Lingnan University 嶺南大學 香港 Hong Kong

School of Data Science 數據科學學院

CDS 533
Statistics for Data Science

Instructor: Lisha Yu

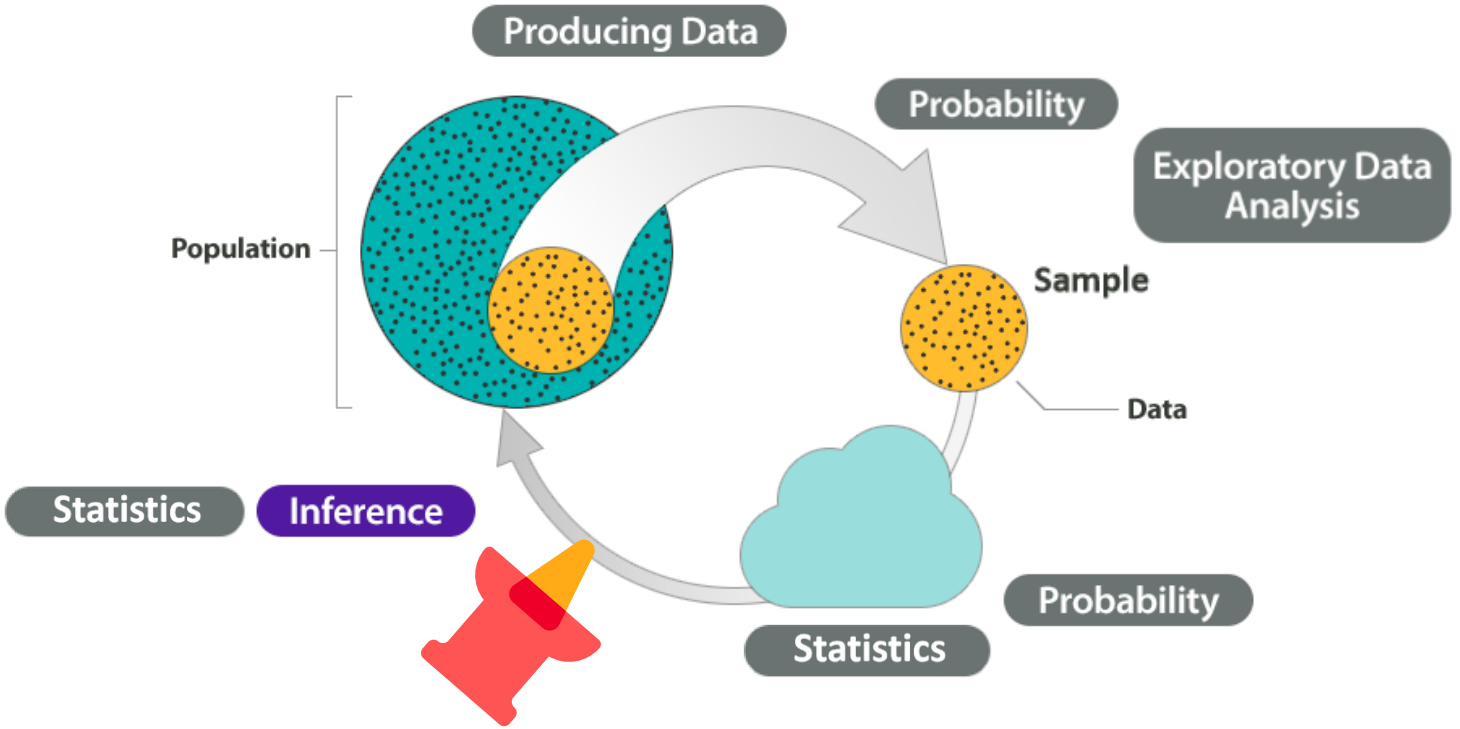
Division of Artificial Intelligence

School of Data Science

Lingnan University

Fall 2024

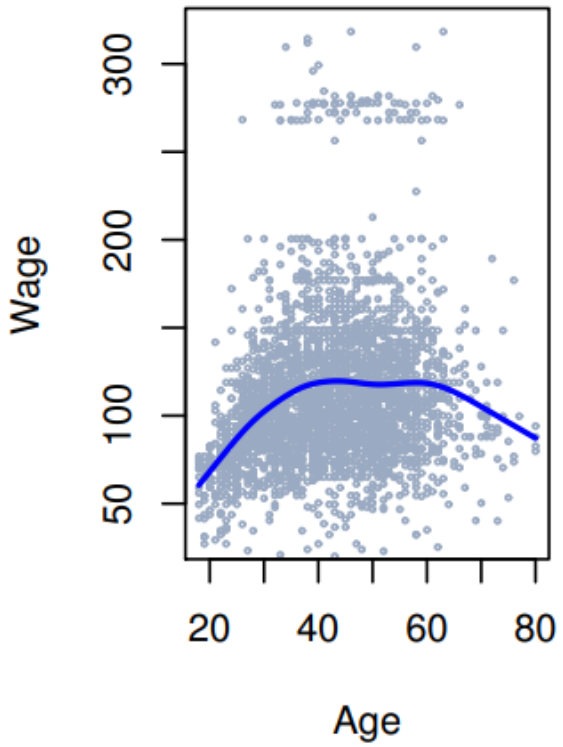
Big Picture of Statistics

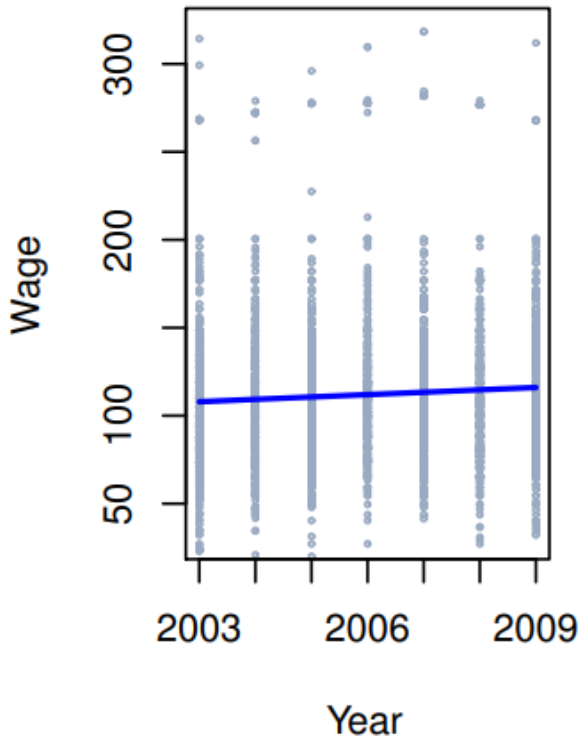


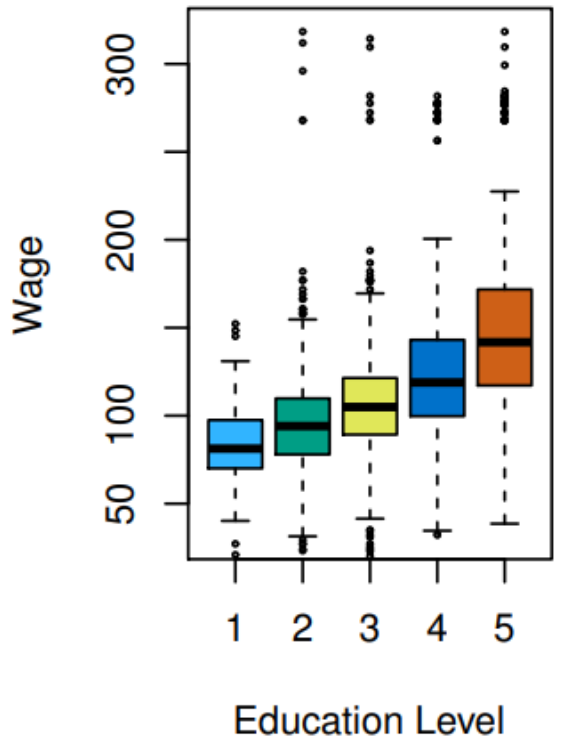
Statistical Inference (Regression)

Motivation

Income survey data for males from the central Atlantic region of the USA in 2009







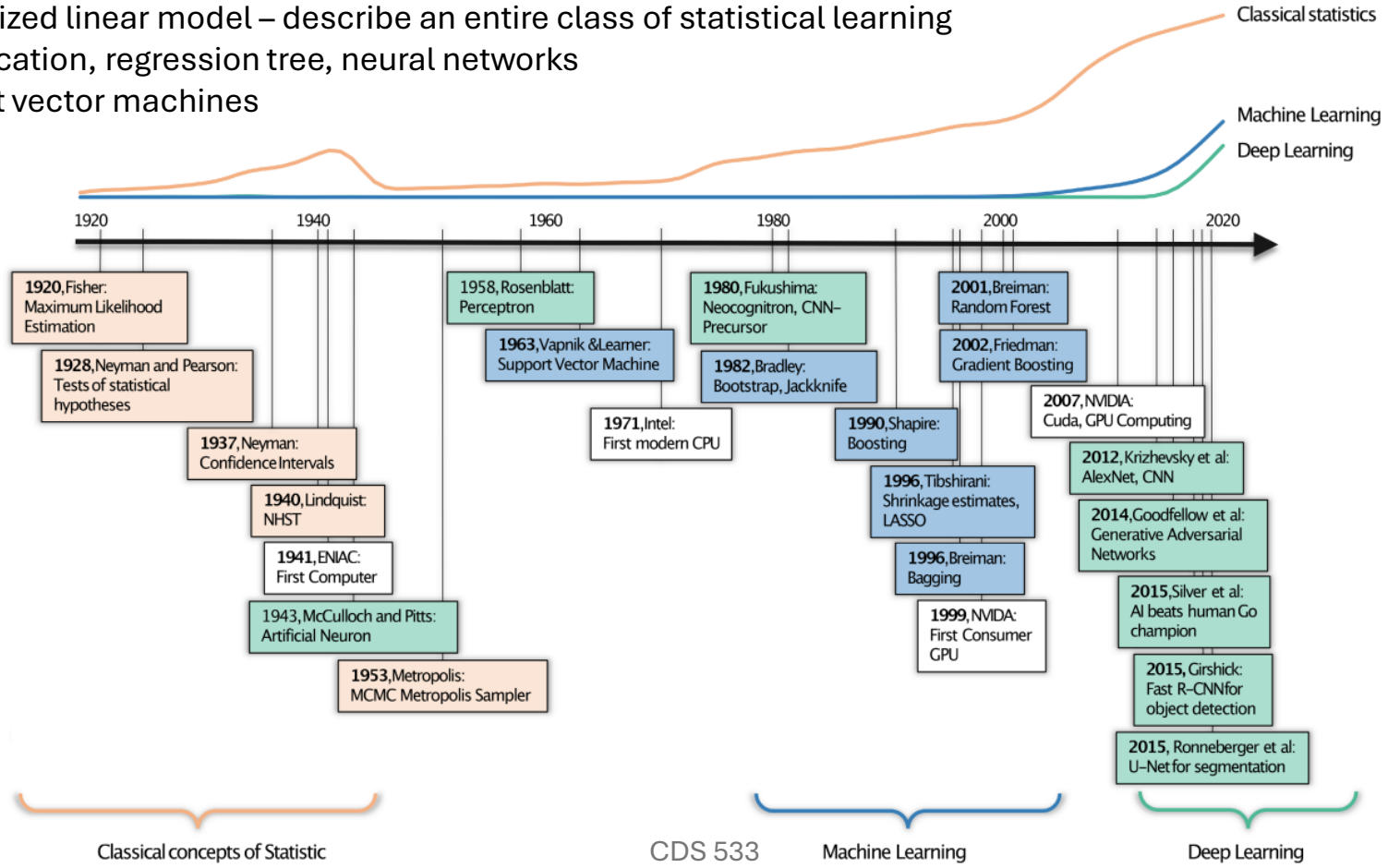
Left: wage as a function of age.

Center: wage as a function of year.

Right: Boxplots displaying wage as a function of education.

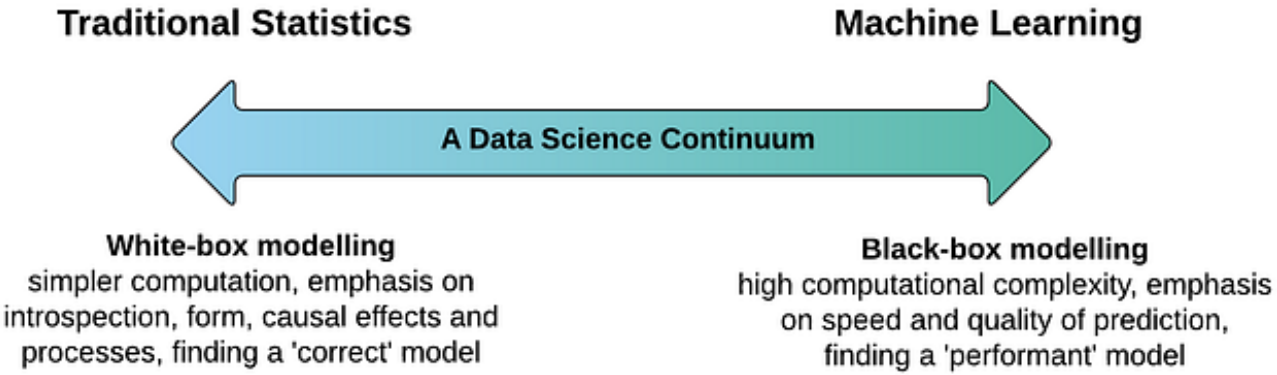
Brief History of Statistical Learning

- Beginning of the 19th century, least squares (earliest form of linear regression)
- Prediction of quantitative values, e.g., individual's salary
- 1936, linear discriminant analysis – qualitative values
- 1940s, logistic regression
- 1970s, generalized linear model – describe an entire class of statistical learning
- 1980s, classification, regression tree, neural networks
- 1990s, support vector machines

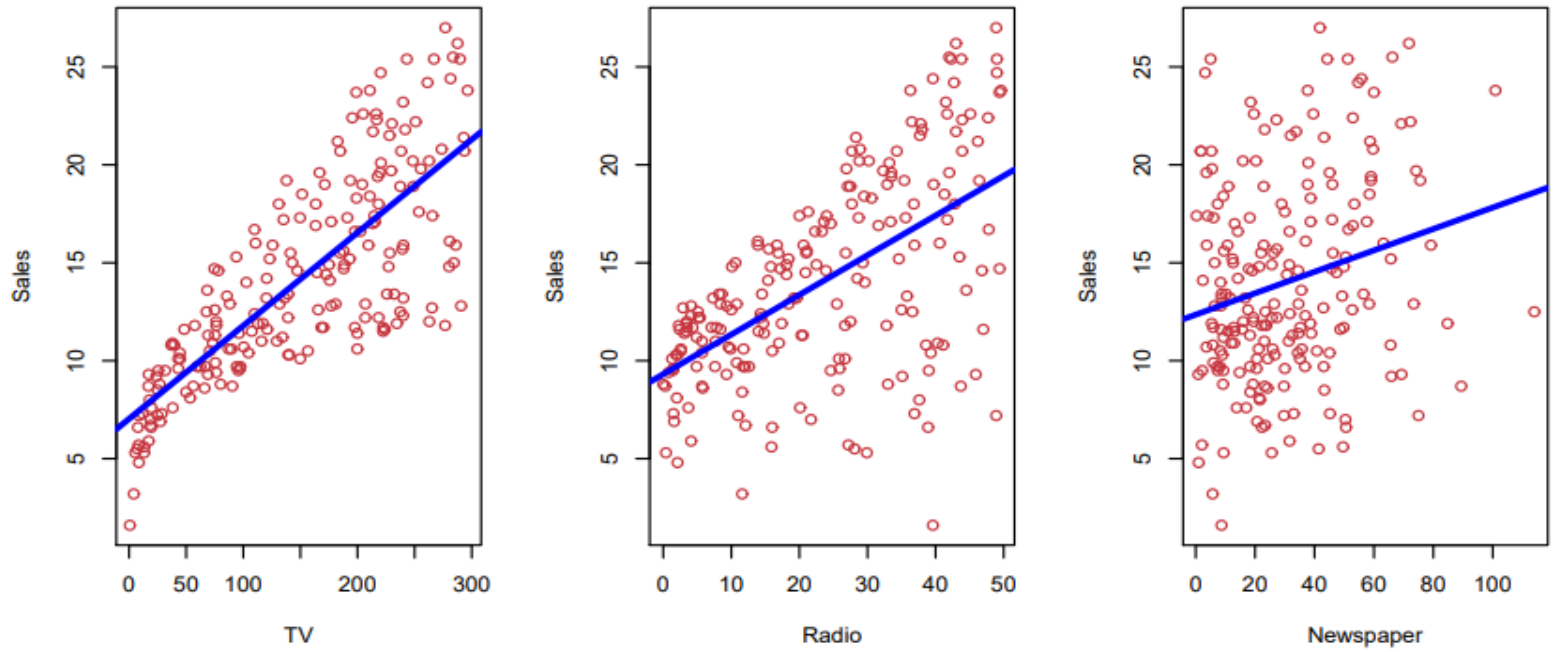


Statistical Learning vs. Machine Learning

- Machine learning arose as a subfield of Artificial Intelligence.
- Statistical learning arose as a subfield of Statistics.
- *There is much overlap* — both fields focus on supervised and unsupervised problems:
  - **Machine learning** has a greater emphasis on **large scale** applications and ***prediction accuracy***.
  - **Statistical learning** emphasizes **models** and their interpretability, and ***precision and uncertainty***.
- But the distinction has become more and more blurred, and there is a great deal of "cross-fertilization".



Statistical Learning



Shown are Sales vs TV, Radio and Newspaper, with a blue linear-regression line fit separately.
Can we predict Sales using these three?
Perhaps we can do better using a model

$$\texttt{Sales} \approx f(\texttt{TV}, \texttt{Radio}, \texttt{Newspaper})$$

Statistical Learning: Notation

- Here Sales is a *response* or *target* that we wish to predict. We generically refer to the response as Y . [**Dependent variable/Response**]
- TV is a *feature, or input*, or *predictor*; we name it X_1. [**Independent variable/Predictors**]
- Likewise name Radio as X_2, and so on.
- We can refer to the *input vector* collectively as

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

Now we write our model as

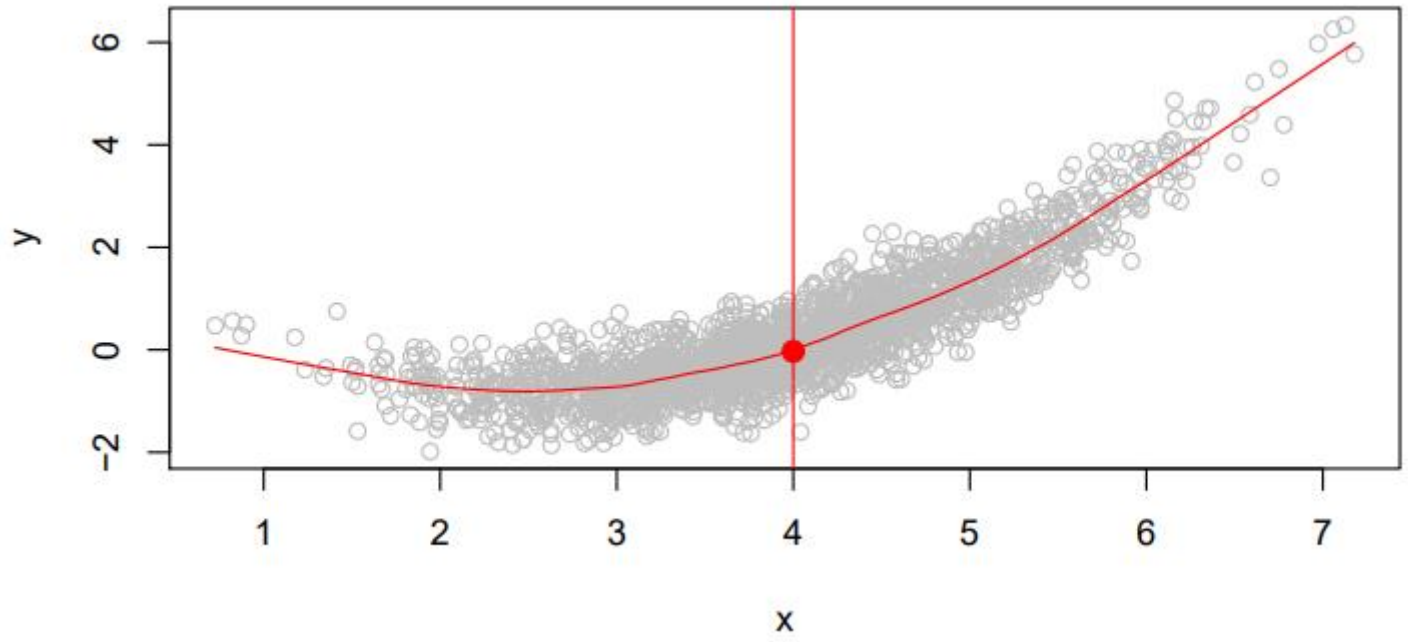
$$Y = f(X) + \epsilon$$

where ε captures measurement errors and other discrepancies.

What is $f(X)$ good for?

- With a good f we can make predictions of Y at new points $X = x$.
- We can understand which components of $X = (X_1, X_2, \ldots, X_p)$ are important in explaining Y, and which are irrelevant. e.g. Seniority and Years of Education have a big impact on Income, but Marital Status typically does not.
- Depending on the complexity of f, we may be able to understand how each component Xj of X affects Y.

What is $f(X)$ good for?



Is there an ideal $f(X)$? In particular, what is a good value for $f(X)$ at any selected value of X, say $X = 4$? There can be many Y values at $X = 4$. A good value is

$$f(4) = E(Y \mid X = 4)$$

$E(Y \mid X = 4)$ means **expected value** (average) of Y given $X = 4$.

This ideal $f(x) = E(Y \mid X = x)$ is called the **regression function**.

The Regression Function $f(x)$

- Is also defined for vector X; e.g.

  $f(x) = f(x_1, x_2, x_3) = E(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3)$
- Is the **ideal** or **optimal** predictor of Y with regard to mean-squared prediction error: $f(x) = E(Y|X = x)$ is the function that minimizes $E[(Y - g(X))^2|X = x]$ over all functions g at all points $X = x$.
- $\varepsilon = Y - f(x)$ is the **irreducible** error — i.e. even if we knew $f(x)$, we would still make errors in prediction, since at each $X = x$ there is typically a distribution of possible Y values.
- For any estimate $\hat{f}(x)$ of $f(x)$, we have

$$E[(Y - \hat{f}(X))^2|X = x] = \underbrace{[f(x) - \hat{f}(x)]^2}_{Reducible} + \underbrace{\mathrm{Var}(\epsilon)}_{Irreducible}$$

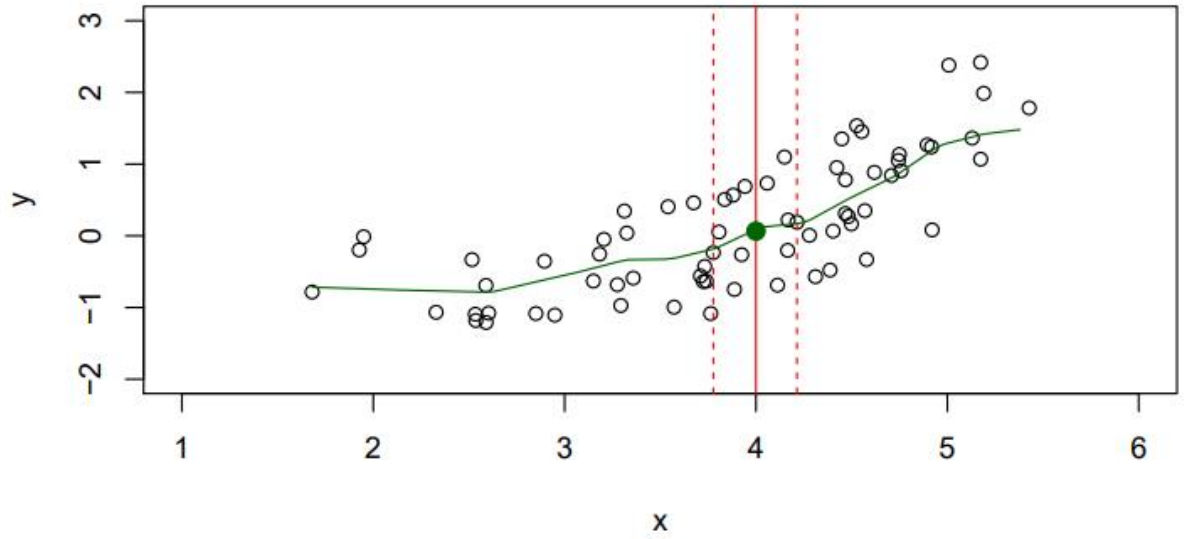
How to estimate f

Example

- Typically we have few if any data points with $X = 4$ exactly.
- So we cannot compute $E(Y|X = x)$!
- Relax the definition and let

$$\hat{f}(x) = \text{Ave}(Y|X \in \mathcal{N}(x))$$

where $\mathcal{N}(x)$ is some *neighborhood* of x.



Parametric and Structured Models

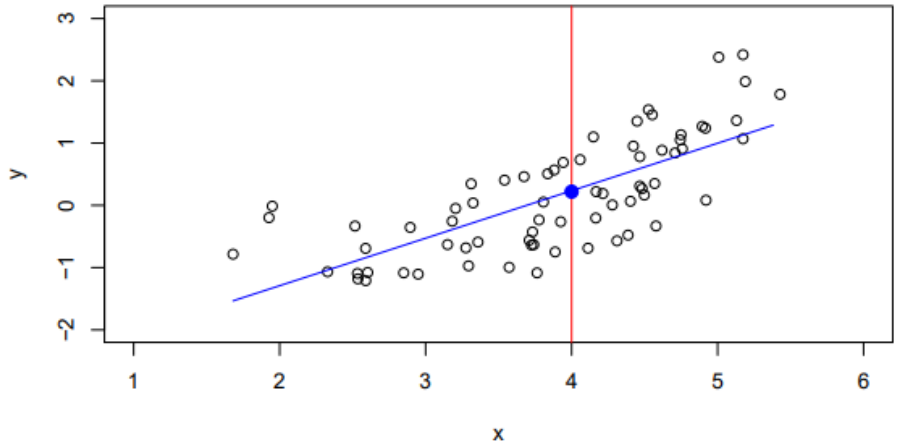
The **linear** model is an important example of a parametric model:

$$f_L(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p.$$

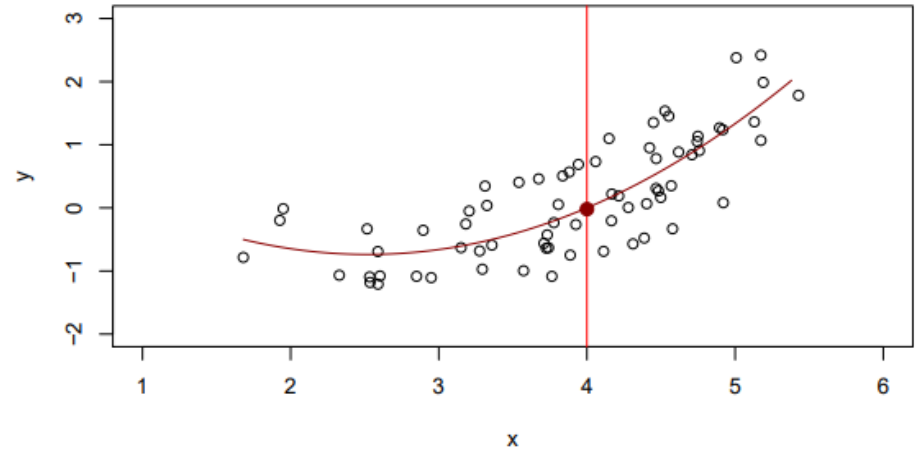
- A linear model is specified in terms of $p + 1$ parameters $\beta_0, \beta_1, \ldots, \beta_p$.
- We estimate the parameters by fitting the model to training data.
- Although it is **almost never correct**, a linear model often serves as a good and interpretable approximation to the unknown true function $f(X)$.

Parametric and Structured Models

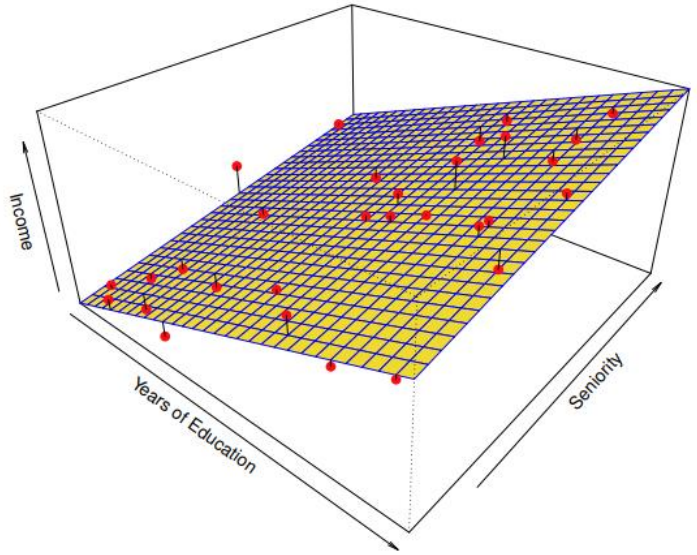
A linear model $\hat{f}_L(x) = \hat{\beta}_0 + \hat{\beta}_1 X$ gives a reasonable fit here



A quadratic model $\hat{f}_Q(x) = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$ gives a reasonable fit slightly better

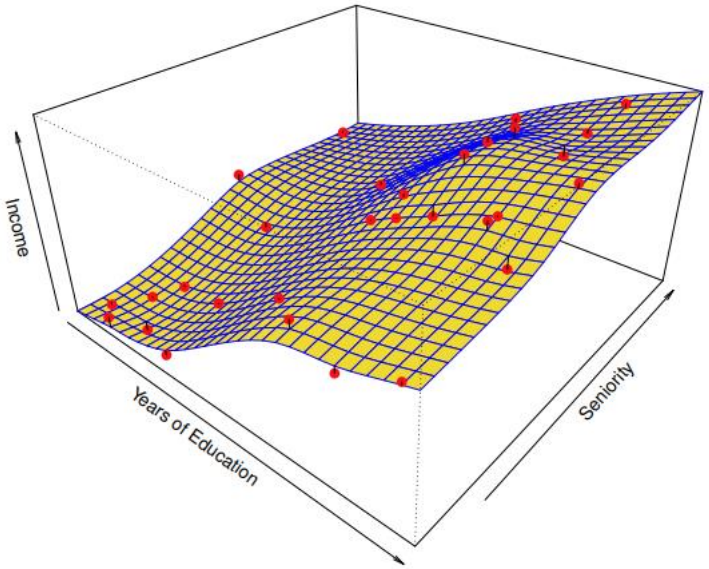


Parametric and Structured Models

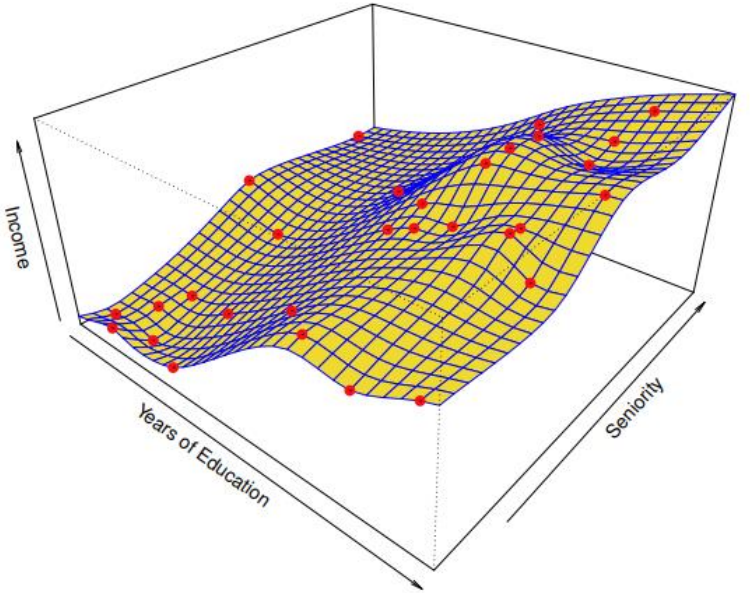


linear model fit by least squares

$\hat{f}_L(\texttt{education}, \texttt{seniority}) = \hat{\beta}_0 + \hat{\beta}_1 \times \texttt{education} + \hat{\beta}_2 \times \texttt{seniority}$



thin-plate spline regression model

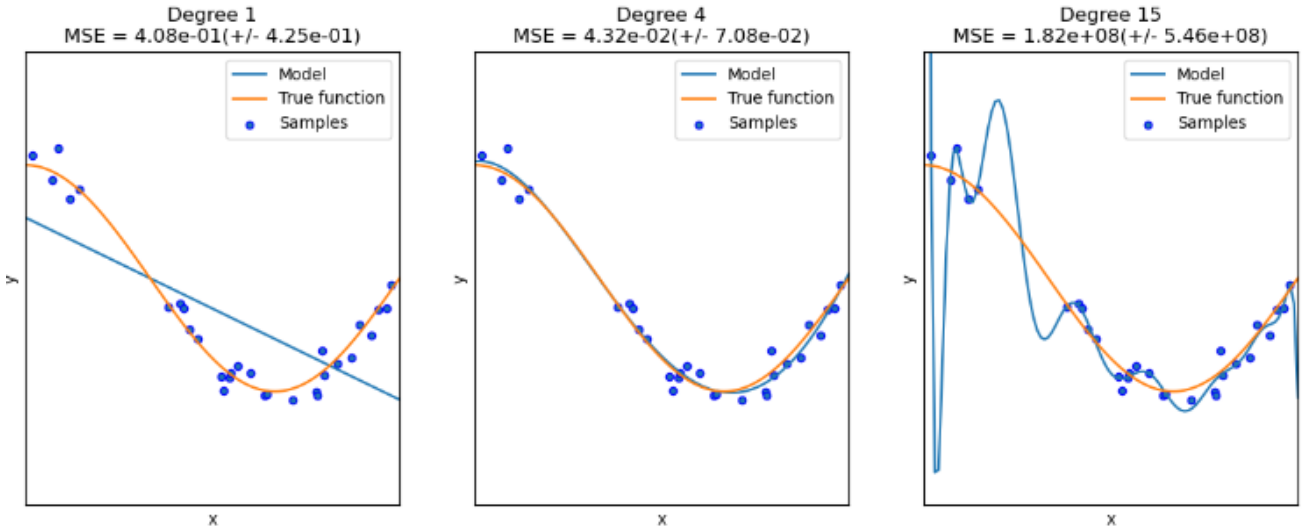


more flexible spline regression model

overfitting

Some Trade-offs

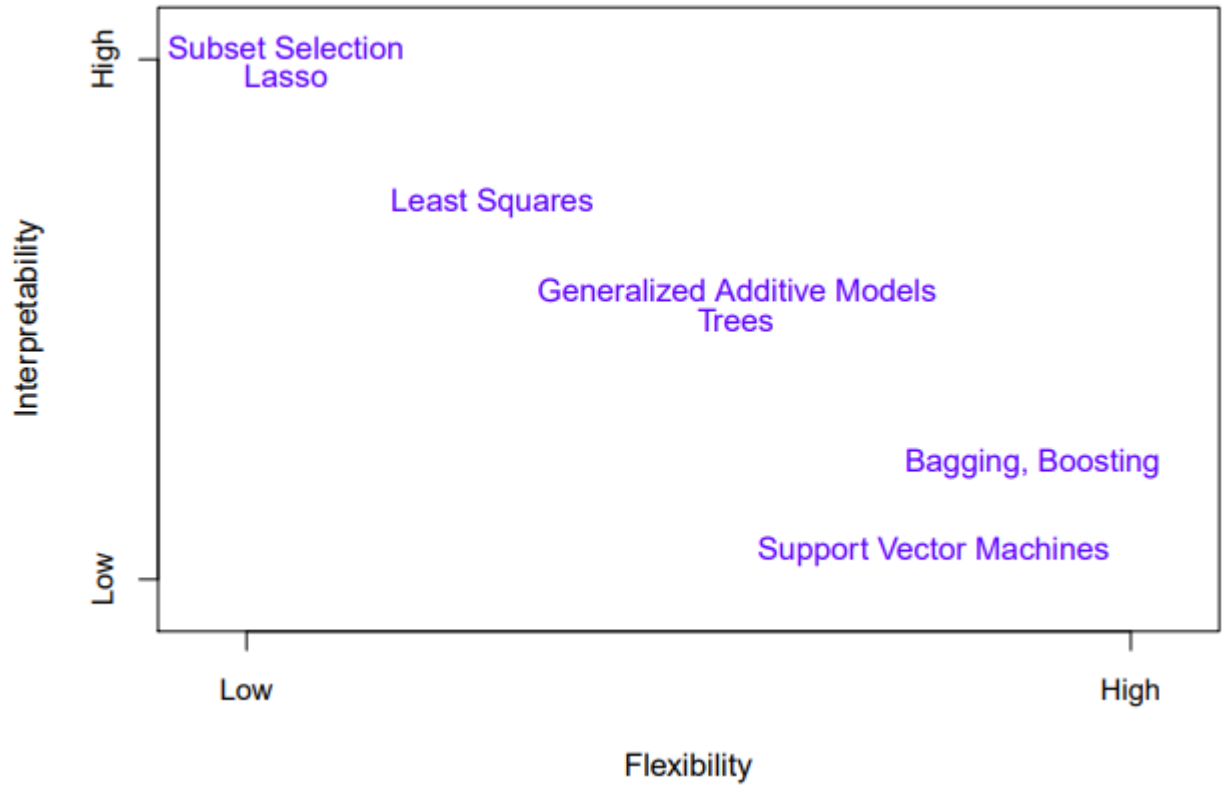
- Prediction accuracy versus interpretability.
  - Linear models are easy to interpret; thin-plate splines are not.
- Good fit versus over-fit or under-fit.
  - How do we know when the fit is just right?



- Parsimony versus black-box.
  - We often prefer a simpler model involving fewer variables over a black-box predictor involving them all.

Statistical Learning Overview (Limited)

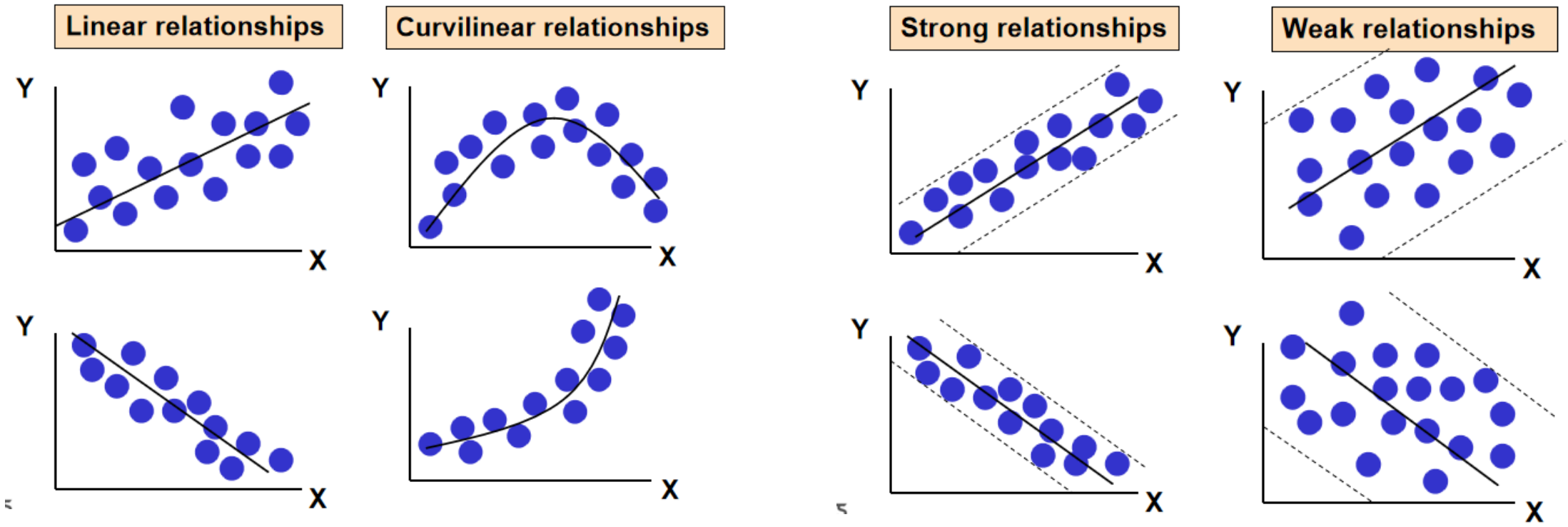
why would we ever choose to use a more restrictive method instead of a very flexible approach?



A representation of the tradeoff between flexibility and interpretability using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.

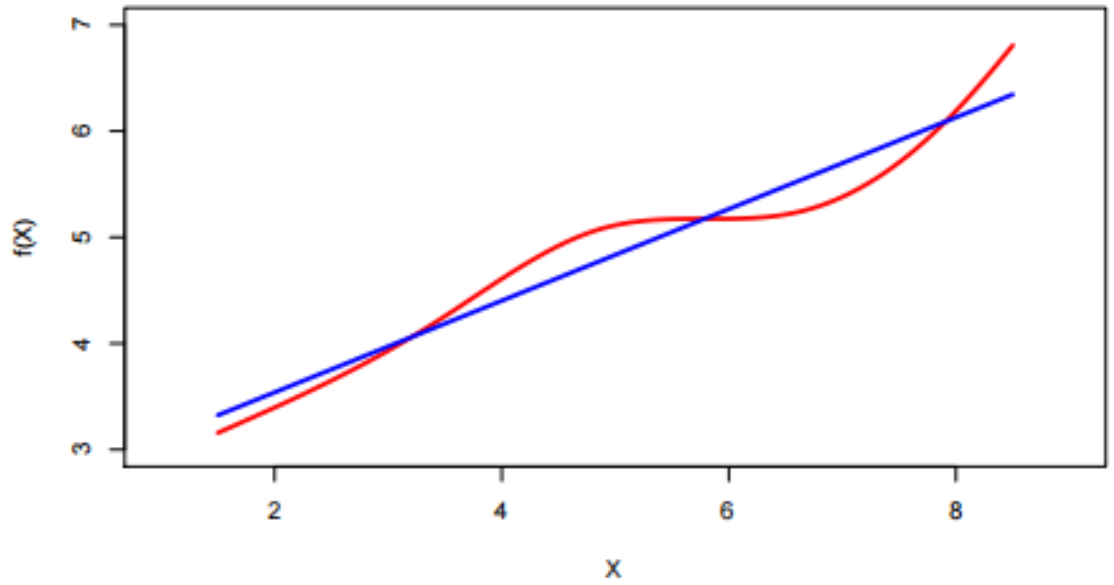
Linear Regression

- Linear regression is a simple approach to supervised learning. It assumes that the dependence of Y on $X_1, X_2, \ldots X_p$ is linear.



Linear Regression

- True regression functions are never linear!

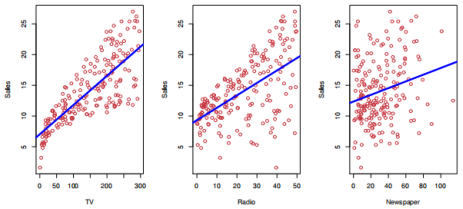


- although it may seem overly simplistic, linear regression is **extremely useful** both conceptually and practically.

Linear Regression: Motivation

Example

$$\texttt{Sales} \approx f(\texttt{TV}, \texttt{Radio}, \texttt{Newspaper})$$



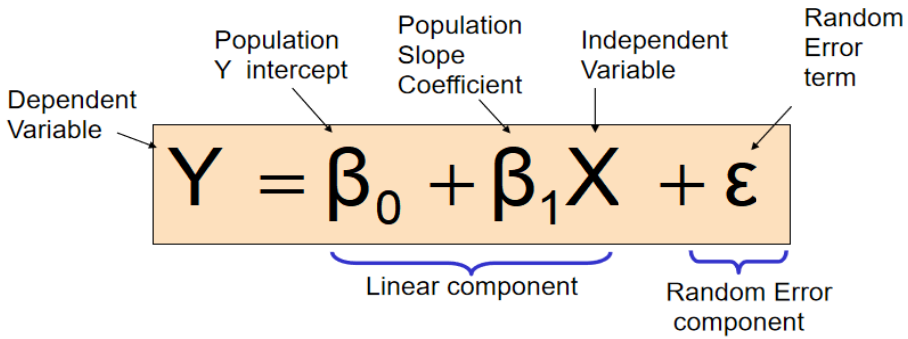
As statistical consultants, a marketing plan for next year that will result in high product sales.

What information would be useful in order to provide such a recommendation?

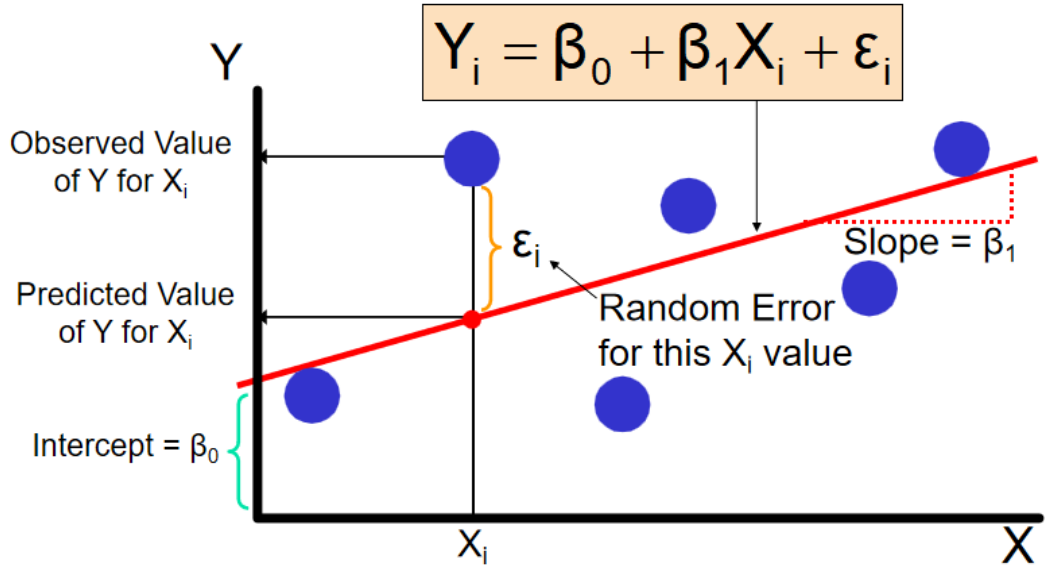
1. *Is there a relationship between advertising budget and sales?*
2. *How strong is the relationship between advertising budget and sales?*
3. *Which media are associated with sales?*
4. *How large is the association between each medium and sales?*
5. *How accurately can we predict future sales?*
6. *Is the relationship linear?*
7. *Is there synergy among the advertising media?*

Simple linear regression (a single predictor)

- We assume a model



where β_0 and β_1 are two unknown constants that represent the **intercept** and **slope**, also known as coefficients or parameters, and ε is the error term.



Simple linear regression (a single predictor)

- Given some estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, we predict future sales using

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

where $\hat{y}$ indicates a prediction of Y on the basis of $X = x$.

The hat symbol denotes an estimated values, some references use b_0, b_1.

Parameters Estimation by Least Squares

- Let $\widehat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X.
- Then $e_i = y_i - \widehat{y_i}$ represents the i^{th} 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

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

Parameters Estimation by Least Squares

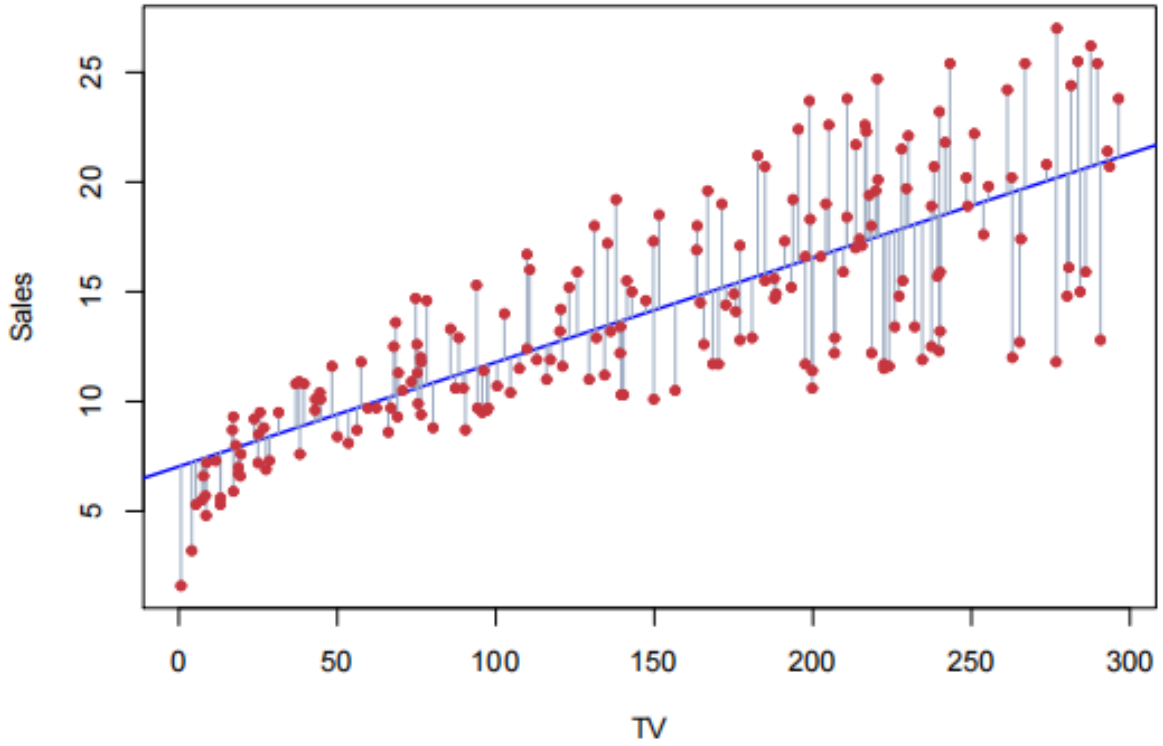
- The **least squares approach** chooses $\hat{\beta}_0$ amd $\hat{\beta}_1$ to minimize the RSS. The minimizing values can be shown to be

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

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

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means.

Parameters Estimation by Least Squares



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

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

Accuracy of the Coefficient Estimates

- 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 $\sigma^2 = Var(\epsilon)$

- [Variance unknown]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

$$\hat{\beta}_1 \pm t_{1-\alpha/2,n-2} SE(\hat{\beta}_1)$$

$$s^2 = \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2}{n-2}$$

Confidence Intervals: Coefficient Estimates

That is, there is approximately a 1-α chance that the interval

$$\left[\hat{\beta}_1 - t_{1-\alpha/2,n-2} SE(\hat{\beta}_1)\,,\, \hat{\beta}_1 + t_{1-\alpha/2,n-2} SE(\hat{\beta}_1)\right]$$

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]

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]\right)$$

Prediction Intervals: Coefficient Estimates

The predicted response y for an individual case given x_h

$$E[Y|X = x_h] = \beta_0 + \beta_1 x_h$$

The 100(1-α)% level prediction interval for a future observation on the response variable y is

$$\hat{\beta}_0 + \hat{\beta}_1 x_h \pm t_{1-\alpha/2,n-2} S \sqrt{\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}$$

$$s^2 = \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2}{n-2}$$

Hypothesis Testing: Coefficient Estimates

- 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

$$H_0: \quad \beta_1 = 0$$
$$H_A: \quad \beta_1 \neq 0$$

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

Hypothesis Testing: Coefficient Estimates

- 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 $\beta_1 = 0$.
- Using statistical software, it is easy to compute the probability of observing any value equal to |t| or larger.

	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

Overall Accuracy of the Model

- We compute the **Residual Standard Error**

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

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

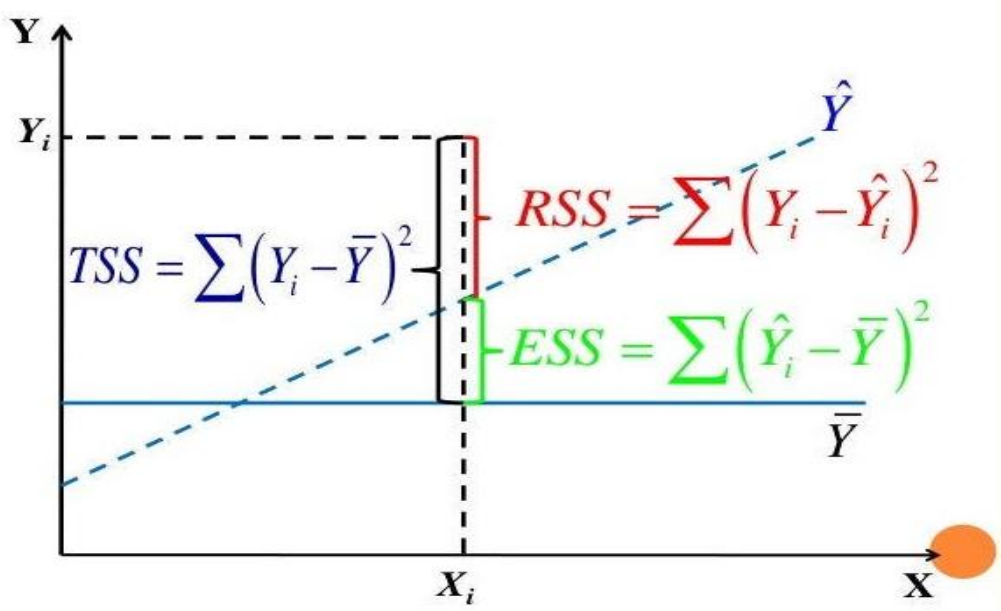
$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

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

- It can be shown that in this simple linear regression setting that $R^2 = r^2$, where r is the correlation between X and Y :

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Overall Accuracy of the Model



Quantity	Value
Residual Standard Error	3.26
R^2	0.612
F-statistic	312.1

Are Low R^2 Values Inherently Bad? → **NO**

- In some fields, it is entirely expected that the R-squared values will be low.
- Furthermore, if your R-squared value is low but you have statistically significant predictors, you can still draw important conclusions about how changes in the predictor values are associated with changes in the response value.
- Judge on R^2 wisely!

Multiple Linear Regression

- Here our model is

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

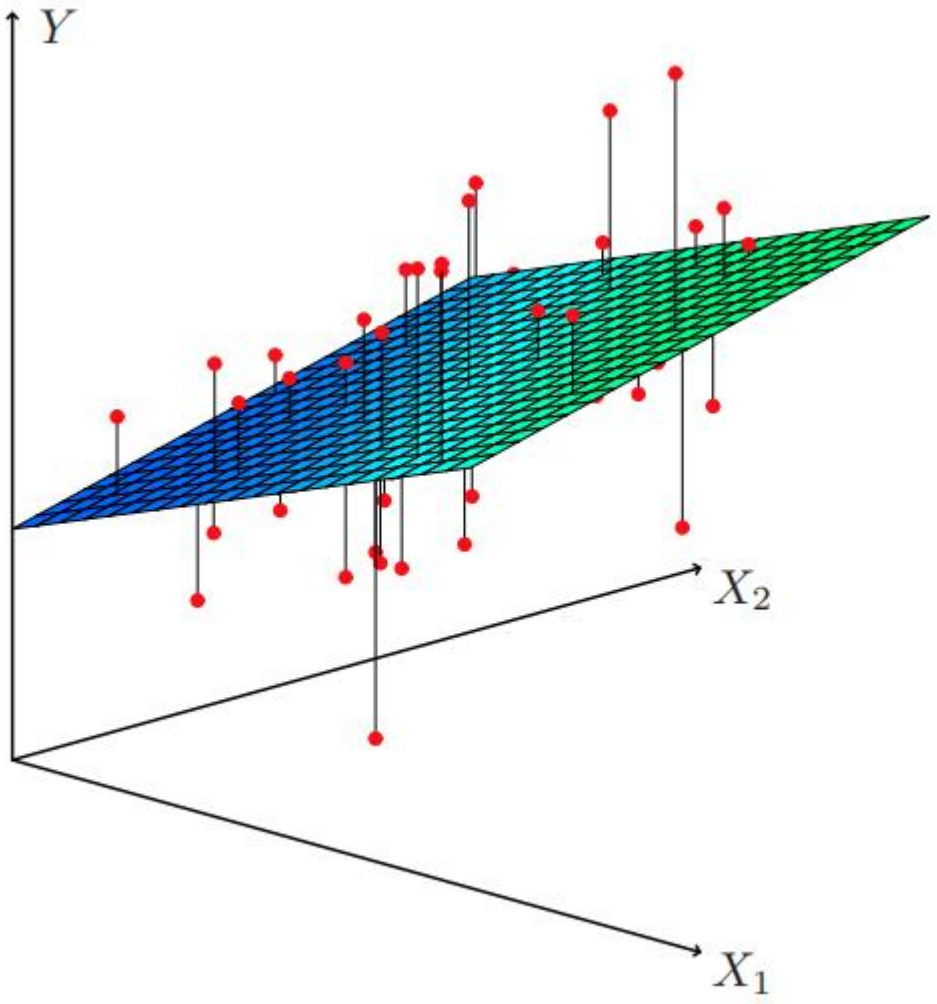
- We interpret β_j as the **average** effect on Y of a one unit increase in X_j , **holding all other predictors fixed**. In the advertising example, the model becomes

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

95% CI of the β_j

$$\left[\hat{\beta}_j - t_{1-\alpha/2,n-p}(\hat{\beta}_j)\,,\, \hat{\beta}_j + t_{1-\alpha/2,n-p}(\hat{\beta}_j)\right]$$

Multiple Linear Regression



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.

- **Claims of causality** should be avoided for observational data.

Two Quotes by Famous Statisticians

"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"

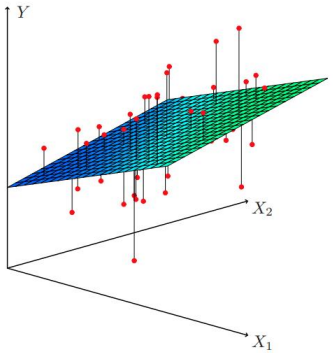
Fred Mosteller and John Tukey, paraphrasing George Box

Estimation and Prediction for Multiple Reg.

- 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 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.

Multiple Regression: R

	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

1. *Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?*

For the first question, we can use the F-statistic

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)} \sim F_{p,n-p-1}$$

Quantity	Value
Residual Standard Error	1.69
R^2	0.897
F-statistic	570

Some Important Questions

1. *Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?*

2. *Do all the predictors help to explain Y , or is only a subset of the predictors useful? [Model Selection → Next class]*

3. *How well does the model fit the data?*

4. *Given a set of predictor values, what response value should we predict, and how accurate is our prediction?*

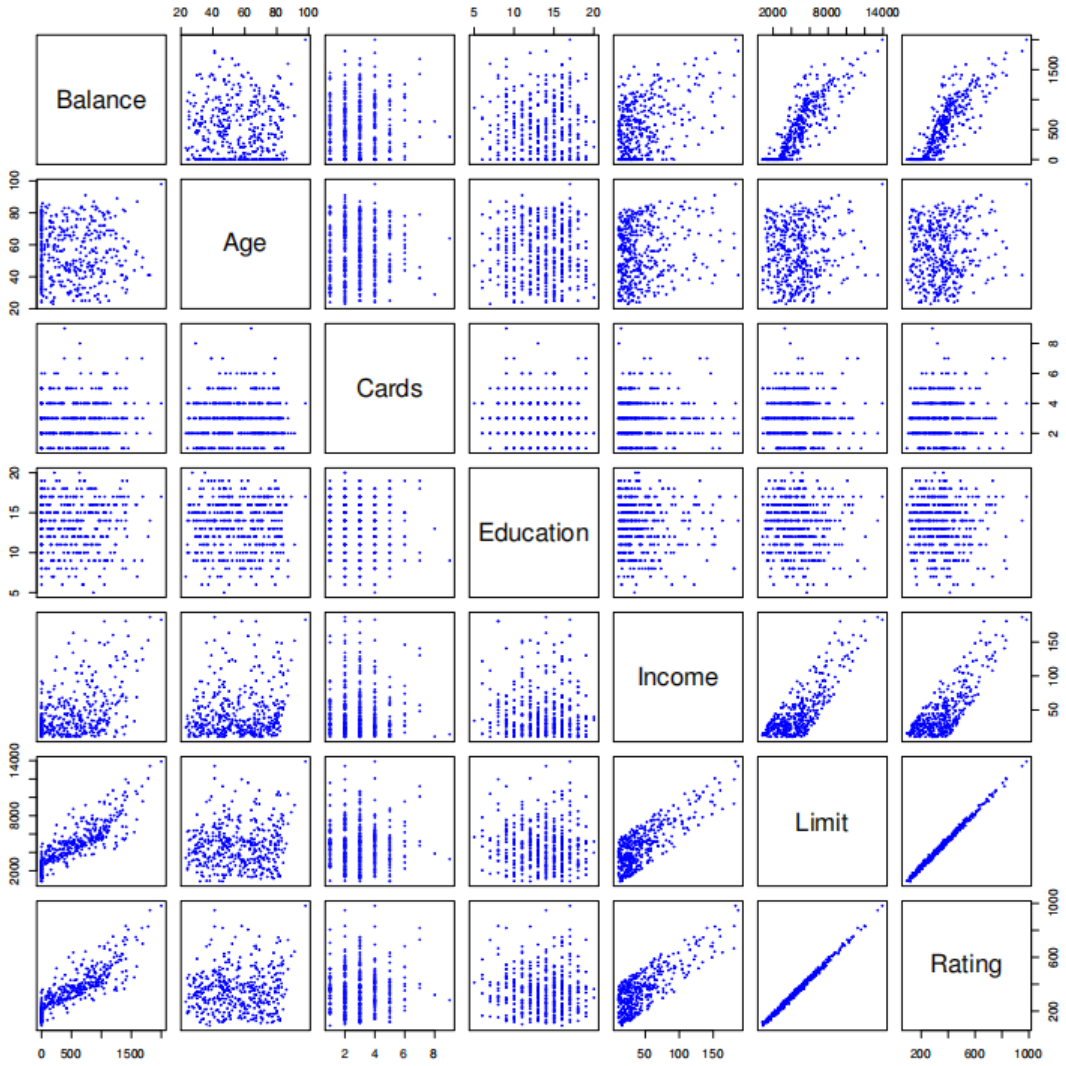
Some Considerations in Regression

Qualitative Predictors

- Some predictors are not quantitative but are qualitative, taking a discrete set of values.
- These are also called **categorical** predictors or **factor variables**.

- See for example the scatterplot matrix of the credit card data in the next slide.
  - In addition to the 7 quantitative variables shown, there are four qualitative variables: gender, student (student status), status (marital status), and ethnicity (Caucasian, African American (AA) or Asian)

Some Considerations in Regression



The Credit data set contains information about balance, age, cards, education, income, limit, and rating for a number of potential customers.

Qualitative Predictors

Example:

investigate differences in credit card balance between males and females, ignoring the other variables. We create a new variable

Dummy 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:

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

Interpretation?

Qualitative Predictors

Results for gender model

Dummy Variable (0,1) – 1 representing female

	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

Extensions of the Linear Model

Removing the additive assumption: **interactions**

Interactions:

- In our previous analysis of the Advertising data, 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

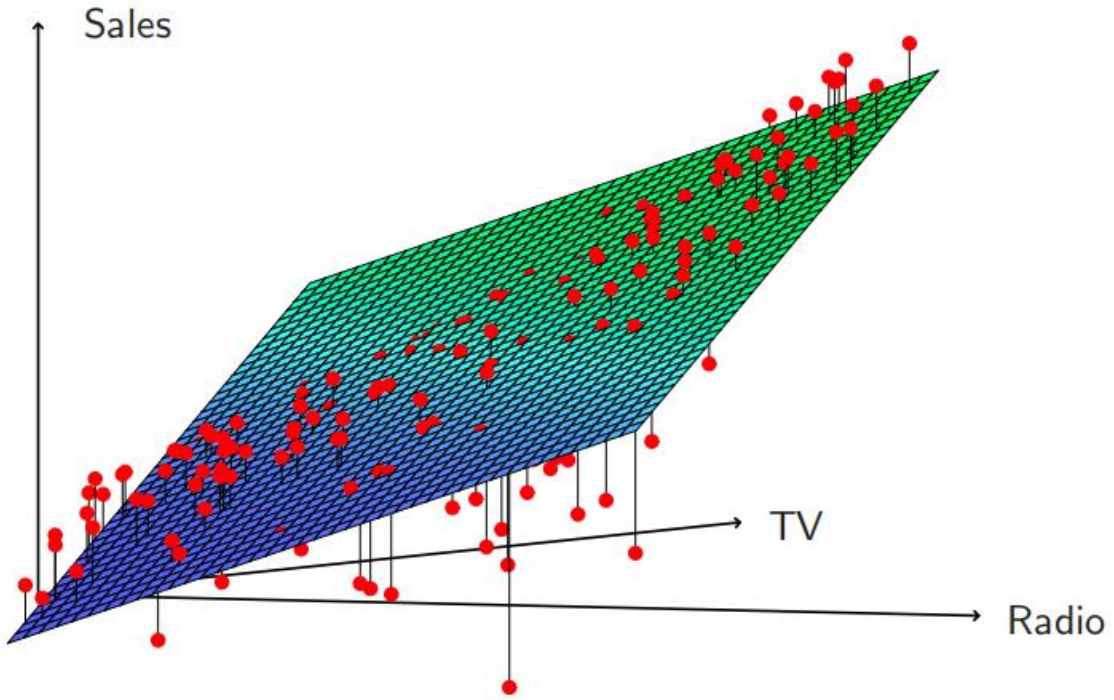
$$\widehat{\texttt{sales}} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper}$$

states that the average effect on sales of a one-unit increase in TV is always β_1, regardless of the amount spent on radio.

Linear Model: Interaction

- But suppose that spending money on radio advertising actually increases the effectiveness of TV advertising, so that the slope term for TV should increase as radio increases.
- In this situation, 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 either TV or to radio.
- In marketing, this is known as a synergy effect, and in statistics it is referred to as an **interaction** effect.

Linear Model: Interaction



- 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.

Linear Model: Interaction

Model takes the form

$$\begin{aligned} \texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{radio} \times \texttt{TV}) + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \texttt{radio}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon. \end{aligned}$$

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

- The results in this table suggests that interactions are important.
- The p-value for the interaction term TV×radio is extremely low, indicating that there is strong evidence for $H_A : \beta 3 = 0$.

Linear Model: Interaction

Model and Results:

$$\begin{aligned}\texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{radio} \times \texttt{TV}) + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \texttt{radio}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon.\end{aligned}$$

	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

- The R^2 for the interaction model is 96.8%, compared to only 89.7% for the model that predicts sales using TV and radio without an interaction term.
- This means that (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 $(\hat{\beta}_1 + \hat{\beta}_3 \times \texttt{radio}) \times 1000 = 19 + 1.1 \times \texttt{radio}$ units.
- An increase in radio advertising of \$1, 000 will be associated with an increase in sales of $(\hat{\beta}_2 + \hat{\beta}_3 \times \texttt{TV}) \times 1000 = 29 + 1.1 \times \texttt{TV}$ units.

Linear Model: Interaction

Interactions between qualitative and quantitative variables

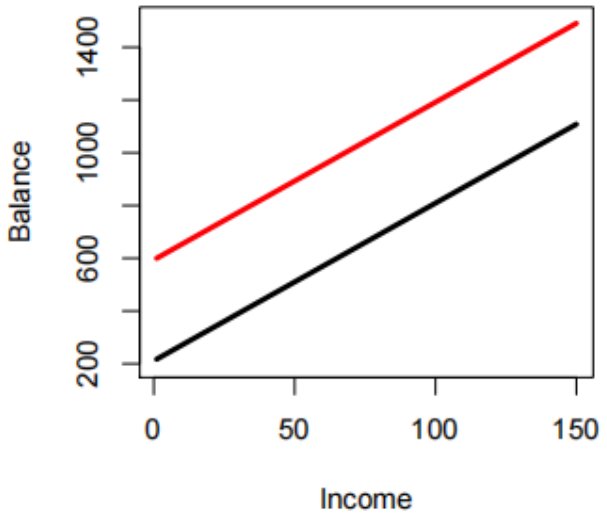
Consider the Credit data set, and suppose that we wish to predict balance using income (quantitative) and student (qualitative).

Without an interaction term, the model takes the form

$$\begin{aligned} \texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases} \\ &= \beta_1 \times \texttt{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{if } i\text{th person is a student} \\ \beta_0 & \text{if } i\text{th person is not a student.} \end{cases} \end{aligned}$$

Linear Model: Interaction

Interactions between qualitative and quantitative variables



Left: no interaction between income and student.

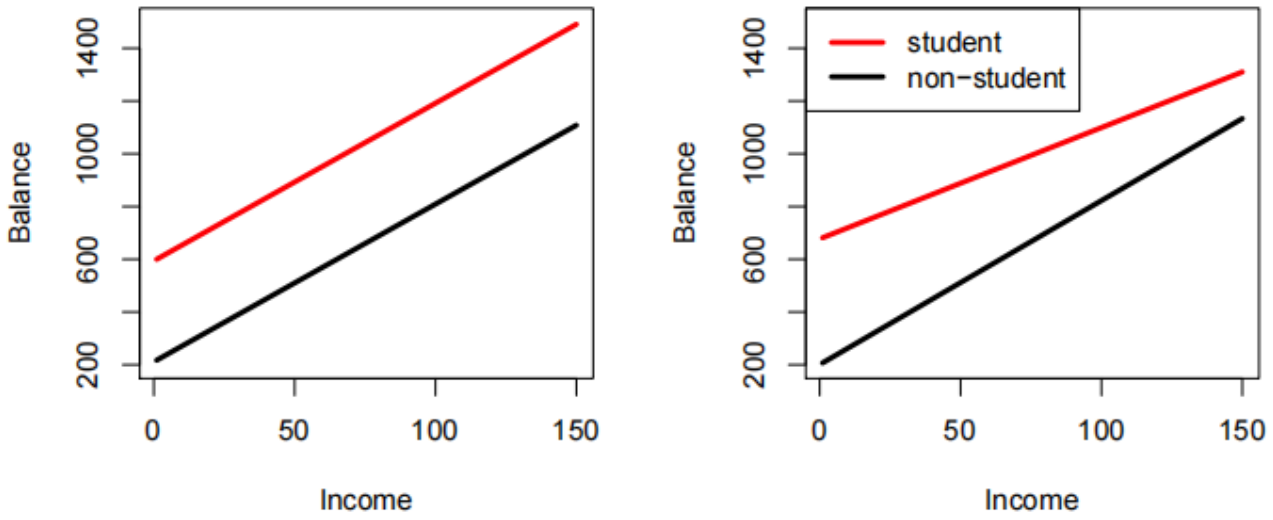
Linear Model: Interaction

Interactions between qualitative and quantitative variables

With an interaction term, the model takes the form

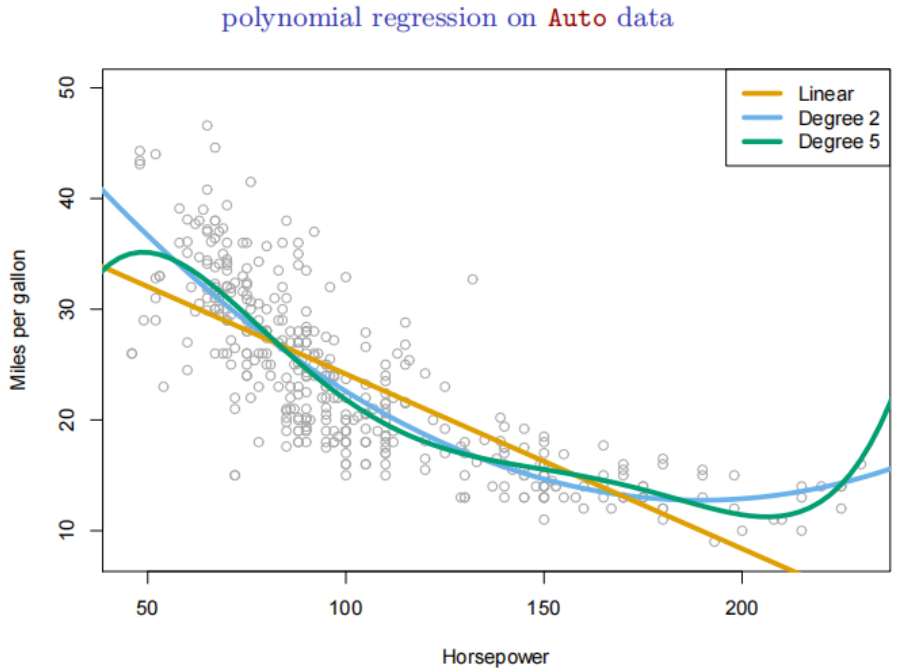
$$\texttt{balance}_i \approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 + \beta_3 \times \texttt{income}_i & \text{if student} \\ 0 & \text{if not student} \end{cases}$$

$$= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \texttt{income}_i & \text{if student} \\ \beta_0 + \beta_1 \times \texttt{income}_i & \text{if not student} \end{cases}$$



Left: no interaction between income and student.

Non-Linear Effects of Predictors



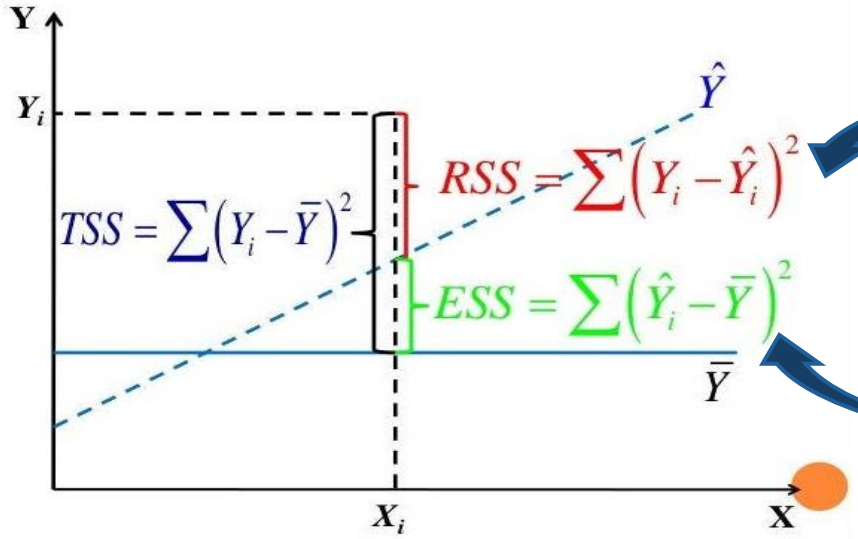
The figure suggests that

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

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
horsepower^2	0.0012	0.0001	10.1	< 0.0001

ANOVA



$$SS_{total} = SS_{reg} + SS_{error}.$$

- SS_{total}, the total sum of squares, measures the total variation in response.
- SS_{reg}, the sum of squares due to regression or, more precisely, due to the inputs, measures variation in response explained by that of the inputs.
- SS_{error}, the sum of squares due to error, measures the size of randomness due to error or noise.

ANOVA

Source of Variation	SumOfSquares	Degree of Freedom	Mean Squared	F-statistic
Regression	SS_{reg}	p	MS_{reg}	MS_{reg}/MS_{error}
Error	SS_{error}	$n-p-1$	MS_{error}	
Total	SS_{total}	$n-1$		

Where $MS_{reg} = SS_{reg}/p$ and $MS_{error} = SS_{error}/(n-p-1)$
And the F-statistic follow $F_{p,n-p-1}$ distribution under the hypothesis that $\beta_1 = \beta_2 = \cdots = \beta_p = 0$

- Linear regression is used to analyze continuous relationships; however, regression is essentially the same as ANOVA.
- In ANOVA, we calculate means and deviations of our data from the means.
- In linear regression, we calculate the best line through the data and calculate the deviations of the data from this line.
- The F ratio can be calculated in both.

ANOVA

Example

We are wanting to understand how to explain the WeightLoss variable from the diet variable.

Exercise	Diet	WeightLoss
Cardio	A	22.6
Cardio	A	18.9
Cardio	B	5.9
Cardio	B	5.8
Weights	A	9.7
Weights	A	7.1
Weights	B	9.8
Weights	B	12.7

Analysis of Variance Table

	Df	SS	Mean SS	F Value	P(>F)
Between-group	1	284.6	284.62	12	0.00133
Within-group	38	901.4	23.72		

OR

Analysis of Variance Table

	Df	SS	Mean SS	F Value	P(>F)
Diet	1	284.6	284.62	12	0.00133
Residuals	38	901.4	23.72		

Observations:

- The F Value is well over 1 indicating that this variable has some explanatory value for WeightLoss.
- The P-Value is statistically significant at the 0.05 level.

Assumptions of Regression

- **Linear Relationship**
- **Equal Variance (Homoscedasticity)**
  - The probability distribution of the errors has constant variance
- **Normality of Error**
  - Error values (ε) are normally distributed for any given value of X
- **Independence of Errors**
  - Error values are statistically independent
- **No or little Multicollinearity**

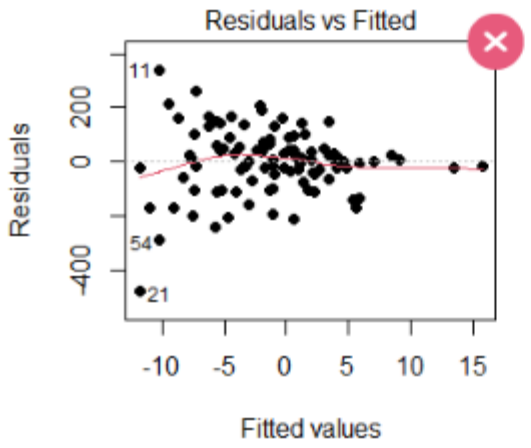
Use Graphical Analysis of Residuals!

Assumptions of Regression: Residual Plots

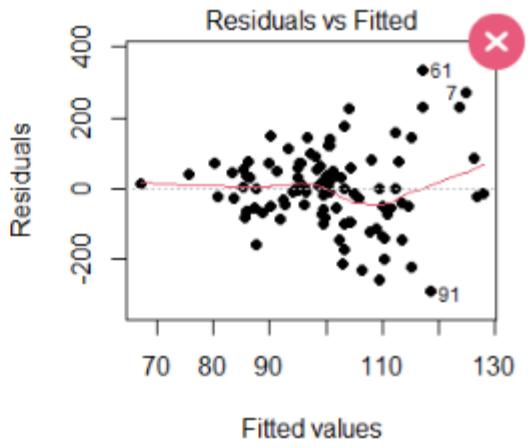
Residual Plots

A residual plot is a scatterplot of the residuals (difference between the actual and predicted value) against the predicted value. $e_i = y_i - \hat{y}_i$

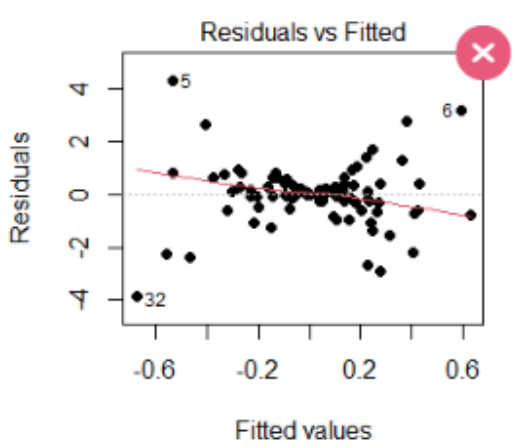
A: Fan-shaped plot



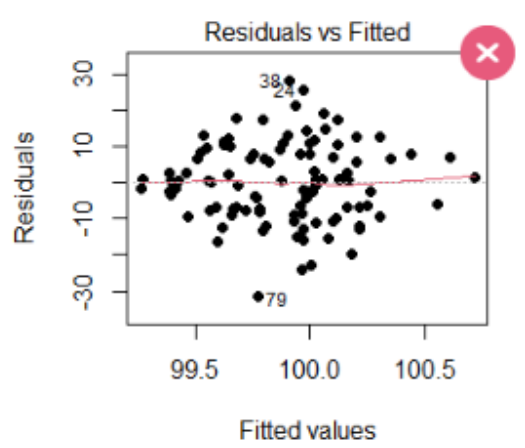
B: Fan-shaped plot



C: Butterfly-shaped plot



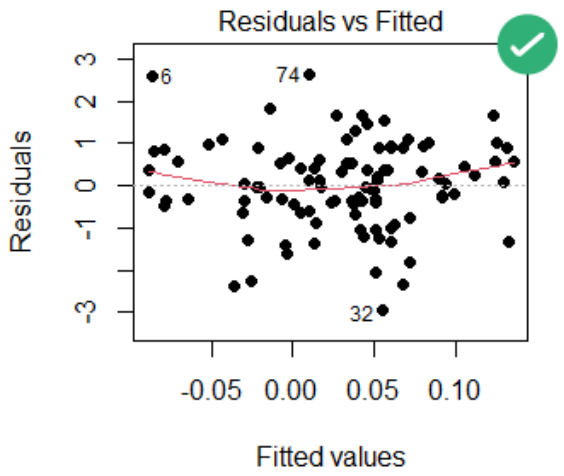
D: Blob-shaped plot



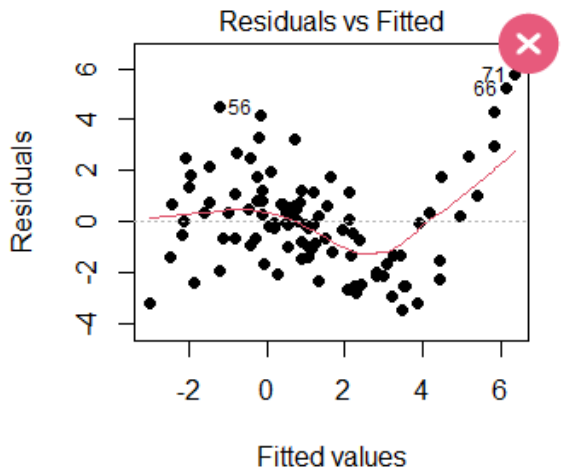
- A proper model will exhibit a **random pattern** for the spread of the residuals with no discernable shape.
- Residual plots are used extensively in linear regression for diagnostics and assumption testing.
- When a pattern is observed, a linear regression model is probably not appropriate for the data.
- If the residuals form a curvature like shape, then we know that a transformation will be necessary.

Assumptions of Regression: Linearity

A: Linearity assumption satisfied:



B: Linearity assumption violated:



Techniques to correct non-linearity:

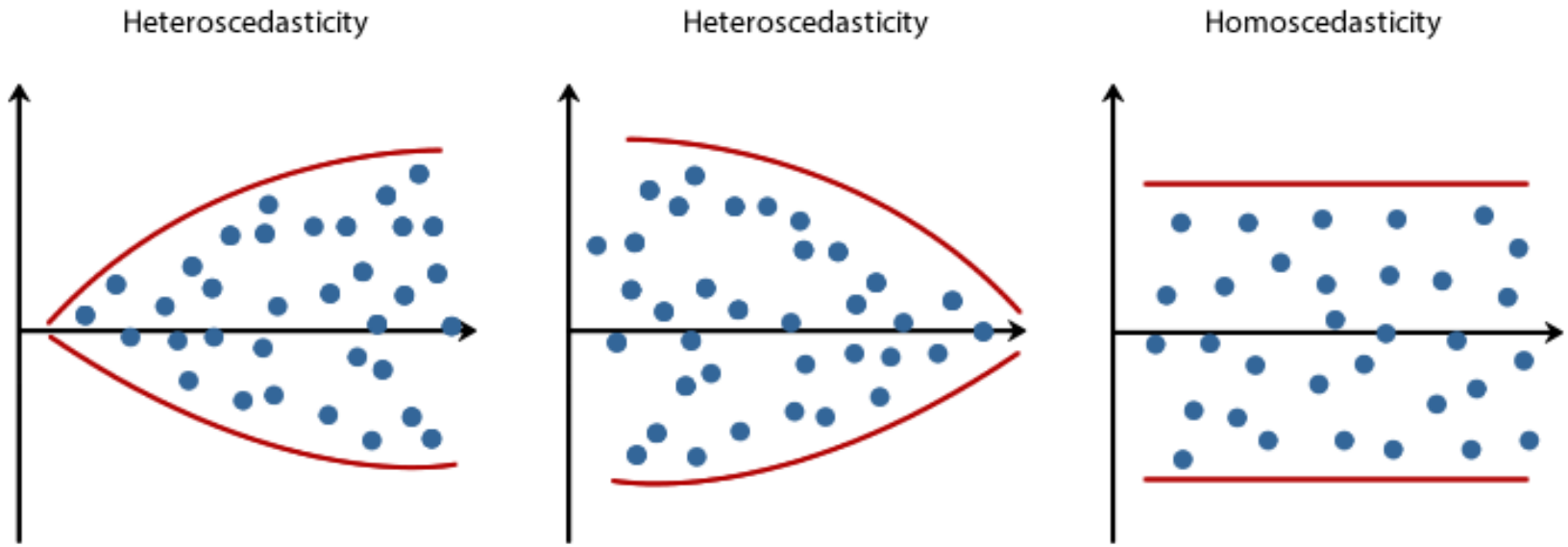
- Transforming the predictor X (log, square root): $Y = \log(X_1) + \log(X_2)$
- Adding an interaction term (since non-linearity can be due to an interaction between predictors): $Y = X1 + X2 + X1 \times X2$
- Categorizing the predictor X (when X is a numeric variable)

Assumptions of Regression: Heteroskedasticity

Heteroskedasticity

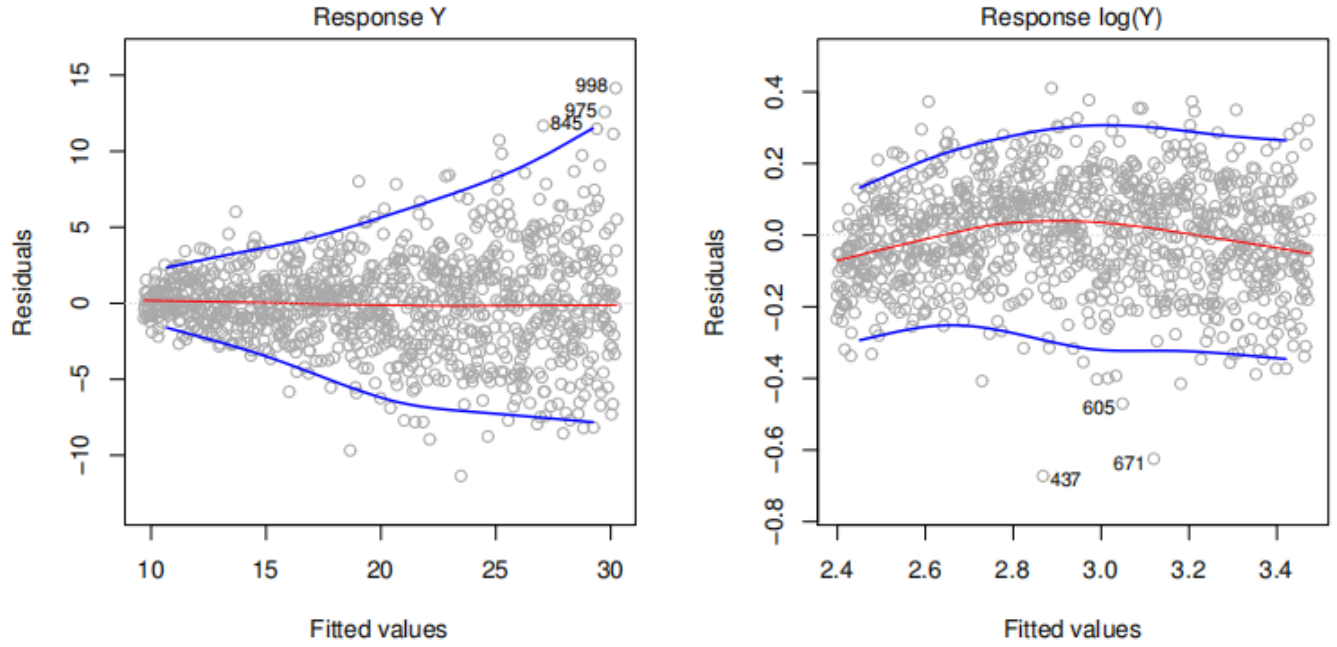
Linear Regression Analysis using OLS contains an assumption that residuals are identically distributed across every X variable. [**Equal Variance - Homoscedastic**]

- Errors have the same scatter regardless of the value of X.



Assumptions of Regression: Heteroskedasticity

Heteroskedasticity



Techniques to correct heteroskedasticity:

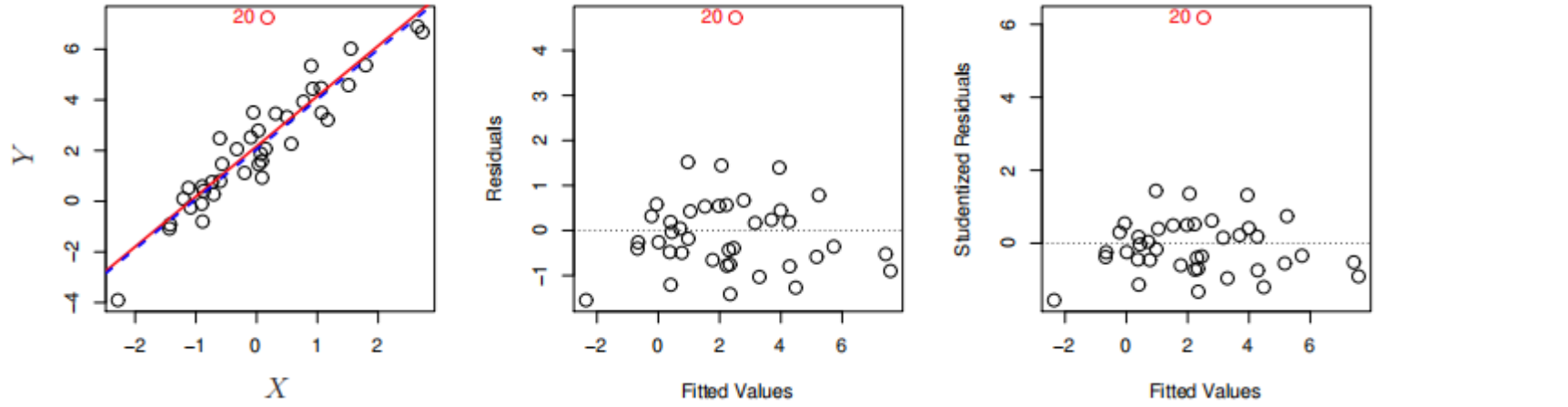
- Transforming the outcome Y (log, square root)
- Use Weighted Least Squares in place of OLS
- Converting the outcome into a binary variable

Assumptions of Regression: Outlier

Outliers

An outlier is a point for which y_i is far from the value predicted by the model.

Standardized Residuals vs Fitted

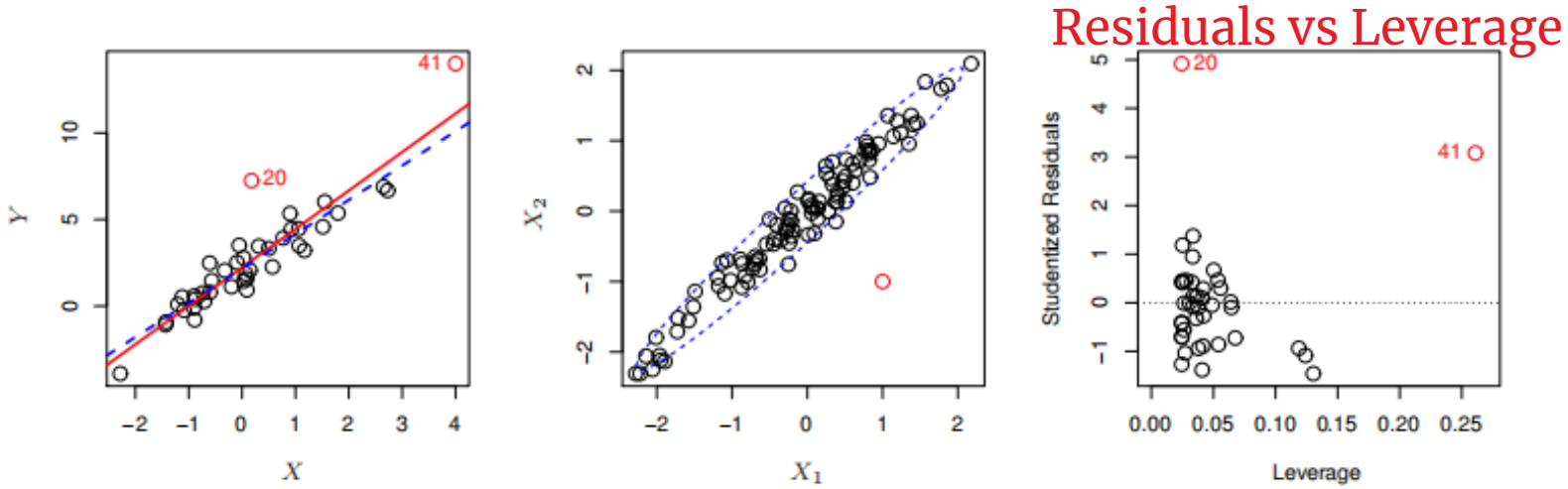


Observations whose studentized residuals are greater than **3** are possible outliers.

Assumptions of Regression: Outlier

Outliers

An outlier is a point for which y_i is far from the value predicted by the model.



- **Heteroskedasticity and non-linearity:**
  - standardized residuals shouldn't change as a function of leverage.

- Points with high leverage may be influential: deleting them would change the model a lot.

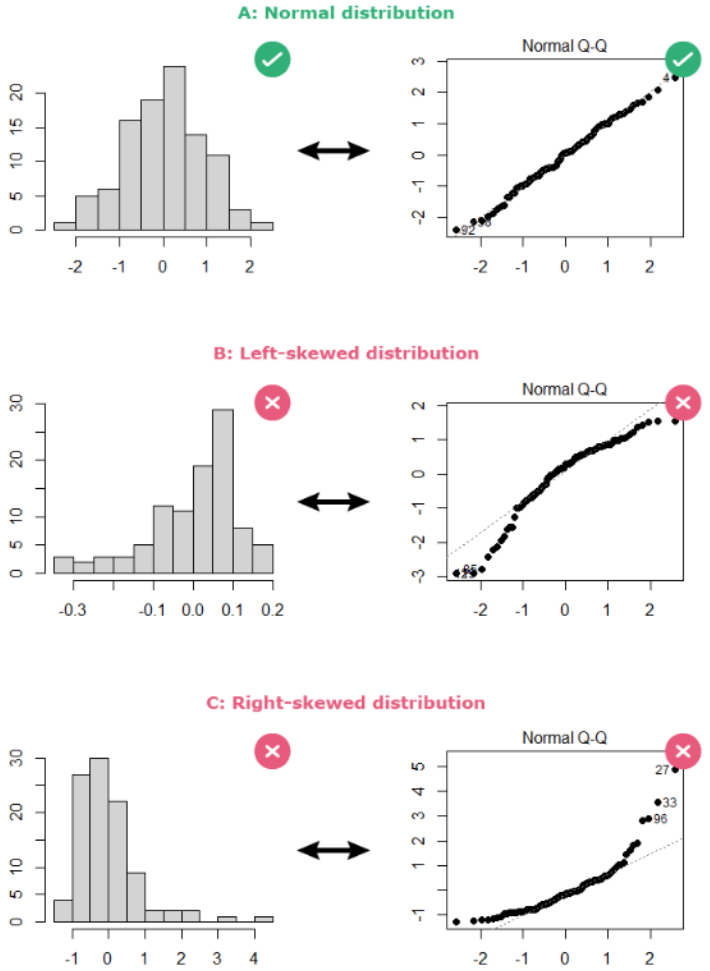
Assumptions of Regression: Normality

Check for Normality

A normal probability plot of the residuals can be used to check for normality:

The normality of the errors can be checked in 2 ways:

- histogram of the residuals: the distribution should be bell-shaped.
- normal Q-Q plot of the residuals: The points should follow the diagonal straight line.



Techniques to correct non-normality:

- Transforming the outcome Y (log, square root)
- Removing outliers (observations with Y values that are far from the regression line)
- Transforming the outcome into a binary variable then using logistic regression

Assumptions of Regression: Multicollinearity

Autocorrelation/ Independence

- Autocorrelation is correlation of the errors (residuals) over time [**Time-series class**]

Multicollinearity

Collinearity (or multicollinearity) is the undesirable situation where the correlations among the independent variables are strong.

- For instance, the model may fit the data well (high F-Test), even though none of the X variables has a statistically significant impact on explaining Y.
- **How is this possible?** When two X variables are highly correlated, they both convey essentially the same information. When this happens, the X variables are collinear and the results show multicollinearity.

Assumptions of Regression: Multicollinearity

Techniques to detect Multicollinearity:

- Formally, variance inflation factors (VIF) measure how much the variance of the estimated coefficients are increased over the case of no correlation among the X variables. If no two X variables are correlated, then all the VIFs will be 1.

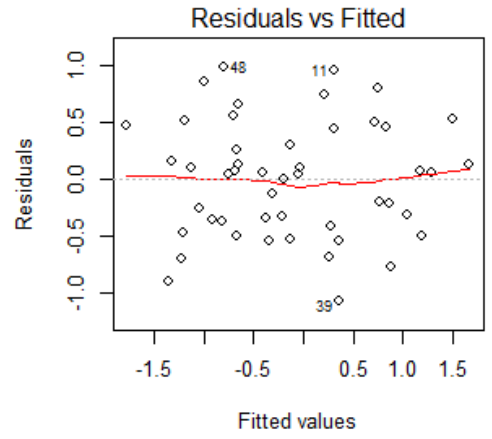
- If VIF for one of the variables is around or greater than 5, there is collinearity associated with that variable.

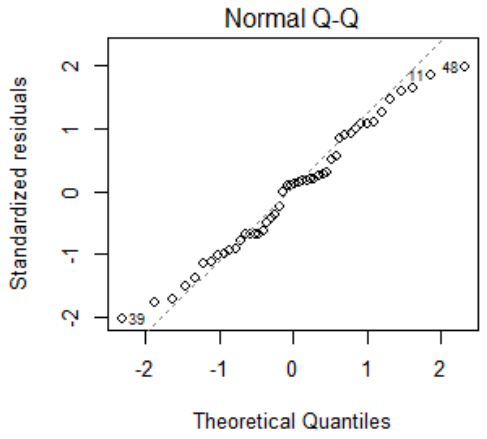
Test for Multicollinearity			
Variables	**VIF**	**Df**	**VIF^(1/(2*Df))**
education	3.97	1	2.44
income	1.68	1	1.30
type	6.10	2	1.57

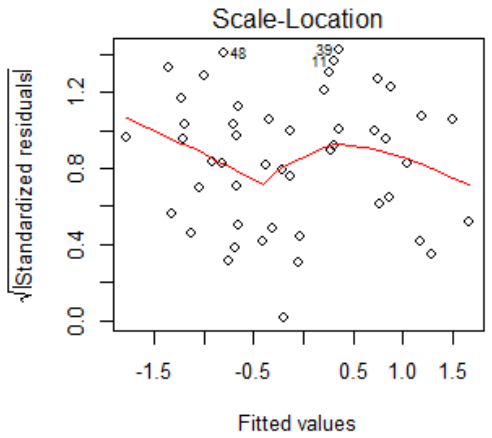
- The easy solution is: If there are two or more variables that will have a VIF around or greater than 5, one of these variables must be removed from the regression model. To determine the best one to remove, remove each one individually. Select the regression equation that explains the most variance (R^2 the highest).

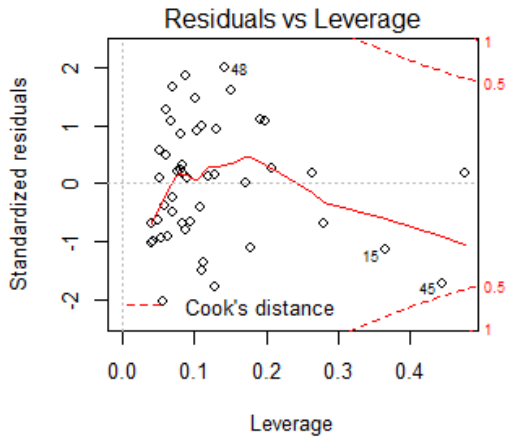
Assumptions of Regression: R Examples

Example 1

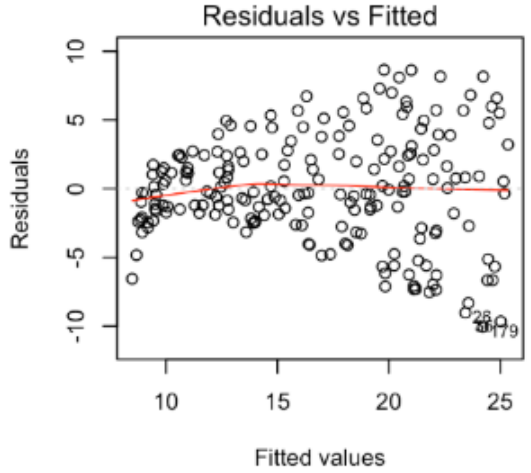


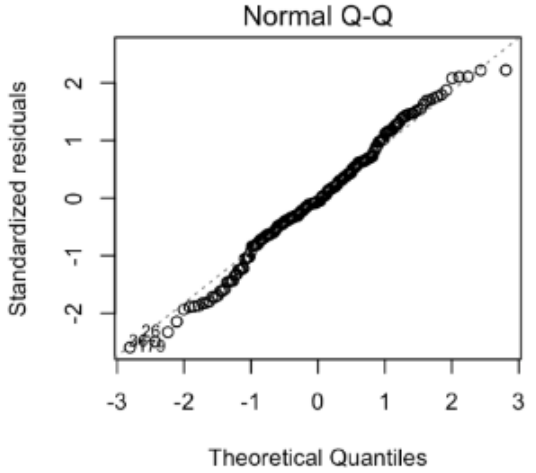


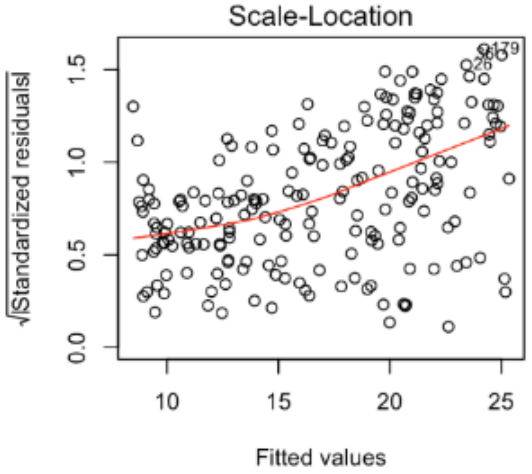


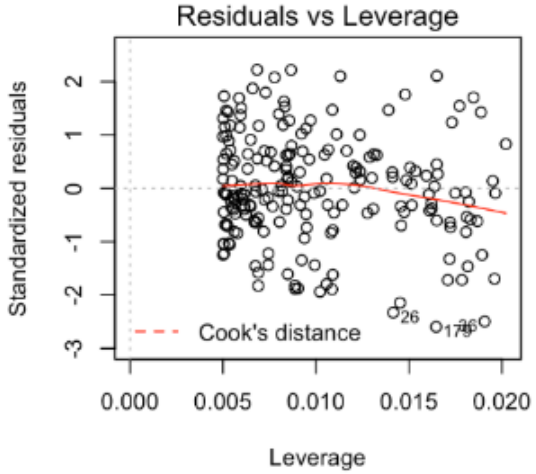


Example 2

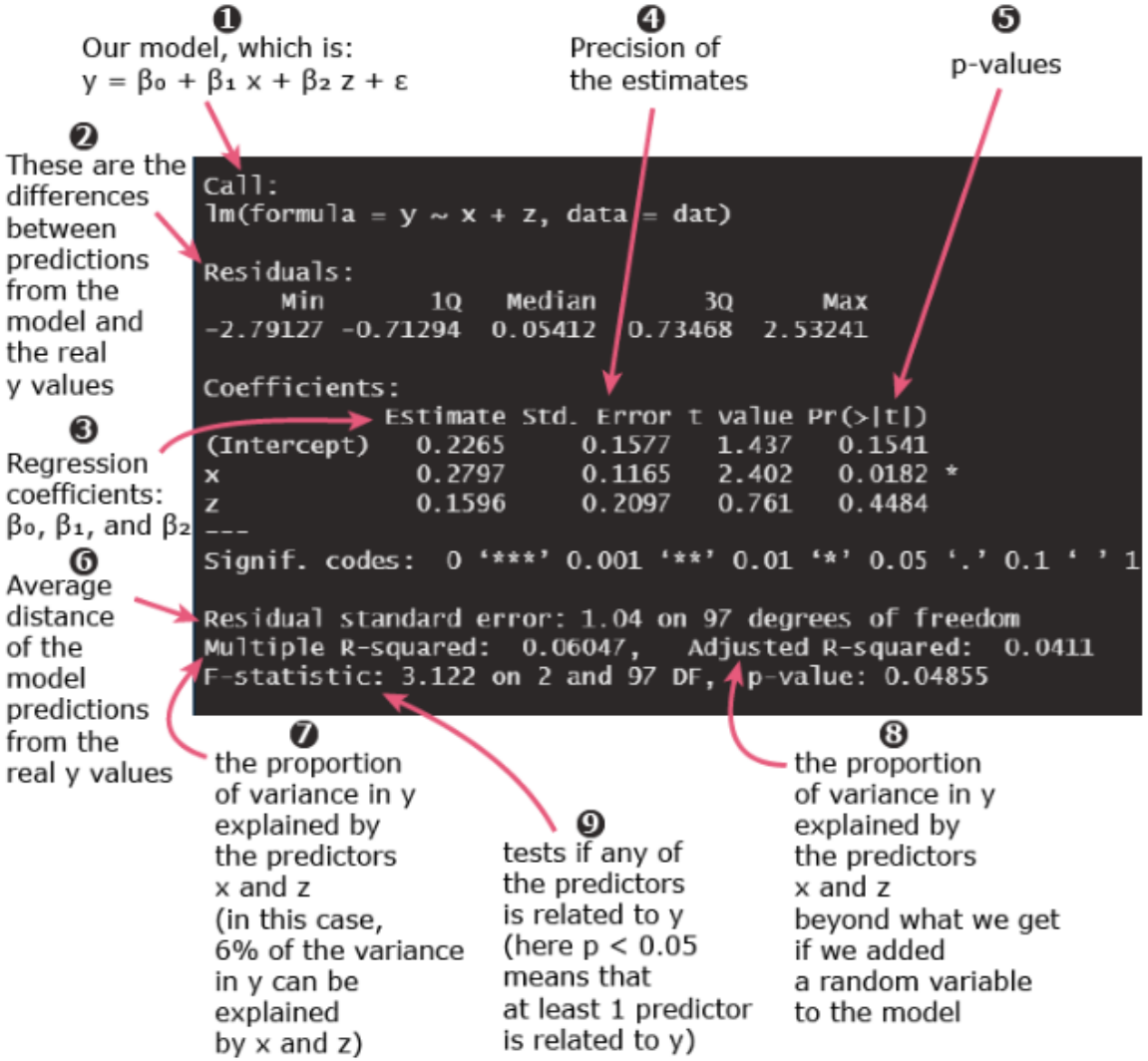








Assumptions of Regression: R Examples



R: https://libguides.princeton.edu/R-linear_regression

1. True-False: Linear Regression is a supervised algorithm.

A) TRUE

B) FALSE

2. Which of the following methods do we use to find the best fit line for data in Linear Regression?

A) Least Square Method

B) Maximum Likelihood

C) Both A and B

3. Which of the following evaluation metrics can be used to evaluate a model while modelling a continuous output variable?

A) AUC - ROC

B) Accuracy

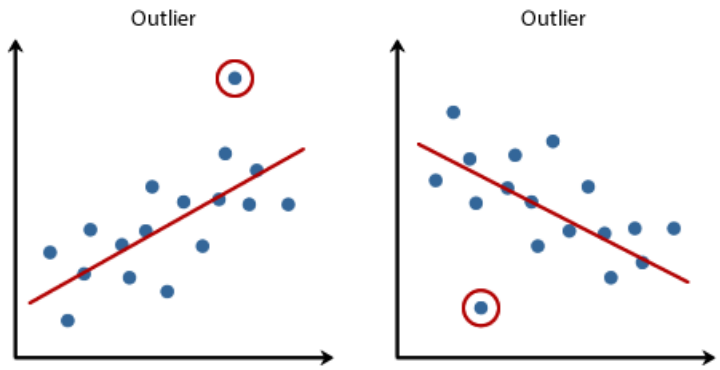
C) Mean Squared Error

4. Which of the following statements is true about outliers in Linear Regression

A) Linear Regression is sensitive to outliers

B) Linear Regression is not sensitive to outliers

C) No Idea



Lab Time!