Tutorial letter 201/2/2018

Forecasting

STA2604

Semester 2

Department of Statistics

Solutions to Assignment 1

Define tomorrow.

UNISA university of south africa

QUESTION 1

(1.1) Forecasting helps in predicting future events and conditions based on present conditions. Here are two examples:

- Demography: Forecasting the population growth in a country for planning new infrastructures (e.g. roads, hospitals, schools).
- Business: forecasting the number of customers for a certain product in a shop for supply planning.

(1.2) The four components of a time series are the following:

- Trend: describes the upward or downward change for a time series during a given period of time.
- Seasonal variations: describes periodic patterns in a time series during a year and repeated in subsequent years.
- Cyclical variations: describes upward or downward changes around the trend that may occur after several years (e.g. five years).
- Irregular variations: describes changes in time series that follow no known (predictable) pattern.

(1.3) Let y_t and $\widehat{y_t}$ denote the actual and predicted costs, respectively, in the table below:

Weeks	Actual cost y_t	Predicted cost $\widehat{y_t}$	e_t	$\lvert e_t \rvert$	e_t^2	APE_t
1	75	73.3	1.7	1.7	2.89	2.2667
2	89.7	86.5	3.2	3.2	10.24	3.5674
3	123.6	129.4	-5.8	5.8	33.64	4.6926
4	102	99.7	2.3	2.3	5.29	2.2549
5	139	142	-3	3	9	2.1583
Total				16	61.06	14.9399

(1.4.1) The forecast error e_t for each day is given in the fourth column of the table.

(1.4.2) MAD $= \dfrac{\sum_{t=1}^{5} |e_t|}{5} = \dfrac{16}{5} = 3.2.$

(1.4.3) MSE $= \dfrac{\sum_{t=1}^{5} e_t^2}{5} = \dfrac{61.06}{5} = 12.212.$

(1.4.4) MAPE $= \dfrac{\sum_{t=1}^{5} APE_t}{5} = \dfrac{14.9399}{5} = 2.9880.$

QUESTION 2

(2.1) First construct a table with necessary information for all parts of the question.

t	y_t	ty_t	t^2	$\widehat{y_t}$	$e_t = y_t - \widehat{y_t}$	e_t^2	$d_t = (e_t - e_{t-1})^2$	$(y_t - \bar{y}_t)^2$
1	98	98	1	99.62	-1.62	2.62		5662.56
2	100	200	4	109.44	-9.44	89.03	61.11	5365.56
3	110	330	9	119.25	-9.25	85.62	0.03	4000.56
4	150	600	16	129.07	20.93	438.03	910.98	540.56
5	140	700	25	138.89	1.11	1.24	392.74	1105.56
6	160	960	36	148.71	11.29	127.55	103.68	175.56
7	200	1400	49	158.52	41.48	1720.29	910.98	715.56
8	150	1200	64	168.34	-18.34	336.40	3578.15	540.56
9	156	1404	81	178.16	-22.16	491.01	14.57	297.56
10	170	1700	100	187.98	-17.98	323.15	17.49	10.56
11	200	2200	121	197.79	2.21	4.87	407.33	715.56
12	190	2280	144	207.61	-17.61	310.17	392.74	280.56
13	220	2860	169	217.43	2.57	6.61	407.33	2185.56
14	210	2940	196	227.25	-17.25	297.45	392.74	1350.56
15	260	3900	225	237.06	22.94	526.04	1614.63	7525.56
16	258	4128	256	246.88	11.12	123.61	139.66	7182.56
136	2772	26900	1496			4883.69	9344.15	37655.00

The values of $\widehat{\beta}_1$ and $\widehat{\beta}_0$ are the following:

$$\widehat{\beta}_1 = \frac{16\sum_{t=1}^{16} ty_t - \sum_{t=1}^{16} t \sum_{t=1}^{16} y_t}{16\sum_{t=1}^{16} t^2 - \left(\sum_{t=1}^{16} t\right)^2} = \frac{16 \times 26900 - 136 \times 2772}{16 \times 1496 - (136)^2} = \frac{53408}{5440} = 9.8176.$$

$$\widehat{\beta}_0 = \bar{y}_t - \widehat{\beta}_1 \bar{t} = \frac{\sum_{t=1}^{16} y_t - \widehat{\beta}_1 \sum_{t=1}^{16} t}{16} = \frac{2772 - 9.8176 \times 136}{16} = \frac{1436.8064}{16} = 89.8004$$

Thus, the fitted model is: $\widehat{y_t} = 89.8004 + 9.8176t$.

(2.2) The point forecast at month 21 is: $\widehat{y}_{17} = 89.8004 + 9.8176 \times 17 = 256.6996$. If there was a seasonal index, this result should be multiplied by that index (or added that index). The computation of the 95% prediction interval needs some extra calculations. The fitted model and the values of t in the first column were used to calculated the fitted values in the fifth column. The residuals e_t in the sixth column are $e_t = y_t - \widehat{y}_t$. The squared residuals e_t^2 may now be used to calculate the mean square residual as: $MSE = \frac{e_t^2}{n-1}$ and thus its square root is $s = \sqrt{\frac{e_t^2}{n-2}} = \sqrt{\frac{4883.7}{14}} = 18.68$ The D distance is: $D = \frac{1}{n} + \frac{(t_0 - \bar{t})^2}{SS_{tt}}$ where $t_0 = 17$, $\bar{t} = \frac{136}{16} = 8.5$, and

$$SS_{tt} = \sum_{t=1}^{16}(t-\bar{t})^2 = \sum_{t=1}^{16} t^2 - \frac{\left(\sum_{t=1}^{16} t\right)^2}{16} = 1496 - \frac{136^2}{16} = 340.$$

Hence,

$$D = \frac{1}{n} + \frac{(t_0-\bar{t})^2}{SS_{tt}} = \frac{1}{16} + \frac{(17-8.5)^2}{340} = 0.275.$$

The 95% prediction interval is: $[\widehat{y}_t \pm t_{[0.025]}^{(16-2)} s\sqrt{1+D}]$. That is: $[256.7 \pm t_{[0.025]}^{(14)} 18.68\sqrt{1+0.275}]$ where $t_{[0.025]}^{(14)} = 2.145$. Hence, the required 95% prediction interval is: $[211.46; 301.94]$.

(2.3) The null and alternative hypotheses are:
H_0: The error terms are not autocorrelated.
H_a: The error terms are positively correlated.
The Durbin-Watson test statistic for positive autocorrelation is:

$$d = \frac{\sum_{t=2}^{16}(e_t - e_{t-1})^2}{\sum_{t=1}^{16} e_t^2} = \frac{9344.15}{4883.69} = 1.91.$$

We will reject H_0 if $d < d_{L,0.05}$, do not reject H_0 if $d > d_{L,0.05}$ and fail to conclude H_0 if $d_{L,0.05} < d <_{U,0.05}$. For $n = 16$, we have: $d_{L,0.05} = 1.10$ and $d_{U,0.05} = 1.37$. Since $d = 1.91 >_{U,0.05} = 1.37$, we do not reject H_0. We conclude that the error terms are not autocorrelated.

(2.4) The mean sales is: $\bar{y}_t = \frac{\sum_{t=1}^{16} y_t}{16} = \frac{2772}{16} = 173.25$.

The total variation is: $\sum_{t=1}^{16}(y_t - \bar{y}_t)^2 = 37655$.

The total unexplained variation is: $\sum_{t=1}^{16} e_t^2 = \sum_{t=1}^{16}(y_t - \widehat{y}_t)^2 = 4883.69$.

Thus, the coefficient of determination is:

$$R^2 = 1 - \frac{\sum_{t=1}^{16}(y_t - \widehat{y}_t)^2}{\sum_{t=1}^{16}(y_t - \bar{y}_t)^2} = 1 - \frac{4883.69}{37655} = 0.8703.$$

The adjusted coefficient of determination is:

$$R^2_{\text{adj}} = \bar{R}^2 = \left(R^2 - \frac{k}{n-1}\right)\left(\frac{n-1}{n-(k+1)}\right) = \left(0.8703 - \frac{1}{15}\right)\frac{15}{14} = 0.8610.$$

(2.5) There is no point of calculating a VIF here since there is only one predictor variable.

QUESTION 3

The first important thing is to correctly capture that data. The dataset should look as follows:

x_1	x_2	x_1x_2	y
0	-2	0	8
0	-1	0	9
0	0	0	9.1
0	1	0	10.2
0	2	0	10.4
1	-2	-2	10
1	-1	-1	10.3
1	0	0	12.2
1	1	1	12.6
1	2	2	13.9

(3.1) Model (obtained using Excel).

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9876
R Square	0.9754
Adjusted R Square	0.9630
Standard Error	0.3490
Observations	10

Note that overall, the model fits the data since R^2 and R^2_{adj} are large. However, this information is often misleading. Residual analysis is a more suitable approach.

ANOVA

	df	SS	MS	F	Significance F
Regression	3	28.9300	9.6433	79.1518	0.00003
Residual	6	0.7310	0.1218		
Total	9	29.6610			

Here, the ANOVA table again confirms that overall, the model fits the data since F is large $(F = 79.1518)$ with $p = 0.00003 < 0.05$. However, again this information is often misleading. Residual analysis is a more suitable approach. The table of parameter estimates is given below, and indicates that both x_1 (Bacteria type) and x_2 (Time), and their interaction (x_1x_2) have positive significant effects on y (growth) since all p-values are small ($\alpha = 0.05$ is often used when no significance level is indicated).

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	9.3400	0.1561	59.8341	0.0000	8.9580	9.7220
X Variable 1 (x_1)	2.4600	0.2208	11.1435	0.0000	1.9198	3.0002
X Variable 2 (x_2)	0.6000	0.1104	5.4359	0.0016	0.3299	0.8701
X Variable 3 (x_1x_2)	0.4100	0.1561	2.6266	0.0392	0.0280	0.7920

Hence the predictive model is:

$$\widehat{y} = 9.34 + 2.46x_1 + 0.6x_2 + 0.41x_1x_2.$$

(3.2) Here $x_1 = 0$ and $x_2 = 0$, and thus $\widehat{y} = 9.34$ which is not very different from the observed value 9.1. The residual is: $9.1 - 9.34 = -0.24$.

(3.3) Residuals are calculated as $e = y - \widehat{y}$ where $\widehat{y} = 9.34 + 2.46x_1 + 0.6x_2 + 0.41x_1x_2$.
Manual calculation in Excel (as requested in the question) gives the following table:

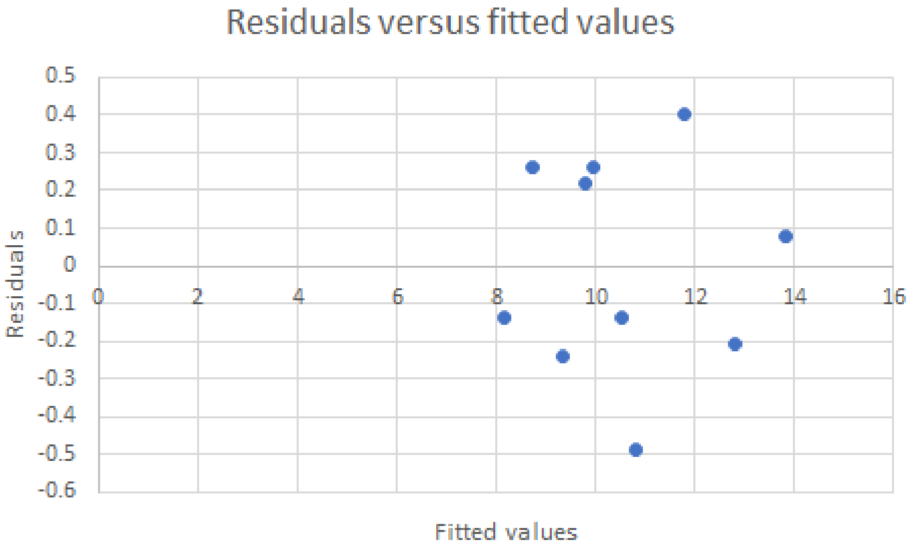
Observation	x_1	x_2	x_1x_2	y	$\widehat{y}$	e
1	0	-2	0	8	8.14	-0.14
2	0	-1	0	9	8.74	0.26
3	0	0	0	9.1	9.34	-0.24
4	0	1	0	10.2	9.94	0.26
5	0	2	0	10.4	10.54	-0.14
6	1	-2	-2	10	9.78	0.22
7	1	-1	-1	10.3	10.79	-0.49
8	1	0	0	12.2	11.8	0.4
9	1	1	1	12.6	12.81	-0.21
10	1	2	2	13.9	13.82	0.08

If an option to calculate residuals was ticked in Excel, we could obtain the following output (correct answer but the aim of the question was to assess if you know how the residuals are computed):

RESIDUAL OUTPUT

Observation	Predicted Y	Residuals
1	8.14	-0.14
2	8.74	0.26
3	9.34	-0.24
4	9.94	0.26
5	10.54	-0.14
6	9.78	0.22
7	10.79	-0.49
8	11.8	0.4
9	12.81	-0.21
10	13.82	0.08

Plot of residuals versus fitted values:



Clearly, the residuals are randomly scattered around the line $e = 0$. Thus, the assumption made in part (3.1) for a linear model with interaction is supported by residual analysis.

(3.4) For $x_1 = 1$, the fitted model is:

$$\begin{aligned}\widehat{y} &= 9.34 + 2.36 + 0.6x_2 + 0.41x_2\\ &= 11.8 + 1.01x_2\end{aligned}$$

The mean value for x_2 is $\bar{x}_2 = \frac{-2-1+0+1+2-2-1+0+1+2}{10} = 0$ and

$$\begin{aligned}S_{x_2x_2} &= \sum_{i=1}^{10}(x_{i2} - \bar{x}_2)^2\\ &= \sum_{i=1}^{10} x_{i2}^2 - \frac{(\sum_{i=1}^{10} x_{i2})^2}{10}\\ &= (-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2 + (-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2 - 0\\ &= 20.\end{aligned}$$

The distance value at $x_2 = 1$ is thus:

$$\begin{aligned}D &= \frac{1}{n} + \frac{(x_{02} - \bar{x}_2)^2}{S_{x_2x_2}}\\ &= \frac{1}{10} + \frac{(1-0)^2}{20}\\ &= \frac{1}{10} + \frac{1}{20}\\ &= 0.15.\end{aligned}$$

The fitted value at $x_2 = 1$ is $\widehat{y} = 11.8 + 1.01 = 12.81$ and for $\alpha = 0.1$, we have $t^{(10-2)}_{[\alpha/2]} = t^{(8)}_{[0.05]} = 1.860$. Hence, the 90% confidence interval of the expected growth (mean value) for type B bacteria at time $x_2 = 1$ is:

$$\begin{aligned}[\widehat{y} \pm t^{(8)}_{[0.05]} s\sqrt{D}] &= [12.81 \pm 1.860 \times 0.349\sqrt{0.15}] \\ &= [12.81 \pm 0.25] \\ &= [12.56; 13.06].\end{aligned}$$

(3.5) The 90% prediction interval of the expected growth (mean value) for type B bacteria at time $x_2 = 1$ is:

$$\begin{aligned}[\widehat{y} \pm t^{(8)}_{[0.05]} s\sqrt{1+D}] &= [12.81 \pm 1.860 \times 0.349\sqrt{1.15}] \\ &= [12.81 \pm 0.70] \\ &= [12.11; 13.51].\end{aligned}$$

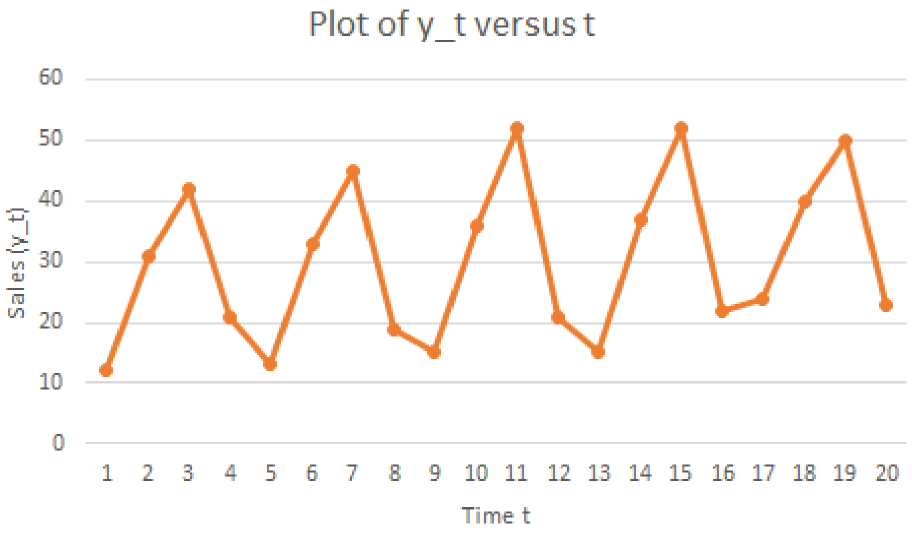
(3.6) The prediction interval is wider than the confidence interval, but this is expected to be so.

QUESTION 4

(4.1) The dummy variables are defined in the following table:

Quarter	Q_2	Q_3	Q_4
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	1

(4.2) The plot of y_t versus t (with $t = 1, 2, \ldots, 20$) is given below:



(4.3) It appears to be a constant seasonal variation since there are similar quarter patterns throughout the five years. However, the trend increases.

(4.4) The output for the fitted model is the following:

SUMMARY OUTPUT Regression Statistics	
Multiple R	0.9889
R Square	0.9779
Adjusted R Square	0.9721
Standard Error	2.2449
Observations	20

ANOVA					
	df	SS	MS	F	Significance F
Regression	4	3350.9577	837.7394	166.2352	0.0000
Residual	15	75.5923	5.0395		
Total	19	3426.5500			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	11.3317	1.2889	8.7919	0.0000	8.5845	14.0789
t	0.4965	0.0898	5.5281	0.0001	0.3051	0.6879
Q_2	19.1035	1.4226	13.4284	0.0000	16.0713	22.1358
Q_3	31.5063	1.4290	22.0484	0.0000	28.4606	34.5521
Q_4	3.9106	1.4451	2.7060	0.0163	0.8304	6.9908

This table indicates that sales significantly increase with time and there is increase seasonal variations for quarters 2, 3, 4 compared to quarter 1 (all p-values very small). The prediction model is:

$$\widehat{y} = 11.3317 + 0.4965t + 19.1035Q_2 + 31.5063Q_3 + 3.9106Q_4.$$

(4.5) $t = 17$ corresponds to the first quarter of year 4, that is $Q_2 = Q_3 = Q_4 = 0$, and thus the fitted model can be written as: $\widehat{y} = 11.3317 + 0.4965t$. Hence, the point forecast is $\widehat{y}_{17} = 11.3317 + 0.4965 \times 17 = 19.7722$. From the regression ANOVA table, $s = \sqrt{MSE} = \sqrt{5.0395} = 2.2449$ (same as standard error), and since there were 5 parameters to estimate, we have

$$t^{(n-p)}_{[\alpha/2]} = t^{(20-5)}_{[0.025]} = t^{(15)}_{[0.025]} = 2.131.$$

The mean value of t is:

$$\bar{t} = \frac{\sum_{t=1}^{20} t}{20} = \frac{1+2+\cdots+20}{20} = 10.5.$$

Then $(17-\bar{t})^2 = (17-10.5)^2 = 42.25$. Calculations for $\sum_{t=1}^{20}(t-\bar{t})^2$ can be read from the following table.

Year	Quarter	t	Q_2	Q_3	Q_4	y_t	$t-\bar{t}$	$(t-\bar{t})^2$
1	1	1	0	0	0	12	-9.5	90.25
1	2	2	1	0	0	31	-8.5	72.25
1	3	3	0	1	0	42	-7.5	56.25
1	4	4	0	0	1	21	-6.5	42.25
2	1	5	0	0	0	13	-5.5	30.25
2	2	6	1	0	0	33	-4.5	20.25
2	3	7	0	1	0	45	-3.5	12.25
2	4	8	0	0	1	19	-2.5	6.25
3	1	9	0	0	0	15	-1.5	2.25
3	2	10	1	0	0	36	-0.5	0.25
3	3	11	0	1	0	52	0.5	0.25
3	4	12	0	0	1	21	1.5	2.25
4	1	13	0	0	0	15	2.5	6.25
4	2	14	1	0	0	37	3.5	12.25
4	3	15	0	1	0	52	4.5	20.25
4	4	16	0	0	1	22	5.5	30.25
5	1	17	0	0	0	24	6.5	42.25
5	2	18	1	0	0	40	7.5	56.25
5	3	18	0	1	0	50	7.5	56.25
5	4	20	0	0	1	23	9.5	90.25
Total								649

Hence, the 95% confidence interval (not prediction interval) of y_{17} is thus:

$$\left[\widehat{y}_{17} \pm s t^{(15)}_{[0.025]} \sqrt{\frac{1}{20} + \frac{(17-\bar{t})^2}{\sum_{t=1}^{20}(t-\bar{t})^2}}\right],$$

that is:

$$\left[19.7722 \pm 2.2449 \times 2.131\sqrt{\frac{1}{20} + \frac{42.25}{649}}\right] = [19.7722 \pm 1.6230] = [18.1492; 21.3952].$$