1. show that $Var(\hat{y}) = 6^2 H$

solution:

for Multiple Linear Regression,

$y = X\beta + \varepsilon$ where, $E(\varepsilon) = 0$

$V(\varepsilon) = 6^2 I$

We know that in multiple linear Regression model,

$\hat{\beta} = (X'X)^{-1} X' y$

$\hat{\beta} = (X'X)^{-1} X' (X\beta + \varepsilon)$

$\hat{\beta} = \beta + (X'X)^{-1} X' \varepsilon$

H: The HAT Matrix

$H = X(X'X)^{-1} X'$

$\hat{y} = X \cdot \hat{\beta}$

$\hat{y} = X \cdot ((X'X)^{-1} X' y)$

$\hat{y} = X \cdot (X'X)^{-1} X' (X\beta + \varepsilon)$

$\hat{y} = X[\beta + (X'X)^{-1} X' \varepsilon]$

$Var(\hat{y}) = Var(X\beta + X(X'X)^{-1} X' \varepsilon)$

$= Var(X(X'X)^{-1} X' \varepsilon)$

$= (X(X'X)^{-1} X' \, Var(\varepsilon))$

$= X(X'X)^{-1} X' \cdot 6^2$

$\boxed{Var(\hat{y}) = H 6^2}$ $\because$ but $X(X'X)^{-1}X' = H$ (Hat matri

2) prove that the matrices H and I-H are symmetric and idempotent.

Solutions:

Idempotent means that a matrix multiplied by itself is equal to itself.
for example, $x'x = X$ and $xx' = X$
The Hat matrix is also a idempotent, we can solve it as follow,

$H \cdot H$ $\quad\quad \therefore H = X(X'X)^{-1}X'$

$X(X'X)^{-1}X' \cdot X(X'X)^{-1}X'$

$X(X'X)^{-1}(X'X)(X'X)^{-1}X'$

$I = (X'X)(X'X)$

$= X(X'X)^{-1} \cdot I \cdot X'$

result is $= X(X'X)^{-1}X'$

To claim that the matrices H & I-H are symmetric & Idempotent, we can write,

$(I-H)(I-H) = I(I-H) - H(I-H)$

$= I - H - H + H \cdot H$

$= I - 2H + H \cdot H$

where $H \cdot H = H$, so the above equation beco

$= I - 2H + H$

$= I - H$

so, we proved, $(I-H)(I-H) = (I-H)$

3) Show that the residuals from linear regression models can be expressed as $e = (I-H)y$ in matrix form, and derive $E(e) = ?$ and $Var(e) = ?$

Solution :

Residuals is defined as the difference between observed value y_i and its fitted value $\hat{y}_i$

$\therefore$ where e_i = residual.

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

$$= y_i - (\hat{\beta}_0 + \hat{\beta}_i x_i) \quad \dots \quad i = 1, 2, \dots n$$

In matrix form,

$$e = (y - \hat{y})$$

$$= (y - X\hat{\beta})$$

$$= (y - X(X'X)^{-1}X'y)$$

$$= (y - Hy)$$

$$e = (1-H)y$$

Mean and variance of residuals,

We have, $E(\varepsilon) = 0$, $V(\varepsilon) = 6^2 I$

1) $E(\varepsilon) = [E(1-H)y]$

$$= (1-H)X\beta$$

$$= X\beta - HX\beta \qquad \therefore HX = X$$

$$= X\beta - X\beta$$

$E(\varepsilon) = 0$ ..... proved.

2) $V(e) = [V(1-H)y]$ where $\varepsilon = (1-H)y$

$$= (1-H)\,V(y)\,(1-H)'$$

$$= (1-H)\,\sigma^2 \qquad \therefore V(y) = \sigma^2$$

$$V(e) = \sigma^2 (I-H)$$

Q.3.1 Consider the National Football League data in Table B.1.

a. Fit a multiple linear regression model relating the number of games won to the

team's passing yardage (x2), the percentage of rushing plays (x7), and the opponents' yards rushing (x8).
b. Construct the analysis-of-variance table and test for significance of a regression.
c. Calculate t statistics for testing the hypotheses H0: β2 = 0, H0: β7 = 0, and H0: β8 = 0. What conclusions can you draw about the roles the variables x2, x7, and x8 the model?
d. Calculate R2 and for this model.
e. Using the partial F test, determine the contribution of x7 to the model. How is this partial F statistic related to the t-test for β7 calculated in part c above?

Solution:
a. Fit a multiple linear regression model relating the number of games won to the team's passing yardage (x2), the percentage of rushing plays (x7), and the opponents' yards rushing (x8).

Regression Equation

y = -1.81 + 0.003598 x2 + 0.1940 x7 - 0.00482 x8

b. Construct the analysis-of-variance table and test for significance of a regression.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	257.09	85.698	29.44	0.000
x2	1	78.03	78.028	26.80	0.000
x7	1	14.07	14.068	4.83	0.038
x8	1	41.40	41.400	14.22	0.001
Error	24	69.87	2.911		
Total	27	326.96			

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-1.81	7.90	-0.23	0.821	
x2	0.003598	0.000695	5.18	0.000	1.12
x7	0.1940	0.0882	2.20	0.038	2.10
x8	-0.00482	0.00128	-3.77	0.001	2.02

Since the p-value <0.05, we reject the Null hypothesis. Therefore the regression is significant.

c. Calculate t statistics for testing the hypotheses H0: β2 = 0, H0: β7 = 0, and H0: β8 = 0. What conclusions can you draw about the roles the variables x2, x7, and x8 the model?

Test for coefficient β2,β7,β8

Solution:

a:H0: β2 = 0

T-value for β2=5.18

P-value decision: Since p-value <0.05, we reject the null hypothesis therefore the regression is significant at 5% of the level of significant

b:H0: β7 = 0

T-value for β7=2.20

P-value decision: Since p-value <0.05, we reject the null hypothesis therefore the regression is significant at 5% of the level of significant

c:H0: β8 = 0

T-value for β8= -3.77

P-value decision: Since p-value <0.05, we reject the null hypothesis therefore the regression is significant at 5% of the level of significant

d. Calculate R2 and for this model.

Solution:

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.70624	78.63%	75.96%	73.25%

R-square=78.63%

R-square(adj)=75.96%

e. Using the partial F test, determine the contribution of x7 to the model. How is this Is partial F statistic related to the t-test for β7 calculated in part c above?
Solution:

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	257.09	85.698	29.44	0.000
x2	1	78.03	78.028	26.80	0.000
x7	1	14.07	14.068	4.83	0.038
x8	1	41.40	41.400	14.22	0.001
Error	24	69.87	2.911		
Total	27	326.96			

B7 is significant hence the model is significant of f-test for X7.

H0: $\beta 7 = 0$
H0: $\beta 7 \neq 0$
F-test= 4.83
Since the P-value<0.05 i.e 0.038<0.05, we reject the null hypothesis and conclude that the regression is significant at the level of significance.

Q.3.10 The quality of Pinot Noir wine is thought to be related to the properties of clarity, aroma, body, flavor, and oakiness. Data for 38 wines are given in Table B.11.
a. Fit a multiple linear regression model relating wine quality to these regressors.
b. Test for significance of a regression. What conclusions can you draw?
c. Use t-tests to assess the contribution of each regressor to the model. Discuss your findings.
d. Calculate R2 and R-square(Adj) for this model. Compare these values to the R2 and R-square(Adj) for the linear regression model relating wine quality to aroma and flavor. Discuss your results.
e. Find a 95 % CI for the regression coefficient for flavor for both models in part d. Discuss any differences.

Solution:

Regression Equation

y = 4.00 + 2.34 x1 + 0.483 x2 + 0.273 x3 + 1.168 x4 - 0.684 x5

b. Test for significance of a regression. What conclusions can you draw?

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	5	111.540	72.06%	111.540	22.3081	16.51	0.000
x1	1	0.125	0.08%	2.458	2.4577	1.82	0.187
x2	1	77.353	49.97%	4.240	4.2397	3.14	0.086
x3	1	6.414	4.14%	0.912	0.9118	0.67	0.418
x4	1	19.050	12.31%	19.899	19.8986	14.72	0.001
x5	1	8.598	5.55%	8.598	8.5978	6.36	0.017
Error	32	43.248	27.94%	43.248	1.3515		
Total	37	154.788	100.00%				

Testing the hypothesis for the population using F-test,
$H0:\beta 1=0$
$H1:\beta 1\neq 0$
The level of significance is 0.05
MSR=22.3081
MSE=1.3515

Fstat= MSR/MSE
Fstat=22.3081/1.3515
Fstat=16.5061

p-value=P(F>16.5061)
p-value=1-P(F<16.5061)
p-value=0
p-value<α(0.005),
Since P-value is less than 0.05 we reject the null hypothesis, we can conclude that there is a linear relationship between the dependent variable quality Y and independent variable X1, X2, X3, X4, X5

c. Use t-tests to assess the contribution of each regressor to the model.
Discuss your findings.

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	4.00	2.23	(-0.55, 8.54)	1.79	0.083	
x1	2.34	1.73	(-1.19, 5.87)	1.35	0.187	1.27
x2	0.483	0.272	(-0.072, 1.038)	1.77	0.086	2.38
x3	0.273	0.333	(-0.404, 0.951)	0.82	0.418	2.06
x4	1.168	0.304	(0.548, 1.789)	3.84	0.001	2.68
x5	-0.684	0.271	(-1.236, -0.132)	-2.52	0.017	1.10

The null and alternative hypothesis of each regressor is as follows,
$H0:\beta1=0$
$H1:\beta1 \neq 0$
The level of significance is 0.05
For X1(Clarity):
P-value =0.187
Since p-value >0.05, we fail to reject the null hypothesis
T-test=1.35
We do not reject the H0 hypothesis,
We can conclude that the independent variable (X1)clarity is not statistically linear of the dependent variable.

For X2(Aroma):
P-value =0.086
Since p-value >0.05, we fail to reject the null hypothesis
T-test=1.77
We do not reject the H0 hypothesis,
We can conclude that the independent variable (X2)Aroma is not statistically linear of the dependent variable.

For X3(Body):
P-value =0.418
Since p-value >0.05, we fail to reject the null hypothesis
T-test=0.82
We do not reject the H0 hypothesis,
We can conclude that the independent variable (X3) Body is not statistically linear of the dependent variable.

For X4(Flavor):
P-value =0.001
Since p-value <0.05, we reject the null hypothesis
T-test=3.84

We reject the H0 hypothesis,
We can conclude that the independent variable (X4) Flavor is statistically linear of the dependent variable.

For X5(oakiness):
P-value =0.017
Since p-value <0.05, we reject the null hypothesis
T-test=-2.52
We reject the H0 hypothesis,
We can conclude that the independent variable (X4) Flavor is statistically linear of the dependent variable.

So only Flavor and oakiness are statistically significant variables to the mode

So we can narrow down the equation as follows,

Y= 1.168X4-0.684X5

d. Calculate R2 and R-square(Adj) for this model. Compare these values to the R2 and R-square(Adj) for the linear regression model relating wine quality to aroma and flavor. Discuss your results.

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
1.16254	72.06%	67.69%	63.9266	58.70%	130.49	138.22

For multiple linear regression R-square=72.06% and the value of adjusted R-square is 67.69%
Now assuming another linear regression model related to wine quality to aroma and flavor.
So the equation becomes,

Regression Equation

$$y = 4.35 + 0.518\ x2 + 1.170\ x4$$

Now the values of R-square and R-square Adj are as follows,

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
1.22885	65.86%	63.90%	60.7436	60.76%	129.59	134.93

Thus, for the above-fitted model, the value of R-square is 65.86% and R-square adj is 63.90%. Thus, the R-square of the multiple linear regression model is more than the adjusted r-square of Y=4.35+0.518X2+1.170 X4 model.
Thus, we can conclude that the first multiple linear regression model is better than the Y=4.35+0.518X2+1.170 X4 model.

e. Find a 95 % CI for the regression coefficient for flavor for both models in part d. Discuss any differences.

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	4.35	1.01	(2.30, 6.39)	4.31	0.000	
x2	0.518	0.276	(-0.042, 1.078)	1.88	0.069	2.19
x4	1.170	0.291	(0.580, 1.760)	4.03	0.000	2.19

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	4.00	2.23	(-0.55, 8.54)	1.79	0.083	
x1	2.34	1.73	(-1.19, 5.87)	1.35	0.187	1.27
x2	0.483	0.272	(-0.072, 1.038)	1.77	0.086	2.38
x3	0.273	0.333	(-0.404, 0.951)	0.82	0.418	2.06
x4	1.168	0.304	(0.548, 1.789)	3.84	0.001	2.68
x5	-0.684	0.271	(-1.236, -0.132)	-2.52	0.017	1.10

95% CI for Model fitted in part A is (0.548,1.789) and 95% CI for the model fitted in pard d is (0.580,1.760)
From the above observations, we can state that both values are nearly identical.

Q.3.25. Consider the multiple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

Using the procedure for testing a general linear hypothesis, show how to test

a. $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta$
b. $H_0: \beta_1 = \beta_2, \beta_3 = \beta_4$

6)

3.25) Consider the multiple linear regression model

$$y = \beta_0 + \beta_1 x + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

using the procedure for testing a general linear hypothesis, show how to test

a) $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta$

b) $H_0 : \beta_1 = \beta_2 \,,\ \beta_3 = \beta_4$

Solution:

consider the multiple linear regression model, i.e

$$Y = \beta_0 + \beta_1 x + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon \rightarrow \text{full model}$$

$$y = X\beta + \varepsilon$$

where, $\beta = (\beta_0 \ \beta_1 \ \beta_2 \ \beta_3 \ \beta_4)^T \ \therefore \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_4 \end{bmatrix}$

$\therefore$ producing for testing a general linear hypothesis:

consider $H_0 : L\beta = 0$

$L \rightarrow$ matrix constants

full model is $y = X\beta + \varepsilon$, with

$$\hat{\beta} = (X'X)^{-1}X'y$$

and SS_{Res} (Full model) = sum of sq. of residuals for full model

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$= y'y - \hat{\beta} x'y$ $\therefore$ with $(n-p)$ df

$\therefore$ $n-5$ df

Reduced model $\rightarrow$ $Y = \beta_0 + \beta_1 X_1 + \beta_4 X_4 + \varepsilon$

$SSRes$ (Reduced model) = sum of square of reduced model

$= y'y - \hat{\beta} z'y$, with $(n-p+r)$ df

$\therefore SSRes\ (RM) \geq SSRes\ (FM)$

$\therefore$ the reduced model contains fewer parameters than the full model.

To test H_0 :

$H_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta$ $\therefore L = [0\ 1\ -1\ -1\ -1]$

$SS(H) = SSRes\ (RM) - SSRes\ (FM)$

$SS(H)$ = sum square due to H_0

The test statistic is

$$F = \frac{SS(H)/r}{SSRes\ (FM)/n-p} \sim F_{r, n-p}$$

We reject $H_0 : L\beta = c$ at level α, if observed $F' > F_{\alpha, r, n-p}$

$\therefore c = [0]$

b) $H_0: \beta_1 = \beta_2, \beta_3 = \beta_4$

$H_0: \beta_1 - \beta_2 = 0$

$H_0: \beta_3 - \beta_4 = 0$

To state this equations in general linear hypothesis,

$$L = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

To test $H_0: L\beta = 0$

general linear hypothesis written in the below form,

$$F_0 = \frac{\hat{\beta}' L' (L (X'X)^{-1} L')^{-1} L\hat{\beta} / 2}{SS_{Res}(FM) / (n-5)}$$

Reject $H_0: L\beta = 0$ if $F_0 > F_{\alpha, 2, n-5}$