1.

(a)

Dependent Variable: WGHT					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2624.670184	2624.670184	137.91	<0.001
Error	96	1827.099916	19.032291		
Corrected total	97	4451.7701			

(b)

The model assumptions behind the model $WGHT = \beta_0 + \beta_1 * TIME + \varepsilon$ are:

(1) existence assumption: assume ε_i has finite first and second moment. In other words, we observe values of random variables with finite variance.

(2) linearity assumption: we assume that the expected values of the weight (WGHT) are linear functions of the average daily exercise times (TIME).

(3) independent assumption: we assume that each element of ε is statistically independent of every other.

(4) homogeneity assumption: we assume that each element of ε has the same variance σ^2.

(5) Gaussian error assumption: we assume that $\varepsilon_i \sim N(0, \sigma^2)$. Note that the normality assumption is needed for the F test, but not needed for the estimates.

(c)

The average daily exercise time is a significant predictor for predicting weight loss. In this specific hypothesis testing, the null hypothesis is that the average daily exercise time is not a significant predictor for predicting weight loss (H_0: $\beta_1 = 0$); and the alternative hypothesis is that the average daily exercise time is a significant predictor for predicting weight loss (H_A: $\beta_1 \neq 0$). According to the ANOVA table, the test statistic $F_{obs} = 137.91$, which follows F distribution with degree of freedom of 1 and 96. The corresponding p-value is less than 0.05. Therefore, we reject the null hypothesis. The result can be interpreted as the following: assuming that the average daily exercise time is not a significant predictor for predicting weight loss, then the probability of

observing the data that are as extreme as ours or more extreme is less than 0.05, which is too small for us to believe that the null hypothesis is true.

(d)

The analysis does not suggest that neither variable is significant, essentially because they are correlated covariates so that the addition of one additional covariate does not provide additional significant information, given the other covariate is in the model. More specifically, the output is for the type III test, which refers to the statistical significance of the added-last test, given that all the other variables are in the model. In other words, the results are interpreted as: given the RUN is in the model, then the p-value for the test of the regression coefficient of TIME after being added to the model is 0.08; and given the TIME is in the model, then the p-value for the test of the regression coefficient of RUN after being added to the model is 0.12.

However, if we add either TIME or RUN to the intercept only model of WGHT, it is still possible that they are significant variables.

7

a.

Source	Df	SS	MS	Fobs	p
Intercept	1.0	4839362527.0	4839362527.0	11694.6	<.0001
Model (Un.)	4.0	4885896692.4	1221474173.1	2951.8	<.0001
Model (Cor.)	3.0	46534165.4	15511388.5	37.5	<.0001
Error (Res.)	166.0	68692765.6	413811.8		
Total (Un.)	170.0	4954589458.0	29144643.9		
Total (Cor.)	169.0	115226931.0	681816.2		

b. $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$ (This set of predictors does not contribute to explaining any of the variability in FVC.) We reject the null hypothesis that these predictors, as a set,

do not significantly predict FVC, as the set of predictors significantly predict FVC, $F(9,160) = 13.90$, $p < 0.0001$.

c. $R_C^2 = \frac{CSS(Regression)}{CSS(Total)} = \frac{50570942}{115226931} = 0.4389$

d. For all of the following calculations, the 2 individuals with any missing data were deleted before running the model.

i. Comparing intercept-only model to full model

$$Full\ Model: FVC_i = \beta_0 + \beta_1(height) + \beta_2(weight) + \beta_3(BMI) + \beta_4(age) + \beta_5(avg\ treadmill\ elevation) + \beta_6(avg\ treadmill\ speed) + \beta_7(temp) + \beta_8(barom) + \beta_9(humid) + \varepsilon_i$$

$$Intercept - Only\ Model: FVC_i = \beta_0 + \varepsilon_i$$

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$$

$$F_{obs} = \frac{\frac{SSE(reduced) - SSE(full)}{dfE(reduced) - dfE(full)}}{SSE(full)/dfE(full)} = \frac{\frac{115226931 - 64655989}{169 - 160}}{64655989/160} = \frac{50570942/9}{404100} = 13.90$$

We reject the null hypothesis that these predictors, as a set, do not significantly predict FVC, as the set of predictors significantly predict FVC, $F(9,160) = 13.90$, $p < 0.0001$.

ii. Comparing intercept-only model to model with height as predictor

$$Larger\ Model: FVC_i = \beta_0 + \beta_1(height) + \varepsilon_i$$

$$Intercept - Only\ Model: FVC_i = \beta_0 + \varepsilon_i$$

$$H_0: \beta_1 = 0 \text{ in } FVC_i = \beta_0 + \beta_1(height) + \varepsilon_i$$

$$F_{obs} = \frac{\frac{SSE(smaller) - SSE(larger)}{dfE(smaller) - dfE(larger)}}{SSE(full)/dfE(full)} = \frac{\frac{115226931 - 76194753}{169 - 168}}{64655989/160} = \frac{39032178}{404100} = 96.59$$

Height provides information about FVC in a simple regression, F(1,160)=96.59, p<0.001. We reject the null hypothesis that height provides no information in predicting FVC, height is significantly related to FVC.

iii. Comparing model with weight and BMI as predictors to model with height, weight, and BMI as predictors

$$Larger\ Model: FVC_i = \beta_0 + \beta_1(height) + \beta_2(weight) + \beta_3(BMI) + \varepsilon_i$$

$$Smaller\ Model: FVC_i = \beta_0 + \beta_2(weight) + \beta_3(BMI) + \varepsilon_i$$

$$H_0: \beta_1 = 0 \text{ in } FVC_i = \beta_0 + \beta_1(height) + \beta_2(weight) + \beta_3(BMI) + \varepsilon_i$$

$$F_{obs} = \frac{\frac{SSE(smaller) - SSE(larger)}{dfE(smaller) - dfE(larger)}}{SSE(full)/dfE(full)} = \frac{\frac{70003745 - 69983719}{167 - 166}}{64655989/160} = \frac{20026}{404100} = 0.0496$$

After adjusting for weight and BMI, height provides no additional information in predicting FVC, F(1,160)=0.0496, p=0.82. We fail to reject the null hypothesis that height provides information about FVC after controlling for weight and BMI.

iv. Comparing full model to full model minus height

$$Full\ Model: FVC_i = \beta_0 + \beta_1(height) + \beta_2(weight) + \beta_3(BMI) + \beta_4(age) + \beta_5(avg\ treadmill\ elevation) + \beta_6(avg\ treadmill\ speed) + \beta_7(temp) + \beta_8(barom) + \beta_9(humid) + \varepsilon_i$$

$$Reduced\ Model: FVC_i = \beta_0 + \beta_2(weight) + \beta_3(BMI) + \beta_4(age) + \beta_5(avg\ treadmill\ elevation) + \beta_6(avg\ treadmill\ speed) + \beta_7(temp) + \beta_8(barom) + \beta_9(humid) + \varepsilon_i$$

$H_0: \beta_1 = 0$ in
$FVC_i = \beta_0 + \beta_1(height) + \beta_2(weight) + \beta_3(BMI) + \beta_4(age) + \beta_5(avg\ treadmill\ elevation) + \beta_6(avg\ treadmill\ speed) + \beta_7(temp) + \beta_8(barom) + \beta_9(humid) + \varepsilon_i$

$$F_{obs} = \frac{\frac{SSE(reduced) - SSE(full)}{dfE(reduced) - dfE(full)}}{SSE(full)/dfE(full)} = \frac{\frac{64719380 - 64655989}{161 - 160}}{64655989/160} = \frac{63391}{404100} = 0.1569$$

After adjusting for all the other predictors in the model, height provides no additional information in predicting FVC, F(1,160)=0.1569, p=0.69. We fail to reject the null hypothesis that height provides information about FVC after controlling for the other variables in the model.

v. Comparing intercept-only model to model with model with height, weight, and BMI as predictors

$$Larger\ Model: FVC_i = \beta_0 + \beta_1(height) + \beta_2(weight) + \beta_3(BMI) + \varepsilon_i$$

$$Intercept - Only\ Model: FVC_i = \beta_0 + \varepsilon_i$$

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$ in $FVC_i = \beta_0 + \beta_1(height) + \beta_2(weight) + \beta_3(BMI) + \varepsilon_i$

$$F_{obs} = \frac{\frac{SSE(smaller) - SSE(larger)}{dfE(smaller) - dfE(larger)}}{SSE(full)/dfE(full)} = \frac{\frac{115226931 - 69983719}{169 - 166}}{64655989/160} = \frac{15081070}{404100} = 37.32$$

Height, weight, and BMI together provide additional information about FVC compared to a model for only the mean level of FVC, F(3,160)=37.32, p<0.001. We reject the null hypothesis that the body size variables provide no more information than the mean-only model. As a set, these three variables provide significant information about FVC.

vi. Comparing model with age, elevation, speed, temp, barometric pressure, and humidity with model with age, elevation, speed, temp, barometric pressure, humidity, height, weight, and BMI

$$Full\ Model: FVC_i = \beta_0 + \beta_1(height) + \beta_2(weight) + \beta_3(BMI) + \beta_4(age) + \beta_5(avg\ treadmill\ elevation) + \beta_6(avg\ treadmill\ speed) + \beta_7(temp) + \beta_8(barom) + \beta_9(humid) + \varepsilon_i$$

$$Reduced\ Model: FVC_i = \beta_0 + \beta_4(age) + \beta_5(avg\ treadmill\ elevation) + \beta_6(avg\ treadmill\ speed) + \beta_7(temp) + \beta_8(barom) + \beta_9(humid) + \varepsilon_i$$

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$ in $FVC_i = \beta_0 + \beta_1(height) + \beta_2(weight) + \beta_3(BMI) + \beta_4(age) + \beta_5(avg\ treadmill\ elevation) + \beta_6(avg\ treadmill\ speed) + \beta_7(temp) + \beta_8(barom) + \beta_9(humid) + \varepsilon_i$

$$F_{obs} = \frac{\frac{SSE(reduced) - SSE(full)}{dfE(reduced) - dfE(full)}}{SSE(full)/dfE(full)} = \frac{\frac{99234383 - 64655989}{163 - 160}}{64655989/160} = \frac{11526131}{404100} = 28.523$$

As a set, height, weight, and BMI together provide additional information about FVC after controlling for all of the other variables in the model, F(3,160)=28.523, p<0.001. We reject the null hypothesis that the body size variables provide no more information after controlling for all of the other variables in the model. As a set, these three variables provide significant information about FVC after controlling for the other variables.

e. *Full Model*: $FVC_i = \beta_0 + \beta_1(height) + \beta_2(weight) + \beta_3(BMI) + \beta_4(age) + \beta_5(avg\ treadmill\ elevation) + \beta_6(avg\ treadmill\ speed) + \beta_7(temp) + \beta_8(barom) + \beta_9(humid) + \varepsilon_i$

Reduced Model: $FVC_i = \beta_0 + \beta_1(height) + \beta_2(weight) + \beta_3(BMI) + \beta_4(age) + \beta_5(avg\ treadmill\ elevation) + \beta_6(avg\ treadmill\ speed) + \beta_7(temp) + \beta_8(barom) + \varepsilon_i$

$H_0: \beta_9 = 0$ in $FVC_i = \beta_0 + \beta_1(height) + \beta_2(weight) + \beta_3(BMI) + \beta_4(age) + \beta_5(avg\ treadmill\ elevation) + \beta_6(avg\ treadmill\ speed) + \beta_7(temp) + \beta_8(barom) + \beta_9(humid) + \varepsilon_i$

$$F_{obs} = \frac{\frac{SSE(reduced) - SSE(full)}{dfE(reduced) - dfE(full)}}{SSE(full)/dfE(full)} = \frac{\frac{64768642 - 64655989}{161 - 160}}{64655989/160} = \frac{112653}{404100} = 0.2788, p = 0.5982$$

We fail to reject the null hypothesis that humidity has no effect on FVC after controlling for the other variables in the model, F(1,160)=0.2788, p=0.5982. After controlling for the other variables in the model, humidity does not significantly relate to FVC.

f. The body size variables together are significantly related to FVC, F(3,160)=37.32, p<0.001. They are significantly related even after controlling for all other variables in the model, F(3,160)=28.523, p<0.001. However, the variables are correlated with each other, as height is significantly related to FVC, F(1,160)=96.59, p<0.001, but not after controlling for weight and BMI, F(1,160)=0.0496, p=0.82.

g. Examining the added-in-order test, we see that height and weight are both highly related to FVC, but controlling for the other variables, neither height nor weight are significantly related (collinearity with BMI). The added-in-order also reveals that average treadmill elevation and average treadmill speed are significantly related to FVC as well. Again, controlling for the other variables in the model, neither variable individually is significantly related to FVC. Age is significantly related to FVC when controlling for the other variables, but is not significantly related to FVC when controlling for only height, weight, and BMI.

Examining parameter estimates from the added-in-order tests, we find that increasing height as well as increasing weight is associated with an increased FVC. Controlling for height, weight, BMI, and age, increasing average treadmill elevation as well as increasing treadmill speed are associated with increased FVC. While age is not significantly related to FVC controlling for height, weight, and BMI, an increased age is associated with increased FVC controlling for all other variables in the model.

3. This problem involves the FEV data described above. We will consider the model $FVC_i = \beta_0 + \beta_1 X_H + \beta_2 X_W + \beta_3 X_{BMI} + \beta_4 X_{Area} + \beta_5 X_{Age} + \beta_6 X_{avtrel} + \beta_7 X_{avtrsp} + \beta_8 X_{avtrel*avtrsp} + \beta_9 X_{temp} + \beta_{10} X_{barm} + \beta_{11} X_{hum} + \epsilon$.

a. Compute the following correlations, giving the interpretation of each, between FVC and age, and report tests of the hypotheses that each correlation equals zero.

i. The correlation between age and FVC, controlling both for all the other variables in the model

- **Hypotheses**:
  - $H_0: \rho_{(age,FVC|other\ variables)} = 0$
  - $H_A: \rho_{(age,FVC|other\ variables)} \neq 0$

- **Test Statistic**: $F_{obs} = \frac{\frac{[SSE(\text{no age}) - SSE(full)]}{[df_e(\text{no age}) - df_e(full)]}}{\frac{SSE(full)}{df_e(full)}} = \frac{\frac{[64828716.44 - 62761458.29]}{[159-158]}}{\frac{62761458.29}{158}} = 5.20426$
- **Degrees of Freedom**: $df_e(\text{no age}) = 159, df_e(full) = 158$
- **P-value**: $\Pr(F_{obs} > F_{(1,158)}) = 0.02387$
- **Decision**: Reject the null hypothesis
- **Interpretation**: There is a nonzero correlation between age and FVC after adjusting both for the other variables in the model.
- **Correlation**:
  - According to SAS, $\rho_{(age,FVC|other\ variables)} = 0.17857$
  - This suggests that, after controlling both variables for all of the other variables, with one increase in standard deviation of age, FVC is expected to increase by 0.17857 standard deviations.

ii. The correlation between age and FVC, controlling only age for all the other variables in the model

- **Hypotheses**:
  - $H_0: \rho_{FVC(age|other\ variables)} = 0$
  - $H_A: \rho_{FVC(age|other\ variables)} \neq 0$
- **Test Statistic**:
  - First, used SAS to model obtain studentized residuals of age=(other variables)
  - Second, used SAS to model avfvc=(studentized residuals)
    - $F_{obs} = \frac{\frac{[SSE(\beta_0) - SSE(full)]}{[df_e(\beta_0) - df_e(full)]}}{\frac{SSE(full)}{df_e(full)}} = \frac{\frac{[CSS(model)]}{[169-168]}}{\frac{SSE(full)}{df_e(full)}} = \frac{\frac{[2067258.2]}{[169-168]}}{\frac{113159672.8}{168}} = 3.07$
- **Degrees of Freedom**: $df_e(\beta_0) = 169, df_e(full) = 168$
- **P-value**: $\Pr(F_{obs} > F_{(1,168)}) = 0.0816$
- **Decision**: Fail to reject the null hypothesis
- **Interpretation**: After controlling the other variables on age, there isn't enough evidence to suggest a significant correlation between age (adjusted) and FVC.
- **Correlation**: $r_{FVC(age|other\ variables)} = r(FVC_i, \hat{e}_{age}) = 0.13394$. If this correlation was statistically significant, it would suggest that with one increase in standard deviation of age (adjusted), FVC (unadjusted) is expected to increase by 0.13570 standard deviations.

iii. The simple correlation between age and FVC (not controlling for any other variables)

- **Hypotheses**: $H_0: \rho_{FVC,age} = 0$ vs. $H_A: \rho_{FVC,age} \neq 0$
- **Test Statistic**:
  - $F_{obs} = \frac{\frac{[SSE(\beta_0) - SSE(\beta_0,age)]}{[df_e(\beta_0) - df_e(\beta_0,age)]}}{\frac{SSE(\beta_0,age)}{df_e(\beta_0,age)}} = \frac{\frac{[115226930.97 - 112547661.64]}{[169-168]}}{\frac{112547661.64}{168}} = 3.99935$
- **Degrees of Freedom**: $df_e(\beta_0) = 169, df_e(full) = 168$
- **P-value**: $\Pr(F_{obs} > F_{(1,168)}) = 0.047129$
- **Decision**: Reject the null hypothesis
- **Interpretation**: There is enough evidence to suggest a simple correlation between age and FVC.
- **Correlation**: $r_{FVC,age} = 0.15249$. This suggests that with one increase in standard deviation of age (unadjusted), FVC (unadjusted) is expected to increase by 0.15249 standard deviations.

b. Provide and interpret the following diagnostics (include subject ID when appropriate) for the regression model.

(b)

i.

the largest 5 studentized residuals in absolute values are:

Subject ID	Studentized residuals in absolute value
60	3.882
185	2.904
49	2.903
181	2.660
99	2.267

ii.

To test whether the studentized residuals are normal, we can use one of:

1) Shapiro-Wilk Test;

2) Kolmogorov-Smirnov Test;

3) Cramer-von Mises Test;

4) Anderson-Darling Test.

The null hypothesis is that the studentized residuals follow normal distribution, and the followings are the corresponding test statistics and the p-values:

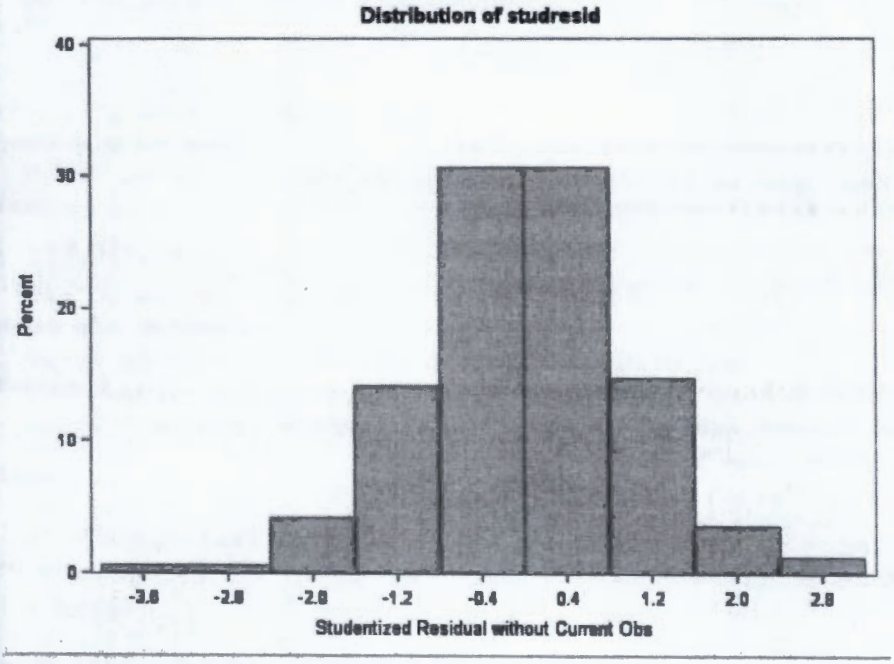
Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.989513	Pr < W	0.2420
Kolmogorov-Smirnov	D	0.044551	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.039587	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.308589	Pr > A-Sq	>0.2500

All the tests show p-values that are greater than 0.05, so that we fail to reject the null hypothesis. The result can be interpreted as the following: assuming that the studentized residuals are normally distributed, then the probability of observing the data that are as extreme as ours or more extreme is larger than 0.05, which is not too small for us to question the null hypothesis.

Note that the type of test should be decided before conducting the test to avoid "p shopping", although the test statistics and p-values are all reported in the table above.

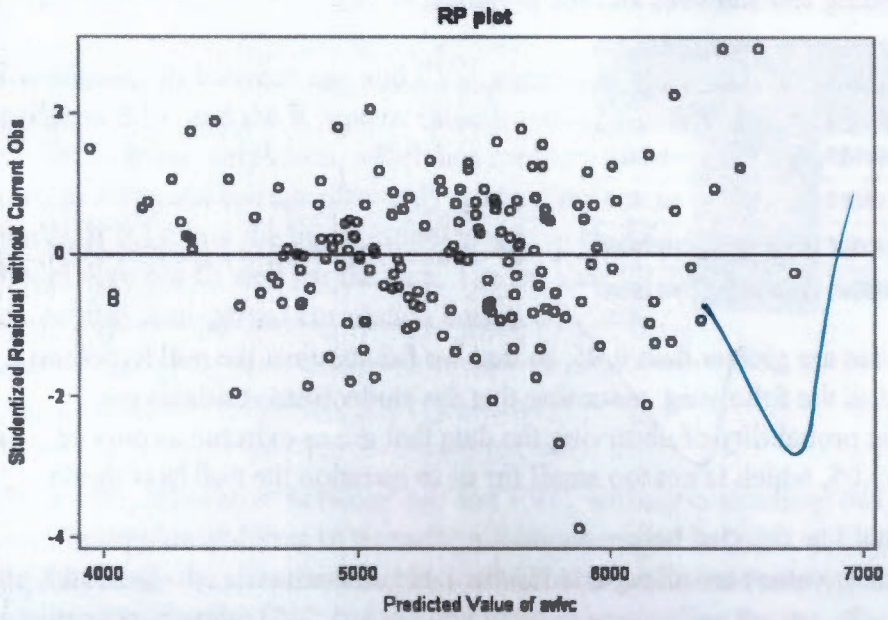
iii.

The histogram of the studentized residuals is plotted as the following:



The histogram can be interpreted as the following: the shape of the studentized residuals appears to be normal in this histogram, which coincide with our hypothesis testing in part ii.

iv. the RP plot of studentized residuals is plotted as the following:



The RP plot can be interpreted as the following: from the RP plot we can see that the linearity assumption, the homogeneity of variance assumption, and Gauss error assumption holds, because we cannot see any nonlinear pattern or heterogeneity of variance of the residuals. Also the studentized residuals appear to be normal as well.