Simple Linear Regression (single variable)

Introduction to Machine Learning

Marek Petrik

January 31, 2017

Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

Last Class

1. Basic machine learning framework

$$Y = f(X)$$

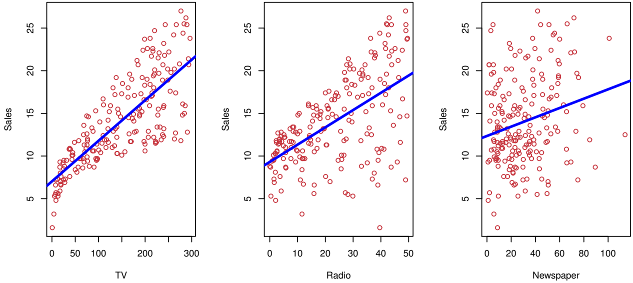
2. Prediction vs inference: predict Y vs understand f
3. Parametric vs non-parametric: linear regression vs k-NN
4. Classification vs regressions: k-NN vs linear regression
5. Why we need to have a test set: overfitting

What is Machine Learning

- Discover unknown function f:

$$Y = f(X)$$

- X = set of features, or inputs
- Y = target, or response



$$\text{Sales} = f(\text{TV}, \text{Radio}, \text{Newspaper})$$

Errors in Machine Learning: World is Noisy

- World is too complex to model precisely
- Many features are not captured in data sets
- Need to allow for errors ϵ in f:

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

How Good are Predictions?

- Learned function $\hat{f}$
- Test data: $(x_1, y_1), (x_2, y_2), \ldots$
- **Mean Squared Error (MSE)**:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2$$

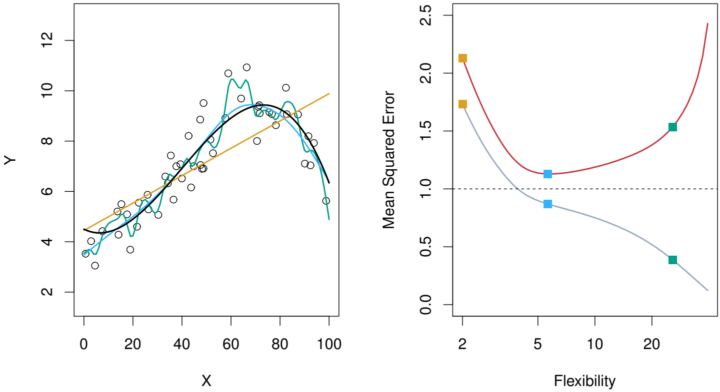
- This is the estimate of:

$$\text{MSE} = \mathbb{E}[(Y - \hat{f}(X))^2] = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} (Y(\omega) - \hat{f}(X(\omega)))^2$$

- Important: Samples x_i are i.i.d.

Do We Need Test Data?

- Why not just test on the training data?

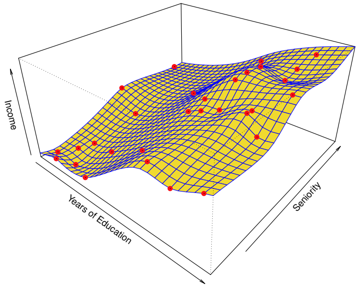


- Flexibility is the degree of polynomial being fit
- Gray line: training error, red line: testing error

Types of Function f

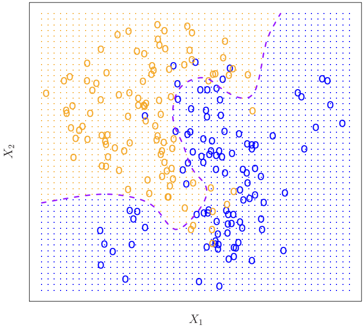
Regression: continuous target

$$f : \mathcal{X} \to \mathbb{R}$$



Classification: discrete target

$$f : \mathcal{X} \to \{1, 2, 3, \ldots, k\}$$



Bias-Variance Decomposition

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

Mean Squared Error can be decomposed as:

$$\text{MSE} = \mathbb{E}(Y - \hat{f}(X))^2 = \underbrace{\mathrm{Var}(\hat{f}(X))}_{\text{Variance}} + \underbrace{(\mathbb{E}(\hat{f}(X)))^2}_{\text{Bias}} + \mathrm{Var}(\epsilon)$$

- **Bias**: How well would method work with infinite data
- **Variance**: How much does output change with different data sets

Today

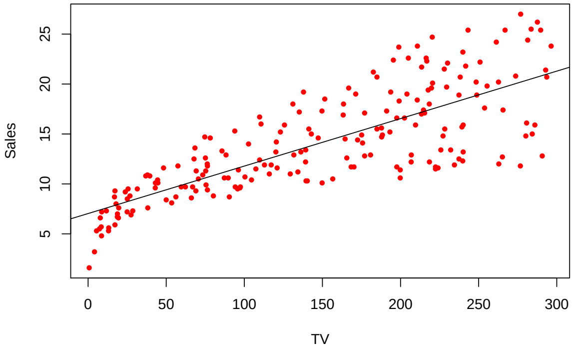
- Basics of linear regression
- Why linear regression
- How to compute it
- Why compute it

Simple Linear Regression

- We have only one feature

$$Y \approx \beta_0 + \beta_1 X \qquad Y = \beta_0 + \beta_1 X + \epsilon$$

- Example:



$$\text{Sales} \approx \beta_0 + \beta_1 \times \text{TV}$$

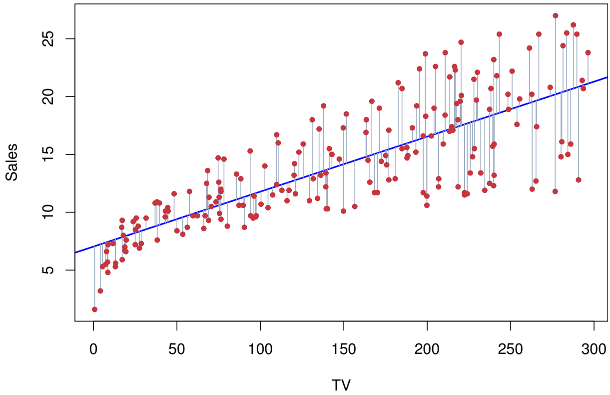
How To Estimate Coefficients

- No line that will have no errors on data x_i
- Prediction:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Errors (y_i are true values):

$$e_i = y_i - \hat{y}_i$$



Residual Sum of Squares

- Residual Sum of Squares

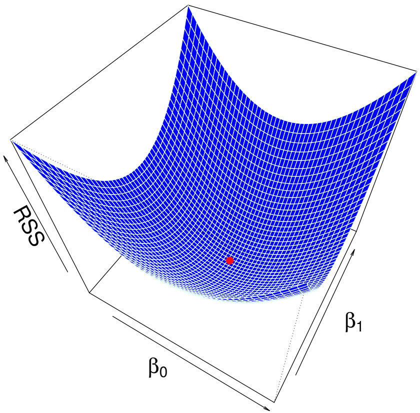
$$\mathrm{RSS} = e_1^2 + e_2^2 + e_3^2 + \cdots + e_n^2 = \sum_{i=1}^{n} e_i^2$$

- Equivalently:

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

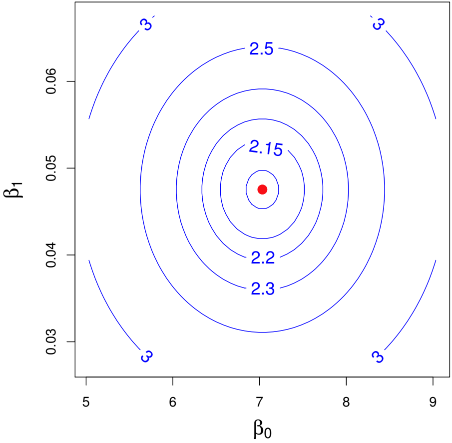
Minimizing Residual Sum of Squares

$$\min_{\beta_0,\beta_1} \text{RSS} = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} e_i^2 = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$



Minimizing Residual Sum of Squares

$$\min_{\beta_0,\beta_1} \text{RSS} = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} e_i^2 = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$



Solving for Minimal RSS

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

- RSS is a **convex** function of β_0, β_1
- Minimum achieved when (recall the chain rule):

$$\frac{\partial \, \mathrm{RSS}}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial \, \mathrm{RSS}}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

Linear Regression Coefficients

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Solution:

$$\begin{aligned} \beta_0 &= \bar{y} - \beta_1 \bar{x} \\ \beta_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sum_{i=1}^{n} x_i(x_i - \bar{x})} \end{aligned}$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Why Minimize RSS

Why Minimize RSS

1. Maximize likelihood when $Y = \beta_0 + \beta_1 X + \epsilon$ when $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Why Minimize RSS

1. Maximize likelihood when $Y = \beta_0 + \beta_1 X + \epsilon$ when $\epsilon \sim \mathcal{N}(0, \sigma^2)$

2. Best Linear Unbiased Estimator (BLUE): Gauss-Markov Theorem (ESL 3.2.2)

Why Minimize RSS

1. Maximize likelihood when $Y = \beta_0 + \beta_1 X + \epsilon$ when $\epsilon \sim \mathcal{N}(0, \sigma^2)$

2. Best Linear Unbiased Estimator (BLUE): Gauss-Markov Theorem (ESL 3.2.2)

3. It is convenient: can be solved in closed form

Bias in Estimation

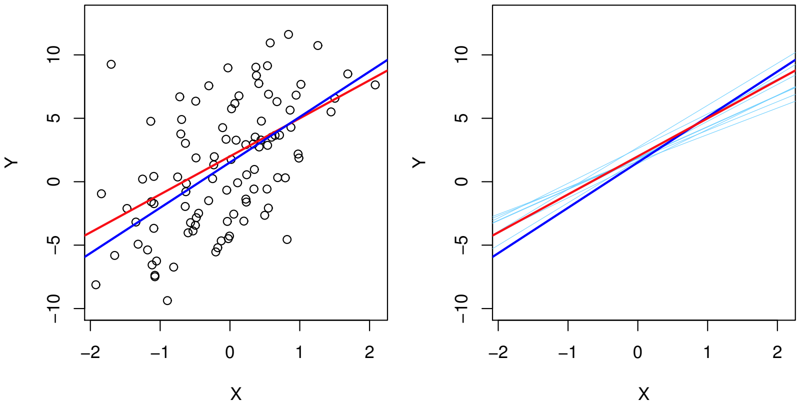
- Assume a true value $\mu^\star$
- Estimate μ is **unbiased** when $\mathbb{E}[\mu] = \mu^\star$
- Standard mean estimate is unbiased (e.g. $X \sim \mathcal{N}(0,1)$):

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = 0$$

- Standard variance estimate is biased (e.g. $X \sim \mathcal{N}(0,1)$):

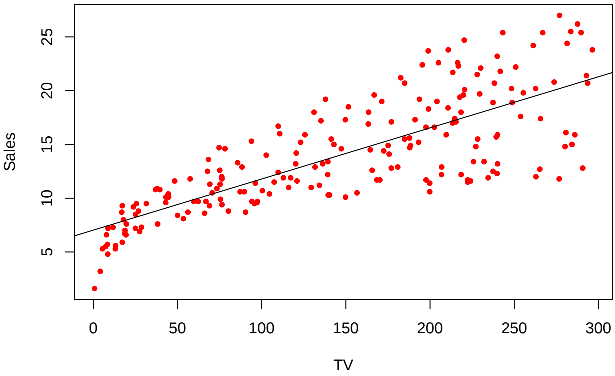
$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2\right] \neq 1$$

Linear Regression is Unbiased



Gauss-Markov Theorem (ESL 3.2.2)

Solution of Linear Regression



How Good is the Fit

- How well is linear regression predicting the training data?
- Can we be sure that TV advertising really influences the sales?
- What is the probability that we just got lucky?

R^2 Statistic

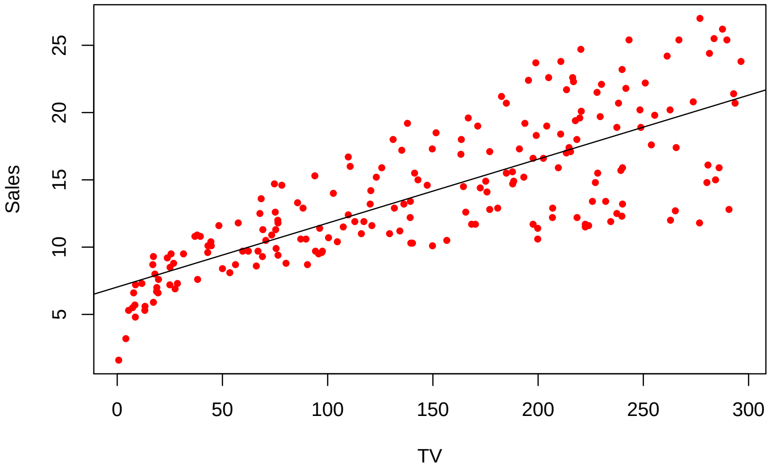
$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- RSS - residual sum of squares, TSS - total sum of squares
- R^2 measures the goodness of the fit as a proportion
- Proportion of data variance explained by the model
- Extreme values:

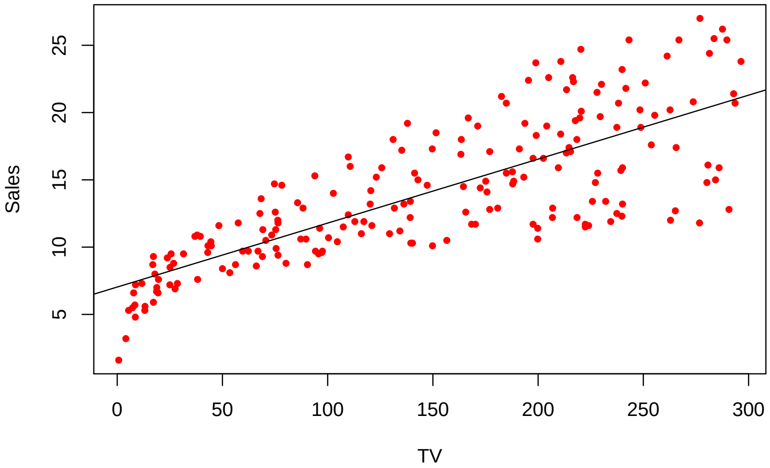
  0: Model does not explain data

  1: Model explains data perfectly

Example: TV Impact on Sales

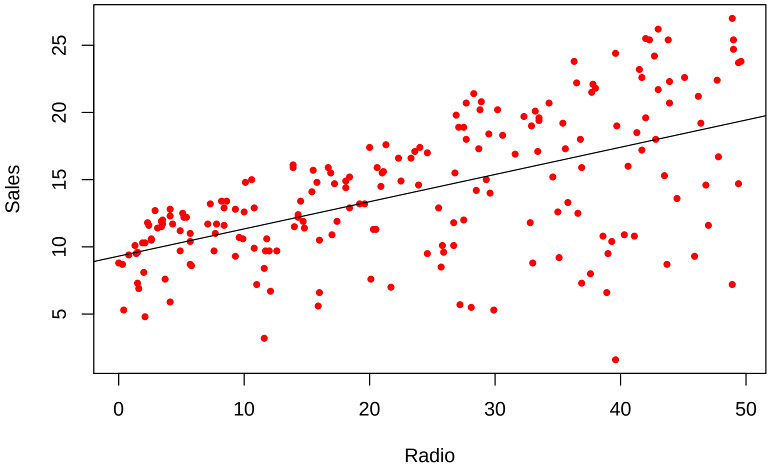


Example: TV Impact on Sales

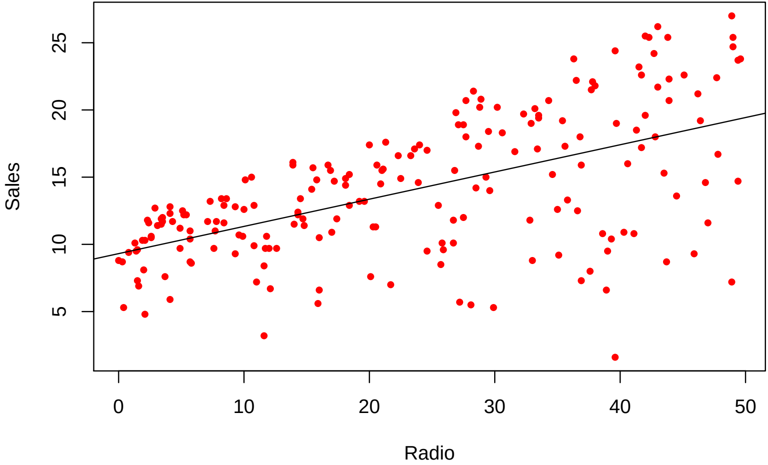


$$R^2 = 0.61$$

Example: Radio Impact on Sales

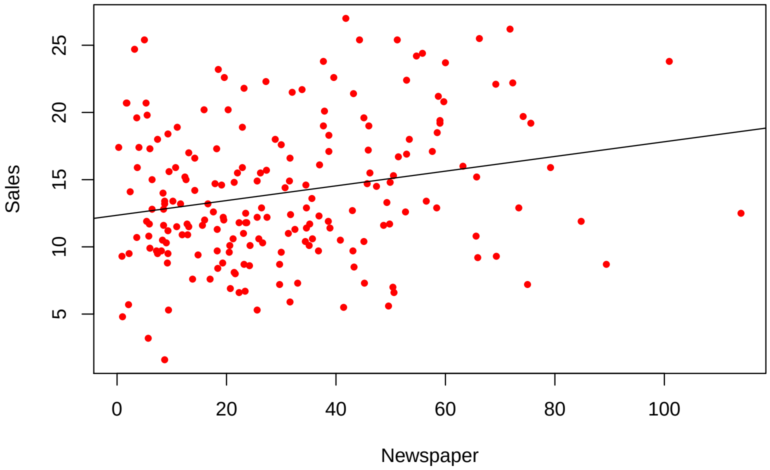


Example: Radio Impact on Sales

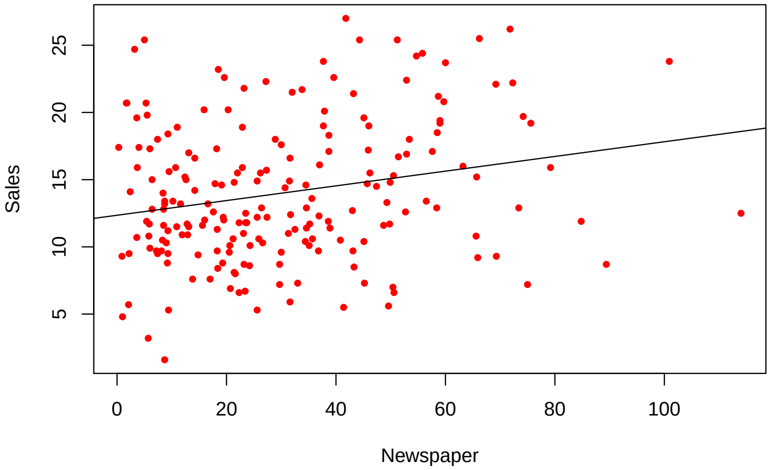


$$R^2 = 0.33$$

Example: Newspaper Impact on Sales



Example: Newspaper Impact on Sales



$$R^2 = 0.05$$

Correlation Coefficient

- Measures dependence between two random variables X and Y

$$r = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

- Correlation coefficient r is between $[-1, 1]$

  0: Variables are not related

  1: Variables are perfectly related (same)

  -1: Variables are negatively related (different)

Correlation Coefficient

- Measures dependence between two random variables X and Y

$$r = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

- Correlation coefficient r is between $[-1, 1]$

  0: Variables are not related

  1: Variables are perfectly related (same)

  -1: Variables are negatively related (different)

- $R^2 = r^2$

Hypothesis Testing

- Null hypothesis H_0:

$$\text{There is no relationship between } X \text{ and } Y$$

$$\beta_1 = 0$$

- Alternative hypothesis H_1:

$$\text{There is some relationship between } X \text{ and } Y$$

$$\beta_1 \neq 0$$

- Seek to reject hypothesis H_0 with small "probability" (p-value) of making a mistake
- Important topic, but beyond the scope of the course

Multiple Linear Regression

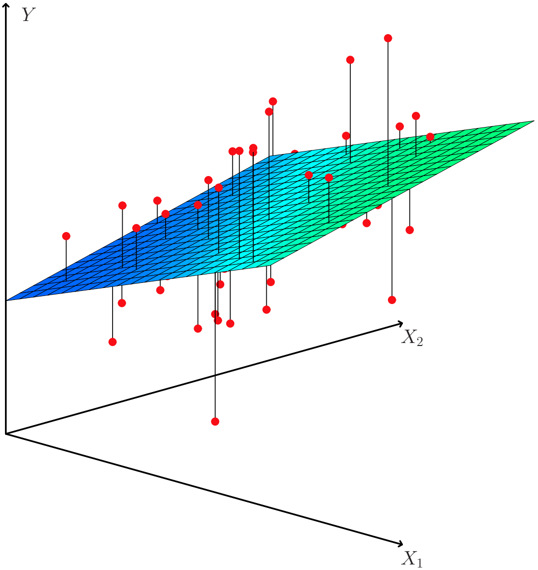
- Usually more than one feature is available

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

- In general:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

Multiple Linear Regression



Estimating Coefficients

- Prediction:

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij}$$

- Errors (y_i are true values):

$$e_i = y_i - \hat{y}_i$$

- Residual Sum of Squares

$$\text{RSS} = e_1^2 + e_2^2 + e_3^2 + \cdots + e_n^2 = \sum_{i=1}^{n} e_i^2$$

- How to minimize RSS? Linear algebra!

Inference from Linear Regression

1. Are predictors $X_1, X_2, \ldots, X_p$ really predicting Y?
2. Is only a subset of predictors useful?
3. How well does linear model fit data?
4. What Y should be predict and how accurate is it?

Inference 1

"Are predictors $X_1, X_2, \ldots, X_p$ really predicting Y?"

- Null hypothesis H_0:

  There is no relationship between X and Y

  $$\beta_1 = 0$$

- Alternative hypothesis H_1:

  There is some relationship between X and Y

  $$\beta_1 \neq 0$$

- Seek to reject hypothesis H_0 with small "probability" (p-value) of making a mistake
- See ISL 3.2.2 on how to compute F-statistic and reject H_0

Inference 2

“Is only a subset of predictors useful?”

- Compare prediction accuracy with only a subset of features

Inference 2

"Is only a subset of predictors useful?"

- Compare prediction accuracy with only a subset of features
- **RSS always decreases with more features!**

Inference 2

"Is only a subset of predictors useful?"

- Compare prediction accuracy with only a subset of features
- **RSS always decreases with more features!**
- Other measures control for number of variables:
  1. Mallows C_p
  2. Akaike information criterion
  3. Bayesian information criterion
  4. Adjusted R^2

Inference 2

"Is only a subset of predictors useful?"

- Compare prediction accuracy with only a subset of features
- **RSS always decreases with more features!**
- Other measures control for number of variables:
  1. Mallows C_p
  2. Akaike information criterion
  3. Bayesian information criterion
  4. Adjusted R^2
- Testing all subsets of features is impractical: 2^p options!

Inference 2

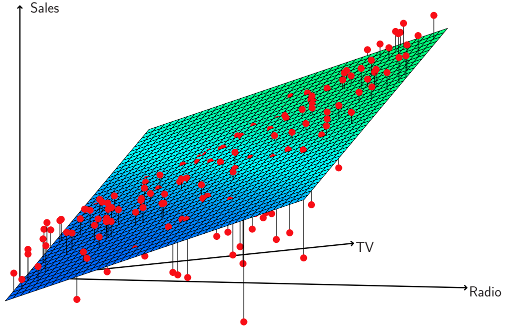
"Is only a subset of predictors useful?"

- Compare prediction accuracy with only a subset of features
- **RSS always decreases with more features!**
- Other measures control for number of variables:
  1. Mallows C_p
  2. Akaike information criterion
  3. Bayesian information criterion
  4. Adjusted R^2
- Testing all subsets of features is impractical: 2^p options!
- More on how to do this later

Inference 3

"How well does linear model fit data?"

- R^2 also always increases with more features (like RSS)
- Is the model linear? Plot it:



- More on this later

Inference 4

"What Y should be predict and how accurate is it?"

- The linear model is used to make predictions:

$$y_{\text{predicted}} = \hat{\beta}_0 + \hat{\beta}_1\, x_{\text{new}}$$

- Can also predict a confidence interval (based on estimate on ϵ):

Inference 4

"What Y should be predict and how accurate is it?"

- The linear model is used to make predictions:

$$y_{\text{predicted}} = \hat{\beta}_0 + \hat{\beta}_1 \, x_{\text{new}}$$

- Can also predict a confidence interval (based on estimate on ϵ):
- **Example**: Spent $\$100\,000$ on TV and $\$20\,000$ on Radio advertising
  - **Confidence interval**: predict $f(X)$ (the average response):

$$f(x) \in [10.985, 11, 528]$$

  - **Prediction interval**: predict $f(X) + \epsilon$ (response + possible noise)

$$f(x) \in [7.930, 14.580]$$

R notebook