analysis.rmd

2025-12-02

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union

Loading required package: carData


Attaching package: 'car'

The following object is masked from 'package:dplyr':

    recode

Loading required package: lattice


Attaching package: 'olsrr'

The following object is masked from 'package:datasets':

    rivers
```

```r
df <- read.csv("main.csv")
df <- na.omit(df)
head(df)
```

```
##                    Player Team  G   MP FG_pct  FTA  TRB  AST stocks  PTS
## 1 Shai Gilgeous-Alexander  OKC 76 34.2  0.519  8.8  5.0  6.4    2.7 32.7
## 2   Giannis Antetokounmpo  MIL 67 34.2  0.601 10.6 11.9  6.5    2.1 30.4
## 3            Nikola Jokić  DEN 70 36.7  0.576  6.4 12.7 10.2    2.4 29.6
## 5         Anthony Edwards  MIN 79 36.3  0.447  6.3  5.7  4.5    1.8 27.6
## 6            Jayson Tatum  BOS 72 36.4  0.452  6.1  8.7  6.0    1.6 26.8
## 7            Kevin Durant  PHO 62 36.5  0.527  5.8  6.0  4.2    2.0 26.6
##   Value_Billions awards_1 awards_2plus avg_salary_millions Age_22_26 Age_27_31
## 1           4.35        0            1             55.3591         1         0
## 2           4.30        0            1             58.4566         0         1
## 3           4.60        0            1             59.0331         0         1
## 5           3.60        0            1             50.6117         1         0
## 6           6.70        0            1             62.7867         1         0
## 7           5.43        1            0             54.7086         0         0
##   Age_32_34 Age_35_plus Pos_PF Pos_PG Pos_SF Pos_SG   Age    Awards
## 1         0           0      0      1      0      0 Age_1 2+ awards
## 2         0           0      1      0      0      0 Age_2 2+ awards
```

```
## 3          0           0      0      0      0      0 Age_2 2+ awards
## 5          0           0      0      0      0      1 Age_1 2+ awards
## 6          0           0      1      0      0      0 Age_1 2+ awards
## 7          0           1      1      0      0      0 Age_4   1 award
```

Log-transformation of Data

```r
df <- df[df$MP > 20, ]
df$log_salary <- log(df$avg_salary_millions + 1)
```

Initial Model Creation

```r
# need to change to sqrt_sal now
initial_model <- lm(log_salary ~ MP + PTS + FG_pct + FTA + TRB + AST + stocks + Value_Billions + (PTS *
summary(initial_model)
```

```
##
## Call:
## lm(formula = log_salary ~ MP + PTS + FG_pct + FTA + TRB + AST +
##     stocks + Value_Billions + (PTS * FTA), data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.56608 -0.30240  0.07474  0.39067  1.06634
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.142483   0.513725   0.277   0.7818
## MP              0.026935   0.016035   1.680   0.0948 .
## PTS             0.103407   0.021991   4.702 5.22e-06 ***
## FG_pct         -0.348649   0.936100  -0.372   0.7100
## FTA             0.099041   0.087697   1.129   0.2603
## TRB             0.028151   0.024461   1.151   0.2514
## AST             0.030023   0.028370   1.058   0.2914
## stocks          0.179301   0.085205   2.104   0.0368 *
## Value_Billions  0.013198   0.022176   0.595   0.5525
## PTS:FTA        -0.006443   0.003506  -1.838   0.0678 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5487 on 174 degrees of freedom
## Multiple R-squared:  0.6151, Adjusted R-squared:  0.5952
## F-statistic:  30.9 on 9 and 174 DF,  p-value: < 2.2e-16
```

VIF Analysis

```r
vif_values <- vif(initial_model)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```r
print(vif_values)
```

```
##             MP            PTS         FG_pct            FTA            TRB
##       3.501201      10.536141       1.951849      17.818292       2.285399
##            AST         stocks Value_Billions        PTS:FTA
##       2.012890       1.395618       1.014153      23.427323
```

```r
# high vifs detected, remove
vif_less <- lm(log_salary ~ MP + PTS + FG_pct + FTA + TRB + AST + stocks + Value_Billions, data = df)
print(vif(vif_less))
```

```
##             MP            PTS         FG_pct            FTA            TRB
##       3.278581       8.043663       1.945680       5.252468       2.273890
##            AST         stocks Value_Billions
##       1.986783       1.394672       1.013286
```

Stepwise

```r
ols_step_both_p(vif_less,p_ent=0.15,p_rem=0.15,details=T)
```

```
## Stepwise Selection Method
## -------------------------
##
## Candidate Terms:
##
## 1. MP
## 2. PTS
## 3. FG_pct
## 4. FTA
## 5. TRB
## 6. AST
## 7. stocks
## 8. Value_Billions
##
##
## Step   => 0
## Model  => log_salary ~ 1
## R2     => 0
##
## Initiating stepwise selection...
##
## Step      => 1
## Selected  => PTS
## Model     => log_salary ~ PTS
## R2        => 0.554
##
## Step      => 2
## Selected  => MP
## Model     => log_salary ~ PTS + MP
## R2        => 0.586
##
## Step      => 3
## Selected  => stocks
## Model     => log_salary ~ PTS + MP + stocks
```

```
## R        => 0.601
##
##
## No more variables to be added or removed.
```

```
##
##
##                               Stepwise Summary
## -------------------------------------------------------------------------
## Step    Variable        AIC        SBC        SBIC        R2       Adj. R2
## -------------------------------------------------------------------------
##  0      Base Model    470.698    477.128     -53.336    0.00000    0.00000
##  1      PTS (+)       324.220    333.865    -198.262    0.55378    0.55133
##  2      MP (+)        312.612    325.472    -209.584    0.58559    0.58101
##  3      stocks (+)    307.557    323.632    -214.339    0.60118    0.59453
## -------------------------------------------------------------------------
##
## Final Model Output
## ------------------
##
##                          Model Summary
## --------------------------------------------------------------
## R                       0.775       RMSE                0.543
## R-Squared               0.601       MSE                 0.295
## Adj. R-Squared          0.595       Coef. Var          19.087
## Pred R-Squared          0.582       AIC               307.557
## MAE                     0.442       SBC               323.632
## --------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##  AIC: Akaike Information Criteria
##  SBC: Schwarz Bayesian Criteria
##
##                                ANOVA
## -------------------------------------------------------------------
##               Sum of
##              Squares         DF    Mean Square      F         Sig.
## -------------------------------------------------------------------
## Regression     81.826         3         27.275    90.444    0.0000
## Residual       54.283       180          0.302
## Total         136.110       183
## -------------------------------------------------------------------
##
##                                   Parameter Estimates
## ----------------------------------------------------------------------------------------
##       model     Beta    Std. Error    Std. Beta       t        Sig      lower    upper
## ----------------------------------------------------------------------------------------
## (Intercept)    0.153         0.295                  0.518    0.605    -0.429    0.734
##         PTS    0.077         0.011        0.534    7.004    0.000     0.055    0.099
##          MP    0.042         0.014        0.233    2.953    0.004     0.014    0.071
##      stocks    0.203         0.076        0.132    2.652    0.009     0.052    0.354
## ----------------------------------------------------------------------------------------
```

New Model as a Result of Test

```r
quant <- lm(log_salary ~ PTS + MP + stocks, data = df)
summary(quant)
```

```
##
## Call:
## lm(formula = log_salary ~ PTS + MP + stocks, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6657 -0.3807  0.1025  0.4096  1.0897
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.15259    0.29459   0.518  0.60511
## PTS          0.07698    0.01099   7.004 4.78e-11 ***
## MP           0.04243    0.01437   2.953  0.00357 **
## stocks       0.20276    0.07644   2.652  0.00870 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5492 on 180 degrees of freedom
## Multiple R-squared:  0.6012, Adjusted R-squared:  0.5945
## F-statistic: 90.44 on 3 and 180 DF,  p-value: < 2.2e-16
```

Adding Qualitative Predictors

```r
quant_and_qual <- lm(log_salary ~ PTS + MP + stocks + awards_1 + awards_2plus + Age_22_26 + Age_27_31 +
summary(quant_and_qual)
```

```
##
## Call:
## lm(formula = log_salary ~ PTS + MP + stocks + awards_1 + awards_2plus +
##     Age_22_26 + Age_27_31 + Age_32_34 + Age_35_plus + Pos_PF +
##     Pos_PG + Pos_SF + Pos_SG, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.59338 -0.38236  0.07087  0.39109  0.85694
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.232475   0.376221   0.618   0.5374
## PTS           0.078019   0.011956   6.526 7.39e-10 ***
## MP            0.037191   0.014811   2.511   0.0130 *
## stocks        0.154281   0.082681   1.866   0.0638 .
## awards_1      0.007647   0.153044   0.050   0.9602
## awards_2plus  0.094168   0.169552   0.555   0.5794
## Age_22_26     0.168827   0.194192   0.869   0.3859
## Age_27_31     0.441569   0.195977   2.253   0.0255 *
## Age_32_34     0.496186   0.229412   2.163   0.0319 *
## Age_35_plus         NA         NA      NA       NA
```

```
## Pos_PF       -0.067416   0.135301  -0.498   0.6189
## Pos_PG       -0.109450   0.140230  -0.781   0.4362
## Pos_SF       -0.182692   0.138017  -1.324   0.1874
## Pos_SG       -0.304188   0.135332  -2.248   0.0259 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5288 on 171 degrees of freedom
## Multiple R-squared:  0.6486, Adjusted R-squared:  0.624
## F-statistic: 26.31 on 12 and 171 DF,  p-value: < 2.2e-16
```

ANOVA Test for reduced model (Awards)

```
reduced_awards <- lm(log_salary ~ PTS + MP + stocks + Age_22_26 + Age_27_31 + Age_32_34 + Age_35_plus +
anova(reduced_awards, quant_and_qual)
```

```
## Analysis of Variance Table
##
## Model 1: log_salary ~ PTS + MP + stocks + Age_22_26 + Age_27_31 + Age_32_34 +
##     Age_35_plus + Pos_PF + Pos_PG + Pos_SF + Pos_SG
## Model 2: log_salary ~ PTS + MP + stocks + awards_1 + awards_2plus + Age_22_26 +
##     Age_27_31 + Age_32_34 + Age_35_plus + Pos_PF + Pos_PG + Pos_SF +
##     Pos_SG
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    173 47.912
## 2    171 47.824  2  0.087488 0.1564 0.8553
```

ANOVA Test for reduced model (Age)

```
reduced_age <- lm(log_salary ~ PTS + MP + stocks + awards_1 + awards_2plus + Pos_PF + Pos_PG + Pos_SF +
anova(reduced_age, quant_and_qual)
```

```
## Analysis of Variance Table
##
## Model 1: log_salary ~ PTS + MP + stocks + awards_1 + awards_2plus + Pos_PF +
##     Pos_PG + Pos_SF + Pos_SG
## Model 2: log_salary ~ PTS + MP + stocks + awards_1 + awards_2plus + Age_22_26 +
##     Age_27_31 + Age_32_34 + Age_35_plus + Pos_PF + Pos_PG + Pos_SF +
##     Pos_SG
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    174 51.818
## 2    171 47.824  3    3.9943 4.7607 0.003256 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA Test for reduced model (Pos)

```
reduced_pos <- lm(log_salary ~ PTS + MP + stocks + awards_1 + awards_2plus + Age_22_26 + Age_27_31 + Ag
anova(reduced_pos, quant_and_qual)
```

```
## Analysis of Variance Table
```

```
##
## Model 1: log_salary ~ PTS + MP + stocks + awards_1 + awards_2plus + Age_22_26 +
##     Age_27_31 + Age_32_34 + Age_35_plus
## Model 2: log_salary ~ PTS + MP + stocks + awards_1 + awards_2plus + Age_22_26 +
##     Age_27_31 + Age_32_34 + Age_35_plus + Pos_PF + Pos_PG + Pos_SF +
##     Pos_SG
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1    175 49.660
## 2    171 47.824  4    1.8358 1.641 0.1661
```

Final Model

```r
final <- lm(log_salary ~ PTS + MP + stocks + Age_22_26 + Age_27_31 + Age_32_34 + Age_35_plus, data = df)
summary(final)
```

```
##
## Call:
## lm(formula = log_salary ~ PTS + MP + stocks + Age_22_26 + Age_27_31 +
##     Age_32_34 + Age_35_plus, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.50724 -0.35409  0.07407  0.38454  0.91928
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.10079    0.35792   0.282   0.7786
## PTS          0.08228    0.01071   7.682 1.03e-12 ***
## MP           0.03258    0.01425   2.287   0.0234 *
## stocks       0.23531    0.07425   3.169   0.0018 **
## Age_22_26    0.08700    0.18681   0.466   0.6420
## Age_27_31    0.39501    0.19132   2.065   0.0404 *
## Age_32_34    0.43997    0.22416   1.963   0.0512 .
## Age_35_plus       NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5301 on 177 degrees of freedom
## Multiple R-squared:  0.6345, Adjusted R-squared:  0.6221
## F-statistic: 51.21 on 6 and 177 DF,  p-value: < 2.2e-16
```

Confidence Intervals

```r
conf <- confint(final, level = 0.95)
conf
```

```
##                    2.5 %     97.5 %
## (Intercept) -0.605546080 0.80711856
## PTS          0.061143234 0.10341931
## MP           0.004463391 0.06069981
## stocks       0.088792918 0.38183369
## Age_22_26   -0.281667367 0.45567479
## Age_27_31    0.017457991 0.77256498
```

```
## Age_32_34   -0.002396352 0.88232976
## Age_35_plus           NA         NA
```

K-Fold Cross Validation

```r
cv_model <- train(
  log_salary ~ PTS + MP + stocks + Age_22_26 + Age_27_31 + Age_32_34 + Age_35_plus, data = df,
  method = "lm",
  trControl = trainControl(method = "cv", number = 5)
)

# display results
print(cv_model)
```

```
## Linear Regression
##
## 184 samples
##   7 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 147, 148, 147, 147, 147
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   0.533174  0.6164662  0.4342761
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

Assumption Plots

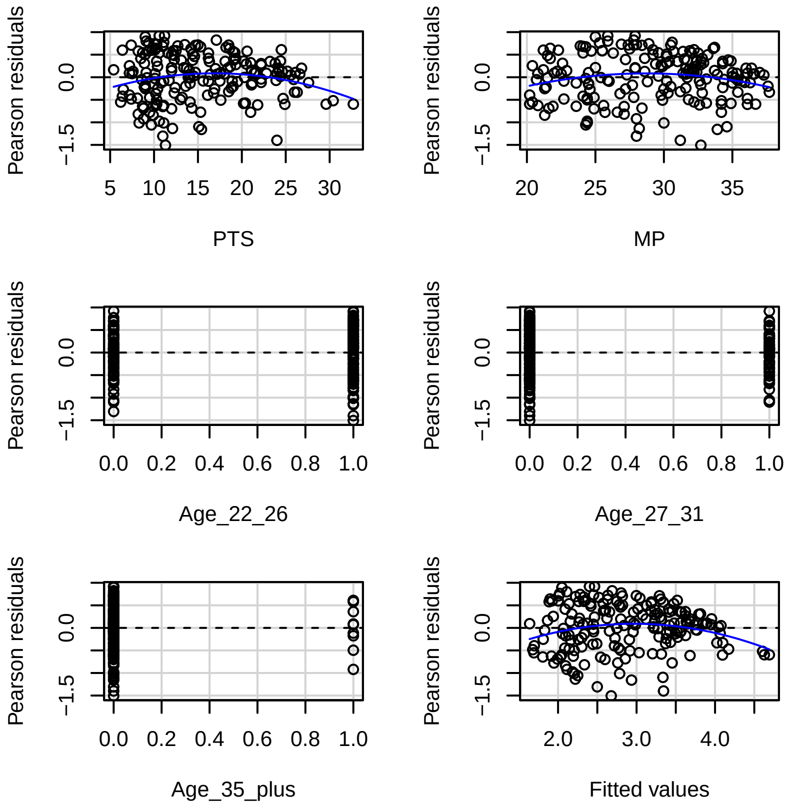
```r
#store residuals from the model
finres<-residuals(final)
sum(finres)
```

```
## [1] 2.418205e-15
```

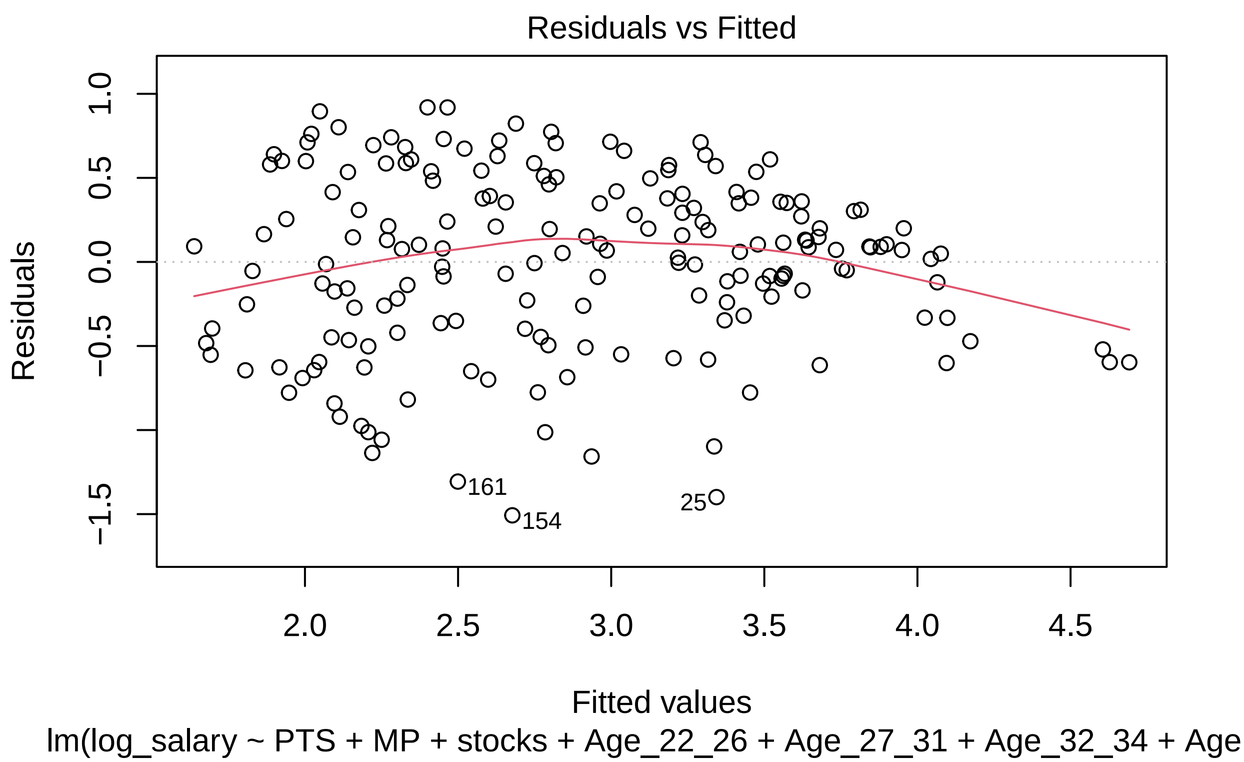
```r
mean(finres)
```

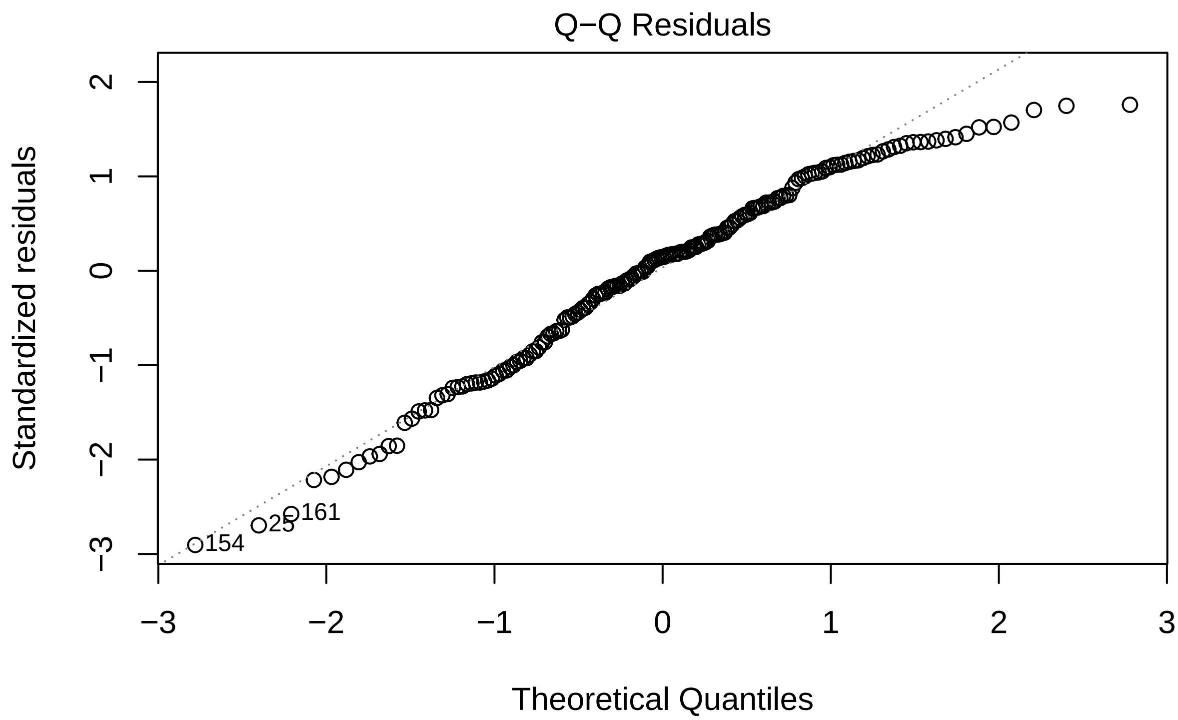
```
## [1] 1.314536e-17
```

```r
#Residuals Plots of explanatory variables vs residuals
residualPlots(final,tests=F)
```



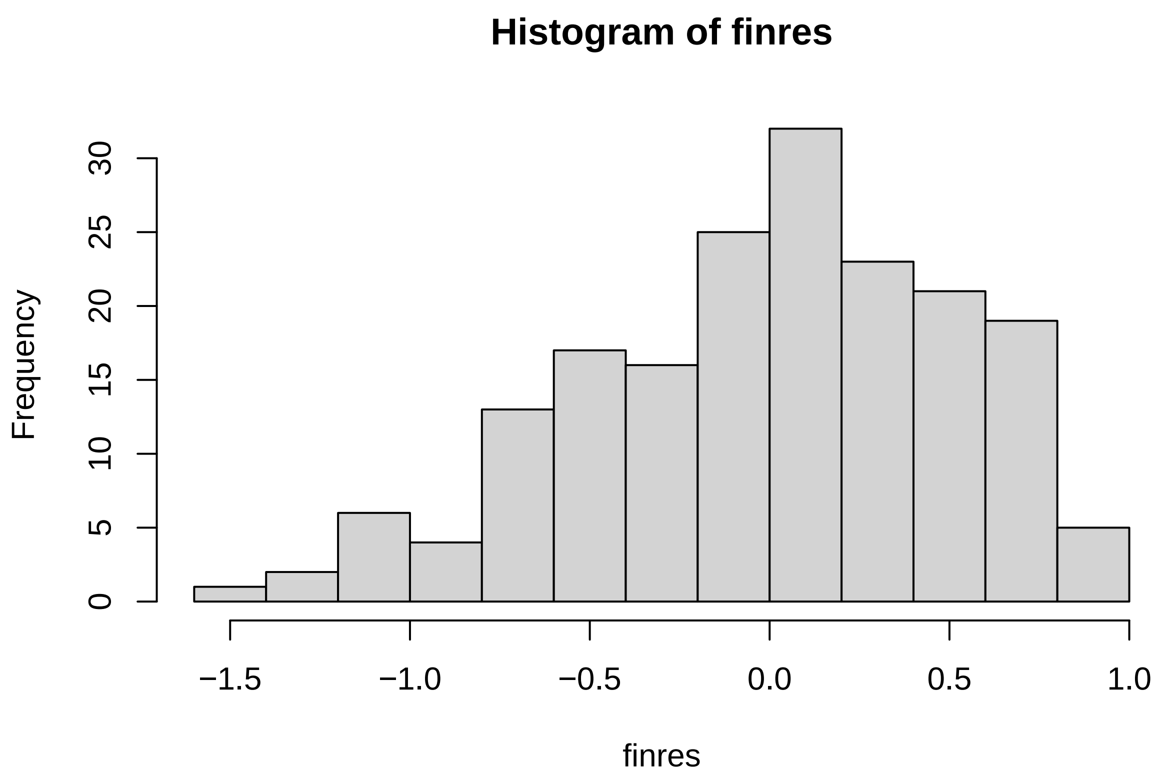

```r
#Residual vs Fitted and QQ plot
plot(final, which=c(1,2))
```





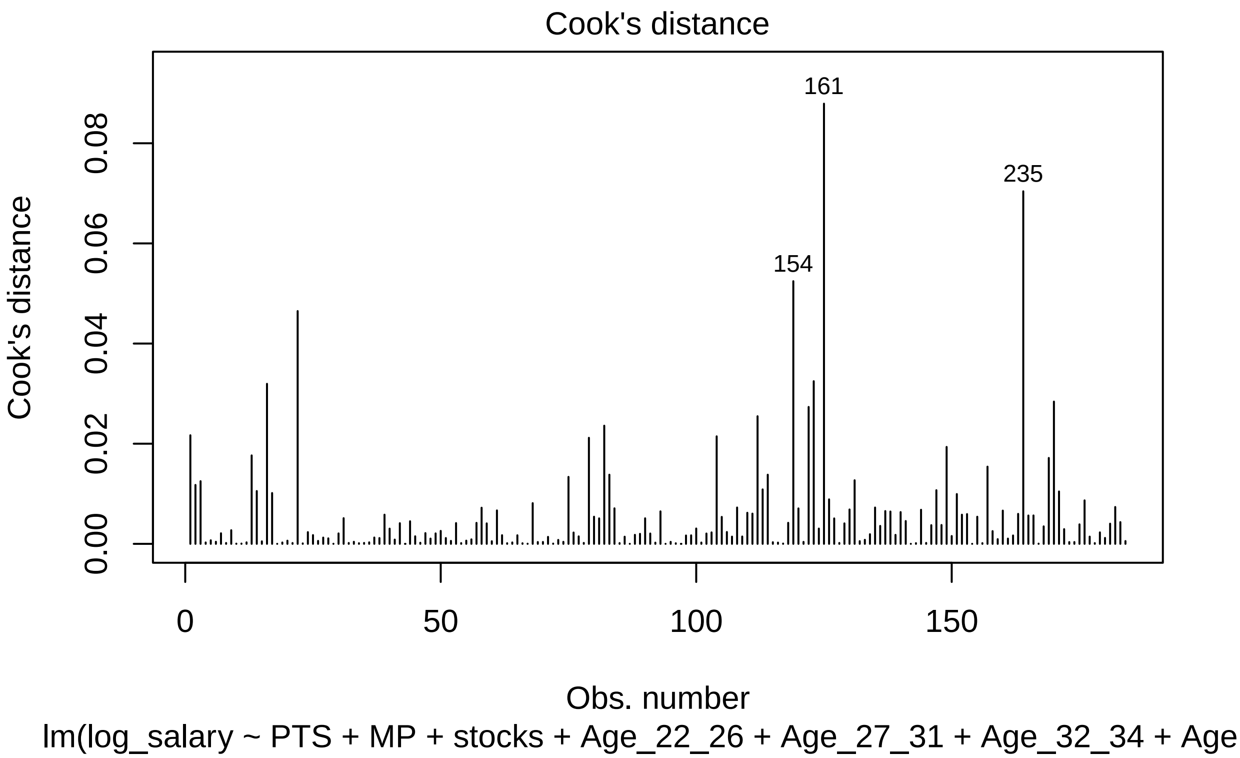
lm(log_salary ~ PTS + MP + stocks + Age_22_26 + Age_27_31 + Age_32_34 + Age

```
#histogram of residuals
hist(finres)
```

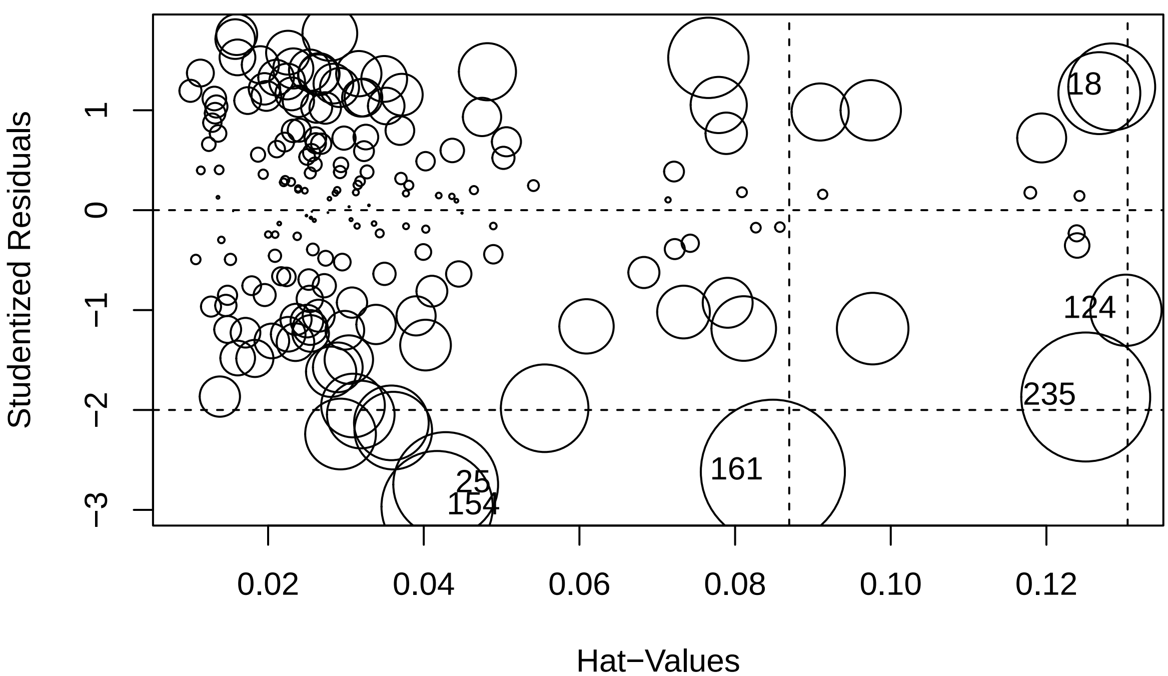


Residual Analysis

```r
# Cooks Distance Thresholds
plot(final,which=4)
```



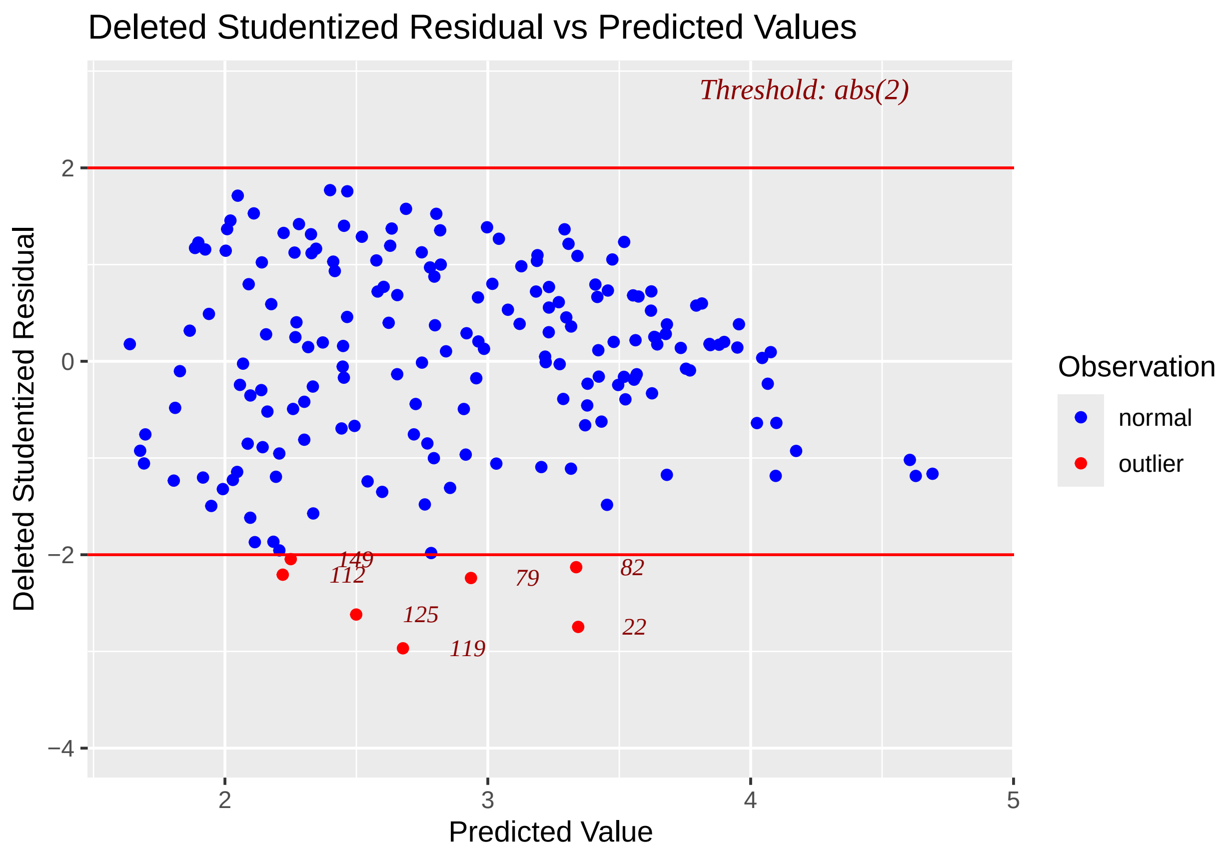

```r
# Leverage vs Studentized Residuals
influencePlot(final,fill=F)
```




```
##         StudRes         Hat       CookD
```

```
## 18    1.234125 0.12839271 0.03195641
## 25   -2.746409 0.04281875 0.04648451
## 124 -1.001901 0.13021793 0.02146854
## 154 -2.967628 0.04171297 0.05245075
## 161 -2.618442 0.08485551 0.08791072
## 235 -1.870001 0.12505150 0.07040578
```

```
# Deleted Studentized Residuals vs Predicted values
ols_plot_resid_stud_fit(final)
```



remove obs 86 and 154 (work in progress)

```
df <- df[-c(86, 154), ]
```