Regression

Topic 4.1. Overview, simple and multiple regression

Sonja Petrović

Created for ITMD/ITMS/STAT 514

Spring 2021.

Goals of this lecture

- What is a 'regression' function
- What is prediction
- Elements of simple&multiple linear regression
- Best approximation, least squares, residual sum of squares
- Basic `R` command to run a regression model
- Looking ahead:
  - basic `Python` command to run a regression model
  - polynomial regression

Section 1

The setup: basics

What is the prediction task?

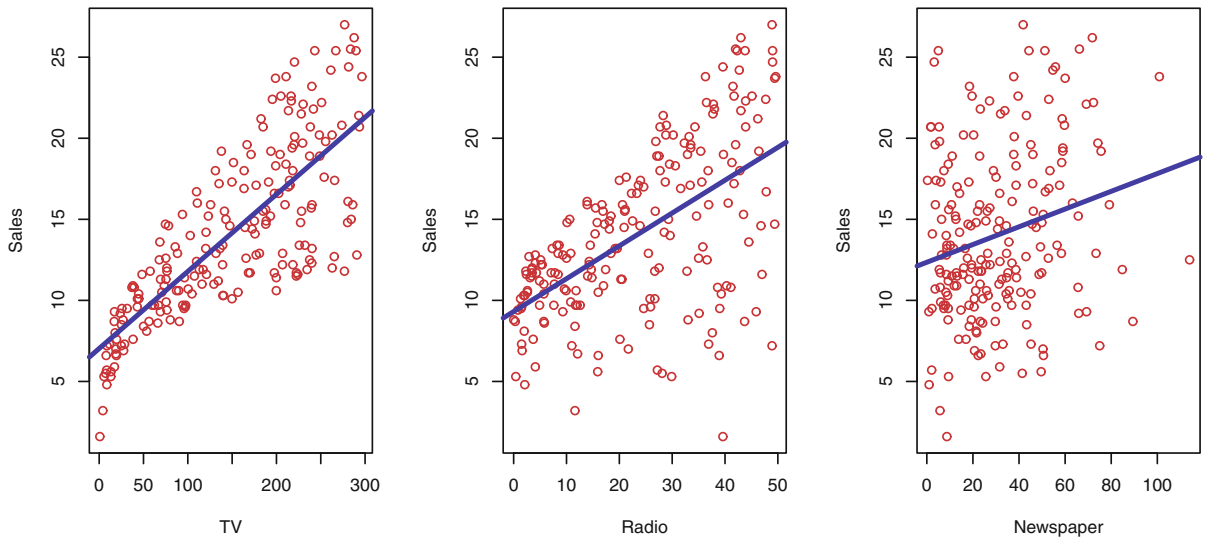


Figure 1: Figure 2.1. from ISLR: Y =`Sales` plottedagainst `TV`,`Radio` and `Newspaper` advertising budgets.

Our goal is to develop an accurate **model** (f) that can be used to predict `sales` on the basis of the three media budgets:

$$Sales \approx f(TV, Radio, Newspaper).$$

Notation

- `Sales` = a reponse, target, or outcome.
  - The variable we want to predicit.
  - Denoted by Y.
- `TV` is one of the features, or inputs.
  - Denoted by X_1.
- Similarly for `Radio` and `Newspaper`.
- We can put all the predictors into a single input vector

$$X = (X_1, X_2, X_3)$$

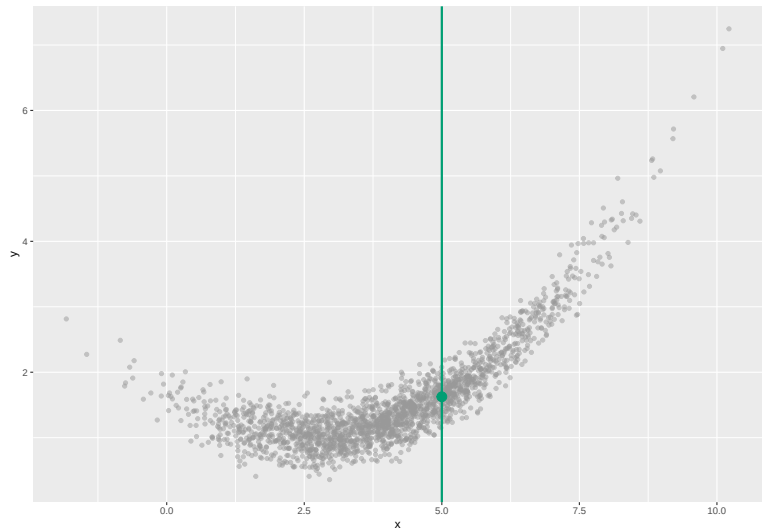
- Now we can write our model as

$$Y = f(X) + \epsilon,$$

where ϵ captures measurement errors and other discrepancies between the response Y and the model f.

What does it mean to *predict* Y?

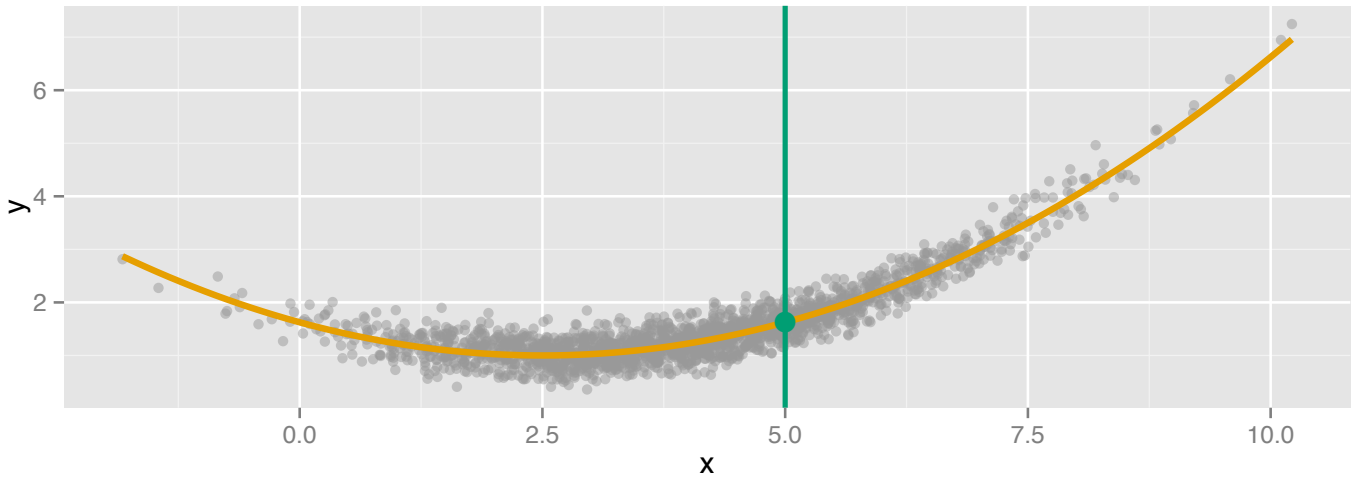
Here's some simulated data.



- Look at $X = 5$. There are many different Y values at $X = 5$.
- When we say **predict** Y at $X = 5$, we're really asking:

What is the expected value (average) of Y at $X = 5$?

The regression function



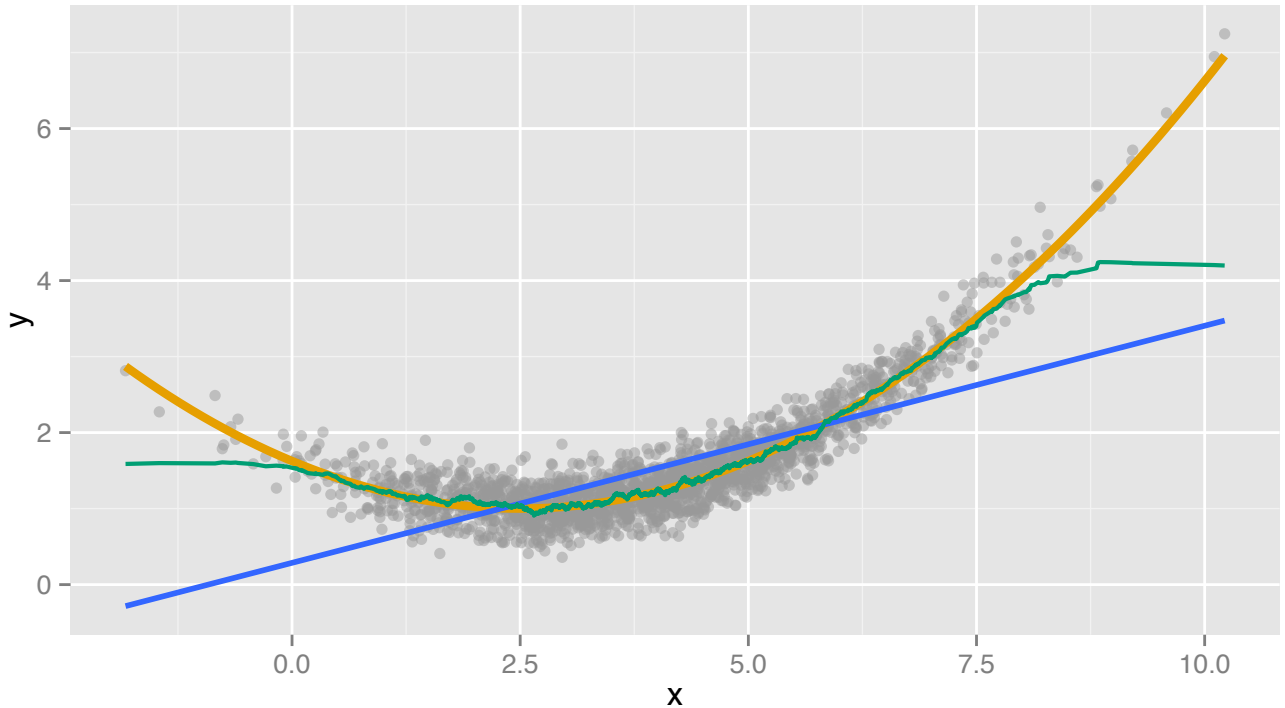
Definition: Regression function

Formally, the regression function is given by $E(Y|X = x)$. This is the *expected value* of Y at $X = x$.

- The ideal or optimal predictor of Y based on X is thus

$$f(x) = E(Y|X = x)$$

The prediction problem



regression function f linear regression $\hat{f}$ 50-nearest-neighbours $\hat{f}$

The prediction problem

We want to use the observed data to construct a predictor $\hat{f}(x)$ that is a good estimate of the regression function $f(x) = E(Y|X = x)$.

To summarize

- The ideal predictor of a response Y given inputs $X = x$ is given by the regression function
$$f(x) = E(Y|X = x)$$

- We don't know what f is, so the prediction task is to estimate the regression function from the available data.

- The various prediction methods are different ways of using **data** to construct estimators $\hat{f}$.

The best method?

Remember: There is no free lunch...

* If the data you work with tends to have linear associations, you may be well-served by a linear model.

* If you know that similar people like similar things, you may be well-served by a nearest-neighbours method.

Simple linear regression

- Predict a quantitative Y by single predictor variable X

Linear relationship:

$$Y \approx \beta_0 + \beta_1 X.$$

- Example: $\textit{sales} \approx \beta_0 + \beta_1 \times \textit{TV}$.
- β_0, β_1 are two unknown constants. [*parameters*, or *coefficients*.]

Prediction

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

$\rightarrow$ discussion in lecture (with notes).

Estimating the coefficients

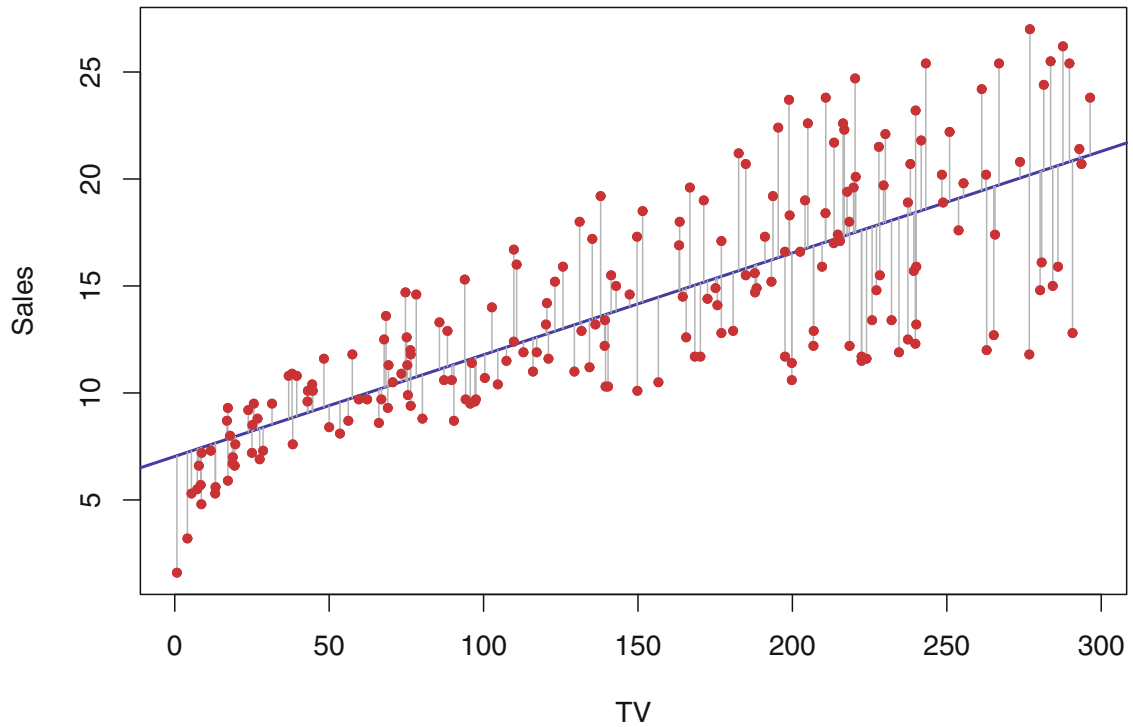


Figure 2: Fig 3.1. ISLR: For the `Advertising` data, the least squares fit for the regression of `sales` onto `TV` is shown. The fit is found by minimizing the sum of squared errors. Each grey line segment represents an error, and the fit makes a compro- mise by averaging their squares. In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.

Many different least squares lines

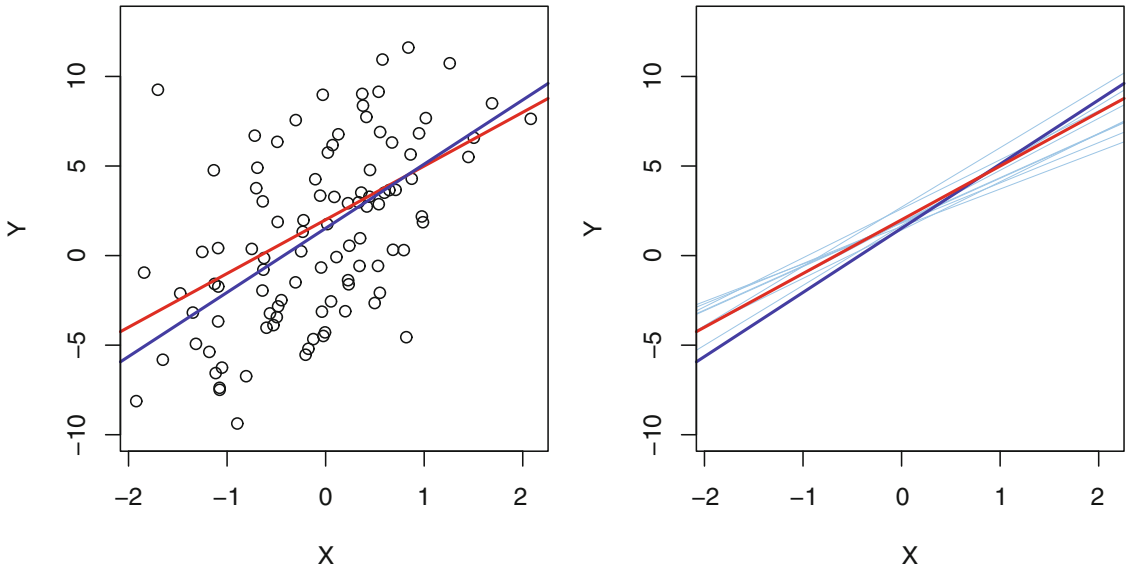


Figure 3: Fig. 3.2. ISLR: A simulated data set. Left: The red line represents the true rela- tionship, f(X) = 2 + 3X, which is known as the population regression line. The blue line is the least squares line; it is the least squares estimate for f(X) based on the observed data, shown in black. Right: The population regression line is again shown in red, and the least squares line in dark blue. In light blue, ten least squares lines are shown, each computed on the basis of a separate random set of observations. Each least squares line is different, but on average, the least squares lines are quite close to the population regression line.

Multiple linear regression

p predictors, $X_1, \ldots, X_p$, for modeling the continuous response variable Y:

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

- $f_L(X) := \beta_0 + \sum_{j=1}^{p} \beta_j X_j$ is the best linear approximation to the true regression function.
- The true regression function may not be linear.

Estimating the coefficients:

. . . same setup as in simple linear regression $\hat{\beta}_j$.

→ discussion.

Section 2

Graphics for the story about multiple linear regression

	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

TABLE 3.1. *For the* `Advertising` *data, coefficients of the least squares model for the regression of number of units sold on TV advertising budget. An increase of* \$1,000 *in the TV advertising budget is associated with an increase in sales by around 50 units (Recall that the* `sales` *variable is in thousands of units, and the* `TV` *variable is in thousands of dollars).*

Figure 4: ISLR table 3.1.

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	9.312	0.563	16.54	< 0.0001
`radio`	0.203	0.020	9.92	< 0.0001

Simple regression of `sales` on `newspaper`

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	12.351	0.621	19.88	< 0.0001
`newspaper`	0.055	0.017	3.30	0.00115

Figure 5: ISLR table 3.3.

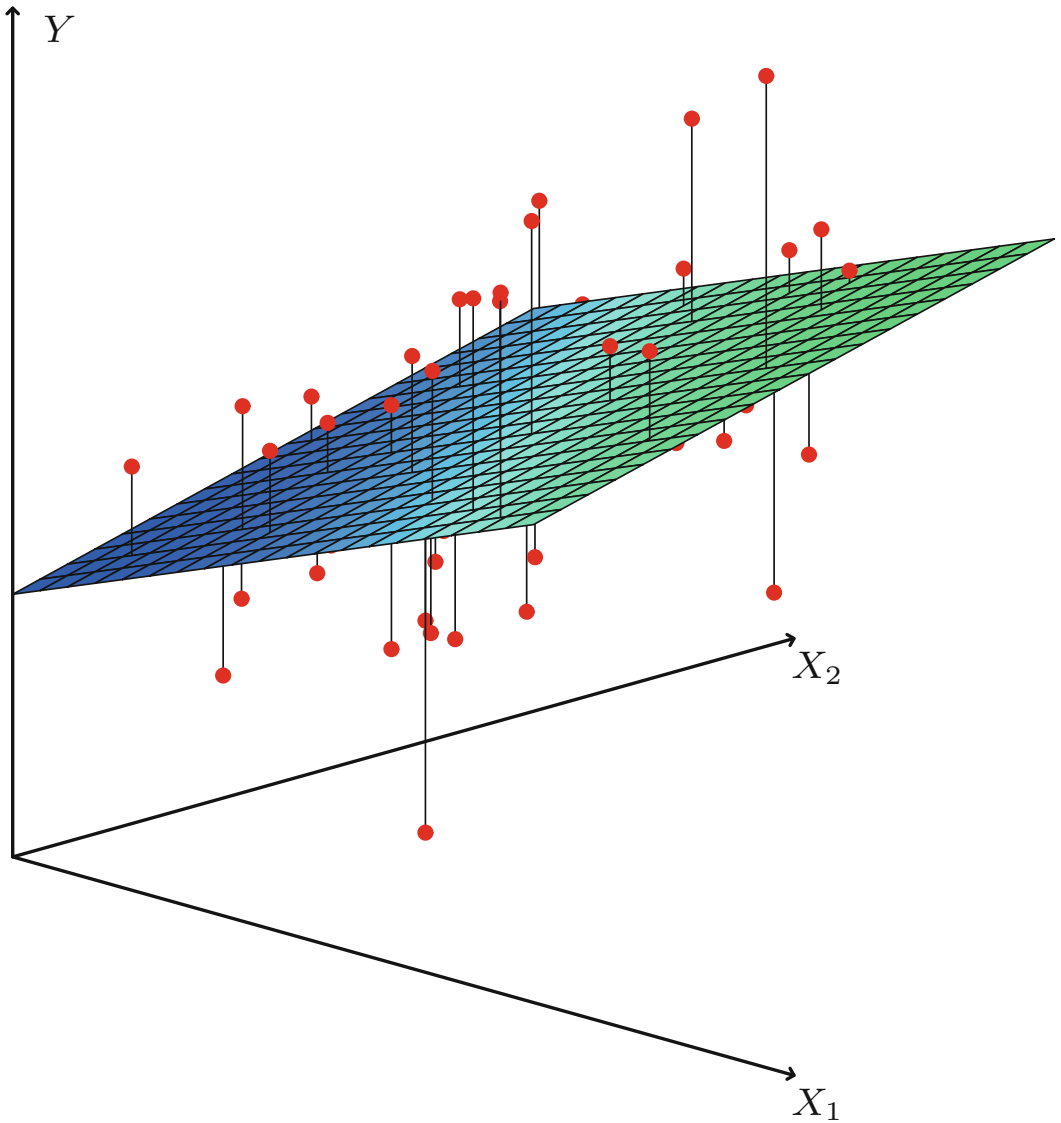


Figure 6: ISLR fig.3.4.

	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

TABLE 3.4. *For the* `Advertising` *data, least squares coefficient estimates of the multiple linear regression of number of units sold on radio, TV, and newspaper advertising budgets.*

Figure 7: ISLR table 3.4.

	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

TABLE 3.5. *Correlation matrix for* TV, radio, newspaper, *and* sales *for the* Advertising *data.*

Figure 8: ISLR table 3.5

License

Contents of this lecture is based on the chapter 3 of the textbook Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, *' An Introduction to Statistical Learning: with Applications in R'*.

Part of this lecture notes are extracted from Prof. Alexandra Chouldechova, released under a Attribution-NonCommercial-ShareAlike 3.0 United States license.