Topic – Simple Linear Regression

(1)

The moisture of the wet mix of a product is considered to have an effect on the finished product density. The moisture of the mix was controlled and finished product densities were measured as shown in the following data:

Mix moisture (Coded)	Density (Coded)
X	Y
4.7	3
5.0	3
5.2	4
5.2	5
5.9	10
4.7	2
5.9	9
5.2	3
5.3	7
5.9	6
5.6	6
5.0	4

a. Fit the model $Y = \beta_0 + \beta_1, X + \in$ to the data.

b. Place 95% confidence limits on β_1.

c. Is there any evidence in that data that's a more complex model should be tried? (Use $\alpha = 0.05$)

Ans. a. $\hat{Y} = -21.33 + 5X$

b. $2.984 \leq \beta_1 \leq 7.016$

c. **ANOVA**

Source	df	ss	Ms	F
Regression	1	52.50	52.50	30.57
Residual	10	17.17	1.717	
Total	11			

Since $F \succ F_{.05,1,10} = 4.96,$ We conclude that the model is significant at $\alpha = .05$.

Topic: Simple Linear Regression

The effect of the temperature of the deodorizing process on the color of the finished products was determined experimentally. The data collected were as follows:

Temperature	Color
X	Y
460	0.3
450	0.3
440	0.4
430	0.4
420	0.6
410	0.5
450	0.5
440	0.6
430	0.6
420	0.6
410	0.7
400	0.6

420	0.6
410	0.6
400	0.6

a. Fit the model $Y = \beta_0 + \beta_1 X + \in$

b. Is this model sensitive? (Use $\alpha = 0.05$)

c. Obtain a 95% confidence interval for the true mean value of Y at any given value of X_1, say X_0.

Answer: a. $\hat{Y} = 2.5372000 - 0.004718X$

b. ANOVA

Source	df	ss	Ms	F
Regression	1	0.110395	0.110395	14.50
Residual	13	0.98938	0.007611	
Total	14	0.20933		

Since $F = 14.50 > F_{.05,1,13} = 4.67,$ We reject $H_0 : \beta_1 = 0$. The regression model is significant.

c. The 95% confidence interval on the true mean value of Y, calculated of four points: X=0, $X = \overline{X}$, $X = 400,\ X = 460$

At $X = 0,$ $\hat{Y} \pm 1.138,$

At $X = \overline{X},$ $\hat{Y} \pm 0.048,$

At $X = 400,$ $\hat{Y} \pm 0.084$

At $X = 460$ $\hat{Y} \pm 0.104$

Topic – Simple Linear Regression

(3) Show that, for a straight line fit, $r_{xy}^2 = R^2 = r_{Y\hat{Y}}^2$.

$$r_{XY}^2 = \frac{\left\{\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)\right\}^2}{\left\{\sum\left(X_i - \overline{X}\right)^2\right\}\left\{\sum\left(Y_i - \overline{Y}\right)^2\right\}}$$

$$= \frac{SS_{Reg}}{\sum\left(X_i - \overline{X}\right)^2} = \frac{SS_{Reg}}{SS_T} = R^2$$

$$r_{Y\hat{Y}} = \frac{\sum\left(Y_i - \overline{Y}\right)\left(\hat{Y}_i - \bar{\hat{Y}}_i\right)}{\left\{\sum\left(\hat{Y}_i - \bar{\hat{Y}}_i\right)^2\right\}^{\frac{1}{2}}\left\{\sum\left(Y_i - \overline{Y}\right)^2\right\}^{\frac{1}{2}}}$$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_i X_i, \qquad \bar{\hat{Y}}_i = \hat{\beta}_0 + \hat{\beta}_i \overline{X} = \overline{Y}$$

Thus, $\left(\hat{Y}_i - \bar{\hat{Y}}_i\right) = \hat{\beta}_1\left(X_i - \overline{X}\right)$

$$r_{Y\hat{Y}} = \frac{\hat{\beta}_1 \sum\left(Y_i - \overline{Y}\right)\left(X_i - \overline{X}\right)}{\hat{\beta}_1\left\{\sum\left(X_i - \overline{X}\right)^2\right\}^{\frac{1}{2}}\left\{\sum\left(Y_i - \overline{Y}\right)^2\right\}^{\frac{1}{2}}}$$

$$= r_{XY}$$

(4)Topic: Multiple Linear Regressions.

Eight runs were made at various conditions of saturation $\left(X_1\right)$ and transisomers $\left(X_2\right)$. The response, SCI, is listed below as Y for the corresponding levels of X_1 and X_2.

Y	X_1	X_2
66.0	38	47.5
43.0	41	21.3
36.0	34	36.5
23.0	35	18.0
22.0	31	29.5

14.0	34	14.2
12.0	29	21.0
7.6	32	10.0

a. Fit the model $Y = \beta_0 + \beta_1 X_1 + \beta_1 X_2 + \in$

b. Is the overall regression significant? (Use $\alpha = 0.05$)

c. How much of the variation in Y about $\overline{Y}$ is explained by X_1 and X_2 ?

Ans. a. $\hat{Y} = -94.552026 + 2.801551 X_1 + 1.072683 X_2$

b. ANOVA

Source	df	ss	Ms	F
Regression	2	2618.98	1309.49	151.74
Residual	5	43.16	8.63	
Total	7	2662.16		

Since $F = 151.74 > F_{.05,2,5} = 5.79$, the overall regression is statistically significant.

c. $R^2 = \frac{ss_{\mathrm{Re}g}}{ss_T} = \frac{2618.98}{2662.14} = 98.38\%$

Topic: Multiple linear regressions

(5) Fit the model $Y = B_0 + B_1 X_1 + B_2 X_2 + \in$ to the data below. Is the overall regression significant? (Use $\alpha = 0.05$). Find the appropriate extra ss F- statistic for testing $H_0 : B_2 = 0$ against $H_1 : B_2 \neq 0$, and find its degree of freedom. Relate this F-statistic numerically to the t-statistic typically used to test the same hypothesis.

X_1	X_2	Y
-1	-1	8
-1	1	13

-1	1	12
-1	1	11
1	-1	9
1	-1	8
1	-1	7
1	1	13
0	0	11
0	0	13

Ans. $\hat{B} = (X'X)^{-1} X'Y = \begin{pmatrix} 10 & 0 & 0 \\ 0 & 8 & -4 \\ 0 & -4 & 8 \end{pmatrix}^{-1} \begin{pmatrix} 105 \\ -7 \\ 17 \end{pmatrix} =$ .....

Fitted model $\hat{Y} = 10.50 + 0.25X_1 + 2.25X_2$

ANOVA

Source	df	ss	Ms	F
Regression	2	36.50	18.25	10.67
Residual	7	12.00	1.714	
Total	9	48.50		

Since $F = 10.67 > F_{.05,2,7} = 4.74$, the overall regression is statistically significant.

Test $H_0 : B_2 = 0$ of $H_1 : B_2 \neq 0$

Test statistic using extra sum of square technique.

$$F = \frac{\left\{ ss_{\mathrm{Reg}}\left(Full \bmod el\right) - ss_{REG}\begin{pmatrix} \mathrm{Re}\, stricted \bmod el \\ Y = B_0 + B_1 X + \in \end{pmatrix} \right\} / 1}{Ms_{\mathrm{Res}}}$$

$$\frac{\{36.50-6.125\}/1}{1.714}$$

$=17.72 > F_{.05,1,7} = 5.59,$ reject H_0

This F is the square of the corresponding t_7 statistic since the df are 1, 7.

To test $H_0 : B_1 = 0$ of $H : B_1 \neq 0$

$$F = \frac{\left\{SS_{\text{Re}g}\left(Full \ \text{mod}el\right) - SS_{\text{Re}g}\left(\text{Re}stricted \ \text{mod}el : Y = B_0 + B_2 X_2 + \in\right)\right\}/1}{MS_{\text{Re}s}}$$

$$= \frac{36.50 - 36.125}{1.714} = 0.219$$

Since $F < F_{.05,1,7} = 5.59,$ H_0 is accepted.

So X_1 could be dropped to give the equation.

$\hat{Y} = 10.5 + 2.125X_2$

Topic: Multiple Linear Regressions

(6) Show that's $X'e = 0$

Ans.

$$\begin{aligned} X'e &= X'\left(I - H\right)Y \\ &= X'\left(I - X\left(X'X\right)^{-1}X'\right)Y \\ &= \left(X' - X'\right)Y = 0 \end{aligned}$$

Topic: Multiple Linear Regressions

(7) Consider the formal regression of the residuals e_i onto a quadratic function $\alpha_0 + \alpha_1 \hat{Y}_i + \alpha_2 \hat{Y}_i^2$ on the fitted values $\hat{Y}_i$, by least squares show that all the three estimated coefficients depend on $T_{12} = \sum e_i \hat{Y}_i^2$. What does this imply?

Ans. The vector on the right hand side of the appropriate normal equations consists of the three elements $\sum e_i = 0,\ \sum e_i \hat{Y} = 0,\ and\ T_{12} = \sum e_i \hat{Y}_i^2$. Thus all three estimated coefficients depend on T_{12} , which is thus a measure of the amount of quadratic trend in the e_i versus $\hat{Y}_i$ plot.

Topic: Regression Model with auto correlated Errors

(8) A regression fit $\hat{Y} = \hat{B_0} + \hat{B_1}\hat{X}_1 + \hat{B_2}\hat{X}_2 + \hat{B_3}\hat{X}_3$

On 85 observations equally spaced in time produces a Durbin- Watson statistic of $d = 2.33$ .Might this indicates serial correlation? Test at a two-tailed $\alpha = 0.05$ level.

Ans. $d = 2.33$ is in upper tail, So use $4 - d = 1.67$. The 2.5% lower tail bounds are $(d_2, du) = (1.51, 1.65)$. So $4 - d$ is not significant.

Topic: Multiple Linear Regressions

(9) We fit a straight line model to a set of data using the formula $\hat{B} = (X'X)^{-1} X'Y$, $\hat{Y} = X\hat{B}$ with the usual definitions. We define $H = X(X'X)^{-1} X'$ show that $SS_{Reg} = Y'HY$.

Ans.

$$\begin{aligned} SS_{Reg} &= \hat{\beta}' X'Y \\ &= \left((X'X)^{-1} X'Y\right)' X'Y \\ &= Y'X(X'X)^{-1} X'Y \\ &= Y'HY \end{aligned}$$

Topic: Multiple Linear Regressions

(10) Prove that

$$R^2 = \frac{v_1 F}{v_1 F + v_2}$$

$v_1 = \text{Re}\, gression\ df.\ and\ v_2 = \text{Re}\, sidualdf.$

Ans. $R^2 = \dfrac{SS_{Reg}}{SS_T}$

$$= \frac{SS_{Reg}}{SS_{Reg} + SS_{Res}}$$

$$= \frac{SS_{Reg} / MS_{Res}}{\frac{SS_{Reg}}{MS_{Res}} + \frac{SS_{Res}}{MS_{Res}}}$$

$$= \frac{v_1 \frac{MS_{Reg}}{MS_{Res}}}{v_2 \frac{MS_{Reg}}{MS_{Res}} + v_2}$$

$$= \frac{v_1 F}{v_1 F + v_2}$$

Topic: Multi-collinearity

(11) Can we use the data below to get a unique fit to the model $Y = B_0 + B_1 X_1 + B_2 X_2 + B_3 X_3 + 6$?

If not, what model can be fitted?

X_1	X_2	X_3	Y
1	-2	4	81
2	-7	11	88
4	3	5	94
7	1	13	95
8	-1	17	123

Answer: No, because the columns are related by $-2X_1 + X_2 + X_3 = 0$. Dropping any one of the three columns allows a unique fit.

Topic : Multicolinearity

(12) Can we use the data below to fit, uniquely, the model
$Y = B_0 + B_1 X_1 + B_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + B_{12} X_1 X_2 + \in$ If not, what model can be fitted?

X_1	X_2	Y
-1	-1	38

1	-1	45
-1	1	41
1	1	40
0	0	47
0	0	42
0	0	48

Answer:

No, we add an $X_0 = 1$ column & columns generated by X_1^2, X_2^2 *and* X_1X_2 to form X . Note that $X_1^2 = X_2^2$ always, so the X matrix is singular. We can fit the model $Y = B_0 + B_1X_1 + B_2X_2 + B\left(X_1^2 + X_2^2\right) + B_{12}X_1X_2 + \in$ but cannot estimate β_{11} & β_{22} individually.

Topic: Model Adequacy Checking

(13) Prove that the plot of e_i versus Y_i always has a slope on size $1 - R^2$ in it.

Ans. Slope = $\dfrac{e'e}{S_{YY}}$

$$e'e = \sum e_i^2 = \text{Re}\,sidual\ \ sum\ of\ \ square$$
$$= SS_{\text{Res}}$$
$$S_{YY} = \sum\left(Y_i - \overline{Y}\right) = SS_T$$

Slope = $\dfrac{e'e}{S_{YY}} = \dfrac{SS_{\text{Res}}}{SS_T}$

$$= \frac{SS_T - SS_{\text{Re}g}}{SS_T}$$
$$= 1 - \frac{SS_{\text{Re}g}}{SS_T} = 1 - R^2$$

The implication of this result is that it is a mistake to attempt to find detective regression by a plot of the residuals e_i versus observations Y_i as this always show a slope.

Topic: Model Adequacy Checking

TRUE / FALSE Question

(a) $\left(Y-\hat{Y}\right)' 1=0$ is true when the model does not contain a B_0 term.

(b) On a normal probability plot, a line should be drawn "middle bulk" of the residuals as a check on normality.

(c) When $R^2=1$, all the residuals must be zero.

(d) When we fit the model $Y=B_0+\in\left(i.e.\ no\ X'S\right)$ to a set of data, $R^2=0$ always.

Answer- a. FALSE b. TRUE c. TRUE d. TRUE

Topic : Model Adequacy Checking

(15) Prove that the vector of fitted values $\hat{Y}$ is always orthogonal to the vector of residuals e.

Ans. We know

$$\hat{Y}=X\hat{B}$$
$$=X\left(X'X\right)^{-1}X'Y$$
$$=HY$$

Where $H=X\left(X'X\right)^{-1}X'$

$$e=Y-\hat{Y}=Y-HY=\left(I-H\right)Y$$
$$\therefore e'\hat{Y}=Y'\left(I-H\right)HY=Y'\left(H-H^2\right)Y=0$$

Test for influential observations

(16) TRUE / FALSE Question

(a) An observation cannot be both influential & leverage.

(b)An observation can be influential but not leverage.

(c)An observation can be a leverage but not influential.

(d) If we fit a straight line $Y=B_0+B_1X+\in$, and we find that $\hat{B}_0=0$ exactly, then the residual may not add to zero.

Answer- a. FALSE b. TRUE c. TRUE d. FALSE

If you are asked to fit a straight line to the data

$\left(X,Y\right)=\left(1,3\right),\left(2,2.5\right),\left(2,1.2\right),\left(3,1\right)\ and\ \left(6,4.5\right)$ What would you say about it?

Answer- The last point $\left(6,4.5\right)$ is very influential. With it, the line has positive slope; without it negative. Some observations between $X=3\ \&\ X=6$ would be useful here.

Topic : Transformation & weighting to correct model in adequacies

(18) Consider the simple linear regression model $y_i = B_0 + B_1 x_i + E_i$, where the variance of ε_i is proportional to x_i^2, i.e. $v(\varepsilon_i) = 6^2 x_i^2$

(a) Suppose that we use the transformation $y' = \frac{y}{x}$ & $x' = \frac{1}{x}$. Is it a variation stabilizing transformation?

(b)What are the relationships between the parameters in the original & the transformed model?

(c)Suppose we use the method of weighted least squares with $\omega_i = \frac{1}{x_i^2}$. Is this equivalent to the transformation introduced in past (a).

Answer:

(a)

$$y_i' = \frac{y_i}{x_i} = \frac{B_0 + B_1 x_i + \in_i}{x_i}$$

$$= \frac{B_0}{x_i} + B_1 + \frac{\in_i}{x_i}$$

$$= B_o x_i + B_1 + \in_i'$$

$$v\left(\in_i'\right) = v\left(\frac{\in_i}{x_i}\right) = \frac{6^2 x_i^2}{2} = 6^2$$

It is a variance stabilizing transformation.

(b) The intercept of original model is the slope of transformed model & the slope of original model is intercept of transformed model.

(c)YES

Topic : transformation & weighting to correct model inadequacies

(19) Suppose we have n observations of variables $X_1, X_2 \cdots, X_K, Y,$ where X 'S are predictors & Y is variable. Suppose we are told that observation Y_i are uncorrected but the last observation has variance 166^2 rather than 6^2. Find the best linear unbiased estimator (BLUE) of B using weighted least square.

Ans.

$$v(\varepsilon) = 6^2 \; Diag(1,1,\cdots 1,16)$$

$$= 6^2 v$$

$$\hat{B} = \left(X' v^{-1} X\right)^{-1} X' v^{-1} Y$$

$$= \left(X' \; Diag\left(1,1,\cdots,\frac{1}{16}\right) X\right)^{-1} \left(X' \; Diag\left(1,1,\cdots,\frac{1}{16}\right) Y\right)$$

Topic: Transformation & weighting to correct model inadequacies.

(20) prove that $\hat{B} = \left(X' v^{-1} X \right)^{-1} X' v^{-1} Y$ is the generalized least squares solution where we assume $\in \sim N\left(0, v6^2\right)$