Quiz 4

Drew Davison

2025-04-23

Context

In 2012, the Colorado Rockies launched **Project 5183**, a novel strategy to improve pitcher health and performance by limiting starters to 75 pitches per outing. This assignment explores that experiment using generalized linear models (GLMs) to determine the impact on pitcher performance. You'll analyze real data collected from FanGraphs and stored in `FinalRockiesdata.csv`.

Objectives

- Use GLMs to model pitcher performance under a pitch limit.
- Interpret and communicate the statistical significance of the pitch count limit.
- Develop understanding of model assumptions and diagnostics.

Data Variables

- `PCL`: 1 if pitch count limit in effect, 0 otherwise
- `ERA`: Earned runs per nine innings
- `K/9`: Strikeouts per nine innings
- `vFA`: Average fastball velocity
- `Pitpct`: Percent of pitches that were strikes
- `Age`: Age of pitcher
- `Coors`: Was the game played at Coors field?

Instructions

1. Load the data into R and ensure variables are appropriately typed.
2. Conduct a descriptive summary comparing pre- and post-limit performances.
3. Fit a GLM for each of: ERA, K/9, vFA, Pitpct, using PCL, Age, and Coors as predictors.
4. Interpret PCL's coefficient in each model.
5. Plot diagnostics and comment on model validity.
6. Reflect on whether Project 5183 was effective.

Rubric

Criteria	Excellent (4)	Good (3)	Fair (2)	Poor (1)
Clarity	Precise answers	Mostly clear	Some confusion	Vague
Correctness	Fully correct	Minor errors	Multiple errors	Major errors
Class Connections	Strong links to GLM topics	Some links	Weak	None
Reproducibility	Runs smoothly, clear code	Mostly reproducible	Needs fixing	Not reproducible
Creativity	Reflective, insightful	Some depth	Minimal insight	No insight

Solutions

Load Packages and Data

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
rockies <- read_csv("~/STA310/Quiz 4/Quiz 4 Final/STA310/Data/FinalRockiesdata.csv", show_col_types
                    = FALSE)
rockies$PCL <- as.factor(rockies$PCL)
rockies$Coors <- as.factor(rockies$Coors)
```

Descriptive Summary

```
rockies %>%
  group_by(PCL) %>%
 summarize(across(c(ERA, "K/9", vFA, Pitpct), mean))
```

```
## # A tibble: 2 x 5
##   PCL     ERA `K/9`   vFA Pitpct
##   <fct> <dbl> <dbl> <dbl>  <dbl>
## 1 n      7.28  6.82  91.5  0.605
## 2 y      5.68  6.40  89.4  0.617
```

Interpretation: This summary highlights pre- and post-limit average performance. There are expected decreases in K/9 and ERA for pitchers with the pitch count in effect.

Model 1: ERA ~ PCL + Age + Coors

```
model_era <- glm(ERA ~ PCL + Age + Coors, data = rockies)
summary(model_era)
```

```
##
## Call:
## glm(formula = ERA ~ PCL + Age + Coors, data = rockies)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.1097     3.6203   0.307   0.7598
## PCLy         -1.4354     0.9558  -1.502   0.1359
## Age           0.1867     0.1290   1.448   0.1503
## Coorsy        2.1356     0.9193   2.323   0.0219 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 24.85215)
##
##     Null deviance: 3088.8  on 117  degrees of freedom
## Residual deviance: 2833.1  on 114  degrees of freedom
## AIC: 719.93
##
## Number of Fisher Scoring iterations: 2
```

Interpretation: A negative PCL coefficient suggests improved run prevention with the limit, thought it is not statistically significant with a p-value of .1359.

Model 2: K.9

```
model_k9 <- glm(`K/9` ~ PCL + Age + Coors, data = rockies)
summary(model_k9)
```

```
##
## Call:
## glm(formula = `K/9` ~ PCL + Age + Coors, data = rockies)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.15482    2.40470   4.639 9.42e-06 ***
## PCLy        -0.49326    0.63485  -0.777   0.4388
## Age         -0.15511    0.08565  -1.811   0.0728 .
## Coorsy      -0.22585    0.61058  -0.370   0.7121
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 10.96446)
##
##     Null deviance: 1292.2  on 117  degrees of freedom
## Residual deviance: 1249.9  on 114  degrees of freedom
```

```
## AIC: 623.37
##
## Number of Fisher Scoring iterations: 2
```

Interpretation: A drop in strikeouts would support the pitch-to-contact theory, where pitchers de-prioritize strikeouts. However, the coefficient again is not statistically significant.

Model 3: vFA

```
model_vfa <- glm(vFA ~ PCL + Age + Coors, data = rockies)
summary(model_vfa)
```

```
##
## Call:
## glm(formula = vFA ~ PCL + Age + Coors, data = rockies)
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 100.71612    1.81227  55.574  < 2e-16 ***
## PCLy         -2.24202    0.47845  -4.686 7.76e-06 ***
## Age          -0.33492    0.06455  -5.188 9.33e-07 ***
## Coorsy       -0.13073    0.46016  -0.284    0.777
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 6.227516)
##
##     Null deviance: 999.30  on 117  degrees of freedom
## Residual deviance: 709.94  on 114  degrees of freedom
## AIC: 556.62
##
## Number of Fisher Scoring iterations: 2
```

Interpretation: An decrease in velocity, with a statistically significant p-value, might imply pitchers are throwing softer per pitch due to de-prioritizing strikeouts with the limit. It hurts the theory that pitchers are throwing faster due to the limit.

Model 4: Pitpct

```
model_pitpct <- glm(Pitpct ~ PCL + Age + Coors, data = rockies)
summary(model_pitpct)
```

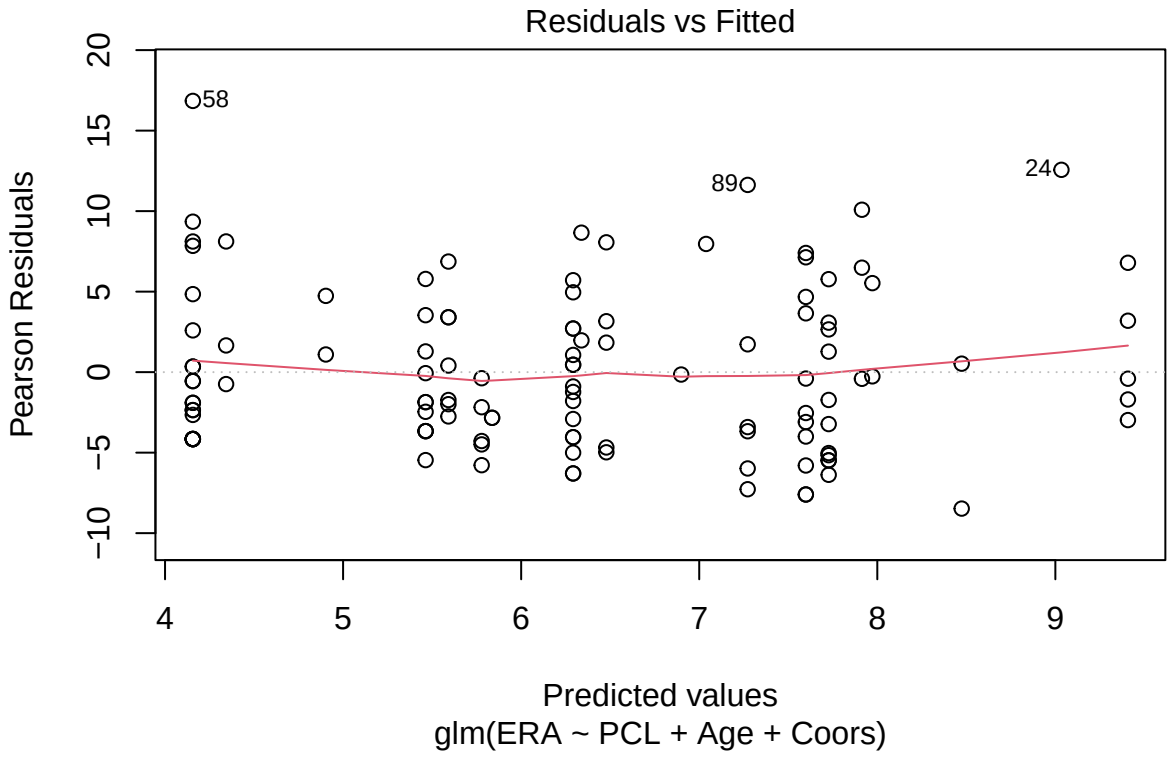
```
##
## Call:
## glm(formula = Pitpct ~ PCL + Age + Coors, data = rockies)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.546114   0.038841  14.060   <2e-16 ***
## PCLy        0.013187   0.010254   1.286   0.2010
```

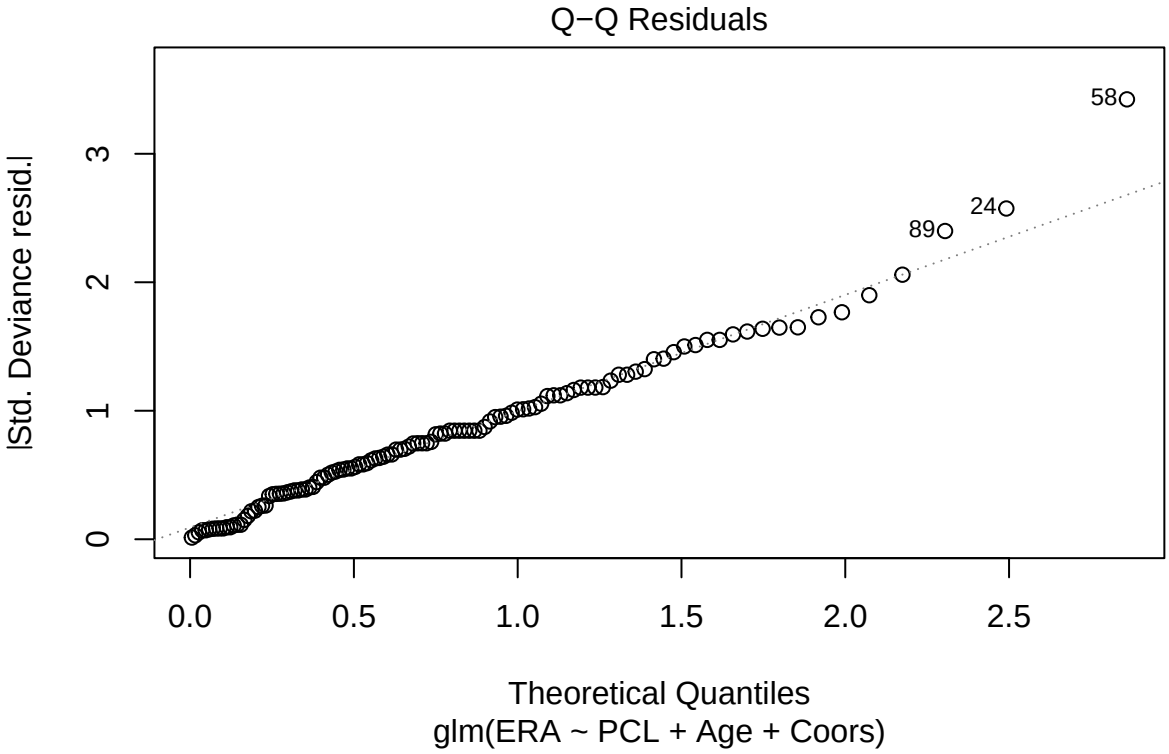
```
## Age         0.001794   0.001383   1.296   0.1975
## Coorsy      0.019652   0.009862   1.993   0.0487 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.00286048)
##
##     Null deviance: 0.34587  on 117  degrees of freedom
## Residual deviance: 0.32609  on 114  degrees of freedom
## AIC: -350.3
##
## Number of Fisher Scoring iterations: 2
```

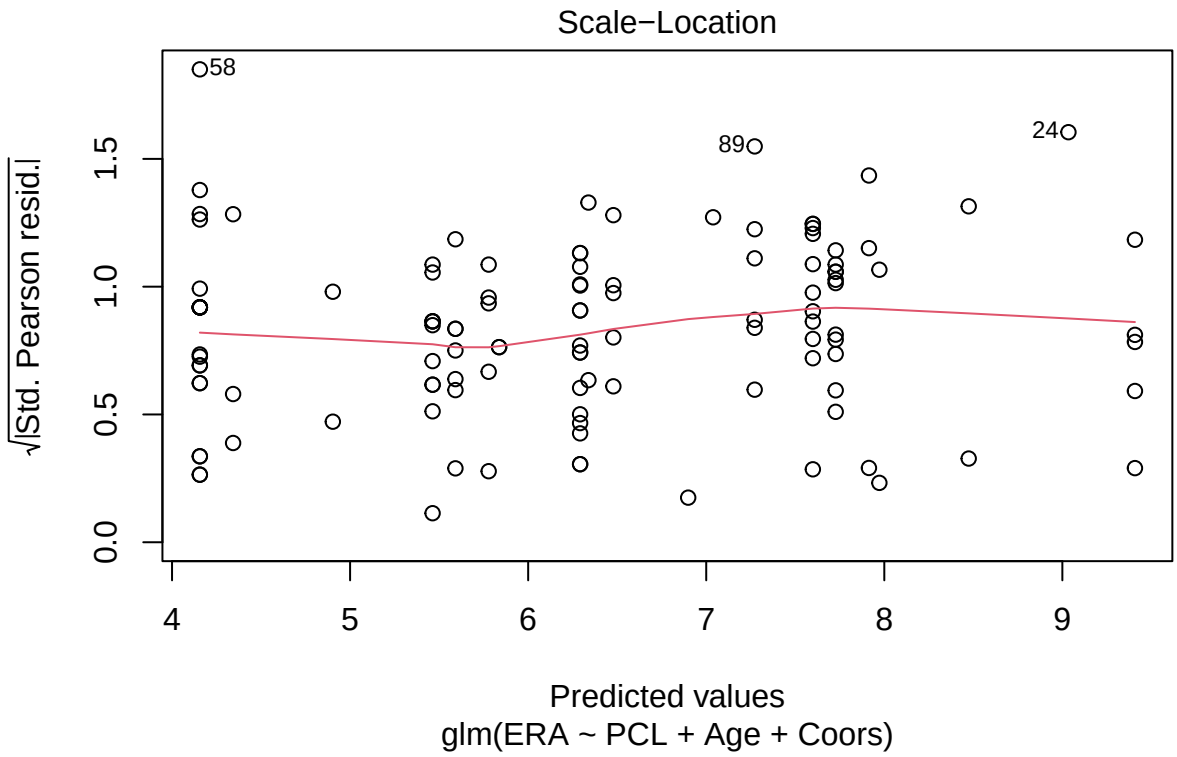
Interpretation: A rise in strike percentage would indicate better command, but the coefficient is relatively small, and it is not statistically significant.

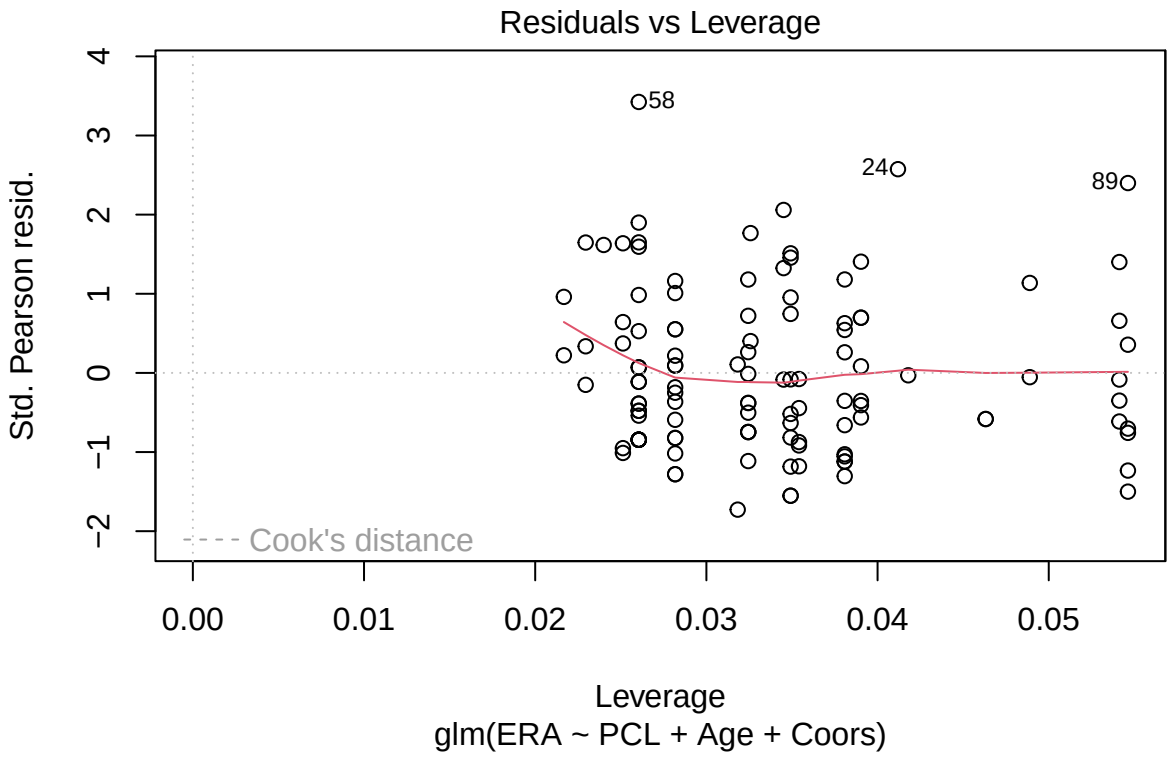
Model Diagnostics

```
plot(model_era)
```




Q−Q Residuals
|Std. Deviance resid.|
0
1
2
3
58
24
89
0.0
0.5
1.0
1.5
2.0
2.5
Theoretical Quantiles
glm(ERA ~ PCL + Age + Coors)


Scale−Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
58
89
24
4
5
6
7
8
9
Predicted values
glm(ERA ~ PCL + Age + Coors)



Diagnostics: Residuals should be roughly normally distributed. Check for heteroscedasticity.

Reflection

Was the pitch count limit effective? Did ERA drop? Was contact pitching successful? Consider tradeoffs.

There was little statistical significance to the coefficients that suggest the limit was effective, using measures like ERA. However, there is statistical significance to the fact the the pitchers *did* change their pitching strategy. There is more data needed to confirm if the change is strategy was successful in limiting runs.

Optional Extension: Try logistic regression if categorizing `good'' vs` bad'' outings.