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 2

multiple linear regression, hypothesis testing, important variables, fit quality, predictions

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 multiple predictors.
- Hypothesis testing, the F-statistic.
- Choosing important variables.
- Assessing the fit & prediction quality.

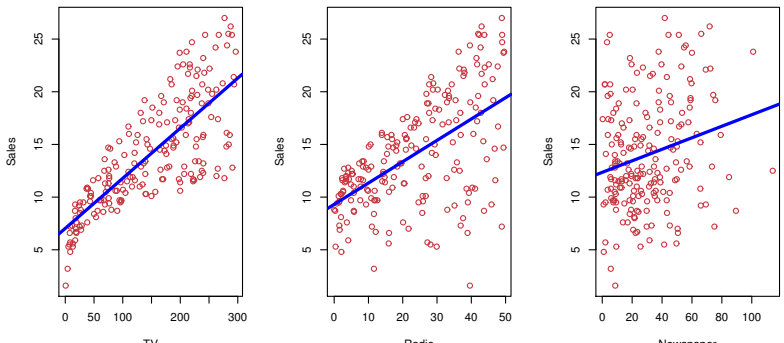
This will require some mathematics.

The Scenario

- We look at the `Advertising` data set again.
- Previously, we only used the `TV` predictor to predict `sales`.
- The data set provides two more predictors, `radio` and `newspaper`.
- We are now interested in using all available information in the data set.
- We can think of two ways of doing that:
    1. Perform simple linear regression on the predictors separately.
    2. Somehow combine all predictors.

As you might guess, option two is more interesting, but let's look at option one first.

Separate Simple Linear Regression



Separate simple linear regression looks promising on this data set.

Separate Simple Linear Regression

Regression of `sales` onto `radio`

	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	9.312	0.563	16.54	< 0.0001
`radio`	0.203	0.020	9.92	< 0.0001

Regression of `sales` onto `newspaper`

	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	12.351	0.621	19.88	< 0.0001
`newspaper`	0.055	0.017	3.30	0.00115

All the separate regressions look reasonable.

Separate Simple Linear Regression

- This is not entirely satisfactory for several reasons.
- It is unclear how could predict a single value of `sales` from the separate models.
- Each of the separate models ignores the other variables.
- This is problematic if there are correlations between the predictors.

A better approach is to use multiple predictors simultaneously.

Multiple Linear Regression

- *Multiple linear regression* uses multiple predictors.
- Each predictor has its own slope coefficient.
- For p predictors the model takes the form

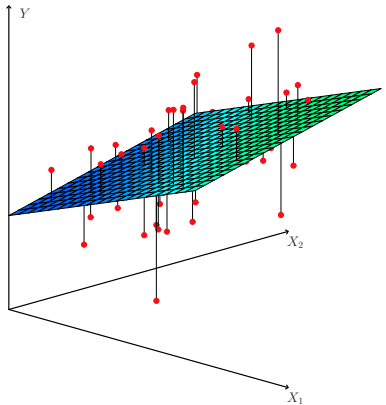
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

  where the β_j are interpreted as the average effect X_j has on Y while *holding all other predictors constant*.
- For example:

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

Now we have one model using all the information.

Example: Two Predictors



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

In general, the result is a *hyperplane*.

Estimating the Regression Coefficients

- As in simple linear regression we want to make predictions using estimated parameters:

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

- We choose the coefficients $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ to minimise the sum of squared residuals:

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2 - \cdots - \hat{\beta}_p x_p)^2 \end{aligned}$$

Deriving the solution involves some matrix manipulations.

Estimating the Regression Coefficients

- Recall the notation for the *feature matrix*:

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \in \mathbb{R}^{n \times p}$$

- We define the vectors:

$$\begin{aligned} \boldsymbol{\beta_0} &= (\beta_0, \beta_0, \dots, \beta_0)^T && \in \mathbb{R}^n \\ \beta &= (\beta_1, \beta_2, \dots, \beta_p)^T && \in \mathbb{R}^p \\ \boldsymbol{\epsilon} &= (\epsilon_1, \epsilon_2, \dots, \epsilon_n)^T && \in \mathbb{R}^n \end{aligned}$$

- With this and the convention for the response vector $\boldsymbol{y} \in \mathbb{R}^n$ we can write the model as

$$\boldsymbol{y} = \boldsymbol{\beta_0} + \boldsymbol{X}\beta + \boldsymbol{\epsilon}$$

These are $\boldsymbol{n}$ simultaneous equations describing the linear model.

Estimating the Regression Coefficients

- We now redefine $\boldsymbol{X}$ to obtain the model's *design matrix*:

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \in \mathbb{R}^{n \times (p+1)}$$

- And the vector β to include the intercept:

$$\beta = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)^T \in \mathbb{R}^{p+1}$$

- We can now write the model in the compact form

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}$$

The linear model can be extended by adding more columns to the design matrix.

Estimating the Regression Coefficients

- We now want to find the estimated coefficient vector $\hat{\beta}$ to make predictions

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta}$$

by minimising the residual sum of squares

$$\begin{aligned} \text{RSS} &= \boldsymbol{e}^2 \\ &= (\boldsymbol{y} - \hat{\boldsymbol{y}})^2 \\ &= (\boldsymbol{y} - \hat{\boldsymbol{y}})^T(\boldsymbol{y} - \hat{\boldsymbol{y}}) \\ &= (\boldsymbol{y} - \boldsymbol{X}\hat{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}) \end{aligned}$$

In this matrix notation the RSS is the squared norm of the residuals vector $\boldsymbol{e}$.

Estimating the Regression Coefficients

- The procedure is the same as in the case of simple linear regression.
- But now the derivative becomes the *gradient*:

$$\frac{\partial \mathrm{RSS}(\beta)}{\partial \beta} = \nabla_\beta \mathrm{RSS}(\beta) = \nabla_\beta (\boldsymbol{y} - \boldsymbol{X}\beta)^T (\boldsymbol{y} - \boldsymbol{X}\beta)$$

- The minimising condition then is:

$$\nabla_\beta (\boldsymbol{y} - \boldsymbol{X}\beta)^T (\boldsymbol{y} - \boldsymbol{X}\beta)\Big|_{\beta=\hat{\beta}} = 0$$

- And the minimising solution is:

$$\hat{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

We will show the details on the blackboard.

Estimating the Regression Coefficients

- The solution

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

  makes the assumption that $\boldsymbol{X}^T\boldsymbol{X}$ can be inverted.
- We can insert the solution into the prediction equation:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

- The matrix $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is called the *hat matrix* (it puts the hat, ˆ , on $\boldsymbol{y}$).
- The hat matrix is also known as the *leverage matrix* (more on that later).

In practice we determine the components of $\hat{\boldsymbol{\beta}}$ from data using a computer.

Example: Advertising

- We fitted the model

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

and obtained the following results:

	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

Note that `newspaper` doesn't seem to have an influence on sales anymore.

Example: Advertising

- Where does the apparent contradiction between multiple and separate linear regression come from?
- The answer lies in the correlation between the `radio` and `newspaper` budgets:

	TV	radio	newspaper	sales
TV	1.0	0.0548	0.0567	0.7822
radio		1.0	0.3541	0.5762
newspaper			1.0	0.2281
sales				1.0

The `newspaper` predictor doesn't add much information because it is correlated with `radio`.

Some Important Questions

1. Is at least one of the predictors useful in predicting the response?
2. Do *all* of the predictors help to explain Y?
3. How well does the model fit the data?
4. How accurate are our predictions?

We will address each question in turn.

Relationship Between Response and Predictors

- In the simple linear regression model we simply asked whether β_1 is zero.
- In multiple linear regression with p predictors our null hypothesis becomes:

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

- And the alternative hypothesis becomes:

$$H_a : \text{at least one } \beta_j \text{ is not zero.}$$

- This hypothesis test is performed using the F-statistic:

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

If H_a is true, we expect F to be greater than one, and close to one otherwise.

Example: Advertising

- Fitting the model

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

we find:

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

There is strong evidence that at least one predictor influences `sales`.

Relation of a Subset of the Predictors to the Response

- We are often interested in whether a specific subset of the predictors is related to the response:

$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0$$

  where we have shuffled the q excluded predictors to the front of the predictor list.

- The F-statistic for this hypothesis test is computed as follows:

$$F = \frac{(\text{TSS} - \text{RSS})/q}{\text{RSS}/(n-p-1)}$$

The *t*-statistics are exactly equivalent to the *F*-statistics with one variable left out.

Why not use the Individual *p*-values?

- One might think that we don't need the F-statistic.
- After all we can easily compute the t-statistic p-value for each predictor.
- So why can't we reject

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

  from the individual p-values?
- The answer depends on the cutoff on the p-value we use to reject the null hypothesis and the number of parameters.
- For example, if we choose $p < 0.05$ and the number of predictors is 20, it is very likely that we would reject the null hypothesis for at least one predictor.

This is related to the dimensionality problem we'll discuss soon.

Which Predictors are Relevant?

- Once we have established that some (at least one) predictors are relevant we would like to know *which* ones.
- There are some established methods to answer this question.

1. *Forward Selection*: start with only the intercept and add predictors with the lowest RSS. Proceed until some cutoff is reached.
2. *Backward Selection*: Start with all variables and keep removing the ones with the largest p-values. Stop when a cutoff is reached.
3. *Mixed Selection*: Like forward selection, but remove variables that acquire large p-values after another variable is added.

Forward selection is greedy. Backward selection requires $p < n$.

Model Fit

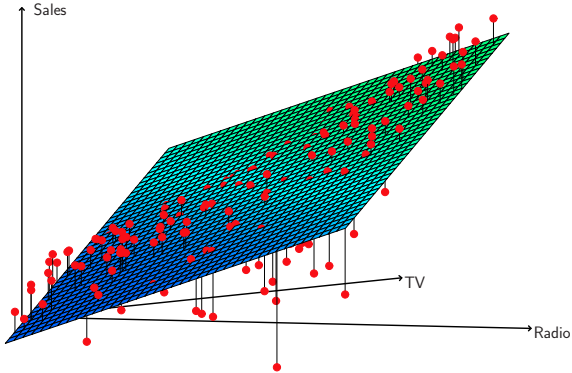
- The main measures of the fit quality are the RSE and R^2.
- The R^2 is computed as in the simple case and the RSE is:

$$\text{RSE} = \sqrt{\frac{1}{n-p-1}\text{RSS}}$$

	R^2	RSE
TV	0.61200	3.260
TV + radio	0.89719	1.681
TV + radio + newspaper	0.89720	1.686

The table confirms our previous finding that `newspaper` does not do much good.

Model Fit



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

Visualisation is our friend.

Prediction Accuracy

- The coefficient estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ are only estimates of for $\beta_0, \beta_1, \ldots, \beta_p$.
- That is, the *least squares plane*

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$$

  is only an estimate to the *population regression plane*

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

- The inaccuracies of these estimates are related to the *reducible errors*.

We can compute *confidence intervals* to quantify the inaccuracies.

Prediction Accuracy

- In practice, assuming a linear model for $f(X)$ is always an approximation.
- This creates another source of reducible error called *model bias*.
- This bias is present even if we knew the population regression parameters.
- Most of the time we act *as if* the model was correct in terms of how we judge the fit quality and so on.

We will learn about ways to improve on this situation.

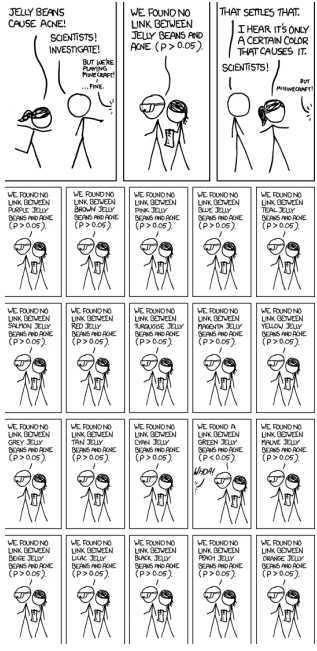
Prediction Accuracy

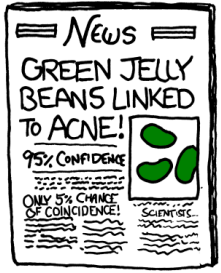
- Even if we knew the true $f(X)$, that is we know the parameters

$$\beta_0, \beta_1, \beta_2, \ldots, \beta_p$$

  we could not make perfect predictions.
- This is due to the *irreducible error* stemming from the error term ϵ.
- The question then is how much do the predictions $\hat{Y}$ deviate from Y?

We use *prediction intervals* to quantify this.


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE RUNNING A PHANTOGRAPH!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MYM-WECHANT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...