Name _____

Spring 2016 **STAT 401** **Final**
(100 points)

Instructions:

- Full credit will be given only if you show your work.
- The questions are not necessarily ordered from easiest to hardest.
- You are allowed to use any resource except aid from another individual.
- Aid from another individual, will automatically earn you a 0.

Suppose the following summary statistics are available for a given data set.

```
length(x); mean(x); sd(x)

## [1] 100
## [1] -0.03654185
## [1] 0.8801217

length(y); mean(y); sd(y)

## [1] 100
## [1] -0.2804129
## [1] 1.176844

cor(x,y)

## [1] 0.7047403
```

Assume the model $y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$. Use the summary statistics above to calculate the following quantities.

1. Maximum likelihood estimate (MLE) for β_0

2. MLE for β_1

3. MLE for σ^2

4. Coefficient of variation R^2

5. Standard error for $\hat{\beta}_1$

State the 4 assumptions in a simple linear regression model and how you would evaluate these assumptions.

- 1

- 2

- 3

- 4

For the questions on this page and the following page, use the code and output on the `R Code - Math Scores` page.

1. How many total observations were used in this analysis?

2. Write down the model used in this analysis making sure to define any notation you introduce.

3. Provide an interpretation for the following quantities.

   (a) 9.3629

   (b) 1.6006

   (c) 4.6012

   (d) 0.4335

4. Construct a 95% confidence/credible interval for the effect of having calculus compared to only having algebra after adjusting for sex.

5. Construct a 95% prediction interval for the score of the next male student who has taken geometry. If you cannot derive the interval, then explain what you would need and why you don't have it.

6. Explain why these data are insufficient to claim that, genetically, women are worse than men at math.

Using the `ex1220` data set in the `Sleuth3` R package, fit a multiple regression model with log of the total number of observed species as the response and area, elevation, and their interaction as the explanatory variables. Answer the following questions based on this model.

1. Provide estimates of all βs in this model.

2. Provide an estimate for the effect of a 100 m increase in elevation when the area is 200 km^2 on the total number of observed species.

3. Aside from considering alternative explanatory variables, explain why this is a poor model for the data.

R Code - Math Scores

```
d %>% group_by(Sex,HighestMath) %>%
  summarize(n = n(), mean = mean(Score), sd = sd(Score))

## # A tibble: 6 x 5
## # Groups:   Sex [?]
##   Sex    HighestMath     n  mean    sd
##   <fct>  <fct>       <int> <dbl> <dbl>
## 1 female Algebra        82  9.07  4.19
## 2 female Geometry      387 14.0   5.00
## 3 female Calculus       54 24.6   4.85
## 4 male   Algebra        48 11.5   5.09
## 5 male   Geometry      223 15.6   4.89
## 6 male   Calculus       67 25.4   5.55

m <- lm(Score ~ Sex+HighestMath, data = d)
summary(m)

##
## Call:
## lm(formula = Score ~ Sex + HighestMath, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.9641  -3.3629   0.0359   3.4354  14.0366
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           9.3629     0.4519  20.718  < 2e-16 ***
## Sexmale               1.6006     0.3479   4.601 4.84e-06 ***
## HighestMathGeometry   4.6012     0.4772   9.642  < 2e-16 ***
## HighestMathCalculus  14.8252     0.6273  23.633  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.94 on 857 degrees of freedom
## Multiple R-squared:  0.4355,Adjusted R-squared:  0.4335
## F-statistic: 220.4 on 3 and 857 DF,  p-value: < 2.2e-16
```