SOLUTIONS TO ASSIGNMENT 4

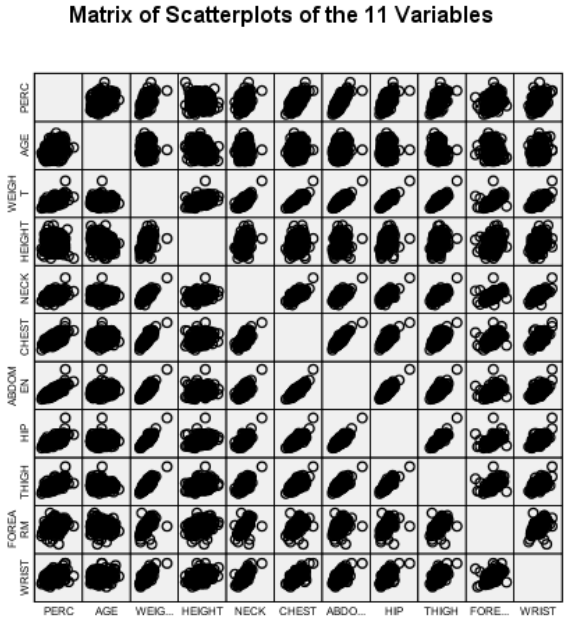
Question 1

(a) The purpose of the study is to establish a generalized body fat percentage prediction equation by using age, height and weight of an individual, and some simple circumference measurements. The experimental units in the study are 252 male volunteers. The percentage of body fat is the response variable and there are 10 explanatory variables including age, weight, height and seven various body circumference measurements. The subjects in the study were randomly selected from a larger pool of volunteers in Europe. In order to extend the results of the study to the population of all males, we should assume that these 252 men are representative of the whole population of interest.

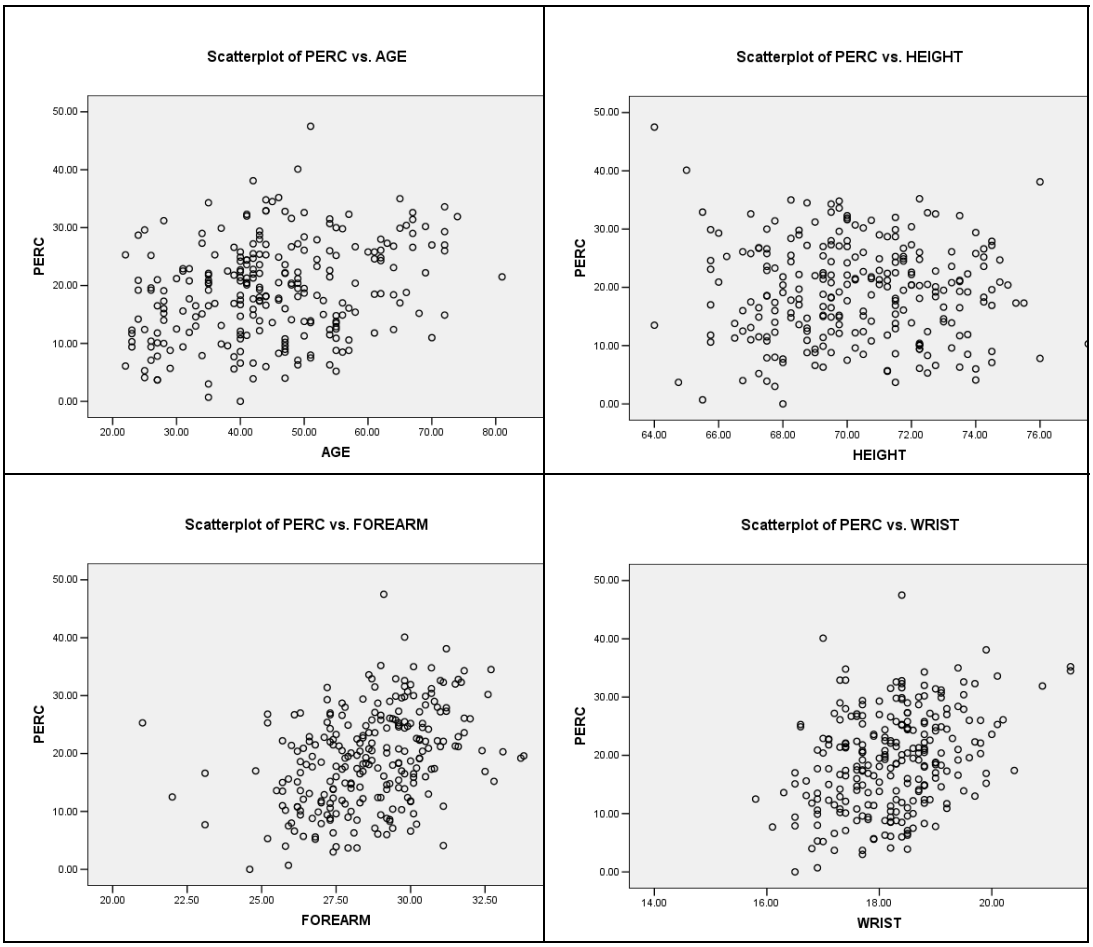
(b) Given the fact that men tend to store fat around their middle, we can expect that, among the seven body circumference measurements, the ABDOMEN variable (Abdomen 2 circumference) should be a better predictor of body fat percentage for men. The accuracy of the measurements plays an extremely important role of producing good estimates of body fat percentage. There are several ways to reduce the measurement errors in practice. For example, we can reduce the measurement error with calibration; buy more expensive test instruments; provide training for examiners and take the average of several readings instead of a single reading.

Question 2

(a) The matrix of scatterplots of the 11 variables is given below.



(b) The scatterplot above shows relatively strong positive linear patterns between PERC and WEIGHT, NECK, CHEST, ABDOMEN, HIP and THIGH. To evaluate the relationship between PERC and the other predictors properly, the separate plots were given below. There seem to be a moderate positive linear pattern between PERC and AGE, and strong positive linear patterns between PERC and FOREARM and WRIST. However, data points are randomly scattered in the scatterplot of PERC and HEIGHT, which means the relationship between PERC and HEIGHT are quite weak.



Overall, a linear model seems to be appropriate for describing the relationship between PERC and the 10 predictors. Among all the predictors, the ABDOMEN variable has the strongest linear relationship with PERC. There are one or two outliers indicated in the scatterplot.

Question 3

(a) According to the correlation matrix, the signs and magnitudes of the correlation coefficients confirm the conclusions based on the examination of the matrix of scatterplots in Question 2. There are statistically significant linear relationship between PERC and all predictors except for HEIGHT.

(b) If we wished to use only one explanatory variable to predict percent body fat, we should choose the ABDOMEN variable, because it has the highest correlation coefficient of 0.804 with PERC.

Multicollinearity is a potential problem in this case, because there are several explanatory variables that are highly correlated. For example, WEIGHT is highly correlated with CHEST (Pearson Correlation=0.894), ABDOMEN (Pearson Correlation=0.893), HIP (Pearson Correlation=0.941) and THIGH (Pearson Correlation=0.869).

The correlation matrix of the 11 variables is displayed below.

Correlations

		PERC	AGE	WEIGHT	HEIGHT	NECK	CHEST	ABDOMEN	HIP	THIGH	FOREARM	WRIST
PERC	Pearson Correlation	1	.291**	.612**	-.025	.491**	.703**	.804**	.625**	.560**	.361**	.347**
	Sig. (2-tailed)		.000	.000	.690	.000	.000	.000	.000	.000	.000	.000
	N	252	252	252	252	252	252	252	252	252	252	252
AGE	Pearson Correlation	.291**	1	-.013	-.245**	.114	.176**	.226**	-.050	-.200**	-.085	.214**
	Sig. (2-tailed)	.000		.840	.000	.072	.005	.000	.426	.001	.178	.001
	N	252	252	252	252	252	252	252	252	252	252	252
WEIGHT	Pearson Correlation	.612**	-.013	1	.487**	.831**	.894**	.893**	.941**	.869**	.630**	.730**
	Sig. (2-tailed)	.000	.840		.000	.000	.000	.000	.000	.000	.000	.000
	N	252	252	252	252	252	252	252	252	252	252	252
HEIGHT	Pearson Correlation	-.025	-.245**	.487**	1	.321**	.227**	.189**	.372**	.339**	.322**	.398**
	Sig. (2-tailed)	.690	.000	.000		.000	.000	.003	.000	.000	.000	.000
	N	252	252	252	252	252	252	252	252	252	252	252
NECK	Pearson Correlation	.491**	.114	.831**	.321**	1	.785**	.759**	.735**	.696**	.624**	.745**
	Sig. (2-tailed)	.000	.072	.000	.000		.000	.000	.000	.000	.000	.000
	N	252	252	252	252	252	252	252	252	252	252	252
CHEST	Pearson Correlation	.703**	.176**	.894**	.227**	.785**	1	.913**	.829**	.730**	.580**	.660**
	Sig. (2-tailed)	.000	.005	.000	.000	.000		.000	.000	.000	.000	.000
	N	252	252	252	252	252	252	252	252	252	252	252
ABDOMEN	Pearson Correlation	.804**	.226**	.893**	.189**	.759**	.913**	1	.881**	.770**	.494**	.620**
	Sig. (2-tailed)	.000	.000	.000	.003	.000	.000		.000	.000	.000	.000
	N	252	252	252	252	252	252	252	252	252	252	252
HIP	Pearson Correlation	.625**	-.050	.941**	.372**	.735**	.829**	.881**	1	.896**	.545**	.630**
	Sig. (2-tailed)	.000	.426	.000	.000	.000	.000	.000		.000	.000	.000
	N	252	252	252	252	252	252	252	252	252	252	252
THIGH	Pearson Correlation	.560**	-.200**	.869**	.339**	.696**	.730**	.770**	.896**	1	.567**	.559**
	Sig. (2-tailed)	.000	.001	.000	.000	.000	.000	.000	.000		.000	.000
	N	252	252	252	252	252	252	252	252	252	252	252
FOREARM	Pearson Correlation	.361**	-.085	.630**	.322**	.624**	.580**	.494**	.545**	.567**	1	.586**
	Sig. (2-tailed)	.000	.178	.000	.000	.000	.000	.000	.000	.000		.000
	N	252	252	252	252	252	252	252	252	252	252	252
WRIST	Pearson Correlation	.347**	.214**	.730**	.398**	.745**	.660**	.620**	.630**	.559**	.586**	1
	Sig. (2-tailed)	.000	.001	.000	.000	.000	.000	.000	.000	.000	.000	
	N	252	252	252	252	252	252	252	252	252	252	252

**. Correlation is significant at the 0.01 level (2-tailed).

Question 4

Define a multiple regression model with PERC as the response variable and the 10 explanatory variables as follows:

$$\mu(PERC) = \beta_0 + \beta_1 \times AGE + \beta_2 \times WEIGHT + \beta_3 \times HEIGHT + \beta_4 \times NECK + \beta_5 \times CHEST + \beta_6 \times ABDOMEN + \beta_7 \times HIP + \beta_8 \times THIGH + \beta_9 \times FOREARM + \beta_{10} \times WRIST$$

or

$$PERC = \beta_0 + \beta_1 \times AGE + \beta_2 \times WEIGHT + \beta_3 \times HEIGHT + \beta_4 \times NECK + \beta_5 \times CHEST + \beta_6 \times ABDOMEN + \beta_7 \times HIP + \beta_8 \times THIGH + \beta_9 \times FOREARM + \beta_{10} \times WRIST + ERROR$$

We assume that ERROR follows a normal distribution for each value of each explanatory variable and the mean of ERROR is zero. Moreover, the variance of the ERROR variable is constant for each value of each explanatory variable.

Question 5

Use the forward model selection procedure for the first 200 observations to fit the model defined in the previous question. The related SPSS outputs are displayed below.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	ABDOMEN	.	Forward (Criterion: Probability-of-F-to-enter <= .050)
2	WEIGHT	.	Forward (Criterion: Probability-of-F-to-enter <= .050)
3	NECK	.	Forward (Criterion: Probability-of-F-to-enter <= .050)
4	FOREARM	.	Forward (Criterion: Probability-of-F-to-enter <= .050)

a. Dependent Variable: PERC

Model Summary[e]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.804[a]	.646	.644	4.96065
2	.835[b]	.697	.694	4.59837
3	.840[c]	.706	.701	4.54677
4	.846[d]	.715	.709	4.48378

a. Predictors: (Constant), ABDOMEN

b. Predictors: (Constant), ABDOMEN, WEIGHT

c. Predictors: (Constant), ABDOMEN, WEIGHT, NECK

d. Predictors: (Constant), ABDOMEN, WEIGHT, NECK, FOREARM

e. Dependent Variable: PERC

ANOVA[e]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8895.817	1	8895.817	361.500	.000[a]
	Residual	4872.396	198	24.608		
	Total	13768.213	199			
2	Regression	9602.654	2	4801.327	227.067	.000[b]
	Residual	4165.559	197	21.145		
	Total	13768.213	199			
3	Regression	9716.284	3	3238.761	156.665	.000[c]
	Residual	4051.928	196	20.673		
	Total	13768.213	199			
4	Regression	9847.871	4	2461.968	122.460	.000[d]
	Residual	3920.342	195	20.104		
	Total	13768.213	199			

a. Predictors: (Constant), ABDOMEN

b. Predictors: (Constant), ABDOMEN, WEIGHT

c. Predictors: (Constant), ABDOMEN, WEIGHT, NECK

d. Predictors: (Constant), ABDOMEN, WEIGHT, NECK, FOREARM

e. Dependent Variable: PERC

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-34.883	2.874		-12.139	.000
	ABDOMEN	.583	.031	.804	19.013	.000
2	(Constant)	-40.718	2.849		-14.295	.000
	ABDOMEN	.921	.065	1.270	14.168	.000
	WEIGHT	-.142	.025	-.518	-5.782	.000
3	(Constant)	-25.613	7.032		-3.643	.000
	ABDOMEN	.934	.065	1.288	14.478	.000
	WEIGHT	-.107	.029	-.390	-3.741	.000
	NECK	-.594	.254	-.171	-2.344	.020
4	(Constant)	-34.757	7.801		-4.455	.000
	ABDOMEN	.968	.065	1.335	14.894	.000
	WEIGHT	-.130	.030	-.475	-4.396	.000
	NECK	-.726	.255	-.209	-2.844	.005
	FOREARM	.529	.207	.126	2.558	.011

a. Dependent Variable: PERC

(a) According to the *Variable Entered/Removed* Table, 4 of the 10 predictors were included by the forward model selection procedure in the order of ABDOMEN, WEIGHT, NECK, and FOREARM.

(b) The estimated regression equation for the final model is

$$\mu(PERC) = -34.757 + 0.968 \times ABDOMEN - 0.130 \times WEIGHT - 0.726 \times NECK + 0.529 \times FOREARM$$

(c) According to the *Model Summary* table, R^2=0.715 for the final model. So 71.5% of the variation in PREC is explained by the explanatory variables in the final model.

(d) In order to see whether the fitted model as a whole is significant, we define the null and alternative hypotheses as follows:

H_o: $\beta_2 = \beta_4 = \beta_6 = \beta_9 = 0$ vs. H_a: at least one $\beta_i \neq 0$, $i = 2,4,6 \text{ or } 9$

In other words, the null hypothesis claims that the model is useless because no explanatory variable contributes to prediction of PREC. The alternative hypothesis claims that at least one explanatory variable contributes to prediction of PREC.

According to the *ANOVA* table, the value of the F test statistic is 122.46, where F follows an F distribution with 4 degrees of freedom for the numerator and 195 degrees of freedom for the denominator. The p value of the test is reported as zero. Therefore, there is strong evidence against the null hypothesis. In other words, the final model as a whole is significant at the 0.05 level.

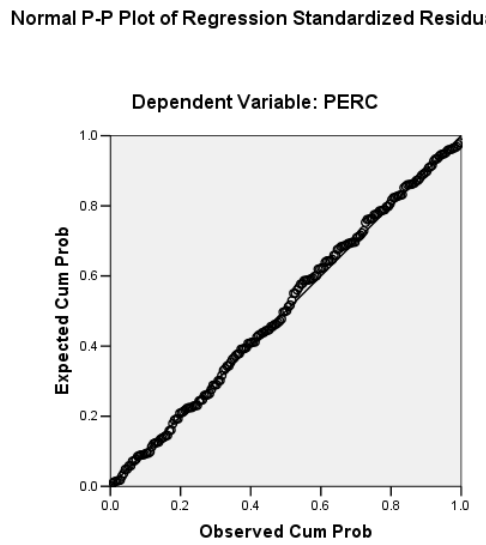
(e) In order to see how significantly each explanatory variable contributes individually given the other variables in the model, we define the null and alternative hypotheses as follows:

H_o: $\beta_i = 0$ vs. H_a: $\beta_i \neq 0$ ($i = 2,4,6 \text{ or } 9$)

Based on the *Coefficients* table, the p value of each of the four tests for the variable ABDOMEN, WEIGHT, NECK, and FOREARM is smaller than the threshold value of 0.05. Therefore, each of the four variables is useful in predicting percent body fat given the other variables in the model.

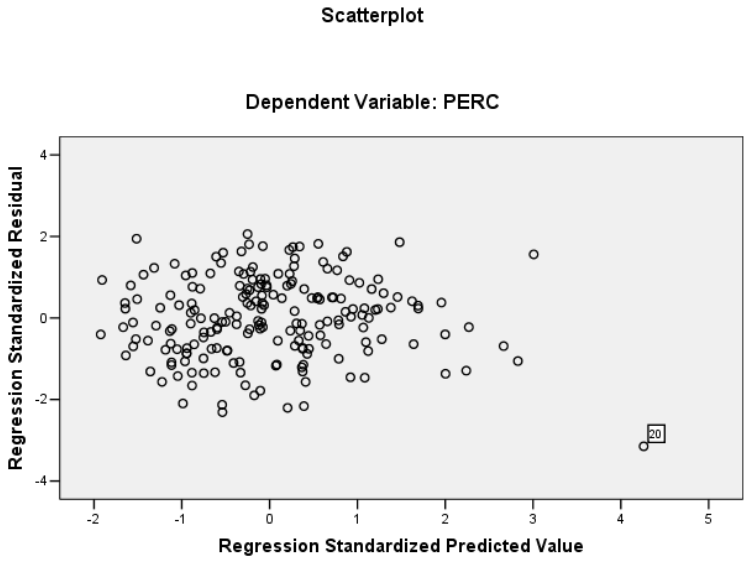
Question 6

(a) The normal probability plot of standardized residuals for the final model obtained in Question 5 is given below.



The above plot shows all points lie reasonably close to a straight line, therefore, it does not indicate any serious departures from the normality assumption.

(b) The plot of standardized residuals versus standardized predicted values for the final model obtained in Question 5 is given below.



All the points are randomly and uniformly scattered around zero. There is no evidence that the variance of the residuals increases or decreases with increasing fitted values. Therefore, the equal variance assumption is approximately satisfied. There are some outliers indicted in the plot.

(c) We find that the case #20 has a relatively high studentized residuals of -3.73430, and large Cook's distance of 1.13760 (>1), and high leverage value of .28472 (>2×5/200=0.05). Therefore, the case #20 is an influential case.

Question 7

The related SPSS outputs without the influential case #20 are given below.

Variables Entered/Removed[d]

Model	Variables Entered	Variables Removed	Method
1	ABDOMEN	.	Forward (Criterion: Probability-of-F-to-enter <= .050)
2	WEIGHT	.	Forward (Criterion: Probability-of-F-to-enter <= .050)
3	WRIST	.	Forward (Criterion: Probability-of-F-to-enter <= .050)

a. Dependent Variable: PERC

Model Summary[d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.826[a]	.683	.681	4.66399
2	.850[b]	.722	.719	4.37870
3	.857[c]	.734	.730	4.29006

a. Predictors: (Constant), ABDOMEN

b. Predictors: (Constant), ABDOMEN, WEIGHT

c. Predictors: (Constant), ABDOMEN, WEIGHT, WRIST

d. Dependent Variable: PERC

ANOVA[d]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9230.234	1	9230.234	424.324	.000[a]
	Residual	4285.303	197	21.753		
	Total	13515.537	198			
2	Regression	9757.626	2	4878.813	254.463	.000[b]
	Residual	3757.911	196	19.173		
	Total	13515.537	198			
3	Regression	9926.644	3	3308.881	179.786	.000[c]
	Residual	3588.892	195	18.405		
	Total	13515.537	198			

a. Predictors: (Constant), ABDOMEN

b. Predictors: (Constant), ABDOMEN, WEIGHT

c. Predictors: (Constant), ABDOMEN, WEIGHT, WRIST

d. Dependent Variable: PERC

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-40.896	2.939		-13.914	.000
	ABDOMEN	.649	.032	.826	20.599	.000
2	(Constant)	-45.068	2.872		-15.693	.000
	ABDOMEN	.935	.062	1.190	15.082	.000
	WEIGHT	-.124	.024	-.414	-5.245	.000
3	(Constant)	-23.525	7.646		-3.077	.002
	ABDOMEN	.924	.061	1.176	15.188	.000
	WEIGHT	-.084	.027	-.280	-3.138	.002
	WRIST	-1.522	.502	-.166	-3.030	.003

a. Dependent Variable: PERC

(a) The estimated regression equation is

$$\mu(PERC) = -23.525 + 0.924 \times ABDOMEN - 0.084 \times WEIGHT - 1.522 \times WRIST$$

73.4% of the variation in PERC is explained by the explanatory variables in the final model, which is higher than the one in Question 5 of 71.5%.

(b) According to the SPSS output, the predicted value of the percentage of body fat for an individual with a body weight of 184.75 pounds, an abdomen 2 circumference of 86.4 cm, and a wrist circumference of 18.2 cm is 13.11%.

(c) The 95% confidence interval is (11.94839, 14.27628) and the 95% prediction interval is (4.57179, 21.65289). The prediction interval is wider. The confidence interval estimates the mean percentage of body fat of all individuals with a body weight of 184.75 pounds, an abdomen 2 circumference of 86.4 cm, and a wrist circumference of 18.2 cm. However, the prediction interval provides an estimate of the percentage of body fat for a new individual having the above numerical characteristics. The uncertainty includes the noise that is inherent in the estimates of the regression parameters and the uncertainty of the new individual. Therefore, the prediction interval for a new individual will be wider than the confidence interval for the value of the regression function.

(d) In order to test whether there is an association between the body fat percentage and the lower body measurements (THIGH and HIP) after accounting for age, weight, height, neck, chest, abdomen, forearm and wrist, it is equivalent to compare two models: the model defined in Question 4 (full model), and the model without the lower body measurements THIGH and HIP (reduced model).

We will defined the null and alternative hypotheses as follows:

H_o: $\beta_7 = \beta_8 = 0$ vs. H_a: $\beta_7 \neq 0$ *or* $\beta_8 \neq 0$

In other words, the null hypothesis claims that the lower body measurements THIGH and HIP are useless to predict the value of PREC. The alternative hypothesis claims that at least one of them contributes to prediction of PREC.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10113.566	10	1011.357	55.890	.000[a]
	Residual	3401.970	188	18.096		
	Total	13515.537	198			

a. Predictors: (Constant), WRIST, AGE, HEIGHT, FOREARM, THIGH, CHEST, NECK, ABDOMEN, HIP, WEIGHT

b. Dependent Variable: PERC

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10077.527	8	1259.691	69.616	.000[a]
	Residual	3438.010	190	18.095		
	Total	13515.537	198			

a. Predictors: (Constant), WRIST, AGE, HEIGHT, FOREARM, ABDOMEN, NECK, CHEST, WEIGHT

b. Dependent Variable: PERC

According to the above SPSS outputs, the F-statistic is

$$F = \frac{(SSR_{reduced} - SSR_{full})/(df_{reduced} - df_{full})}{SSR_{full}/df_{full}}$$

$$= \frac{(3438.010 - 3401.970)/(190 - 188)}{3401.970/188} = 0.9958 \sim F^2_{188}$$

The F statistic follows an F distribution with 2 degrees of freedom for the numerator and 188 degrees of freedom for the denominator. Using the value of the cumulative probabilities of the F distribution with degrees of freedom 2 and 188 in the *Compute* feature in SPSS, we will get

P-value = 1- CDF.F (0.9958, 2,188) = 0.37

Therefore, there is no convincing evidence to show that the full model is significantly better than the reduced model. That is, adding the variable THIGH and HIP does not significantly contribute to the predictive ability of the model. In other words, there is no convincing evidence of an association between the body fat percentage and the lower body measurements (THIGH and HIP) after accounting for age, weight, height, neck, chest, abdomen, forearm and wrist.

Question 8

To test the validity of the fit obtained in Question 7 Part (a), we use a paired t-test on the mean difference between actual and predicted percent fat for the remaining 52 observations in the data set. The SPSS output is given below.

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	PERC - Fitted	-.14290	4.43844	.61550	-1.37857	1.09277	-.232	51	.817

The value of the t test is –0.232. The t test statistic follows a t distribution with 51 degrees of freedom and the p value is 0.817. The 95% confidence interval for the mean difference is (-1.37857, 1.09277).

Therefore, there is no significant difference between the actual and the predicted percent fat for the remaining 52 observations in the data set. Hence, the model obtained in Question 7 Part (a) is valid.

LAB ASSIGNMENT 4 MARKING SCHEMA

Proper Header: 10 points

Question 1

(a) Purpose of the study: 1 point
Experimental units: 1 point
Response variable: 1 point
Explanatory variables: 1 point
Assumptions: 1 point
(b) Better predictor: 1 point
Accuracy of measurements: 1 point
Reduce measurement errors: 2 point

Question 2

(a) Scatterplot: 6 points
(b) Describe the relationships: 2 points
Validity of a linear model: 1 point
Predictor has the strongest relationship with PERC: 1 point
Outliers: 1 point

Question 3

Correlation matrix: 6 points
(a) Describe the relationships: 3 points
(b) One explanatory to predict PERC: 1 point
Multicollinearity: 2 points

Question 4

Define regression model: 3 points
Model assumptions: 3 points

Question 5

(a) Number of predictors: 1 point
Order: 2 points
(b) Estimated regression equation: 3 points
(c) R2: 2 points
(d) Null and alternative hypotheses: 2 points
Value of the F-statistic and p-value: 2 points
Null distribution: 2 points
Conclusion: 2 points
(e) Conclusion: 2 points

Question 6

(a) Normal probability plot: 6 points
Normality assumption: 2 points
(b) Residual plot: 6 points
Equal variance assumption: 2 points
Outliers: 2 points
(c) Influential case: 1 point

Case statistics: 3 points

Question 7

SPSS outputs: 3 points
(a) Estimated regression equation: 3 points
R2: 2 points
Comparison: 1 point
(b) Predicted value: 2 points
(c) Confidence interval: 2 points
Prediction interval: 2 points
Wider and why: 3 points
(d) Null and alternative hypotheses: 2 points
Value of the F-statistic: 6 points
P-value: 2 points
Conclusion: 2 points

Question 8

Paired t-test: 2 points
Confidence interval: 2 points
P-value: 2 points
Conclusion: 2 points

TOTAL= 126