Matriculation number: 2383746W

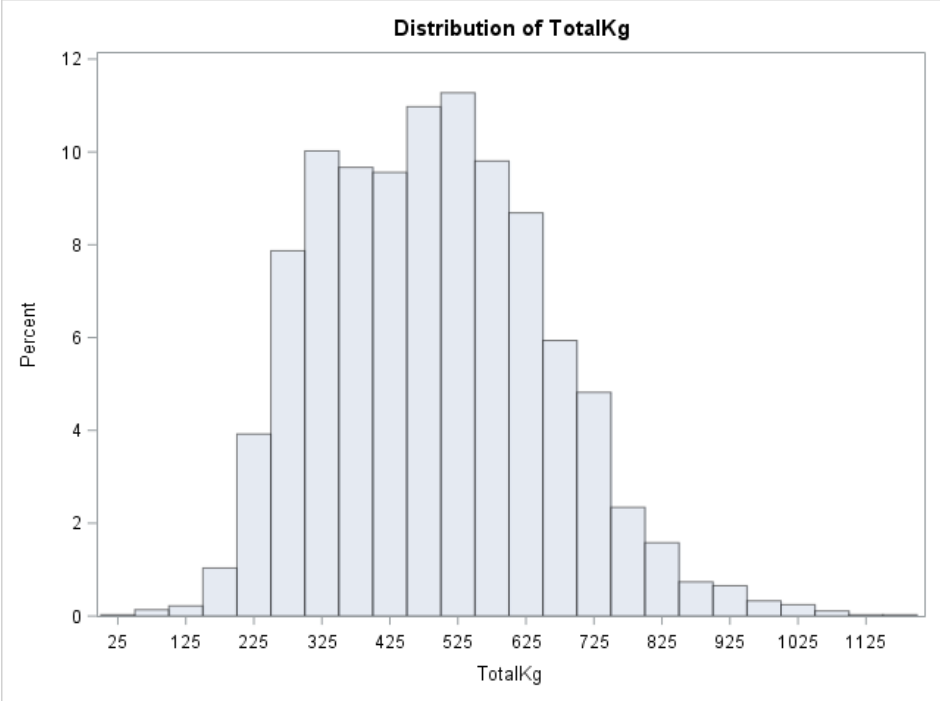
SAS Analytics Report - December 2018

1. Part One: Powerlifting Data

Question A)

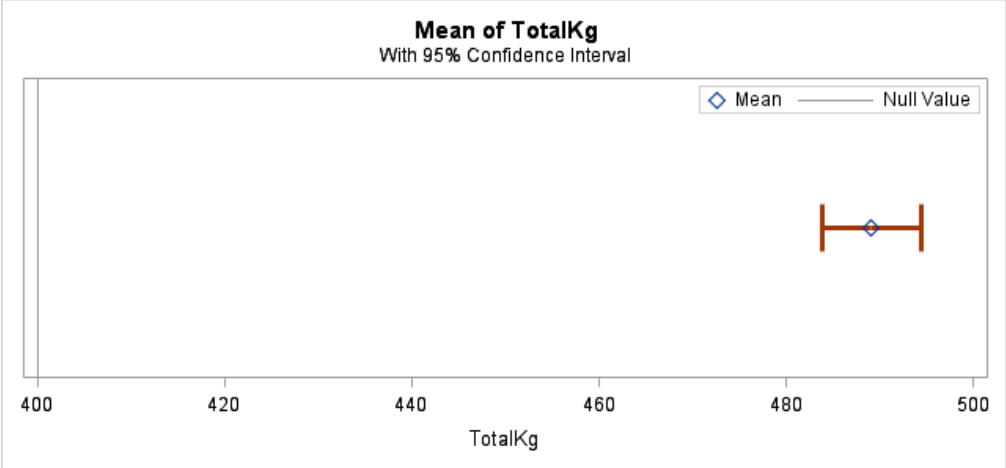
```
data Powerlifting;
set Powerlifting;
TotalKg = BestSquatKg + BestBenchKg + BestDeadliftKg;
run;

proc univariate data = Powerlifting plots;
var TotalKg;
histogram TotalKg;
run;
```



Question B)

```
proc ttest data = Powerlifting H0 = 400 plots(shownull)= all;
var TotalKg;
title "One-sample t-test. Is the population mean different than 400?";
run;
```



Question C)

```
proc ttest data = Powerlifting plots(shownull) = interval;
class Sex;
var TotalKg;
title "Two-sample t-test with Class=Sex";
run;
```

Question D)

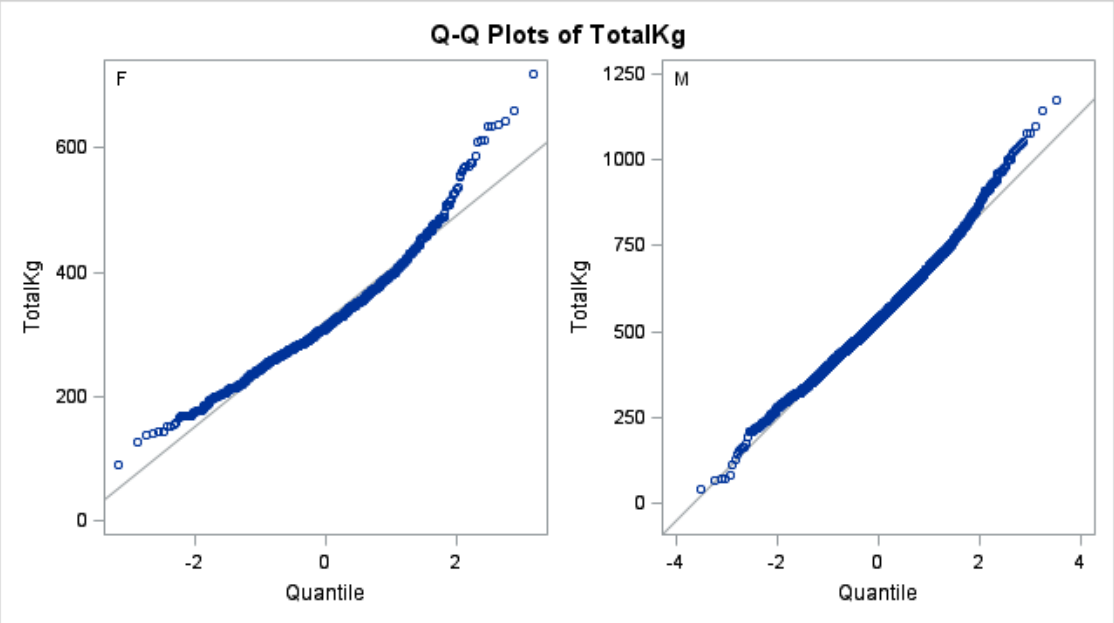
Based on the output it is clear that the variances between males and females are not equal (the result of F-test is not statistically significant at $\alpha = 0.05$) therefore the Satterthwaite result, which does not assume homogenous variances, should be used to read the 95% confidence interval of (-229.2 , -213.5).

Since this CI for the mean difference between males and females is wholly negative (i.e. on average females lift less total kg than males), thus it is evident to reject the null hypothesis of equal population means (H_0: $\mu_F = \mu_M$).

Question E)

Check of assumptions for two-sample t-test:

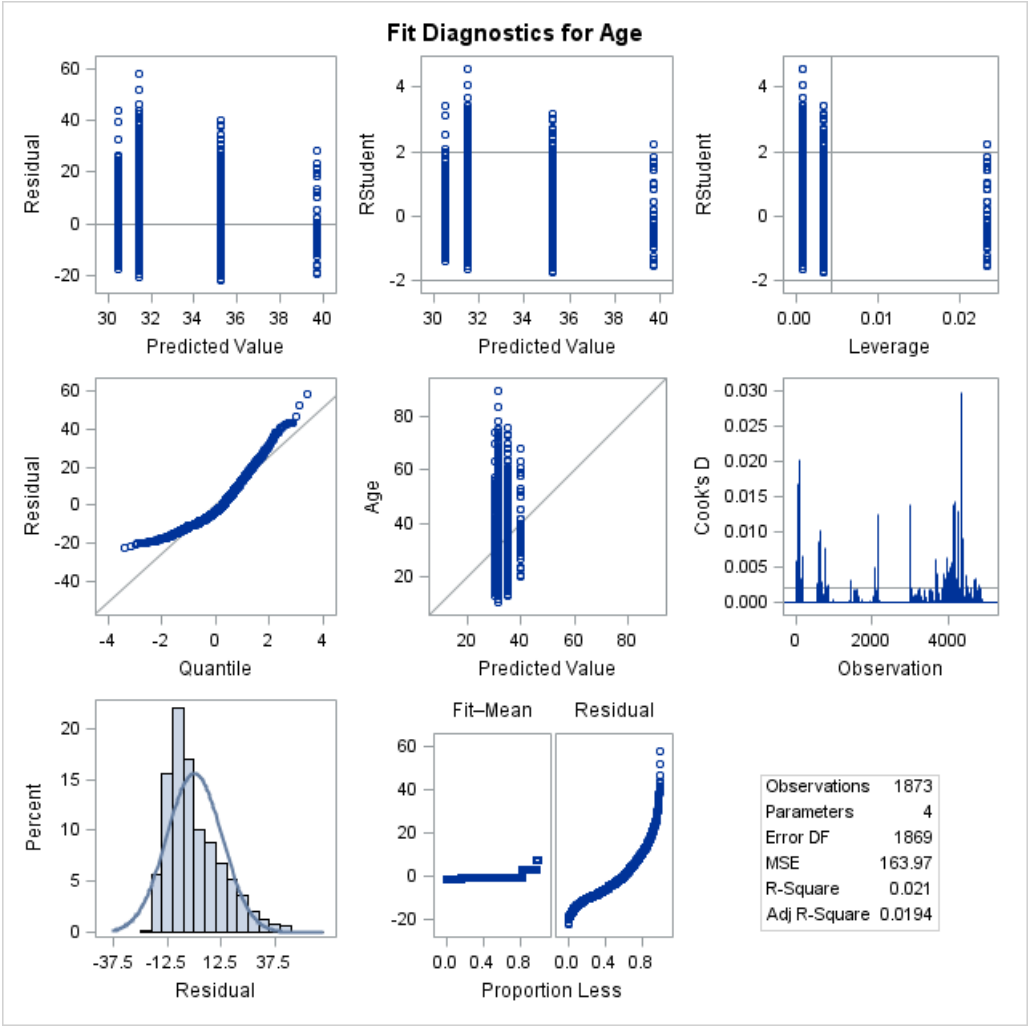
The information for the powerlifting dataset explicitly mentions independence and the normal probability plot indicates a slight curvature in the female population (test for clarification would be required), but a clear straight-line alignment of residuals for the male population, overall both appear to satisfy normality. The F-value is 3.08 with a very low p-value, this provides evidence that the null hypothesis of equal group variances (Female-Male) will be rejected at a 5% significance level (violated assumption).

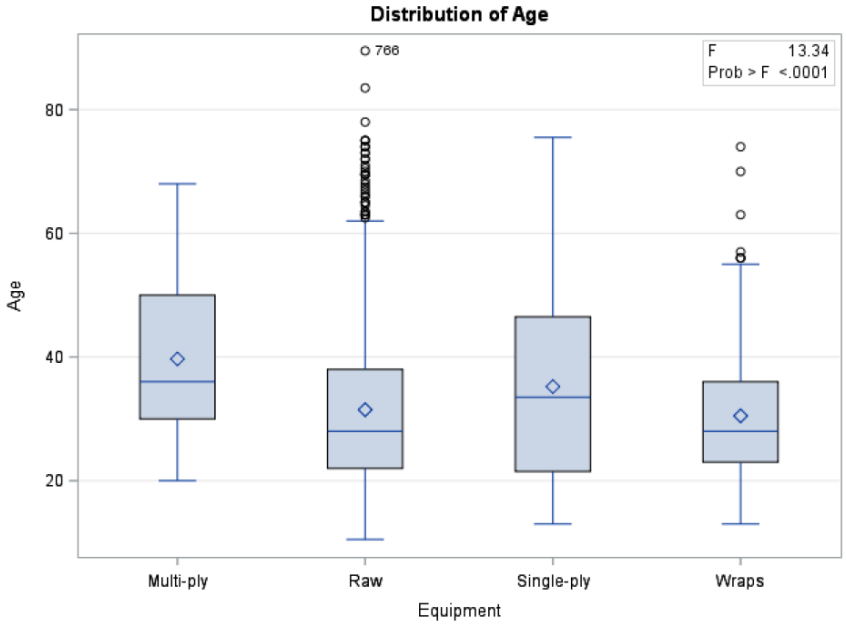


Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	2788	882	3.08	<.0001

Question F)

```
proc glm data = Powerlifting plots = all;
class Equipment;
model Age = Equipment;
title "One-Way ANOVA with Equipment as Explanatory";
run;
```





Conclusions:

The diagnostics output for the dependent variable age reports a large F-value (13.34) together with a statistically significant p-value (< 0.001). Thus, there is evidence to reject the null hypothesis of equal mean age among the four types of equipment and at least one of the means differ from the others.

Check the validity of modelling assumptions for One-Way ANOVA:

1. Since the powerlifting dataset is being analysed, we already know from question E) that the observations are independently selected from their respective populations.
2. The Q-Q plot shows a notable deviation of the residuals from the straight line (most extreme in the tails), which is an indication that the normality assumption might be violated. A test for normality would be required to clarify this with certainty.
3. The scale of variability is clearly different as shown in the residual against predicted value plot, hence group variances do not seem to be constant (presence of heteroscedasticity).

In order to check for equal variances for all types of equipment, a Lavene's test should be run.

```
proc glm data = Powerlifting plots = diagnostics;
class Equipment;
model Age = Equipment;
means Equipment / hottest = Levene;
run;
```

Levene's Test for Homogeneity of Age Variance
ANOVA of Squared Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Equipment	3	1861049	620350	8.41	<.0001
Error	1869	1.3787E8	73769.3		

Since the p-value of the Lavene's test is below the significance level of 5%, the null hypothesis of homogenous group variances is rejected, which confirms the violation of equal variances amongst the four equipment types.

```
proc glm data = Powerlifting plots = diagnostics;
class Equipment;
model Age = Equipment;
means Equipment / Welch;
run;
```

Welch's ANOVA for Age

Source	DF	F Value	Pr > F
Equipment	3.0000	12.33	<.0001
Error	180.9		

Because the homogeneity of group variances has been rejected, we perform Welch's ANOVA and find that the p-value is very low. As a result, the null hypothesis of equal group means is rejected at $\alpha = 0.05$ and we conclude that there is a statistically significant difference in average age for each type of equipment.

Question G)

```
proc glm data = Powerlifting;
class Equipment;
model age = Equipment;
lsmeans Equipment / pdiff = all adjust = tukey;
run;
```

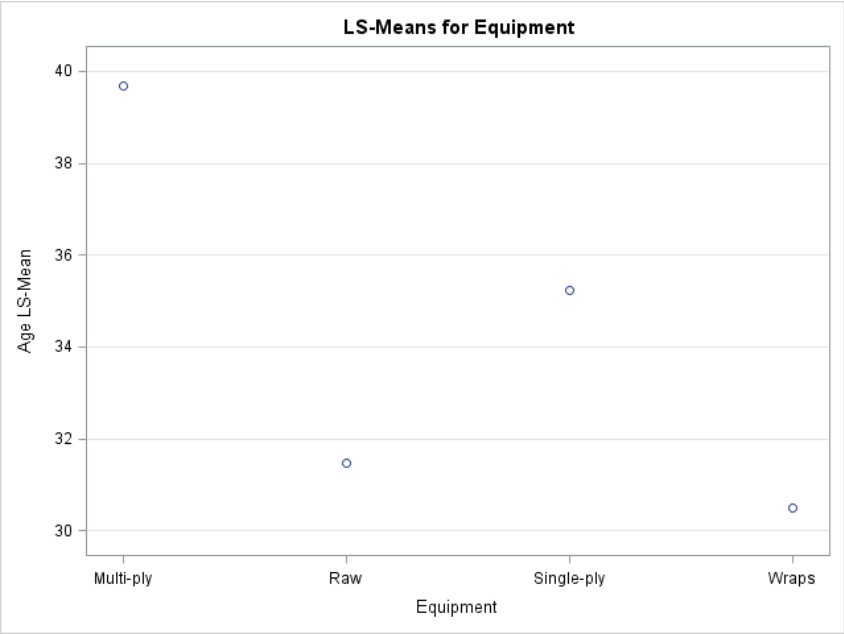
The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

Equipment	Age LSMEAN	LSMEAN Number
Multi-ply	39.6860465	1
Raw	31.4695269	2
Single-ply	35.2206897	3
Wraps	30.4829352	4

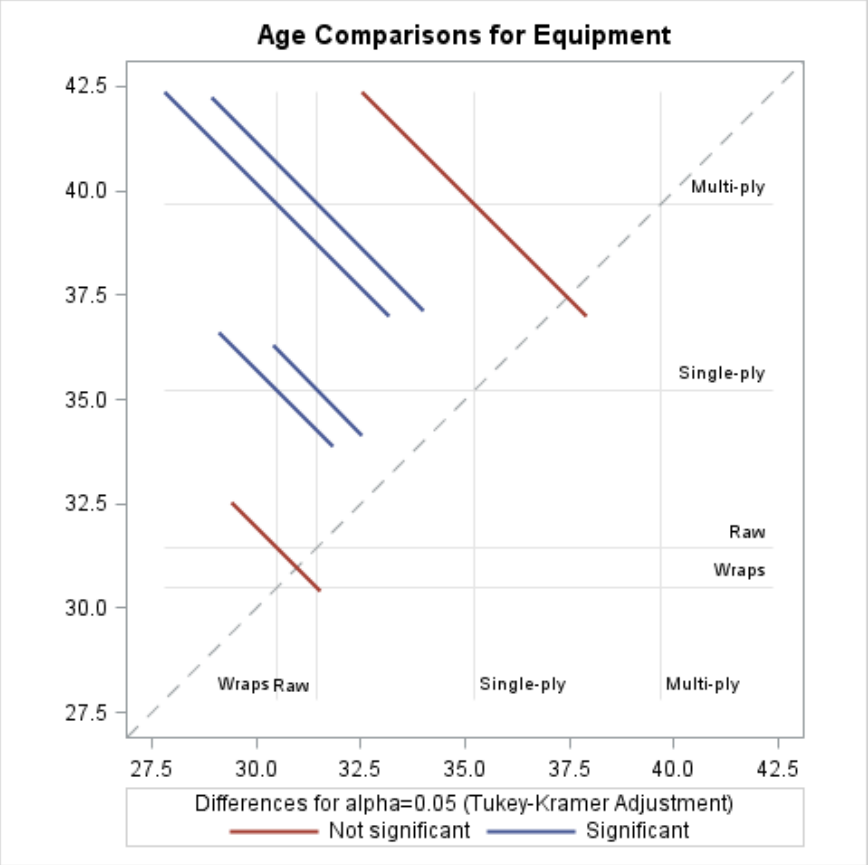
Least Squares Means for effect Equipment
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: Age

i/j	1	2	3	4
1		0.0002	0.1427	<.0001
2	0.0002		<.0001	0.6353
3	0.1427	<.0001		<.0001
4	<.0001	0.6353	<.0001	

The table above reports the p-values of pairwise mean comparisons. High p-values identify a statistically significant difference between the means of interest, while low p-values (below 5%) imply that there is no significant difference between the means. Thus, we conclude that only the pairs (Wraps, Raw) and (Single-ply, Multi-ply) do not have significantly different means.



The LS-Means graph shows the location of the means for each type of equipment.



The above diffogram displays which of the six pairwise comparisons of means are statistically significant under the Tukey-Kramer Adjustment. It can be observed that the blue line segments do not intersect the dashed reference line which indicates that the means for the pairs

- (Wraps, Single-ply)
- (Raw, Single-ply)
- (Wraps, Multi-ply)
- (Raw, Multi-ply)

are significantly different. Whereas the red line segments intersect the dashed line, as a result, the means for the pairs (Wraps, Raw) and (Single-ply, Multi-ply) are not significantly different. Hence, the diffogram and the table of p-values reach the same conclusion.

Question H)

```
proc reg data = Powerlifting;
model Wilks = Age/clb;
run;
```

Check model assumptions for linear regression:

The assumptions of independent errors, normally distributed errors and zero mean of errors seems to be clearly satisfied as the residual vs. predicted and the Q-Q plots indicates, whereas there seems to be the absence of a linear relationship between Wilks and the predicted values as well as evidence of non constant variance (increase with predicted values in the response).

Association between Wilks and Age:

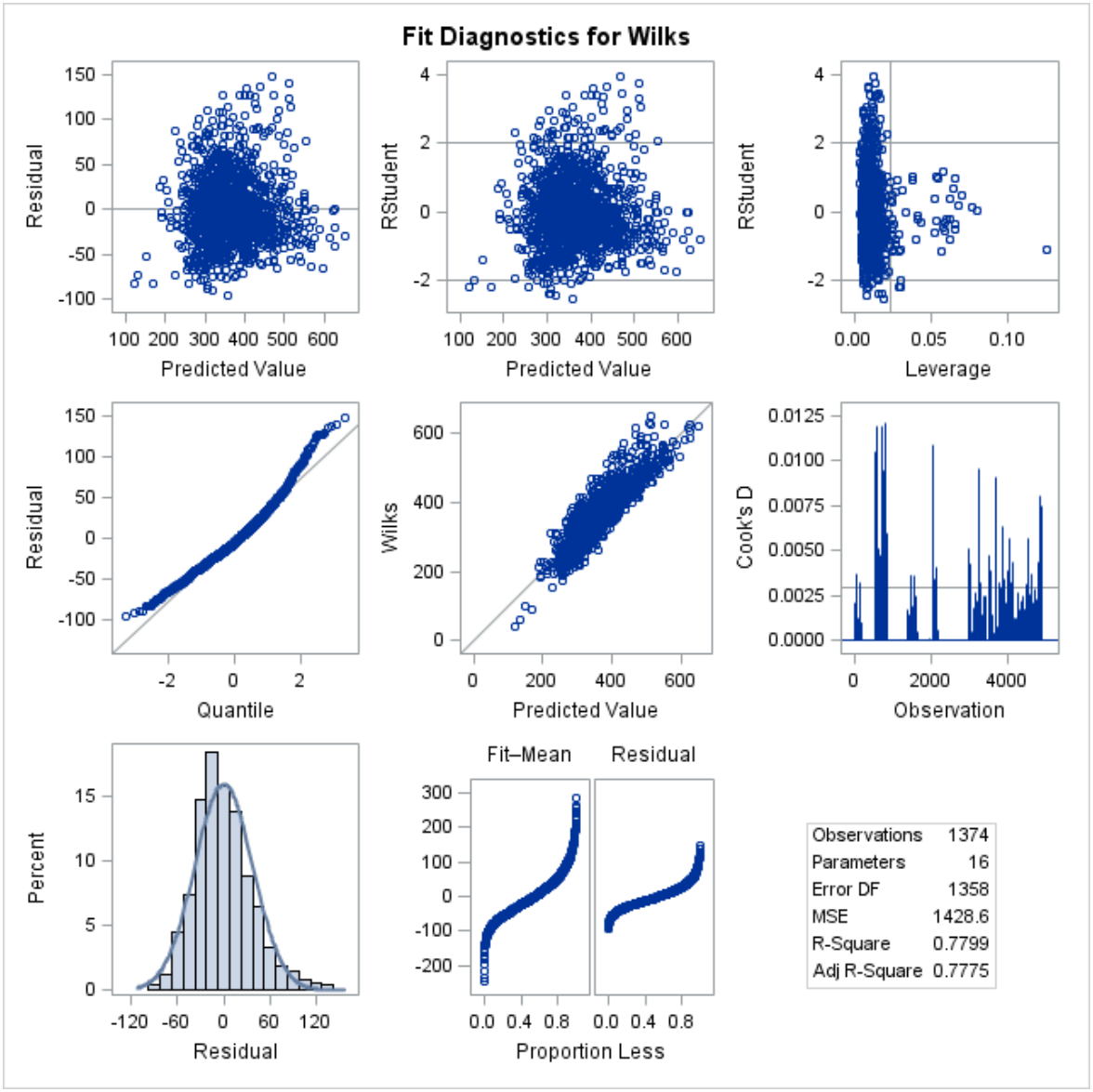
Fitting the linear regression where Age is the explanatory and Wilks is the dependent variable, returns a very low R-squared value of 0.0698, which declares a weak linear association between Wilks and Age (6.98% of the variation in the response is explained by age).

Question I)

```
data Powerlifting;
set Powerlifting;
if Place not in (1 2 3) then
Place2 = "Other";
else
Place2 = Place;
run;
```

Question J)

```
proc glm data = Powerlifting plots(only) = diagnostics;
class Sex Equipment Schedule Place2;
model Wilks = Sex Equipment LiquidConsumed Schedule Place2 GymCost Age
BestSquatKg BestBenchKg BestDeadLiftKg AverageTime / solution clparm;
run;
```



Question K)

Code for full model:

```
proc glmselect data = Powerlifting plots = all;
class Sex Equipment Schedule Place2;
model Wilks = Sex Equipment LiquidConsumed Schedule Place2 GymCost Age
BestSquatKg BestBenchKg BestDeadLiftKg AverageTime / selection = backward
select = SBC slstay = 0.05 showpvalues;
run;
```

The backward elimination method (using the Schwarz' Bayesian criterion) deletes one covariate at a time (those with the smallest contribution i.e. highest p-values) from the full model until the remaining variables in the model produce a statistically significant F-value at the 0.05 level.

Subsequent deletion of statistically insignificant explanatory variables lead to the removal of

- LiquidConsumed, Schedule, Place2, GymCost, and AverageTime

from the full model in five steps, hence the final (reduced) model returns the best covariates to be included. We notice the following changes after adding new variables to the final model:

- The progression for the adjusted R-Squared reaches the best criterion value after the third step, that is when the variables LiquidConsumed and Schedule are added to the final model.
- The progression of the Average Squared Errors show a slightly decreasing trend and reaches the minimum at 1412 in the full model.
- There seems to be a notable increase in the Schwarz' Bayesian information criterion from approximately 10032 to about 10080, reaching the best criterion value in the final model.
- According to the coefficient progression for Wilks, the following six variables: Sex, Equipment, Age, BestSquatKg, BestBenchKg, and BestDeadLiftKg are retained in the final model.

From the output we conclude that Equipment types multi-ply and raw do not seem to be statistically significant in the prediction of Wilks (high p-values), however, since Single-ply is significant the 5% level, the variable Equipment should not be dropped from the model. An Adjusted R-Squared value of 0.7775 in the final model indicates that 77.75% of the variation in the response variable (Wilks) is explained by the covariates in the model (good fit to the data).

Question L)

Code for final model:

```
proc glm data = Powerlifting plots(only) = diagnostics;
class Sex Equipment;
model Wilks = Sex Equipment Age BestSquatKg BestBenchKg BestDeadLiftKg /
solution clparm;
run;
```

Check model assumptions for multiple linear regression:

The residuals vs predicted plot appear to have a random scatter around 0 without a systematic pattern (errors are independent and have zero mean). The Q-Q plot suggests that residuals are closely aligned to the reference line, the normality of errors seems plausible. As the scale of variability changes for different predicted values, it is likely that the error variance is not constant. Finally, there seems to be a clear linear relationship between Wilks and the predicted values.

2. Part Two: Orthopaedics Data

Question M)

```
proc ttest data = Orthopaedics plots(shownull) = diagnostics;
class class;
var degree_spondylolisthesis;
run;
```

The equality of variance table suggests that the variances between the levels Normal and Abnormal are not equal (the result of F-test is not statistically significant at $\alpha = 0.05$) thus, the Satterthwaite approximation should be used to read the 95% confidence interval of (29.9198, 41.2624).

Since this CI for the mean difference (Abnormal – Normal) is wholly positive (i.e. the mean for Abnormal is likely to be higher than the mean for Normal), thus it is evident to reject the null hypothesis of equal population means between the levels (H_0: $\mu_{Abnorm} = \mu_{Norm}$).

Check of assumptions for two-sample t-test:

The orthopaedics dataset information explicitly mentions independence, furthermore, we do not have equal group variances (see above) and deviations in the tails of the Q-Q plot for each level make the normality assumption questionable (further test required for verification).

Question N)

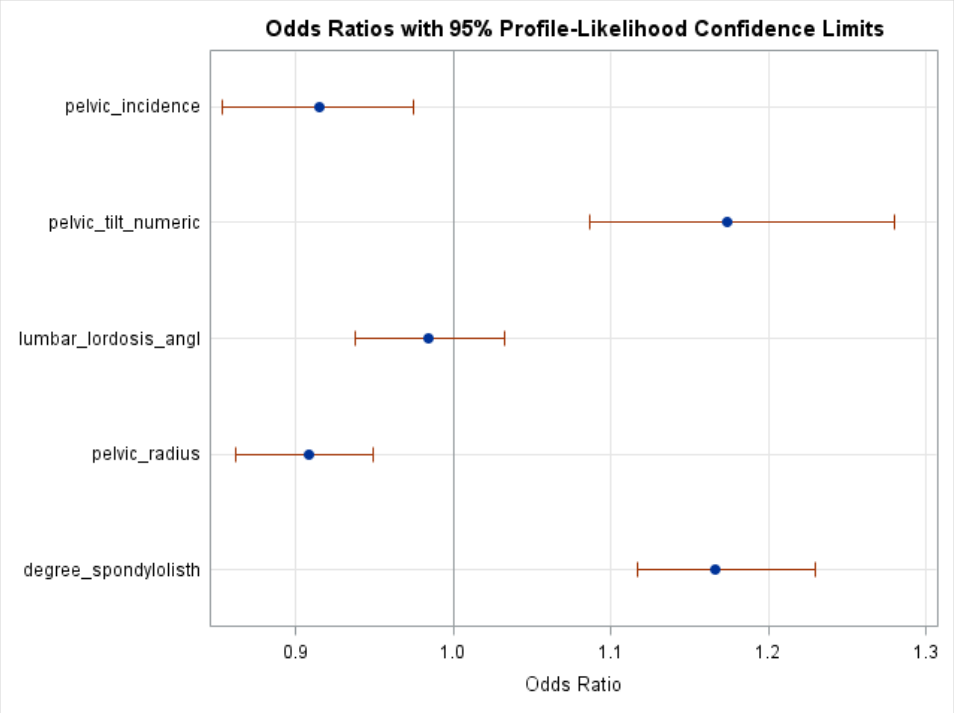
```
proc surveyselect noprint data = Orthopaedics samprate = 0.75 outall out =
Orthopaedics;
strata class;
run;


data train(drop = selected SelectionProb SamplingWeight) test(drop = selected
SelectionProb SamplingWeight);
set Orthopaedics;
if selected then output train;
else output test;
run;
```

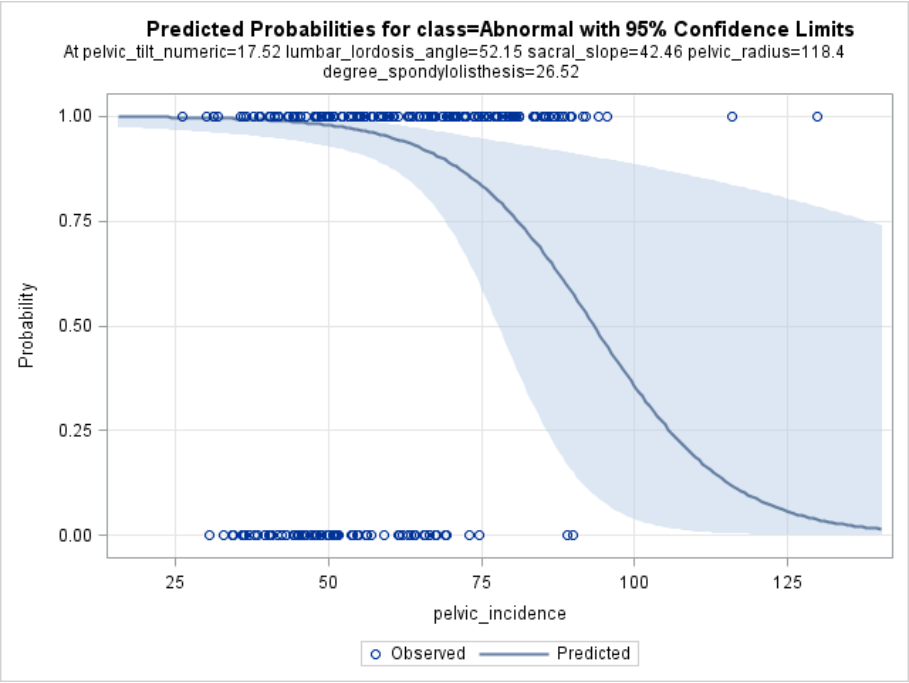
Question O)

Code for the model with degree spondylolisthesis

```
proc logistic data = Train plots(only)=(effect oddsratio);
class class(ref = 'Normal');
model class = pelvic_incidence pelvic_tilt_numeric lumbar_lordosis_angle
sacral_slope pelvic_radius degree_spondylolisthesis / clodds = pl;
run;
```



Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	13.5557	3.5002	14.9993	0.0001
pelvic_incidence	1	-0.0887	0.0338	6.8944	0.0086
pelvic_tilt_numeric	1	0.1607	0.0415	14.9870	0.0001
lumbar_lordosis_angl	1	-0.0155	0.0243	0.4077	0.5231
sacral_slope	0	0	.	.	.
pelvic_radius	1	-0.0960	0.0244	15.4813	<.0001
degree_spondylolisth	1	0.1541	0.0244	39.9638	<.0001

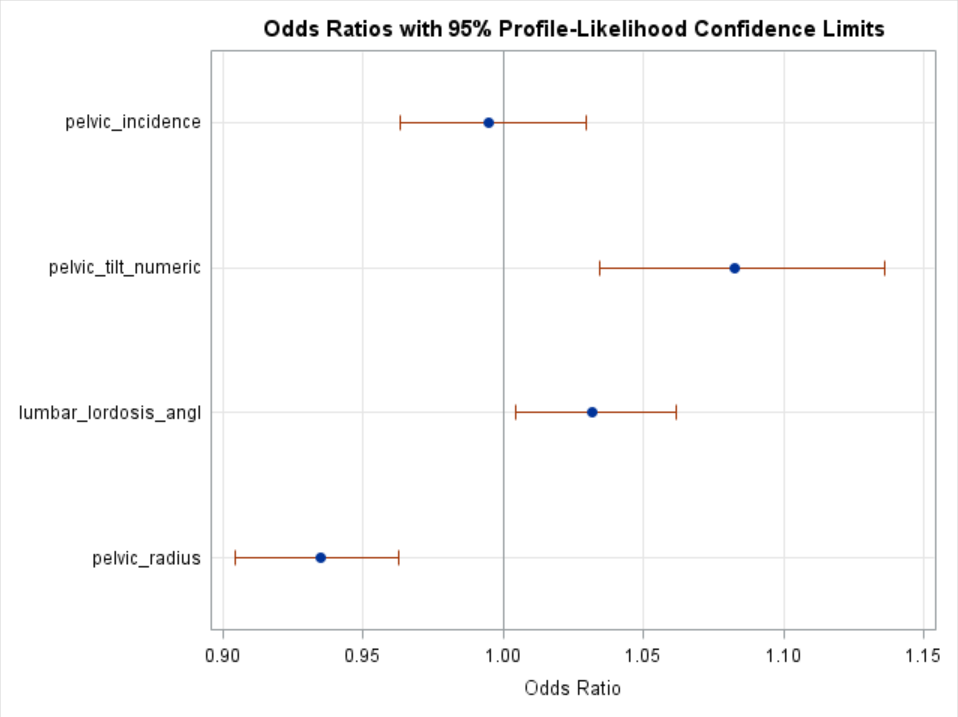


The above graph of the odds ratios with 95% profile-likelihood confidence limits shows that only lumbar_lordosis_angle crosses the vertical line and is not statistically significant, thus this predictor can be removed from the model.

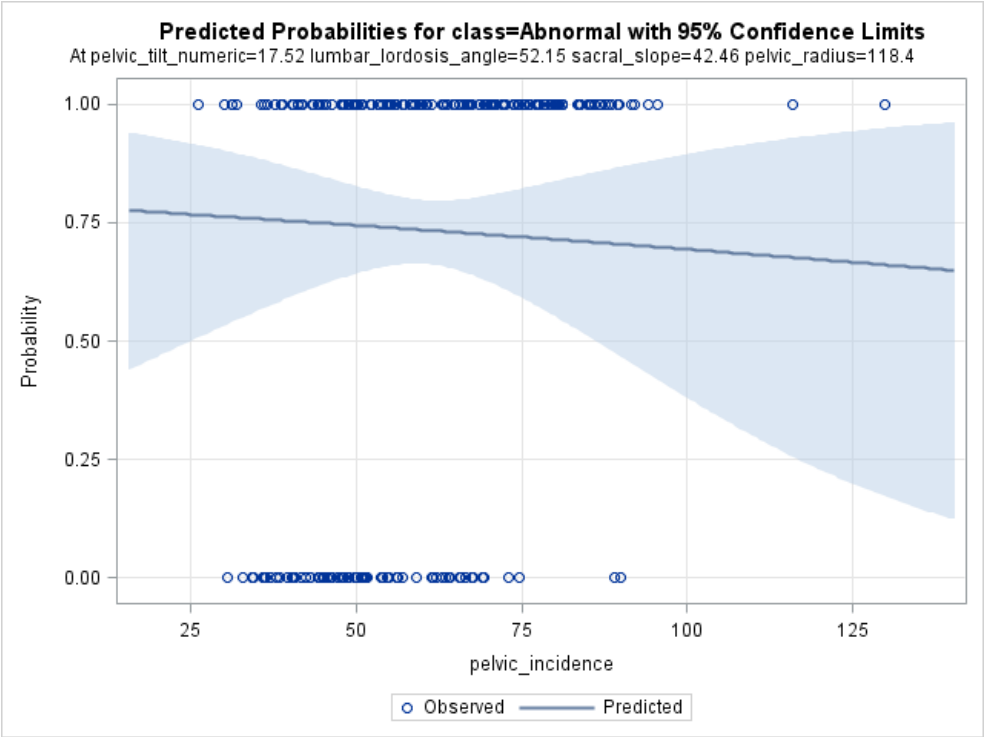
This result is confirmed by the Analysis of Maximum Likelihood Estimates (see table below) where lumbar_lordosis_angle has an associated p-value of 0.5231, whereas all other predictors are statistically significant with very low p-values (these should be kept in the model).

Code for the model without degree spondylolisthesis

```
proc logistic data = train plots(only) = (effect oddsratio);
class class(ref = 'Normal');
model class = pelvic_incidence pelvic_tilt_numeric lumbar_lordosis_angle
sacral_slope pelvic_radius / clodds = pl;
run;
```



Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	6.2941	2.0337	9.5784	0.0020
pelvic_incidence	1	-0.00503	0.0165	0.0935	0.7597
pelvic_tilt_numeric	1	0.0792	0.0238	11.0367	0.0009
lumbar_lordosis_angl	1	0.0313	0.0141	4.9304	0.0264
sacral_slope	0	0	.	.	.
pelvic_radius	1	-0.0675	0.0159	18.0918	<.0001



The graph of the odds ratios with 95% profile-likelihood confidence limits shows that only pelvic_incidence crosses the vertical line and is not statistically significant, thus this predictor can be excluded from the model.

This result is further reinforced by the Analysis of Maximum Likelihood Estimates where pelvic_incidence has an associated p-value of 0.7597, whereas all other predictors have statistically significant p-values and should be retained in the model. Since sacral_slope is set to 0 in both models, it is a redundant predictor.

Question P)

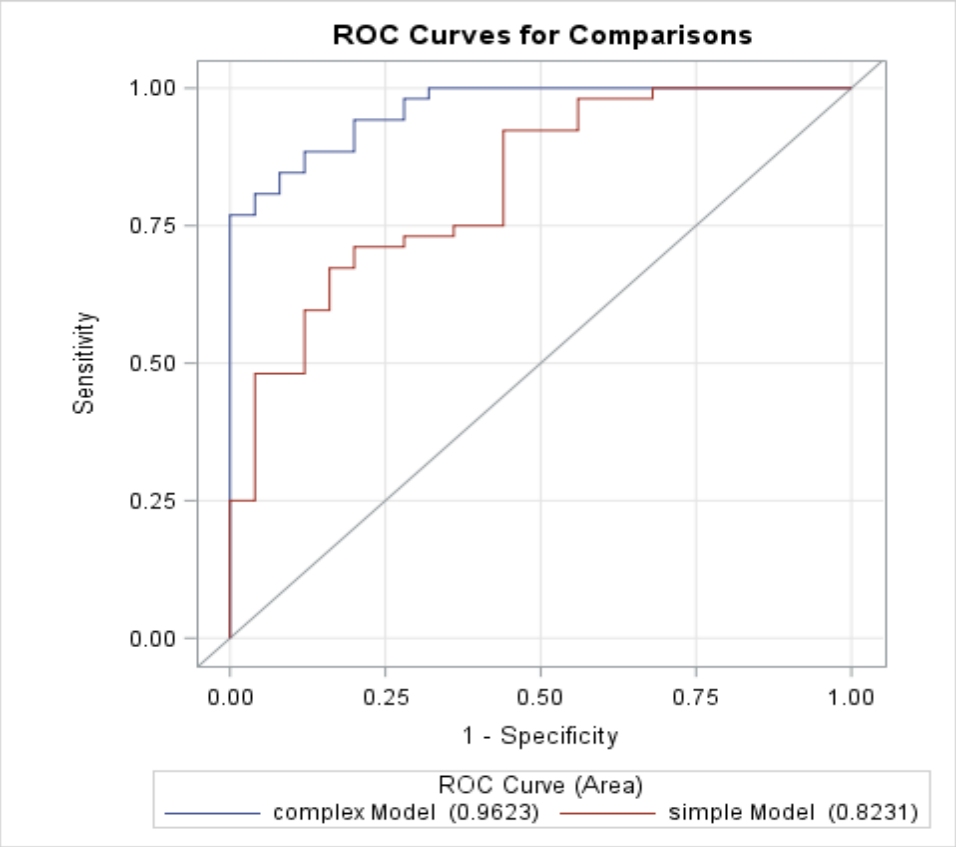
Code for the complex model:

```
proc logistic data = train plots = all;
class class (ref = "Normal");
model class = pelvic_incidence pelvic_tilt_numeric lumbar_lordosis_angle
sacral_slope pelvic_radius degree_spondylolisthesis / clodds = pl;
score data = test out = testAssess (rename = (P_Normal = p_complex)) outroc =
roc;
run;
```

Code for the simple model:

```
proc logistic data = train plots = all;
class class (ref = "Normal");
model class = pelvic_incidence pelvic_tilt_numeric lumbar_lordosis_angle
sacral_slope pelvic_radius / clodds = pl;
score data = testAssess out = testAssess (rename = (P_Normal = p_simple))
outroc = roc;
run;
```

```
proc logistic data = testAssess;
model class = p_complex p_simple / nofit;
roc "Complex Model" p_complex;
roc "Simple Model" p_simple;
roccontrast "Comparing Models";
run;
```



ROC Association Statistics

ROC Model	Mann-Whitney Area	Standard Error	95% Wald Confidence Limits		Somers' D	Gamma	Tau-a
complex Model	0.9623	0.0179	0.9273	0.9973	0.9246	0.9246	0.4108
simple Model	0.8231	0.0500	0.7251	0.9211	0.6462	0.6462	0.2871

Generally, models which cover a greater area under the ROC Curve are preferred, because they perform better with regards to predictive accuracy.

As can be seen from the ROC Association Statistics the complex model covers approximately 14% more area than the simple model, thus it is advisable to keep the explanatory variable degree_spondylolisthesis in the final model (we prefer the complex model).