simple linear

We define some sum of squares as follows
SST (T for total): $SST = SS_{yy} = \sum(Y_i - \bar{Y})^2 = \sum Y_i^2 - \frac{1}{n}\sum Y_i^2$
SSR: $\sum(\hat{Y}_i - \bar{Y})^2 = \sum(\bar{Y} - b_1\bar{X} + b_1X_i - \bar{Y})^2 = b_1^2 SS_{XX}$
SSE (E for error): $\sum(Y_i - \hat{Y}_i)^2 = SST - SSR$
GaussMarkov theorem, states: Under the conditions of [[Simple Linear Regression Model]], the least squares estimators b_0 and b_1 are unbiased and have minimum variance among all unbiased linear estimators.
let

$$k_i = \frac{X_i - \bar{X}}{\sum(X_i - \bar{X})^2} \to \sum k_i^2 = \frac{\sum(X_i - \bar{X})^2}{(\sum(X_i - \bar{X})^2)^2} = SS_{XX}$$

then

$$b_1 = \sum k_i(Y_i - \bar{Y}) = \sum k_i Y_i$$

hence

$$\text{Var}[b_1] = \text{Var}\left[\sum k_i Y_i\right] = \sum k_i^2 \text{Var}[Y_i] = \frac{\sigma^2}{SS_{XX}}$$

which can be estimated by

$$s^2(b_1) = s^2/SS_{XX} = (SSE/n-2)(1/SS_{XX})$$

where s^2 is the unbiased estimator of σ^2
Since b_1 is normally distributed, then the statistic

$$b_1 - \beta_1/s(b_1) \sim t_{n-2}$$

Hence the confidence interval of $1-\alpha$ is

$$b_1 - s(b_1)t_{n-2,\alpha/2} < \beta_1 < b_1 + s(b_1)t_{n-2,\alpha/2}$$

$$\text{Var}[b_0] = \text{Var}[\bar{Y} - b_1\bar{X}] = \frac{\sigma^2}{n} + \frac{\bar{X}^2\sigma^2}{S_{XX}} = \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right)$$

and the estimator is

$$s^2(b_0) = \frac{SSE}{n-2}\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right)$$

similarly,

$$(b_0 - \beta_0)/s(b_0) \sim t(n-2)$$

hence the confidence interval is $b_0 - s(b_0)t_{n-2,\alpha/2} < \beta_0 < b_1 + s(b_0)t_{n-2,\alpha/2}$
If ε is not exactly normally distributed but not depart seriously, then b estimation will be approximately normal.
Even if the distri of Y are far from normal, the estimators b_0, b_1 generally have the property of **asymptotic normallity as sample increase**
Thus, with sufficiently large samples, the confidence interval and decision rules given earlier still apply even if the probability distributions of Y depart far from normality

$$\mathbb{E}[\hat{Y}] = \mathbb{E}[b_0 + b_1X] = \beta_0 + \beta_1X$$

$$\text{Var}[\hat{Y}] = \text{Var}[b_1(X - \bar{X}) + \bar{Y}] = \sigma^2\left(\frac{1}{n} + \frac{(X-\bar{X})^2}{S_{XX}}\right)$$

The estimated variance is

$$s^2(\hat{Y}_i) = \frac{SSE}{n-2}\left(\frac{1}{n} + \frac{(X_i - \bar{X})^2}{S_{XX}}\right)$$

The confidence interval is then

$$\hat{Y}_i - s(\hat{Y}_i)t_{n-2,\alpha/2} < \mathbb{E}[\hat{Y}_i] < \hat{Y}_i + s(\hat{Y}_i)t_{n-2,\alpha/2}$$

If we know the parameters of β_0, β_1, σ, then with observation X_i, Y_i, we can determine how Y_i is distributed and where is its mean value, then we can use confidence interval to determine whether accept Y_i
When the regression parameters are unknown, they must be estimated. Thus the position of Y_i needed to be estimated by $\hat{Y}_i$, and we will get two boundary limits. Based on these two limits, we separately determine the confidence interval and combine together. Then we get our prediction interval, which is larger than confidence interval
In theorem, we have

$$(Y_i - \hat{Y}_i)/s(\hat{Y}_i) \sim t_{n-2}$$

then the variance for predict number is

$$\sigma^2(pred) = \sigma^2(Y_i) + \sigma^2(\hat{Y}_i) = \sigma^2\left(1 + \frac{1}{n} + \frac{(X-\bar{X})^2}{S_{XX}}\right)$$

which consists of two parts 1. The variance of the distribution of Y_i, which is σ^2 2. The variance of the sampling distribution of $\hat{Y}_i$, which is $\sigma^2\left(\frac{1}{n} + \frac{(X-\bar{X})^2}{S_{XX}}\right)$
the prediciton interval of Y_i is

$$\hat{Y}_i - s(Y_{pred})t_{n-2,\alpha/2} < pred < \hat{Y}_i + s(Y_{pred})t_{n-2,\alpha/2}$$

Prediction intervals resemble confidence intervals. However, they differ conceptually. A confidence interval represents an inference on a parameter and is an interval that is intended to cover the value of the parameter. A prediction interval, on other hand, is a statement about the value to be taken by a random variable, the new observation Y_i

R

The R^2, called coefficient of determination, provides a summary measure of how well the regression line fits the sample. R^2 is the proportion of the variability in the dependent variable that is explained by the independent variable The additional independent variables may not contribute significantly to the explanation of the dependent variable Y, but they do increase R^2. Adding more independent variables in the regression equation for the purpose of increasing R^2 often results in overfitting and result in worse models rather than better ones.
To help prevent overfitting in regression analysis, we use the so called Adjusted R Square
R_a^2 is defined by

$$R_a^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} = 1 - \frac{n-1}{n-k-1}(1-R^2)$$

$R_a^2 = 48.3\%$ means that approximately 48.3% of the total variation in the values of Ycan be explained by a linear relationship with independent variables after adjusting the number of independent variables

rho

Usually we treat X as fixed value, but some time it can also be trated as rvs.Since both X and Y are rvs, then we need to consider their correlation or covariance
The correlation coefficient is defined by

$$\rho = \sigma_{XY}/\sigma_X\sigma_Y$$

Note that ρ measures only linear relationship. The variables may be perfectly correlated in a curvilinear relationship
The sample correlation coefficient is defined by

$$r = (\sum(X_i - \bar{X})(Y_i - \bar{Y}))/\sqrt{\sum(X_i - \bar{X})^2\sum(Y_i - \bar{Y})^2}$$

In regression, $r^2 = R^2$ is simply a measure of closeness of fit with $H_0 : \rho = 0$, The statistic is

$$t = r/\sqrt{(1-r^2)(n-2)}$$

residual

we need to analyze e_i to check whether Gauss-Markov Theorem is satisfied. (zero mean, constant variance, no correlation and $\sum X_i e_i = 0$)
Scatter plots can also be used to detect whether the assumption of **constant variance** of Y for all values of X is being violated. If residuals increases with X or Y, then the assumption of **homogeneity** of variance is being violated
When the sample size is large in comparison to the number of parameters in the regression model, the **dependency effect** among the residuals e_i is relatively unimportant and can be ignored for most purposes
We have many things to test residual
1.**Independence**: We use Durbin-Watson Test to test autocorrelation. Usually this assumption is relative easy to meet since observations appear in a random position, and hence successive error terms are also likely to be random. However, in time series data or repeated measures data, this problem of dependence between successive error terms often occurs.
2.**Normality**: we use, Shapiro-Wilk Test, Chi-Square Test, Kolmogorov Test. and we can draw box-plot, histgram, P-P plot
3.**Equality**: we can use Runs Test, If the spread of the residuals increases or decreases with the values of the independent variable or with the predicted values, then the assumption of homogeneity of variance is being violated.
4.**linearity**: we use Brown-Forsythe Test, Cook-Weiberg(Breusch-Pegan) Test. If the assumptions of linearity and homogeneity of variance are met, there should be no relationship between the predicted and residual values, i.e. the residuals should be randomly distributed around the horizontal line through zero.
5.**Outliers**: we use F-test. Test whether a linear regression function is a good fit for the data

ANOVA

Source of Variation	SS	df	MS
Regression	$SSR = \mathbf{b'X'Y} - \left(\frac{1}{n}\right)\mathbf{Y'JY}$	$p-1$	$MSR = \frac{SSR}{p-1}$
Error	$SSE = \mathbf{Y'Y} - \mathbf{b'X'Y}$	$n-p$	$MSE = \frac{SSE}{n-p}$
Total	$SSTO = \mathbf{Y'Y} - \left(\frac{1}{n}\right)\mathbf{Y'JY}$	$n-1$	

suppose there are k variables and n samples, then DF of SSR is k, DF of SSE is $n-k-1$, DF of SST is $n-1$

Extra Sum of Squares and Coefficients of Partial Determination

one can view an extra sum of squares as measuring the marginal increase in the regression sum of squares when one or several predictor variables are added to the regression model The extra sum of squares $SSR(X_2|X_1)$ equivalently can be viewed as the marginal increase in the regression sum of squares:

$$SSR(X_2|X_1) = SSR(X_1, X_2) - SSR(X_1)$$

since $SST = SSR - SSE$, the equation is equivalent to

$$SSR(X_2|X_1) = SSE(X_1) - SSE(X_1, X_2)$$

A coefficient of partial determination, in contrast, measures the marginal contribution of one X variable when all others are already included in the model
Suppose $SSE(X_1)$ measures the variation in Y when X_1 is included in the model. $SSE(X_1, X_2)$ measures the variation in Y when both X_1, X_2 is included in the model. Then the relative marginal reduction in the variation in Y associated with X_2 when X_1 is already in the model is

$$r^2_{Y_{2|1}} = \frac{SSR(X_2|X_1)}{SSE(X_1)} = \frac{SSE(X_1) - SSE(X_1, X_2)}{SSE(X_1)}$$

The above is the coefficient of partial determination between Y and X_2, given that X_1 is in the model
That is to say, for a given model with X_1, the unexplained part is $SSE(X_1)$, now we add a new variable X_2, then the new explained part is $SSR(X_2|X_1)$. Naturally, the coefficient $r^2_{Y_{2|1}}$ is defined

Partial F test

Given two model
$Y = \beta_0 + \beta_1X_1 + \cdots + \beta_gX_g + \varepsilon$ reduced form
$Y = \beta_0 + \beta_1X_1 + \cdots + \beta_gX_g + \beta_{g+1}X_{g+1} + \cdots + \beta_kX_k + \varepsilon$ complete
We with to test two hypothesis

$$H_0 : \beta_{g+1} = \cdots = \beta_k = 0 \qquad H_1 : H_1 \text{ is not true}$$

Then the partial F-Test statistic is

$$F = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/(n-k-1)} = \frac{MSR(X_{g+1}, \ldots X_k|X_1, \ldots, X_g)}{MSE}$$

where $F \sim F(k-g, n-k-1)$, - SSE_R: sum of squared errors for the reduced model - SSE_C: sum of squared errors for the complete model - MSE_C: mean square error for the complete model - $k-g$: number of β parameters tested in H_0

Multicollinearity

When the predictor variables are correlated among themselves, **intercorrelation** or **multicollinearity** among them is said to exist
First, high correlation among the independent variables increase the likelihood of rounding errors in the calculations of the β_i estimates, standard errors; **Second**, and more important, the regression results may be confusing and misleading
In general, when two or more predictor variables are uncorrelated, the marginal contribution of one predictor variable in reducing the error sum of squares when the other predictor variables are in the model is exactly the same as when this predictor variable is in the model alone
That is to say

$$SSR(X_1|X_2) = SSE(X_2) - SSE(X_1, X_2) = SSR(X_1)$$

$SSR(X_1)$ is the part explained by X_1, while $SSR(X_1|X_2)$ is the the part explained by X_1 when X_2 is already in the model. If the two variables are uncorrelated, then the explained part should be equal no matter how many othere predictor variables are already in the model
Methods to detect multicollinearity:
1. see if there are Significant correlation between pairs of independent variables in the model
2. Nonsignificant t tests for all (or nearly all) of the individual β parameters when the F-test for overall model adequacy
3. VIF method
4. A more sophisticated method is to use Principal Components Analysis
One of the commonly used simple methods to solve the multicollinearity is to drop one or more of the highly correlated independent variables from the multiple regression model.

Responsibility

By drawing scatter plots, histograms, PP graphs, etc., we can visually observe whether the residuals meet various assumptions.
Using Form methods(various test) to test the assumption of residual

Model Selection

The The R_p^2 criterion is equivalent to using the error sum of squares SSE_p as the criterion

$$R_p^2 = 1 - SSE_p/SST$$

where SST is constant for all possible regression models
When we add additional X variables, the R_p^2 will increase, so when the model contains all potential $P-1$ variables, it is maximum. R_p^2's aim is to find the point where adding more variables is not **worthwhile** because it leads to a **very small increase**
The $R_{a,p}^2$ or MSE_p criterion is

$$R_{a,p}^2 = 1 - \frac{n-1}{n-p}\frac{SSE_p}{SST} = 1 - \frac{MSE_p}{SST/(n-1)}$$

Users of the $R_{a,p}^2$ criterion seek to find a few subsets for which $R_{a,p}^2$ is at the maximum or so close to the maximum that adding more variables is not worthwhile
The Γ_p is the Total Mean Squared Error divided by σ^2, then

$$\Gamma_p = \frac{1}{\sigma^2}\left[\sum \mathbb{E}[\hat{Y}_i] - \mu_i)^2 + \sum \sigma^2(\hat{Y}_i)\right]$$

The model which includes all $P-1$ potential X variables is assumed to have been carefully chosen so that $MSE(X_1, ..., X_{P-1})$ is an unbiased estimator of σ^2. It can then be shown that an **estimator** of Γ_p is C_p

$$C_p = \frac{SSE_p}{MSE(X_1, \ldots, X_{P-1})} - (n - 2p)$$

When there is no bias in the regression model with $P-1$ X variables so that $\mathbb{E}[\hat{Y}_i] = \mu_i$, the expected value of C_p is approximately p:

$$\mathbb{E}[C_p] \approx p \qquad \text{when } \mathbb{E}[\hat{Y}_i] = \mu_i$$

Thus, when the C_p values for all possible regression models are plotted against p, those models with little bias will tend to fall near the line $C_p = p$. Models with substantial bias will tend to fall considerably above this line
In using the C_p criterion, we seek to identify subsets of X variables for which
1. the C_p value is small
2. the C_p value is near p
The two criterions are defined by
$AIC_p = n \ln SSE_p - n \ln n + 2p \quad SBC_p = n \ln SSE_p - n \ln n + p \ln n$
For first term SSE_p, it will decrease as p increase, the 2nd term $n \ln n$ is fixed which only rely on the sample size. the 3rd term increase as p increase.
So small AIC or SBC perform better since we have low SSE and low p
The PRESS prediction error for the ith case is

$$Y_i - \hat{Y}_{i(i)}$$

and the $PRESS_p$ criterion is the sum of the squared prediction errors over all n cases

$$PRESS_p = \sum_{i=1}^{n}(Y_i - \hat{Y}_{i(i)})^2$$

Forward Selection
- Independent variables should be inserted one at a time until a satisfactory regression equation is found.
- The procedure starts with the independent variable that has the highest correlation with the response variable y.
- Once the independent variable is in the model, it stays.
- At the second step, the remaining $k-1$ variables are examined, and the variable for which the partial F-statistic is a maximum is added to the equation.
- The procedure goes on until there are no remaining independent variables that significantly increase R^2.

Partial Regression Plot

Added-variable plots, also called partial regression plots and adjusted variable plots, are refined residual plots that provide graphic information about the marginal importance of a predictor variable X_k , given the other predictor variables already in the model
Consider two variable X_1, X_2, Suppose we are concerned about the nature of the regression effect for X_1, given that X_2 is already in the model. We regress Y on X_2 and obtain the fitted values and residual

$$\hat{Y}_i(X_2) = b_0 + b_2 X_2 \quad e_i(Y|X_2) = Y_i - \hat{X}_{i1}(X_2)$$

We also regress X_1 on X_2 and obtain

$$\hat{X}_{i1}(X_2) = b_0^{\#} + b_2^{\#} X_{i2} \quad e_i(X_1|X_2) = X_{i1} - \hat{X}_{i1}(X_2)$$

The added-variable plot for predictor variable X_1 consists of a plot of the Y residuals $e(Y|X_2)$ against the X1 residuals $e(X_1|X_2)$.
The regression beta of $e(Y|X_2) = \beta e(X_1|X_2)$is just the beta of $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, i.e. $\beta = \beta_2$

Outlier Y

Recall PRESS, we define deleted residual as

$$d_i = Y_i - \hat{Y}_{i(i)}$$

An algebraically equivalent expression for di that does not require a recomputation of the fitted regression function omitting the ith case is:

$$d_i = e_i/(1 - h_{ii})$$

If Y_i is an outlier, then d_i will have large value, since the regreesion line will far form Y_i
Based on Deleted Residual and Studentized Residual, **Studentized Deleted Residual** is defined by

$$t_i = d_i/s(d_i)$$

where

$$s(d_i) = \sqrt{MSE_i/(1 - h_{ii})}$$

and we have relation

$$(n-p)MSE = (n-p-1)MSE_i + e_i^2/(1 - h_{ii})$$

Then

$$t_i = e_i\sqrt{\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2}}$$

We identify as outlying Y observations those cases whose Studentized Deleted Residual are large in absolute value
The formal test of outliers is Bonferroni Test, The appropriate Bonferroni critical value therefore is

$$t(1 - \alpha/2n; n - p - 1)$$

Note that the test is two-sided since we are not concerned with the direction of the residuals but only with their absolute values
The **Hat Matrix** is defined by

$$H = X(X^{\mathrm{T}}X)X^{\mathrm{T}}$$

hence in Multiple Linear Regression Model,

$$\hat{Y} = HY$$

$$0 \le h_{ii} \le 1 \qquad \sum h_{ii} = p$$

where p is the number of regression parameters.
In addition, it can be shown that h_{ii} is a measure of the distance between X for the ith case and the means of the X values for all n cases. if h_{ii} is large, it indicates ith case is distant from the centre of X.
If the ith case is outlying in terms of X and therefore has a large **leverage value** h_{ii}, it exercise substantial leverage in determining the fitted value $\hat{Y}_i$. This is because
1. $\hat{Y} = HY$, h_{ii} is the weight of Y_i in determining fitted value, the larger is h_{ii}, the more important is Y_i
2. The larger is h_{ii}, the samller is variance of e_i
A h_{ii} is considered to be large if it is more than twice as large as the mean leverage value, denoted by $\bar{h}$, which is

$$\bar{h} = \sum h_{ii}/n = p/n$$

hence, **leverage value over $2p/n$ should be regarded as outliers**

WLS

Let $W = \sigma^{-2} diag(w_1, \ldots, w_n) = diag(\sigma_1^{-2}, \ldots, \sigma_n^{-2})$, then

$$X^{\mathrm{T}}WY = (X^{\mathrm{T}}WX)b_W \implies b_W = (X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}WY$$

and

$$s^2(b_w) = MSE_w(X^{\mathrm{T}}WX)^{-1}$$

where

$$MSE_w = \left[\sum w_i(Y_i - \hat{Y}_i)^2\right]/(n-p) = \left(\sum w_i e_i^2\right)/(n-p)$$

hence $\mathrm{Var}[YW^{1/2}] = W^{1/2}\mathrm{Var}[Y]W^{1/2} = \sigma^2 I$

Outlier X

A useful measure of the influence that case i has on the fitted value $\hat{Y}_i$ is given by:

$$DFFITS_i = \hat{Y}_i - \hat{Y}_{i(i)}/\sqrt{MSE_i h_{ii}}$$

It can be shown that the **DFFITS** values can be computed by using only the results from fitting the entire data set, as follows:

$$DFFITS_i = e_i\sqrt{\left[\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2}\right]\left[\frac{h_{ii}}{1-h_{ii}}\right]} = t_i\sqrt{\left[\frac{h_{ii}}{1-h_{ii}}\right]}$$

As a guideline for identifying influential cases, we suggest considering a case influential if the absolute value of **DFFITS** exceeds 1 for small to medium data sets and $2\sqrt{p/n}$ for large data sets
In contrast to the DFFITS, which considers the influence of the ith case on the fitted value $\hat{Y}_i$ for this case, **Cook's distance** measure considers the influence of the ith case on all n fitted values. Cook's distance measure, denoted by D_i , is an aggregate influence measure, showing the effect of the ith case on all n fitted values

$$D_i = \sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(j)})^2/pMSE$$

An algebraically equivalent expression is:

$$D_i = \frac{e_i}{pMSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$

If the percentile value is less than about 10% or 20% the ith case has little influence on the fitted values. If the percentile value is near 50% or more, the fitted values obtained with and without the ith case should be considered to differ substantially, implying that the ith case has a major influence on the fit of the regression function
A measure of the influence of the ith case on each regression coefficient b_k is the difference between the estimated regression coefficient b_k based on all n cases and the regression coefficient obtained when the ith case is omitted, to be denoted by $b_{k(i)}$. When this difference is divided by an estimate of the standard deviation of bk , we obtain the measure DFBETAS:

$$DFBETAS_{k(i)} = \frac{b_k - b_{k(i)}}{\sqrt{MSE_i c_{kk}}}$$

where c_{kk} is the kth diagonal ele of $(X^{\mathrm{T}}X)^{-1}$
A large absolute value of DFBETAS indicates large impact of the ith case on the kth coefficient
As a guideline for identifying influential cases, we recommend considering a case influential if the absolute value of DFBETAS exceeds 1 for small to medium data sets and $2/\sqrt{n}$ for large data sets.

Example

i	(1) e_i	(2) h_{ii}	(3) t_i
1	−1.683	.201	−.730
2	3.643	.059	1.534
3	−3.176	.372	−1.656

two rvs, SSE = 109.95, MSE = 109.95 / (20 - 3) = 6.47

$$t_1 = -1.683\left[\frac{20-3-1}{109.95(1-.201) - (-1.683)^2}\right]^{1/2} = -.730$$

two largest $h_{33} = .372$ and $h_{15} = .333$. Both exceed the criterion of twice the mean leverage value, $\frac{2p}{n} = \frac{2(3)}{20} = .30$, and both are separated by a substantial gap from the next largest leverage values, $h_{55} = .248$ and $h_{11} = .201$. Having identified cases 3 and 15 as outlying in terms of their X values, we shall need to ascertain how influential these cases are

$$(DFFITS)_3 = -1.656\,(.372/(1 - .372))^{1/2} = -1.27$$

This value is somewhat larger than 1. However, the value is close enough to 1 that the case may not be influential enough to require remedy

$$D_3 = \frac{(-3.176)^2}{3(6.47)}\left[\frac{.372}{(1-.372)^2}\right] = .490$$

To assess the magnitude of the influence of case 3 ($D_3 = .490$), we refer to the corresponding F distribution, namely, $F(p, n-p) = F(3, 17)$. We find that .490 is the 30.6th percentile of this distribution. Hence, it appears that case 3 does influence the regression fit, but the extent of the influence may not be large enough to call for consideration of remedial measures.