**QUESTION 1**

> BP<- read.csv(file.choose(),quote="",header=TRUE)

> attach(BP)

>

> model1<- lm(Systolic ~ Age + Weight)

>

> # part A

>

> summary(model1)

Call:

lm(formula = Systolic ~ Age + Weight)

Residuals:

Min 1Q Median 3Q Max

-3.4640 -1.1949 -0.4078 1.8511 2.6981

Coefficients:

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

(Intercept) 30.9941 11.9438 2.595 0.03186 \*

Age 0.8614 0.2482 3.470 0.00844 \*\*

Weight 0.3349 0.1307 2.563 0.03351 \*

---

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

Residual standard error: 2.318 on 8 degrees of freedom

Multiple R-squared: 0.9768, Adjusted R-squared: 0.9711

F-statistic: 168.8 on 2 and 8 DF, p-value: 2.874e-07

> # the p-value of Weight is 0.03351

> # the alpha value is 0.05

> #since the p-value is less than alpha, Weight is a statistically significant predictor od Systolic.

>

> #part B

>

> # R^2 = 0.9768 the value is high, which means it is good.

> # R^2 adj = 0.9711 the value is high, which means it is good.

> # s = 2.318

> # F p-value = 2.874e-07 (the value is low, which is good)

> # part C

>

> # For every 1 year of Age increased, the blood pressure will increase by 0.8614, if all the other variables are constant.

> # part D

> rstandard<-rstandard(model1)

> leverages<-hatvalues(model1)

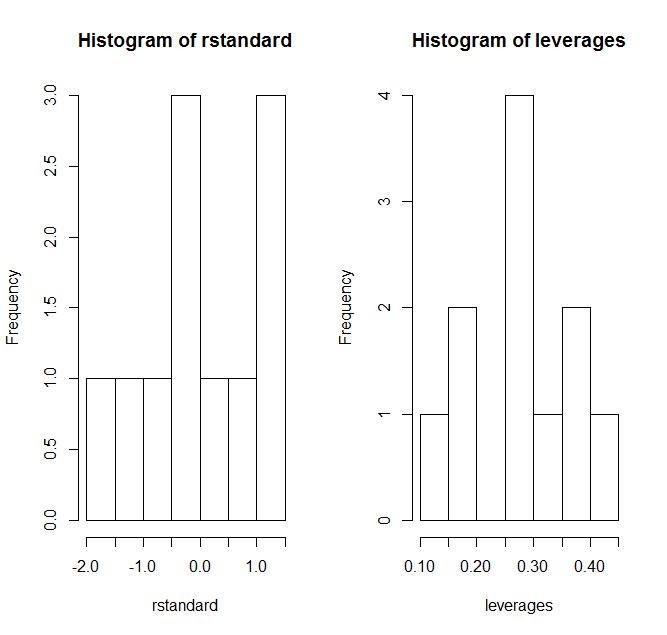
>

> par(mfrow=c(1,2))

> hist(rstandard)

> hist(leverages)

>

>

What about the cutoffs for outliers and the high leverage point values?

> # part E

>

> model2 <- lm(Systolic ~ Age + Weight + Age\*Weight)

> summary(model2)

Call:

lm(formula = Systolic ~ Age + Weight + Age \* Weight)

Residuals:

Min 1Q Median 3Q Max

-3.7346 -1.1805 -0.1372 1.9545 2.4635

Coefficients:

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

(Intercept) 11.348196 70.417505 0.161 0.877

Age 1.143904 1.030303 1.110 0.304

Weight 0.445377 0.413656 1.077 0.317

Age:Weight -0.001587 0.005597 -0.284 0.785

Residual standard error: 2.464 on 7 degrees of freedom

Multiple R-squared: 0.9771, Adjusted R-squared: 0.9673

F-statistic: 99.6 on 3 and 7 DF, p-value: 4.187e-06

> # H0 : B3 = 0, H1 : B3 not equal to zero.

> qt(0.025,7)

[1] -2.364624

>

> # since the t-statistic is less than t-critical, we accept H0 and reject H1. So, Age\*Weight is not a statistically significant predictor of Systolic.

What about the p-value?

>

> # part F

>

> new <- data.frame(Age=60, Weight=195)

> predict(model1, new, interval="prediction", level=0.95)

fit lwr upr

1 147.9765 142.2189 153.7341

>

> # prediction of Blood Pressure level at 95% confidence is 147.98.

> # we are 95% confident that blood level will fall between 142.22 and 153.73.

> # part G

> We are 95% confident that the average blood pressure for a person with the given characteristics will fall between 142.2189 and 153.7341.

**QUESTION 2**

> ball<-read.csv(file.choose(),quote="",header=TRUE)

> attach(ball)

>

> # QUestion 2

>

> # part A

>

> modelA<-lm(Wins~ERA+Runs+Hits+Walks+Saves+Errors)

> summary(modelA)

Call:

lm(formula = Wins ~ ERA + Runs + Hits + Walks + Saves + Errors)

Residuals:

Min 1Q Median 3Q Max

-10.1203 -1.7654 0.0048 1.6691 7.1764

Coefficients:

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

(Intercept) 55.048654 23.612328 2.331 0.028858 \*

ERA -10.918838 1.583130 -6.897 4.95e-07 \*\*\*

Runs 0.072993 0.017892 4.080 0.000461 \*\*\*

Hits 0.012268 0.021554 0.569 0.574756

Walks -0.008788 0.013568 -0.648 0.523595

Saves 0.290179 0.126220 2.299 0.030926 \*

Errors -0.061607 0.079268 -0.777 0.444962

---

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

Residual standard error: 3.758 on 23 degrees of freedom

Multiple R-squared: 0.9046, Adjusted R-squared: 0.8797

F-statistic: 36.34 on 6 and 23 DF, p-value: 1.295e-10

> # The R^2 value is 0.9045,which is high (very good)

> # The R^2 adjusted value is 0.8797 which is high ( very good)

> # Low f p-value of 1.295e-10 is very small (very good)

>

> # part B

>

> error1<-1\*(Errors>99)

> check=data.frame(Errors,error1)

> fix(check)

>

>

> modelB<-lm(Wins~ERA+Runs+Hits+Walks+Saves+Errors+error1)

> summary(modelB)

Call:

lm(formula = Wins ~ ERA + Runs + Hits + Walks + Saves + Errors +

error1)

Residuals:

Min 1Q Median 3Q Max

-10.3898 -1.7552 0.0128 1.7824 7.0117

Coefficients:

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

(Intercept) 51.384470 25.731764 1.997 0.058352 .

ERA -11.203406 1.762000 -6.358 2.13e-06 \*\*\*

Runs 0.071857 0.018446 3.895 0.000778 \*\*\*

Hits 0.014342 0.022559 0.636 0.531488

Walks -0.010745 0.014659 -0.733 0.471287

Saves 0.266469 0.141526 1.883 0.073015 .

Errors -0.007795 0.156600 -0.050 0.960751

error1 -1.287888 3.211183 -0.401 0.692242

---

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

Residual standard error: 3.828 on 22 degrees of freedom

Multiple R-squared: 0.9053, Adjusted R-squared: 0.8751

F-statistic: 30.03 on 7 and 22 DF, p-value: 7.646e-10

> # The binary variable for Errors has p value = 0.692242. which is larger than alpha = 0.05. So it is not statistically significant in predicting Wins.

>

What about any other model changes before and after?

> #part C

>

> # there will be 1.2878 decrease in Wins for every 1 unit error increase in a team that has 100 errors or more, given that other x variables held constant.

This should start: If a team has 100 errors or more, ….

> #part D

>

> cor(cbind(Wins,ERA,Runs,Hits,Walks,Saves,Errors)

+ cor(cbind(Wins,ERA,Runs,Hits,Walks,Saves,Errors))

Error: unexpected symbol in:

"cor(cbind(Wins,ERA,Runs,Hits,Walks,Saves,Errors)

cor"

> cor(cbind(Wins,ERA,Runs,Hits,Walks,Saves,Errors))

Wins ERA Runs Hits Walks Saves

Wins 1.0000000 -0.77026385 0.467280894 0.30648524 -0.6139753 0.60478114

ERA -0.7702639 1.00000000 0.075396226 0.14978668 0.4806223 -0.55326083

Runs 0.4672809 0.07539623 1.000000000 0.76428208 -0.3238141 0.03116950

Hits 0.3064852 0.14978668 0.764282082 1.00000000 -0.1234054 0.02744051

Walks -0.6139753 0.48062231 -0.323814111 -0.12340538 1.0000000 -0.38597120

Saves 0.6047811 -0.55326083 0.031169501 0.02744051 -0.3859712 1.00000000

Errors -0.5441757 0.59891965 -0.004142096 0.14817498 0.4070057 -0.41995888

Errors

Wins -0.544175691

ERA 0.598919654

Runs -0.004142096

Hits 0.148174981

Walks 0.407005718

Saves -0.419958878

Errors 1.000000000

>

>

We only want the correlations between the explanatory variables, so we are not interested in the Wins correlations.

>

>

> install.packages("car")

--- Please select a CRAN mirror for use in this session ---

trying URL 'http://watson.nci.nih.gov/cran\_mirror/bin/macosx/contrib/3.1/car\_2.0-20.tgz'

Content type 'application/octet-stream' length 1327641 bytes (1.3 Mb)

opened URL

==================================================

downloaded 1.3 Mb

The downloaded binary packages are in

/var/folders/2z/843dcpq55wl\_xy1f1m7fxy4c0000gn/T//RtmpBit6se/downloaded\_packages

> library(car)

> vif(modelA)

ERA Runs Hits Walks Saves Errors

2.119078 2.894361 2.644476 1.665735 1.533633 1.681815

>

> # ERA and Wins has the highest negative correlation with -0.770

> # Runs and Hits has the highest positive correlation (0.7643)

> # Collinearity does not seem t be an issue for this model. Why do you think that?

What about the variance inflation factors? What is the cutoff for a problem?

> # part E

>

> model1<-lm(Wins~ERA+Runs+Hits+Walks+Saves+Errors)

> summary(model1)

Call:

lm(formula = Wins ~ ERA + Runs + Hits + Walks + Saves + Errors)

Residuals:

Min 1Q Median 3Q Max

-10.1203 -1.7654 0.0048 1.6691 7.1764

Coefficients:

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

(Intercept) 55.048654 23.612328 2.331 0.028858 \*

ERA -10.918838 1.583130 -6.897 4.95e-07 \*\*\*

Runs 0.072993 0.017892 4.080 0.000461 \*\*\*

Hits 0.012268 0.021554 0.569 0.574756

Walks -0.008788 0.013568 -0.648 0.523595

Saves 0.290179 0.126220 2.299 0.030926 \*

Errors -0.061607 0.079268 -0.777 0.444962

---

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

Residual standard error: 3.758 on 23 degrees of freedom

Multiple R-squared: 0.9046, Adjusted R-squared: 0.8797

F-statistic: 36.34 on 6 and 23 DF, p-value: 1.295e-10

> model2<-lm(Wins~ERA+Walks+Saves+Errors)

> summary(model2)

Call:

lm(formula = Wins ~ ERA + Walks + Saves + Errors)

Residuals:

Min 1Q Median 3Q Max

-10.9007 -3.7948 0.4089 3.1629 14.1493

Coefficients:

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

(Intercept) 131.80034 20.02480 6.582 6.78e-07 \*\*\*

ERA -8.35460 2.61335 -3.197 0.00374 \*\*

Walks -0.04531 0.02086 -2.172 0.03956 \*

Saves 0.32132 0.21172 1.518 0.14165

Errors -0.04703 0.13219 -0.356 0.72502

---

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

Residual standard error: 6.377 on 25 degrees of freedom

Multiple R-squared: 0.7014, Adjusted R-squared: 0.6536

F-statistic: 14.68 on 4 and 25 DF, p-value: 2.686e-06

> anova(model1)

Analysis of Variance Table

Response: Wins

Df Sum Sq Mean Sq F value Pr(>F)

ERA 1 2019.62 2019.62 143.0092 2.368e-11 \*\*\*

Runs 1 944.87 944.87 66.9064 2.924e-08 \*\*\*

Hits 1 4.82 4.82 0.3412 0.56483

Walks 1 18.41 18.41 1.3034 0.26533

Saves 1 82.95 82.95 5.8734 0.02365 \*

Errors 1 8.53 8.53 0.6040 0.44496

Residuals 23 324.81 14.12

---

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

> anova(model2)

Analysis of Variance Table

Response: Wins

Df Sum Sq Mean Sq F value Pr(>F)

ERA 1 2019.62 2019.62 49.6668 2.185e-07 \*\*\*

Walks 1 263.04 263.04 6.4687 0.01754 \*

Saves 1 99.62 99.62 2.4498 0.13011

Errors 1 5.15 5.15 0.1266 0.72502

Residuals 25 1016.58 40.66

---

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

>

>

> anova(model1,model2)

Analysis of Variance Table

Model 1: Wins ~ ERA + Runs + Hits + Walks + Saves + Errors

Model 2: Wins ~ ERA + Walks + Saves + Errors

Res.Df RSS Df Sum of Sq F Pr(>F)

1 23 324.81

2 25 1016.58 -2 -691.77 24.492 2.003e-06 \*\*\*

---

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

>

> qf(0.95,2,23)

[1] 3.422132

> # H0 : B runs = B hits = 0

> # H1 : at least 1 B not equal to zero.

> # The t-test statistic=24.492 which is more than t-critical value=3.422132. So, we reject H0 and believe that at least one B not equal to 0. This means that there is a significant difference between the model with these two variables and the one without.

> # Therefore, we can’t use the reduced model that does not include these two variables.

> # Runs and Hits are jointly important explanatory variables for understanding Wins.

> # part F

> modelA<-lm(Wins~ERA+Runs+Hits+Walks+Saves+Errors)

> summary(modelA)

Call:

lm(formula = Wins ~ ERA + Runs + Hits + Walks + Saves + Errors)

Residuals:

Min 1Q Median 3Q Max

-10.1203 -1.7654 0.0048 1.6691 7.1764

Coefficients:

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

(Intercept) 55.048654 23.612328 2.331 0.028858 \*

ERA -10.918838 1.583130 -6.897 4.95e-07 \*\*\*

Runs 0.072993 0.017892 4.080 0.000461 \*\*\*

Hits 0.012268 0.021554 0.569 0.574756

Walks -0.008788 0.013568 -0.648 0.523595

Saves 0.290179 0.126220 2.299 0.030926 \*

Errors -0.061607 0.079268 -0.777 0.444962

---

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

Residual standard error: 3.758 on 23 degrees of freedom

Multiple R-squared: 0.9046, Adjusted R-squared: 0.8797

F-statistic: 36.34 on 6 and 23 DF, p-value: 1.295e-10

> rstandard<-rstandard(modelA)

> rstandard[order(rstandard)]

14 13 22 6 11 26 20 24 23 10

-2.99114287 -1.54112000 -1.26696190 -0.78531102 -0.76215599 -0.71786116 -0.62284825 -0.55618455 -0.51805619 -0.28772345

8 19 17 16 25 5 21 1 12 2

-0.16143275 -0.13808056 -0.10402458 -0.09612549 -0.05872370 0.06418547 0.07679928 0.08678275 0.16174255 0.20462875

18 15 29 30 9 27 7 3 28 4

0.24554323 0.34152082 0.56705901 0.65855249 0.88794394 0.97911343 1.03748531 1.11418758 1.91372254 2.05775136

> # Outliers exist when standard residual > |2|

> # Outliers = 14, 4

> leverages<-hatvalues(modelA)

> leverages[order(leverages)]

2 28 15 25 4 19 27 21 14 26 10 5

0.1055353 0.1174018 0.1242659 0.1343348 0.1387734 0.1421522 0.1428659 0.1662606 0.1893911 0.1911118 0.1912453 0.2057758

22 24 20 9 1 30 17 13 16 29 3 11

0.2084533 0.2128207 0.2199622 0.2283998 0.2314322 0.2470464 0.2485462 0.2520360 0.2633061 0.2664340 0.3003419 0.3025964

12 6 18 7 23 8

0.3171384 0.3199179 0.3375755 0.3624353 0.4075585 0.4248855

> # cutoff for identifying high leverage = 0.7

> newmodel<-lm(Wins~ERA+Runs+Hits+Walks+Saves+Errors,subset=-c(14,4))

> summary(newmodel)

Call:

lm(formula = Wins ~ ERA + Runs + Hits + Walks + Saves + Errors,

subset = -c(14, 4))

Residuals:

Min 1Q Median 3Q Max

-5.9993 -1.0867 -0.2449 0.9832 6.4462

Coefficients:

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

(Intercept) 49.48757 17.56916 2.817 0.0103 \*

ERA -10.04989 1.19776 -8.391 3.80e-08 \*\*\*

Runs 0.07355 0.01331 5.525 1.75e-05 \*\*\*

Hits 0.01296 0.01608 0.806 0.4295

Walks -0.01202 0.01012 -1.188 0.2481

Saves 0.36528 0.10536 3.467 0.0023 \*\*

Errors -0.07209 0.05887 -1.224 0.2343

---

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

Residual standard error: 2.789 on 21 degrees of freedom

Multiple R-squared: 0.9468, Adjusted R-squared: 0.9316

F-statistic: 62.26 on 6 and 21 DF, p-value: 2.739e-12

> # The R^2 value for new model is higher than Model A.

> # The R^2 adjusted for new model is higher than Model A.

> # The f p-value for new model is lower than Model A.

> # After removing the outlier from the model, the R^2 value has increased and this shows that the new model is the better model than model A.

What about removing insignificant terms?

I would have also liked to see you make plots to see if any variable needs a log transformation, squared term, etc.