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

   Answer: To find the MLE, we have $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$. See below for $\hat{\beta}_1$.

   ```
   mean(y) - cor(x,y)*sd(y)/sd(x) * mean(x)

   ## [1] -0.2459782
   ```

2. MLE for β_1

   Answer:

   $$\hat{\beta}_1 = \frac{SXY}{SXX} = \frac{r_{XY} s_X s_Y}{s_X^2} = \frac{r_{XY} s_Y}{s_X}$$

   ```
   cor(x,y)*sd(y)/sd(x)

   ## [1] 0.9423351
   ```

3. MLE for σ^2

   Answer: We have $\sigma^2 = SSE/(n-2)$ and $SSE = (1 - R^2)SYY = (1 - R^2) \times (n-1) * s_Y^2$.

   ```
   n = length(y)
   SSE = (1-cor(x,y)^2)*(n-1)*sd(y)^2
   (sigma2 <- SSE/(n-2))

   ## [1] 0.7042218
   ```

4. Coefficient of variation R^2

   Answer: The coefficient of variation is the correlation squared, i.e.

```
cor(x,y)^2

## [1] 0.4966589
```

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

   Answer: The standard error for $\hat{\beta}_1$ is $\hat{\sigma}\sqrt{1/(n-1)s_X^2}$.

```
sqrt(sigma2)*sqrt(1/((n-1)*sd(x)^2))

## [1] 0.09582843
```

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

- 1

  Answer: There is a linear relationship between the expected response and the explanatory variable. The best way to evaluate this assumption is to plot the response vs the explanatory variable or the residuals vs the explanatory variable. If these show curvature, then there is a departure from this assumption.

- 2

  Answer: The errors are normally distributed. In the normal Q-Q plot normality can be determined by how well the points comparing standardized residuals versus the theoretical quantiles fall along the line. If points don't generally fall along the line then normality is violated.

- 3

  Answer: The errors errors have a constant variance. If a plot of residuals vs fitted values shows a funnel pattern or if the (square root of absolute values) of standardized residuals vs fitted values shows an increasing/decreasing trend, these are both indications of lack of constant variance.

- 4

  Answer: The errors are independent. This is the most difficult assuption to evaluate, but one plot that should be done is a residuals vs row index (or time, if available) plot. If this plot shows a pattern, then there is a violation of this assumption.

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?

   Answer: 861

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

   Answer: Define

   - Y_i be the IQ score for student i
   - S_i be the sex for student i (Male/Female)
   - HM_i be the highest math taken by student i (Algebra/Geometry/Calculus)

   The model is

   $$Y_i \overset{ind}{\sim} N(\mu_i, \sigma^2) \qquad \mu_i = \beta_0 + \beta_1 \mathrm{I}(S_i = male) + \beta_2 \mathrm{I}(HM_i = Geometry) + \beta_3 \mathrm{I}(HM_i = Calculus)$$

3. Provide an interpretation for the following quantities.

   (a) 9.3629

   Answer: This is the estimated mean IQ score for female students whose highest math is algebra.

   (b) 1.6006

   Answer: This is the estimated difference in IQ score between males and females across all levels of highest math.

   (c) 4.6012

   Answer: This is the estimated difference in IQ score between those whose highest math is geometry compared to algebra across both sexes.

   (d) 0.4335

   Answer: The model accounts for those amount of variation in IQ score.

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

   Answer:

$$\hat{\beta}_3 \pm t_{0.975,857} SE(\hat{\beta}_3) = 14.8252 \pm 1.962736 \times 0.6273 = (13.59398, 16.05642) \approx (14, 16).$$

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.

   Answer: The point estimate here is $\beta_0 + \beta_1 + \beta_2$ and the t critical value is the same as in the previous problem, i.e. $t_{0.975,857} = 1.962736$. The difficulty is in obtaining a standard error for $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2$ and this standard error is a function of the explanatory variable values (male and geometry) and the means of these levels. Although we could probably derive it based on the table provided, it would take a while.

   In addition, the standard error for prediction is slightly different than the standard error of the mean. If $SE(\hat{\mu})$ is the standard error of the mean where $\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2$, then $SE(Pred) = \hat{\sigma}\sqrt{1 + SE(\hat{\mu})^2/\hat{\sigma}^2}$.

   Thus our interval is

$$\begin{aligned} &\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 \pm t_{0.975,857} SE(Pred) \\ &= 9.3629 + 1.6006 + 4.6012 \pm 1.962736 SE(Pred) \\ &= 15.5647 \pm 1.962736 \hat{\sigma}\sqrt{1 + SE(\hat{\mu})^2/\hat{\sigma}^2}. \end{aligned}$$

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

   Answer: The main point is that sex is not randomized. So, despite the positive and significant β_1, this is not evidence of a causal relationship. One issue here is that our environment in our educational system and elsewhere may play a role in the process that results in, on average, women scoring worse on standardized tests. In addition, the standardized tests themselves may be biased in their ability to estimate mathematically ability.

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.

   Answer:

```
(m <- lm(log(Total) ~ Area*Elev, data = Sleuth3::ex1220))

##
## Call:
## lm(formula = log(Total) ~ Area * Elev, data = Sleuth3::ex1220)
##
## Coefficients:
## (Intercept)         Area         Elev    Area:Elev
##   2.544e+00    5.117e-03    1.949e-03   -3.022e-06
```

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.

   Answer: The effect of a 100 m increase in elevation when the are is 200 km^2 is $e^{100\beta_2+200\times100\beta_3}$. In R, we can campute this with.

```
exp(sum(coef(m)*c(0,0,100,200*100)))

## [1] 1.143926
```

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

   Answer: Observation 16 has huge leverage and Cook's distance due to its extremely large area and elevation.

```
library("dplyr")
Sleuth3::ex1220 %>%
  mutate(leverage = hatvalues(m),
         cooksd   = cooks.distance(m)) %>%
  filter(Island == "Isabela")

##    Island Total Native    Area Elev DistNear DistSc AreaNear  leverage
## 1 Isabela   347     89 4669.32 1707      0.7   28.1   634.49 0.9948369
##     cooksd
## 1 237.0484

# summary(Sleuth3::ex1220 %>% select(Total, Area, Elev))
```

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
```