SA3 Individual Assignment

Name – Vineet Kapoor - 11910076

1. Obtain summary statistics for the two whale calves, Hudson and Casey and make comparative statements.

As per the summary stats, skewness for lockons for Hudson is more than 1 not at daytime.Variance is greater than mean for both types of calves , whether its day or not.

> describe.by(df\_new,df$Daytime.Hudson)

Descriptive statistics by group

group: 0

vars n mean sd median trimmed mad min max range skew kurtosis se

Period.Hudson 1 111 111.50 64.38 112 111.51 83.03 1 221 220 0.00 -1.23 6.11

Bouts.Hudson 2 111 8.54 3.42 8 8.58 2.97 0 18 18 -0.07 0.12 0.32

Lockons.Hudson 3 111 15.79 9.72 14 14.83 7.41 0 49 49 1.08 1.48 0.92

Period.Casey 4 111 111.50 64.38 112 111.51 83.03 1 221 220 0.00 -1.23 6.11

Bouts.Casey 5 111 11.25 3.13 12 11.21 2.97 0 24 24 0.25 2.55 0.30

Lockons.Casey 6 111 28.48 10.72 28 27.93 11.86 0 60 60 0.44 0.22 1.02

---------------------------------------------------------------------------------------

group: 1

vars n mean sd median trimmed mad min max range skew kurtosis se

Period.Hudson 1 112 112.50 64.95 112.5 112.50 83.03 2 223 221 0.00 -1.23 6.14

Bouts.Hudson 2 112 8.81 3.58 9.0 8.79 3.71 0 17 17 0.02 -0.18 0.34

Lockons.Hudson 3 112 15.55 8.62 14.0 14.88 7.41 0 42 42 0.72 0.16 0.81

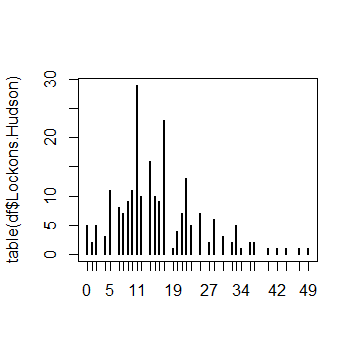
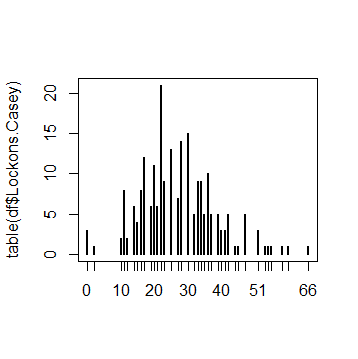
Period.Casey 4 112 112.50 64.95 112.5 112.50 83.03 2 223 221 0.00 -1.23 6.14

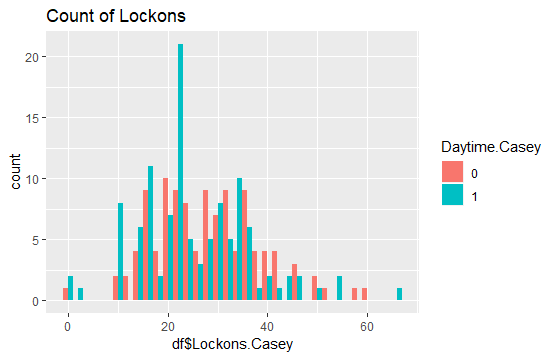
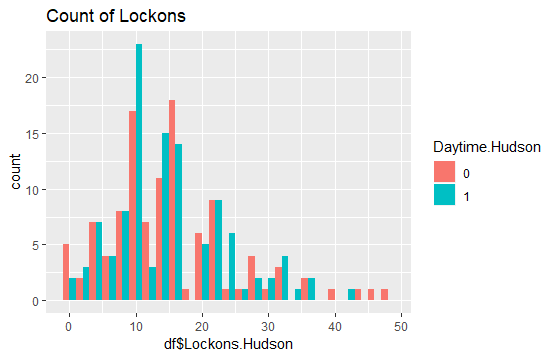
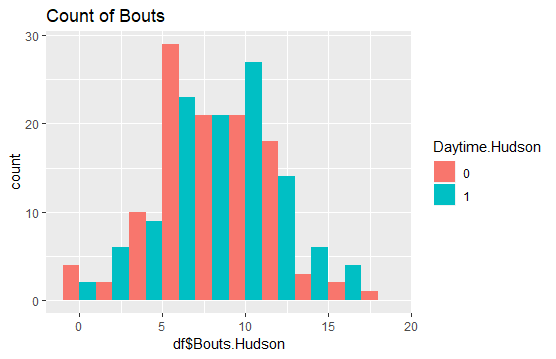
Bouts.Casey 5 112 10.42 3.35 11.0 10.49 2.97 0 18 18 -0.35 0.85 0.32

Lockons.Casey 6 112 25.79 11.30 23.0 25.20 10.38 0 66 66 0.59 0.93 1.07

2. Make comparative statements on the underlying probability distributions of the number of lockons of the two calves. With justification, propose appropriate regression model for number of lockons for each data set.

The Distribution is highly right-skewed for Hudson and Casey calves. For Hudson, it is more skewed. Poisson Regression model is used as the distribution follows poisson distribution.





3. For each data set, construct the regression model you have proposed in (2) above for the number of lockons in each period as a function of time, number of nursing bouts, and time of the day. Interpret your results.

> poisson <- glm(Lockons.Hudson ~ Period.Hudson+Daytime.Hudson + Bouts.Hudson, family = 'poisson',data=df)

> summary(poisson)

Call:

glm(formula = Lockons.Hudson ~ Period.Hudson + Daytime.Hudson +

Bouts.Hudson, family = "poisson", data = df)

Deviance Residuals:

Min 1Q Median 3Q Max

-4.0159 -1.2830 -0.2867 0.7933 5.1118

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.7311910 0.0598733 28.914 < 2e-16 \*\*\*

Period.Hudson -0.0009706 0.0002596 -3.739 0.000184 \*\*\*

Daytime.Hudson1 -0.0570726 0.0338902 -1.684 0.092173 .

Bouts.Hudson 0.1228921 0.0048903 25.130 < 2e-16 \*\*\*

---

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

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 1199.68 on 222 degrees of freedom

Residual deviance: 559.99 on 219 degrees of freedom

AIC: 1539.2

Number of Fisher Scoring iterations: 4

**Interpretation::**

1. in daytime, the Number of times the calf attached itself to the mother while suckling milk from a mammary gland decreases by 5.7% for hudson whales

2. For each unit increase in bouts, the Number of times the calf attached itself to the mother while suckling milk from a mammary gland increases by 12.2 % for Hudson whales.

3. For Hudson whales: the number of lockons decreases by less amount with increase in the time period of duration.

> anova(poisson)

Analysis of Deviance Table

Model: poisson, link: log

Response: Lockons.Hudson

Terms added sequentially (first to last)

Df Deviance Resid. Df Resid. Dev

NULL 222 1199.68

Period.Hudson 1 4.71 221 1194.97

Daytime.Hudson 1 0.19 220 1194.78

Bouts.Hudson 1 634.79 219 559.99

Interpret -> The residual deviance for Bouts in Hudson whales is very high.

> with(poisson, cbind(res.deviance = deviance, df = df.residual,

+ p = pchisq(deviance, df.residual, lower.tail=FALSE)))

res.deviance df p

[1,] 559.9859 219 9.707037e-32

Using Chi-square test, p-value is significant, and model does not fit properly.

> summary(poisson2)

Call:

glm(formula = Lockons.Casey ~ Period.Casey + Bouts.Casey + Daytime.Casey,

family = "poisson", data = df)

Deviance Residuals:

Min 1Q Median 3Q Max

-4.6903 -1.0117 -0.0307 0.9982 5.0066

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 2.3989934 0.0559184 42.902 < 2e-16 \*\*\*

Period.Casey -0.0011639 0.0002056 -5.661 1.5e-08 \*\*\*

Bouts.Casey 0.0919255 0.0038966 23.591 < 2e-16 \*\*\*

Daytime.Casey1 -0.0265118 0.0259256 -1.023 0.306

---

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

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 1061.86 on 222 degrees of freedom

Residual deviance: 461.15 on 219 degrees of freedom

AIC: 1586.4

Number of Fisher Scoring iterations: 4

**Interpretation::**

1. in daytime, the Number of times the calf attached itself to the mother while suckling milk from a mammary gland decreases by 2.6% for casey whales

2. For each unit increase in bouts, the Number of times the calf attached itself to the mother while suckling milk from a mammary gland increases by 9.1 % for Casey whales.

3. For Casey whales: the number of lockons decreases by 0.11% with increase in the time period of duration.

Residual deviance by ANOVA test shows that Bouts variable and Period variables are the main variables which cause high residual deviance.

> anova(poisson2)

Analysis of Deviance Table

Model: poisson, link: log

Response: Lockons.Casey

Terms added sequentially (first to last)

Df Deviance Resid. Df Resid. Dev

NULL 222 1061.86

Period.Casey 1 45.08 221 1016.79

Bouts.Casey 1 554.58 220 462.20

Daytime.Casey 1 1.05 219 461.15

P-value is significant and it shows that model does not fit the data properly.

with(poisson2, cbind(res.deviance = deviance, df = df.residual,

+ p = pchisq(deviance, df.residual, lower.tail=FALSE)))

res.deviance df p

[1,] 461.1542 219 2.29086e-19

4. Specifically answer following questions for **each data set**.

4.1. Do the variables provide predictive power? Justify.

Yes, The number of Bouts in each model provide predictive power but not other variables. We need to fit the data using another model.

4.2. How would you interpret the coefficient for time period for the model?

For casey whales: the number of lockons decreases by 0.11% with increase in the time period of duration.

For Hudson whales: the number of lockons decreases by less amount with increase in the time period of duration.

4.3. Does this model suffer from over dispersion? Justify with appropriate analysis.

If the selected model suffers from over dispersion, propose and fit an alternative

model which will take care of over dispersion.

Yes , the model suffer from over-dispersion as per the analysis, we can see that the alpha value is greater than 0. P-value is highly significant and I will use Negative Binomial Model

> dispersiontest(poisson, trafo=1)

Overdispersion test

data: poisson

z = 5.4106, p-value = 3.14e-08

alternative hypothesis: true alpha is greater than 0

sample estimates:

alpha

1.544171

> dispersiontest(poisson2, trafo=1)

Overdispersion test

data: poisson2

z = 4.4633, p-value = 4.035e-06

alternative hypothesis: true alpha is greater than 0

sample estimates:

alpha

0.9363754

> hudson\_NB <- glm.nb(Lockons.Hudson ~ Period.Hudson+Daytime.Hudson + Bouts.Hudson,data = df)

> summary(hudson\_NB)

Call:

glm.nb(formula = Lockons.Hudson ~ Period.Hudson + Daytime.Hudson +

Bouts.Hudson, data = df, init.theta = 10.93891714, link = log)

Deviance Residuals:

Min 1Q Median 3Q Max

-3.1530 -0.8462 -0.2131 0.4867 2.5725

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.5883514 0.0927358 17.128 <2e-16 \*\*\*

Period.Hudson -0.0007760 0.0004193 -1.851 0.0642 .

Daytime.Hudson1 -0.0531937 0.0541252 -0.983 0.3257

Bouts.Hudson 0.1355832 0.0079789 16.993 <2e-16 \*\*\*

---

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

(Dispersion parameter for Negative Binomial(10.9389) family taken to be 1)

Null deviance: 532.88 on 222 degrees of freedom

Residual deviance: 250.33 on 219 degrees of freedom

AIC: 1419.1

Number of Fisher Scoring iterations: 1

Theta: 10.94

Std. Err.: 1.85

2 x log-likelihood: -1409.053

casey\_NB <- glm.nb(Lockons.Casey ~ Period.Casey+Bouts.Casey+Daytime.Casey, data=df)

> summary(casey\_NB)

Call:

glm.nb(formula = Lockons.Casey ~ Period.Casey + Bouts.Casey +

Daytime.Casey, data = df, init.theta = 29.16910176, link = log)

Deviance Residuals:

Min 1Q Median 3Q Max

-4.1688 -0.7408 0.0216 0.7229 3.3550

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 2.3158006 0.0794822 29.136 < 2e-16 \*\*\*

Period.Casey -0.0010278 0.0002861 -3.593 0.000327 \*\*\*

Bouts.Casey 0.0980853 0.0057166 17.158 < 2e-16 \*\*\*

Daytime.Casey1 -0.0295272 0.0364269 -0.811 0.417601

---

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

(Dispersion parameter for Negative Binomial(29.1691) family taken to be 1)

Null deviance: 583.74 on 222 degrees of freedom

Residual deviance: 262.33 on 219 degrees of freedom

AIC: 1532.5

Number of Fisher Scoring iterations: 1

Theta: 29.17

Std. Err.: 6.11

2 x log-likelihood: -1522.462

4.4. Does this alternative model have different implications than the initial model you

selected in (3) above? Is it a better fit to the data? Justify your answer with proper

analysis.

YES, by applying the Negative Binomial model, we get the model which fitted the data very well as p- value is not very significant and residual deviance has also decreased.For both Case and Hudson whales, the Number of lockons can be predicted using this negative binomial model.

> with(hudson\_NB,cbind(res.deviance=deviance, df=df.residual,

+ p=pchisq(deviance, df.residual, lower.tail=FALSE)))

res.deviance df p

[1,] 250.3263 219 0.07181872

with(casey\_NB,cbind(res.deviance=deviance, df=df.residual,

+ p=pchisq(deviance, df.residual, lower.tail=FALSE)))

res.deviance df p

[1,] 262.3312 219 0.02390058

5. Using the two models for two calves you have finally selected, make comparative statements on how the predictors are affecting the number of lockons.

The predictors

1. Explore the data. What is the basic difference you are noticing between the two groups?

|  |
| --- |
| > summary(df)  survival.times status logWBC Rx sex\_1  Min. : 1.00 0:12 Min. :1.450 0:21 Min. :0.0000  1st Qu.: 6.00 1:30 1st Qu.:2.303 1:21 1st Qu.:0.0000  Median :10.50 Median :2.800 Median :0.0000  Mean :12.88 Mean :2.930 Mean :0.4762  3rd Qu.:18.50 3rd Qu.:3.490 3rd Qu.:1.0000  Max. :35.00 Max. :5.000 Max. :1.0000 |
|  |
| |  | | --- | |  | |

2. Compute Kaplan-Meier estimate of survival function and Nelson-Allen estimates of cumulative hazard rate.

> nelson\_aalen\_fit <- basehaz(coxph(Surv(time,event)~group,data=leukemia))

> plot(nelson\_aalen\_fit)

> nelson\_aalen\_fit

hazard time

1 0.03667141 1

2 0.07655417 2

3 0.09788573 3

4 0.14380031 4

5 0.19486426 5

6 0.27950087 6

7 0.30884130 7

8 0.44157474 8

9 0.44157474 9

10 0.48257106 10

11 0.57213669 11

12 0.68661853 12

13 0.75681084 13

14 0.82932221 15

15 0.91555735 16

16 1.00531939 17

17 1.00531939 19

18 1.00531939 20

19 1.29716868 22

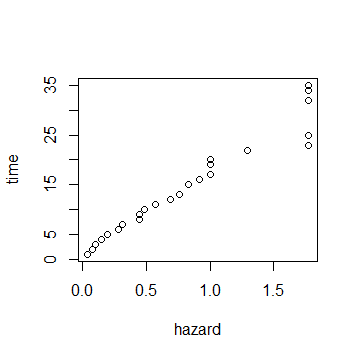
20 1.77758666 23

21 1.77758666 25

22 1.77758666 32

23 1.77758666 34

24 1.77758666 35



> # Nelson-Aalen non-parametric analysis

> nasurvival <- survfit(coxph(Surv(time,event)~X), type="aalen")

> summary(nasurvival)

Call: survfit(formula = coxph(Surv(time, event) ~ X), type = "aalen")

time n.risk n.event survival std.err lower 95% CI upper 95% CI

1 42 2 0.9852 0.0120 0.96196 1.000

2 40 2 0.9667 0.0205 0.92733 1.000

3 38 1 0.9531 0.0261 0.90325 1.000

4 37 2 0.9242 0.0359 0.85638 0.997

5 35 2 0.8913 0.0455 0.80647 0.985

6 33 3 0.8357 0.0593 0.72725 0.960

7 29 1 0.8132 0.0648 0.69567 0.951

8 28 4 0.6844 0.0839 0.53816 0.870

10 23 1 0.6429 0.0893 0.48974 0.844

11 21 2 0.5578 0.0969 0.39684 0.784

12 18 2 0.4595 0.1013 0.29819 0.708

13 16 1 0.4102 0.1013 0.25283 0.666

15 15 1 0.3608 0.0999 0.20974 0.621

16 14 1 0.3152 0.0967 0.17272 0.575

17 13 1 0.2549 0.0945 0.12327 0.527

22 9 2 0.1245 0.0793 0.03574 0.434

23 7 2 0.0376 0.0434 0.00391 0.362

3. Plot the KM estimate for the two groups with confidence intervals. Can you notice any major differences between two groups? Specify

> summary(kmsurvival1)

Call: survfit(formula = Surv(time, event) ~ group)

group=0

time n.risk n.event survival std.err lower 95% CI upper 95% CI

6 21 3 0.857 0.0764 0.720 1.000

7 17 1 0.807 0.0869 0.653 0.996

10 15 1 0.753 0.0963 0.586 0.968

13 12 1 0.690 0.1068 0.510 0.935

16 11 1 0.627 0.1141 0.439 0.896

22 7 1 0.538 0.1282 0.337 0.858

23 6 1 0.448 0.1346 0.249 0.807

group=1

time n.risk n.event survival std.err lower 95% CI upper 95% CI

1 21 2 0.9048 0.0641 0.78754 1.000

2 19 2 0.8095 0.0857 0.65785 0.996

3 17 1 0.7619 0.0929 0.59988 0.968

4 16 2 0.6667 0.1029 0.49268 0.902

5 14 2 0.5714 0.1080 0.39455 0.828

8 12 4 0.3810 0.1060 0.22085 0.657

11 8 2 0.2857 0.0986 0.14529 0.562

12 6 2 0.1905 0.0857 0.07887 0.460

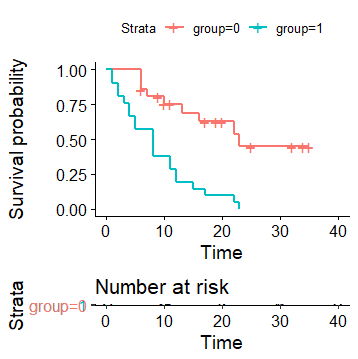
15 4 1 0.1429 0.0764 0.05011 0.407

17 3 1 0.0952 0.0641 0.02549 0.356

22 2 1 0.0476 0.0465 0.00703 0.322

23 1 1 0.0000 NaN NA NA

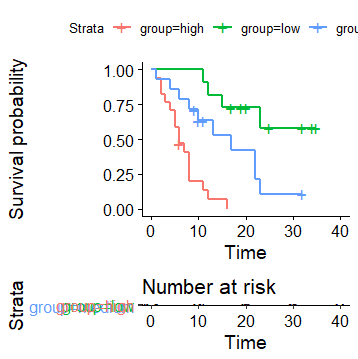
##Survival rate for the group which does not recieves treatment is more likely to survive as compared to other group, It can be seen that after 10 periods , the survival probability go down to 28.57% for the group which took treatment, as compared to the group which has 75.3% of survival rate.



4. Suppose we wish to compare KM estimates given the variable logWBC, for which we categorize logWBC into 3 classes---low, medium, high---as follows:

low(0-2.3), medium(2.31-3) & high(>3).

Compare the 3 KM plots you obtained. How are they different?



> summary(kmsurvival1)

Call: survfit(formula = Surv(time, event) ~ g)

g=high

time n.risk n.event survival std.err lower 95% CI upper 95% CI

1 17 1 0.9412 0.0571 0.8357 1.000

2 16 2 0.8235 0.0925 0.6609 1.000

3 14 1 0.7647 0.1029 0.5875 0.995

4 13 1 0.7059 0.1105 0.5194 0.959

5 12 2 0.5882 0.1194 0.3952 0.876

6 10 2 0.4706 0.1211 0.2842 0.779

7 7 1 0.4034 0.1210 0.2241 0.726

8 6 3 0.2017 0.1022 0.0747 0.544

11 3 1 0.1345 0.0875 0.0376 0.481

12 2 1 0.0672 0.0646 0.0102 0.442

16 1 1 0.0000 NaN NA NA

g=low

time n.risk n.event survival std.err lower 95% CI upper 95% CI

11 11 1 0.909 0.0867 0.754 1

12 10 1 0.818 0.1163 0.619 1

15 9 1 0.727 0.1343 0.506 1

23 5 1 0.582 0.1687 0.330 1

g=medium

time n.risk n.event survival std.err lower 95% CI upper 95% CI

1 14 1 0.929 0.0688 0.8030 1.000

4 13 1 0.857 0.0935 0.6921 1.000

6 12 1 0.786 0.1097 0.5977 1.000

8 11 1 0.714 0.1207 0.5129 0.995

10 9 1 0.635 0.1308 0.4240 0.951

13 6 1 0.529 0.1457 0.3085 0.908

17 5 1 0.423 0.1501 0.2112 0.848

22 4 2 0.212 0.1297 0.0636 0.704

23 2 1 0.106 0.0990 0.0169 0.662

The survival probability for 3 Logwbc groups Low has highest survival probability

Order of Survival probability is : low>medium>high

5. Fit the Cox PH model that can be used to assess the relationship of interest, which considers the potential confounders Sex and logWBC.

> # Cox proportional hazard model - coefficients and hazard rates

> coxph <- coxph(Surv(time,event) ~ X, method="breslow")

> summary(coxph)

Call:

coxph(formula = Surv(time, event) ~ X, method = "breslow")

n= 42, number of events= 30

coef exp(coef) se(coef) z Pr(>|z|)

XlogWBC 1.6089 4.9971 0.3003 5.358 8.42e-08 \*\*\*

Xsex\_1 -0.2343 0.7911 0.4081 -0.574 0.566

---

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

exp(coef) exp(-coef) lower .95 upper .95

XlogWBC 4.9971 0.2001 2.7741 9.001

Xsex\_1 0.7911 1.2640 0.3556 1.760

Concordance= 0.806 (se = 0.062 )

Rsquare= 0.548 (max possible= 0.989 )

Likelihood ratio test= 33.36 on 2 df, p=6e-08

Wald test = 28.71 on 2 df, p=6e-07

Score (logrank) test = 34.7 on 2 df, p=3e-08

Interpretation: The patients having sex as 1 will have die slowly and pateints having higher logwbc count , will die faster.

6. Compare the fits using KM and Cox PH plots for estimated survival function.

Thw model is fitted well more in the case of CoxPHmodel

**Students Weight Data:**

1. Carry out summary statistics and identify the missing values, if any.

> summary(df)

Age Weight Sex Height

Min. :18.00 Min. :100.0 Min. :0.0000 Min. :55.00

1st Qu.:19.00 1st Qu.:119.0 1st Qu.:0.0000 1st Qu.:64.00

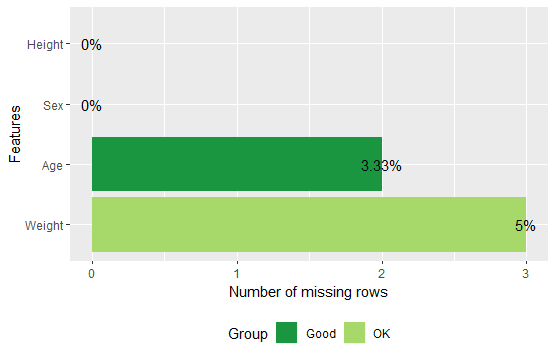
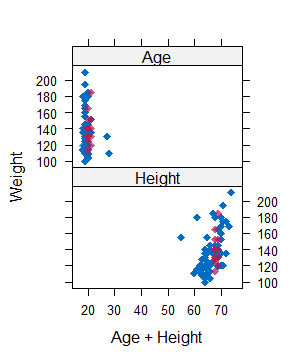
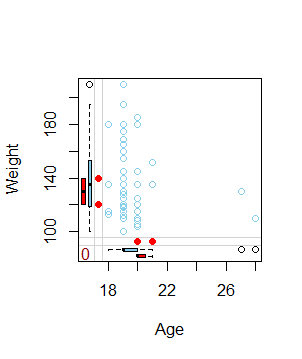
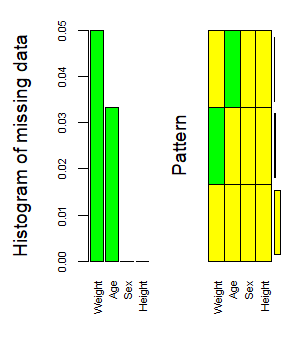
Median :19.00 Median :135.0 Median :1.0000 Median :67.00

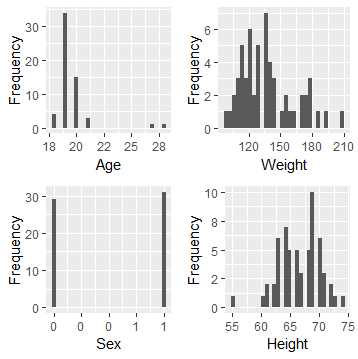
Mean :19.59 Mean :137.5 Mean :0.5167 Mean :66.62

3rd Qu.:20.00 3rd Qu.:152.0 3rd Qu.:1.0000 3rd Qu.:69.00

Max. :28.00 Max. :210.0 Max. :1.0000 Max. :74.00

NA's :2 NA's :3





2. Fit a linear regression model to the data with missing values. Interpret your results.

> summary(m3)

Call:

lm(formula = Weight ~ Age + Height + Sex\_1, data = df)

Residuals:

Min 1Q Median 3Q Max

-37.741 -11.038 -2.756 9.271 60.600

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 52.8982 58.1504 0.910 0.3673

Age -0.5416 1.6309 -0.332 0.7412

Height 1.6448 0.8093 2.032 0.0474 \*

Sex\_1 -27.8643 6.1462 -4.534 3.54e-05 \*\*\*

---

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

Residual standard error: 19.18 on 51 degrees of freedom

(5 observations deleted due to missingness)

Multiple R-squared: 0.4804, Adjusted R-squared: 0.4498

F-statistic: 15.72 on 3 and 51 DF, p-value: 2.297e-07

44.98% of variance is covered as by R-square value.

3. Impute the missing values by multiple imputation techniques using MICE. Give summary of the imputed data.

> imputed\_Data$imp$Age

1 2 3 4 5

14 20 19 21 18 19

26 19 19 19 20 19

> imputed\_Data$imp$Weight

1 2 3 4 5

33 152 145 175 152 140

49 145 180 152 175 180

54 120 118 118 135 115

> summary(imputed\_Data)

Class: mids

Number of multiple imputations: 5

Imputation methods:

Age Weight Height

"pmm" "pmm" ""

PredictorMatrix:

Age Weight Height

Age 0 1 1

Weight 1 0 1

Height 1 1 0

4. Fit liner regression models to the imputed data decks and get the combined result.

> summary(m1)

Call:

lm(formula = Weight ~ Age + Height + Sex\_1, data = completeData)

Residuals:

Min 1Q Median 3Q Max

-39.030 -11.352 -2.126 8.747 59.294

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 52.6872 57.7220 0.913 0.365273

Age -0.9552 1.6128 -0.592 0.556037

Height 1.7419 0.8064 2.160 0.035073 \*

Sex\_1 -24.5825 5.9142 -4.156 0.000112 \*\*\*

---

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

Residual standard error: 19.32 on 56 degrees of freedom

Multiple R-squared: 0.4385, Adjusted R-squared: 0.4085

F-statistic: 14.58 on 3 and 56 DF, p-value: 3.919e-07

summary(pool(modelFit1))

estimate std.error statistic df p.value

(Intercept) -36.910919 61.2527909 -0.6025998 54.14407 5.492694e-01

Age -2.824120 1.7555128 -1.6087153 54.60571 1.134456e-01

Height 3.448853 0.7940538 4.3433491 54.59309 6.137003e-05

5. Briefly compare the two models, one with missing data and second with imputed data.

