Statistics 608-621 Waiver Exam

August 20, 2007

On the answer sheet *before* the exam begins …

- **Write in your name and Penn student id number**.
- **Mark the "bubbles"** under the letters of your name *and* student id number on the form. Failure to do so will lead to a score of zero.
- Use a **#2 pencil**. Erase any changes 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**, but no laptops or computers are allowed. You may not use a cell phone during the exam for any purpose.

Turn in the solution page only; keep the 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.

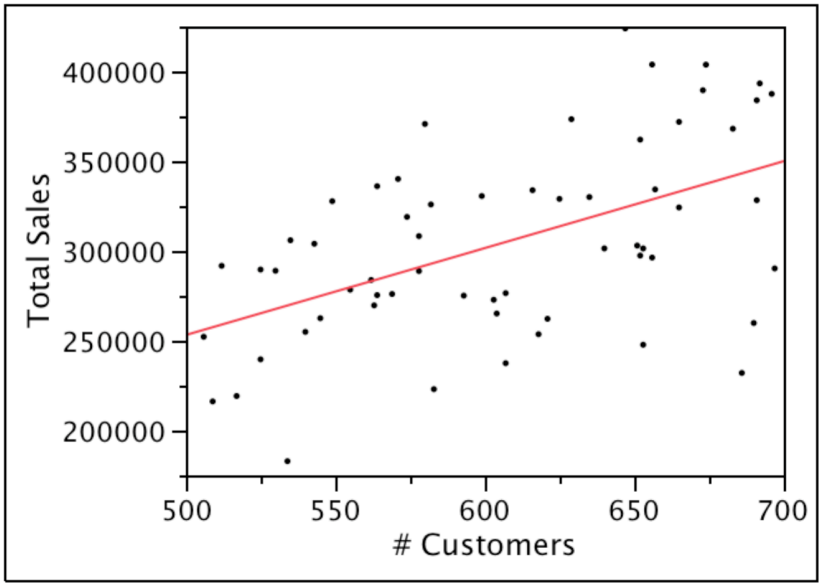
Your **score** is the number of correct answers. The multiple-choice questions are equally weighted. Some questions may be dropped and not counted as part of the overall score. There is no deduction for incorrect answers.

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 grade 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.*

(Questions 1–12) A chain of retail stores that sell custom home furnishings collected data from various outlets. Each observation denotes the number of customers who made a purchase in a given month (*# Customers*), and the total level of sales (*Total Sales*, in dollars) made to these customers. This analysis uses data from 60 retail stores for May 2006.



RSquare	0.27
Root Mean Square Error	45892
Observations	60

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	10876	63323	0.17	0.86
# Customers	484	104	4.66	<.0001

(1) To cover costs, a store needs to generate $205,000 in sales per month. Based on the fitted model, how many customers are required, on average, to achieve this level of sales?
(a) About 400 customers.
(b) About 451 customers.
(c) About 484 customers.
(d) About 500 customers.
(e) About 600 customers.

(2) Based on this regression analysis and the standard assumptions of the SRM, if a store has 500 customers in a month, what is the probability of it not breaking even (*i.e.*, having less than $205,000 in sales)?
(a) About $1/6$.
(b) About $1/3$.
(c) About $1/2$.
(d) About $2/3$.
(e) About $5/6$.

(3) The slope of the fitted model implies that, on average,
(a) Customers spend about $484 on purchases.
(b) A store makes a sale to about 1 customer out of every 484 who visit.
(c) 484 customers visit a store in this chain during a month.
(d) Sales increase by $484 for each additional purchasing customer.
(e) The number of customers who visit stores of this chain does not affect total sales.

(4) The intercept of the fitted model
(a) Implies that on average stores make sales of $10,876 when closed.
(b) Is consistent with the claim that sales are proportional to the number of customers.
(c) Implies that these data are a sample from a typical population.
(d) Represents the effect of an influential outlier and is not reliable.
(e) Implies that on average $10,876 of sales each month goes to taxes.

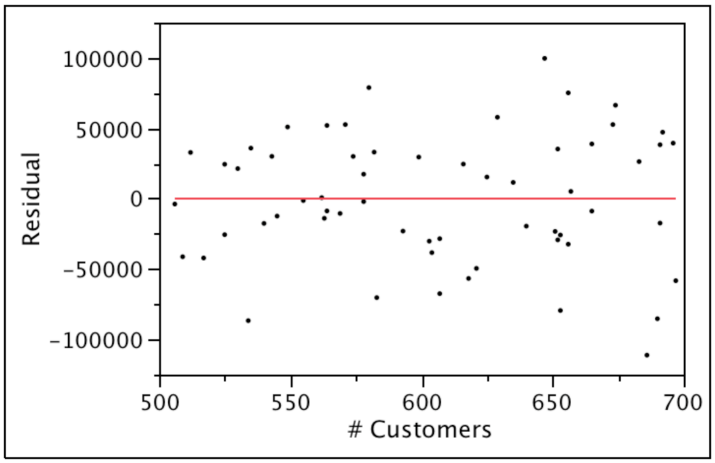
(5) Financial analysts expect retail stores of this type to generate more that $500 in sales per additional purchasing customer. Based on the shown results and the assumptions of the SRM, this model implies that stores in this retail chain generate
(a) Statistically significantly less sales per additional purchasing customer.
(b) Statistically significantly more sales per additional purchasing customer.
(c) More sales per additional purchasing customer, but not by a statistically significant margin.
(d) Less sales per additional purchasing customer, but not by a statistically significant margin.
(e) $500 in sales per additional purchasing customer.

(6) A new promotion has been designed to attract additional customers to visit stores in the chain and make purchases. If 100 more customers make a purchase at a store in this chain, then the fit of this model indicates that on average sales would
(a) Not be increased by a statistically significant amount of dollars.
(b) Increase by 48.4% above sales at the prior level.
(c) Increase by 27% above sales at the prior level.
(d) Increase from $27600 to $69200, with 95% confidence.
(e) Increase from $48192 to $48608, with 95% confidence.

(7) Retail sales tax is 6% in the region where these stores are located. The fit of this model, assuming the SRM, implies that the regression of the variable *Sales Tax* (measured in dollars, computed as 0.06 times the response in this regression) on # *Customers* would have
(a) A smaller value of the R^2 statistic.
(b) A larger value of the R^2 statistic.
(c) A larger estimated slope.
(d) A larger estimated intercept.
(e) A smaller estimated RMSE.

(8) The *RMSE* for the shown model, assuming the SRM, implies that the fitted model
(a) Accurately predicts sales at about ¼ of these stores.
(b) Predicts sales to within about $92,000 for 95% of these stores.
(c) Predicts sales to within about $46,000 for almost all of these stores.
(d) Predicts sales to within about $46,000 for about ½ of these stores.
(e) Generates predictions that increase by about $92,000 over the range of x.

(9) Had the fit of this model been based on a larger sample of 120 stores that are similar to these 60, then we should expect that the
(a) RMSE of the model would have been substantially less than 45,892.
(b) R^2 of the fitted model would be substantially larger than 0.27.
(c) *P*-value of the intercept would have been substantially closer to one.
(d) Slope of the fitted model would have been substantially larger than 484.
(e) Standard error of the slope would be substantially smaller than 104.



(10) The plot of the residuals from the fitted model shown above indicates that
(a) The underlying model errors are not normally distributed.
(b) The fitted model violates the assumption of equal-variance errors.
(c) The underlying model errors are not independent.
(d) The data are clustered and not appropriate for this type of model.
(e) The data appear to be consistent with the SRM.

(11) In order to verify the assumption of normality for the shown simple regression, the best approach is to
(a) Inspect the normal quantile plot of *Total Sales*.
(b) Inspect the normal quantile plot of the number of customers.
(c) Inspect the normal quantile plot of the residuals.
(d) Plot the residuals in groups defined by several sizes of the explanatory variable.
(e) Find the value of the Durbin-Watson statistic.

(12) Which of the following actions should be expected to improve the precision of the estimated slope in the regression model the most?
(a) Add 10 more stores with 500 customers.
(b) Add 10 more stores with 600 customers.
(c) Add 10 more stores with 700 customers.
(d) Add 10 more stores, ½ with 500 and ½ with 700 customers.
(e) Add 10 more stores with *Sales* larger than $375,000.

(Questions 13-24) An analysis was done to model annual profits of 32 companies in the pharmaceutical and computer industries. The explanatory variables were the assets of the company, the sales of the company, and a categorical variable that indicates the industry (*Type* = computer or *Type* = phamaceutical). The variables sales, assets, and profits are all measured in millions of dollars (*Assets* = 1 implies $1,000,000). Questions follow the summary of the fitted model.

Correlations

	Profits ($M)	Sales ($M)	Assets($M)
Profits ($M)	1.0000	0.8745	0.8301
Sales ($M)	0.8745	1.0000	0.9846
Assets ($M)	0.8301	0.9846	1.0000

Summary of Fit: Response = Profits ($M)

RSquare	0.875445
Root Mean Square Error	289.7584
Observations	32

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	16523407	5507802	65.6004
Error	28	2350878	83960	Prob > F
C. Total	31	18874286		<.0001

Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	98.825083	60.72692	1.63	0.1149
Sales ($M)	0.1154404	0.027432	4.21	0.0002
Assets ($M)	-0.043795	0.022443	-1.95	0.0611
Type[Computer]	-226.2494	53.59964	-4.22	0.0002
Type[Pharmaceutical]	226.24935	53.59964	4.22	0.0002

(13) Do companies with larger assets average more profits than companies with smaller assets?
(a) No, because the slope of *Assets* is negative in this multiple regression.
(b) No, because the sample size is too small to make such an inference.
(c) Yes, because the *p*-value of *Assets* is positive.
(d) Yes, because the slope in the simple regression of *Profit* on *Assets* is positive.
(e) Cannot be judged from the shown output.

(14) Based on the numerical summary of the multiple regression shown above, has the model using these variables explained significant variation in the level of profits?
(a) Yes, because the R^2 statistic is close to 1.
(b) No, because the model contains an insignificant explanatory variable.
(c) No, because the residual standard deviation is too large to allow useful prediction.
(d) Yes, because the *p*-value of the overall *F* statistic is quite small.
(e) No, because the fit relies on too little data to judge the level of significance.

(15) Had all three variables *Profit*, *Sales* and *Assets* been measured in units of dollars rather than in millions of dollars, then
(a) The slope for *Assets* would have been larger.
(b) The slope for *Assets* would have been the same.
(c) The slope for *Assets* would have been smaller.
(d) The *t*-statistic for *Assets* would have been larger.
(e) None of the above would have occurred.

(16) Based on the fitted model, the predicted profits for a computer company with $4 billion (4000 million) in sales and $2 billion in assets is approximately
(a) $250 million
(b) – $250 million (*i.e.*, a loss)
(c) $750 million
(d) $150 million
(e) $500 million

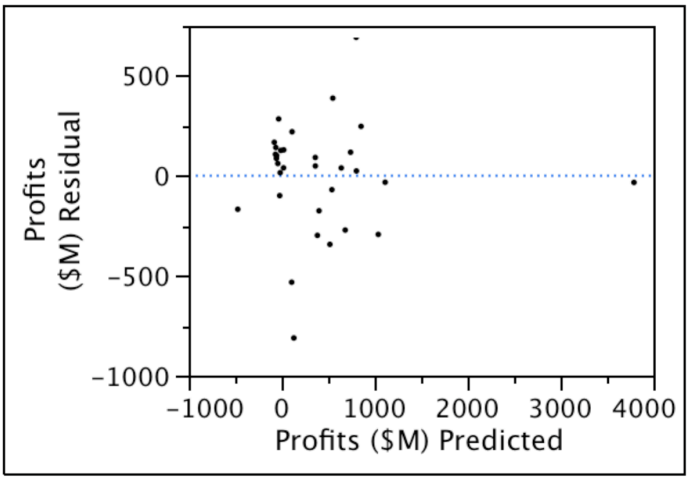
(17) An industry analyst interpreted this model as meaning that the effect of sales on profits is the same in the computer and pharmaceutical industries for companies with comparable assets. Is this an appropriate interpretation?
(a) No, because the coefficient of *Type* is significantly different from zero.
(b) No, because the model forces a common slope of *Sales* for both industries.
(c) Yes, because the *Type* factor is not statistically significant.
(d) Yes, because the estimated interaction is so close to zero.
(e) Yes, because the intercept is not significantly different from zero.

(18) The estimated coefficient for *Assets* is negative. The *best* explanation for the sign of *Assets* is that
(a) The correlation between *Assets* and *Profits* is negative.
(b) The correlation between *Assets* and *Sales* is nearly 1.
(c) The true coefficient for this variable in the underlying population is zero.
(d) The model as a whole does not explain significant variation in profits.
(e) The observations in this sample violate the assumption of independence.

(19) The estimated coefficient of *Type*[*computer*] is best interpreted as meaning that
(a) Firms in the computer industry average about $226 million less profits than those in the pharmaceutical industry.
(b) Firms in the computer industry average about $452 million less profits than those in the pharmaceutical industry.
(c) Firms in the computer industry average about $226 million less profits than those with the same sales and assets in the pharmaceutical industry.
(d) Firms in the computer industry average about $452 million less profits than those with the same sales and assets in the pharmaceutical industry.
(e) The sample is too small to estimate the difference between these two industries.

(20) The *p*-value of the intercept in this model implies, given the assumptions of the MRM and the truth of the null hypothesis H_0: $\beta_0 = 0$, that the
(a) Population intercept is 0.
(b) Probability that the population intercept is 0 is 0.1149.
(c) Probability that the population intercept is not 0 is 0.1149.
(d) Probability of an estimated intercept so far (or farther) from 0 is 0.1149.
(e) Probability that the fitted model would improve if this term is removed is 0.1149.

(21) The fitted model implies that among companies in the computer industry with $4 billion in assets, those with $2 billion in sales have profits that average about
(a) $88 million to $143 million more than companies with sales of $1 billion.
(b) $60 million to $170 million more than companies with sales of $1 billion.
(c) $115.4 million to $115.5 million more than companies with sales of $1 billion.
(d)–$464.4 million to $695.5 million more than companies with sales of $1 billion.
(e) The same level of profits as companies with sales of $1 billion.

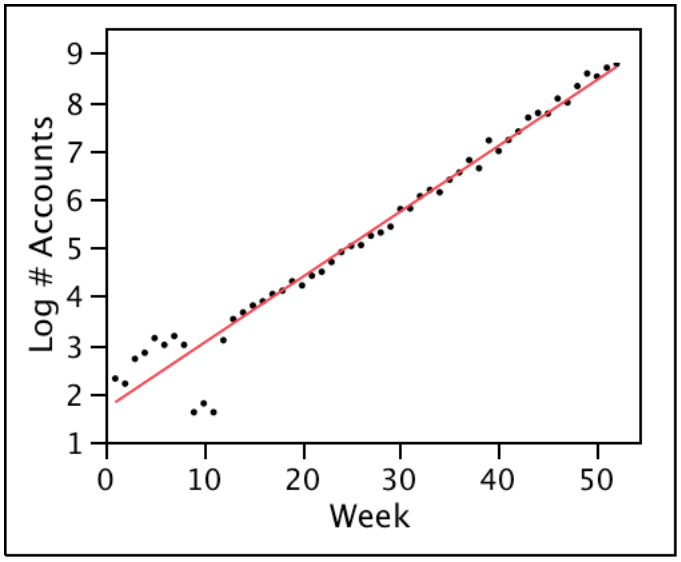


(22) The plot of the residuals shown above indicates that
(a) The model requires a transformation to obtain a more linear fit.
(b) The model has omitted an important underlying explanatory variable.
(c) The underlying model errors violate the assumption of independence.
(d) The underlying model errors violate the assumption of normality.
(e) The fitted model is consistent with assumptions of the MRM.

(23) If this analysis were recomputed *without* the observation with the largest predicted price, then we should expect that the
(a) RMSE of the model would decrease.
(b) The slope for *Assets* would be closer to 0.
(c) The slope for *Sales* would be farther from 0.
(d) The R^2 statistic would decrease.
(e) All of the above.

(24) A claim has been made that companies in the pharmaceutical industry retain a higher amount of sales as profits than companies of comparable assets in the computer industry. To perform a hypothesis test of this claim, an analyst should
(a) Compute the confidence interval for the estimated slope of *Sales*.
(b) Compute a confidence interval for twice the estimated coefficient of *Type*[computer].
(c) Add an interaction variable between *Sales* and *Type* to the model.
(d) Remove outliers that adversely distort the fitted model.
(e) Reformulate the explanatory variables in the model to reduce the collinearity.

(Questions 25-31) A startup network communication company would like to forecast the future level of customer activity. This company provides high-speed network connections for business customers. The data used in this analysis describe the most recent 52 weeks of activity. Near week 10 of these data, the company experienced a variety of system failures associated with problems caused by lightning striking the building that houses its servers. The response is the natural logarithm of the number of user accounts served each week.



RSquare	0.961913
Root Mean Square Error	0.412292
Mean of Response	5.266599
Observations (or Sum Wgts)	52

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	1.6792054	0.116019	14.47	<.0001
Week	0.1353734	0.00381	35.54	<.0001

Durbin-Watson

Durbin-Watson	AutoCorrelation	Prob<DW
0.6602356	0.6557	0.0000

(25) The fitted model suggests that the trend in the number of accounts is:
(a) A steady, upward linear trend with a few early outliers.
(b) Increasing exponentially, with a few early outliers.
(c) Growing linearly, but with substantial variation about the trend.
(d) Not well captured by the fitted model.
(e) Decreasing since the log is growing so slowly over time.

(26) The Durbin-Watson statistic indicates that
(a) There is a strong, statistically significant trend in the number of calls.
(b) A pattern in the residuals indicates the errors are not independent.
(c) The log transformation is incorrect; a reciprocal should have been used.
(d) The underlying model errors violate the assumption of normality.
(e) The fitted regression is consistent with the assumptions of the SRM.

(27) The autocorrelation shown in the output estimates the
(a) Correlation between *Week* and the number of accounts.
(b) Correlation between *Week* and the log of the number of accounts.
(c) Elasticity of the number of accounts with respect to week changes.
(d) Correlation between the error at week t and at week t+1.
(e) Correlation between the log of the number of accounts in adjacent weeks.

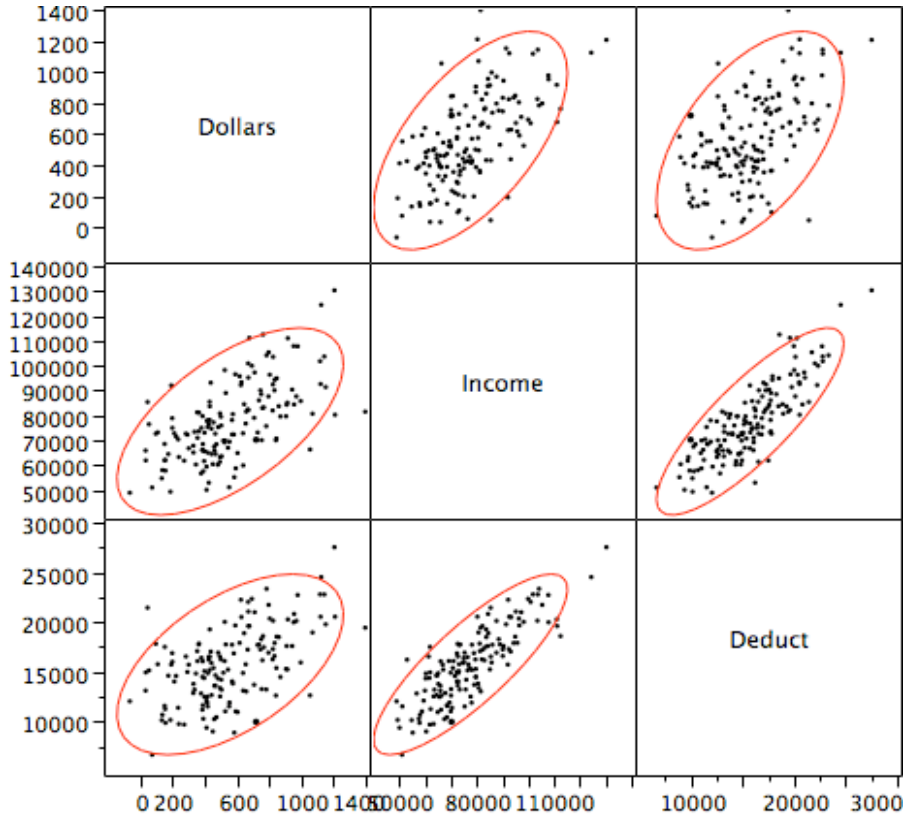
(28) The slope of the estimated regression equation indicates that
(a) The number of accounts is growing at about 14% per week.
(b) The number of accounts is growing at a rate of about 140 accounts per week.
(c) The number of accounts is nearly constant at about 1,679.
(d) The elasticity of the number of accounts with respect to week is about 0.14.
(e) There is not a statistically significant upward trend in the number of accounts.

(29) Based on this model, the predicted number of accounts expected during Week 53 (the week immediately following this period) is approximately
(a) 1,679 accounts.
(b) 7,002 accounts.
(c) 8,854 accounts.
(d) 714,489 accounts.
(e) More than 1,000,000 accounts.

(30) A 95% prediction interval was computed to go with the forecast for Week 53 (see the prior question). Based on the fitted model and the assumptions of the SRM, the length of this interval (that is, the upper limit minus the lower limit) is approximately
(a) 5.2 accounts.
(b) 1.65 accounts.
(c) 12,900 accounts.
(d) 0.82 accounts.
(e) 466 million accounts.

(31) If the analysis were to exclude the data for the weeks 1-15 before and during the lightning strike, then we can be sure that the revised analysis would produce a
(a) Wider prediction interval because over ¼ of the data would have been excluded.
(b) Smaller estimated slope for the explanatory variable.
(c) Larger estimated slope for the explanatory variable.
(d) More narrow 95% prediction interval for the number of accounts in week 53.
(e) Fitted model with a smaller value of R^2 since the outliers had been excluded.

(Questions 32-44) The U.S. Internal Revenue Service occasionally examines a random sample of personal income tax submissions for auditing, as suggested by the regression analysis that follows. The dependent variable *Dollars* in this analysis is the dollar amount recovered from the taxpayer as a result of the audit; this variable is negative if the audit indicates that the taxpayer overpaid. The explanatory variables in the analysis are the gross income of the taxpayer (*Income*, in dollars) and the itemized deductions (*Deduct*, in dollars). In addition, the regression model includes a categorical variable (*PreparedBy*) that indicates if the form was prepared by the taxpayer ("Payer"), a certified public accountant ("CPA"), or a tax-preparation service ("Service"). Further questions follow summaries of *two* fitted models on the next page. The second model adds several variables to the first model.



(32) Based upon the shown output, the average amount of unpaid taxes uncovered in these audits in these data is approximately
(a) Less than $100
(b) $200
(c) $600
(d) $1000
(e) Cannot be identified from the shown output.

(33) Based upon the shown output, the largest correlation is between the variables
(a) *Dollars* and *Income*.
(b) *Dollars* and *Deduct*.
(c) *Income* and *Deduct*.
(d) *PreparedBy* and *Dollars*.
(e) Cannot be identified from the shown output.

Summary of Fit, Model #1

RSquare	0.616832
Root Mean Square Error	170.7837

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	6808305	1702076	58.3561
Error	145	4229226	29167	**Prob > F**
C. Total	149	11037531		<.0001

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Income	1	1	149801.0	5.1360	0.0249
Deduct	1	1	142768.0	4.8948	0.0285
PreparedBy	2	2	3067978.8	52.5932	<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	124.04821	82.99104	1.49	0.1372
Income	0.0036853	0.001626	2.27	0.0249
Deduct	0.0157184	0.007105	2.21	0.0285
PreparedBy[CPA]	31.921774	24.82197	1.29	0.2005
PreparedBy[Payer]	156.85313	21.60189	7.26	<.0001
PreparedBy[Service]	-188.7749	20.65401	-9.14	<.0001

Summary of Fit, Model #2

RSquare	0.626981
Root Mean Square Error	170.8803

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	119.9538	92.99616	1.29	0.1992
Income	0.0050454	0.001809	2.79	0.0060
Deduct	0.0092495	0.007941	1.16	0.2461
PreparedBy[CPA]	19.600661	154.7862	0.13	0.8994
PreparedBy[Payer]	175.92176	121.7484	1.44	0.1507
PreparedBy[Service]	-195.5224	114.4913	-1.71	0.0899
PreparedBy[CPA]*Income	0.0048374	0.002976	1.63	0.1062
PreparedBy[Payer]*Income	-0.001537	0.002361	-0.65	0.5163
PreparedBy[Service]*Income	-0.003301	0.002279	-1.45	0.1497
PreparedBy[CPA]*Deduct	-0.02317	0.011117	-2.08	0.0371
PreparedBy[Payer]*Deduct	0.0064142	0.010279	0.62	0.5336
PreparedBy[Service]*Deduct	0.0167557	0.010033	1.67	0.0971

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Income	1	1	227264.02	7.7830	0.0060
Deduct	1	1	39614.60	1.3567	0.2461
PreparedBy	2	2	128195.86	2.1951	0.1151
PreparedBy*Income	2	2	86141.88	1.4750	0.2323
PreparedBy*Deduct	2	2	105845.96	1.8124	0.1670

(34) The overall *F*-statistic of Model #1 indicates that
(a) Higher incomes and deductions produce statistically significantly higher audit amounts.
(b) At least one slope has a *t*-statistic that is larger than 2.
(c) At least one *p*-value must be less than 0.05.
(d) The R^2 is significantly larger than expected by chance.
(e) The fitted model suffers from substantial amounts of collinearity.

(35) If we were to remove the explanatory variable *Deduct* from Model #1, then we can be sure from the output that
(a) The estimated slope for *Income* would remain the same.
(b) The R^2 of the model would decrease by a statistically significant amount.
(c) The differences implied by coefficients of *PreparedBy* would become smaller.
(d) The estimated slope for *Income* would no longer be statistically significant.
(e) The resulting fitted model would suffer from worse collinearity.

(36) Based on Model #1, for taxpayers with $100,000 in income and $20,000 in deductions, should auditors expect different amounts of recovered taxes, on average, among the three types of preparation methods?
(a) Yes, because this model explains statistically significant variation in *Dollars*.
(b) Yes, because the estimated coefficient for *PreparedBy*[Payer] is statistically significant.
(c) Yes, because the partial effect test for *PreparedBy* is statistically significant.
(d) No, because the estimated coefficient for *PreparedBy*[CPA] is not statistically significant.
(e) No, because this output does not show the average taxes paid by the 3 groups.

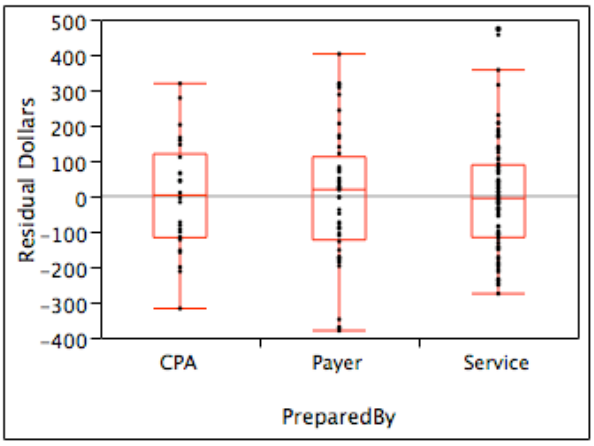
(37) Two groups of returns were audited. Both were prepared by CPAs and included taxpayers with income of about $125,000. One group of returns claimed deductions that average $20,000 whereas the other claimed deductions of $30,000. Based on Model #1 and the assumptions of the MRM, compared to those with lower deductions, audits of the tax returns with $30,000 in deductions yield on average
(a) About the same recovered amounts as those claiming $20,000 in deductions.
(b) About $15 to $300 more in recovered taxes than those claiming $20,000 in deductions.
(c) Precisely $157 more in recovered taxes than those claiming $20,000 in deductions.
(d) About $32 more in recovered taxes than those claiming $20,000 in deductions.
(e) About $157 more in recovered taxes than those claiming $20,000 in deductions.

(38) Based on Model #2, for payer-prepared tax returns that show $100,000 in income and $20,000 in deductions, an auditor can expect to recover about
(a) nothing (that is, $0)
(b) $176
(c) $300
(d) $960
(e) $985

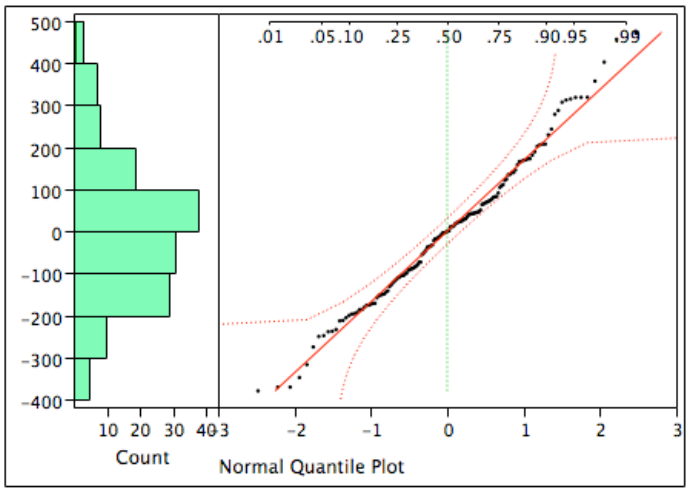
(39) The comparison of Model #1 to Model #2 shows that
(a) The effect of *Income* depends statistically significantly on the type of preparation.
(b) The effect of *Deduct* depends statistically significantly on the type of preparation.
(c) Model #2 has statistically significantly less collinearity than Model #1.
(d) There is no evidence of statistically significant interaction.
(e) Model #2 explains significantly more variation than Model #1.

(40) The explanatory variable *Deduct* is statistically significant in Model #1 but not in Model #2. The *best* explanation for this change in statistical significance is
(a) *Deduct* was really not so useful in the first place in Model #1.
(b) Both models omit too many explanatory variables to be credible.
(c) Outliers conceal an important difference between these two models.
(d) The MRM is not well-suited to these data.
(e) Interactions added to form Model #2 produce substantial collinearity.

(41) In order to assess the impact of possible outliers on the slope of *Deduct* in Model #1, in addition to the shown plots, we should specifically look at a
(a) Comparison (side-by-side) boxplot of the residuals grouped by preparation method.
(b) Normal quantile plot of the residuals.
(c) Leverage plot for *Deduct*.
(d) Scatterplot of the residuals on *Deduct*.
(e) Scatterplot of the residuals on the predicted values.



(42) The plot of the residuals from Model #1 shown immediately above indicates that
(a) These data violate the assumption of independence.
(b) The model requires a transformation to normality.
(c) We should fit separate models for each of these three subsets of these data.
(d) The underlying model errors have comparable levels of variation in these groups.
(e) The underlying model errors in some groups are not normally distributed.



(43) The plot of the residuals from Model #1 shown immediately above indicates that

(a) The fitted model explains statistically significant variation in dollars recovered.
(b) It is appropriate to use a linear equation in the regression model.
(c) The underlying model errors are dependent.
(d) The underlying model errors meet the assumption of equal variation.
(e) The underlying model errors meet the assumption of a normal distribution

(44) In a review of this analysis, it was discovered that audits for all of the taxpayers who submitted returns prepared by a CPA were performed by Auditor A, those prepared by the taxpayers were audited by Auditor B, and the rest audited by Auditor C. The categorical variable *Auditor* records which auditor performed the audit ("A", "B" or "C"). This discovery suggests that

(a) The effects of preparation method are confounded with the auditor.
(b) We should add the categorical explanatory variable *Auditor* to Model #1.
(c) We should add the categorical explanatory variable *Auditor* to Model #2.
(d) We should check for consistency of the residuals when grouped by *Auditor*.
(e) The analysis requires more cases for each type of preparation method.