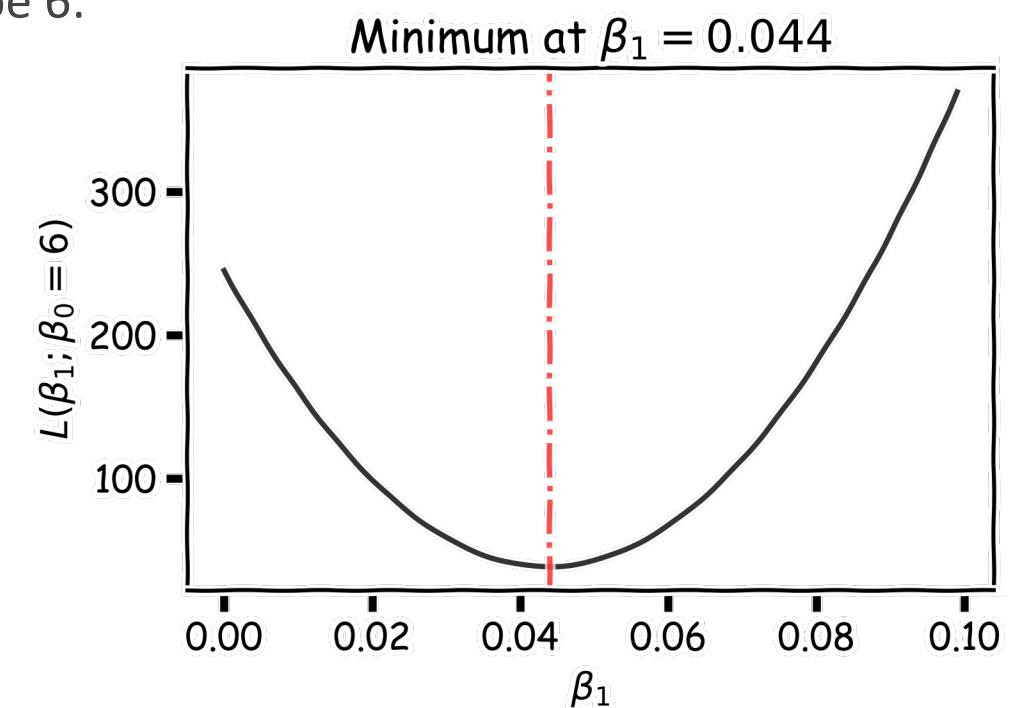
- **Brute Force:** Try every combination
- **Exact:** Solve the above equation
- **Greedy Algorithm:** Gradient Descent

Optimization: Estimate of the regression coefficients

Brute force

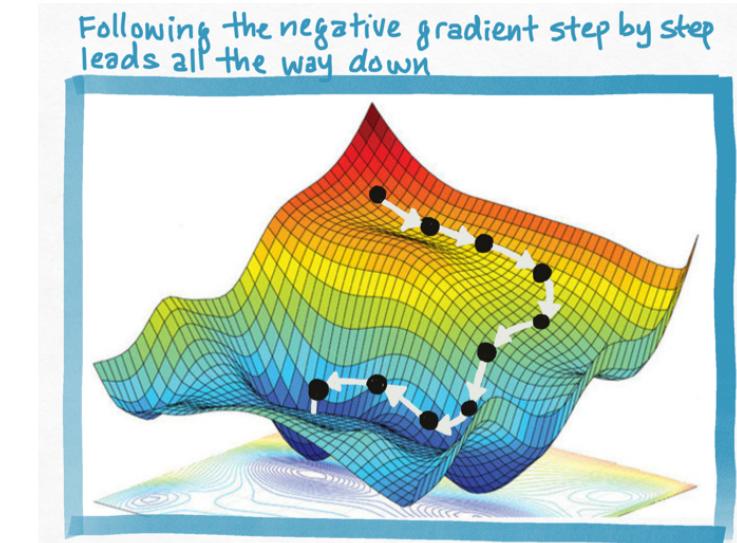
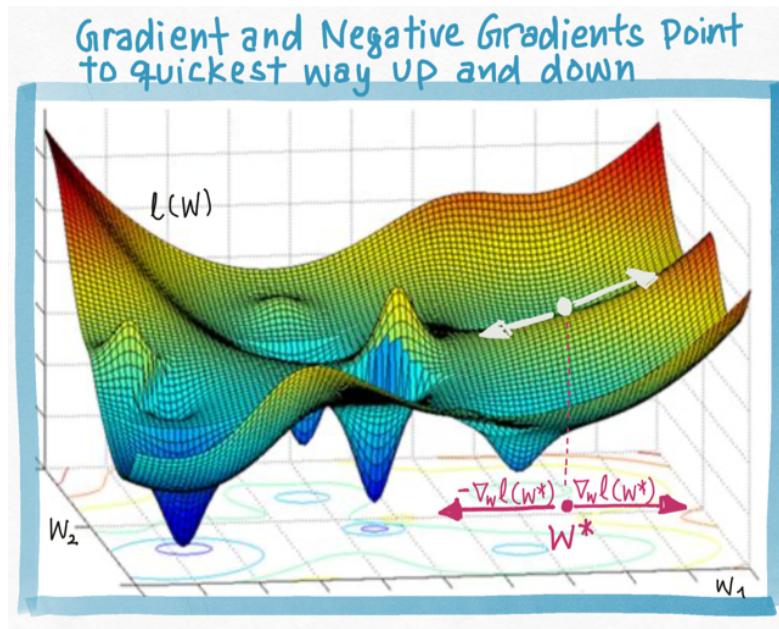
- A way to estimate $\operatorname{argmin}_{\beta_0, \beta_1} L$ is to calculate the loss function for every possible β_0 and β_1 . Then select the β_0 and β_1 where the loss function is minimum.
- E.g. the loss function for different β_1 when β_0 is fixed to be 6:

Very computationally expensive with many coefficients



Gradient Descent

- When we can't analytically solve for the stationary points of the gradient, we can still exploit the information in the gradient.
- The gradient ∇L at any point is the **direction of the steepest increase**. The negative gradient is the **direction of steepest decrease**.
- By following the -ve gradient, we can eventually find the lowest point.
- This method is called **Gradient Descent**.



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

This does not usually yield to a close form solution. However, **for linear regression** this procedure gives us explicit formulae for $\hat{\beta}_0$ and $\hat{\beta}_1$:

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

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

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

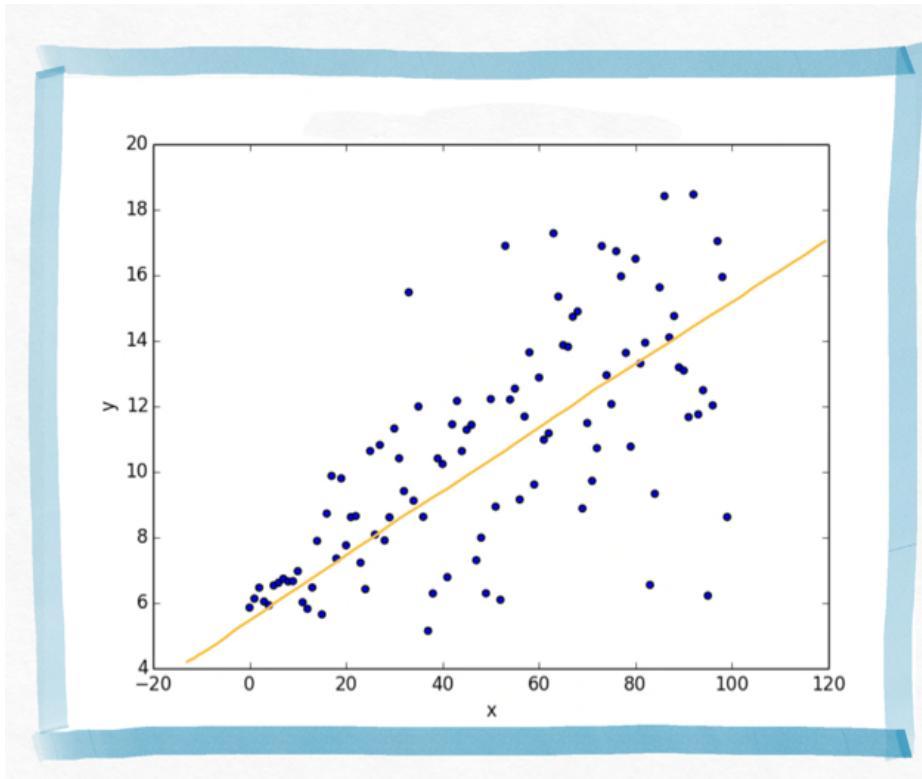
The line:

$$\hat{Y} = \hat{\beta}_1 X + \hat{\beta}_0$$

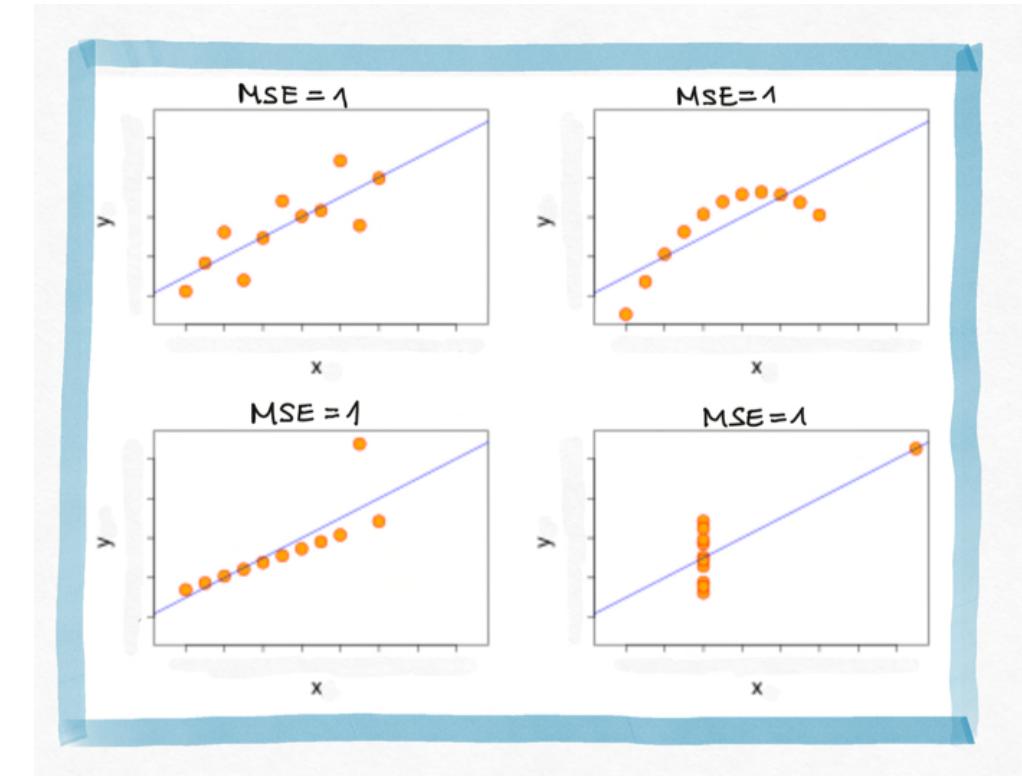
is called the **regression line**.

Evaluation: Training Error

Just because we found the model that minimizes the squared error it doesn't mean that it's a good model. We investigate the R² but also:



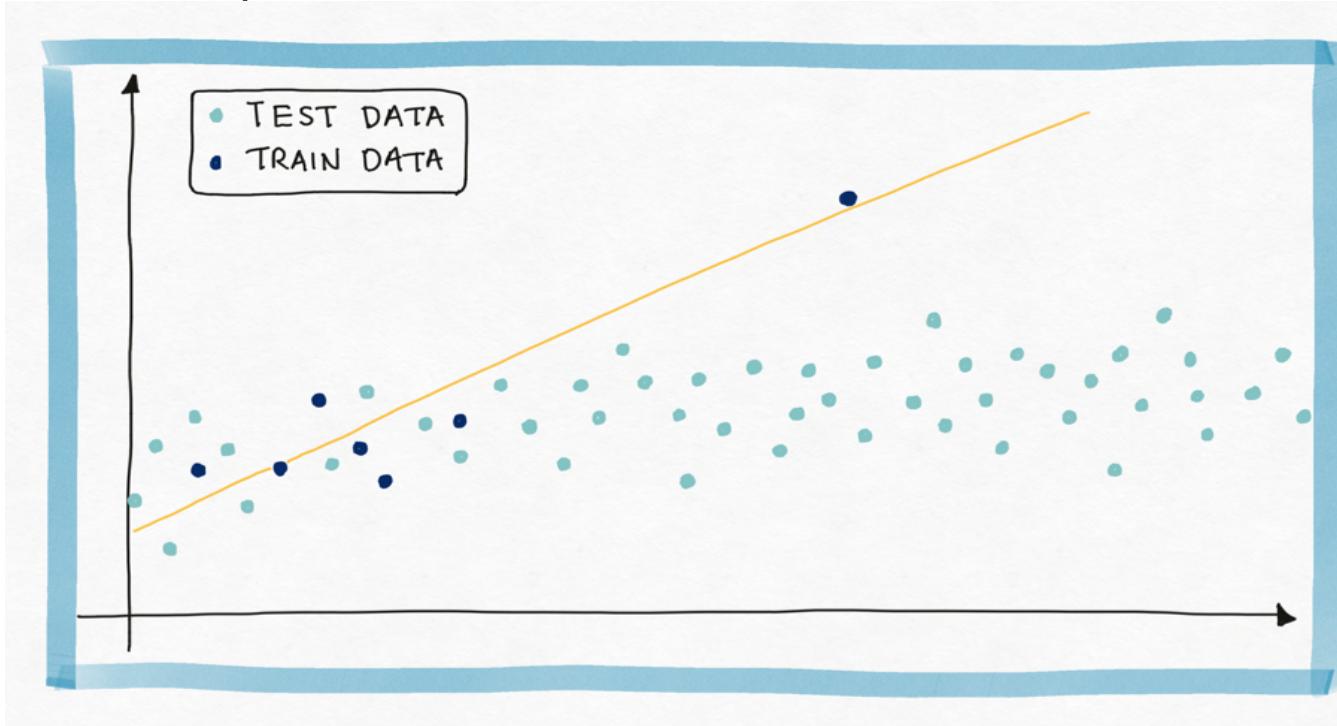
The MSE is high due to noise in the data.



The MSE is high in all four models but the models are not equal.

Evaluation: Test Error

We need to evaluate the fitted model on new data, data that the model did not train on, the **test data**.



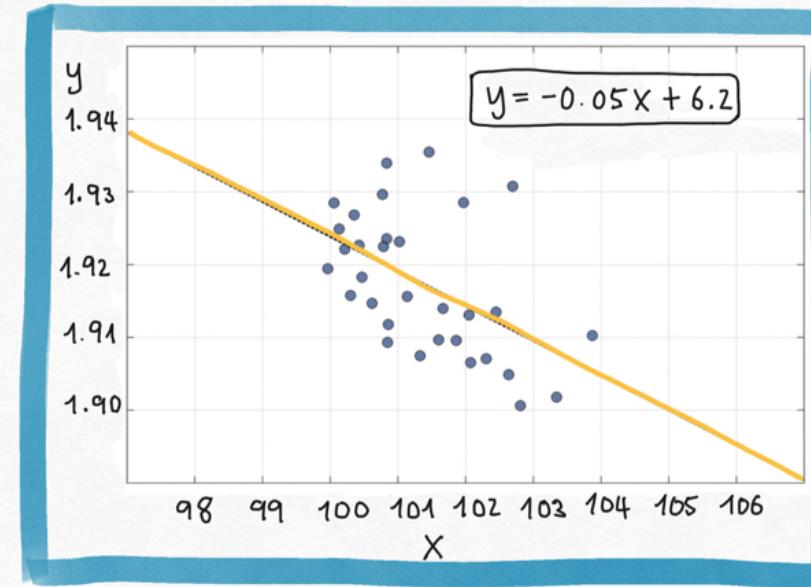
The **training** MSE here is 2.0 where the **test** MSE is 12.3.

The training data contains a strange point – **an outlier** – which confuses the model.

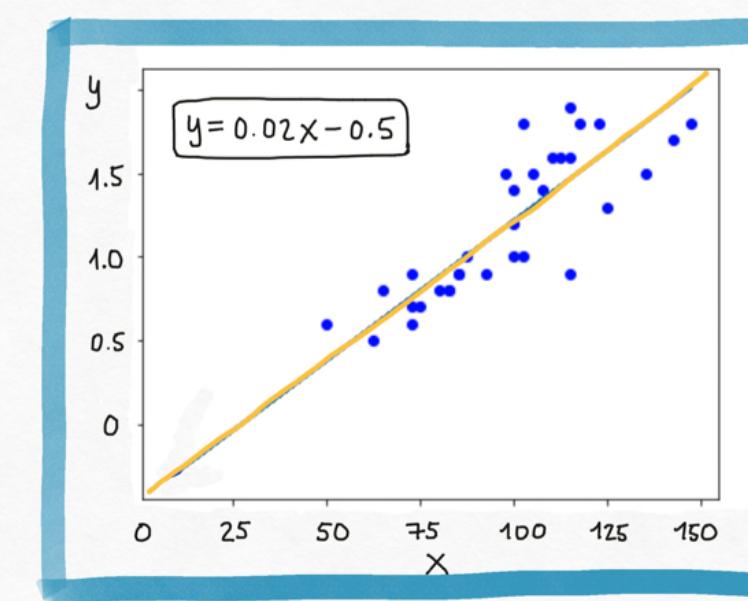
Fitting to meaningless patterns in the training is called **overfitting**.

Evaluation: Model Interpretation

For linear models it's important to interpret the parameters



The MSE of this model is very small. But the slope is -0.05. That means the larger the budget the less the sales.



The MSE is very small but the intercept is -0.5 which means that for very small budget we will have negative sales.

Break



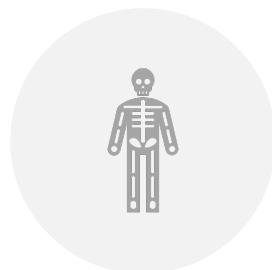
WATER



WALK



COFFEE OR TEA



MAKE FRIENDS

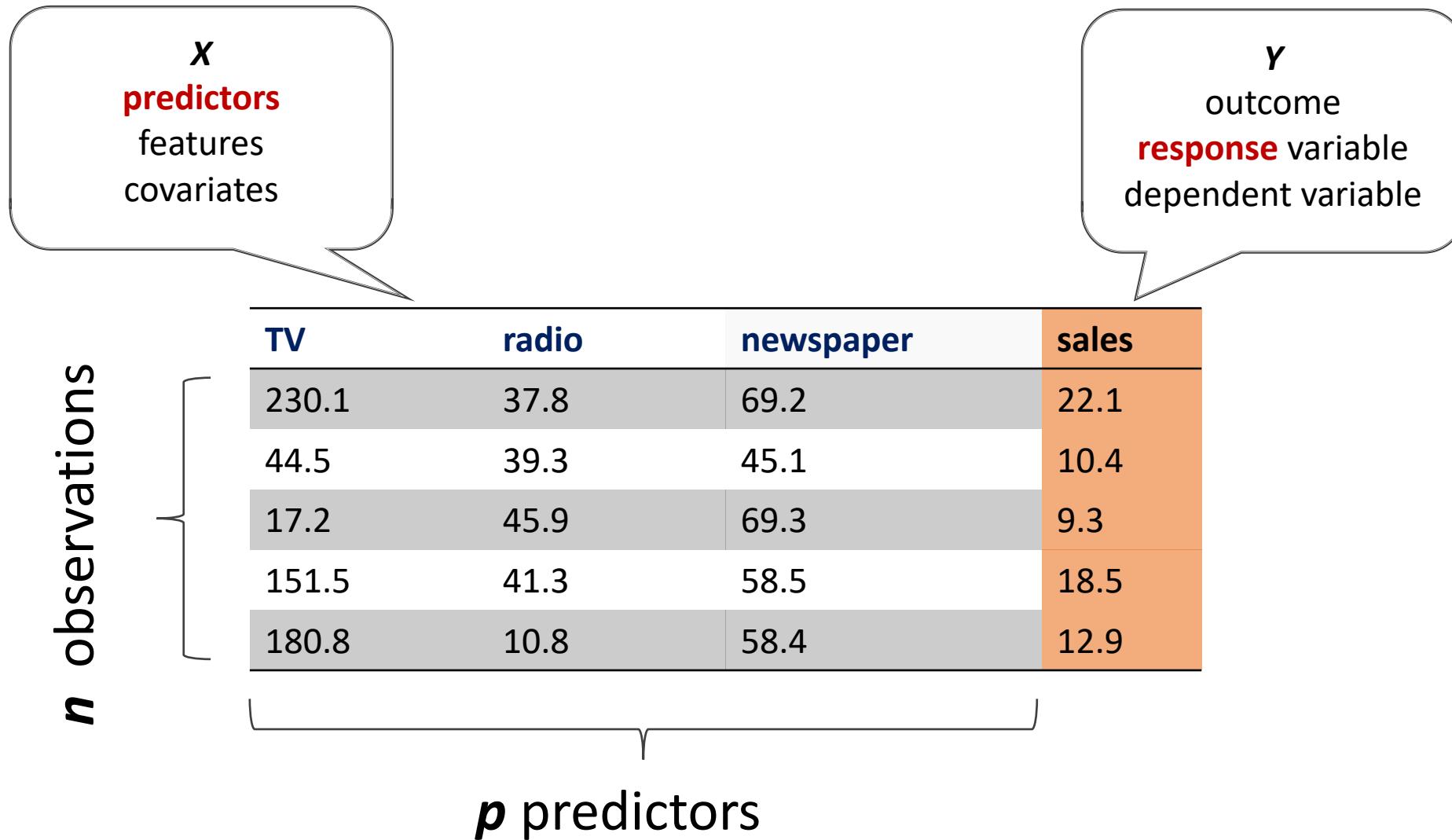
Multiple Linear Regression

If you must guess someone's height, would you rather be told

- Their weight, only
- Their weight and gender
- Their weight, gender, and income
- Their weight, gender, income, and favorite number

Of course, you'd always want as much data about a person as possible. Even though height and favorite number may not be strongly related, at worst you could just ignore the information on favorite number. We want our models to be able to take in lots of data as they make their predictions.

Response vs. Predictor Variables



X
predictors
features
covariates

Y
outcome
response variable
dependent variable

	TV	radio	newspaper	sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9

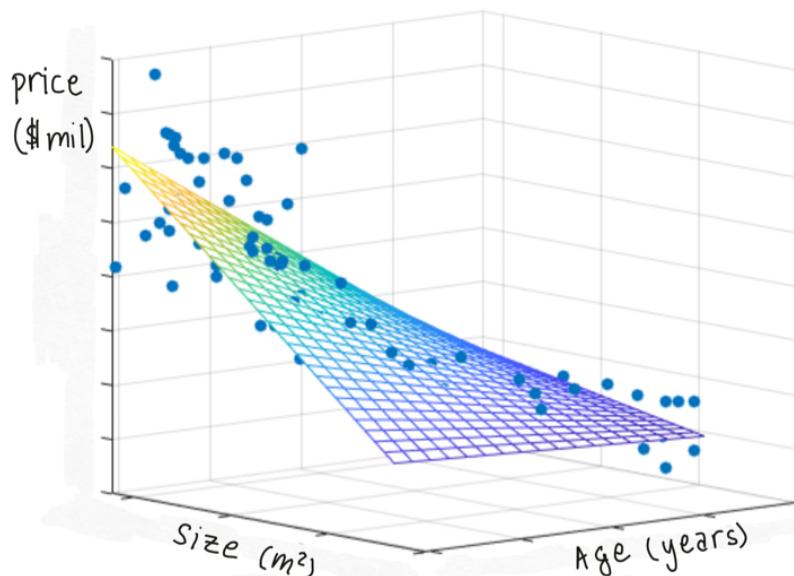
n observations

p predictors

Multilinear Models

In practice, it is unlikely that any response variable Y depends solely on one predictor x . Rather, we expect that Y is a function of multiple predictors $f(X_1, \dots, X_J)$. Using the notation we introduced last lecture,

$$Y = y_1, \dots, y_n, \quad X = X_1, \dots, X_J \quad \text{and} \quad X_j = x_{1j}, \dots, x_{ij}, \dots, x_{nj},$$



we can still assume a simple form for f -a multilinear form:

$$f(X_1, \dots, X_J) = \beta_0 + \beta_1 X_1 + \dots + \beta_J X_J$$

Hence, \hat{f} , has the form:

$$\hat{f}(X_1, \dots, X_J) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_J X_J$$

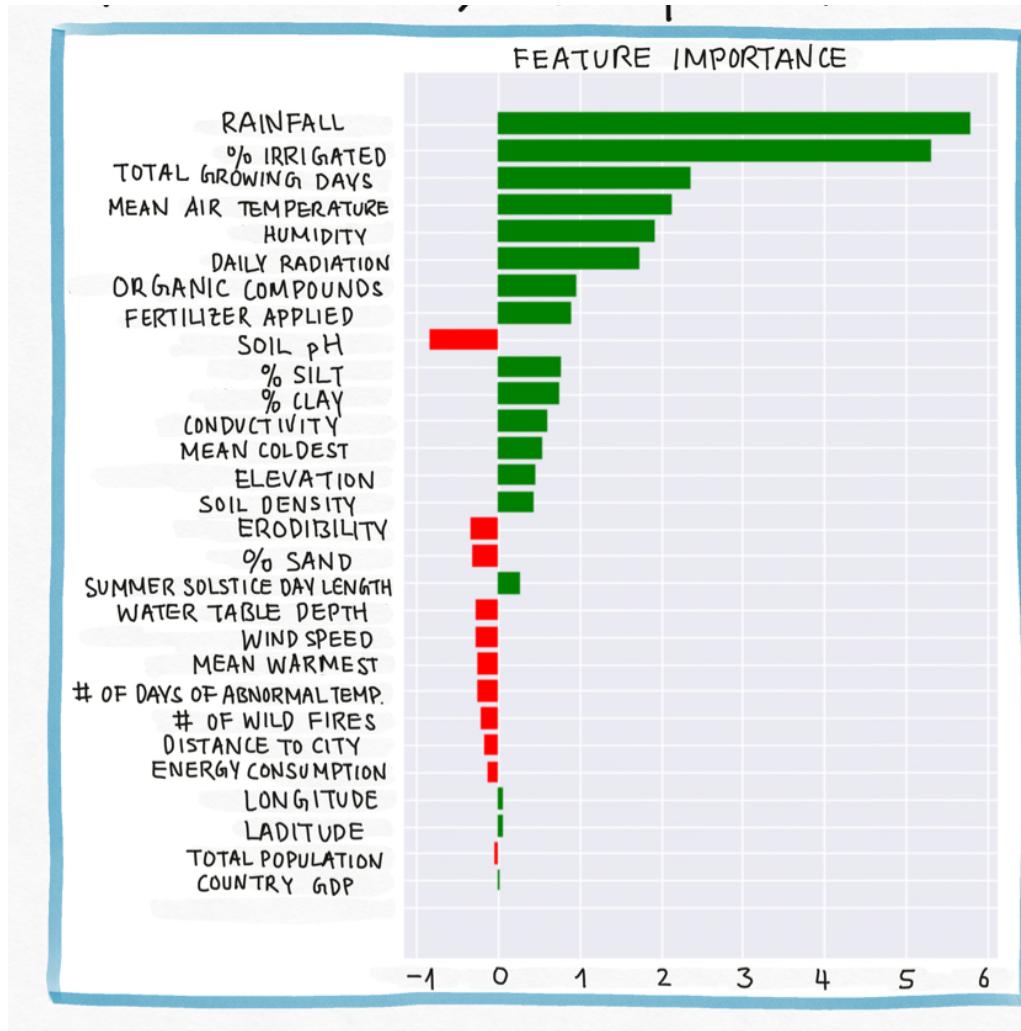
Multilinear Model, example

For our data

$$\text{Sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{Radio} + \beta_3 \times \text{Newspaper}$$

Interpreting multi-linear regression

For linear models, it is easy to interpret the model parameters.



When we have many predictors: X_1, \dots, X_J , there will be many model parameters, $\beta_1, \beta_2, \dots, \beta_J$.

Looking at the values of β 's is impractical, so we visualize these values in a [feature importance](#) graph.

The feature importance graph shows which predictors has the most impact on the model's prediction.

Qualitative Predictors

So far, we have assumed that all variables are quantitative. But in practice, often some predictors are **qualitative**.

Example: The *credit data set* contains information about balance, age, cards, education, income, limit , and rating for several potential customers.

Income	Limit	Rating	Cards	Age	Education	Gender	Student	Married	Ethnicity	Balance
14.890	3606	283	2	34	11	Male	No	Yes	Caucasian	333
106.02	6645	483	3	82	15	Female	Yes	Yes	Asian	903
104.59	7075	514	4	71	11	Male	No	No	Asian	580
148.92	9504	681	3	36	11	Female	No	No	Asian	964
55.882	4897	357	2	68	16	Male	No	Yes	Caucasian	331

Qualitative Predictors

- If the predictor takes only two values, then we create an **indicator** or **dummy variable** that takes on two possible numerical values.
- For example for the gender, we create a new variable:

$$x_i = \begin{cases} 1 & \text{if } i \text{ th person is female} \\ 0 & \text{if } i \text{ th person is male} \end{cases}$$

- We then use this variable as a predictor in the regression equation.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i \text{ th person is female} \\ \beta_0 + \epsilon_i & \text{if } i \text{ th person is male} \end{cases}$$

Qualitative Predictors

Question: What is interpretation of β_0 and β_1 ?

Qualitative Predictors

Question: What is interpretation of β_0 and β_1 ?

- β_0 is the **average** credit card balance among **males**,
- $\beta_0 + \beta_1$ is the **average** credit card balance among **females**,
- and β_1 the average **difference** in credit card balance between **females** and **males**.

Example: Calculate β_0 and β_1 for the Credit data.

You should find $\beta_0 \sim \$509$, $\beta_1 \sim \$19$

More than two levels: One hot encoding

- Often, the qualitative predictor takes more than two values (e.g. ethnicity in the credit data).
- In this situation, a single dummy variable cannot represent all possible values.
- We create **additional** dummy variable as:

$$x_{i,1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian} \end{cases}$$

$$x_{i,2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian} \end{cases}$$

More than two levels: One hot encoding

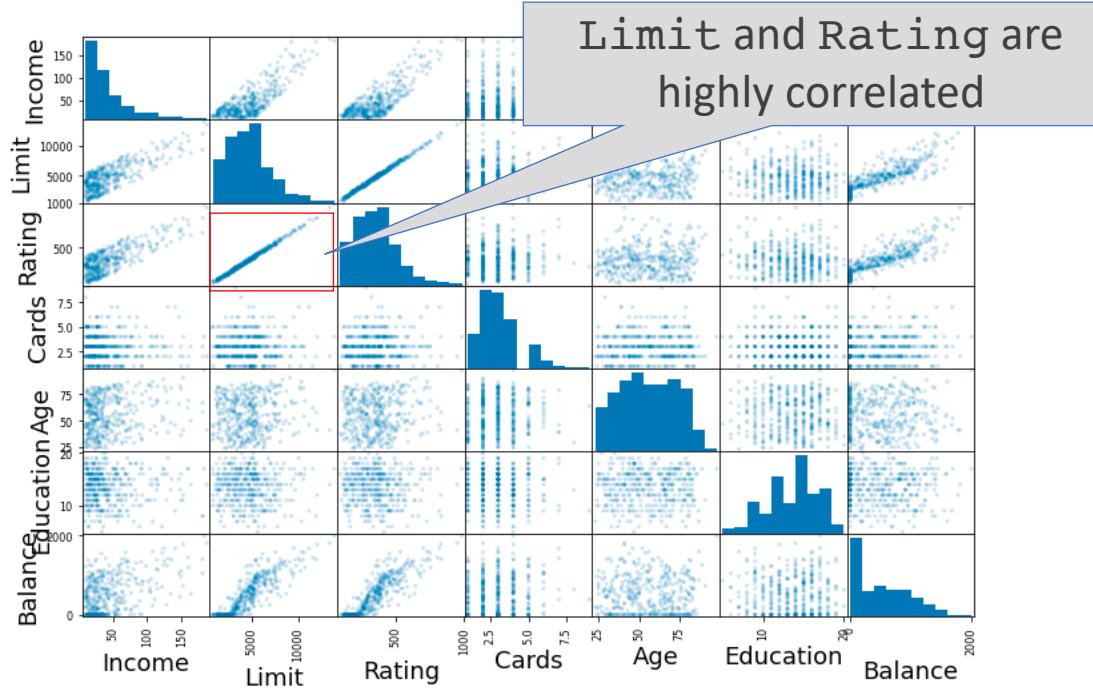
We then use these variables as predictors, the regression equation becomes:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is AfricanAmerican} \end{cases}$$

Question: What is the interpretation of $\beta_0, \beta_1, \beta_2$?

Collinearity

Collinearity and **multicollinearity** refers to the case in which two or more predictors are correlated (related).



	Columns	Coefficients
0	Income	-7.802001
1	Limit	0.193077
2	Rating	1.102269
3	Cards	17.923274
4	Age	-0.634677
5	Education	-1.115028
6	Gender	10.406651
7	Student	426.469192
8	Married	-7.019100

	Columns	Coefficients
0	Income	-7.770915
1	Rating	3.976119
2	Cards	4.031215
3	Age	-0.669308
4	Education	-0.375954
5	Gender	10.368840
6	Student	417.417484
7	Married	-13.265344

The regression coefficients are not uniquely determined. In turn it hurts the **interpretability** of the model as then the regression coefficients are **not unique** and have influences from other features.

Both **limit** and **rating** have positive coefficients, but it is hard to understand if the balance is higher because of the rating or is it because of the limit? If we remove limit then we achieve almost the same model performance but the coefficients change.

Beyond linearity

So far we assumed:

- linear relationship between X and Y
- the residuals $r_i = y_i - \hat{y}_i$ were uncorrelated (taking the average of the square residuals to calculate the MSE implicitly assumed uncorrelated residuals).

These assumptions need to be verified using the data and **visually inspecting the residuals.**

Residual Analysis

If the correct model is **not linear** then,

$$y = \beta_0 + \beta_1 x + \phi(x) + \epsilon$$

our model assuming linear relationship is:

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

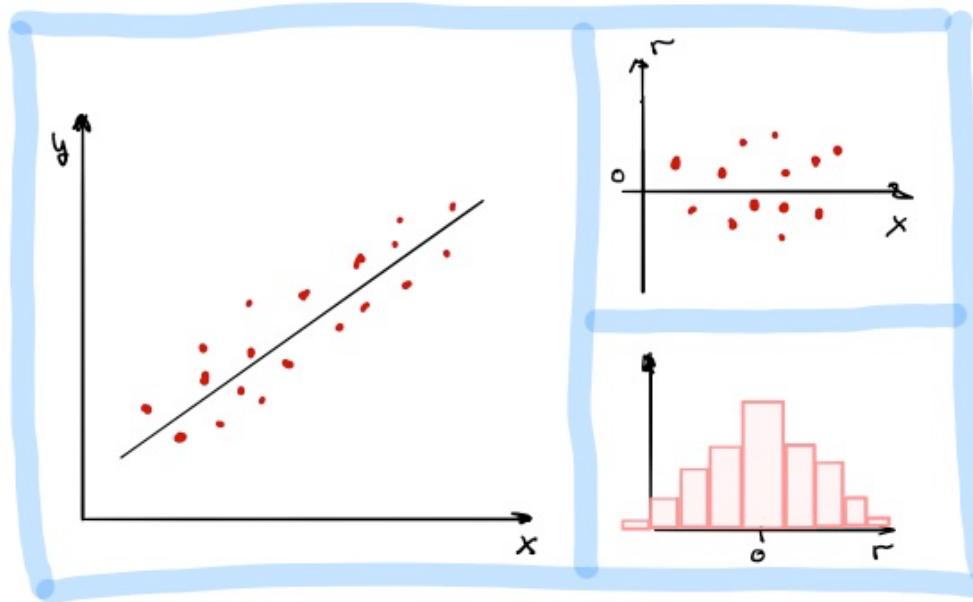
Then the residuals, $r = (y - \hat{y}) = \epsilon + \phi(x)$, are **not independent** of x

Residual Analysis

In residual analysis, we typically create two types of plots:

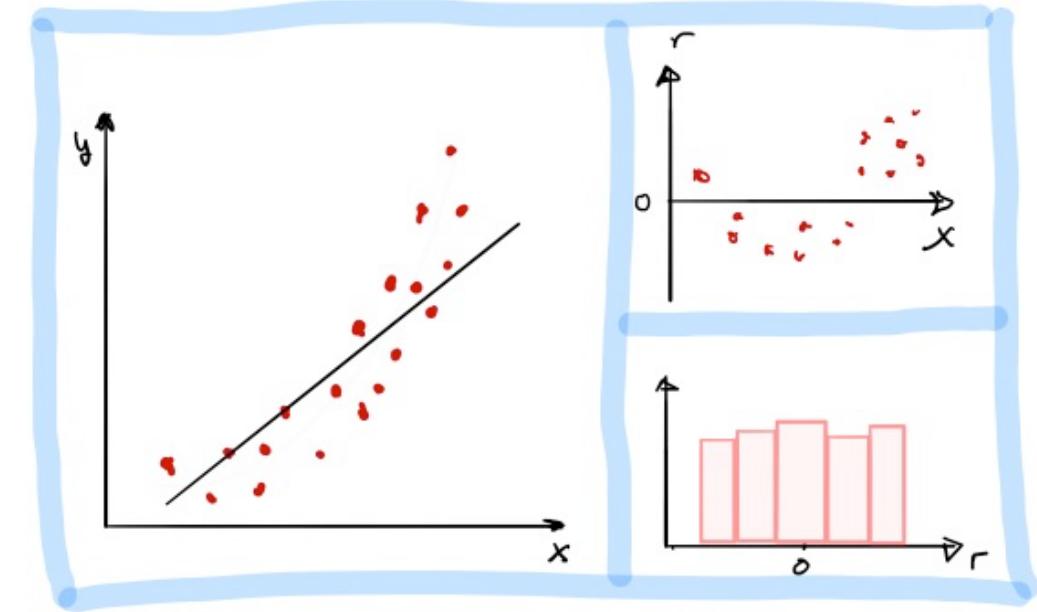
1. a plot of r_i with respect to x_i or \hat{y}_i . This allows us to compare the distribution of the noise at different values of x_i or \hat{y}_i .
2. a histogram of r_i . This allows us to explore the distribution of the noise independent of x_i or \hat{y}_i .

Residual Analysis



Linear assumption is correct. There is no obvious relationship between residuals and x . Histogram of residuals is symmetric and normally distributed.

Note: For multi-regression, we plot the residuals vs predicted y , \hat{y} , since there are too many x 's and that could wash out the relationship.



Linear assumption is incorrect. There is an obvious relationship between residuals and x . Histogram of residuals is symmetric but not normally distributed.

Beyond linearity: **synergy effect** or **interaction effect**

- We also assume that the average effect on sales of a one-unit increase in TV is always β_1 regardless of the amount spent on radio.
- **Synergy effect** or **interaction effect** states that when an increase on the radio budget affects the effectiveness of the TV spending on sales.

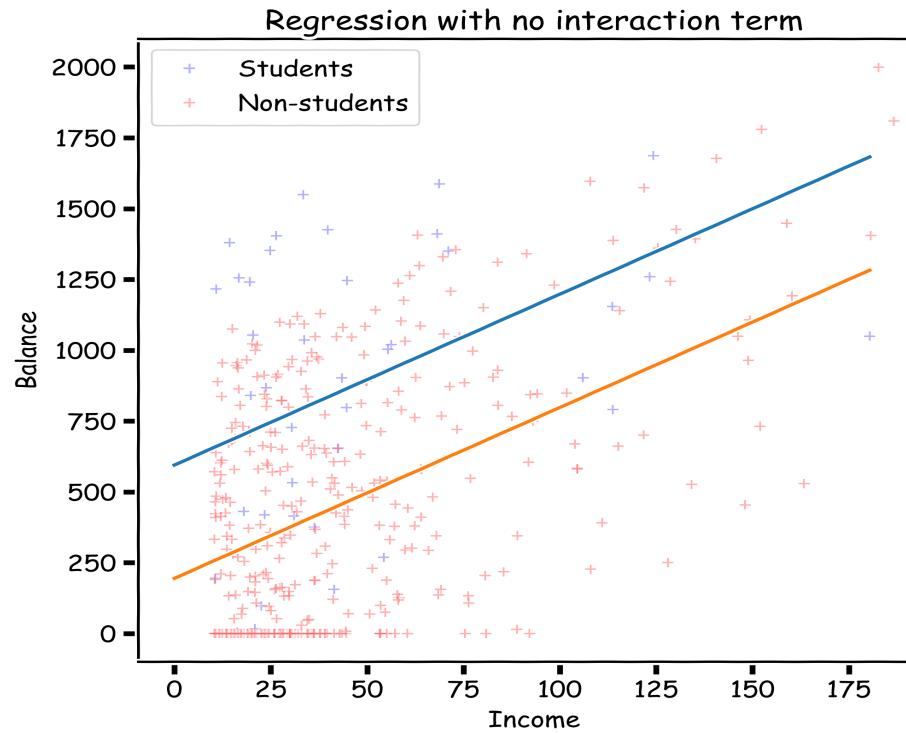
We change

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

to:

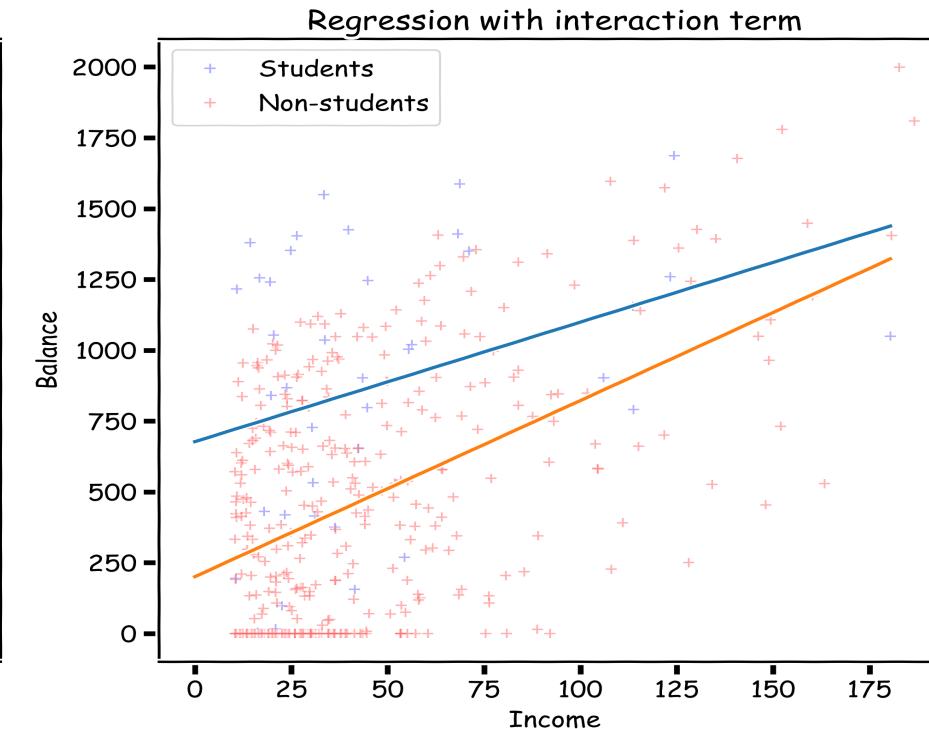
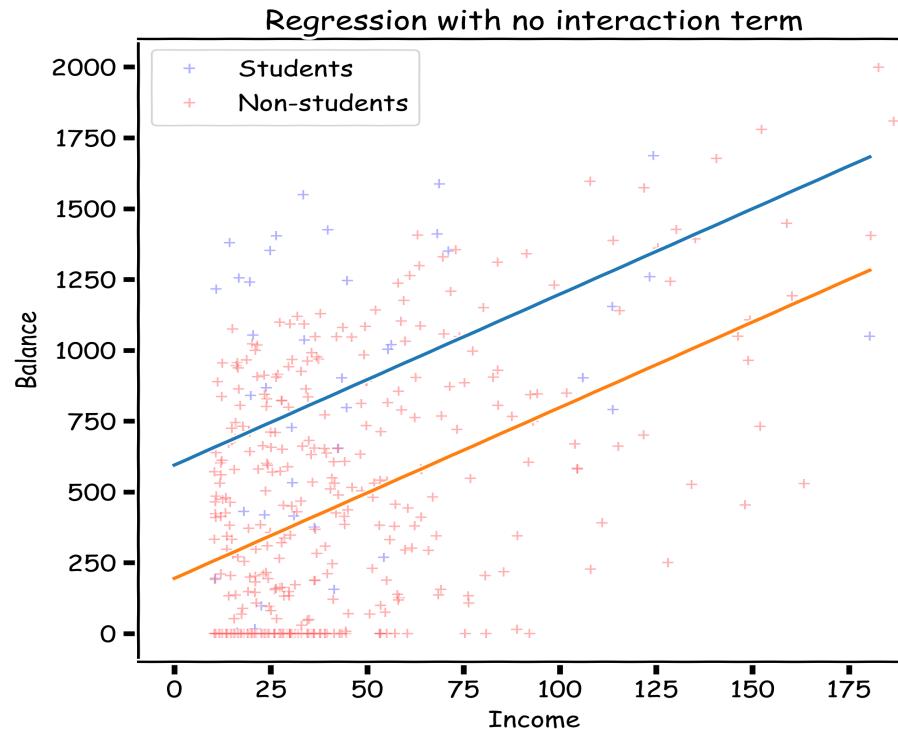
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \boxed{\beta_3 X_1 X_2} + \epsilon$$

What does it mean?



$$x_{Student} = \begin{cases} 0 & Balance = \beta_0 + \beta_1 \times Income. \\ 1 & Balance = (\beta_0 + \beta_2) + (\beta_1) \times Income. \end{cases}$$

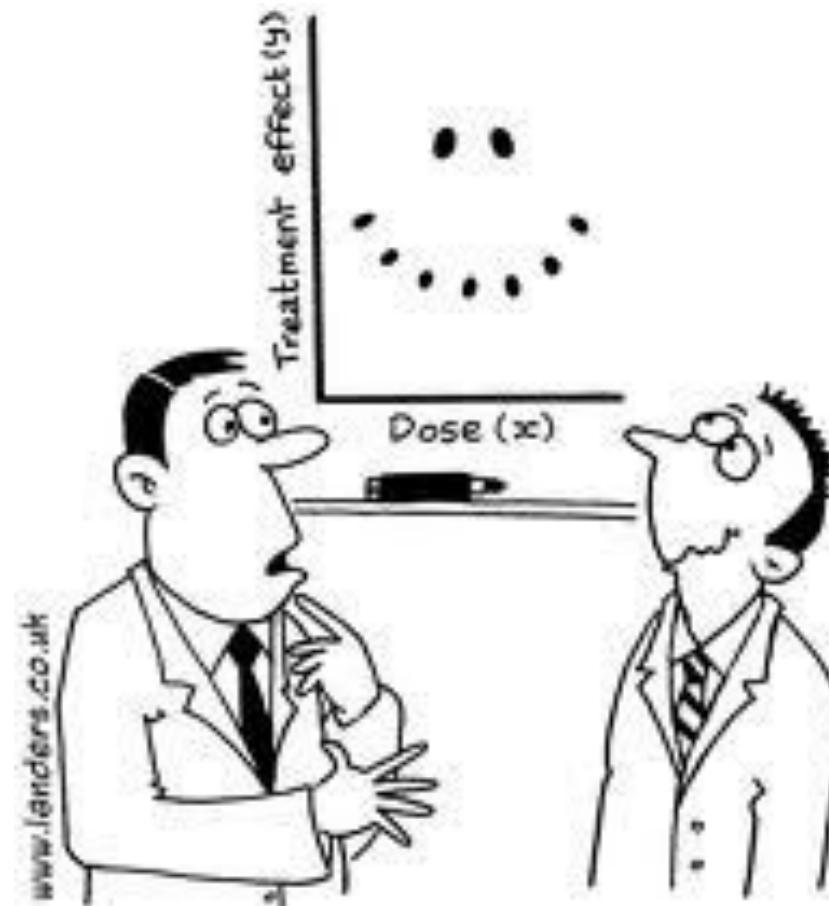
What does it mean?



$$x_{Student} = \begin{cases} 0 & Balance = \beta_0 + \beta_1 \times Income. \\ 1 & Balance = (\beta_0 + \beta_2) + (\beta_1) \times Income. \end{cases}$$

$$x_{Student} = \begin{cases} 0 & Balance = \beta_0 + \beta_1 \times Income. \\ 1 & Balance = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times Income \end{cases}$$

Too many predictors, collinearity and too many interaction terms leads to **OVERFITTING!**

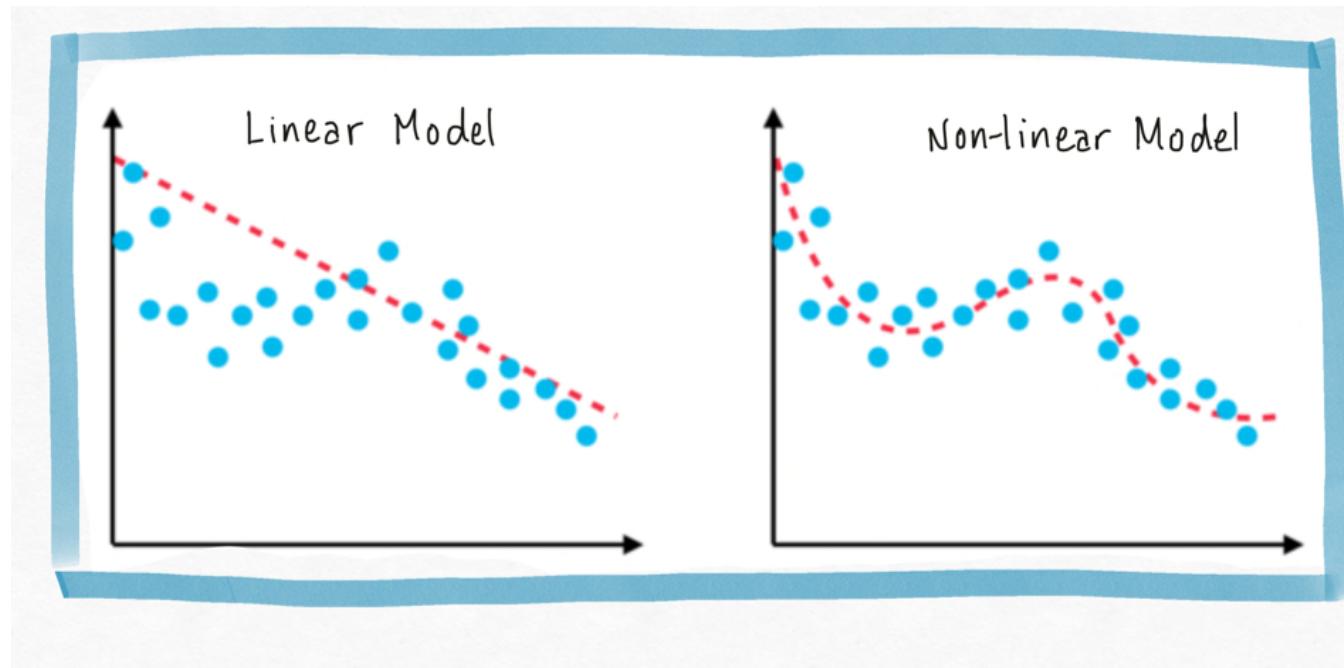


"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."

Polynomial Regression

Fitting non-linear data

Multi-linear models can fit large datasets with many predictors. But the relationship between predictor and target isn't always linear.



We want a model:

$$y = f_{\beta}(x)$$

Where f is a non-linear function and β is a vector of the parameters of f .

Polynomial Regression

The simplest non-linear model we can consider, for a response Y and a predictor X , is a polynomial model of degree M ,

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_M x^M$$

Just as in the case of linear regression with cross terms, polynomial regression is a special case of linear regression - we treat each x^m as a separate predictor. Thus, we can write

This looks a lot like multi-linear regression where the predictors are powers of x !

Model Training

Given a dataset $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, we find the optimal polynomial model:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_M x^M$$

We transform the data by adding new predictors:

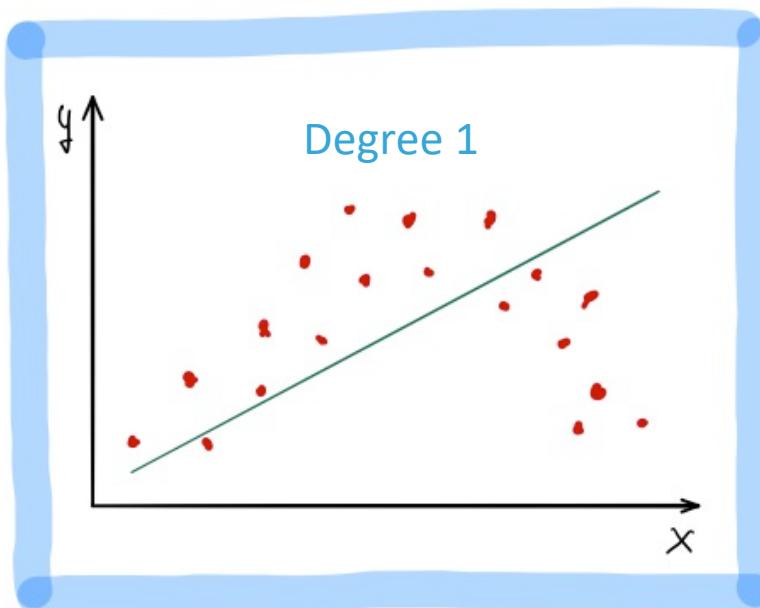
$$\tilde{x} = [1, \tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_M]$$

where $\tilde{x}_k = x^k$

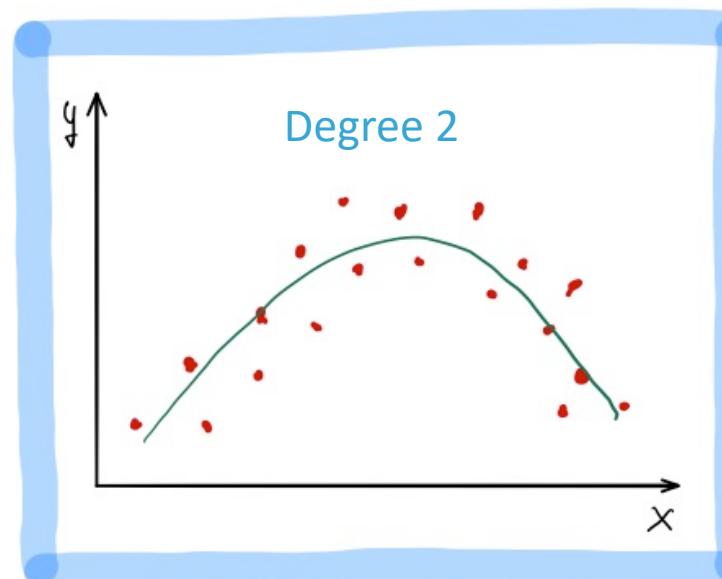
Fit the parameters by minimizing the MSE using vector calculus. As in multi-linear regression:

Polynomial Regression (cont)

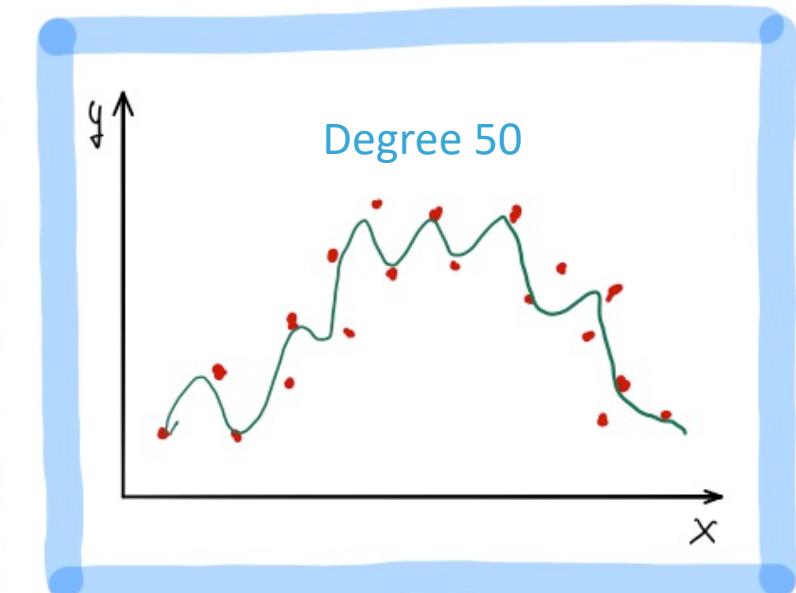
Fitting a polynomial model requires choosing a degree.



Underfitting: when the degree is too low, the model cannot fit the trend.



We want a model that fits the trend and ignores the noise.



Overfitting: when the degree is too high, the model fits all the noisy data points.

Feature Scaling

Do we need to scale out features for polynomial regression?

Linear regression, $Y = X\beta$, is **invariant** under scaling. If X is called by some number λ then β will be scaled by $\frac{1}{\lambda}$ and MSE will be identical.

However if the range of X is low or large then we run into troubles. Consider a polynomial degree of 20 and the maximum or minimum value of any predictor is large or small. Those numbers to the 20th power will be problematic.

- It is always a good idea to **scale** X when considering polynomial regression:

$$X^{norm} = \frac{X - \bar{X}}{\sigma_X}$$

Note: sklearn's `StandardScaler()` can do this.

High degree of polynomial
leads to **OVERFITTING!**

Recapitulation

- Consider collinearity
- Consider interaction effects
- Consider residual analysis
- Too many predictors, collinearity and too many interaction terms leads to **OVERFITTING!**
- High degree of polynomial leads to **OVERFITTING!**

For next class..



Finish Lab 06 to practice programming



Submit Homework 06 for peer review on Peer



Submit Assignment 3 – due in **Week 7** on Sunday at **2330**



See “To do before class” for every lecture (~ 1 hour of self study)



Read paper for **Discussion** session before every Friday



Post questions on the **Discussion** forum on Brightspace (especially on **ML** for this week)