Linear Regression Lines (Part 1)

Instructor: Hamed Hasheminia

Lecture 3

Agenda

- What are linear regression lines?
- Explore single variable regression lines
- Explore multi-variable regression lines
- How to capture non-linearity with linear regression lines
- Learn how to interpret regression coefficients
- How to deal with dummy variables

- Lab
- Use sklearn library

Linear Regression Lines

- Linear regression is a simple approach to supervised learning. It assumes that the dependence of Y (Your quantitative output) on $X_1, X_2, \ldots., X_p$ *(your inputs) is* *linear*
- **Linear Regression Lines are:**
  - Simple to explain
  - Highly interpretable
  - Model training and prediction are fast
  - No tuning is required (excluding regularization)
  - (Input) Features don't need scaling
  - Can perform well with a small number of observations
  - Well-understood
  - *Not too flexible*

Simple Linear regression Using a single predictor

- We assume a model

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

- Where beta0 and beta1 are two unknown constants that represent the intercept and slope, also known as coefficients or parameters and epsilon is the error term.

- we predict y using:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

- Y_hat indicates a prediction of Y on the basis of X = x. The hat symbol denotes an estimated value.

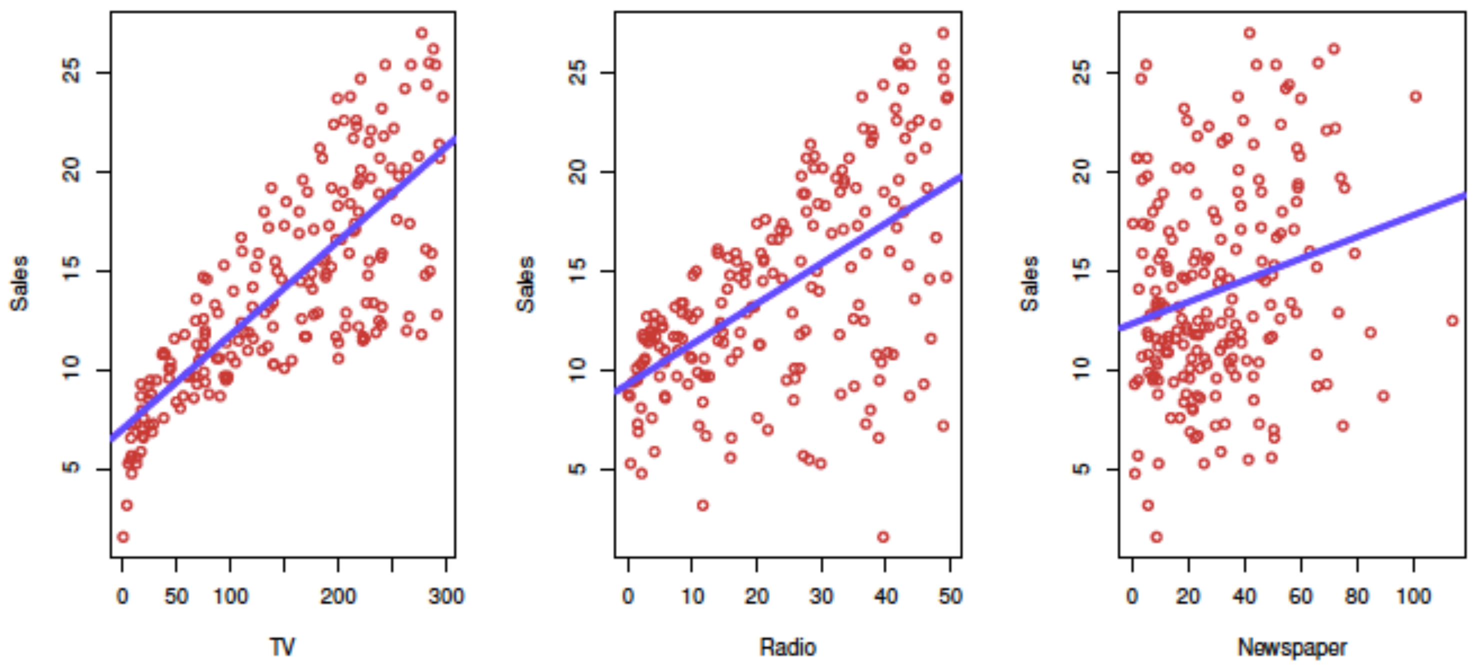
Linear Regression for the advertising data

Consider the advertising data we worked on last session:

Questions we might ask:

- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget band sales?
- Which media does contribute to sales the most?
- How accurately can we predict future sales?
- How good does our linear model perform?
- Is there synergy among the advertising media? (We talk about this next session)

Advertising Data



Estimation of the parameters by least squares

- 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

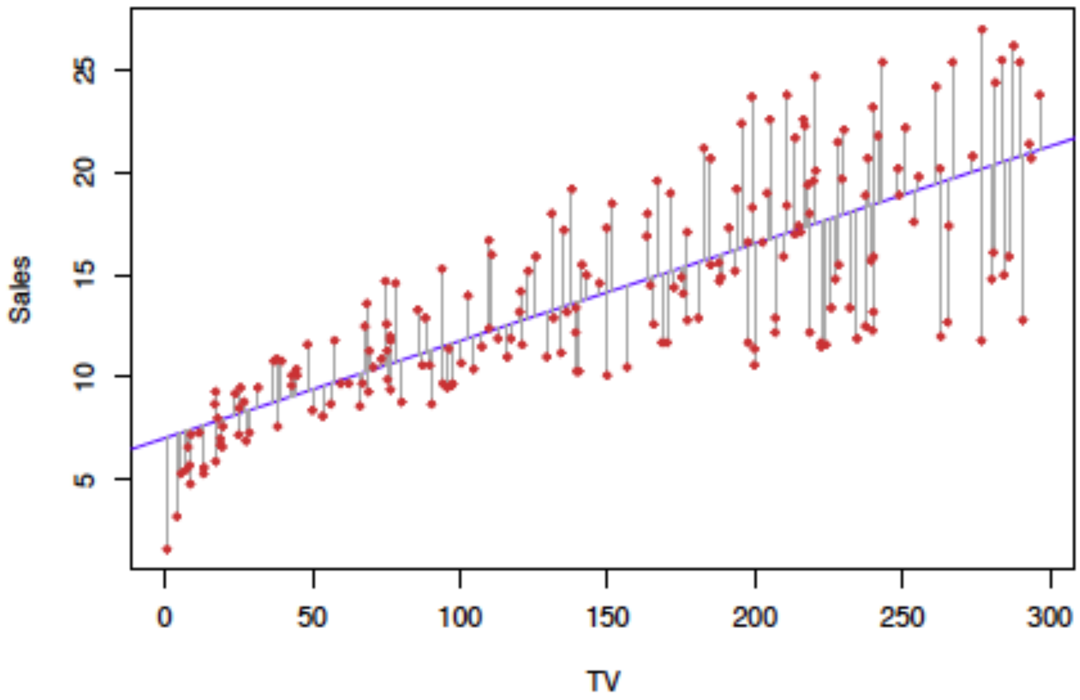
$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

- Or equivalently as

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

- The least square approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS.

Example: Advertising Data



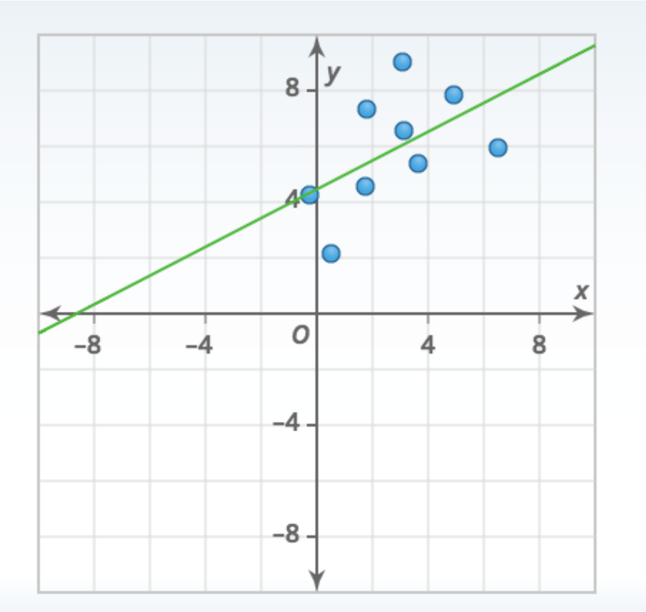
The least squares fit for the regression of Sales onto TV.

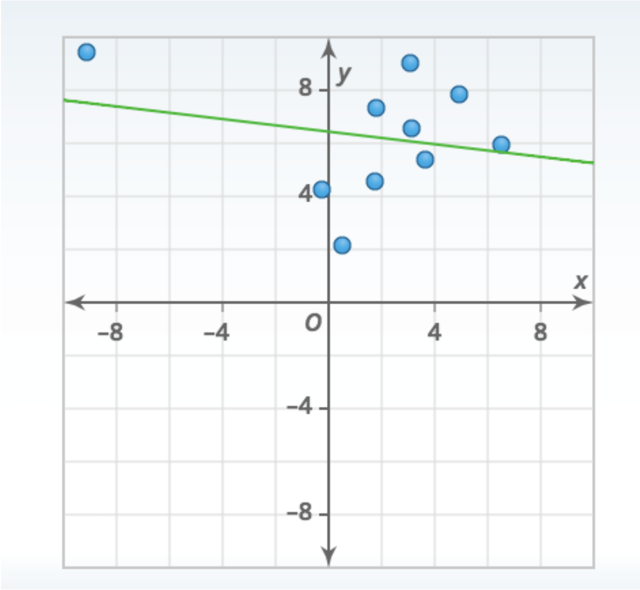
- In this case linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.

Results for the advertising Data

	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

The effect of outliers on Regression lines





Multiple Linear Regression

- Here is our model is

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

- We interpret Beta_j as the *average* effect on Y of a one unit increase in X_j, *holding 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 Beta_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.
- *Claims of causality* should be avoided for observational data.

The woes of (interpreting) regression coefficients

- *"Data analysis and regression" Mosteller and Tukey 1977*
  - A regression coefficient Beta_j estimates the expected change in Y per unit change in X_j, *with other predictors held fixed*. But predictors usually change together.
  - Example: Y = Number of tackles by a football player in a season; W and H are his weight and height. Fitted regression model is:

    $\hat{Y} = b_0 + .50W - .10H$ How do we interpret -0.1?
- *"Essentially, all models are wrong, but some are useful"* George Box
- *"The only way to find out what will happen when a complex system is disturbed is to disturb the system, not merely to observe it passively."* Fred Mosteller and John Tukey, paraphrasing George Box

Estimation and Prediction for Multiple Regression

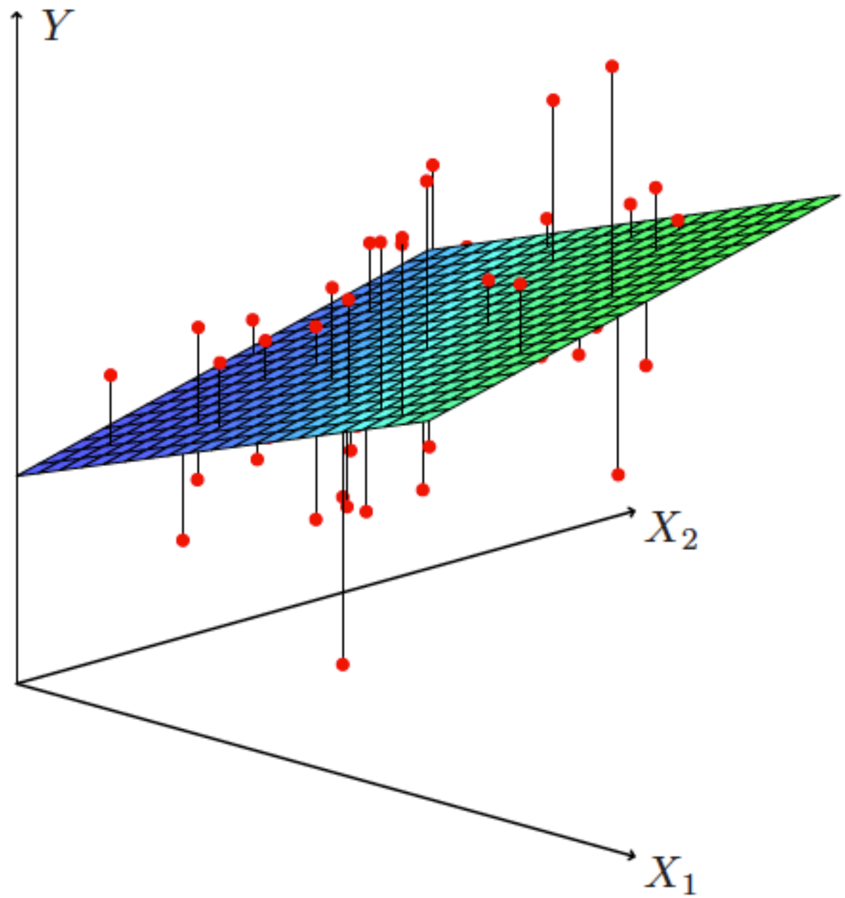
- Given Estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_p,$ we can make predictions using the formula

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p.$$

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

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2 \end{aligned}$$

- This is done using any standard statistical software.



Y
X_2
X_1

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

Some important Questions

- Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?
- Do all the predictors help to explain Y, or is only a subset of the predictors useful?
- How well does the model fit the data?
- Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Linear Regression Models – adding non-linear terms

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

$$Y = \beta_0 e^{\beta_1 X} \leftrightarrow \Upsilon = \mathrm{B}_0 + \beta_1 X \quad where \;\; \Upsilon = log_e(Y) \; and \;\; \mathrm{B}_0 = log_e(\beta_0)$$

Qualitative Predictors

- Example: Let's investigate differences in credit card balances between males and females.
  - Our output, Y, is credit card balance
  - We assume we only want to use gender as a predictor
  - We create a new variable:

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases}$$

Resulting model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

Intrepretation?

Let's interpret the results

Results for gender model:

Qualitative Predictors

	Coefficient	Std. Error	t-statistic	p-value
Intercept	509.80	33.13	15.389	< 0.0001
gender[Female]	19.73	46.05	0.429	0.6690

Qualitative predictors with more than two levels

- With more than two levels, we create additional dummy variables. For example, for the `ethnicity` variable we create two dummy variables. The first could be

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian,} \end{cases}$$

and the second could be

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian.} \end{cases}$$

Dummy Variables and Results for ethnicity

- Then both of these variables can be used in the regression equation, in order to obtain the model

$$y_i = \beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is AA.} \end{cases}$$

	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

Summary

- We learned:
  - What are regression lines – advantages and limitations
  - Single variable and multi-variable regression lines
  - How to interpret coefficients of regression lines
  - Different classes of non-linear functions that can be handled using regression lines.
  - Using dummy variables – how to interpret dummy variables

- What we are going to learn next session:
  - Hypothesis test – test of significance on regression coefficients
  - P-Value
  - Different types of error
  - Interaction effects

Python - sklearn library

- Before leaving class – please make sure you have sklearn library. Type “import sklearn” on your ipython notebook. If it runs, perfect, otherwise, you need to install sklearn on your computer.
- To install sklearn library type “conda install scikit-learn” on your terminal window/command line
- Also, please check if you have seaborn package.
- If not type “conda install seaborn” on your terminal window/command line