

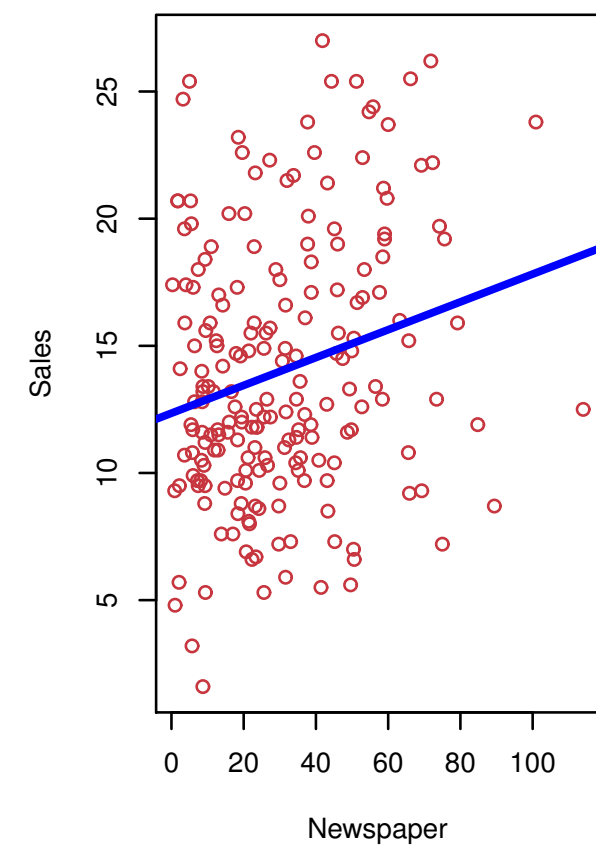
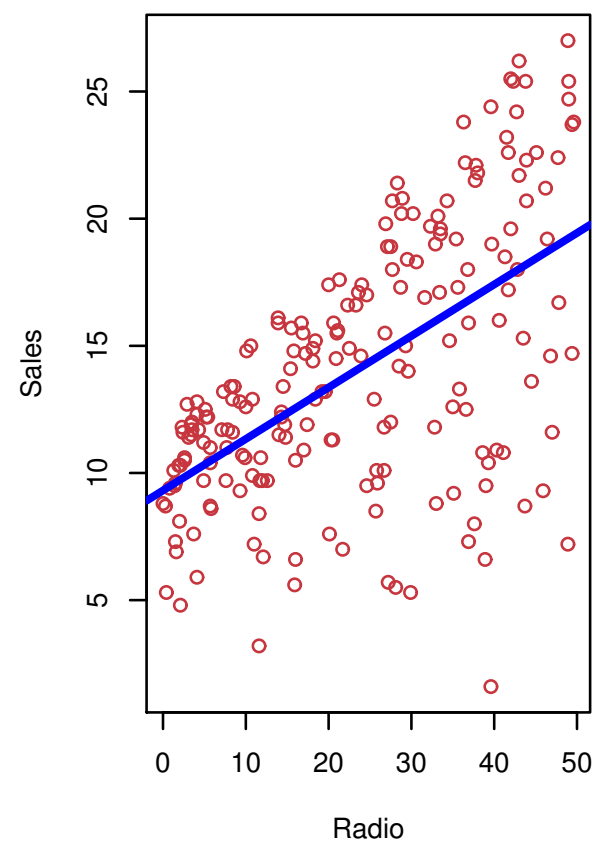
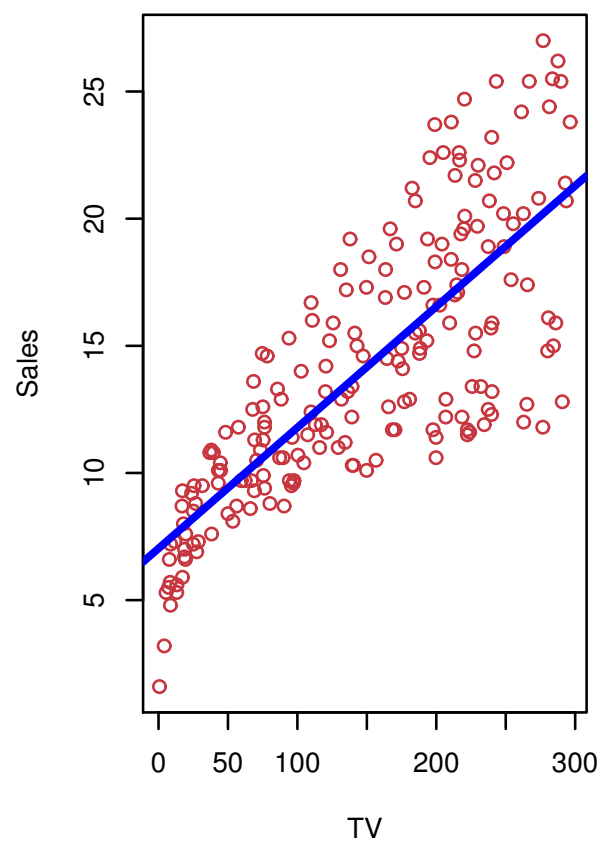
# Linear Regression

Santi Seguí | 2019-2020

# Linear Regression

- Linear regression is a **simple** approach to supervised learning.
- It **assumes linear dependence** of  $Y$  on  $X_1, X_2, \dots, X_p$ .
- True linear regression functions are "**never**" linear!
- Although it may seem overly simplistic, linear regression is **extremely useful** both conceptually and practically.

# Linear Regression



# Linear regression for advertising data

- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media contribute to sales?
- How accurately can we predict future sales?
- Is the relationship linear?
- Is there synergy among the advertising media?

# Linear Regression

- Widely used due to its simplicity but also because its interpretability.
- However, when features are correlations predictions can cause problems:
  - The variance of all coefficients tends to increase dramatically
  - Interpretations becomes hazardous- When  $x_j$  changes, everything else changes.
- Causes of **causality** must be **avoided** for observational data.

# Simple Linear regression using a single predictor $X$

- We assume a model:

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

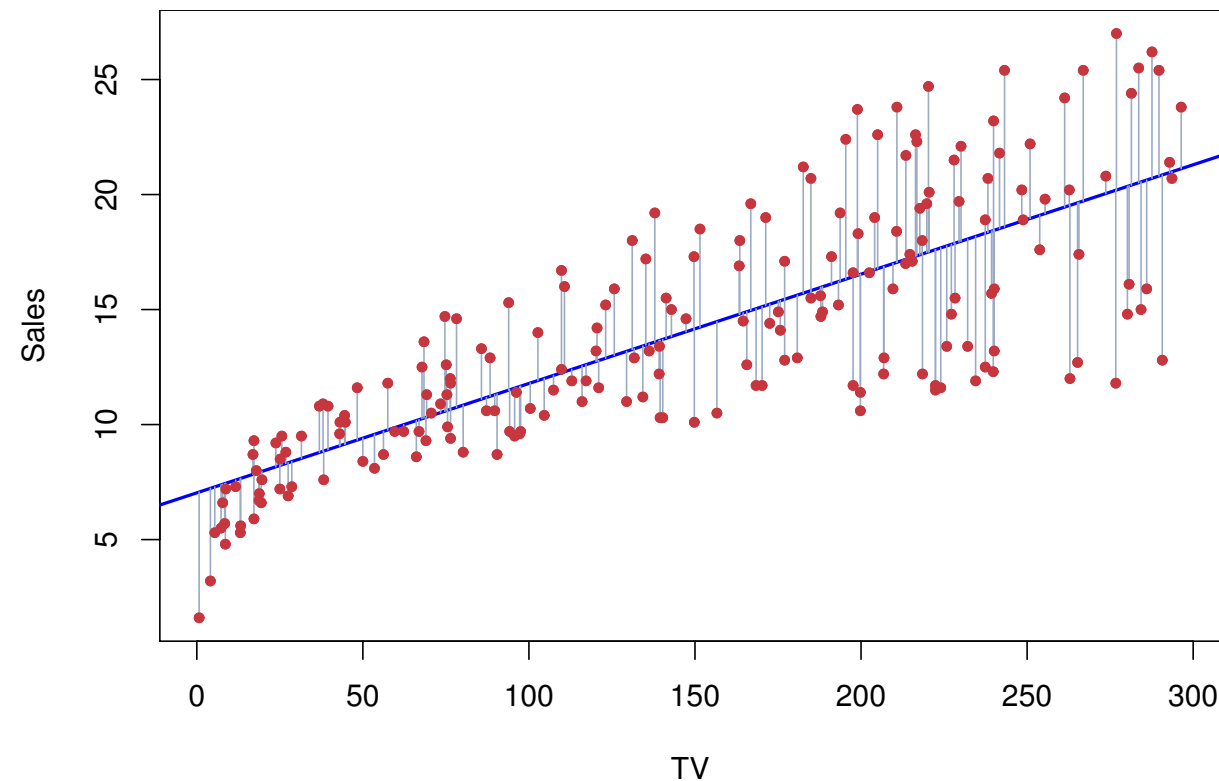
where  $\beta_0$  and  $\beta_1$  are two unknown constants that represent the **intercept** and **slop**, also known as **parameters** or **coefficients**, and  $\epsilon$  is the error term.

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

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

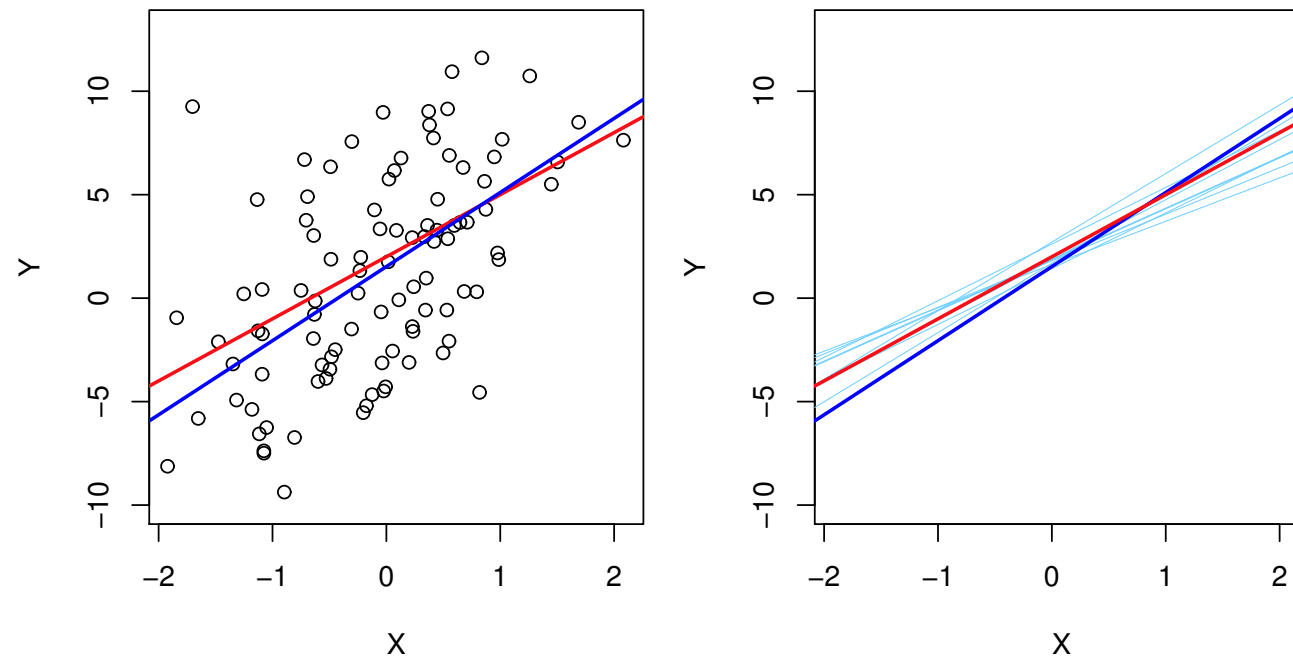
where  $\hat{y}$  indicates a prediction of  $Y$  on the basis of  $X = x$ .

# Example: Advertising data



The least square 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.

# Example: Synthetic data



- Simulated data set. *Left*: the **red** line represents the true relationship  $f(X) = 2 + 3X$ , called population line and the **blue** line is the least squares line. *Right*: The red and blue line represents again the population line and least squares line, while light blue are ten least squares lines computed with a separate random set of observations.



# Assessing the Accuracy of the Coefficient Estimates

- The standard error of an estimator reflects how it varies under **repeated sampling**. We have

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, SE(\hat{\beta}_0)^2 = \sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

where  $\sigma^2 = \text{Var}(\epsilon)$

- 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

$$\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1)$$

# Confidence Interval

- That is, there is approximately a 95% chance that the interval

$$\left[ \hat{\beta}_1 - 2 \cdot SE(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot SE(\hat{\beta}_1) \right]$$

- will contain the true value of  $\beta_1$  (under a scenario where we got repeated samples like the present sample)
- For the advertising data, the 95% confidence interval for  $\beta_1$  is  $[0.042, 0.053]$

# Hypothesis Testing

- 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**.
- $H_1$ : There is some relationship between  $X$  and  $Y$ .

- Mathematically, this corresponds to testing

$$H_0 : \beta_1 = 0$$

$$\text{versus } H_A : \beta_1 \neq 0$$

- Some of  $\beta_1 = 0$  then the model reduces to  $Y = \beta_0 + \epsilon$ , and  $X$  is not associated with  $Y$ .

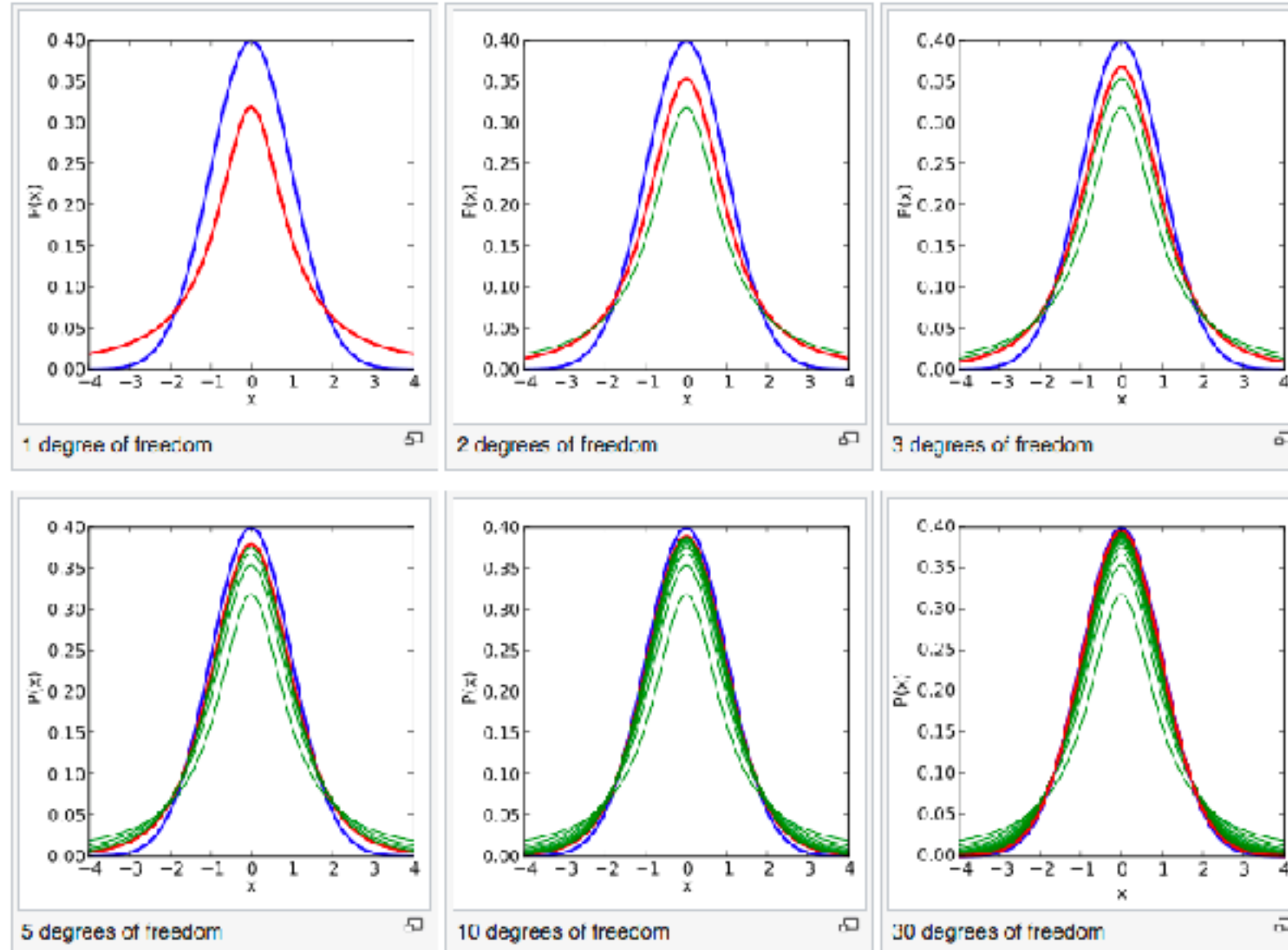
# Hypothesis Testing

- To test the null hypothesis, we compute a **t-statistic**, given by

- $$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

- 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. We call this probability the **p-value**.

# T-Distribution



# P-value

A small p-value indicates that it is **unlikely** to observe such a substantial association between the predictor and response.

If the null hypothesis if the p-value is small enough. Typically p-value cutoffs for rejecting the null hypothesis are 5 or 1 %.

# Linear Regression

- Parameter Estimation:
- Given estimates  $\hat{B}_1, \hat{B}_2, \dots, \hat{B}_p$ , we can make predictions using the formula:

$$\hat{y} = \hat{B}_0 + \hat{B}_1 x_1 + \dots + \hat{B}_p x_p$$

- We can estimate  $B_0, B_1, \dots, B_p$  as the values that minimize the sum of the squared residuals:

$$RSS = \sum_{i=0}^n (y_i - \hat{y}_i)^2$$

- This can be done using standard statistical software.

# 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

$$\widehat{sales} = 7.0325 + 0.0475 * TV$$



# Assessing the Overall Accuracy of the Model

- We compute the **Residual Standard Error**

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

- where the **residual sum-of-squares** is  $RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ .
- RSE provides an absolute measure of lack of the fit of the model to the data. Since, it is measured in the units of  $Y$ , it is not always clear what constitutes a good RSE.

# Assessing the Overall Accuracy of the Model

- **R-squared** or fraction of variance explained is

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

where  $TSS = \sum_{i=1}^n (y_i - \bar{y})^2$  is the **total sum of squares** and measures the total variance in the response.

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

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

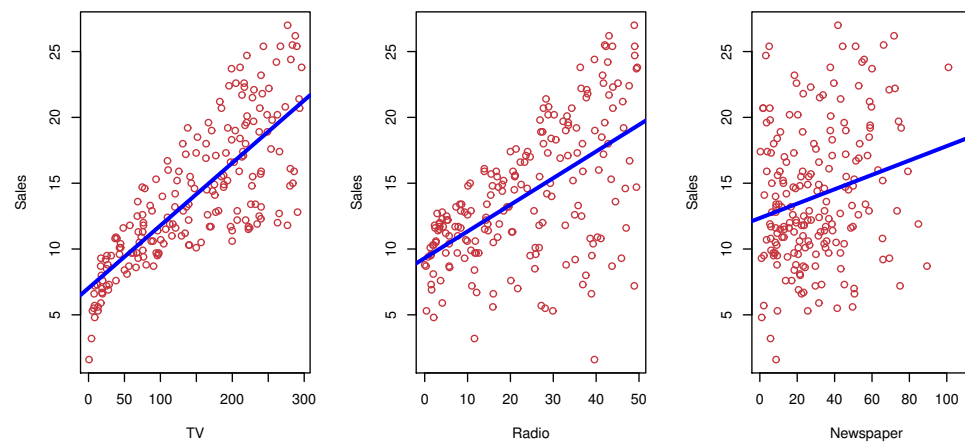
# Advertising data results

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

# Multiple Linear Regression

- The model:

$$Y = \beta_0 + \beta_1 X_1 + \dots \beta_p X_p + \epsilon$$



- We interpret  $\beta_j$  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:

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

# Interpreting regression coefficients

- The ideal scenario is when the predictors are uncorrelated
  - 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.**

# Estimation and Prediction for Multiple Regression

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

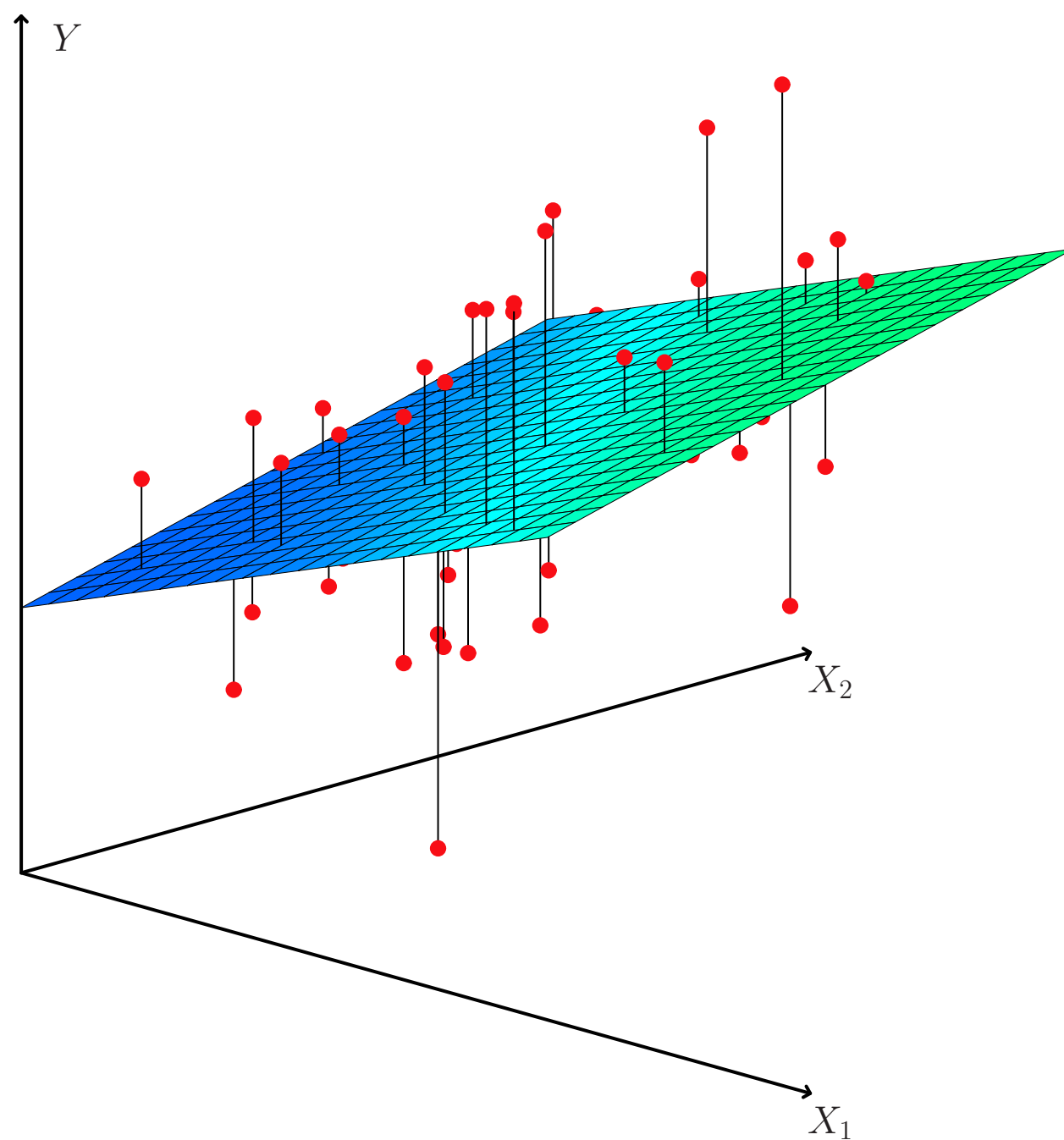
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p$$

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

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$RSS = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_p x_{ip})^2$$

- This is done using standard statistical software. The values  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$  that minimize  $RSS$  are the multiple least squares regression coefficient estimates.



# Linear Regression

- 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

	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



# Algunes preguntes interessants

Q1) Is at least one of the predictors  $X_1, X_2, \dots, X_p$  useful in predicting the response?

Q2) Do all the predictors help to explain  $Y$ , or is only a subset of the predictors useful?

Q3) How well does the model fit the data?

Q4) Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

# 1) Is at least one predictor useful?

- In the multiple regression setting with  $p$  predictors, we need to ask whether all of the regression coefficients are zero. We test the **null hypothesis**:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$$

versus the **alternative**:

- $H_a$  : at least one of  $\beta_j$  is non-zero
- This hypothesis test is performed by computing the F-statistic

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

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

# Q2: Deciding on the important variables

- The most direct approach is called all subsets or best subsets regression: we compute the least squares fit for all possible subsets and then choose between them based on some criterion that balances training error with model size.
- However we often can't examine all possible models, since there are  $2^p$  of them; for example when  $p = 40$  there are over a billion models! Instead we need an automated approach that searches through a subset of them. We discuss two commonly used approaches next.

# Q2: Forward selection

- Begin with the null model - a model that contains an intercept but no predictors.
- Fit  $p$  simple linear regressions and add to the null model the variable that results in the lowest RSS.
- Add to that model the variable that results in the lowest RSS amongst all two-variable models.
- Continue until some stopping rule is satisfied, for example when all remaining variables have a p-value above some threshold.

# Q2: Backward selection

- Start with all variables in the model.
- Remove the variable with the largest p-value - that is, the variable that is the least statistically significant.
- The new  $(p - 1)$ -variable model is fit, and the variable with the largest p-value is removed.
- Continue until a stopping rule is reached. For instance, we may stop when all remaining variables have a significant p-value defined by some significance threshold.
- **Intead of using p-value, other measures or methodologies can be used**

# Q3: How well the model fits the 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

Quantity	Value
Residual Standard Error	1.69
$R^2$	0.8972
F-statistic	570

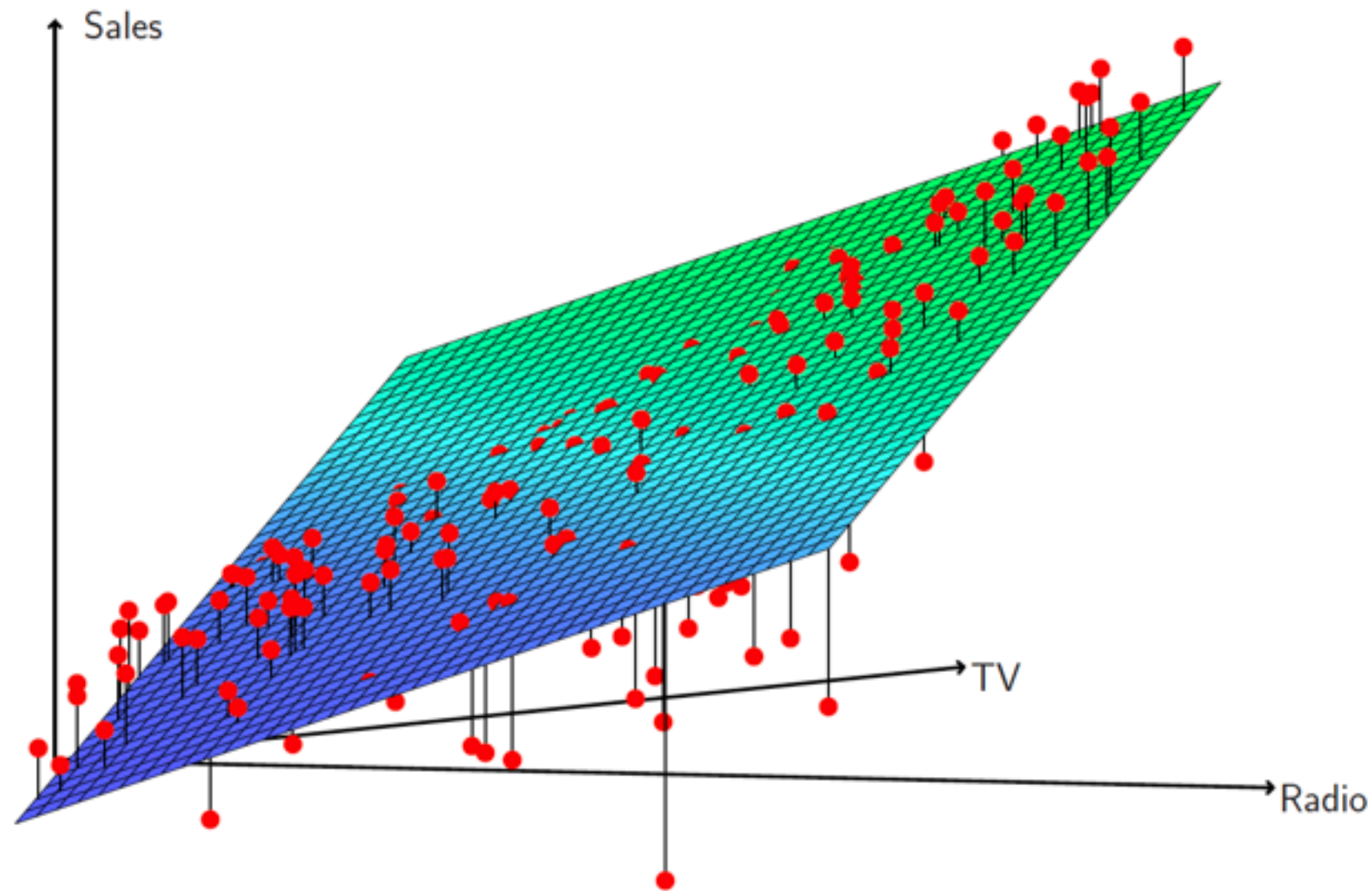
# Q3: How well the model fits the data

with newspaper:

Quantity	Value
Residual Standard Error	1.686
$R^2$	0.8972

without newspaper:

Quantity	Value
Residual Standard Error	1.681
$R^2$	0.89719



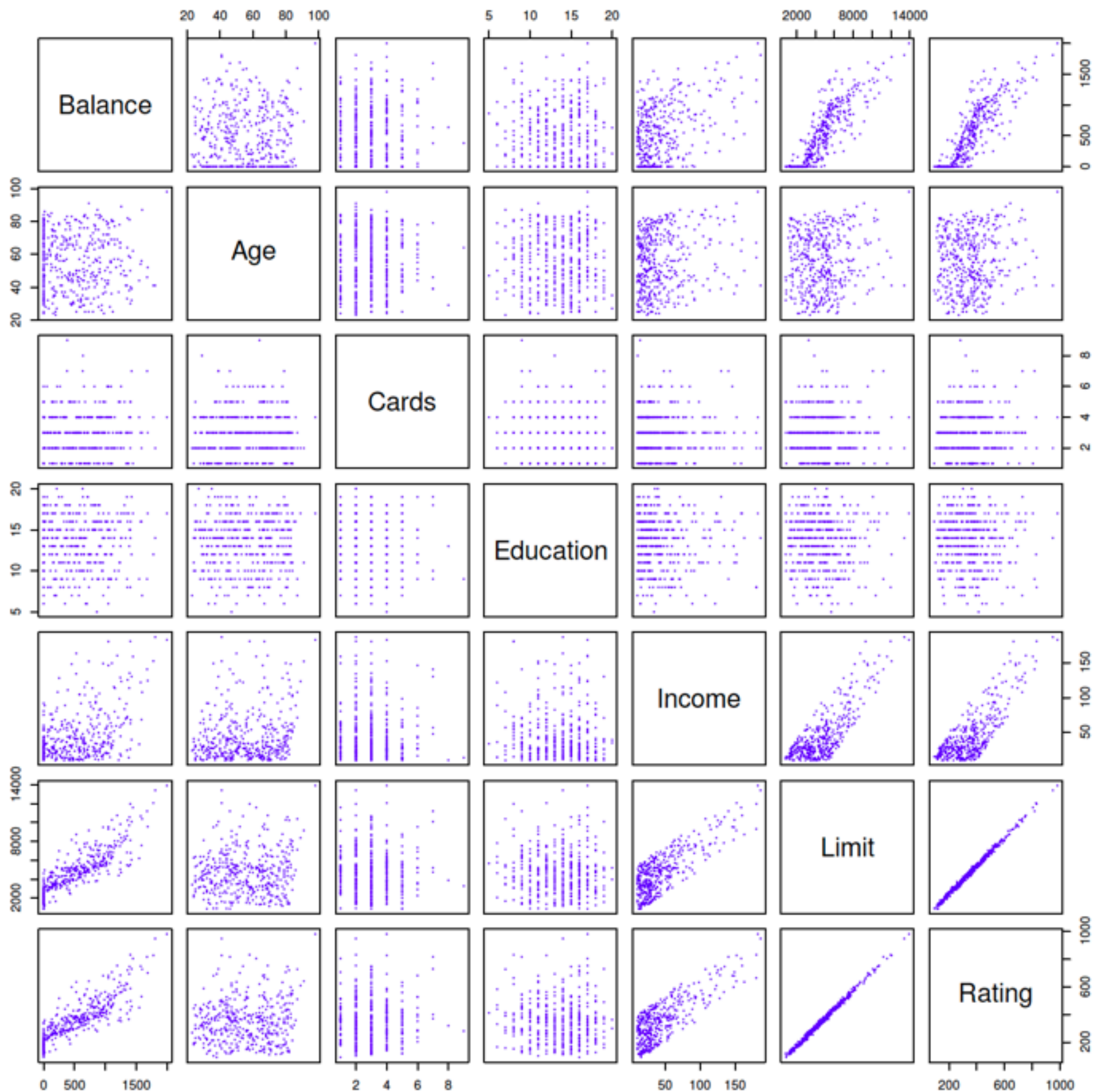
The model seems to overestimate estimate **sales** for instances in which most of the advertising money was spend exclusively on either **TV** OR **radio**. It underestimates **sales** for instances where the budget was split between to medias



**Categorical data?**

# Categorical data

- Some times our data is not quantitative but qualitative. For example: sex, ethnicity, city,...
- 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: (student status), **status** (marital status), and **ethnicity** (Caucasian, African American (AA) or Asian).



# Qualitative Predictors

- Example: investigate differences in credit card balance between males and females, ignoring the other variables. 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:

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

# Credit card data

Results for gender model:

	Coefficient	Std. Error	t-statistic	p-value
Intercept	509.80	33.13	15.389	<0.0001
gender[Female]	19.73	46.08	0.429	0.6690

# Qualitative predictors with more than two levels

- What happens if there is more than two levels?
- For example, for the **ethnicity**?

# 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:

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

# Qualitative predictors with more than two levers

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

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

- There will always be one fewer dummy variable than the number of levels. The level with no dummy variable - African American in this example - is known as the baseline.



# Results for ethnicity

	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

# Extensions of the linear model

- Removing the additive assumption: **interactions** and nonlinearity  
**interactions:**
  - In the previous analysis on the Advertising data, it is 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

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

states that the average effect on sales of a one-unit increase in TV is always  $\beta_1$ , regardless of the amount spent on radio

# Interactions

- 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.

# Modeling interactions - Advertising data

- Model takes the form:

$$sales = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times (radio \times TV) + e$$

	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 x radio	0.0011	0.0000	20.73	<0.0001

# Interpretation

- The results in this table suggests that interactions are important.
- The p-value for the interaction term  $TV \times radio$  is extremely low, indicating that there is strong evidence for  $H_A : \beta_3 \neq 0$ .
- 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.

# Interactions

- Sometimes it is the case that an interaction term has a very small p-value, but the associated main effects (in this case, TV and radio) do not.
- The hierarchy principle: ***If we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.***

# 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:

$$\begin{aligned} \text{balance}_i &\approx \beta_0 + \beta_1 \times \text{income}_i + \begin{cases} \beta_2 & \text{if } \text{student} \\ 0 & \text{if } \text{notstudent} \end{cases} \\ &= \beta_1 \times \text{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{if } \text{student} \\ \beta_0 & \text{if } \text{notstudent} \end{cases} \end{aligned}$$

# 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:

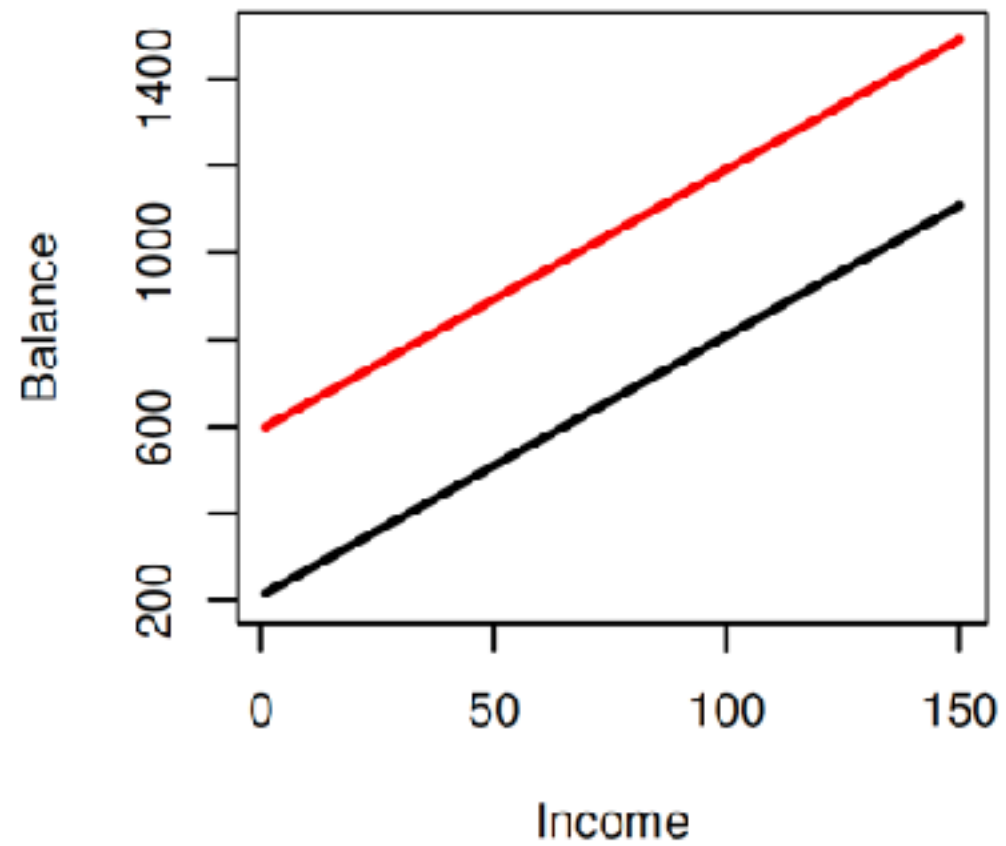
$$\begin{aligned} \text{balance}_i &\approx \beta_0 + \beta_1 \times \text{income}_i + \begin{cases} \beta_2 & \text{if student} \\ 0 & \text{if notstudent} \end{cases} \\ &= \beta_1 \times \text{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{if student} \\ \beta_0 & \text{if notstudent} \end{cases} \end{aligned}$$

- With interactions, it takes the form:

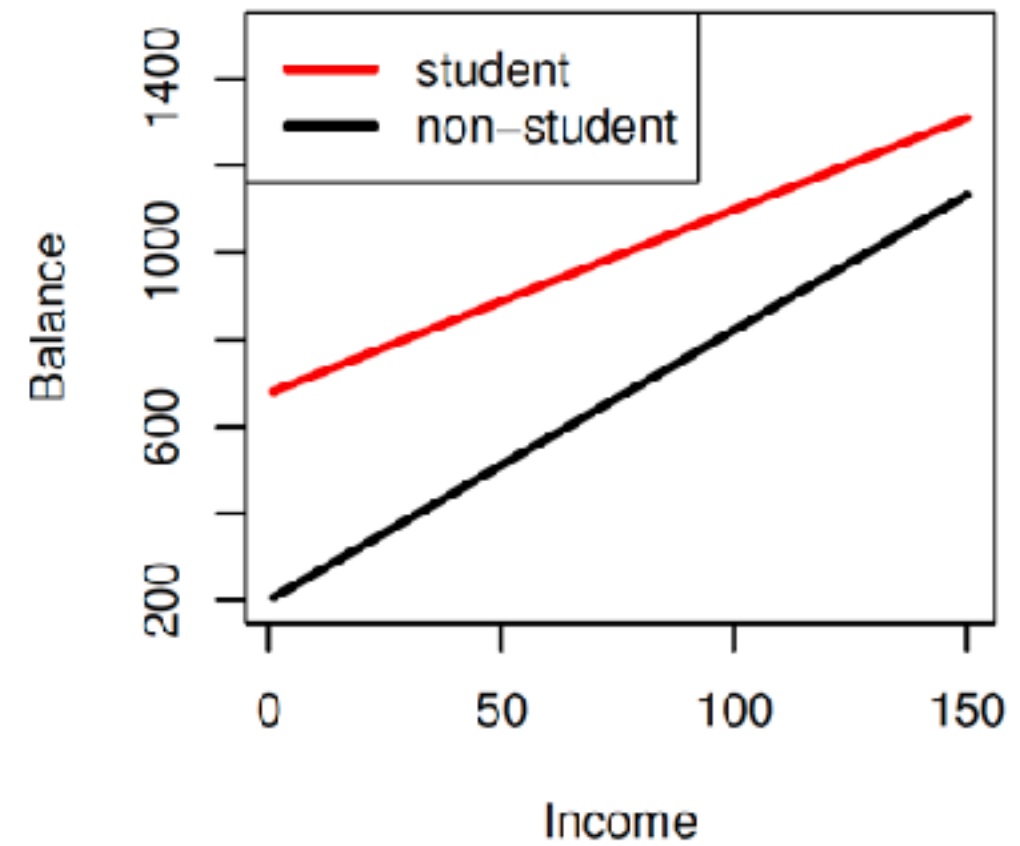
$$\begin{aligned} \text{balance}_i &\approx \beta_0 + \beta_1 \times \text{income}_i + \begin{cases} \beta_2 + \beta_3 \times \text{income}_i & \text{if student} \\ 0 & \text{if notstudent} \end{cases} \\ &= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \text{income}_i & \text{if student} \\ \beta_0 + \beta_1 \times \text{income}_i & \text{if notstudent} \end{cases} \end{aligned}$$



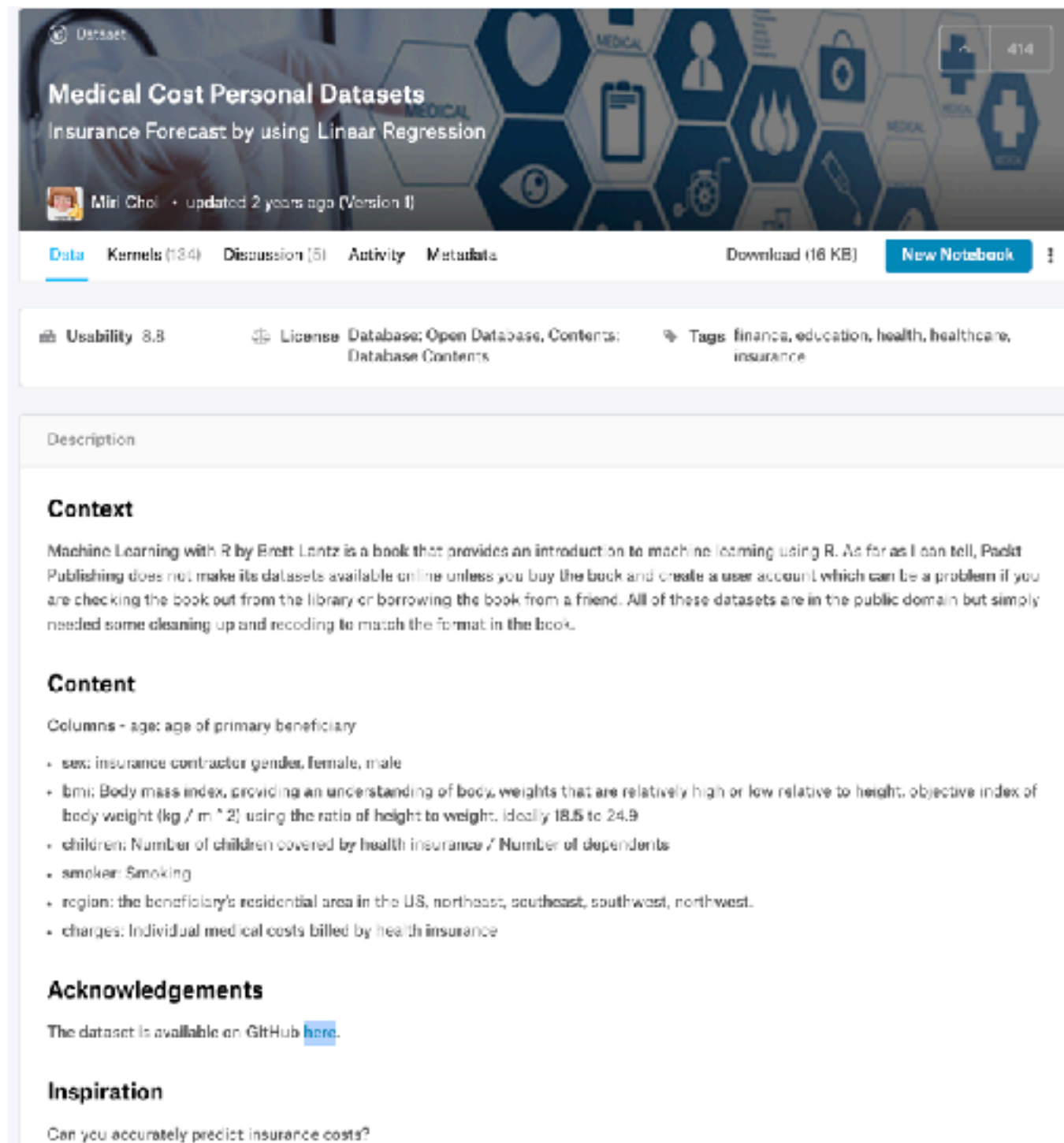
**Without interactions**



**With interactions**



# Example: Patient cost estimation



The screenshot shows a Kaggle dataset page for 'Medical Cost Personal Datasets'. The header includes the dataset title, subtitle 'Insurance Forecast by using Linear Regression', and the creator 'Miri Choi' with a note 'updated 2 years ago (Version 1)'. Below the header is a navigation bar with links for 'Data', 'Kernels (134)', 'Discussion (5)', 'Activity', and 'Metadata'. A 'Download (16 KB)' button and a 'New Notebook' button are also present. The 'Usability' is 8.8, and the license is 'Database: Open Database, Contents: Database Contents'. Tags include 'finance, education, health, healthcare, insurance'. The 'Description' section contains three sub-sections: 'Context', 'Content', and 'Acknowledgements'. The 'Context' section mentions 'Machine Learning with R' by Ernst Lantz. The 'Content' section lists columns: age, sex, bmi, children, smoker, region, and charges. The 'Acknowledgements' section states the dataset is available on GitHub. The 'Inspiration' section asks 'Can you accurately predict insurance costs?'.

**Medical Cost Personal Datasets**  
Insurance Forecast by using Linear Regression

Miri Choi • updated 2 years ago (Version 1)

[Data](#) [Kernels \(134\)](#) [Discussion \(5\)](#) [Activity](#) [Metadata](#) [Download \(16 KB\)](#) [New Notebook](#)

**Usability** 8.8 **License** Database: Open Database, Contents: Database Contents **Tags** finance, education, health, healthcare, insurance

**Description**

**Context**

Machine Learning with R by Ernst Lantz is a book that provides an introduction to machine learning using R. As far as I can tell, Packt Publishing does not make its datasets available online unless you buy the book and create a user account which can be a problem if you are checking the book out from the library or borrowing the book from a friend. All of these datasets are in the public domain but simply needed some cleaning up and recoding to match the format in the book.

**Content**

Columns - age: age of primary beneficiary

- sex: insurance contractor gender, female, male
- bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height; objective index of body weight ( $\text{kg} / \text{m}^2$ ) using the ratio of height to weight; ideally 18.5 to 24.9
- children: Number of children covered by health insurance / Number of dependents
- smoker: Smoking
- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance

**Acknowledgements**

The dataset is available on GitHub [here](#).

**Inspiration**

Can you accurately predict insurance costs?

# Example: Patient cost estimation

**Task: Can you accurately predict insurance costs?**

**Experience:**

**1338 examples with the following features:**

**sex:** insurance contractor gender, female, male

**bmi:** Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight ( $\text{kg} / \text{m}^2$ ) using the ratio of height to weight, ideally 18.5 to 24.9

**children:** Number of children covered by health insurance / Number of dependents

**smoker:** Smoking

**region:** the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.

**charges:** Individual medical costs billed by health insurance

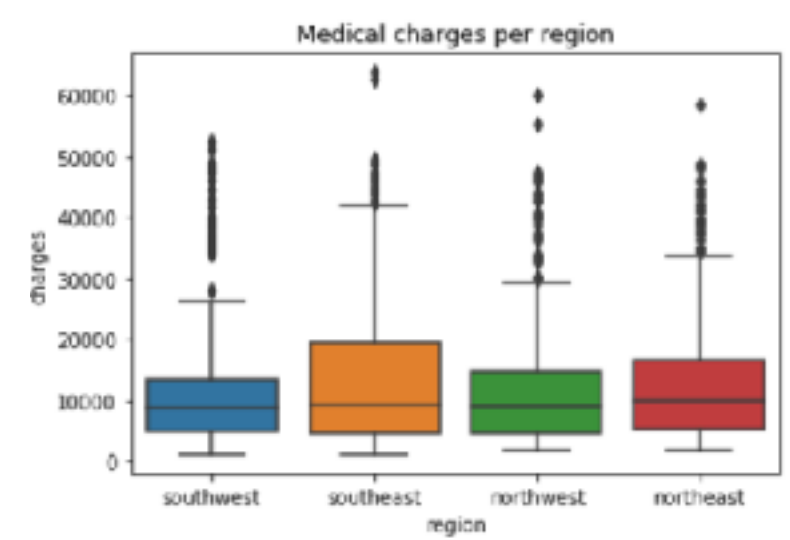
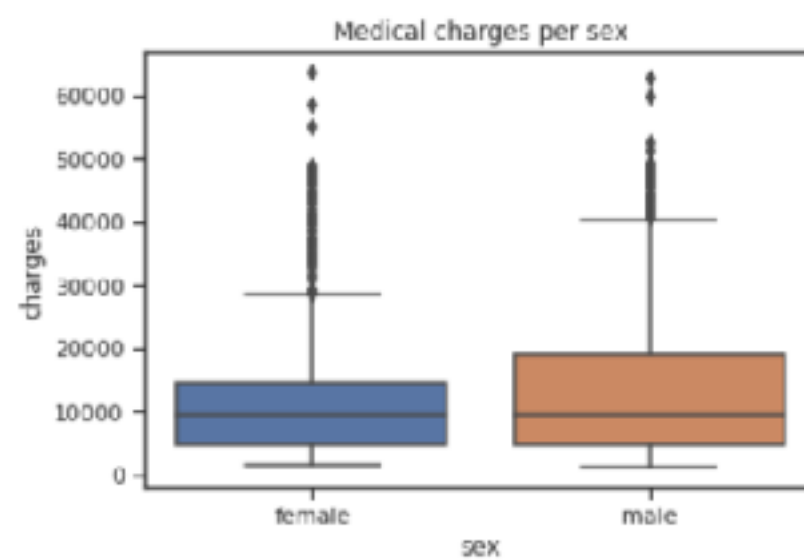
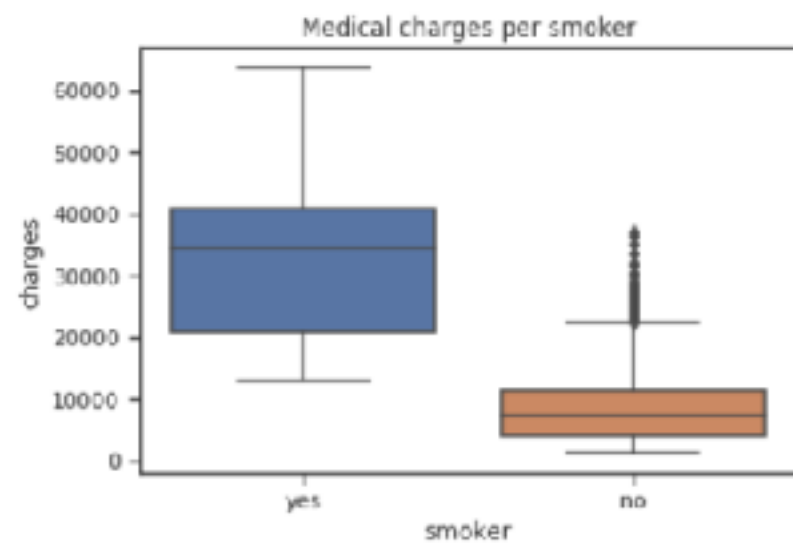
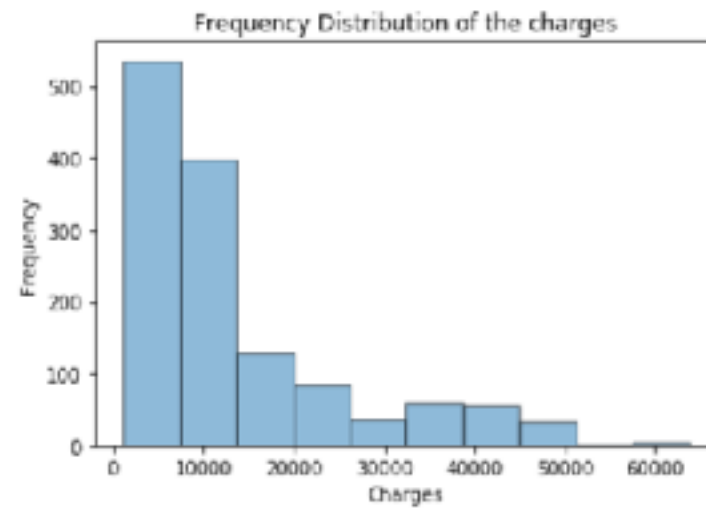
# Example: Patient cost estimation

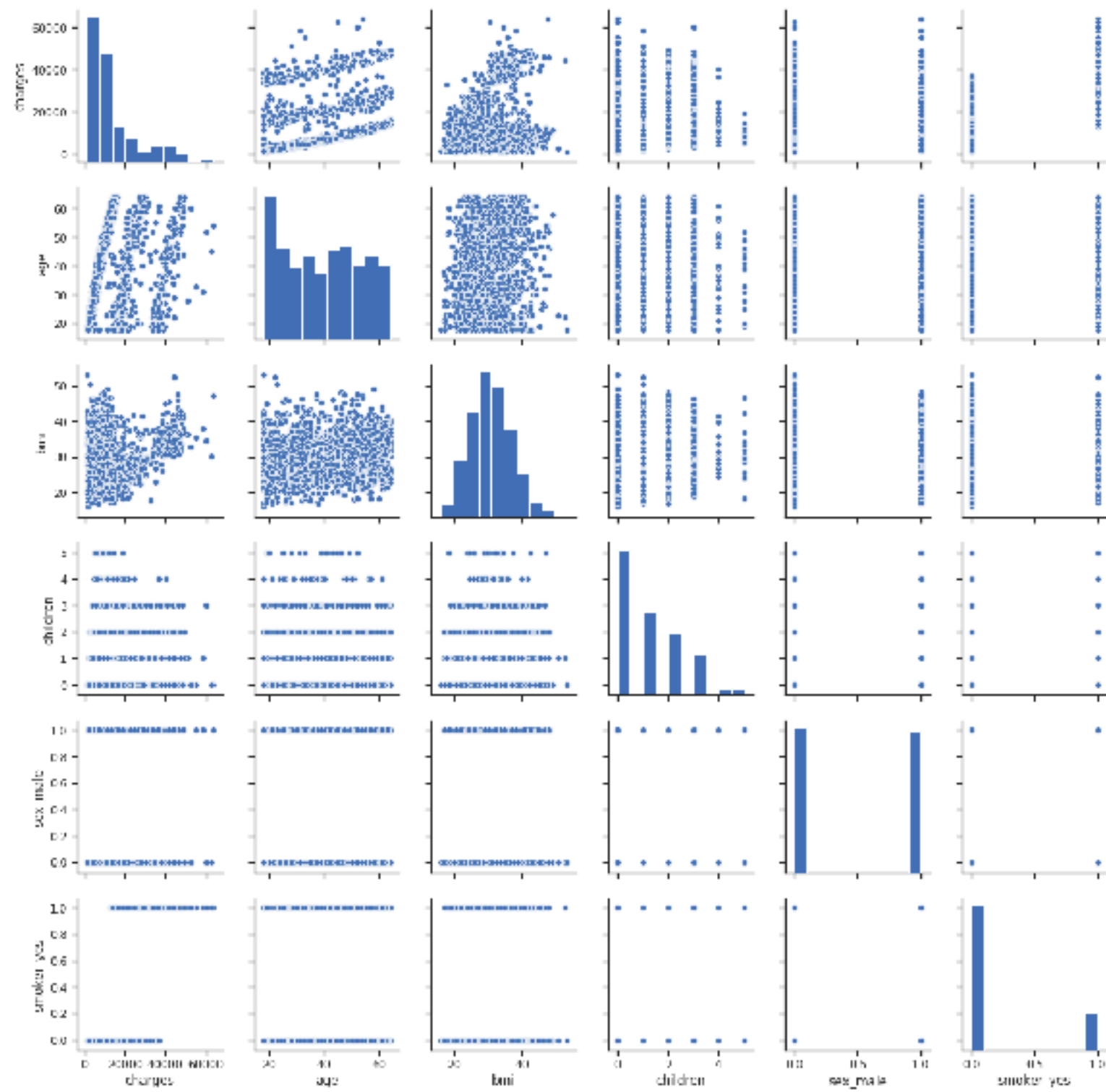
**Is data preparation needed here?  
What should we consider?**

We have 3 categorical features: "sex", "smoker", "region"

Is there any missing value?

Which is the feature distribution?





# Example: Patient cost estimation

	charges	age	bmi	children	sex_0	smoker_0
<b>count</b>	1338.000000	1338.000000	1338.000000	1338.000000	1338.000000	1338.000000
<b>mean</b>	13270.422265	39.207025	30.663397	1.094918	0.494768	0.795217
<b>std</b>	12110.011237	14.049960	6.098187	1.205493	0.500160	0.403694
<b>min</b>	1121.873900	18.000000	15.960000	0.000000	0.000000	0.000000
<b>25%</b>	4740.287150	27.000000	26.296250	0.000000	0.000000	1.000000
<b>50%</b>	9382.033000	39.000000	30.400000	1.000000	0.000000	1.000000
<b>75%</b>	16639.912515	51.000000	34.693750	2.000000	1.000000	1.000000
<b>max</b>	63770.428010	64.000000	53.130000	5.000000	1.000000	1.000000

There is no missing values! :)

# Example: Patient cost estimation

	age	bmi	children	sex_male	smoker_yes
charges	0.299008	0.198341	0.067998	-0.057292	0.787251
age		0.109272	0.042469	0.020856	-0.025019
bmi			0.012759	-0.046371	0.003750
children				-0.017163	0.007673
sex_0					-0.076185

Feature Correlations

	Coefficient
age	255.765700
bmi	325.533716
children	586.352547
sex_male	392.430723
smoker_yes	23770.069109

Linear Regression Features



# Example: Patient cost estimation

We are ready to learn a **regression MODEL**.

Let's divide the data into two sets: training(75%) and test (25%)

## Linear regression MODEL

$$\hat{y} = B_0 + B_1x_1 + \dots + B_px_p$$

	Coefficient
age	252.277070
bmi	308.930902
children	359.000639
sex_male	257.252678
smoker_yes	24069.017691

$R^2 = 0.7314$   
RMSE = 6009.79

Linear Regression Features


# Regularization methods

Regularization means making the model less complex which can allow it to generalize better (i.e. avoid overfitting) and perform better on a new data.

- Linear Regression:

$$\text{minimize}(\sum_{i=0}^n (y_i - B_0 - \sum_{j=1}^p B_j x_{ij}))$$

L2 penalty / Penalty Term /  
Regularisation Term



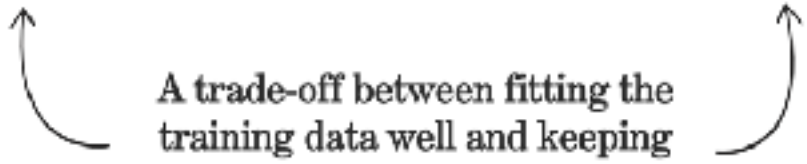
- Ridge Regression:

$$\text{minimize}(\sum_{i=0}^n (y_i - B_0 - \sum_{j=1}^p B_j x_{ij})^2 + \lambda \sum_{j=1}^p B_j^2)$$

Fit training data well (OLS)

Keep parameters small

A trade-off between fitting the  
training data well and keeping  
parameters small



# Regularization methods

- Linear Regression:

$$\text{minimize}(\sum_{i=0}^n (y_i - B_0 - \sum_{j=1}^p B_j x_{ij}))$$

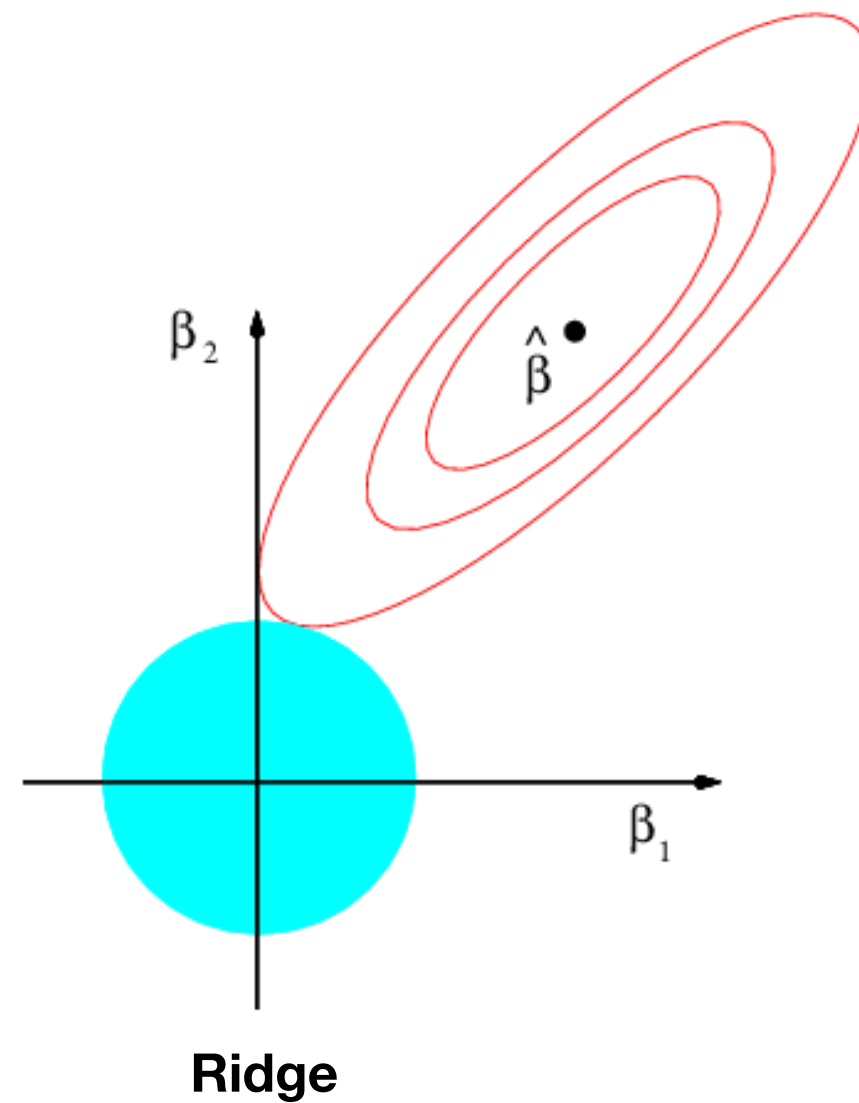
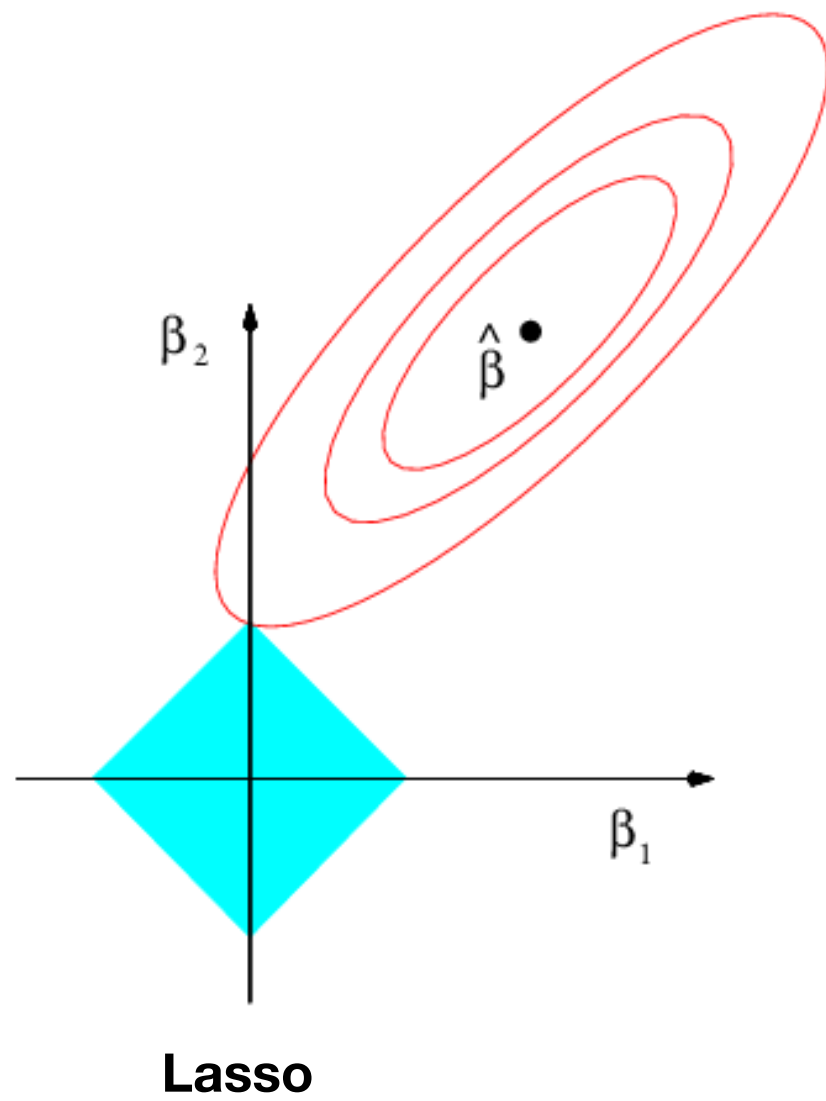
- Ridge Regression:

$$\text{minimize}(\sum_{i=0}^n (y_i - B_0 - \sum_{j=1}^p B_j x_{ij}) - \lambda \sum_{j=1}^p B_j^2)$$

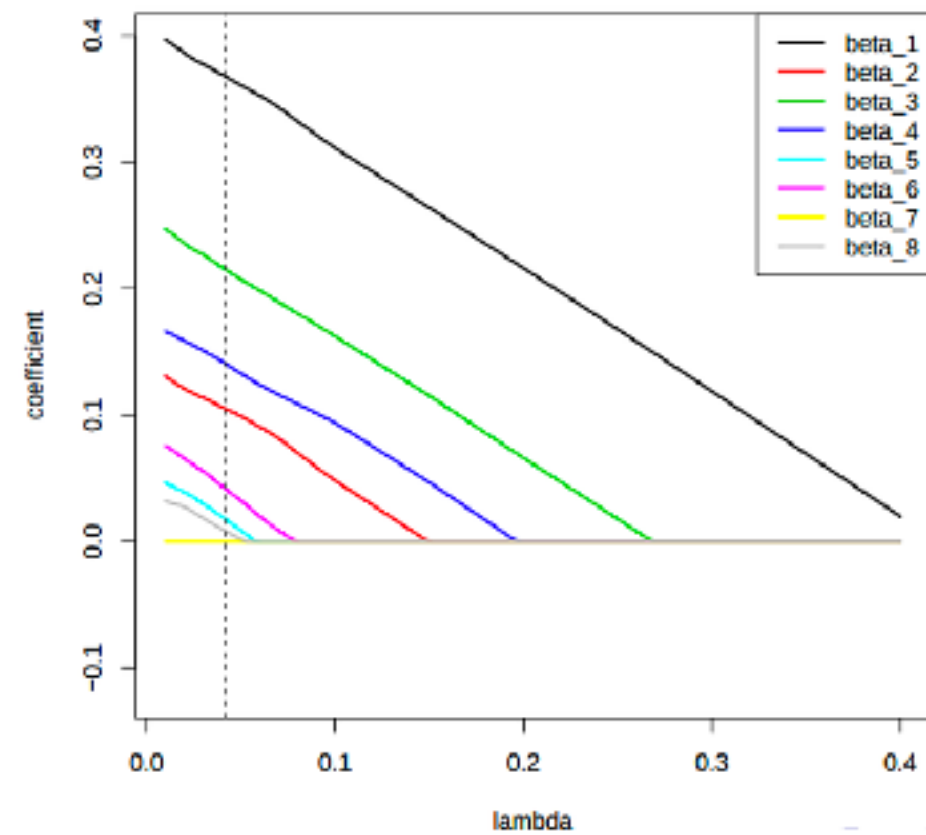
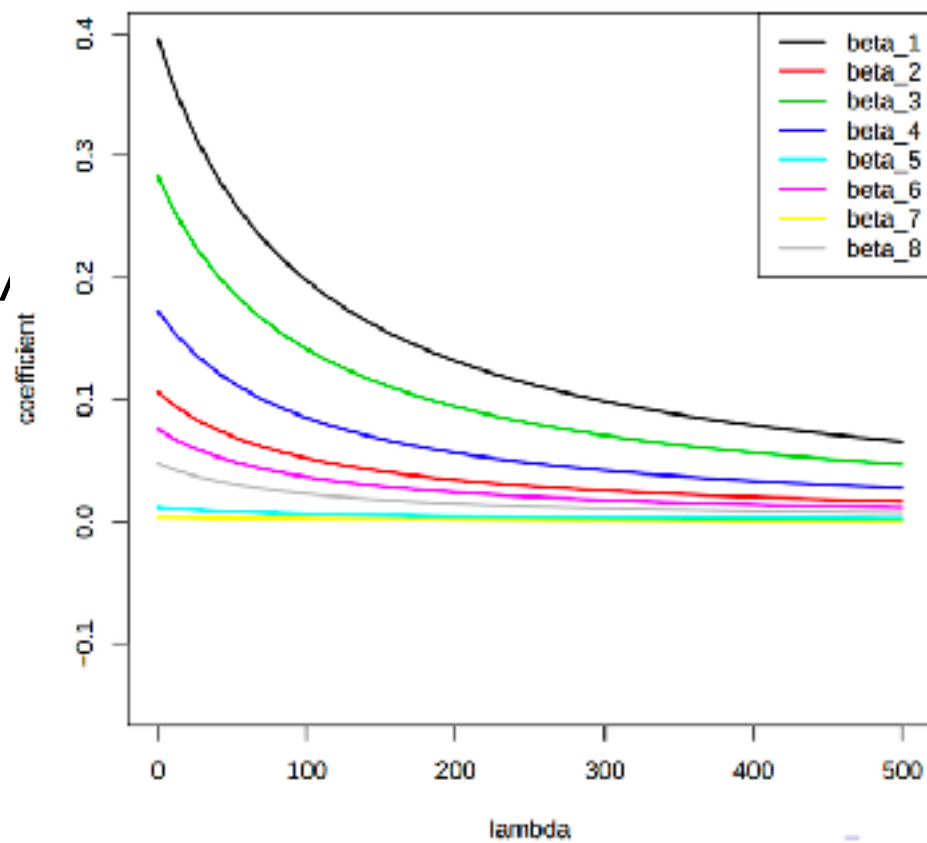
- Lasso

$$\text{minimize}(\sum_{i=0}^n (y_i - B_0 - \sum_{j=1}^p B_j x_{ij}) - \lambda \sum_{j=1}^p |B_j|)$$

# Regularization methods



# Regularization methods



# Elastic Net

- The LASSO method has some limitations:
  - In small-n-large-p dataset (high-dimensional data with few examples), the LASSO selects at most  $n$  variables before it saturates.
  - If there is a group of highly correlated variables, LASSO tends to select one variable from a group and ignore the others.
- To overcome these limitations, the elastic net adds a quadratic part to the L1 penalty, which when used alone is a ridge regression

**Any Idea of non Linear  
Methods?**