Draft Write-Up

Michael Li

April 12, 2021

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v tibble  3.0.6     v dplyr   1.0.3
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## v purrr   0.3.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readr)
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
library(broom)
```

```
players <- read_csv("data/players.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_character(),
##   career_AST = col_double(),
##   career_G = col_double(),
##   career_PTS = col_double(),
##   career_WS = col_double()
## )
## i Use `spec()` for the full column specifications.
```

```
## Warning: 2 parsing failures.
##  row       col expected actual               file
## 2274 career_WS a double      - 'data/players.csv'
## 4370 career_WS a double      - 'data/players.csv'
```

```
salaries <- read_csv("data/salaries_1985to2018.csv")
```

```
##
```

```
## -- Column specification --------------------------------------------------------
## cols(
##   league = col_character(),
##   player_id = col_character(),
##   salary = col_double(),
##   season = col_character(),
##   season_end = col_double(),
##   season_start = col_double(),
##   team = col_character()
## )
```

```
colnames(players)[1] <- "player_id"

teams <- salaries %>%
  group_by(player_id) %>%
  count(team) %>%
  mutate(years_with_team = max(n)) %>%
  subset(n == years_with_team) %>%
  slice(1) %>%
  select(player_id, team, years_with_team)

# df of aggregate salaries
agg_salaries <- salaries %>%
  group_by(player_id) %>%
  summarise(career_salary = sum(salary),
            career_start = min(season_start),
            career_end = max(season_end))

agg_salaries <- agg_salaries %>%
  merge(teams, by = "player_id")
```

```
df <- players %>%
  merge(agg_salaries, by = "player_id") %>%
  separate(col = birthDate, into = c("MonthDay", "birthYear"), sep = ", ") %>%
  separate(col = birthPlace, into = c("City", "birthPlace"), sep = ", ") %>%
  separate(col = draft_pick, into = c("draft_pick", "overall"), sep = "[thrdndst]") %>%
  separate(col = height, into = c("feet", "inches"), sep = "-") %>%
  mutate(height = as.double(feet) * 12 + as.double(inches)) %>%
  separate(col = position, into = c("primary_pos", "secondary_pos", "tertiary_pos", "quarternary_pos"),
           sep = " and ") %>%
  mutate(num_positions = if_else(is.na(primary_pos), 0, 1) +
           if_else(is.na(secondary_pos), 0, 1) +
           if_else(is.na(tertiary_pos), 0 , 1) +
           if_else(is.na(quarternary_pos), 0, 1)) %>%
  separate(col = weight, into = c("weight", "metric"), sep = "l") %>%
  select(-c(MonthDay, City, overall, draft_round, feet, inches, metric)) %>%
  mutate(years_played = career_end - career_start) %>%
  mutate(averageWS = career_WS / years_played)

#df$birthYear <- as.Date(df$birthYear, "%Y")
df$`career_FG%` <- as.double(df$`career_FG%`)
df$`career_FG3%` <- as.double(df$`career_FG3%`)
df$`career_FT%` <- as.double(df$`career_FT%`)
df$career_TRB <- as.double(df$career_TRB)
df$`career_eFG%` <- as.double(df$`career_eFG%`)
```

```r
df$draft_year <- as.double(df$draft_year)
df$weight <- as.double(df$weight)
df$career_PER <- as.double(df$career_PER)
df$draft_pick <- as.integer(df$draft_pick)
#df$career_start <- as.Date(as.character(df$career_start), "%Y")
#df$career_end <- as.Date(as.character(df$career_end), "%Y")
df <- df %>%
  mutate(average_salary  = (career_salary / years_played)/1000000) # salary in millions
```

Linear Model

```r
lm_sal <- lm(average_salary ~ career_AST +
             + `career_G` + `career_PER` + career_PTS + career_TRB + averageWS +
               `career_eFG%` + draft_pick + primary_pos +
               num_positions + draft_year,
             data  = df)
summary(lm_sal)
```

```
##
## Call:
## lm(formula = average_salary ~ career_AST + +career_G + career_PER +
##     career_PTS + career_TRB + averageWS + `career_eFG%` + draft_pick +
##     primary_pos + num_positions + draft_year, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8349 -0.9665 -0.1053  0.8567  7.6585
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -2.299e+02  7.454e+00 -30.842  < 2e-16 ***
## career_AST                   1.419e-01  4.339e-02   3.270 0.001095 **
## career_G                     2.684e-03  1.800e-04  14.913  < 2e-16 ***
## career_PER                   5.685e-02  1.495e-02   3.803 0.000148 ***
## career_PTS                   1.943e-01  1.618e-02  12.007  < 2e-16 ***
## career_TRB                   1.860e-01  3.258e-02   5.709 1.32e-08 ***
## averageWS                   -5.302e-02  6.405e-03  -8.278 2.35e-16 ***
## `career_eFG%`               -4.729e-02  7.645e-03  -6.186 7.58e-10 ***
## draft_pick                   2.521e-04  1.907e-03   0.132 0.894857
## primary_posPoint Guard      -7.035e-01  1.818e-01  -3.870 0.000112 ***
## primary_posPower Forward    -2.677e-01  1.172e-01  -2.284 0.022504 *
## primary_posShooting Guard   -5.091e-01  1.496e-01  -3.403 0.000681 ***
## primary_posSmall Forward    -4.520e-01  1.335e-01  -3.386 0.000725 ***
## num_positions                4.411e-02  7.441e-02   0.593 0.553442
## draft_year                   1.154e-01  3.744e-03  30.833  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.626 on 1867 degrees of freedom
##   (526 observations deleted due to missingness)
## Multiple R-squared:  0.6496, Adjusted R-squared:  0.647
```

```
## F-statistic: 247.2 on 14 and 1867 DF,  p-value: < 2.2e-16
```

```
lm_sal_out <- tidy(lm_sal, conf.int = TRUE)
lm_sal_out$term <- c(
  "(Intercept)",
  "APG", "CareerGames", "PER", "PPG", "RPG",
  "WinShares", "eFGPercentage", "Draft Pick", "PrimaryPositionPG", "PrimaryPositionPF",
  "PrimaryPositionSG", "PrimaryPositionSF", "NumberOfPositions", "DraftYear"
)
knitr::kable(lm_sal_out, digits = 3, caption = "Average Salary OLS Model Output", col.names = c('Term',
  kable_styling(latex_options = "HOLD_position")
```

Table 1: Average Salary OLS Model Output

Term	Estimate	Standard Error	Statistic	P-Value	CI (low)	CI (high)
(Intercept)	-229.912	7.454	-30.842	0.000	-244.531	-215.292
APG	0.142	0.043	3.270	0.001	0.057	0.227
CareerGames	0.003	0.000	14.913	0.000	0.002	0.003
PER	0.057	0.015	3.803	0.000	0.028	0.086
PPG	0.194	0.016	12.007	0.000	0.163	0.226
RPG	0.186	0.033	5.709	0.000	0.122	0.250
WinShares	-0.053	0.006	-8.278	0.000	-0.066	-0.040
eFGPercentage	-0.047	0.008	-6.186	0.000	-0.062	-0.032
Draft Pick	0.000	0.002	0.132	0.895	-0.003	0.004
PrimaryPositionPG	-0.703	0.182	-3.870	0.000	-1.060	-0.347
PrimaryPositionPF	-0.268	0.117	-2.284	0.023	-0.498	-0.038
PrimaryPositionSG	-0.509	0.150	-3.403	0.001	-0.803	-0.216
PrimaryPositionSF	-0.452	0.133	-3.386	0.001	-0.714	-0.190
NumberOfPositions	0.044	0.074	0.593	0.553	-0.102	0.190
DraftYear	0.115	0.004	30.833	0.000	0.108	0.123

```
car::vif(lm_sal)
```

```
##                    GVIF Df GVIF^(1/(2*Df))
## career_AST     3.361886  1        1.833545
## career_G       2.756200  1        1.660181
## career_PER     3.516675  1        1.875280
## career_PTS     4.799689  1        2.190819
## career_TRB     3.677863  1        1.917775
## averageWS      1.444584  1        1.201908
## `career_eFG%` 2.009769  1        1.417663
## draft_pick     1.313078  1        1.145896
## primary_pos    3.411033  4        1.165764
## num_positions 1.223556  1        1.106145
## draft_year     1.359736  1        1.166077
```

```
lm_ws <- lm(averageWS ~ career_AST +
             + `career_G` + `career_PER` + career_PTS + career_TRB +
               `career_eFG%` + draft_pick + primary_pos +
               num_positions + draft_year,
             data  = df)
summary(lm_ws)
```

```
##
```

```
## Call:
## lm(formula = averageWS ~ career_AST + +career_G + career_PER +
##     career_PTS + career_TRB + `career_eFG%` + draft_pick + primary_pos +
##     num_positions + draft_year, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -9.751  -2.086  -0.350   1.123 136.060
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 2.388e+02  2.636e+01   9.061  < 2e-16 ***
## career_AST                  1.298e-01  1.567e-01   0.828   0.4077
## career_G                    3.384e-04  6.501e-04   0.521   0.6027
## career_PER                  8.453e-03  5.400e-02   0.157   0.8756
## career_PTS                  4.901e-01  5.735e-02   8.545  < 2e-16 ***
## career_TRB                  5.717e-01  1.169e-01   4.890 1.10e-06 ***
## `career_eFG%`               5.692e-03  2.762e-02   0.206   0.8367
## draft_pick                  4.099e-02  6.823e-03   6.007 2.27e-09 ***
## primary_posPoint Guard     -7.964e-01  6.564e-01  -1.213   0.2251
## primary_posPower Forward   -7.679e-01  4.231e-01  -1.815   0.0697 .
## primary_posShooting Guard -4.195e-01  5.404e-01  -0.776   0.4377
## primary_posSmall Forward   -6.478e-01  4.820e-01  -1.344   0.1791
## num_positions              -2.702e-01  2.687e-01  -1.006   0.3148
## draft_year                 -1.213e-01  1.323e-02  -9.170  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.876 on 1868 degrees of freedom
##   (526 observations deleted due to missingness)
## Multiple R-squared:  0.3078, Adjusted R-squared:  0.3029
## F-statistic: 63.88 on 13 and 1868 DF,  p-value: < 2.2e-16
```

```
lm_ws_out <- tidy(lm_ws, conf.int = TRUE)
lm_ws_out$term <- c(
  "(Intercept)",
  "APG", "CareerGames", "PER", "PPG", "RPG",
  "eFGPercentage", "Draft Pick", "PrimaryPositionPG", "PrimaryPositionPF",
  "PrimaryPositionSG", "PrimaryPositionSF", "NumberOfPositions", "DraftYear"
)
knitr::kable(lm_ws_out, digits = 3, caption = "Average Win Shares OLS Model Output", col.names = c('Ter
  kable_styling(latex_options = "HOLD_position")
```

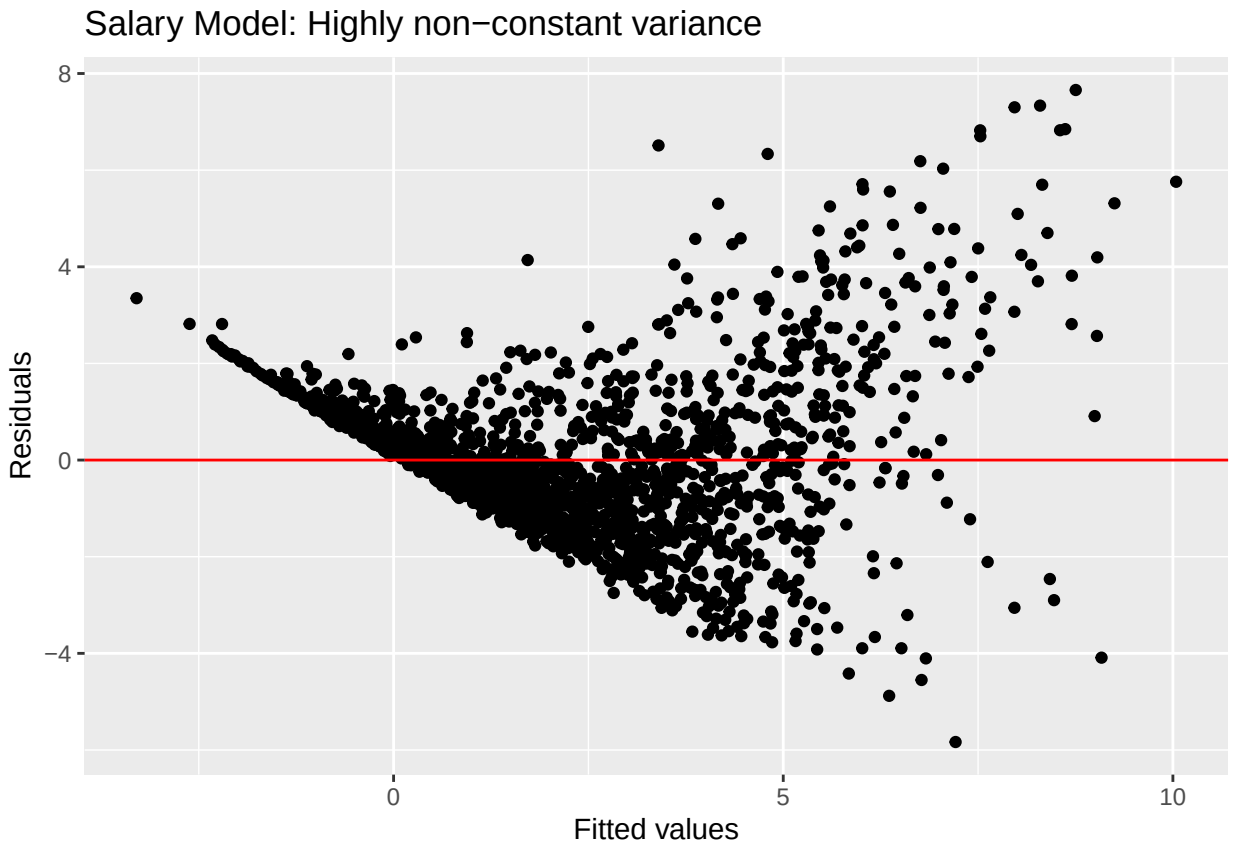
Table 2: Average Win Shares OLS Model Output

Term	Estimate	Standard Error	Statistic	P-Value	CI (low)	CI (high)
(Intercept)	238.806	26.356	9.061	0.000	187.116	290.495
APG	0.130	0.157	0.828	0.408	-0.178	0.437
CareerGames	0.000	0.001	0.521	0.603	-0.001	0.002
PER	0.008	0.054	0.157	0.876	-0.097	0.114
PPG	0.490	0.057	8.545	0.000	0.378	0.603
RPG	0.572	0.117	4.890	0.000	0.342	0.801
eFGPercentage	0.006	0.028	0.206	0.837	-0.048	0.060
Draft Pick	0.041	0.007	6.007	0.000	0.028	0.054
PrimaryPositionPG	-0.796	0.656	-1.213	0.225	-2.084	0.491
PrimaryPositionPF	-0.768	0.423	-1.815	0.070	-1.598	0.062
PrimaryPositionSG	-0.419	0.540	-0.776	0.438	-1.479	0.640
PrimaryPositionSF	-0.648	0.482	-1.344	0.179	-1.593	0.297
NumberOfPositions	-0.270	0.269	-1.006	0.315	-0.797	0.257
DraftYear	-0.121	0.013	-9.170	0.000	-0.147	-0.095

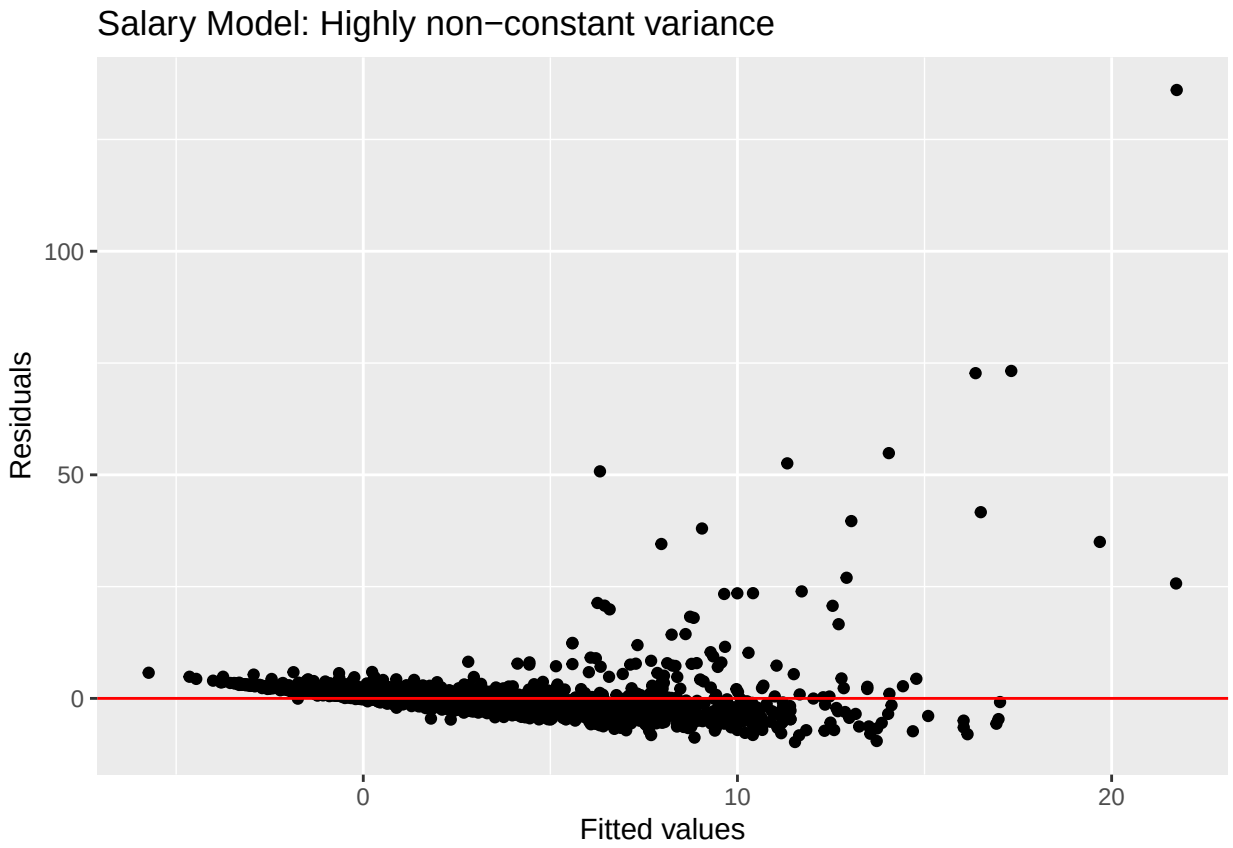
```
car::vif(lm_sal)
```

```
##                   GVIF Df GVIF^(1/(2*Df))
## career_AST    3.361886  1        1.833545
## career_G      2.756200  1        1.660181
## career_PER    3.516675  1        1.875280
## career_PTS    4.799689  1        2.190819
## career_TRB    3.677863  1        1.917775
## averageWS     1.444584  1        1.201908
## `career_eFG%` 2.009769  1        1.417663
## draft_pick    1.313078  1        1.145896
## primary_pos   3.411033  4        1.165764
## num_positions 1.223556  1        1.106145
## draft_year    1.359736  1        1.166077
```

```
temp_lm <- tibble(res = lm_sal$residuals,
                  fitted = lm_sal$fitted.values)
ggplot(data = temp_lm, aes(x = fitted, y = res)) +
  geom_point() +
  labs(x = "Fitted values", y = "Residuals",
       title = "Salary Model: Highly non-constant variance") +
  geom_hline(yintercept = 0, color = "red")
```




```
temp_ws <- tibble(res = lm_ws$residuals,
                  fitted = lm_ws$fitted.values)
ggplot(data = temp_ws, aes(x = fitted, y = res)) +
  geom_point() +
  labs(x = "Fitted values", y = "Residuals",
       title = "Salary Model: Highly non-constant variance") +
  geom_hline(yintercept = 0, color = "red")
```


Salary Model: Highly non−constant variance
Residuals
100
50
0
0
10
20
Fitted values