Machine Learning & Data Mining

Linear Regression

Kyung-Ah Sohn

Ajou University

Content

- Linear regression model
- Variable selection
- Regularized linear model

LINEAR REGRESSION

Supervised Learning

Feature Space $\mathcal{X}$

Label Space $\mathcal{Y}$

"Sports"
"News"
"Science"
...

Discrete Labels
Classification

Share Price
"$ 24.50"

Continuous Labels
Regression

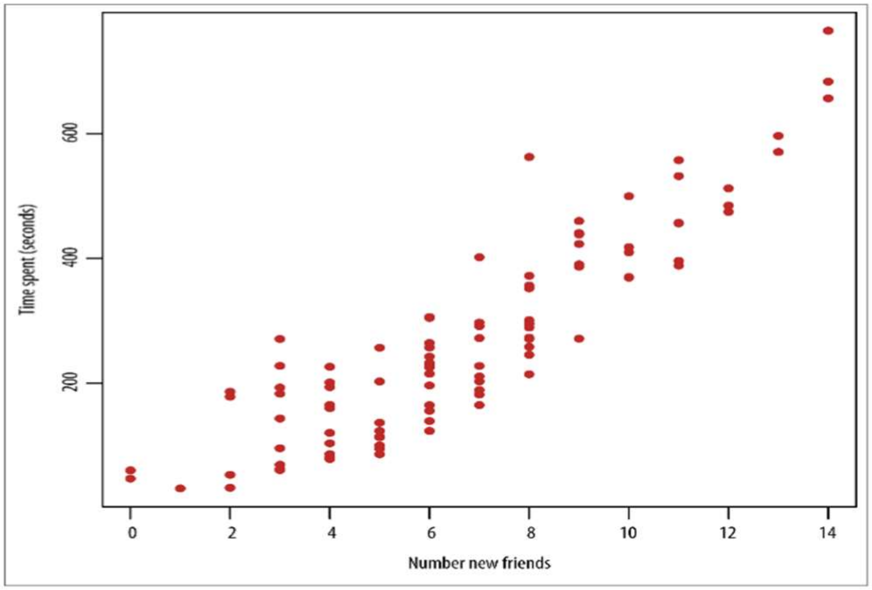
Task: Given $X \in \mathcal{X}$, predict $Y \in \mathcal{Y}$.

Regression

- Supervised learning: predict one variable Y given a set of other variables X
  - Classification: Y is categorical
  - Regression: Y is numeric
- Assume a linear or non-linear model
- Applications
  - Sales forecasting
  - Stock market prediction

Example: user behavior at social networking site

Number of new friends	Time spent (sec)
7	276
3	43
4	82
6	136
10	417
9	269
…	…



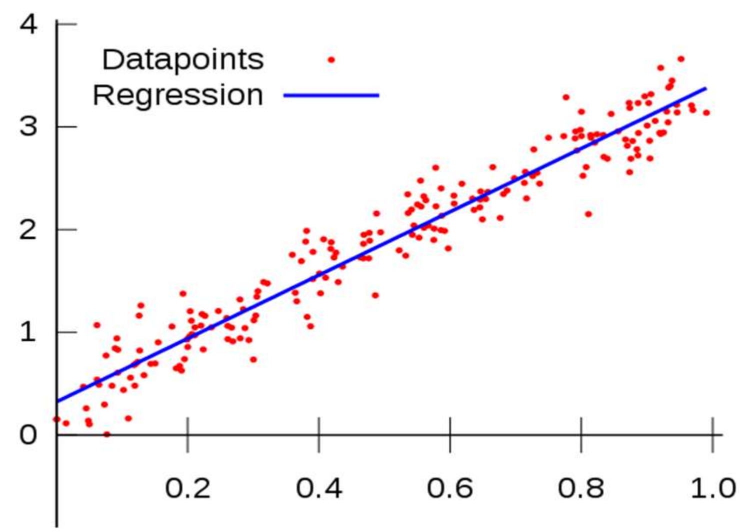
Looking kind of linear

Regression in general

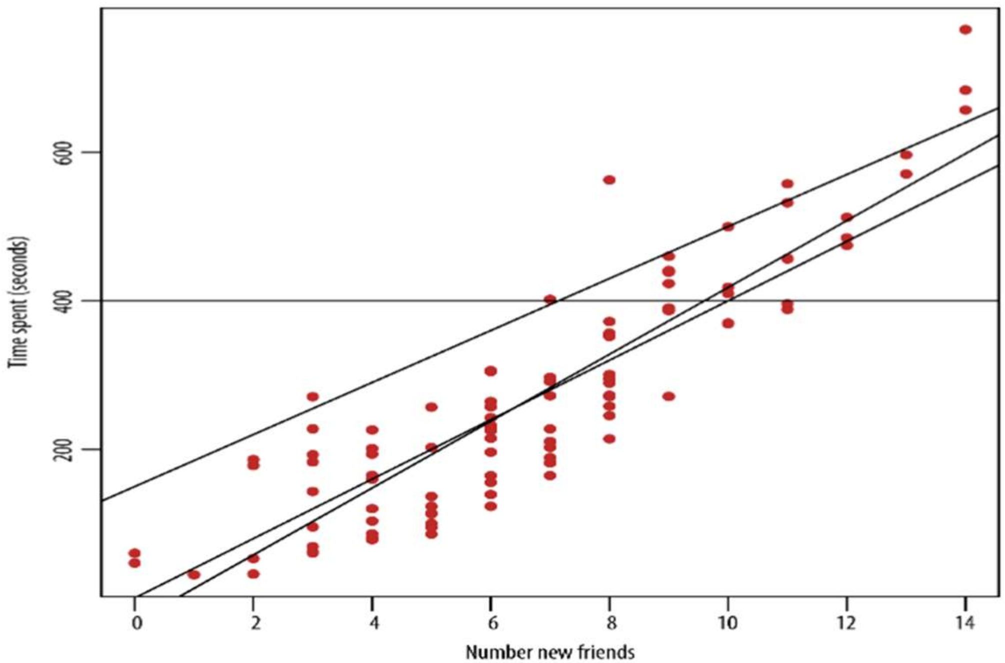
- Technique used for the modeling and analysis of numerical data
- Exploits the relationship between two or more variables so that we can gain information about one of them through knowing values of the other

Simple linear regression

- Find a linear line that best fits the data points (x_1,y_1), (x_2,y_2), …,(x_n,y_n)



Which line is the best fit?



Basic idea:

Estimate the parameters that minimize the sum of differences between the actual y value and the predicted value

Simple Linear Regression Model

- For each training data (x_i, y_i)
- Typically, it is assumed $\epsilon \in N(0, \sigma^2)$

Y-Intercept

Slope

Random Error

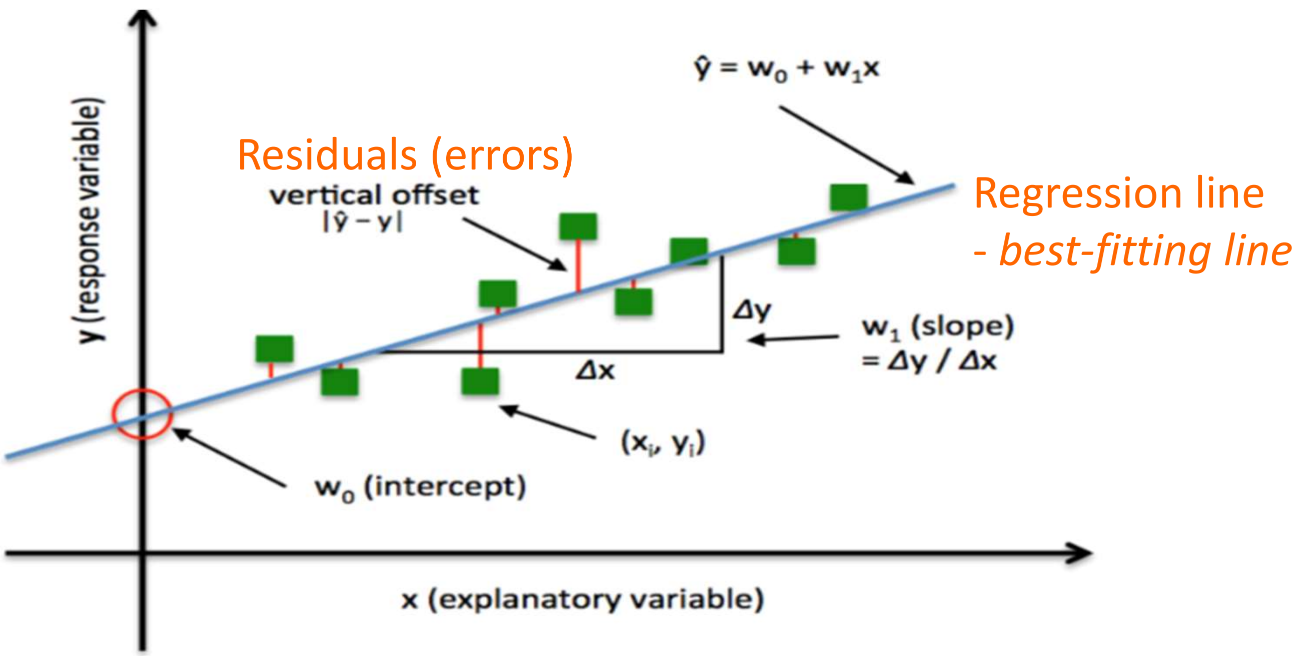
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Regression coefficients

Dependent (Response) Variable (e.g., scores)

Independent (Explanatory) Variable (e.g., studying hours)

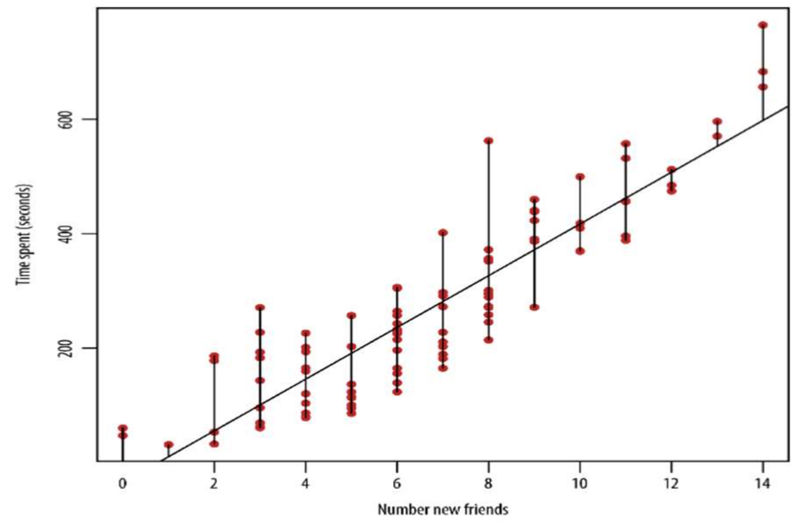
Simple linear regression



Least squares

- “Best fit” means a Difference Between Actual Y Values & Predicted Y Values is a Minimum

- Find the coefficients that minimizes Residual Sum of Squares



Estimating model parameter

- Training data: n samples

- Goal: minimize the difference between the actual and predicted target

$$\min \sum_{i=1}^{n} \epsilon_i^2 = \min \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$
$$= \min \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

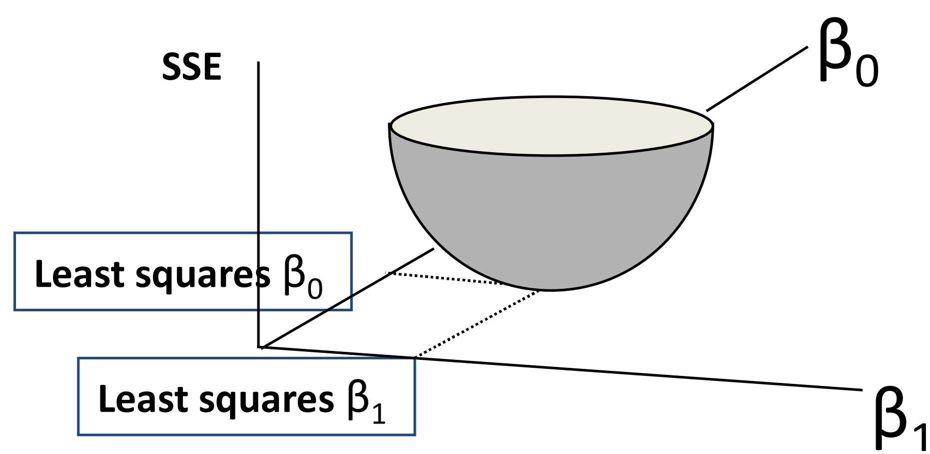
x_1	y_1
x_2	y_2
x_3	y_3
…	…
x_n	y_n

Least Squares Regression

- Find β_0, β_1 that minimize the following objective function (Sum of Squared Errors, SSE)

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

- How to get such coefficients?
  - Convex function
  - Unique minimum point



Derivation of parameters

- Minimize

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

Second order polynomial equation

- Take the derivative of *f* with respect to each coefficient:

Derivation of Parameters (1)

- Least Squares (L-S):

  Minimize squared error

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

$$0 = \frac{\partial \sum \varepsilon_i^2}{\partial \beta_0} = \frac{\partial \sum (y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_0}$$

$$= -2\left(n\overline{y} - n\beta_0 - n\beta_1\overline{x}\right)$$

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

Derivation of Parameters (1)

- Least Squares (L-S):

  Minimize squared error

$$0 = \frac{\partial \sum \varepsilon_i^2}{\partial \beta_1} = \frac{\partial \sum (y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_1}$$

$$= -2 \sum x_i (y_i - \beta_0 - \beta_1 x_i)$$

$$= -2 \sum x_i (y_i - \overline{y} + \beta_1 \overline{x} - \beta_1 x_i)$$

$$\beta_1 \sum x_i (x_i - \overline{x}) = \sum x_i (y_i - \overline{y})$$

$$\beta_1 \sum (x_i - \overline{x})(x_i - \overline{x}) = \sum (x_i - \overline{x})(y_i - \overline{y})$$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

Coefficient Equations

- Prediction equation

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

- Sample slope

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

- Sample Y - intercept

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

Multiple Linear Regression

- Fit a linear equation between a dependent variable Y and a set of predictors X=(X_1, ..., X_p)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon$$

Coefficients

Noise, unexplained

- Parameter estimation:

$$\begin{aligned} \min \sum_{i=1}^{n} \epsilon_i^2 &= \min \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \\ &= \min \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi}))^2 \end{aligned}$$

Multivariate linear regression

Hours studying (x_1)	Hours sleeping (x_2)	Exam scores (y)
4	2	60
1	5	10
10	0.5	64
14	0.1	75
4	3	50
7	1	70
22	0.1	95
3	4	27
8	1.5	49

$$\underbrace{\begin{bmatrix} 60 \\ 10 \\ 64 \\ 75 \\ 50 \\ 70 \\ 95 \\ 27 \\ 49 \end{bmatrix}}_{\mathbf{Y}} = \underbrace{\begin{bmatrix} 1 & 4 & 2 \\ 1 & 1 & 5 \\ 1 & 10 & 0.5 \\ 1 & 14 & 0.1 \\ 1 & 4 & 3 \\ 1 & 7 & 1 \\ 1 & 22 & 0.1 \\ 1 & 3 & 4 \\ 1 & 8 & 1.5 \end{bmatrix}}_{\mathbf{X}} \mathbf{x} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}}_{\boldsymbol{\beta}}$$

Least Squares Estimation

- Choose the value of **β** that minimizes the sum of squared errors

$$(Y - X\beta)'(Y - X\beta)$$

- The least squares estimate of **β**

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Exploring the Housing Dataset

https://archive.ics.uci.edu/ml/datasets/Housing

- Information about houses in the suburbs of Boston collected by D. Harrison and D.L. Rubinfeld in 1978
- 506 samples
- Features:

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.00632	18	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
1	0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
2	0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
						6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
						7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2

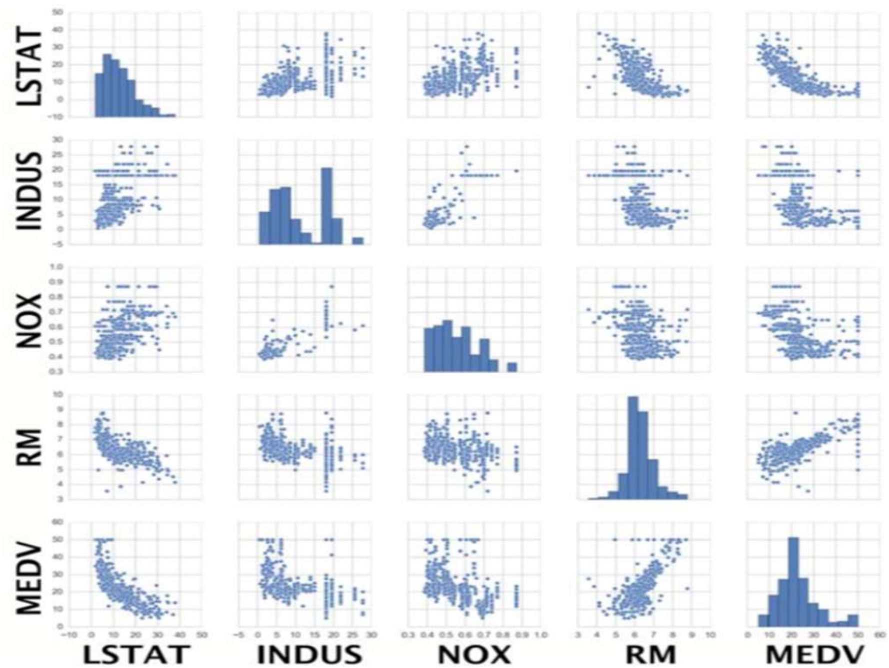
- **CRIM**: This is the per capita crime rate by town
- **ZN**: This is the proportion of residential land zoned for lots larger than 25,000 sq.ft.
- **INDUS**: This is the proportion of non-retail business acres per town
- **CHAS**: This is the Charles River dummy variable (this is equal to 1 if tract bounds river; 0 otherwise)
- **NOX**: This is the nitric oxides concentration (parts per 10 million)
- **RM**: This is the average number of rooms per dwelling
- **AGE**: This is the proportion of owner-occupied units built prior to 1940
- **DIS**: This is the weighted distances to five Boston employment centers
- **RAD**: This is the index of accessibility to radial highways
- **TAX**: This is the full-value property-tax rate per $10,000
- **PTRATIO**: This is the pupil-teacher ratio by town
- **B**: This is calculated as *1000(Bk - 0.63)^2*, where Bk is the proportion of people of African American descent by town
- **LSTAT**: This is the percentage lower status of the population
- **MEDV**: This is the median value of owner-occupied homes in $1000s

Target variable (y)

Housing dataset: EDA

```
>>> import matplotlib.pyplot as plt
>>> import seaborn as sns
>>> sns.set(style='whitegrid', context='notebook')
>>> cols = ['LSTAT', 'INDUS', 'NOX', 'RM', 'MEDV']
>>> sns.pairplot(df[cols], size=2.5);
>>> plt.show()
```

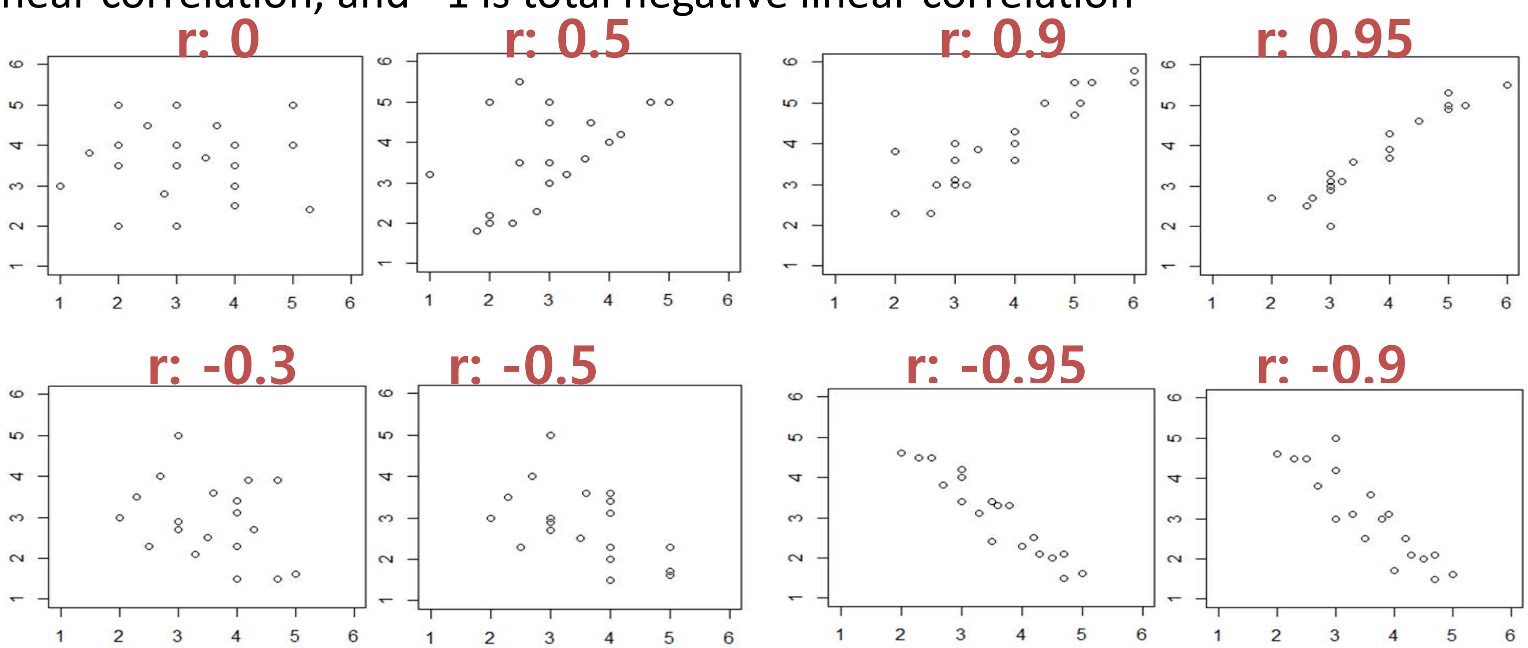
Scatter plot



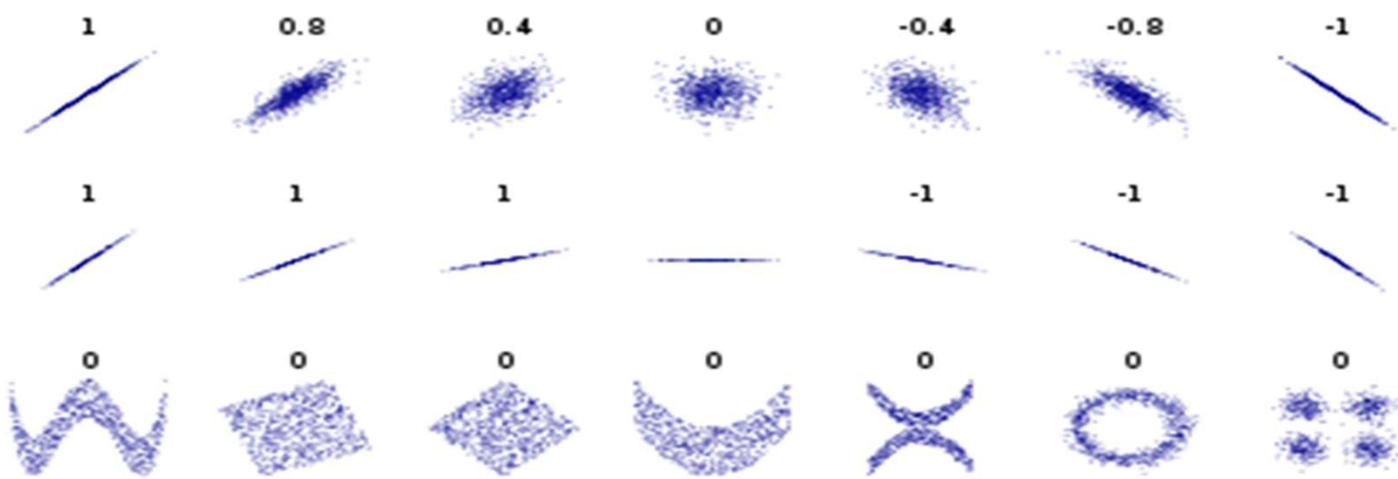
Pearson's Correlation coefficient

$$r = \frac{\sum_{i=1}^{n}\left[\left(x^{(i)}-\mu_x\right)\left(y^{(i)}-\mu_y\right)\right]}{\sqrt{\sum_{i=1}^{n}\left(x^{(i)}-\mu_x\right)^2}\sqrt{\sum_{i=1}^{n}\left(y^{(i)}-\mu_y\right)^2}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

- A measure of the linear correlation between two variables *X* and *Y*
- A value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation

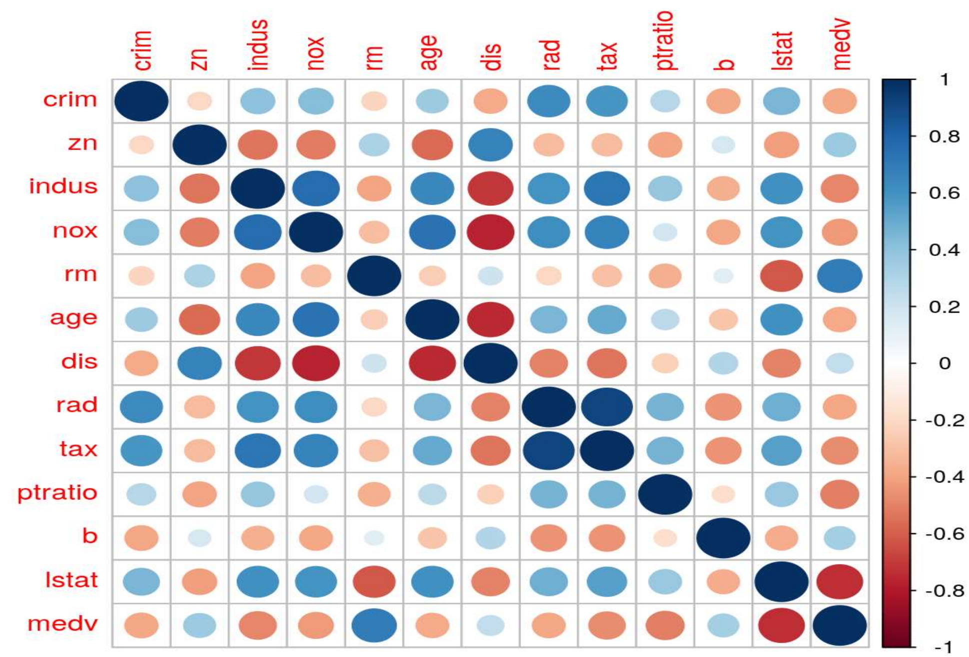


Pearson’s Correlation coefficient



Housing dataset: correlation


```
>>> import numpy as np
>>> cm = np.corrcoef(df[cols].values.T)
>>> sns.set(font_scale=1.5)
>>> hm = sns.heatmap(cm,
...             cbar=True,
...             annot=True,
...             square=True,
...             fmt='.2f',
...             annot_kws={'size': 15},
...             yticklabels=cols,
...             xticklabels=cols)
>>> plt.show()
```

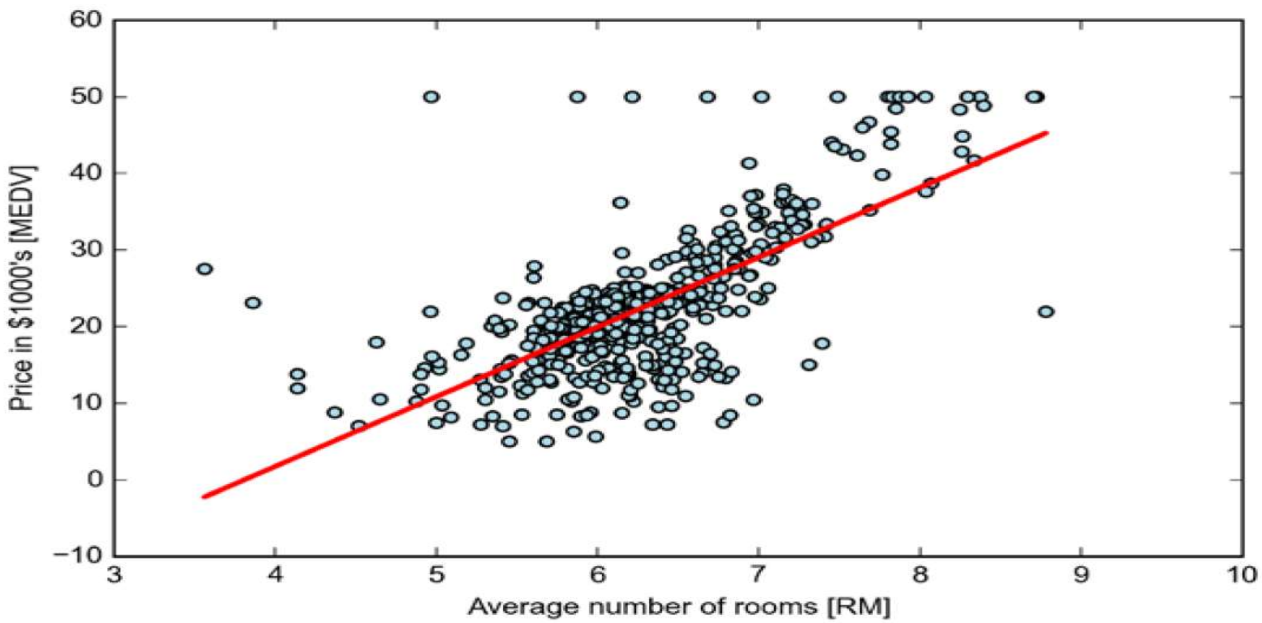


Housing dataset: Regression Model fitting

X : RM

Y : MEDV

Regression line: Y = $\beta_0 + \beta_1 x$



β_1 = 9.102

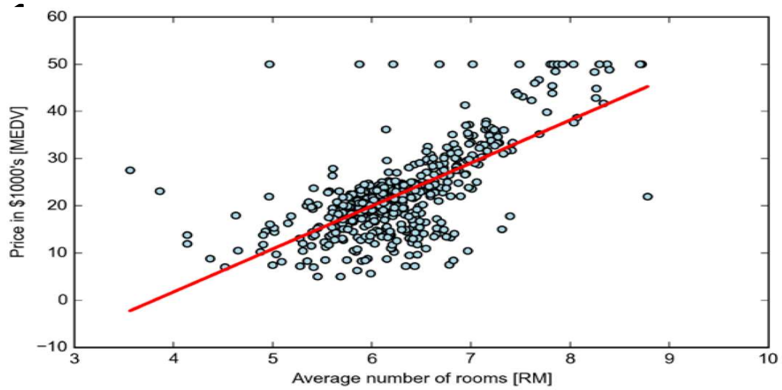
β_0 = 34.671

β_1 = 9.102
β_0 = -34.671

Coefficient Interpretation

1. Slope (β_1)

   Price (Y) is expected to increase by 9.102 units (in $1000) for each 1 increase in average number of rooms (X)



2. Intercept (β_0)
   - Price when X = 0
     - difficult to explain in many cases

Performance measure

$$MSE = \frac{1}{n}\sum_i (y_i - \hat{y}_i)^2$$

$$= \frac{1}{n}\sum_i (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + ... + \hat{\beta}_p x_{pi})^2)$$

RMSE = $(MSE)^{1/2}$

$$MAE = \frac{1}{n}\sum_i |y_i - \hat{y}_i|$$

$$\boxed{\begin{aligned} R^2 &= 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \\ R^2_{adj} &= 1 - \frac{n-1}{n-p-1}(1 - R^2) \end{aligned}}$$

The proportion of variance explained by our model

- P-values: if the p-value is low (e.g. below 0.05), then the coefficient is highly likely to be nonzero and therefore significant

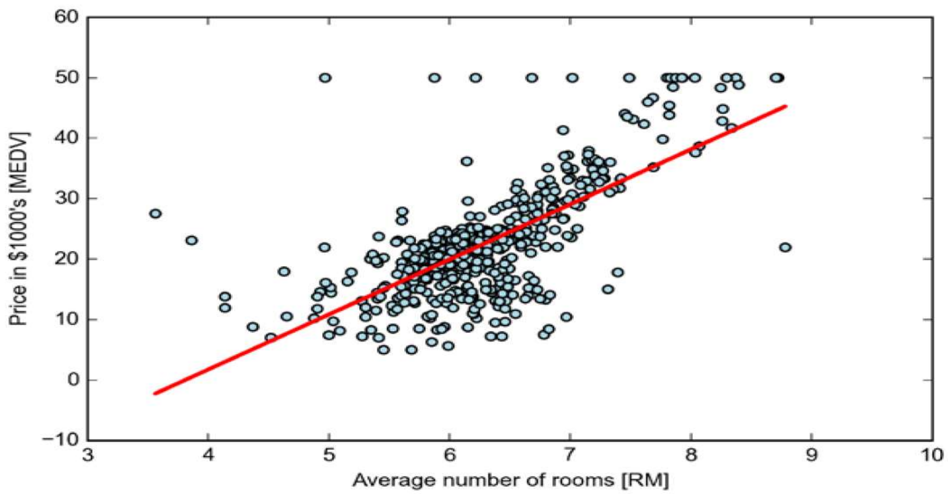
Measures of Fit: R^2

- Some of the variation in Y can be explained by variation in the X's and some cannot.
- R^2 tells you the fraction of variance that can be explained by X.

$$R^2 = 1 - \frac{RSS}{\sum (Y_i - \bar{Y})^2} \approx 1 - \frac{\text{Ending Variance}}{\text{Starting Variance}}$$

R^2 is always between 0 and 1. Zero means no variance has been explained. One means it has all been explained (perfect fit to the data).

Model evaluation



MSE on the training set = 19.96
MSE of the test set = 27.20,
which is an indicator that our model is overfitting the training data.

$R^2_{training}$ = 0.765
R^2_{test} = 0.673

Or do **cross-validation** by looking at the Mean Squared Error (MSE)

Model assumption

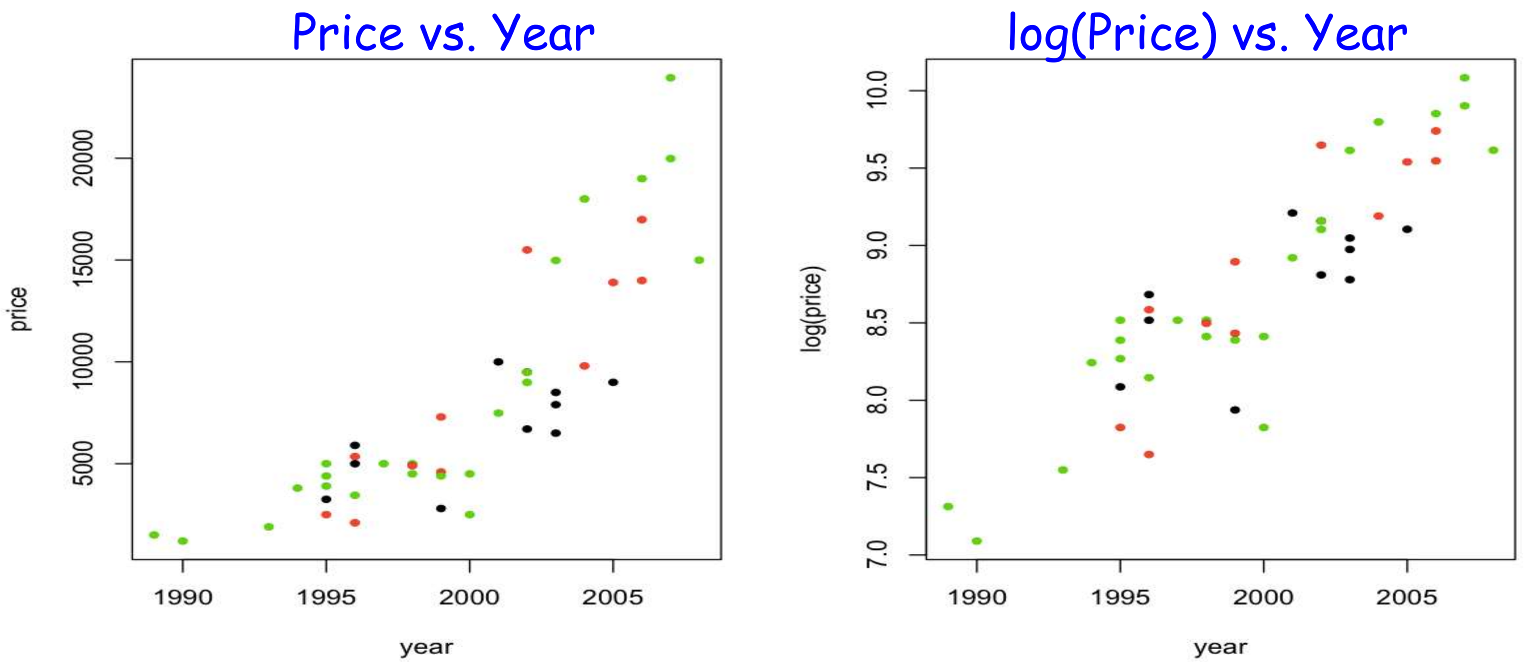
- If the model assumptions do not hold
  - Prediction can be systematically biased

- Variance Stabilizing Transformation
  - log(Y): If Y has only positive values, take log(Y)
  - Sqrt(Y) is another useful transformation
  - Or Y/X

In general, think about in what scale you expect linearity

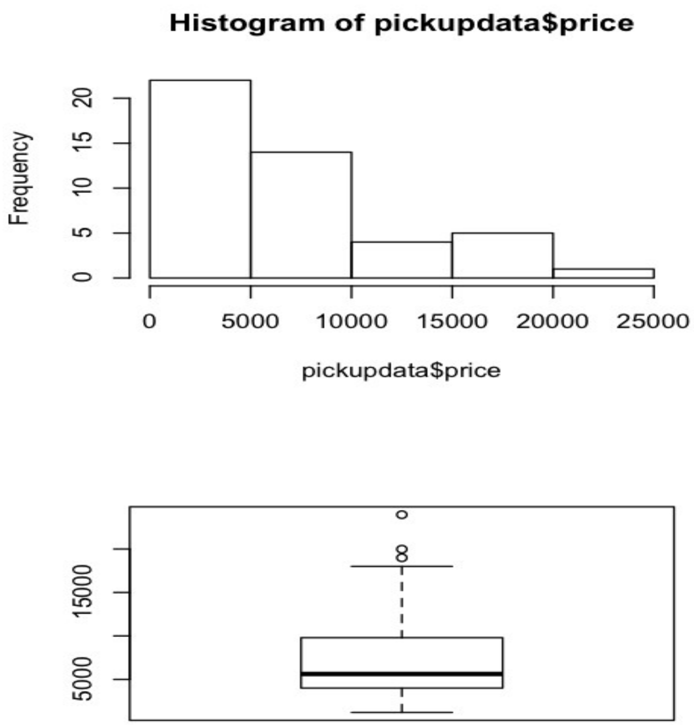
Log Transform example

- Reconsider the regression of truck price onto year

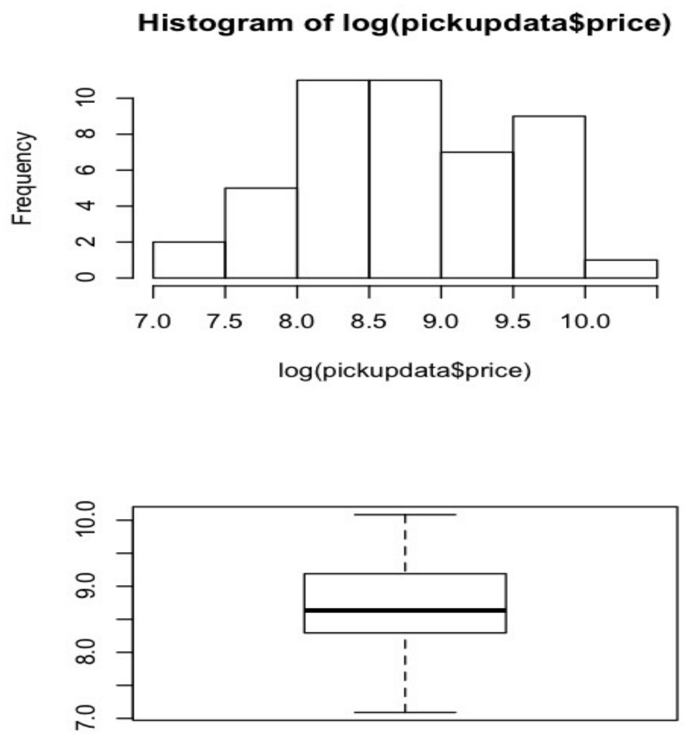


Log-transformation

Before



After

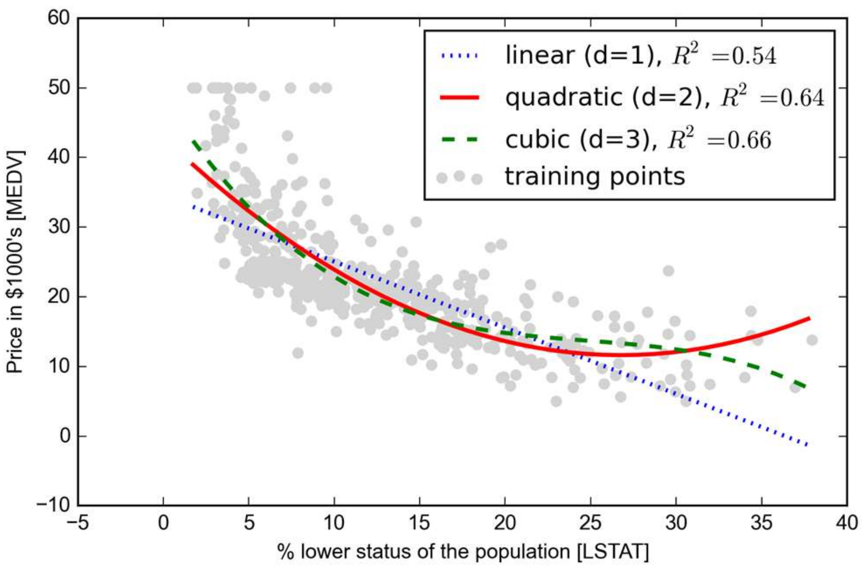


Modeling nonlinear regression

- Polynomial regression

$$y = w_0 + w_1 x + w_2 x^2 + \cdots + w_d x^d$$

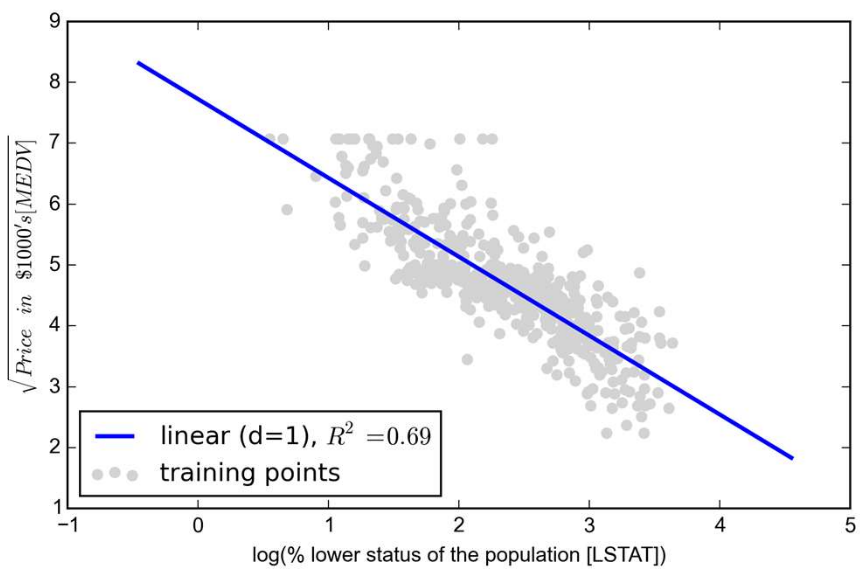
d: degree of the polynomial



- Smallest R^2 for cubic regression, but with increased complexity, so careful about overfitting

Modeling nonlinear regression

- Linear regression combined with feature transformation

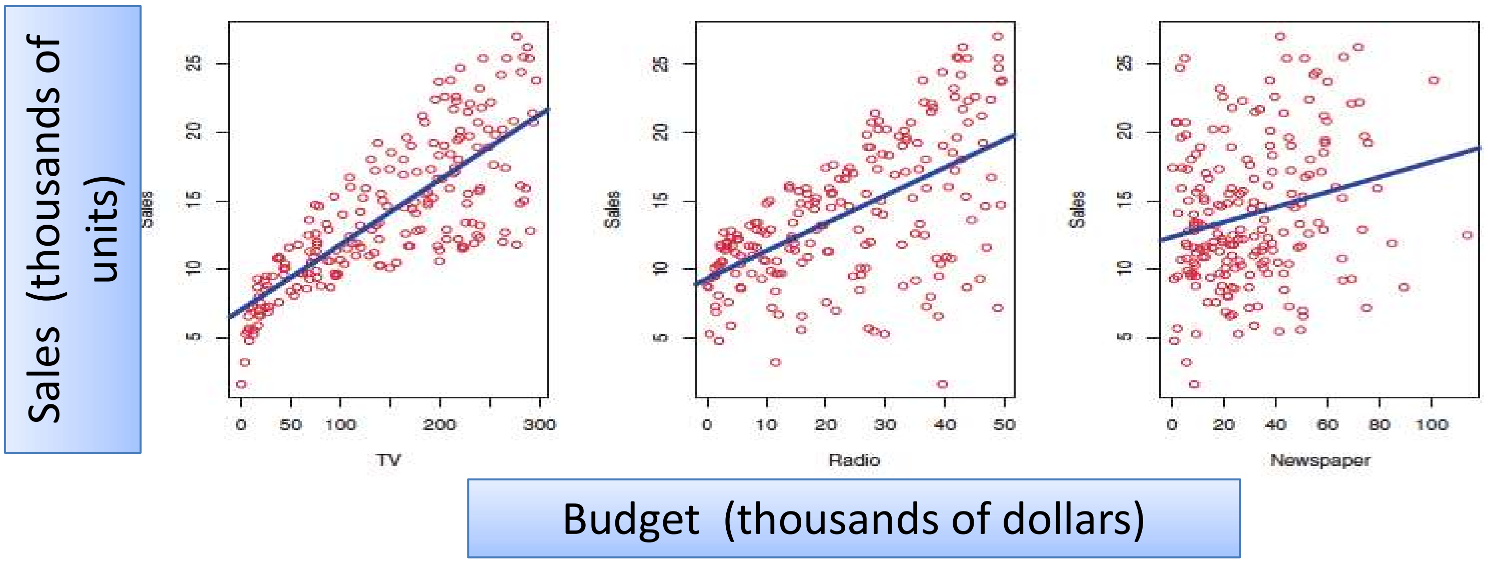


- Log-transformation
- Square root

Nonlinear regression models

- Decision tree/Random forest regression
- KNN regression
- Support vector regression 등

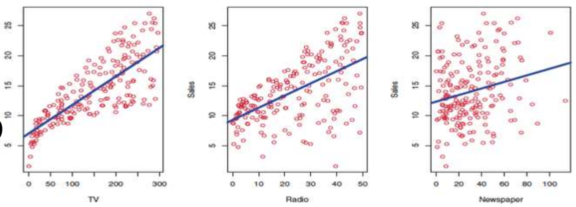
Example: *Advertising* data set



- Suppose that in our role as statistical consultants we are asked to suggest, on the basis of this data, 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?

Some important questions



- 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 estimate the effect of each medium on sales
- How accurately can we predict future sales
- Is the relationship linear
- Is there synergy among the advertising media?

Linear model

- $$sales \approx \beta_0 + \beta_1 \times \mathrm{TV}$$

$$sales \approx \beta_0 + \beta_1 \times \mathrm{TV} + \beta_2 \times Radio + \beta_3 \times Newspaper$$

Estimate the coefficients by minimizing the least squares

Is β_j=0 i.e. is X_j an important variable?

- We use a hypothesis test to answer this question
- H_0: $\beta_j=0$ vs H_a: $\beta_j \neq 0$
- Calculate $\boxed{t = \hat{\beta}_j \big/ SE(\hat{\beta}_j)}$
- If t is large (equivalently p-value is small) we can be sure that $\beta_j \neq 0$ and that there is a relationship

Regression coefficients

	Coefficient	Std Err	t-value	p-value
Constant	7.0326	0.4578	15.3603	0.0000
TV	0.0475	0.0027	17.6676	0.0000

$\hat{\beta}_1$ $\quad SE(\hat{\beta}_1)$

$\hat{\beta}_1$ is 17.67 SE's from 0

P-value

Testing Individual Variables

Is there a (statistically detectable) linear relationship between Newspapers and Sales after all the other variables have been accounted for?

Regression coefficients

	Coefficient	Std Err	t-value	p-value	
Constant	2.9389	0.3119	9.4223	0.0000	
TV	0.0458	0.0014	32.8086	0.0000	
Radio	0.1885	0.0086	21.8935	0.0000	
Newspaper	-0.0010	0.0059	-0.1767	0.8599	← No: big p-value

Regression coefficients

	Coefficient	Std Err	t-value	p-value	
Constant	12.3514	0.6214	19.8761	0.0000	
Newspaper	0.0547	0.0166	3.2996	0.0011	← Small p-value in simple regression

Almost all the explaining that Newspapers could do in simple regression has already been done by TV and Radio in multiple regression!