Statistics Waiver Exam

August XX, 2013

Bring your Penn student ID, #2 pencils, eraser, and calculator to the exam. You may also bring *one page of handwritten notes* (8.5”x11” or A4, using both sides as you like) to the exam.

When you receive an answer sheet *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 **50 questions**.  
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 basic **calculator** or graphing calculator.  
No laptops, phones, or computers are allowed.

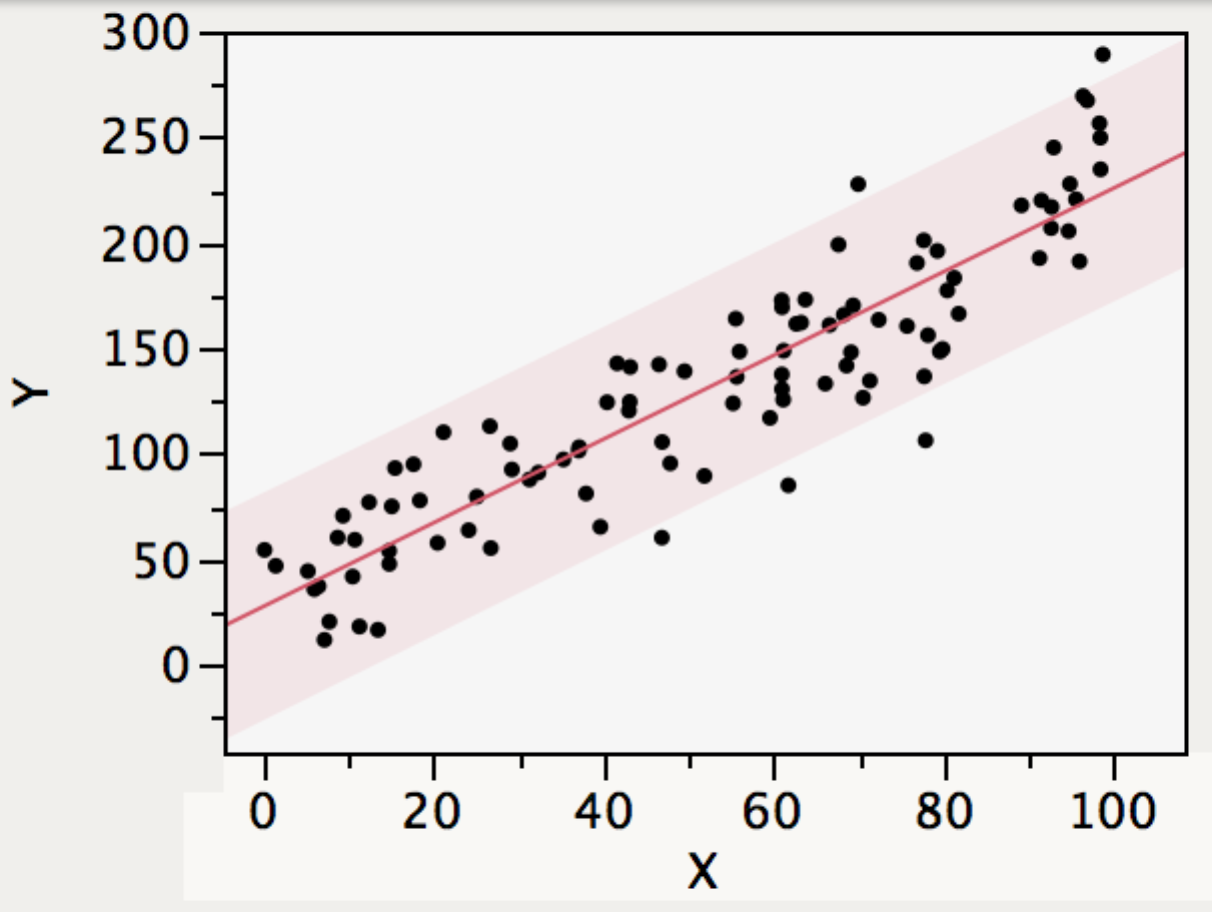
You have **two hours** for the exam. The **computer output** associated with one or more items should be considered an essential part of the questions. Throughout, the word “significant” implies “statistically significant”. The abbreviation SRM stands for standard definition of a ‘simple regression model’;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

This exam has XX questions. Your **score** is the number of correct answers. The XX 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.*

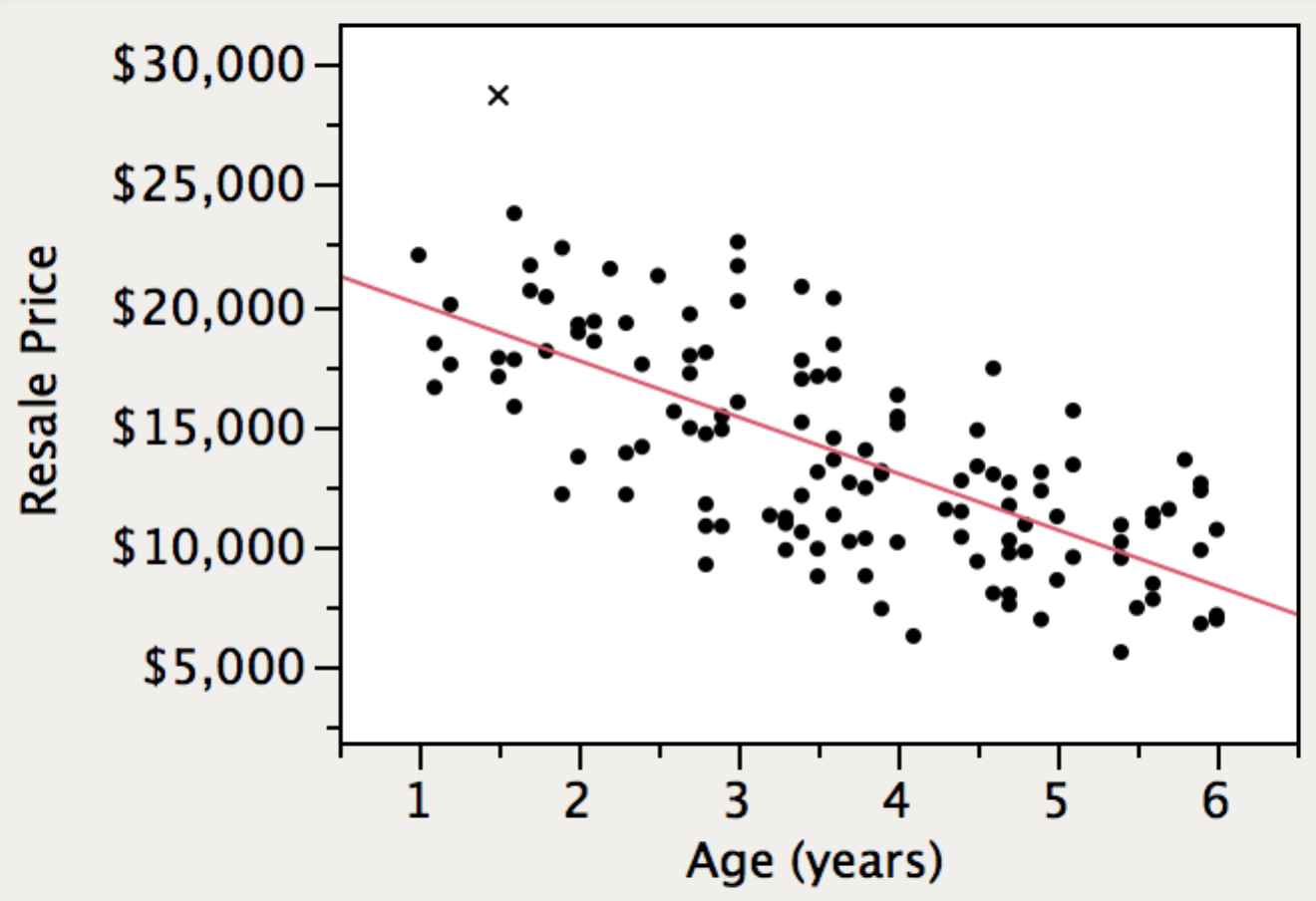
1. Mark the answer to question #1 on your answer form **A**.
2. The alpha-level of a test of the null hypothesis *H*0 versus the alternative hypothesis *H*a is the
   1. **Probability of rejecting *H*0 when *H*0 is true.**
   2. Probability of rejecting *H*a when *H*a is true.
   3. Probability of correctly rejecting *H*a when *H*a is false.
   4. Probability of correctly rejecting *H*0 when *H*0 is false.
   5. Probability that the test produces the correct decision.
3. The *p*-value of 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 the probability of rejecting *H*0 when *H*0 is true.
   5. Is small when the assumptions of the test are not met.
4. The estimated 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. The calculation of 95% confidence intervals relies on the Central Limit Theorem which implies that, when drawing simple random samples from a population with mean *μ* of sufficient size,
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 become closer to *μ*.
16. A manufacturer produces electronic circuits whose output voltage is normally distributed with mean 5 volts and standard deviation 0.05 volts. 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 evenly distributed.
41. Scatterplots of Y versus each explanatory variable are linear.



(Q 12 – 14) The figure shown to the left displays the fit of a simple regression to *n*=100 cases, with the shaded region denoting 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 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. Implies 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. Indicates 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.**

**(QXX-XX)** Representing a large auto dealer, a buyer attends car auctions. To help with the bidding, the buyer built a fit a simple regression to predict the resale value purchased at the auction to data that describe *n* = 125 prior sales of a specific model of cars. The accompanying output describes 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 type 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 what proportion of 2-year old cars of this type has resale value larger than $21,000?
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 type of car is sold in two model types that represent collections of options, denoted SE and R. It is believed that these types depreciate at different rates. The model types are held in the categorical variable *Type*. To investigate whether types depreciate at statistically significantly different rates are, 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 a categorical variable *Type* to the model.
41. **Add a categorical variable *Type* and the 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.**

**(QXX-XX)** A retail food market used data obtained from its loyalty shopper program to estimate the effects of sending coupons to customers. These data describe purchase amounts (in dollars) made by 130 households; households range in size from 1 to 6 persons, indicated by the variable *Household Size*. The data also include a variable *Coupon* that indicates whether the customer took advantage of a discount coupon sent to households that participate in the loyalty shopper program.

|  |  |
| --- | --- |
| RSquare | 0.636951 |
| Root Mean Square Error | 21.95849 |
| *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\* |

**Parameter Estimates**

| **Term** | **Estimate** | **Std Error** | **t Ratio** | **Prob>|t|** |
| --- | --- | --- | --- | --- |
| Intercept | 33.541092 | 5.170663 | 6.49 | <.0001\* |
| Household Size | 22.638643 | 1.689984 | 13.40 | <.0001\* |
| Coupon[No] | 9.6815015 | 5.170663 | 1.87 | 0.0635 |
| Coupon[No]\*Household Size | -4.873977 | 1.689984 | -2.88 | 0.0046\* |

1. The scatterplot indicates that
2. A regression analysis
3. with the number of newspaper ads.
4. It was additional minute
5. Is significantly
6. commercial is shown.
7. Given a than others?
8. Yes, sales are
9. of *Sales* by *Region*.
10. If the type sales would on average
11. Increase $12,620.
13. Increase $38,534.
14. Two average sales
15. Are higher
16. 123,800.
17. An that
18. The amount of spending on
19. comparable.
20. The that
21. These data do not
22. clusters.
23. Which inappropriate?)
24. Add
25. this analysis.

**(Q39-44)** A seller of *ntacts*).

|  |  |
| --- | --- |
| *R*2 | 0.842824 |
| Root Mean Square Error | 6.21078 |

1. The out
2. 16 contacts.
3. None of the above.
4. Based about
5. The same number of contacts as those shown on site G.
6. 3.7 more contacts than those shown on site G.
7. The about
8. 18 more contacts.
9. 0 more contacts.
10. The be about
11. Not statistically
12. in the presence of an interaction.
13. An should
14. Inspect
15. two sites.
16. Because is false?
17. The
18. distributed.

**(Q45-50)** A seller of children’s toys has been studying the allocation of its historical advertising between printed advertisements, such as those mailed directly to consumers and in newspapers, and television advertising, particularly concentrated during Saturday morning cartoons. The data give the annual sales of its products and the spending on television and printed advertisements over the past 30 years. All amounts are expressed on the scale of the natural log of US dollars. (The natural log is the logarithm to base e). The output shows two simple regressions and a multiple regression.

Log(Sales) = 21.8 – 0.28 Log(Print Ads) Log(Sales) = 12.0 + 0.40 Log(TV Ads)

Standard error for slope = 0.12 Standard error for slope = 0.08  
SD(residuals) = 0.50 R2=0.19 SD(residuals) = 0.40 R2=0.49

**Mult Regression, Response Log(Sales)**

RSquare 0.58  
SD(residuals) 0.37

**Parameter Estimates**

**Term Estimate Std Error t Ratio Prob>|t| VIF**Intercept 3.89 3.69 1.05 0.3020 0.00  
Log(Print Ads) 0.33 0.15 2.28 0.0309 2.89  
Log(TV Ads) 0.63 0.12 5.17 <.0001 2.89

1. The estimated slope in the regression of Log(Sales) on Log(TV Ads) implies that, on average,
2. Changes in spending for television ads have no significant effect upon sales.
3. For each additional dollar spent on television ads, sales increase by $0.40.
4. For each additional dollar spent on television ads, sales increase by 0.40%.
5. For each 1% increase in dollars spent on television ads, sales increase by 0.40%.
6. For each 1% increase in dollars spent on television ads, sales increase by $40.
7. Given the standard assumptions for the regression of Log(Sales) on Log(TV Ads), if the seller spends $1.2 million on television advertising, then the probability of its sales being less than $20 million is
8. Approximately 2.5%.
9. Approximately 5%.
10. Approximately 16%.
11. Approximately 33%.
12. More than 50%.
13. Does the estimated multiple regression explain statistically significantly more variation in the log of sales than the model which uses only the log of the television advertising? (Assume that these models meet the required assumptions of regression.)
14. No, because the change in *R*2 is too small to be useful.
15. No, because the *t*-statistic for the slope is smaller in the multiple regression.
16. No, the SD of the residuals is not much smaller than using Log(TV Adv) alone.
17. Yes, Log(Print Ads) adds a significant improvement.
18. Cannot tell without further output.
19. From the shown regression output indicates that if the seller retains the current level of advertising and increases the amount of printed advertising next year that it can expect (assume the conditions of regression modeling are satisfied)
20. Sales to fall by a statistically significant amount.
21. Sales to fall, but not by a statistically significant amount.
22. Sales to remain at the current level.
23. Sales to increase, but not by a statistically significant amount.
24. Sales to increase, by a statistically significant amount.
25. The partial regression leverage plot from the multiple regression shown immediately below and to the right indicates that
26. Several leveraged observations distort the   
    slope of Log(TV ads).
27. Log(TV ads) makes a significant contribution  
     to the model.
28. Log(TV ads) does not make an significant   
    contribution to the model.
29. The data are contaminated by outliers because   
    many points lie outside the dashed bands.
30. The fitted model omits time trends that would   
    improve prediction.
31. The plot of the residuals from the multiple regression shown immediately below and to the right indicates
32. Larger samples are needed in order to find   
    important effects.
33. The data are autocorrelated.
34. The data are heteroscedastic.
35. The data are not normally distributed.
36. No deviation from the usual assumptions.