

## Statistics 621 Waiver Exam

August 27, 2000

This is an **open-book** test. You have **two hours** for the exam. The computer output associated with one or more items should be considered an essential part of the questions. The multiple-choice questions are equally weighted. Please also note the following:

- **Fill in your name and student id number** on the answer form.
- **Mark the “bubbles”** under your name and student id number on the form.
- Choose **one best answer** by marking the item on the answer form.
- Mark the answer form using only a **#2 pencil**. Erase all changes completely.

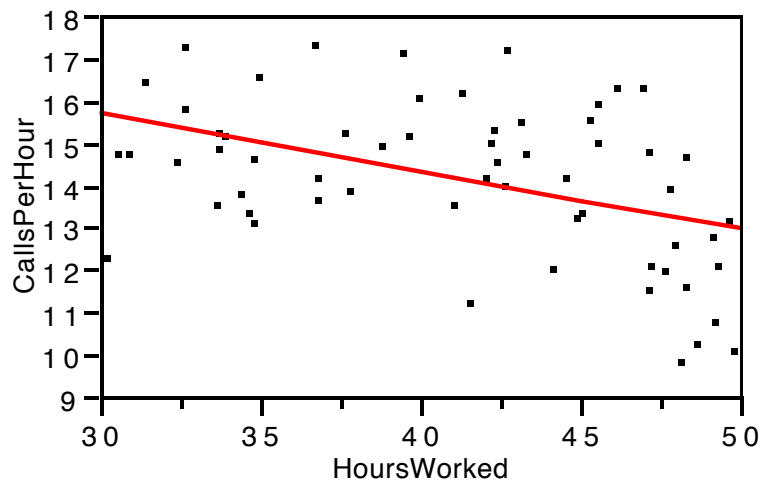
Turn in only the answer form; you may keep the test. Solutions will be posted on the Statistics 608 web page. Mark your copy of the exam copy in order to see how well you did. The list of those passing this exam will be available in 111 Vance.

---

### Question 1. Mark your solution form A for question #1.

---

(Questions 2-9) A service firm has experienced rapid growth. Because of this growth, some of the employees who handle customer calls have had to work additional hours (overtime). The firm is concerned that over-worked employees are less productive and handle fewer calls per hour than employees who work less demanding schedules. Most employees who work the “conventional” schedule put in 30-40 hours a week, depending upon demand. The firm constructed the regression model shown next relating the number of hours worked (X) to the number of calls serviced per hour (Y) for 60 employees.



#### Summary of Fit

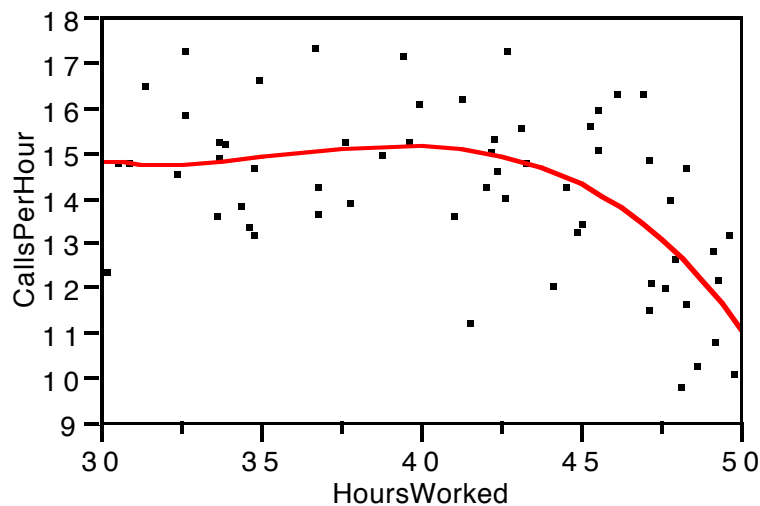
RSquare	0.19
Root Mean Square Error	1.71

Term	Parameter Estimates		t Ratio	Prob> t
	Estimate	Std Error		
Intercept	19.8	1.5	13.16	<.0001
HoursWorked	-0.14	0.04	-3.74	0.0004

- (2) The fitted model implies that an employee who works a 30 hour week processes on average about
- (a) 20 calls per hour.
  - (b) 16 calls per hour.
  - (c) 14 calls per hour.
  - (d) 10 calls per hour.
  - (e) 5 calls per hour.
- (3) The fitted model implies that for each additional 10 hours of work, the number of calls processed per hour, on average,
- (a) Drops by about 3.
  - (b) Drops by about 1.5.
  - (c) Does not change.
  - (d) Increases by about 1.5.
  - (e) Increases by about 3.
- (4) Based on the fitted model, an increase in the hours worked
- (a) Has an inconclusive effect on productivity because the sample is too small.
  - (b) Has a substantively interesting, but insignificant effect on productivity.
  - (c) Decreases the productivity by a significant amount, on average.
  - (d) Increases the productivity by a significant amount, on average.
  - (e) May affect some workers, but not most.
- (5) When the model was used to predict the number of calls handled by a new employee, the model's prediction of the number of calls per hour was 3 calls per hour *more* than the actual productivity. This prediction error implies that
- (a) The model is a poor match to the data and should be discarded.
  - (b) The model is very significant and can be used to identify valuable employees.
  - (c) This employee is performing significantly worse than his colleagues.
  - (d) This employee is performing significantly better than his colleagues.
  - (e) This employee is performing about the same as his colleagues.

- (6) This fitted model uses data from 60 employees. If the sample size were increased to 240 (i.e., fit to a sample that is four times larger), then
- (a) The confidence interval for the slope would be 1/4 as long.
  - (b) The prediction error would drop since the RMSE would be smaller.
  - (c) The fitted slope and intercept would be estimated more precisely.
  - (d) The model's  $R^2$  would increase, indicating a better fit.
  - (e) All of the above.
- (7) The  $R^2$  statistic for this model implies that
- (a) The model explains about 20% of the variation in productivity.
  - (b) The model predicts about 20% of the employees accurately.
  - (c) The model predicts about 80% of the employees accurately.
  - (d) Only 20% of the observations lie within 2 RMSEs of the fitted line.
  - (e) Explains a statistically insignificant portion of the variation in the response.

The firm hired a consultant to investigate the fitted regression model shown above. The consultant suggested as an improvement the following polynomial regression model.



### Summary of Fit

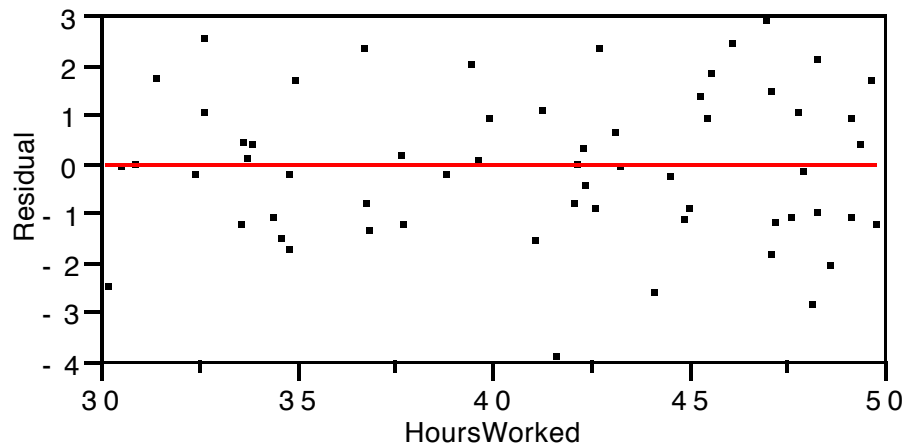
RSquare	0.35
Root Mean Square Error	1.52

### Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	18.057	3.451	5.23	<.0001
HoursWorked	-0.072	0.084	-0.86	0.3948
(HoursWorked-41.1251)^2	-0.028	0.007	-3.71	0.0005
(HoursWorked-41.1251)^3	-0.002	0.001	-1.30	0.1991

- (8) Predictions from the polynomial model proposed by the consultant
- (a) Are unreliable since the model contains terms that are not statistically significant.
  - (b) Are unreliable since polynomial models are poor representations of real phenomena.
  - (c) Are not significantly more accurate than those offered by the initial linear regression.
  - (d) Are significantly more accurate than those of the initial linear regression.
  - (e) Cannot be interpreted easily since the model is a nonlinear fit to the data.

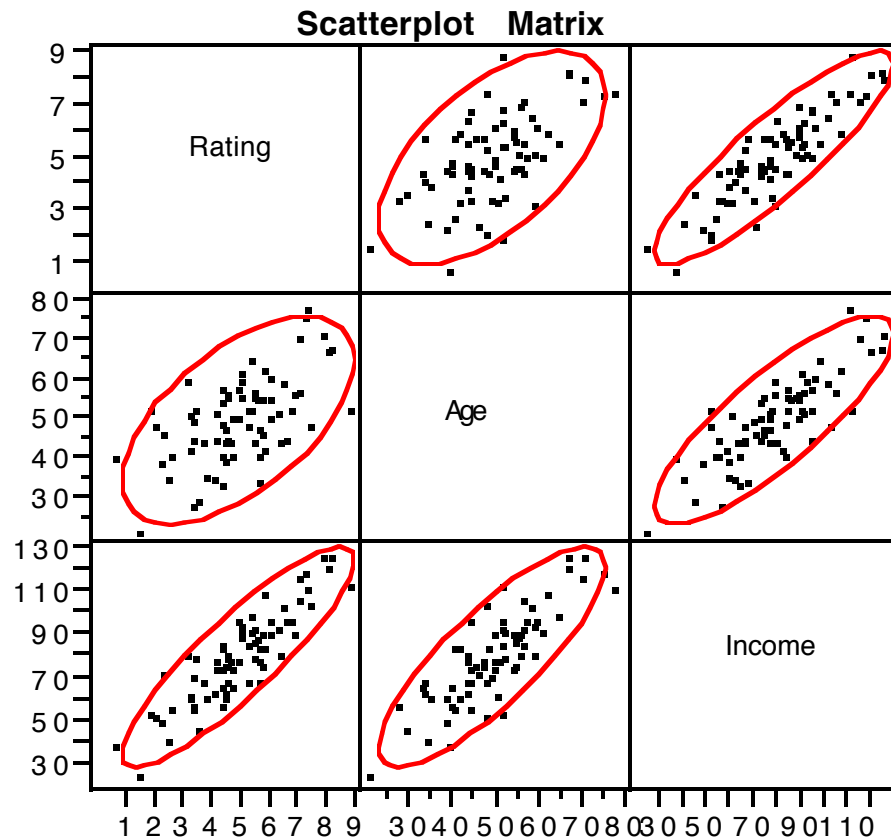
- (9) The residual plot shown below was generated by one of the previous two models. Which?



- (a) By the original linear model.
- (b) By the polynomial regression model.
- (c) Cannot be determined since the fit of the two models is so similar.
- (d) Cannot be determined since neither model explains much variation in the data.
- (e) Cannot be determined since both models leave outliers in the data.

**(Questions 10-15)** A marketing firm performed a study to assess interest in a new type of personal digital assistant (PDA). The firm obtained a sample of 75 consumers and showed each of the consumers (individually) the new device. It then asked each consumer to rate their “likelihood of purchase” on a scale of 1–10, with 1 implying low chance of purchase and 10 indicating high chance of purchase. The following three regression models all use this likelihood of purchase rating as the response, and each employs one or both of the following predictors: the age (in years) and the income (in thousands of \$’s) of the respondent.

Correlations			
	Rating	Age	Income
Rating	1.000	0.587	0.885
Age	0.587	1.000	0.829
Income	0.885	0.829	1.000



<b>Model 1 Regression Estimates</b>				
<b>Term</b>	<b>Estimate</b>	<b>Std Error</b>	<b>t Ratio</b>	<b>Prob&gt; t </b>
Intercept	-0.596	0.352	-1.69	0.0951
Income	0.070	0.004	16.20	<.0001

<b>Model 2 Regression Estimates</b>				
<b>Term</b>	<b>Estimate</b>	<b>Std Error</b>	<b>t Ratio</b>	<b>Prob&gt; t </b>
Intercept	2.067	0.487	4.24	<.0001
Age	0.059	0.009	6.19	<.0001

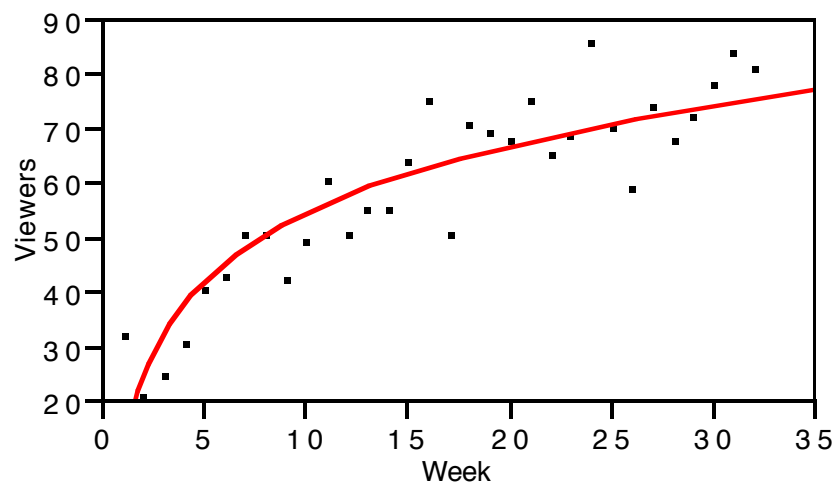
<b>Model 3 Regression Estimates</b>				
<b>Term</b>	<b>Estimate</b>	<b>Std Error</b>	<b>t Ratio</b>	<b>Prob&gt; t </b>
Intercept	-0.736	0.295	-2.50	0.0149
Age	-0.047	0.008	-5.74	<.0001
Income	0.101	0.006	15.63	<.0001

(10) The scatterplot matrix shows that

- (a) The predictors are positively correlated.
- (b) The range of ages is about 20 to 75 years.
- (c) *Income* is a better predictor of *Rating* than *Age* (when used separately).
- (d) The data is roughly normally distributed.
- (e) All of the above.

- 
- (11) Which of the following correctly explains the negative intercept in Model 3 (*Rating* on *Age* and *Income*)?
- (a) The variance of the error terms increases near the origin.
  - (b) The intercept represents an extrapolation far from observed data.
  - (c) The estimate is not significantly different from zero.
  - (d) The intercept is never of interest in multiple regression without categorical terms.
  - (e) All of the above.
- (12) Does adding *Age* to the regression of *Rating* on *Income* significantly improve the predictive ability of the model?
- (a) Yes, because adding *Age* caused the coefficient of *Income* to increase in size.
  - (b) No, because adding *Age* caused the t-statistic of *Income* to decrease in size.
  - (c) Yes, because the t-statistic of age is larger than 2 (absolute size) in Model 3.
  - (d) No, because the t-statistic of age is negative in Model 3.
  - (e) Cannot be answered without the  $R^2$  statistic from the multiple regression.
- (13) Among consumers with incomes of \$100,000, we would expect those at age 50 to differ from those at age 40 in what way?
- (a) Those at age 50 will give an average rating about 0.5 lower.
  - (b) Those at age 50 will give an average rating about 0.6 higher.
  - (c) Both groups will have comparable average values.
  - (d) Those at age 50 will have more variable responses.
  - (e) Those at age 40 will have more variable responses.
- (14) The regression of *Rating* on *Age* has a positive slope (Model 2), whereas in Model 3 the slope for *Age* is negative. We can explain this change in sign as arising because
- (a) Of random sampling variation.
  - (b) The values are estimates with overlapping confidence intervals.
  - (c) The fitted models are affected by autocorrelation.
  - (d) The predictors in Model 3 are collinear.
  - (e) Both of these coefficients are close to zero and the differences are not important.
- (15) Based on this regression analysis, which of the following groups is most likely a good target for marketing this PDA?
- (a) Older consumers.
  - (b) Affluent consumers.
  - (c) Younger, affluent consumers.
  - (d) Older, affluent consumers.
  - (e) Without a categorical grouping, we cannot assess market reaction.

**(Questions 16-25)** A network television show features a plot line that gradually eliminates characters from the cast as the season progresses. In some episodes, one of the original cast of 20 disappears, sometimes happily, sometimes tragically. To keep the audience engaged, viewers are given “hints” in previous episodes that suggest who will be next. Viewers can win cash prizes by predicting who will go next. The theme has been very successful, and the show’s audience grew dramatically over its first 16-week season. The success of the show led to it being continued for a second 16-week season that immediately followed the first. The following model describes the growth in number of millions of viewers over this 32-week period. The number of viewers is expressed in millions, and the week of the episode is consecutively labeled 1,2,...32.



#### Summary of Fit

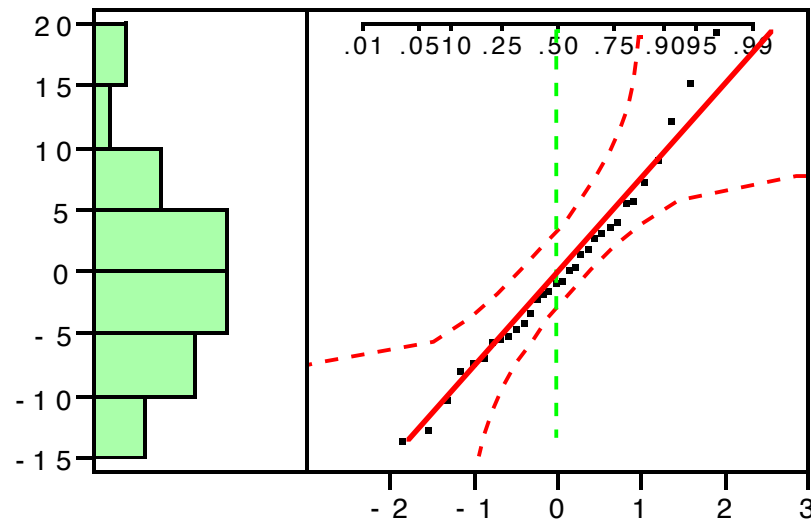
RSquare	0.81
Root Mean Square Error	7.675
Mean of Response	59.067
Observations	32

Term	Parameter	Estimates		
	Estimate	Std Error	t Ratio	Prob> t
Intercept	12.8	4.3	2.95	0.0061
Log(Week)	18.2	1.6	11.28	<.0001

- (16) If we replace the natural log of time as used in this model by the base 10 log of time, then how would the fitted model change?
- The fit would be improved significantly.
  - The intercept would be larger by a factor of about 2.3.
  - The slope would be larger by a factor of about 2.3.
  - The RMSE would drop by a factor of 2.3.
  - The t-statistic of the slope would increase by about 2.3.

- 
- (17) The intercept in the fitted model implies that
- (a) About 13 million watched this network during week 0 (week zero).
  - (b) About 13 million were predicted to watch during the first week of this program.
  - (c) The model is a poor fit since the intercept is not significantly different from zero.
  - (d) The model is a poor fit since the standard error of the intercept is so large.
  - (e) The number of viewers grows by about 13% per week of the series.
- (18) The executive producer of this program claimed that the week-to-week growth of the number of viewers was at least 2 million per week. Based upon the fitted model, we can conclude that the claim is
- (a) Indeed true over these 32 weeks.
  - (b) Not true for any of these weeks.
  - (c) True only for the first 10 weeks.
  - (d) True only for the first 15 weeks.
  - (e) True only after week 15.
- (19) If the program were to continue into week 33 and the trend represented by this model continue to hold, then the probability that at least 85 million viewers would tune in to watch in week 33 is roughly
- (a)  $1/6$
  - (b)  $1/3$
  - (c)  $1/2$
  - (d)  $2/3$
  - (e)  $5/6$
- (20) From the following plot of the residuals from this model, we can conclude that
- (a) The residuals are approximately normally distributed.
  - (b) The residuals lack constant variation.
  - (c) The tracking pattern evident in the sequence of residuals indicates autocorrelation.
  - (d) The mean value of the residuals is positive, indicating a lack of fit.
  - (e) The model has explained little variation in the data and is not useful.





Normal Quantile Plot

It has been claimed that the logarithmic model fit above is inappropriate. An alternative multiple regression was used instead. This model uses a categorical variable for the season (coded as “First” for the first 16 weeks, and “Second” thereafter) and is summarized below.

### Summary of Fit

RSquare	0.857
Root Mean Square Error	6.876

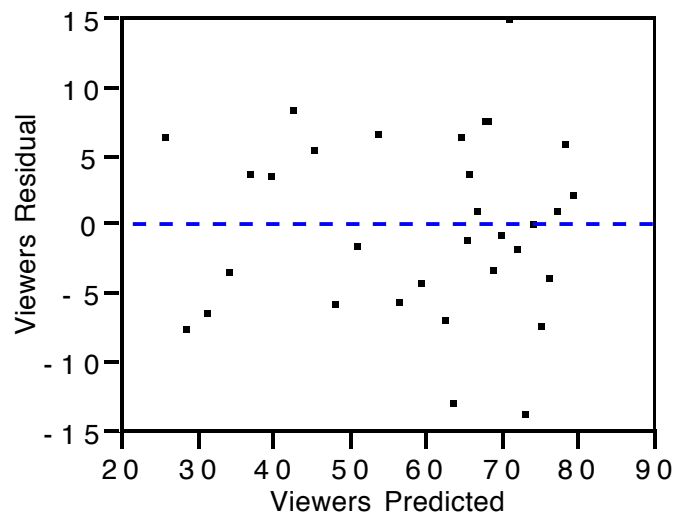
### Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	7937.12	2645.71	55.9532
Error	28	1323.96	47.28	<b>Prob &gt; F</b>
C. Total	31	9261.08		<.0001

### Expanded Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	34.31	4.99	6.88	<.0001
Week	1.90	0.25	7.60	<.0001
Season[First]	3.08	2.43	1.27	0.2159
Season[Second]	-3.08	2.43	-1.27	0.2159
(Week-16.5)*Season[First]	0.90	0.25	3.60	0.0023
(Week-16.5)*Season[Second]	-0.90	0.25	-3.60	0.0023

- (21) Based on the revised model, does the rate of growth decrease significantly during the second season?
- (a) Yes, the coefficient of  $(\text{Week}-16.5)*\text{Season}[\text{First}]$  is positive and significant.
  - (b) Yes, the coefficient of  $\text{Season}[\text{First}]$  is positive and significant.
  - (c) Yes, the coefficient of “ $\text{Season}[\text{First}]$ ” is positive.
  - (d) No, the coefficient of “ $\text{Season}[\text{First}]$ ” is positive.
  - (e) The added terms do not improve the model and no conclusions are warranted.
- (22) From this fitted model, we would conclude that during the second season on average the audience for this program
- (a) Increased by about 1.9 million per week.
  - (b) Increased by about 0.9 million per week.
  - (c) Decreased by about 0.9 million per week.
  - (d) Increased by about 1 million per week.
  - (e) Increased by about 2.8 million per week.
- (23) According to the fitted model, the audience grew on average by about
- (a) 0.4 to 1.4 million per week faster in the first season.
  - (b) 0.4 to 1.4 million per week slower in the first season.
  - (c) 0.8 to 2.8 million per week faster in the first season.
  - (d) 0.8 to 2.8 million per week slower in the first season.
  - (e) The same amount during both seasons.
- (24) The plot shown on the next page displays residuals from the multiple regression. From this plot we can conclude that
- (a) The errors in the underlying model violate the assumption of independence.
  - (b) The errors in the underlying model violate the assumption of equal variance.
  - (c) The errors in the underlying model are not normally distributed.
  - (d) An outlier has distorted the analysis.
  - (e) That this view of the residuals suggests no problems with the model.



- (25) Which of the following would *not* be an appropriate next step to the analysis? (i.e., which of the following would be a pretty dumb thing to do next?)
- (a) Review the accuracy of the data used in building these models.
  - (b) Investigate the use of other predictors, such as the ratings of the preceding program.
  - (c) Check for autocorrelation in the residuals.
  - (d) Remove the confusing interaction term from the model.
  - (e) Check for the equality of variances of the errors in the first and second seasons.

**(Questions 26-34)** An e-business is considering a plan that would utilize a web site that is customized to the individual tastes of each regular visitor to the site. The business is concerned, though, that it will not be able to generate (i.e., “serve”) such individualized pages quickly. As a preliminary stage in the development of its business plan, it has run a competition among its staff to see if such pages can be constructed quickly. Each competitor, using the computer language of choice wrote a program to generate such customized pages; 16 used “C”, 24 used “Java”, and 40 used “Script”. The time to generate such a page (on average) was obtained under a collection of conditions; this time is measured in seconds and is the response in this analysis. Other predictors of the speed of the program include the age of the programmer (Age), the sex of the programmer (“male” or “female”), the number of lines of code written, and the length of time (in hours) required to write the demo program.

#### Summary of Fit

RSquare	0.60
Root Mean Square Error	4.5
Mean of Response	10.4
Observations	80

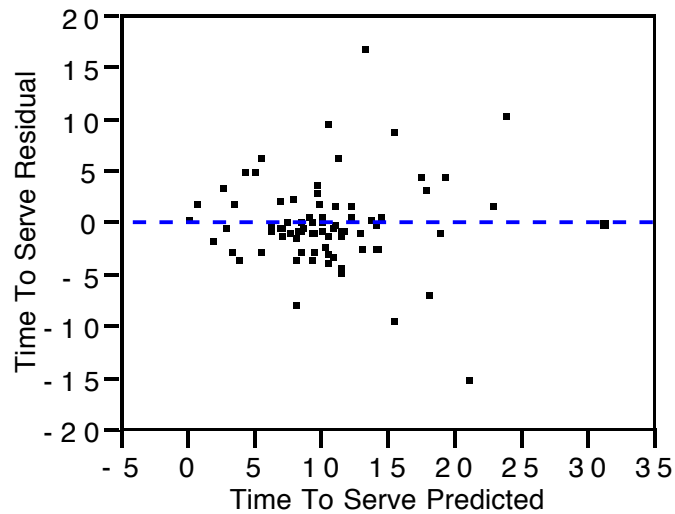
Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	8	2157.2579	269.657	13.2619
Error	71	1443.6613	20.333	<b>Prob &gt; F</b>
C. Total	79	3600.9191		<.0001

Expanded Estimates				
Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	4.469	2.217	2.02	0.0476
Language[C]	-8.021	1.195	-6.71	<.0001
Language[Java]	0.721	1.064	0.68	0.5002
Language[Script]	7.300	1.181	6.18	<.0001
Lines Of Code	0.023	0.007	3.06	0.0031
Coding Time	0.367	0.067	5.47	<.0001
Sex[Female]	0.227	0.638	0.36	0.7232
Sex[Male]	-0.227	0.638	-0.36	0.7232
Age	-0.026	0.039	-0.67	0.5052
Language[C]*(Lines Of Code-187.875)	-0.012	0.010	-1.20	0.2358
Language[Java]*(Lines Of Code-187.875)	-0.029	0.008	-3.67	0.0005
Language[Script]*(Lines Of Code-187.875)	0.041	0.011	3.78	0.0003

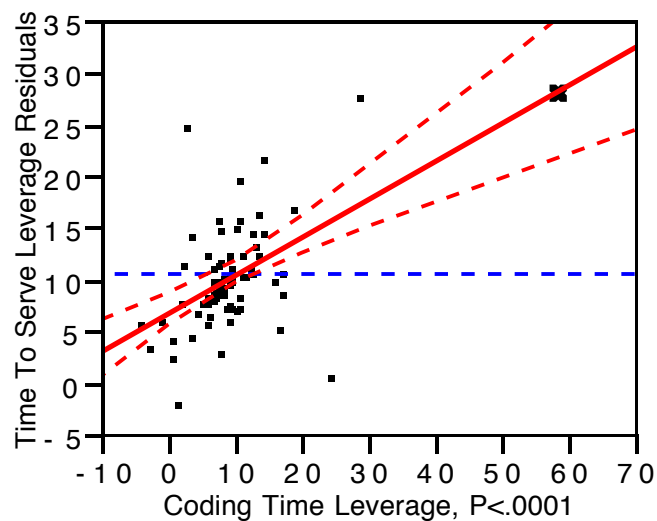
Effect Tests					
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Language	2	2	1064.79	26.18	<.0001
Lines Of Code	1	1	190.54	9.37	0.0031
Coding Time	1	1	607.47	29.88	<.0001
Sex	1	1	2.57	0.13	0.7232
Age	1	1	9.12	0.45	0.5052
Language*Lines Of Code	2	2	376.87	9.27	0.0003

- (26) Does the model developed by the business explain significant variation in the measured time to prepare the test web pages?
- (a) Yes, several variables have t-statistics that are larger than 2 (in absolute size).
  - (b) No, several variables in the model are not statistically significant.
  - (c) Yes, the overall F statistic is significant.
  - (d) Yes, the  $R^2$  statistic is larger than 50%.
  - (e) No, the RMSE is too large to have explained significant variation.
- (27) According to the model, do men write faster programs, on average, than women?
- (a) Yes, the coefficient for Sex[Female] is positive.
  - (b) No, the coefficient for Sex[Male] is negative.
  - (c) No, the effect of sex in this model is not significant.
  - (d) No, the coefficient for Sex[Male] is the negative of the coefficient for Sex[Female].
  - (e) No, the model is missing an important interaction of Sex with Age.

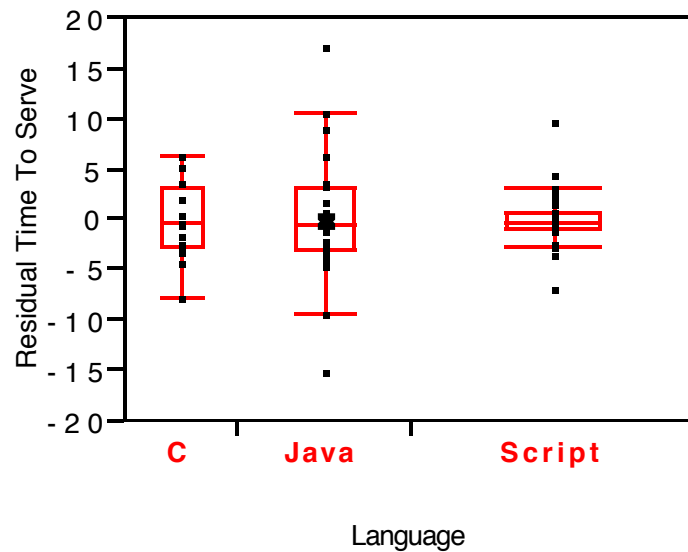
- (28) It has been claimed that the business should strive to have programmers write the shortest program because “short programs run faster.” With regard to this claim, the fitted model implies that
- (a.) The claim is true since the coefficient of *Lines Of Code* is significantly positive.
  - (b.) The claim is false since the coefficient of *Lines Of Code* is not significant.
  - (c.) The claim is only true for programs written in “Script”.
  - (d.) The claim is not true for programs written in “Java”.
  - (e.) The claim is not true for programs written in “C” or “Java”.
- (29) Assuming that the program is written using 200 lines of code, which language is most likely to provide the fastest solution, on average, based on this model?
- (a.) C
  - (b.) Java
  - (c.) Script
  - (d.) It does not matter which language is used.
  - (e.) This conclusion cannot be addressed by the fitted model.
- (30) Based on the tabular summary shown, a logical next step in the development of this model would be to remove
- (a.) The categorical indicating sex of the programmer.
  - (b.) Both age and the variable indicating sex of the programmer.
  - (c.) *Lines Of Code* from the model since its coefficient is small.
  - (d.) *Language[Java]* from the model.
  - (e.) All of the terms whose associated p-value is less than 0.05.
- (31) The p-value for Sex[Female]
- (a.) Implies that about 72% of this sample of programmers are women.
  - (b.) Implies that the probability that the population slope is zero is about 0.72.
  - (c.) Implies that about 72% of women in the population have zero slope, on average.
  - (d.) Requires that we assume that the population slope is zero.
  - (e.) Requires that half of the data be women.
- (32) The following plot shows the residuals from the multiple regression. This plot suggests that
- (a.) The model is likely to over-predict the length of time for long programs.
  - (b.) The model is likely to under-predict the length of time for long programs.
  - (c.) The prediction accuracy of the model decreases for short programs.
  - (d.) The prediction accuracy of the model increases for short programs.
  - (e.) The model is well-matched to the data and no problems are suggested.



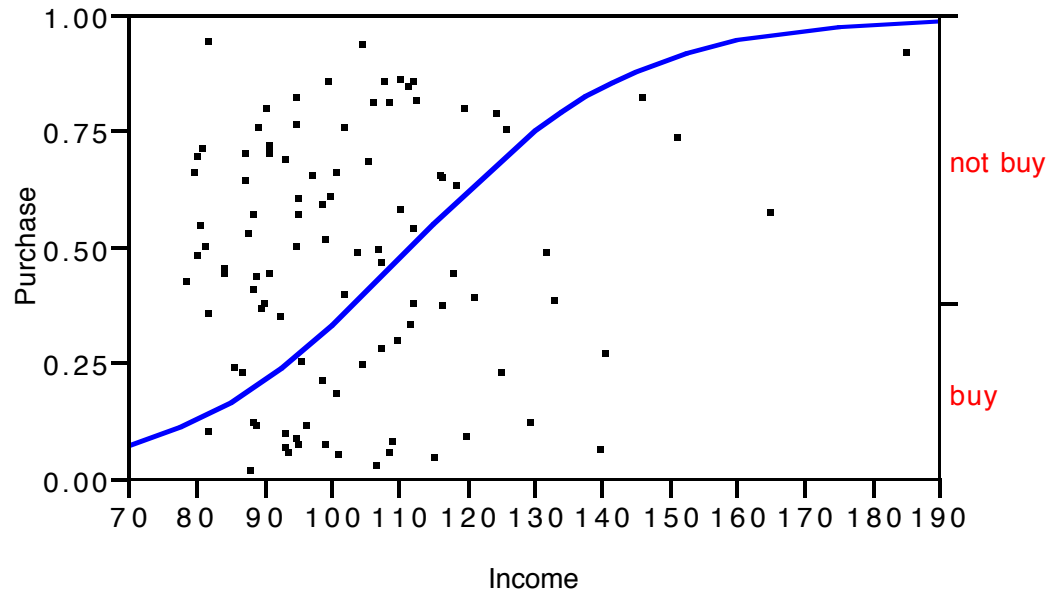
- (33) The following plot is a leverage plot for the variable *Coding Time*. If the model were run again without the point marked with an x in the plot, we would expect that
- (a.) The p-value for *Coding Time* would not change.
  - (b.) The  $R^2$  summary would increase.
  - (c.) The RMSE would decrease.
  - (d.) The standard error for *Coding Time* would increase.
  - (e.) The estimated slope for *Coding Time* would decrease.



- (34) The following plot of the residuals from this model indicates that
- (a.) The residuals are not from a normal distribution.
  - (b.) The model consistently over-predicts time required for Script programs.
  - (c.) The residuals are dependent, with dependence introduced by language type.
  - (d.) The model is most accurate at predicting times for Script programs.
  - (e.) The model violates a crucial assumption and must be discarded.



(Questions 35-37) An automobile manufacturer has conducted a study of the types of customers who purchase a particular option, namely its version of a computer-automated direction system (so you don't get lost while driving). The choice of this option adds an additional \$2000 to the purchase price of the car. The following model summarizes one model that attempts to explain the purchase choice (buy/not buy) in terms of the income of the purchaser. The income of the purchasers is measured in thousands of US dollars.



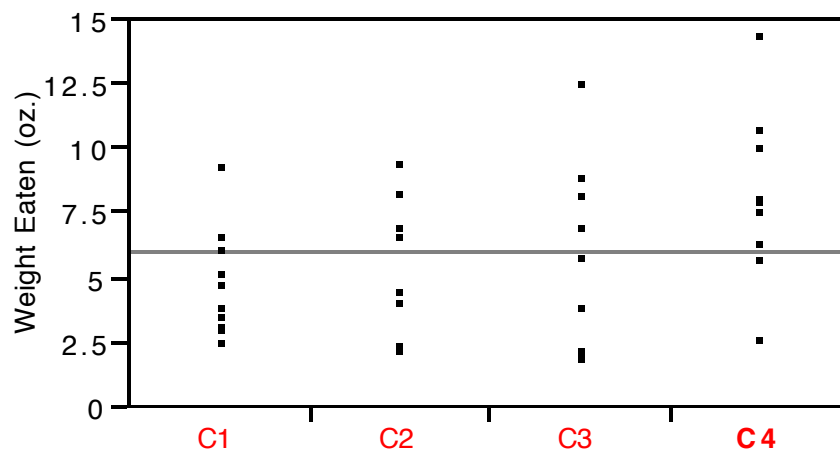
Parameter Estimates				
Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-6.710	1.666	16.23	<.0001
Income	0.060	0.016	14.13	0.0002
For log odds of buy/not buy				

- (35) Does income of the customer significantly impact on the probability of selecting this option?
- (a.) No, the effect is in the right direction but not significantly large.
  - (b.) No, the effect is near zero.
  - (c.) No, the effect is in the wrong direction and not meaningful.
  - (d.) Yes, the effect is significant and in the expected direction.
  - (e.) We have too little data to evaluate the effect of income.
- (36) If the probability of the customer purchasing this option is small, the company would prefer its sales personnel to avoid the lengthy promotional discussion. Assuming that the customer's income is known from a credit application, at what income does the probability drop below 0.25 for all smaller incomes?
- (a.) \$90,000.
  - (b.) \$95,000.
  - (c.) \$110,000.
  - (d.) \$115,000.
  - (e.) \$125,000.



- (37) From the fitted model, does an increase in income from \$100,000 to \$110,000 have the same impact on the probability of choosing this option as an increase from \$150,000 to \$160,000?
- (a.) Yes, both changes in income have equivalent effect.
  - (b.) Yes, both changes are equivalent since the model suggests no impact for income.
  - (c.) No, the impact of the change from \$100,000 to \$110,000 is smaller.
  - (d.) No, the impact of the change from \$100,000 to \$110,000 is larger.
  - (e.) No, the constant term in the model implies increasing odds of purchase.

**(Questions 38-47)** A manufacturer of dog treats would like to be able to claim that dogs like its treats better than those of competitors. It conducted the following experiment. A collection of 40 dogs was randomly divided into 4 groups, 10 in each. Dogs in each group were offered treats manufactured by one of the different companies, here C1, C2, C3 and C4 (the company running the trial is C4). The weight of treats consumed was recorded and is the response in this analysis.



Manufacturer							
Level	Number	Mean	Std Dev	Std Err	Mean	Lower 95%	Upper 95%
C1	10	4.80	2.09		0.66	3.46	6.14
C2	10	4.85	2.77		0.87	3.08	6.63
C3	10	6.09	3.63		1.15	3.77	8.42
C4	10	8.40	3.28		1.04	6.30	10.50

Analysis of Variance						
Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F	
Manufacturer	3	85.05	28.35	3.1540	0.0365	
Error	36	323.60	8.99			
C. Total	39	408.65				

**Comparisons for all pairs using Tukey-Kramer HSD**

Abs(Dif)-LSD	<b>C4</b>	<b>C3</b>	<b>C2</b>	<b>C1</b>
C4	-3.61	-1.31	-0.07	-0.02
C3	-1.31	-3.61	-2.37	-2.32
C2	-0.07	-2.37	-3.61	-3.56
C1	-0.02	-2.32	-3.56	-3.61

**Comparisons with the best using Hsu's MCB**

Mean[i]-Mean[j]-LSD	<b>C4</b>	<b>C3</b>	<b>C2</b>	<b>C1</b>
C4	-2.86	-0.55	0.69	0.74
C3	-5.16	-2.86	-1.62	-1.57
C2	-6.41	-4.10	-2.86	-2.81
C1	-6.46	-4.15	-2.91	-2.86

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

Mean[i]-Mean[j]+LSD	<b>C4</b>	<b>C3</b>	<b>C2</b>	<b>C1</b>
C4	2.86	5.16	6.41	6.46
C3	0.55	2.86	4.10	4.15
C2	-0.69	1.62	2.86	2.91
C1	-0.74	1.57	2.81	2.86

If a column has any negative values, the mean is significantly greater than the min.

- (38) Using this study, can any of the four manufacturers make the claim that their brand is preferred significantly more than any of the others (using weight eaten to measure preference)?
- Yes, the Tukey-Kramer procedure finds significant differences.
  - No, the Tukey-Kramer procedure finds no significant differences.
  - Yes, the Hsu procedure finds significant differences.
  - No, the Hsu procedure finds no significant differences.
  - No, the group means are too close relative to the background random variation.
- (39) This analysis used a collection of dogs of various breeds, of varying size. If the experiment were repeated using 40 dogs of the *same* breed (say, a breed of “typical” size), then we could expect that
- The appropriate test would find several significant effects.
  - The appropriate test would not find any significant effects.
  - The random variation in food eaten would be reduced.
  - The analysis would be invalidated since the data would then be dependent.
  - The data would be closer to normally distributed.

- (40) A professor had a different question in mind that utilized only the difference in mean values from the treats prepared by C2 and C4. The appropriate 95% confidence interval for this professor to use to measure the mean difference between C2 and C4 is
- (a.)  $[3.55 \pm 2.25]$
  - (b.)  $[3.55 \pm 2.68]$
  - (c.)  $[3.55 \pm 3.62]$
  - (d.)  $[3.55 \pm 12.1]$
  - (e.) Cannot be determined from the available summary information.
- (41) The manufacturer who ran this study is considering a larger study, and would like to get an idea of “what would happen if” it ran the study with twice as many dogs (80 altogether rather than 40, with 20 dogs per group rather than 10). To investigate the potential gains, an analyst simply duplicated the data (added 40 rows to the data by copying the first 40). This approach
- (a.) Is foolish because duplication of the data introduces heteroscedasticity.
  - (b.) Is foolish because duplication of the data introduces a lack of normality.
  - (c.) Is foolish because duplication of the data is just too easy to do.
  - (d.) Is foolish because duplication of the data introduces dependence.
  - (e.) Is reasonable and will convey a good sense of what would be found in a larger study.

**(Questions 42-45, continued)** A rival dog treat manufacturer learned of the plans of the company described in the previous group of questions. This manufacturer took an alternative approach. It too used 40 dogs in its study, but used 8 of each of five species (chihuahua, spaniel, poodle, shepherd, St. Bernard). A randomly selected pair of dogs of each breed were fed one of the four brands of dog treats (the same four brands, as above). Again, weight consumed was measured.

#### Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	19	1159.52	61.03	16.0
Error	20	76.28	3.81	<b>Prob &gt; F</b>
C. Total	39	1235.79		<.0001

#### Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Manufacturer	3	3	45.44	3.97	0.0226
Breed	4	4	1077.48	70.63	<.0001
Manufacturer*Breed	12	12	36.59	0.80	0.6475

- (42) Was the analysis improved by structuring the experiment to account for differences in food consumption across breeds?
- (a.) Yes, the addition of *Breed* and its interaction are significant.
  - (b.) No, the interaction term of *Breed* and *Manufacturer* is not significant.
  - (c.) Yes, the addition of *Breed* to the model explains significant additional variation.
  - (d.) Yes, the overall F statistic reported in the Anova Table is significant.
  - (e.) Cannot tell without the results for an analysis using only *Manufacturer* (for this data).
- (43) The results of this model imply that
- (a.) Some breeds prefer the products of different manufacturers.
  - (b.) Breeds show no differences in preference among these products.
  - (c.) All of the breeds consume about the same amount of the treats.
  - (d.) There are no significant differences among the products of these manufacturers.
  - (e.) A St. Bernard will eat more than a chihuahua, on average.
- (44) The tabular summary of this model implies that we can predict the “treat consumption” of a dog within what range with 95% probability?
- (a.) Within  $\pm 3.9$ , if we know only the breed.
  - (b.) Within  $\pm 3.9$ , if we know only the manufacturer.
  - (c.) Within  $\pm 3.9$ , if we know both the manufacturer and breed.
  - (d.) Within  $\pm 7.6$ , if we know both the manufacturer and breed.
  - (e.) Cannot be estimated without the root mean squared error.
- (45) The *symmetry* of the residuals around the horizontal line at zero in the following residual plot from the analysis of variance
- (a.) Indicates the presence of collinearity in the model.
  - (b.) Indicates the presence of dependence in the underlying measurements.
  - (c.) Indicates that the underlying data are not normally distributed.
  - (d.) Is produced by the use of two factors in the design of this regression model.
  - (e.) Is an artifact of using two dogs for each combination of manufacturer and breed.

