### Multiple Linear Regression Diagnostics

Packages

#install.packages("tidyverse","GGally","gridExtra","car")  
library(tidyverse)  
library(tidymodels)  
library(GGally)  
library(gridExtra) #used for a little fancy arranging of plots  
library(car) #for the VIF function  
library(glmnet)

Read-in data. For this work we will use a subset of the Lahman Baseball Database. The full database is available online here: <http://www.seanlahman.com/baseball-archive/statistics/>.

teams = read\_csv("Teams.csv")

## Rows: 2895 Columns: 48  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (13): lgID, teamID, franchID, divID, DivWin, WCWin, LgWin, WSWin, name, ...  
## dbl (35): yearID, Rank, G, Ghome, W, L, R, AB, H, 2B, 3B, HR, BB, SO, SB, CS...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

Examine the teams data frame.

str(teams)

## spc\_tbl\_ [2,895 × 48] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ yearID : num [1:2895] 1871 1871 1871 1871 1871 ...  
## $ lgID : chr [1:2895] NA NA NA NA ...  
## $ teamID : chr [1:2895] "BS1" "CH1" "CL1" "FW1" ...  
## $ franchID : chr [1:2895] "BNA" "CNA" "CFC" "KEK" ...  
## $ divID : chr [1:2895] NA NA NA NA ...  
## $ Rank : num [1:2895] 3 2 8 7 5 1 9 6 4 2 ...  
## $ G : num [1:2895] 31 28 29 19 33 28 25 29 32 58 ...  
## $ Ghome : num [1:2895] NA NA NA NA NA NA NA NA NA NA ...  
## $ W : num [1:2895] 20 19 10 7 16 21 4 13 15 35 ...  
## $ L : num [1:2895] 10 9 19 12 17 7 21 15 15 19 ...  
## $ DivWin : chr [1:2895] NA NA NA NA ...  
## $ WCWin : chr [1:2895] NA NA NA NA ...  
## $ LgWin : chr [1:2895] "N" "N" "N" "N" ...  
## $ WSWin : chr [1:2895] NA NA NA NA ...  
## $ R : num [1:2895] 401 302 249 137 302 376 231 351 310 617 ...  
## $ AB : num [1:2895] 1372 1196 1186 746 1404 ...  
## $ H : num [1:2895] 426 323 328 178 403 410 274 384 375 753 ...  
## $ 2B : num [1:2895] 70 52 35 19 43 66 44 51 54 106 ...  
## $ 3B : num [1:2895] 37 21 40 8 21 27 25 34 26 31 ...  
## $ HR : num [1:2895] 3 10 7 2 1 9 3 6 6 14 ...  
## $ BB : num [1:2895] 60 60 26 33 33 46 38 49 48 29 ...  
## $ SO : num [1:2895] 19 22 25 9 15 23 30 19 13 28 ...  
## $ SB : num [1:2895] 73 69 18 16 46 56 53 62 48 53 ...  
## $ CS : num [1:2895] 16 21 8 4 15 12 10 24 13 18 ...  
## $ HBP : num [1:2895] NA NA NA NA NA NA NA NA NA NA ...  
## $ SF : num [1:2895] NA NA NA NA NA NA NA NA NA NA ...  
## $ RA : num [1:2895] 303 241 341 243 313 266 287 362 303 434 ...  
## $ ER : num [1:2895] 109 77 116 97 121 137 108 153 137 166 ...  
## $ ERA : num [1:2895] 3.55 2.76 4.11 5.17 3.72 4.95 4.3 5.51 4.37 2.9 ...  
## $ CG : num [1:2895] 22 25 23 19 32 27 23 28 32 48 ...  
## $ SHO : num [1:2895] 1 0 0 1 1 0 1 0 0 1 ...  
## $ SV : num [1:2895] 3 1 0 0 0 0 0 0 0 1 ...  
## $ IPouts : num [1:2895] 828 753 762 507 879 ...  
## $ HA : num [1:2895] 367 308 346 261 373 329 315 431 371 573 ...  
## $ HRA : num [1:2895] 2 6 13 5 7 3 3 4 4 3 ...  
## $ BBA : num [1:2895] 42 28 53 21 42 53 34 75 45 63 ...  
## $ SOA : num [1:2895] 23 22 34 17 22 16 16 12 13 77 ...  
## $ E : num [1:2895] 243 229 234 163 235 194 220 198 218 432 ...  
## $ DP : num [1:2895] 24 16 15 8 14 13 14 22 20 22 ...  
## $ FP : num [1:2895] 0.834 0.829 0.818 0.803 0.84 0.845 0.821 0.845 0.85 0.83 ...  
## $ name : chr [1:2895] "Boston Red Stockings" "Chicago White Stockings" "Cleveland Forest Citys" "Fort Wayne Kekiongas" ...  
## $ park : chr [1:2895] "South End Grounds I" "Union Base-Ball Grounds" "National Association Grounds" "Hamilton Field" ...  
## $ attendance : num [1:2895] NA NA NA NA NA NA NA NA NA NA ...  
## $ BPF : num [1:2895] 103 104 96 101 90 102 97 101 94 106 ...  
## $ PPF : num [1:2895] 98 102 100 107 88 98 99 100 98 102 ...  
## $ teamIDBR : chr [1:2895] "BOS" "CHI" "CLE" "KEK" ...  
## $ teamIDlahman45: chr [1:2895] "BS1" "CH1" "CL1" "FW1" ...  
## $ teamIDretro : chr [1:2895] "BS1" "CH1" "CL1" "FW1" ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. yearID = col\_double(),  
## .. lgID = col\_character(),  
## .. teamID = col\_character(),  
## .. franchID = col\_character(),  
## .. divID = col\_character(),  
## .. Rank = col\_double(),  
## .. G = col\_double(),  
## .. Ghome = col\_double(),  
## .. W = col\_double(),  
## .. L = col\_double(),  
## .. DivWin = col\_character(),  
## .. WCWin = col\_character(),  
## .. LgWin = col\_character(),  
## .. WSWin = col\_character(),  
## .. R = col\_double(),  
## .. AB = col\_double(),  
## .. H = col\_double(),  
## .. `2B` = col\_double(),  
## .. `3B` = col\_double(),  
## .. HR = col\_double(),  
## .. BB = col\_double(),  
## .. SO = col\_double(),  
## .. SB = col\_double(),  
## .. CS = col\_double(),  
## .. HBP = col\_double(),  
## .. SF = col\_double(),  
## .. RA = col\_double(),  
## .. ER = col\_double(),  
## .. ERA = col\_double(),  
## .. CG = col\_double(),  
## .. SHO = col\_double(),  
## .. SV = col\_double(),  
## .. IPouts = col\_double(),  
## .. HA = col\_double(),  
## .. HRA = col\_double(),  
## .. BBA = col\_double(),  
## .. SOA = col\_double(),  
## .. E = col\_double(),  
## .. DP = col\_double(),  
## .. FP = col\_double(),  
## .. name = col\_character(),  
## .. park = col\_character(),  
## .. attendance = col\_double(),  
## .. BPF = col\_double(),  
## .. PPF = col\_double(),  
## .. teamIDBR = col\_character(),  
## .. teamIDlahman45 = col\_character(),  
## .. teamIDretro = col\_character()  
## .. )  
## - attr(\*, "problems")=<externalptr>

summary(teams)

## yearID lgID teamID franchID   
## Min. :1871 Length:2895 Length:2895 Length:2895   
## 1st Qu.:1921 Class :character Class :character Class :character   
## Median :1965 Mode :character Mode :character Mode :character   
## Mean :1957   
## 3rd Qu.:1994   
## Max. :2018   
##   
## divID Rank G Ghome   
## Length:2895 Min. : 1.000 Min. : 6.0 Min. :44.00   
## Class :character 1st Qu.: 2.000 1st Qu.:154.0 1st Qu.:77.00   
## Mode :character Median : 4.000 Median :158.0 Median :81.00   
## Mean : 4.073 Mean :150.7 Mean :78.56   
## 3rd Qu.: 6.000 3rd Qu.:162.0 3rd Qu.:81.00   
## Max. :13.000 Max. :165.0 Max. :84.00   
## NA's :399   
## W L DivWin WCWin   
## Min. : 0.00 Min. : 4.00 Length:2895 Length:2895   
## 1st Qu.: 66.00 1st Qu.: 66.00 Class :character Class :character   
## Median : 77.00 Median : 76.00 Mode :character Mode :character   
## Mean : 74.94 Mean : 74.94   
## 3rd Qu.: 87.00 3rd Qu.: 87.00   
## Max. :116.00 Max. :134.00   
##   
## LgWin WSWin R AB   
## Length:2895 Length:2895 Min. : 24.0 Min. : 211   
## Class :character Class :character 1st Qu.: 615.0 1st Qu.:5142   
## Mode :character Mode :character Median : 691.0 Median :5404   
## Mean : 683.6 Mean :5155   
## 3rd Qu.: 764.0 3rd Qu.:5520   
## Max. :1220.0 Max. :5781   
##   
## H 2B 3B HR BB   
## Min. : 33 Min. : 1.0 Min. : 0.00 Min. : 0 Min. : 1.0   
## 1st Qu.:1303 1st Qu.:195.0 1st Qu.: 30.00 1st Qu.: 43 1st Qu.:427.0   
## Median :1393 Median :233.0 Median : 41.00 Median :109 Median :494.0   
## Mean :1348 Mean :229.1 Mean : 46.51 Mean :104 Mean :475.2   
## 3rd Qu.:1466 3rd Qu.:272.0 3rd Qu.: 59.00 3rd Qu.:153 3rd Qu.:554.0   
## Max. :1783 Max. :376.0 Max. :150.00 Max. :267 Max. :835.0   
## NA's :1   
## SO SB CS HBP   
## Min. : 3.0 Min. : 1 Min. : 3.00 Min. : 7.00   
## 1st Qu.: 513.5 1st Qu.: 64 1st Qu.: 34.00 1st Qu.: 32.00   
## Median : 755.0 Median : 94 Median : 45.00 Median : 43.00   
## Mean : 751.0 Mean :111 Mean : 47.69 Mean : 45.36   
## 3rd Qu.: 980.5 3rd Qu.:139 3rd Qu.: 57.00 3rd Qu.: 56.00   
## Max. :1594.0 Max. :581 Max. :191.00 Max. :160.00   
## NA's :16 NA's :126 NA's :832 NA's :1158   
## SF RA ER ERA   
## Min. :18.00 Min. : 34.0 Min. : 23.0 Min. :1.220   
## 1st Qu.:39.00 1st Qu.: 611.0 1st Qu.: 504.0 1st Qu.:3.360   
## Median :45.00 Median : 689.0 Median : 594.0 Median :3.830   
## Mean :45.05 Mean : 683.6 Mean : 574.1 Mean :3.823   
## 3rd Qu.:51.00 3rd Qu.: 765.5 3rd Qu.: 669.0 3rd Qu.:4.310   
## Max. :77.00 Max. :1252.0 Max. :1023.0 Max. :8.000   
## NA's :1541   
## CG SHO SV IPouts   
## Min. : 0.00 Min. : 0.000 Min. : 0.00 Min. : 162   
## 1st Qu.: 11.00 1st Qu.: 6.000 1st Qu.: 9.00 1st Qu.:4080   
## Median : 43.00 Median : 9.000 Median :25.00 Median :4251   
## Mean : 48.98 Mean : 9.686 Mean :24.22 Mean :4033   
## 3rd Qu.: 77.00 3rd Qu.:13.000 3rd Qu.:38.00 3rd Qu.:4342   
## Max. :148.00 Max. :32.000 Max. :68.00 Max. :4518   
##   
## HA HRA BBA SOA E   
## Min. : 49 Min. : 0 Min. : 1.0 Min. : 0.0 Min. : 47.0   
## 1st Qu.:1290 1st Qu.: 49 1st Qu.:430.0 1st Qu.: 506.5 1st Qu.:114.0   
## Median :1392 Median :112 Median :496.0 Median : 754.0 Median :143.0   
## Mean :1348 Mean :104 Mean :475.4 Mean : 750.5 Mean :184.1   
## 3rd Qu.:1470 3rd Qu.:152 3rd Qu.:554.0 3rd Qu.: 985.5 3rd Qu.:212.0   
## Max. :1993 Max. :258 Max. :827.0 Max. :1687.0 Max. :639.0   
##   
## DP FP name park   
## Min. : 0.0 Min. :0.7610 Length:2895 Length:2895   
## 1st Qu.:117.0 1st Qu.:0.9650 Class :character Class :character   
## Median :141.0 Median :0.9770 Mode :character Mode :character   
## Mean :133.5 Mean :0.9658   
## 3rd Qu.:157.0 3rd Qu.:0.9810   
## Max. :217.0 Max. :0.9910   
##   
## attendance BPF PPF teamIDBR   
## Min. : 6088 Min. : 60.0 Min. : 60.0 Length:2895   
## 1st Qu.: 544782 1st Qu.: 97.0 1st Qu.: 97.0 Class :character   
## Median :1185781 Median :100.0 Median :100.0 Mode :character   
## Mean :1380458 Mean :100.2 Mean :100.2   
## 3rd Qu.:2065338 3rd Qu.:103.0 3rd Qu.:103.0   
## Max. :4483350 Max. :129.0 Max. :141.0   
## NA's :279   
## teamIDlahman45 teamIDretro   
## Length:2895 Length:2895   
## Class :character Class :character   
## Mode :character Mode :character   
##   
##   
##   
##

Let’s restrict our analysis to more recent years (from 1969 to 2018).

teams = teams %>% filter(yearID >= 1969)  
summary(teams$yearID)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1969 1982 1995 1995 2007 2018

Let’s work toward building a multiple linear regression model to predict the number of games that a team will win each year (recorded in the “W” column in the data). To expedite things a bit, let’s choose a few simple variables.

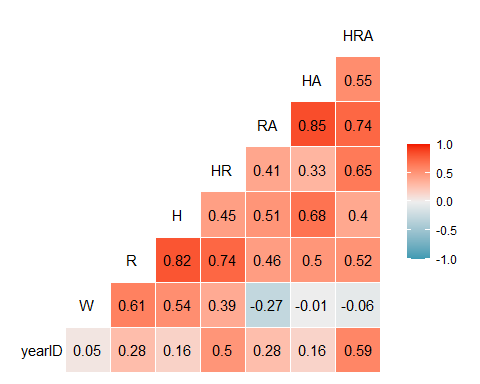
teams = teams %>% select(c("yearID","teamID","W","R","H","HR","RA","HA","HRA"))  
#These variables are W = Wins, R = Runs, H = Hits, HR = Home Runs, RA = Runs Against, HA = Hits Against, HRA = Home Runs Against  
summary(teams)

## yearID teamID W R   
## Min. :1969 Length:1378 Min. : 37.00 Min. : 329.0   
## 1st Qu.:1982 Class :character 1st Qu.: 71.00 1st Qu.: 650.0   
## Median :1995 Mode :character Median : 80.00 Median : 710.0   
## Mean :1995 Mean : 79.68 Mean : 709.9   
## 3rd Qu.:2007 3rd Qu.: 89.00 3rd Qu.: 772.0   
## Max. :2018 Max. :116.00 Max. :1009.0   
## H HR RA HA HRA   
## Min. : 797 Min. : 32.0 Min. : 331.0 Min. : 827 Min. : 40.0   
## 1st Qu.:1361 1st Qu.:119.0 1st Qu.: 648.0 1st Qu.:1357 1st Qu.:123.0   
## Median :1422 Median :147.0 Median : 707.5 Median :1425 Median :148.5   
## Mean :1414 Mean :148.2 Mean : 709.9 Mean :1414 Mean :148.2   
## 3rd Qu.:1487 3rd Qu.:176.0 3rd Qu.: 774.0 3rd Qu.:1490 3rd Qu.:173.0   
## Max. :1684 Max. :267.0 Max. :1103.0 Max. :1734 Max. :258.0

Examine the correlation between the quantitative variables.

ggcorr(teams,label = TRUE,label\_round = 2)

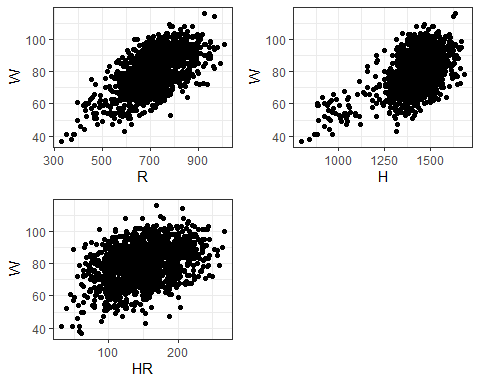
## Warning in ggcorr(teams, label = TRUE, label\_round = 2): data in column(s)  
## 'teamID' are not numeric and were ignored



The correlation matrix shows that many of the variables are correlated with each other. This is a pretty common occurrence in many datasets (especially sports-related datasets). The variable that is most strongly correlated with our response variable (W) is R (Runs) with a correlation of 0.61. The next most correlated variable with W is H (Hits) with a correlation of 0.54. Note that R and H are strongly correlated with each other (0.82). The next strongest variable (with respect to correlation with W) is HR (Home Runs) with a correlation of 0.39. Note that HR is also correlated with R and H.

To demonstrate multicollinearity, let’s build a model with R, H, and HR to predict W. Before we do so, visually examine the relationship between R, H, and HR and W.

p1 = ggplot(teams, aes(x=R,y=W)) + geom\_point() + theme\_bw()  
p2 = ggplot(teams, aes(x=H,y=W)) + geom\_point() + theme\_bw()  
p3 = ggplot(teams, aes(x=HR,y=W)) + geom\_point() + theme\_bw()  
grid.arrange(p1,p2,p3, ncol = 2) #arranging ggplot objects in a grid

 Each of these variables appears (as suggested by the correlation matrix and the plots) to have a positive relationship with W. We would expect each of these variables to contribute in a positive manner to the number of expected wins (i.e., more R, H, and HRs leads to more W).

Building the linear regression model.

recipe1 = recipe(W ~ R + H + HR, teams)  
  
lm\_model = #give the model type a name   
 linear\_reg() %>% #specify that we are doing linear regression  
 set\_engine("lm") #specify the specify type of linear tool we want to use   
  
lm\_wflow =   
 workflow() %>%   
 add\_model(lm\_model) %>%   
 add\_recipe(recipe1)  
  
lm\_fit = fit(lm\_wflow, teams)

summary(lm\_fit$fit$fit$fit)

##   
## Call:  
## stats::lm(formula = ..y ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -30.6354 -6.7215 0.5404 6.7770 25.8725   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 18.300666 3.250501 5.630 2.18e-08 \*\*\*  
## R 0.084840 0.006849 12.387 < 2e-16 \*\*\*  
## H 0.005023 0.004141 1.213 0.225349   
## HR -0.040160 0.010341 -3.884 0.000108 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.66 on 1374 degrees of freedom  
## Multiple R-squared: 0.3863, Adjusted R-squared: 0.385   
## F-statistic: 288.3 on 3 and 1374 DF, p-value: < 2.2e-16

Each of the variable coefficients (for R, H, and HR), as shown in the Estimate column, should be positive. However, the coefficient for HR is negative! This is a clear indication of multicollinearity. Be sure to “sanity check” each of your coefficients to make sure that their signs are oriented correctly.

We can also assess multicollinearity via a statistic known as the Variance Inflation Factor (VIF). We use the vif function from the car package.

car::vif(lm\_fit$fit$fit$fit) #Using the vif function from the the car package

## R H HR   
## 6.589245 3.717506 2.688779

In general, seeing variables with VIF values greater than 4 indicates the presence of multicollinearity. This is just a “rule of thumb” though and should not be taken as an absolute.

Ridge regression should help or we can drop the HR variable.

What happens if we drop the HR variable?

recipe2 = recipe(W ~ R + H, teams)  
  
lm\_model = #give the model type a name   
 linear\_reg() %>% #specify that we are doing linear regression  
 set\_engine("lm") #specify the specify type of linear tool we want to use   
  
lm\_wflow =   
 workflow() %>%   
 add\_model(lm\_model) %>%   
 add\_recipe(recipe2)  
  
lm\_fit2 = fit(lm\_wflow, teams)

summary(lm\_fit2$fit$fit$fit)

##   
## Call:  
## stats::lm(formula = ..y ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -31.453 -6.865 0.522 6.798 26.618   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16.733423 3.241829 5.162 2.81e-07 \*\*\*  
## R 0.065438 0.004709 13.895 < 2e-16 \*\*\*  
## H 0.011664 0.003791 3.077 0.00213 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.709 on 1375 degrees of freedom  
## Multiple R-squared: 0.3796, Adjusted R-squared: 0.3787   
## F-statistic: 420.7 on 2 and 1375 DF, p-value: < 2.2e-16

car::vif(lm\_fit2$fit$fit$fit) #Using the vif function from the the car package

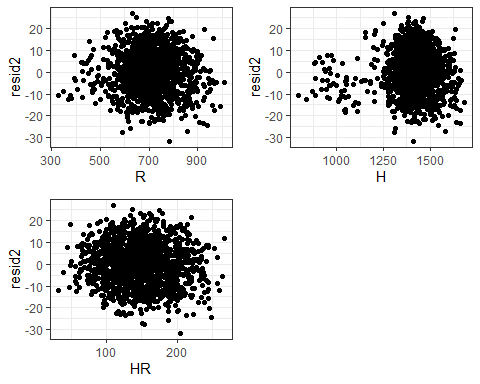
## R H   
## 3.083699 3.083699

As we did with simple linear regression, we also need to examine residuals (to assess the linear regression model assumptions related to residuals).

teams = teams %>% mutate(resid2 = lm\_fit2$fit$fit$fit$residuals)

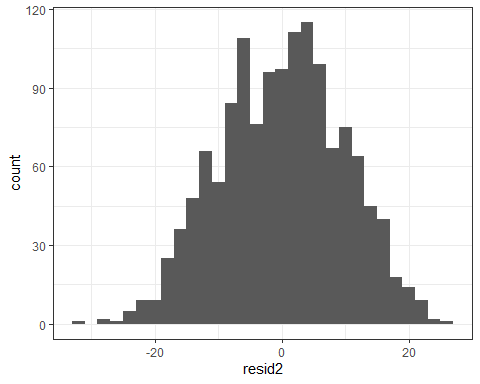
We create separate residual plots for each of the three predictor variables in our model.

p1 = ggplot(teams, aes(x=R,y=resid2)) + geom\_point() + theme\_bw()  
p2 = ggplot(teams, aes(x=H,y=resid2)) + geom\_point() + theme\_bw()  
p3 = ggplot(teams, aes(x=HR,y=resid2)) + geom\_point() + theme\_bw()  
grid.arrange(p1,p2,p3, ncol = 2) #arranging ggplot objects in a grid



ggplot(teams,aes(x=resid2)) + geom\_histogram() + theme\_bw()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

 The residual plots are not indicative of unequal variance or non-Normal residuals.