Lecture 3: Linear Regression, kNN Regression and Inference

S-109A Introduction to Data Science

Pavlos Protopapas and Kevin Rader

Lecture Outline

Data

Statistical Modeling I

k-Nearest Neighbors (kNN)

Statistical Modeling II

Linear Regression

Bootstrap

Standard Errors

Hypothesis Testing

Confidence Intervals

Model Fitness

How does the model perform predicting?

Comparison of Two Models

How do we choose from two different models?

Evaluating Significance of Predictors

Does the outcome depend on the predictors?

How well we know $\hat{f}$

Let’s start

Predicting a Variable

Let's image a scenario where we'd like to predict one variable using another (or a set of other) variables.

Examples:

- Predicting the amount of view a YouTube video will get next week based on video length, the date it was posted, previous number of views, etc.
- Predicting which movies a Netflix user will rate highly based on their previous movie ratings, demographic data etc.

Data

Data

The **Advertising data set** consists of the sales of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper. Everything is given in units of $1000.

TV	radio	newspaper	sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9

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 "

Response vs. Predictor Variables

There is an asymmetry in many of these problems:

The variable we'd like to predict may be more difficult to measure, is more important than the other(s), or may be directly or indirectly influenced by the values of the other variable(s).

Thus, we'd like to define two categories of variables:

- variables whose value we want to predict
- variables whose values we use to make our prediction.

Response vs. Predictor Variables

X
predictors
features
covariates

Y
outcome
response variable
dependent variable

n observations

TV	radio	newspaper	sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9

p predictors

Response vs. Predictor Variables

$X = X_1, \dots, X_p$

$X_j = x_{1j}, \dots, x_{ij}, \dots, x_{nj}$

predictors

features

covariates

$Y = y_1, \dots, y_n$

outcome

response variable

dependent variable

n observations

TV	radio	newspaper	sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9

p predictors

Definition

We are observing $p + 1$ number variables and we are making n sets of observations. We call:

- the variable we'd like to predict the **outcome** or **response variable**; typically, we denote this variable by Y and the individual measurements y_i.
- the variables we use in making the predictions the **features** or **predictor** variables; typically, we denote these variables by $X = X_1, \dots, X_p$ and the individual measurements $x_{i,j}$.

Note: i indexes the observation ($i = 1, \dots, n$) and j indexes the value of the j-th predictor variable (j $= 1, \dots, p$).

Statistical Model

True vs. Statistical Model

We will assume that the response variable, Y, relates to the predictors, X, through some unknown function expressed generally as:

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

Here, f is the unknown function expressing an underlying rule for relating Y to X, ε is the random amount (unrelated to X) that Y differs from the rule $f(X)$.

A ***statistical model*** is any algorithm that estimates f. We denote the estimated function as $\widehat{f}$.

Statistical Model

Statistical Model

How do we find $\hat{f}(x)$?

Statistical Model

How do we find $\hat{f}(x)$?

Statistical Model

Simple idea is to take the mean of all y's, $\hat{f}(x) = \frac{1}{n}\sum_1^n y_i$

Prediction vs. Estimation

For some problems, what's important is obtaining $\hat{f}$, our estimate of f. These are called ***inference*** problems.

When we use a set of measurements, $(x_{i,1}, \dots, x_{i,p})$ to predict a value for the response variable, we denote the ***predicted*** value by:

$$\hat{y}_i = \hat{f}(x_{i,1}, \dots, x_{i,p}).$$

For some problems, we don't care about the specific form of $\hat{f}$, we just want to make our prediction $\hat{y}$ as close to the observed value y as possible. These are called ***prediction problems.***

Simple Prediction Model

What is $\hat{y}_q$ at some x_q ?

Find distances to all other points $D(x_q, x_i)$

Find the nearest neighbor, (x_p, y_p)

Predict $\hat{y}_q = y_p$

Simple Prediction Model

Do the same for "all" $x's$

Extend the Prediction Model

What is $\hat{y}_q$ at some x_q ?

Find distances to all other points $D(x_q, x_i)$

Find the k-nearest neighbors, $x_{q_1}, \ldots, x_{q_k}$

Predict $\widehat{y_q} = \frac{1}{k}\sum_i^k y_{q_i}$

Simple Prediction Models

Simple Prediction Models

Same models on more data

k-Nearest Neighbors

The ***k-Nearest Neighbor (kNN) model*** is an intuitive way to predict a quantitative response variable:

> *to predict a response for a set of observed predictor values, we use the responses of other observations most similar to it*

Note: this strategy can also be applied in classification to predict a categorical variable. We will encounter kNN again later in the course in the context of classification.

k-Nearest Neighbors - kNN

For a fixed a value of *k*, the predicted response for the i-th observation is the average of the observed response of the *k*-closest observations:

$$\hat{y}_n = \frac{1}{k}\sum_{i=1}^{k} y_{n_i}$$

where $\{x_{n1}, \dots, x_{nk}\}$ are the *k* observations most similar to x_i (*similar* refers to a notion of distance between predictors).

Things to Consider

Model Fitness

How does the model perform predicting?

Comparison of Two Models

How do we choose from two different models?

Evaluating Significance of Predictors

Does the outcome depend on the predictors?

How well we know $\hat{f}$

The confidence intervals of our $\hat{f}$

Error Evaluation

Error Evaluation

Start with some data.

Error Evaluation

Hide some of the data from the model. This is called **train-test** split.

We use the train set to estimate $\hat{y}$, and the test set to evaluate the model.

Error Evaluation

Estimate $\hat{y}$ for *k=1* .

Error Evaluation

Now, we look at the data we have not used, the **test data** (red crosses).

Error Evaluation

Calculate the **residuals** $(y_i - \hat{y}_i)$.

Error Evaluation

Do the same for *k=3.*

Error Evaluation

In order to quantify how well a model performs, we define a ***loss*** or ***error function.***

A common loss function for quantitative outcomes is the **Mean Squared Error (MSE):**

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

The quantity $y_i - \hat{y}_i$ is called a ***residual*** and measures the error at the *i*-th prediction.

Error Evaluation

Caution: The MSE is by no means the only valid (or the best) loss function!

Question: What would be an intuitive loss function for predicting categorical outcomes?

Note: The square **R**oot of the **M**ean of the **S**quared **E**rrors (RMSE) is also commonly used.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}$$

Things to Consider

Comparison of Two Models

How do we choose from two different models?

Model Fitness

How does the model perform predicting?

Evaluating Significance of Predictors

Does the outcome depend on the predictors?

How well do we know $\hat{f}$

The confidence intervals of our $\hat{f}$

Model Comparison

Model Comparison

Do the same for all *k*'s and compare the RMSEs. *k=3* seems to be the best model.

Things to Consider

Comparison of Two Models

How do we choose from two different models?

Model Fitness

How does the model perform predicting?

Evaluating Significance of Predictors

Does the outcome depend on the predictors?

How well do we know $\hat{f}$

The confidence intervals of our $\hat{f}$

Model Fitness

Model fitness

For a subset of the data, calculate the RMSE for *k=3*. Is RMSE=5.0 good enough?

Model fitness

What if we measure the Sales in cents instead of dollars?

RMSE is now 5004.93.
Is that good?

Model fitness

It is better if we compare it to something.

We will use the simplest model:

$$\hat{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

R-squared

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

- If our model is as good as the mean value, $\bar{y}$, then $R^2 = 0$
- If our model is perfect then $R^2 = 1$
- R^2 can be negative if the model is worst than the average. This can happen when we evaluate the model in the test set.

Linear Regression

Linear Models

Note that in building our kNN model for prediction, we did not compute a closed form for $\hat{f}$.

What if we ask the question:

"how much more sales do we expect if we double the TV advertising budget?"

Alternatively, we can build a model by first assuming a simple form of f:

$$Y = f(X) + \epsilon = \beta_1 X + \beta_0 + \epsilon.$$

Linear Regression

If our statistical model is:

$$Y = f(X) + \epsilon = \beta_1^{\text{true}} X + \beta_0^{\text{true}} + \epsilon,$$

then it follows that our estimate is:

$$\widehat{Y} = \widehat{f}(X) = \widehat{\beta_1} X + \widehat{\beta_0}$$

where $\hat{\beta}_1$ and $\hat{\beta}_0$ are estimates of β_1 and β_0 respectively, that we compute using observations.

Linear Regression

For a given data set

Linear Regression

Is this line good?

Linear Regression

Maybe this one?

Linear Regression

Or this?

Inference for Linear Regression

We choose $\hat{\beta}_1$ and $\hat{\beta}_0$ in order to minimize the predictive errors made by our model, i.e. minimize our loss function.

Again we use MSE as our loss function,

$$L(\beta_0, \beta_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - (\beta_1 X + \beta_0)\right]^2.$$

Then the optimal values for $\hat{\beta}_0$ and $\hat{\beta}_1$ should be:

$$\widehat{\beta}_0, \widehat{\beta}_1 = \underset{\beta_0, \beta_1}{\operatorname{argmin}} L(\beta_0, \beta_1).$$

Inference for Linear Regression

First calculate the residuals

Linear Regression

And calculate the loss function for every possible β_0 and β_1. Find the β_0 and β_1 where the loss function is minimum.

E.g. the loss function for different β_1 (β_0 is fixed to be 6).

Alternatively:

Take the partial derivatives of L with respect to β_0 and β_1, set to zero, and find the minimum. This procedure will give us explicit formulae for $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}$$

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

where $\bar{y}$ and $\bar{x}$ are sample means.

The line:

$$\widehat{Y} = \widehat{\beta}_1 X + \widehat{\beta}_0$$

is called the **regression line**.

More on Loss Function

There are multiple ways to measure the fitness of a model, i.e. there are multiple **loss functions**.

Max absolute deviation: Count only the biggest *error*

$$\max_i |y_i - \widehat{y}_i|$$

Sum of absolute deviations: Add up the *error*

$$\sum_i |y_i - \widehat{y}_i| \quad \text{or} \quad \frac{1}{n}\sum_i |y_i - \widehat{y}_i|$$

Sum of squared errors: Add up the squared *error*

$$\sum_i |y_i - \widehat{y}_i|^2 \quad \text{or} \quad \frac{1}{n}\sum_i |y_i - \widehat{y}_i|^2$$

Loss Functions Revisited

Question: In what scenarios would you choose Mean Absolute Deviation or Max Absolute Deviation over MSE?

Things to Consider

Comparison of Two Models

How do we choose from two different models?

Model Fitness

How does the model perform predicting?

Evaluating Significance of Predictors

Does the outcome depend on the predictors?

How well do we know $\hat{f}$

The confidence intervals of our $\hat{f}$

Evaluating Significance of Predictors

Evaluating Significance of Predictors

We interpret the ε term in our observation

$$y = f(x) + \epsilon$$

to be noise introduced by random variations in natural systems or imprecisions of our scientific instruments.

If we knew the exact form of $f(x)$, for example, $f(x) = \beta_0 + \beta_1 x$, and there was no ε then estimating the $\hat{\beta}'s$ would have been exact.

Evaluating Significance of Predictors

However, three things happen:

- we do not know the exact form of $f(x)$
- $\boldsymbol{\varepsilon}$ is always there
- limited sample size

We first address ε

We call ε the measurement error or ***irreducible error***. Since even predictions made with the actual function f will not match observed values of y.

Because of ε, every time we measure the response Y for a fix value of X we will obtain a different observation, and hence a different estimate of $\hat{\beta}'s$.

Evaluating Significance of Predictors

Start with a model

Evaluating Significance of Predictors

For a some values of $X, Y = f(X)$

Evaluating Significance of Predictors

But due to error, every time we measure the response *Y* for a fixed value of X we will obtain a different observation.

Evaluating Significance of Predictors

One set of observations, ‘one realization’, means one set of *Y*s (red crosses).

Evaluating Significance of Predictors

Another set of observations, ‘another realization’ means another set of *Y*s (green crosses).

Evaluating Significance of Predictors

Another set of observations, ‘another realization’ means another set of *Y*s (black crosses).

Evaluating Significance of Predictors

For each one of those 'realizations', we could fit a model and estimate, $\hat{\beta}_0$ and $\hat{\beta}_1$.

Evaluating Significance of Predictors

For each one of those 'realizations', we could fit a model and estimate, $\hat{\beta}_0$ and $\hat{\beta}_1$.

Evaluating Significance of Predictors

For each one of those 'realizations', we could fit a model and estimate, $\hat{\beta}_0$ and $\hat{\beta}_1$.

Evaluating Significance of Predictors

Repeat this on real data. We now select a sub-sample from all the data.

Evaluating Significance of Predictors

Another sub-sample.

Evaluating Significance of Predictors

Another sub-sample.

Evaluating Significance of Predictors

Another sub-sample.

Evaluating Significance of Predictors

Repeat this for 100 times.

Evaluating Significance of Predictors

We can estimate the mean and standard deviation of the estimate $\hat{\beta}_1$

Evaluating Significance of Predictors

We can calculate the confidence intervals, which are the ranges of values such that the true value of β_1 is contained in this interval with *n* percent probability.

Now we can answer the question, “how significant are the predictors?” Here we show the same analysis for all three predictors.

Question: Which ones are important?

We examine how consistent the estimated values of the coefficients are with zero. Looking at the area under the curve below zero.

This is called hypothesis testing which we will visit again later next lecture.

Evaluating Significance of Predictors

However, three things happen:

- we do not know the exact form of $f(x)$
- $\boldsymbol{\varepsilon}$ is always there
- Sample size

We now address sample size

As we saw in Lecture 1b, sample estimates are approximation to population means … blah balh

Let's think how this affects our estimate of predictors and how this affects are conclusions on the significance of the predictors.

Estimate Uncertainty for Coefficients

Once we have the distribution,

Estimate Uncertainty for Coefficients

Once we have the distribution, we can calculate the mean and variance.

Estimate Uncertainty for Coefficients

And then we calculate the confidence intervals and the importance of the predictors.

Sample size

Let’s re-examine this more carefully.

Sample size

Let's re-examine this more carefully.

In statistics, any estimate value is subject to sample size:

$$E[x] = \lim_{n \to \infty} \frac{1}{n} \sum_{i}^{n} x_i$$

The estimated value, is called the **sample mean**

$$\mu_x = \hat{x} = \frac{1}{n} \sum_{i}^{n} x_i$$

Whereas $E[x]$ is called the **population mean**

TV	sales
230.1	22.1
44.5	10.4
17.2	9.3
151.5	18.5
180.8	12.9
8.7	7.2
57.5	11.8
120.2	13.2
8.6	4.8
199.8	10.6
66.1	8.6
214.7	17.4
23.8	9.2
97.5	9.7
204.1	19.0
195.4	22.4
67.8	12.5
281.4	24.4
69.2	11.3
[illegible]3	14.6

Bigger the sample less variation in the prediction

How do μ_β and σ_β depend on ε, $\hat{f}$, and number of samples?

Assume we know the functional form of f, we examine the relationship to number of samples.

How do μ_β and σ_β depend on ε, $\hat{f}$, and number of samples?

How do μ_β and σ_β depend on ε, $\hat{f}$, and number of samples?

We used sub-sample size 50 before

How do μ_β and σ_β depend on ε, $\hat{f}$, and number of samples?

However, we should have used sub-sample size 200.

Bootstrap

Bootstrap

Imagine we have 5 billiard balls in a bucket.

Bootstrap

We first pick randomly a ball and replicate it. This is called **sampling with replacement.** We move the replicated ball to another bucket.

Bootstrap

Bootstrap

We then pick randomly another ball and again we replicate it.

Again, we then move the replicated ball to the other bucket.

Bootstrap

Bootstrap

We repeat this process.

Bootstrap

Again

Bootstrap

And again

Bootstrap

Until the other bucket has the same number of balls as the original one.

Bootstrap

We repeat the same process and acquire another sample

Bootstrap

We repeat the same process and acquire another sample

Bootstrapping for Estimating Sampling Error

Definition

Bootstrapping is the practice of estimating properties of an estimator by measuring those properties by, for example, sampling from the observed data.

For example, we can compute $\hat{\beta}_0$ and $\hat{\beta}_1$ multiple times by randomly sampling from our data set. We then use the variance of our multiple estimates to approximate the true variance of $\hat{\beta}_0$ and $\hat{\beta}_1$.

Standard Errors

The variance of β_0 and β_1 are also called their ***standard errors***, $SE(\hat{\beta}_0), SE(\hat{\beta}_1)$.

If our data is drawn from a larger set of observations then we can empirically estimate the **standard errors**, $SE(\hat{\beta}_0), SE(\hat{\beta}_1)$ of β_0 and β_1 through bootstrapping.

If we know the variance σ^2 of the noise ϵ, we can compute $SE(\hat{\beta}_0), SE(\hat{\beta}_1)$ analytically, using the formulae below:

$$SE\left(\widehat{\beta}_0\right) = \sigma\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_i (x_i - \overline{x})^2}}$$

$$SE\left(\widehat{\beta}_1\right) = \frac{\sigma}{\sqrt{\sum_i (x_i - \overline{x})^2}}$$

Standard Errors

In practice, we do not know the theoretical value of σ since we do not know the exact distribution of the noise ϵ. However, if we make the following assumptions,

- the errors $\epsilon_i = y_i - \hat{y}_i$ and $\epsilon_j = y_j - \hat{y}_j$ are uncorrelated, for $i \neq j$,
- each ϵ_i is normally distributed with mean 0 and variance σ^2,

then, we can empirically estimate σ^2, from the data and our regression line:

$$\sigma \approx \sqrt{\frac{n \cdot \mathrm{MSE}}{n-2}} = \sqrt{\frac{\sum_i (y_i - \widehat{y}_i)^2}{n-2}}$$

Standard Errors

Exercise: Duplicate the following results for the coefficients for TV advertising.

Method	$SE(\hat{\beta}_0)$	$SE(\hat{\beta}_1)$
Analytic Formula	0.353	0.0023
Bootstrap	0.328	0.0028

Importance of predictors

We have discussed finding the importance of predictors, by determining the cumulative distribution from ∞ to 0.

.

Importance of predictors

This is equivalent in looking the distance of the estimated value of the coefficient in units of SE($\hat{\beta}_1$) = $\sigma_{\hat{\beta}_1}$.

.

Hypothesis Testing

Hypothesis testing is a formal process through which we evaluate the validity of a statistical hypothesis by considering evidence for or against the hypothesis gathered by random sampling of the data.

1. State the hypotheses, typically a **null hypothesis**, H_0 and an **alternative hypothesis**, H_1, that is the negation of the former.
2. Choose a type of analysis, i.e. how to use sample data to evaluate the null hypothesis. Typically this involves choosing a **single test statistic**.
3. **Sample** data and compute the test statistic.
4. Use the value of the test statistic to either **reject** or **not reject** the null hypothesis.

Hypothesis testing

1. State Hypothesis:

Null hypothesis:

H_0: There is no relation between *X* and *Y*

The alternative:

H_a: There is some relation between *X* and *Y*

2: **Choose test statistics**

To test the null hypothesis, we need to determine whether, our estimate for $\hat{\beta}_1$, is sufficiently far from zero that we can be confident that $\hat{\beta}_1$ is non-zero. We use the following test statistic:

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

Hypothesis testing

3. **Sample**:

Using bootstrap we can estimate $\hat{\beta}_1$.

4. **Reject or not reject the hypothesis:**

If there is really no relationship between *X* and *Y* , then we expect that will have a *t-distribution with n-2 degrees* of freedom.

To compute the probability of observing any value equal to $|t|$ or larger, assuming $\hat{\beta}_1 = 0$ is easy. We call this probability the p-value.

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

Hypothesis testing

P-values for all three predictors done independently

Method	$SE(\hat{\beta}_0)$	$SE(\hat{\beta}_1)$
Analytic Formula	0.353	0.0023
Bootstrap	0.328	0.0028

Method	$SE(\hat{\beta}_0)$	$SE(\hat{\beta}_1)$
Analytic Formula	0.353	0.0023
Bootstrap	0.328	0.0028

Method	$SE(\hat{\beta}_0)$	$SE(\hat{\beta}_1)$
Analytic Formula	0.353	0.0023
Bootstrap	0.328	0.0028

Things to Consider

Comparison of Two Models

How do we choose from two different models?

Model Fitness

How does the model perform predicting?

Evaluating Significance of Predictors

Does the outcome depend on the predictors?

How well do we know $\hat{f}$

The confidence intervals of our $\hat{f}$

How well do we know $\hat{f}$?

Our confidence in f is directly connected with the confidence in βs. So for each β we can make a prediction.

How well do we know $\hat{f}$?

Here we show two difference set of predictions given the fitted coefficioents for a given subsample

How well do we know $\hat{f}$?

There is one such regression line for every imaginable sub-sample.

How well do we know $\hat{f}$?

Below we show all regression lines for a thousand of such sub-samples.

For a given x, we examine the distribution of $\hat{y}$, and determine the mean and standard deviation.

How well do we know $\hat{f}$?

Below we show all regression lines for a thousand of such sub-samples.

For a given x, we examine the distribution of $\hat{y}$, and determine the mean and standard deviation.

How well do we know $\hat{f}$?

Below we show all regression lines for a thousand of such sub-samples.

For a given x, we examine the distribution of $\hat{y}$, and determine the mean and standard deviation.

How well do we know $\hat{f}$?

For every x, we calculate the mean of the predictions, $\hat{y}$ (shown with dotted line) and the 95% CI of those predictions (shaded area).