https://ttpoll.com/p/911058

# COSC 325: Introduction to Machine Learning

Dr. Hector Santos-Villalobos

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

# Lecture 07: Linear Regression

THE UNIVERSITY OF TENNESSEE KNOXVILLE

# Class Announcements

## Homework:

Reduced homework assignments from seven to six.

Homework #2 is due this Sunday.

## Course Project:

Check groups in Canvas.

PRFAQ is due 09/19 (10 days)

## Lectures:

Absences: In your email's subject, include the following text "[COSC325 ABSENCE]"

## Exams:

I had to move things around due to exam policy 5 days before study day. Please check the new schedule.

Exam #1: Thursday, 10/03

Exam #2:  Thursday, 11/21

What: UTK Machine Learning Club

Where: **MK 525**

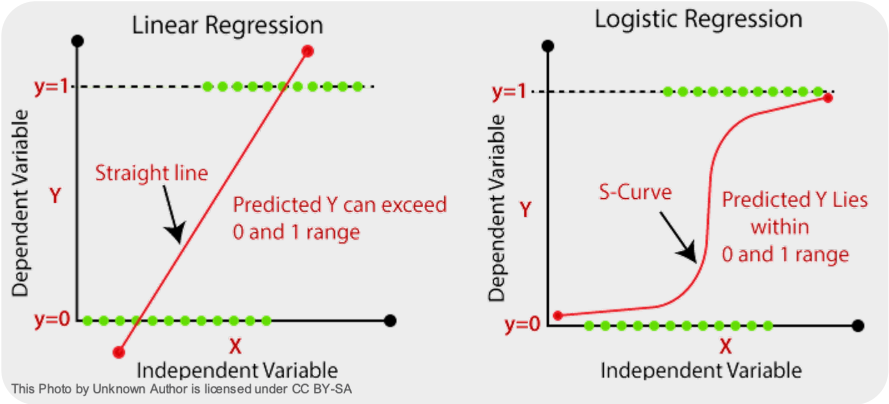
When: **Tuesday** at **5:00**
(including today)

Who: Any experience level

Everyone is welcome to the first meeting of ML club today. Whether you are a beginner looking to learn from our intro to ML lesson series, experienced practitioner who wants to learn from and discuss with other enthusiasts in our reading groups, or you just want to hear from our industry guest speakers and seminars, utkML can help you scratch your machine learning itch!

# Today’s Topics

*Regression Techniques*

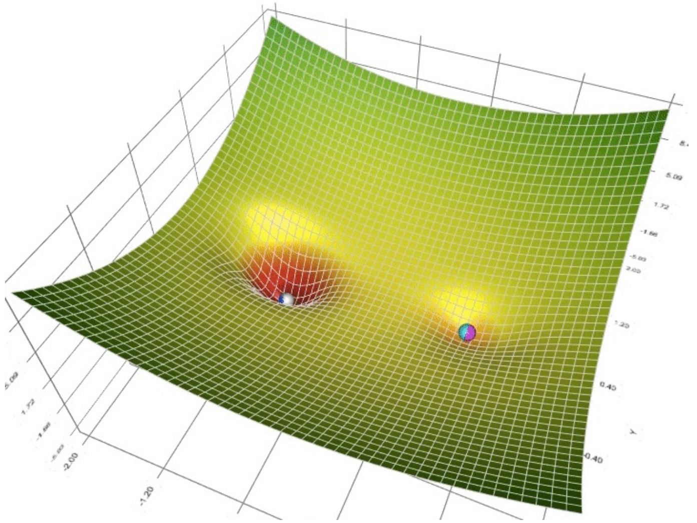


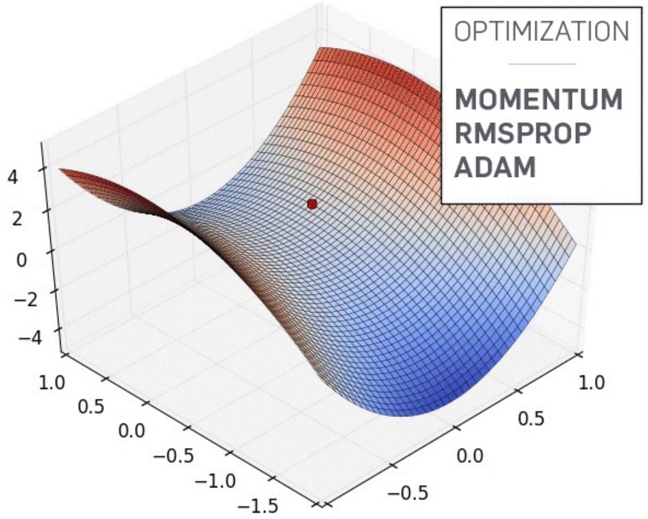
This Photo by Unknown Author is licensed under CC BY-SA

# Last Lecture

- Gradient descent algorithm and concepts
  - Convexity, learning rate, saddle point, global vs. local minimum, etc.
- Derivatives measure the influence of a variable on the function output.
- Computational graphs and the chain rule
- Linear regression
  - Close form vs. GD solutions
  - Python implementation.

$$\theta = (X^T X)^{-1} X^T y$$





OPTIMIZATION

MOMENTUM
RMSPROP
ADAM

# Linear Regression

https://ttpoll.com/p/911058

# Gradient Descent for Linear Regression and MSE Cost Function

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_m x_m$$

```
$X \coloneqq$ data features
$y \coloneqq$ data targets
$\theta = \theta_0$
Repeat:
        $\hat{y} = h_\theta(X)$
        $cost = J_\theta(y, \hat{y})$
        $d\theta = \frac{\partial J_\theta(y,\hat{y})}{\partial \theta}$
        $\theta \coloneqq \theta - \alpha(d\theta)$
Until a fixed number of iterations or $d\theta$ very small.
```

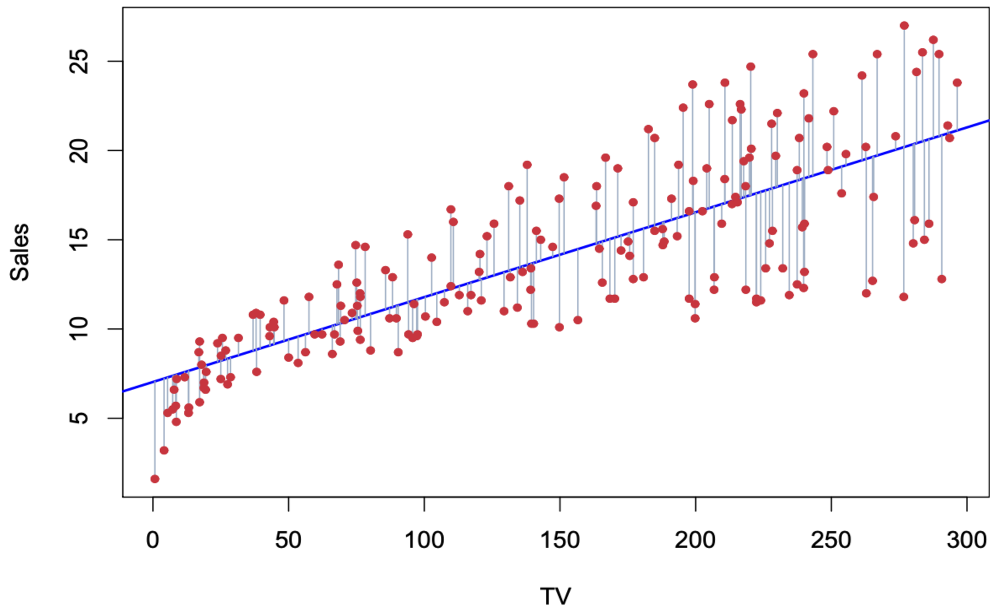
First column: [ones].

$$\hat{y} = h_\theta(X) = X\theta$$

$$J(y, \hat{y}) = \frac{1}{n}\Sigma\left(\hat{y}^{(i)} - y^{(i)}\right)^2$$

$$\frac{dJ}{d\theta} = \frac{2}{n}\left(X^T(\hat{y} - y)\right)$$

Advertising Data*



*Source: James, et. al., An Intro to Statistical Learning, 2023.

https://ttpoll.com/p/911058

# Notebook Time

# Assessing the Accuracy of the Coefficients

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

$$SE(\theta_0)^2 = \frac{\sigma^2}{\sum(x^{(i)} - E[x])^2} \qquad SE(\theta_1)^2 = \frac{1}{n} + \frac{E[x]^2}{\sum(x^{(i)} - E[x])^2} \qquad \sigma = Var(\epsilon)$$

- These standard errors can be used to compute 95% confidence intervals. Confidence intervals 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

$$\theta_k \pm 2 \cdot SE(\theta_k)$$

Assumes $\hat{y}^{(i)} = \theta_0 + \theta_1 x_1^{(i)}$

*Slide Credit: James, et. al., An Intro to Statistical Learning, 2023.*

# Confidence Intervals

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

$$[\theta_1 - 2 \cdot SE(\theta_1), \theta_1 + 2 \cdot SE(\theta_1)]$$

will contain the actual value of $\theta_1$.

For the advertising data, the 95% confidence interval for $\theta_1$ is $[0.042, 0.053]$.

*Slide Credit: James, et. al., An Intro to Statistical Learning, 2023.*

# Hypothesis Testing of Coefficients.

- Standard errors can also be used to perform hypothesis tests on the coefficients.
- The most common hypothesis test involves testing the null hypothesis of
  - $H_0$: 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
  - $H_0 : \theta_1 = 0$ versus
  - $H_A : \theta_1 \neq 0$,

since if $\theta_1 = 0$ then the model reduces to $y = \theta_0 + \epsilon$, and $X$ is not associated with $y$.

Slide Credit: James, et. al., An Intro to Statistical Learning, 2023.

# Hypothesis Testing (Continue)

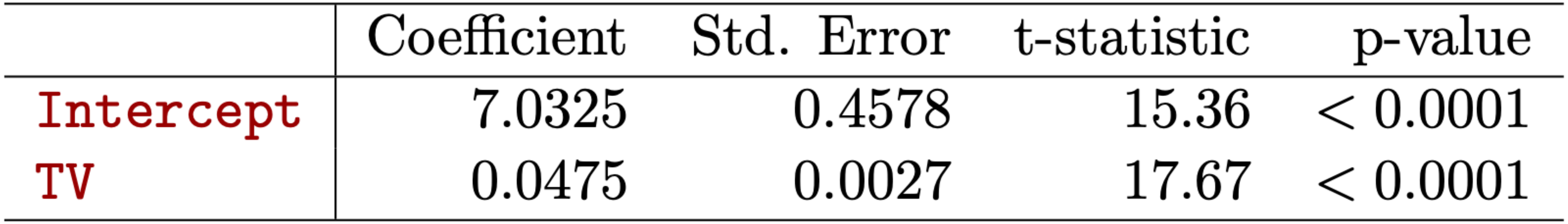
- To test the null hypothesis, we compute a ***t-statistic***, given by

$$t = \frac{\theta_1 - 0}{SE(\theta_1)}$$

- This will have a t-distribution with $n - 2$ degrees of freedom, assuming $\theta_1 = 0$.
- 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***.

*Slide Credit: James, et. al., An Intro to Statistical Learning, 2023.*

# Hypothesis Testing Advertising Model



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

*Slide Credit: James, et. al., An Intro to Statistical Learning, 2023.*

https://ttpoll.com/p/911058

# Notebook Time

# Multiple Linear Regression

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_m x_m$$

We interpret $\theta_j$ as the average effect on $\hat{y}$ of a one unit increase in $x_j$, holding all other predictors fixed.

- The ideal scenario is when the predictors are uncorrelated
  - Each coefficient can be estimated and tested separately.
- Correlations amongst predictors cause problems:
  - The variance of all coefficients tends to increase.
  - Interpretations become hazardous—when $x_j$ changes, everything else changes.
- ***Claims of causality*** should be avoided for observational data.

*Slide Credit: James, et. al., An Intro to Statistical Learning, 2023.*

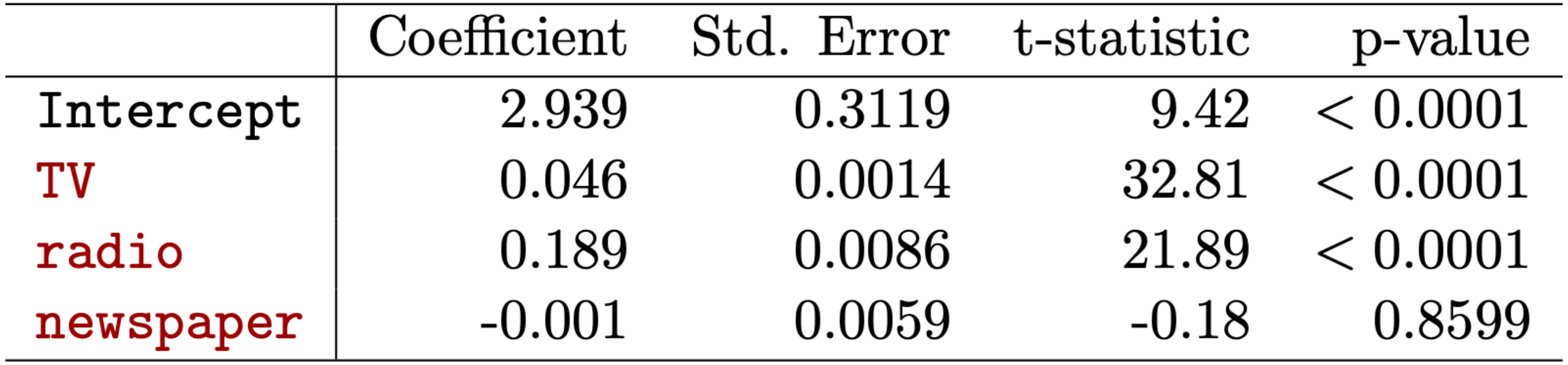
# Predictors usually change together!

- Example #1: $y$ is the total amount of change in your pocket; $\mathrm{x}_1$ = # of coins; $x_2$ = # of pennies, nickels, and dimes. By itself, the regression coefficient of $y$ on $x_2$ will be $> 0$. But how about with $x_1$ in the model?

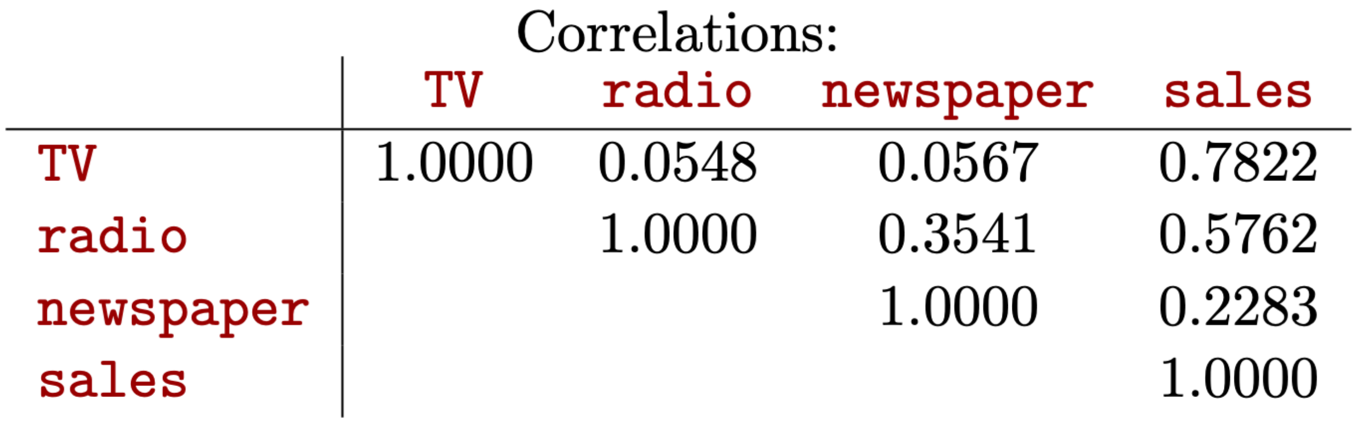
- Example #2: $y$ = # of tackles by a football player in a season; $w$ and $h$ are his weight and height. Fitted regression model is $\hat{y} = \theta_0 + 0.5w - 0.1h$. How do we interpret $\theta_2 < 0$?

*Slide Credit: James, et. al., An Intro to Statistical Learning, 2023.*

# Results for advertising data



| | Coefficient | Std. Error | t-statistic | p-value |
|---|---:|---:|---:|---:|
| Intercept | 2.939 | 0.3119 | 9.42 | $< 0.0001$ |
| TV | 0.046 | 0.0014 | 32.81 | $< 0.0001$ |
| radio | 0.189 | 0.0086 | 21.89 | $< 0.0001$ |
| newspaper | -0.001 | 0.0059 | -0.18 | 0.8599 |



Correlations:

| | TV | radio | newspaper | sales |
|---|---:|---:|---:|---:|
| TV | 1.0000 | 0.0548 | 0.0567 | 0.7822 |
| radio | | 1.0000 | 0.3541 | 0.5762 |
| newspaper | | | 1.0000 | 0.2283 |
| sales | | | | 1.0000 |

*Slide Credit: James, et. al., An Intro to Statistical Learning, 2023.*

# Qualitative Features

- Example: investigate differences in credit card balance between males and females, ignoring the other variables.
  - We map gender to zero/one (male=1, female=0)
  - $y^{(i)} = \theta_0 + \theta_1 x^{(i)}$

- Resulting model

$$y^{(i)} = \begin{cases} \theta_0 + \theta_1, & if \quad ith\, person\, male \\ \theta_0 & if \quad ith\, person\, female \end{cases}$$

*Slide Credit: James, et. al., An Intro to Statistical Learning, 2023.*

# Qualitative Features

- Example: investigate differences in credit card balance between Asian, Caucasian, and African American.
  - We create an extra dummy variables
    - Dummy variables = Qualitative classes – 1
    - Variable 1: Asian
    - Variable 2: Caucasian
    - Baseline: African American

1s for Asian samples, 0s otherwise

$$y = \theta_0 + \theta_1 x_{1,1} + \theta_2 x_{1,2}$$

1s for Caucasian samples, 0s otherwise

Slide Credit: James, et. al., An Intro to Statistical Learning, 2023.

# Qualitative Features

- Example: investigate differences in credit card balance between Asian, Caucasian, and African American.

$$y = \theta_0 + \theta_1 x_{1,1} + \theta_2 x_{1,2}$$

- Resulting model

$$y = \begin{cases} \theta_0 + \theta_1, & \text{if ith person is Asian} \\ \theta_0 + \theta_2, & \text{if ith person is Caucasian} \\ \theta_0, & \text{if ith person is African American} \end{cases}$$

Baseline

| | Coefficient | Std. Error | t-statistic | p-value |
|---|---|---|---|---|
| `Intercept` | 531.00 | 46.32 | 11.464 | $< 0.0001$ |
| `ethnicity[Asian]` | -18.69 | 65.02 | -0.287 | 0.7740 |
| `ethnicity[Caucasian]` | -12.50 | 56.68 | -0.221 | 0.8260 |

*Slide Credit: James, et. al., An Intro to Statistical Learning, 2023.*

# Extensions to Linear Model

- Removing the additive assumption: ***interactions*** and ***nonlinearity***
- Interactions:
  - In our previous *Advertising* data analysis, we assumed that the effect on sales of increasing one advertising medium is independent of the amount spent on the other media.
- For example, the linear model

$$sales = \theta_0 + \theta_1 TV + \theta_2 radio + \theta_3 \, newspaper$$

states that the average effect on sales of a one-unit increase in TV is always $\theta_1$, regardless of the amount spent on radio.

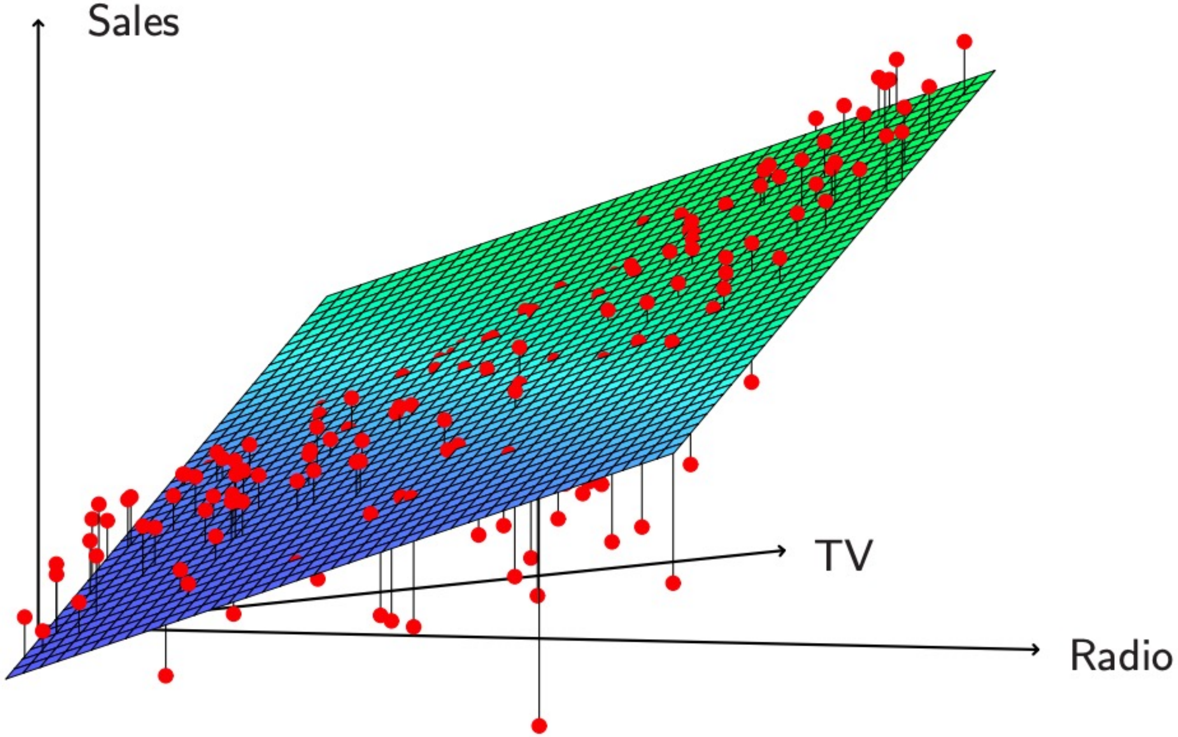
Slide Credit: James, et. al., An Intro to Statistical Learning, 2023.

# Interactions

- However, spending money on radio advertising increases the effectiveness of TV advertising, so the slope term for TV should increase as radio advertising increases.
- Given a fixed budget of $100,000, spending half on radio and half on TV may increase sales more than allocating the entire amount to TV or radio.
- In marketing, this is known as a ***synergy*** effect; in statistics, it is referred to as an ***interaction*** effect.

*Slide Credit: James, et. al., An Intro to Statistical Learning, 2023.*

https://ttpoll.com/p/911058



- When levels of either **TV** or **radio** are low, then the true **sales** are lower than predicted by the linear model.
- But when advertising is split between the two media, then the model tends to underestimate **sales**

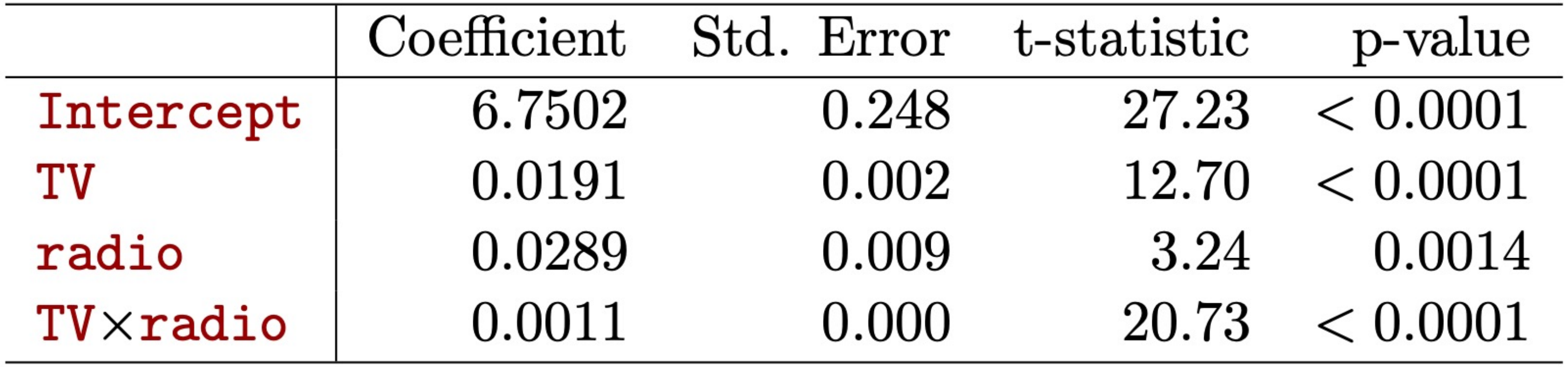
*Slide Credit: James, et. al., An Intro to Statistical Learning, 2023.*

# Modeling Interactions

- Model takes the form

$$\begin{aligned} sales &= \theta_0 + \theta_1 TV + \theta_2 radio + \theta_3 (radio \times TV) \\ &= \theta_0 + (\theta_1 + \theta_3 radio) TV + \theta_2 radio \end{aligned}$$

- Results:



| | Coefficient | Std. Error | t-statistic | p-value |
|---|---|---|---|---|
| Intercept | 6.7502 | 0.248 | 27.23 | $< 0.0001$ |
| TV | 0.0191 | 0.002 | 12.70 | $< 0.0001$ |
| radio | 0.0289 | 0.009 | 3.24 | 0.0014 |
| TV×radio | 0.0011 | 0.000 | 20.73 | $< 0.0001$ |

*Slide Credit: James, et. al., An Intro to Statistical Learning, 2023.*

# Interpretation

- The results in this table suggest that interactions are essential.
- The ***p-value*** for the interaction term $\boldsymbol{TV \times radio}$ is extremely low, indicating that there is strong evidence for

$$HA : \theta_3 \neq 0$$

- The $R^2$ for the interaction model is 96.8%, compared to only 89.7% for the model that predicts sales using $\boldsymbol{TV}$ and $\boldsymbol{radio}$ without an interaction term.

*Slide Credit: James, et. al., An Intro to Statistical Learning, 2023.*

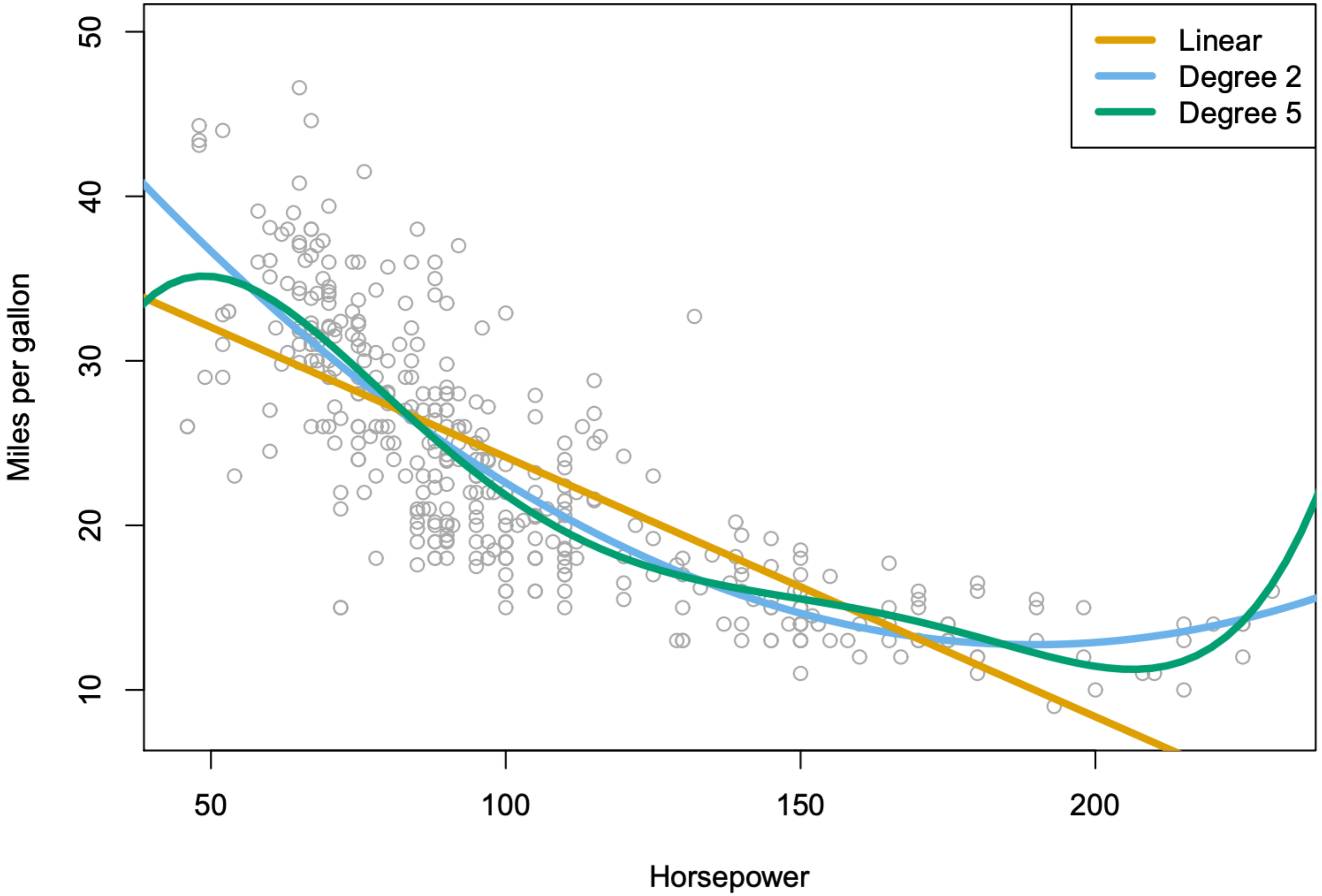
# Interpretation (Continue)

- This means that $\frac{(96.8-89.7)}{100-89.7} = 69\%$ of the variability in sales that remains after fitting the additive model has been explained by the interaction term.
- The coefficient estimates in the table suggest that an increase in TV advertising of $1, 000 is associated with increased sales of
$$(\theta_1 \ + \theta_3 radio) \cdot 1000 = 19 + 1.1 radio\, \boldsymbol{units}$$
- An increase in radio advertising of $1, 000 will be associated with an increase in sales of $(\theta_2 \ + \theta_3 TV) \cdot 1000 = 29 + 1.1\, TV\, \boldsymbol{units}$.

*Slide Credit: James, et. al., An Intro to Statistical Learning, 2023.*

https://ttpoll.com/p/911058

# Polynomial Regression



Auto Dataset

*Slide Credit: James, et. al., An Intro to Statistical Learning, 2023.*

https://ttpoll.com/p/911058

# Matrix Design

- We have an input matrix $X$ with shape $(n, m)$
- We want to fit a polynomial of degree $d$
- Polynomial feature extraction process
  - Columns for each feature polynomial power (e.g., $x_1^3, x_4^5$)
  - Plus, columns for each feature interaction up to $d - 1$(e.g., $x_1 x_4, x_1^2 x_4$)
- Example for data with $\mathrm{n}$ samples, $\mathrm{m} = 2$ features, and polynomial degree $d = 3$.

$$X_{new} = [1 \quad x_1 \quad x_2 \quad x_1^2 \quad x_1 x_2 \quad x_2^2 \quad x_1^3 \quad x_1^2 x_2 \quad x_1 x_2^2 \quad x_2^3]$$

- Then, apply linear regression algorithm on $X_{new}$

Slide Credit: James, et. al., An Intro to Statistical Learning, 2023.

# Example

- $X = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, and we want to fit a polynomial of order $d = 2$

$$X_t = \begin{bmatrix} 1 & x_1 & x_2 & x_1^2 & x_1 x_2 & x_2^2 \end{bmatrix}$$

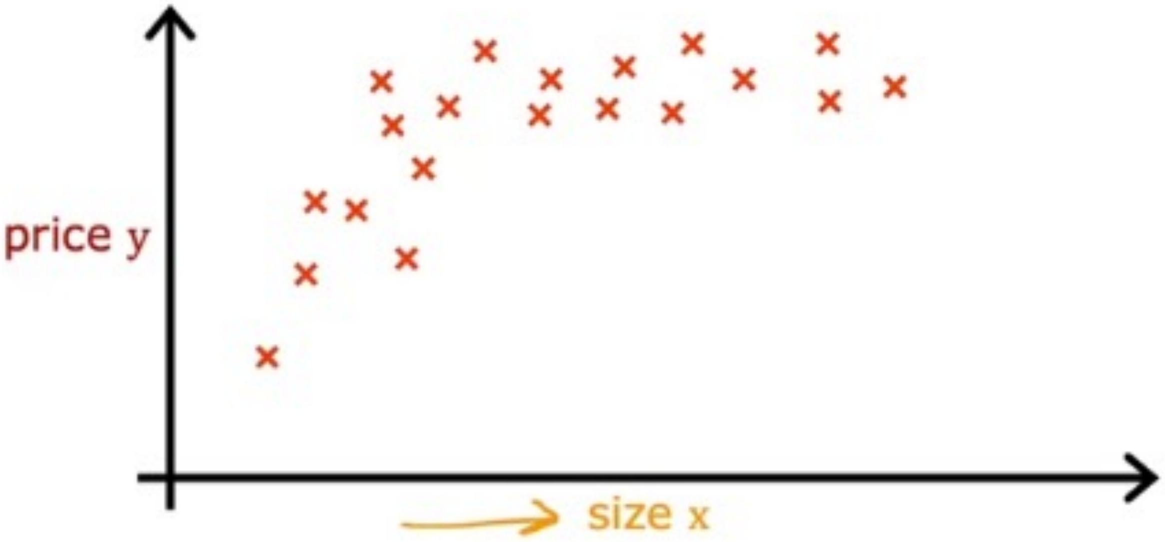
***Caution:*** $X$ # of columns given by $\frac{(m+d)!}{m!d!}$

$$X_t = \begin{bmatrix} 1 & 1 & 2 & 1 & 2 & 4 \\ 1 & 3 & 4 & 9 & 12 & 16 \end{bmatrix}$$

https://ttpoll.com/p/911058

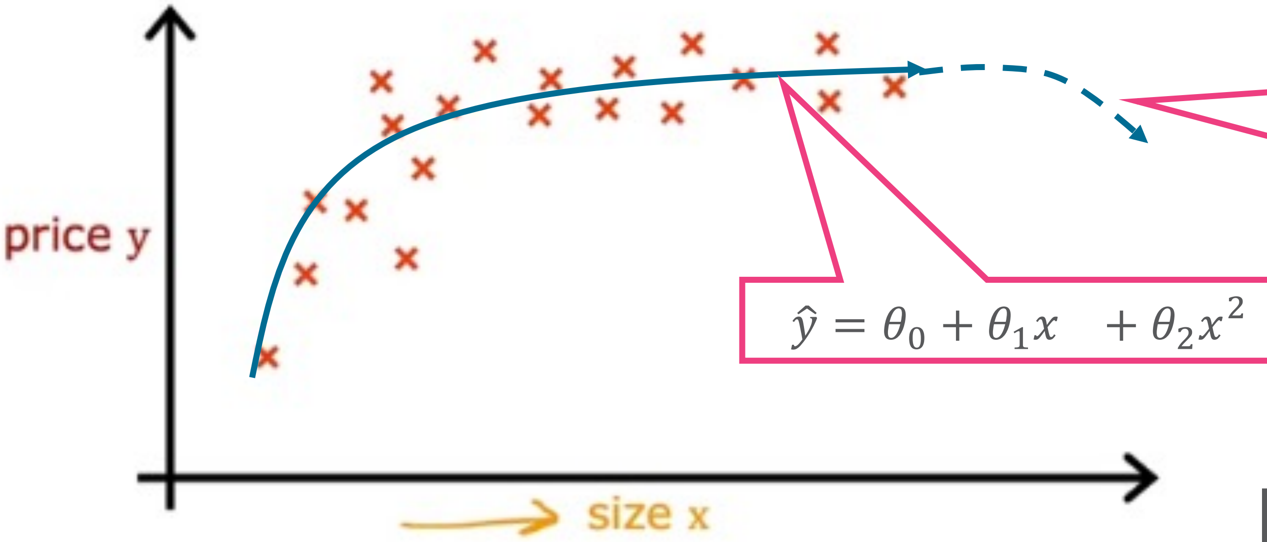
# Thinking outside the box

- Look at your data
- Learn about the domain
- Use equations that match your data and domain knowledge



# Thinking outside the box

- Look at your data
- Learn about the domain
- Use equations that match your data and domain knowledge

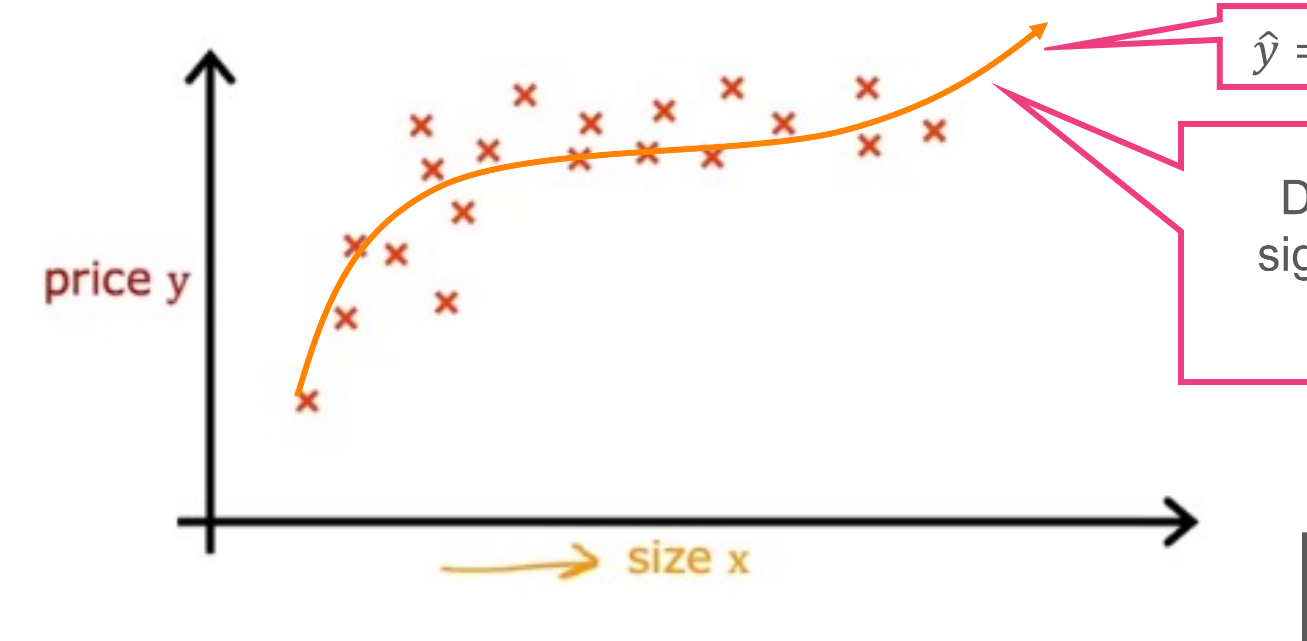


$\hat{y} = \theta_0 + \theta_1 x \quad + \theta_2 x^2$

Well, not the best choice given prices should not go down as the house size increases.

# Thinking outside the box

- Look at your data
- Learn about the domain
- Use equations that match your data and domain knowledge

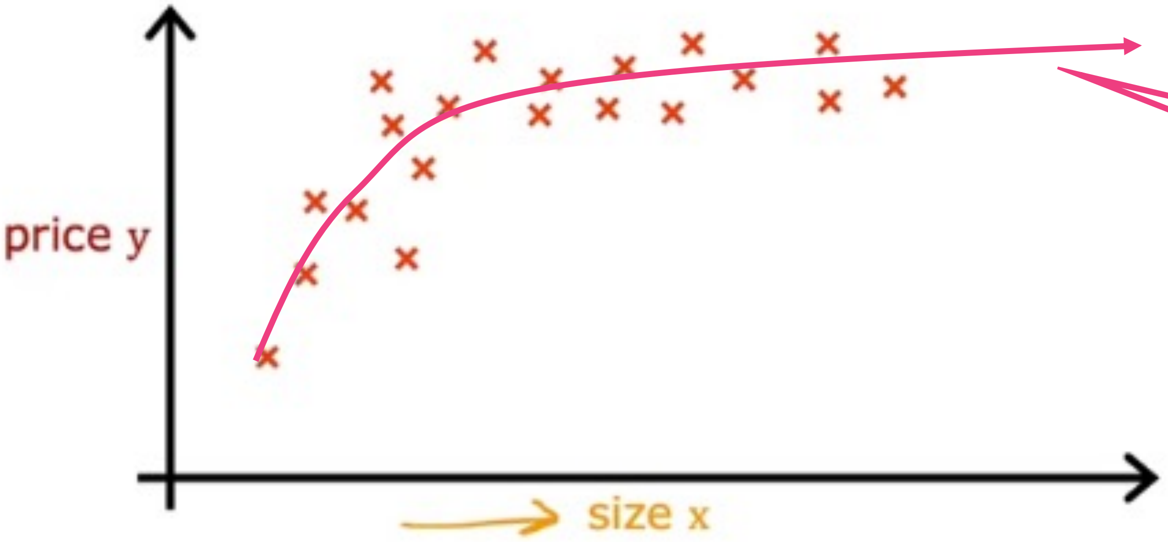


$\hat{y} = \theta_0 + \theta_1 x \quad + \theta_2 x^2 \; + \theta_3 x^3$

Do prices increase significantly more for larger houses?

# Thinking outside the box

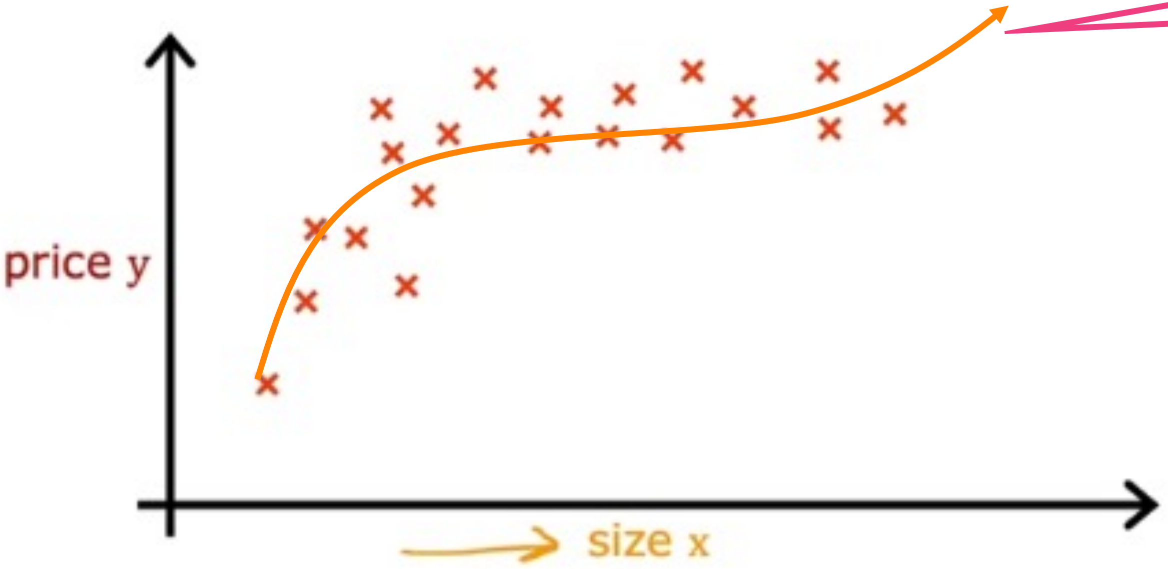
- Look at your data
- Learn about the domain
- Use equations that match your data and domain knowledge



$$\hat{y} = \theta_0 + \theta_1 x \quad + \theta_2 \sqrt{x}$$

# Thinking outside the box

- Look at your data
- Learn about the domain
- Use equations that match your data and domain knowledge



$$\boxed{\hat{y} = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3}$$

| size | size$^2$ | size$^3$ |
|---|---|---|
| $1 - 10^3$ | $1 - 10^6$ | $1 - 10^9$ |

***Feature scaling***

https://ttpoll.com/p/911058

# Notebook Time

# Review

- Regression Techniques
  - Linear
    - Parameter confidence intervals
    - Parameter hypothesis testing
    - Interactions
    - Polynomial regression

# Next Lectures

- Logistic Regression
- Logistic Regression Loss
- Overfitting/Underfitting
- Bias-Variance Tradeoff
- Regularization