Introduction to Statistical Learning
with applications in Python

Based on "Introduction to Statistical Learning, with applications in R" by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

Linear Regression, Part 1

linear models, least square fit, simple linear regression, parameter & model accuracy

Kurt Rinnert

Physics Without Frontiers

The Abdus Salam International Centre for Theoretical Physics

UNIVERSITY OF LIVERPOOL

Abstract

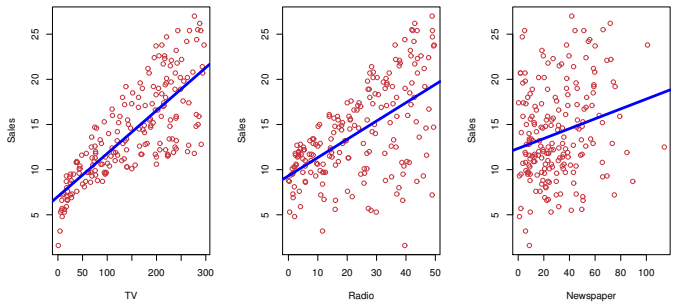
Linear models are an important topic in statistical learning.
The true relationships between predictors and responses are rarely linear. But linear models often provide reasonable approximation. They provide high interpretability and have low variance, mitigating the risk of over-fitting. Linear models can be extended to include (some) non-linear relationships.
Linear models also provide an excellent baseline to compare other models against: if our sophisticated model does not do much better than a linear model we might consider trading some bias for lower variance.

Overview

- Linear models with one predictor.
- Least squares fit.
- Accuracy of parameter estimates.
- Model Accuracy.

This will require some mathematics.

The Advertising Data Set



We want to understand how `sales` depends on `TV`, `radio` and `newspaper`.

We will use this data set to illustrate the concepts in this lecture.

Interesting Questions

1. Is there a relationship between advertising budget and sales?
2. How strong is the relationship between advertising budget and sales?
3. Which media contribute to sales?
4. How accurately can we estimate the effect of each medium on sales?
5. How accurately can we predict future sales?
6. Is the relationship linear?
7. Is there synergy among the advertising media?

Linear regression can answer all of these questions.

Simple Linear Regression

- Simple linear regression assumes an approximate simple linear relationship between one predictor and the response:

$$Y \approx \beta_0 + \beta_1 X$$

- For example, X might represent the `TV` budget and Y might represent `sales`:

$$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV}$$

- The *coefficients*, or *parameters*, β_0 and β_1 are the *intercept* and *slope* of a line.
- We can *estimate* the parameters from the training data and *predict* `sales` from the `TV` budget.

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

We use the *hat* symbol, ˆ, to denote estimates and predictions.

Estimating the Coefficients

- In practice, β_0 and β_1 are unknown.
- Given the training data

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

we want to obtain estimates of the coefficients β_0 and β_1 such that

$$y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad \forall i = 1, \ldots, n$$

- In other words we want to find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the resulting line is as close as possible to the $n = 200$ observations in the `Advertising` data set.

We need to define what we mean by "close".

Residual Sum of Squares

- Given the predictions of Y for the ith value of X

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

we define the $i$ith *residual* as

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

- The *residual sum of squares* is then

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

or equivalently

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

A least squares fit minimises the RSS.

Least Squares Fit

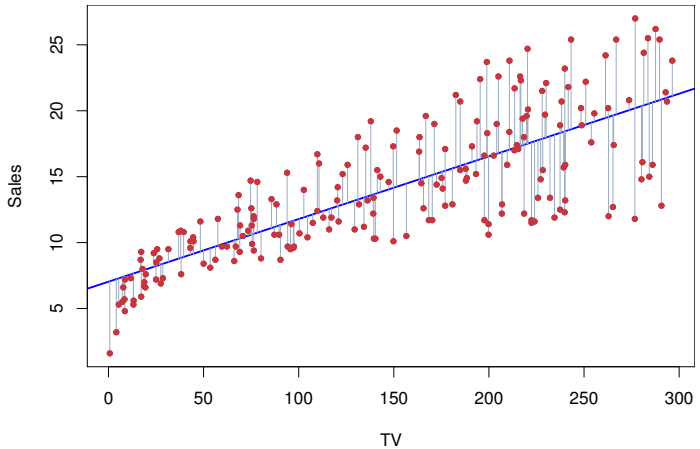
- The *least squares* approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimise the RSS.
- Using some calculus and sum manipulation we can show that the optimal parameter estimates are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

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

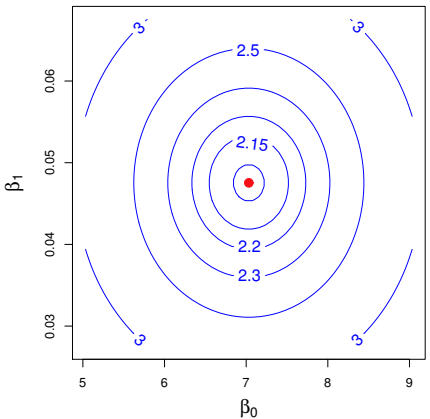
Buckle up, we are going to prove it on the blackboard.

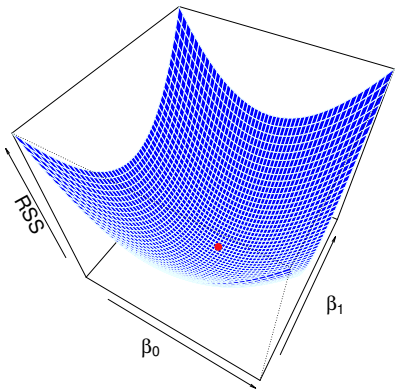
Least Squares Fit



Result of the least squares fit for the regression of `sales` onto `TV`.

Least Squares Fit





The RSS with `sales` as the response and `TV` as the predictor. The red dot marks the optimum.

Assessing Accuracies

- We want to *quantitatively* answer two rather important questions:

    1. *Is there a relationship between the predictor and the response?*
    2. *To what extend does our model fit the data?*

- To this end, we need to clarify a few concepts. In particular the notions of *population*, *sample* and *degrees of freedom*.
- Then we introduce a number of so-called *statistics*. Simply put, these are useful quantities we can compute from our data.
- It is beyond the scope of this course to formally justify all of this, but we will now spend some time to develop some intuition.

For this interlude we will look at the simpler problem of estimating a mean.

Populations

- We (somewhat sloppily) define a *population* as the *full set* of potential observations.
- That could be an actual population (as in everyday language) like, say, the population of *all* blue whales alive on Earth today.
- It can also be the full set of instances of any abstract entity.
- We then might be interested in measuring a certain property of the instances in the population.
- For example, the `length` of all blue whales alive on Earth today.
- Often we want to summarise our findings rather than presenting a raw data table (or graphical representation thereof).
- A common way to summarise the data is reporting the *mean* and the *standard error* of the quantity of interest.

Note that there is nothing wrong with reporting the full distribution!

Samples

- In practice, it is generally not possible to access the whole population.
- Quite literally, we don't have access to *all* blue whales.
- But we can look at a *sample* of the full population.
- The sample is a subset of the population we have access to.
- We generally assume the sample to be a *random sample*.
- Then we can try to *estimate* the mean and variance of the quantity of interest from the sample.
- Clearly, our estimates will *not* yield the *true* values of the mean and the variance.

Our goal is to avoid any *bias* in the estimates and evaluate their *accuracy*.

Estimating the Population Mean

- If we had access to the whole population we could calculate the *true* mean μ.
- Sadly, we have only access to a sample of n observations of the random variable Y.
- So all we can do is produce an estimate $\hat{\mu}$:

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

- This is the best possible *unbiased* estimate we can draw from a random sample.
- Intuitively, this is obvious from the fact that $\hat{\mu} = \mu$ if the sample covers the entire population.

It is *very* important that the sample is random. Beware of *selection bias*!

The Error of the Mean

- Let's assume our sample was drawn from a normal (Gaussian) distribution with *unkown* mean μ but *known* variance σ^2.
- Then the squared *standard error* of our estimate $\hat{\mu}$ is:

$$\mathrm{Var}(\hat{\mu}) = \mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n}$$

- It is then common to report our findings in the form

$$\mu = \hat{\mu} \pm \frac{\sigma}{\sqrt{n}}$$

  or give a specific *confidence interval*.

In pratice we are rarely that lucky.

Estimating the Population Variance

- More often than not, neither the population mean μ nor the variance σ^2 are kown.
- Then we have to leverage the the sample mean $\hat{\mu}$ to compute the *sample variance*

$$\hat{\sigma}^2 = \langle (y_i - \hat{\mu})^2 \rangle = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\mu})^2$$

as an estimate of the true population variance.

Unfortunately, this $\hat{\sigma}^2$ is *biased* because in general $\hat{\mu} \neq \mu$!

The Sample Variance Bias

- We have simulated a data set of blue whale `length`s.
- The sample mean is $\hat{\mu} = 26.0$.
- The sample variance is:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu})^2 = 17.37$$

- The *true* sample variance is (in general we don't have access to μ):

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2 = 23.97$$

$\hat{\mu} = 26.0$
$\mu = 22.5$
$n = 5$

observation	length
1	24.23
2	29.57
3	25.14
4	30.10
5	20.97

The sample variance is *systematically biased* to lower values.

Source of the Sample Variance Bias

- The source of the bias is that the estimate $\hat{\mu}$ does *not* have n *degrees of freedom*.
- Intuitively, this can be understood from the extreme case of $n = 1$.
- With $n = 1$ the sample variance would be zero, resulting in a zero standard error.
- Clearly it does not make sense to assume absolute confidence in $\hat{\mu}$.
- We need *at least two observations* to even start estimating the standard error of $\hat{\mu}$. (If we don't know the true μ, which is generally the case).
- We therefore apply Bessel's correction when computing the estimate $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \hat{\mu})^2$$

For large *n* this distinction is less relevant. The idea extends to more than one parameter.

The Population Regression Line

- Recall that the *true* relationship between X and Y takes the general form

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

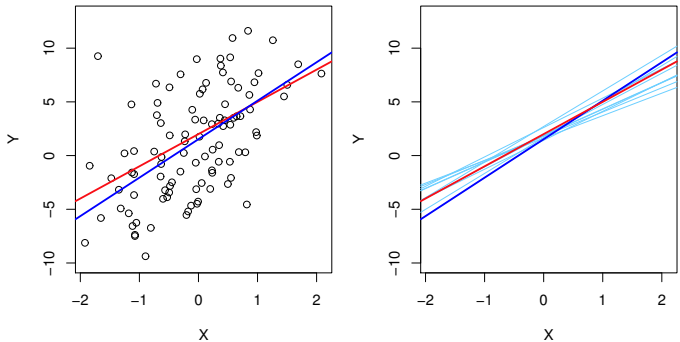
  where ϵ is a random error term with mean zero.
- If we approximate f as a linear function the relationship becomes:

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

- The error term is a catch-all for what we miss with this simplified model.
- This linear model defines the *population regression line*.
- This the best linear approximation to the true relationship between X and Y.
- It is analogous to the *population mean* we discussed earlier.

Note that the true relationship is *not* required to be linear.

The Population Regression Line



- We created a random sample from the model $Y = 2 + 3X + \epsilon$.
- The population regression line is shown in red.
- *Sample regression lines* are shown in blue.

The mean over many samples is close to the population regression line.

Accuracy of the Coefficient Estimates

- The *sample regression line* is given by

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

  with the estimated coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$.
- The associated standard errors are then (this is a bit tedious to show):

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

- This assumes the ϵ_i are uncorrelated with variance σ^2.

In general the population variance $\boldsymbol{\sigma^2}$ is not known.

Estimating the Population Variance

- If we don't know the population variance σ^2 we need to estimate it.
- This works like in the simple case of estimating a population mean.
- The estimated $\hat{\sigma}^2$, or *residual standard error*, is

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

with the *residual sum of squares*

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

Note that number of degrees of freedom is now $n-2$!

Confidence Intervals

- Given the standard errors for the coefficient estimate, we can compute *confidence intervals*.

  A 95% confidence interval is a range of values such that the interval will contain the true parameter value with a probability of 95%.

- This is a somewhat sloppy but very practical definition.
- In simple linear regression the 95% confidence interval for the parameters β is approximately:

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

- That is, with 95% probability the true value β lies in the interval

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

This assumes Gaussian errors and large *n*. In practice we use computers to get it right.

Hypothesis Testing

- We are interested in whether X is related to Y.
- This question is answered by a *hypothesis test*.
- Most commonly, we want to distinguish the *null hypothesis*

$$H_0 : \beta_1 = 0$$

from the *alternative hypothesis*

$$H_a : \beta_1 \neq 0$$

- If $\beta_1 = 0$ then $Y = \beta_0 + \epsilon$ and there is no relationship between X and Y.
- If $|\beta_1|$ is large, we can conclude that there is a relationship.

We need to quantify what we mean by "large" in this context.

The t-statistic

- Clearly, simply checking whether $\hat{\beta}_1 \neq 0$ is not good enough.
- The answer must depend on our confidence in the estimated value.
- This confidence depends on the value of the standard error $\text{SE}(\hat{\beta}_1)$.
- The *t-statistic* weighs the deviation of $\hat{\beta}_1$ from zero by the standard error:

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

- If there is *truly* no relationship between X and Y, this quantity follows a t-distribution with $n - 2$ degrees of freedom.

The t-distribution approaches the normal distribution for large n.

The *p*-value

- We can compute the probability to find a value of $|t|$ or larger from the t-distribution, assuming that there is truly no relationship between X and Y.
- This probability is called the *p-value*.
- A high p-value indicates that we should accept the null hypothesis.
- A low p-value indicates we should reject the null hypothesis.
- Common values for rejecting the null hypothesis are $p < 0.05$ and $p < 0.01$.

Some disciplines require *much* lower *p*-values to reject the null hypothesis.

Example: Advertising Data Set

Regression of `sales` onto `TV`

	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	7.0325	0.4578	15.36	< 0.0001
`TV`	0.0475	0.0027	17.62	< 0.0001

- We can reject $H_0 : \beta_0 = 0$ (that is, the mean of Y is not zero).
- We can reject $H_0 : \beta_1 = 0$ (that is, `TV` does influence `sales`).

The probabilities of observing these values if H_0 were true is virtually zero.

Model Accuracy

- We now want to quantify the *extend to which the model fits the data*.
- For simple linear regression we use two quantities to do this.

1. The residual standard error, RSE.
2. The R^2 statistic.

- The RSE can be considered a measure of *lack of fit*.
- The R^2 statistic can be considered a normalised measure of the quality of fit.
- R^2 always takes on values between 0 and 1. It is independent of the scale of Y.

The R^2 statistic represents the *proportion of variance explained*.

The R^2 Statistic

- The R^2 statistic is defined as

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

  with the *total sum of squares*

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- We can think of the TSS as the total variance of the response Y before the regression is performed.
- An R^2 close to zero indicates the regression does not explain much of the variability.
- An R^2 close to one indicates the regression explains a lot of the variability.

R^2 measures the proportion of variability in Y that can be explained by X.

Example: Advertising Data Set

Regression of `sales` onto `TV`

Residual standard error	3.26
R^2 statistic	0.612

- On average, `sales` deviate by 3,260 units from the regression line.
- The R^2 indicates that just under two thirds of the variability in `sales` is explained by the regression onto `TV`.

The R^2 tells us that the simple linear regression model is reasonable.

DID THE SUN JUST EXPLODE?

(IT'S NIGHT, SO WE'RE NOT SURE.)