Tutorial Worksheet 5 - Advanced Regression Analysis (with Solutions)

Problem 1.

A study was made on the effect of temperature on the yield of a chemical process, the following data were collected:

X	-5	-4	-3	-2	-1	0	1	2	3	4	5
Y	1	5	4	7	10	8	9	13	14	13	18

(a) Assuming a model, $Y = \beta_0 + \beta_1 X + \epsilon$, what are the least square estimates of β_0 and β_1? What is the fitted equation?

(b) Construct the ANOVA table and test the hypothesis $H_0 : \beta_1 = 0$ with $\alpha = 0.05$.

(c) What are the confidence limits for β_1 at $\alpha = 0.05$?

(d) What are the confidence limits for the true mean value of Y when $X = 3$ at $\alpha = 0.05$?

(e) What are the confidence limits at $\alpha = 0.05$ level of significance for the difference between the true mean value of Y when $X_1 = 3$ and the true mean value of Y when $X_2 = -2$?

[Given, $F_{0.05,1,9} = 5.12,\ t_{0.05,9} = 1.833,\ t_{0.025,9} = 2.263,$]

Solution:

(a) The given data is $(x_i, y_i);\ i = 1, 2, \ldots, 11$. The linear regression model is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \text{ thus, } S = \sum_{i=1}^{11} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{11} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{11} x_i^2 - n\bar{x}^2} = \frac{158}{110} = 1.44; \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 9.27$$

Thus, $\hat{y_i} = 9.27 + 1.44 x_i$ is the fitted model.

(b) The sum of squares can be calculated as follows:

$$SS_T = \sum_{i=1}^{11} (y_i - \bar{y})^2 = 248.18,\ SS_{Reg} = \hat{\beta}_1^2 S_{xx} = 226.94,\ SS_{Res} = \sum_{i=1}^{11} (y_i - \hat{y_i})^2 = SS_T - SS_{Reg} = 22.23$$

Thus, the ANOVA table is

Source	DF	SS	MS = SS/DF	F-obs=MS$_{Reg}$/MS$_{Res}$	F-tab
Regression	1	226.94	226.94	96.17	$F_{0.05,1,9} = 5.12$
Residual	9	22.23	2.36		
Total	10	248.18			

To test the hypothesis $H_0 : \beta_1 = 0$ **vs** $H_1 : \beta_1 \neq 0$.

Here, we reject H_0, since $F_{obs} > F_{tab}$. Y and X are linearly dependent.

Alternative method for testing:

To test the hypothesis $H_0 : \beta_1 = 0$ **vs** $H_1 : \beta_1 \neq 0$.

Under $H_0 : \beta_1 = 0, \quad t = \frac{\hat{\beta}_1}{\sqrt{\frac{MS_{Res}}{S_{xx}}}} \sim t_{n-2}$. Hence, $t = \frac{1.44}{\sqrt{\frac{2.36}{110}}} = 9.83 > t_{0.05,9} = 1.833$. Hence, we reject $H_0 : \beta_1 = 0$.

Hence, $F \equiv t^2$, under H_0.

(c) The confidence interval of β_1 can be computed as follows:

$$\hat{\beta}_1 \sim \mathrm{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right) \Rightarrow \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim \mathrm{N}(0,1) \quad [\sigma^2 \text{ is unknown and estimated by } MS_{Res}]$$

$$\Rightarrow \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_{Res}}{S_{xx}}}} \sim \mathrm{t}_{n-2}$$

$$\Rightarrow P\left[-\mathrm{t}_{\alpha/2,n-2} \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_{Res}}{S_{xx}}}} \leq \mathrm{t}_{\alpha/2,n-2}\right] = 1-\alpha$$

$$\Rightarrow \hat{\beta}_1 - \mathrm{t}_{\alpha/2,n-2}\sqrt{\frac{MS_{Res}}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + \mathrm{t}_{\alpha/2,n-2}\sqrt{\frac{MS_{Res}}{S_{xx}}}$$

$$\Rightarrow 1.44 - 2.263 * 0.146 \leq \beta_1 \leq 1.44 + 2.263 * 0.146 \;\Rightarrow 1.11 \leq \beta_1 \leq 1.77.$$

(d) The confidence limits for $E(Y|X=3)$ can be computed as follows:

95% CI for $E(Y \text{ at } X = x_0)$ is $\beta_0 + \beta_1 x_0$. An unbiased estimator $E(Y \text{ at } X = x_0)$ is $\hat{\beta}_0 + \hat{\beta}_1 x_0$.

Therefore, $\hat{\beta}_0 + \hat{\beta}_1 x_0 \sim \mathrm{N}\left(\beta_0 + \beta_1 x_0, \sigma^2\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]\right)$.

$$A = \frac{\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right) - (\beta_0 + \beta_1 x_0)}{\sqrt{MS_{Res}\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]}} \sim t_{n-2},\ \therefore P\left\{-t_{\alpha/2,n-2} \leq A \leq t_{\alpha/2,n-2}\right\} = 1-\alpha$$

$$\therefore \left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right) \pm t_{\alpha/2,n-2} \times \sqrt{MS_{Res}\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]} \text{ is the confidence interval for } \beta_0 + \beta_1 x_0.$$

Thus, replacing the values we get the required confidence interval is $12.15 \leq \beta_0 + \beta_1 x_0 \leq 15.03$.

(e) The confidence limits for the difference between the true mean value of Y when $X_1 = 3$ and the true mean value of Y when $X_2 = -2$ can be computed as follows:

$$E(Y \text{ at } X_1 = 3) - E(Y \text{ at } X_2 = -2) \equiv Z_1 - Z_2.$$

Unbiased estimators of Z_1 and Z_2 are: $\hat{Z}_1 = \hat{\beta}_0 + \hat{\beta}_1 3$ and $\hat{Z}_2 = \hat{\beta}_0 + \hat{\beta}_1(-2)$.
Thus, $\hat{Z}_1 - \hat{Z}_2 = (\hat{\beta}_0 + \hat{\beta}_1 3) - (\hat{\beta}_0 + \hat{\beta}_1(-2)) = 5\hat{\beta}_1 = 7.20$ (point estimation).
Confidence interval for $Z_1 - Z_2$:
Compute $Var(\hat{Z}_1 - \hat{Z}_2) = Var(5\hat{\beta}_1) = \frac{25\sigma^2}{S_{xx}} = \frac{25\sigma^2}{110}$; thus $\hat{Z}_1 - \hat{Z}_2 \sim \mathrm{N}(Z_1 - Z_2, \frac{25\sigma^2}{110}) \Rightarrow \frac{(\hat{Z}_1-\hat{Z}_2)-(Z_1-Z_2)}{\sqrt{\frac{25 MS_{Res}}{110}}} \sim \mathrm{t}_{n-2}$.
So, C.I. of $Z_1 - Z_2$ is:

$$(\hat{Z}_1 - \hat{Z}_2) - \mathrm{t}_{\alpha/2,9}\sqrt{\frac{25 \times 2.36}{110}} \leq Z_1 - Z_2 \leq (\hat{Z}_1 - \hat{Z}_2) + \mathrm{t}_{\alpha/2,9}\sqrt{\frac{25 \times 2.36}{110}} \;\Rightarrow 5.54 \leq Z_1 - Z_2 \leq 8.86.$$

Problem 2.

There are a few occasions where it makes sense to fit a model without an intercept β_0. If there were an occasion to fit the model $y = \beta x + \epsilon$ to a set of data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the least square estimate of β would be

$$\hat{\beta} = b = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$

Suppose you have a programmed calculator that will fit only the intercept model $y = \beta_0 + \beta_1 x + \epsilon$, but you want to fit a non-intercept model. By adding one more fake data point $(m\bar{x}, m\bar{y})$ to the data above, where $m = \frac{n}{(n+1)^{1/2}-1} = \frac{n}{a}$, say, and letting the calculation fit $y = \beta_0 + \beta_1 x + \epsilon$, can you estimate β by using b?

Solution:

Given the data set $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, we fit: $y = \beta x + \epsilon$, where the least square estimate of β is calculated as follows:

$$\text{SSE} = \text{S} = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{\beta} x_i)^2; \quad \frac{\partial S}{\partial \beta} = 0 \Rightarrow \sum (y_i - \hat{\beta} x_i) x_i = 0 \Rightarrow \hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}.$$

For the computer-fitted model with intercept, we have:

$$y = \beta_0 + \beta_1 x + \epsilon \text{ where, the LSEs are } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \text{ and } \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

Adding one more data point: $(m\bar{x}, m\bar{y})$; $m = \frac{n}{(n+1)^{1/2} - 1} = \frac{n}{a} \Rightarrow (a+1)^2 = (n+1)$.
Thus the new dataset is given by $(u_1, v_1), (u_2, v_2), \ldots, (u_n, v_n), (u_{n+1}, v_{n+1}) = (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n), (m\bar{x}, m\bar{y})$. Hence we have

$$\bar{u} = \frac{n\bar{x} + (n/a)\bar{x}}{n+1} = \frac{n\bar{x} + m\bar{x}}{n+1} = \frac{n\bar{x}(\frac{a+1}{a})}{n+1} = \frac{n\bar{x}}{(a+1)^2} \times \frac{a+1}{a} = \frac{n\bar{x}}{a(a+1)} \text{ similarly } \bar{v} = \frac{n\bar{y}}{a(a+1)}$$

$$S_{uu} = \sum_{i=1}^{n} x_i^2 + (m\bar{x})^2 - (n+1)\bar{u}^2 = \sum_{i=1}^{n+1} u_i^2 - (n+1)\bar{u}^2 = \sum_{i=1}^{n} x_i^2 + \frac{n^2}{a^2}\bar{x}^2 - (n+1)\frac{n^2\bar{x}^2}{a^2(a+1)^2} = \sum_{i=1}^{n} x_i^2$$

$$S_{uv} = \left(\sum_{i=1}^{n} x_i y_i (m\bar{x} m\bar{y})\right) - (n+1)\left[\frac{n\bar{x}}{a(a+1)}\right]\left[\frac{n\bar{y}}{a(a+1)}\right] = \sum_{i=1}^{n} x_i y_i + m^2\bar{x}\bar{y} - \frac{(n+1)m^2\bar{x}\bar{y}}{(a+1)^2}$$

$$= \sum_{i=1}^{n} x_i y_i + m^2\bar{x}\bar{y} - m^2\bar{x}\bar{y} = \sum_{i=1}^{n} x_i y_i$$

For the programmed calculation: $\hat{\beta}_1 = \frac{S_{uv}}{S_{uu}} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$

Problem 3.

Fit the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ for the data given below. Provide an ANOVA table and perform the partial F-tests to test $H_0 : \beta_i = 0$ vs $H_1 : \beta_i \neq 0$ for $i = 1, 2$; given the other variable is already in the model. Comment on the relative contributions of the variables X_1 and X_2, depending on whether they enter the model first or second. Find the regression equation.

X_1	-5	-4	-1	2	2	3	3
X_2	5	4	1	-3	-2	-2	-3
Y	11	11	8	2	5	5	4

Solution:

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

Fit the model with both X_1, X_2. OLS estimates can be obtained using $\hat{\beta} = (X^T X)^{-1} X^T Y$

$$X = \begin{pmatrix} 1 & -5 & 5 \\ 1 & -4 & 4 \\ 1 & -1 & 1 \\ 1 & 2 & -3 \\ 1 & 2 & -2 \\ 1 & 3 & -2 \\ 1 & 3 & -3 \end{pmatrix}; \quad \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 7 & 0 & 0 \\ 0 & 68 & -67 \\ 0 & -67 & 63 \end{pmatrix}^{-1} \begin{pmatrix} 46 \\ -66 \\ 69 \end{pmatrix} = \begin{pmatrix} \frac{46}{7} \\ 1 \\ 2 \end{pmatrix} \therefore \hat{Y} = \frac{46}{7} + X_1 + 2X_2$$

$SS_T = \sum \left(Y_i - \bar{Y}\right)^2 = 73.71$, $\quad e_i = y_i - \hat{y}_i$, $\quad SS_{Res} = \sum_{i=1}^{n} e_i^2$, $\quad H_0 : \beta_1 = \beta_2 = 0$ vs. $H_1 : H_0$ is not true. To test this hypothesis, we have $F = 83.72 > F_{0.05,2,4} = 6.94 \Rightarrow H_0$ is rejected.

Partial F-tast:- Fit the model with X_1 as $Y = \beta_0 + \beta_1 X_1 + \epsilon$, $\hat{Y} = \frac{46}{7} - \frac{66}{68} X_1$ and test the hypothesis $H_0 : \beta_2 = 0$ vs $H_1 : \beta_2 \neq 0$

Source	DF	SS	MS	F-obs	F-tab
Regression	2	72.00	36.00	83.72	$F_{0.05,2,4} = 6.94$
Residual	4	1.71	0.43		
Total	6	73.71			

$$F = \frac{\{SS_{\text{Reg}}\,(\text{Full}) - SS_{\text{Reg}}\,(\text{Restricted Model})\}\,/1}{MS_{\text{Res}}} = \frac{72 - 64.06}{0.43} = 18.53 > F_{0.05,1,4} = 7.71$$

$\therefore H_0$ is rejected at 5% level of significance, i.e., X_2 is significant in the presence of X_1.

Partial F-tast:- Fit the model with X_2 as $Y = \beta_0 + \beta_2 X_2 + \epsilon$, $\hat{Y} = \frac{46}{7} - \frac{69}{68}X_2$ and test the hypothesis $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$

$$F = \frac{\{SS_{\text{Reg}}\,(\text{Full}) - SS_{\text{Reg}}\,(\text{Restricted Model})\}\,/1}{MS_{\text{Res}}} = \frac{72 - 70.01}{0.43} = 4.64 < F_{0.05,1,4} = 7.71$$

$\therefore H_0$ is not rejected at 5% level of significance, i.e., X_1 is not significant in the presence of X_2.

Implication:- If X_2 is in the model, we donot need X_1. If X_1 is in the model, X_2 helps out significantly. Then X_2 is clearly more useful variable and it explains $R^2 = \frac{70.01}{73.71} = 95\%$ of the total variability in Y about mean, where as X_1 alone explains $R^2 = \frac{64.06}{79.71} = 86\%$ of total variability in Y about mean. And X_1 and X_2 together explain $\frac{72.00}{73.71} = 97\%$ of total variability

NOTE: In this problem, $X_1 + X_2 \approx 0$ and the presence of multicollinearity (X_1 and X_2 are not independent) and that is why Partial F-test suggests $\beta_0 = 0$ is accepted.

Problem 4.

Given a 2-variables linear regression problem $Y = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + \epsilon$, yield the following

$$X^T X = \begin{bmatrix} 33 & 0 & 0 \\ 0 & 40 & 20 \\ 0 & 20 & 60 \end{bmatrix}, \quad X^T Y = \begin{bmatrix} 132 \\ 24 \\ 92 \end{bmatrix} \quad \text{and} \quad \sum (Y - \bar{Y})^2 = 150.$$

(a) What is the sample size?

(b) Write the normal equations and solve for the regression coefficients.

(c) Estimate the standard error of β_2 and test the hypothesis that $\beta_2 = 0$

(d) Compute R^2 and interpret it. Also, interpret the value of regression coefficients.

(e) Predict the value of y given $x_1 = -4$ and $x_2 = 2$

(f) Comment on the possibilities of any regressors being a dummy variable.

Solution:

(a) The variance-covariance matrix of the design matrix $X = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix}$ is given by

$$X^T X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix} = \begin{bmatrix} n & \sum x_{i1} & \sum x_{i2} \\ \sum x_{i1} & \sum x_{i1}^2 & \sum x_{i1}x_{i2} \\ \sum x_{i2} & \sum x_{i1}x_{i2} & \sum x_{i2}^2 \end{bmatrix} = \begin{bmatrix} 33 & 0 & 0 \\ 0 & 90 & 20 \\ 0 & 20 & 60 \end{bmatrix}$$

$\therefore$ The sample size, $n = 33$.

$$X^T Y = \begin{bmatrix} 1 & \ldots & 1 \\ x_{21} & \ldots & x_{n1} \\ x_{12} & \ldots & x_{n2} \end{bmatrix}_{3\times n} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} \sum y_i \\ \sum x_{i1}y_i \\ \sum x_{i2}y_i \end{bmatrix} = \begin{bmatrix} 132 \\ 24 \\ 92 \end{bmatrix}$$

(b) We know the normal equation is given by $\hat{\beta} = (X^TX)^{-1}X^TY \Rightarrow (X^TX)\hat{\beta} = X^TY$. Thus we have

$$(X^TX)\hat{\beta} = X^TY \Rightarrow \begin{bmatrix} 33 & 0 & 0 \\ 0 & 40 & 20 \\ 0 & 20 & 60 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 132 \\ 24 \\ 92 \end{bmatrix} \Rightarrow \begin{bmatrix} 33 & 0 & 0 \\ 0 & 40 & 20 \\ 0 & 0 & 50 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 832 \\ 24 \\ 80 \end{bmatrix}$$

$$\left[\text{using Gaussian elimination:} R_3' = R_3 - \frac{1}{2}R_2\right]$$

Thus we obtain $b_3 = 1.60, 40b_2 + 20b_3 = 24 \Rightarrow 40b_2 = -8 \Rightarrow b_2 = -0.20, 33b_1 = 132 \Rightarrow b_1 = 4$. Hence, the regression equation is $\hat{y} = 4 - 0.20x_1 + 1.60x_2$

(c)

$$\left|X^TX\right| = 33(2400 - 400) = 66000,$$

$$\text{S.E. of } b_2 = S\sqrt{c_{22}} \text{ where } S = \sqrt{\frac{RSS}{n-k}} = \sqrt{\frac{7.6}{2}} = \sqrt{3.8} \text{ and } C_{22} = \frac{\det(\text{co-factor})}{|X^TX|} = \frac{1980}{66000} = 0.003$$

$$\text{SE}(b_2) = \sqrt{0.003 \times 3.8} = 0.106$$

To test the hypothesis $H_0 : b_2 = 0$ vs. $H_1 : b_2 \neq 0$, we use the following test statistic $t = \frac{b_2}{\text{SE}(b_2)} = \frac{-0.20}{0.106} = 1.88 < 4.303$. We reject H_0.

(d) $\text{TSS} = \text{RSS} + \text{SS}_{\text{Reg}}$

Given that $TSS = 150$, we can calculate $\text{SS}_{\text{Reg}} = \beta^{*T}(X^TY)^* = \begin{bmatrix} -0.20 & 1.60 \end{bmatrix} \begin{bmatrix} 24 \\ 92 \end{bmatrix} = 142.4$ and thus we have $RSS = 150 - 142.4 = 7.6$

Thus the coefficient of determination R^2 is given by $R^2 = 142.4/150 = 0.9493$ explains 94.93% of the total variability in the response variable.

(e) The predicted value of y for the given values of $x_1 = -4$ and $x_2 = 2$ is given by $\hat{y} = 4 - 0.20(-4) + 1.60(2) = 8$.

Problem 5.

Can we use the data given below to get a unique fit to the model $Y = \beta_0X_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \epsilon$

X_1	-4	3	1	4	-3	-1
X_2	1	2	3	4	5	6
X_3	3	-5	-4	-8	-2	-5
Y	7.4	14.7	13.9	18.2	12.1	14.8

Solution:

The LSE of the regression model $Y = X\beta + \epsilon$ is given by $\hat{\beta} = \left(X^TX\right)^{-1}X^TY$

$$X = \begin{bmatrix} x_0 & x_1 & x_2 & x_3 \\ 1 & -4 & 1 & 3 \\ 1 & 3 & 2 & -5 \\ 1 & 1 & 3 & -4 \\ 1 & 4 & 4 & -8 \\ 1 & -3 & 5 & -2 \\ 1 & -1 & 6 & -5 \end{bmatrix}$$

Note that, $X_1 + X_2 + X_3 = 0$ which implies that $\left(X^TX\right)$ is a singular matrix, i.e., $|\left(X^TX\right)| = 0$. 'NO', we can't compute $\hat{\beta}$ uniquely here.

Problem 6.

Show that in linear regression with a β_0 term in the model:

(a) The correlation between the vector e and Y is $(1 - R^2)^{1/2}$. This result implies that it is a mistake to find defective regressions by a plot of residuals e_i versus observations Y_i as this always shows a slope.

(b) Show further that the correlation between e and $\hat{Y}$ is zero.

Solution:

(a)

$$\text{Cor}(e, Y) = \frac{\sum (e_i - \bar{e}) \left(Y_i - \bar{Y}\right)}{\sqrt{\sum (e_i - \bar{e})^2 \sum \left(Y_i - \bar{Y}\right)^2}}$$

$$\sum (e_i - \bar{e}) \left(Y_i - \bar{Y}\right) = \sum e_i \left(Y_i - \bar{Y}\right) = \sum e_i Y_i = Y^T e \text{ [since, } \bar{e} = 0 \text{ if } \beta_0 \text{ is in the model]}$$

Note that,

$$Y = X\beta + \epsilon,\ \hat{\beta} = \left(X^T X\right)^{-1} X^T Y \ \Rightarrow \hat{Y} = X \left(X^T X\right)^{-1} X^T Y = X\hat{\beta} = HY, \text{ where } H = X\left(X^T X\right)^{-1} X^T$$
$$e = Y - \hat{Y} = (I - H)Y \Rightarrow e^T e = Y^T (I - H)^T (I - H) Y = Y^T (I - H) Y = Y^T e \text{ [since, } H^2 = H]$$

Thus we have

$$\text{Cor}(e, Y) = \frac{Y^T e}{\sqrt{(e^T e) SS_T}} = \sqrt{\frac{e^T e}{SS_T}} = \sqrt{\frac{SS_{Res}}{SS_T}} = \sqrt{1 - \frac{SS_{Reg}}{SS_T}} = \sqrt{1 - R^2}.$$

Implication: This is why we plot $\hat{Y}_i$ and e_i but not Y_i and e_i since they are correlated.

(b) The correlation between e and $\hat{Y}$ can be computed as follows:

$$\text{Cov}(e, \hat{Y}) = \sum (e_i - \bar{e}) \left(\hat{Y}_i - \bar{\hat{Y}}\right) = e^T \hat{Y} = Y^T (I - H) H Y = Y^2 (H - H^2) Y = 0 \text{ [since, } \hat{Y} = HY, e = (I - H)Y]$$
$$\text{Cor}(e, \hat{Y}) = 0.$$

Problem 7.

Prove that the multiple coefficients R^2 is equal to the square of the correlation between Y and $\hat{Y}$.

Solution:

We have to prove that $[\text{Cor}(Y, \hat{Y})]^2 = R^2 = \frac{SS_{Reg}}{SS_T}$

Note that, $\sum e_i = 0 \Rightarrow \sum \left(Y_i - \hat{Y}_i\right) = 0 \Rightarrow \sum Y_i = \sum \hat{Y}_i \Rightarrow \bar{Y} = \bar{\hat{Y}}$ [since, $Y_i = \hat{Y}_i + e_i$]

$$\text{Cor}\left(e, \hat{Y}\right) = 0 \Rightarrow \text{Cov}\left(e, \hat{Y}\right) = 0 \Rightarrow \sum \left(\hat{Y}_i - \bar{Y}\right) e_i = 0$$

$$\text{r}_{Y\hat{Y}} = \frac{\sum \left(\hat{Y}_i - \bar{\hat{Y}}\right) \left(Y_i - \bar{Y}\right)}{\sqrt{\sum \left(\hat{Y}_i - \bar{\hat{Y}}\right)^2 \sum \left(Y_i - \bar{Y}\right)^2}} = \frac{\sum \left(\hat{Y}_i - \bar{Y}\right)\left(\hat{Y}_i - \bar{Y}\right) + \sum \left(\hat{Y}_i - \bar{Y}\right) e_i}{\sum \left(Y_i - \bar{Y}\right) \left(\hat{Y}_i - \bar{Y}\right)^2} = \sqrt{\frac{\sum \left(\hat{Y}_i - \bar{Y}\right)^2}{\sum \left(Y_i - \bar{Y}\right)^2}} = \frac{SS_{Reg}}{SS_T} = \sqrt{R^2}$$

Problem 8.

A new born baby was weighted weekly. Twenty such weights are shown below, recorded in ounces. Fit to the data, using orthogonal polynomials, a polynomial model of degree justified by the accuracy of the figures, that is, test as you go along for the significance of the linear, quadratic and so fourth, terms.

No. of weeks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Weights	141	144	148	150	158	161	166	170	175	181	189	194	196	206	218	229	234	242	247	257

Solution:

We wish to fit the model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \ldots + \beta_k X^k + \epsilon$$

Polynomial fitting using orthogonal polynomial:

$Y = \alpha_0 + \alpha_1 P_1(X) + \alpha_2 P_2(X) + \ldots + \alpha_k P_k(X) + \epsilon;\quad \left[P_i(x) \text{ are } i^{th} \text{ orthogonal polynomial}\right]$. The parameters can be estimated as $\hat{\alpha_0} = \bar{y}$ and $\hat{\alpha_j} = \frac{\sum P_j(x_i) y_i}{\sum P_j^2(x_i)}$. Given the data (x_i, y_i), we compute total variability (SS_T), then build a model to explain this variability. Suppose that $SS_{Reg}(\alpha_1)$ represents how much of the total variability in Y is explained by the linear term, similarly for cubic and quadratic.

$$SS_{Reg}(\alpha_1) = \hat{\alpha}_1 \sum_{i=1}^{n} y_i P_1(x_i) = 25438.75 \text{ (linear term)}, \; SS_{Reg}(\alpha_2) = \hat{\alpha}_2 \sum_{i=1}^{n} y_i P_2(x_i) = 489 \text{ (quadratic term)}$$

$$SS_{Reg}(\alpha_3) = \hat{\alpha}_3 \sum_{i=1}^{n} y_i P_3(x_i) = 1.15 \text{ (cubic term)}, \; SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 26018$$

Source	DF	SS	MS	F-obs	F-tab
Regression (α_1)	1	25438.75	25438.75	4558.98	$F_{0.05,1,6} = 4.49$
Regression (α_2)	1	489	489	84.63	
Regression (α_3)	1	1.15	1.15	0.21	
Residual	16	89.30	5.58		
Total	19	26018			

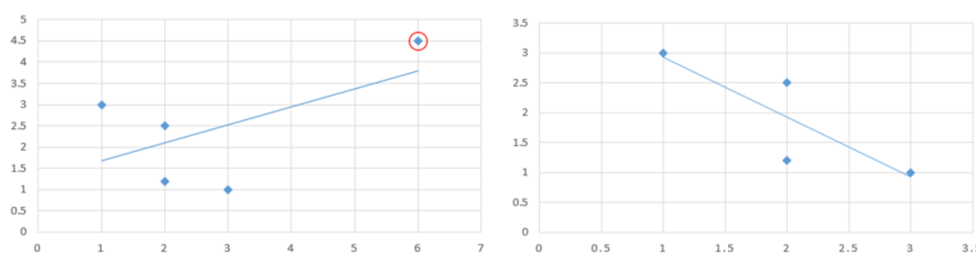
As we can see, the linear term explains the major part of variability, and α_2 is also significant (F-value). Thus the final model is $\hat{y} = 136.227 + 2.68x + 0.167x^2$, (quadratic fit). If residual SS_{Res} is large, you may check for fourth-degree polynomial fitting.

Problem 9.

If you are asked to fit a straight line to the data $(X, Y) = (1, 3), (2, 2.5), (2, 1.2), (3, 1),$ and $(6, 4.5)$. What would you do about it?

Solution:

Here from the figure below, we can observe that $(6, 4.5)$ is an influential observation. Including the influential observation, we have a positive slope in the regression model, and without the influential observation, we have a negative slope in the regression model.



With influential observation

Without influential observation

Recommendation: You can ignore influential observation if it's small in number. Some observations between $X = 3$ and $X = 6$ would be useful here.

Problem 10.

Your friend says he/she has fitted a plane to $n = 33$ observations on (X_1, X_2, Y) and his/ her overall regression (given β_0) is just significant at the $\alpha = 0.05$ level. You ask him/ her for R^2 value but s/he does not know. You work it out for him/ her based on what s/he has told you.

Solution:

The equation of the plane is

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

Here we have n = 33 observations

Source	DF	SS	MS	F-obs	F-tab
Regression	2	SS_{Reg}	MS_{Reg}	F	$F_{0.05,2,30} = 3.32$
Residual	30	SS_{Res}	MS_{Res}		
Total	32	SS_T			

$$R^2 = \frac{SS_{Reg}}{SS_T} = \frac{SS_{Reg}}{SS_{Reg} + SS_{Res}} = \frac{SS_{Reg}/MS_{Res}}{\frac{SS_{Reg}}{MS_{Res}} + \frac{SS_{Res}}{MS_{Res}}} = \frac{2MS_{Reg}/MS_{Res}}{\frac{2MS_{Reg}}{MS_{Res}} + \frac{30MS_{Res}}{MS_{Res}}} = \frac{2F}{2F+30} = \frac{2 \times 3.32}{(2 \times 3.32) + 30} = 0.1812$$

$\therefore$ 18% of the total variability is explained by the model.

Implication: Thus, R^2 is a "good" measure to measure the "goodness of fit" even when the statistical test suggests that the regression is significant.

Problem 11.

You are given a regression printout that shows a planar to fit X_1, X_2, X_3, X_4, X_5 plus an intercept term obtained from a set of 50 observations. The overall F for regression is ten times as high as the 5% upper-tail F percentage point. How big is R^2 ?

Solution:

The regression equation is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$$

The ANOVA table is We know,

Source	DF	SS	MS	F-obs	F-tab
Regression	5	SS_{Reg}	MS_{Reg}	F	$F_{0.05,5,44} = 2.43$
Residual	44	SS_{Res}	MS_{Res}		$10 \times F_{0.05,5,44} = 24.3$
Total	49	SS_T			

$$R^2 = \frac{SS_{Reg}}{SS_T} = \frac{SS_{Reg}}{SS_{Reg} + SS_{Res}} = \frac{SS_{Reg}/MS_{Res}}{\frac{SS_{Reg}}{MS_{Res}} + \frac{SS_{Res}}{MS_{Res}}} = \frac{5MS_{Reg}/MS_{Res}}{\frac{5MS_{Reg}}{MS_{Res}} + \frac{44MS_{Res}}{MS_{Res}}} = \frac{5F}{5F+44} = \frac{5 \times 24.3}{(5 \times 24.3) + 44} = 0.7343.$$

Conclusion: 73.43% of the total variability in the response variable is explained by the fitted model.

Problem 12.

Consider the simple linear regression model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where the variance of ϵ_i is proportional to x_i^2, i.e., $\mathrm{V}(\epsilon_i) = \sigma^2 x_i^2$ (assumption of constant variance is NOT satisfied).

(a) Suppose that we use these transformation $y' = \frac{y}{x}$ and $x' = \frac{1}{x}$. Is this a variance-stabilizing transformation?

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

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

Solution:

(a)

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \longrightarrow \text{ original model}$$

$$\frac{y_i}{x_i} = \frac{\beta_0}{x_i} + \beta_1 + \frac{\epsilon_i}{x_i},\ y_i' = \beta_0 x_i' + \beta_1 + \epsilon_i' \longrightarrow \text{ Transformed model}$$

Now, $\mathrm{Var}(y_i') = \mathrm{Var}\left(\frac{\epsilon_i}{x_i}\right) = \frac{\sigma^2 x_i^2}{x_i^2} = \sigma^2$. Yes, it's a variance-stabilizing transformation.

(b) Slope in the transformed model became an intercept in the transformed model and vice-versa.

(c) **Weighted LS function:**

$$S_{OLS}(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \Longrightarrow \text{ Oridinary Least Squares}$$

$$S_{WLS}(\beta_0, \beta_1) = \sum_{i=1}^{n} w_i (y_i - \beta_0 - \beta_1 x_i)^2 \Longrightarrow \text{ Weighted Least Squares}$$

$$= \sum_{i=1}^{n} \frac{1}{x_i^2} (y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n} \left(\frac{y_i}{x_i} - \frac{\beta_0}{x_i} - \beta_1\right)^2$$

Calculate β_0 and β_1 by minimizing $S_{WLS}(\beta_0, \beta_1)$ in Weighted Least Squares method.

For the transformed model, OLS estimates are:

$$S^*(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i' - \beta_0 x_i' - \beta_1)^2 = \sum_{i=1}^{n} \left(\frac{y_i}{x_i} - \frac{\beta_0}{x_i} - \beta_1\right)^2$$

Calculate β_0 and β_1 by minimizing $S^*(\beta_0, \beta_1)$ for the transformed model.

Here, $S^*(\beta_0, \beta_1) \equiv S_{WLS}(\beta_0, \beta_1)$.

Problem 13.

Consider the simple linear regression model $y_t = \beta_0 + \beta_1 x_t + \epsilon_t$ where the errors are generated by second-order autoregressive process

$$\epsilon_t = \rho_1 \epsilon_{t-1} + \rho_2 \epsilon_{t-2} + z_t,$$

where z_t is an NID $(0, \sigma_z^2)$ random variable, and ρ_1 and ρ_2 are autocorrelation parameters. Discuss how the Cochrane-Orcutt iterative procedure could be used in this situation. What transformations would be appropriate on the variables y_t and x_t? How would you estimate the parameters ρ_1 and ρ_2?

Solution:

$y_t = \beta_0 + \beta_1 x_t + \epsilon_t$ where $\epsilon_t = \rho_1 \epsilon_{t-1} + \rho_2 \epsilon_{t-2} + z_t;\ z_t \overset{\text{ind}}{\sim} N(0, \sigma_z^2)$ [In OLS-based SLR, we assume $\epsilon_t \overset{\text{ind}}{\sim} \text{N}(0, \sigma^2)$ which is Not true here.]

$$\begin{aligned} y_t \to y_t' = y_t - \rho_1 y_{t-1} - \rho_2 y_{t-2} &= (\beta_0 + \beta_1 x_t + \epsilon_t) - \rho_1 (\beta_0 + \beta_1 x_{t-1} + \epsilon_{t-1}) - \rho_2 (\beta_0 + \beta_1 x_{t-2} + \epsilon_{t-2}) \\ &= (\beta_0 - \rho_1 \beta_0 - \rho_2 \beta_0) + \beta_1 (x_t - \rho_1 x_{t-1} - \rho_2 x_{t-2}) + (\epsilon_t - \rho_1 \epsilon_{t-1} - \rho_2 \epsilon_{t-2}) = \beta_0' + \beta_1 x_t' + z_t \end{aligned}$$

Now z_t's are independent and $z_t \sim N(0, \sigma_z^2)$. But, (y_t', x_t') cannot be directly used as

$$y_t' = y_t - \rho_1 y_{t-1} - \rho_2 y_{t-2} \text{ and } x_t' = x_t - \rho_1 x_{t-1} - \rho_2 x_{t-2}$$

are functions of unknown parameters ρ_1 and ρ_2.

We know $\epsilon_t = \rho_1 \epsilon_{t-1} + \rho_2 \epsilon_{t-2} + z_t$; (How to estimate ρ_1 and ρ_2?).

- Fit $y_t = \beta_0 + \beta_1 x_t + \epsilon_t$ using OLS and obtain e_i (ignoring autocorrelation)
- Regress e_i on e_{i-1} and e_{i-2}, i,e., $e_i = \rho_1 e_{i-1} + \rho_2 e_{i-2} + z_t$ (MLR with two regressors)
- Compute $S(\rho_1, \rho_2) = \sum (e_i - \rho_1 e_{i-1} - \rho_2 e_{i-2})^2$ and minimize $S(\rho_1, \rho_2)$ and obtain $\hat{\rho}_1$ and $\hat{\rho}_2$.

$$\frac{\partial S}{\partial \rho_1} = 0 \Rightarrow \sum (e_i - \rho_1 e_{i-1} - \rho_2 e_{i-2}) e_{i-1} = 0,\ \frac{\partial S}{\partial \rho_2} = 0 \Rightarrow (e_i - \rho_1 e_{i-1} - \rho_2 e_{i-2}) e_{i-2} = 0$$

  These will generate the LSE of $\hat{\rho}_1$ and $\hat{\rho}_2$.
- $y_t' = y_t - \hat{\rho}_1 y_{t-1} - \hat{\rho}_2 y_{t-2}$ and $x_t' = x_t - \hat{\rho}_1 x_{t-1} - \hat{\rho}_2 x_{t-2}$ and apply OLS to the transformed data $y_t' = \beta_0' + \beta_1 x_t' + z_t$, where $z_t \overset{\text{ind}}{\sim} N(0, \sigma_z^2)$.
- Final fitted model is: $\hat{y}_t' = \hat{\beta}_0' + \hat{\beta}_1 x_t$.

Problem 14.

The following 24 residuals from a straight line fit are equally spaced and are given in time sequential order. Is there any evidence of lag-1 serial correlation?

$$8, -5, 7, 1, -3, -6, 1, -2, 10, 1, -1, 8, -6, 1, -6, -8, 10, -6, 9, -3, 3, -5, 1, -9$$

Use a two-sided test at level $\alpha = 0.05$

Solution:

The correlation between the residuals, $e_i, \quad i = 1(1)24$ is $\text{Cor}(e_u, e_{u+1}) = \rho$. (if $\rho \neq 0$, there is autocorrelation).

We test the hypothesis $H_0 : \rho = 0$ vs $H_1 : \rho \neq 0$, thus H_0 indicates that there is no lag-1 autocorrelation. To perform this test we compute the Durbin - Watson test statistic:

$$d = \frac{\sum_{u=2}^{24} (e_u - e_{u-1})^2}{\sum_{u=1}^{24} e_u^2} = \frac{2225}{834} = 2.67 \Rightarrow 4 - d = 1.33$$

Now, compare with d_L and d_U values from d table. For $\alpha = 0.025$. (two-sided test) $n = 24, k = 1$ (since straight line fit with one regressor variable) $d_L = 1.16, d_U = 1.33$

- If $d < d_L$ and $4 - d < d_L$: reject H_0. Here, we accept H_0 as $d = 2.67 > 1.16$ (then is no lag-1 autocorrelation).
- If $d > d_U$ and $4 - d > d_U$: accept H_0. Here, we accept H_0 as $d = 2.67 > 1.33$ (then is no lag-1 autocorrelation).

Thus, there is no lag-1 autocorrelation/serial correlation in the data.

Problem 15.

Estimate the parameters α & β in the non-linear model $Y = \alpha + (0.49 - \alpha)e^{-\beta(X-8)} + \epsilon$ from the following observations:

X	8	10	12	14	16	18	20	22	24	26	28	30	32	34	36	38	40	42
Y	0.490	0.475	0.450	0.433	0.458	0.423	0.407	0.407	0.407	0.405	0.393	0.405	0.400	0.395	0.400	0.390	0.407	0.390

Solution:

The problem is to estimate α, β of the non-linear model using the data, the residual sum of square can be written as

$$S(\alpha, \beta) = \sum_u (y_u - f(x_u, \alpha, \beta))^2 = \sum_u (y_u - \alpha - (0.49 - \alpha)e^{-\beta(x_u - 8)}))^2$$

$$f(x_u, \alpha, \beta) = \alpha + (0.49 - \alpha)e^{-\beta(x_u-8)}$$

[Since, f is nonlinear, we solve the system of nonlinear eqns by using the Taylor series approximation of nonlinear into linear one]

$$\frac{\partial f}{\partial \alpha} = 1 - e^{-\beta(x_u-8)}, \quad \frac{\partial f}{\partial \beta} = -(0.49 - \alpha)e^{-\beta(x_u-8)}(x_u - 8)$$

Linearization: Taylor series expansion of $f(x_u, \alpha, \beta)$ about the point (α_0, β_0) is

$$\begin{aligned} f(x_u; \alpha, \beta) &= f(x_u, \alpha_0, \beta_0) + (1 - e^{-\beta_0(x_u-8)})(\alpha - \alpha_0) + [-(0.49 - \alpha_0)e^{-\beta_0(x_u-8)}(x_u - 8)](\beta - \beta_0) \\ &= f_u^0 + z_{1u}^0(\alpha - \alpha_0) + z_{2u}^0(\beta - \beta_0) \text{ [linear function from nonlinear function using Taylor approximmation]} \\ Y_u &= f_u^0 + z_{1u}^0(\alpha - \alpha_0) + z_{2u}^0(\beta - \beta_0) + \epsilon_u \\ \Rightarrow Y_u - f_u^0 &= z_{1u}^0(\alpha - \alpha_0) + z_{2u}^0(\beta - \beta_0) + \epsilon_u \longrightarrow \text{[MLR model]} \\ \Rightarrow Y_0 &= z_0\theta_0 + \epsilon \longrightarrow \text{[In matrix form]} \\ \Rightarrow \hat{\theta_0} &= (z_0^T z_0)^{-1} z_0^T Y_0 \text{ is the least square estimate} \end{aligned}$$

where

$$Y_0 = \begin{bmatrix} Y_1 - f_1^0 \\ \vdots \\ Y_n - f_n^0 \end{bmatrix}, \; z_0 = \begin{bmatrix} z_{11}^0 & z_{21}^0 \\ \vdots & \vdots \\ z_{1n}^0 & z_{2n}^0 \end{bmatrix}, \; \theta_0 = \begin{bmatrix} \alpha - \alpha_0 \\ \beta - \beta_0 \end{bmatrix}, \; \epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

- If we begin the iteration with initial guess $\alpha_0 = 0.30, \beta_0 = 0.02$

$$z_0 = \begin{bmatrix} 1 - e^{-\beta_0(x_1-8)} & -(0.49 - \alpha_0)(x_1 - 8)e^{-\beta_0(x_1-8)} \\ \vdots & \vdots \\ 1 - e^{-\beta_0(x_n-8)} & -(0.49 - \alpha_0)(x_n - 8)e^{-\beta_0(x_n-8)} \end{bmatrix}$$

iteration	α_j	β_j
0	0.30	0.02
1	0.84	0.10

iteration	α_j	β_j
2	0.3901	0.1004
3	0.3901	0.1016
4	0.3901	0.1016

- Iteration continues and obtain α_{j+1} and β_{j+1}.
- This process continue until $|\alpha_{j+1} - \alpha_j| < \delta$ and $|\beta_{j+1} - \beta_j| < \delta = 0.0001$. So, we stop here.

Problem 16.

Look at these data. I don't know whether to fit two straight lines, one straight line or what. How to solve this dilemma?

X	8	0	12	2
Y	5.3	0.9	7.1	2.4

(a) Set A

X	9	7	8	6
Y	5.1	4.4	5.2	3.8

(b) Set B

Solution:

If we attach a dummy variable Z to distinguish the two groups (such that $Z = 0$ for set A and $Z = 1$ for set B), we can look at all 4 possibilities.

$$Y = (\beta_0 + \beta_1 X) + Z(\alpha_0 + \alpha_1 X) + \epsilon = \beta_0 + \beta_1 X + \alpha_0 Z + \alpha_1 XZ + \epsilon$$

Thus the $\underset{\sim}{X}$ matrix becomes

$$\underset{\sim}{X} = \begin{bmatrix} \mathbf{1} & \mathbf{X} & \mathbf{Z} & \mathbf{XZ} \\ 1 & 8 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 12 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 1 & 9 & 1 & 9 \\ 1 & 7 & 1 & 7 \\ 1 & 8 & 1 & 8 \\ 1 & 6 & 1 & 6 \end{bmatrix}$$

Thus we have $Y = X\beta + \epsilon \Rightarrow \beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T$ and $\hat{\beta} = (X^T X)^{-1} X^T Y \Rightarrow \hat{Y} = 1.142 + 0.506X - 0.0418Z - 0.036XZ$

Case: Test if a single line is sufficient i.e., $H_0 : \alpha_0 = \alpha_1 = 0$

$$F = \frac{\{SS_{Reg}(\textbf{Full}) - SS_{Reg}(\textbf{Restricted Model})\} \,/\, \{df(\textbf{Full}) - df(\textbf{Restricted Model})\}}{MS_{Res}}$$
$$= \frac{0.1818/(3-1)}{0.3272/4} = 1.11 < F_{0.05,2,4}$$

Hence, we fail to reject H_0 and can go for a single straight-line fit.

Problem 17.

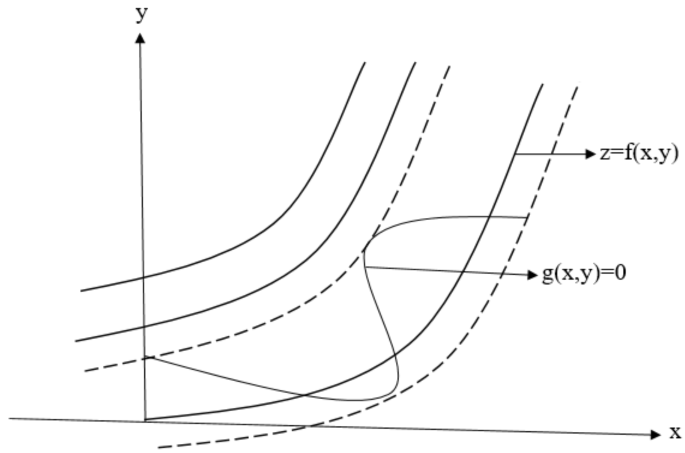
Let $\underset{\sim}{x}$ be a vector of p random variables and α_k is a vector of p constants and we write $\alpha_k^T \underset{\sim}{x} = \sum_{j=1}^{p} \alpha_{kj} x_j$. Also, let S be the (known) sample covariance matrix for the random variable $\underset{\sim}{x}$. For $k = 1, 2$, show that the k^{th} principal component is given by $z_k = \alpha_k^T \underset{\sim}{x}$ where α_k is an eigenvector of S corresponding to its k^{th} largest eigenvalue λ_k. [Principal component Regression]

Solution:

The solution is given in the teaching material, the method is based on Lagrange's multipliers. We will explain the technique below.

METHOD OF LAGRANGE'S MULTIPLIERS: Suppose we wish to minimize or maximize a function of two variables $z = f(x, y)$ where (x, y) is constrained to satisfy $g(x, y) = 0$. Assuming that these functions have continuous derivatives, we can visualize $g(x, y) = 0$ as a curve along with the level curve of $z = f(x, y)$.

Intuitively, if we move the level curve in the direction of increasing z, the largest or smallest z occurs at a point where a level curve touches $g(x, y) = 0$. The quadrants of 'f' and 'g ' should be in the same or opposite direction. Then $\nabla f = -\lambda \nabla g$ for some constant $\lambda \ni \nabla\{f + \lambda g\} = 0$.



Proof. $g(x,y)=0$, $-\frac{dy}{dx}=-\frac{g_x}{g_y}$ and for $f(x,y)=c$, $\frac{dy}{dx}=-\frac{f_x}{f_y}$.

$$\text{At the point of tangency, } -\frac{f_x}{f_y}=\frac{dy}{dx}=-\frac{g_x}{g_y} \Rightarrow \quad \frac{f_x}{f_x}=\frac{f_y}{g_y}=-\lambda \text{ (say) } \therefore (f_x,f_y)=-\lambda\,(g_x,g_y)$$

□

Hence, to find the maximum on minimum of $f(x,y)$ subject to $g(x,y)=0$, we find all the solution of equation,

$$\nabla\{f+\lambda g\}=\underset{\sim}{0} \text{ and } g(x,y)=0$$
$$\Rightarrow \frac{\partial F}{\partial x}=0=\frac{\partial F}{\partial y},\ g(x,y)=0;\ \text{where } F(x,y)=f(x,y)+\lambda g(x,y).$$

Local maxima and minima will be among the solutions. If the curve $g(x,y)=0$ is closed and bounded, then the absolute maxima and minima of $f(x,y)$ exist and are among these solutions.

General Case: To maximize or minimizes $z=f\left(x_1,x_2,\ldots,x_n\right)$ subject to the constraints $g_i\left(x_1,x_2,\ldots,x_n\right)=0; i=1(1)k$, solve the following equations simultaneously,

$$\nabla\left\{f+\sum_{i=1}^{k}\lambda_i g_i\right\}=0 \text{ and } g_i\left(x_1,x_2,\ldots,x_n\right)=0, i=1(1)k.$$

The numbers $\lambda_1,\lambda_2,\ldots,\lambda_k$ are called the Lagrange's multipliers. The method for finding the extrema of a function subject to some constraints is called the "method of Lagrange's Multipliers".

EXAMPLE OF LAGRANGE'S MULTTPUERS:
Maximize $f(x,y)=x^2y$ subject to $x^2+xy=12$.
We let $F(x,y)=x^2y+\lambda\left(x^2+xy-12\right)$

$$0=\frac{\partial F}{\partial x}=2xy+\lambda(2x+y)\rightarrow\text{(i)},\ 0=\frac{\partial F}{\partial y}=x^2+\lambda x\rightarrow\text{(ii)},\ x^2+xy=12\rightarrow\text{(iii)}$$

From (ii) $\rightarrow x(x+\lambda)=0\Rightarrow x=-\lambda$ as $x=0$ is not a solution of $x^2+xy=12$.
From (i) $\rightarrow -2\lambda y+\lambda 2(-\lambda)+\lambda y=0\Rightarrow -\lambda y=2\lambda^2\Rightarrow y=-2\lambda$
From (iii) $\rightarrow x=-\lambda, y=-2\lambda$, then $x^2+xy=12$ gives $\lambda=\pm 2$.
$\therefore (x,y)=(-2,-4)$ or $(2,4)$. Hence $\max\{xy\}=16, \min\{xy\}=-16$.

Problem 18.
Find the maximum and minimum value of $\underset{\sim}{x}'A\underset{\sim}{x}$ subject to $\underset{\sim}{x}'\underset{\sim}{x}=1$.
Solution:

$$F(\underset{\sim}{x}) = \underset{\sim}{x}^T A \underset{\sim}{x} - \lambda(\underset{\sim}{x}^T \underset{\sim}{x} - 1)$$

$$\frac{\partial F(\underset{\sim}{x})}{\partial \underset{\sim}{x}} = \underset{\sim}{0}$$

$$\Rightarrow \frac{\partial}{\partial \underset{\sim}{x}} \{\underset{\sim}{x}^T A \underset{\sim}{x} - \lambda(\underset{\sim}{x}^T \underset{\sim}{x} - 1)\} = \underset{\sim}{0} \Rightarrow 2A\underset{\sim}{x} - 2\lambda \underset{\sim}{x} = \underset{\sim}{0} \Rightarrow A\underset{\sim}{x} = \lambda \underset{\sim}{x} \Rightarrow \underset{\sim}{x}^T A \underset{\sim}{x} = \underset{\sim}{x}^T \lambda \underset{\sim}{x} = \lambda \underset{\sim}{x}^T \underset{\sim}{x} = \lambda \text{ [since, } \underset{\sim}{x}^T \underset{\sim}{x} = 1]$$

$$\therefore \text{Max}\{\underset{\sim}{x}^T A \underset{\sim}{x}\} = \max\{\lambda_i\} = \lambda_{(n)} = \text{Largest eigenvalue of } \underset{\sim}{x} \ni \underset{\sim}{x}^T \underset{\sim}{x} = 1.$$

$$\text{Min}\{\underset{\sim}{x}^T A \underset{\sim}{x}\} = \min\{\lambda_i\} = \lambda_{(1)} = \text{Smallest eigenvalue of } \underset{\sim}{x} \ni \underset{\sim}{x}^T \underset{\sim}{x} = 1.$$

Problem 19.

Show that $\|\hat{\beta}_{\text{Ridge}}\|$ increases as its tuning parameter $\lambda \to 0$. Does the same property hold for the LASSO regression?

Solution

SVD Decomposition of $X = U_{n\times p} D_{p\times p} V_{p\times p}^T$

$$\hat{\beta}_{\text{Ridge}} = \left(X^T X + \lambda I\right)^{-1} X^T Y = \left(V D^2 V^T + \lambda I\right)^{-1} V D U^T Y = \left(V\left(D^2 + \lambda I\right) V^T\right)^{-1} V D U^T Y = V^T \left(D^2 + \lambda I\right)^{-1} D U^T Y.$$

$$\|\hat{\beta}_{\text{Ridge}}\|_2^2 = Y^T U D (D^2 + \lambda I)^{-1} (D^2 + \lambda I)^{-1} D U^T Y = (U^T Y)^T [D(D^2 + \lambda I)^{-2} D](U^T Y) = \sum_{j=1}^{p} \frac{d_j^2 (U^T Y)_j^2}{(d_j^2 + \lambda)^2};$$

where $D\left(D^2 + \lambda I\right)^{-1} D$ represents a diagonal matrix with elements $\frac{d_j^2}{\left(d_j^2 + \lambda\right)^2}$.

Therefore, we see that $\|\hat{\beta}_{\text{Ridge}}\|$ increases as its tuning parameter $\lambda \to 0$. Recall the dual form of LASSO as defined below:

$$\hat{\beta}_{\text{Lasso}} = \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2$$

$$\text{s.t.} \quad \sum_{j=1}^{p} |\beta_j| \leq t$$

$$\hat{\beta}_{\text{Lasso}} = \arg\min_{\beta} \left\{\frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|\right\}$$

It is easy to see that t and λ have an inverse relationship; therefore, as $\lambda \to 0, t$ increases and so does the norm of optimal solutions.

Problem 20.

Consider a two-class logistic regression problem with $x \in \mathbb{R}$. Characterize the maximum-likelihood estimates of the slope and intercept parameter if the sample x_i for the two classes are separated by a point $x_0 \in \mathbb{R}$. Generalize this result to (a) $x \in \mathbb{R}^p$ and (b) more than two classes.

Solution

Without loss of generality, suppose that $x_0 = 0$ and that the coding is $y = 1$ for $x_i > 0$ and $y = 0$ for $x_i < 0$. Now, suppose that

$$p(x; \beta) = \frac{\exp\{\beta x + \beta_0\}}{1 + \exp\{\beta x + \beta_0\}}$$

so that

$$1 - p(x; \beta) = \frac{1}{1 + \exp\{\beta x + \beta_0\}}$$

Since $x_0 = 0$ is the boundary then $p(x_0) = 1 - p(x_0)$ then $\beta_0 = 0$. Therefore,

$$p(x; \beta) = \frac{\exp\{\beta x\}}{1 + \exp\{\beta x\}}$$

so that

$$1 - p(x; \beta) = \frac{1}{1 + \exp\{\beta x\}}.$$

Therefore, the likelihood function

$$L(\beta; y, x) = \prod_{i=1}^{N} p\left(x_i; \beta\right)^{y_i} \left[1 - p\left(x_i; \beta\right)\right]^{1-y_i} = \prod_{i=1}^{N} \left[\frac{p\left(x_i; \beta\right)}{1 - p\left(x_i; \beta\right)}\right]^{y_i} \left[1 - p\left(x_i; \beta\right)\right] = \prod_{i=1}^{N} \left[\exp\left\{\beta x_i\right\}\right]^{y_i} \left[1 + \exp\left\{\beta x_i\right\}\right]$$

so that the log-likelihood function

$$l(\beta; y, x) = \sum_{i=1}^{N} y_i \left[\beta x_i\right] - \log\left[1 + \exp\left\{\beta x_i\right\}\right]$$

Taking the derivative with respect to β and substituting in the proper coding of y_i gives

$$\begin{aligned}
\frac{dl(\beta; x, y)}{d\beta} &= \sum_{i=1}^{N} x_i \left(y_i - \frac{\exp\left\{\beta x_i\right\}}{1 + \exp\left\{\beta x_i\right\}}\right) \\
&= \sum_{x_i > 0} x_i \left(1 - \frac{\exp\left\{\beta x_i\right\}}{1 + \exp\left\{\beta x_i\right\}}\right) - \sum_{x_i < 0} x_i \left(\frac{\exp\left\{\beta x_i\right\}}{1 + \exp\left\{\beta x_i\right\}}\right) \\
&= \sum_{x_i > 0} x_i - \sum_{x_i > 0} x_i \left(\frac{\exp\left\{\beta x_i\right\}}{1 + \exp\left\{\beta x_i\right\}}\right) - \sum_{x_i < 0} x_i \left(\frac{\exp\left\{\beta x_i\right\}}{1 + \exp\left\{\beta x_i\right\}}\right).
\end{aligned}$$

Setting the above equal to zero gives

$$\sum_{x_i > 0} x_i = \sum_{i=1}^{N} x_i \left(\frac{\exp\left\{\beta x_i\right\}}{1 + \exp\left\{\beta x_i\right\}}\right).$$

Clearly, for any data set $\{x_i\}_{i=1}^{N}$ we must have that $\beta \to \infty$ for the above equality to hold.

(b) Now, suppose that there are K classes such that x_1 seperates classes one and two, x_2 seperates classes two and three, and so on to x_{K-1} that seperates classes $K-1$ and K with $-\infty = x_0 < x_1 < x_2 < \ldots < x_{K-1} < x_K = \infty$. Now, define probabilities

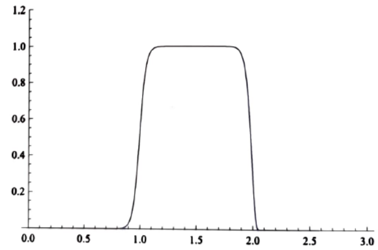
$$\begin{aligned}
p_1(x; \beta) &= \frac{\exp\left\{\beta_1 x + \beta_{01}\right\}}{1 + \sum_{j=1}^{K-1} \exp\left\{\beta_j x + \beta_{0j}\right\}} \\
p_2(x; \beta) &= \frac{\exp\left\{\beta_2 x + \beta_{02}\right\}}{1 + \sum_{j=1}^{K-1} \exp\left\{\beta_j x + \beta_{0j}\right\}} \\
&\vdots \\
p_{K-1}(x; \beta) &= \frac{1}{1 + \sum_{j=1}^{K-1} \exp\left\{\beta_j x + \beta_{0j}\right\}}.
\end{aligned}$$

Now, suppose that the coding is $y_i = 1$ if $x_{j-1} < x_i < x_j$ and $y_i = 0$ otherwise for observation $i = 1, \ldots, N$ and class $j = 1, \ldots, K$. Therefore, the likelihood function

$$L(\beta; y, x) = \prod_{j=1}^{K} \prod_{i=1}^{N_j} \left[p_j\left(x_i; \beta\right)\right]^{y_i}$$

where N_j is the number of observations in class j, so that the log-likelihood function

$$\begin{aligned}
l(\beta; y, x) &= \sum_{j=1}^{K-1} \sum_{i=1}^{N_j} y_i \log\left[\frac{\exp\left\{\beta_j x_i + \beta_{0j}\right\}}{1 + \sum_{j=1}^{K-1} \exp\left\{\beta_j x + \beta_{0j}\right\}}\right] + \sum_{i=1}^{N_k} y_i \log\left[\frac{1}{1 + \sum_{j=1}^{K-1} \exp\left\{\beta_j x_i + \beta_{0j}\right\}}\right] \\
&= \sum_{j=1}^{K-1} \sum_{i=1}^{N_j} y_i \left[\beta_j x_i + \beta_{0j}\right] - \sum_{j=1}^{K} \sum_{i=1}^{N_j} y_i \log\left[1 + \sum_{j=1}^{K-1} \exp\left\{\beta_j x_i + \beta_{0j}\right\}\right]
\end{aligned}$$



Now, we determine the values of β_{0j}. First, note that β_{0j} is a function of β_j, x_{j-1}, and x_j. So that the expression $p(x;\beta)$ maintains proper form, for $x_{j-1} < x < x_j$ we define

$$
\begin{aligned}
p(x;\beta_j) &= \frac{\exp\{\beta_j(x - x_{j-1})\} - \exp\{\beta_j(x - x_j)\}}{1 + \sum_{j=1}^{K-1} \exp\{\beta_j x_i + \beta_{0j}\}} \\
&= \frac{\exp\{\beta_j x\}[\exp\{\beta_j x_{j-1}\} - \exp\{\beta_j x_j\}]}{1 + \sum_{j=1}^{K-1} \exp\{\beta_j x_i + \beta_{0j}\}} \\
&= \frac{\exp\{\beta_j x + \beta_{0j}\}}{1 + \sum_{j=1}^{K-1} \exp\{\beta_j x_i + \beta_{0j}\}}
\end{aligned}
$$

where $\beta_{0j} = \log[\exp\{\beta_j x_{j-1}\} - \exp\{\beta_j x_j\}]$. The reason for the begining step of the formulation above is due to the fact that, for example, when $x \in (x_1, x_2)$ so that x classifies to class two, the probability function appears as in the following figure, where it was assumed that $x_1 = 1$ and $x_2 = 2$.

Now, taking the derivative with respect to $\beta = (\beta_1, \ldots, \beta_{K-1})$ and substituting in the proper coding of y_i gives

$$
\begin{aligned}
\frac{dl(\beta;x,y)}{d\beta_j} &= \sum_{i=1}^{N_j} x_i + \sum_{i=1}^{N_j} \frac{\exp\{\beta_j x_{j-1}\} x_{j-1} - \exp\{\beta_j x_j\} x_j}{\exp\{\beta_j x_{j-1}\} - \exp\{\beta_j x_j\}} \\
&\quad - \sum_{i=1}^{N_j} \left[x_i + \frac{\exp\{\beta_j x_{j-1}\} x_{j-1} - \exp\{\beta_j x_j\} x_j]}{\exp\{\beta_j x_{j-1}\} - \exp\{\beta_j x_j\}}\right] \left(\frac{\exp\{\beta_j x_i - \beta_{0j}\}}{1 + \sum_{j=1}^{K-1} \exp\{\beta_j x_i + \beta_{0j}\}}\right)
\end{aligned}
$$

Note that the $\frac{\exp\{\beta_j x_{j-1}\} x_{j-1} - \exp\{\beta_j x_j\} x_j}{\exp\{\beta_j x_{j-1}\} - \exp\{\beta_j x_j\}}$ term in the above is a constant in the sum over $i = 1, \ldots, N_j$. Therefore, setting the above equal to zero for each $j = 1, \ldots, K-1$ and solving for β_j gives the maximum likelihood estimators in a similar fashion to the two-class case that $\beta_j \to \infty$