hw3.Rmd

Joshua Burkhart

November 17, 2015

Homework 3: Multiple Linear Regression

Overview

Export/import wrangled dataset from midterm

```
# for comma separated values
teams_hw3 <- read.table("~/SoftwareProjects/Probability/hw3/teams_midterm.csv",
                        header = TRUE, sep = ",", row.names = 1)

# wrapper function for .csv files
teams_hw3_2 <- read.csv("~/SoftwareProjects/Probability/hw3/teams_midterm.csv",
                        row.names = 1)

# for tab delimited files
teams_hw3_3 <- read.table("~/SoftwareProjects/Probability/hw3/teams_midterm3.tsv",
                          header = TRUE, sep = "\t", quote = "")

# wrapper function for tab delimited files
teams_hw3_4 <- read.delim("~/SoftwareProjects/Probability/hw3/teams_midterm3.tsv")
```

```
# install.packages(readr)
library(readr)
```

```
##
## Attaching package: 'readr'
##
## The following objects are masked from 'package:scales':
##
##     col_factor, col_numeric
```

```
# exporting datasets
?write_csv

# importing datasets
?read_csv
?read_tsv
```

HLO old friend

Confirm that all variables imported as you expected. Any surprises?

> It appears all four teams_hw3 tables are the same.

Simple linear regression (SLR)

```
# read in teams_bat from midterm, store as teams_hw3
teams_iid <- teams_hw3 %>%
  group_by(teamID) %>%
  summarise(avg_w = mean(W, na.rm = TRUE), # avg games won
            avg_payz = mean(z_pay, na.rm = TRUE), # avg payroll z-scores
            avg_onb = mean(ob_perc, na.rm = TRUE), # avg perc on base
            avg_runs = mean(R, na.rm = TRUE), # avg runs scored by team
            avg_rsc = mean(RA, na.rm = TRUE)) # avg runs scored by opponents
head(teams_iid)
```

```
## Source: local data frame [6 x 6]
##
##   teamID    avg_w    avg_payz   avg_onb avg_runs  avg_rsc
##   (fctr)    (dbl)       (dbl)     (dbl)    (dbl)    (dbl)
## 1    ARI 79.33333 -0.13301743 0.2546464 724.0000 745.0667
## 2    ATL 89.40000  0.36832453 0.2536220 747.5333 662.6667
## 3    BAL 73.80000 -0.12976614 0.2571155 732.8000 808.6000
## 4    BOS 89.06667  1.41556759 0.2695047 837.4000 740.4667
## 5    CHA 83.00000  0.08729364 0.2690874 769.1333 752.2667
## 6    CHN 76.86667  0.40296376 0.2447524 710.7333 731.4000
```

```
bball1 <- lm(avg_w ~ avg_payz, data = teams_iid)
bball1_vars <- bball1 %>%
  augment() %>% # broom::augment
  mutate(teamID = teams_iid$teamID, # add teamID as column
         .ext.resid = rstudent(bball1)) %>% # add externally studentized residuals
  dplyr::select(teamID, everything()) # just reorders the columns nicely
```

```
# library(MBESS)
# need 2 degrees of freedom bc we're not removing the intercept
ci.R2(glance(bball1)$r.squared, df.1 = 2, df.2 = 28, conf.level = .95)
```

```
## Loading required package: gsl
## Loading required package: MASS
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```
## $Lower.Conf.Limit.R2
## [1] 0.1667487
##
## $Prob.Less.Lower
## [1] 0.025
##
## $Upper.Conf.Limit.R2
## [1] 0.6857296
```

```
##
## $Prob.Greater.Upper
## [1] 0.025
```

What proportion of variability in mean wins does average payroll z-scores "explain"?

```
glance(bball1)$r.squared
```

```
## [1] 0.4773283
```

47.73%. The R-squared value is 0.4773, indicating the proportion of variance of our sample's mean wins is explained by our sample's average payroll z-scores is 47.73%.

What is the 95% confidence interval for multiple R2 in this simple linear regression model?

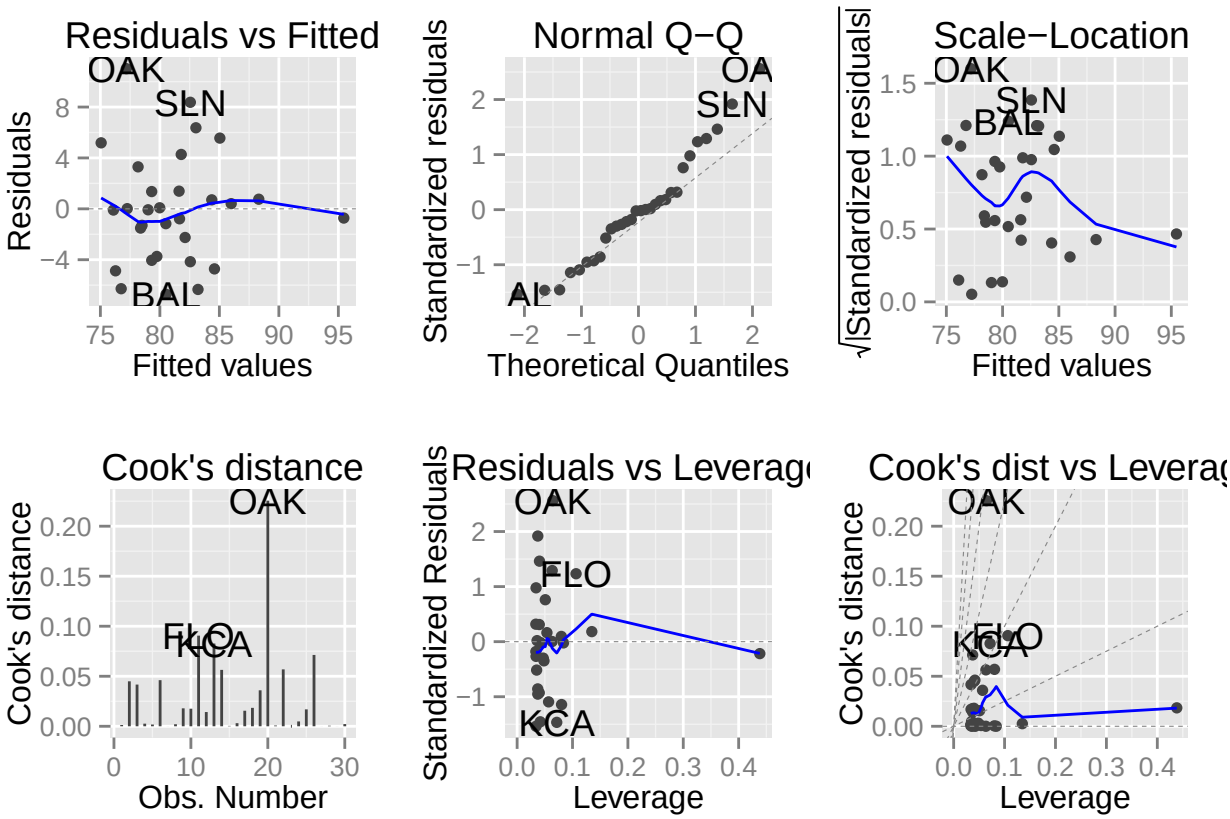
0.1667487 - 0.6857296

What does that mean in words (i.e., write a sentence!)?

There is a probability of .95 that our population's R-squared value is between 0.1667487 and 0.6857296

SLR diagnostics

```
# plot(bball1) # this leads you through a series of 4 plots interactively
# library(ggfortify)
autoplot(bball1, which = 1:6, ncol = 3, label.n = 3, data = teams_iid,
         label.label = "teamID")
```



Visually inspect and comment on the SLR residuals.

Residual for OAK appears furthest from fitted value. This looks right.

Are any externally studentized residuals significant at $p < .05$ according to the Bonferroni outlier test?

```
# library(car)
outlierTest(bball1, labels = teams_iid$teamID)
```

```
##
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferonni p
## OAK 2.873348          0.0078189      0.23457
```

No. The Bonferonni p reported is 0.235

Any teams with leverage values greater than 3 times the mean leverage value?

```
lev_meanx3 <- hatvalues(bball1) %>% mean() * 3
bball1_vars %>% filter(.hat > lev_meanx3) %>% dplyr::select(teamID)
```

```
##   teamID
## 1    NYA
```

Yes. NYA

Any teams with Cook's distance values greater than the "rule of thumb" we talked about in class?

```
ro_thumb <- 4/(nrow(bball1_vars) - 2 - 1)
bball1_vars %>% filter(.cooksd > ro_thumb) %>% dplyr::select(teamID)
```

```
##   teamID
## 1    OAK
```

Yes. OAK

Are any teams flagged as "influential" according to more than one of these indices (leverage, externally studentized residuals, and Cook's d)?

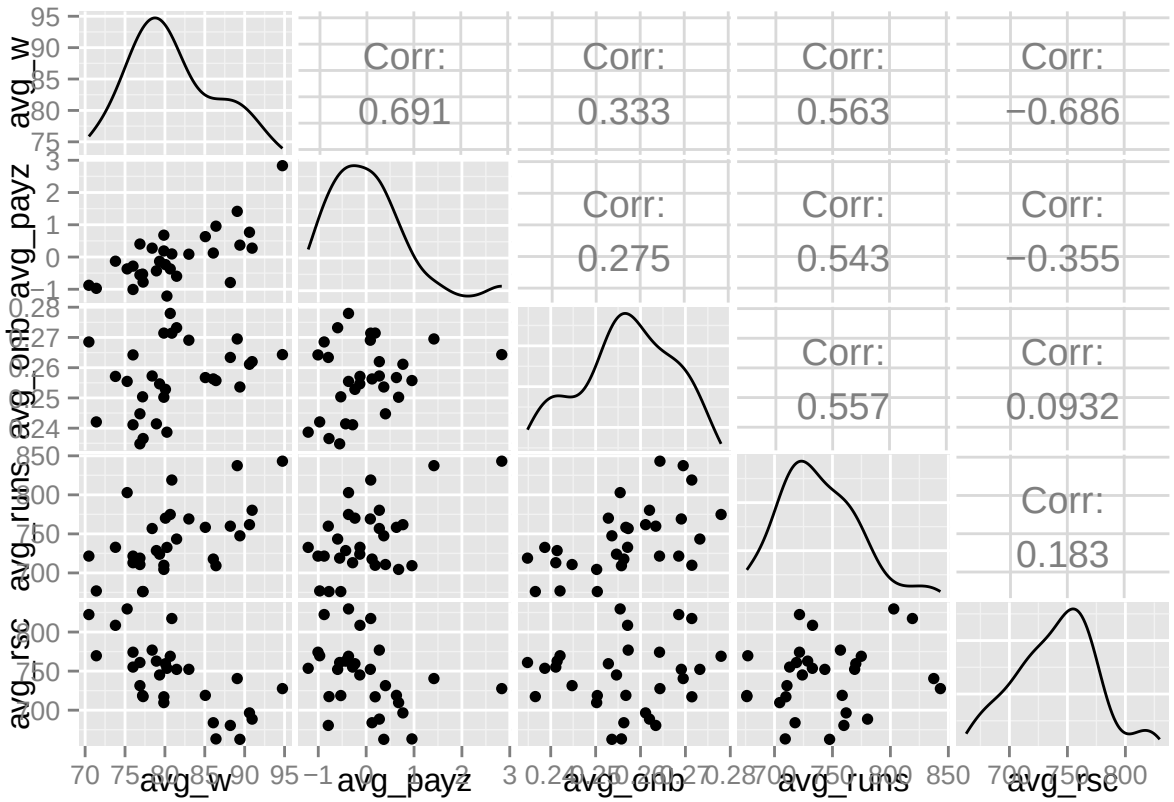
```
bball1_vars %>% filter(.ext.resid > 2) %>% dplyr::select(teamID)
```

```
##   teamID
## 1    OAK
## 2    SLN
```

Yes. OAK is influential according to both externally studentized residuals and Cook's d.

Multiple linear regression (MLR)

```
# library(GGally)
ggpairs(teams_iid, 2:6)
```



Discuss zero-order correlations among predictors (that is, all variables other than average number of wins).

> Predictors appear to be positively correlated with each other except for avg_rsc (average runs scored against team).

Are the direction/strength of the linear relationships as you might expect?

> avg_runs is positively correlated with avg_w (as expected). avg_rsc is negatively correlated with avg_w (as expected). avg_runs is positively correlated with avg_onb (as expected). This all looks right.

That is, do they make sense?

> Yes.

Run two linear regressions. The first model should be your model predicting average games won from average payroll z-scores. Your second model should add at least one other predictor (you can add them all if you want).

```
bball1 <- lm(avg_w ~ avg_payz, data = teams_iid)
bball2 <- update(bball1, . ~ . + avg_onb + avg_runs)
anova(bball1, bball2)
```

```
## Analysis of Variance Table
##
## Model 1: avg_w ~ avg_payz
## Model 2: avg_w ~ avg_payz + avg_onb + avg_runs
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     28 554.13
## 2     26 499.32  2    54.812 1.4271 0.2582
```

Explain why you chose those predictors (i.e., you have a hypothesis about something, or perhaps you had an idea based on the scatterplot matrix?)

> The predictors, avg_onb and avg_runs, were both positively correlated with the outcome, avg_w.

Compare the two nested models.

```
bball1 %>% summary()
```

```
##
## Call:
## lm(formula = avg_w ~ avg_payz, data = teams_iid)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7168 -3.3719 -0.0858  1.3803 11.0187
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  81.1713     0.8122  99.937  < 2e-16 ***
## avg_payz      5.0438     0.9974   5.057 2.37e-05 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.449 on 28 degrees of freedom
## Multiple R-squared:  0.4773, Adjusted R-squared:  0.4587
## F-statistic: 25.57 on 1 and 28 DF,  p-value: 2.375e-05
```

```
bball2 %>% summary()
```

```
##
## Call:
## lm(formula = avg_w ~ avg_payz + avg_onb + avg_runs, data = teams_iid)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7743 -2.6257 -0.6153  2.1768  9.4790
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 50.12126   20.33128   2.465  0.02061 *
## avg_payz     4.00168    1.17101   3.417  0.00209 **
## avg_onb     25.37667   83.76912   0.303  0.76435
## avg_runs     0.03301    0.02576   1.281  0.21146
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.382 on 26 degrees of freedom
## Multiple R-squared:  0.529,  Adjusted R-squared:  0.4747
## F-statistic: 9.735 on 3 and 26 DF,  p-value: 0.000176
```

Interpret!

> Comparing R-squared values, 47.73% of the variance in avg_w is predicted by avg_payz alone but 52.9% of the variance is predicted by avg_payz, avg_onb, and avg_runs together.

Does adding these additional variables to your model add any predictive value (that is, does it explain any more variability in average wins) above and beyond payroll alone?

> Yes. (see above)

Which model do you think is "better" and why- according to what criterion/criteria would you judge or compare two linear regression models?

> The second model, bball2, appears to have more predictive power based on the R-squared value alone. Both p-values are low ($< .05$) and both F-statistics are high ($>> 1$).

Did the coefficients for any of your predictor variables reverse sign from their zero-order Pearson Product moment correlations?

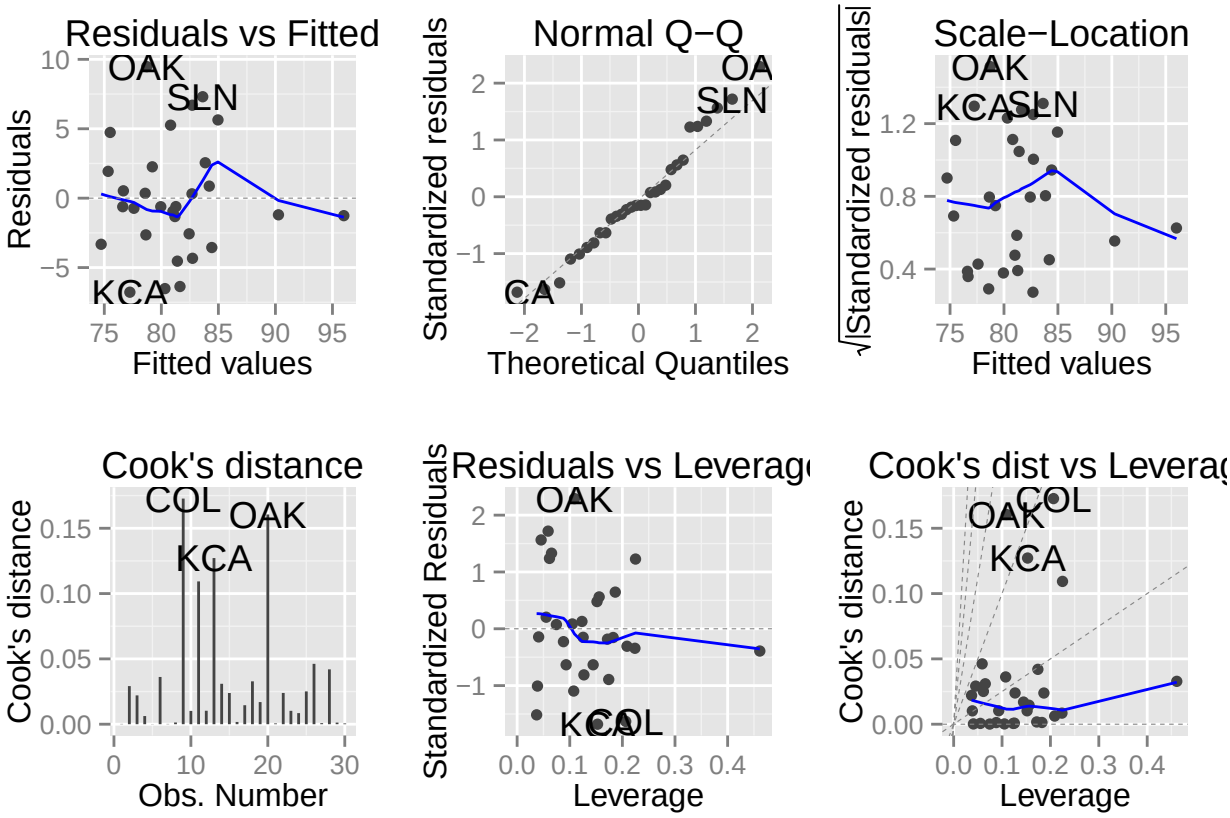
> No. All remained positive.

(for example, were any predictors positively correlated with wins initially but in the MLR the partial regression coefficient is negative rather than positive? Or vice versa?)

> No.

MLR diagnostics

```
# plot(bball2)
# library(ggfortify)
autoplot(bball2, which = 1:6, ncol = 3, label.n = 3, data = teams_iid,
         label.label = "teamID")
```



Do you identify any specific influential observations?

```
# copying from bball1_vars generation above
bball2_vars <- bball2 %>%
  augment() %>% # broom::augment
  mutate(teamID = teams_iid$teamID, # add teamID as column
         .ext.resid = rstudent(bball2)) %>% # add externally studentized residuals
  dplyr::select(teamID, everything()) # just reorders the columns nicely
```

```
lev_meanx3 <- hatvalues(bball2) %>% mean() * 3
bball2_vars %>% filter(.hat > lev_meanx3) %>% dplyr::select(teamID)
```

```
##   teamID
## 1    NYA
```

```
ro_thumb <- 4/(nrow(bball2_vars) - 2 - 1)
bball2_vars %>% filter(.cooksd > ro_thumb) %>% dplyr::select(teamID)
```

```
##   teamID
## 1    COL
## 2    OAK
```

```
bball2_vars %>% filter(.ext.resid > 2) %>% dplyr::select(teamID)
```

```
##   teamID
## 1    OAK
```

NYA, COL, and OAK seem to be most influential in this model.

Is the influence on x, y, or on the overall regression line?

Influence is on all three, as below. on x (leverage): NYA on y (studentized residual): OAK on overall regression line (cook's d): COL, OAK

Are there any observations you would consider excluding from analyses?

Yes. NYA (Yankees) and possibly OAK (A's). I suppose we'd label both of these datapoints 'special snowflakes'.

Why?

The Yankees seem to be in a league of their own in that both models used here indicate they have high influence, labelling them as outliers. The behavior of the Yankees may indeed model how any team would behave if it had the same payroll, avg wins, etc. but there isn't much data to support that, thus removing NYA may allow us to better predict the behavior of other teams in the league.

The A's were also shown to have disproportional influence in both models. If our goal was to predict the avg wins of a team, given the other measures in our models, removing the A's should reduce the error in our prediction.

Are any observations you flagged as influential in your SLR model also unduly influencing your MLR model?

Yes. NYA and OAK (see above).

Examine the following diagnostics for the multiple linear regression object you created above, bball2:

Variance inflation: a problem?

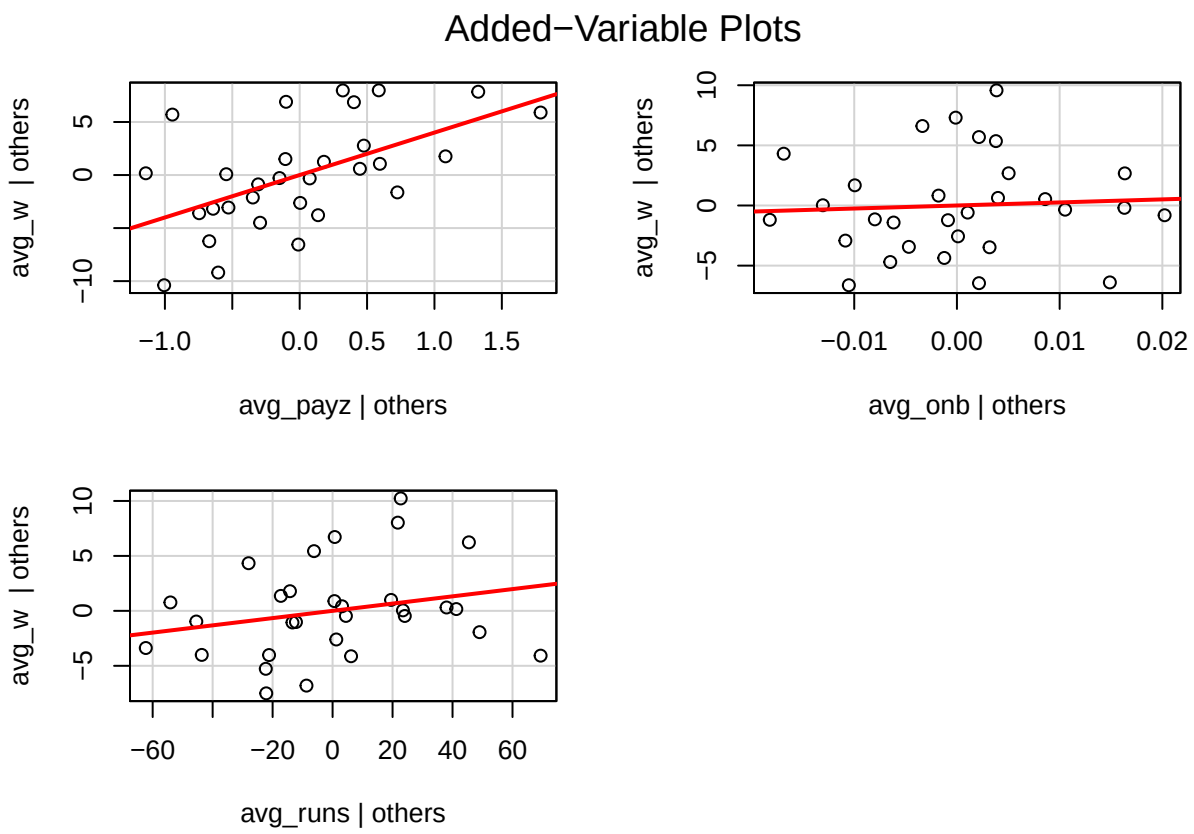
```
vif(bball2)
```

```
## avg_payz  avg_onb avg_runs
## 1.420382 1.453101 1.904555
```

This doesn't seem to be a problem as all values are fairly close to 1.

Added variable plot: comment on outliers

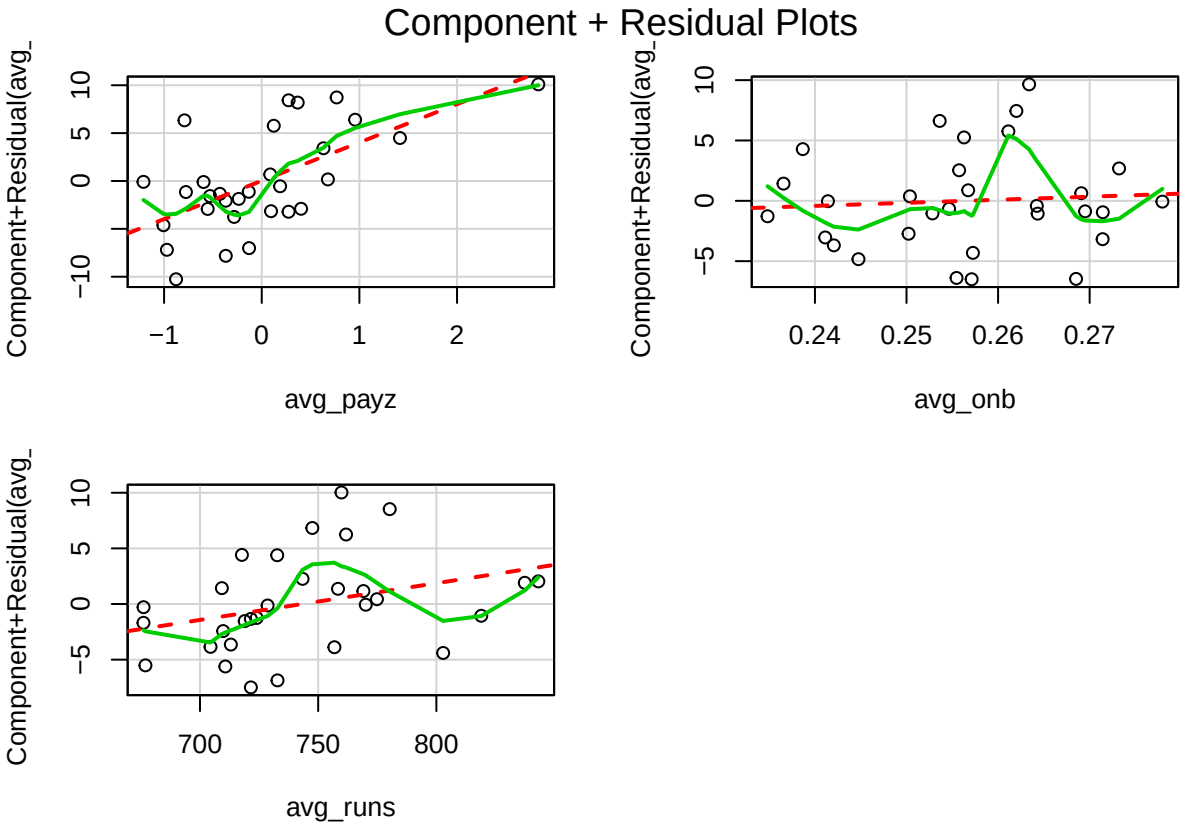
```
avPlots(bball2)
```



There are a few suspicious looking dots but the relationships do all appear linear. No red flags here.

Component residual plot: comment on linearity

```
crPlots(bball2)
```



avg_onb and avg_runs both appear to have some silliness going on. Perhaps a more complex model (polynomial?) would help to capture some of this complexity. It looks as if a linear model fits avg_payz just fine.

Report your process

Reflect on what was hard/easy, problems you solved, helpful tutorials you read, etc.

It's not clear to me how to best choose the features of a linear model. Also, if any of diagnostics failed for the MLR, I'd be unsure what next steps to take.

Is simple or multiple linear regression part of your replication/extension project?

Yes. We're doing four simple linear regressions.

Are there any elements that are part of a good SLR/MLR model analysis/diagnostic that were not reported in your article?

Yes. None of leverage, studentized residuals, cook's d were reported in the main text, though the residuals were shown to be roughly normally destributed in Figure 4. The diagnostic measures described in this homework were not reported in the main text.