SDS 358 Group5 RP#3

## I. Group Member

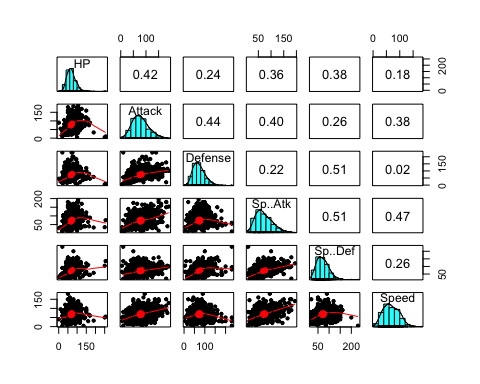
Xiaohan Wu (xw4822)  
Ningxin Liu (nl8385)  
Binglin Zhang (bz3856)

## II. Loading Data

We only keep the data of five predictors and the response we need.

## Correlation Matrix Plot

pairs.panels(data, method = "pearson")



## Interaction Analysis

# Calculate the interaction term:  
data <- data %>%  
 mutate(def.spdf = Defense\*Sp..Def)  
# Fit the regression model with 2 predictors and the interaction effect  
reg2 <- lm(HP~Attack+Defense+Sp..Atk+Sp..Def+Speed + def.spdf, data)  
# Display the summary table for the regression model  
summary(reg2)

##   
## Call:  
## lm(formula = HP ~ Attack + Defense + Sp..Atk + Sp..Def + Speed +   
## def.spdf, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -55.941 -12.305 -3.212 7.876 176.944   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.7982541 4.3522909 2.481 0.013306 \*   
## Attack 0.2764647 0.0294542 9.386 < 2e-16 \*\*\*  
## Defense 0.2187298 0.0585761 3.734 0.000202 \*\*\*  
## Sp..Atk 0.0817300 0.0304205 2.687 0.007368 \*\*   
## Sp..Def 0.6102262 0.0668753 9.125 < 2e-16 \*\*\*  
## Speed -0.1021945 0.0314174 -3.253 0.001191 \*\*   
## def.spdf -0.0038912 0.0006321 -6.156 1.19e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 21.31 on 793 degrees of freedom  
## Multiple R-squared: 0.3088, Adjusted R-squared: 0.3035   
## F-statistic: 59.04 on 6 and 793 DF, p-value: < 2.2e-16

# Calculate the interaction term:  
data <- data %>%  
 mutate(spat.spdf = Sp..Atk\*Sp..Def)  
# Fit the regression model with 2 predictors and the interaction effect  
reg3 <- lm(HP~Attack+Defense+Sp..Atk+Sp..Def+Speed + spat.spdf, data)  
# Display the summary table for the regression model  
summary(reg3)

##   
## Call:  
## lm(formula = HP ~ Attack + Defense + Sp..Atk + Sp..Def + Speed +   
## spat.spdf, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -61.837 -12.263 -2.847 8.237 189.587   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 28.7018963 4.6641158 6.154 1.20e-09 \*\*\*  
## Attack 0.2962777 0.0299643 9.888 < 2e-16 \*\*\*  
## Defense -0.0866873 0.0327698 -2.645 0.00832 \*\*   
## Sp..Atk 0.1570769 0.0697717 2.251 0.02464 \*   
## Sp..Def 0.3058496 0.0665541 4.596 5.02e-06 \*\*\*  
## Speed -0.0967130 0.0322776 -2.996 0.00282 \*\*   
## spat.spdf -0.0005749 0.0007663 -0.750 0.45333   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 21.81 on 793 degrees of freedom  
## Multiple R-squared: 0.2763, Adjusted R-squared: 0.2708   
## F-statistic: 50.45 on 6 and 793 DF, p-value: < 2.2e-16

# Calculate the interaction term:  
data <- data %>%  
 mutate(atk.df = Attack\*Defense)  
# Fit the regression model with 2 predictors and the interaction effect  
reg4 <- lm(HP~Attack+Defense+Sp..Atk+Sp..Def+Speed + atk.df, data)  
# Display the summary table for the regression model  
summary(reg4)

##   
## Call:  
## lm(formula = HP ~ Attack + Defense + Sp..Atk + Sp..Def + Speed +   
## atk.df, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -61.560 -12.584 -2.988 8.564 178.113   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 44.0424026 4.5324101 9.717 < 2e-16 \*\*\*  
## Attack 0.1128202 0.0598085 1.886 0.059612 .   
## Defense -0.2764196 0.0633693 -4.362 1.46e-05 \*\*\*  
## Sp..Atk 0.1110967 0.0305416 3.638 0.000293 \*\*\*  
## Sp..Def 0.2781959 0.0370529 7.508 1.62e-13 \*\*\*  
## Speed -0.0898969 0.0319077 -2.817 0.004962 \*\*   
## atk.df 0.0023514 0.0006659 3.531 0.000438 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 21.64 on 793 degrees of freedom  
## Multiple R-squared: 0.287, Adjusted R-squared: 0.2816   
## F-statistic: 53.19 on 6 and 793 DF, p-value: < 2.2e-16

# Calculate the interaction term:  
data <- data %>%  
 mutate(spat.speed = Sp..Atk\*Speed)  
# Fit the regression model with 2 predictors and the interaction effect  
reg5 <- lm(HP~Attack+Defense+Sp..Atk+Sp..Def+Speed + spat.speed, data)  
# Display the summary table for the regression model  
summary(reg5)

##   
## Call:  
## lm(formula = HP ~ Attack + Defense + Sp..Atk + Sp..Def + Speed +   
## spat.speed, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -59.497 -12.175 -2.768 7.901 191.643   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 25.5282930 4.6140764 5.533 4.29e-08 \*\*\*  
## Attack 0.2940241 0.0299486 9.818 < 2e-16 \*\*\*  
## Defense -0.0826107 0.0325360 -2.539 0.01131 \*   
## Sp..Atk 0.2058801 0.0662220 3.109 0.00194 \*\*   
## Sp..Def 0.2563528 0.0373995 6.854 1.44e-11 \*\*\*  
## Speed -0.0053298 0.0632473 -0.084 0.93286   
## spat.speed -0.0011913 0.0007296 -1.633 0.10290   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 21.78 on 793 degrees of freedom  
## Multiple R-squared: 0.2782, Adjusted R-squared: 0.2727   
## F-statistic: 50.93 on 6 and 793 DF, p-value: < 2.2e-16

The significant interactions are Attack\*Defense and Defense\*SpecialAttack.

## Model Selection

#stepwise regression forward selection  
data <- data %>%  
mutate(atk.df = Attack\*Defense, def.spdf = Defense\*Sp..Def)  
FitStart <-lm(HP ~1, data)  
FitAll <- lm(HP~Attack+Defense+Sp..Atk+Sp..Def+Speed + def.spdf +atk.df, data)  
step(FitStart,direction="forward", scope =formula(FitAll))

## Start: AIC=5185.06  
## HP ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + Attack 1 92945 428018 5029.8  
## + atk.df 1 81017 439947 5051.8  
## + Sp..Def 1 74720 446243 5063.2  
## + Sp..Atk 1 68412 452551 5074.4  
## + def.spdf 1 36915 484049 5128.3  
## + Defense 1 29913 491050 5139.8  
## + Speed 1 16129 504835 5161.9  
## <none> 520963 5185.1  
##   
## Step: AIC=5029.85  
## HP ~ Attack  
##   
## Df Sum of Sq RSS AIC  
## + Sp..Def 1 39985 388034 4953.4  
## + Sp..Atk 1 23493 404526 4986.7  
## + def.spdf 1 9023 418996 5014.8  
## + atk.df 1 3746 424272 5024.8  
## + Defense 1 1904 426114 5028.3  
## <none> 428018 5029.8  
## + Speed 1 136 427883 5031.6  
##   
## Step: AIC=4953.39  
## HP ~ Attack + Sp..Def  
##   
## Df Sum of Sq RSS AIC  
## + def.spdf 1 13639.9 374394 4926.8  
## + Sp..Atk 1 4934.8 383099 4945.1  
## + Defense 1 2844.7 385189 4949.5  
## <none> 388034 4953.4  
## + Speed 1 589.4 387444 4954.2  
## + atk.df 1 76.3 387957 4955.2  
##   
## Step: AIC=4926.76  
## HP ~ Attack + Sp..Def + def.spdf  
##   
## Df Sum of Sq RSS AIC  
## + atk.df 1 10923.4 363470 4905.1  
## + Defense 1 7997.2 366397 4911.5  
## + Speed 1 3904.9 370489 4920.4  
## + Sp..Atk 1 1843.9 372550 4924.8  
## <none> 374394 4926.8  
##   
## Step: AIC=4905.07  
## HP ~ Attack + Sp..Def + def.spdf + atk.df  
##   
## Df Sum of Sq RSS AIC  
## + Speed 1 2696.59 360774 4901.1  
## + Sp..Atk 1 1823.00 361647 4903.1  
## <none> 363470 4905.1  
## + Defense 1 364.44 363106 4906.3  
##   
## Step: AIC=4901.12  
## HP ~ Attack + Sp..Def + def.spdf + atk.df + Speed  
##   
## Df Sum of Sq RSS AIC  
## + Sp..Atk 1 3643.4 357130 4895.0  
## <none> 360774 4901.1  
## + Defense 1 332.2 360442 4902.4  
##   
## Step: AIC=4895  
## HP ~ Attack + Sp..Def + def.spdf + atk.df + Speed + Sp..Atk  
##   
## Df Sum of Sq RSS AIC  
## <none> 357130 4895.0  
## + Defense 1 183.74 356947 4896.6

##   
## Call:  
## lm(formula = HP ~ Attack + Sp..Def + def.spdf + atk.df + Speed +   
## Sp..Atk, data = data)  
##   
## Coefficients:  
## (Intercept) Attack Sp..Def def.spdf atk.df Speed   
## 24.893443 0.121892 0.577826 -0.003343 0.002059 -0.099120   
## Sp..Atk   
## 0.085953

#stepwise regression backward selection  
step(FitAll,direction="backward", scope =formula(FitStart))

## Start: AIC=4896.58  
## HP ~ Attack + Defense + Sp..Atk + Sp..Def + Speed + def.spdf +   
## atk.df  
##   
## Df Sum of Sq RSS AIC  
## - Defense 1 184 357130 4895.0  
## <none> 356947 4896.6  
## - Attack 1 2597 359544 4900.4  
## - atk.df 1 3157 360104 4901.6  
## - Sp..Atk 1 3495 360442 4902.4  
## - Speed 1 4440 361387 4904.5  
## - def.spdf 1 14523 371470 4926.5  
## - Sp..Def 1 35914 392861 4971.3  
##   
## Step: AIC=4895  
## HP ~ Attack + Sp..Atk + Sp..Def + Speed + def.spdf + atk.df  
##   
## Df Sum of Sq RSS AIC  
## <none> 357130 4895.0  
## - Attack 1 2638 359769 4898.9  
## - Sp..Atk 1 3643 360774 4901.1  
## - Speed 1 4517 361647 4903.1  
## - atk.df 1 9306 366436 4913.6  
## - def.spdf 1 23253 380383 4943.5  
## - Sp..Def 1 41709 398840 4981.4

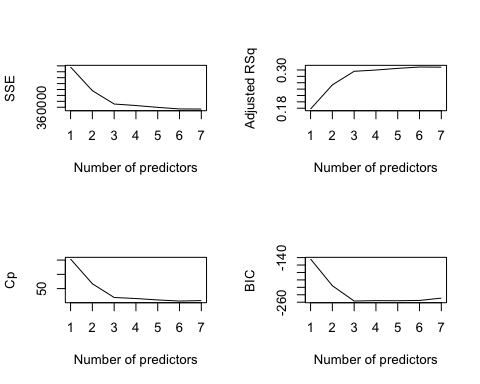
##   
## Call:  
## lm(formula = HP ~ Attack + Sp..Atk + Sp..Def + Speed + def.spdf +   
## atk.df, data = data)  
##   
## Coefficients:  
## (Intercept) Attack Sp..Atk Sp..Def Speed def.spdf   
## 24.893443 0.121892 0.085953 0.577826 -0.099120 -0.003343   
## atk.df   
## 0.002059

The results indicate that we should include all predictors expcept the Defense in the model. However, by the hierarchy principle, since we include the interaction term Defense\*Attack, we have to include the Defense in our model.

#best subset regression  
models <- regsubsets(HP~Attack+Defense+Sp..Atk+Sp..Def+Speed + def.spdf +atk.df, data, nvmax =7)  
summary(models)

## Subset selection object  
## Call: regsubsets.formula(HP ~ Attack + Defense + Sp..Atk + Sp..Def +   
## Speed + def.spdf + atk.df, data, nvmax = 7)  
## 7 Variables (and intercept)  
## Forced in Forced out  
## Attack FALSE FALSE  
## Defense FALSE FALSE  
## Sp..Atk FALSE FALSE  
## Sp..Def FALSE FALSE  
## Speed FALSE FALSE  
## def.spdf FALSE FALSE  
## atk.df FALSE FALSE  
## 1 subsets of each size up to 7  
## Selection Algorithm: exhaustive  
## Attack Defense Sp..Atk Sp..Def Speed def.spdf atk.df  
## 1 ( 1 ) "\*" " " " " " " " " " " " "   
## 2 ( 1 ) "\*" " " " " "\*" " " " " " "   
## 3 ( 1 ) " " " " " " "\*" " " "\*" "\*"   
## 4 ( 1 ) " " " " "\*" "\*" " " "\*" "\*"   
## 5 ( 1 ) " " " " "\*" "\*" "\*" "\*" "\*"   
## 6 ( 1 ) "\*" " " "\*" "\*" "\*" "\*" "\*"   
## 7 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"

models.sum <-summary(models)  
par(mfrow =c(2,2))  
# SSE  
plot(models.sum$rss, xlab ="Number of predictors", ylab ="SSE", type ="l")  
# R2  
plot(models.sum$adjr2, xlab ="Number of predictors", ylab ="Adjusted RSq", type ="l")  
# Mallow's Cp  
plot(models.sum$cp, xlab ="Number of predictors", ylab ="Cp", type ="l")  
# BIC  
plot(models.sum$bic, xlab ="Number of predictors", ylab ="BIC", type ="l")



The best subset selection method indicates that we shall use the model with seven predictors since the model has highest Adjusted and lowest SSE and , although the BIC of the model is a little bit higher.

## Diagnostic

reg <- lm(HP~ Attack+Defense+Sp..Atk+Sp..Def+Speed+def.spdf+atk.df, data)  
summary(reg)

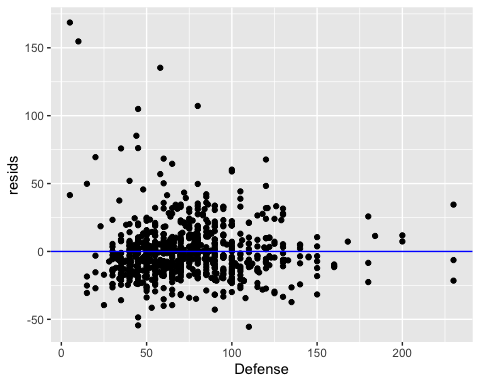
##   
## Call:  
## lm(formula = HP ~ Attack + Defense + Sp..Atk + Sp..Def + Speed +   
## def.spdf + atk.df, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -55.501 -12.449 -3.104 8.226 168.619   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 21.5865299 5.9508702 3.627 0.000305 \*\*\*  
## Attack 0.1413511 0.0588795 2.401 0.016594 \*   
## Defense 0.0544028 0.0852043 0.638 0.523334   
## Sp..Atk 0.0844415 0.0303233 2.785 0.005485 \*\*   
## Sp..Def 0.5965252 0.0668242 8.927 < 2e-16 \*\*\*  
## Speed -0.0983456 0.0313329 -3.139 0.001760 \*\*   
## def.spdf -0.0036215 0.0006380 -5.677 1.93e-08 \*\*\*  
## atk.df 0.0017514 0.0006617 2.647 0.008286 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 21.23 on 792 degrees of freedom  
## Multiple R-squared: 0.3148, Adjusted R-squared: 0.3088   
## F-statistic: 51.99 on 7 and 792 DF, p-value: < 2.2e-16

We decide to choose the model

### Residual Plot

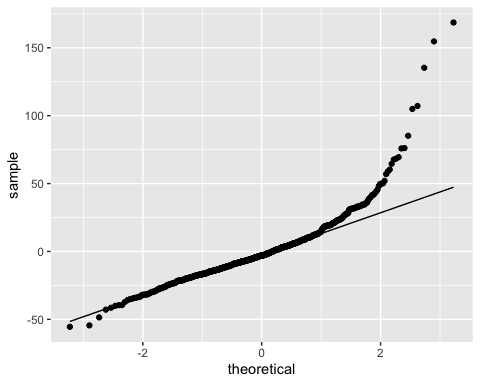
data$resids <- residuals(reg)  
data$predicted <- predict(reg)

ggplot(data, aes(x=Defense, y=resids)) +   
 geom\_point() + geom\_hline(color='blue', yintercept = 0)



The residual plot shows that the model does not violate the linearity condition. However the condition of equal variance is not met. Moreover, there are some outliers.

ggplot(data, aes(sample=resids)) +  
 stat\_qq() + stat\_qq\_line()



From the normal probability plot, we can see that the normal assumption of the normality of the errors seems to be approximately met except some outliers on the right tail. Therefore, we shall consider a log transformation on the response.

### Log Transformation on Response

data <- data %>%  
 mutate(lnHP = log(HP))

reg\_ln\_old <- lm(lnHP ~ Attack+Defense+Sp..Atk+Sp..Def+Speed+def.spdf+atk.df, data)  
summary(reg\_ln\_old)

##   
## Call:  
## lm(formula = lnHP ~ Attack + Defense + Sp..Atk + Sp..Def + Speed +   
## def.spdf + atk.df, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.9457 -0.1561 -0.0027 0.1522 1.3330   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.238e+00 8.827e-02 36.679 < 2e-16 \*\*\*  
## Attack 3.500e-03 8.734e-04 4.007 6.73e-05 \*\*\*  
## Defense 3.734e-03 1.264e-03 2.954 0.003229 \*\*   
## Sp..Atk 1.605e-03 4.498e-04 3.569 0.000379 \*\*\*  
## Sp..Def 9.049e-03 9.912e-04 9.129 < 2e-16 \*\*\*  
## Speed -1.081e-03 4.648e-04 -2.326 0.020294 \*   
## def.spdf -6.367e-05 9.463e-06 -6.729 3.29e-11 \*\*\*  
## atk.df 8.522e-06 9.815e-06 0.868 0.385497   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3149 on 792 degrees of freedom  
## Multiple R-squared: 0.348, Adjusted R-squared: 0.3423   
## F-statistic: 60.4 on 7 and 792 DF, p-value: < 2.2e-16

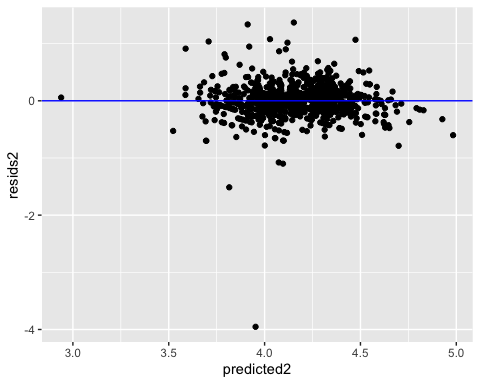
reg\_ln <- lm(lnHP ~ Attack+Defense+Sp..Atk+Sp..Def+Speed+def.spdf, data)  
summary(reg\_ln)

##   
## Call:  
## lm(formula = lnHP ~ Attack + Defense + Sp..Atk + Sp..Def + Speed +   
## def.spdf, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.9529 -0.1550 0.0003 0.1488 1.3690   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.185e+00 6.431e-02 49.532 < 2e-16 \*\*\*  
## Attack 4.157e-03 4.352e-04 9.552 < 2e-16 \*\*\*  
## Defense 4.533e-03 8.655e-04 5.238 2.08e-07 \*\*\*  
## Sp..Atk 1.592e-03 4.495e-04 3.543 0.000419 \*\*\*  
## Sp..Def 9.116e-03 9.881e-04 9.225 < 2e-16 \*\*\*  
## Speed -1.100e-03 4.642e-04 -2.369 0.018086 \*   
## def.spdf -6.499e-05 9.340e-06 -6.958 7.25e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3149 on 793 degrees of freedom  
## Multiple R-squared: 0.3474, Adjusted R-squared: 0.3425   
## F-statistic: 70.36 on 6 and 793 DF, p-value: < 2.2e-16

data$resids2 <- residuals(reg\_ln)  
data$predicted2 <- predict(reg\_ln)

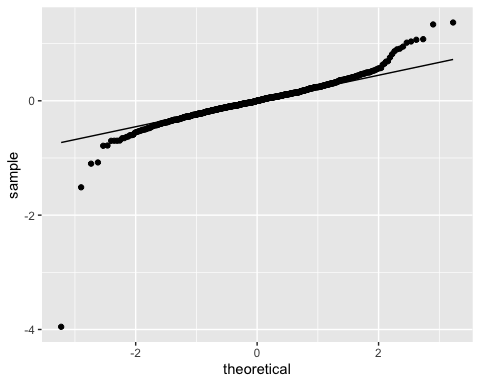
After the log transformation for y, the atk.df has p-value larger than 0.05, so we omitted it from the model and did diagnostic analysis.

ggplot(data, aes(x=predicted2, y=resids2)) +   
 geom\_point() + geom\_hline(color='blue', yintercept = 0)



The residual plot indicates that the conditions of constant variance for error terms and linearity are met, except some outliers.

ggplot(data, aes(sample=resids2)) +  
 stat\_qq() + stat\_qq\_line()

 The assumption of the normality of the errors seems to be approximately met and is better than the previous model.

## Reference

[1]Barradas, A. (2016, August 29). Pokemon with stats. Retrieved October 13, 2020, from <https://www.kaggle.com/abcsds/pokemon>