For this assignment use the Pulse.txt dataset used in the class examples this week. Use Resting pulse rate as the response variable (Y = Rest). The explanatory variables you will use are:
X₁ = Height (Hgt)
X₂ = Weight (Wgt)
X₃ = Smoke (1 for Smokers, 0 for Nonsmokers)

1. Write the population model for this situation, including any necessary conditions. Use standard notation for all parameters, and use X₁, etc., instead of the names of the variables.

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \]
where the condition is that the \( \epsilon \) are independent and each from a N(0, \( \sigma_\epsilon \)) distribution. (The assumption of a linear relationship between Y and Height and Weight is built into the model.)

2. Fit the regression model using all 3 explanatory variables and show the results of the “summary” command. Use it to answer the following questions:

```r
> Mod.hw<-lm(Rest~Hgt+Wgt+Smoke, data=Pulse)
> summary(Mod.hw)
```

```
Call:
  lm(formula = Rest ~ Hgt + Wgt + Smoke, data = Pulse)

Residuals:
   Min     1Q Median     3Q    Max
-25.68  -5.94  -1.09   5.85  34.29

Coefficients:  Estimate Std. Error t value Pr(>|t|)
(Intercept) 111.02244   14.25754   7.787 2.39e-13 ***
Hgt          -0.60484    0.25644  -2.359  0.01919 *
Wgt          -0.01295    0.03039  -0.426  0.67031
Smoke         5.80259    2.01400   2.881  0.00434 **

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

Residual standard error: 9.543 on 228 degrees of freedom
Multiple R-squared:  0.09193, Adjusted R-squared:  0.07998
F-statistic: 7.694 on 3 and 228 DF,  p-value: 6.414e-05
```

a. Write the (sample) regression equation that would be used to predict resting pulse rate using these 3 explanatory variables. Fill in the numerical values for the coefficients; round each one to 3 decimal places. Use the names of the variables (Hgt, etc.) instead of \( X_i \), etc.

\[ \hat{Y} = 111.022 - 0.605(Hgt) - 0.013(Wgt) + 5.803(Smoke) \]

b. Interpret the value of the coefficient for “Smoke.”

For a smoker and non-smoker of the same weight and height we would predict the smoker to have a resting pulse rate that is about 5.8 beats a minute higher than the non-smoker.

c. Either using R or a calculator, find the predicted resting pulse rate for a smoker who is 65 inches tall and weighs 150 pounds.

\[ \hat{Y} = 111.022 - 0.605(65) - 0.013(150) + 5.803 = 75.55 \text{ beats per minute} \]

Using R you would get 75.56714.
d. State and test the hypotheses to determine whether any one or more of the three predictors are useful in predicting resting pulse rate. Use appropriate parameter notation in your hypotheses (not words).

\[ H_0: \beta_1 = \beta_2 = \beta_3 = 0; \text{ Ha: At least one of } \beta_1, \beta_2, \beta_3 \text{ is not } 0. \]

From R, the test statistic and \( p \)-value are \( F = 7.694, p = 6.414 \times 10^{-5} \). Reject the null hypothesis, and conclude that at least one of the predictors is useful (i.e. at least one of \( \beta_1, \beta_2, \beta_3 \) is not 0).

3. Show the results of the “anova” command for this model, and use it to answer the following questions.

```r
> anova(Mod.hw)
Analysis of Variance Table
Response: Rest
Df  Sum Sq Mean Sq  F value    Pr(>F)
Hgt         1  1346.2 1346.18 14.7814 0.0001566 ***
Wgt         1     0.0    0.03  0.0003 0.9857373
Smoke       1   756.0  755.99  8.3009 0.0043405 **
Residuals 228 20764.5   91.07
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

a. What are the numerical values of SSModel, SSE and SSTotal?

\[
\text{SSModel} = 1346.2 + 0.0 + 756.0 = 2102.2, \quad \text{SSE} = 20764.5, \quad \text{SSTotal} = 2102.2 + 20764.5 = 22866.7
\]

b. If Height was the only predictor in the model, what would be the values of SSModel, SSE and SSTotal? (You do not need to rerun the model to answer this. You can find the appropriate values from the results of the anova command.)

SSModel is given as the SumSq for Hgt, so it is 1346.2. SSTotal remains the same, so it is 22866.7. SSE can be found either as SSTotal – SSModel = 21520.5, or by adding the SumSq for Wgt and Smoke to the old SSE, to get 20764.5 + 0.0 + 756.0 = 21520.5.

4. For this question, you will investigate the role of weight in this situation.

a. State and test the hypotheses to determine whether weight is a significant predictor of resting pulse rate in the absence of any other predictors. Make sure you define any parameters you use in your hypotheses.

There are two ways you can do this. You can run a separate model with just weight or you can use the cor.test command in R. The results will be the same. Using the correlation test, the hypotheses are \( H_0: \rho = 0, \text{ Ha: } \rho \neq 0 \) where \( \rho \) is the population correlation between Rest and Weight. Command and results from R are:

```r
> cor.test(Pulse$Rest,Pulse$Wgt)

Pearson's product-moment correlation
data:  Pulse$Rest and Pulse$Wgt
t = -2.8378, df = 230, p-value = 0.004948 alternative hypothesis: true correlation is not equal to 0
```

With a \( p \)-value of 0.004948, reject the null hypothesis and conclude that weight is a significant predictor of resting pulse rate.

b. State and test the hypotheses to determine whether weight is a significant predictor of resting pulse rate after accounting for height and smoking status. Again make sure you define any parameters used.

For this test, using the model as defined in question 1, the hypotheses are \( H_0: \beta_2 = 0 \) and \( \text{ Ha: } \beta_2 \neq 0 \), but conditional on Height and Smoke being in the model. The test statistic and \( p \)-value are given in the summary; \( t = -0.426, p-value = 0.67031 \). Do not reject the null hypothesis. Weight is not a significant predictor of resting pulse rate after accounting for height and smoking status.
c. Discuss whether your results differed for parts (a) and (b). If you conclude that your results were similar, justify your conclusion statistically. If instead you conclude that they were different, explain why the results would differ in this situation.

The results were very different. The reason is that height and weight are highly correlated, and once height is in the equation, weight does not contribute much additional information in predicting resting pulse rate.

Now you will use a model with only Weight and Smoke as predictors (not Height).

5. Write the population model. (Use variable names). Identify what each coefficient in the model represents. As a hint and to get you started, $\beta_0 =$ the intercept of the regression line relating $Y =$ Rest and $X =$ Weight for non-smokers.

$$Y = \beta_0 + \beta_1(Wgt) + \beta_2(Smoke) + \varepsilon; \beta_1$$ is the slope for the line relating $Y =$ Rest and $X =$ Weight, for both smokers and non-smokers. $\beta_2$ is the additional intercept for smokers, so the intercept for the line for smokers is $Y = \beta_0 + \beta_2$.

6. Run the regression in R, and show the results of the summary command. State and test the hypotheses to determine if the intercepts differ for the regression lines relating $Y =$ Rest and $X =$ Weight for smokers and non-smokers.

```R
> Mod.hw3<-lm(Rest~Wgt+Smoke, data=Pulse)
> summary(Mod.hw3)
```

```
Call:
  lm(formula = Rest ~ Wgt + Smoke, data = Pulse)

Residuals:
     Min      1Q  Median      3Q     Max
-25.872  -6.207  -0.719   5.794  37.128

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  78.24692    3.22061  24.296  < 2e-16 ***
Wgt         -0.06697    0.02017  -3.319  0.00105 **
Smoke       -2.03136    0.21608  -9.392  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.638 on 229 degrees of freedom
Multiple R-squared:  0.06978, Adjusted R-squared:  0.06165
F-statistic:  8.589 on 2 and 229 DF,  p-value: 0.0002531
```

Using the notation in the model in question 5, the hypotheses are $H_0: \beta_2 = 0$ and $H_a: \beta_2 \neq 0$. In words, the null hypothesis is that the additional intercept term for smokers is not needed. The test statistic and $p$-value are found in the row for “Smoke” and are $t = 2.975$, $p$-value = 0.00325. Reject the null hypothesis and conclude that the intercepts are different for smokers and non-smokers.

7. Now write the population model that would include different intercepts and different slopes for smokers and non-smokers in the regression lines relating $Y =$ Rest and $X =$ Weight. Interpret each of the coefficients.

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_1X_2 + \varepsilon.$$ Now $\beta_0$ is the intercept for the line for non-smokers, $\beta_1$ is the slope of the line for non-smokers, $\beta_2$ is the additional intercept for the line for smokers (so the intercept is $\beta_0 + \beta_2$), and $\beta_3$ is the addition to the slope for the line for smokers (so the slope for smokers is $\beta_1 + \beta_3$).

8. Using the notation from your model in question 7, write the hypotheses that would be used to test whether the regression lines are the same (intercepts and slopes) for smokers and non-smokers. (You do not have to carry out the test.)

$H_0: \beta_2 = \beta_3 = 0; H_a: \text{At least one of } \beta_2 \text{ and } \beta_3 \text{ is not 0.}$