COMP47350: Data Analytics (Conv)

Dr. Georgiana Ifrim
georgiana.ifrim@ucd.ie

Insight Centre for Data Analytics
School of Computer Science
University College Dublin

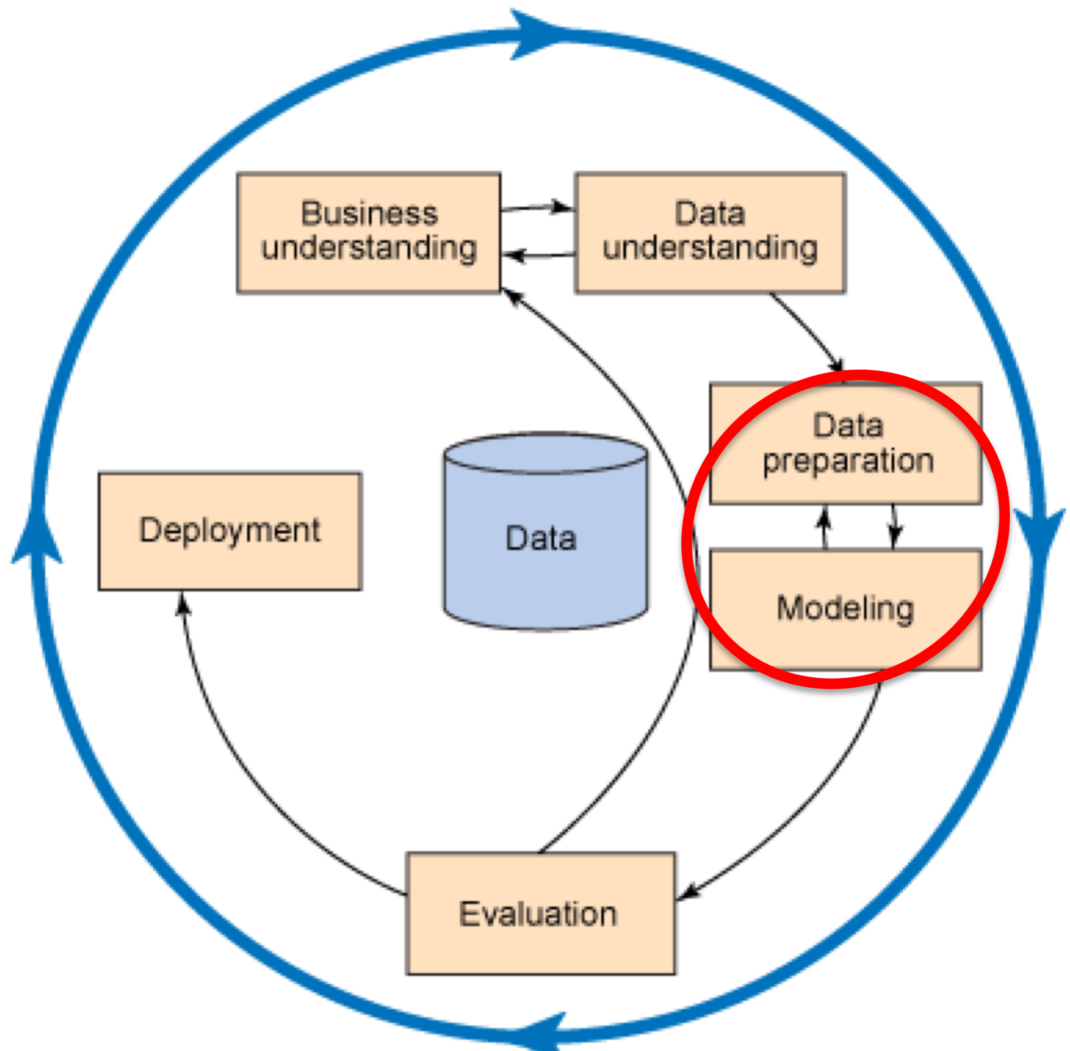
2018/19

Module Topics

- **Python Environment** (Anaconda, Jupyter Notebook)
- **Getting Data** (Web scrapping, APIs, DBs)
- **Understanding Data** (slicing, visualisation)
- **Preparing Data** (cleaning, transformation)
- **Modeling & Evaluation** (machine learning)

Data Analytics Project Lifecycle: CRISP-DM

CRISP-DM: **CR**oss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining



Modeling Data

- Modeling:
  - **How to build prediction models**
  - How to evaluate prediction models

- **Regression:** predicting a numeric target feature

Supervised Machine Learning

- **Regression:** Automatically learn/estimate a **model (i.e., function)** for the relationship between a set of **descriptive features** and a **numeric target feature**
  - E.g., learn the relationship between <u>descriptive features</u>, **SIZE, LOCATION, FLOOR SPACE** and <u>target feature</u>, **RENTAL PRICE.**

Recap: Linear Regression

- Assumes a **linear relationship** between descriptive features and target feature

  predicted_target = w_0 + w_1 *feature_1 + w_2*feature_2 + +…+ w_n*feature_n

- The model is described by a set of **parameters** also known as **weights** (e.g., w_0, w_1, …, w_n, where **n** is the number of features)
- Training stage: Estimate (aka learn) the parameters w_0, w_1, …, w_n
- Prediction stage: Apply the weights learned during training, to the descriptive features of each example, to get a predicted target
- Evaluation: Measure the difference between actual target and the predicted target to get a measure for how well the model works

Linear Regression

Topics covered in this lecture:

1. Model Interpretability
2. Categorical Features
3. Non-linear Relationship

Linear Regression

Topics covered in this lecture:

1. **Model Interpretability** (interpreting a linear regression model)

Linear Regression: Example

Table: The **office rentals dataset**: a dataset that includes office rental prices and a number of descriptive features for 10 Dublin city-centre offices.

ID	Size	Floor	Broadband Rate	Energy Rating	Rental Price
1	500	4	8	C	320
2	550	7	50	A	380
3	620	9	7	A	400
4	630	5	24	B	390
5	665	8	100	C	385
6	700	4	8	B	410
7	770	10	7	B	480
8	880	12	50	A	600
9	920	14	8	C	570
10	1,000	9	24	B	620

Can we predict the **rental price**, given the descriptive features (**size, floor**, broadband rate, energy rating) for an office?

Simple Linear Regression (1 feature)

ID	SIZE	RENTAL PRICE
1	500	320
2	550	380
3	620	400
4	630	390
5	665	385
6	700	410
7	770	480
8	880	600
9	920	570
10	1,000	620

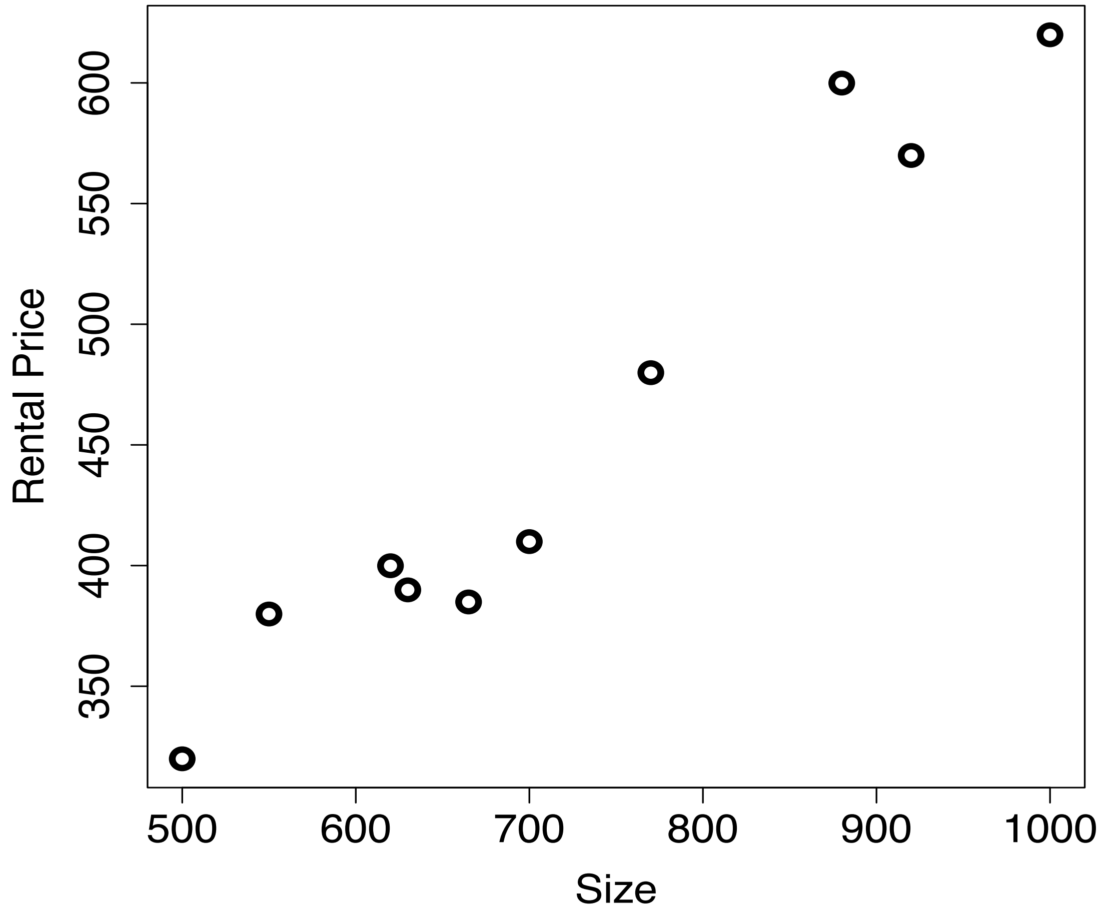
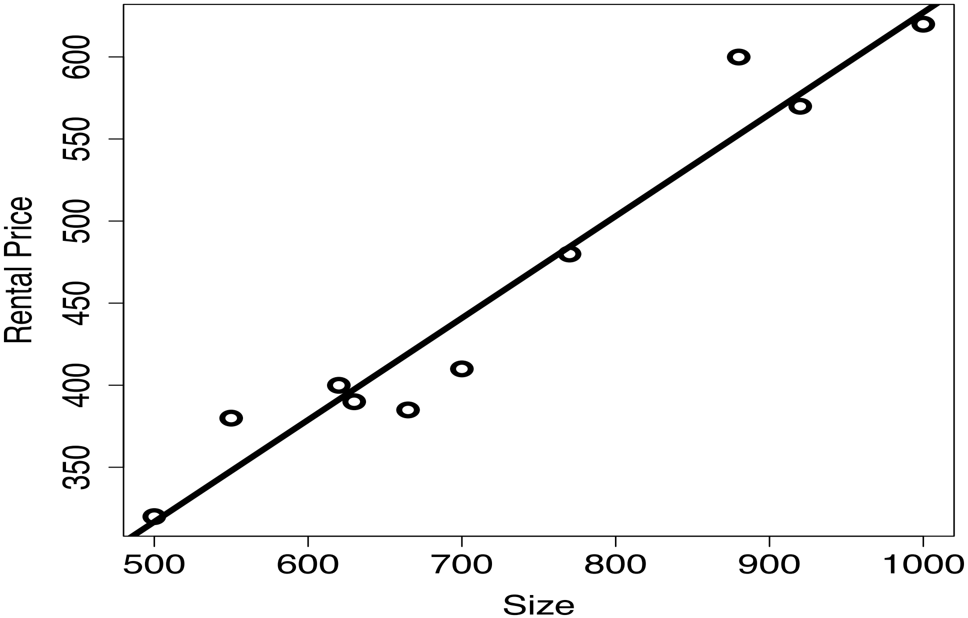


Figure: A scatter plot of the SIZE and RENTAL PRICE features from the office rentals dataset.

Simple Linear Regression

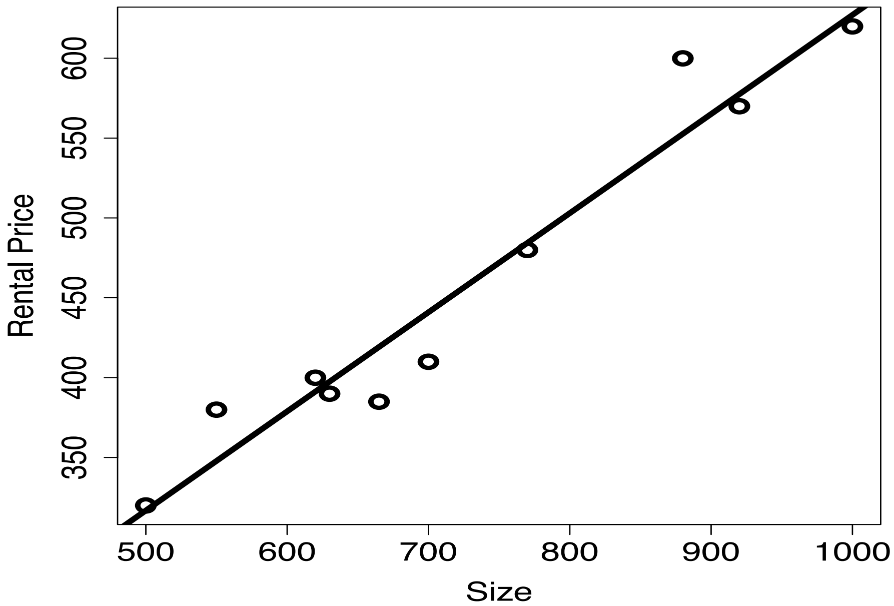
- **Regression line** estimates relationship between SIZE and RENTAL PRICE
- **Learned model** using our training set with 10 examples is: w_0 = 6.47, w_1 = 0.62

$$\text{Rental Price} = 6.47 + 0.62 \times \text{Size}$$



Simple Linear Regression: Interpretation

- **Interpretation:** $\boxed{\text{RENTAL PRICE} = 6.47 + 0.62 \times \text{SIZE}}$
  - w_0 = 6.47: Start is from a base price of 6.47 euro (adjustment parameter called bias).
  - w_1 = 0.62: For every increase of a square foot in SIZE, the RENTAL PRICE increases by 0.62 euro.



Multiple Linear Regression: Interpretation

Linear regression model using all continuous features:

$$
\begin{aligned}
\text{RENTAL PRICE} = \mathbf{w}[0] \quad & + \quad \mathbf{w}[1] \times \text{SIZE} + \mathbf{w}[2] \times \text{FLOOR} \\
& + \quad \mathbf{w}[3] \times \text{BROADBAND RATE}
\end{aligned}
$$

Multiple linear regression model learned from our 10 training examples:

$$
\begin{aligned}
\text{RENTAL PRICE} = -0.1513 \quad & + \quad 0.6270 \times \text{SIZE} \\
& - \quad 0.1781 \times \text{FLOOR} \\
& + \quad 0.0714 \times \text{BROADBAND RATE}
\end{aligned}
$$

- For every unit increase in SIZE (everything else being fixed), the PRICE increases by 0.627 euro.
- For every unit increase in FLOOR, the PRICE decreases by 0.1781 euro.
- For every unit increase in BROADBANDRATE, the PRICE increases by 0.0714 units.

Multiple Linear Regression: Prediction

- Using this model:

$$
\begin{aligned}
\textsc{Rental Price} = -0.1513 \quad & + \quad 0.6270 \times \textsc{Size} \\
& - \quad 0.1781 \times \textsc{Floor} \\
& + \quad 0.0714 \times \textsc{Broadband Rate}
\end{aligned}
$$

- we can, for example, predict the expected rental price of a 690 square foot office on the 11^{th} floor of a building with a broadband rate of 50 Mb per second as:

$$
\begin{aligned}
\textsc{Rental Price} \quad & = \quad -0.1513 + 0.6270 \times 690 \\
& \quad\quad -0.1781 \times 11 + 0.0714 \times 50 \\
& = \quad 434.0896
\end{aligned}
$$

Linear Regression

Topics covered in this lecture:

2. **Categorical Features** (linear regression expects numeric feature values; need to transform categorical features values into numeric values)

Dealing with Categorical Features

Table: The **office rentals dataset**: a dataset that includes office rental prices and a number of descriptive features for 10 Dublin city-centre offices.

ID	Size	Floor	Broadband Rate	Energy Rating	Rental Price
1	500	4	8	C	320
2	550	7	50	A	380
3	620	9	7	A	400
4	630	5	24	B	390
5	665	8	100	C	385
6	700	4	8	B	410
7	770	10	7	B	480
8	880	12	50	A	600
9	920	14	8	C	570
10	1,000	9	24	B	620

Can we predict the rental price, given the descriptive features (size, floor, broadband rate, **energy rating**) for an office?

ENERGY RATING is a categorical feature, can we still use it?

Dealing with Categorical Features

Two popular approaches:

1. **Binary encoding** (aka dummy encoding, one-hot-encoding)

- Turn each categorical feature into many binary continuous features that encode the levels of the categorical feature
- Example: ENERGY RATING feature (with levels A, B, C) turns into 3 (or 2) continuous features.
- A dummy variable is a variable created to assign numerical value to levels of categorical variables. For L levels, create L-1 variables (first level is reference)

ID	SIZE	FLOOR	BROADBAND RATE	ENERGY RATING A	ENERGY RATING B	ENERGY RATING C	RENTAL PRICE
1	500	4	8	0	0	1	320
2	550	7	50	1	0	0	380
3	620	9	7	1	0	0	400
4	630	5	24	0	1	0	390
5	665	8	100	0	0	1	385
6	700	4	8	0	1	0	410
7	770	10	7	0	1	0	480
8	880	12	50	1	0	0	600
9	920	14	8	0	0	1	570
10	1 000	9	24	0	1	0	620

Dealing with Categorical Features

- New model with the categorical feature turned into several continuous features; does not assume ordering of categories
- Downside: if the categorical feature has many levels, we will create many new features (need to handle more features)

$$\begin{aligned}\text{RENTAL PRICE} = \mathbf{w}[0] &+ \mathbf{w}[1] \times \text{SIZE} + \mathbf{w}[2] \times \text{FLOOR} \\ &+ \mathbf{w}[3] \times \text{BROADBAND RATE} \\ &+ \mathbf{w}[4] \times \text{ENERGY RATING A} \\ &+ \mathbf{w}[5] \times \text{ENERGY RATING B} \\ &+ \mathbf{w}[6] \times \text{ENERGY RATING C}\end{aligned}$$

ID	SIZE	FLOOR	BROADBAND RATE	ENERGY RATING A	ENERGY RATING B	ENERGY RATING C	RENTAL PRICE
1	500	4	8	0	0	1	320
2	550	7	50	1	0	0	380
3	620	9	7	1	0	0	400
4	630	5	24	0	1	0	390
5	665	8	100	0	0	1	385
6	700	4	8	0	1	0	410
7	770	10	7	0	1	0	480
8	880	12	50	1	0	0	600
9	920	14	8	0	0	1	570
10	1 000	9	24	0	1	0	620

Dealing with Categorical Features

- Dummy encoding with dropping the refence column
- RENTAL PRICE = w[0] + w[1] * SIZE + w[2]*FLOOR + w[3]*BROABAND RATE + w[4] * ENERGY RATING B + w[5] * ENERGY RATING C
- Interpretation for categorical features: a change from the reference level (e.g., ENERGY RATING A) to level B results in w[4] change in PRICE

ID	SIZE	FLOOR	BROADBAND RATE	~~ENERGY RATING A~~	ENERGY RATING B	ENERGY RATING C	RENTAL PRICE
1	500	4	8	~~0~~	0	1	320
2	550	7	50	~~1~~	0	0	380
3	620	9	7	~~1~~	0	0	400
4	630	5	24	~~0~~	1	0	390
5	665	8	100	~~0~~	0	1	385
6	700	4	8	~~0~~	1	0	410
7	770	10	7	~~0~~	1	0	480
8	880	12	50	~~1~~	0	0	600
9	920	14	8	~~0~~	0	1	570
10	1 000	9	24	~~0~~	1	0	620

Dealing with Categorical Features

2. **Integer encoding:**
   - Assign an integer number for each category (preserve meaning of original feature values, e.g., ranking order)
   - E.g., Energy Rating A = 1, Energy Rating B = 2, Energy Rating C = 3, or other coding, e.g., Energy Rating A = 1, Energy Rating B = 10, Energy Rating C = 100
   - Downside: introduces a (potentially arbitrary) ordering of categories; we can control the category to number mapping (e.g., assign random numbers, assign numbers that preserve the order on categories if any)
   - Tends to work well in practice
   - Does not increase the number of features
   - Interpretation as for regular continuous features

Linear Regression

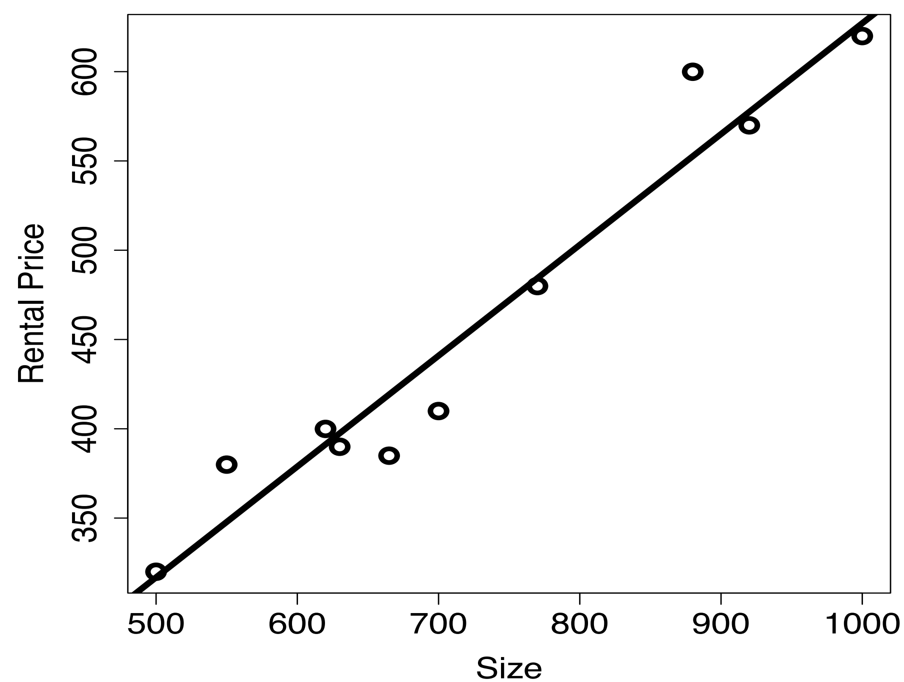
Topics covered in this lecture:

3. **Non-linear Relationship** (the case where the dependency between descriptive features and target feature is not linear)

Modeling Non-linear Relationships

- Linear regression assumes a linear relationship between features and target (i.e., the function learned is a line that best approximates the training data)
- Trained model for the Office dataset: RENTAL PRICE = w_0 + w_1 * SIZE

$$\text{RENTAL PRICE} = 6.47 + 0.62 \times \text{SIZE}$$



Modeling Non-linear Relationships

- What if the relationship is <u>not linear</u>?

Table: A dataset describing grass growth on Irish farms during July 2012.

ID	RAIN	GROWTH	ID	RAIN	GROWTH	ID	RAIN	GROWTH
1	2.153	14.016	12	3.754	11.420	23	3.960	10.307
2	3.933	10.834	13	2.809	13.847	24	3.592	12.069
3	1.699	13.026	14	1.809	13.757	25	3.451	12.335
4	1.164	11.019	15	4.114	9.101	26	1.197	10.806
5	4.793	4.162	16	2.834	13.923	27	0.723	7.822
6	2.690	14.167	17	3.872	10.795	28	1.958	14.010
7	3.982	10.190	18	2.174	14.307	29	2.366	14.088
8	3.333	13.525	19	4.353	8.059	30	1.530	12.701
9	1.942	13.899	20	3.684	12.041	31	0.847	9.012
10	2.876	13.949	21	2.140	14.641	32	3.843	10.885
11	4.277	8.643	22	2.783	14.138	33	0.976	9.876

Modeling Non-linear Relationships

- What if the relationship is <u>not linear</u>?

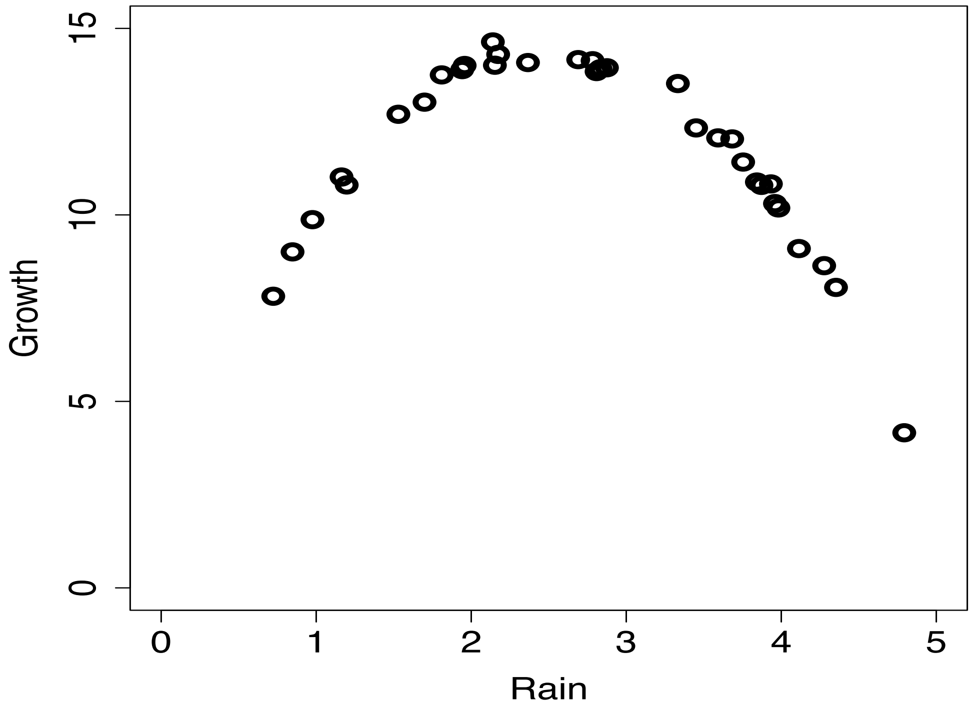


Figure: A scatter plot of the RAIN and GROWTH feature from the grass growth dataset.

Modeling Non-linear Relationships

- Linear model GROWTH = w0 + w1*RAIN
- The best linear model for this dataset: w0 = 13.510, w1 = -0.667
- Best linear model for this dataset: GROWTH = 13.510 - 0.667 * RAIN
- We can train a better linear model using basis functions on features: GROWTH = w0 + w1*RAIN + w2* RAIN^2

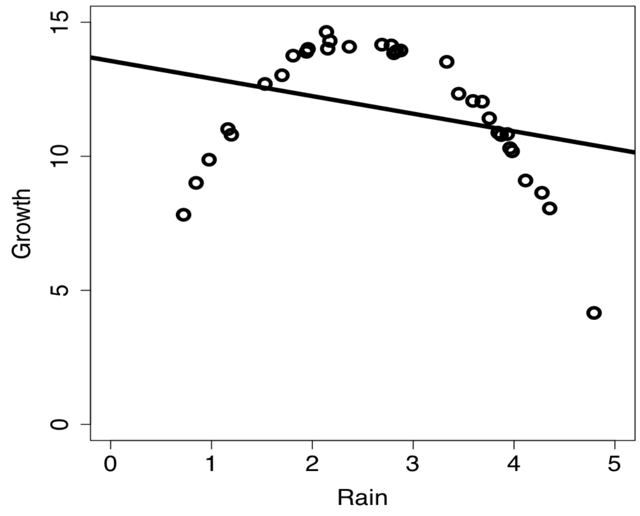


Figure: A simple linear regression model trained to capture the relationship between the grass growth and rainfall.

Modeling Non-linear Relationships

Common solutions (basis functions applied to features):

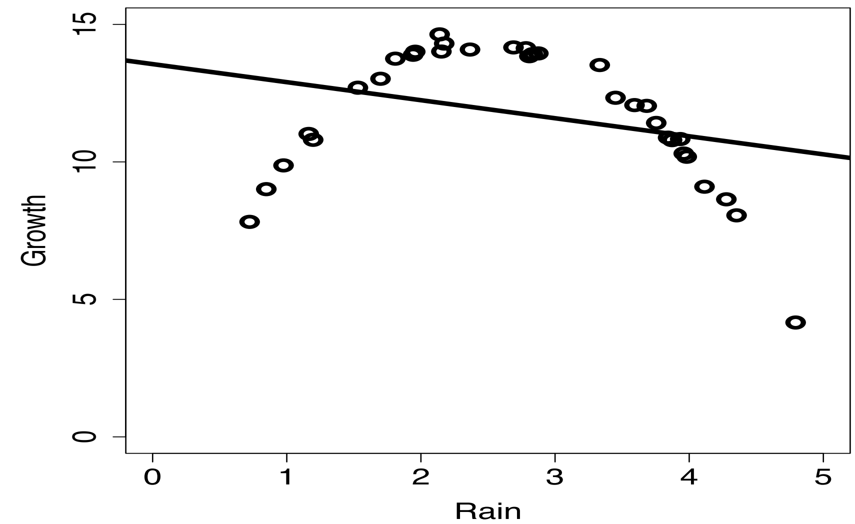
- **Solution1:** Create new features that capture non-linear polynomials of original features
  - E.g., original descriptive feature: RAIN. Create a new feature (quadratic polynomial): $RAIN^2$
- **Solution2:** Create feature interactions
  - E.g., original descriptive features: SALARY, HOUSE_PRICE. Create a new feature: the ratio of the two features SALARY/HOUSE_PRICE
- **Finally:** Build linear regression model with original features + new (derived) features that aim to capture non-linear behavior

Modeling Non-linear Relationships

A linear model using **original** features:

GROWTH = 13.510 - 0.667 * RAIN

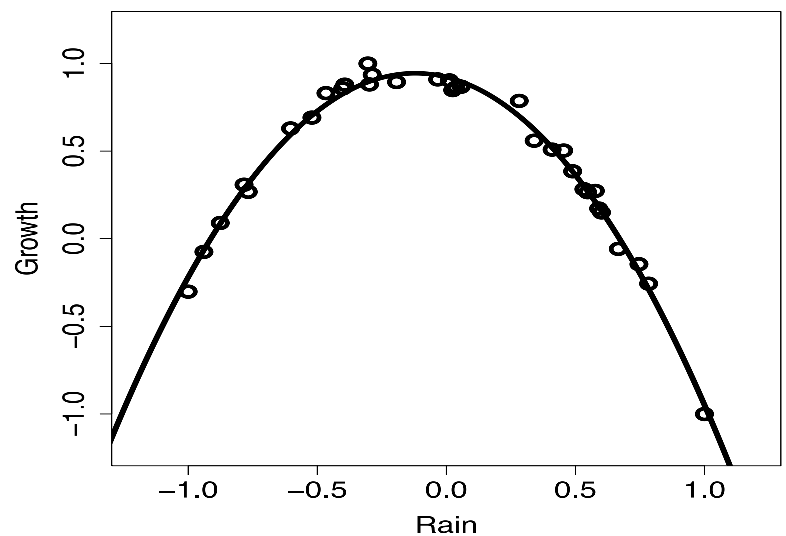
(This model has a large error!)



A linear model using **original** features <u>and</u> **quadratic** features:

GROWTH = 0.3707 + 0.8475 * RAIN - 1.717 * RAIN^2

(This model fits the data better and has lower error)



Linear Regression: Summary

- Main assumption: linear relationship between descriptive features and target feature
- Estimates a linear model from given examples (a set of weights w_0, w_1,…, w_n)

- Expects numeric feature values
- Categorical features can be transformed into continuous features
- Interpretation: proceed with caution (continuous vs categorical features; sanity check interpretation: correlation does not imply causation)

- If scatter plots show non-linear relationship between descriptive and target feature, we can introduce new features to capture non-linearity
- Trade-off: improving model fit (training error) vs increasing model complexity (more features, higher polynomials); evaluate this trade-off using error on a separate test set

- Not covered (advanced topics):
  - Statistical significance testing of feature coefficients
  - Collinearity (effect of correlated features on linear regression)

References

- Chapter7 from **FMLPDA Book: Fundamentals of Machine Learning for Predictive Data Analytics**, by J. Kelleher, B. Mac Namee and A. D'Arcy, MIT Press, 2015 (machinelearningbook.com)
- Chapter3 from **An Introduction to Statistical Learning**, by G. James, D. Witten, T. Hastie, R. Tibshirani, Springer, 2016 (free book: http://www-bcf.usc.edu/~gareth/ISL/)
- A friendly introduction to linear regression (using Python): http://www.dataschool.io/linear-regression-in-python/
- Feature Selection: http://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf