Balancing Performance and Longevity: Modeling Injury Risk in MLB Pitchers

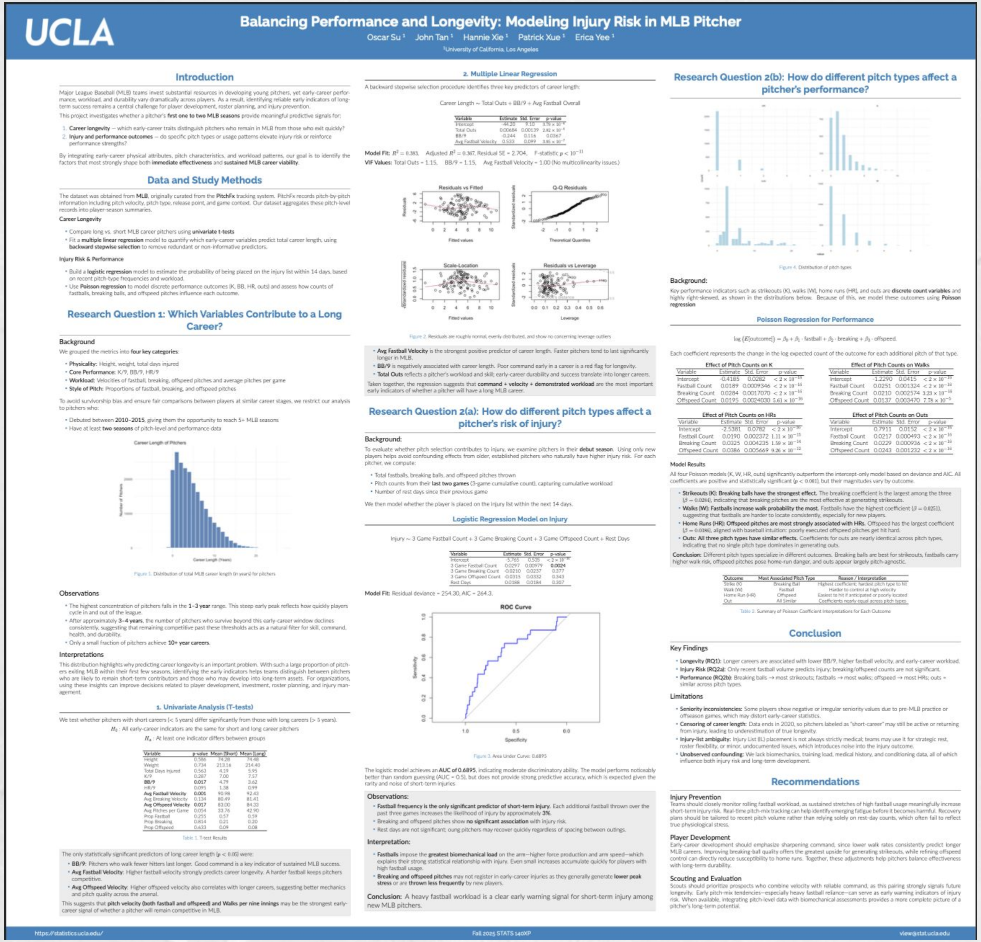
STATS140XP
Group 5: Oscar Su, John Tan, Hannie Xie, Patrick Xue, Erica Yee

Table of contents

01

Introduction

Project Poster

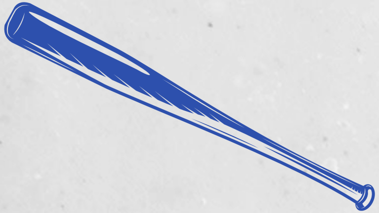


Motivation

Performance ↔ health: Research shows higher pitching velocity and heavier workloads are strongly associated with higher injury risk in MLB pitchers.

Increased velocity raises stress on the shoulder/elbow → **UCL injuries** that may require Tommy John surgery (Krause).

Why it matters: Recovery commonly keeps pitchers out for ~17 months, which can stall development for younger players and end careers for veterans



Source: Krause, Matthew. “Matthew Krause.” Professional Baseball Strength, 5 Jan. 2024, pbsccs.org/the-effects-of-high-velocity-on-players-health-safety-and-performance/.

Research Questions

Question 1: Which variables contribute the most to a long career length?

Question 2: Are there specific pitches that lead to

(a) Higher chance of injury

(b) Better performance

02

RQ1: Which variables contribute the most to a long career length?

Variables

Type	Metric	Description
Physicality	Height	Height of player in inches. We take the first recorded height when the player enters the dataset (assume no drastic change in height)
	Weight	Weight of player in lbs. We average their weight across the two seasons.
	Total_Days_Injured	Total cumulative number of days injured
Core Performance	K_per_9	Number of strikeouts per 9 innings (a measure of performance for a pitcher; outs are good)
	BB_per_9	Number of walks per 9 innings (a measure of blunders for a pitcher; walks are bad and load the bases)
	HR_per_9	Number of home runs given per 9 innings (a measure of blunders for a pitcher; home runs are really bad)

Variables

Workload	Avg_Fastball_Overall	Average speed of a fastball over their first 2 seasons
	Avg_Breaking_Overall	Average speed of a breaking ball over their first 2 seasons
	Avg_Offspeed_Overall	Average speed of an offspeed ball over their first 2 seasons
	Avg_Pitches_Per_Game	Average number of pitches per game over their first 2 seasons
Style of Pitch	Prop_Fastball	Average proportion of fastballs thrown by a player over their first 2 seasons
	Prop_Breaking	Average proportion of breaking balls thrown by a player over their first 2 seasons
	Prop_Offspeed	Average proportion of offspeed balls thrown by a player over their first 2 seasons

RQ1 Methods: t-tests + stepwise linear regression

- Sample: MLB pitchers who debuted 2010–2015
- Predictors window: summarize each pitcher’s first 2 seasons (early-career metrics + pitch mix)
- Outcome label: career length > 5 years = long, otherwise short (to reduce survivorship bias)

Univariate t-tests → **Multiple Linear Regression** → **Backward Stepwise Selection**

RQ1 t-test Results

	Variable <chr>	P_Value <dbl>	Mean_Short_Career <dbl>	Mean_Long_Career <dbl>	Significant <chr>
1	Height	0.58603	74.279	74.476	NO
2	Weight	0.73358	213.162	214.397	NO
3	Total_Days_Injured	0.56250	4.191	5.952	NO
4	K_per_9	0.28747	7.001	7.566	NO
5	BB_per_9	0.01740	4.785	3.624	YES
6	HR_per_9	0.09470	1.381	0.990	NO
7	Avg_Fastball_Overall	0.00095	90.983	92.426	YES
8	Avg_Breaking_Overall	0.13390	80.489	81.409	NO
9	Avg_Offspeed_Overall	0.01660	83.003	84.334	YES
10	Avg_Pitches_Per_Game	0.05437	33.760	42.903	NO
11	Prop_Fastball	0.25529	0.570	0.590	NO
12	Prop_Breaking	0.81394	0.205	0.201	NO
13	Prop_Offspeed	0.63333	0.089	0.082	NO

- **BB_per_9:** Players who give lower walks per 9 on average have longer careers.

- **Avg_Fastball_Overall**: Important to have a faster than average fastball.

- **Avg_Offspeed_Overall:** Also important to have a faster offspeed ball.

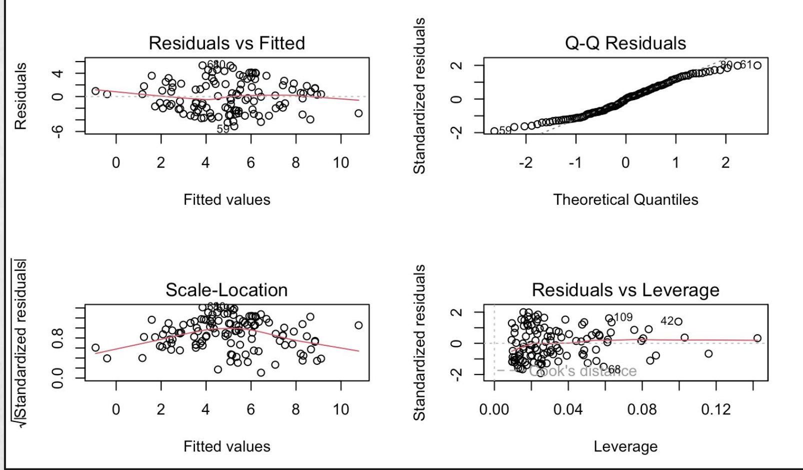
RQ1 MLR and Backward Stepwise Selection

Career Length = -44.2 + 0.00684(Total Outs) - 0.244(BB_per_9) + 0.533(Avg Fastball Velocity)

```
Call:
lm(formula = Career_Length ~ Total_Outs + BB_per_9 + Avg_Fastball_Overall,
    data = model_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.1094 -2.2065  0.0238  2.1264  5.3589 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)         -44.200341   9.096232  -4.859 3.79e-06 ***
Total_Outs            0.006840   0.001387   4.930 2.82e-06 ***
BB_per_9             -0.244467   0.115651  -2.114   0.0367 *  
Avg_Fastball_Overall  0.533436   0.099077   5.384 3.95e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.704 on 114 degrees of freedom
Multiple R-squared:  0.3829,	Adjusted R-squared:  0.3667 
F-statistic: 23.58 on 3 and 114 DF,  p-value: 6.036e-12
```

```
vif(backward_model)
```

```
          Total_Outs             BB_per_9 Avg_Fastball_Overall 
            1.149446             1.149504             1.000127 
```



No multicollinearity, OK diagnostic plots

RQ1 Interpretation: Key early-career predictors of longevity (debut 2010–2015)

Command matters: Long-career pitchers issue fewer walks (BB/9) early on.

Velocity matters: Long-career pitchers have higher average fastball velocity (and offspeed velocity shows up in group differences too).

Early dominance matters: In the final OLS model, the strongest predictors retained are Total Outs (+), BB/9 (-), Avg Fastball Velocity (+)

Coaches should prioritize developing and investing in young pitchers who, early in their careers, record more outs, issue fewer walks, and sustain higher fastball velocity, since these factors are associated with greater long-term value and a higher likelihood of career longevity.

03

RQ2: How do different pitch types affect pitcher's
(a) risk of injury
(b) performance?

RQ2(a) How do different pitch types affect a pitcher’s risk of injury?

- Focus on new players in their **debut** season
- Predictors: **3-game cumulative frequency** of different **pitch types** (fastball / breaking / offspeed) and **number of rest days**
- Outcome: **potential injuries** (placed on the injury list within the next 14 days)

Logistic Regression
On Injury

RQ2(a) Analysis: Logistic Regression on Injury

```
Call:
glm(formula = fml, family = "binomial", data = new_players_pitch_count)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -5.764673   0.535299 -10.769  < 2e-16 ***
roll3_fastball   0.029684   0.009787   3.033  0.00242 **
roll3_breaking  -0.020956   0.023737  -0.883  0.37731
roll3_offspeed  -0.031495   0.033183  -0.949  0.34255
rest_days        0.018842   0.018428   1.022  0.30655
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 265.19  on 2709  degrees of freedom
Residual deviance: 254.30  on 2705  degrees of freedom
AIC: 264.3

Number of Fisher Scoring iterations: 8
```

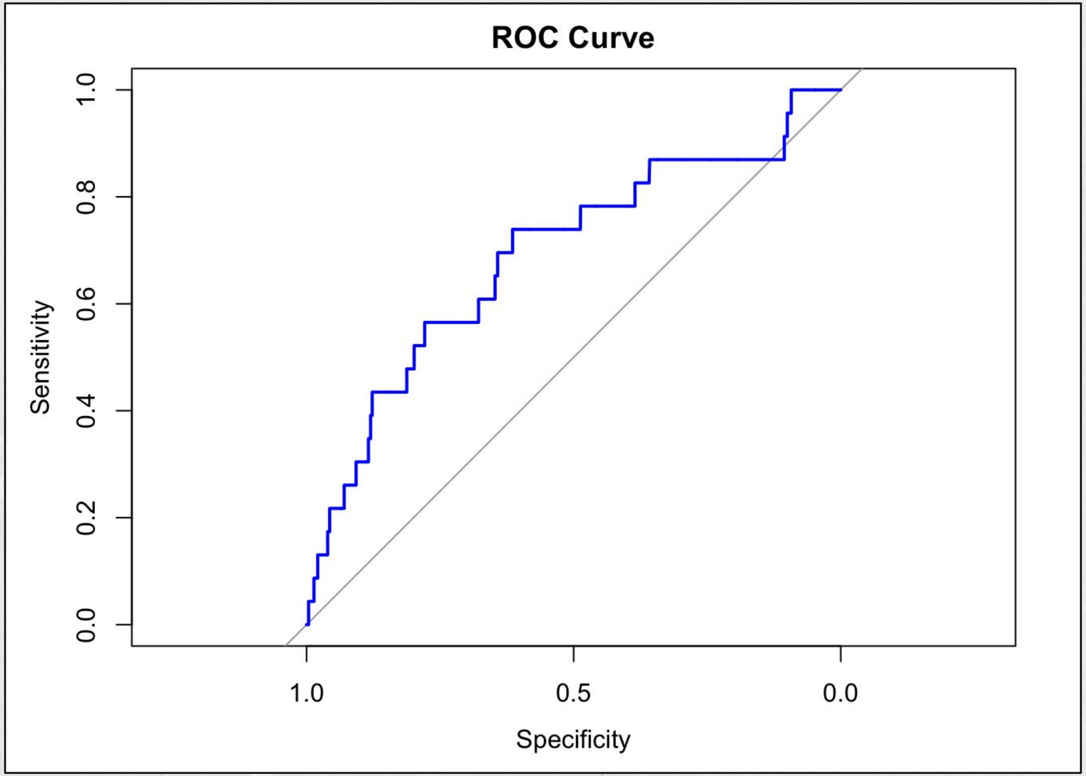
roll3_fastball_count is positive and significant ($\beta \approx 0.0297$, $p \approx 0.0024$)

RQ2(a) Analysis Interpretation

Odds Ratio = exp(0.0297) ≈ **1.03**

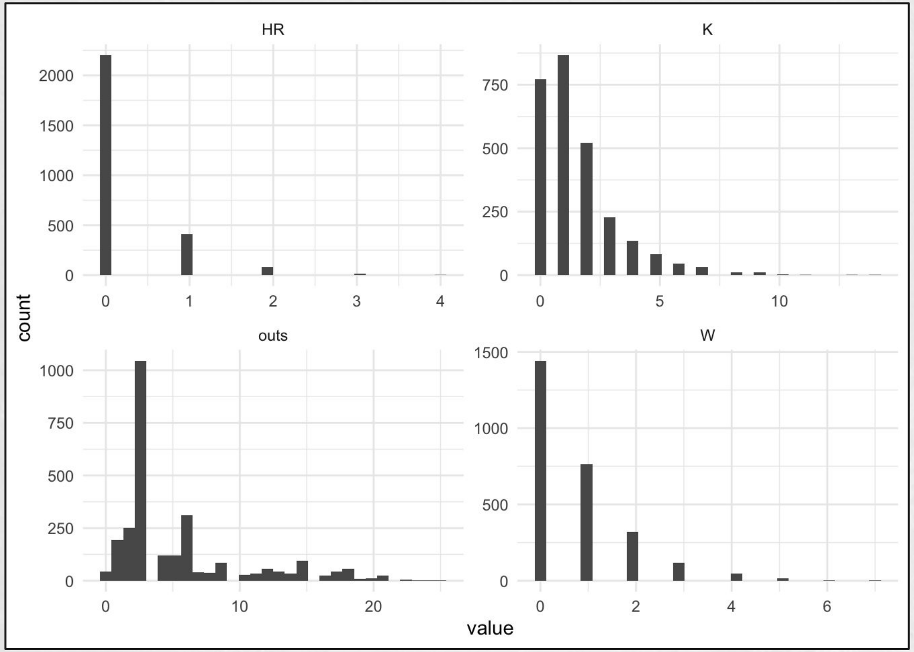
- Odds interpretation: each extra fastball in the last 3 games → ~3% higher odds of injury
- Breaking/offspeed: not significant in this short-term window
- Rest days: not significant (may be too coarse / young pitchers recover quickly)

RQ2(a) Analysis #1: Logistic Regression Model Quality



AUC ≈ 0.689

RQ2(b) How do different pitch types affect a pitcher's performance?



Why Poisson: KPIs (outs, walks, HR, strikeouts) are discrete + right-skewed

Model form: log(E[outcome]) = β0 + β1·fastball + β2·breaking + β3·offspeed

larger coefficient ⇒ stronger association with that outcome

Poisson Regression on Performance

RQ2(b) Analysis #2: Poisson Regression on Performance

Effect of Pitch Counts on K

Variable	Estimate	Std. Error	p-value
Intercept	-0.4185	0.0282	$< 2 \times 10^{-16}$
Fastball Count	0.0189	0.0009346	$< 2 \times 10^{-16}$
Breaking Count	0.0284	0.0017070	$< 2 \times 10^{-16}$
Offspeed Count	0.0195	0.0024030	5.61×10^{-16}

Effect of Pitch Counts on Walks

Variable	Estimate	Std. Error	p-value
Intercept	-1.2290	0.0415	$< 2 \times 10^{-16}$
Fastball Count	0.0251	0.001324	$< 2 \times 10^{-16}$
Breaking Count	0.0210	0.002574	3.23×10^{-16}
Offspeed Count	0.0137	0.003470	7.78×10^{-5}

Effect of Pitch Counts on HRs

Variable	Estimate	Std. Error	p-value
Intercept	-2.5381	0.0782	$< 2 \times 10^{-16}$
Fastball Count	0.0190	0.002372	1.11×10^{-15}
Breaking Count	0.0325	0.004235	1.59×10^{-14}
Offspeed Count	0.0386	0.005669	9.26×10^{-12}

Effect of Pitch Counts on Outs

Variable	Estimate	Std. Error	p-value
Intercept	0.7911	0.0152	$< 2 \times 10^{-16}$
Fastball Count	0.0217	0.000493	$< 2 \times 10^{-16}$
Breaking Count	0.0229	0.000936	$< 2 \times 10^{-16}$
Offspeed Count	0.0243	0.001232	$< 2 \times 10^{-16}$

All coefficients are positive and statistically significant, but their magnitudes vary by outcome

RQ2(b) Analysis #2: Interpretation

Outcome	Pitch type with strongest association	Direction
Strikeouts (K)	Breaking	↑ more breaking → ↑ expected Ks
Walks (W)	Fastball	↑ more fastballs → ↑ expected BBs (walks)
Home Runs (HR)	Offspeed	↑ more offspeed → ↑ expected HRs
Outs	About the same across all	similar association across pitch types

RQ2 Analysis #2: Poisson Regression on Performance

```
summary(final_model_results$K$Best_Model_Object)


Call:
glm(formula = fml, family = poisson, data = data)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.4184550  0.0282053 -14.836  < 2e-16 ***
fastball     0.0188992  0.0009346  20.221  < 2e-16 ***
breaking     0.0284338  0.0017070  16.657  < 2e-16 ***
offspeed     0.0194577  0.0024030   8.097 5.61e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 4753.1  on 27
Residual deviance: 2765.9  on 27
AIC: 7675.5

Number of Fisher Scoring iterati
```

```
summary(final_model_results$W$Best_Model_Object)


Call:
glm(formula = fml, family = poisson, data = data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.229033   0.041485 -29.626  < 2e-16 ***
fastball     0.025095   0.001324  18.955  < 2e-16 ***
breaking     0.021017   0.002574   8.164 3.23e-16 ***
offspeed     0.013711   0.003470   3.951 7.78e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3675.3  on 2709  degrees
Residual deviance: 2582.6  on 2706  degrees
AIC: 5526.6

Number of Fisher Scoring iterations: 5
```

```
summary(final_model_results$outs$Best_Model_Object)


Call:
glm(formula = fml, family = poisson, data = data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.7911495  0.0152265   51.96   <2e-16 ***
fastball    0.0217320  0.0004930   44.09   <2e-16 ***
breaking    0.0228794  0.0009356   24.45   <2e-16 ***
offspeed    0.0243477  0.0012328   19.75   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 9543.7  on 2709  degrees of freedom
Residual deviance: 1693.1  on 2706  degrees of freedom
AIC: 10574

Number of Fisher Scoring iterations: 4
```

```
summary(final_model_results$HR$Best_Model_Object)


Call:
glm(formula = fml, family = poisson, data = data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.538112   0.078153 -32.476  < 2e-16 ***
fastball     0.019012   0.002372   8.014 1.11e-15 ***
breaking     0.032527   0.004235   7.680 1.59e-14 ***
offspeed     0.038646   0.005669   6.817 9.26e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2177.3  on 2709  degrees of freedom
Residual deviance: 1757.7  on 2706  degrees of freedom
AIC: 2847.1

Number of Fisher Scoring iterations: 6
```

04

Conclusion

Limitations

Data quirks:

- Some players have negative seniority years (not official MLB debut; practice/offseason games), which may inflate performance.

Incomplete window:

- Missing games, and the dataset cuts off at 2020 → careers may continue past the data (or players may disappear due to breaks/injury).

Injury label isn't purely medical:

- "Injury list" can reflect roster strategy (resting, demotion), not just true injury incidence.

Unobserved confounding:

- Lack player-level health/lifestyle/conditioning info that could affect both injuries and longevity.

Recommendations

Injury Prevention

- Closely monitor rolling fastball workload
- Recovery plans should be tailored to recent pitch volume rather than relying solely on rest-day counts

Player Development

- Emphasize sharpening command as lower walk rates consistently predict longer careers
- Improve breaking-ball quality → greatest upside for generating strikeouts, refining offspeed pitches → reduce susceptibility to home runs
- Help pitchers balance effectiveness with long-term durability.

Scouting and Evaluation

- Prioritize identifying prospects with strong fastball quality as this pitch is the foundation of pitching success

Thanks!