Ch. 3. Linear Regression

Assumptions

Four main assumptions:

- **linearity and additivity of the relationship** between dependent and independent variables
- **statistical independence of the errors** (in particular, no correlation between consecutive errors in the case of time series data)
- **homoscedasticity** (constant variance) of the errors
- **normality** of the error distribution

Introduction

LR is a useful tool for predicting a quantitative response. Many statistical learning approaches can be seen as generalizations or extensions of LR.

Questions that we can answer using LR:

1. Is there a relationship between our target variable and our explanatory variable(s)?
2. How strong is this relationship?
3. Which explanatory variables contribute to the value of our target variable?
4. How accurately can we estimate the effect of each explanatory variable on the target variable?
5. How accurately can we predict the target variable for future observations?
6. Is the relationship linear?
7. Is there synergy between explanatory variables?

Simple Linear Regression

Simple LR is a straightforward approach for predicting a quantitative response Y on the basis of a single predictor variable X. It assumes that there is approximately a linear relationship between X and Y.

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

We are *regressing* Y on X (or Y onto X).

β_0 and β_1 are two unknown constants that represent the *intercept* and the *slope* terms in the linear model. Together they are known as the model *coefficients* or *parameters*. Once we have used our training data to produce *estimates* $\widehat{\beta}_0$ and $\widehat{\beta}_1$ for the model coefficients, we can predict future sales on the basis of a particular value x by computing:

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ *least squares line*

where $\hat{y}$ indicates a prediction of Y on the basis of X = x.

Estimating the Coefficients

In practice the real value of the coefficients is unknown, we use data to get an estimate. Our goal is to obtain coefficient estimates such that the linear model fits the available data well. In other words we want to find an intercept and a slope such that the resulting line is as *close* as possible to our data points. There are a number of ways of measuring *closeness*. The most common approach involves minimizing the *least squares* criterion.

*i*th residual → difference between *i*th observed response value and the *i*th response value predicted by our model:

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

We define the Residual Sum of Squares as the squared sum of all residuals:

$RSS = e_1^2 + ... + e_n^2$

or equivalently:

$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + ... + (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$

The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. Using some calculus, one can show that the minimizers are:

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

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

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

Assessing the Accuracy of the Coefficient estimates

We assume that the *true relationship* between X and Y takes the form $Y = f(X) + \varepsilon$ for some unknown function f, where ε is a mean-zero random error term. If f is to be approximated by a linear function, then we can write this relationship as

$Y = \beta_0 + \beta_1 X + \varepsilon$ *population regression line*

The error term is a catch-all for what we miss with this simple model: the true relationship is probably not linear, there may be other variables that cause variation in Y, and there may be measurement error. We typically assume that the error term is independent of X.
This model defines the *population regression line*, which is the best linear approximation possible to the true relationship between X and Y. The assumption of linearity is often a useful working model. However, we seldom believe that the true relationship is linear.

The least square regression coefficient estimates characterize the *least squares line*.
The *true relationship* is generally not known for real data, but the *least squares line* can always be computed using the coefficient estimates.

In the real world we have access to a series of observations from which we can compute the least squares line; however, the population regression line is unobserved.
Different data sets(samples) generated from the same true model will result in slightly different least squares lines, but the unobserved population regression line does not change.

The coefficient estimates are unbiased: they don't systematically under or over estimate the real values and, given a huge number of sets of observations and they will exactly equal the real values.
If we estimate our coefficients from a particular data set, their values will not be exactly equal to the true values. But if we could average the estimates obtained from a huge number of data sets, then the average of these estimates would be spot on.

How accurate are our estimates of the coefficients?
The variance of the population mean of a random variable Y is the square of its standard error: $Var(\hat{\mu}) = SE(\hat{\mu})^2 = \frac{\sigma^2}{n}$
Where sigma is the standard deviation of each of the realizations y_i of Y (providing the n observations are uncorrelated).
Roughly speaking, the Standard Error tells us the average amount that this estimate differs from its actual value. The equation tells us also how this deviation shrinks with n, the more observations we have, the smaller the standard error of the estimator.

In a similar vein, we can wonder how close our estimates of the coefficients to the real values. To compute their Standard Errors we use the following formulas:

$$SE(\hat{\beta}_0)^2 = \sigma^2 \frac{1}{n} \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where σ^2 is the variance of the noise. For these formulas to be valid we need to assume that the errors for each observation are uncorrelated with common variance σ^2.
In general, σ^2 is not known, but can be estimated from the data. The estimate of σ^2 is known as the *residual standard error*, and is given by the formula

$$RSE = \sqrt{RSS/(n-2)}$$

Strictly speaking, when σ^2 is estimated from the data we should write $\widehat{SE}(\hat{\beta}_1)^2$ to indicate that an estimate has been made, for simplicity of notation the extra "hat" will be dropped.

Standard Errors can be used to compute *confidence intervals*.

A 95% CI is defined as the range of values such that, with 95% probability, the range will contain the *true unknown value* of the parameter. The range is defined in terms of lower and upper limits computed from the sample of data.
For LR, the $(1 - \frac{\alpha}{2})x100\%$ confidence intervals for and take the form of:
$\hat{\beta}_1 \pm z_{1-\frac{\alpha}{2}} \cdot SE(\hat{\beta}_1)$ and $\hat{\beta}_0 \pm z_{1-\frac{\alpha}{2}} \cdot SE(\hat{\beta}_0)$

Where $z_{1-\frac{\alpha}{2}}$ is the $1 - \frac{\alpha}{2}$ quantile of a t-distribution with n-2 degrees of freedom. This two equations rely on the assumption that the errors are gaussian.

Standard Errors can also be used to perform *hypothesis tests* on the coefficients.

The most common hypothesis test involves testing the *null hypothesis* of

H_0 : there is no relationship between X and Y

versus the *alternative hypothesis*

H_a : There is some relationship between X and Y

Mathematically, this corresponds to testing $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$ since if $\beta_1 = 0$ then the model reduces to $Y = \beta_0 + \varepsilon$ and X is not associated with Y.
To test the null hypothesis we need to determine whether $\widehat{\beta_1}$, our estimate for β_1,is sufficiently far from zero that we can be confident that is non-zero. How far is enough? It depends on the accuracy of our estimations → depends on the Standard Error of our estimator. If $SE(\widehat{\beta_1})$ is small, then even relatively small values may provide strong evidence that $\beta_1 \neq 0$ and hence that there is a relationship between X and Y. In contrast, if $SE(\widehat{\beta_1})$ is large, then β_1 must be large in absolute value in order for us to reject the null hypothesis.
In practice we compute a t-statistic, given by

$$t = \frac{\widehat{\beta_1} - 0}{SE(\widehat{\beta_1})}$$

which measures the number of standard deviations that $\widehat{\beta_1}$ is away from 0.
If $\beta_1 = 0$ (there is no relationship between X and Y, the true value of the coefficient is 0), then the sampling distribution of t is a t-student with n-2 degrees of freedom and therefore we can easily compute the probability of observing any number **equal or larger** to the **absolute value of t**, assuming $\beta_1 = 0 \rightarrow P(x \geq |t| \mid \beta_1 = 0)$.
Given that $\beta_1 = 0$, if the probability of finding a value more extreme than $|t|$ (meaning that it's on the tails) is lower than the alpha level, we reject the null hypothesis.
The **alpha level** is the bigger probability we can accept that we reject the null hypothesis by mistake.
E.g. alpha = 0.05 → we accept that we could reject H_0 when H_0 is true with probability = at most 5%.

Clearly one should choose the alpha based on the problem: sometimes we can accept a bigger chance of getting to the wrong conclusion (permissive alpha), sometimes a mistake costs a lot so we try to avoid it at all costs (strict alpha).

This probability is called the **p-value** and it is roughly interpreted as follows: a small p-value indicates that it's unlikely (small probability) to observe such a substantial association between the predictor and the response due to chance, in the absence of any real association between the predictor and the response.

small p-value → we can infer that there is an association between the predictor and the response. We *reject the null hypothesis* - that is, we declare a relationship exists between X and Y if the p-value is small enough.

Assessing the Accuracy of the Model

Once we have rejected the null hypothesis in favor of the alternative hypothesis, we want to quantify the extent to which the model fits the data. The quality of a linear regression fit is typically assessed using two related quantities: the *residual standard error* (RSE) and the R^2 statistic.

Residual Standard Error

Remember that we defined the Residual Sum of Squares (RSS) as the squared sum of all residuals:

$RSS = e_1^2 + ... + e_n^2$ or $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

and the Residual Standard Error as:

$RSE = \sqrt{RSS/(n-2)}$

The RSE is an estimate of the standard deviation of . Roughly speaking, it is the average amount that the response will deviate from the true regression line. It is considered a measure of the *lack of fit* of the model to the data. Small → model fits well, Large → model does not fit well.

R^2 statistic

Since the RSE is measured in the units of Y, it is not always clear what constitutes a good RSE. The R^2 statistic provides an alternative measure of fit. It takes the form of a proportion -- the proportion of variance explained -- and so it always ranges from 0 to 1.

it is computed as $R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$

where TSS is the Total Sum of Squares $\sum_{i=1}^{n}(y_i - \bar{y})^2$. TSS measures the total variance in the response Y and can be thought as the amount of variability inherent in the response before the regression is performed. In contrast, RSS measures the amount of variability that is left unexplained after performing the regression.
Hence TSS - RSS measures the amount of variability in the response that is explained (or removed) by performing the regression, and R^2 measures the proportion of variability in Y that can be explained using X.
R^2 close to 1 → almost all the variability of Y is explained by our linear model
R^2 close to 0 → the regression did not explain much of the variability in the response; this might occur because the linear model is wrong, or the inherent error σ^2 is high, or both.

Even if the range is fixed [0,1], it can still be hard to define a good R^2 and in general it depends on the application.

In the simple linear regression setting, $R^2 = r^2$ where r is an estimate of Cor(X,Y). In the multiple linear regression setting R^2 fills the role of Cor as a measure of correlation between the predictors and the response since Cor can be only computed pairwise.

Multiple Linear Regression

In practice, we often have more than one predictor. How can we extend our analysis to accommodate additional predictors? One option is to run several separate simple linear regressions, each with a different variable as predictor. However, this approach is not satisfactory.

1. it's unclear how to make a single prediction given the levels of the several predictors, since each is associated with a different regression equation.
2. each of the several regression equations ignores the remaining others in forming estimates for the regression coefficients.

Instead of fitting separate simple linear regression models for each predictor, a better approach is to extend the simple linear regression model so that it can directly accommodate multiple predictors. We will give each predictor a separate slope coefficient in a single model:
suppose *p* predictors, the model will be:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + + \beta_p X_p + \varepsilon$$

Where X_j represents the j^{th} predictor and β_j quantifies the association between that variable and the response. We interpret β_j as the *average* effect on Y of a one unit increase in X_j, *holding all other predictors fixed*.

The coefficients can be estimated using the least squares method analogously as in the simple linear case.

note: the simple and multiple regression coefficients can be quite different.

When we perform Multiple Linear Regression, we usually are interested in answering a few important questions.

1. Is <u>at least one</u> of the predictors useful in predicting the response?
2. Do all the predictors help to explain Y, <u>or is only a subset of the predictors useful</u>?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

1.
Same setting as with the Simple Linear Regression but H_0 now is "ALL coefficients are equal to 0" while H_a is "**at least one** coefficient different from zero".

This hypothesis test is performed by computing the F-statistic: $F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$
If the linear model assumptions are correct, one can show that
$E\{RSS/(n-p-1)\} = \sigma^2$ and that, provided H_0 is true, $E\{TSS - RSS/p\} = \sigma^2$
Hence, if there is no relationship between the response and predictors (accept H_0), one would expect the F-statistic to take on a value close to 1.
On the other hand, if H_0 is rejected, then $E\{TSS - RSS/p\} > \sigma^2$, so we expect F to be greater than 1.
How larger than 1 an F-statistic has to be for us to reject H_0? It depends on n and p. When n is large an F-statistic even slightly larger than 1 might still provide evidence against H_0.

When H_0 is true and the errors have normal distribution, the F-statistic follows an F-distribution. For any given value of n and p, any statistical software package can be used to compute the p-value associated with the F-statistic using this distribution. Based on this p-value, we can determine whether or not to reject H_0.

t-statistic and p-values for each individual predictor in Multiple Linear Regression provide information whether each individual predictor is related to the response, after adjusting for the other predictors. They report the *partial effect* of adding that variable to the model (e.g. adding the predictor X_3 to the model composed by predictors X_1 and X_2).

Given these individual p-values for each variable, why do we need to look at the overall F-statistic? It seems likely that if any one of the p-values for the individual variables is very small, then *at least one of the predictors is related to the response*. However, this logic is flawed, especially when the number of predictors is large.

example: p = 100 and H_0 is true (at least one coefficient is not 0) with alpha = 0.05. In this situation, about 5% of the p-values associated to each predictor will be below 0.05 by chance. In other words we expect to see approximately (0.05 x 100 =)**5** *small* p-values **even** in the absence of any true association between the predictors and the response. In fact, we are almost guaranteed to observe a p-value < 0.05 by chance. **High chance of making a mistake if we use individual t-statistics and p-values!**
However, the F-statistic does not suffer from this problem because it adjusts for the number of predictors.

2. p(78-79)
forward selection
backward selection
mixed selection

3.

Two of the most common numerical measures of model fit are the RSE and R^2, the fraction of variance explained. In multiple linear regression R^2 equals , the square of the correlation between the response and the fitted linear model.
R^2 will always increase when more variables are added to the model, even if those variables are only weakly associated with the response. This is due to the fact that adding another variable to the least squares estimation must allow us to fit the training data (though not necessarily the testing data) more accurately. Thus the R^2 statistic, which is also computed on the training data, must increase.

4.
Once we have fit the multiple regression model, it is straightforward to apply in order to predict the response Y on the basis of a set of values for the predictors $X_1, \dots, X_p$. However, there are three sorts of uncertainty associated with this prediction.

1. The coefficient estimates, are estimators for the true value of the coeff. That is, the *least squares plane* $\widehat{Y}$ is only an estimate for the *true population regression plane* $f(X)$. The inaccuracy in the coefficient estimates is related to the *reducible error*. We can compute a ***confidence interval*** in order to determine how close $\widehat{Y}$ will be to $f(X)$
2. In practice, assuming a linear model for $f(X)$ is almost always an approximation of reality, so there is an additional source of potentially reducible error which we call *model bias*. So when we use a linear model, we are n fact estimating the best linear approximation to the true surface. However, we will ignore this discrepancy. and operate as if the linear model were correct.
3. Even if we knew $f(X)$ -- that is, even if we knew the real values for the coefficients -- the response value cannot be predicted perfectly because of the random error ε in the model. How much Y will vary from $\widehat{Y}$? We use ***prediction intervals*** to answer this question. Prediction intervals are always wider than confidence intervals, because they incorporate both the error in the estimate for $f(X)$ (the reducible error) and the uncertainty as to how much an individual point will differ from the population regression plane (the irreducible error).

Confidence interval → quantifies the uncertainty surrounding the *average* response value over a large number of observations.

Prediction interval → quantifies the uncertainty surrounding the response value for a *particular* observation.

Qualitative predictors

p(82-86)
Use dummy(indicator) variables

Extensions of the Linear Model

p(86-92)
Interaction terms

hierarchical principle: *if we include an interaction term, we should also include the main effects(of the interaction term) even if the p-values associated with their coefficients are not significant*
non linear relationships

Potential problems

When we fit a linear regression model to a particular data set, many problems may occur. Most common among these are the following:

1. Non-linearity of the response-predictor relationships
2. Correlation of error terms
3. Non-constant variance of error terms
4. Outliers
5. High-leverage points
6. Collinearity between the predictors

Regression Diagnostic Plots Explained

1. Non linearity of the Data

If the true relationship is far from linear, then virtually all of the conclusions that we draw from the fit are suspect. In addition, the prediction accuracy of the model can be significantly reduced.
To detect use *Residual plots* → there should not be a detectable pattern

2. Correlation of error terms

e.g. time series data or predicting a person height from his weight and some observations are from the same family or eat the same diet.

3. Non-constant variance of error terms

if we have an idea of the variance of each response we can use a weighted least squares to estimate our coefficients.
To detect use *Residual plots* → the spread of the residuals should be ~ constant for all fitted values, no funnel shape.

4. Outliers

Outliers are points for which y_i is far from the value predicted by the model for a variety of reasons, such as the incorrect recording of an observation during data collection.
They affect the RSE, used to compute all confidence intervals and p-values, affecting the interpretation of the fit.
If they have also high-leverage, they can drastically alter the fit.

To detect we may use *Residual plots* → outliers are clearly visible outside the rest of the points.

Using residuals it's difficult to decide a cutoff → usage of studentized residuals (residuals divided by their estimated standard error). Observations whose studentized residuals are greater than 3 in absolute value are possible outliers.
If we believe that an outlier has occurred due to an error in data collection or recording, then one solution is to simply remove the observation.
However, care should be taken, since an outlier may instead indicate a deficiency with the model, such as a missing predictor.

5. high leverage points

High leverage points are values with an unusual value for x_i.
High leverage observations tend to have a sizable impact on the estimated regression line. It is cause for concern if the least squares line is heavily affected by just a couple of observations, because any problems with these points may invalidate the entire fit. For this reason, it is important to identify high leverage observations.

In a simple linear regression they are easy to identify → just look for observations for which the only predictor value x is outside of the normal range of each individual predictor values.

In a multiple linear regression with many predictors, it is possible to have an observation that is well within the range of each individual predictor values, but that is unusual in terms of the full set of predictors → hard to notice if # of predictors is > 2, dimensionality is too high for plotting.

In order to quantify an observation's leverage, we compute the *leverage statistic*.

Large value → high leverage observation.
Range is between 1/n and 1, the average leverage for all the observations is always equal to (p+1)/n. If a given observation has a leverage statistic that greatly exceeds (p+1)/n, then we may suspect that the corresponding point has high leverage.

6. Collinearity

Collinearity refers to the situation in which two or more predictor variables are closely related to one another.
The presence of collinearity can pose problems in the regression context,, since it can be difficult to separate out the individual effects of collinear variables on the response → estimation of coefficients will be off.
Since collinearity reduces the accuracy of the estimates of the regression coefficients, it causes the standard error for the coefficient estimators to grow. ⇒ collinearity results in a decline in the t-statistic.
As a result, in the presence of collinearity we may fail to reject H_0 even if H_0 is wrong.
This means that the *power* of the hypothesis test -- the prob. of correctly detecting a *non-zero* coefficient -- is reduced by collinearity.

Simple way to detect collinearity → look at the correlation matrix of the predictors.
Unfortunately this can only detect pairwise collinearity.
Collinearity can exist between three or more variables even if no pair of variables has a particularly high correlation → *multicollinearity*.

A better way to assess multicollinearity is to compute the *variance inflation factor*.
The VIF is the ratio of the variance of when fitting the full model divided by the variance if fit on its own. Smallest possible value for VIF is 1 → complete absence of collinearity.

Rule of thumb: a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity.

When faced with the problem of collinearity, there are two simple solutions:

1. Drop one of the problematic variables from the regression (no great consequence since the presence of collinearity implies that the info provided by this variable about the response is redundant in the presence of other variables)
2. Combine the collinear variables together into a single, more meaningful, predictor.

Ch. 4. Classification

odds log odds

Multivariate Gaussian