



DS-UA 112

Introduction to Data Science

Week 11: Lecture 1

Features - Extending Linear Models





How can we model non-linear trends with linear regression?

DS-UA 112

Introduction to Data Science

Week 11: Lecture 1

Features - Extending Linear Models

Adapted from Nolan, Speed, Gonzalez, Lau



Announcements

- ▶ Please check Week 11 agenda on NYU Classes
 - ▶ Lab 10
 - ▶ Due on Friday April 10 at 11:59PM EST
 - ▶ Homework 4
 - ▶ Due on Saturday April 18 at 11:59PM EST
 - ▶ Survey
 - ▶ [Link to Qualtrics](#)



Review

Question 1

Correlation measures the strength of the _____ between two quantitative variables

- ☐ association
- ☐ relationship
- ☐ linear association
- ☐ causal relationship

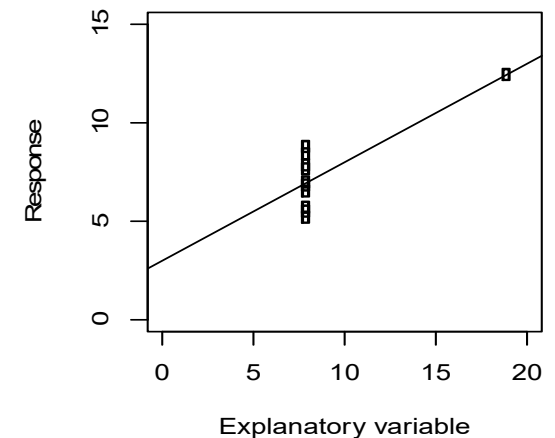
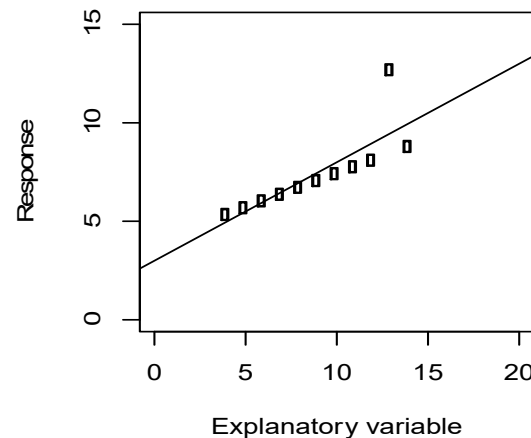
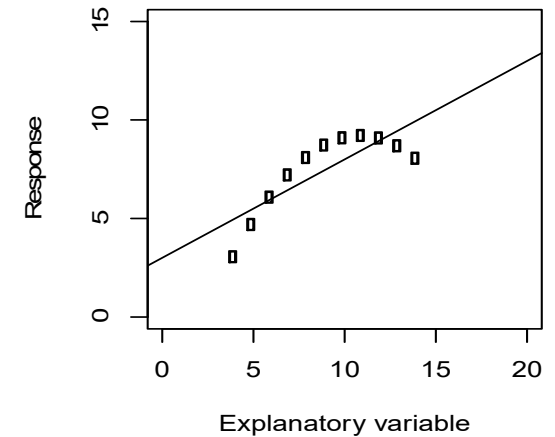
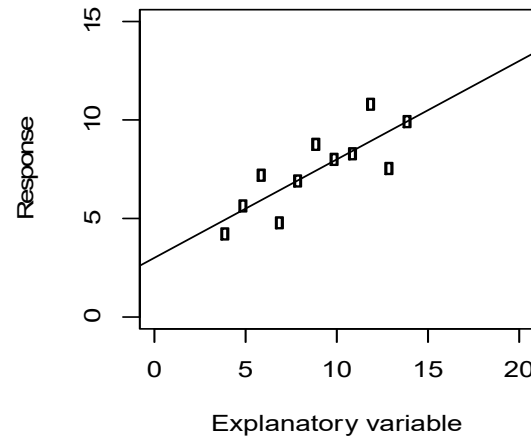
Question 2

Suppose the scatter-plot of two datasets concentrates along a line. Select all possible values for the correlation

- ☐ 1
- ☐ 0
- ☐ -1

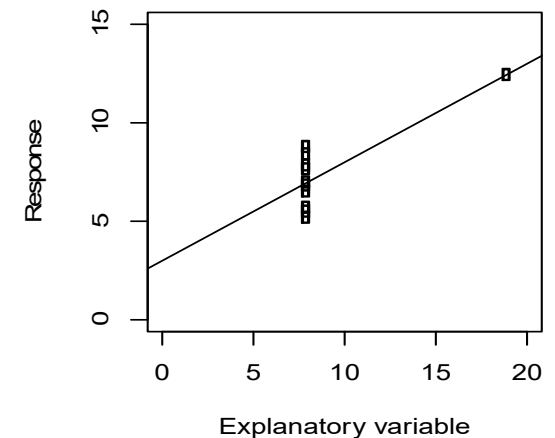
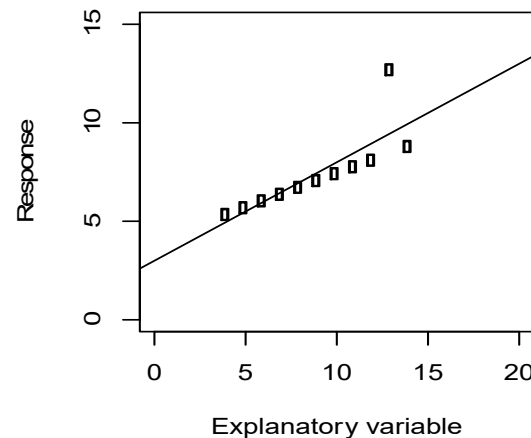
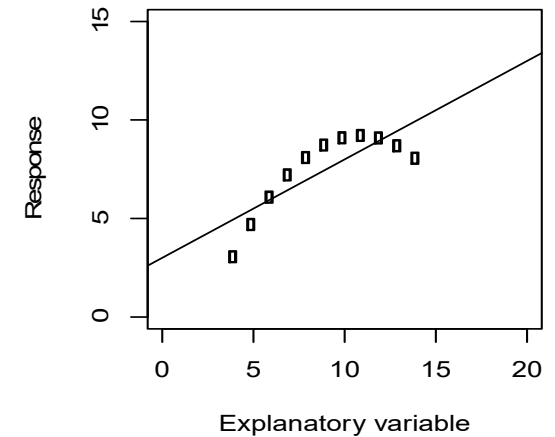
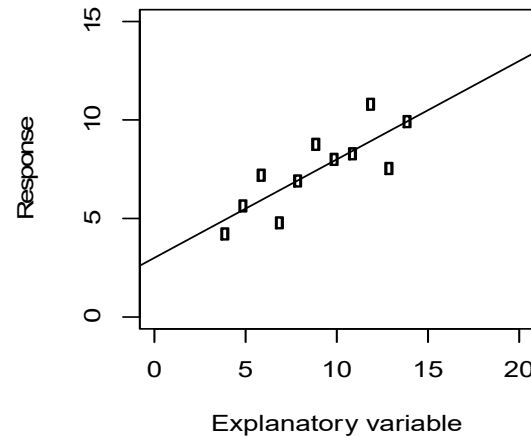
Review

- ▶ Correlation measures the concentration around a line in a scatter-plot of the independent and dependent variables.
- ▶ Correlation is a number between -1 and 1. Around 1 or -1 we have strong positive or negative correlation. Around 0 we have no correlation.
- ▶ Correlation measures association between variables. We cannot measure causation.



Review

- ▶ If we use linear regression for these four datasets, then we obtain the same
 - ▶ correlation
 - ▶ slope
 - ▶ intercept
- ▶ We learn that
 - ▶ slope and intercept are not robust to outliers
 - ▶ correlation cannot capture non-linear associations



Review

Question 3

Which of the following are properties of linear regression?

- ☐ The two parameters in the model are slope and intercept
- ☐ The variance of the predicted values is less than the value of the observed values
- ☐ The correlation of the residuals and independent variables is 0
- ☐ The average of the residuals is 0

Review

- ▶ Remember that we could determine expressions in linear regression for the slope and intercept from the data.

$$\hat{y} = \bar{y} + rSD_y \frac{(x - \bar{x})}{SD_x}$$

- ▶ We could use the derivative of the average loss to solve for the parameters

$$\min_{a,b} \sum_{i=1}^n [y_i - (a + bx_i)]^2$$

- ▶ Derivative with respect to a

$$-2 \sum_i (y_i - a - bx_i)$$

- ▶ Derivative with respect to b

$$-2 \sum_i (y_i - a - bx_i)x_i$$

- ▶ Set to 0 and solve for a and b . Note that the equation for a shows the summation of residuals is 0

Review

See Week 11
Lecture 1 Notes

- ▶ How can we interpret these two equations?
 - ▶ Setting the first expression equal to 0, we learn that the sum of the residuals is 0.
 - ▶ So the average error between observed values and predicted values is 0
 - ▶ Setting the second expression equal to 0, we learn that the residuals and the independent variables have correlation 0.
 - ▶ So the errors and the independent values lack a linear relationship

- ▶ Derivative with respect to a

$$-2 \sum_i (y_i - a - bx_i)$$

- ▶ Derivative with respect to b

$$-2 \sum_i (y_i - a - bx_i)x_i$$

- ▶ Set to 0 and solve for a and b . Note that the equation for a shows the summation of residuals is 0

Review

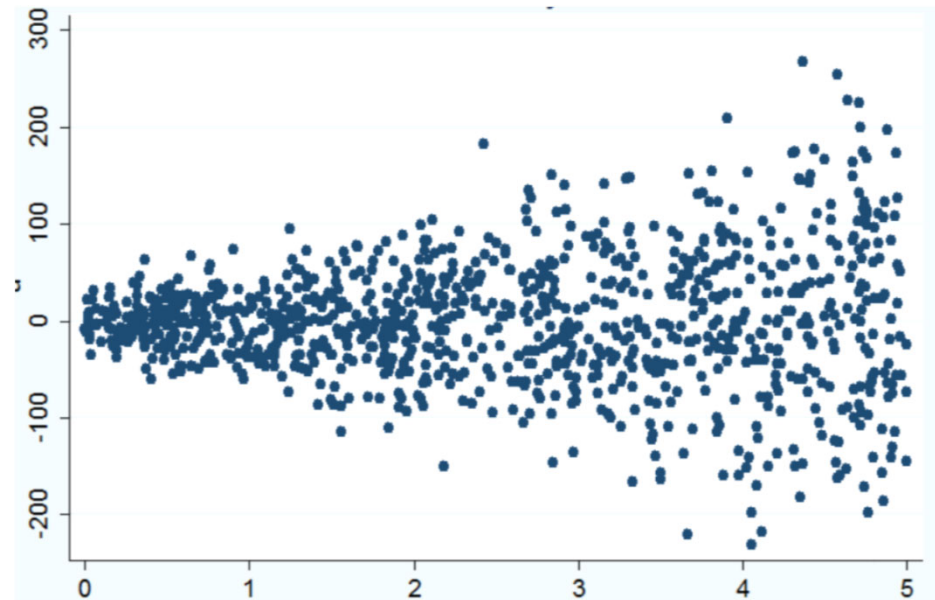
Question 4

Suppose we fit a linear model to data. We make a scatter-plot of the independent variables and residuals. If the scatter-plot has a pattern, then which of the following are True?

- ☐ We have a linear relationship between the dependent variable and independent variable.
- ☐ We cannot use a linear model to predict the independent variable from the dependent variable
- ☐ We should try to add variables or transform variables to remove the pattern in the residuals.

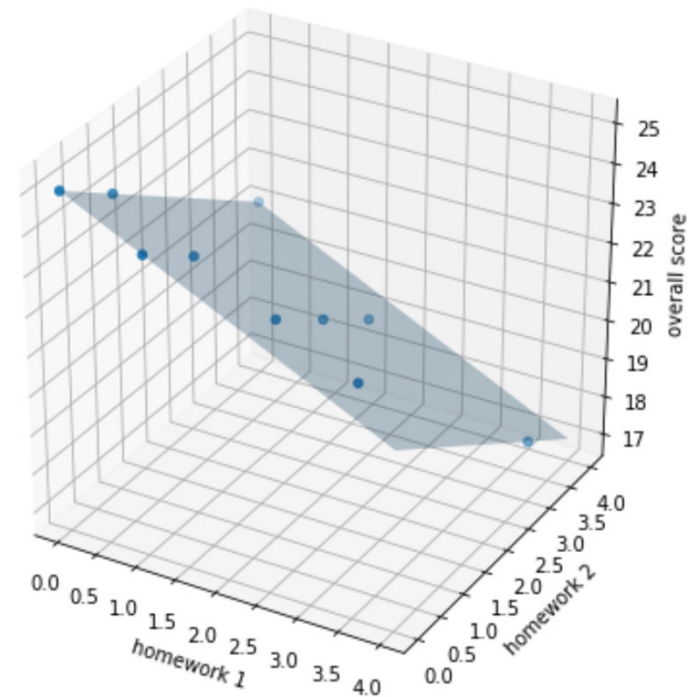
Agenda

- Features
 - Polynomial Transformations
 - One-Hot Encoding
 - Logarithmic Transformations



Multiple Independent Variables

- ▶ We have been studying linear regression with one dependent variable and one independent variable
 - ▶ We nickname the independent variable the explanatory variable
 - ▶ We nickname the dependent variable the response variable
- ▶ However some problems require multiple independent variables to explain the response.



Multiple Independent Variables

- ▶ We have been studying linear regression with one dependent variable and one independent variable
 - ▶ We nickname the independent variable the **explanatory variable**
 - ▶ We nickname the dependent variable the **response variable**
- ▶ However some problems require multiple independent variables to explain the response.

- ▶ Single variable linear model

$$\hat{y} = a + b x$$

- ▶ Two variable linear model

$$\hat{y} = a + b x_1 + c x_2$$

- ▶ Many variable linear model

$$\hat{y} = c_0 + \sum_{i=1}^k c_i x_i$$

Linear and Non-Linear

- Suppose we have a function f with n inputs. If we have the property

$$f(x_1+y_1, \dots, x_n+y_n) = f(x_1, \dots, x_n) + f(y_1, \dots, y_n)$$

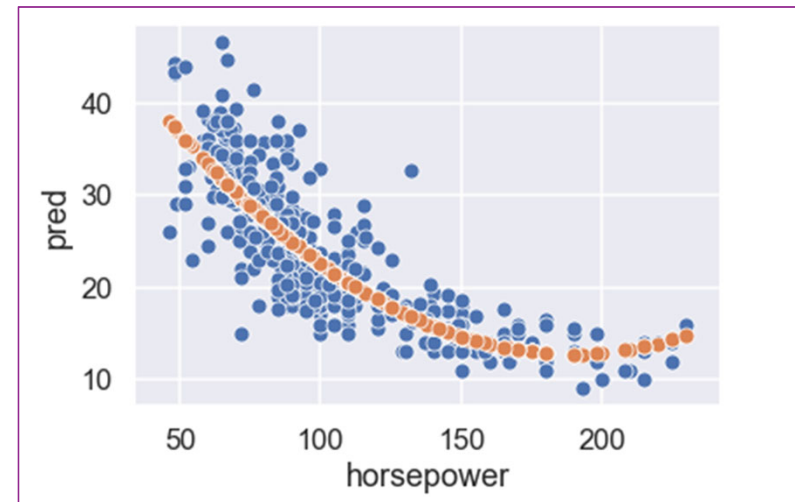
for any inputs, then we call the function **linear**

- Here the output from a sum of inputs is the sum of the respective outputs
- Note that linear models are linear in the independent variables

- Before fitting a model to data, we can apply a transformation to the independent variables. These transformations allow us to incorporate additional **features** in the model.
- Instead of x_1, \dots, x_n we might have $g(x_1), \dots, g(x_n)$. We fit a linear model to the transformed independent variables.
- While the new model may not be linear in the variables x_1, \dots, x_n , it will be linear in the variables $g(x_1), \dots, g(x_n)$

Polynomial Transformations

- If we use transformations to generating features, then we can obtain more accurate predictions with linear models. Sometimes we can capture non-linear relationships between dependent variables and independent variables
- If we replace an independent variable x with powers $1, x, x^2, x^3, \dots$ then we have a polynomial transformation



- If we have multiple independent variables then we can multiply them to model interactions between the features.

One-Hot Encoding

- ▶ If we have qualitative data, then we must transform it to quantitative data. However we must be careful with the **encoding** of the categories.

- ▶ If we encoded categories


usa, japan, europe

in a geographic dataset by

1,2,4

then the japan would contribute twice as much as use to the predicted value

c, 2c, 4c



origin	origin=usa	origin=europe	origin=japan
usa	1	0	0
usa	1	0	0
europe	0	1	0
...
usa	1	0	0
japan	0	0	1
japan	0	0	1

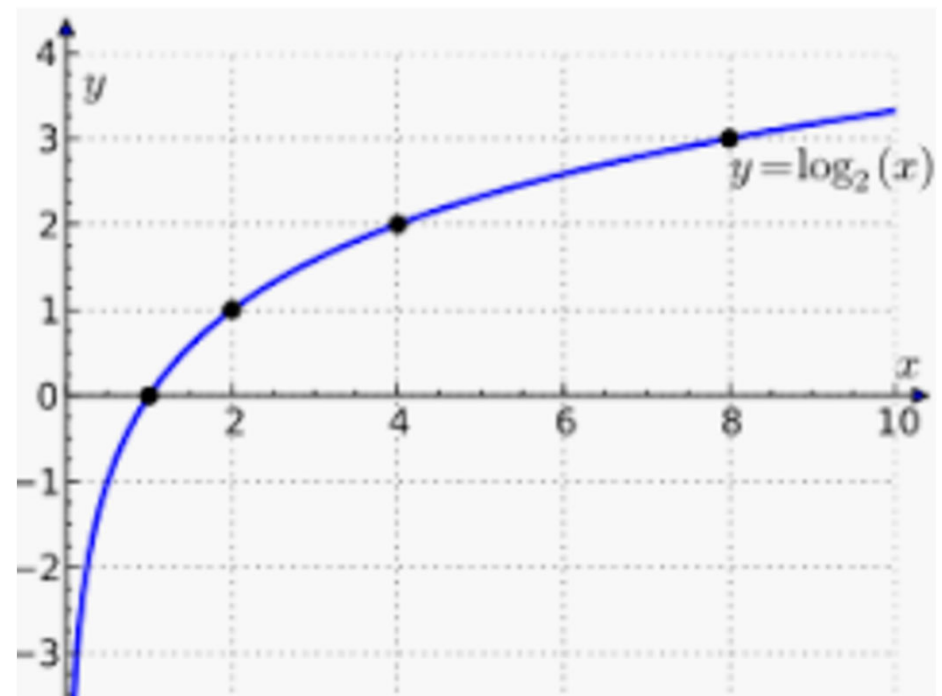
- ▶ Instead we use a one-hot encoding. We add another independent variable for each category. The additional variables take the value 0 or 1.

Logarithmic Transformations

- Remember that logarithmic transformations help us with visualization. We can transform a large range of numbers to a small range of numbers suitable for a chart.

$$\log_b(a) = c \iff b^c = a$$

- If the independent variable and dependent variable have different scales, then we can apply logarithms to help with linear regression



Logarithmic Transformations

► Logarithmic Growth

$$\text{Dependent Variable} = \text{Intercept} - \text{Slope} \times \log(\text{Independent Variable})$$

► Exponential Growth

$$\log(\text{Dependent Variable}) = \text{Intercept} - \text{Slope} \times \text{Independent Variable}$$

► Power Law

$$\log(\text{Dependent Variable}) = \text{Intercept} - \text{Slope} \times \log(\text{Independent Variable})$$

Summary

► Features

- One-Hot Encoding
- Polynomial Transformations
- Logarithmic Transformations

Goals

- Treat qualitative variables in linear regression with one-hot encoding.
- Use polynomial transformations to capture non-linear relationships between variables
- Apply logarithm transformations to study power laws

Questions

- ▶ Questions on Piazza?
 - ▶ Please provide your feedback along with questions
- ▶ Question for You!

Why are logarithms useful in models for exponential growth?

IDEAS | EVERYDAY MATH

When a Virus Spreads Exponentially

The key to stopping the Covid-19 pandemic lies in lowering the rate at which infections multiply.



By *Eugenia Cheng*

April 2, 2020 2:01 pm ET



PRINT



TEXT

Fighting a pandemic like Covid-19 requires experts in many fields: epidemiologists who study the spread of disease, doctors who treat the sick, scientists who work on finding a vaccine. There is math involved in all of these specialties, but math can also help us to make sense of the barrage of information that we're receiving daily.

The starting point is the math of exponential growth. The word "exponential" is sometimes

