Linear Regression

Lu Haibo

Mondy, Feb 25, 2019

Linear Regression

- a very simple approach for supervised learning
- mainly useful for predicting a quantitative response
- though very simple, still a useful and widely used statistical learning method

Example

Suppose that in our role as statistical consultants we are asked to suggest, on the basis of this data, a marketing plan for next year that will result in high product sales. What information would be useful in order to provide such a recommendation? Here are a few important questions that we might seek to address:

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
8.7	48.9	75.0	7.2

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? (*synergy* effect, *interaction* effect)

Simple Linear Regression

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

- $\approx$ as *is approximately modeled as*
- `sales` $\approx \beta_0 + \beta_1 \times$ `TV`
- β_0, β_1: *coefficients* or *parameters*

- β_0: *intercept*, β_1: *slope*

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

- $\hat{\beta}_0$, $\hat{\beta}_1$: *estimated value* using our training data
- $\hat{y}$: *predicted value* of the response

Estimating the Coefficients

In practice, β_0 and β_1 are unknown. So before making predictions, we must use data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ to estimate the coefficients.

- measuring *closeness*
- minimizing the *least squares*: $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

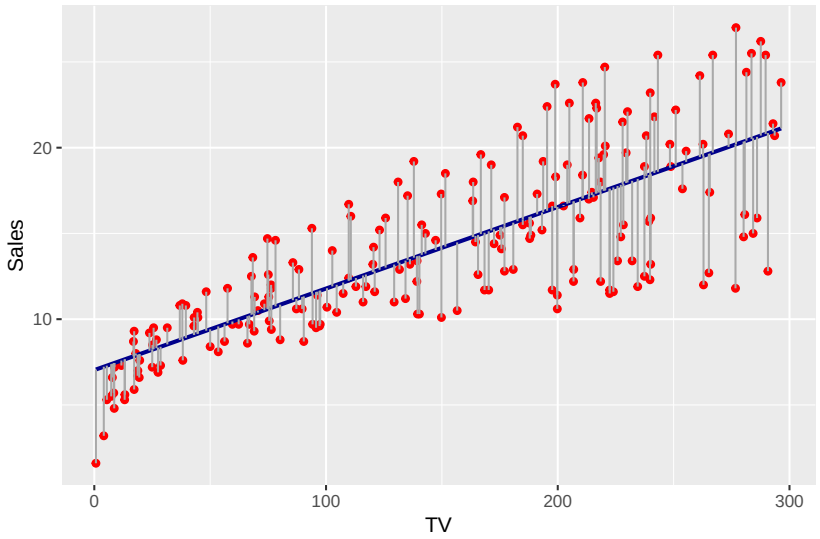


Figure 1: For the Advertising data, the least squares fit for the regression of sales onto TV is shown. The fit is found by minimizing the sum of squared errors. Each grey line segment represents an error, and the fit makes a compromise by averaging their squares. In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot

- *residual*: $e_i = y_i - \hat{y}_i$
- *residual sum of squares* (RSS)

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2$$

- *least squares coefficient estimates* for simple linear regression

$$\begin{aligned}\hat{\beta}_1 =& \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},\\ \hat{\beta}_0 =& \bar{y} - \hat{\beta}_1 \bar{x}\end{aligned}$$

Assessing the Accuracy of the Coefficient Estimates

- Assume the *true* relationship between X and Y takes the form

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

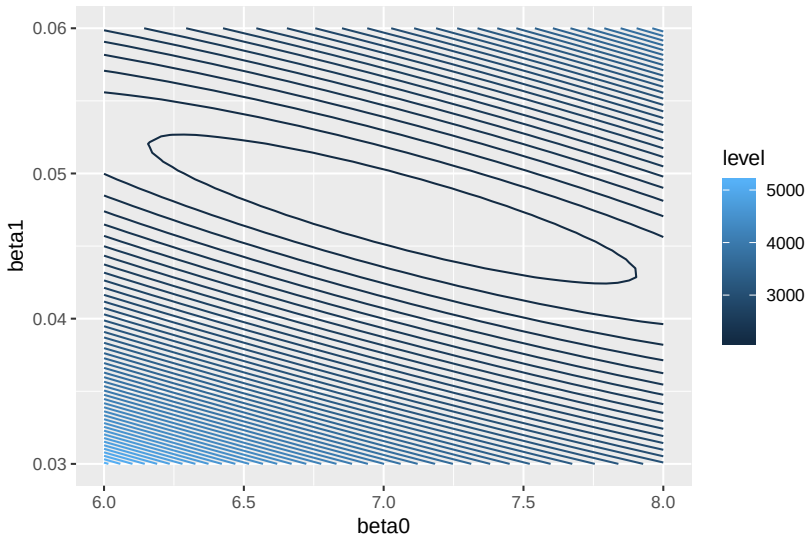


Figure 2: Contour plots of the RSS on the Advertising data, using sales as the response and TV as the predictor. The red dots correspond to the least squares estimates β_0, β_1.

- If f is to be *approximated* by a linear function, the *population regression line*

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

- The error term ϵ catches 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
  - there may be measurement error
- Typically assume that the error term is independent of X

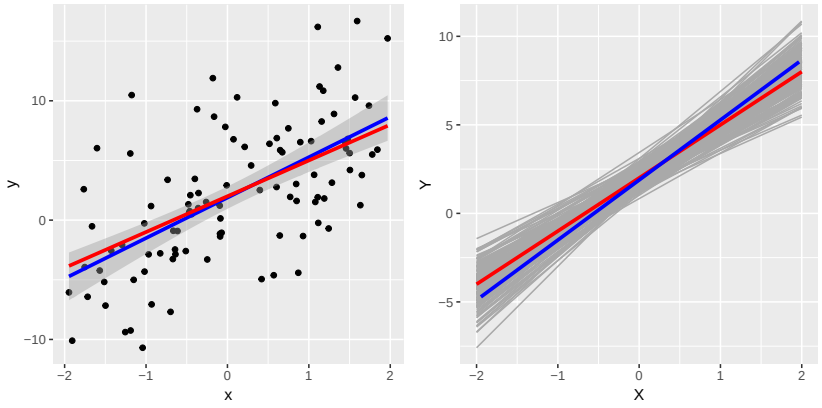


Figure 3: A simulated data set. Left: The red line represents the true relationship, $f(X) = 2 + 3X$, which is known as the population regression line. The blue line is the least squares line; it is the least squares estimate for f(X) based on the observed data, shown in black. Right: The population regression line is again shown in red, and the least squares line in dark blue. In light blue, ten least squares lines are shown, each computed on the basis of a separate random set of observations. Each least squares line is different, but on average, the least squares lines are quite close to the population regression line.

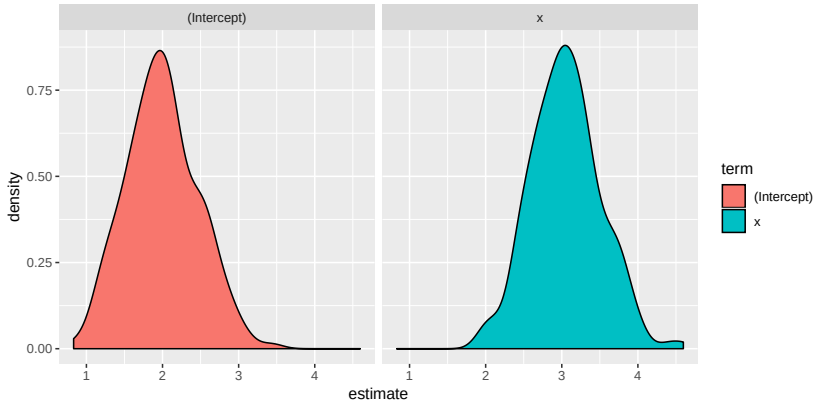
Population regression line VS. Least squares line

Using information from a sample to estimate characteristics of a large population

Confidence interval & Hypothesis test

Suppose that we are interested in how close $\hat{\beta}_0$ and $\hat{\beta}_1$ are to the true values β_0 and β_1. Unfortunately, β_0, β_1 is unknown, but in general the $\hat{\beta}_0, \hat{\beta}_1$ calculated from the sample will provide a good estimate of them. In fact, this estimate is *unbiased*:

$$E(\hat{\beta}_i) = \beta_i, \quad i = 1, 2.$$



- how accurate is $\hat{\beta}_1$ as an estimate of β_1?
  - *standard error* of $\hat{\beta}_1$

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

    where $\sigma^2 = Var(\epsilon)$
  - *residual standard error* (RSE)
    - estimate for σ
    - $RSE = \sqrt{RSS/(n-2)}$
  - A 95% *confidence interval*
    - a 95% confidence interval for β_1 approximately takes the form

$$\hat{\beta}_1 \pm 2 \times SE(\hat{\beta}_1)$$

    - a range of values such that with 95% probability, the range will contain the true unknown value of the parameter

- *hypothesis test* on the coefficients

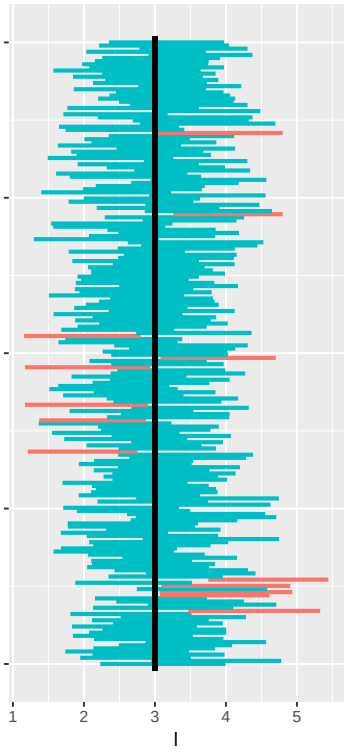


Figure 4: Confidence interval calculated from 100 different sample. The black vertical line denotes the true pa-

	Dependent variable:
	y
x	3.390***
	(0.442)
Constant	1.884***
	(0.472)
Observations	100
R^2	0.375
Adjusted R^2	0.369
Residual Std. Error	4.705 (df = 98)
F Statistic	58.886*** (df = 1; 98)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 1: Coefficients of the least squares model for the regression of $Y = 2 + 3X + \epsilon$

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

$$H_0 : \textit{There is no relationship between } X \textit{ and } Y$$

versus the *alternative hypothesis*

$$H_a : \textit{There is some relationship between } X \textit{ and } Y$$

- Mathematically, this corresponds to testing

$$H_0 : \beta_1 = 0$$

versus

$$H_a : \beta_1 \neq 0$$

- To test the null hypothesis, we need to determine whether $\hat{\beta}_1$, our estimate for β_1, is *sufficiently far from zero* that we can be confident that β_1 is non-zero
- *t-statistic*

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

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

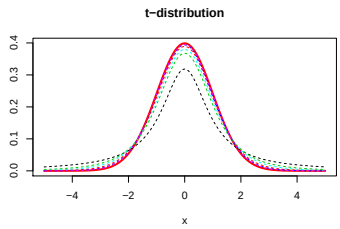


Figure 5: The red curve denotes the Normal distribution $N(0, 1)$. The dashed curve denotes the t-distribution with different degrees of freedom. As df increases, the t-distribution approaches to $N(0, 1)$.

- *p-value* - the probability of observing any value equal to $|t|$ or larger, assuming $\beta_1 = 0$ - a small p-value indicates that it is unlikely

term	estimate	std.error	statistic	p.value
(Intercept)	7.03	0.46	15.36	0
TV	0.05	0.00	17.67	0

Table 2: Regression result for the Advertising data: Sales TV

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

- *significant level* α
  - the criterion used for rejecting the null hypothesis
  - chosen before data collection and is usually set to 0.05

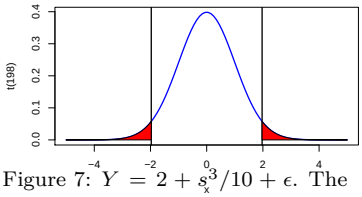


Figure 7: $Y = 2 + s_x^3/10 + \epsilon$. The t-statistics shows that there is no enough evidence shows that there is a relationship between X and Y.

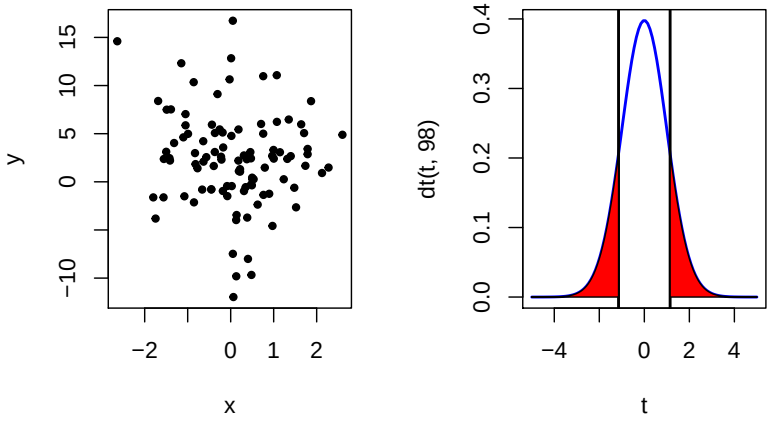


Figure 6: The red area shows the area for $Pr(>|t|) = 0.05$

- Types of error in Hypothesis test
  - type I error (*false positive*)
    - incorrect rejection of a true null hypothesis
    - leads one to conclude that a supposed effect or relationship exists when in fact it doesn't
    - measured by α
  - type II error (*false negative*)
    - failure to reject a false null hypothesis
    - fail to believe relationship exists
    - measured by β, usually can not be calculated
  - *power* of a test: $1 - \beta$

term	estimate	std.error	statistic	p.value
(Intercept)	2.38	0.49	4.84	0.00
x	-0.54	0.48	-1.15	0.25

Table 3: Regression result for Fig. 7

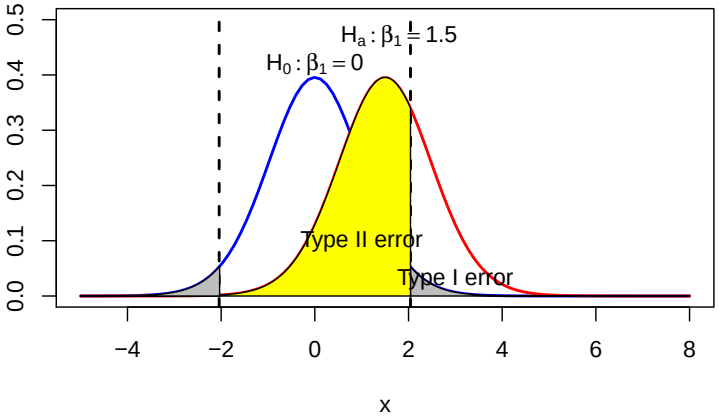


Figure 8: The blue curve shows the distribution of β_1 under H_0. The red curve shows the distribution of β_1 under H_a. The dashed black line shows the quantile for $Pr(> |t|) = 0.05$. The gray area shows the Type I error. The yellow area shows the Type II error.

Assessing the Accuracy of the Model

Once we have rejected the null hypothesis in favor of the alternative hypothesis, it is natural to want to quantify *the extend to which the model fits the data*

- *residual standard error* (RSE)
  - Due to the presence of error term ϵ, even if we knew the true regression line, we would not be able to perfectly predict Y from X
  - RSE is an estimate of $\sigma(\epsilon)$, a measure of the *lack of fit* of the model to the data
- R^2 statistic
  - measures the *proportion of variability in* Y *that can be explained using* X

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

where $TSS = \sum(y_i - \bar{y})^2$ is the *total sum of squares*, $ESS = TSS - RSS$ is the *explained sum of squares*

		Null hypothesis (H_0) is	
		Valid/True	**Invalid/False**
Judgement of Null Hypothesis (H_0)	**Reject**	Type I error (α)/False Positive	Correct inference/True Positive
	Fail to reject	Correct inference/True Negative	Type II error (β)/False Nagative

- An R^2 statistic that is close to 1 indicates that a large proportion of the variability in the response has been explained by the regression
- A number near 0 indicates that the regression did not explain much of the variability in the response, this might occur because
  * the linear model is wrong
  * the inherent error σ^2 is high
  * in the simple linear regression setting, $R^2 = r^2$, where $r = Cor(X, Y)$

	Dependent variable:
	Sales
TV	0.048***
	(0.003)
Constant	7.033***
	(0.458)
Observations	200
R^2	0.612
Adjusted R^2	0.610
Residual Std. Error	3.259 (df = 198)
F Statistic	312.145*** (df = 1; 198)
Note:	*p<0.1; **p<0.05; ***p<0.01

Predictions

- The inaccuracy in the coefficient estimates is related to the *reducible error*
- *confidence interval*: determine how close $\hat{Y}$ will be to $f(X)$, uncertainty surrounding the *average*
- *prediction interval*: determine how close $\hat{Y}$ will be to $f(X) + \epsilon$, uncertainty surrounding for one particular input

Exercise

This question involves the use of simple linear regression on the `Auto` data set. (`library(ISLR)`)

a. Use the `lm()` function to perform a simple linear regression with `mpg` as the response and `horsepower` as the predictor. Use the

Sales	TV	.fitted	.se.fit	.resid	.hat	.sigma	.cooksd	.std.resid
22.1	230.1	17.97	0.32	4.13	0.01	3.25	0.01	1.27
10.4	44.5	9.15	0.36	1.25	0.01	3.27	0.00	0.39
9.3	17.2	7.85	0.42	1.45	0.02	3.27	0.00	0.45
18.5	151.5	14.23	0.23	4.27	0.01	3.25	0.00	1.31
12.9	180.8	15.63	0.25	-2.73	0.01	3.26	0.00	-0.84
7.2	8.7	7.45	0.44	-0.25	0.02	3.27	0.00	-0.08

Table 4: Predicted values for the model: Sales TV

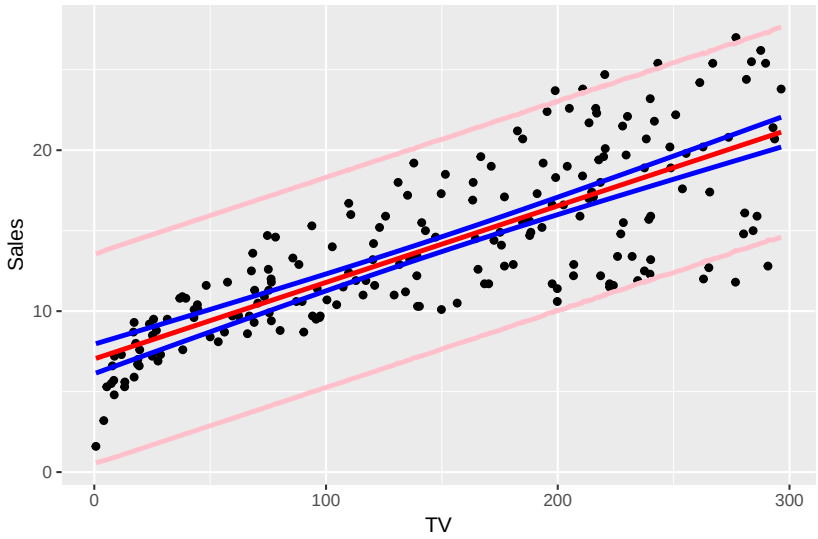


Figure 9: The red line shows the regression line between Sales and TV. The blue lines are the 95% confidence interval, and the pink lines are the 95% prediction interval.

`stargazer()` function (`library(stargazer)`) to print the results. Comment on the output. For example:

1. Is there a relationship between the predictor and the response?
2. How strong is the relationship between the predictor and the response?
3. Is the relationship between the predictor and the response positive or negative?
4. What is the predicted `mpg` associated with a `horsepower` of 98? What are the associated 95% cofidence and prediction intervals?

b. Plot the response and the predictor. Use `ggplot` (`library(ggplot2)`) to display the least squares regression line.

c. Use `ggplot` to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

Multiple Linear Regression

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

term	estimate	std.error	statistic	p.value
(Intercept)	2.94	0.31	9.42	0.00
TV	0.05	0.00	32.81	0.00
Radio	0.19	0.01	21.89	0.00
Newspaper	0.00	0.01	-0.18	0.86

Table 5: For the Advertising data, least squares coefficients estimates of the multiple linear regress of number of units sold on radio, TV, and newspaper advertising budgets.

- As was the case in simple linear regression setting, the regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$ are unknown, but can be estimated using the same least squares approach.
- Given estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, we can make predictions using

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

- Does it make sense for the multiple regression to suggest no relationship between `Sales` and `Newspaper` while the simple linear regression implies the opposite? Notice that the correlation between `Radio` and `Newspaper` is 0.3541038.

	TV	Radio	Newspaper	Sales
TV	1.00	0.05	0.06	0.78
Radio	0.05	1.00	0.35	0.58
Newspaper	0.06	0.35	1.00	0.23
Sales	0.78	0.58	0.23	1.00

Some Important Questions

When we perform multiple linear regression, we usually are interested in answering a few important questions.

1. Is at least one of the predictors $X_1, X_2, \ldots, X_p$ usful in predicting the response?
2. Do all the predictors help to explain Y, or is only a subset of the predictors useful?
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?

One: Is There a Relationship Between the Response and Predictors?

As in the simple linear regression setting, we use a hypothesis test to answer this question:

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

versus the alternative

$$H_a : \textit{at least one } \beta_j \textit{ is non} - \textit{zero}$$

This hypothesis test is performed by computing the *F-statistic*

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

- When there is no relationship between the response and predictors, one would expect the F-statistic to take on a value close to 1. On the other hand, if H_a is true, we expect F to be greater than 1.

- Given the individual p-values for each variable, **why do we need to look at the overall F-statistic?** [1]

[1] Remember any hypothesis test can make mistakes. If H_0 is true, there is only a 5% chance that the F-statistic will result in a p-value below 0.05. While if we use t-statistic for every variable, the probability that we did wrong will be much larger than 5% when p is large.

Two: Deciding on Important Variables

The task of determining which predictors are associated with the response, in order to fit a single model involving only those predictors, is referred to as *variable selection*

- Ideally, we would like to perform variable selection by trying out a lot different models, each containing a different subset of the predictors. We can then select the *best* model out of all of the models that we have considered.
- How do we determine which model is best?
  - *Mallows's* C_p
  - *Akaike information criterion* (AIC)
  - *Bayesian information criterion* (BIC)
  - *adjusted* R^2
- Unfortunately, there are a total of 2^p models that contain subsets of p variables. There are three classical approaches
  - *Forward selection*: begin with the *null model*
  - *Backward selection*: begin with all variables in the model
  - *Mixed selection*

Three: Model Fit

Two of the most common numerical measures of model fit

- RSE
- R^2 VS *adjusted* R^2

Four: Predictions

- confidence interval
- prediction interval

Table 6:

	Dependent variable:			
	Sales			
	(1)	(2)	(3)	(4)
TV	0.048***	0.046***	0.047***	0.046***
	(0.003)	(0.001)	(0.003)	(0.001)
Radio		0.188***		0.189***
		(0.008)		(0.009)
Newspaper			0.044***	−0.001
			(0.010)	(0.006)
Constant	7.033***	2.921***	5.775***	2.939***
	(0.458)	(0.294)	(0.525)	(0.312)
AIC	1044.09	780.39	1027.78	782.36
BIC	1053.99	793.59	1040.97	798.85
Mallows's Cp	544.08	2.03	481.33	4
Observations	200	200	200	200
R^2	0.612	0.897	0.646	0.897
Adjusted R^2	0.610	0.896	0.642	0.896
Residual Std. Error	3.259 (df = 198)	1.681 (df = 197)	3.121 (df = 197)	1.686 (df = 196)
F Statistic	312.145*** (df = 1; 198)	859.618*** (df = 2; 197)	179.619*** (df = 2; 197)	570.271*** (df = 3; 196)
Note:				*p<0.1; **p<0.05; ***p<0.01

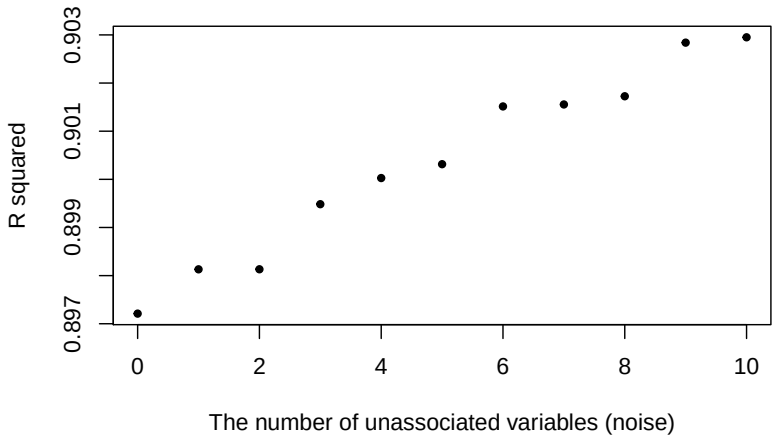


Figure 10: R^2 will always increase when more variables are added to the model, even if those variables are not associated with the response.

year	age	maritl	race	education	jobclass	health	health_ins	wage
Min. :2003	Min. :18.00	1. Never Married: 648	1. White:2480	1. < HS Grad :268	1. Industrial :1544	1. <=Good : 858	1. Yes:2083	Min. : 20.09
1st Qu.:2004	1st Qu.:33.75	2. Married :2074	2. Black: 293	2. HS Grad :971	2. Information:1456	2. >=Very Good:2142	2. No : 917	1st Qu.: 85.38
Median :2006	Median :42.00	3. Widowed : 19	3. Asian: 190	3. Some College :650	NA	NA	NA	Median :104.92
Mean :2006	Mean :42.41	4. Divorced : 204	4. Other: 37	4. College Grad :685	NA	NA	NA	Mean :111.70
3rd Qu.:2008	3rd Qu.:51.00	5. Separated : 55	NA	5. Advanced Degree:426	NA	NA	NA	3rd Qu.:128.68
Max. :2009	Max. :80.00	NA	NA	NA	NA	NA	NA	Max. :318.34

Table 7: Wage and other data for a group of 3000 male workers in the Mid-Atlantic region. (WageISLR)

Other Considerations in the Regression Model

Qualitative Predictors

$$jobclass2.Information = \begin{cases} 0, & \text{if jobclass} = 1.\text{Industrial} \\ 1, & \text{if jobclass} = 2.\text{Information} \end{cases}$$

	Dependent variable:
	wage
year	1.241***
	(0.307)
age	0.271***
	(0.062)
maritl2. Married	17.177***
	(1.720)
maritl3. Widowed	2.052
	(8.005)
maritl4. Divorced	3.967
	(2.887)
maritl5. Separated	11.530**
	(4.844)
race2. Black	−5.096**
	(2.146)
race3. Asian	−2.814
	(2.603)
race4. Other	−6.059
	(5.666)
education2. HS Grad	7.759***
	(2.369)
education3. Some College	18.340***
	(2.520)
education4. College Grad	31.240***
	(2.548)
education5. Advanced Degree	53.949***
	(2.811)
jobclass2. Information	3.571***
	(1.324)
health2. >=Very Good	6.515***
	(1.421)
health_ins2. No	−17.513***
	(1.403)
Constant	−2,423.329***
	(616.543)
Observations	3,000
R^2	0.340
Adjusted R^2	0.336
Residual Std. Error	34.001 (df = 2983)
F Statistic	95.886*** (df = 16; 2983)
Note:	*p<0.1; **p<0.05; ***p<0.01

Extenion of Linear Model

1. *synergy effect* or *interaction effect*

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

It is sometimes the case that an interaction term has a very small p-value, but the associated main effects (in this case, TV and radio) do not. The *hierarchical principle* states that *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.*

	Dependent variable:		
	Sales		
	(1)	(2)	(3)
TV	0.046***	0.019***	0.046***
	(0.001)	(0.002)	(0.001)
Radio	0.188***	0.029***	0.189***
	(0.008)	(0.009)	(0.009)
TV:Radio		0.001***	
		(0.0001)	
Newspaper			−0.001
			(0.006)
Constant	2.921***	6.750***	2.939***
	(0.294)	(0.248)	(0.312)
AIC	780.39	550.28	782.36
BIC	793.59	566.77	798.85
Mallows's Cp	2.03	-130.58	4
Observations	200	200	200
R^2	0.897	0.968	0.897
Adjusted R^2	0.896	0.967	0.896
Residual Std. Error	1.681 (df = 197)	0.944 (df = 196)	1.686 (df = 196)
F Statistic	859.618*** (df = 2; 197)	1,963.057*** (df = 3; 196)	570.271*** (df = 3; 196)
Note:			*p<0.1; **p<0.05; ***p<0.01

2. Non-linear relationships

	Dependent variable:		
	mpg		
	(1)	(2)	(3)
horsepower	−0.158***		
	(0.006)		
poly(horsepower, 2)1		−120.138***	
		(4.374)	
poly(horsepower, 2)2		44.090***	
		(4.374)	
poly(horsepower, 5)1			−120.138***
			(4.326)
poly(horsepower, 5)2			44.090***
			(4.326)
poly(horsepower, 5)3			−3.949
			(4.326)
poly(horsepower, 5)4			−5.188
			(4.326)
poly(horsepower, 5)5			13.272***
			(4.326)
Constant	39.936***	23.446***	23.446***
	(0.717)	(0.221)	(0.218)
AIC	2363.32	2274.35	2268.66
BIC	2375.24	2290.24	2296.46
Mallows's Cp	1430.43	1053.82	1011.46
Observations	392	392	392
R^2	0.606	0.688	0.697
Adjusted R^2	0.605	0.686	0.693
Residual Std. Error	4.906 (df = 390)	4.374 (df = 389)	4.326 (df = 386)
F Statistic	599.718*** (df = 1; 390)	428.018*** (df = 2; 389)	177.366*** (df = 5; 386)

Note: *p<0.1; **p<0.05; ***p<0.01

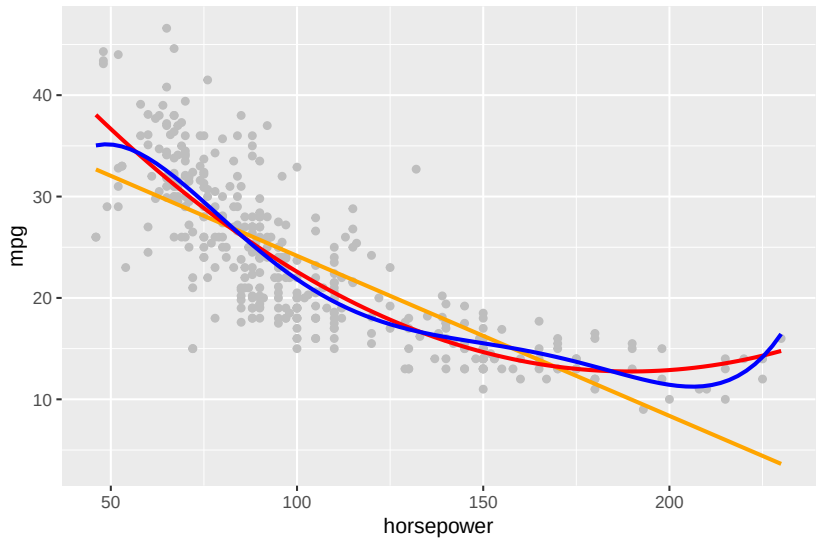


Figure 11: The Auto data set. For a number of cars, mpg and horsepower are shown. The linear regression fit is shown in orange. The linear regression fit for a model that includes $horsepower^2$ is shown as a red curve. The linear regression fit for a model that includes all polynomials of horsepower up to fifth-degree is shown in blue.

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

In practice, identifying and overcoming these problems is as much an art as a science.

1. Nonlinearity of the Data

The linear regression model assumes that there is a straight-line relationship between the predictors and the response. If the true relationship is far from linear, then virtually all of the conclusions that we draw from the fit are suspect.

Residual plots are a useful graphical tool for identifying non-linearity.

2. Correlation of Error Terms

An Important assumption of the linear regression model is that the error terms, $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$, are uncorrelated. This means that the fact ϵ_i is positive provides little or no information about the sign of ϵ_{i+1}.

If in fact there is correlation among the error terms, then **the estimated standard errors will tend to underestimate the true standard errors**. As a result, confidence and prediction intervals will be narrower than they should be.

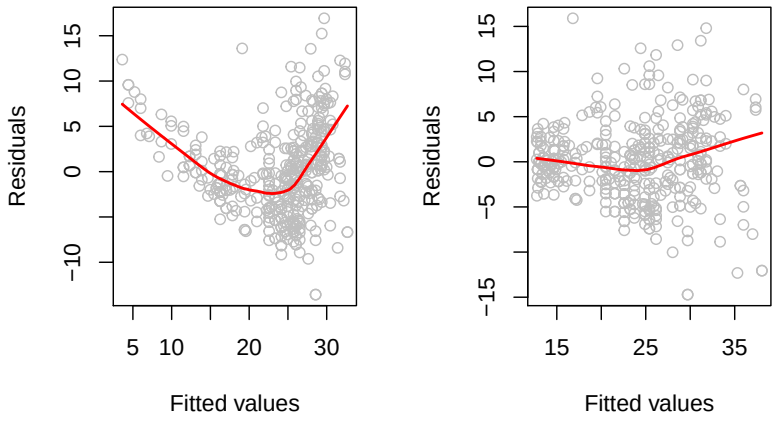


Figure 12: Plots of residuals versus predicted (or fitted) values for the Auto data set. In each plot, the red line is a smooth fit (use the 'smooth()' function) to the residuals, intended to make it easier to identify a trend. Left: A linear regression of mpg on horsepower. A strong pattern in the residuals indicates non-linearity in the data. Right: A linear regression of mpg on horsepower and horsepower2. There is little pattern in the residuals.

Such correlations frequently occur in the context of *time series* data, which consists of observations for which measurements are obtained at discrete points in time.

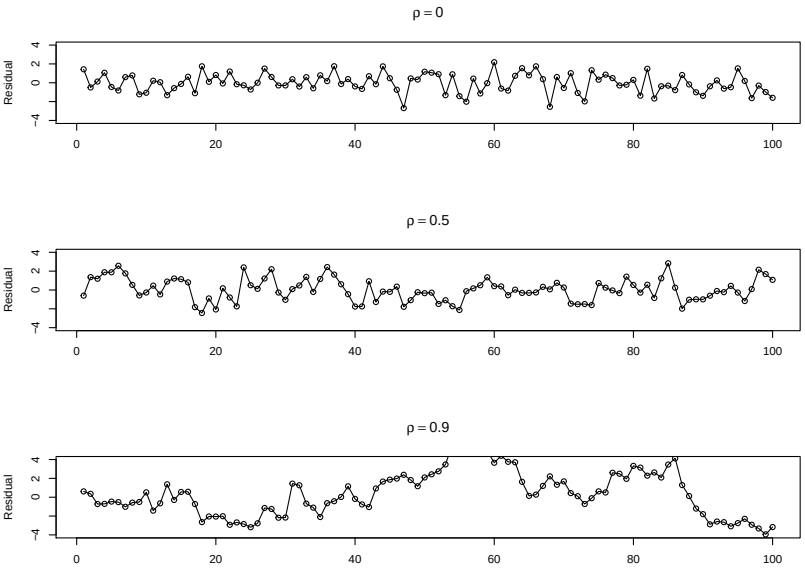


Figure 13: Plots of residuals from simulated time series data sets generated with differing levels of correlation ρ between error terms for adjacent time points

3. Non-constant Variance of Error Terms (*heteroscedasticity*)

Another important assumption of the linear regression model is that the error terms have a constant variance, $Var(\epsilon_i) = \sigma^2$. The standard errors, confidence intervals, prediction intervals, and hypothesis tests associated with the linear model rely upon this assumptions.

4. Outliers

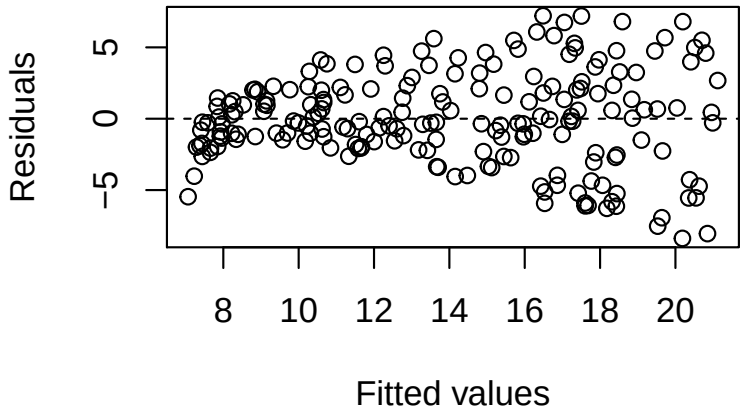


Figure 14: The funnel shape indicates heteroscedasticity.

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

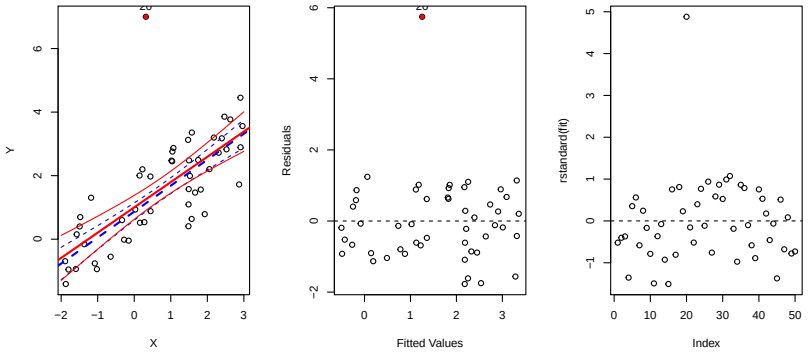


Figure 15: Left: The least squares regression line is shown in red, and the regression line after removing the outlier is shown in blue. Center: The residual plot clearly identifies the outlier. Right: The outlier has a studentized residual above 4; typically we expect values between -3 and 3.

Removing the outlier has little effect on the least squares line, however, the outlier can have a dramatic increase to the RSE, and cause the R^2 to decline.

We can plot the *studentized residuals* using `rstandard()` function. Observations whose studentized residuals are greater than 3 in absolute value are possible outliers.

5. High Leverage Points

The outliers are observations for which the response y_i is unusual given the predictor x_i. In contrast, observations with *high leverage* have an unusual value for x_i.

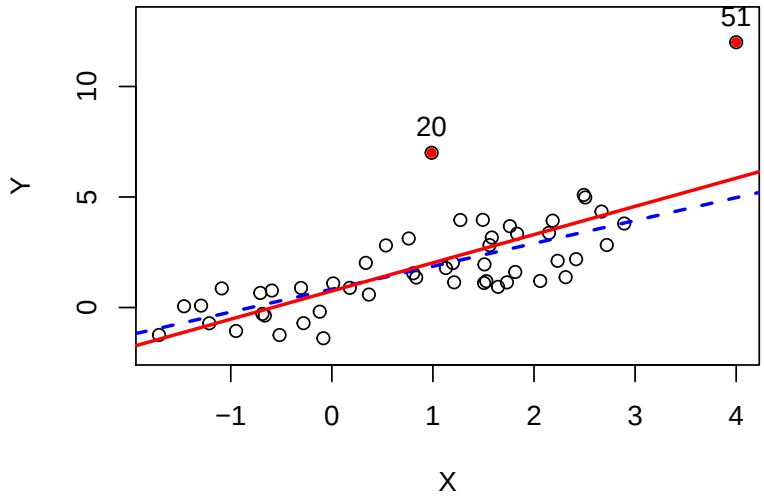


Figure 16: Observation 51 is a high leverage point, while 20 is not. The red line is the fit to all the data, and the blue line is the fit with observation 51 removed.

In fact, 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.

Leverage statistics can be computed using the `hatvalues()` function.

6. Colinearity

Collinearity refers to the situation in which two or more predictor variables are closely related to one another.

	Income	Limit	Rating	Cards	Age	Education	Gender	Student	Married	Ethnicity	Balance
1	14.89	3606	283	2	34	11	Male	No	Yes	Caucasian	333
2	106.03	6645	483	3	82	15	Female	Yes	Yes	Asian	903
3	104.59	7075	514	4	71	11	Male	No	No	Asian	580
4	148.92	9504	681	3	36	11	Female	No	No	Asian	964
5	55.88	4897	357	2	68	16	Male	No	Yes	Caucasian	331
6	80.18	8047	569	4	77	10	Male	No	No	Caucasian	1151

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. In other words, since `Limit` and `Rating` tend to increase or decrease together, it can be difficult to determine how each one separately is associated with the response, `balance`.

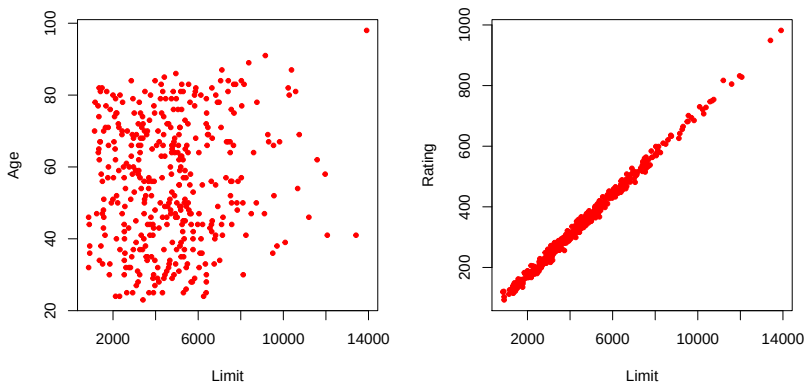


Figure 17: Scatterplots of the observations from the Credit data set. Left: A plot of age versus limit. These two variables are not collinear. Right: A plot of rating versus limit. There is high collinearity.

Table 8: The results for three multiple regression models involving the Credit data set are shown. The standard error of $\hat{\beta}_{limit}$ increases 12-fold in the second regression, due to collinearity.

	Dependent variable:		
	Balance		
	(1)	(2)	(3)
Age	−2.291***	−2.351***	−2.346***
	(0.672)	(0.668)	(0.669)
Limit	0.173***		0.019
	(0.005)		(0.063)
Rating		2.593***	2.310**
		(0.074)	(0.940)
Constant	−173.411***	−269.581***	−259.518***
	(43.828)	(44.806)	(55.882)
Observations	400	400	400
R^2	0.750	0.754	0.754
Adjusted R^2	0.749	0.752	0.752
Residual Std. Error	230.532 (df = 397)	228.818 (df = 397)	229.080 (df = 396)
F Statistic	594.988*** (df = 2; 397)	606.920*** (df = 2; 397)	403.718*** (df = 3; 396)
Note:			*p<0.1; **p<0.05; ***p<0.01

Since collinearity reduces the accuracy of the estimates of the regression coefficients, it causes the standard error for $\hat{\beta}_j$ to grow.

The Marketing Plan

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? (*synergy* effect, *interaction* effect)