DATA7001
INTRODUCTION TO DATA SCIENCE

Module 4 Making the Data Confess Part 3

What Will be Covered

- Overview of machine learning
- Regression
    - Nearest neighbor regression
    - Simple linear regression
    - Multiple linear regression
    - Nonlinear regression via basis expansion
- Classification
- Clustering
- Model selection

Regression

- In regression, we want to find out the relationship between the predictors and a numerical output.
  - E.g. determine how sales is related to advertisement costs
- In the simplest case, there is a linear relationship between the predictors and the output.
  - Linear models are easy to interpret.
- In general the relationship is nonlinear.
  - E.g. the daily maximum temperature is not a linear function of the day in a year.

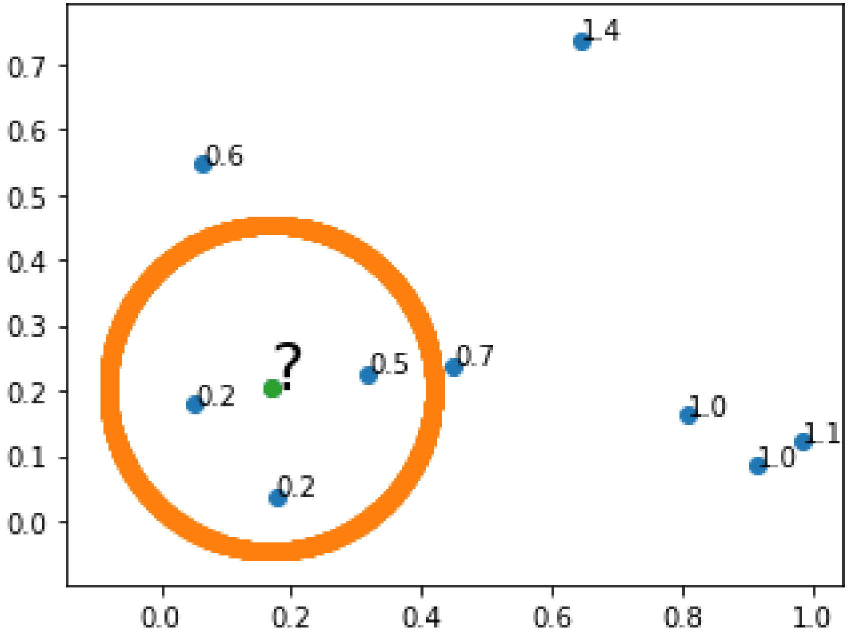
Nearest Neighbor Regression

- The idea in a Chinese proverb

近朱者赤，近墨者黑

(English: one takes on the color of one’s company)

- Formally, given a set of training examples, to predict output for x
  - Find the k nearest neighbors of x (using some distance measure)
  - Predict the average output value for these k examples



- Quiz: what does 1-NN predict for (1,1) given the following training set?

x	**y**
(0, 0)	10
(1, 1)	11
(2, 2)	12

Simple Linear Regression

- This is one of the simplest regression problem
  - There is only a single predictor
  - The output is assumed to be a linear function of the predictor
- Mathematically,

  $Y = f(X) + \varepsilon, where$

  - $f(X) = \beta_0 + \beta_1 X$
  - Errors independent and identically distributed (iid) according to a Normal distribution with mean zero and constant variance σ^2.

- We fit a model $\widehat{\boldsymbol{\beta}} = \left(\widehat{\beta_0}, \widehat{\beta_1}\right)$ by minimising the sum of squared residuals

$$\mathrm{RSS} = \left(y_1 - \widehat{\beta_0} - \widehat{\beta_1} x_1\right)^2 + \cdots + \left(y_n - \widehat{\beta_0} - \widehat{\beta_1} x_n\right)^2 \equiv n\ \mathrm{MSE}$$

  - ○ Equivalently, by *Maximum Likelihood*

- We have closed-form formula for computing the parameter estimates.

- The error term ε is assumed to follow a normal distribution, and its standard deviation shows how uncertainty the data is.
- Thus, besides estimating $\boldsymbol{\beta}$,we are also interested in estimating the standard deviation of the error term.
- This is estimated using the residual standard error, which is the square root of SSE/(n-2)
- We can use this to construct a confidence interval for the predicted value.

Goodness of Fit

- We are often interested in how well a model fit the training data.
- For a regression model, the R^2 statistic is often used as a goodness of fit measure
  - It is the proportion of variance explained by the model.
  - Takes a value between 0 and 1;
    - 0: no variance explained
    - 1: all variance explained

Model Validation

- A linear model makes very strong assumptions.
- We often want to check whether the assumptions are valid.
- There are a few common checks
    - Plotting the residuals themselves
    - Constructing a *quantile-quantile (qq) plot*
    - Examine summary statistics of the residuals

Uncertainty in Model Parameters

- Model is trained using a random sample, and thus the model parameters are random variables themselves
- For simple linear regression, we can estimate the standard errors of the parameter estimates
- We can also construct approximate confidence intervals for the parameters.
  - 95% confidence interval: [parameter estimate – 1.96 * standard error, parameter estimate + 1.96 * standard error]
  - Use this only when model assumptions have not been shown to invalid!

Significance of A Predictor

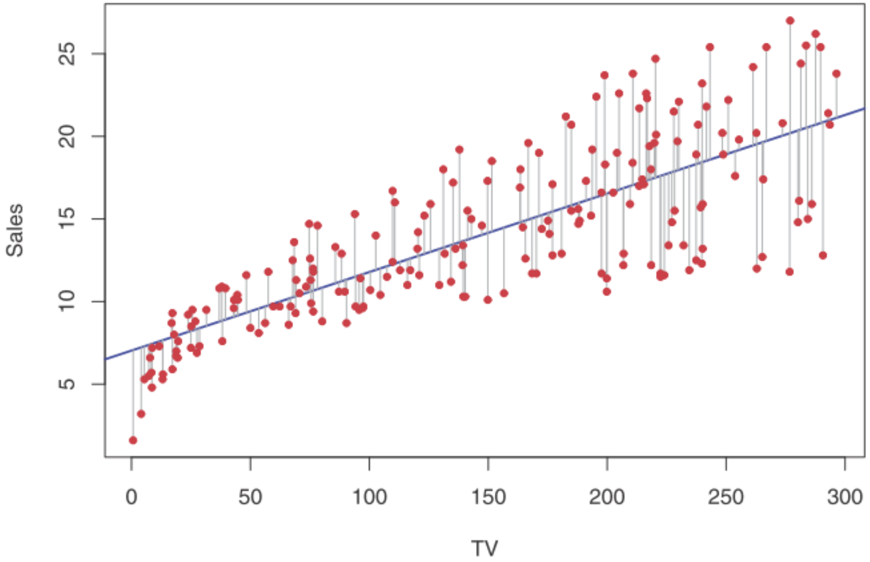
- We are often interested to know whether a predictor has a significant effect on the output
    - That is, whether the weight of the predictor is different from 0
- The p-value of a parameter measures the probability of observing the given parameter estimate if the true value is in fact zero.
- The smaller the p-value, the more unlikely that the true parameter value is 0.

Significance of Several Predictors

- Testing whether a single predictor is significant can give misleading result, especial when there are many predictors.
- Another question of interest is whether all predictors have 0 weights.
- The F statistic is used to answer this question
  - The larger the statistic, the more unlikely that the true model only has a bias term.
  - A value far larger than 1 indicates that some predictor has nonzero weight.
  - Again, use this only when the model assumptions have not been shown to be invalid.

Example

- The diagram below shows a plot of sales against TV advertisement budget.



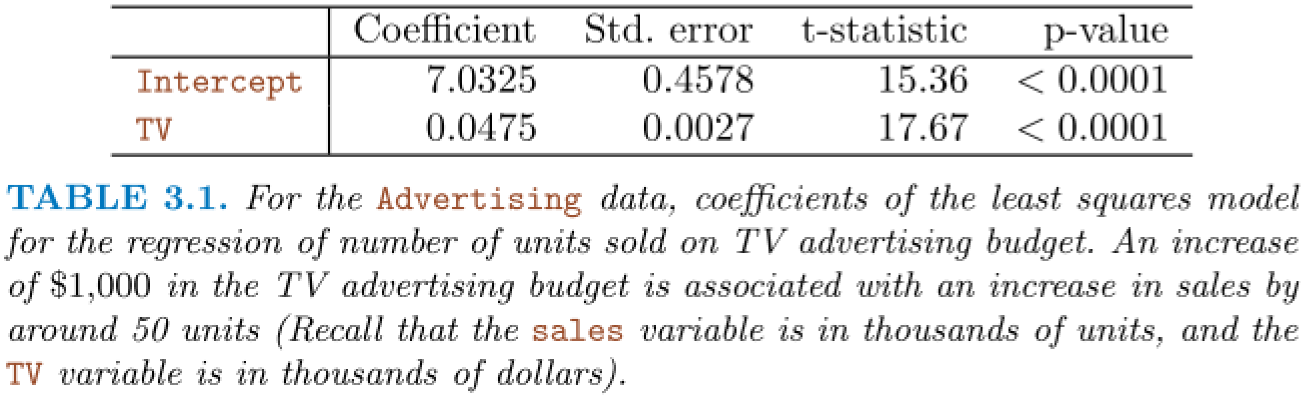
- This suggests an approximately linear relationship

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \epsilon.$$

- Question: does the residuals satisfy the model assumption?

Figure reproduced from: G. James, D. Witten, T. Hastie, R. Tibshirani. An Introduction to Statistical Learning: with Applications in R. Corrected 6th printing, 2015. Springer, New York.

- The fitted model parameters are shown below



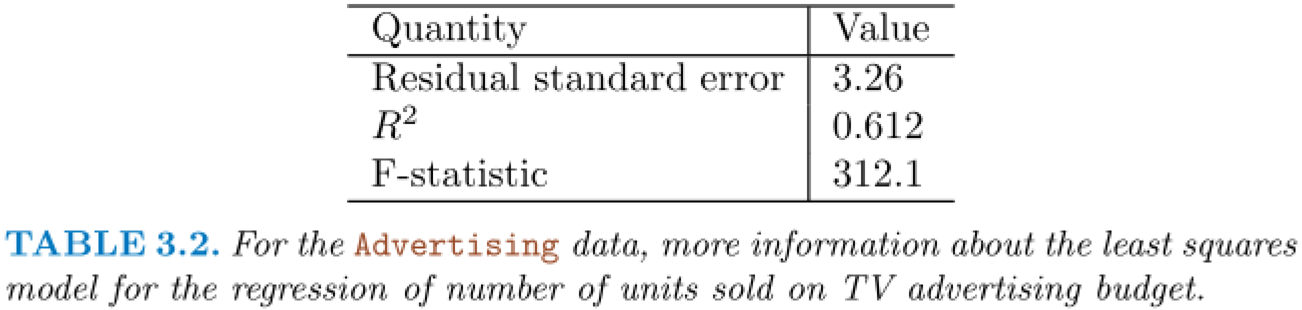
	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

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

- We can see that the p-values for the parameters are very small, and thus both the TV predictor and the bias are significant.
    - Note that strictly speaking, we should not do this as the model assumption is not correct (the residuals violate the model assumption)
- Questions
    - What is the predicted average level of sales for a TV budget of 100?
    - What is a 95% confidence interval for the intercept?

Figure reproduced from: G. James, D. Witten, T. Hastie, R. Tibshirani. An Introduction to Statistical Learning: with Applications in R. Corrected 6th printing, 2015. Springer, New York.

- The estimated standard deviation for the residual ε is 3.26.
- The goodness of fit as measured by R^2 is 0.612.
- The F statistic is far larger than 1, thus the TV predictor is important.



Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

TABLE 3.2. *For the* Advertising *data, more information about the least squares model for the regression of number of units sold on TV advertising budget.*

Multiple Linear Regression

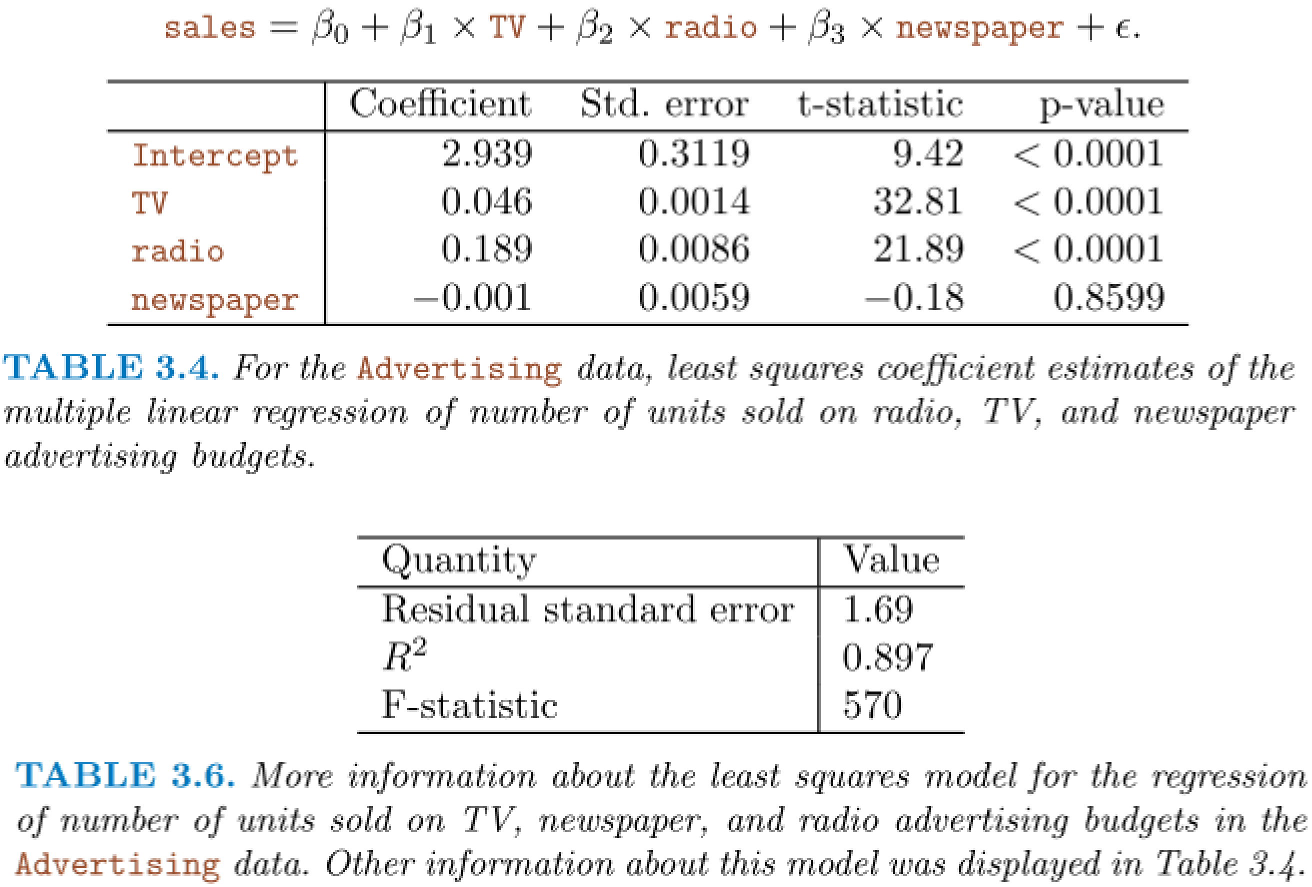
- In general, we have more than one predictors.
- Multiple linear regression generalizes simple linear regression to this case.
- The model assumptions are the same, except that to handle multiple predictors, we assume that $f(X)$ is linear in the predictors

$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Nonlinear Regression

- For many problems, the response is not a linear function of the predictors
- A useful trick to learn a nonlinear relationships is to include interaction terms, and then perform linear regressio
$$f(\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \beta_{1:2} X_1 X_2 + \cdots + \beta_{1:p} X_1 X_p + \cdots + \beta_{1:\cdots:p} X_1 X_2 \cdots X_p$$
- In R, `lm(Y~X1+X2*X3)` corresponds to:
$$f(\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{2:3} X_2 X_3$$

Number of Sales vs. Ad. Budgets



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

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	−0.001	0.0059	−0.18	0.8599

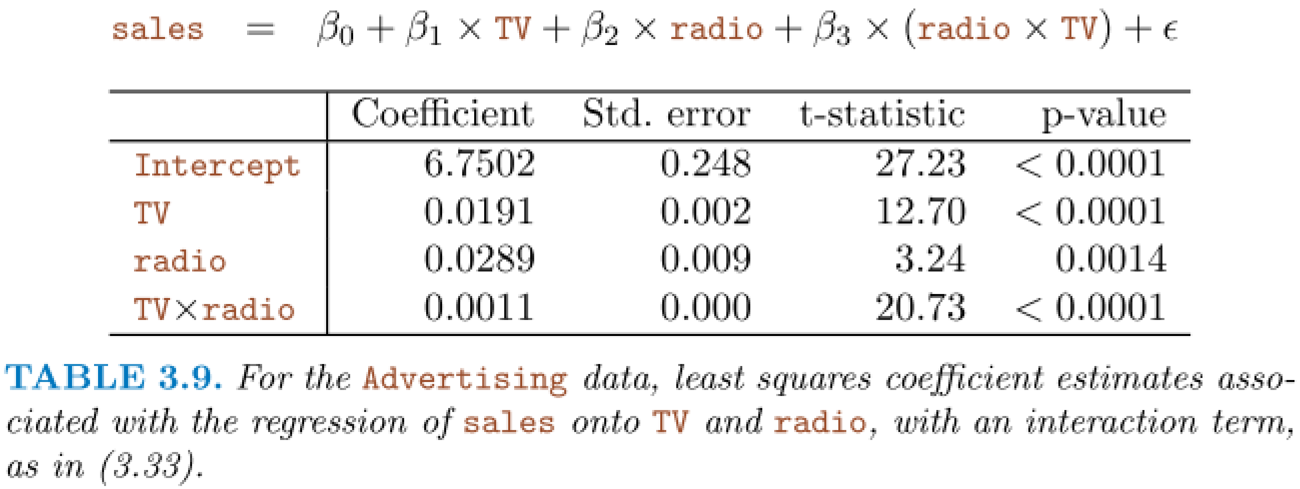
TABLE 3.4. *For the* Advertising *data, least squares coefficient estimates of the multiple linear regression of number of units sold on radio, TV, and newspaper advertising budgets.*

Quantity	Value
Residual standard error	1.69
R^2	0.897
F-statistic	570

TABLE 3.6. *More information about the least squares model for the regression of number of units sold on TV, newspaper, and radio advertising budgets in the* Advertising *data. Other information about this model was displayed in Table 3.4.*

Tables reproduced from: G. James, D. Witten, T. Hastie, R. Tibshirani. An Introduction to Statistical Learning: with Applications in R. Corrected 6th printing, 2015. Springer, New York.

Including Interaction Terms



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

	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

TABLE 3.9. *For the* Advertising *data, least squares coefficient estimates associated with the regression of* sales *onto* TV *and* radio*, with an interaction term, as in (3.33).*

Table reproduced from: G. James, D. Witten, T. Hastie, R. Tibshirani. An Introduction to Statistical Learning: with Applications in R. Corrected 6th printing, 2015. Springer, New York.

Summary

- What can we make the data confess about
  - We can’t prove a model is correct using data, but we can get a useful model
- Questions
  - Diagnostic analysis (insight), predictive analysis (foresight), prescriptive analysis
- Data-driven approach
- Many applications

Summary

- Overview of machine learning
  - Machine learning approaches
  - How a learning algorithm works
  - What is the objective of learning
  - Statistical learning and prediction
- Regression
- Classification
- Clustering
- Model selection

Summary

- Overview of machine learning
- Regression
  - Nearest neighbor regression
  - Simple linear regression
  - Multiple linear regression
  - Nonlinear regression via basis expansion
- Classification
- Clustering
- Model selection

POLL QUESTIONS - MAKING THE DATA CONFESS - REGRESSION