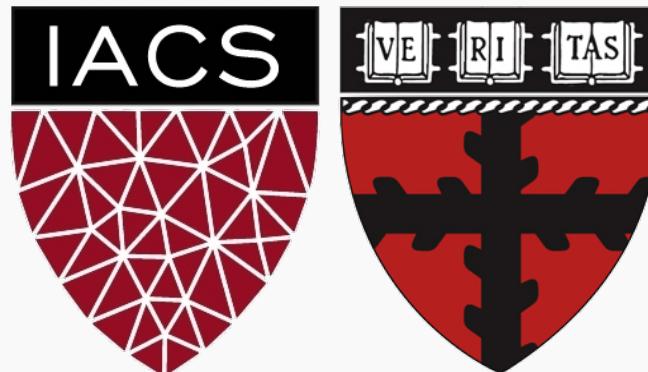


Multi, Poly Regression and Model Selection

Part B: Multi-regression

CS109A Introduction to Data Science
Pavlos Protopapas, Kevin Rader and Chris Tanner



GETTING VALUES FROM PANDAS



Multiple Linear Regression

If you have to 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

The diagram illustrates the components of a dataset:

- X predictors**: features, covariates
- Y outcome**: response variable, dependent variable
- n observations**: The number of observations in the dataset.
- p predictors**: The number of predictor variables in the dataset.

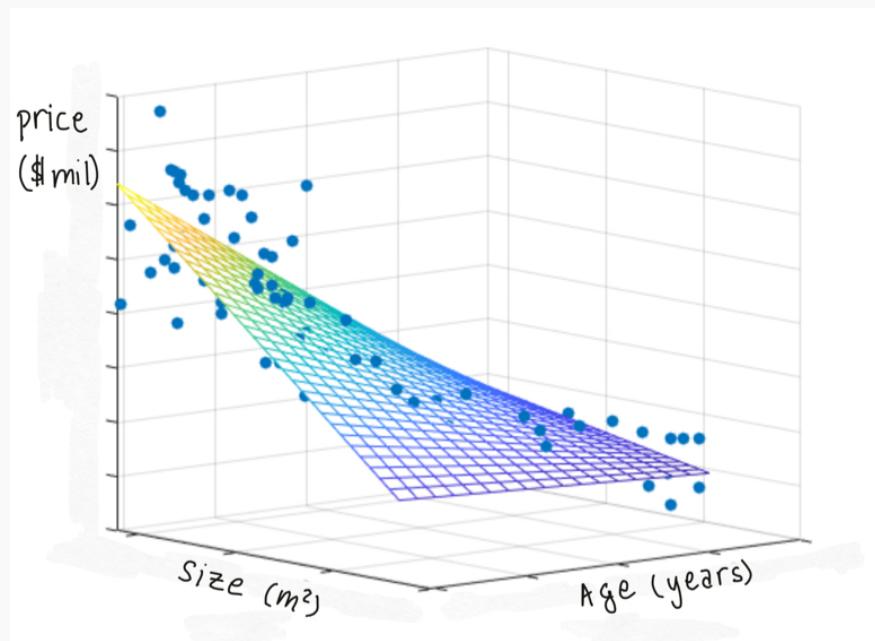
A bracket labeled "n observations" spans the five rows of the table below. A bracket labeled "p predictors" spans the four columns of the table below, except for the last column which is labeled "sales".

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

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:

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

Multiple Linear Regression

Given a set of observations,

$$\{(x_{1,1}, \dots, x_{1,J}, y_1), \dots, (x_{n,1}, \dots, x_{n,J}, y_n)\},$$

the data and the model can be expressed in vector notation,

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_y \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,J} \\ 1 & x_{2,1} & \dots & x_{2,J} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,J} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_J \end{pmatrix},$$

Multilinear Model, example

For our data

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

In linear algebra notation

$$Y = \begin{pmatrix} Sales_1 \\ \vdots \\ Sales_n \end{pmatrix}, X = \begin{pmatrix} 1 & TV_1 & Radio_1 & News_1 \\ \vdots & & \vdots & \vdots \\ 1 & TV_n & Radio_n & News_n \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_3 \end{pmatrix}$$

$$Sales_1 = [1 \quad TV_1 \quad Radio_1 \quad News_1] \times \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_3 \end{pmatrix}$$



Multiple Linear Regression

The model takes a simple algebraic form:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

We will again choose the **MSE** as our loss function, which can be expressed in vector notation as

$$\text{MSE}(\beta) = \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\beta\|^2$$

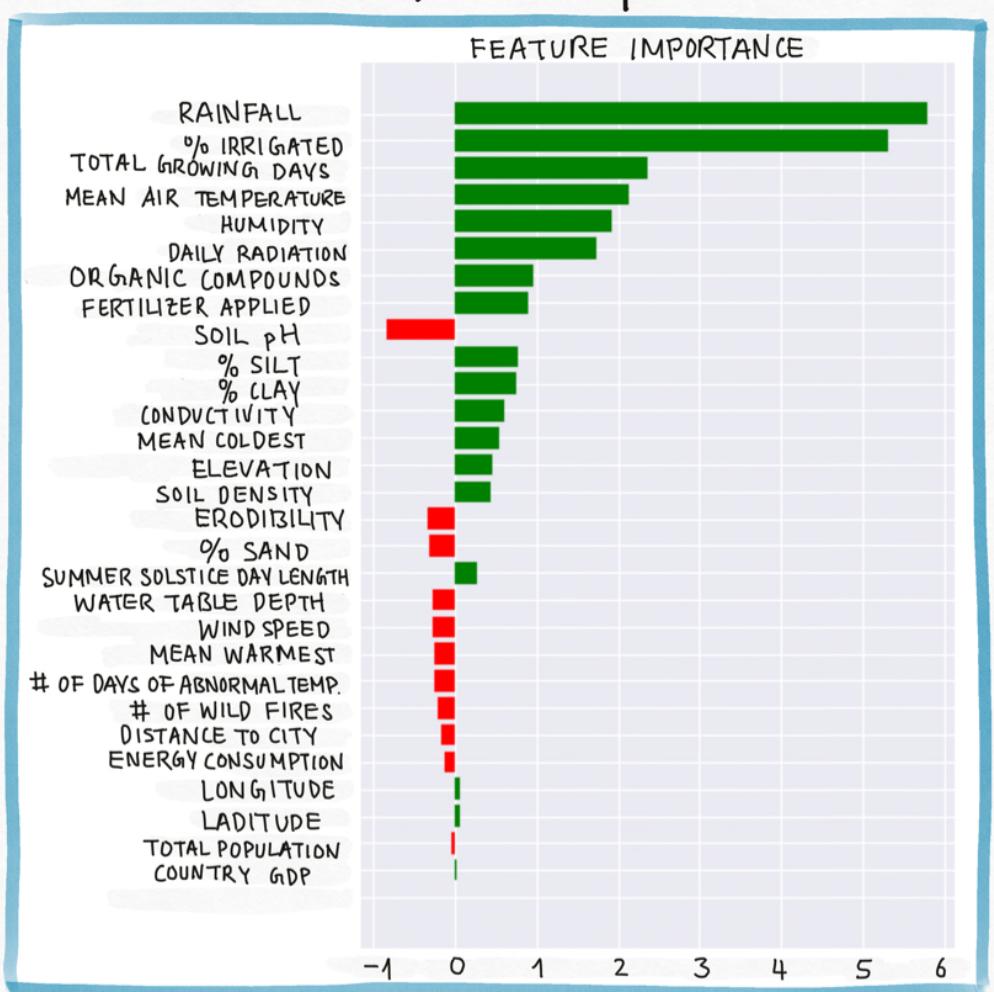
Minimizing the MSE using vector calculus yields,

$$\widehat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} = \underset{\beta}{\operatorname{argmin}} \text{MSE}(\beta).$$



Interpreting multi-linear regression

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



When we have a large number of predictors: X_1, \dots, X_J , there will be a large number of 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 a number of 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

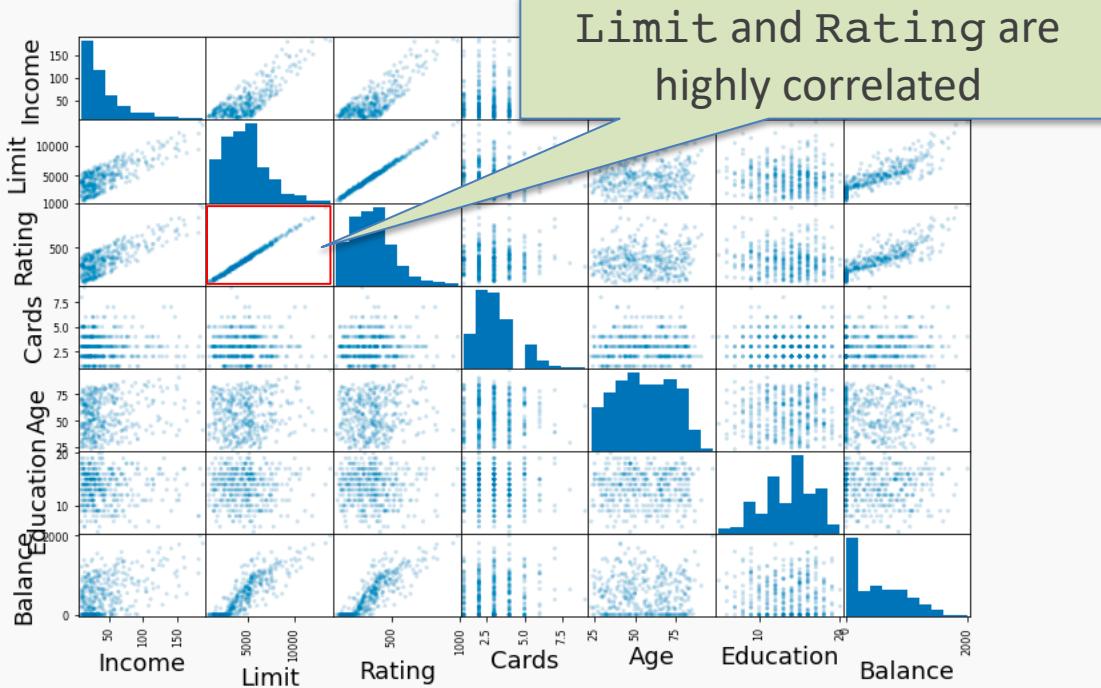
We then use these variables as predictors, the regression equation becomes:

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

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).



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.

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

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:

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

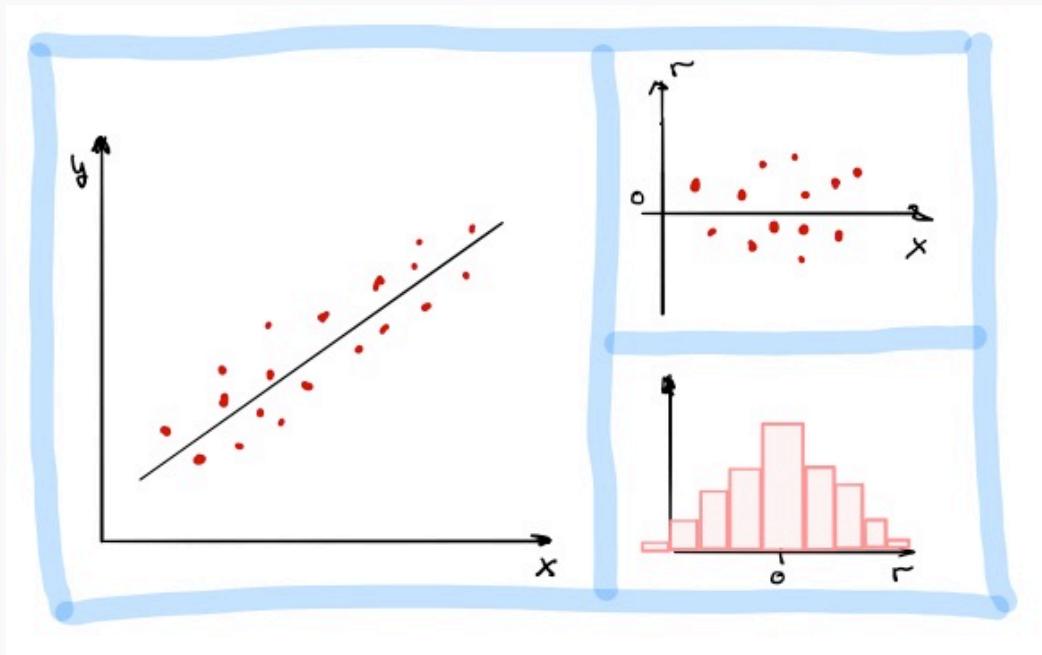
Then the residuals, $r = (y - \hat{y}) = \epsilon + \phi(x)$, are **not independent** of x

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 .

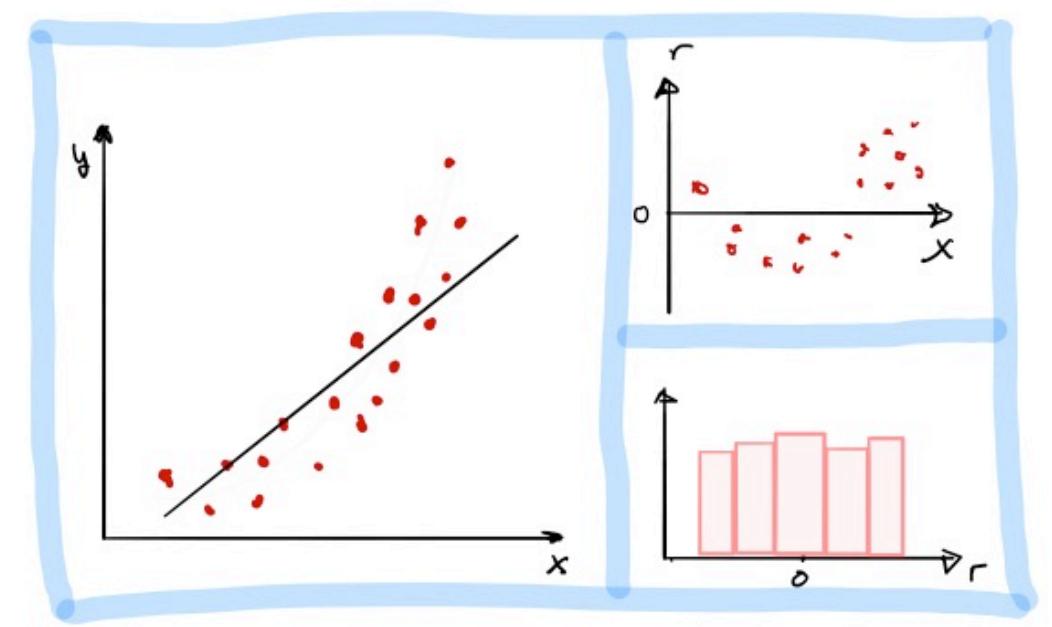


Residual Analysis



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.

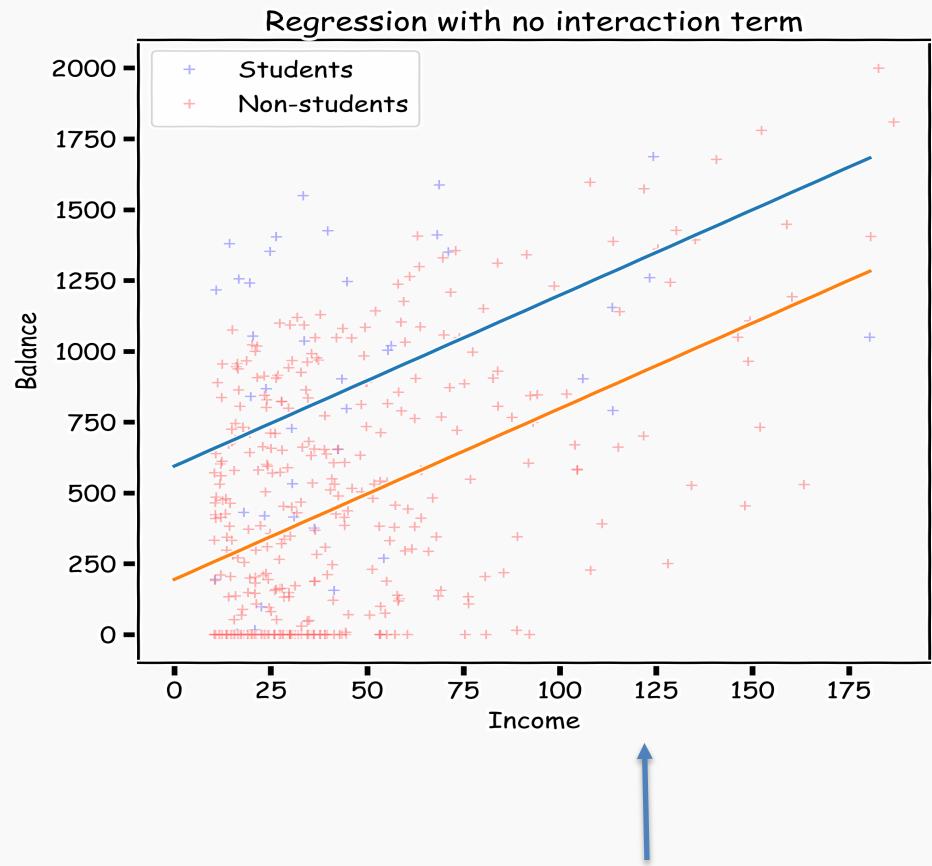
We change

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

to:

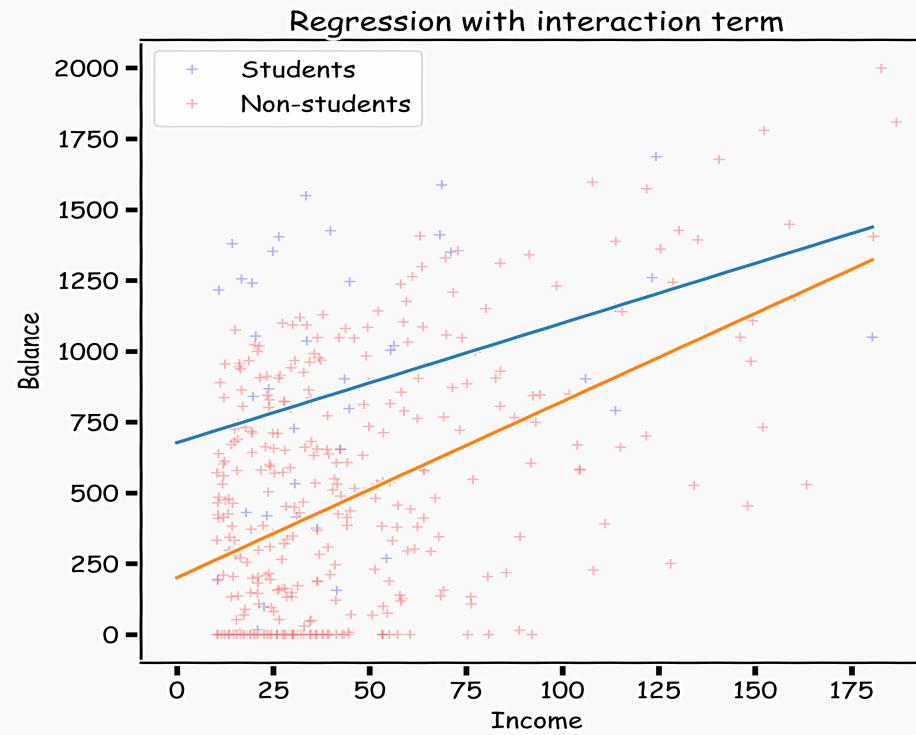
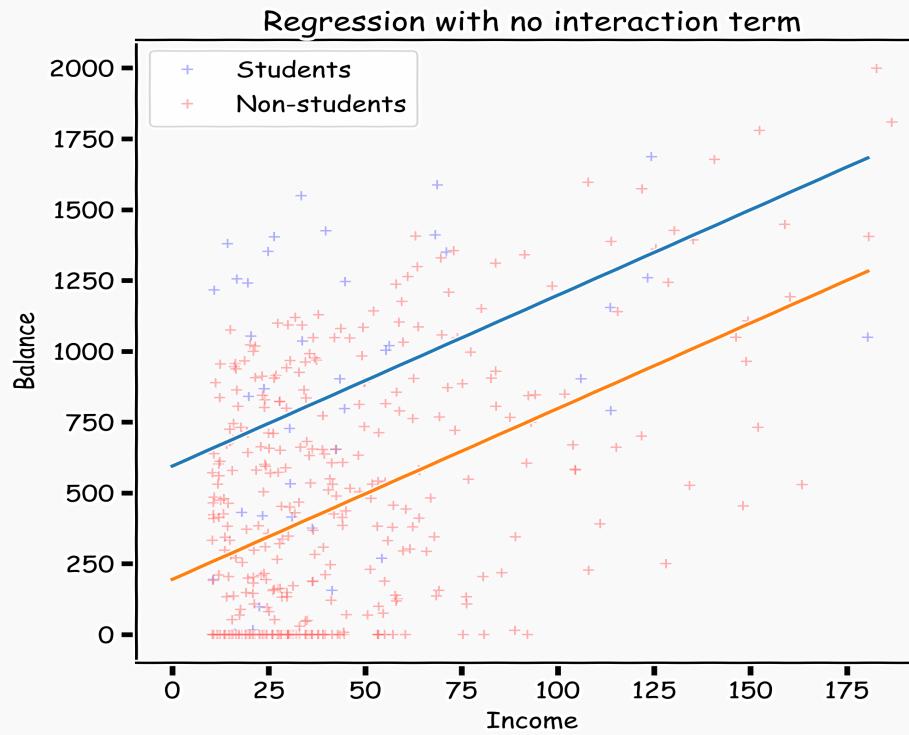
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \boxed{\beta_3 X_1 X_2} + \epsilon$$

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?

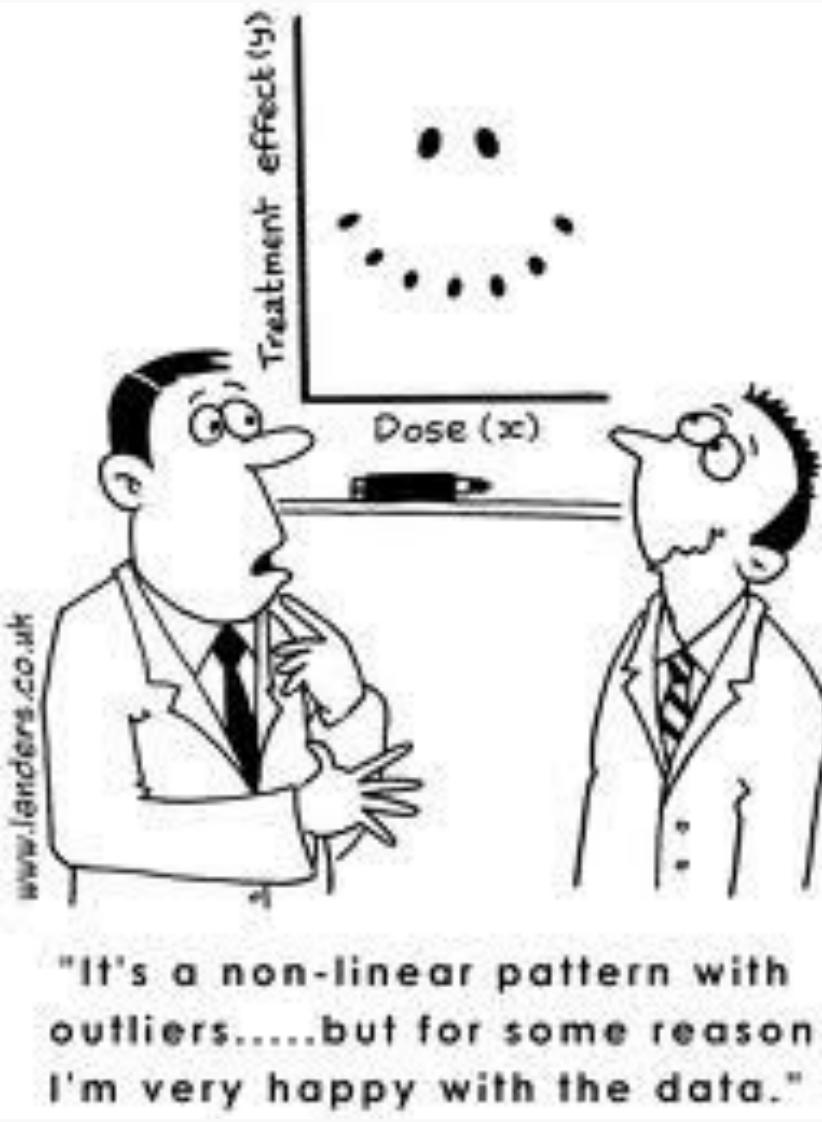


$$x_{Student} = \begin{cases} 0 & \text{Balance} = \beta_0 + \beta_1 \times \text{Income}. \\ 1 & \text{Balance} = (\beta_0 + \beta_2) + (\beta_1) \times \text{Income}. \end{cases}$$

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

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



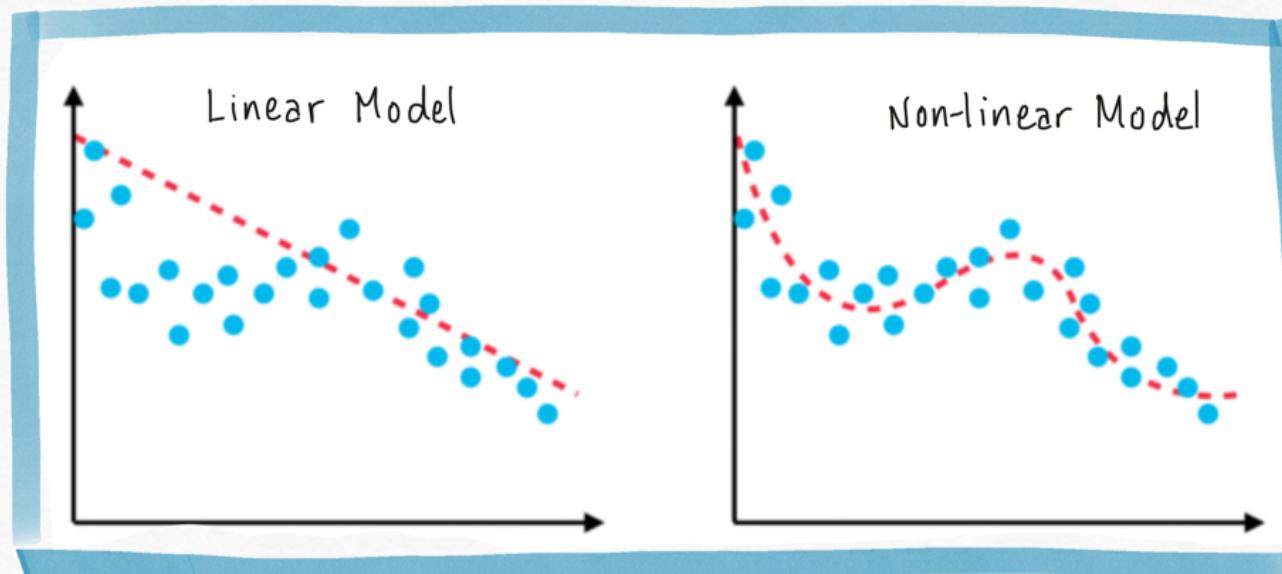


Polynomial Regression



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

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1^1 & \cdots & x_1^M \\ 1 & x_2^1 & \cdots & x_2^M \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^M \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_M \end{pmatrix}.$$



Polynomial Regression

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

Multi-Regression

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,J} \\ 1 & x_{2,1} & \dots & x_{2,J} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,J} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_J \end{pmatrix},$$

Poly-Regression

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1^1 & \dots & x_1^M \\ 1 & x_2^1 & \dots & x_2^M \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \dots & x_n^M \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_M \end{pmatrix}.$$



Model Training

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

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

1. We transform the data by adding new predictors:

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

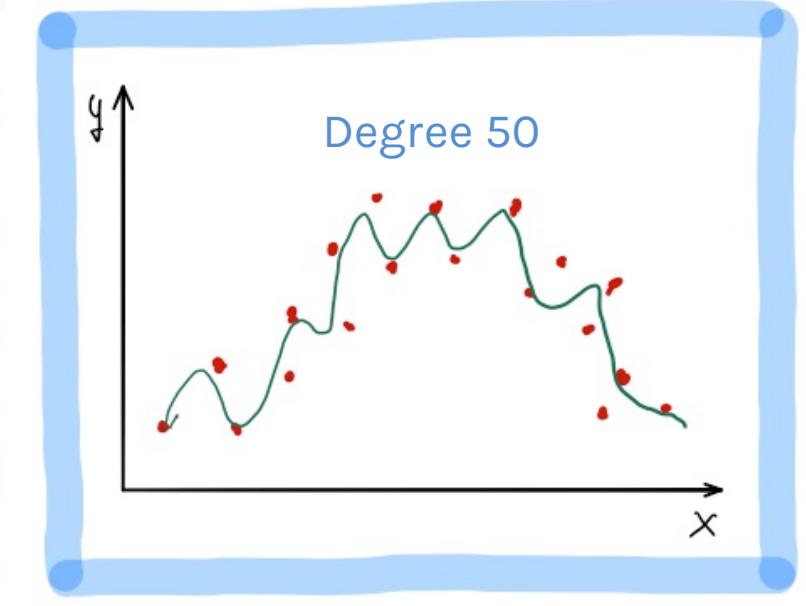
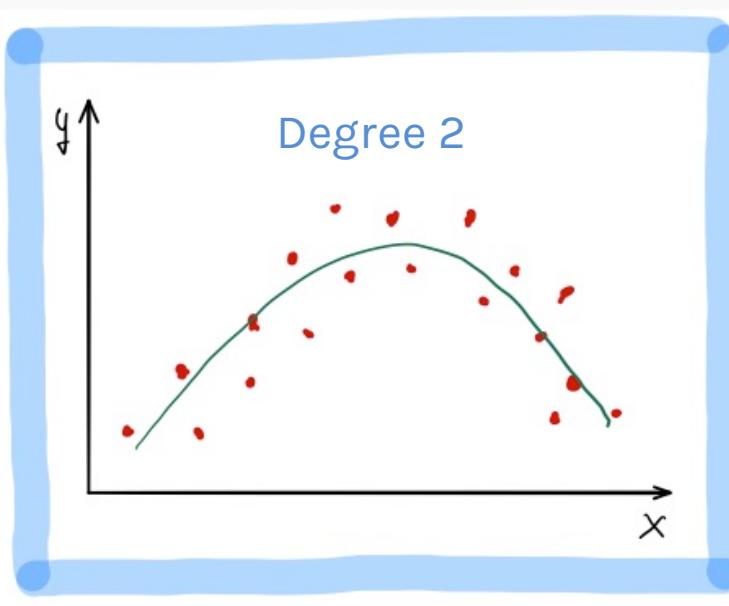
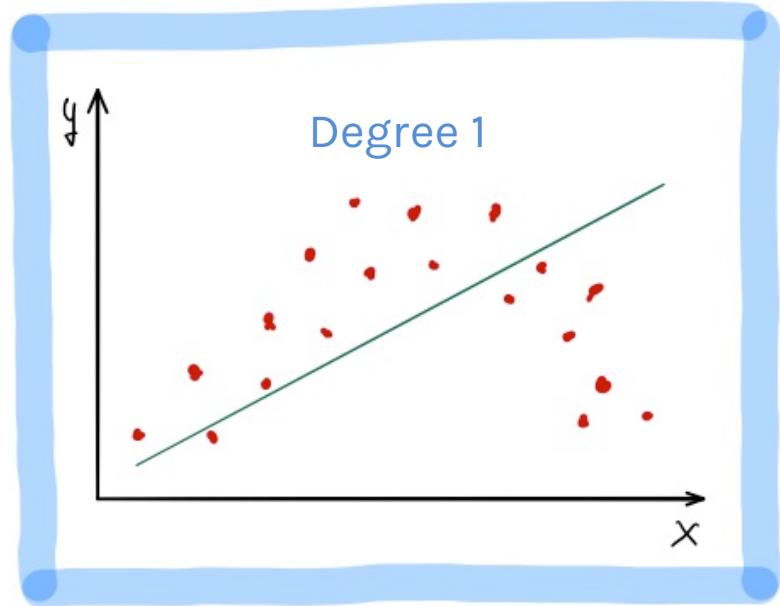
where $\tilde{x}_k = x^k$

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

$$\hat{\beta} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T y$$

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:

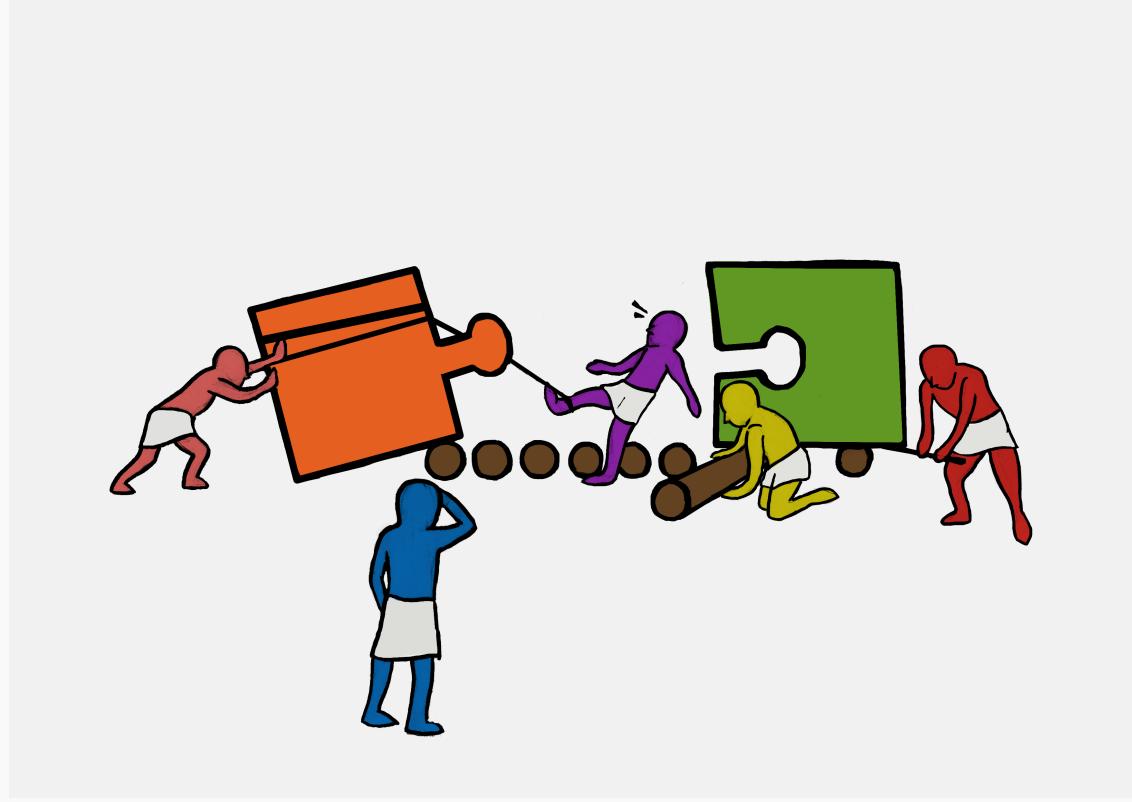
$$X^{norm} = \frac{X - \bar{X}}{\sigma_X}$$

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



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





Ex B.1, B.2 & B.3

