Statistics 608-621 Waiver Exam
August 23, 2005

On the answer sheet *before* the exam begins …

- Use a **#2 pencil**. Erase any changes completely.
- **Fill 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.

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 textbook and 2 pages of notes** during the exam. No other reference materials are permitted.
- You may use a **calculator**, but no laptops or computers are allowed.

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.

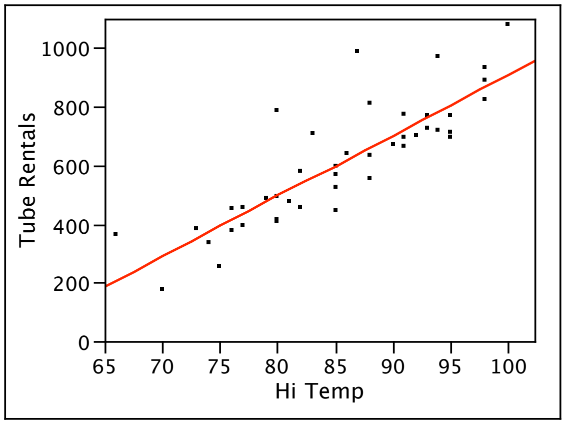
Your **score** is the number of correct answers. The multiple-choice questions are equally weighted. Some questions may be dropped and not counted as part of the overall score. There is no deduction for incorrect answers.

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 marked answers on the grade form will be considered. You can use the "My Grades" feature to find your score as determined by your answers on the answer form.

STOP

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

(Questions 1–11). A small company offers tubing trips on the Delaware River. For $10, the company provides an inflatable raft (or doughnut-shaped tube) and a ride to a destination up-river. Customers are dropped off and then float downstream with the current. The company operates during the summer months and collected data on the number of tubes rented daily during the week (Mon-Fri, so long as *not* a holiday) for the period June through August. The 45 days used in this analysis exclude weekdays that were rained out. The one predictor is the high temperature of the day, in degrees Fahrenheit. (If x is the temperature in Fahrenheit, then $5(x\text{-}32)/9$ gives the temperature in Celsius.)



RSquare	0.721
Root Mean Square Error	107.847
Mean of Response	618.060
Observations	45

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1153.46	168.62	-6.84	<.0001
Hi Temp	20.65	1.96	10.55	<.0001

(1) According to the fitted model, when compared to a 90 degree day, a day with high temperature 95 degrees would average about
- (a) 100 more tube rentals than a day with high temperature 90 degrees.
- (b) 20 more tube rentals than a day with high temperature 90 degrees.
- (c) 10 more tube rentals than a day with high temperature 90 degrees.
- (d) 5 more tube rentals than a day with high temperature 90 degrees.
- (e) the same number of tube rentals as a day with high temperature 90 degrees.

(2) The fitted model suggests that on a day with high temperature 80 degrees the company should expect, on average,
- (a) more than 1000 tube rentals.
- (b) about 650 tube rentals.
- (c) about 500 tube rentals.
- (d) about 350 tube rentals.
- (e) fewer than 300 tube rentals.

(3) The company used this model to predict the tube rentals. The model predicted 800 rentals, based on the high temperature. If the company has 900 tubes available this day, then this model along with the usual regression assumptions implies that there is about a
(a) 50% chance that it will have enough tubes to meet demand.
(b) 67% chance that it will have enough tubes to meet demand.
(c) 84% chance that it will have enough tubes to meet demand.
(d) 95% chance that it will have enough tubes to meet demand.
(e) 99% chance that it will have enough tubes to meet demand.

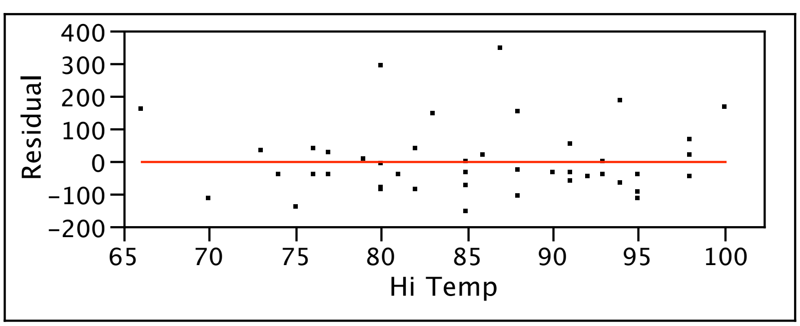
(4) The intercept in the fitted model
(a) implies that the company will lose money if it tries to operate in cold weather.
(b) should be removed to improve the form of the model.
(c) implies that these data violate the assumption of independence.
(d) implies that these data come from a population with zero for the intercept.
(e) is a gross extrapolation and is not reliable as a predicted value.

(5) The fit of this model implies that, on average with 95% confidence, that sales in dollars on an 81 F^{o} day exceed those on an 80 F^{o} day by
(a) $20 to $25.
(b) $0 to $425.
(c) $203 to $211.
(d) $167 to $246.
(e) Cannot be determined from this output without knowing the temperatures.

(6) Were the temperature data recorded in degrees Celsius rather than Fahrenheit, then the slope would have been estimated to be ($C^{o} = 5(F^{o}-32)/9$)
(a) 52.65
(b) 37.17
(c) 11.47
(d) 6.299
(e) 20.65

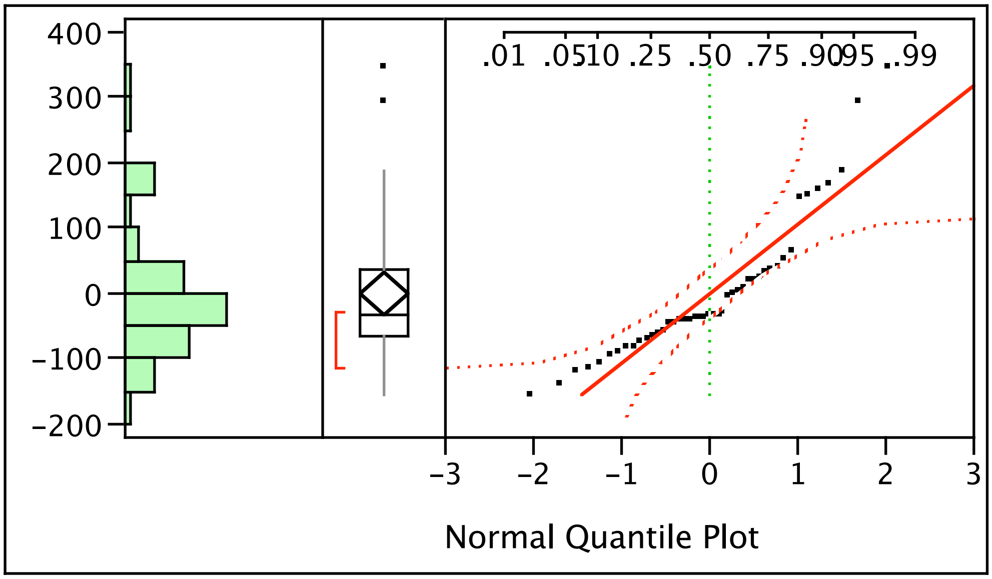
(7) The correlation between temperature (in degrees Celsius) and gross sales (at $10 per rental) is
(a) 0.21
(b) 0.52
(c) 0.72
(d) 0.85
(e) Not available from the shown output.

(8) The following plot of the residuals from the fitted model indicates that



(a) The specified model has missed a nonlinear pattern.
(b) The underlying errors are autocorrelated.
(c) The underlying errors lack constant variance.
(d) The model tends to under-predict rentals, on average.
(e) Tube rentals are occasionally more numerous than predicted.

(9) The following plot of the residuals from the fitted model implies that



(a) The specified model has missed a nonlinear pattern.
(b) The underlying errors are autocorrelated.
(c) The underlying errors lack constant variance.
(d) The model tends to under-predict rentals, on average.
(e) The underlying errors have a non-normal distribution.

(10) In his quest for a more accurate model, the manager of this company decided to include data from weekends and holidays during this period. The addition of this data to the model would most likely have what effect?
(a) Produce a model with a smaller standard error for the slope.
(b) Produce a model with a smaller RMSE.
(c) Produce a model with a larger R^2 summary statistic.
(d) Require a categorical predictor to distinguish weekdays from weekends.
(e) The added data would have little noticeable effect on the estimated model.

(11) An important aspect of this data that is not apparent in the shown output is the possible presence of
(a) autocorrelation.
(b) homoscedasticity.
(c) heteroscedasticity.
(d) outliers.
(e) nothing. All important aspects have been explored.

(Questions 12-20). An analyst in the purchasing group at a corporation built a regression to estimate the cost of expanding the amount of disk space in the desktop machines at this firm. To build the model, the analyst collected prices from advertisements at several vendors of 18 disk drives of various sizes from 200 GB (gigabytes) up to 1,000 GB. Some were internal disk drives (which must be installed into the desktop computer) whereas others were external drives (which need only be connected to the desktop machine via a cable).

Summary of Fit

RSquare	0.979
Root Mean Square Error	32.646
Mean of Response	340.778
Observations (or Sum Wgts)	18

Expanded Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-45.586	26.220	-1.74	0.1040
Type[external]	40.836	9.619	4.25	0.0008
Type[internal]	-40.836	9.619	-4.25	0.0008
Capacity (GB)	0.936	0.074	12.68	<.0001
(Capacity (GB)-400)*Type[ext]	-0.020	0.074	-0.27	0.7920
(Capacity (GB)-400)*Type[int]	0.020	0.074	0.27	0.7920

(12) This output implies that, following the usual regression assumptions with 95% confidence, we should conclude that for disks in this range of sizes
- (a) External disks cost significantly more per GB than internal disks.
- (b) External disks cost significantly less per GB than internal disks.
- (c) The cost per GB is identical for internal and external disks.
- (d) The cost per GB is similar for internal and external disks.
- (e) Capacity has no effect on price for either type of disk.

(13) Based on the fit of this model, for internal disk drives of about 400 GB, the average cost per additional gigabyte of disk space is approximately
- (a) \$41 per GB.
- (b) \$82 per GB.
- (c) \$1 per GB.
- (d) \$0.02 per GB, with external drives costing statistically significantly more per GB.
- (e) \$0.02 per GB, with internal drives costing statistically significantly more per GB.

(14) Based on the fit of this model and making the standard assumptions, a 400 GB internal disk can be expected to cost approximately
- (a) \$378
- (b) \$296
- (c) \$337
- (d) \$288
- (e) \$370

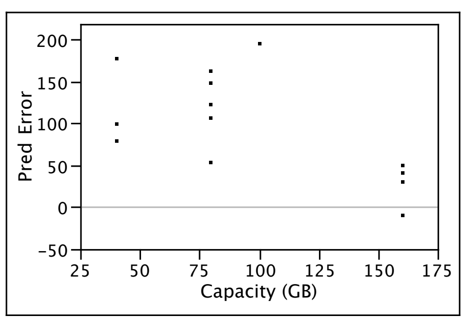
(15) The fitted model implies that a 400 GB internal disk costs on average between
- (a) \$22 to \$60 less than a 400 GB external disk.
- (b) \$31 to \$50 less than a 400 GB external disk.
- (c) \$43 to \$120 less than a 400 GB external disk.
- (d) \$62 to \$101 less than a 400 GB external disk.
- (e) About the same as a 400 GB external disk.

(16) Vendor "A" offers an internal 300 GB disk for \$35 less than vendor "B". When comparing these prices, the results of the fitted model suggest that
(a) These prices are within random pricing variation for disks of this size.
(b) Vendor "A" is charging an unusually low price for this size disk.
(c) The disk offered by vendor "B" must be of higher quality than that of vendor "A".
(d) The model is flawed because it predicts the same price for both disks.
(e) The price offered by vendor "A" is obviously an outlier and probably in error.

(17) It costs the corporation \$50 in labor and parts to install an internal disk drive. External disks do not incur this additional cost. The results shown indicate that on average for disks of about 400 GB, when installation costs of internal disks are included,
(a) External disks cost \$11 more to deploy than internal disks.
(b) External disks cost \$11 less to deploy than internal disks.
(c) External disks cost \$32 more to deploy than internal disks.
(d) External disks cost \$42 more to deploy than internal disks.
(e) External disks cost the same to deploy as internal disks.

(18) In order to verify that the variance of prediction errors from this model is comparable for both internal and external disks, we would be best served by checking the
(a) Plot of residuals on fitted values.
(b) Plot of observed prices on fitted values.
(c) Normal quantile plot of the residuals.
(d) A sequence plot of the residuals.
(e) Comparison boxplot of residuals by *Type*.

(19) An associate of this analyst used her model to estimate the cost of several compact, lightweight disk drives (both internal and external) intended for use with laptops. The following plot shows the prediction errors when this model was used to estimate the prices of these smaller disks.



This plot suggests that
(a) The fitted model should not be used to price disks larger than 200 GB.
(b) Internal laptop disks cost more than external laptop disks.
(c) Internal laptop disks cost less than external laptop disks.
(d) These data should be added to the prior analysis and the modeling redone.
(e) Many laptop disks cost significantly more per GB than larger format disks.

(20) These 18 prices were gathered from advertised prices of 4 vendors, using the prices of 4-5 disks from each. This information suggests that

(a) The data may not be independent.
(b) We should consider adding a categorical term for vendor to the model.
(c) We should consider adding vendor interactions to the model.
(d) The data may not have comparable variation over the vendors.
(e) All of the above.

(Questions 21-26) A recent Wharton graduate decided that it was time to get a better car. She wanted a BMW, but decided it made more sense to purchase a used (i.e., "certified pre-owned") model rather than a new car. She collected data from the official BMW web site that show the asking price (in $US), model year (2001-2005), and mileage of 69 3-Series BMW's offered for sale by BMW dealers within about 50 miles of Philadelphia. She fit a multiple regression model to this data; the response is the listed price for the car.

Correlations

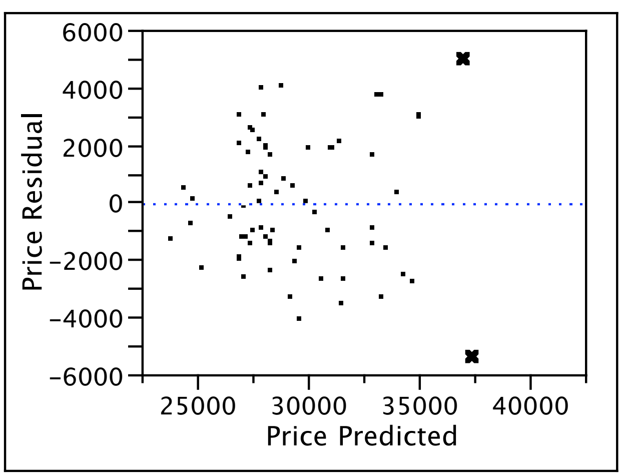
	Price	**Year**	**Mileage**
Price	1.000	0.765	-0.661
Year	0.765	1.000	-0.628
Mileage	-0.661	-0.628	1.000

Summary of Multiple Regression

RSquare	0.639
Root Mean Square Error	2280

Parameter Estimates

Term	**Estimate**	**Std Error**	**t Ratio**	**Prob>|t|**
Intercept	-4.453e+6	739054.4	-6.03	<.0001
Year	2239.556	368.807	6.07	<.0001
Mileage	-0.081	0.026	-3.13	0.0026



(21) Her results suggest that if she buys a 2002 model with 20,000 miles rather than a comparable 2003 model with 20,000 miles, she can expect to save, on average,
(a) Less than \$1000
(b) About \$1600.
(c) About \$2400.
(d) About \$3200.
(e) More than \$3500.

(22) If she were to fit a simple regression with the model year as the one predictor (keeping the listed price as the response), then we can see that
(a) The slope in simple regression for *Year* would be smaller than 2239.556.
(b) The RMSE would be smaller.
(c) The R^2 statistic would drop, but not by a statistically significant amount.
(d) The R^2 statistic would drop by a statistically significant amount.
(e) The *p*-value slope of *Year* in the simple regression would be larger than 0.05.

(23) In a simple regression of *Price* on *Mileage*, the estimated slope would be
(a) -0.06
(b) -0.12
(c) -0.36
(d) -0.66
(e) Cannot be obtained from the shown results.

(24) A friend of the modeler claimed that each additional mile of driving on average reduces the resale value of a car of this type by 10 cents (i.e., \$0.10). These results
(a) Are too inaccurate to be used to verify this claim.
(b) Imply that her friend's claim is basically right.
(c) Imply that the incremental mileage cost is significantly *more* than her friend's claim.
(d) Imply that the incremental mileage cost is significantly *less* than her friend's claim.
(e) Do not address her claim; we need the regression of *Price* on *Mileage*.

(25) If she were to remove the two points marked with "*x*" in the shown residual plot, then we can be sure that
(a) The slope for *Mileage* would increase.
(b) The slope for *Year* would increase.
(c) The model would suffer from increased collinearity.
(d) The RMSE would decrease.
(e) The fit would be about the same because these points cancel each other out.

(26) The residual variation not represented in this model can best be attributed to
(a) Other factors, such as engine size, not used in this model.
(b) The presence of a nonlinearity that is not captured by this model.
(c) The limited sample size of only 69 observations.
(d) The restriction of this data to cars available in the Philadelphia area.
(e) The variation associated with a sample from a normal distribution.

(Questions 27-41) An analyst at an insurance company used the following sample of 2,099 households from the 2003 American Community Survey, a survey of households in the US The analyst used these data to build a regression model. The response in the model is the annual cost for hazard insurance model paid by residents who either own their home or are paying off a mortgage. The model includes several other characteristics of each household:

Income	Measured in dollars
Value	Of home, coded as 1=$5,000, 2=$10,000, 3=$15,000…
Rooms	Number of rooms, in total
Bedrooms	Number of bedrooms
Owned	Code as "Yes" or "No" (no indicates paying a mortgage)

In addition, homes in the sample were divided into four categories, based on the age of the construction of the home:

New	Less than 10 years old
Recent	More than 10 and less than 25 years old
Aging	More than 25 years old, but built since 1945
Old	Earlier construction

The following table shows the counts and average annual payment for hazard insurance for homes in each category of the variable *Age*:

Level	Number	Mean Annual Payment	Std Dev of Annual Payment
New	113	612.57	373.44
Recent	772	600.38	406.50
Aging	1083	566.75	388.92
Old	131	590.84	373.41

Correlations

	Hazard_Ins	Income	Value	Rooms	Bedrooms
Hazard_Ins	1.000	0.347	0.495	0.338	0.284
Income	0.347	1.000	0.485	0.393	0.330
Value	0.495	0.485	1.000	0.535	0.431
Rooms	0.338	0.393	0.535	1.000	0.726
Bedrooms	0.284	0.330	0.431	0.726	1.000

Summary of Multiple Regression

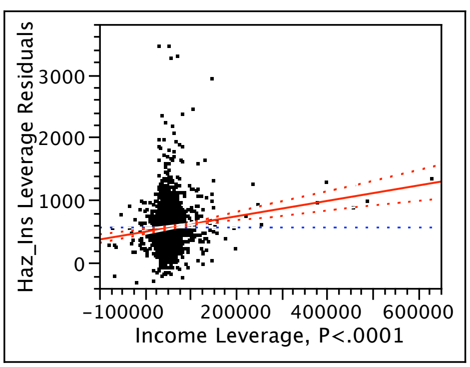
RSquare	0.28
Root Mean Square Error	325
Mean of Response	583
Observations (or Sum Wgts)	2099

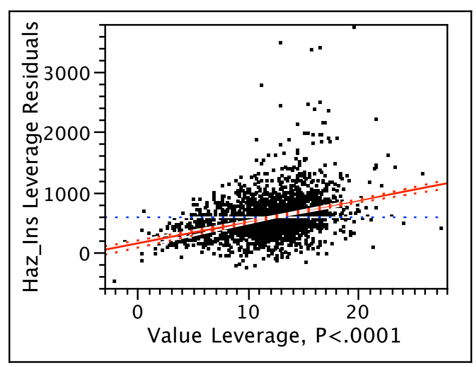
Effect Tests

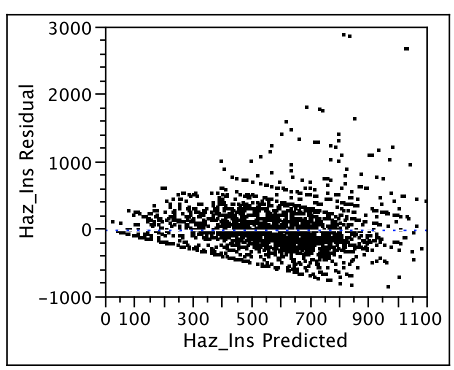
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Age	3	3	1469507	4.3351	0.0047
Income*Age	3	3	1813180	5.3489	0.0011

Expanded Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-54.2325	36.5015	-1.49	0.1375
Income	0.0012	0.0002	5.29	<.0001
Value	35.8466	2.0920	17.13	<.0001
Rooms	10.8040	7.7798	1.39	0.1651
Bedrooms	25.3080	14.5861	1.74	0.0829
Owned[No]	-19.046	8.0536	-2.4	0.0164
Owned[Yes]	19.046	8.0536	2.4	0.0164
Age[New]	-71.0125	26.4439	-2.69	0.0073
Age[Recent]	-16.1311	14.7998	-1.09	0.2759
Age[Aging]	12.6262	14.0069	0.90	0.3675
Age[Old]	74.5175	24.5864	3.03	0.0025
(Income-55276)*Age[New]	-0.0003	0.0005	-0.70	0.4824
(Income-55276)*Age[Recent]	-0.0009	0.0003	-3.15	0.0017
(Income-55276)*Age[Aging]	0.0002	0.0003	0.65	0.5171
(Income-55276)*Age[Old]	0.0010	0.0004	2.56	0.0104







(27) Other factors, unaccounted for by this model,
 (a) Account for more than 2/3 of the variation in insurance payments.
 (b) Account for less than 2/3 of the variation in insurance payments.
 (c) Would lead to statistically significant improvements in this model.
 (d) Explain the statistically insignificant predictive ability of this model.
 (e) Are unimportant because this model explains significant amounts of variation.

(28) When compared to using the overall mean $583 for predicting the amount paid for hazard insurance, this model offers a prediction whose margin for error is about
 (a) 28% of that obtained by using the mean.
 (b) 53% of that obtained by using the mean.
 (c) 72% of that obtained by using the mean.
 (d) 85% of that obtained by using the mean.
 (e) The same as that obtained by using the mean.

(29) The p-value for the intercept in the shown model indicates that, given the usual assumptions of multiple regression,
 (a) About 14% of the observations come from a population with intercept zero.
 (b) About 14% of the observations have intercept as small as that observed here.
 (c) There is about a 14% chance that the population intercept is zero.
 (d) There is about a 14% chance that the population intercept is not zero.
 (e) None of the above.

(30) This model implies that when comparing otherwise similar homes and households and with 95% confidence, the amount of annual hazard insurance payments rises on average by (Note that the predictor *Value* is measured in increments of $5,000)
 (a) $6.33 to $8.01 per $1,000 increase in the value of the home.
 (b) $6.75 to $7.59 per $1,000 increase in the value of the home.
 (c) $31.66 to $40.03 per $1,000 increase in the value of the home.
 (d) $33.76 to $37.94 per $1,000 increase in the value of the home.
 (e) $158 to $200 per $1,000 increase in the value of the home.

(31) This analysis suggests that on average for a given type of home and household (home size, age and household income) households who own their home pay
 (a) A significantly higher premium than those paying a mortgage.
 (b) A higher premium than those paying a mortgage, but not significantly more.
 (c) About the same annual premium as those paying a mortgage.
 (d) A lower premium than those paying a mortgage, but not significantly less.
 (e) A significantly lower premium than those paying a mortgage.

(32) Based on the fit of this model, a household with income $80,000 living in an "old" home valued at $200,000 with 10 rooms and 3 bedrooms would be expected on average to spend for hazard insurance annually about
 (a) $177 more than a comparable household living in a "new" home.
 (b) $145 more than a comparable household living in a "new" home.
 (c) $75 more than a comparable household living "new" home.
 (d) $71 less than a comparable household living "new" home.
 (e) About the same as a comparable household living in a "new" home.

(33) Taken by itself alone, the Effect Test for *Age*
- (a) Implies that older homes are more expensive to insure than newer homes.
- (b) Implies that newer homes are more expensive to insure than older homes.
- (c) Implies that newer homes cost about the same to insure as older homes.
- (d) Implies that some slopes representing *Age* in the fit are statistically significant.
- (e) Should not be interpreted directly because of an interaction.

(34) The best interpretation of the regression coefficient for the predictor *Bedrooms* is that with regard to annual payments for hazard insurance for households like those used to build this model that on average
- (a) Adding a bedroom to a house increases the insurance premium by $10.
- (b) Adding a bedroom to a house increases the insurance premium by $25.
- (c) Bedrooms are statistically significantly more expensive to insure than most rooms.
- (d) Additional bedrooms can be insured at no additional cost.
- (e) Ten-room households with 4 bedrooms pay about $50 more than if they had 2 bedrooms.

(35) It has been claimed within this company that, all other things comparable, households with higher incomes are willing to – or perhaps need to – pay higher premiums, with the amount of the premium rising by about $1 per thousand in income. With regard to this claim, this analysis (if we accept this shown model)
- (a) Rejects the claim: rising income produces greater increases in the premium.
- (b) Rejects the claim: rising income produces smaller increases in the premium.
- (c) Cannot reject the claim because of the effects of collinearity in this model.
- (d) Cannot reject the claim because of the effects of outliers in this model.
- (e) Implies that we cannot address the claim without knowing the age of the home.

(36) If we accept the fitted regression model, then we have assumed that households living in more expensive homes (those with higher value)
- (a) Prefer homes constructed more recently.
- (b) Prefer older homes, particularly those constructed before 1945.
- (c) Pay the same for their insurance per dollar value as owners with less income.
- (d) Pay more for their insurance per dollar value than owners with less income.
- (e) Pay less for their insurance per dollar value than owners with less income.

(37) The predictor *Rooms* is noticeably correlated with the amount paid for hazard insurance, but is not statistically significant in this regression. The best explanation for this discrepancy is
- (a) Outliers have distorted the multiple regression.
- (b) The effects of collinearity.
- (c) The underlying relationship is not linear and should be transformed to use logs.
- (d) The presence of autocorrelation.
- (e) The absence of an important interaction.

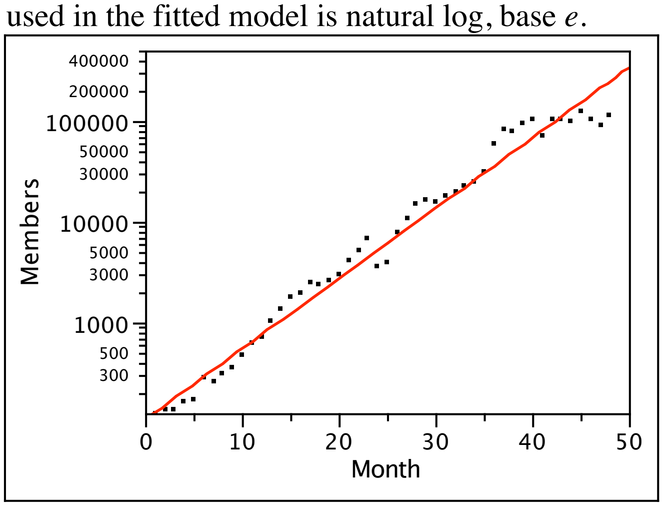
(38) The shown diagnostic plots of this model indicate that
- (a) The slope for the predictor *Income* has been distorted by collinearity.
- (b) The slope for the predictor *Value* has been distorted by collinearity.
- (c) Several leveraged points influence the slope for the predictor *Income*.
- (d) Several leveraged points influence the slope for the predictor *Value*.
- (e) No problems in the estimation of the slopes for *Value* and *Income*.

(39) This company sells insurance policies in 35 states, all of which appear in this sample. In order to improve the fit of this model, an analyst added a categorical variable indicating the state of residence to the shown model. In order to obtain a statistically significant effect
(a) At least one of the added dummy variable coefficients would have to be significant.
(b) More than one added dummy variable coefficient would have to be significant.
(c) The R^2 of the revised model would have to increase by at least 0.005.
(d) The R^2 of the revised model would have to increase by at least 0.03.
(e) The R^2 of the revised model would have to increase by at least 0.10.

(40) Two brothers live in different states, but both have hazard policies written by this company. The brothers have similar \$150,000 incomes and own comparable, "new" homes with 12 rooms and 4 bedrooms, valued at \$400,000. One brother pays \$650 more than the other for his insurance policy. Accepting the fit of this model and the usual assumptions of regression, a difference of at least this size between two households
(a) Happens less than 1% of the time.
(b) Happens about 2.5% of the time.
(c) Happens about 5% of the time.
(d) Happens about 15% of the time.
(e) Happens more than 50% of the time.

(41) The *most* useful next step to take in the analysis of this data would be to
(a) Remove outliers that have distorted the slope for *Income*.
(b) Remove outliers that have distorted the slope for *Value*.
(c) Use sequence plots to isolate the presence of dependence.
(d) Reformulate the model to use the amount of insurance per home value.
(e) Remove the insignificant predictor that indicates home ownership.

(Questions 42-45) A commercial web site opened in January 2000. In order to keep track of it's visitors, it asks them to sign up. Those that do receive better services and access to more of the information at this web site. The data in this analysis track the growth in the membership of this web site over the 48 months from January 2000 through December 2003. The log scale used in the fitted model is natural log, base *e*.



Log(Members) = 4.732 + 0.161 Month

RSquare	0.979
Root Mean Square Error	0.334
Observations	48

Parameter Estimates

Term	**Estimate**	**Std Error**	**t Ratio**	**Prob>|t|**
Intercept	4.732	0.098	48.34	<.0001
Month	0.161	0.003	46.24	<.0001

(42) The slope in the shown model implies that the membership at this web site has been growing by about
(a) 0.16% per month.
(b) 1.6% per month.
(c) 16% per month.
(d) 16,100 per month.
(e) 161,000 per month.

(43) Given the usual assumptions of regression, with 95% confidence, this model predicts the membership in January 2004 (the next month) to be in the range
(a) 108,000 to 412,000.
(b) 155,000 to 591,000.
(c) 234,000 to 368,000.
(d) 267,000 to 334,000.
(e) 302,851 to 302,853.

(44) If a linear model were fit to the data with the response as the number of members and the predictor *Month* as used here, then predictions from the linear model would
(a) Over-predict the membership in January 2004.
(b) Under-predict the membership in January 2004.
(c) Be more accurate than the predictions from the shown model.
(d) Give results that are comparable to those of the shown model.
(e) One cannot identify from what is shown here the nature of the linear model.

(45) The most useful diagnostic that should be examined in order to check the fit of the shown model would be to examine a plot that shows
(a) The fit of the model on the actual scale of members versus month.
(b) Normal quantile plot of the residuals.
(c) The residuals versus the month.
(d) Log of the membership versus the log of month.
(e) Comparison boxplots of residuals grouped by year.