Exam 3 Solutions

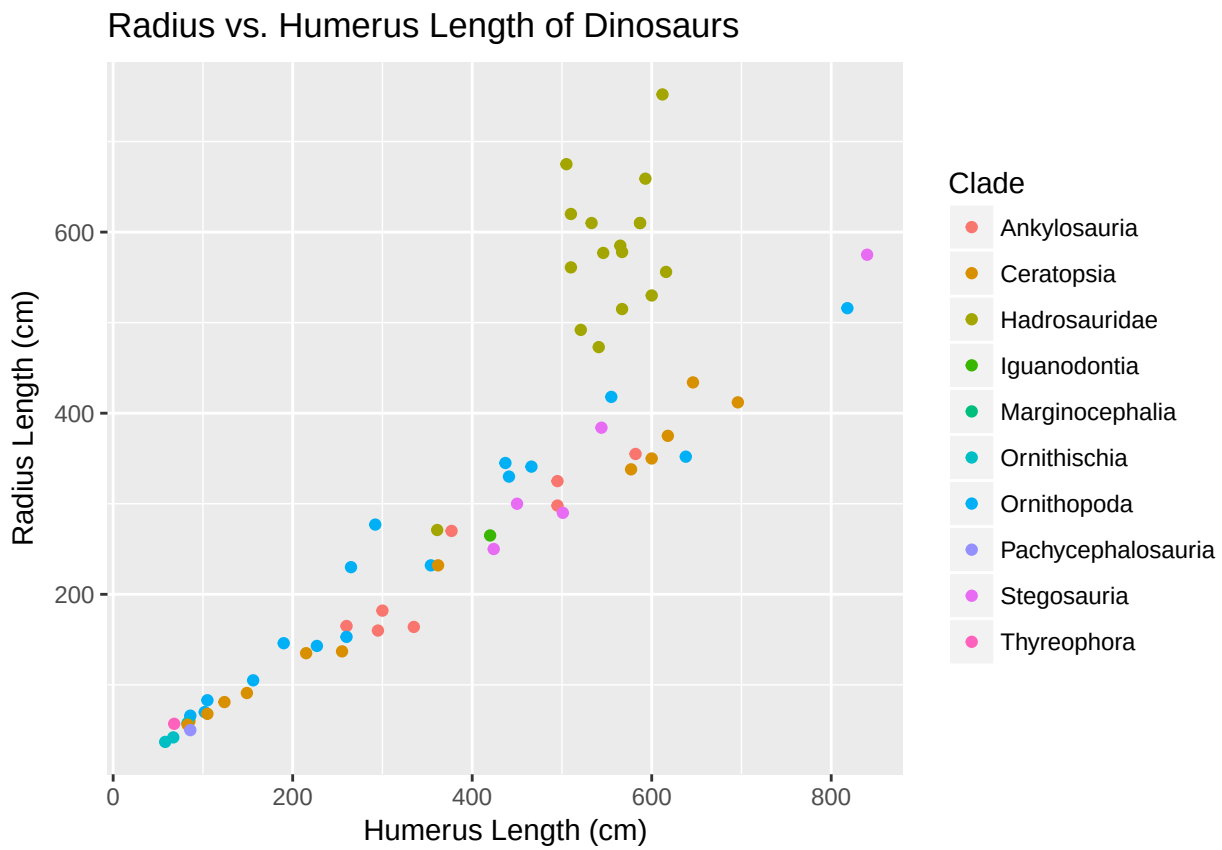
Problem 1

A) [4 pts] Create a scatterplot of radius length (response) vs. humerus length (predictor). Give every Clade its own color and/or shape (hint: you can just use the default colors/shapes). Don't forget to give the plot appropriate title/axis labels. Paste the resulting scatterplot, as well as any R code, below.

```
# Read the data set into RStudio
library(readr)
dinosaur <- read_csv("~/Math 338 Summer 2017/dinosaur-bone-lengths.csv")
head(dinosaur)  # view the first few entries in the data set
```

```
## # A tibble: 6 × 11
##           Clade        GenusFinal SpeciesFinal    Specimen Scapula Humerus
##           <chr>             <chr>        <chr>       <chr>   <int>   <int>
## 1   Ornithopoda     Abrictosaurus      consors NHM RU B.54      NA      NA
## 2  Ankylosauria        Aletopelta      coombsi SDNHM 33909      NA      NA
## 3 Hadrosauridae Brachylophosaurus   canadensis    CMN 8893     857     593
## 4   Ornithopoda      Camptosaurus       dispar   USNM 4282     476     354
## 5    Ceratopsia      Centrosaurus      apertus   AMNH 5351     700     600
## 6    Ceratopsia        Cerasinops   hodgskissi     MOR 300     277     260
## # ... with 5 more variables: Radius <int>, `MC III L` <dbl>, Femur <int>,
## #   Tibia <int>, `MT III L` <dbl>
```

```
# Make the plot
library(ggplot2)
dinosaur_plot <- ggplot(data = dinosaur, mapping = aes(x = Humerus,
    y = Radius, color = Clade)) + geom_point() + labs(x = "Humerus Length (cm)",
    y = "Radius Length (cm)", title = "Radius vs. Humerus Length of Dinosaurs")
print(dinosaur_plot)
```



B) [1.5 pts] Based only on your scatterplot, do you believe humerus length and radius length to be positively correlated, negatively correlated, or uncorrelated? Justify your answer.

Positively correlated, because as humerus length increases, radius length also increases.

C) [1.5 pts] Do you believe there is a causal relationship between humerus length and radius length? Why or why not?

No, there is not a causal relationship between humerus length and radius length. There is not even an explanatory-response relationship between humerus length and radius length, since it does not make sense that changes in the length of one bone would explain changes in the length of the other bone (but not vice versa).

D) [1 pt] The data appear to follow a linear relationship, except for dinosaurs in one Clade. Which clade?

Hadrosauridae.

Problem 2

Remove from the data set the clade that was your answer to problem 1D, and save the new data set.

```
library(dplyr)
dinosaur_new <- dinosaur %>% filter(Clade != "Hadrosauridae")
```

A) [3 pts] Report the equation of the least-squares regression line that best fits the relationship between humerus length (predictor) and radius length (response). Include the Parameter Estimates (Coefficients:) table and any relevant R code below.

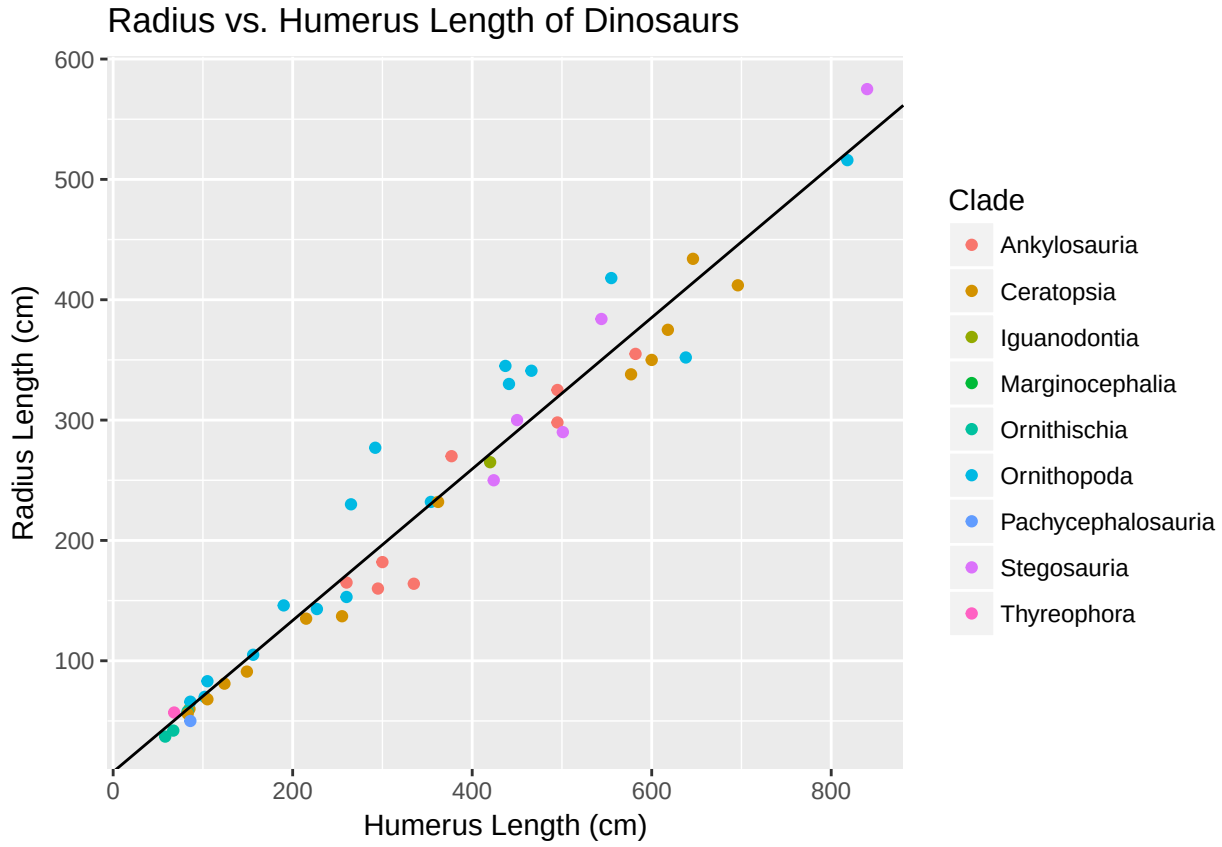
```
lm1 <- lm(Radius ~ Humerus, data = dinosaur_new)
summary(lm1)
```

```
##
## Call:
## lm(formula = Radius ~ Humerus, data = dinosaur_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.052 -18.377  -5.187   9.245  85.637
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.64857    8.28319   0.923    0.361
## Humerus      0.62916    0.02039  30.860   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.18 on 46 degrees of freedom
##   (23 observations deleted due to missingness)
## Multiple R-squared:  0.9539, Adjusted R-squared:  0.9529
## F-statistic: 952.3 on 1 and 46 DF,  p-value: < 2.2e-16
```

Radius = 7.65 + 0.629(Humerus)

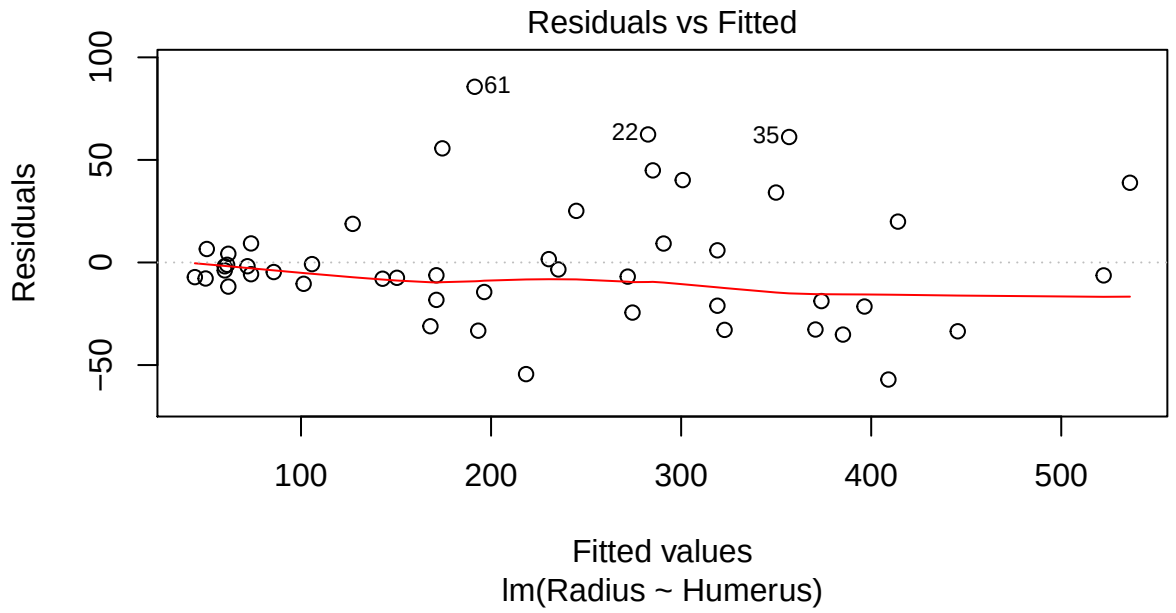
B) [1.5 pts] In the space below, paste the following three plots (along with any R code used to create them). Label which plot is which.

```
coefs_2B <- coef(lm1)
scatterplot_2B <- ggplot(data = dinosaur_new, mapping = aes(x = Humerus,
    y = Radius, color = Clade)) + geom_point() + geom_abline(intercept = coefs_2B[1],
    slope = coefs_2B[2]) + labs(x = "Humerus Length (cm)",
    y = "Radius Length (cm)", title = "Radius vs. Humerus Length of Dinosaurs")
print(scatterplot_2B)
```



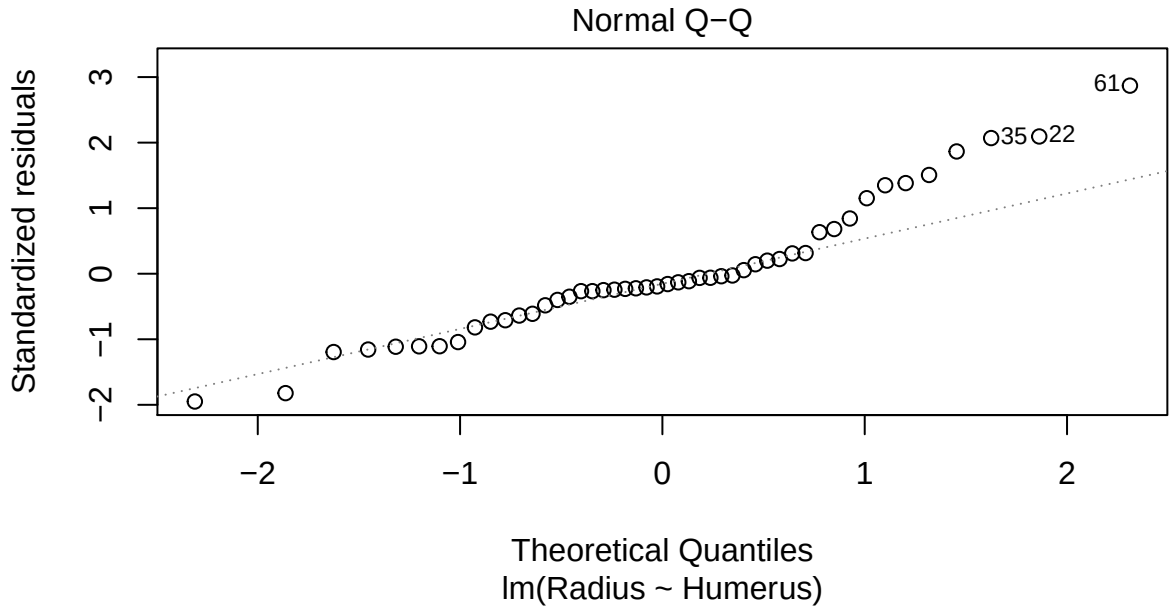
Above: Scatter plot

```
plot(lm1, which = 1)
```



Above: Residual plot

```
plot(lm1, which = 2)
```



Above: Normal q-q plot

C) [2 pts] Do the diagnostic plots in part 2B suggest that we are okay to perform inference using this model? Justify your answer by checking each assumption of the model.

Clearly, the linear model is appropriate. However, I have some reservations about the remainder of the assumptions. The mean of the residuals (as shown by the red line) is consistently getting more negative, and the absolute value of the residuals tends to be much smaller at very low fitted values than at higher fitted values. Therefore, it is difficult to argue that the assumptions of zero mean and constant standard deviation are met. For normality, the normal q-q plot is problematic on the right side of the graph. Either of those arguments would suggest we are not okay to perform inference using this model. Or, you could argue that these problems are not problematic enough to suggest inference is inappropriate.

D) [3 pts] Regardless of your answer to part 2C, report and interpret a 95% confidence interval for the true slope of the linear relationship between radius length and humerus length. Include all relevant (code and) output below.

```
confint(lm1)
```

```
##                   2.5 %     97.5 %
## (Intercept) -9.0246368 24.3217765
## Humerus      0.5881215  0.6701975
```

The confidence interval for the slope is (0.588, 0.670). We are 95% confident that for every 1 cm increase in humerus length, the population mean radius length increases by between 0.588 and 0.670 cm.

E) [1 pt] What is the correlation between radius length and humerus length?

```
cor(dinosaur_new$Humerus, dinosaur_new$Radius, use = "pairwise.complete.obs")
```

```
## [1] 0.9766901
```

By either running the above code or reading the R^2 value from the output in part 2A and taking the square root, we find that the correlation is about 0.977.

F) [2 pts] Predict the radius length of a dinosaur whose humerus has length 400 cm. If you use Rguroo or RStudio to make the prediction, include all relevant (code and) output below. Otherwise, show your work.

```
dino_humerus <- data.frame(Humerus = 400)
predict(lm1, newdata = dino_humerus)
```

```
##        1
## 259.3124
```

We predict that a dinosaur with humerus length 400 cm will have a radius length of about 259 cm.

G) [1.5 pt] Was the prediction in part 2D an example of interpolation or extrapolation? Justify your answer.

```
summary(dinosaur_new$Humerus)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    34.0   105.0   295.0   324.5   495.0   840.0      14
```

Interpolation, because 400 is between the minimum and maximum values of Humerus in the data set.

H) [2 pts] Would a dinosaur with a humerus length of 400 cm and a radius length of 250 cm have high influence, a high residual, both, or neither? Justify your answer.

A dinosaur with a humerus length of 400 cm and a radius length of 250 cm would not have a high influence because 400 cm is reasonably close to the mean of humerus length and definitely not an outlier. It would also not have a high residual because 250 cm is close to the predicted value of 259 cm; a residual of -9.31 is not that high compared to the values seen in the residual plot from part 2B.

Problem 3

In the data set you used in Problem 2, add two new variables, log_radius and log_humerus, that represent the natural logarithm of radius length and humerus length, and save the new data set.

```
dinosaur_log_rh <- dinosaur_new %>% mutate(log_radius = log(Radius),
    log_humerus = log(Humerus))
```

A) [3 pts] Report the equation of the least-squares regression line that best fits the relationship between natural logarithm of humerus length (predictor) and natural logarithm of radius length (response). Include the Parameter Estimates (Coefficients:) table and any relevant R code below.

```
lm2 <- lm(log_radius ~ log_humerus, data = dinosaur_log_rh)
summary(lm2)
```

```
##
## Call:
## lm(formula = log_radius ~ log_humerus, data = dinosaur_log_rh)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28653 -0.08655 -0.02081  0.08624  0.36993
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.21359    0.13563  -1.575    0.122
## log_humerus  0.96317    0.02402  40.105   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1281 on 46 degrees of freedom
##   (23 observations deleted due to missingness)
## Multiple R-squared:  0.9722, Adjusted R-squared:  0.9716
## F-statistic:  1608 on 1 and 46 DF,  p-value: < 2.2e-16
```

log(Radius) = -0.214 + 0.963 log(Humerus)

B) [1 pt] In this new model, what is the value of our estimate of σ, the population standard deviation of the residuals?

Our estimate of σ is s, the residual standard error, which has a value of 0.128 according to the summary above.

C) [4 pts] Suppose that the skeleton of a new dinosaur species is unearthed. The humerus is measured to be 600 cm long, but the radius is not found. Using this new model, report and interpret a 95% prediction interval for the radius length of the new dinosaur. Include all relevant (code and) output below.

```
dino_log <- data.frame(log_humerus = log(600))
predict(lm2, newdata = dino_log, interval = "prediction")
```

```
##        fit      lwr     upr
## 1 5.947734 5.684338 6.21113
```

The 95% prediction interval is then $(e^{5.68}, e^{6.21})$ or (294.2, 498.3).

```
dino_log <- data.frame(log_humerus = log(600))
dino_pi <- predict(lm2, newdata = dino_log, interval = "prediction")
exp(dino_pi)
```

```
##        fit      lwr     upr
## 1 382.8847 294.2229 498.264
```

We are 95% confident that this new dinosaur, with a 600 cm long humerus, will have a radius length between 294.2 and 498.3 cm.

Problem 4

Using the original data set (dinosaur-bone-lengths), create a multiple linear regression model, with Femur as the response and Humerus, Tibia, and Scapula as predictors (in that order).

```
lm3 <- lm(Femur ~ Humerus + Tibia + Scapula, data = dinosaur)
```

A) [3 pts] Paste the Parameter Estimates (Coefficients:) table below, and use it to write out the full least-squares regression equation for this model. Include any relevant code used to make the table.

```
summary(lm3)
```

```
##
## Call:
## lm(formula = Femur ~ Humerus + Tibia + Scapula, data = dinosaur)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -187.870  -24.818   -1.951   29.187  281.607
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -9.0309    24.6267  -0.367 0.715632
## Humerus       0.1684     0.1844   0.913 0.366207
## Tibia         0.3862     0.1063   3.634 0.000741 ***
## Scapula       0.7111     0.1636   4.345 8.35e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 74.74 on 43 degrees of freedom
##   (42 observations deleted due to missingness)
## Multiple R-squared:  0.9678, Adjusted R-squared:  0.9655
## F-statistic: 430.5 on 3 and 43 DF,  p-value: < 2.2e-16
```

Femur = -9.03 + 0.168 (Humerus) + 0.386 (Tibia) + 0.711 (Scapula)

B) [2 pts] Interpret the slope corresponding to the variable Tibia.

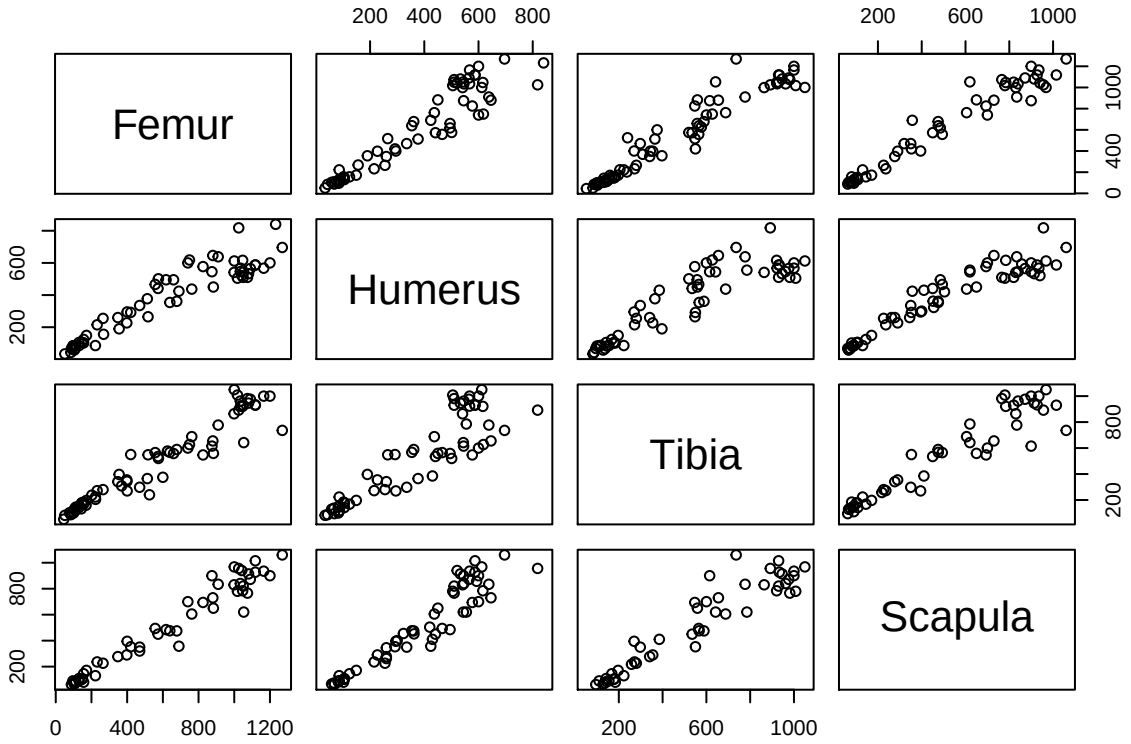
For every 1 cm increase in tibia length, we expect the mean femur length to increase by 0.386 cm, holding the values of Humerus and Scapula constant.

C) [1.5 pts] Overall, is this model significant at the 1% significance level? How do you know? Paste below any output from Rguroo/RStudio that supports your answer.

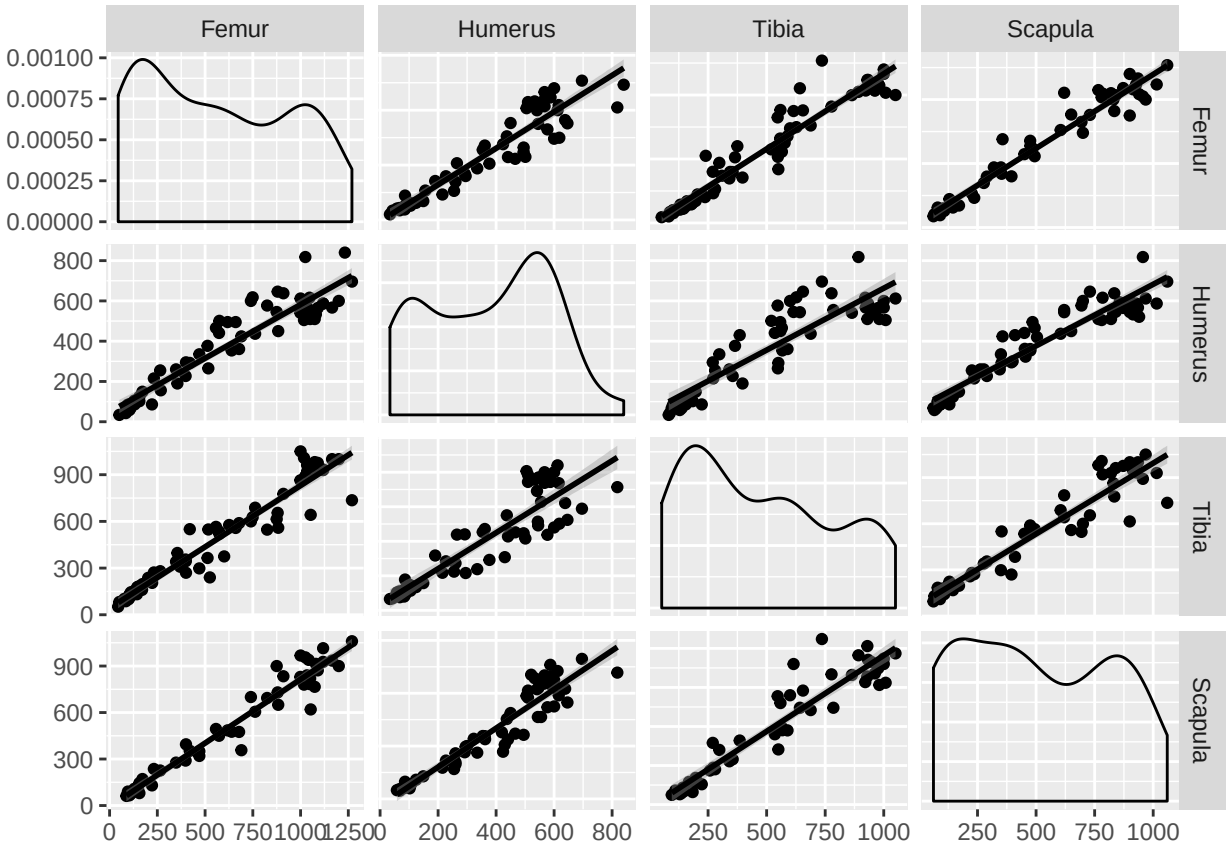
From the bottom line of the summary output above, the p-value for the ANOVA F test for multiple linear regression is less than 2×10^{-16}, so the overall model is significant at the 1% significance level. (You can paste the summary again, or just refer to the summary from Part 4A.)

D) [2 pts] Is there evidence of collinearity? If so, paste below an appropriate table or plot and explain how the table/plot shows that collinearity exists. If not, paste below an appropriate table or plot and explain how the table/plot shows that it does not exist.

```
dinosaur_femur <- dinosaur %>% select(Femur, Humerus,
    Tibia, Scapula)
pairs(dinosaur_femur)
```



```
library(GGally)
ggpairs(dinosaur_femur, lower = list(continuous = "smooth"),
    upper = list(continuous = "smooth"))
```



From either of the above two plots, it is clear that we have linear relationships between Humerus and Tibia, Humerus and Scapula, and Tibia and Scapula. Therefore there is evidence of collinearity.

E) [2 pts] If you were to perform backward selection using this initial model, which variable would you remove first? Why would you remove that variable?

We would remove Humerus first, because it is the only one with a nonsignificant p-value, and therefore the least significant predictor. Alternatively, we could run the stepwise regression command to verify that Humerus is the first (and only) variable to be removed:

```
step(lm3)
```

```
## Start:  AIC=409.33
## Femur ~ Humerus + Tibia + Scapula
##
##           Df Sum of Sq    RSS    AIC
## - Humerus  1      4658 244831 408.23
## <none>                 240173 409.33
## - Tibia    1     73744 313917 419.92
## - Scapula  1    105457 345630 424.44
##
## Step:  AIC=408.23
## Femur ~ Tibia + Scapula
##
##           Df Sum of Sq    RSS    AIC
## <none>                 244831 408.23
## - Tibia    1     70122 314953 418.07
## - Scapula  1    369115 613946 449.44
##
## Call:
## lm(formula = Femur ~ Tibia + Scapula, data = dinosaur)
##
## Coefficients:
## (Intercept)        Tibia      Scapula
##      1.6470       0.3732       0.8280
```

F) [1.5 pts] If you were to perform backward selection using this initial model, would the next model you created have a higher or lower (Multiple) R^2 value? Explain your answer.

As the number of variables, increases R^2 increases. Therefore, removing a variable will make R^2 lower. We could also verify this:

```
lm4 <- lm(Femur ~ Tibia + Scapula, data = dinosaur_femur)
summary(lm4)
```

```
##
## Call:
## lm(formula = Femur ~ Tibia + Scapula, data = dinosaur_femur)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -195.065  -28.950   -5.053   26.580  298.365
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.6470    21.6331   0.076 0.939660
## Tibia         0.3732     0.1051   3.550 0.000931 ***
## Scapula       0.8280     0.1017   8.145  2.5e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 74.59 on 44 degrees of freedom
##   (42 observations deleted due to missingness)
## Multiple R-squared:  0.9672, Adjusted R-squared:  0.9657
## F-statistic: 647.8 on 2 and 44 DF,  p-value: < 2.2e-16
```

The original model has an R^2 value of 0.9678. The new model has an R^2 value of 0.9672. Note that in this instance, the Adjusted R^2 value actually increases when we remove the variable Humerus from the model (from 0.9655 to 0.9657).