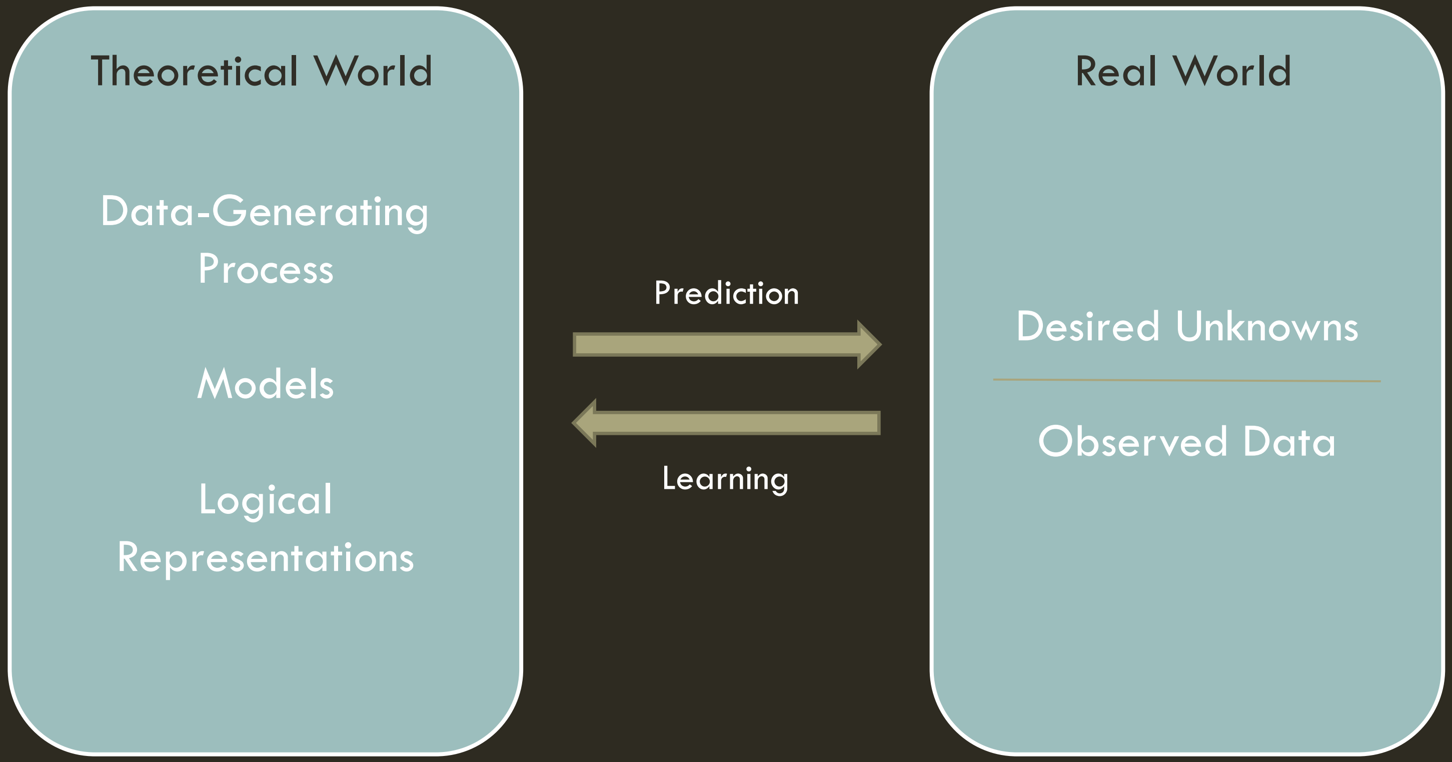
NOT-SO-SIMPLE LINEAR REGRESSION

Ryann Sim
SUTD

CONTENTS

INTRODUCTION

WHY STATISTICS?



WHY STUDY LINEAR REGRESSION AT ALL?

Linear regression is well established historically, but has been overtaken in terms of performance by many newer, flashier models. Why study it at all?

- Many more flexible and useful methods of regression can be seen, mathematically, as extensions or generalizations of linear regression.
- Many ideas behind better methods (such as regularization or cross-validation) can be illustrated using linear models.
- Easy to explain and understand the parameters of a linear regression model.
- Linear regression is part of the shared culture of data analysis, so we should appreciate it!

SIMPLE LINEAR REGRESSION

REGRESSION

Assume we have I.I.D. samples $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n) \sim F_{X,Y}$

X – independent variable (covariate)

Y – dependent variable (response)

Regression Function: $r(x) = \mathbb{E}(Y|X = x)$

Given what we know about the x values, what's our best guess for y?

Regression – estimating r from the given observations

Prediction – estimating Y from a new observation X

THE SIMPLE LINEAR REGRESSION MODEL

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Details:

1. The slope of the line is β_1
2. The y-intercept of the line is β_0
3. Generally, you can check for linearity graphically!

IMPORTANT EXPRESSIONS

Parameters	β_0, β_1, σ
Estimates	$\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}$
Regression Function	$r(x) = \beta_0 + \beta_1 x$
Fitted Line	$\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$
Fitted Values	$\hat{Y}_i = \hat{r}(X_i)$
Residuals	$\hat{\epsilon}_i = Y_i - \hat{Y}_i$
Residual Sum of Squares	$RSS = \sum_i \hat{\epsilon}_i^2$

How far away each prediction is from the real value

Total squared errors made by the linear model

LEAST SQUARES ESTIMATE

The least squares estimates are values of $\hat{\beta}_0, \hat{\beta}_1$ that minimize the RSS.

Theorem. The least squares estimates are given by

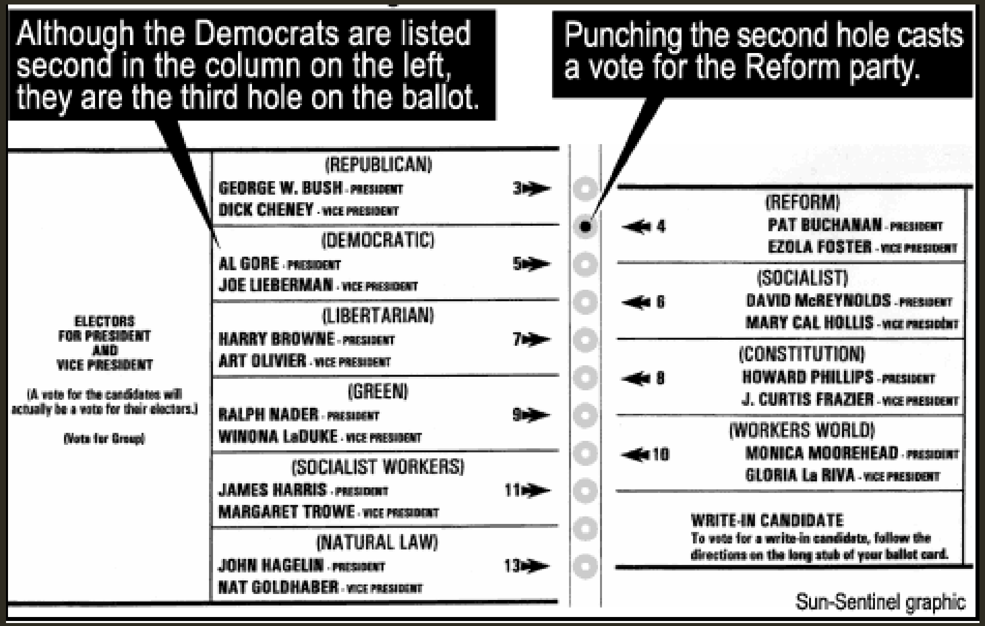
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2} = \frac{Cov(X,Y)}{Var(X)}$$

$$\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n$$

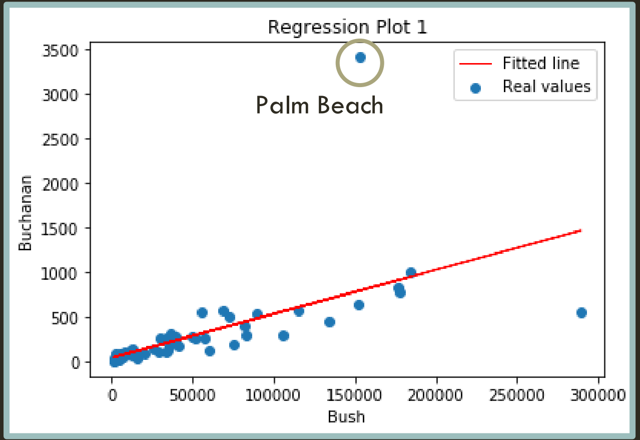
What does this mean? There is a simple formula that can be plugged into any programming language which returns an unbiased estimator for the 'best' fitted line.

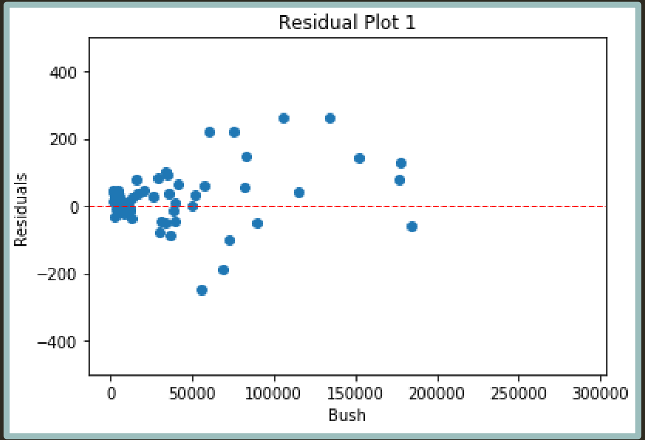
EXAMPLE – ELECTION DATA

- The 2000 US Presidential election (Al Gore vs George Bush) was controversial because its outcome was decided by the results in Florida, where Bush won by only 537 votes out of 6 million cast.
- Critics claim that in Palm Beach, thousands of votes went to Buchanan (Reform Party) instead of Al Gore due to confusing ballot design.
- Let’s check that hypothesis!

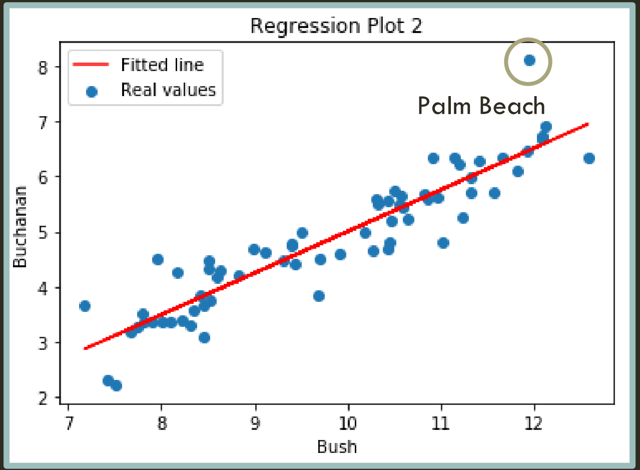


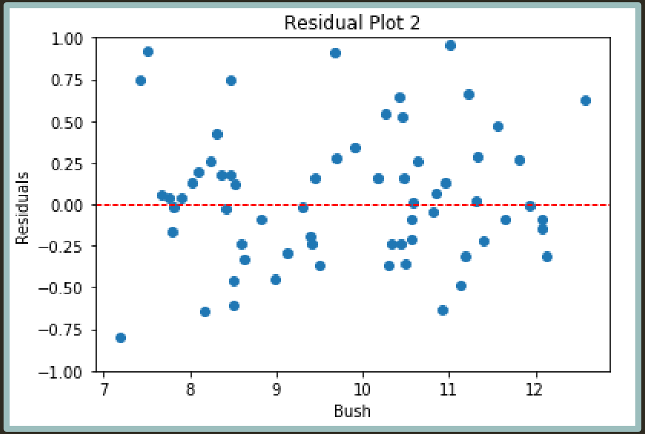
EXAMPLE – ELECTION DATA





$$Buchanan = 0.0049 * Bush + 45.29$$





$$\log(Buchanan) = 0.758 * \log(Bush) - 2.577$$

PREDICTION AND UNCERTAINTY

POINT PREDICTION

Step 1. Training/Fitting/Regression

- Given: Data $(X_1, Y_1), \ldots, (X_n, Y_n)$
- Estimate: Regression function $\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$

Step 2. Testing/Prediction

- Given: New observation x_*
- Estimate: Prediction $\hat{Y}_* = \hat{r}(X_*) = \hat{\beta}_0 + \hat{\beta}_1 x_*$

We assumed that there is a linear relationship between variables, and use our fitted line to ‘guess’ what the y value should be given an x value. How good is this prediction?

PREDICTION INTERVAL

Theorem. Let

$$\hat{\xi}_n^2 = \left(\frac{\sum_{i=1}^n (X_i - X_*)^2}{n \sum_{i=1}^n (X_i - \bar{X})^2} + 1 \right)$$

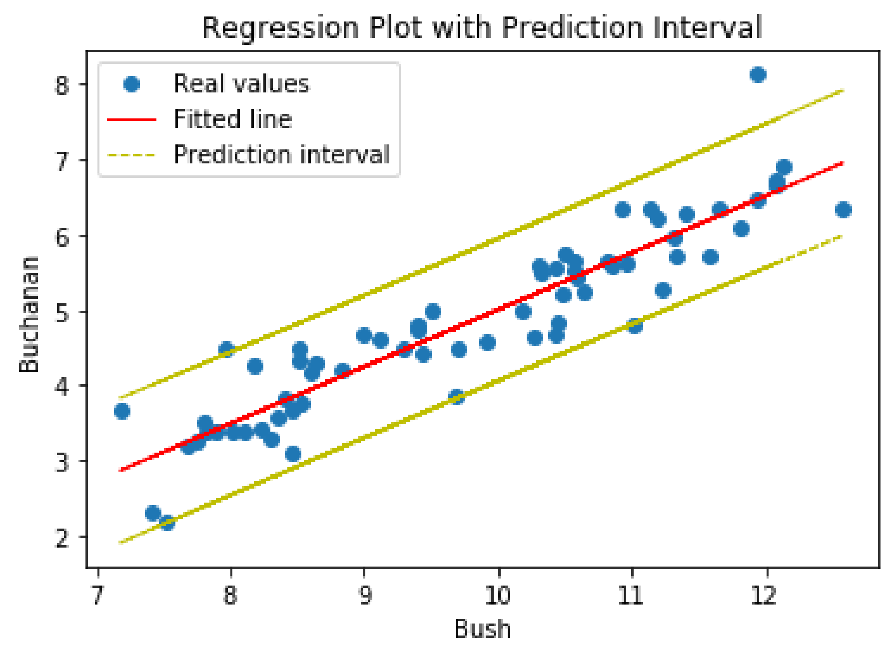
An approximate $1 - \alpha$ prediction interval for Y_* is

$$\hat{Y}_* \pm z_{\alpha/2} \hat{\xi}_n$$

A prediction interval is a set where you can say: 'With α probability, Y_* will be inside the interval'. Again, one can easily implement this in code.

EXAMPLE – ELECTION DATA

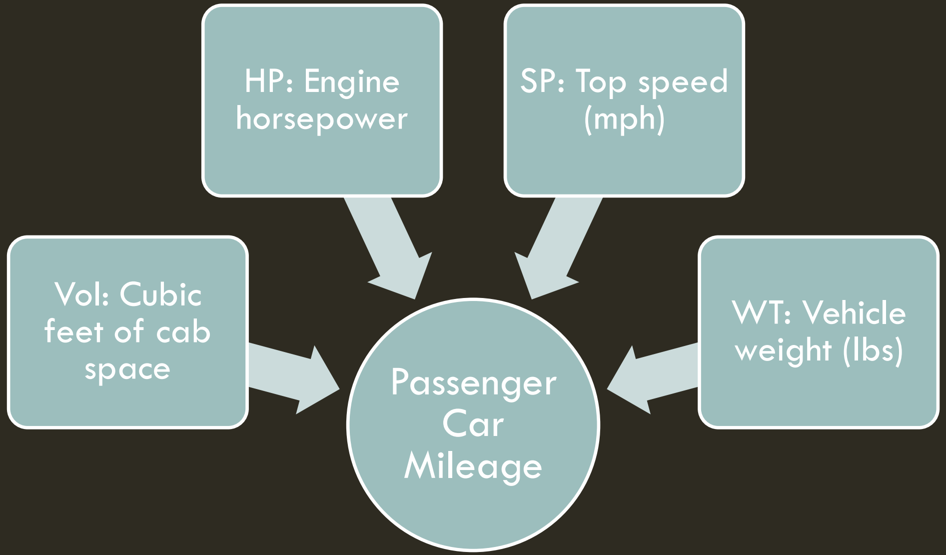
	Predicted	Prediction Interval	Actual Value
Log(# Buchanan Votes)	6.388	(5.200, 7.578)	8.15



MODEL SELECTION

MULTIPLE LINEAR REGRESSION

We have seen linear regression with just one covariate and one response variable. We can extend the definition to multiple covariates!



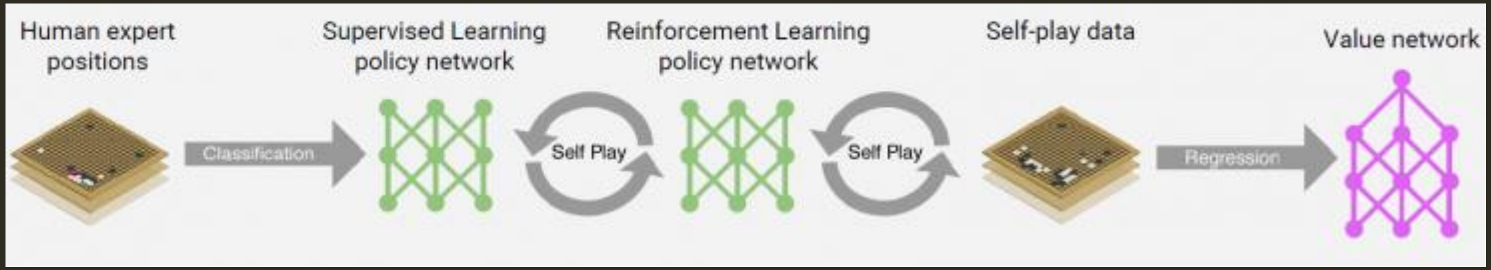
Can we just use all of the covariates in our linear model? Or do we need to be more selective?

EXAMPLE – CRIME RATES

- Should we eliminate some variables from the model? If so, which ones?
- Should we interpret these relationships as causal?
- Watch out for multicollinearity!

Covariate	$\widehat{\beta}_j$	$\widehat{\text{se}}(\widehat{\beta}_j)$	t value	p-value
(Intercept)	-589.39	167.59	-3.51	0.001 **
Age	1.04	0.45	2.33	0.025 *
Southern State	11.29	13.24	0.85	0.399
Education	1.18	0.68	1.7	0.093
Expenditures	0.96	0.25	3.86	0.000 ***
Labor	0.11	0.15	0.69	0.493
Number of Males	0.30	0.22	1.36	0.181
Population	0.09	0.14	0.65	0.518
Unemployment (14–24)	-0.68	0.48	-1.4	0.165
Unemployment (25–39)	2.15	0.95	2.26	0.030 *
Wealth	-0.08	0.09	-0.91	0.367

EXAMPLE – ALPHAGO



Factors considered in model selection

- Number of neurons in each layer
- Activation functions used by each layer
- Batch size, learning rate, exit criteria for training
- And many more!

Previously, we measured a model's ability to fit the data. Now, how can we measure a model's ability to generalize?

TRADE-OFF IN MODEL SELECTION

Simple Model

Few covariates
Few parameters
High bias
Underfitting

Complex Model

Many covariates
Many parameters
High variance
Overfitting

Occam's Razor

"Entities must not be multiplied beyond necessity."

"The simplest explanation is usually the correct one."

COMMON SCORES & METHODS

- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)
- k-Fold Cross Validation
- Forward/Backward Stepwise Regression
- LASSO Regression

Most of these can be implemented easily in Python/R/Matlab!

LASSO

In simple linear regression, we minimize the error given by:

$$\sum_i^n (Y_i - \hat{Y}_i)^2$$

LASSO (Least Absolute Shrinkage and Selection Operator) penalizes complexity by adding a regularization term as follows:

$$min \sum_i^n \overbrace{(Y_i - \hat{Y}_i)}^{\text{How off we were}}{}^2 + \lambda \sum_j^p |\hat{\beta}_j|$$

True value

Model's guess

How harshly we penalize

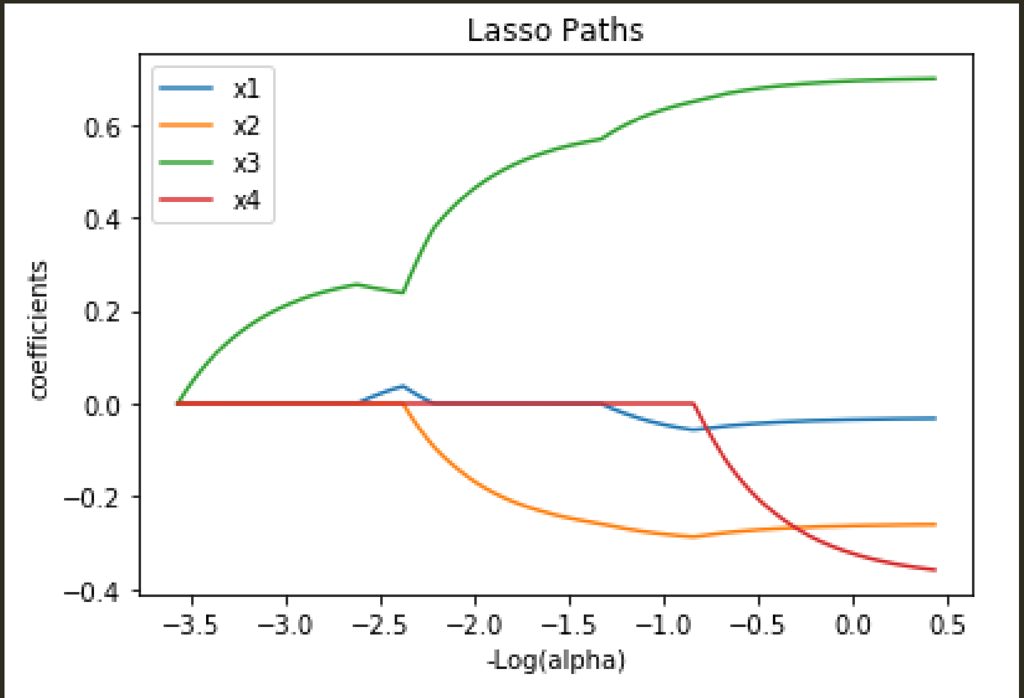
How big the coefs are

EXAMPLE – PASSENGER CAR MILEAGE

OLS Regression Results

Dep. Variable:	MPG	R-squared:	0.883
Model:	OLS	Adj. R-squared:	0.877
Method:	Least Squares	F-statistic:	143.8
Date:	Wed, 15 Jul 2020	Prob (F-statistic):	1.21e-34
Time:	11:30:39	Log-Likelihood:	-214.38
No. Observations:	81	AIC:	438.8
Df Residuals:	76	BIC:	450.7
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	188.7035	22.747	8.296	0.000	143.399	234.008
x1	-0.0121	0.022	-0.549	0.585	-0.056	0.032
x2	0.3806	0.079	4.838	0.000	0.224	0.537
x3	-1.2528	0.237	-5.293	0.000	-1.724	-0.781
x4	-1.8553	0.206	-9.013	0.000	-2.265	-1.445



Alpha = 10

const	x1	x2	x3	x4
63.26	-0.077	-0.092	0	-0.352

EXTENSIONS AND CONCLUSIONS

EXTENSIONS AND EQUIVALENCES

- Logistic regression: Categorical version of linear regression, used for classification problems.
- Generalized Linear Models: Does not assume that the response variable is normally distributed.
- Error Term: Can be generalized as loss/cost function, which is a function to be minimized in machine learning problems.
- Regularization: Often used in modern machine learning.
- Interpretability: LIME

CONCLUSION

Despite being considered somewhat outdated, studying Linear Regression is still important. Some final tips for using regression in practice:

- Understand when and where to use regression as opposed to more complex methods.
- Gain intuition for the mathematics behind regression, and why it works in practice.
- Use model selection techniques when designing models to improve performance.
- Use graphical methods to get a feel for your data before applying regression.

THANK YOU!

Any questions?