

**Statistics 3540**  
**Assignment #3: Solutions**

1. Refer back to the DJIA data used in Assignment #2, and answer the following:

(a) Fit the regression model

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \epsilon_t$$

See output below.

```
The regression equation is
DJIA = 1150 - 124 Time + 9.51 Timesq
```

We see that  $\hat{y} = 1150 - 124t + 9.51t^2$ .

(b) Is this model useful for predicting the DJIA closing values? Base your conclusion on the p-value of the appropriate test.

$H_o : \beta_i = 0$  for all  $i$  (model not useful)

$H_a : \text{at least 1 } \beta_i \neq 0$  (model useful)

From the output below:  $F_{obs} = MSR/MSE = 12302619/16941 = 726.18$ .

p-value =  $P(F > 726.18) \approx 0$  using  $k = 2$  and  $(n - k - 1) = 22$  df.

Very strong evidence against  $H_o$ .

Therefore model appears useful in predicting DJIA closing values.

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	2	24605238	12302619	726.18	0.000
Error	22	372713	16941		
Total	24	24977950			

(c) Test at  $\alpha = 0.10$  if the errors in this model are correlated.

$H_o : \phi = 0$  (Errors are not autocorrelated)

$H_a : \phi \neq 0$  (Errors are autocorrelated)

Test Statistic:  $d_{obs} = 1.91$  (output)

Durbin-Watson statistic = 1.91

Test at  $\alpha = 0.1$ , where  $n = 25$  and  $k = 2$ .

Table A5, using  $\alpha/2 = 0.05$ :  $d_L = 1.21$ ,  $d_U = 1.55$

Since  $1.91 > 1.55$  and  $(4 - 1.91) > 1.55$ , we do not reject  $H_0$ .

Therefore the errors do not appear to be autocorrelated.

---

2. In this problem you will analyze a time series of Canadian monthly unemployment rates (numbers and percentages), observed from Jan. 1992 to Dec. 2007.

(a) Plot the unemployment rate versus time and describe any features you see in the plot.

**NOTE:** To plot this time series, it would be preferable to go into the graph's time/date options to tell Minitab this is monthly data. This makes it easier to identify at what particular times of year the imports are high (or low).

The data exhibits both seasonal variation and an decreasing (probably linear) trend. There is a peak in unemployment rates around January each year, and a drop each summer.

(b) Fit the model

$$y_t = \beta_0 + \beta_1 t + \beta_{s1} x_{s1,t} + \beta_{s2} x_{s2,t} + \dots + \beta_{s11} x_{s11,t} + e_t$$

to the data, where  $x_{s1,t} = 1$  if observation is in January, ...,  $x_{s11,t} = 1$  if observation is in November.

Part of the output is below. We see the least squares regression line is

$$\hat{y} = 10.6 - 0.0283t + 1.05x_{s1,t} + \dots + 0.009x_{s11,t}$$

where  $x_{s1,t} = 1$  if January (0 otherwise), ...,  $x_{s11,t} = 1$  if November (0 otherwise).

The regression equation is

The regression equation is

Unemploy = 10.6 - 0.0283 Time + 1.05 Jan + 0.905 Feb +

1.00 Mar + 0.742 Apr + 0.458 May - 0.051 June + 0.534 July

+ 0.512 Aug - 0.297 Sept - 0.313 Oct + 0.009 Nov

S = 0.644756    R-Sq = 87.7%    R-Sq(adj) = 86.9%

- (c) Is the model useful in predicting the unemployment rate? Base your conclusion on the p-value of the appropriate test.

$H_o : \beta_i = 0$  for all  $i$  (model not useful)

$H_a : \text{at least 1 } \beta_i \neq 0$  (model useful)

From the output below:  $F_{obs} = MSR/MSE = 106.29$ .

P-value:  $P(F > 106.29) \approx 0$  using F-distribution with  $k = 12$  and  $n - k - 1 = 179$  df.

Very strong evidence against  $H_o$ ; model appears useful.

#### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	12	530.236	44.186	106.29	0.000
Residual Error	179	74.412	0.416		
Total	191	604.649			

- (d) Test at  $\alpha = 0.05$  whether we can drop the seasonal terms from the model.

$H_o : \beta_{si} = 0$  for all  $i$  (drop seasonal terms)

$H_a : H_o$  not true

We will use the partial F-test. We have 2 ways to find  $SSE_R - SSE_C$ . I will fit 2 different models (one with, one without seasonal terms) to get the  $SSE$  values. It will give the same result as using the **Seq SS** values in Minitab.

From the output above for the complete model:  $SSE_C = 74.412$ ,  $MSE_C = 0.416$ .

A portion of the output for the reduced model (only includes trend) is below:

The regression equation is  
 Unemploy = 11.1 - 0.0287 Time

#### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	487.20	487.20	788.17	0.000
Residual Error	190	117.45	0.62		
Total	191	604.65			

#### Analysis of Variance

Therefore  $SSE_R = 117.45$ , giving us

$$SSE_R - SSE_C = 117.45 - 74.412 = 43.038.$$

$$F_{obs} = \frac{SSE_R - SSE_C / (k - g)}{MSE_C} = \frac{43.038 / 11}{0.416} = 9.41$$

where  $k - g = 11$  because we are dropping 11 terms from the model.

Test at  $\alpha = 0.05$ : Reject  $H_o$  since  $9.41 > F_{.05} = 1.91$ , using F-distribution with 10 (since  $k - g = 11$  not in table) and 120 (since  $n - k - 1 = 179$  not in table) df. So we need to keep the seasonal terms.

**NOTE:** It's fine if 10 (instead of 12) or  $\infty$  (instead of 120) degrees of freedom are used.

- (e) Find a 99% prediction interval for the Canadian unemployment rate in January 2008.

To get Minitab to find this PI, we first have to decide what  $t$  is. Our data contain  $n = 192$  observations, ending with Dec. 2007. We are looking for a PI 1 month after that. That means  $t = 192 + 1 = 193$ .

Next, to get Minitab to find the PI, under **Options**, we must tell Minitab to find the prediction when  $t = 192$ , and we will be in Jan. That means the box must have

```
193 1 0 0 0 0 0 0 0 0 0 0
```

The output is below, giving us the 99% PI as (4.4948, 7.9842).

Predicted Values for New Observations

New

Obs	Fit	SE Fit	99% CI	99% PI
1	6.2395	0.1826	(5.7641, 6.7149)	(4.4948, 7.9842)

Values of Predictors for New Observations

New

Obs	Time	Jan	Feb	Mar	Apr	May	Jun	Jul
1	193.0	1.00	0.00	0.0000	0.0000	0.0000	0.0000	0.0000

New

Obs	Aug	Sep	Oct	Nov
1	0.000000	0.000000	0.000000	0.000000

3. For the data in question #2:

- (a) Calculate the least squares regression line for the model

$$y_t = \beta_0 + \beta_1 t + \beta_2 \sin(2\pi t/L) + \beta_3 \cos(2\pi t/L) + e_t$$

using the appropriate value for  $L$ .

We use  $L = 12$ , since we have monthly data.

A portion of the output is below. We see that

$$\hat{y} = 11.0 - 0.0284t + 0.506 \sin(2\pi t/12) - 0.0366 \cos(2\pi t/12)$$

The regression equation is

$$\text{Unemploy} = 11.0 - 0.0284 \text{ Time} - 0.0366 \text{ costerm} + 0.506 \text{ sinterm}$$

Predictor	Coef	SE Coef	T	P
Constant	11.0407	0.1019	108.33	0.000
Time	-0.0284277	0.0009161	-31.03	0.000
costerm	-0.03665	0.07173	-0.51	0.610
sinterm	0.50556	0.07180	7.04	0.000

$$S = 0.702731 \quad R\text{-Sq} = 84.6\% \quad R\text{-Sq}(\text{adj}) = 84.4\%$$

- (b) How does the  $R^2$  value for this model compare to the  $R^2$  value found for the model in Question #2?

From the output above, we see that  $R^2 = 0.846$  in this model, compared with  $R^2 = 0.877$  in the model which uses indicator variables. It appears that the previous model is better, since the  $R^2$  is larger. However, the difference isn't very large.

This may mean we should include more sin and cos terms in the model in this question. However, a big difference between the 2 models is the number of terms. The one in this question has only 2 terms to account for seasonal variation, while the previous model has 11 terms. One issue we briefly mentioned in class is that the  $R^2$  value will not decrease (and will usually increase) if we include more variables in our model. So part (perhaps a lot) of what we are observing can be explained by this fact.

- (c) Test at  $\alpha = 0.05$  whether or not we can drop the seasonal terms from the model.  
 $H_o : \beta_2 = \beta_3 = 0$  (drop seasonal terms)

$H_a : H_o$  not true

This time, I will use the Seq SS values to find  $SSE_R - SSE_C$ . From the output below, we find that

$$SSE_R - SSE_C = 0.13 + 24.48 = 24.61, \quad MSE_C = 0.49$$

Then

$$F_{obs} = \frac{24.61/2}{0.49} = 25.11$$

where we are dropping  $k - g = 2$  terms from the model.

Test at  $\alpha = 0.05$ : Reject  $H_o$  since  $25.11 > F_{.05} = 3.07$ , using F-distribution with  $k - g = 2$  and 120 (since  $n - k - 1 = 188$  not in table) df.

Therefore we need the seasonal terms.

### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	511.81	170.60	345.47	0.000
Residual Error	188	92.84	0.49		
Total	191	604.65			

Source	DF	Seq SS
Time	1	487.20
costerm	1	0.13
sinterm	1	24.48

- (d) Plot the residuals (or standardized residuals) vs. time, and construct a QQ plot. Interpret these plots.

There is some evidence of positive correlation in the error terms, based on the plot of the residuals vs. time (observation order). It appears that we have groups of positive residuals together, then negative, then positive.

The QQ-plot appears linear. Therefore the assumption of normally distributed errors is reasonable.

4. The following are the number of reported cases of a new disease over the last 11 months:

```

Month (t):  1 2 3 4 5 6 7  8  9 10 11
-----
Cases (yt): 1 1 2 3 4 6 8 13 21 27 45

```

(a) Plot  $y_t$  vs. time. Does the use of a growth curve model for forecasting  $y_t$  seem appropriate? Explain.

A growth curve model does seem appropriate because the response (number of cases) is increasing rapidly with time.

(b) Using natural logs, define a transformed growth curve model that will be linear in its parameters.

An appropriate growth curve model would be  $y_t^* = \beta_0^* + \beta_1^*t + e_t$ , where  $y^* = \ln(y)$ .

(c) Plot the natural logs of the  $y_t$  values vs. time. Has the log transformation linearized the data?

The log transformation does appear to have yielded a linear relationship between this variable and time.

(d) Find the least squares regression line for the equation

$$y_t^* = \beta_0 + \beta_1 t + e_t$$

where  $y_t^* = \ln(y_t)$ .

From the output:  $\hat{y}_t^* = -0.543 + 0.390t$ .

```

The regression equation is
logcases = - 0.543 + 0.390 Time

```

Predictor	Coef	SE Coef	T	p
Constant	-0.54334	0.07344	-7.40	0.000
Time	0.38997	0.01083	36.01	0.000

5. In class, we defined the test statistic for the overall F-test in regression as  $F_{obs} = MSR/MSE$ . Given that we know  $R^2 = 1 - (SSE/SS_{yy})$ , show that we can write

$$F_{obs} = \frac{R^2/k}{(1 - R^2)/(n - k - 1)}$$

*Hint:* You may use the fact that  $SSR + SSE = SS_{yy}$ .

**Solution:** As the question points out, we defined  $F_{obs} = MSR/MSE$ . We also know that

$$R^2 = SSR/SS_{yy} = (SS_{yy} - SSE)/SS_{yy} = 1 - (SSE/SS_{yy}).$$

Then

$$\begin{aligned} F_{obs} &= \frac{SSR/k}{SSE/(n-k-1)} \\ &= \frac{(n-k-1)SSR}{kSSE} \\ &= \frac{(n-k-1)SSR}{k(SS_{yy} - SSR)} \quad \text{Since } SSR + SSE = SS_{yy} \\ &= \frac{(n-k-1)SSR/SS_{yy}}{k(SS_{yy} - SSR)/SS_{yy}} \quad \text{Divide top, bottom by } SS_{yy} \\ &= \frac{(n-k-1)SSR/SS_{yy}}{k[1 - (SSR/SS_{yy})]} \\ &= \frac{(n-k-1)R^2}{k(1-R^2)} \quad \text{By definition of } R^2 \\ &= \frac{R^2/k}{(1-R^2)/(n-k-1)} \end{aligned}$$

6. A cubic trend model was fit to the yearly closing stock price of an oil company over a 20 year period, assuming the errors were autocorrelated. The least squares regression line was found to be

$$\hat{y}_t = 160 - 100t + 1.2t^2 + 0.18t^3 + 0.3\hat{e}_{t-1}$$

where  $\hat{\sigma}^2 = 49$  and the closing value in year 20 was 70.

Find a 95% prediction interval for the price of the stock at the end of year 21.

PI at year 21:  $T = 20$ ,  $\tau = 1$ .  $y_{20} = 70$ .

$$e_T = e_{20} = y_{20} - (160 - 100(20) + 1.2(20)^2 + 0.18(20)^3) = 70 - 80 = -10$$

Then

$$\hat{y}_{T+\tau} = \hat{y}_{21} = 160 - 100(21) + 1.2(21)^2 + 0.18(21)^3 - 0.3(-10) = 253.18$$

95% PI:  $t_{\alpha/2} = t_{0.025} = 2.12$  using  $(n - k - 1) = (20 - 3 - 1) = 16$  df.  $k = 3$  since there are 3 explanatory variable terms in the model ( $t$ ,  $t^2$  and  $t^3$ ).

Then 95% PI for  $y_{21}$  is

$$\hat{y}_{21} \pm t_{0.025}\hat{\sigma} = 253.18 \pm 2.12(\sqrt{49}) = (238.34, 268.02)$$