modeling

Ruoyuan Qian

4/20/2022

```
fifa_model = read_csv2("fifa_model.csv")
```

```
## Using ',' as decimal and '.' as grouping mark. Use read_delim() for more control.
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   sofifa_id = col_double(),
##   short_name = col_character(),
##   age = col_double(),
##   league_name = col_character(),
##   overall = col_double(),
##   value_eur = col_double(),
##   preferred_foot = col_character(),
##   team_position = col_character(),
##   pace = col_double(),
##   shooting = col_double(),
##   passing = col_double(),
##   dribbling = col_double(),
##   defending = col_double(),
##   physic = col_double(),
##   BMI = col_double(),
##   log_value_eur = col_double()
## )
```

```
fit.all =
lm(log_value_eur ~ age + league_name + overall + preferred_foot + team_position + pace + shooting + pass

null = lm(log_value_eur ~ 1, data = fifa_model)
full = lm(log_value_eur ~ age + league_name + overall + preferred_foot + team_position + pace + shootin

#stepAIC(object = null, scope = list(upper = full),
#        direction = "forward", k = 2)
#
#stepAIC(object = null, scope = list(upper = full),
#        direction = "forward", k = log(2756))
#
#stepAIC(object = full, scope = list(upper = full),
#        direction = "backward", k = log(2756))
#
#stepAIC(object = null, scope = list(upper = full),
#        direction = "both", k = log(2756))
```

Final model

```
#FinaL MODEL FOR stepwise

#fit.AIC.fw = lm(formula = log_value_eur ~ overall + age + shooting + team_position +
#    league_name + physic + defending + dribbling, data = fifa_model)

fit.BIC.fw = lm(formula = log_value_eur ~ overall + age + shooting + team_position,
    data = fifa_model)

fit.BIC.bw = lm(formula = log_value_eur ~ age + overall + team_position +
    shooting + defending + physic, data = fifa_model)

fit.BIC.sw = lm(formula = log_value_eur ~ overall + age + shooting + team_position,
    data = fifa_model)

# Stepwise is same as forward
```

```
require(broom)
```

```
## Loading required package: broom
```

```
## Warning: package 'broom' was built under R version 3.6.2
```

```
Model_sum<-bind_rows(glance(null),
                     glance(full),
                     glance(fit.BIC.fw),
                     glance(fit.BIC.bw),
                     glance(fit.BIC.sw))%>%
  round(.,3)
Model_sum$names<-c("Intercept",
                   "Full",
                   "BIC Forward",
                   "BIC Backward",
                   "BIC Stepwise")
names(Model_sum)
```

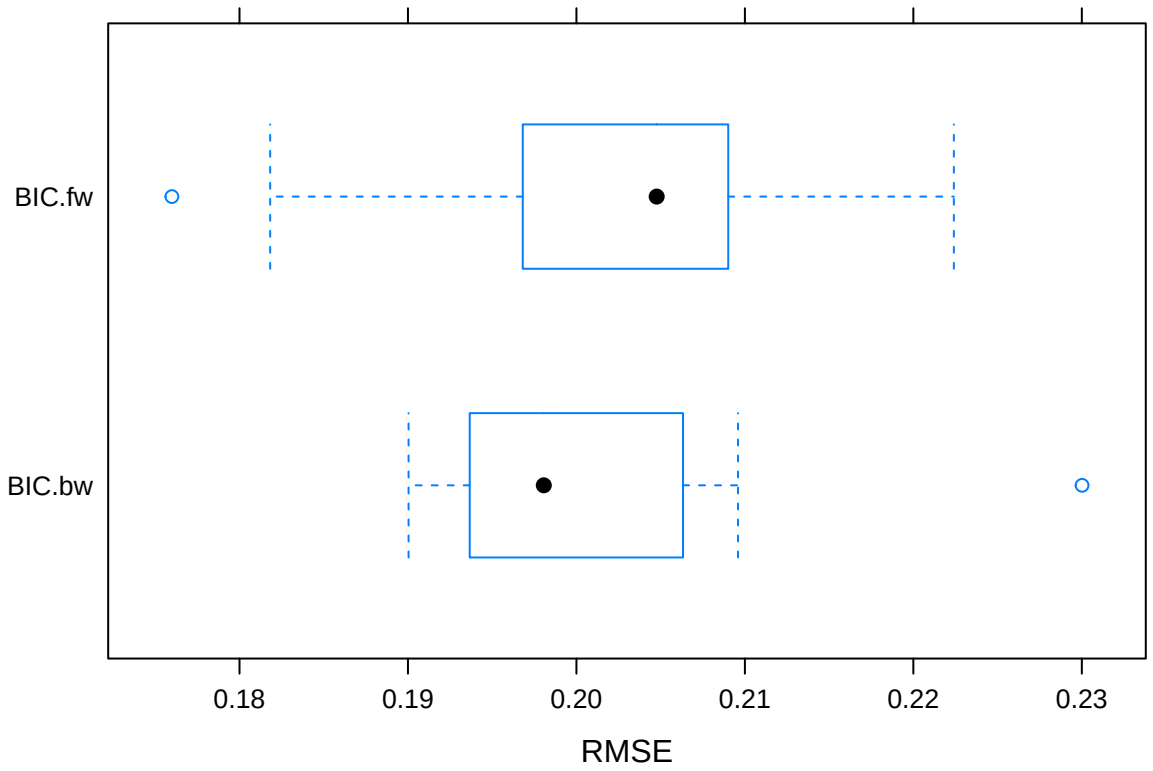
```
##  [1] "r.squared"     "adj.r.squared" "sigma"         "statistic"
##  [5] "p.value"       "df"            "logLik"        "AIC"
##  [9] "BIC"           "deviance"      "df.residual"   "nobs"
## [13] "names"
```

```
Model_sum %>%
  dplyr::select(names, r.squared, adj.r.squared,AIC, BIC, sigma,df, df.residual)
```

```
## # A tibble: 5 x 8
##   names        r.squared adj.r.squared    AIC   BIC sigma    df df.residual
##   <chr>            <dbl>         <dbl>  <dbl> <dbl> <dbl> <dbl>       <dbl>
## 1 Intercept         0             0     9736. 9748. 1.42     NA        2755
## 2 Full              0.98          0.98 -1023. -917. 0.2      16        2739
## 3 BIC Forward       0.98          0.98  -995. -954. 0.202     5        2750
## 4 BIC Backward      0.98          0.98 -1007. -953. 0.201     7        2748
## 5 BIC Stepwise      0.98          0.98  -995. -954. 0.202     5        2750
```

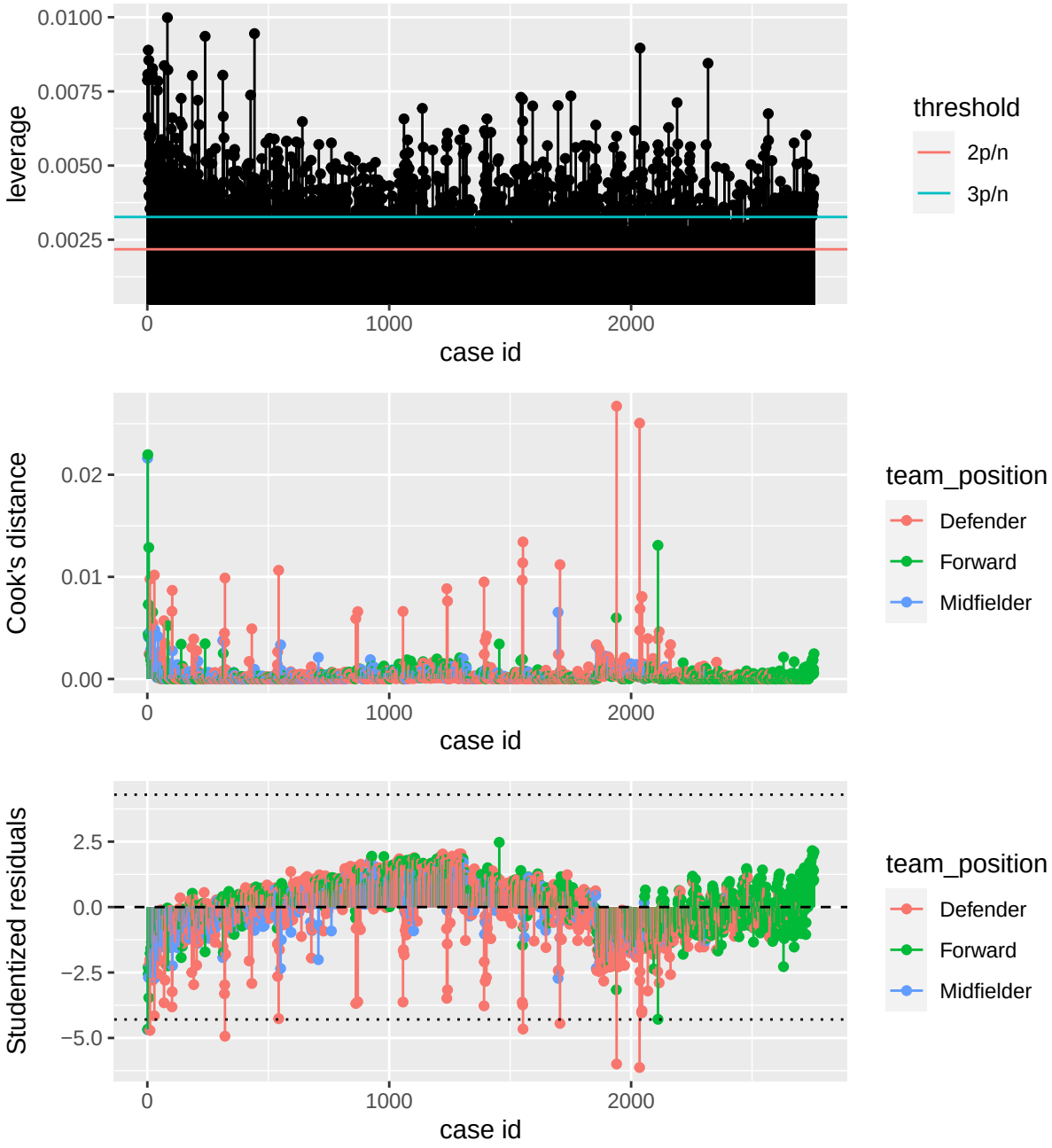
cv

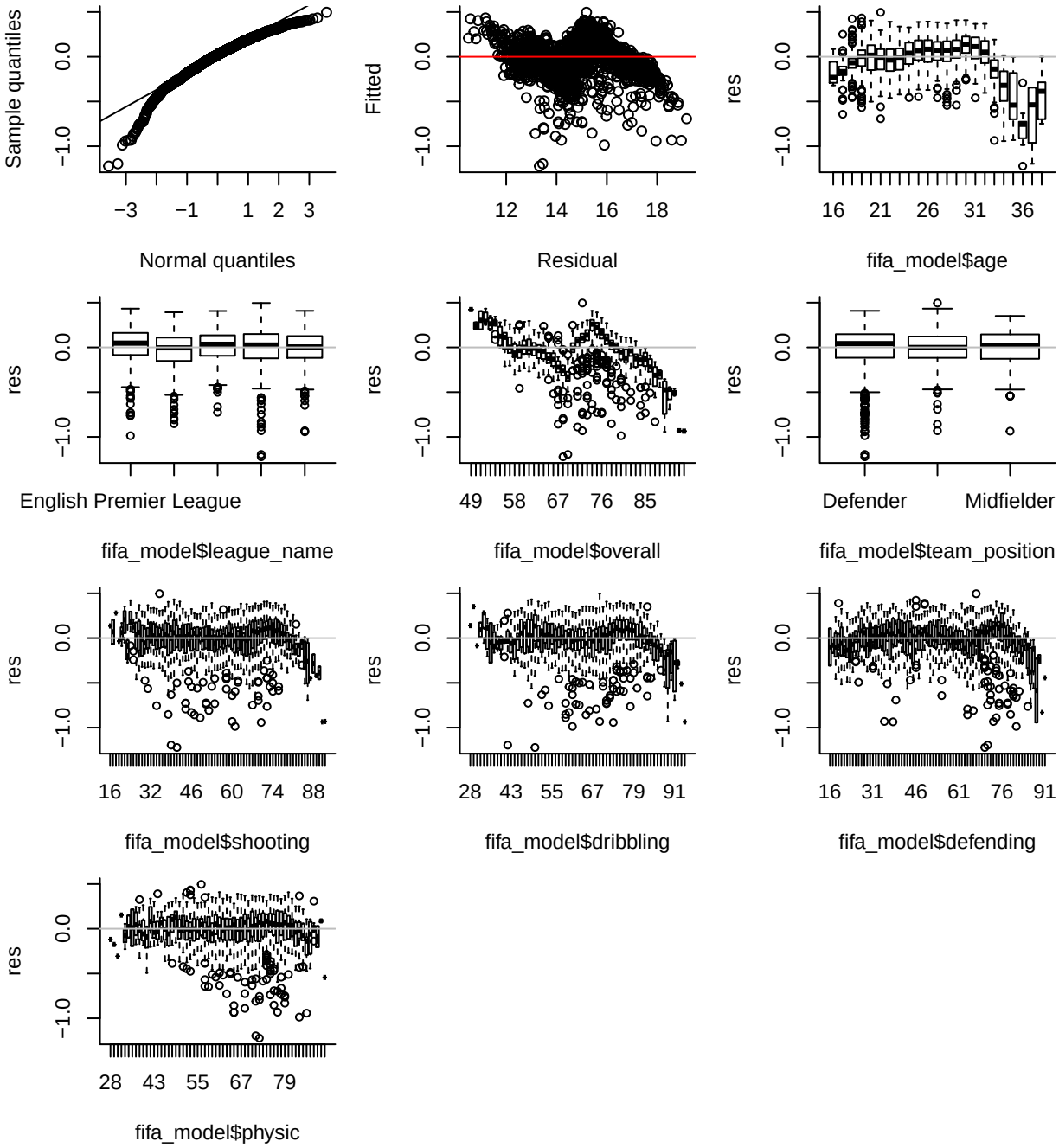
```
set.seed(6950)
tran.control = trainControl(method = "cv",number = 10)

fit.BIC.fw.cv = train(
  log_value_eur ~ overall + age + shooting + team_position,
              fifa_model, method = "lm",
              trControl = tran.control)

fit.BIC.bw.cv = train(
  log_value_eur ~ age + overall + team_position + shooting + defending + physic,
              fifa_model, method = "lm",
              trControl = tran.control)

resamp <- resamples(list(BIC.fw = fit.BIC.fw.cv,
                         BIC.bw = fit.BIC.bw.cv))
#summary(resamp)

bwplot(resamp, metric = "RMSE")
```




```
# select bardward : fit.BIC.bw
# param 6

n.param = dim(fifa_model)[1]- fit.BIC.bw$df.residual
```

Model diagnosis





All age >30, and majority of them are defender. Therefore, we want to investigate whether there are interaction effects between team postion and age.

```
fit.BIC.quart<-lm(formula = log_value_eur ~ poly(age,2) + league_name +  poly(overall,2) + team_positio
    shooting + dribbling + defending + physic, data = fifa_model)

summary(fit.BIC.quart)
```

```
##
## Call:
## lm(formula = log_value_eur ~ poly(age, 2) + league_name + poly(overall,
##     2) + team_position + shooting + dribbling + defending + physic,
##     data = fifa_model)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.96059 -0.09940  0.02403  0.12367  0.69403
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       1.499e+01  6.745e-02 222.184  < 2e-16 ***
## poly(age, 2)1                     -1.606e+01  2.483e-01 -64.682  < 2e-16 ***
## poly(age, 2)2                     -4.237e+00  2.023e-01 -20.948  < 2e-16 ***
## league_nameFrench Ligue 1         -7.150e-02  1.108e-02  -6.452 1.30e-10 ***
## league_nameGerman 1. Bundesliga   -2.323e-02  1.121e-02  -2.073  0.03826 *
## league_nameItalian Serie A        -4.616e-02  1.083e-02  -4.263 2.08e-05 ***
## league_nameSpain Primera Division -4.104e-02  1.076e-02  -3.813  0.00014 ***
## poly(overall, 2)1                  7.937e+01  4.138e-01 191.798  < 2e-16 ***
## poly(overall, 2)2                 -1.730e+00  2.008e-01  -8.617  < 2e-16 ***
## team_positionForward              -4.356e-03  9.186e-03  -0.474  0.63544
## team_positionMidfielder            3.459e-02  1.087e-02   3.183  0.00147 **
## shooting                           2.644e-03  4.849e-04   5.453 5.39e-08 ***
## dribbling                         -4.430e-04  7.315e-04  -0.606  0.54485
## defending                         -7.052e-04  3.317e-04  -2.126  0.03359 *
## physic                            -4.054e-04  5.717e-04  -0.709  0.47833
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.181 on 2741 degrees of freedom
## Multiple R-squared:  0.9837, Adjusted R-squared:  0.9836
## F-statistic: 1.182e+04 on 14 and 2741 DF,  p-value: < 2.2e-16
```

Since the estimate of quadratic term of overall is negative, we drop it.

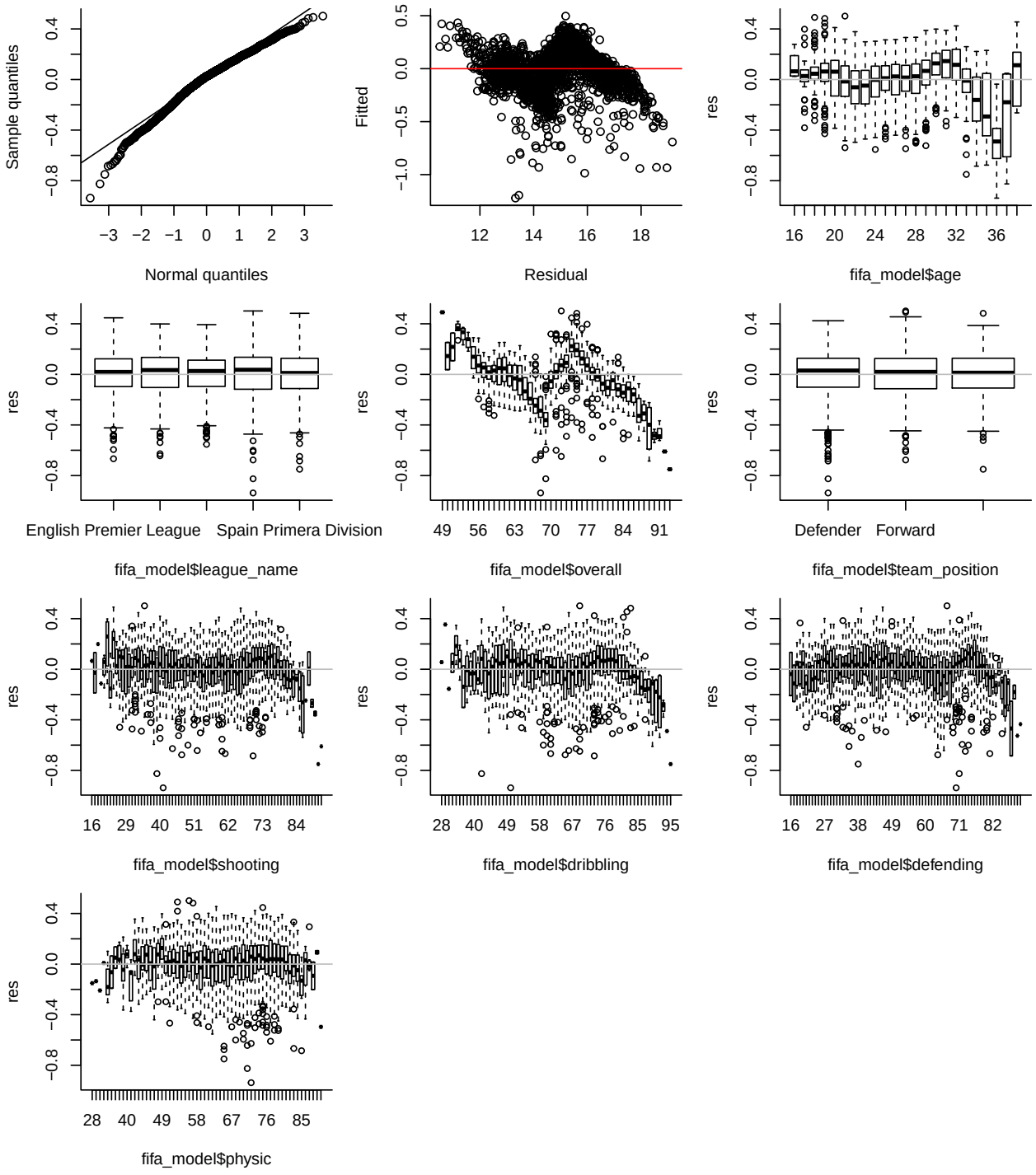
```
fit.BIC.quart<-lm(formula = log_value_eur ~ poly(age,2) + league_name +  overall + team_position +
    shooting + dribbling + defending + physic, data = fifa_model)

summary(fit.BIC.quart)
```

```
##
## Call:
## lm(formula = log_value_eur ~ poly(age, 2) + league_name + overall +
##     team_position + shooting + dribbling + defending + physic,
##     data = fifa_model)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.93781 -0.10756  0.02452  0.12707  0.50152
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       1.587e-01  5.238e-02   3.030 0.002468 **
## poly(age, 2)1                     -1.566e+01  2.471e-01 -63.368  < 2e-16 ***
## poly(age, 2)2                     -4.598e+00  2.005e-01 -22.931  < 2e-16 ***
## league_nameFrench Ligue 1         -6.097e-02  1.116e-02  -5.463 5.10e-08 ***
## league_nameGerman 1. Bundesliga   -1.464e-02  1.131e-02  -1.295 0.195562
## league_nameItalian Serie A        -3.722e-02  1.092e-02  -3.408 0.000665 ***
## league_nameSpain Primera Division -3.296e-02  1.087e-02  -3.033 0.002442 **
```

```
## overall                             2.052e-01  1.067e-03 192.245  < 2e-16 ***
## team_positionForward               -2.538e-02  8.974e-03  -2.828 0.004713 **
## team_positionMidfielder             2.585e-02  1.096e-02   2.358 0.018444 *
## shooting                            2.323e-03  4.899e-04   4.743 2.21e-06 ***
## dribbling                           4.015e-04  7.345e-04   0.547 0.584657
## defending                          -9.294e-04  3.351e-04  -2.774 0.005581 **
## physic                              2.369e-04  5.743e-04   0.412 0.680075
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1834 on 2742 degrees of freedom
## Multiple R-squared:  0.9833, Adjusted R-squared:  0.9832
## F-statistic: 1.239e+04 on 13 and 2742 DF,  p-value: < 2.2e-16
```

```r
res <- residuals(fit.BIC.quart)
fits <- fitted(fit.BIC.quart)

#par(mfrow=c(5,2))
par(mfrow=c(4,3), cex=0.75, mar=c(4,4,1,1), bty="L")

qqnorm(fit.BIC.quart$residuals,
       xlab="Normal quantiles", ylab="Sample quantiles", main="")
qqline(fit.BIC.quart$residuals)
#hist(fit.BIC.bw$residuals, xlab="Residuals", main="")
plot(y=model_diag %>% pull(.resid) , x=model_diag %>% pull(.fitted),xlab="Residual", ylab="Fitted")
abline(h=0,col="red")

#par(mfrow=c(3,3), cex=0.75, mar=c(4,4,1,1), bty="L")

#plot(fits, res)
#abline(h=0, col="gray")

boxplot(res ~ fifa_model$age)
abline(h=0, col="gray")

boxplot(res ~ fifa_model$league_name)
abline(h=0, col="gray")

boxplot(res ~ fifa_model$overall)
abline(h=0, col="gray")

boxplot(res ~ fifa_model$team_position)
abline(h=0, col="gray")

boxplot(res ~ fifa_model$shooting)
abline(h=0, col="gray")

boxplot(res ~ fifa_model$dribbling)
abline(h=0, col="gray")

boxplot(res ~ fifa_model$defending)
abline(h=0, col="gray")

boxplot(res ~ fifa_model$physic)
```

```
abline(h=0, col="gray")
```




```
# Appendix XX
model_diag %>%
  dplyr::select(short_name,log_value_eur, age, league_name, overall,team_position,shooting,
         dribbling,defending,physic,.ti) %>%
  filter(abs(.ti) > abs(qt((0.05/(2*dim(fifa_model)[1])), df = dim(fifa_model)[1]-9 - 1))) %>%
  knitr::kable()
```

short_name	log_value_eur	age	league_name	overall	team_position	shooting	dribbling	defending	physic	.ti
L. Messi	18.02764	33	Spain Primera Division	93	Midfielder	92	95	38	65	-4.685423
Cristiano Ronaldo	17.64415	35	Italian Serie A	92	Forward	93	89	35	77	-4.663520
Sergio Ramos	17.01418	34	Spain Primera Division	89	Defender	70	73	88	85	-4.709715
B. Ivanović	14.91412	36	English Premier League	80	Defender	61	61	81	83	-4.933934
Pedro López	13.26213	36	Spain Primera Division	72	Defender	60	69	73	65	-4.660271
F. Peluso	13.07107	36	Italian Serie A	71	Defender	49	59	74	68	-4.445786
N. Spolli	12.25486	37	Italian Serie A	69	Defender	39	42	71	71	-5.996256
C. Terzi	12.10071	36	Italian Serie A	68	Defender	41	50	70	72	-6.129129