Insurance 260
Time Series Component
Final Exam

The Wharton School
University of Pennsylvania
April 28, 2009

Name:

For multiple choice items, circle the letter that identifies one best answer. For other questions, fill in your answer in the space provided. Answers outside this space will not be graded.

1. **Assume the SRM holds. If the 95% confidence interval for the slope β_1 in an estimated simple regression is the interval [10,22], then**

   The true slope β_1 is not zero.

   a) The standard error of b_1 is approximately 3. (This question had a typo on the exam; all got credit.)
   b) The t-statistic for the estimated slope is approximately 2.
   c) The p-value of the estimated slope is more than 0.05.
   d) The F-statistic of the fitted regression model is less than 4.

2. **Extrapolation occurs in a multiple regression model when**

   a) We interpret b_0 but the observed Xs lie far above zero.
   b) We interpret a slope b_1 from an estimated model when the SDs of X_1 is larger than the SD of Y.
   c) The fitted model is used to predict new values of the response.
   d) The sample size used to estimate the fitted model is small relative to the number of Xs.
   e) Explanatory variables in the fitted are highly collinear.

3. **A narrow column of points in the center of a leverage plot indicates that**

   a) The fitted model has been affected by two highly leveraged outliers.
   b) The residuals of the fitted model are not normally distributed.
   c) The fitted model will not be able to predict accurately many periods beyond the observed data.
   d) This explanatory variable is collinear with other explanatory variables in the model.
   e) This explanatory variable must be transformed in order to improve the fit of the model to the data.

4. **The MRM assumes that the explanatory variables that appear in the model**

   a) Are uncorrelated with each other.
   b) Are linearly related to the response.
   c) Have normal distributions.
   d) Have equal variation.
   e) Represent random samples from the relevant population.

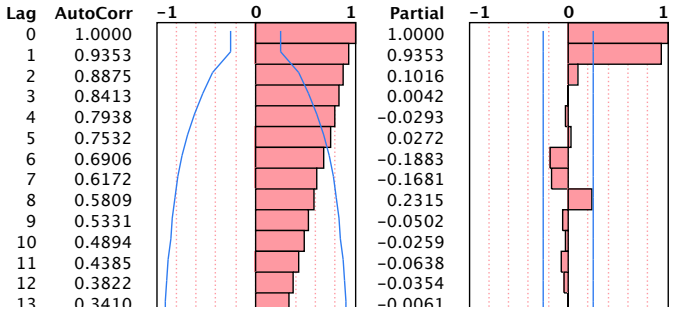
5. **The most useful way to check the assumption of equal error variance in a fitted multiple regression is to**

   a) Check the p-value of the Durbin-Watson statistic.
   b) Plot the residuals from the regression in time order (ie, consider a sequence plot of the residuals).
   c) Plot the residuals versus each explanatory variable.
   d) Inspect the leverage plots for each explanatory variable.
   e) Inspect the plot of the residuals on the fitted values from the model.

6. **The best predictor of Y_{n+3} when modeling a time series $Y_1,\ldots,Y_n$ using exponential smoothing is**

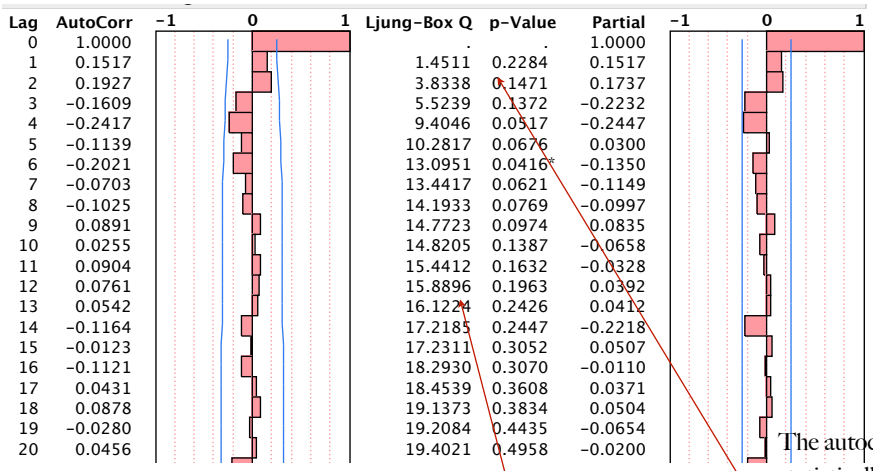
   a) Equal to the best predictor of Y_{n+1}.
   b) The last observation, Y_n.
   c) The last observation plus 3 times the difference between the last two, $Y_n + 3\,(Y_n - Y_{n-1})$.
   d) The last observation plus 2 times the difference between the last two, $Y_n + 2\,(Y_n - Y_{n-1})$.
   e) A weighted average of prior observations of the form $wY_{n+2} + w^2 Y_{n+1} + w^3 Y_n + \ldots$

7. **The figure at right shows the estimated autocorrelation and partial autocorrelations of a time series of n=60 observations. Based on these plots, we should**
   a) Transform the data by taking logs.
   b) Difference the series to obtain stationary data.
   c) Fit an AR(1) model to the time series. (1/2 credit)
   d) Fit an MA(1) model to the time series.
   e) Fit a linear time trend to the time series. (1/2 credit)



Lag	AutoCorr	Partial
0	1.0000	1.0000
1	0.9353	0.9353
2	0.8875	0.1016
3	0.8413	0.0042
4	0.7938	-0.0293
5	0.7532	0.0272
6	0.6906	-0.1883
7	0.6172	-0.1681
8	0.5809	0.2315
9	0.5331	-0.0502
10	0.4894	-0.0259
11	0.4385	-0.0638
12	0.3822	-0.0354
13	0.3410	-0.0061

(Q 8-10) The following output shows statistics that summarize properties of residuals e_t , t = 1,...,n, from a least squares regression fit to 5 years of monthly data.



Lag	AutoCorr	Ljung-Box Q	p-Value	Partial
0	1.0000	.	.	1.0000
1	0.1517	1.4511	0.2284	0.1517
2	0.1927	3.8338	0.1471	0.1737
3	-0.1609	5.5239	0.1372	-0.2232
4	-0.2417	9.4046	0.0517	-0.2447
5	-0.1139	10.2817	0.0676	0.0300
6	-0.2021	13.0951	0.0416*	-0.1350
7	-0.0703	13.4417	0.0621	-0.1149
8	-0.1025	14.1933	0.0769	-0.0997
9	0.0891	14.7723	0.0974	0.0835
10	0.0255	14.8205	0.1387	-0.0658
11	0.0904	15.4412	0.1632	-0.0328
12	0.0761	15.8896	0.1963	0.0392
13	0.0542	16.1224	0.2426	0.0412
14	-0.1164	17.2185	0.2447	-0.2218
15	-0.0123	17.2311	0.3052	0.0507
16	-0.1121	18.2930	0.3070	-0.0110
17	0.0431	18.4539	0.3608	0.0371
18	0.0878	19.1373	0.3834	0.0504
19	-0.0280	19.2084	0.4435	-0.0654
20	0.0456	19.4021	0.4958	-0.0200

The autocorrelation at lag 1 is small, and not statistically significant.

8. **These results imply that**
   a) The Durbin-Watson statistic would not find statistically significant autocorrelation in the residuals.
   b) The fitted model omits dummy variables needed to capture seasonal patterns.
   c) The fitted model used dummy variables to capture seasonal patterns.
   d) The R^2 statistic of the fitted model is close to 1, implying a statistically significant model.
   e) The Durbin-Watson statistic would find statistically significant autocorrelation in the residuals.

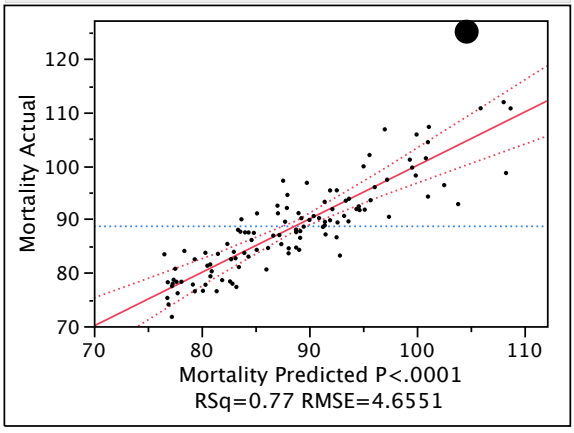
9. **If we are concerned about autocorrelation with lags up to a year, we should**
   a) Add a lag of the residuals (e_{t-1}) to the fitted model to capture the evident dependence.
   b) Identify an ARMA model to capture the evident dependence.
   c) Add seasonal dummy variables to the model to capture evident seasonal patterns.
   d) Difference the data prior to fitting the regression model to obtain a stationary time series.
   e) Recognize that there is not statistically significant residual autocorrelation.

Q statistic is not significant at lag 12.

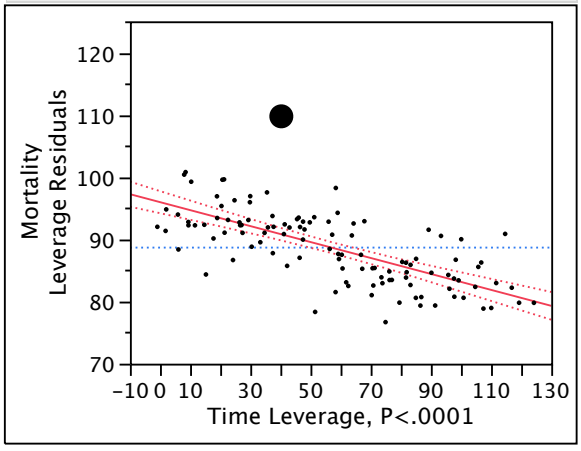
10. **In a least squares regression of the residuals e_t on lags e_{t-1} and e_{t-2}, the coefficient of e_{t-2} is**
    a) About 0.1517
    b) About 0.1737 (This is the second partial autocorrelation.)
    c) About 0.1927
    d) About 0.0
    e) Not revealed by the information given.

(Q11-18) The following JMP output summarizes a regression model for the mortality rate in Los Angeles County during the 1980s. The response is the average daily cardiovascular mortality rate. The explanatory variables are temperature (average degrees Fahrenheit), particulate pollution (micrograms per cubic meter of air), month of the year, and time (a monthly index from 1 to n = 113).

Actual by Predicted Plot



Leverage Plot



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	14	7070.2601	505.019	23.3046
Error	98	2123.6954	21.670	Prob > F
C. Total	112	9193.9554		<.0001*

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Particulates	1	1	369.1624	17.0354	<.0001*
Temperature	1	1	450.6912	20.7976	<.0001*
Month	11	11	496.2203	2.0817	0.0286*
Time	1	1	1910.2706	88.1513	<.0001*

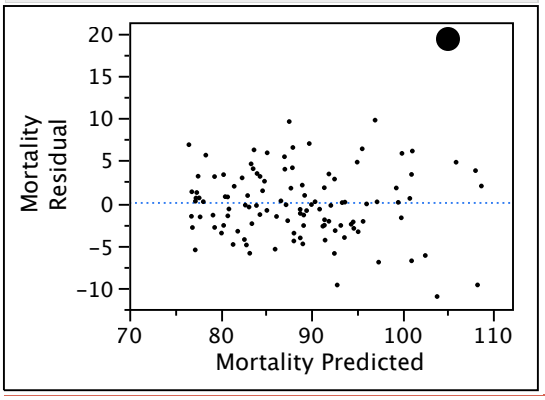
Durbin-Watson

Durbin-Watson	Number of Obs.	AutoCorrelation	Prob<DW
1.4308573	113	0.2838	0.0011*

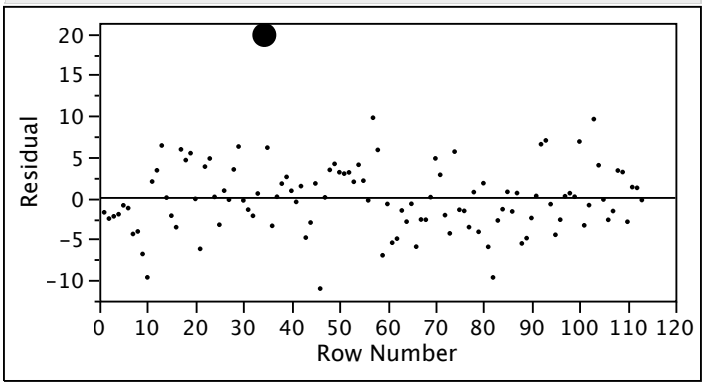
Indicator Function Parameterization

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|
Intercept	119.170	7.118	98.00	16.74	<.0001*
Particulates	0.211	0.051	98.00	4.13	<.0001*
Temperature	-0.432	0.095	98.00	-4.56	<.0001*
Month[Jan]	-1.573	2.182	98.00	-0.72	0.4728
Month[Feb]	-2.606	2.259	98.00	-1.15	0.2513
Month[Mar]	-1.863	2.395	98.00	-0.78	0.4387
Month[Apr]	-3.189	2.449	98.00	-1.30	0.1959
Month[May]	-2.875	2.477	98.00	-1.16	0.2485
Month[June]	-4.151	2.416	98.00	-1.72	0.0889
Month[July]	-2.914	2.516	98.00	-1.16	0.2496
Month[Aug]	-2.796	2.427	98.00	-1.15	0.2520
Month[Sep]	0.859	2.491	98.00	0.34	0.7308
Month[Oct]	4.245	2.324	98.00	1.83	0.0707
Month[Nov]	4.036	2.246	98.00	1.80	0.0754
Time	-0.128	0.014	98.00	-9.39	<.0001*

Residual by Predicted Plot



Residual by Row Plot



11. Assuming the conditions of the MRM, we can see that this model explains statistically significantly more variation in mortality rates than a simple model that fits a constant level alone because:

Use the overall F-test from the Anova Table. The F-statistic is 23.3 with p-value less than 0.0001. This shows that the R^2 statistic is larger than we'd expect by chance alone in a regression with this many observations and explanatory variables; we can reject the null hypothesis that **all** of the slopes in the regression are zero.

12. Interpret the estimated coefficient of the explanatory variable time (-0.128).

Adjusting for changes in particulate levels and temperature over these years, the mortality rate is falling at a rate of about 0.128 per month on average over the period of this study. This decrease beyond the effects of changes in temperature and particulates could be explained by changing patterns of health care over these 10 years.

(The effect is highly statistically significant. The key feature that you had to express in your answer is the notion that this is not the marginal trend in mortality.)

13. An outlier occurs in month 34 of these data, highlighted by the large circle in the figures. Is this outlier unusually large? Explain briefly.

Yes, the outlier is unusually large. The easiest way to see this is to use the RMSE to count how many SDs of the residuals separate this point from the fitted line. (Notice that the RMSE is shown in the plot of the data on the actual values; you can also compute it from the information in the Anova Table.) The residual at this point is about 20 (clearest in the plot of residuals on fitted values) and the RMSE is 4.6551; hence the residual is about 20/4.6551 ≈ 4.3 SD away from the regression. If the data are normally distributed around the line (and that seems reasonable from the plots), then this point is quite far from the fitted line.

14. Which of the following is a correct interpretation of estimated coefficient of Month[June]=-4.151?

a) Mortality rates are the same in December and in June.
b) Mortality rates are lowest in June when adjusted for other explanatory variables.
c) Mortality rates are lower in June than typical mortality rates by about 4.151.
d) Mortality rates are lower in June than December mortality rates by about 4.151
e) Mortality rates in June are more variable than in other months when adjusted for explanatory variables.

None of the other answers correctly mentions the control for the other variables. "d" for instance is a marginal comparison rather than adjusting for differences in weather.

15. If the Month were removed from this model, then (assuming the conditions of the MRM) the R^2 statistic would

a) Increase by a statistically significant amount ($\alpha = 0.05$)
b) Increase, but not by a statistically significant amount ($\alpha = 0.05$).
c) Remain the same.
d) Decrease, but not by a statistically significant amount ($\alpha = 0.05$).
e) Decrease by a statistically significant amount ($\alpha = 0.05$). (use the partial F in the effect test table; p=0.0286)

16. Do these results suggest that the fitted multiple regression meets the conditions specified by the MRM? If so, explain why the model is okay. If not, explain which condition(s) is violated.

The Durbin-Watson statistic is significant ($p = 0.0011 < 0.05$). This is the most important flaw in the fitted model. The one outlier is noticeable, but exerts relatively small influence on the fitted model. (For example, the outlier is not highly leveraged in the leverage plot for time.)

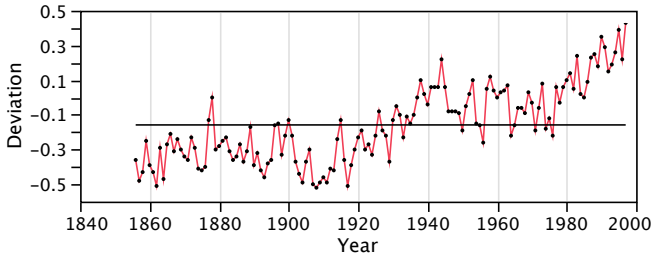
17. Assume that the fitted model satisfies the assumptions of the MRM. If the fitted model is used to predict mortality rates in LA County during the months of the next year, then how accurate do you anticipate those predictions to be?

Use the RMSE to approximate the size of prediction intervals. The RMSE indicates that predictions are accurate to within about $\pm 2(4.6551) \approx 9.3$.

A useful secondary point to make concerns the relative size of these errors. Seeing that mortality rates average about 90 (see the plot of Y on $\hat{Y}$), that's a range of about ±10% of a typical rate.

Do not use the remainder of this page.

The remaining questions consider a time series model for annual global temperature. The data for the time series in this analysis begin in 1856 and run through 1997 (n = 142). The measurements give the deviation from typical temperature in degrees Celsius. (Zero would be considered consistent with the long-run average.)



(Q18-19) These results summarize the fit of a simple exponential smooth to the time series.

Model Summary

DF	140.0000
Sum of Squared Errors	1.7726
Variance Estimate	0.0127
Standard Deviation	0.1125
Akaike's 'A' Information Criterion	-214.4648
Schwarz's Bayesian Criterion	-211.5160
RSquare	0.7328

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Level Smoothing Weight	0.39680005	0.0900926	4.40	<.0001*

	Actual Deviation	Year	Predicted Deviation
133	0.25	1988	0.13245007
134	0.18	1989	0.17909389
135	0.35	1990	0.17945343
136	0.29	1991	0.24712632
137	0.15	1992	0.2641386
138	0.19	1993	0.2188484
139	0.26	1994	0.20740135
140	0.39	1995	0.2282725
141	0.22	1996	0.29244598
142	0.43	1997	0.26369941

18. Use the estimated exponential smooth to predict temperature for the next 3 years (1998-2000). Show your work.

The predicted value from the exponential smooth is the same for all 3 years, so all we need is the value for next year. The expression for the smooth is

$$\text{smooth}_t = \text{smooth}_{t-1} + \alpha\,(y_t - \text{smooth}_{t-1})$$

Hence, for the next point, the next value of the smooth (the prediction for the next observation) is

$$\text{smooth}_n = \text{smooth}_{n-1} + \alpha\,(y_n - \text{smooth}_{n-1}) = 0.2637 + (0.3968)*(0.43-0.2637)$$

$$= \mathbf{0.3297}$$

19. Find 95% prediction intervals for the predictions of temperature in 1998-2000. Show your work.

The sd of the prediction errors is

1 period out 0.1125

2 periods out $0.1125\ \text{sqrt}(1+\alpha^2) = 0.1125 * \text{sqrt}(1+ 0.3968^2) \approx 0.121$

3 periods out $0.1125\ \text{sqrt}(1+ 2\ \alpha^2) = 0.1125 * \text{sqrt}(1+ 2*0.3968^2) \approx 0.129$

Hence the approximate 95% intervals are

$0.3297 \pm 2 * 0.1125$

$0.3297 \pm 2 * 0.121$

$0.3297 \pm 2 * 0.129$

(Q20 - Q22) The following results summarize an ARIMA(1,1,1) model fit to the same global temperature data.

Model Summary	
DF	138
Sum of Squared Errors	1.63017187
Variance Estimate	0.01181284
Standard Deviation	0.10868689
Akaike's 'A' Information Criterion	-222.12472
Schwarz's Bayesian Criterion	-213.27844
RSquare	0.75391022

	Year	Temperature Deviation	Differences	Predicted Differences
136	1991	0.29	-0.06	-0.0909
137	1992	0.15	-0.14	-0.0433
138	1993	0.19	0.04	0.0346
139	1994	0.26	0.07	0.0126
140	1995	0.39	0.13	-0.0199
141	1996	0.22	-0.17	-0.0756
142	1997	0.43	0.21	0.0222

Parameter Estimates

Term	Lag	Estimate	Std Error	t Ratio	Prob>|t|	Constant Estimate
AR1	1	0.348	0.111	3.13	0.0021*	0.00322166
MA1	1	0.828	0.065	12.81	<.0001*	
Intercept	0	0.005	0.002	1.99	0.0481*	

20. Does this ARIMA model offer a better model for the temperature data? Offer 2 reasons to justify your preference. (There are many reasons, some more important and valid than others.)

Good reasons are:

(1) Exponential smoothing is an IMA(1,1) model. This model adds an AR(1) term that is statistically significant and hence a useful addition to the prior model (as if adding another variable to a regression).

(2) Both AIC and SBC, the two model selection criteria, prefer this model as well to the prior model. It's got smaller values for both. (AIC and SBC resemble residual SS; smaller values are better.)

Other reasons include: (at some loss of points)

(1) This model has a higher R^2 or a smaller residual variance.

(2) ARIMA models are more flexible than exponential smoothing.

21. Find the 95% prediction interval for the temperature in 1998 based on this ARIMA model.

All you need is the predicted value, since the software gives you the SD to use (0.0118). To find the predicted temperature, note that the model predicts changes. So, the predicted temperature is the last observed temperature (0.43) plus the predicted change.

Because this model describes the differences in temperature, the predicted temperature in 1998 is the sum of the last value y_n plus the predicted difference:

$$\hat{y}_{n+1} = y_n + 0.00322 + 0.348(0.21) - 0.828(0.21-0.0222) \approx 0.43 - 0.0792 \approx 0.3508$$

The prediction interval is then

$$0.3508 \pm 2*(0.1087)$$

22. Qualitatively (i.e., don't find specific numbers), what is the most important difference between the predictions of global temperature produced by this ARIMA model and those of the prior exponential smooth (Q19-20)?

The exponential smooth predicts a constant future temperature. The ARIMA model predicts eventually increasing temperatures because of the positive constant term. The figure to the right shows JMP's predictions from the ARIMA model.