STAT537: Statistics for Research I: Final

Due on Nov. 1, 2016

Dr. Schmidhammer TR 11:10am – 12:25pm

Wenqiang Feng

Contents

Problem 1

Ads data set analysis

Solution. (a) **Perform an analysis of variance on these data, and test the hypothesis:** From the summary of the ANOVA test,we get the p-vaule is $8.85e-08 < 0.05$, hence reject H_0. Therefore, there is at least one of the means is different from the others.

```
              Df Sum Sq Mean Sq F value   Pr(>F)
factor(ad)     3   5866    1955   13.48 8.85e-08 ***
Residuals    140  20303     145
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```

(b) **Use Levene's Test to test the assumption that the variances in the 4 groups are equal. State your conclusions.** From the following Levene's Test, we can see that the p-vaule is $0.481 > 0.05$, hence there is no enough information to reject H_0. Therefore, the assumption that the variances in the 4 groups are equal is valid.

```
> leveneTest(sales~factor(ad), rawdata)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   3  0.8271  0.481
      140
```

(c) **Check to see if the residuals are modeled well by a normal distribution. State your conclusions.** From the Normal QQ plot of the residuals in Figure.1, we can conclude that residuals are modeled well by a normal distribution, since most of the point located on the reference line.

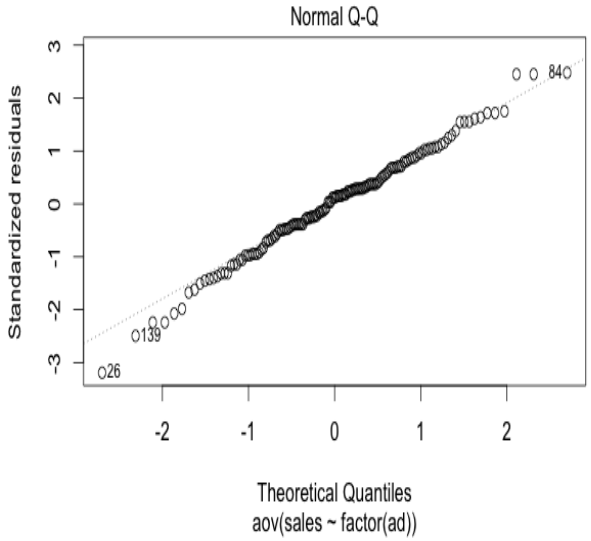

Figure 1: Normal QQ plot of the residuals.

(d) **If you determine that the means are not equal in part (a), use Tukey's HSD procedure to determine which means are different. State your conclusions.** From the p_{adj} value we can see that the means of paper-display (p-value=0.0000002), people-display (p-value=0.0029955) and radio-display (p-value=0.0000080) are significantly different from each other at 95% level. Moreover, we can see the difference from the figure.2

```
> TukeyHSD(onewayAOV,conf.level=0.95)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = sales ~ factor(ad), data = rawdata)

$`factor(ad)`
                     diff        lwr        upr     p adj
paper-display  16.666667   9.286236 24.0470971 0.0000002
people-display 10.055556   2.675125 17.4359860 0.0029955
radio-display  14.333333   6.952903 21.7137637 0.0000080
people-paper   -6.611111 -13.991542  0.7693193 0.0963573
radio-paper    -2.333333  -9.713764  5.0470971 0.8439578
radio-people    4.277778  -3.102653 11.6582082 0.4360116
```

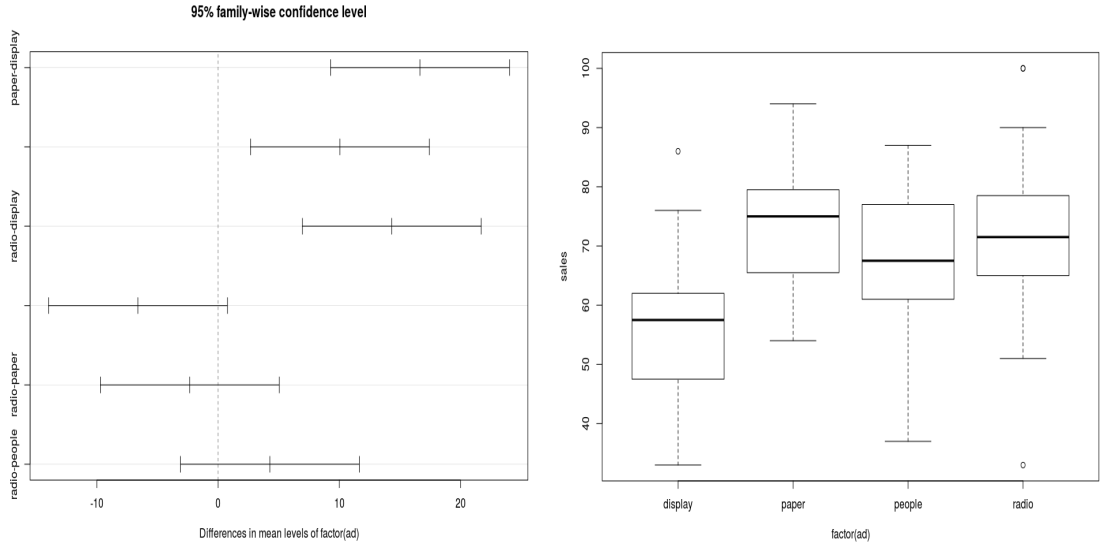

Figure 2: Differences in mean levels of ad and boxplots of ad.

(e) **Rank the data, and perform analysis of variance on the ranks. State the hypothesis being tested, and state your conclusions. Since this is an alternative way to compare the four types of advertising, comment on its appropriateness.** The null hypothesis is that the distributions are the same in the four advertising. From the summary of the Kruskal-Wallis rank sum test, we can see that the p-value is $2.358e-07 < 0.05$, hence reject H_0. Therefore, at least two of these four advertising have different distributions.

```
> kruskal.test(sales~factor(ad), rawdata)
```

```
Kruskal-Wallis rank sum test

data:  sales by factor(ad)
Kruskal-Wallis chi-squared = 33.642, df = 3, p-value = 2.358e-07
```

Based on the results, we know there is a difference among the groups. However, just like ANOVA, we do not know which groups were different. We have to do a post-hoc test in order to determine where the difference in means is among the three groups.

(f) **If you determine that the distributions are not equal in part (e), use Tukey's HSD procedure to determine which groups are different. State your conclusions.** To do this we apply "PCMR" package and do a new analysis. Ran the function "posthoc.kruskal.nemenyi.test" and place the appropriate variables in their place and then indicated the type of posthoc test 'Tukey', we get the following results which indicate that display - paper (p-value=3.8e-07), display - people (p-value=0.0034) and display-radio (p-value =3.9e-05) are different to each other.

```
> posthoc.kruskal.nemenyi.test(x=sales, g=factor(ad), dist='Tukey')

Pairwise comparisons using Tukey and Kramer (Nemenyi) test
                    with Tukey-Dist approximation for independent samples

data:  sales and factor(ad)

       display paper  people
paper  3.8e-07 -      -
people 0.0034  0.1942 -
radio  3.9e-05 0.8034 0.7014
```

□

Problem 2

Fitness data set analysis. Since the Stepwise Regression results from R and SAS are totally different, hence I present both of those results.

Solution. (a) **Fit the model with all predictors**

(i) State the assumptions of the regression model. Plot the residuals against each predictor, and against the predicted values, and determine if the residual plots confirm the assumptions of the regression model. State your conclusions.

- **Linear regression makes several key assumptions:**
  - Linear relationship
  - Multivariate normality
  - No or little multicollinearity
  - No auto-correlation
  - Homoscedasticity

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	89.33638	37.36740	2.39	0.0258
Gender	1	-0.84581	1.22146	-0.69	0.4959
Runtime	1	-2.04253	2.25517	-0.91	0.3749
Age	1	-0.19463	0.10889	-1.79	0.0877
Weight	1	-0.04747	0.07192	-0.66	0.5160
Run_Pulse	1	-0.35830	0.12493	-2.87	0.0089
Rest_Pulse	1	-0.00383	0.07341	-0.05	0.9589
Maximum_Pulse	1	0.29930	0.14174	2.11	0.0463
Performance	1	0.25055	1.03556	0.24	0.8111

Figure 3: Fitted model with all predictors via SAS.

```
Call:
glm(formula = Oxygen_Consumption ~ Gender + Runtime + Age + Weight +
    Run_Pulse + Rest_Pulse + Maximum_Pulse + Performance, data = rawdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.8747  -0.7791  -0.0437   0.9450   5.6979

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   89.336380  37.367403   2.391  0.02580 *
GenderM       -0.845807   1.221460  -0.692  0.49590
Runtime       -2.042532   2.255171  -0.906  0.37490
Age           -0.194627   0.108888  -1.787  0.08766 .
Weight        -0.047472   0.071917  -0.660  0.51605
Run_Pulse     -0.358297   0.124927  -2.868  0.00894 **
Rest_Pulse    -0.003828   0.073409  -0.052  0.95888
Maximum_Pulse  0.299299   0.141741   2.112  0.04631 *
Performance    0.250553   1.035565   0.242  0.81106
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

(Dispersion parameter for gaussian family taken to be 5.761786)

    Null deviance: 851.55  on 30  degrees of freedom
Residual deviance: 126.76  on 22  degrees of freedom
AIC: 151.63

Number of Fisher Scoring iterations: 2
```

- **Plot the residuals against each predictor, and against the predicted values**

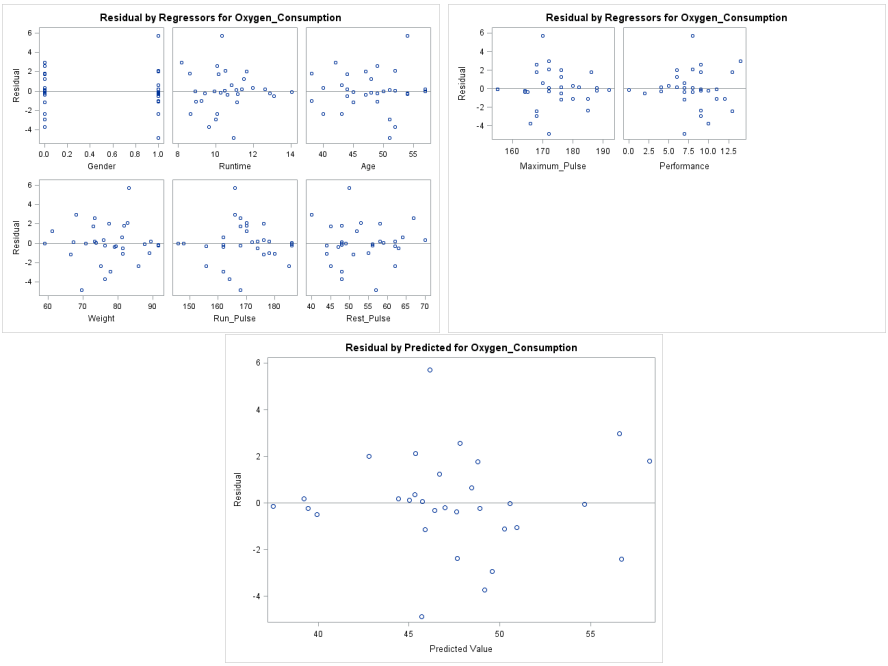

Figure 4: The residuals against each predictor and against the predicted values via SAS.

- **Determine if the residual plots confirm the assumptions of the regression model. State your conclusions.** From the residual plots in Figure 4 and 6, we may conclude that the residual plots confirm the assumptions of the regression model. Since residual plots in Figure 4 indicate residuals have linear patterns. And the Normal Q-Q plots in Figure 6 show the residuals are normally distributed.

(ii) **Plot the studentized residuals and Cook's D against the observation number, and determine if any influential observations exist. State your conclusions.** I apply the cutoff line $4/n = 0.1290323$ to identify D values which is above the cutoff line. Figure 5 indicates that the observation 2, 15, 20 are influential observations.

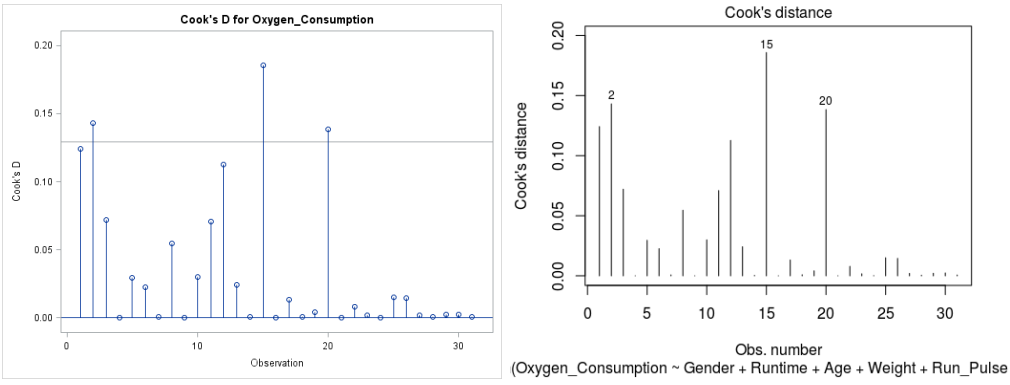

Figure 5: The studentized residuals and Cook's D against the observation number via SAS and R. Left: via SAS; Left: via R.

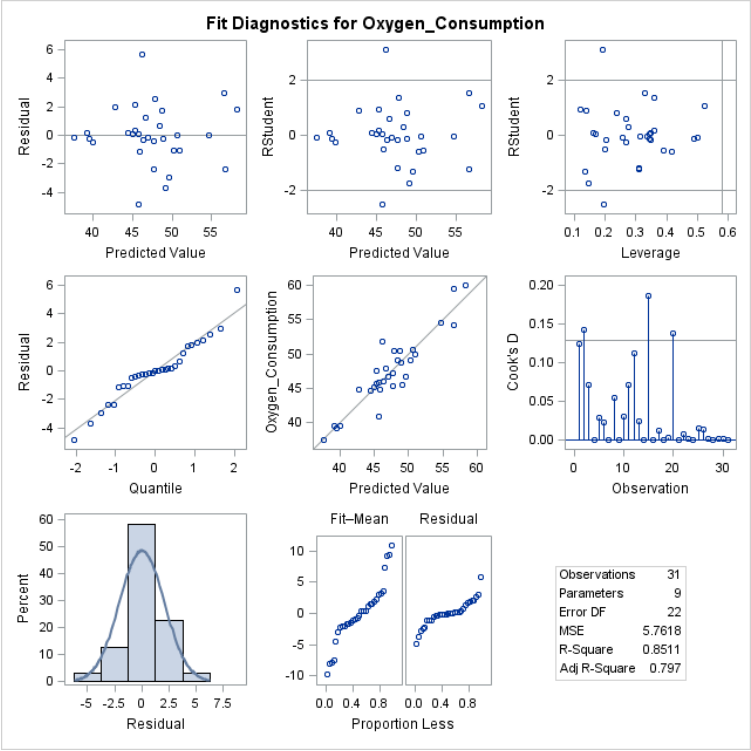

Figure 6: Fitted plots via SAS.

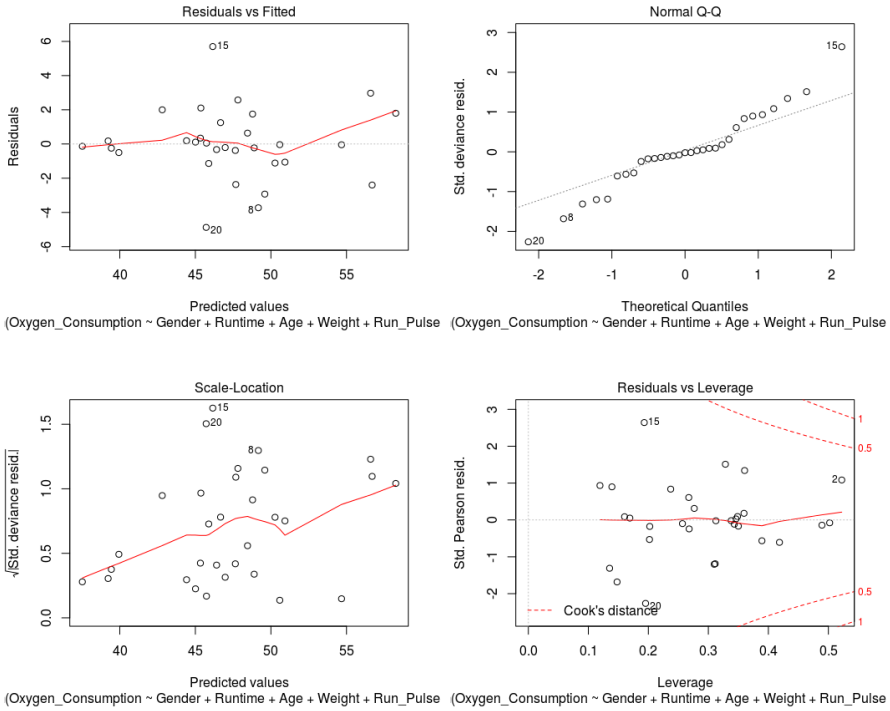

Figure 7: Fitted plots via R.

(b) **Use Stepwise Regression to select variables important to the model. Fit the final model that you choose, and interpret this model.** The results from SAS and R are totally different for this part.

- **Results from SAS:**
  - **Without 2, 15, 20 influential observations** The SAS results indicate that the only significant predictor variable is *Runtime, Run_pulse, Age, Weight, Maximum_Pulse*. The final model is

$$\begin{aligned} Oxygen\_Consumption &= 107.83506 - 2.47167 \cdot Runtime - 0.24248 \cdot Age - 0.11899 \cdot Weight \\ &\quad -0.29215 \cdot Run\_pluse + 0.20670 \cdot Maximum\_Pulse. \end{aligned}$$

    Thus, we may conclude that the ability to use oxygen in the blood stream will
    - decrease 2.47167 units, when the Runtime increases 1 unit and with the other variables fixed;
    - decrease 0.24248 units, when the Age increases 1 unit and with the other variables fixed;
    - decrease 0.11899 units, when the Weight increases 1 unit and with the other variables fixed;
    - decrease 0.29215 units, when the Run_pluse increases 1 unit and with the other variables fixed;
    - increase 0.20670 units, when the Maximum_Pulse increases 1 unit and with the other variables fixed;

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Runtime		1	0.7863	0.7863	20.1248	95.66	<.0001
2	Run_Pulse		2	0.0342	0.8205	15.0711	4.76	0.0388
3	Age		3	0.0402	0.8607	8.7681	6.93	0.0146
4	Weight		4	0.0315	0.8922	4.2661	6.72	0.0163
5	Maximum_Pulse		5	0.0125	0.9046	3.6881	2.88	0.1038

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	107.83506	8.74748	410.96628	151.97	<.0001
Runtime	-2.47167	0.25145	261.29183	96.62	<.0001
Age	-0.24248	0.07142	31.17213	11.53	0.0026
Weight	-0.11899	0.04006	23.85331	8.82	0.0071
Run_Pulse	-0.29215	0.10428	21.22536	7.85	0.0104
Maximum_Pulse	0.20670	0.12179	7.79008	2.88	0.1038

Figure 8: Stepwise Regression via SAS without influential observations.

  - **With influential observations** The SAS results indicate that the only significant predictor variable is *Performance*, with a coefficient of 1.47507. The final model is

$$Oxygen\_Consumption = 35.57526 + 1.47507 \cdot Performance.$$

    Thus, we may conclude that the ability to use oxygen in the blood stream will increase 1.47507 units, when the overall fitness increases 1 unit.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	635.34150	635.34150	85.22	<.0001
Error	29	216.21305	7.45562		
Corrected Total	30	851.55455			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	35.57526	1.36917	5033.48080	675.13	<.0001
Performance	1.47507	0.15979	635.34150	85.22	<.0001

Figure 9: Stepwise Regression via SAS with influential observations.

  - **Results from R:**

* **Without 2, 15, 20 influential observations**

```
Step:  AIC=114.5
Oxygen_Consumption ~ Age + Weight + Run_Pulse + Maximum_Pulse +
    Performance

                Df Deviance    AIC F value    Pr(>F)
<none>               59.35 114.50
+ Runtime        1    57.76 115.74  0.5782  0.455470
+ Rest_Pulse     1    58.72 116.19  0.2282  0.637780
+ Gender         1    59.29 116.47  0.0234  0.879983
- Maximum_Pulse  1    69.59 116.95  3.7931  0.064338 .
- Run_Pulse      1    83.44 122.03  8.9259  0.006786 **
- Age            1    87.21 123.27 10.3262  0.004003 **
- Weight         1    91.13 124.50 11.7766  0.002384 **
- Performance    1   320.79 159.74 96.9017 1.606e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> model_step$coefficients
  (Intercept)           Age        Weight     Run_Pulse Maximum_Pulse   Perfo
   71.3921403    -0.2304163    -0.1362464    -0.3105311     0.2369593     1.1
```

* **With influential observations** The R results indicate that significant predictor variables are *Age, Maximum_Pulse, Run_Pulse, Runtime*. The results from R is as follows:

```
Step:  AIC=146.27
Oxygen_Consumption ~ Gender + Runtime + Age + Run_Pulse + Maximum_Pulse

                Df Deviance    AIC F value    Pr(>F)
<none>              129.40 146.27
- Gender         1   140.10 146.73  2.0674  0.162876
+ Weight         1   127.20 147.74  0.4159  0.525094
+ Performance    1   129.28 148.24  0.0218  0.883795
+ Rest_Pulse     1   129.40 148.27  0.0000  0.999804
- Age            1   149.65 148.78  3.9110  0.059095 .
- Maximum_Pulse  1   153.22 149.51  4.6012  0.041855 *
- Run_Pulse      1   175.55 153.73  8.9152  0.006247 **
- Runtime        1   398.04 179.10 51.9004 1.502e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> model_step$coefficients
  (Intercept)     GenderM    Runtime       Age     Run_Pulse Maximum_Pulse
   93.6490107  -1.3167163   -2.5882956  -0.1831225  -0.3459081     0.2829902
```

(c) **Are there any other models that perform nearly as well?** The forward and backward selected, and the model consider the interactions will perform nearly as well. I will use AIC as the criteria to judge the goodness of the fit. From the following comparison results and Table 1, we may conclude that:

- the models without influential observations is better than the models with influential observations

- backward selected models are a little bit better than the forward selected models
- the model with Age* Maximum_Pulse (without influential observations) and Gender * Age (with influential observations) interaction are the best one according to the AIC scores.

Table 1: AIC scores for different models

variable	all	1 (Performance)	forward selected	backward selected	interaction
AIC(wo)	119.5673	130.06	119.5673	114.5	113.24
AIC(w)	151.63	154.18	149.64	146.27	144.42

(i) **The following are for without influential observations**

- **Results from SAS:**

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	89.04470	25.27827	34.13156	12.41	0.0020
Runtime	-1.21757	1.60123	1.59044	0.58	0.4555
Age	-0.23497	0.07265	28.77369	10.46	0.0040
Weight	-0.12735	0.04176	25.58240	9.30	0.0061
Run_Pulse	-0.30077	0.10573	22.25847	8.09	0.0097
Maximum_Pulse	0.22221	0.12437	8.78048	3.19	0.0884
Performance	0.57176	0.72081	1.73068	0.63	0.4365

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	71.39214	9.90665	140.11172	51.93	<.0001
Age	-0.23042	0.07170	27.85915	10.33	0.0040
Weight	-0.13625	0.03970	31.77222	11.78	0.0024
Run_Pulse	-0.31053	0.10394	24.08138	8.93	0.0068
Maximum_Pulse	0.23696	0.12167	10.23343	3.79	0.0643
Performance	1.11294	0.11306	261.43208	96.90	<.0001

Summary of Forward Selection

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Runtime	1	0.7863	0.7863	20.1248	95.66	<.0001
2	Run_Pulse	2	0.0342	0.8204	15.0711	4.76	0.0388
3	Age	3	0.0402	0.8607	8.7681	6.93	0.0146
4	Weight	4	0.0315	0.8922	4.2661	6.72	0.0163
5	Maximum_Pulse	5	0.0125	0.9046	3.6881	2.88	0.1038
6	Performance	6	0.0028	0.9074	5.1154	0.63	0.4365

Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Gender	7	0.0001	0.9078	7.0281	0.03	0.8686
2	Rest_Pulse	6	0.0004	0.9074	5.1154	0.09	0.7651
3	Runtime	5	0.0025	0.9049	3.6417	0.58	0.4555

Figure 10: Forward and backward Stepwise Regression via SAS without influential observations.

- **Results from R:**

```
> model_step =step(fit11,direction = "forward",test="F")
Start:  AIC=119.57
Oxygen_Consumption ~ Gender + Runtime + Age + Weight + Run_Pulse +
    Rest_Pulse + Maximum_Pulse + Performance

> extractAIC(model_step, scale)
[1]   9.0000 119.5673
> model_step$coefficients
  (Intercept)    GenderM    Runtime        Age         Weight     Run_Pulse
  85.15557728 -0.16244719   -1.06729264 -0.22648084 -0.12141017   -0.3026162
   Rest_Pulse Maximum_Pulse   Performance
   0.01720141    0.22370992    0.65195244

 > model_step =step(fit11,direction = "backward",test="F")
Step:  AIC=114.5
Oxygen_Consumption ~ Age + Weight + Run_Pulse + Maximum_Pulse +
    Performance

                Df Deviance    AIC F value    Pr(>F)
```

```
<none>                59.35 114.50
- Maximum_Pulse  1    69.59 116.95  3.7931  0.064338 .
- Run_Pulse      1    83.44 122.03  8.9259  0.006786 **
- Age            1    87.21 123.27 10.3262  0.004003 **
- Weight         1    91.13 124.50 11.7766  0.002384 **
- Performance    1   320.79 159.74 96.9017 1.606e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> extractAIC(model_step, scale)
[1]   6.0000 114.4974
> model_step$coefficients
  (Intercept)       Age    Weight     Run_Pulse Maximum_Pulse   Performance
   71.3921403    -0.2304163    -0.1362464    -0.3105311     0.2369593     1.1
```

(ii) **The following are for with influential observations**

- **Results from SAS:**

Summary of Forward Selection

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Performance	1	0.7461	0.7461	10.5254	85.22	<.0001
2	Run_Pulse	2	0.0179	0.7640	9.8764	2.13	0.1559
3	Maximum_Pulse	3	0.0432	0.8072	5.4925	6.05	0.0206
4	Age	4	0.0181	0.8253	4.8212	2.69	0.1130
5	Weight	5	0.0168	0.8421	4.3383	2.66	0.1155
6	Runtime	6	0.0056	0.8476	5.5167	0.88	0.3587
7	Gender	7	0.0035	0.8511	7.0027	0.54	0.4710

Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Rest_Pulse	7	0.0000	0.8511	7.0027	0.00	0.9589
2	Performance	6	0.0005	0.8506	5.0762	0.08	0.7842
3	Weight	5	0.0026	0.8480	3.4587	0.42	0.5251
4	Gender	4	0.0126	0.8355	3.3160	2.07	0.1629

Figure 11: Forward and backward Stepwise Regression via SAS.

- **Results from R:**

```
> model_step =step(fit1,direction = "forward",test="F")
Start:  AIC=151.63
Oxygen_Consumption ~ Gender + Runtime + Age + Weight + Run_Pulse +
    Rest_Pulse + Maximum_Pulse + Performance

> model_step =step(fit1,direction = "backward",test="F")
Step:  AIC=146.27
Oxygen_Consumption ~ Gender + Runtime + Age + Run_Pulse + Maximum_Pulse

                Df Deviance    AIC F value    Pr(>F)
<none>               129.40 146.27
- Gender         1   140.10 146.73  2.0674  0.162876
- Age            1   149.65 148.78  3.9110  0.059095 .
- Maximum_Pulse  1   153.22 149.51  4.6012  0.041855 *
- Run_Pulse      1   175.55 153.73  8.9152  0.006247 **
- Runtime        1   398.04 179.10 51.9004 1.502e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> model_step$coefficients
  (Intercept)  GenderM     Runtime       Age     Run_Pulse Maximum_Pulse
   93.6490107  -1.3167163  -2.5882956 -0.1831225  -0.3459081     0.2829902
```

- **Comparison:**
  - **With all variables:** AIC = 151.63

```
Call:
glm(formula = Oxygen_Consumption ~ Gender + Runtime + Age + Weight +
    Run_Pulse + Rest_Pulse + Maximum_Pulse + Performance, data = rawdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.8747  -0.7791  -0.0437   0.9450   5.6979

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   89.336380  37.367403   2.391  0.02580 *
GenderM       -0.845807   1.221460  -0.692  0.49590
Runtime       -2.042532   2.255171  -0.906  0.37490
Age           -0.194627   0.108888  -1.787  0.08766 .
Weight        -0.047472   0.071917  -0.660  0.51605
Run_Pulse     -0.358297   0.124927  -2.868  0.00894 **
Rest_Pulse    -0.003828   0.073409  -0.052  0.95888
Maximum_Pulse  0.299299   0.141741   2.112  0.04631 *
Performance    0.250553   1.035565   0.242  0.81106
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 5.761786)

    Null deviance: 851.55  on 30  degrees of freedom
Residual deviance: 126.76  on 22  degrees of freedom
AIC: 151.63

Number of Fisher Scoring iterations: 2
```

  - **With all variables *Performance*:** AIC = 154.18

```
Call:
glm(formula = Oxygen_Consumption ~ Performance, data = rawdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.0607  -2.0732   0.0942   1.7569   5.3089

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  35.5753     1.3692  25.983  < 2e-16 ***
Performance   1.4751     0.1598   9.231 3.92e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 7.455623)
```

```
    Null deviance: 851.55  on 30  degrees of freedom
Residual deviance: 216.21  on 29  degrees of freedom
AIC: 154.18

Number of Fisher Scoring iterations: 2
```

- **With selected variables by Forward method:** AIC = 149.64

```
Call:
glm(formula = Oxygen_Consumption ~ Gender + Runtime + Weight +
    Age + Run_Pulse + Maximum_Pulse + Performance, data = rawdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.8772  -0.7837  -0.0383   0.9423   5.7092

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  88.61271   33.93409   2.611  0.01561 *
GenderM      -0.85841    1.17106  -0.733  0.47095
Runtime      -2.01439    2.14162  -0.941  0.35669
Weight       -0.04689    0.06948  -0.675  0.50653
Age          -0.19281    0.10090  -1.911  0.06855 .
Run_Pulse    -0.35871    0.12195  -2.942  0.00733 **
Maximum_Pulse  0.29949    0.13859   2.161  0.04135 *
Performance    0.26711    0.96407   0.277  0.78421
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 5.511955)

    Null deviance: 851.55  on 30  degrees of freedom
Residual deviance: 126.77  on 23  degrees of freedom
AIC: 149.64

Number of Fisher Scoring iterations: 2
```

- **With selected variables by Backward method:** AIC = 146.27

```
glm(formula = Oxygen_Consumption ~ Gender + Runtime + Age + Run_Pulse +
    Maximum_Pulse, data = rawdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.3730  -0.7563  -0.1686   1.1909   5.4558

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   93.6490    11.6843   8.015 2.27e-08 ***
GenderM       -1.3167     0.9157  -1.438  0.16288
Runtime       -2.5883     0.3593  -7.204 1.50e-07 ***
```

```
Age            -0.1831     0.0926  -1.978  0.05910 .
Run_Pulse      -0.3459     0.1158  -2.986  0.00625 **
Maximum_Pulse   0.2830     0.1319   2.145  0.04185 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 5.176099)

    Null deviance: 851.55  on 30  degrees of freedom
Residual deviance: 129.40  on 25  degrees of freedom
AIC: 146.27
```

– **With interactions:** AIC = 144.42

```
glm(formula = Oxygen_Consumption ~ Gender * Age + Gender + Runtime +
    Age + Run_Pulse + Maximum_Pulse, data = rawdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.7614  -1.0249   0.0043   1.2024   4.6356

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   103.6772    12.5410   8.267 1.76e-08 ***
GenderM       -14.9960     7.7278  -1.941   0.0641 .
Age            -0.3472     0.1279  -2.714   0.0121 *
Runtime        -2.6626     0.3471  -7.671 6.59e-08 ***
Run_Pulse      -0.2763     0.1178  -2.346   0.0276 *
Maximum_Pulse   0.2065     0.1336   1.545   0.1353
GenderM:Age     0.2880     0.1617   1.782   0.0875 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 4.761918)

    Null deviance: 851.55  on 30  degrees of freedom
Residual deviance: 114.29  on 24  degrees of freedom
AIC: 144.42
```

□

Appendix

R code for Final

Listing 1: Source code for problem 1

```
# reference: http://www.stat.columbia.edu/~martin/W2024/R3.pdf
rm(list = ls())
# set the path or enverionment
```

```r
#linux
#setwd("/home/feng/Dropbox/UTK_Course/Stat537/Final/data")
#Mac
setwd("/Users/wenqiangfeng/Dropbox/UTK_Course/Stat537/Final/data")

#install.packages("readxl") # CRAN version
library(readxl)

rawdata = read_excel("Ads.xlsx",sheet = 1)
attach(rawdata)

# 1.a
onewayAOV=aov(sales~factor(ad), rawdata)
summary(onewayAOV)

# 1.b
library(car)
leveneTest(sales~factor(ad), rawdata)

# 1.c
plot(onewayAOV, which=2)

# 1.d
T_HSD<-TukeyHSD(onewayAOV,conf.level=0.95)
plot(T_HSD)
plot(onewayAOV)
plot(sales~factor(ad), data=rawdata)

# 1.e
k_test<-kruskal.test(sales~factor(ad), rawdata)

# 1.f
#nstall.packages("agricolae")
#install.packages('PMCMR')
library(PMCMR)
kruskal<-posthoc.kruskal.nemenyi.test(x=sales, g=factor(ad), dist='Tukey')
kruskal
```

Listing 2: Source code for problem 2

```r
# reference: http://www.stat.columbia.edu/~martin/W2024/R3.pdf
rm(list = ls())
# set the path or enverionment
#linux
setwd("/home/feng/Dropbox/UTK_Course/Stat537/Final/data")
#Mac
#setwd("/Users/wenqiangfeng/Dropbox/UTK_Course/Stat537/Final/data")

#install.packages("readxl") # CRAN version
library(readxl)

rawdata = read_excel("Fitness.xlsx",sheet = 1)
attach(rawdata)
```

```
rawdata$Gender <- as.factor(rawdata$Gender)
# 2. (a)
fit1 = glm(Oxygen_Consumption~Gender+Runtime
           +Age+Weight+Run_Pulse+Rest_Pulse
           +Maximum_Pulse+Performance,data=rawdata)
summary(fit1)
extractAIC(fit1, scale)

plot(fit1)
library(car)
#fit1$residuals
# Cook's D plot
# identify D values > 4/(n-k-1)
#cutoff <- 4/((nrow(rawdata)-length(fit1$coefficients)-2))
cutoff <- 4/nrow(rawdata)
cutoff
plot(fit1, which=4, cook.levels=cutoff)
# Influence Plot
influencePlot(fit1, main="Influence Plot",
              sub="Circle size is proportial to Cook's Distance" )
# 2.(b)
model_step =step(fit1,direction = "both",test="F")
model_step$coefficients
extractAIC(model_step, scale)
model_step =step(fit1,direction = "forward",test="F")
extractAIC(model_step, scale)
model_step$coefficients
model_step =step(fit1,direction = "backward",test="F")
extractAIC(model_step, scale)
model_step$coefficients
#2.(c)

fit_one = glm(Oxygen_Consumption~Performance,data=rawdata)
summary(fit_one)

fit_for = glm(Oxygen_Consumption~Gender + Runtime + Weight+Age
              + Run_Pulse + Maximum_Pulse+Performance,data=rawdata)
summary(fit_for)

fit_back = glm(Oxygen_Consumption~Gender + Runtime + Age
               + Run_Pulse + Maximum_Pulse,data=rawdata)
summary(fit_back)

fit_inter = glm(Oxygen_Consumption~Gender*Age+Gender + Runtime
                + Age + Run_Pulse + Maximum_Pulse,data=rawdata)
summary(fit_inter)

fit_back_sas = glm(Oxygen_Consumption~Gender + Run_Pulse
                   + Weight+Performance+Rest_Pulse,data=rawdata)
summary(fit_back_sas)

# without  influential observation
```

```
data=rawdata[-c(2,15,20), ]
attach(data)
fit11 = glm(Oxygen_Consumption~Gender+Runtime
            +Age+Weight+Run_Pulse+Rest_Pulse
            +Maximum_Pulse+Performance,data=data)
summary(fit11)
extractAIC(fit11, scale)
#
model_step =step(fit11,direction = "both",test="F")
model_step$coefficients
extractAIC(model_step, scale)
model_step =step(fit11,direction = "forward",test="F")
extractAIC(model_step, scale)
model_step$coefficients
model_step =step(fit11,direction = "backward",test="F")
extractAIC(model_step, scale)
model_step$coefficients
#
fit_one = glm(Oxygen_Consumption~Performance,data=data)
summary(fit_one)

fit_for = glm(Oxygen_Consumption~Gender + Runtime + Weight+Age
              + Run_Pulse + Maximum_Pulse+Performance,data=data)
summary(fit_for)

fit_back = glm(Oxygen_Consumption~Performance + Weight + Age
               + Run_Pulse + Maximum_Pulse,data=data)
summary(fit_back)

fit_inter = glm(Oxygen_Consumption~Age*Maximum_Pulse+Performance + Weight + Age
                + Run_Pulse + Maximum_Pulse,data=data)
summary(fit_inter)
```