Statistics Waiver Exam

August 3, 2013

*Before* the exam begins …

**Fill in your name and Penn student ID number**.  
Your ID number appears in bold on your ID.

**Mark the “bubbles”** under the letters of your name *and* student ID number.   
Failure to do so will lead to a score of zero.

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

**Turn off your phone.**   
You are not allowed to make or receive a call during the exam. Use of your phone (for calls or messages) during the exam is grounds for dismissal from the exam.

*Once* the exam begins …

Choose the **one** ***best* answer** for each question. The exam has **44 questions**.  
Picking more than one answer or leaving an answer blank is an error.

You may consult **1 page of handwritten notes** during the exam (8.5”x11” or A4, both).  
No other reference materials are permitted.

You may use a basic **calculator** or graphing calculator.  
No laptops, phones, or computers are allowed.

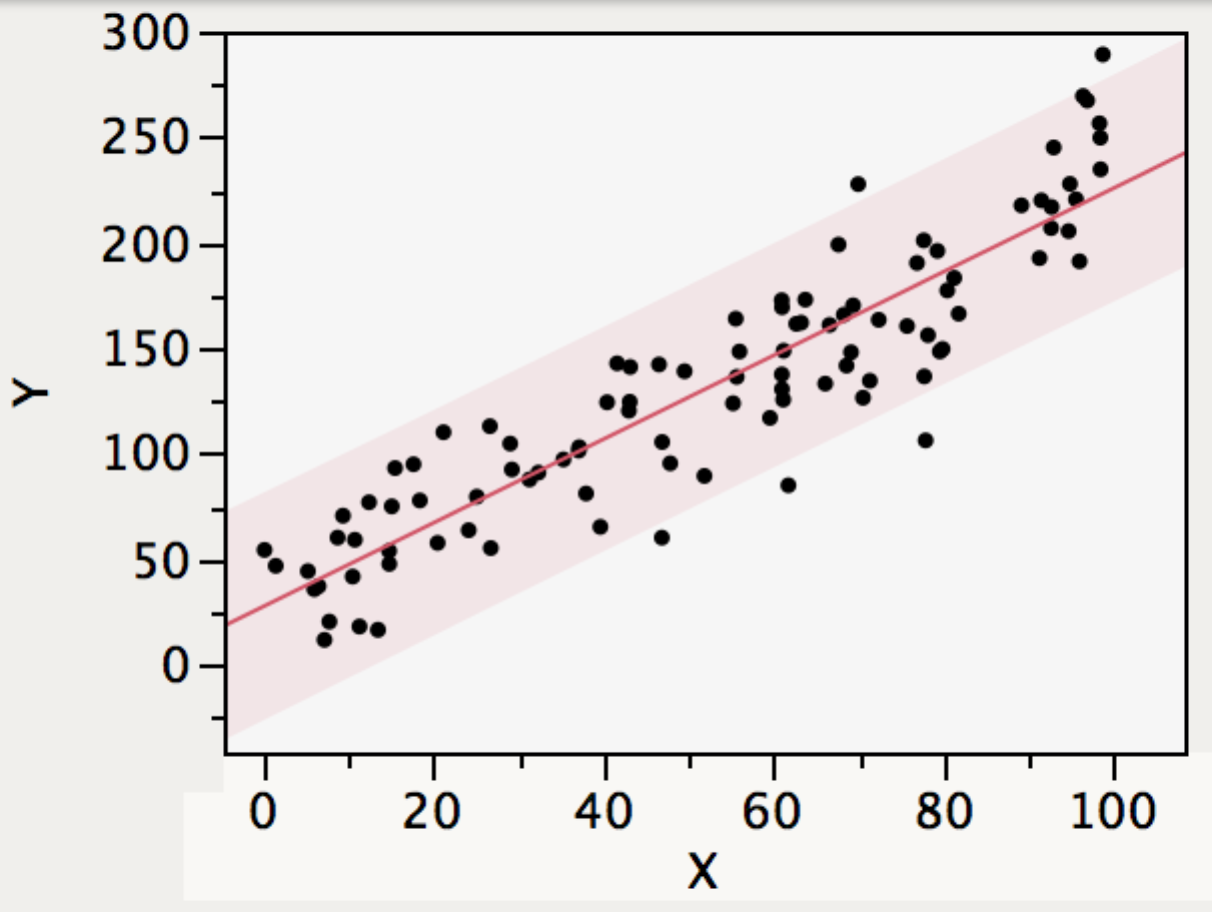
You have **two hours** for the exam. The **computer output** associated with one or more items should be considered an essential part of the questions. Throughout, the word “significant” implies “statistically significant”. The abbreviation SRM stands for standard definition of a ‘simple regression model’; MRM for ‘multiple regression model.’ All logs are natural logs (logs using base *e*) unless otherwise indicated.

Your **score** is the number of correct answers. The 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. Regardless of what you write on your copy of the exam, only the marked answers on the grade form will be considered. The first question is not scored; it indicates which exam you are taking. Mark the answer form as instructed.

**STOP**

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

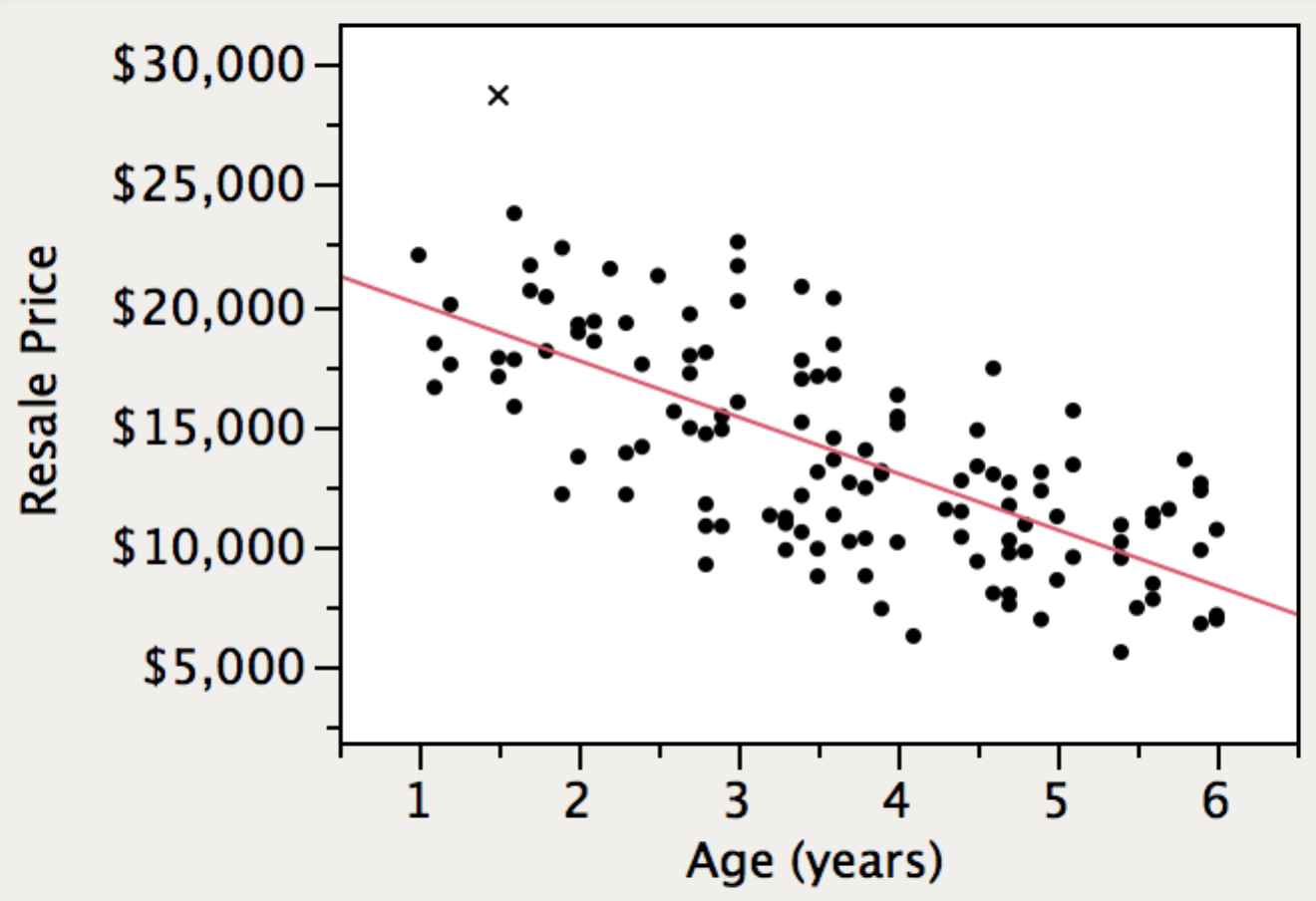
1. Mark the answer to question #1 on your answer form **A**. This answer identifies your version of the exam. Please do as indicated. This question is not scored.
2. The alpha-level of a test of the null hypothesis *H*0 versus the alternative hypothesis *H*a is the
   1. Probability of rejecting *H*0 when *H*0 is true.
   2. Probability of rejecting *H*a when *H*a is true.
   3. Probability of rejecting *H*a when *H*0 is true.
   4. Probability of rejecting *H*0 when *H*a is true.
   5. Probability that the test produces the correct decision.
3. The *p*-value of a test of the null hypothesis *H*0 versus the alternative hypothesis *H*a
   1. Is the probability of rejecting *H*0.
   2. Is the probability of rejecting *H*a.
   3. Rejects *H*0 if the *p*-value is less than the alpha-level.
   4. Is large when the assumptions of the test are not met.
   5. Is small when the assumptions of the test are not met.
4. The sample standard error of the mean of a simple random sample estimates
5. Closeness of the data to a normal distribution.
6. Standard deviation of the sampling distribution of the average.
7. Standard deviation of the data in the population.
8. The expected size of the deviation between x-bar and *μ*.
9. The presence of data entry errors in the observed sample.
10. Confidence intervals rely on the Central Limit Theorem which implies that, when drawing simple random samples from a population with mean *μ* of sufficient size, on average
11. The data within the sample is normally distributed.
12. 95% of samples are normally distributed.
13. The means of 95% of samples lie within 2 standard errors of *μ*.
14. The estimated sample standard deviation approaches the population SD *σ*.
15. The sample averages are closer to *μ* if the samples are from a normal population.
16. A manufacturer produces electronic circuits whose output voltage is normally distributed with mean 5 volts and standard deviation 0.05 volts, independently. The probability that the mean voltage of a simple random sample of 25 circuits lies between 4.9 to 5.1 volts is approximately
17. 0.95
18. 0.05
19. 1-0.9525
20. 1
21. 1-0.0525
22. US Air wants a 95% confidence interval for the average weight of travelers’ checked suitcases. If *σ* = 5 kg, then the sample size necessary for the length of the confidence interval to be 1 kg is approximately
23. 4
24. 20
25. 100
26. 40
27. 400
28. Two airlines build 95% confidence intervals for the mean weight *μ* of luggage carried by passengers. US Air samples 500 passengers, and Lufthansa samples 1,000 passengers. Assume both take simple random samples from the *same* population. Then we should expect that Lufthansa’s confidence interval is
    1. More likely to contain *μ* than the confidence interval used by US Air.
    2. Less likely to contain *μ* than the confidence interval used by US Air.
    3. Shorter than the confidence interval used by US Air.
    4. Longer than the confidence interval used by US Air.
    5. Virtually identical to the confidence interval used by US Air.
29. The simple regression model (SRM) with equation Y = β0 + β1 X + ε presumes that
    1. Observed values of Y form a sample from a normal distribution.
    2. The error terms form a sample from a normal distribution.
    3. Observed values of X and Y form samples from normal distributions.
    4. Observed values of X are independent of each other.
    5. The random variables Y and ε are independent of each other.
30. An advertiser builds a simple regression model to predict sales from the level of promotion. By increasing the sample size from *n*=100 to *n*=400, then we can be assured that the
31. *R*2 of the fitted model will increase.
32. *RMSE* of the fitted model will decrease.
33. Estimated slope of the fitted model will decrease.
34. Estimated intercept of the fitted model will increase.
35. Standard error of the intercept of the fitted model will decrease.
36. The assumption of homoscedasticity in regression analysis implies that
37. Prediction intervals from fitted model have the same length everywhere.
38. The unobserved error variation follows a bell-shaped normal distribution.
39. Unobserved variation is comparable for every observation.
40. Observed values of the explanatory variables are uniformly distributed.
41. Scatterplots of Y versus each explanatory variable are linear.



(Q12 – 14) The figure shown to the right displays a simple regression fit to *n*=100 cases. The shaded region denotes the 95% prediction bands around the fitted line.

1. The standard error of the estimated slope of the fitted model is approximately
2. 0.01
3. 5
4. 0.1
5. 20
6. 1
7. The correlation between X and Y shown in the plot is approximately
8. 1.0
9. 0.6
10. 0.0
11. 0.3
12. 0.9
13. Five observations lie outside the 95% prediction bands in the figure. The presence of these outlying values
14. Is as expected from a model that conforms to the SRM.
15. Indicates that the error variation is not normally distributed.
16. Implies that the data are not homoscedastic.
17. Suggests the presence of underlying dependence.
18. Suggests increasing the coverage of prediction intervals from 95% to 99%.
19. A dummy variable in regression modeling
20. Is a temporary variable used until more precise data is obtained.
21. Indicates which cases belong to a specific category.
22. Signals the presence of an outlying value.
23. Refers to data that is contaminated by random variation.
24. Is used to describe a nonlinear equation.
25. When an additional explanatory variable is added to a multiple regression, we can be sure that
    1. The RMSE will decrease.
    2. The *p*-values of other explanatory variables will decrease.
    3. The *p*-value of the overall *F* statistic will decrease.
    4. The *p*-value of the overall *F* statistic will increase.
    5. The *R*2 statistic will not decrease.

**(Q17-24)** Representing a large auto dealer, a buyer attends car auctions. To help with the bidding, the buyer fit a simple regression to predict the resale value of cars purchased at the auction. The data that the buyer used describe *n* = 125 prior sales of a specific model of cars. The accompanying output summarizes the fitted model.



|  |  |
| --- | --- |
| RSquare | 0.490904 |
| Root Mean Square Error | 3254.812 |
| n | 125 |

| **Term** | **Estimate** | **Std Error** | **t Ratio** | **Prob>|t|** |
| --- | --- | --- | --- | --- |
| Intercept | 22390.00 | 831.0325 | 26.94 | <.0001\* |
| Age (years) | -2348.96 | 215.6875 | -10.89 | <.0001\* |

1. The fitted equation estimates that a new car of this model sells for about
2. $22,400
3. $20,000
4. Cannot be determined from the fitted model.
5. $13,900
6. $32,500
7. The fitted model estimates that the average resale price of a 1-year old car is about
8. The same as the resale price of a 3-year old car.
9. $7,050 more than the resale price of a 3-year old car.
10. $4,700 more than the resale price of a 3-year old car.
11. Less than the resale price of a 3-year old car.
12. $2,350 more than the resale price of a 3-year old car.
13. It has been claimed that cars of this type depreciate in resale value by $2,000 per year. Based on the fit of this model, we can conclude that (assuming the SRM holds)
14. Cars depreciate more than $2,000 per year, but not by a significant amount.
15. Cars depreciate less than $2,000 per year, but not by a significant amount.
16. Cars depreciate less than $2,000 per year, by a significant amount.
17. Cars depreciate more than $2,000 per year, by a significant amount.
18. The estimated model does not address this claim.
19. Assuming that the SRM holds in this example, then the proportion of 2-year old used cars (of this model) that have resale value larger than $21,000 is about
20. 33%
21. 5%
22. 2.5%
23. 16%
24. 50%
25. If the observation marked with the symbol × in the figure were removed from the analysis, then
26. The *R*2 statistic would increase.
27. The estimated intercept would increase.
28. The estimated slope would get closer to zero.
29. The standard error of the slope would increase.
30. The RMSE would increase.
31. If the resale price were expressed in thousands of dollars rather than dollars (so that, for example, the labels on the y-axis in the scatterplot were 5, 10, 15, 20, 25, and 30 rather than 5000, 10000, etc.) then how would the output change?
32. The *R*2 statistic would be larger.
33. The RMSE would be smaller.
34. The *t*-statistic for the intercept would be smaller.
35. The standard error of the slope would be larger.
36. The *p*-value associated with the slope would be smaller.
37. This model of car is sold in two types that represent collections of options, denoted SE and R. It is believed that these types depreciate at different rates. The categorical variable *Type* identifies the types. To investigate whether types depreciate at statistically significantly different rates, we should
38. Separate the data into 2 groups and fit 2 simple regressions.
39. Collect another sample and add it to the shown simple regression.
40. Add the categorical variable *Type* to the model.
41. Add the categorical variable *Type* and its interaction with *Age* to the model.
42. Run a two-sample *t*-test of *Resale Price* on *Type*.
43. In order to obtain a more precise, legitimate estimate of the slope (*i.e.*, a narrower confidence interval), the most effective approach would be to
44. Remove outliers from the shown data.
45. Limit the analysis to cars that are between 0 and 2 years old.
46. Add more old cars to the data.
47. Add more nearly new cars to the data.
48. Add more old and nearly new cars to the data.

**(Q25-30)** A retail food market used data obtained from its loyalty shopper program to estimate the effects of distributing coupons to customers. These data describe purchase amounts (in dollars) made by 130 households; these households range in size from 1 to 6 persons, indicated by the variable *Household Size*. The data also include the dummy variable *Coupon* that indicates whether the customer took advantage of a discount coupon sent to households that participate in the loyalty shopper program (*Coupon[No]=*1 for those who did not use a coupon and 0 otherwise). The response in the model is *Sales*, the dollar value purchased by a household.

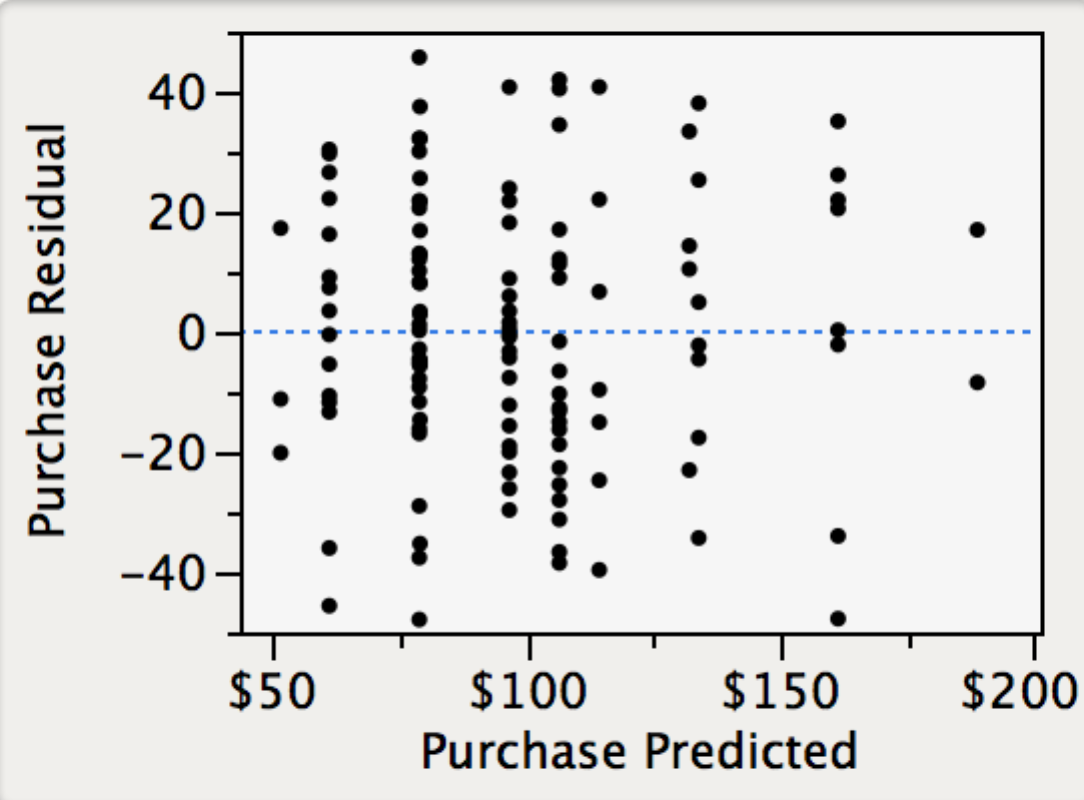
|  |  |
| --- | --- |
| RSquare | 0.6369 |
| Root Mean Square Error | 21.9585 |
| *n* | 130 |

**Analysis of Variance**

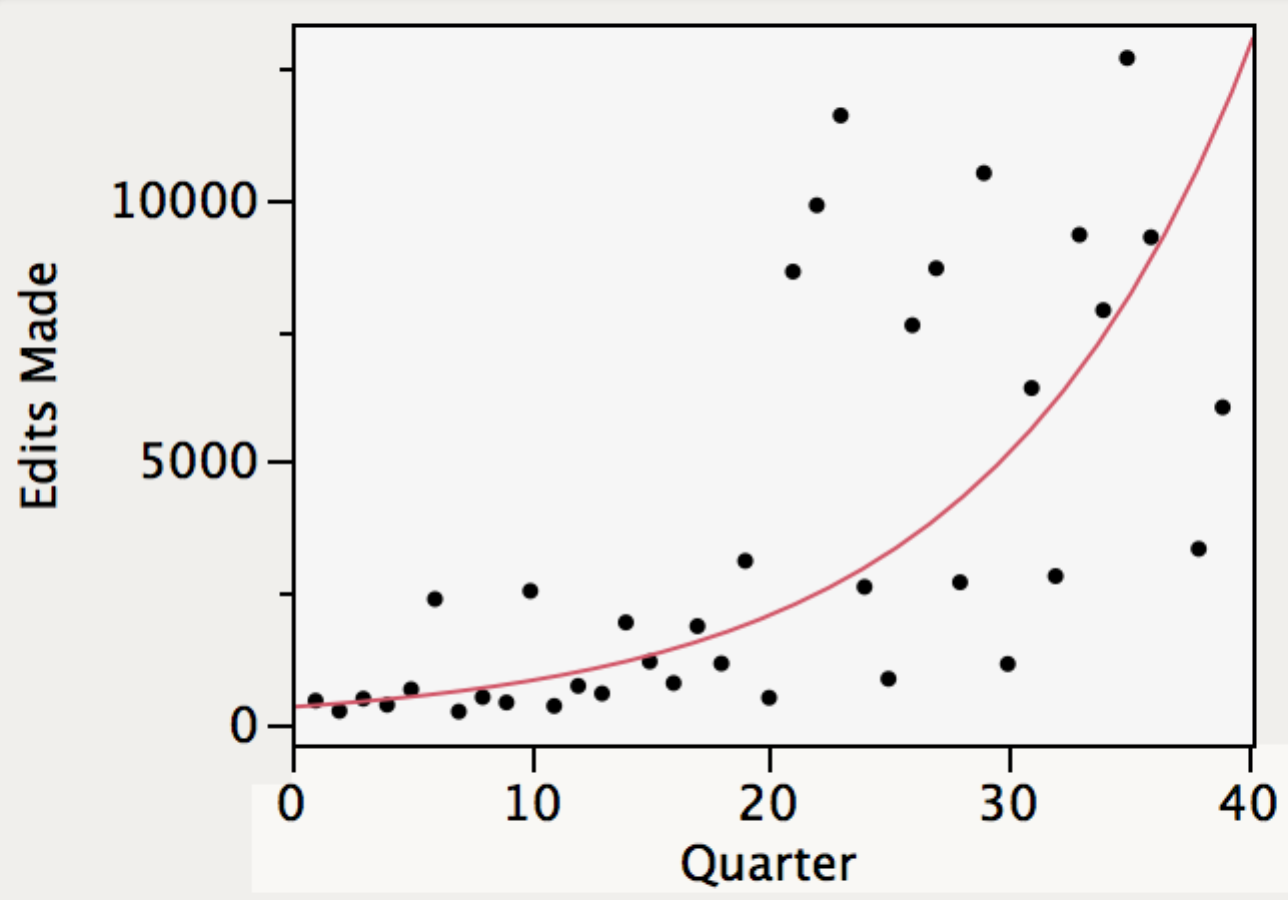
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Ratio** |
| --- | --- | --- | --- | --- |
| Model | 3 | 106589.70 | 35529.9 | 73.6867 |
| Error | 126 | 60754.06 | 482.2 | **Prob > F** |
| C. Total | 129 | 167343.77 |  | <.0001\* |

**Indicator Function Parameterization**

| **Term** | **Estimate** | **Std Error** | **t Ratio** | **Prob>|t|** |
| --- | --- | --- | --- | --- |
| Intercept | 23.860 | 8.099 | 2.95 | 0.0038\* |
| Coupon[No] | 19.363 | 10.341 | 1.87 | 0.0635 |
| Household Size | 27.513 | 2.412 | 11.41 | <.0001\* |
| Coupon[No]\*Household Size | -9.748 | 3.380 | -2.88 | 0.0046\* |

1. The overall *F*-statistic implies that, assuming the conditions of the multiple regression model (MRM) hold,
2. *Household Size* statistically significantly influences *Sales*.
3. The slope of *Household Size* is statistically significantly influenced by *Coupon*.
4. *Coupon* statistically significantly influences *Sales*.
5. All of the explanatory variables statistically significantly influence *Sales*.
6. The fitted model explains statistically significant variation in *Sales*.
7. *FIX THIS ONE FOR INDICATOR OUTPUT  
   The fitted equation of the model estimates that among those who do not use a coupon the average change in sales per added household member is about*
8. $22.64
9. $33.54
10. $17.76
11. $27.52
12. –$4.87
13. The fitted equation of this model predicts that the average sale to a household of size 4 that uses a coupon is approximately
14. $133.80
15. $124.10
16. $114.30
17. $104.60
18. $89.74
19. The fitted equation of this model indicates that on average sales are highest for
20. Small households that use a coupon.
21. Large households that use a coupon.
22. Small households that do not use a coupon.
23. Large households that do not use a coupon.
24. Small households, regardless of whether a coupon is used.
25. The plot to the right of this question shows residuals from   
    the fitted model. This plot indicates
26. Clusters of the data lack constant variance.
27. The data are autocorrelated.
28. The residual do not have a normal distribution.
29. No problem with the assumptions of the MRM.
30. Some households have more than 8 members.
31. To check the assumption of equal error variances it would be most useful to see the
32. Normal quantile plot.
33. Side-by-side boxplots of residuals grouped by coupon use.
34. Histogram of the residuals.
35. Residual leverage plot for coupon use.
36. Scatterplot matrix of *Sales*, *Coupon*, and *Household Size*.

**(Q31-35)** A user-designed social web site contains information about US foreign relations. (The site resembles Wikipedia, but emphasizes foreign relations of the US.) Managers built a regression model to quantify its growth in edits made over the past 40 quarters (10 years). The response in this model is the number of edits made to on-line content. *Quarter* is coded 1,2,3,…40.

****  
Log(Edits Made) = 5.8282978 + 0.0905669\*Quarter

|  |  |
| --- | --- |
| RSquare | 0.6433 |
| Root Mean Square Error | 0.7988 |
| *n* | 40 |

| **Term** | **Estimate** | **Std Error** | **t Ratio** | **Prob>|t|** |
| --- | --- | --- | --- | --- |
| Intercept | 5.8282978 | 0.257409 | 22.64 | <.0001\* |
| Quarter | 0.0905669 | 0.010941 | 8.28 | <.0001\* |

1. The fitted equation estimates that
2. Edits are growing about 9% per quarter.
3. Edits are growing about 9% annually.
4. Edits are growing by about 9,000 annually.
5. Edits are growing by about 90 quarterly.
6. Edits are growing about 5.8% per quarter.
7. Assuming the SRM holds, this model predicts with 95% probability that the hits in quarter 41 (the next quarter) will be in the interval
8. 8,740 to 10,340
9. 7,940 to 11,140
10. 13,925 to 13,928
11. 6,270 to 30,960
12. 2,820 to 68,810
13. The summary graph of this model would be most improved by
14. Counting hits in the thousands rather than individually.
15. Putting the horizontal axis on a log scale.
16. Putting the vertical axis on a log scale.
17. Putting both the horizontal and vertical axes on log scales.
18. Omitting the fitted equation since it misrepresents the data.
19. Managers of the web site suspect that the growth rate is higher in 4th quarters (quarters 4, 8, 12, …) than other quarters because of the timing of elections in the US. The variable *Fall* is coded 1 for these quarters and zero otherwise. To test this claim,
20. Add *Fall* to the shown regression model.
21. Run a two-sample *t*-test of *Edits Made* with groups defined by *Fall*.
22. Remove the non-fall quarters and re-estimate the fitted model.
23. Add *Fall* and its interaction with *Quarter* to the model.
24. Remove the log transformation and fit a better model using a quadratic function.
25. The most important diagnostic check of this model that is not shown is the
26. Durbin-Watson statistic.
27. Normal quantile plot of the residuals.
28. Comparison side-by-side boxplots of the residuals, grouped by quarter.
29. Leverage plot for the effect of quarter.
30. Color-coded scatterplot, colored by Winter, Spring, Summer and Fall quarters.

**(Q36-44)** A realtor that sells residential properties constructed a regression model that describes the price per square foot (Y, in dollars per square foot) of recently sold properties within its territory. The homes reside in one of three locations, loosely described as city, rural, and suburban (*Location*). Other features of the data describe the number of bathrooms in the home (*Number Baths*), the size of the home in square feet (*Square Feet*), and the distance from the nearest school (*Distance*, in miles).

|  |  |
| --- | --- |
| RSquare | 0.850382 |
| Root Mean Square Error | 8.613523 |
| Mean of Response | 171.6016 |
| *n* | 120 |

**Analysis of Variance**

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Ratio** |
| --- | --- | --- | --- | --- |
| Model | 9 | 46385.904 | 5153.99 | 69.4675 |
| Error | 110 | 8161.205 | 74.19 | **Prob > F** |
| C. Total | 119 | 54547.109 |  | <.0001\* |

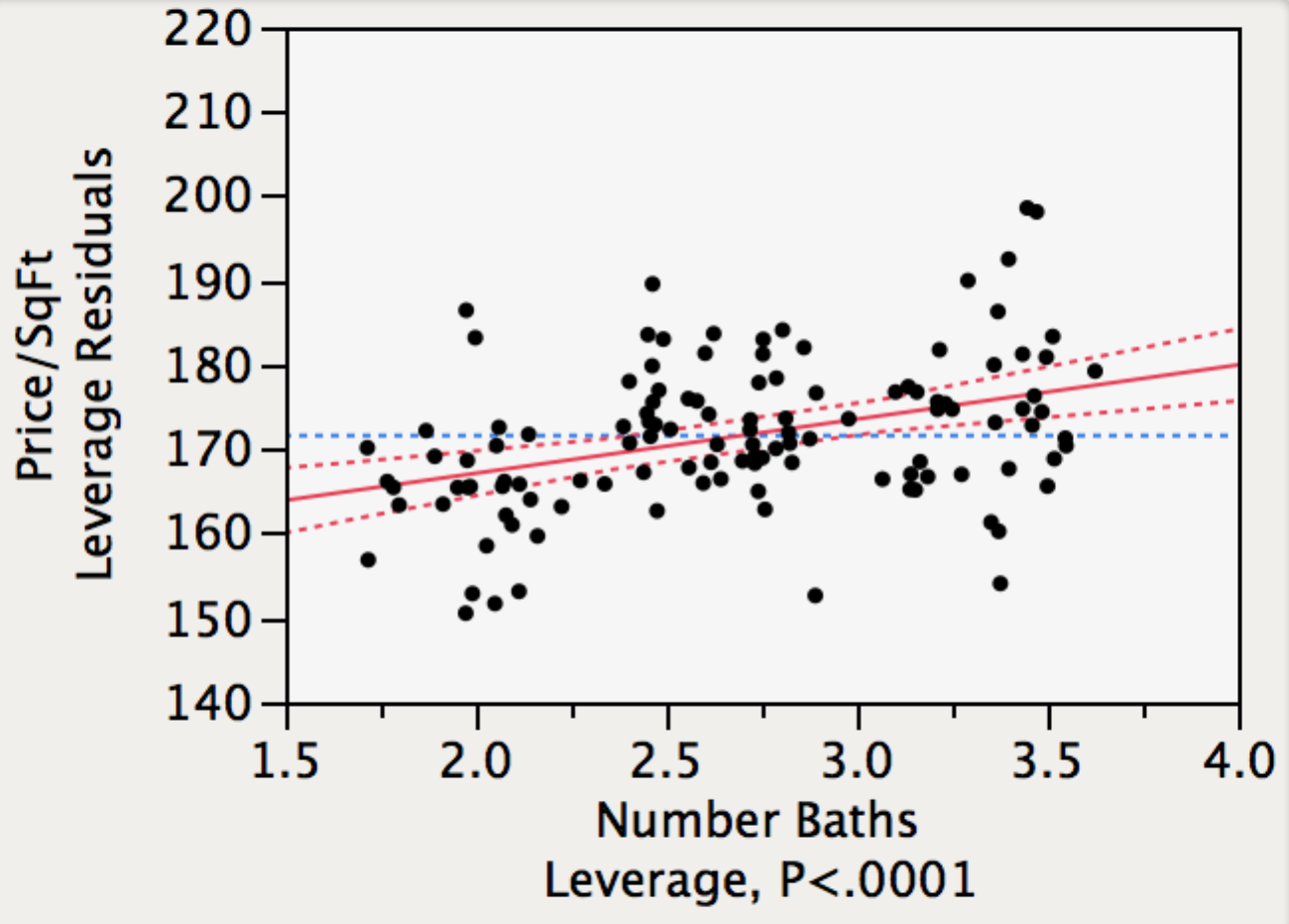
**Effect Tests**

| **Source** | **DF** | **Sum of Squares** | **F Ratio** | **Prob > F** |
| --- | --- | --- | --- | --- |
| 1/Square Feet | 1 | 8.9896 | 0.1212 | 0.7284 |
| Number Baths | 1 | 1389.0214 | 18.7218 | <.0001\* |
| Distance | 1 | 25.6417 | 0.3456 | 0.5578 |
| Location | 2 | 171.8500 | 1.1581 | 0.3179 |
| Number Baths\*Location | 2 | 726.4924 | 4.8960 | 0.0092\* |
| Distance\*Location | 2 | 104.2396 | 0.7025 | 0.4976 |

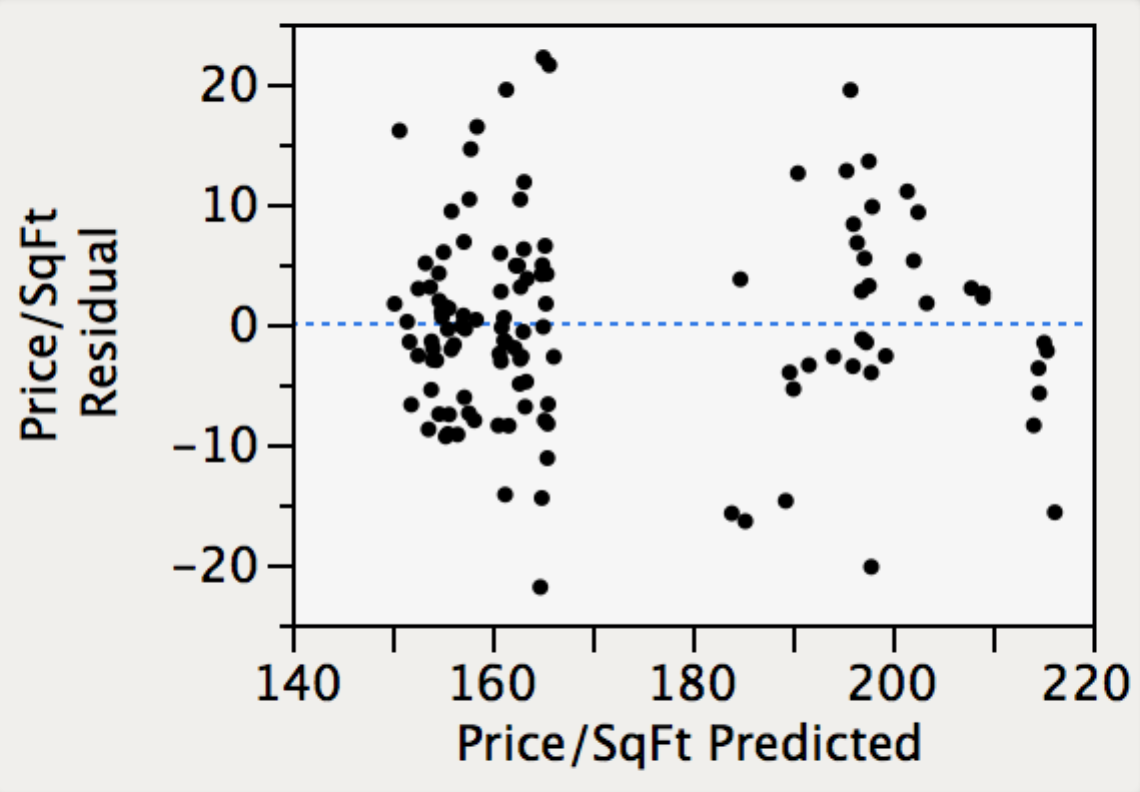
**Indicator Function Parameterization**

| **Term** | **Estimate** | **Std Error** | **t Ratio** | **Prob>|t|** |
| --- | --- | --- | --- | --- |
| Intercept | 162.409 | 8.387 | 19.37 | <.0001\* |
| 1/Square Feet | 2105.522 | 6048.807 | 0.35 | 0.7284 |
| Number Baths | 12.280 | 2.169 | 5.66 | <.0001\* |
| Distance | 1.515 | 1.716 | 0.88 | 0.3793 |
| Location[City] | -8.720 | 9.380 | -0.93 | 0.3546 |
| Location[Rural] | -16.028 | 10.676 | -1.50 | 0.1361 |
| Number Baths\*Location[City] | -8.306 | 3.910 | -2.12 | 0.0359\* |
| Number Baths\*Location[Rural] | -9.275 | 3.142 | -2.95 | 0.0039\* |
| Distance\*Location[City] | -1.097 | 2.431 | -0.45 | 0.6528 |
| Distance\*Location[Rural] | -1.989 | 1.789 | -1.11 | 0.2688 |

1. The fitted model estimates the price of a 3,000 square foot rural home with 3 bathrooms that is 1 mile from the closest school to be about
2. $487,000
3. $441,000
4. $439,000
5. $556,000
6. $467,000
7. The fitted slope of the variable 1/*Square Feet* can be interpreted to estimate that, when describing homes that resemble these,
8. The number of square feet does not influence the selling price.
9. The reciprocal of square feet is uncorrelated with the price per square foot.
10. Fixed costs comprise about $2,000 of the total selling price.
11. The price per square foot is poorly measured to be so unrelated to price.
12. The price per square foot grows more rapidly with size for larger homes.
13. When comparing prices of a homes with 3,000 square feet with 3 bathrooms, the equation of this model estimates that increasing the distance from the nearest school
14. Decreases the average price per square foot in a suburban location.
15. Increases the average price per square foot in a city location.
16. Has no effect on the average price per square foot in any location.
17. Decreases the average price per square foot in a city location.
18. Increases the average price per square foot in a rural location.
19. Within this collection of homes, those that are farther from schools tend to have more bathrooms (corr(*Number Baths*, *Distance*) ≈ 0.5). This property of the data implies that
20. Collinarity increases standard errors of the slopes in the fitted model.
21. These data violate the assumption of independent in the MRM.
22. The model requires an interaction between *Number Baths* and *Distance.*
23. Either *Number Baths* or *Distance* should be removed from the fitted model.
24. Children need to use bathrooms prior to taking longer bus rides to school.
25. The value of the *R*2 statistic implies that
26. The model accurately predicts about 85% of the home prices.
27. The model poorly predicts about 85% of the home prices.
28. The correlation between predicted and actual prices is about 0.92.
29. The prices of 85% of homes are within ±2 RMSE of the predicted prices.
30. Removing 15% of the homes would produce a perfect fit.



1. The leverage plot shown above indicates that
2. The model requires an interaction term to model *Number Baths*.
3. The partial slope for *Number Baths* is significant.
4. *Number Baths* should be transformed to a log scale.
5. A leveraged outlier has reduced the partial slope for *Number Baths*.
6. The simple regression of *Price/SqFt* on *Number* *Baths* has significant slope.
7. A homeowner is considering converting an extra room in a home into a bathroom. Assuming the conditions of the MRM, this model implies that this conversion will
8. Not significantly increase the selling price.
9. Significantly increase the selling price, but only in the city.
10. Significantly increase the selling price, but only in suburban locations.
11. Significantly increase the selling price, with an amount varying by location.
12. Significantly increase the selling price, but only in rural locations.



1. The plot of the residuals from this model shown above indicates that
2. A cluster of homes has higher predicted prices than elsewhere.
3. Rural homes have higher predicted prices than in other locations.
4. The fitted residuals lack constant variation.
5. Suburban have more varying prices than homes elsewhere.
6. The underlying model errors violate the assumption of independent observations.
7. A reasonable next step in building a model for the price per square foot would be to
8. Remove both interactions to simplify the interpretation of the model.
9. Remove the indicator variable *Location[City].*
10. Remove the component, *Distance*\**Location*[*City*].
11. Remove the categorical variable *Location*.
12. Remove the interaction *Distance*\**Location*.