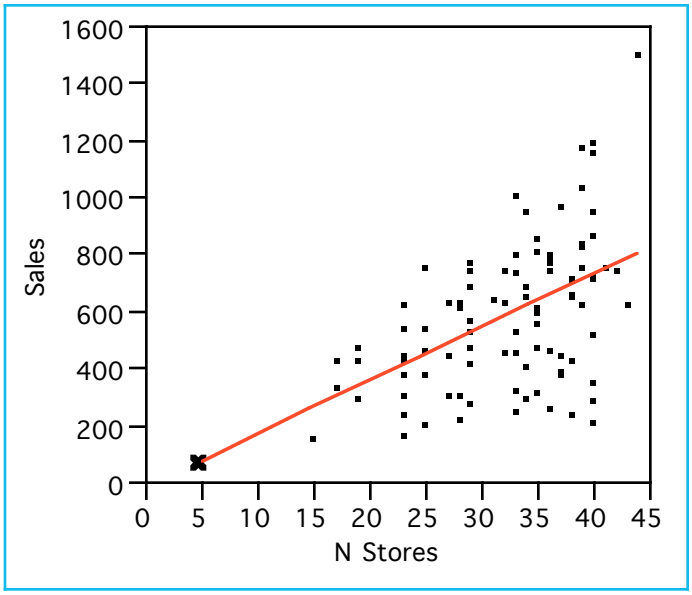
Statistics 621 Waiver Examination
August 31, 1998

This is an open-book, open-notes exam. You have two hours to complete the exam. The computer output associated with one or more items should be considered an essential part of the question. The questions are equally weighted. The exam solution sheets are scored electronically, so keep these issues in mind:

- *Be sure to use a #2 pencil.* If you do not have a pencil, we will supply you with one. *Do not* mark your solutions with a pen.
- Before starting, be sure to *fill in your name and student id* number.
- Choose the *one best possible answer*.
- The exam is scored by *counting the number of correct answers*.

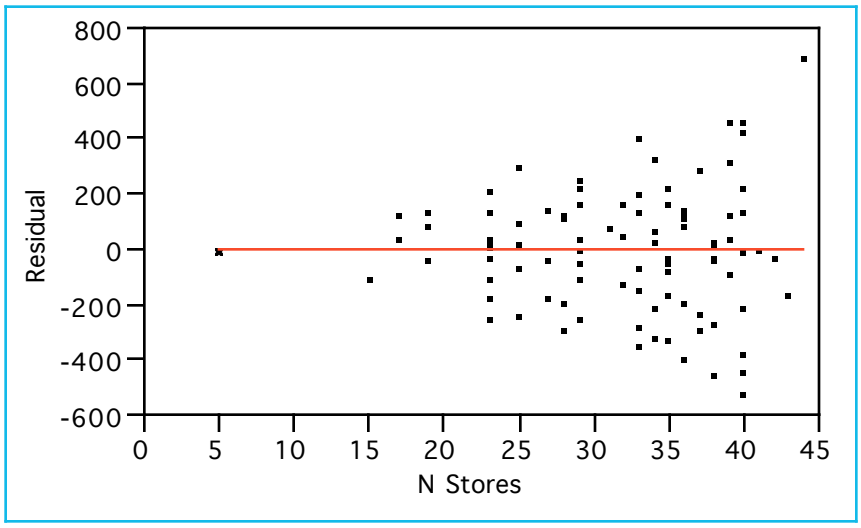
(Questions 1-6) A chain which manages software outlets has collected data for its stores over the past month. The data shown here is aggregated into the chain's 100 sales districts. Sales figures shown are multiples of $10,000, so that for example a sales value of 10 implies monthly sales of $100,000. Note that these values are sales, not net profits, so the values are not negative. The districts do not have the same number of outlets; the number of outlets in each district ranges from 5 to 43. Plots and analysis of the data are followed by several questions. The highlighted observation (a district with 5 stores marked with an x) is included throughout.

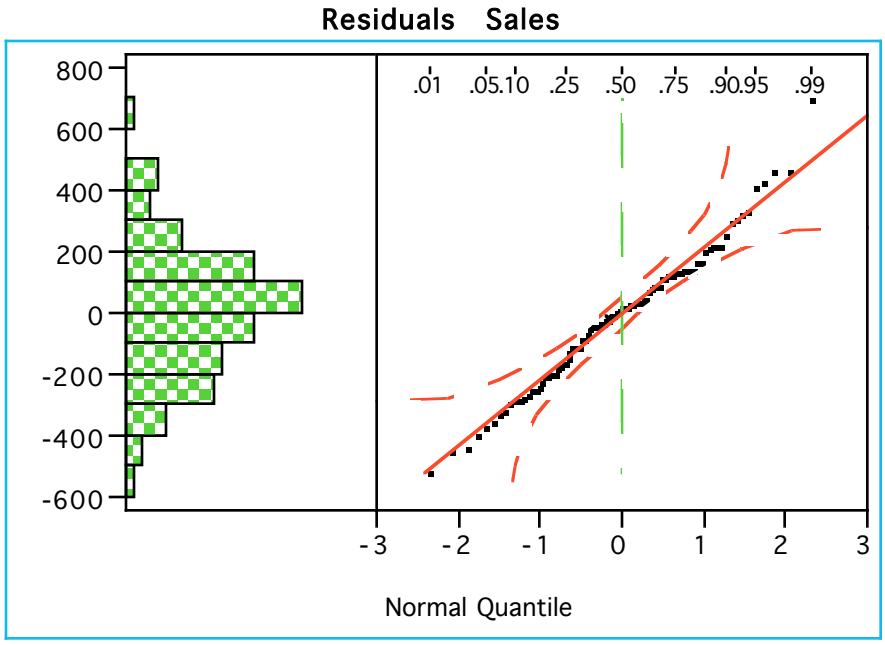


RSquare	0.287
Root Mean Square Error	217.950
Mean of Response	580.695

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-15.13	97.38	-0.16	0.8768
N Stores	18.79	2.99	6.28	<.0001





(1) Does the number of stores in a district explain a significant amount of variation in district sales, according to this fitted model?
(a) Yes, the R^2 of 29% is large.
(b) No, the R^2 of 29% is small.
(c) Yes, the root mean squared error is small.
(d) No, the root mean squared error is large.
(e) Yes, the t-statistic for the slope is significant.

(2) From the model, what volume of sales does the typical store generate in a month?
(a) $150,000
(b) $190,000
(c) $970,000
(d) $1,879,000
(e) $5,810,000

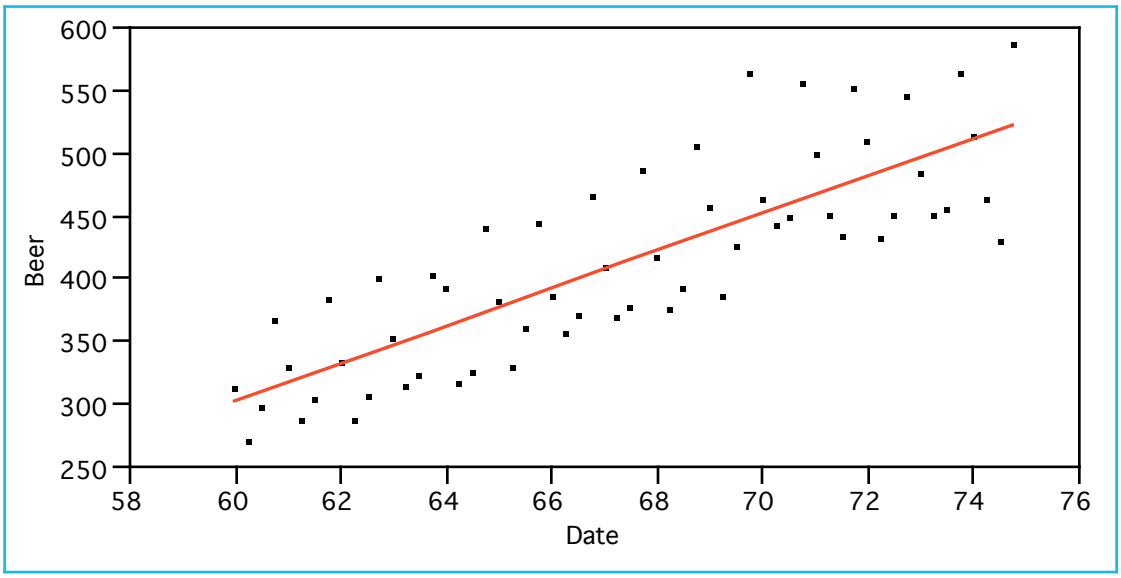
(3) The highlighted observation in the plots (a district with 5 outlets) is
(a) leveraged and very influential.
(b) leveraged but not very influential.
(c) not leveraged but influential.
(d) not leveraged and not influential.
(e) a coding error.

(4) Based on this fitted model, which of the following is the 95% prediction interval for the sales of a district with 10 outlets?
(a) [$ 120,000 – $ 240,000]
(b) [$ 167,000 – $ 191,000]
(c) [$1,200,000 – $2,400,000]
(d) [$1,000,000 – $3,000,000]
(e) [0 – $6,300,000]

(5) Does the fitted model evidently violate an assumption that would threaten the validity of the prediction interval constructed in the previous question (#4).
(a) No.
(b) Yes, the data are dependent and thus the standard error expressions are wrong.
(c) Yes, the data lack constant variance and the interval is too wide.
(d) Yes, the data are not normally distributed so we cannot apply the empirical rule.
(e) Yes, the fit of the model is too weak to use for prediction.

(6) The intercept of the fitted model is negative, suggesting impossible levels of sales in a district with no outlets. Does this imply that the model is flawed?
(a) Yes, it indicates the presence of a nonlinear relationship.
(b) Yes, it indicates that the model fits too poorly to be used for prediction.
(c) Yes, it confirms the presence of autocorrelation.
(d) No, the confidence interval for the intercept includes zero.
(e) No, the intercept is not needed in this model and cannot indicate a flaw.

(Questions 7-10) The following plots and statistics are based on the quarterly Australian beer production in megaliters from the first quarter of 1960 through the end of 1974. The initial plot and analysis are followed by summaries of the residuals from the fitted model.



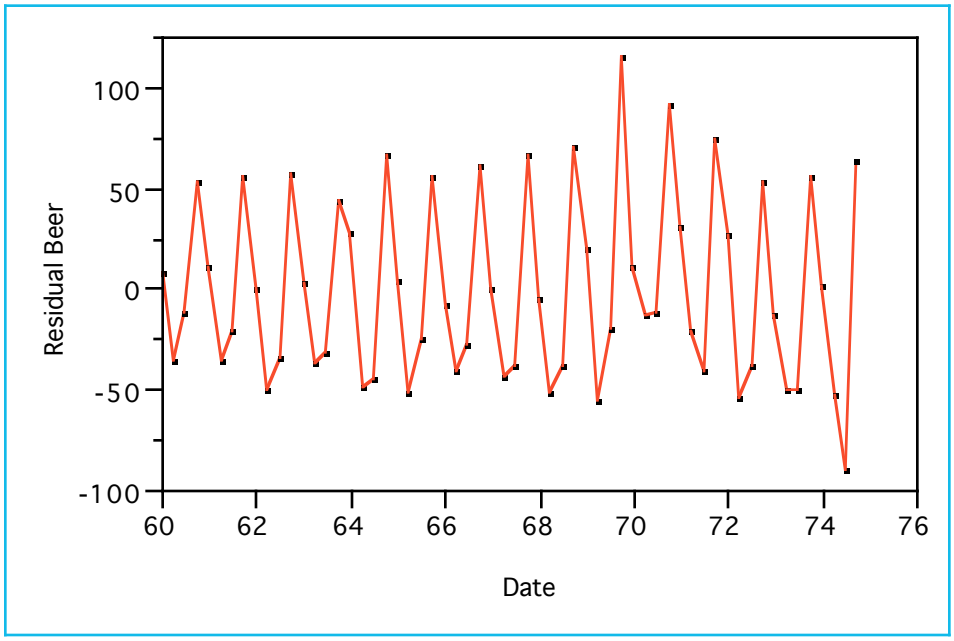
RSquare	0.667
Root Mean Square Error	46.409
Mean of Response	413.682
Observations (or Sum Wgts)	60

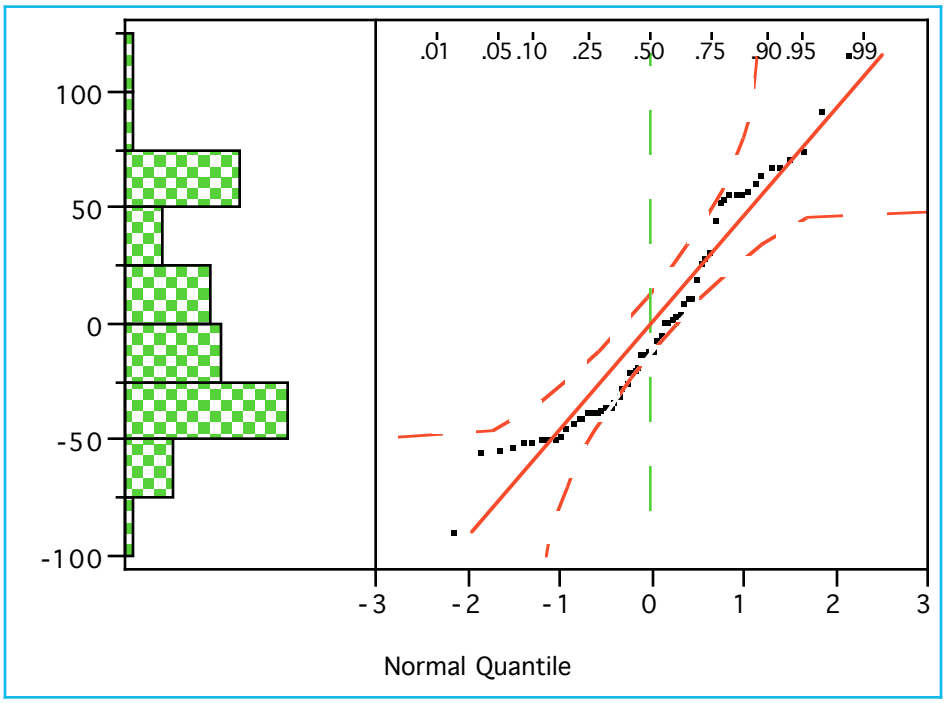
Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-590.7	93.43	-6.32	<.0001
Date	14.9	1.38	10.77	<.0001

Durbin-Watson	Number of Obs.	AutoCorrelation
2.0626944	60	-0.0485

Residuals





(7) Based on the fitted regression model, which of the following is a 95% prediction interval for the first quarter of 1975 (for which the value of Date is 75)?
(a) [524, 530]
(b) [503, 551]
(c) [477, 577]
(d) [431, 623]
(e) Cannot be computed from given information which does not include JMP limits.

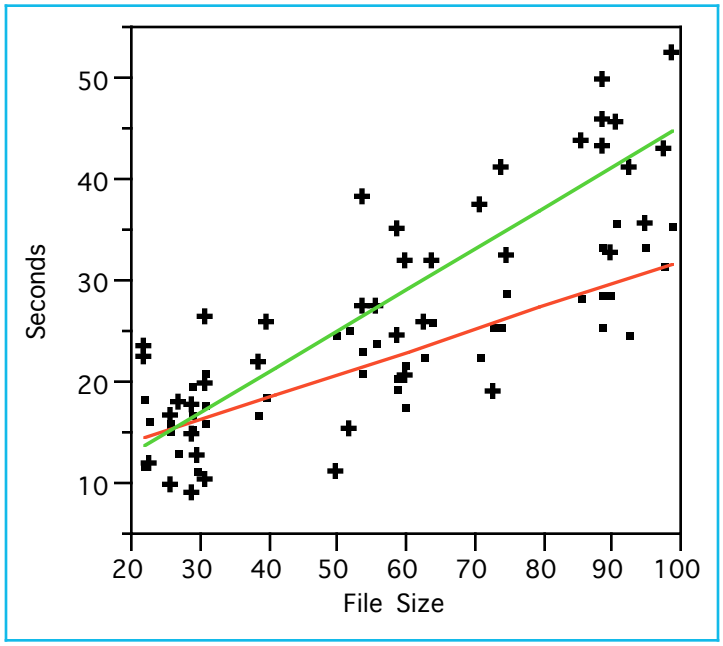
(8) The fitted model implies that beer production in Australia during this period
(a) increased by about 414 megaliters.
(b) rose, on average by about 15 megaliters per year.
(c) rose, on average by about 30 megaliters per year.
(d) rose, on average by about 46 megaliters per year.
(e) did not change significantly.

(9) What assumption of regression modeling appears to be most violated in this data?
(a) Independence
(b) Constant variation
(c) Normal distribution for the error terms
(d) Collinearity
(e) The model does not violate any of the important modeling assumptions.

(10) What would be a natural next step in the analysis of this data?
(a) Add a lagged variable to the regression equation.
(b) Compensate for the unequal variation seen in the data.
(c) Add a categorical predictor that indicates the quarter.
(d) Fit a high-order polynomial to capture residual structure.
(e) Work with the differenced data to remove autocorrelation.

(Questions 11-16) Before adopting one of two competing Internet service providers (MS and NP), a company ran several tests to investigate which provider was better. The specific test was to transfer files using both services. Forty files were transferred, once via MS and once via NP. The transfers were done under similar conditions (time of day, day of week, etc). They recorded the time in seconds to transfer each file. The variable FileSize holds the size of the file, measured in 1000's of characters. A regression analysis was used to summarize the results. Questions follow the output.

Seconds By File Size (MS = •, NP = +)



RSquare	0.75
Root Mean Square Error	5.138
Mean of Response	25.12
Observations (or Sum Wgts)	80

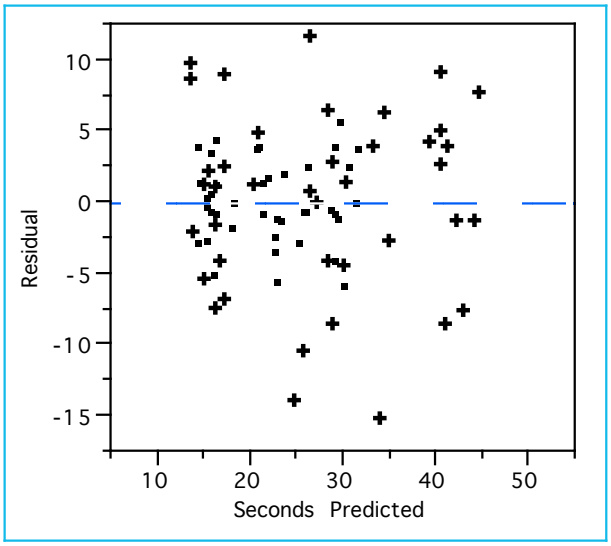
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	7.274	1.411	5.15	<.0001
File Size	0.313	0.022	13.84	<.0001
Provider[MS-NP]	2.381	1.411	1.69	0.0956
Provider[MS-NP]*File Size	-0.0904	0.023	-3.99	0.0001

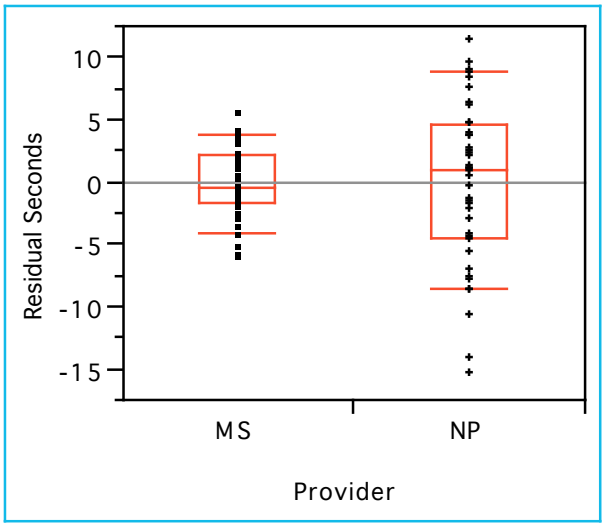
Effect Test

Source	Nparm	DF	Sum of Squares	F Ratio	Prob>F
File Size	1	1	5057.6437	191.5718	<.0001
Provider	1	1	75.1835	2.8478	0.0956
Provider*File Size	1	1	421.2087	15.9544	0.0001

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	6091.5769	2030.53	76.9116
Error	76	2006.4586	26.40	Prob>F
C Total	79	8098.0355		<.0001





(11) On average, to transfer a file of 50,000 characters using the MS provider would take about
(a) 7 seconds
(b) 21 seconds
(c) 23 seconds
(d) 25 seconds
(d) 50 seconds

(12) The company is connected to each Internet service provider using a telephone modem that is theoretically capable of transferring 7,000 characters per second. The better rate of transfer, on average, observed here is
(a) 2,381 characters per second
(b) 2,500 characters per second
(c) 3,200 characters per second
(d) 4,500 characters per second
(e) 7,274 characters per second

(13) Can you conclude from this analysis that there is a difference in the transfer rates of the two providers?
(a) Yes, because the plot of the two fits makes the difference clear.
(b) Yes, because the interaction term is significant.
(c) No, because the plot of the two fits shows too much overlap of the data.
(d) No, because the interaction term is not significant.
(e) This question cannot be answered without doing Hsu's comparison.

(14) Which of the following criticisms of this model is *valid*?
(a) The sample sizes are too small to draw inferences about which is the better provider.
(b) Since the data were paired (each file sent twice), the analysis violates the assumption of independence and cannot be used.
(c) The fitted model does not accommodate the heteroscedasticity seen in the plot of seconds on file size, with the diverging fitted lines.
(d) The fitted model does not accommodate the heteroscedasticity seen in the residuals.
(e) Since the model contains a coefficient which is not significant, it should be run without this and analyzed further.

(15) A consultant recommended a different type of analysis. Rather than use regression, use a paired t-test (using the natural pairing of this data). Would the use of such a test omit an important feature of the regression analysis?
(a) No, since regression with indicators is equivalent to a t-test.
(b) No, the two samples are large enough to find a significant difference.
(c) Yes, the pairing violates the independence of the t-test.
(d) Yes, since the mean difference in the t-test would only be 1.4 SE's from zero.
(e) Yes, since such a test would ignore the interaction in this data.

(16) Suppose that instead of having been based on 80 observations, this regression analysis had been conducted using 120. What would be the effect of this additional data upon the fitted model, assuming that the larger sample is obtained from the same population?
(a) R^2 would be 20% larger than in this fit, with a smaller RMSE.
(b) The t-statistics of the fitted coefficients would be 2/3 of the size in this fit.
(c) The standard errors of the coefficient estimates would be about 80% of those here.
(d) The F-ratio would decrease by 2/3 since more data would increase the DF's.
(e) The p-values for the fitted coefficients would increase by 20%.

(Questions 17-23) A pharmaceutical firm has developed yet another drug to treat patterned hair loss in men. To help determine how to package the drug for over-the-counter sales, a small study with 60 subjects has been run. The 60 men were assigned randomly to one of six groups. Three groups were given the medication in capsules, and the three other groups were given tablets. For those receiving capsules, the doses were 10, 20 and 30 mgs in the three groups. Similarly, the tablets were given in these same three doses. The drug is known to be safe in such doses. The recorded data measures new hair growth.

RSquare	0.550
Root Mean Square Error	14.230

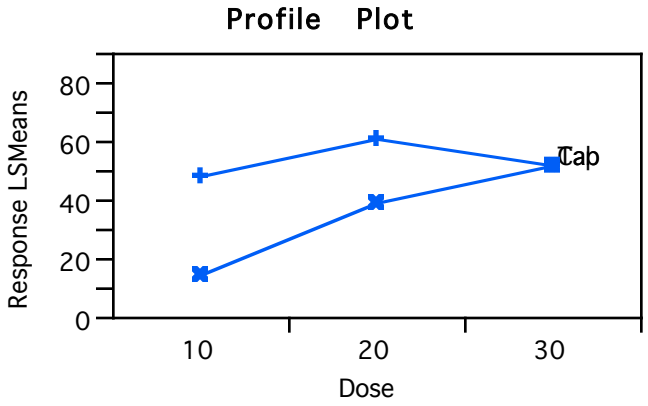
Parameter Estimates

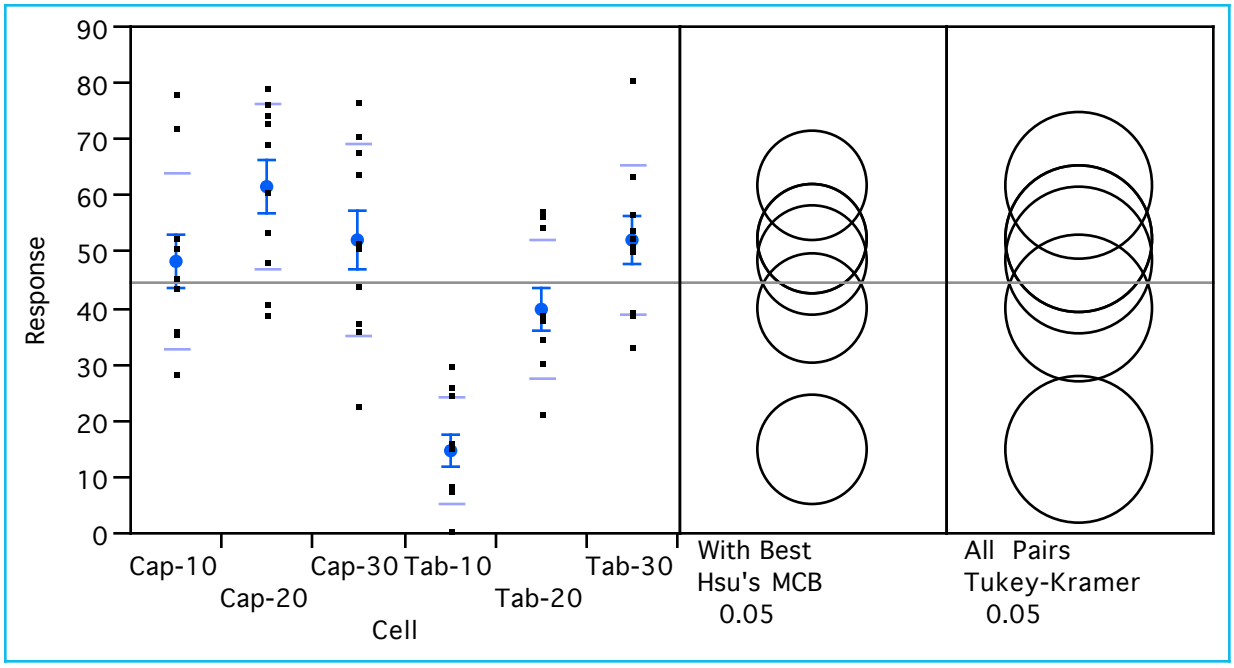
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	31.79	3.18	9.99	<.0001
Dose[20]	18.92	4.50	4.20	<.0001
Dose[30]	1.60	4.50	0.35	0.7244
Type[Cap-Tab]	16.98	3.18	5.34	<.0001
Type[Cap-Tab]*Dose[20]	-6.07	4.50	-1.35	0.1830
Type[Cap-Tab]*Dose[30]	-10.84	4.50	-2.41	0.0194

Source	Nparm	DF	Sum of Squares	F Ratio	Prob>F
Dose	2	2	5209.17	12.86	<.0001
Type	1	1	5769.80	28.50	<.0001
Type*Dose	2	2	2937.17	7.25	0.0016

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	13361.8	2672.36	13.1981
Error	54	10934.0	202.48	Prob>F
C Total	59	24295.8		<.0001





Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean
Cap-10	10	48.78	15.80	5.00
Cap-20	10	61.63	15.12	4.78
Cap-30	10	52.38	17.48	5.53
Tab-10	10	14.81	9.69	3.06
Tab-20	10	39.80	12.33	3.90
Tab-30	10	52.24	13.60	4.30

Comparisons using Hsu's MCB

Mean[i]-Mean[j]-LSD	Cap-20	Cap-30	Tab-30	Cap-10	Tab-20	Tab-10
Cap-20	-14.6	-5.3	-5.2	-1.7	7.3	32.3
Cap-30	-23.8	-14.6	-14.4	-11.0	-2.0	23.0
Tab-30	-24.0	-14.7	-14.6	-11.1	-2.1	22.9
Cap-10	-27.4	-18.2	-18.0	-14.6	-5.6	19.4
Tab-20	-36.4	-27.1	-27.0	-23.5	-14.6	10.4
Tab-10	-61.4	-52.1	-52.0	-48.5	-39.6	-14.6

If a column has any positive values, the mean is significantly less than the max.

Comparisons using Tukey-Kramer HSD

Abs(Dif)-LSD	Cap-20	Cap-30	Tab-30	Cap-10	Tab-20	Tab-10
Cap-20	-18.8	-9.6	-9.4	-6.0	3.0	28.0
Cap-30	-9.6	-18.8	-18.7	-15.2	-6.2	18.8
Tab-30	-9.4	-18.7	-18.8	-15.3	-6.4	18.6
Cap-10	-6.0	-15.2	-15.3	-18.8	-9.8	15.2
Tab-20	3.0	-6.2	-6.4	-9.8	-18.8	6.2
Tab-10	28.0	18.8	18.6	15.2	6.2	-18.8

Positive values show pairs of means that are significantly different.

(17) Assuming normality and the typical assumptions required by analysis of variance and regression, the estimated the probability of a patient who is taking the 10mg tablet growing new hair (i.e., having a positive value) is best estimated to be
(a) 0.94
(b) 0.84
(c) essentially zero.
(d) less than 1/2.
(e) cannot be computed without further output.

(18) Because of the size of the interaction in the fitted model, the company can conclude that
(a) the study must be redone with a larger sample size to eliminate this problem.
(b) since the interaction is not significant, the main effects can be interpreted marginally.
(c) since the interaction is significant, the main effects *cannot* be interpreted marginally**.**
(d) since the interaction is significant, the main effects *can* be interpreted marginally.
(e) the resulting lack of normality means that the study is flawed and cannot be used.

(19) If the objective of the company is to isolate the best dose and type of formulation (capsule or tablet), the results of this study show that
(a) since Dose[20] has the largest coefficient, they should use this dose with either type.
(b) the capsule with 20mg is best, performing better than any other combination.
(c) the capsule with 20mg is best, but Cap-10, Cap-30 and Tab-30 are comparable.
(d) the capsule with 20mg is best, but Cap-30 is comparable.
(e) the proportion of explained variance is too small to make any useful selection.

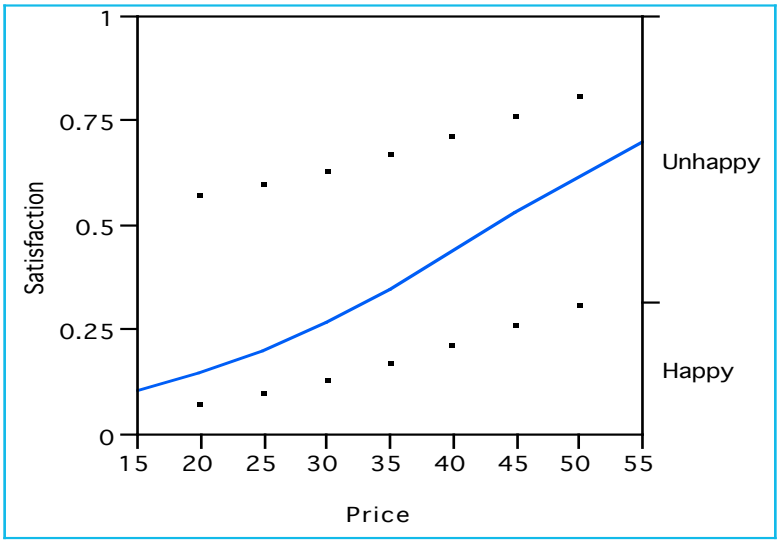
(20) In comparing the results for capsules vs. tablets, the results show that the company can conclude that with 95% confidence
(a) capsules are better than tablets for 10mg doses, but not otherwise.
(b) capsules are better than tablets for 10mg and 20mg doses, but not otherwise.
(c) capsules are better than tablets at every dose.
(d) tablets are better for most subjects, though not on average.
(e) the presence of significant interaction means that we cannot compare the averages.

(21) An MBA from a well-known California school analyzed this data using multiple regression with a dummy variable for type (capsule vs. tablet), dose as a continuous predictor, and an interaction term. How would her results differ from those shown here?
(a) Since anova is a special case of regression, her results would essentially be identical.
(b) The absence of Effects Tests would make it hard for her to interpret the output.
(c) The regression leverage plots would allow a careful look for possible outliers.
(d) Her regression would not obtain as good a fit.
(e) The interpretation of her regression would be difficult due to extreme collinearity.

(22) A logical next step for this analysis would be to
(a) Remove the non-significant coefficients and refit the model.
(b) Eliminate the interaction term since it introduces too much collinearity.
(c) Run separate one-way models for dose and type effects.
(d) Compute further multiple comparisons using the LSD procedure.
(e) Plot residuals by group, checking for normality and constant variance.

(23) An alternative type of analysis might be done in which each subject received each dose. That is, one would use only 10 men in the capsule group, and each would be treated with all three doses, presumably in a randomly assigned order. If the data had been collected in this manner, then
(a) there would be no change in the analysis, and we could proceed as in the above.
(b) the interaction term would be meaningless and have to be excluded from the analysis.
(c) the resulting dependence would make the data useless.
(d) the resulting dependence would require a different type of analysis.
(e) the two-way analysis could be kept, but the multiple comparisons would be invalid.

(Questions 24-27) A clothing manufacturer is interested in how the price paid for a shirt affects the sense of satisfaction of its customers. The company was able to survey 100 customers who has purchased this item. The only difference among the shirts sold were differences in the labeling on the package. Otherwise, the shirts were the same. The shirts were sold at various prices, ranging from \$20 to \$50 in \$5 increments. Most were sold at the lower prices, with few at the higher prices. The response recorded for each customer indicates whether the customer is "happy" with the item or "unhappy."



Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	5.473	1	10.945	0.0009
Full	57.214			
Reduced	62.687			

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-3.210	0.829	14.99	0.0001
Price	0.074	0.023	10.03	0.0015

(24) How does price affect the reaction of customers to the product?
(a) Panelists are *more* likely to be happy as the price increases, but the effect is *not* significant.
(b) Panelists are *more* likely to be happy as the price increases, and the effect is significant.
(c) Panelists are *less* likely to be happy as the price increases, but the effect is *not* significant.
(d) Panelists are *less* likely to be happy as the price increases, and the effect is significant.
(e) Cannot determine from the shown output.

(25) This model can be interpreted to mean that for each additional dollar paid for a shirt,
(a) the proportion of happy customers *increases* by about 7%.
(b) the proportion of happy customers *decreases* by about 7%.
(c) the odds of the customer being happy increases by about 7%.
(d) the number of happy customers increases by about 7%.
(e) there is no significant effect upon whether the customer will be happy.

(26) From this model, which of the following prices makes 50% of customers happy about their purchase?
(a) $34
(b) $37
(c) $40
(d) $43
(e) $46

(27) If the company were to raise the price from $30 to $50, what would be the impact upon the odds of purchase.
(a) The odds would increase by a multiplicative factor of 4.4.
(b) The odds would increase by an additive term of 4.4.
(c) The odds would not change significantly.
(d) The odds would decrease by a multiplicative factor of 3.2.
(e) The odds would decrease by an additive term of 3.2.

(Questions 28-37) The following two regression models (labeled Model1 and Model2) investigate several conjectures about the factors that affect wages in the US. The data are 534 randomly selected adult persons from the 1985 Current Population Survey conducted by the US Census Bureau. The variables in this analysis are:

LnWage	Natural log of the average hourly earnings, in US dollars
YrsEduc	Years of education
PotExpr	Years of potential experience (calculated as Age – YrsEduc – 6)
Union	Coded as "Yes" if the individual works on a union job, otherwise "No"
Married	Coded as "Yes" if the individual is married, otherwise "No"
South	Coded as "Yes" if the individual resides in the southern US, else "No"
Race	Denotes the race of the respondent, coded as AA – African American Hisp – Hispanic White – Caucasian

Model 1

RSquare	0.311
Mean of Response	2.059

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	0.655	0.120	5.44	0.000	0.00
Race[AA-White]	-0.036	0.047	-0.78	0.436	2.74
Race[Hisp-White]	-0.025	0.060	-0.41	0.680	2.76
Union[No-Yes]	-0.100	0.026	-3.91	0.000	1.07
Married[No-Yes]	-0.039	0.021	-1.87	0.062	1.09
Female[No-Yes]	0.116	0.020	5.95	0.000	1.04
South[No-Yes]	0.052	0.021	2.43	0.015	1.05
YrsEduc	0.092	0.008	11.43	0.000	1.20
PotExpr	0.011	0.002	6.09	0.000	1.27

Effect Test

Source	Nparm	DF	Sum of Squares	F Ratio	Prob>F
Race	2	2	0.665	1.706	0.1827
Union	1	1	2.976	15.268	0.0001
Married	1	1	0.682	3.499	0.0620
Female	1	1	6.890	35.351	<.0001
South	1	1	1.153	5.918	0.0153

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	8	46.12	5.765	29.5777
Error	525	102.32	0.195	Prob>F
C Total	533	148.44		<.0001

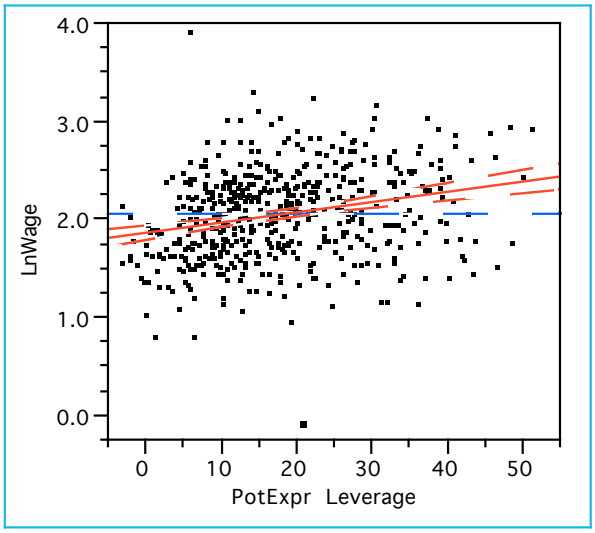
---

(28) Model 1 implies that, holding fixed the other factors in this model, average hourly earnings
(a) increase by \$0.09 for each additional year of education.
(b) increase by \$0.66 for each additional year of education.
(c) increase by 9% for each additional year of education.
(d) increase by 0.09% for each additional year of education.
(e) are not affected by the number of years of education.

(29) From Model 1, do wages in the surveyed population appear to be affected by race as identified by the indicator variable *Race*?
(a) Yes, the individual coefficients are statistically significant.
(b) No, neither t-statistic is significant.
(c) Yes, the partial F-test for *Race* is significant.
(d) No, the partial F-test for *Race* is not significant.
(e) Cannot be determined without further output from a model which omits this variable.

(30) A factory owner is considering moving her manufacturing plant from a northern US city to a southern area. Based on Model 1, what can she assume about the cost of replacing her northern employees with comparably skilled labor in the South?
(a) She can expect to pay significantly higher in wages in the South.
(b) She can expect to pay significantly lower in wages in the South.
(c) She should adjust for differing education patterns in the US before drawing such a conclusion.
(d) She can expect to pay about \$0.10/hour less in the South.
(e) Since the response is on a log scale, we cannot draw such comparisons.

(31) The following leverage plot for *PotExpr* in Model 1 shows that
(a) the slope for this variable is distorted by an influential negative outlier.
(b) the standard error of this coefficient is magnified extensively by collinearity.
(c) the relationship between this variable and log wages is nonlinear.
(d) the slope is significant and not distorted by deviations from assumptions.
(e) the slope estimate is biased toward large wages.



Model 2

RSquare 0.345

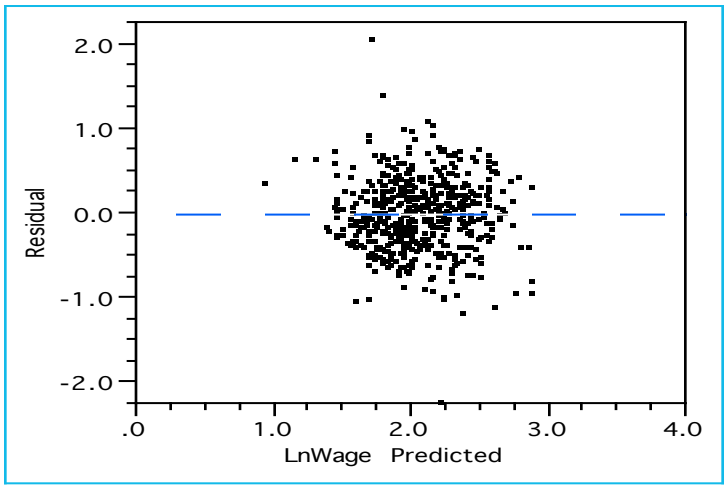
Term	Parameter Estimates Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.844	0.153	5.52	0.000
Race[AA-White]	-0.055	0.046	-1.21	0.227
Race[Hisp-White]	-0.010	0.059	-0.17	0.865
Union[No-Yes]	-0.339	0.126	-2.69	0.007
Married[No-Yes]	-0.016	0.021	-0.74	0.458
Female[No-Yes]	0.113	0.019	5.92	0.000
South[No-Yes]	-0.250	0.103	-2.42	0.016
YrsEduc	0.065	0.011	5.93	0.000
PotExpr	0.033	0.006	5.84	0.000
South[No-Yes]*YrsEduc	0.023	0.008	2.94	0.003
Union[No-Yes]*YrsEduc	0.018	0.010	1.93	0.054
PotExpr ^{2}	-0.001	0.000	-4.24	0.000

Effect Test

Source	Nparm	DF	Sum of Squares	F Ratio	Prob>F
Race	2	2	0.899	2.41	0.0905
Union	1	1	1.349	7.24	0.0074
Married	1	1	0.103	0.55	0.4578
Female	1	1	6.541	35.10	<.0001
South	1	1	1.095	5.88	0.0157

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	11	51.181	4.653	24.9718
Error	522	97.261	0.186	Prob>F
C Total	533	148.442		<.0001



---

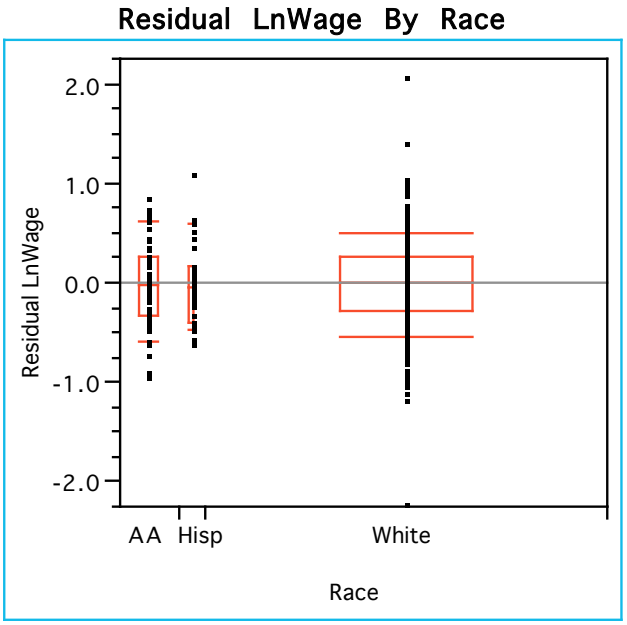
(32) Does Model 2 offer significantly better "explanatory power" than Model 1?
(a) Yes, the R^2 is significantly larger as shown by a partial F test.
(b) Yes, the R^2 is significantly larger since all three added coefficients are significant.
(c) No, the change in R^2 is too small to provide a significant improvement.
(d) No, the overall F-statistic has become smaller.
(e) Further Anova output is needed to answer this question.

(33) In Model 2, the observed increase in the p-value for the coefficient of the *Married* indicator variable is probably due to which of the following causes:
(a) The true marriage effect in the population is zero.
(b) The addition of the variables to form Model 2 from Model 1 has introduced collinearity.
(c) The use of a sample that is too small for estimating the additional coefficients.
(d) The proportion of explained variation is too small.
(e) The model lacks a more complete regional factor, identifying other parts of the US.

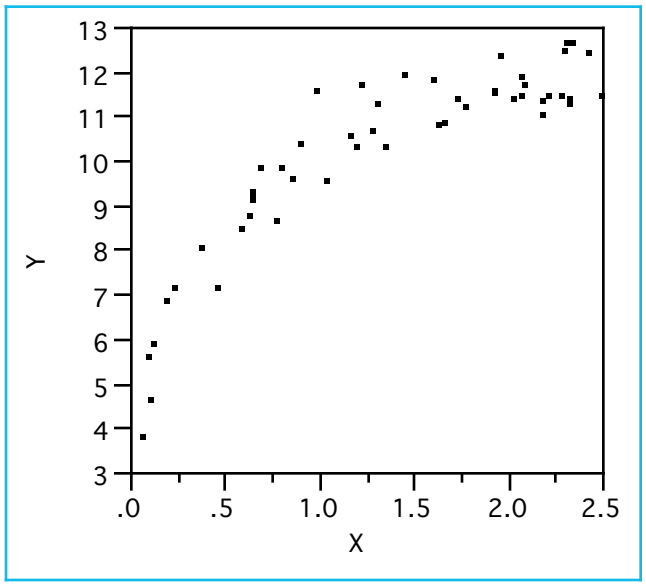
(34) Does Model 2 support the conjecture that the seniority system of unions in the US dilutes (weakens) the value of additional years of education?
(a) Yes, but the effect is marginally significant.
(b) No, but the effect is marginally significant.
(c) Yes, and the effect is very significant.
(d) No, and the effect is very significant.
(e) This model has too much collinearity to permit the interpretation of coefficients.

(35) From the given summary of Model 2, what is the approximate margin for error (half width of a 95% prediction interval) when predicting the wages of a single, southern, white woman with 14 years of education, 10 years of experience, and a non-union job?
(a) 0.070
(b) 0.372
(c) 0.862
(d) 1.100
(e) Cannot be determined without further information.

(36) The following plot shows residuals from Model 2, grouped by *Race*. *From this plot*, we can conclude that
(a) the proportion of white respondents violates the assumption of equal variance.
(b) variation among the white respondents is larger than for the other groups.
(c) the residual variance in these groups is comparable.
(d) the large outliers in the white group have distorted our regression results.
(e) there are no differences in average log wages among these groups.



(37) Which of the following actions would *not* be appropriate to take as a next step in this analysis using Model 2?
(a) Check for heteroscedasticity among other sub-groups in the model.
(b) Obtain a larger sample from the Current Population Survey.
(c) Check for normality of the residuals using a quantile plot.
(d) Remove the *Race* variable since its t-statistics have such large p-values.
(e) Use the variance inflation factors for Model 2 to judge the collinearity.



(38) Which of the following regression models would *not* capture the type of nonlinearity seen in the scatter plot shown above?
(a) Regress Y on log X.
(b) Regress log Y on X.
(c) Regress Y on 1/X.
(d) Regress Y^2 on X.
(e) Regress Y on X and X^2.

(Questions 39-42) These questions refer to the summary statistics shown below (a correlation matrix and the associated summary for a regression of Y on X1 and X2).

Correlations

Variable	Y	X1	X2
Y	1.0000	0.8697	0.6316
X1	0.8697	1.0000	0.7049
X2	0.6316	0.7049	1.0000

Root Mean Square Error	1.068
Observations	50

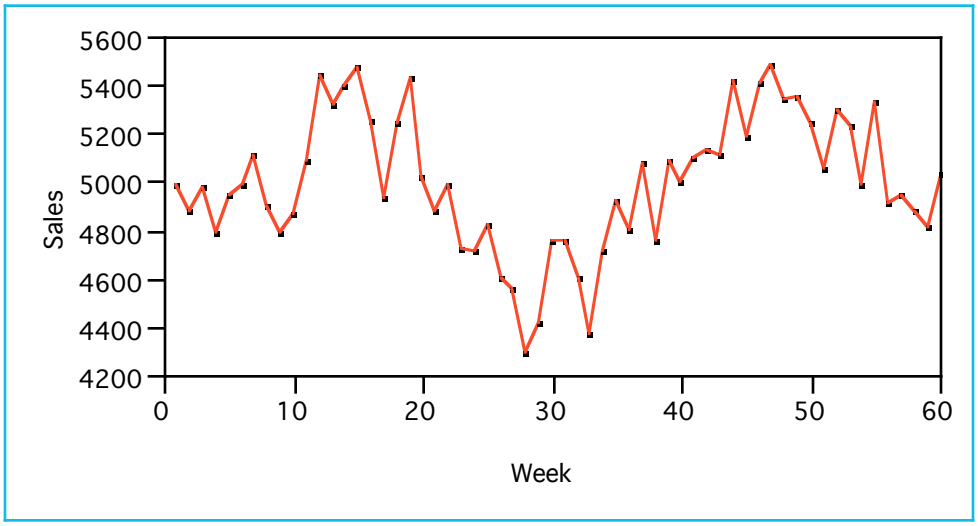
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	7.003	0.312	22.46	<.0001
X1	2.331	0.280	8.32	<.0001
X2	0.065	0.179	0.36	0.7178

(39) Which of the following values is the R^2 for this multiple regression of Y on X1 and X2?
(a) 0.397
(b) 0.705
(c) 0.757
(d) 0.812
(e) 0.873

(40) How is it possible that the response Y and the predictor X2 are highly correlated, but the multiple regression coefficient for X2 is not significant?
(a) Collinearity between X1 and X2.
(b) An outlier is affecting the slope in the multiple regression.
(c) The relationship between Y and X2 is non-linear.
(d) We need to check the Durbin-Watson statistic for independence.
(e) The underlying data is heteroscedastic and hence the t-statistics are not reliable.

(Questions 41-43) A small investment firm has been tracking its sales volume over a recent 60 week period. Management would like to use forecasts to anticipate active periods. The following regression model was developed to assist them. The variable *Week* is a simple time index.



Response Sales

RSquare	0.540
Root Mean Square Error	196.019
Mean of Response	5009.25
Observations	59

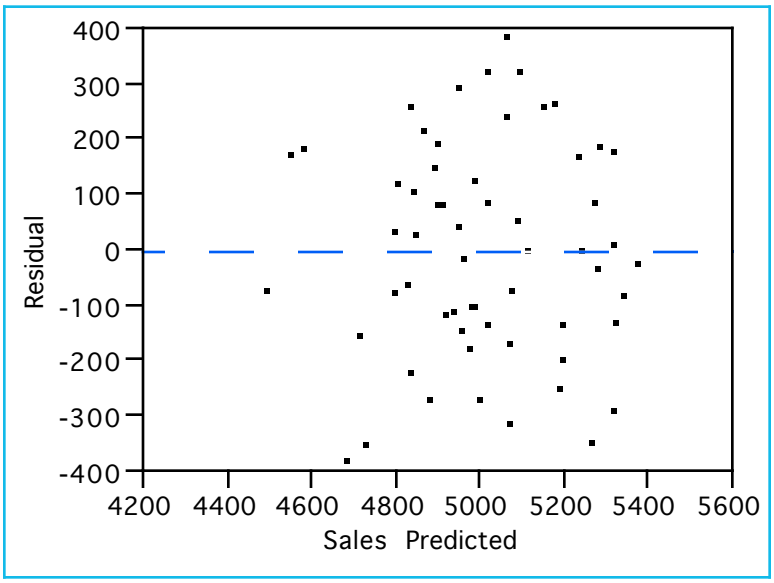
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	1337.714	454.712	2.94	0.0047
Week	0.643	1.512	0.43	0.6722
Lag Sales	0.729	0.091	7.98	<.0001

Durbin-Watson	Number of Obs.	AutoCorrelation
2.24	59	-0.1275

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	2530192.2	1265096	32.9251
Error	56	2151716.1	38424	Prob>F
C Total	58	4681908.3		<.0001



(41) Management has decided that in order to be useful, the model must be able to predict the coming week's sales within 5%. Is this model capable of achieving that level of performance?
(a) Yes, the overall fit is very significant with $R^2 > 0.5$.
(b) Yes, previous errors have been quite small.
(c) No, there is too much unexplained variation.
(d) No, the model has a non-significant term that weakens the forecast.
(e) In order to be within a specified % error, we need to transform to logs.

(42) The fitted coefficient 0.73 of lagged sales in this model indicates that
(a) the model captures 73% of the week-to-week variation.
(b) 73% of the explained variation is represented by the lagged variable.
(c) the sales data are autocorrelated with autocorrelation near 0.73.
(d) there is substantial collinearity between lagged sales and a time trend.
(e) None of the above.

(43) If the time trend variable *Week* were omitted from this model, then
(a) the model would lose its ability to track time trends seen in the sequence plot.
(b) there would be little change in the fit.
(c) the coefficient of lagged sales would change by several standard errors.
(d) the RMSE would increase by 65%.
(e) None of the above.

(44) Which of the following would *not* be a consequence of extending the length of this series to 120 weeks rather than just 60 weeks, assuming that the nature of the relationship observed here is stationary?
(a) The t-statistic for *Week* would increase in absolute size.
(b) The standard error for lagged sales would decrease.
(c) The p-value for the intercept would become smaller.
(d) The residual variance estimate would be smaller.
(e) None of the above (that is, items a-d would occur)