Statistics 608-621 Waiver Exam

August 18, 2008

Before the exam begins, on the answer sheet…

- **Write in your name and Penn student id number**. Your Penn ID number appears in bold numerals on your ID card.
- **Mark the "bubbles"** under the letters of your name *and* under your student id number on the form.
- **Use a #2 pencil**. Erase changes on the answer form completely.

Once the exam begins …

- Choose the **one *best* answer** for each question. Picking more than one answer is scored as an error.
- You may consult **1 page of handwritten notes** during the exam. No other reference materials are permitted.
- You may use a **calculator**. No laptops or telephones are allowed. If you have a phone with you, please silence it prior to the exam.

Turn in the solution page only.

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 word "significant" means "statistically significant" in these questions.

Your **score** is the number of correct answers. The multiple-choice questions are equally weighted. Some questions may be dropped and not counted as part of the overall score. There is no deduction for incorrect answers.

The solutions will be posted in WebCafé. If you wish to compare your answers to the solutions, then mark your choices on your copy of the exam. Regardless of what you write on your copy of the exam, however, only the answers marked on the graded answer form will be considered. You can use the "My Grades" feature of WebCafé to find the score determined by your answers on the answer form.

STOP

*Do **not** turn the page until you are instructed to proceed.*

(1) **Mark the answer for Question #1 on your answer form "a".** (This question identifies your version of the exam; you must mark this item as instructed.)

(2) Two independent surveys of the same population were used to estimate the simple regression model $y = \beta_0 + \beta_1 x + \varepsilon$. Survey 1 is a sample of $n_1 = 100$ cases whereas survey two is a sample of $n_2 = 400$ cases. In both cases, the 95% confidence interval (CI) for β_1 was constructed. It follows that
   (a) The CI from Survey 2 is more likely to contain β_1 than the CI from Survey 1.
   (b) Either the CI from Survey 1 or the CI from Survey 2 contains β_1.
   (c) If the CI from Survey 1 contains β_1, then so will the CI from Survey 2.
   (d) Both confidence intervals either contain β_1 or do not contain β_1.
   (e) The probability that both CIs contain β_1 is less than 0.95.

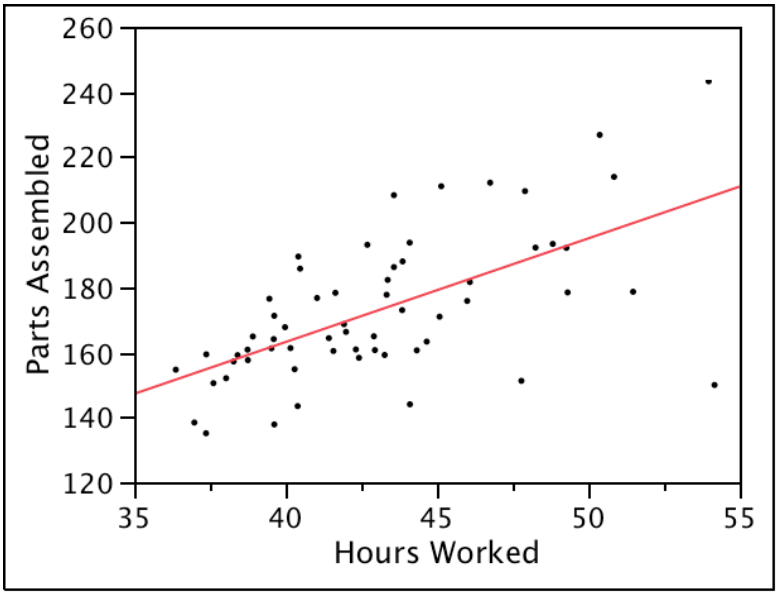
(3) If the test of a slope in a multiple regression model with 5 explanatory variables fails to reject the null hypothesis H_0: $\beta_1 = 0$ (at a 5% level of statistical significance), then we know that
   (a) The data violate an assumption of the multiple regression model.
   (b) The observations are independent of each other.
   (c) The parameter β_1 is zero.
   (d) Zero lies inside the 95% confidence interval for β_1.
   (e) The sample size was too small.

(4) If a leveraged outlier is removed from the data used to estimate a simple regression model, then we should anticipate that if the regression is estimated without this leveraged outlier then
   (a) R^2 will increase.
   (b) R^2 will decrease.
   (c) *RMSE* will decrease.
   (d) The 95% confidence interval for the slope will be wider.
   (e) The estimated slope will become larger and the estimated intercept smaller.

(5) Consider two simple regression models: the first regresses monthly sales on monthly promotion spending and the second regresses monthly sales on the natural log of the monthly promotion spending. Then
   (a) The second model allows diminishing marginal returns.
   (b) The second model will have a larger R^2.
   (c) The second model will have a smaller R^2.
   (d) We cannot compare R^2 between these models.
   (e) The second model estimates the elasticity of sales with respect to promotion.

(6) The categorical variable *Region* identifies 4 geographic regions. If *Region* is added to a regression model, then the addition of *Region* to the model is statistically significant at the $\alpha = 0.05$ level if
   (a) The RMSE of the model decreases when *Region* is added to the model.
   (b) The absolute size of at least one *t*-statistic associated with *Region* is larger than 2.
   (c) The R^2 statistic increases by at least 0.04.
   (d) The overall *F* statistic of the model increases when *Region* is added to the model.
   (e) The *p*-value of the associated partial *F* statistic is less than 0.05.

(Questions 7-19) A manufacturer produces parts for customized motorcycles. Though similar, each part requires manual assembly, and an individual employee assembles each part (rather than use an assembly line). A recent surge of orders required some employees to work more than the typical 40-hour week. Management is concerned that such an increase in the number of hours worked affects the time required for assembly. The time allotted for assembling this type of part is 15 minutes. These data show the number of hours worked and the number of parts assembled by 60 employees during a recent week.



RSquare	0.3765
Root Mean Square Error	17.7106
Mean of Response	172.7951

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	35.581107	23.29815	1.53	0.1321
Hours Worked	3.1874899	0.538605	5.92	<.0001

(7) The R^2 statistic for this model implies that
- (a) The model explains about 38% of the variation in parts assembled.
- (b) The model accurately predicts parts assembled for about 38% of the employees.
- (c) About 38% of these employees are performing as management expects.
- (d) About 38% of the observations lie within 2 RMSEs of the least squares line.
- (e) The fitted line explains statistically insignificant variation in the response.

(8) If the explanatory variable in the model had been the number of minutes worked rather than the number of hours, then
- (a) R^2 would have been larger.
- (b) *RMSE* would have been larger.
- (c) The estimated intercept would be closer to zero.
- (d) The estimated slope would be closer to zero.
- (e) The *t*-statistic for the slope would be closer to zero.

(9) The fitted equation implies that an employee who works a 40 hour week can be expected on average to assemble about
(a) 173 parts.
(b) 36 parts.
(c) 3.2 parts.
(d) 18 parts.
(e) 163 parts.

(10) Accepting the assumptions of the simple regression model, the shown model explains statistically significant variation in *Parts Assembled* because
(a) The F-statistic determined by R^2 is substantially larger than 4.
(b) The p-value of the t-statistic for the slope is less than 0.05.
(c) The 95% confidence interval for the slope excludes zero.
(d) The t-statistic is larger than 2.
(e) All of the above statements are true.

(11) Management believes that the intercept of the regression model in the population is zero. Accepting the assumptions of the SRM, these results indicate that
(a) The population intercept is zero.
(b) The estimated intercept is consistent with management's belief.
(c) The estimated intercept significantly larger than zero.
(d) The estimated intercept significantly smaller than zero.
(e) The intercept is larger than zero for about 13% of employees.

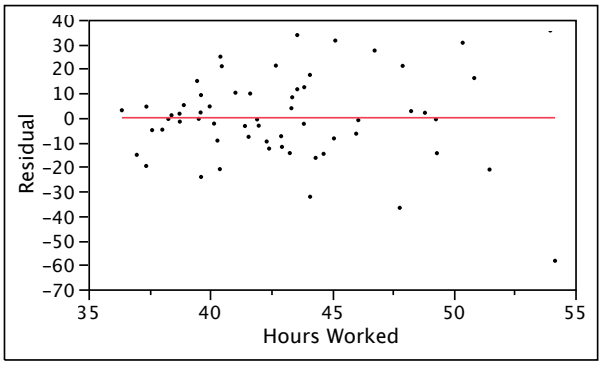
(12) Accepting the assumptions of the SRM, these results indicate that the length of time required to assemble one more additional part is
(a) Significantly more than 15 minutes.
(b) Significantly less than 15 minutes.
(c) More than 15 minutes, but not significantly more.
(d) Less than 15 minutes, but not significantly less.
(e) Not known with adequate precision to address the claim.

(13) If the manufacturer were to extend the standard work week from 40 to 45 hours for every employee, then the number of items produced per week by each employee would be expected on average with 95% confidence (assuming the SRM) to
(a) Increase between 10.6 to 21.3 parts.
(b) Increase between 14.9 to 17.0 parts.
(c) Increase between 15.4 to 16.5 parts.
(d) Stay about the same, with no particular increase or decrease.
(e) Change somewhere between a drop of 19 parts to an increase of 51 parts.

(14) This regression was used to predict the number of parts assembled by a new employee who works a 40-hour week. The employee produced 10 more items than predicted by the fitted model. Accepting the assumptions of the SRM,
(a) The regression is a poor match to this case and should be discarded.
(b) The new employee should be re-trained to match the prediction of the regression.
(c) The new employee is performing significantly worse than his colleagues.
(d) The new employee is performing significantly better than his colleagues.
(e) The new employee is performing comparably to his colleagues.

(15) This fitted regression uses data from 60 employees. If the sample had consisted of 30 observations rather than 60, then we should anticipate that
(a) The confidence interval for the slope would be about twice as long.
(b) The prediction error would increase since the RMSE would be larger.
(c) The standard error of the slope and intercept would be larger.
(d) The model's R^2 would decrease, indicating a worse fit.
(e) All of the other answers to this question are correct.

The following scatterplot shows data from the previous regression model.



(16) The scatterplot shown immediately above suggests that
(a) Outliers have distorted the estimated slope, particularly for long workweeks.
(b) The model requires a transformation to capture a nonlinear trend.
(c) The number of parts assembled becomes more variable with more hours worked.
(d) The data are not normally distributed.
(e) Relatively few employees work more than the standard 40 hours.

(17) It was later learned that rather than represent the performance of 60 employees, these data measure the performance of 30 employees, each observed for two weeks. (Each employee produces two of the points seen in the figures.) This structure of the data implies that
(a) We can continue to rely upon the fitted model and associated inferences.
(b) The fitted data are not normally distributed, violating the SRM.
(c) The fitted data lack constant variance, violating the SRM.
(d) The fitted data are not independent, violating the SRM.
(e) The fitted slope is too large and should be reduced by about half.

(18) Having discovered that the data consist of two measures for each of 30 employees, an analyst built a new data set with 30 rows. Each row in the new data table holds the average hours worked and average parts assembled for each of the 30 employees. In the regression of the average number of parts assembled on the average number of hours worked, we can anticipate that
(a) The slope in the resulting fitted model will be larger than 3.187.
(b) The slope in the resulting fitted model will be smaller than 3.187.
(c) The slope in the resulting fitted model will be near 0.
(d) The RMSE in the resulting fitted model will be about 13.
(e) The R^2 in the resulting fitted model will be about 0.188.

(Questions 19-25) The Human Resources (HR) department of a news agency evaluated the salary paid to 84 full-time reporters. For each reporter, the HR department determined the number of stories that were picked up by the Associated Press and major newspapers (*Num of By-Lines*). HR also obtained the number of years of experience in the news business (either at this news agency or another) and the current annual salary (dollars per year) of each reporter.

Correlations

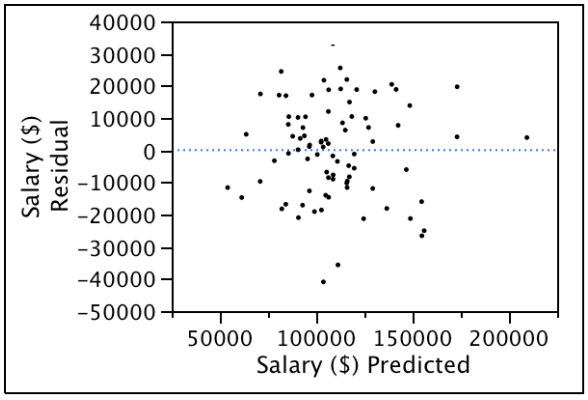
	Salary ($)	Years Exp	By-Lines
Salary ($)	1.0000	0.7812	0.8684
Years of Experience	0.7812	1.0000	0.9021
Num of By-Lines	0.8684	0.9021	1.0000

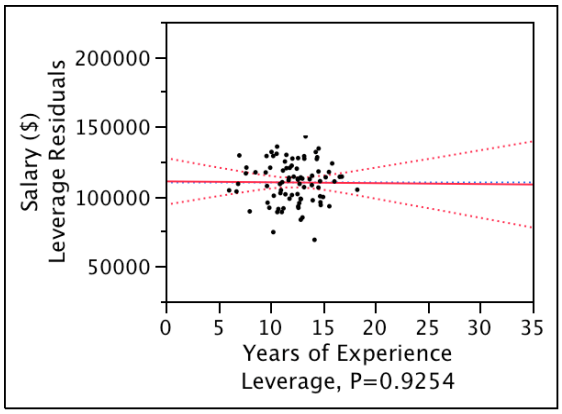
Summary of Regression for Salary

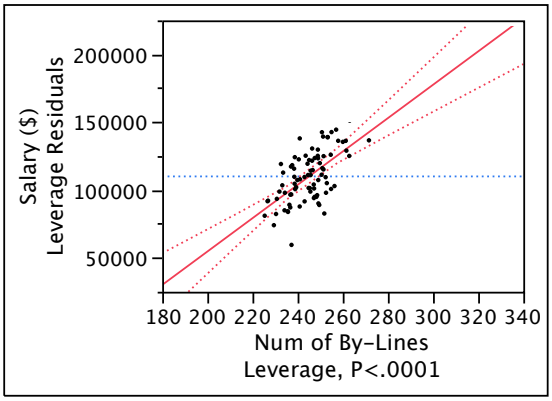
RSquare	0.7541
Root Mean Square Error	15061

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-190487.4	36529.05	-5.21	<.0001
Years of Experience	-63.48647	675.9268	-0.09	0.9254
Num of By-Lines	1229.3339	178.5761	6.88	<.0001







(19) These results show that, in the data used to build this model,
   (a) None of these reporters has more than 20 years experience.
   (b) The largest salary is about $150,000.
   (c) None of these reporters has more than 300 by-lines.
   (d) *Years of Experience* is positively associated with *Salary*.
   (e) The fitted model explains less than half of the variation in *Salary*.

(20) The explanatory variables in the regression (accepting the assumptions of the MRM)
   (a) Together explain statistically significant variation in the salary of reporters.
   (b) Do not explain statistically significant variation in *Salary* due to collinearity.
   (c) Imply that the agency ignores the number of years of experience in setting salary.
   (d) Accurately predict the salary of about ¾ of the data in this sample.
   (e) Require a transformation to a log scale in order to correct for a nonlinear trend.

(21) Which of the following correctly explains the negative intercept in the regression?
   (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 statistically significantly different from zero.
   (d) The intercept is not of interest in regression without categorical variables.
   (e) Reporters owe substantial debts to pay for the cost of traveling.

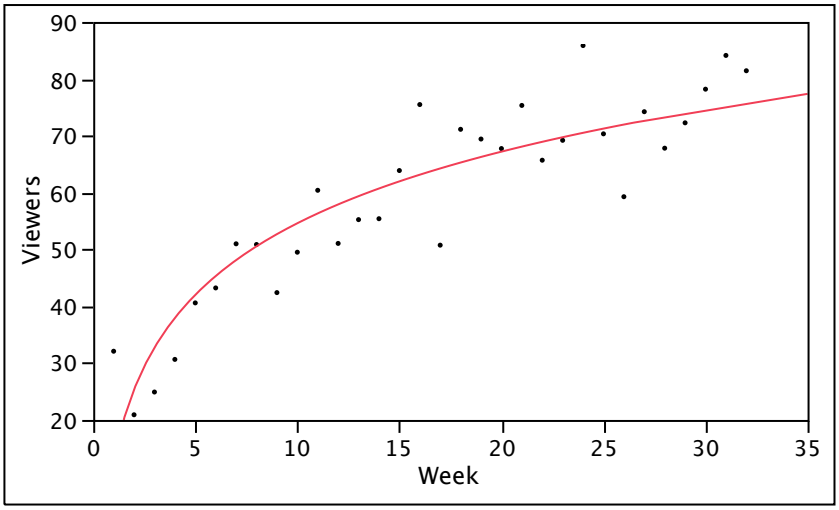
(22) Does adding *Years of Experience* to the simple regression of *Salary* on *Number of By-Lines* significantly improve the predictive ability of the model?
   (a) Yes, because *Years of Experience* has a large correlation with *Salary*.
   (b) No, because *Years of Experience* has a large correlation with *Salary*.
   (c) Yes, because of the size of the *t*-statistic of *Years of Experience* in this regression.
   (d) No, because of the size of the *t*-statistic of *Years of Experience* in this regression.
   (e) Cannot be answered without the R^2 statistic from the simple regression.

(23) Among reporters with 10 years of experience, the fitted model implies that we should expect those with 250 by-lines to differ from those with 240 by-lines in what way?
   (a) Those with 250 by-lines earn about the same salary as those with 240 by-lines.
   (b) Those with 250 by-lines earn about $1,200 more than those with 240 by-lines.
   (c) Those with 250 by-lines earn about $12,000 more than those with 240 by-lines.
   (d) Salaries of reporters with 250 by-lines are more variable those with 240 by-lines.
   (e) Salaries of reporters with 250 by-lines are less variable those with 240 by-lines.

(24) If a reporter at this agency works for another year but produces no stories that earn a by-line credit, then we should expect the salary of this reporter after this year to
   (a) Remain the same.
   (b) Increase, but not by a statistically significant amount.
   (c) Increase, by a statistically significant amount.
   (d) Decrease by $63.49.
   (e) Decrease, by a statistically significant amount.

(25) The plots shown with this model indicate that
   (a) The data are normally distributed.
   (b) Collinearity reduces the precision of the slope estimates.
   (c) A high-leverage point has distorted the slope estimates.
   (d) The data are not normally distributed.
   (e) The observations are not independent.

Questions (26-36) The following data track the number of viewers who watch a network television program. The program was very successful during its first 16-week season and its audience grew dramatically. The success of the program led to it being shown for a second 16-week season that immediately followed the first. The following plot graphs the millions of viewers over this 32-week period. The number of viewers is expressed in millions, and the weeks during which each the episode is shown are consecutively labeled 1, 2, …, 32.



RSquare	0.81
Root Mean Square Error	7.675
Mean of Response	59.067

Parameter Estimates

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

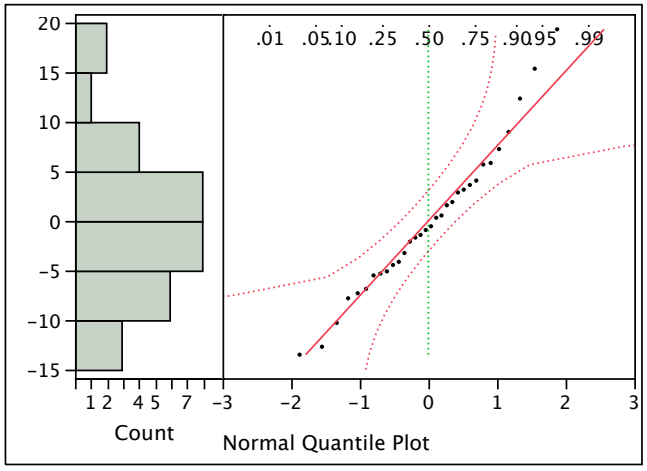
(26) If we were to replace the natural log of week in the shown regression model by the base 10 log of week ($\log_e x \approx 2.3 \log_{10} x$), then how would the fitted model change?
(a) R^2 would increase by a statistically significant amount.
(b) The intercept would be larger by a factor of about 2.3.
(c) The slope would be larger by a factor of about 2.3.
(d) The RMSE would drop by a factor of about 2.3.
(e) The *t*-statistic of the slope would increase by about 2.3.

(27) The estimated intercept in the fitted model implies that
(a) About 13 million watched this network during week 0 (week zero).
(b) The model is a poor fit since the intercept is not significantly different from zero.
(c) The model is a poor fit since the standard error of the intercept is so large.
(d) About 13 million were predicted to watch during the first week of this program.
(e) The number of viewers grows by about 13% per week of the series.

(28) 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) True throughout these 32 weeks.
(b) Not true for any of these weeks.
(c) True only for the first 9 weeks.
(d) True throughout the first 16-week season.
(e) True only during the second 16-week season (weeks 17 through 32).

(29) If the program were to continue into week 33 and the trend represented by this model were to 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

The following plot shows residuals associated with this regression.



(30) The plot of the residuals shown immediately above indicates 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.

A consultant claims that the previous regression is flawed. In its place, she recommends an alternative regression. Her regression, summarized by the following tables, combines the week number with a categorical variable for the season (coded as "First" for the first 16 weeks, and "Second" for weeks 17-32).

Summary of Fit

RSquare	0.857
Root Mean Square Error	6.876

Expanded Estimates

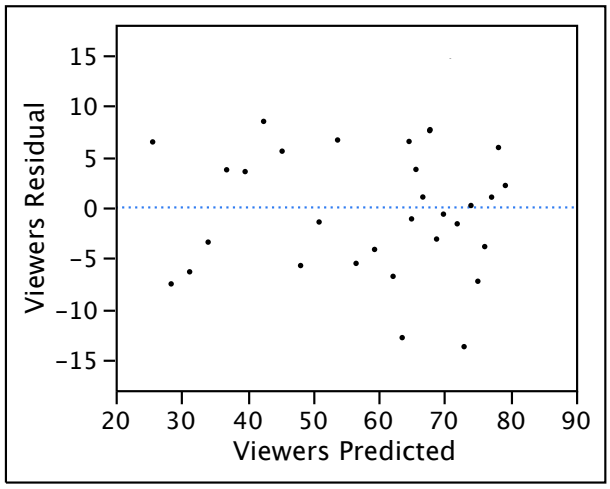
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	34.309012	4.985882	6.88	<.0001
Week	1.9285975	0.263696	7.31	<.0001
Season[First]	-11.48625	4.985882	-2.30	0.0289
Season[Second]	11.486251	4.985882	2.30	0.0289
Season[First]*Week	0.8829581	0.263696	3.35	0.0023
Season[Second]*Week	-0.882958	0.263696	-3.35	0.0023

(31) Based on the multiple regression proposed by the consultant, does the rate of growth of the number of viewers decrease significantly during the second season?
(a) Yes, the coefficient of "Season[Second]*Week" is negative and significant.
(b) Yes, the coefficient of "Season[Second]" is negative and significant.
(c) Yes, the coefficient of "Season[Second]" is negative and not significant.
(d) No, the coefficient of "Season[Second]" is negative.
(e) The added terms do not improve the model and no conclusions are warranted.

(32) The estimated coefficient of Season[First] implies that, given the SRM,
(a) The fitted model estimates a baseline audience of about 22.8 million in week zero.
(b) On average, 11.49 million fewer watch the program in the first season than the second.
(c) On average, 22.98 million fewer watch the program in the first season than the second.
(d) The fitted model estimates a fall in the audience of about 11.5 million in week zero.
(e) The same number of viewers watched the program during the first week of each season.

(33) From the fitted model recommended by the consultant, we would conclude that during the second season the audience for this program
(a) Increased by about 1.9 million on average per week.
(b) Increased by about 0.9 million on average per week.
(c) Decreased by about 0.9 million on average per week.
(d) Increased by about 1 million on average per week.
(e) Increased by about 2.8 million on average per week.

(34) The model recommended by the consultant predicts viewers in Week 33, assuming the shown trends continue, to be about
(a) 98 million.
(b) 80 million.
(c) 112 million.
(d) 85 million.
(e) 57 million.



(35) The plot shown immediately above displays residuals from the multiple regression recommended by the consultant. From this plot, we should 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) This view of the residuals suggests no problems with the model.

(36) Which of the following would *not* be an appropriate next step to the analysis? (*i.e.*, which of the following would be a pretty foolish 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 other programs.
(c) Check for autocorrelation in the residuals.
(d) Remove the confusing interaction from the model.
(e) Confirm the equality of variances of the errors in the first and second seasons.

(Questions 37-45) A company tests applicants for programming positions. Before promoting an applicant to a full-time position, the company requires the applicant to develop a web-based program. The task is rather generic, and programmers can choose to write the software in whatever programming language they want. Among the data for this question, the languages used are "C", "Java" and "Other." The response variable in the following model is the amount of time (in seconds) required to render a web page when requested by a web browser. Smaller times are better. Other variables in the model are

Age	of the applicant, in years,
Lines of code	the length of the submitted program
Sex	of the applicant, (male, female), and
Coding time	hours required to prepare the program that was submitted.

Summary of Fit for Amount of Time

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	Prob > F
C. Total	79	3600.9191		<.0001

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Language	2	2	202.57222	4.9813	0.0095
Coding Time	1	1	607.47038	29.8757	<.0001
Sex	1	1	2.57169	0.1265	0.7232
Age	1	1	9.11867	0.4485	0.5052
Lines Of Code	1	1	190.54352	9.3710	0.0031
Language*Lines Of Code	2	2	376.87164	9.2674	0.0003

Expanded Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	4.4687232	2.217219	2.02	0.0476
Language[C]	-5.837148	2.495178	-2.34	0.0221
Language[Java]	6.202472	1.966131	3.15	0.0024
Language[Other]	-0.365324	1.743467	-0.21	0.8346
Coding Time	0.3666612	0.067082	5.47	<.0001
Sex[Female]	0.227025	0.638364	0.36	0.7232
Sex[Male]	-0.227025	0.638364	-0.36	0.7232
Age	-0.026307	0.039284	-0.67	0.5052
Lines Of Code	0.0229277	0.00749	3.06	0.0031
Language[C]*Lines Of Code	-0.011623	0.009721	-1.20	0.2358
Language[Java]*Lines Of Code	-0.029177	0.007956	-3.67	0.0005
Language[Other]*Lines Of Code	0.0408004	0.010796	3.78	0.0003

(37) Does the shown regression model explain statistically significant variation in the amount of time needed to render 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 statistically significant.
(d) Yes, the R^2 statistic is larger than 50%.
(e) No, the RMSE is too large to have explained significant variation.

(38) Based on these results, would a randomly selected male applicant be expected to write a program that runs faster than a randomly selected female applicant?
(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) The shown results do not provide an answer to this question.

(39) It has been claimed that the business should strive to have programmers write the programs with few lines of code 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 "Other".
(d) The claim is not true for programs written in "Java".
(e) The claim is not true for programs written in "C" or "Java".

(40) Assume that a 40-year-old male programmer writes a 200-line program in 4 hours. Which language is most likely to provide the fastest solution, based on this model?
(a) C
(b) Java
(c) Other
(d) It does not matter which language is used.
(e) The fitted model does not address this question.

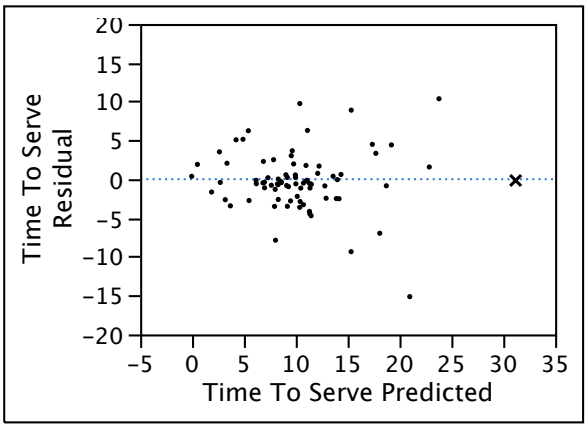
(41) Assume this is the initial fitted model. Based on the summary of the fitted model, a logical next step in the development of this model would be to remove
(a) The categorical variable indicating sex of the programmer.
(b) Both age and the categorical variable indicating sex of the programmer.
(c) *Lines Of Code* since its coefficient is small.
(d) *Language[Other]*.
(e) All of the terms whose associated p-value is less than 0.05.

(42) The calculation of the *p*-value for *Sex*[Female]
(a) Implies that about 72% of the data in this sample 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.

(43) If the variable *Coding Time* were removed from the fitted model, then we can be sure that (assuming the MRM)
(a) The R^2 of the resulting model would be statistically significantly smaller than 0.60.
(b) The intercept of the resulting model would be larger than 4.469.
(c) The intercept of the resulting model would be smaller than 4.469.
(d) The RMSE of the resulting model would be significantly smaller than 4.5.
(e) The residuals of the resulting model would not be normally distributed.

(44) In order to check that the fitted model predicts program time for male applicants as well as it predicts program time for female applicants, we need to
(a) Inspect the normal quantile plot of the residuals from the model.
(b) Inspect the leverage plot for *Sex*.
(c) Verify the sums of squares reported in the Analysis of Variance.
(d) Compute the *p*-value of the partial *F*-test for *Sex*.
(e) Inspect side-by-side boxplots of residuals grouped by *Sex*.

The following plot shows data associated with this regression.



(45) In the diagnostic plot shown immediately above, the observation highlighted with an "x" in the plot
(a) Represents a case that the fitted model under-predicts.
(b) Represents a case that the fitted model over-predicts.
(c) Is accurately predicted by the fitted model.
(d) Suggests that the fitted model does not produce normally distributed residuals.
(e) Indicates an applicant whose program runs very quickly.