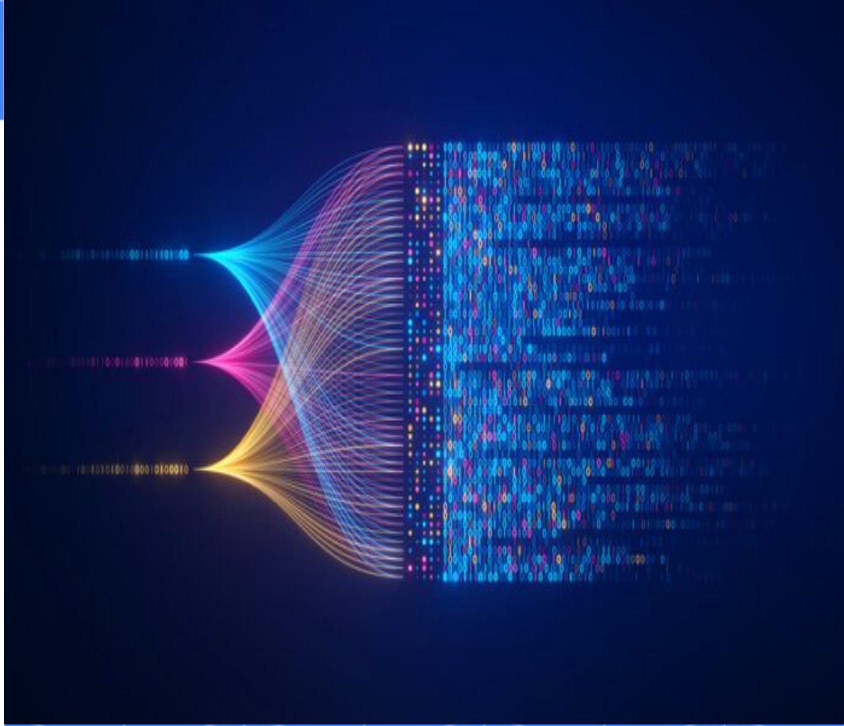
Starting At 6:30 PM

Google Developer Student Clubs
University of Toronto Mississauga

Foundations of Machine Learning: Linear Regression



Presented by;
Hamza & Rahul

Google Developer Student Clubs
University of Toronto Mississauga

Introduction to the 4 part series on ML

```
lookup.KeyValue
f.constant(['em
=tf.constant([0
.lookup.StaticV

_buckets=5)
```

What, who, and when?

- Second of a planned 4 part series on ML
- The next topics are not set in stone so feel free to give us cool ideas
- The next two will also be presented by the two of us
- Dates not yet set in stone, but should be decided soon for workshop 3 so monitor the socials!

Important Notice

- This workshop is largely interactive with a large emphasis on group code along exercises
- So please scan the QR code to join the UTM GDSC discord where you can download the starter code
- Now you should've downloaded 'Linear_Regression_Workshop.ipynb' this is a jupyter notebook and can be opened without an IDE by googling UofT Jupyter Hub or going to jupyter.utoronto.ca/hub/login

SCAN ME

What really is Machine Learning...?

- Essentially just statistical methods that **learn** from a given data set
- Patterns in data **features** are captured in **parameters**
- **NOTE:** Parameters are measurable factors that define a system and determine its behaviour
- More technical definition: Given a simple mathematical objective (eg. the distance between your predictions and the label) we apply iterative mathematical optimization on the parameters to optimize for/minimize this objective

Google Developer Student Clubs
University of Toronto Mississauga

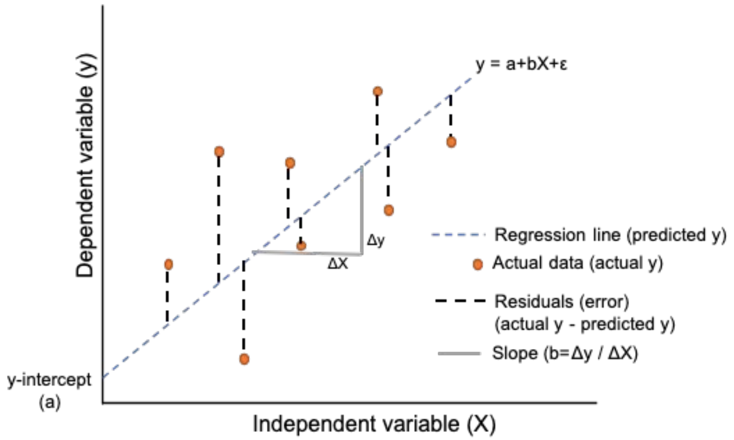
Simple Linear Regression

```
lookup.KeyValue
f.constant(['em
=tf.constant([0
.lookup.StaticV

_buckets=5)
```

What is Linear Regression ?

- Essentially fitting a line through points
- More technically; Models the relationship between variables via a linear equation
- Goal; best line of fit means least deviations from line
- Minimizing the difference between observed and predicted values



Why Linear Regression over more powerful forms?

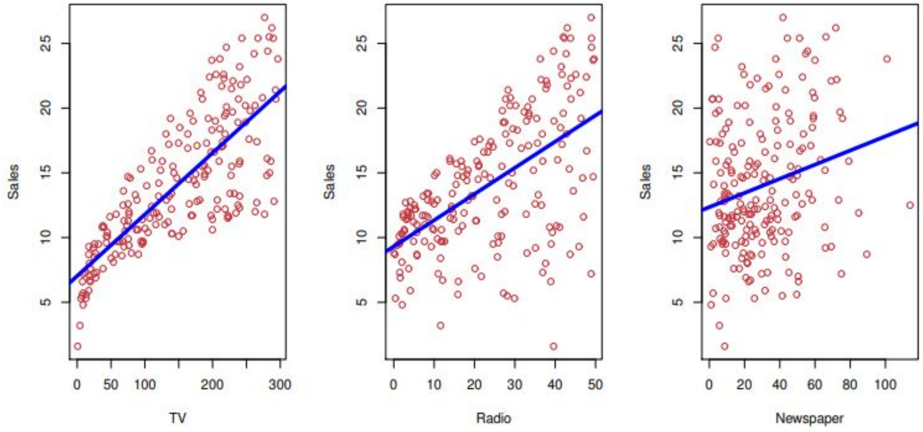
- Simplicity and Interpretability:
  a. Easy to understand and explain, even to non-technical audiences
  b. Clear relationship between independent and dependent variables
- Efficient Computation:
  a. Fast to compute, even with large datasets
  b. Less computationally expensive than many complex models
- Performance in Well-Structured Data:
  a. Highly effective with linear relationships
  b. Can outperform complex models when data structure is straightforward

Why Linear Regression over more powerful forms? - continued

- Less Prone to Overfitting:
    a. Simplicity reduces the risk of overfitting
    b. Easier to generalize to new data
- Good Starting Point for Analysis:
    a. Useful for initial analysis to understand data trends
    b. Provides a baseline for comparison with complex models
- Flexibility with Enhancements:
    a. Can be extended with polynomial and interaction terms
    b. Adaptable for various types of data through transformations

Linear Regression Theory - Form

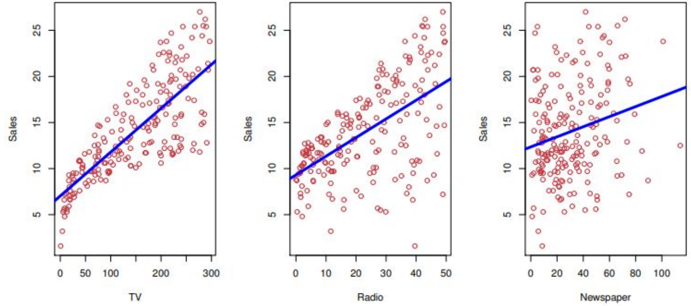
- $Y = \beta_0 + \beta_1 X + \epsilon$
- Y = independent variable
- X = dependant variable
- β_0 = Intercept
- β_1 = Slope
- ϵ = Error term
- Example: $Y = 5 + 7x + \epsilon$



Linear Regression Theory - Form

- Prediction Representation Per Datapoint : $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- Example:

TV (x)	Sales (y)
50	8
75	10



- If Y = 6 + 0.1x (Our prediction for best fit line)
- y-hat = 6 + 0.1*50 = 11 (not accurate, but this is what our equation gives us)

Linear Regression Theory - RSS

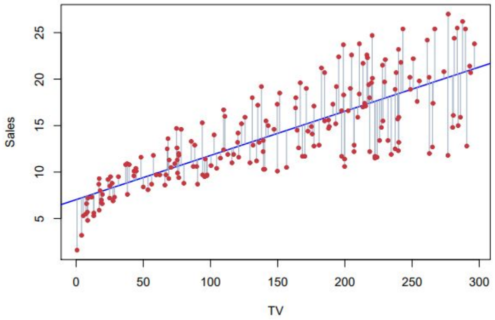
- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X. Then $e_i = y_i - \hat{y}_i$ represents the ith *residual*
- We define the *residual sum of squares* (RSS) as

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

or equivalently as

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

TV (x)	Sales (y)
50	8
75	10



Lab Exercise

10 min

Collaboration is encouraged!

Google Developer Student Clubs
University of Toronto Mississauga

Standard Error and Confidence Intervals

Standard Error

- The standard error of an estimator reflects how it varies under repeated sampling. We have

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

where $\sigma^2 = \mathrm{Var}(\epsilon)$

$$\sigma^2 = \frac{SSR}{n-p} = \frac{\sum_{i=1}^{n}(e_i)^2}{n-p}$$

Confidence Intervals

- These standard errors can be used to compute *confidence intervals.* A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter. It has the form

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1).$$

Confidence Intervals - continued

That is, there is approximately a 95% chance that the interval

$$\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\right]$$

will contain the true value of β_1 (under a scenario where we got repeated samples like the present sample)

For the advertising data, the 95% confidence interval for β_1 is $[0.042, 0.053]$

Lab Exercise

5 min

Collaboration is encouraged!

Google Developer Student Clubs
University of Toronto Mississauga

Hypothesis Testing

```
lookup.KeyValue
f.constant(['em
=tf.constant([0
.lookup.StaticV

_buckets=5)
```

What is Hypothesis Testing?

- Statistical tool used to make inferences about a population based on sample data
- Standard Errors are used to perform hypothesis tests on the coefficients
- The most common hypothesis test involves testing the null hypothesis...

Null Hypothesis

H_0 : There is no relationship between X and Y

versus the *alternative hypothesis*

H_A : There is some relationship between X and Y.

- Mathematically, this corresponds to testing

$$H_0 : \beta_1 = 0$$

versus

$$H_A : \beta_1 \neq 0,$$

since if $\beta_1 = 0$ then the model reduces to $Y = \beta_0 + \epsilon$, and X is not associated with Y.

t-statistic

- To test the null hypothesis, we compute a *t-statistic*, given by

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)},$$

- This will have a t-distribution with $n - 2$ degrees of freedom, assuming $\beta_1 = 0$.

p-value

- Using statistical software, it is easy to compute the probability of observing any value equal to $|t|$ or larger. We call this probability the *p-value*.

	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	7.0325	0.4578	15.36	< 0.0001
`TV`	0.0475	0.0027	17.67	< 0.0001

Lab Exercise

10 min

Collaboration is encouraged!

Google Developer Student Clubs
University of Toronto Mississauga

RSE, R^2, F-statistic

RSE & R^2

- We compute the *Residual Standard Error* Used to calculate the lack of fit

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

where the *residual sum-of-squares* is $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

- *R-squared* or fraction of variance explained is

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

where $\mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the *total sum of squares*.

RSE = actual values deviate from the true regression line by approximately RSE units
TSS = measures the total variance in the response variable that we're trying to explain
RSS = measures the variance in the response variable that the model does not explain

F-statistic

- Definition: The F statistic is a ratio used in statistical testing to compare the model's fit against a model with no predictors.
- Purpose: It tests whether the group of variables in a model are jointly significant. Essentially, it asks, "Is the model better than nothing?"

F-statistic - continued

- Interpretation:
  a. A higher F statistic indicates that the model explains a significant amount of variance in the dependent variable.
  b. A lower F statistic suggests the model does not provide a better fit than one without independent variables.
- f_{value} = ((TSS - RSS) / (p-1)) / (RSS / (n - p))

Lab Exercise

10 min

Collaboration is encouraged!

Google Developer Student Clubs
University of Toronto Mississauga

Multiple Linear Regression

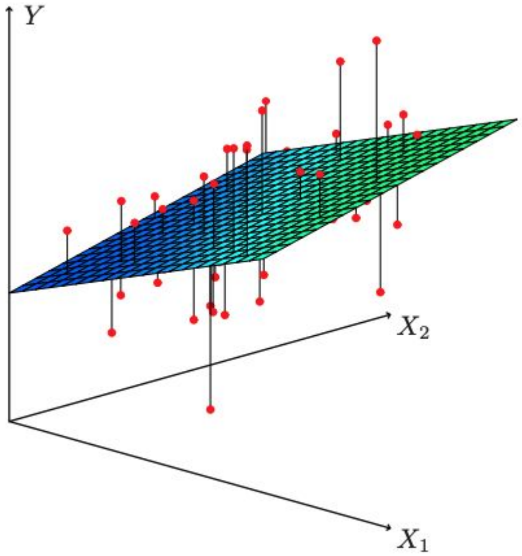
Multiple Linear Regression

- Here our model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

- We interpret β_j as the *average* effect on Y of a one unit increase in X_j, *holding all other predictors fixed*. In the advertising example, the model becomes

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon.$$



Interpreting Regression Coefficients

- The ideal scenario is when the predictors are uncorrelated — a *balanced design*:
  - Each coefficient can be estimated and tested separately.
  - Interpretations such as *"a unit change in X_j is associated with a β_j change in Y, while all the other variables stay fixed"*, are possible.
- Correlations amongst predictors cause problems:
  - The variance of all coefficients tends to increase, sometimes dramatically
  - Interpretations become hazardous — when X_j changes, everything else changes.

Estimation and Prediction

- Given estimates $\hat\beta_0, \hat\beta_1, \ldots \hat\beta_p$, we can make predictions using the formula

$$\hat{y} = \hat\beta_0 + \hat\beta_1 x_1 + \hat\beta_2 x_2 + \cdots + \hat\beta_p x_p.$$

- We estimate $\beta_0, \beta_1, \ldots, \beta_p$ as the values that minimize the sum of squared residuals

$$\begin{aligned} \mathrm{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat\beta_0 - \hat\beta_1 x_{i1} - \hat\beta_2 x_{i2} - \cdots - \hat\beta_p x_{ip})^2. \end{aligned}$$

This is done using standard statistical software.
The values $\hat\beta_0, \hat\beta_1, \ldots, \hat\beta_p$ that minimize RSS are the multiple least squares regression coefficient estimates.

Lab Exercise

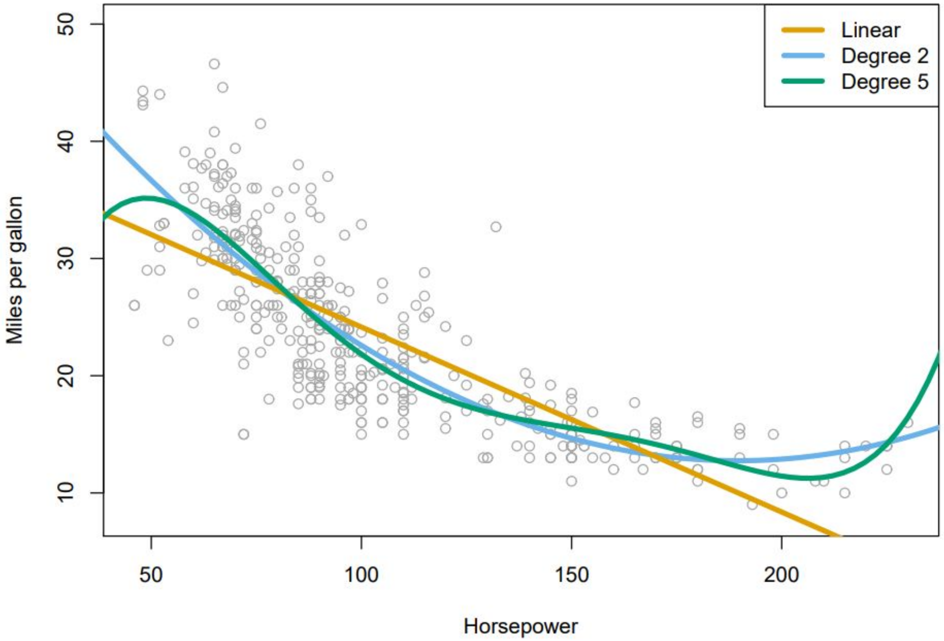
10 min

Collaboration is encouraged!

Google Developer Student Clubs
University of Toronto Mississauga

Polynomial Regression

Polynomial Regression Figure



Polynomial Regression

The figure suggests that

$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$

may provide a better fit.

	Coefficient	Std. Error	t-statistic	p-value
Intercept	56.9001	1.8004	31.6	< 0.0001
horsepower	-0.4662	0.0311	-15.0	< 0.0001
$\texttt{horsepower}^2$	0.0012	0.0001	10.1	< 0.0001

Lab Exercise (HW)

10 min

Collaboration is encouraged!

Google Developer Student Clubs
University of Toronto Mississauga

Thank you!

```
lookup.KeyValue
f.constant(['em
=tf.constant([0
.lookup.StaticV

_buckets=5)
```