The Element of Surprise: An Instrumental Variables Approach for Identifying Usage Curves in Professional Football

Daniel Putman* Tor Tolhurst†

This Draft: November 26, 2022

Abstract

What is the optimal rate of passing in professional football? We derive a simple model of play efficiency where efficiency is highest when the play comes as a surprise and falls when a more expected play is called. That is, playcalling efficiency is determined by a usage curve. We propose using fumbles lost as an instrumental variable to estimate the effect of passing, expected passing, and their interaction on the per play efficiency on early downs. More specifically, we argue that conditional on total cumulative fumbles by each team, cumulative fumbles lost serves as a valid instrument in each of these cases. Using fifteen years of play-by-play data, we use this strategy to estimate linear usage curves and are able to recover the the optimal passing rate on early downs. On first and second down, coaches have their quarterbacks drop back to pass on about 52% of the time, whereas the optimal rate is around 71%. These results suggest that coaches deviate considerably from the optimal pass rate, overestimating the element of surprise.

Keywords: American Football, Sports Economics, Instrumental Variables, Usage Curves, Strategy, Football Analytics

JEL Codes: L83, Z20, L19, D81

*Postdoctoral Research Fellow, University of Pennsylvania Center for Social Norms and Behavioral Dynamics (putmands@gmail.com)

†Assistant Professor, Purdue University Department of Agricultural Economics

1 Introduction

The passing premium puzzle, introduced in Alamar (2006) is as follows: despite higher returns to passing plays,[1] that an approximately equal number of passing and running plays are run by National Football League (NFL) teams. Over the years, many approaches have been proposed to rationalize the apparent difference in optimal pass rates and the observed rate of passing. These explanations include decisions-maker risk aversion and the role of defensive adjustments in determining pass rate (Rockerbie, 2008; Jordan et al., 2009; McGarrity and Linnen, 2010). The availability of play-by-play data to the public has increased dramatically since the majority of work was done. Open source packages in R as well as other programming languages now allow anyone to pull data directly from league sources. This increased access has allowed for the development of a robust public analytics discussion.[2]

Since 2006, passing has increased dramatically in the NFL, around 4.5 percentage points (figure 1). Absent any increases in passing efficiency (relative to rushing), this might suggest that the passing premium puzzle has been solved. That is, that coaches have become wise to the work of analysts and have responded by passing more. However, given that the efficiency of passing has also improved as compared to rushing, a clear alternative explanation exists: coaches have selected into passing as the returns have increased. Given this large increase in both pass rate and passing efficiency, does the passing premium puzzle persist?

To answer this question, we first need to return to a centrally important question in football analytics: what is the optimal pass rate? Following work which posits defensive adjustments as a primary reason for persistent differences, we use a simple model of per play efficiency which allows the efficiency of passing and running to vary with their respective usage. Estimating the "usage effects" on passing and rushing efficiency, we can determine the optimal pass rate. How-

[1] Measured in yards or more advanced statistics like Expected Points Added.

[2] Additionally, more data is available about what is happening on any given play, though often not to the public. This includes personnel, alignments, passing and coverage concepts, and even more recently, player tracking data.

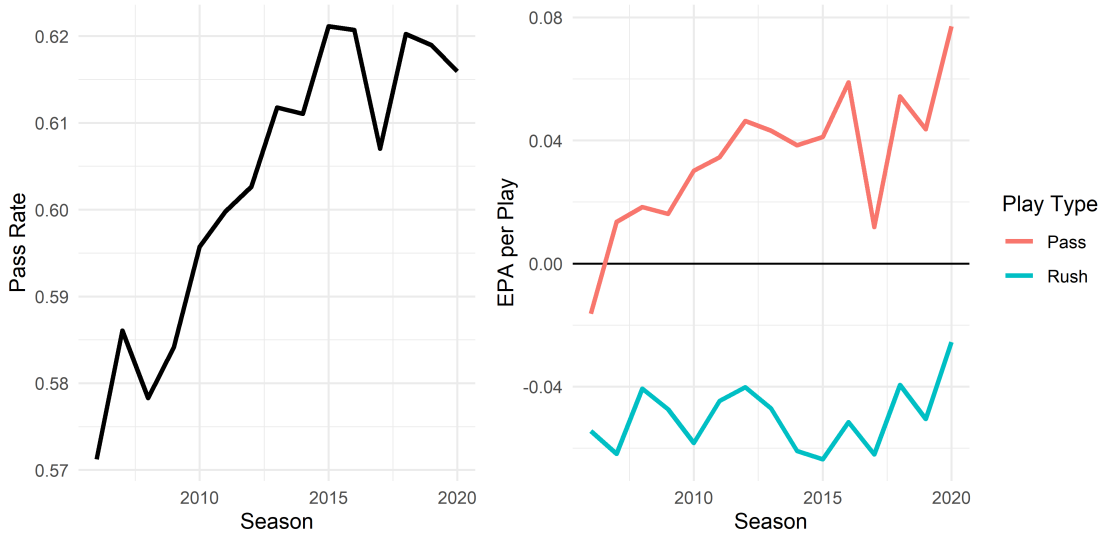

Figure 1: Evolution of NFL Pass rate and per play efficiency over time, 2006-2020

ever, just as selection into passing over the past fifteen years may be driven by improvements in passing offense, estimating usage effects means one must contend with how teams select into passing. For example, coaches with strong passing offense may pass more often than their counterparts, which would bias the estimation of usage effects.

To remove selection bias in estimating the parameters of this model, we introduce a new instrumental variable for passing and expected pass rate. In particular, we argue that conditional on the number of fumbles by a team, fumbles lost is a important and essentially random determinant of game state, measured through win probability or net score. This change in game state forces teams to become more aggressive than they would otherwise would be, passing more often in order to score the points necessary to win the game. Importantly, both the fumbling team and the opposing team realize this necessity, meaning that our instrument forces changes in passing expectations for the defense as well. In this way, our approach mimics the defensive adjustments of a long term increase in passing probability.

We draw on the play-by-play data from the NFL from 2006 to 2020 to investigate these ques-

tions, as well as predictive models of expected passing and expected points have been developed in the public sphere (Yurko et al., 2019; Baldwin, 2021). Using several instruments generated from fumbles lost, we estimate (1) the casual effect of passing, (2) the causal effect of expected pass rate, and (3) the interaction of the two variables on per play efficiency using two-stage least squares, conditional on the number of total fumbles (lost or otherwise). We argue that our instruments and specification fulfill the conditions outlined in Angrist and Krueger (1999) and Blandhol et al. (2022). That is, they can be interpreted as local average treatment effects (LATE).

We find that coaches actually deviate substantially from the optimal usage of passing when usage curves are linear. Using the 2SLS estimates on early downs, when the opponent expects a sure run, passing increasing EPA by 0.48 points/play relative to a rush. However, when passing is expected, EPA actually falls 0.18 points per play. Embedding our instrumental variables results with the model of usage curves returns and optimal pass rate of 71% compared to an actual pass rate of 52%, a considerable deviation. However, this deviation is smaller than would be recovered by naively estimating OLS. We attribute this to the fact that the 2SLS do not suffer from the same issues of selection into passing as the OLS estimation. This result suggests that coaches may overestimate the element of surprise in football strategy, opting to run too often on early downs.

2 A Simple Model of Play Calling Efficiency

The intuition of usage curves is straightforward, the relative efficiency of a given play is a function of how often it is used. More specifically, the greater frequency with which a play is used, the more opponents will prepare for that play and the lower the efficiency of this play. A stylized model of playcalling can formalize this intuition about usage curves. In this model, a play caller can call a run or a pass. At a macro level, the playcaller is looking to maximize efficiency, which is the decision-weighted sum of passing efficiency and rushing efficiency. While passing efficiency tends to be higher for NFL teams, we will allow the efficiency of both passing and rushing to

vary with the decision weights. If usage curves are negatively sloped, e.g., passing efficiency falls when pass rate increases, this might suggest an optimal pass rate between zero and one. [3]

2.1 Optimization Problem

I set up the optimization problem as so:

$$\max_{p_i} \mathrm{E} = p \times \mathrm{PE}(p) + (1-p) \times \mathrm{RE}(p) \tag{1}$$

where E is overall efficiency, RE is rushing efficiency and PE is passing efficiency, p is the pass rate, and $1-p$ is the rush rate. We allow passing efficiency and rushing efficiency to rely on rush rate, and call these efficiencies usage curves.

2.2 General Case

In the general case, the first order condition will be

$$p \times \frac{\partial \mathrm{RE}_i(p)}{\partial p} - \mathrm{RE}(p) + p \times \frac{\partial \mathrm{PE}(p)}{\partial p} + \mathrm{PE}(p) = 0 \tag{2}$$

Without further assumptions we won't be able to find a closed form, explicit solution for p^*. Moreover to do something like the implicit function theorem, we would need to introduce some parameters that should influence the optimal rate of p. Eshewing this, we decide to work within a linear case.

[3] I recognize that this simple and abstract model may be lacking in realism: various aspects of pre-play call situation that impact the optimality of the play called. Largely, I present this model as a model of playcalling all else held equal. Therefore, we can return to this realism in our empirical work. For example, this model might apply to a specific situations, e.g., 1st & 10, or to a number of similar situations controlling for pre-play observables (specifically, down, distance, and field position).

2.3 Linear Case

For simplicity of exposition (and later estimation), we assume affine usage curves, i.e., $\frac{\partial \text{PE}(p)}{\partial p}$ is constant. Written another way: RE $= a + bp$, PE $= c + dp$ where $\frac{\partial \text{PE}(p)}{\partial p} = d$ and $\frac{\partial \text{RE}(p)}{\partial p} = b$. Solving this optimization problem we get to an optimal rush rate,[4]

$$p^* = \frac{1}{2}\left(\frac{a-c}{d-b} - \frac{b}{d-b}\right) \tag{3}$$

How do we interpret this optimal pass rate? We can break it out into two effects:

1. Efficiency Effect: $\frac{1}{2}\left(\frac{a-c}{d-b}\right)$. The ratio of the efficiency of the run game less efficiency of the pass game to the relative decrease in the passing efficiency premium yielded from always passing. This effect tells us that the better the pass game and worse the run game are overall, the more we should pass.

2. Responsiveness Effect: $-\frac{1}{2}\left(\frac{b}{d-b}\right)$. Ratio of the absolute increase in rushing efficiency from always passing to the relative decrease in the passing efficiency premium yielded from always passing. The less the rush game responds to rushing, or the more the pass game responds to passing, the less we should pass.

Finally, are there ever times that we should only pass or only run? Yes. We can characterize corner solutions based on this optimal rush rate.

1. If $d + c \leq c + 2d$, then $p^* = 1$. I.e., it will be maximize the efficiency of the offense to only pass.

[4] Intermediate steps:

$$\begin{aligned}(1-p) \times \frac{\partial RE(p)}{\partial p} - RE + p \times \frac{\partial PE(p)}{\partial p} + PE &= 0 \\ (1-p)b - (a+bp) + pd + (c+dp) &= 0 \\ 2p(d-b) + c + b - a &= 0\end{aligned}$$

2. If $a \leq a - b$, then $p^* = 0$. I.e., it will be maximize the efficiency of the offense to only run.

Given credible estimates that allow us to recover $\hat{a}$, $\hat{b}$, $\hat{c}$, and $\hat{d}$, we can gain a good idea of what the optimal rush rate would be, holding situation constant.

3 Data

3.1 Play-by-Play Data

We use play-by-play data from games in the National Football League over the years 2006 to 2020. This data includes comprehensive contextual and outcome variables that document the state of the game before the play and describe what happened on each play. Contextual variables include things like down, distance to earn a first down, time remaining in the half, time remaining in the game, yard line, the current score, as well as information about the where the game is being played and one which date. Additionally, outcomes of plays include whether a play was called as a pass or a run, the yardage gained, if the pass was intercepted, if the ball was fumbled and who recovered, whether a sack was recorded, or if the team scored. This data was obtained using the nflfastR package in R (Carl and Baldwin, 2021). Included in the data are outputs from several public analytics models, which we will described in the next section.

3.2 Public Analytics Models

These include expected points and expected pass probability. We will describe these models briefly, but urge readers to consult more comprehensive sources for deeper understanding if they are not familiar. In particular, for details on how to each of these variables were computed, see Baldwin (2021) which draws on the framework of Yurko et al. (2019).

Baldwin (2021) uses XGBoost to predict the relevant outcomes for each model. The expected points model predicts the probabilities of a set of scoring plays (touchdown, field goal, opponent

touchdown, etc.) using variables related to time in game, yard line, home, field type, down and distance, era, and timeouts remaining. Summing up the expected value of the next score using the estimated probabilities leaves us with expected points. Finally, taking the difference in expected points before and after a play yields that play's Expected Points Added (EPA). This will serve as our main measure of efficiency. Second, to estimate usage curves, we will need an estimate of the probability of a pass. Similarly, the Expected Pass (XPass) Model estimates this using variables related to yard line, home, field type, time in game, down and distance, score differential, times, win probability, and era. These values are provided for each play in the play-by-play dataset.

4 Empirical Strategy

4.1 Identifying Pass-Run Usage Curves

4.1.1 Ideal Variation: A Play-Calling Experiment

In preparing for future games, coaches watch and chart game film from previous games to understand opponent "tendencies." Anecdotally, this is often the last three games an opponent has played or the last time a team played that team, coach, or key player. Then, these tendencies are used in setting future game plans. If teams run more, for example, the game plans would be adjusted to be "stout against the run." As an example, defensive alignments might tend to put put more men in the box.[5]

Suppose we could run an experiment, where teams were randomly chosen to pass more (for example, by a certain amount). We might run this experiment for the first four weeks of the season, to set tendencies. If teams passes more often, based on the game planning process commonly instituted in the NFL, the opponent will tend to devote more defenders to coverage responsibil-

[5]From Football Outsiders Glossary, the box is "the defensive area between the offensive tackles extending approximately seven yards deep in the defensive backfield. The defense will put more players "in the box" the more intent they are on stopping a running play." See https://www.footballoutsiders.com/info/glossary_general

ities. Given this type of exogenous variation, we could estimate the causal impact of usage on efficiency by regressing efficiency in that fourth game on pass rate from these first three games.

Aggregating the linear model we specified above, we construct an efficiency regression,[6]

$$E_{ig} = \alpha + \beta \bar{p}_{i,-g} + \gamma Pass_i + \delta Pass_i \bar{p}_{i,-g} \tag{4}$$

where i indexes team, g indexes game, and $\bar{r}_{i,-g}$ is the average rush rate leaving out game g and future games. This research design would estimate the effect of passing, and the effect of passing rate on rushing and passing efficiency.

However, in the real world any such experiments in play calling are few and far between. Moreover, if and when they do take place, they would be difficult to detect. Moreover, in the absence of these experiments, there is real and serious selection bias to contend with: teams who are better at passing will pass more. Even when there are changes in offensive philosophy, these might be driven by improvements in passing offense. Therefore, the endogenous nature of passing will likely bias observational estimates.

4.1.2 Quasi-Experimental Variation: Fumbles Lost as an Instrument

While coaches likely would respond the experiment described above, its just one possible cause for which we could estimate an effect. Just as coaches might respond to a persistent change in run-pass rate, there are many other causes for coaches to adjust their defensive decision-making to counter the pass or the run. Coaches do not just rigidly choose a gameplan and then execute it (in fact, this may be the hallmark of a bad coach). Instead, they make adjustments on the fly as the game goes on, often implemented in response to the game state, which might encompass the net score and the probability a given team will win the football game. It is well documented that an

[6]Note $\alpha = a$, $\beta = b$, $\gamma = c - a$, and $\delta = d - b$. This means the efficiency effect could be restated as $-\gamma/2\delta$ and the the responsiveness effect as $-\beta/2\delta$ and $p^* = -\frac{1}{2}\left(\frac{\gamma+\beta}{\delta}\right)$

increase in a teams win probability decreases propensity to pass and increases their propensity to run. However, using the natural variation in game state within a game will face a similar issue with reverse causality. Teams who are behind likely already have worse efficiency on offense, which is what caused them to be behind. To this end, I propose to use a set of instruments based on cumulative fumbles lost to the opposing team (conditional on cumulative fumbles overall) to remove the selection bias in expected pass rate and actual decision to pass. This set of instruments includes fumbles lost by the possessing team, fumbles lost by the opposing team to the possessing team, and their interaction with actual passing calls.

Fumbles – and who recovers them – have been identified by football analysts as a random, but crucial component of winning football games. For example, Massey-Peabody Analytics, who prepare predictive team ratings based on historical data, down weight statistics related to recovered fumbles in their team ratings: "recovered fumbles, which greatly influence the outcome of games, [...] are completely random."[7] The randomness and importance of fumbles is noted elsewhere. Bill Connelly identifies fumbles and fumble recoveries as a one of the five main factors influencing the outcome of college football games. More specifically, he identifies it as a *random* factor in winning football games. As he so adeptly puts, if you want to win football games, "you want that damned, pointy ball to bounce in a favorable way."[8]

Therefore, cumulative fumbles lost is an attractive candidate for an instrument for pass rate

[7] http://massey-peabody.com/methodology/ Full quote: "Our chief way of doing this is to weigh performance statistics by their predictive ability. That is, their ability to predict out-of-sample performance rather than describe in-sample performance. The canonical example is recovered fumbles, which greatly influence the outcome of games, yet are completely random. Because they are random they have no predictive power, and any stat heavily influenced by fumbles recovered (or many other chance events) will carry less weight in our model than in models based on descriptive analysis."

[8] https://www.footballstudyhall.com/2014/1/24/5337968/college-football-five-factors Full quote: "Over time, I've come to realize that the sport comes down to five basic things, four of which you can mostly control. You want to be efficient when you've got the ball, because if you fall behind schedule and into passing downs, you're far less likely to make a good play. You want to eat up chunks of yardage with big plays, because big plays mean both points and fewer opportunities to make mistakes. When you get the opportunity to score, you want to score. And when you give the ball back to your opponent, you want to give them to have to go as far as possible. And you want that damned, pointy ball to bounce in a favorable way. Again, you control four of the five."

and expected pass rate. In fact, the properties of cumulative past fumbles lost, relate closely to three of the conditions needed for a valid instrument.[9] First, if the number of past fumbles lost are random (conditional on total cumulative fumbles by that team), this satisfies the independence assumption, which states that the instrument is as good as randomly assigned. Second, the importance of fumbles as driving swings in the game state and the importance of game state in passing rate, suggest that fumbles lost will serve as a strong instrument. That is, they will be strongly correlated with the endogenous variable (Stock and Yogo, 2005). This condition can of course be checked in the data. Third, since fumbles themselves make a poor predictor of future performance, this helps in build a case for the exclusion restriction – that the instrument is related to the outcome through the endogenous variable(s) – is satisfied. In addition to these three conditions, the last condition needed for instrumental variables to estimate the local average treatment effect (LATE) is monotonicity. If there is heterogeneous response to the fumbles lost, we need everyone to response weakly in the same direction. That is, while not all coaches need to raise their expected pass rate in response to past fumbles lost, but coaches cannot be pushed toward rushing when fumbles have been lost in the past.

4.1.3 Fixed Effects

For additional robustness, we also propose the use of panel fixed effects to control for offensive quality. With panel data, fixed effects are widely used to control for time invariant observable characteristics. In this case we want to control for team level propensity to pass (a function of underlying talent) as a time invariant unobservable, so the decision of defining these fixed effects is important. I propose to use team-year fixed effects to control for the propensity of a given team with a quarterback to pass or to run.[10]

[9] See, for example Theorem 4.5.1 in (Angrist and Pischke, 2009).

[10] While one might want to utilize a great deal of fixed effects to control for various situations, many of these will overlap. In particular, it seems reasonable that a quarterbacks' talent, a coaches' offensive system, an offenses' supporting cast would affect the probability of calling a run or calling a pass. However, many coach-quarterback

4.2 Main Specifications

4.2.1 Instrumental Variables Specification

Our approach to the instrumental variables specification is to replicate an experimental set-up as closely as possible. Our intuition is reflected in a simple approach where we take games where a fumble was lost as the "treatment" group and those games where the offense fumbled but did not lose it as the "control" group. However, such a set-up would neglect the richness of the data. Therefore, we construct indicator variables for each value of cumulative fumbles lost and each value of cumulative fumbles for both the possession and opposing team. Therefore, instead of one treatment, this experiment has multiple. For example, for those teams that fumble three times, we have three possible treatments corresponding to whether one, two, or three of these fumbles were lost. Finally, we interact our instrumental variables with whether a pass was called on this play. We use the resulting set of instruments in our two-stage least squares estimator and the number of cumulative fumbles as controls.

Before presenting the estimating equations we use to estimate usage curves, we build intuition

combinations are themselves invariant for long periods of time, so estimating coach-quarterback combinations might be a bit fruitless. Moreover, while contracts might last five years in the NFL, talent surrounding the quarterback fluctuates at a faster rate than the quarterback does. Using quarterback-year fixed effects controls for injuries to starting quarterbacks, which tend to adjust the propensity to pass and using the year scale tends to proxy for the turnover of the rest of team by coinciding with free agency. In addition, this coincides with installations of new playbooks, a key point where offensive philosophy might change. Of course, we could cut these a finer, estimating QB-coach-year fixed effects that differ from QB-year fixed effects only when there is a mid-season firing, injury, or quarterback change.

by estimating the causal effect of calling a passing play on efficiency.

$$\begin{aligned}\text{Pass}_{igt} &= \alpha_1 \\ &+ \sum_{k=1}^{K} \eta_{1k} \text{Post-}k \text{ Team Fumbles}_{igt} + \sum_{l=1}^{L} \theta_{1l} \text{Post-}l \text{ Opp. Fumbles}_{igt} \\ &+ \sum_{n=1}^{N} \lambda_{1n} \text{Post-}n \text{ Team Fumbles Lost}_{igt} + \sum_{m=1}^{M} \pi_{1m} \text{Post-}m \text{ Opp. Fumbles Lost}_{igt} \\ &+ \varepsilon_{1igt}\end{aligned} \tag{5}$$

where Pass_{1igt} is an indicator (equal to one) if team i passed on play t of game g. The fitted values of this variable appears in the second stage, denoted with a hat:

$$\begin{aligned}y_{2igp} &= \alpha_2 + \beta_2 \widehat{\text{Pass}}_{igt} \\ &+ \sum_{k=1}^{K} \eta_{2k} \text{Post-}k \text{ Team Fumbles}_{igt} + \sum_{l=1}^{L} \theta_{2l} \text{Post-}l \text{ Opp. Fumbles}_{igp} + \varepsilon_{2igt}\end{aligned} \tag{6}$$

where y_{2igp} is an efficiency variable of choice. This might be EPA, win probability added (WPA), play success (defined as passing some yards or expected points threshold), but we default to EPA. Based on the assumptions presented in Section 4.1.2, β_2 is identified as the LATE of passing on efficiency.

Our main results feature a more complex estimation. Since we have two endogenous variables as well as their interaction, we instrument using both fumbles lost, and fumbles lost specifically

on passing plays as instruments. We specify the first stage regressions:

$$
\begin{aligned}
y_{1igt} &= \alpha_1 \\
&+ \sum_{k=1}^{K} \eta_{1k} \text{Post-}k \text{ Team Fumbles}_{igt} + \sum_{l=1}^{L} \theta_{1l} \text{Post-}l \text{ Opp. Fumbles}_{igt} \\
&+ \sum_{n=1}^{N} (\lambda_{1n} \text{Post-}n \text{ Team Fumbles Lost}_{igt} + \mu_{1n} \text{Post-}n \text{ Team Fumbles Lost} \times \text{Pass}_{igt}) \\
&+ \sum_{m=1}^{M} \Big(\pi_{1m} \text{Post-}m \text{ Opp. Fumbles Lost}_{igt} + \phi_{1m} \text{Post-}m \text{ Opp. Fumbles Lost}_{igt} \times \text{Pass}_{igt} \Big) \\
&+ \varepsilon_{1igp}
\end{aligned}
\tag{7}
$$

where y_{1igp} stands in for the two endogenous variables and their interaction. Again, the fitted values of these endogenous variables appear in the second stage, denoted with a hat:

$$
\begin{aligned}
y_{2igt} &= \alpha_2 + \beta_2 \widehat{\text{Pass}}_{igt} + \gamma_2 \widehat{\text{XPass}}_{igt} + \delta_2 \widehat{\text{Pass}_{igt} \times \text{XPass}}_{igt} + \\
&+ \sum_{k=1}^{K} \eta_{2k} \text{Post-}k \text{ Team Fumbles}_{igt} + \sum_{l=1}^{L} \theta_{2l} \text{Post-}l \text{ Opp. Fumbles}_{igt} + \varepsilon_{2igt}
\end{aligned}
\tag{8}
$$

where y_{2igp} is an efficiency variable of choice. We identify β_2, γ_2, and δ_2 as their respective LATEs. Moreover, this gives us credible estimates to fill our our model of optimal play calling. This will serve as our preferred specification for estimating parameters of the usage curves model.

4.2.2 Fixed Effects Specification

Additionally, we can use fixed effects estimation to control for unobservable differences in teams. This might control for quarterback skill or offensive play calling philosophy. For the uninstru-

mented specification, we estimate:

$$y_{2igp} = \alpha_{2i} + \beta_2 \text{Pass}_{igp} + \gamma_2 \text{XPass}_{igp} + \delta_2 \text{Pass}_{igp} \times \text{XPass}_{igp} + \varepsilon_{2igp} \tag{9}$$

Of course, making a similar adjustment to specification XX and XX allows us to estimate the 2SLS specification with Fixed Effects. We will do so as a robustness check.

5 Results

5.1 The Causal Effect of Passing on Early Downs

Estimating the naive regressions (Table 1 columns 1 and 2) first, we find a positive relationship between QB Dropbacks (i.e., called passes) and offensive efficiency on early down plays. Qualitatively, this the effect we expect – passing has consistently outperformed rushing in this efficiency metric. However, when we estimate the Two Way Least Squares Estimator using our Cumulative Fumbles Lost Instruments we find a much larger effect. In particular, while the "naive" effect was an increase of 0.16 EPA per called pass, 2SLS estimates a 0.26 EPA increase in EPA (Table 1 column 3.2). How can we understand this difference in effect sizes? If teams with better passing offenses also have better rushing offenses (say, because they have good offensive lines), this could lead to the kind of negative omitted variable bias seen here.

Second, this illustration shows the behavioral relevance of the instruments for passing. Cumulative fumbles lost seem to drive pass rate as we would expect them to. In particular, fumbles lost, which put a coach into a worse game state, increases their probability of dropping back to pass while opponent fumbles lost (i.e., fumbles gained) decreases their probability of dropping back. When coaches get lucky with fumbles, they are induced to pass less. Moreover, the instruments tell a coherent story. The effect size of fumbles lost and gained is relatively symmetric (opponent fumbles have tiny bit of a stronger effect) and the effects increase as fumbles mount up

Table 1: The Effects of Called QB Dropbacks on Efficiency. First and second down plays outside of the two minute warning from 2006-2020. Controls omitted from table.

	OLS	FE	2SLS		FE 2SLS	
			1st Stage	2nd Stage	1st Stage	2nd Stage
	EPA	EPA	Pass	EPA	Pass	EPA
	(1)	(2)	(3.1)	(3.2)	(4.1)	(4.2)
1 Pos. Fum. Lost			0.021***		0.023***	
			(0.005)		(0.004)	
2 Pos. Fum. Lost			0.045***		0.047***	
			(0.009)		(0.009)	
3 Pos. Fum. Lost			0.098***		0.098***	
			(0.019)		(0.018)	
1 Opp. Fum. Lost			−0.024***		−0.023***	
			(0.004)		(0.004)	
2 Opp. Fum. Lost			−0.054***		−0.059***	
			(0.009)		(0.009)	
3 Opp. Fum. Lost			−0.129***		−0.139***	
			(0.020)		(0.020)	
Pass	0.161***	0.161***		0.258*		0.307**
	(0.004)	(0.004)		(0.144)		(0.134)
FE	No	Yes	No	No	Yes	Yes
Observations	328367	328367	328369	328367	328369	328367
R^2	0.005	0.009	0.004	0.003	0.016	0.005
Adjusted R^2	0.005	0.008	0.004	0.003	0.015	0.004

Note: *p<0.1; **p<0.05; ***p<0.01

Table 2: Estimates of Usage Curves in Professional Football. First and second down plays outside of the two minute warning from 2006-2020. Controls omitted from table.

	EPA			
	OLS	FE	2SLS	FE 2SLS
	(1)	(2)	(3)	(4)
Pass	0.243^{***}	0.238^{***}	0.484^{***}	0.484^{***}
	(0.014)	(0.014)	(0.117)	(0.117)
XPass	-0.014	0.030^{**}	0.458^{**}	0.540^{***}
	(0.014)	(0.015)	(0.203)	(0.200)
Pass × XPass	-0.136^{***}	-0.137^{***}	-0.663^{***}	-0.680^{***}
	(0.025)	(0.025)	(0.198)	(0.199)
Constant	-0.079^{***}		-0.294^{***}	
	(0.007)		(0.087)	
Fixed Effects	No	Yes	No	Yes
Observations	328367	328367	328367	328367
R^2	0.005	0.009	0.003	0.007
Adjusted R^2	0.005	0.008	0.003	0.005
Note:				$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

(two fumbles lost makes us pass more than one, and three more than two). Formally testing the instruments, we estimate a first stage Craig Donald statistic of 134.4 from the non-fixed effects 2SLS regression, comparable to a critical value of 19.28 (for one endogenous variable, six instruments, and 2SLS bias of 0.05) (Stock and Yogo, 2005), reducing concerns about issues related to weak instruments.

5.2 Estimating Usage Curves on Early Downs

5.2.1 Naive Estimates

While the value of passing fell as the expectation of passing increases, rushing did not become more valuable as it becomes more surprising. Placing these estimates into our simple model of optimal pass rates, the naive estimates suggest an optimal pass rate of approximately 84% on early downs.

5.2.2 Two Stage Least Squares Estimates

The main effects on early downs (all first and second down plays) are presented in table 2. Focusing on the pooled 2SLS regression, estimates differ considerably from the results including later downs. In particular, usage curves flatten substantially. When the probability of passing approaches zero, 2SLS estimates a premium of 0.48 EPA. However, when passing probability is close to the average pass rate (52%), the passing premium is estimated at around 0.14 EPA (close to the unconditional premium of 0.16). Finally, in situations where the probability of passing approaches one, the premium estimated falls to -0.18 EPA. These results suggest that offenses should pass about 71% of early down plays as opposed to the actual pass rate of 52% on early downs. These results suggest that the average coach overestimate the element of surprise (at least on early downs) and tends to run to too much as a result.

5.3 Addressing Threats to Validity

5.3.1 Relevance with Multiple Endogenous Variables

While we demonstrated relevance in single equation models, we also provide evidence here around the relevance of the instruments when we estimate regressions with multiple endogenous variables. In particular, when estimating the Two-Stage Least Squares with pass rate, expected

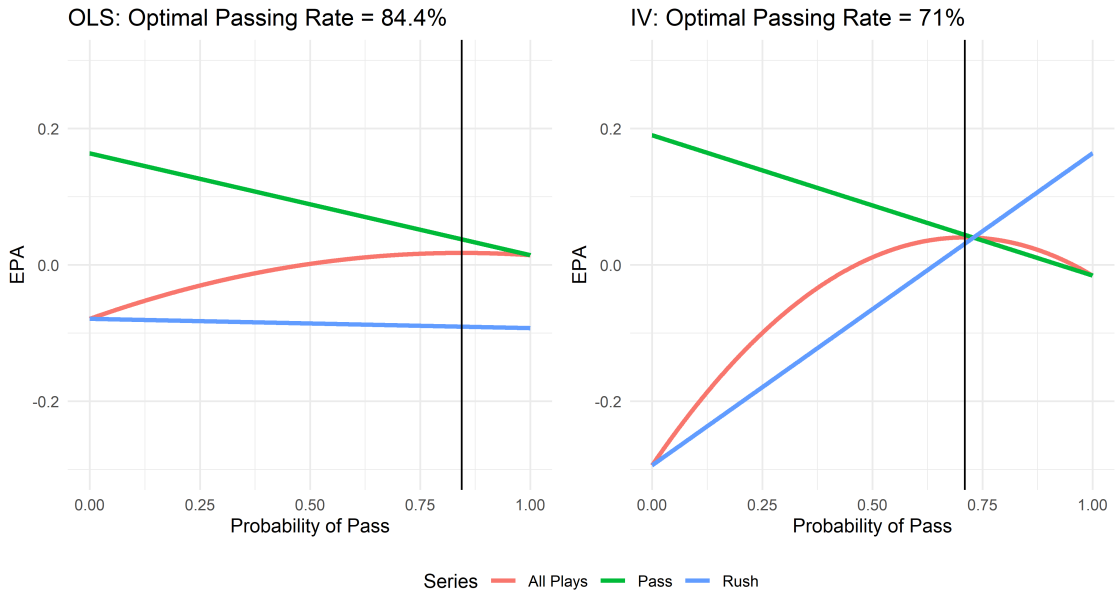

Figure 2: Optimal pass rates with linear usage effects OLS and IV estimates

pass rate, and their interaction, we obtain a Craig-Donald Statistic of 134.4 in the early down sample (our preferred sample). This is comparable to a critical value of 17.8 from Stock and Yogo (2005) (for three endogenous regressors and twelve instruments). Results from the first stage of the Pooled 2SLS estimation are presented in Table 3.

5.3.2 When Does 2SLS Estimate LATE?

Recent work has found that 2SLS does not always return estimates of LATE when controls are included, which might complicate interpretation of results. However, the "saturated" design we use is consistent with the non-parametric design recommended in Blandhol et al. (2022), which they show is necessary and sufficient for 2SLS to return LATE.

Table 3: First Stage Results from 2SLS Usage Curves Estimation. First and second down plays outside of the two minute warning from 2006-2020. Controls omitted from table.

	Dependent variable:		
	Pass	XPass	XPass x Pass
	(1)	(2)	(3)
1 Pos. Fum. Lost	-0.392^{***}	-0.055^{***}	-0.246^{***}
	(0.004)	(0.003)	(0.003)
2 Pos. Fum. Lost	-0.391^{***}	-0.053^{***}	-0.264^{***}
	(0.009)	(0.006)	(0.006)
3 Pos. Fum. Lost	-0.360^{***}	-0.019	-0.258^{***}
	(0.020)	(0.017)	(0.012)
1 Opp. Fum. Lost	-0.406^{***}	-0.087^{***}	-0.246^{***}
	(0.004)	(0.003)	(0.003)
2 Opp. Fum. Lost	-0.392^{***}	-0.118^{***}	-0.244^{***}
	(0.008)	(0.006)	(0.005)
3 Opp. Fum. Lost	-0.376^{***}	-0.162^{***}	-0.237^{***}
	(0.016)	(0.012)	(0.010)
1 Pos. Fum. Lost x Pass	0.755^{***}	0.132^{***}	0.490^{***}
	(0.006)	(0.002)	(0.004)
2 Pos. Fum. Lost x Pass	0.742^{***}	0.157^{***}	0.529^{***}
	(0.013)	(0.005)	(0.009)
3 Pos. Fum. Lost x Pass	0.686^{***}	0.148^{***}	0.538^{***}
	(0.033)	(0.015)	(0.023)
1 Opp. Fum. Lost x Pass	0.766^{***}	0.128^{***}	0.444^{***}
	(0.006)	(0.002)	(0.004)
2 Opp. Fum. Lost x Pass	0.747^{***}	0.157^{***}	0.427^{***}
	(0.013)	(0.005)	(0.009)
3 Opp. Fum. Lost x Pass	0.711^{***}	0.173^{***}	0.374^{***}
	(0.032)	(0.017)	(0.027)
Observations	328369	328369	328369
R^2	0.391	0.128	0.381
Adjusted R^2	0.391	0.128	0.381
Note:			$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

6 Conclusion

6.1 Future Work

6.1.1 Further Exploration of Context and Heterogeneity

We hope to continue this research by further exploring the context that might drive play calling. First, do game scenarios beyond those that we described augment optimal choice? Second, we hope to explore heterogeneity in incentives. In this work, we have not addressed the fact that some teams may better or worse passing and rushing games. For example, a team with an elite passer like prime Patrick Mahomes will have a better passing offense than others, which would mean that team should have a better passing offense. Alternatively, a team with an elite rushing quarterback like Lamar Jackson may tilt more towards the run game based on the efficiency of quarterback runs (as compared to running backs).

6.1.2 What Explains Departures from Optimal Passing Rates?

It is often hard to find clear mistakes in decision-making. Because football features fixed, known rules and computable incentives, in a highly competitive environment with high stakes, studying deviations from optimal play-calling can be an enlightening look into decision-making under uncertainty. Deviations from optimal decision-making have been explored before, often in limited situations. For example, Romer (2006) documents systematic departures from fourth down decisions that would maximize teams changes of winning. Football also provides a setting to move beyond "do firms maximize?" to study "what do firms maximize?" Slade and Tolhurst (2019) studies how the coaches incentives might drive decision-making under uncertainty in these settings, finding that coaches are more risk loving when their job is very secure or when they are close to being fired.

While past work has worked to do this in the context of pass-rush ratios,[11] none have had such convincing causal identification in determining optimal decision-making. Therefore, there may be considerable value added to reexamine these explanations within this context. In future work, we hope to study decision-making leveraging the clarity with which we can identify both optimal playcalling and therefore deviations from optimal playcalling.

[11] See Alamar (2006), Rockerbie (2008), Jordan et al. (2009), and McGarrity and Linnen (2010)

References

B. C. Alamar. The Passing Premium Puzzle. *Journal of Quantitative Analysis in Sports*, 2(4), 2006. ISSN 2194-6388. doi: 10.2202/1559-0410.1051.

J. D. Angrist and A. B. Krueger. Empirical strategies in labor economics. *Handbook of Labor Economics*, 3(23), 1999. URL papers3://publication/uuid/27A06988-75A6-4DC1-A4A0-05BDA7B31A71.

J. D. Angrist and J.-S. Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. 2009.

B. Baldwin. Open Source Football: nflfastR EP, WP, CP xYAC, and xPass models, 2021. URL https://www.opensourcefootball.com/posts/2020-09-28-nflfastr-ep-wp-and-cp-models/.

C. Blandhol, J. Bonney, M. Mogstad, and A. Torgovitsky. When is TSLS Actually LATE? 2022.

S. Carl and B. Baldwin. nflfastR: Functions to Efficiently Access NFL Play by Play Data, 2021. URL https://www.nflfastr.com/,%0Ahttps://github.com/nflverse/nflfastR.

J. D. Jordan, S. H. Melouk, and M. B. Perry. Optimizing Football Game Play Calling. *Journal of Quantitative Analysis in Sports*, 5(2), 2009. ISSN 2194-6388. doi: 10.2202/1559-0410.1176.

J. P. McGarrity and B. Linnen. Pass or Run: An Empirical Test of the Matching Pennies Game Using Data from the National Football League. *Southern Economic Journal*, 76(3):791–810, 2010. ISSN 0038-4038. doi: 10.4284/sej.2010.76.3.791.

D. W. Rockerbie. The Passing Premium Puzzle Revisited. *Journal of Quantitative Analysis in Sports*, 4(2), 2008. ISSN 2194-6388. doi: 10.2202/1559-0410.1093.

D. Romer. Do firms maximize? Evidence from professional football. *Journal of Political Economy*, 114(2):340–365, 2006. ISSN 00223808. doi: 10.1086/501171.

P. Slade and T. Tolhurst. Job Security and Risk-Taking: Theory and Evidence From Professional Football. *Southern Economic Journal*, 85(3):899–918, 2019. ISSN 00384038. doi: 10.1002/soej.

12313.

J. H. Stock and M. Yogo. Testing for weak instruments in linear IV regression. (February):80–108, 2005.

R. Yurko, S. Ventura, and M. Horowitz. nflWAR: A reproducible method for offensive player evaluation in football. *Journal of Quantitative Analysis in Sports*, 15(3):163–183, 2019. ISSN 15590410. doi: 10.1515/jqas-2018-0010.

A Robustness

A.1 Relevance with Multiple Endogenous Variables: Fixed Effects Results

Table 4 presents evidence of instrumental relevance from the fixed effects specifications.

Table 4: First Stage Results from FE 2SLS Usage Curves Estimation. First and second down plays outside of the two minute warning from 2006-2020. Controls omitted from table.

	Dependent variable:		
	Pass	XPass	XPass x Pass
	(1)	(2)	(3)
1 Pos. Fum. Lost	-0.390^{***}	-0.052^{***}	-0.243^{***}
	(0.004)	(0.003)	(0.003)
2 Pos. Fum. Lost	-0.388^{***}	-0.050^{***}	-0.261^{***}
	(0.009)	(0.006)	(0.006)
3 Pos. Fum. Lost	-0.358^{***}	-0.016	-0.256^{***}
	(0.020)	(0.016)	(0.012)
1 Opp. Fum. Lost	-0.404^{***}	-0.086^{***}	-0.244^{***}
	(0.004)	(0.003)	(0.003)
2 Opp. Fum. Lost	-0.393^{***}	-0.119^{***}	-0.245^{***}
	(0.008)	(0.006)	(0.005)
3 Opp. Fum. Lost	-0.379^{***}	-0.156^{***}	-0.236^{***}
	(0.016)	(0.012)	(0.010)
1 Pos. Fum. Lost x Pass	0.751^{***}	0.130^{***}	0.487^{***}
	(0.006)	(0.002)	(0.004)
2 Pos. Fum. Lost x Pass	0.738^{***}	0.154^{***}	0.525^{***}
	(0.013)	(0.005)	(0.009)
3 Pos. Fum. Lost x Pass	0.686^{***}	0.146^{***}	0.537^{***}
	(0.032)	(0.014)	(0.023)
1 Opp. Fum. Lost x Pass	0.763^{***}	0.126^{***}	0.442^{***}
	(0.006)	(0.002)	(0.004)
2 Opp. Fum. Lost x Pass	0.744^{***}	0.155^{***}	0.425^{***}
	(0.013)	(0.005)	(0.009)
3 Opp. Fum. Lost x Pass	0.710^{***}	0.168^{***}	0.370^{***}
	(0.032)	(0.016)	(0.026)
Observations	328,369	328,369	328,369
R^2	0.396	0.147	0.387
Adjusted R^2	0.395	0.146	0.386
Note:			$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

A.2 Instrumental Validity: Passing Aggression

Other margins of aggression in decision-making by coaches may threaten the exclusion restriction if they are correlated with efficiency. A concerning alternative mechanism might be depth of passes thrown when in a negative game state. Table 5 presents results of regressing passing strategy and efficiency on our instruments. However, we find that on early down pass attempts, cumulative fumbles lost do not increase the share of deep passes, nor the EPA per dropback on called passes.

Table 5: Effect of Cumulative Fumbles Lost on Passing Scheme Aggression. First and second down plays outside of the two minute warning from 2006-2020. Controls omitted from table.

	Dependent variable:			
	Deep Pass		Dropback EPA	
	(1)	(2)	(3)	(4)
1 Pos. Fum. Lost	−0.002	−0.002	−0.001	0.002
	(0.003)	(0.003)	(0.012)	(0.011)
2 Pos. Fum. Lost	0.0003	−0.002	0.002	0.006
	(0.006)	(0.006)	(0.022)	(0.022)
3 Pos. Fum. Lost	0.00004	−0.006	0.052	0.079*
	(0.013)	(0.013)	(0.046)	(0.046)
1 Opp. Fum. Lost	−0.004	−0.005	0.007	0.011
	(0.003)	(0.003)	(0.011)	(0.011)
2 Opp. Fum. Lost	−0.003	−0.003	0.003	0.013
	(0.007)	(0.007)	(0.023)	(0.023)
3 Opp. Fum. Lost	−0.008	−0.003	0.023	0.022
	(0.016)	(0.016)	(0.056)	(0.056)
Constant	0.179***		0.085***	
	(0.002)		(0.006)	
Fixed Effects	No	Yes	No	Yes
Observations	154572	154572	169466	169466
R^2	0.0002	0.008	0.0002	0.007
Adjusted R^2	0.0001	0.005	0.0001	0.004
Note:				*p<0.1; **p<0.05; ***p<0.01