Chapter 3 Linear Regression

1 Simple Linear Regression

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

Simple linear regression is a straightforward approach for predicting a quantitative response on the basis of a single predictor variable.

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

β0 and β1 are two unknown constants that represent the intercept and slope terms in the linear model.

$\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates of the model coefficients.

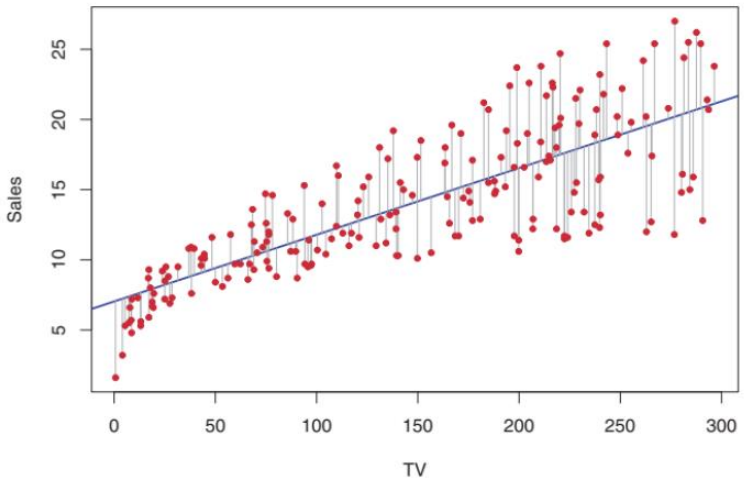
where $\hat{y}$ indicates a prediction of Y on the basis of $X = x$.

1.1 Estimating the Coefficients

In practice, $\hat{\beta}_0$ and $\hat{\beta}_1$ are unknown. Therefore, we must use data to estimate the coefficients.

Our goal is to obtain coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the linear model $Y \approx \beta_0 + \beta_1 X.$ fits the available data well.

There are many approaches to measure *closeness*. We use *least squares* criterion to derive $\hat{\beta}_0$ and $\hat{\beta}_1$.



Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X.

Then $e_i = y_i - \hat{y}_i$ represents the ith residual(difference between the ith observed response value and the ith response value that is predicted by our model)

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

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

We can find $\hat{\beta}_0$ and $\hat{\beta}_1$ which minimize RSS most. (if you want to derive equation below, send kakao talk to me)

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

From the Advertising problem, we can get $\hat{\beta}_0$ = 7.03, $\hat{\beta}_1$ = 0.0475

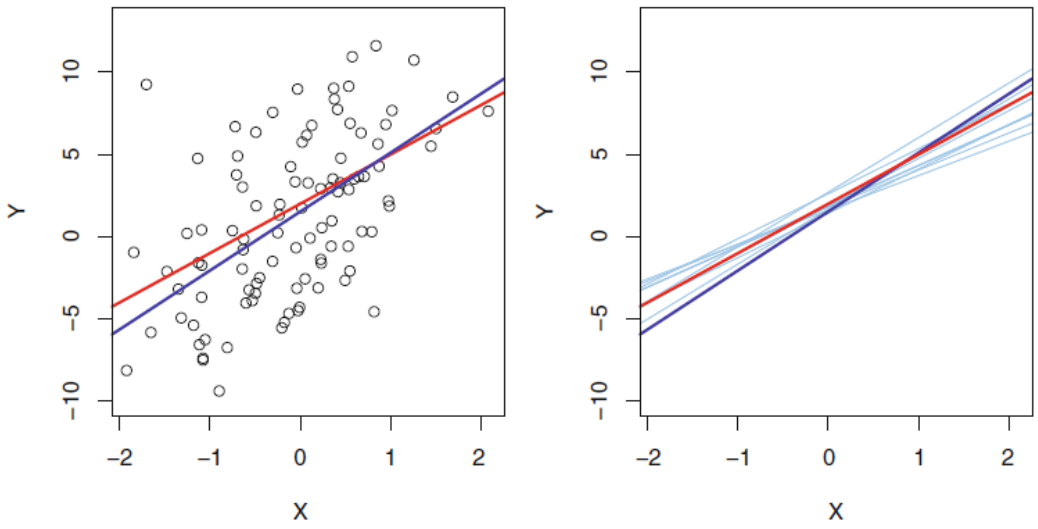
This means '$1000 increase in advertising would approximately lead to 47.5 increase in sales.'

1.2 Assessing the Accuracy of the CoefficientEstimates

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

The error term represents what we miss this simple model.

1. The true relationship is probably not linear
2. there may be other variables that cause variation in Y
3. There may be measurement error



Left image:
simulation result (100 simulated obs)
Generated with $Y = 2 + 3X + \epsilon$,
red line: true relationship; blue line: least squares line
In general, the true relationship of X and Y is not known. However, we can compute the least squares line.

Right image:

Generated 10 data set from $Y = 2 + 3X + \epsilon$,
plotted the resulting least squares line from each of the data set. (the light blue lines)

1. Each set resulted in different least squares line.
2. The population regression line does not change.

a natural extension of the standard statistical approach: Using information from a sample to estimate a large population.
$\hat{\mu} = \bar{y}$ is reasonable.

In the same way, in linear regression we are estimating the unknown βs.

$\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased.

$\hat{\beta}$ = β for all i

To answer the question of how close $\hat{\beta}_0$ and $\hat{\beta}_1$ are to the true values of $\hat{\beta}_0$ and $\hat{\beta}_1$, we need to compute the standard error (SE) of the parameter.

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

We need σ^2 in order to find the standard error. We don't know that. However, we can estimate σ^2 from data.

The estimate of σ is the *residual standard error* (RSE): $\mathrm{RSE} = \sqrt{\mathrm{RSS}/(n-2)}$.

From SE we can get confidence interval of β like:

$$\hat{\beta}_0 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_0). \quad \hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1).$$

Standard error can also be used for hypothesis testing on the coefficients where:

H_0 : There is no relationship between X and Y

H_a : There is some relationship between X and Y.

With SE of β we can compute t-statistic with d.f n-2. Through this computation we can get p-value. We will declare hypothesis if p-values is small emough. (which means X and Y are related)

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

1.3 Assessing the Accuracy of the Model

We will talk about accuracy of the model

1. RSE

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

it is the average amount that response will deviate from the true regression line.

2. R^2

RSE is measured in units of Y, it is not always clear what constitutes a good RSE. Therefore R^2 can be a useful measurement.

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} \quad \text{where} \quad \text{TSS} = \sum(y_i - \bar{y})^2$$

2 Multiple Linear Regression

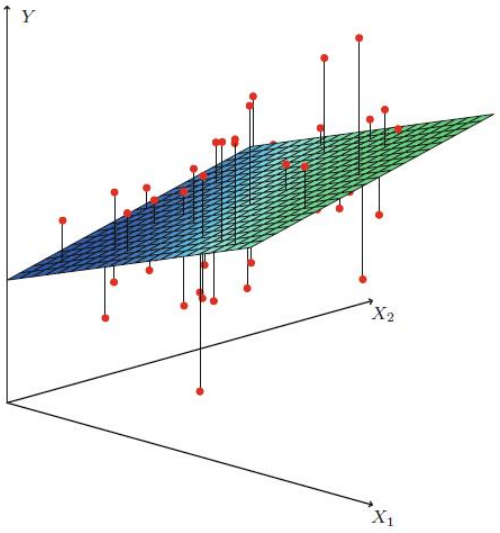
Multiple Linear Regression is an extension of simple linear regression to accommodate multiple predictors.

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

2.1 Estimating the Regression Coefficients

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2. \end{aligned}$$

Unlike the simple linear regression estimates given in (3.4), the multiple regression coefficient estimates have somewhat complicated forms that are most easily represented using matrix algebra. (They don't provide Googling gogo. Plz don't ask me)



	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

This difference stems from the fact that in the simple regression case, the slope term represents the average effect of a $1,000 increase in newspaper advertising, ignoring other predictors such as TV and radio.

In a simple linear regression which only examines sales versus newspaper, we will observe that higher values of newspaper tend to be associated with higher values of sales, even though newspaper advertising does not actually affect sales. So newspaper sales are a surrogate for radio advertising; newspaper gets "credit" for the effect of radio on sales.

2.2 Some Important Questions

1. Is at least one of the predictors useful in terms of predicting ?
2. Do all the predictors help to explain ?
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?

2.2.1 One: Is There a Relationship Between the Response?

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

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

We can decide whether to declare or not bt using F-statistic,

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

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

In this example the F-statistic is 570. Since this is far larger than 1, it provides compelling evidence against the null hypothesis H_0.

F-statistic suggests that at least one of the advertising media must be related to sales.

When n is large, an F-statistic that is just a little larger than 1 might still provide evidence against H_0. In contrast, a larger F-statistic is needed to reject H_0 if n is

small.

Sometimes we want to test that a particular subset of q of the coefficients are zero.

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

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/q}{\mathrm{RSS}/(n-p-1)}.$$

By calculating F-test, we get p-value and we can choose hypothesis to accept

2.2.2 Two: Deciding on Important Variables

1. *Forward selection*

   Starting from null model(just intercept no predictors). We then fit p simple linear regressions and add to the null model the variable that results in the lowest RSS. This approach is continued until some stopping rule is satisfied.

2. *Backward selection*

   We start with all variables in the model, and remove the variable with the largest p-value This procedure continues until a stopping rule is reached.

3. *Mixed selection*

   This is a combination of forward and backward selection. We start with no variables in the model, and as with forward selection, we add the variable that provides the best fit. if at any point the p-value for one of the variables in the model rises above a certain threshold, then we remove that variable from the model.

2.2.3 Three: Model Fit

Two of the most common numerical measures of model fit are RSE and R^2.

R^2 will always increase when more variables are added to the model, even if those variables are only weakly associated with the response.
adding newspaper advertising to the model containing only TV and radio advertising leads to just a tiny increase in R^2 provides additional evidence that newspaper can be dropped from the model.

The model that contains only TV and radio as predictors has an RSE of 1.681, and the model that also contains newspaper as a predictor has an RSE of 1.686). In contrast, the model that contains only TV has an RSE of 3.26

The observant reader may wonder how RSE can increase when newspaper is added to the model given that RSS must decrease. In general RSE is defined as

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

2.2.4 Four: Predictions

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

   This is only an estimate of the true line.

   The inaccuracy in the coefficient estimates is related to the reducible error.

   We can compute a Confidence interval.

   2. In practice, assuming a linear model f(x) for is almost always an approximation of reality. This is called *model bias*.
   However, we will ignore this discrepancy and operate as if the linear model is correct.

3. Even if we can perfectly predict f(x), the Y value cannot be perfectly predicted due to ε

prediction interval is alway wider than confidence intervals.

3.3 Other Considerations in the Regression Model

3.3.1 Qualitative Predictors

In practice, there exists qualitative(categorical) predictors. (e.g gender, status, ethnicity)

Predictors with only 2 levels

If a qualitative predictor only has two levels(possible values). We simply create an indicator or dummy variable. takes on two possible numerical values.

e.g dummy variable of gender.

$x_i = 1$ if ith person is female

$= 0$ if ith person is male.

we use this variable as a predictor in the regression equation.

$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \beta_0 + \beta_1 + \epsilon_i$ if ith person is female

$= \beta_0 + \epsilon_i$ if ith person is male.

It does not matter how we encode the categorical variables. if male = 1 and female = 0. There is no difference in terms of regression fit.

However, there will be interpretation difference for β.

Alternatively, we could also encode gender as: female = 1, male = -1

Predictors with more than 2 levels

When there are more than 2 possible values of a qualitative predictor, a single dummy variable cannot represent al possible values.

we need to create addition dummy variable.

e.g ethnicity: Asian, Caucasian, African American

$x_i1 = 1$ (if ith person is Asian)

$= 0$ (if ith person is not Asian)

$x_i2 = 1$ (if ith person is Caucasian)

$= 0$ (if ith person is not Caucasian)

then the regression equation becomes:

$y_i = \beta_0 + \beta_1 x_i1 + \beta_2 xi2 + \epsilon$

There will always be one fewer, dummy variable than the number of levels.

3.3.2 Extensions of the Linear Model

The standard linear regression model provides interpretable results and works quite well on many real world problems.

However, it makes several highly restrictive assumptions that are often violated in practice.

2 important assumptions

1. the predictors and response are additive:

   predictors are independent of each other.

2. the predictor and response are linear:

   the change in Y cause by 1 unit of X is constant.

Removing the Additive Assumption

In our previous analysis of the Advertising data. We assumed that the effect effect on sales of increasing one ad-medium is independent of other ad-mediums. (The effect of a variable is constant)

However, this simple model may be incorrect. (notice, when levels of either TV or radio are low. the true sales are lower than predicted by the linear model.) a value of X can alter the effect (β).

In marketing, this is called synergy effect.

In statistics, this is called interaction effect.

We can account for this interaction effect by adding an *interaction term*.

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

$= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon$

$= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon$

Since, $\tilde{\beta}_1 = \beta_1 + \beta_3 X_2$, it changes with X_2, the effect of X_1 on Y is no longer constant.

	Coefficient	Std. error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

Adding a interaction term can increase the performance of a model. $R^2 = 0.897 \rightarrow R^2 = 0.968$

This mean that $(96.8 - 0.897)/(100 - 0.897) = 69\%$ of the variability in sales that remains after fitting the additive model has been explained by the interaction term.

the p-value of the interactive term suggest that $\beta_3 \neq 0$, in other words, it is clear that the true relationship is not additive.

Note: Sometimes the p-value of individual Xs can be large. However, their interaction term's p-value can be small.

Then we must include both the Xs in the linear model.

The concept of interactions can also be applied to qualitative variables.

In fact the interaction between a qualitative and a quantitative variable is particularly nice.

e.g

$$\begin{aligned}\texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases}\beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student}\end{cases} \\ &= \beta_1 \times \texttt{income}_i + \begin{cases}\beta_0 + \beta_2 & \text{if } i\text{th person is a student} \\ \beta_0 & \text{if } i\text{th person is not a student.}\end{cases}\end{aligned}$$

MLR with categorical variable without interaction term. (it yields 2 parallel lines)

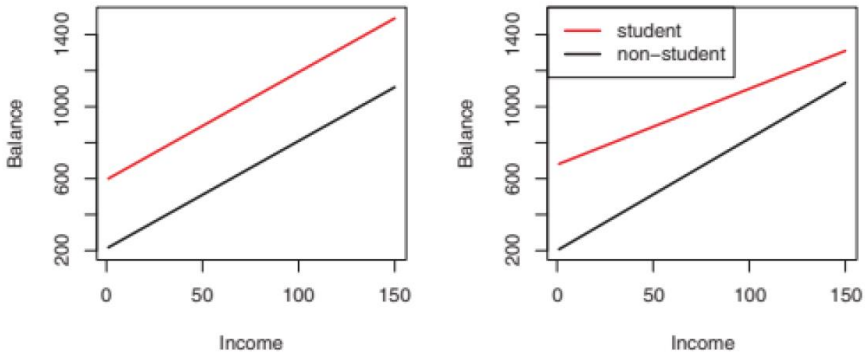
$$\begin{aligned}\texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases}\beta_2 + \beta_3 \times \texttt{income}_i & \text{if student} \\ 0 & \text{if not student}\end{cases} \\ &= \begin{cases}(\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \texttt{income}_i & \text{if student} \\ \beta_0 + \beta_1 \times \texttt{income}_i & \text{if not student}\end{cases}\end{aligned} \tag{3.35}$$

MLR with categorical variable and with interaction term.

Likewise, there are two different regression lines for the student and non-student.

However, the two lines have different intercept as well as different slopes.

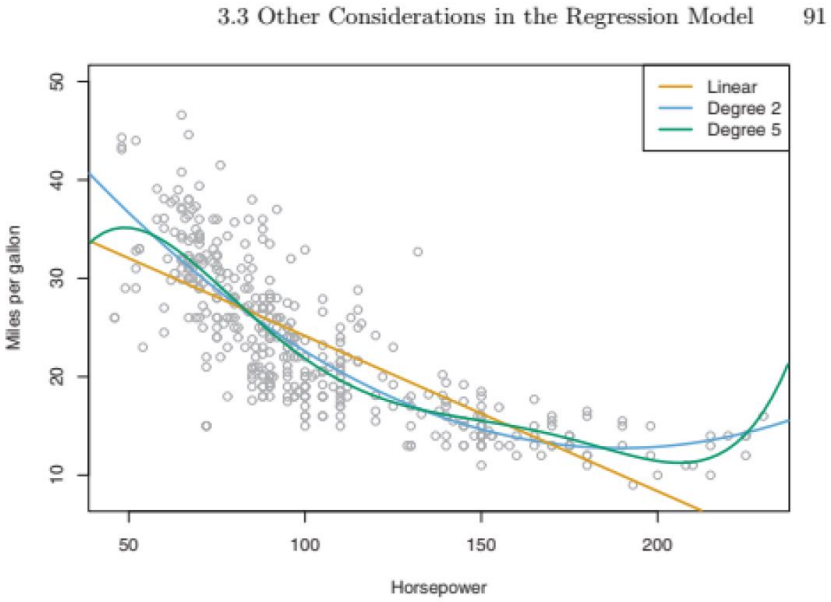
This allows for the change in income(Y) to be different among student and non-students.



Non-linear Relationship

The linear regression model assumes a linear relationship between Y and Xs.

e.g



The example above suggest that the relationship between Miles/gallon and Horsepower is nonlinear.

A simple approach to deal with non-linear relation is to include transformed versions of the predictors in the model.

example)

$mpg = \beta_0 + \beta_1 horsepower + \beta_2 horsepower^2 + \epsilon$

polynomial regression methods will be discussed in **Chapter 7**

3.3.3 Potential Problems

When we fit a linear regression model to a particular data set, many problems occur.

These are the 6 most common problems:

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

Note: overcoming these problems is more art-like than scientific.

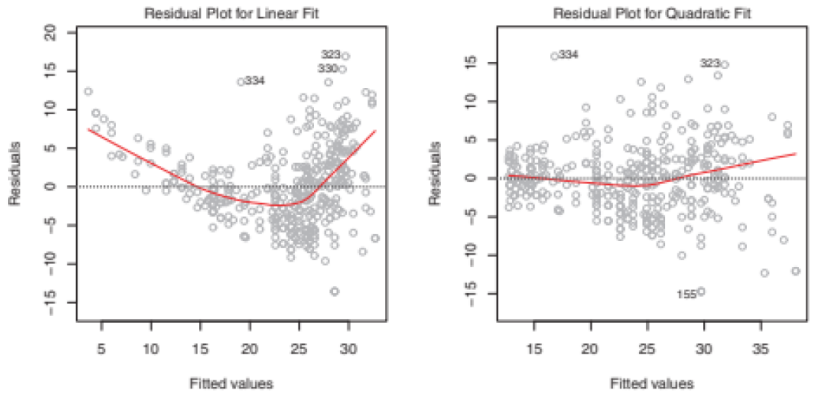
1.Non-linearity of the Data

If the true relationship of the response and the predictors are far from linear.

The insight we gained from the regression model is uncertain.

Also, the prediction accuracy of the model is significantly reduced.

Residual plots can be used to identify non-linearity.



The residual plot of mpg and horsepower. (left is vanilla simple linear regression, right is regression with $horsepower^2$ term.

The U shape of the residual plot suggests there is non-linear associations.

The transformed residual plot seems to have reduced this effect.

Note: there are more than 1 Fitted values for MLR.

Therefore in MLR we plot the residual e *versus* $\hat{y}_i$

When non-linear association is suggested by the residual plot. Then, we can use simple transformations of the predictors (*e.g* $logX, \sqrt{X}, X^2$)

More will be discussed in Chapter 7.

---

2. Correlation of Error Terms

linear regression assumes that the error terms $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are uncorrelated.

that is the sign of the value of ϵ_i has no effect sign of the ϵ_{i+1}.

If there is correlation among the error terms, the estimated standard error will tend to underestimate the true standard error.

As a result, confidence and prediction intervals will be narrower than they should be.

e.g 95% confidence interval may in reality have a much lower probability than 0.95 of containing the true value of the parameter. This may lead us to erroneously conclude that a parameter is statistically significant.

Correlations among error terms frequently occur in the context of time series data. (It occurs in other places as well)

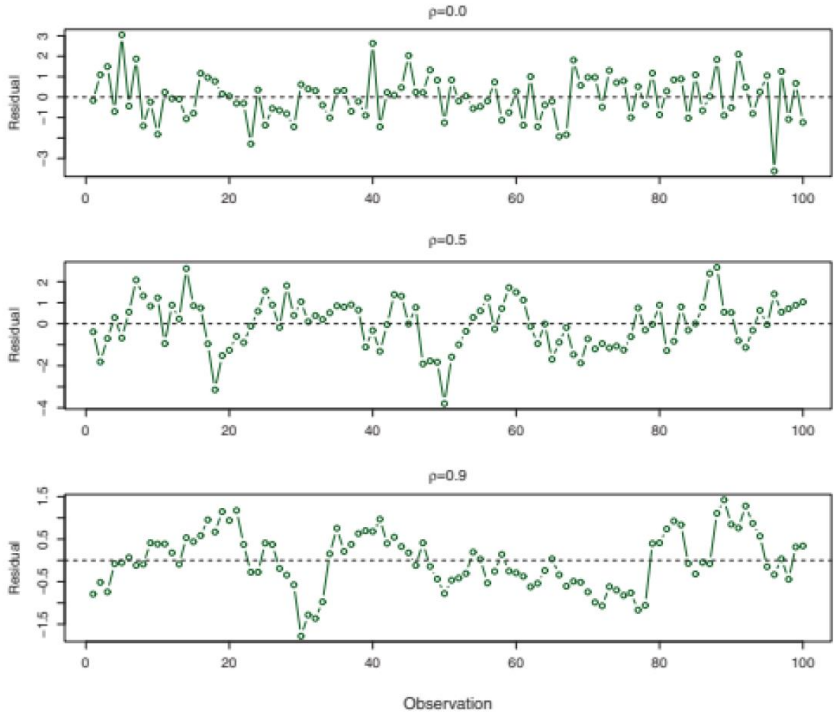


FIGURE 3.10. *Plots of residuals from simulated time series data sets generated with differing levels of correlation ρ between error terms for adjacent time points.*

Plotting residuals with observation:

1. check for tracking, that is if adjacent residual have similar values.

Many methods have been developed to properly take account of correlations in the error terms in time series data.

---

3. Non-constant variance of error terms.

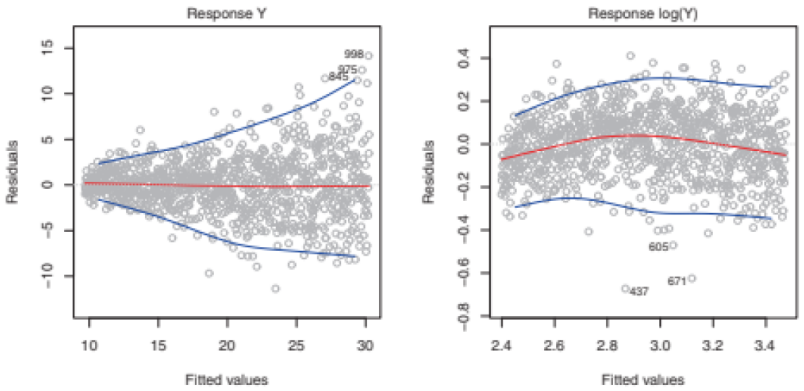
Linear regression model also assumes the error term to have a constant variance.

that is, $Var(\epsilon_i) = \sigma^2$

standard error(RSE), confidence intervals, and hypothesis tests used in linear model rely on this assumption.

However, often the variances of the error terms are non-constant.

We can identify non-constant variances in the errors, (*heteroscedasticity*), by looking at the residual plot.



The funnel shape indicates heteroscedasticity (not constant residual).

We can solve this problem by transforming Y using a concave function such as $logY$ or $\sqrt{Y}$.

(this results a greater amount of shrinkage of the larger responses.)

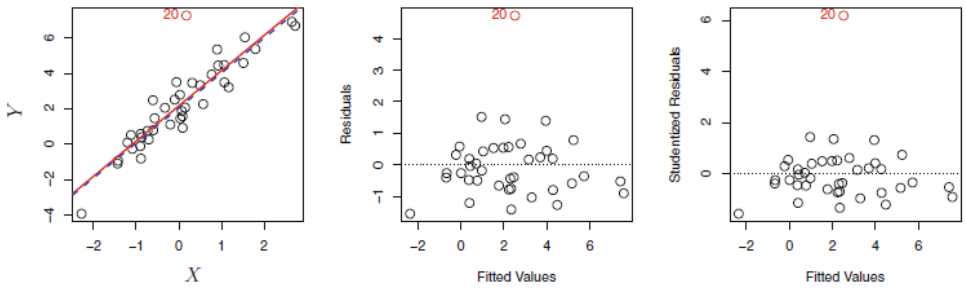
This can be confirmed on the left plot.

---

4. Outliers

An outlier is a point, where the y_i value is far from the value predicted by the model.

this might occur for a variety of reason. (incorrect recording, missing 0, etc...)



left plot:

The red point in the left panel illustrates a typical outlier.

Red line is is the regression model. and the red point is a outlier.

(Blue line)In this case, removing the outlier does not have a big effect on the regression line. (this is typical)

Caution: removing an outlier will increase R^2 or decrease RSE.

Center plot:

Residual plots can be used to identify outliers.

But in practice, it is difficult to decide how large does a residual has to be.

Right plot:

To solve the problem of residual plot. we plot the studentized residuals.

(simply divide each e_i by its estimated standard error.)

The obs where the studentized residual exceed 3 in absolute value is classified as outliers.

The red dot is clearly an outlier by this standard

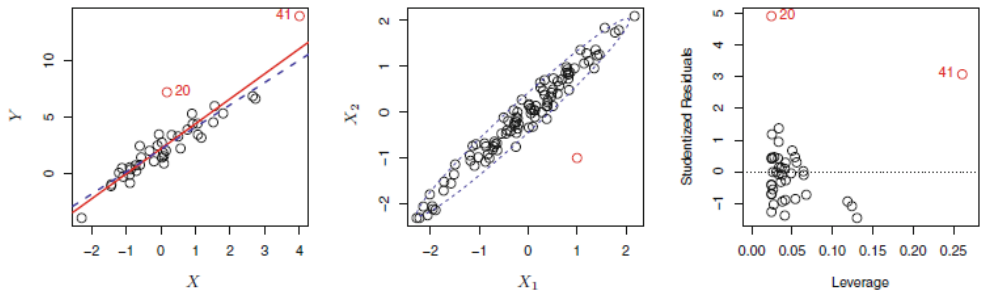
We can simply delete the outliers, if we believe them to be measurement errors.

However, be careful. an outlier may indicate problems within the model. (missing predictor. etc...)

---

5. High-leverage points.

High-leverage points are points with unusual x_i values.



Left picture:

Obs 41 is a high leverage point. Removing it has a large impact on the regression line.

Removing a high leverage point $\rightarrow$ big impact on regression line.

Center picture:

In SLR, a point with unusual X value is the high leverage point.

But, in MLR it is less obvious.

In the picture, by plotting X_1, X_2. we identify the red dots as the high leverage point. unusual in terms of the full set of predictors.

Note: its X_1, X_2 range is acceptable.

This is a problem, we cannot always plot all Xs in MLR.

to quantify an obs' leverage, compute the *leverage statistic*:

$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}$ (leverage statistic for single predictor) $1/n < h_i < 1$

There is a simple extension for multiple predictors (not provided in book)

if h_i greatly exceeds (p+1)/n, we suspect that the corresponding obs has high leverage.

Right picture:

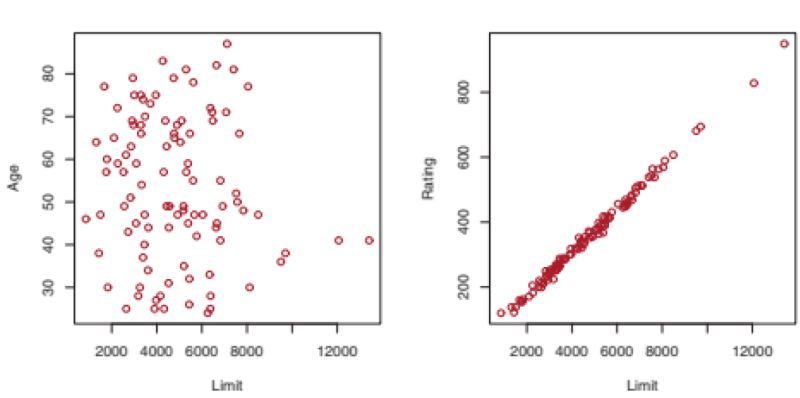
Studentized residual vs h_i.

obs 41 has high values of sr and h_i. (both an outlier and a high leverage point)

---

6. Collinearity

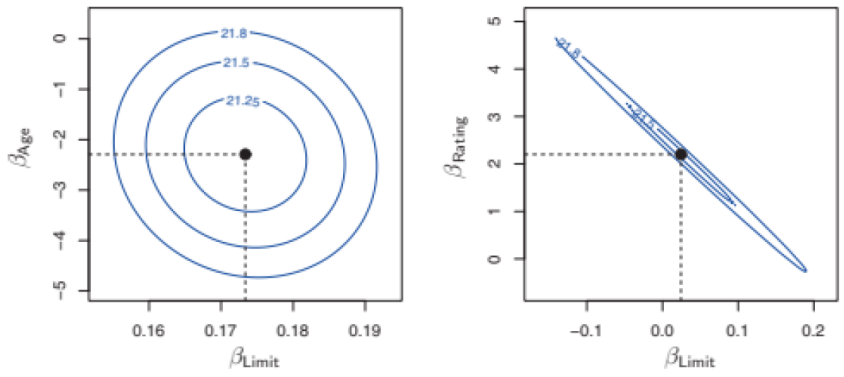
Collinearity is the situation where two or more predictor variables are closely related to one another.



Left: no obvious relation. Right: collinear.

The presence of collinearity can be a problem in linear regression.

Since, X_1 and X_2 increase and decrease together. it is difficult to determine how each one is separately associated with Y.



Right: contour plot of collinear Xs.

Each ellipse represents the set of βs that yields the same value of RSS. closer to the center the smaller the RSS.

collinearity exists $\rightarrow$ small change in the data could cause the optimal coefficient values to move anywhere along this long valley.

(because, the range of βs that yields the same RSS is now much larger)

Notice, that the β_{Limit} 's range has increased significantly.

		Coefficient	Std. error	t-statistic	p-value
Model 1	Intercept	−173.411	43.828	−3.957	< 0.0001
	age	−2.292	0.672	−3.407	0.0007
	limit	0.173	0.005	34.496	< 0.0001
Model 2	Intercept	−377.537	45.254	−8.343	< 0.0001
	rating	2.202	0.952	2.312	0.0213
	limit	0.025	0.064	0.384	0.7012

TABLE 3.11. *The results for two multiple regression models involving the* `Credit` *data set are shown. Model 1 is a regression of* `balance` *on* `age` *and* `limit`*, and Model 2 a regression of* `balance` *on* `rating` *and* `limit`*. The standard error of* $\hat{\beta}_{\text{limit}}$ *increases 12-fold in the second regression, due to collinearity.*

collinearity increase the standard error for $\hat{\beta}_j$ to increase.

collinearity also decrease the power of t-tests. (probability of correctly detecting a non-zero coefficient is reduced) (t = $\hat{\beta}_j / SE$)

We might make wrong decisions (picture above p-value)

A simple way to detect collinearity $\rightarrow$ correlation matrix. (However, it cannot detect multicollinearity)

A better way to assess multicollinearity: *vairance inflation factor* (VIF)

$$\text{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}} \geq 1 \text{ ; } R^2 \text{ here: } Y \text{ is } X_j \text{ , } Xs \text{ are other } Xs$$

VIF is the ratio of the variance of $\hat{\beta}_j$ in full model devided by variance of $\hat{\beta}_j$ when it is fit on its own.

$VIF = 1$ indicate the complete absence of collinearity. (does not happen in practice)

Rule of thumb:

if $VIF \geq 5,\ or \geq 10 \rightarrow$ problematic collinearity.

When faced with collinearity.

1. drop one of the problematic variables.
2. combine the collinear variables. (take average)

3.5 Comparison of Linear Regression with K-NN

Linear regression is a *parametric* approach. (assumes a linear form of $f(X)$)

adv: easy to fit. (only need to estimate a small number of coefficients)

(linear regression)easy to interpret

statistical tests can be easily performed

dis-adv:

make strong assumptions about the form of $f(X)$, if not true $\rightarrow$ low performance.

non-parametric methods:

no assumption about the form of $f(X)$.

more flexible.

e.g K-nearest neighbors regression. (KNN regression)

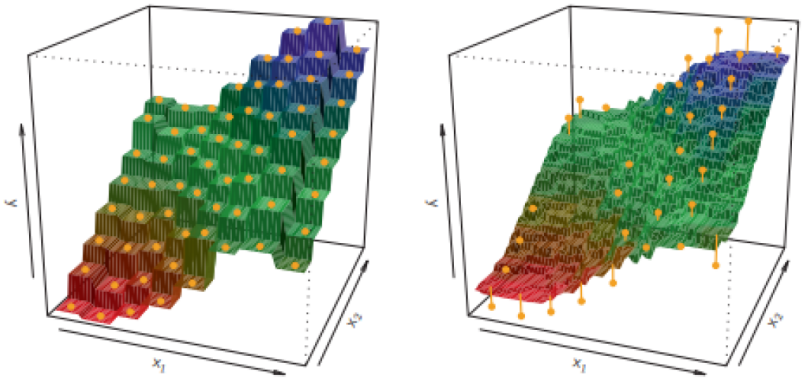


FIGURE 3.16. *Plots of $\hat{f}(X)$ using KNN regression on a two-dimensional data set with 64 observations (orange dots).* Left: *$K = 1$ results in a rough step function fit.* Right: *$K = 9$ produces a much smoother fit.*

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in \mathcal{N}_0} y_i.$$

the parametric approach will outperform the nonparametric approach if the parametric form that has been selected is close to the true form of f.

Also, when there are noise, parametric methods will perform better.

KNN-regression is vulnerable to high dimensional data.

Note: Even when dimensions are low, KNN is hard to interpret. Therefore, linear regression can be preferred.

Generally, parametric methods will tend to outperform non-parametric approaches when there is a small number of observation per predictor.