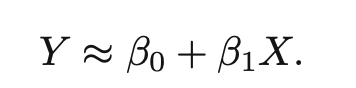
3. Linear Regression

Linear regression: Useful for predicting a quantitative response

1. Determine whether the data provide evidence of an association.
2. Determine the strength of the relationship(strong~Weak)
3. Distinguish the effects of each factors
4. Find a way to separate out the individual effects of each medium
5. How accurate can we estimate the effects
6. How accurate can we predict future
7. Is the relationship linear🡪linear regression is an appropriate tool
8. Is there synergy among the factors🡪in stats, called interaction effect

3.1 Simple Linear Regression

Simple Linear Regression: A very straightforward approach for predicting a quantitative response Y on the basis of a single predictor variable X.



1. “Approximately models as”
2. : Intercept term
3. : slop
4. :coefficients or parameters.

A picture containing object, clock

Description automatically generated

1. Hat symbol: , denite the estimated value for an unknown parameter or coefficient. Denote the predicted value of the response.

3.1.1 Estimating the Coefficients

Goal is to find an intercept such that the resulting line is as close as possible

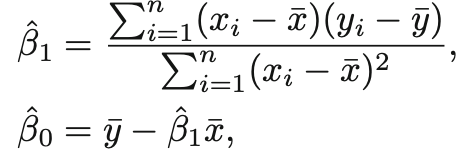
Residual Sum of squares(RSS)



Minimize the difference between the ith observed value and the ith response value that is predicted by the linear model.



Least Squares Coefficient Estimates:

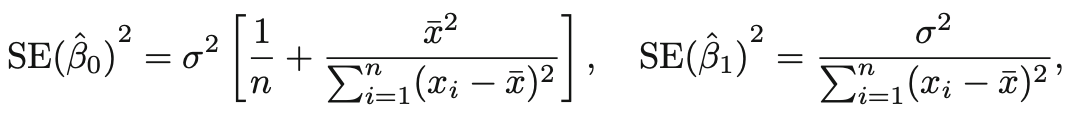


3.1.2 Assessing the Accuracy of the Coefficient Estimates



1. The expected value Y when X=0 and is the slope—the average increase in Y associated with a one-unit increase in X
2. Error term is assumed to be independent of X
3. The true relationship is generally not known for real data, but the least squares can always be computed using coefficient estimate.
4. The unknown coefficients in linear regression define the population regression line, the coefficient estimates define the least squares line.
5. Property of Unbiased estimator: Unbiased estimator does not systematically over- or under-estimate the true parameter.
   1. Property of unbiasedness holds for the least squares coefficient estimates.

Stanadard Errors of the Coefficients:

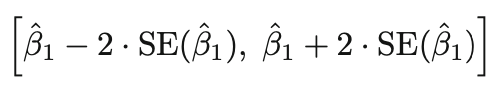




1. The errors for each observation are uncorrelated with common variance, .
2. From the second equation, error is smaller when x is more spread out
3. Residual standard error: estimate of .



1. 95% confidence interval:



1. The approximately a 95% chance the interval will contain the true value of .
2. Hypothesis test:
   1. Null hypothesis: : No relationship between X and Y,
   2. Alternative hypothesis: : Some relationship between X and Y,
   3. Determine whether is sufficiently different from zero
   4. In practice, use t-statistics:

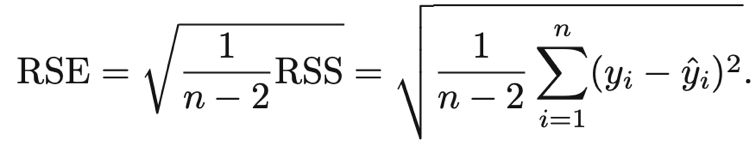
A picture containing object, clock

Description automatically generated

* 1. P-Value: A small p-value indicates that it is unlikely to observe substantial association between the predictor and the response due to chance

3.1.3 Assessing the Accuracy of the Model

Residual Standard Error: An estimate of the standard deviation of .



RSE is a measure of the lack of fit of the model: if the predictions obtained using the model are very close to the true outcome values, then RSE will be small.

Statistic:

A screenshot of a cell phone

Description automatically generated

1. measures the proportion of variability in Y that can be explained using X
2. statistic is a measure of the linear relationship between X and Y

Residual Sum of Squares:

A close up of a clock

Description automatically generated

1. The amount of variability that left unexplained after performing the regression

Total Sum of Squares:

A picture containing drawing

Description automatically generated

1. Total variance in the response Y
2. Amount of variability inherent in the response before the regression is performed

3.2 Multiple Linear Regression



Where represents the jth predictor and quantifies the association between that variable and the response

Interpretation:

is the average effect on of a one unit increase in , holding all other predictors fixed.

3.2.1 Estimating the Regression Coefficients

The parameters are estimated using the least squares approach

A picture containing clock

Description automatically generated

Each predictor a separate slope coefficient in a single model, doing that we have a p distinct predictors.



Simple and multiple regression coefficients can be quite different

1. In simple regression, the slope term represents the average effect
2. In the multiple regression setting, the coefficient for one factor is the average holding the other factors constant
   * 1. Some Important Questions

Question1: Is at least one of the predictors ,,…,useful in predicting the response?

1. Need to ask whether all of the regression coefficients are zero🡪Use a hypothesis test

Null Hypothesis:



Alternative Hypothesis:



*F-statistic: Hypothesis test is performed by computing the F-statistic*

A close up of a logo

Description automatically generated

\*When there is no relationship between the response and the predictors, one would expect the F-statistic to take on a value close to 1.

\*If the alternative hypothesis is trye, we expect F to be greater than 1.

\*When n is large, an F-statistic that is just a little larger than 1 might still provide evidence against

\*A large F-statistic is needed to reject if n is small

Test a particular subset of q of the coefficients are zero:

Null Hypothesis:



\*Fit a model that uses all the variables except those last q.

A close up of a logo

Description automatically generated

\*Reports the partial effect of adding a variable to the model

\*The approach of using an F-statistic to test for any association between the predictors and the response works when p is relatively small.

Question 2: Do all the predictors help to explain Y, or is only a subset of the predictors useful.

Variable Selections: The task of determining which predictors are associated with the response, in order to fit a single model involving only those predictors.

1. Trying out a lot of different models, each containing a different set of predictors.(total of models)
2. Forward Selection:
3. Begin with the null model- a model that contains a intercept but no predictors.
4. Then we fit p simple linear regressions and add to the null model the variable that results in the lowest RSS.
5. Add to the model that results in the lowest RSS.
6. Backward Selection: Can not be used if p is larger than n
   1. Stat with all variables in the model
   2. Remove the variable with the largest p-value(variable that is least statistically significant)
   3. The new (p-1) variable model is fit, and the variable with the largest p-value is removed.
7. Mixed Selection:
   1. Start with no variables in the mode
   2. Add the variable that provides the best fit
   3. Continue to add variables one-by-one
   4. If at any point the p-value for one of the variables in the model rises above certain threshold, then remove the variable from the model.
   5. Continue the procedure until all variables in the model have low p value, all variables outside the model have large p values

Question 3: How well does the model fit the data?

:

1. In simple linear model, the square of the correlation of the response and the variable
2. In multilinear model, the square of the correlation between the response and the fitted linear model
3. is close to 1 indicates that the model explains a large portion of the variance
4. will always increase when more variables are added to the model, even if those models are weakly associated with the response.

A close up of a clock

Description automatically generated

Question 4: Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Three sources of uncertainty associated with the prediction:

1. The inaccuracy on the coefficient estimates(reducible error)
2. Linear model for is almost an approximation of reality(model bias/reducible error)
3. Prediction intervals: Wider than confidence intervals because they incorporate both the error in the estimate for and the uncertainty as to how much an individual point will differ from the population regression plane.
   1. Other Considerations in the Regression Model
      1. Qualitative Predictors

Predictors with Only Two Levels

1. Create a dummy variable that takes on two possible numerical values

Qualitative Predictors with More than Two Levels

1. Create more than one dummy variables
2. There will always be one fewer dummy variable than the number of levels.
3. The level with no dummy variable is the baseline.
4. Dummy variable approach can still be used when incorporating both quantitative and qualitative predictors
   * 1. Extensions of the Linear Model

Additive Assumptions: The effect of changes in a predictor on the response Y is independent of the values of the other predictors

Linear Assumptions: Changes in the response Y due to a one-unit change in is constant, regardless of the value of

Removing the Additive Assumption

Synergy effect/Interaction Effect:

1. Add an interaction term:



Hierarchical principle: If we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant

Non-linear Relationships

Polynomial Regression:

Incorporating non-linear associations in a linear model by including transformed versions of the predictors in the model(still a linear model)

* + 1. Potential Problems

1. Non-linearity of the Data

Linear regression model assumes that there is a straight-line relationship between the predictors and the response

1. Residual plots: Useful graphical tool for identifying non-linearity

\*In simple linear regression, plot the residuals versus the predictor,

\*In multi linear regression, plot the residuals versus the predicted(or fitted) values

\*Ideally, the residual plot will show no discernible pattern

1. Correlation of Error Terms

* If there is correlation among the error terms, then the estimated standard error will tend to underestimate the true standard errors.(p-value associated with the model will be lower than they should be)
* Correlations frequently occur in the context of time series data, which consists of observations for which measurements are obtained at discrete points in time.
* Tracking in the residuals: Adjacent residuals may have similar values.

1. Non-constant Variance of Error Terms

Error terms should have a constant variance,

Heteroskedasticity: Non-constant variances in the errors

1. Can be identified form the presence of a funnel shape in the residual plot
2. One possible solution is to transform the response Y using a concave function such as
3. Fit the model by weighted least squares, with wights proportional to the inverse variances
4. Outliers

An outlier is a point for which is far from the value predicted by the model

1. It is typical for an outlier that does not have an unusual predictor value to have little effect on the least squares fit
2. Outliers can cause complications for the interpretation of the fit
3. Studentized residuals: computed by dividing each residual by its estimated standard error.
   1. Studentized residuals that are greater than 3 in absolute value are possible outliers
4. High Leverage Points

Observations with high leverage have an unusual value for

1. Removing the high leverage observation has a much more substantial impact on the least squares line than removing the outlier
   1. High leverage observations tend to have a sizable impact on the estimated regression line.
   2. Use leverage statistics to quantify an observation’s leverage(A large value indicates an observation with high leverage)

A picture containing object, clock

Description automatically generated

\*The leverage statistic is always between and 1, and the average leverage for all the observations is always equal to

1. Collinearity

Two or more predictor variables are closely related to one another.

1. It can be difficult to separate out the individual effects of collinear variables on the response
2. Collinearity reduces the accuracy of the estimates of the regression coefficients, it causes the standard error for to grow.
3. Collinearity results in a decline in the t-statistic: Power of the hypothesis test-the probability of correctly detecting a non-zero coefficient-is reduced by collinearity.
4. A simple way to detect collinearity is to look at the correlation matrix of the predictors
5. A matrix that is large in absolute value indicates a pair of highly correlated variables, and therefore a collinearity problem in the data

Multicollinearity: It is possible for collinearity to exist between three or more variables even if no pair of variables has a particularly high correlation

A close up of a clock

Description automatically generated

 is the from a regression onto all of the other predictors

1. The smallest possible value for VIF is 1, which indicates the complete absence of collinearity.

Solutions to Multicollinearity:

1. Drop one of the problematic variables from the regression
2. Combine the collinear variables together in a single predictor
   1. The Marketing Plan
3. Is there a relationship?

Fit a regression model and test the null hypothesis

1. How strong is the relationship?

Use the RSE estimates or use R squared

1. Which media contribute to sales?

Examine the p-value associated with each predictor’s t-statistic

1. How large is the effect of each medium on sales?

Use the confidence interval, sort out the collinearity problem

1. How accurately can we predict future sales?

Prediction intervals will always be wider than confidence intervals because they account for the uncertainty associated with , the irreducible error

1. Is the relationship linear?

Use the residual plots to identify non-linearity

1. Is there synergy among the advertising media?

Additive model: The effect of each predictor on the response is unrelated to the values of the other predictors.

Interaction term: Accommodate the non-additive relationships

* 1. Comparison of Linear Regression with K-Nearest Neighbors

1. Linear regression is an example of a parametric approach because it assumes a linear functional form for
   1. Easy to fit
   2. Estimate only a small number of coefficients
   3. Disadvantages: Make strong assumptions about the form of
2. Non-parametric methods: Do not explicitly assume a parametric form for , provides an alternative and more flexible approach for performing regression.

KNN(K-nearest neighbors):

1. KNN regression first identifies the K training observations that are closest to , represented by
2. Then estimates using the average of all the training responses in

A close up of a clock

Description automatically generated

\*Optimal value for K will depend on the bias-variance tradeoff

\*Small value for K provides the most flexible fit, which will have low bias but high variance

\*Larger values of K provide a smoother and less variable fit; the smoothing may cause bias by masking some of the structure in f(x)

Parametric vs. Non-Parametric:

* + 1. The parametric approach will outperform the nonparametric approach if the parametric form that has been selected Is close to the true form of .
    2. As the extent of non-linearity increases, there is little change in the test set MSE for the non-parametric KNN method, but there is a large increase in the test set MSE of linear regression.
    3. Curse of dimensionality: higher dimensions effectively reduce sample size.
    4. General rule: Parametric methods will tend to outperform non-parametric approaches where there is a small number of observations per predictor.
  1. Lab: Linear Regression

3.6.2 Simple Linear Regression

Lm(y~x,data) or

Attach(data)

Lm(y~x)

Summary(): standard error for the coefficients, as well as the R squared statistic and F-statistic

Coef(model):Extract coefficients

Confint(model): obtain a confidence interval

predict(lm.fit,data.frame(lstat=c(5,19,15)),

interval = "confidence"): produce 95% confidence interval for a given value of lstat

predict(lm.fit,data.frame(lstat=c(5,19,15)),

interval = "prediction"): 95% prediction interval for a given value of lstat

plot(y,x): scatter plot

abline(fitted model): regression fit

abline(a,b): a line with intercept a and slope b

A screenshot of a cell phone

Description automatically generated

Diagnostic Plots:plot(fitted model)

1. Residuals vs Fitted(Used to detect nonlinearity): Fitted values vs. residuals
2. Normal Q-Q: Theoretical Quantiles vs Standardized Residuals
3. Scale-Location: Fitted values vs.
4. Residuals vs. Leverage: Leverage vs. Standardized residuals

Residuals(): used to compute linear regression fit

Rstudent(): return the studentized residuals

plot(predict(lm.fit),residuals(lm.fit)):Plot residuals against the fitted values

Hatvalues(): compute the leverage statistics

Which.max(): identifies the index of the largest element of s vector

Multiple Linear Regression

Lm(y~x1+x2+x3): fit a multiple linear regression using least squares

Lm(y~.): Fit all of the parameters into the model

To access the individual components of the summary:

1. ?summary.lm: see all that is available components of the summary
2. Summary(lm.fit)$r.sq
3. Vif(): used to compute the variance inflation factors, part of the car package
4. Lm(y~-.variable): perform a regression using all of the variables except one
5. Update(lm.fit,~.-age): update the function and leave one of the factors

3.6.4 Interaction Terms

To include interaction term:

Lm(medv~lstat:black): includes an interaction term between lstat and black

Lm(medv~lstat\*age)=lm(medc~lstat+age+lstat:age)

3.6.5 Non-linear Transformations of the Predictors

I(X^2): Raise X to the power 2

Perform non-linear regression:

Lm(med~lstat+I(stat^2)): perform a polynomial regression on lm.

Anova(): produce an analysis of variance of table

1. Performs hypothesis test comparing the two models,
2. Null hypothesis is that the two models fit the data equally well
3. The alternative hypothesis is that the full model is superior

Poly(lstat,n): including additional polynomial terms, up to nth order

* + 1. Qualitative Predictors

Given a qualitative variable, R generates dummy variables automatically.

Contrasts(): returns the coding R uses for the dummy variables

* + 1. Writing Functions

Funcname=function(){

Commands

}