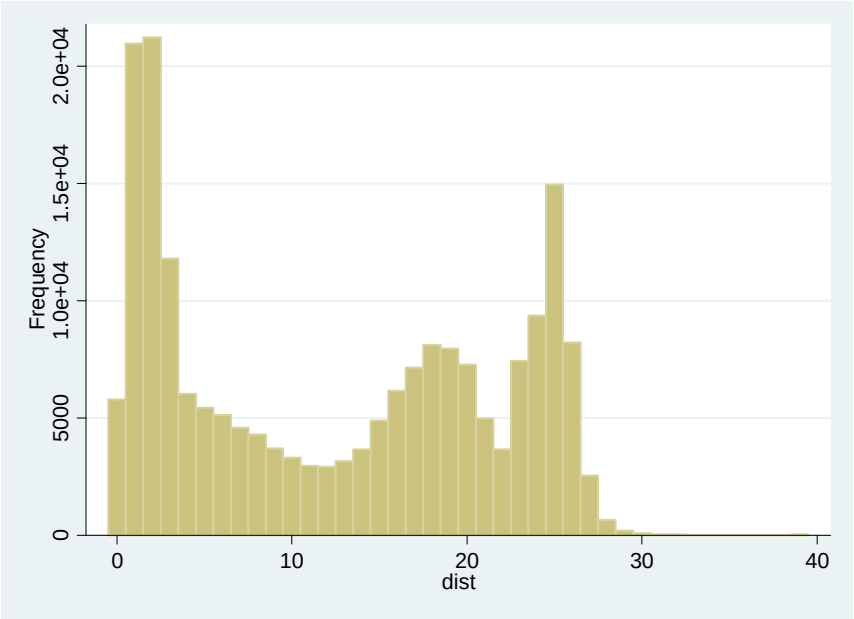
NBA Shots 2011: Success v. Distance

```
. bcuse nbashots11

Contains data from http://fmwww.bc.edu/ec-p/data/wooldridge/nbashots11.dta
  obs:       199,119
 vars:            15                          25 Oct 2012 09:08
 size:    24,690,756
-------------------------------------------------------------------------------------------
storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------------------
mademiss        str6   %9s
qtr             byte   %12.0g
shotsid         str9   %9s
minsleft        byte   %12.0g
secsleft        double %12.0g
player          str21  %21s
shotpts         byte   %12.0g
dist            byte   %12.0g
score           str32  %32s
home            str3   %9s
away            str3   %9s
gtime           str8   %8s
gameid          str14  %14s
team            str13  %13s
team3           str3   %9s
-------------------------------------------------------------------------------------------
Sorted by:  shotsid  gameid

. histogram dist, discrete frequency
(start=0, width=1)
```

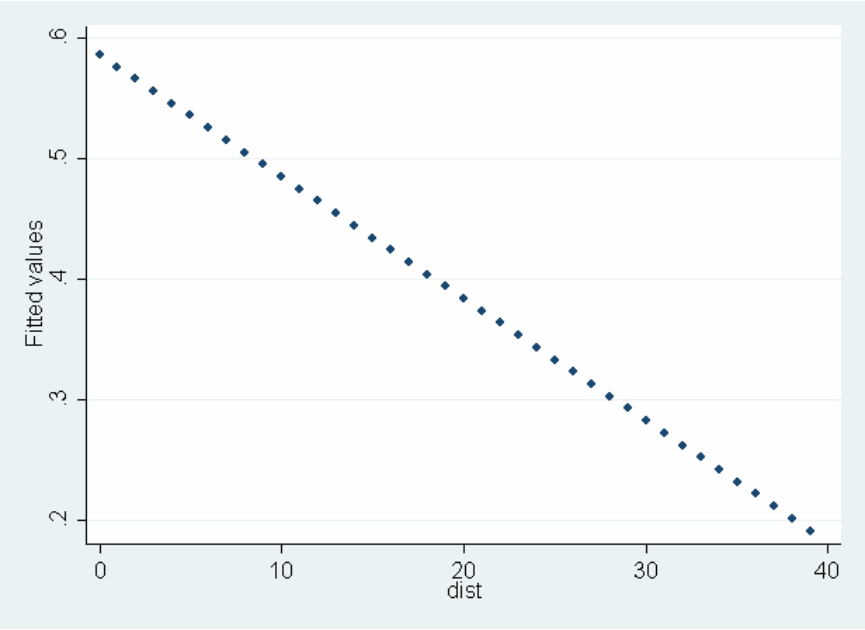


Exploring Functional Forms: *NBA Shots*

Linear Probability Models (LPMs)

Bring on the Linear Probability Model (LPM), an OLS model that predicts probabilities of success. Just regress swish (success) on, say, distance:

```
. gen swish=(mademiss=="made")

. reg swish dist

      Source |       SS       df       MS              Number of obs =  199119
-------------+------------------------------           F(  1,199117) = 7351.64
       Model |  1761.15932     1  1761.15932           Prob > F      =  0.0000
    Residual |  47700.5009199117  .239560163           R-squared     =  0.0356
-------------+------------------------------           Adj R-squared =  0.0356
       Total |  49461.6602199118  .248403762           Root MSE      =  .48945

------------------------------------------------------------------------------
       swish |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dist |  -.0101387   .0001182   -85.74   0.000    -.0103704   -.0099069
       _cons |   .5867489   .0018405   318.81   0.000     .5831417    .5903562
```



But does it really make sense to impose a linear relationship between shot success and distance? Let's explore functional forms.

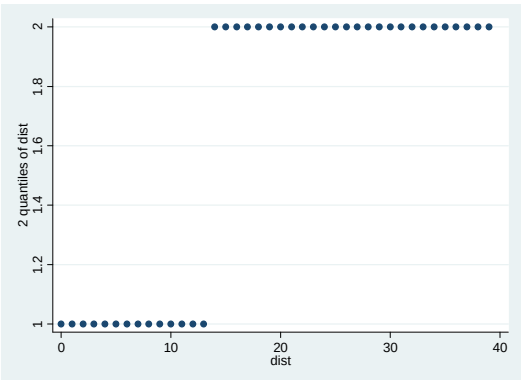
Exploring Functional Forms: *NBA Shots*

Looking at Functional Forms

A. Generate Quantile Dummies I: *Median-tile* Dummies

Let's start with two (equally sized) quantile distance groups:
Percentile dummies for 0%-50% and 50%-100%

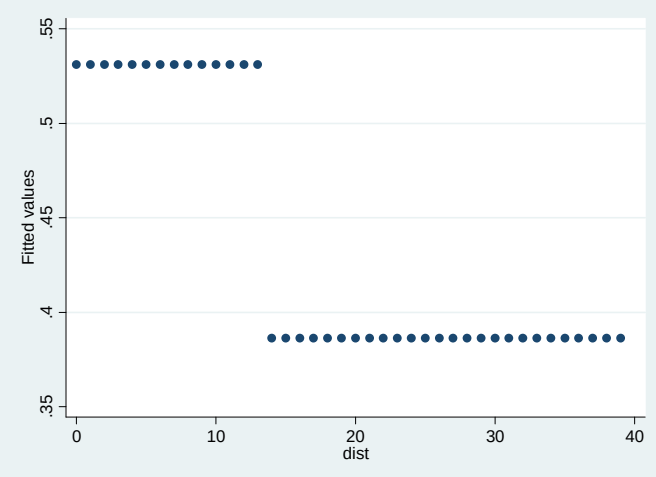
```
. xtile dist2=dist,n(2)
. tab dist2

2 quantiles |
    of dist |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |    101,365       50.91       50.91
          2 |     97,754       49.09      100.00
------------+-----------------------------------
      Total |    199,119      100.00
```




```
. bysort dist2: summ dist

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
   dist2 = 1 |    101365    4.173324    3.598438          0         13
   dist2 = 2 |     97754    21.13107    3.930779         14         39

. reg swish i.dist2

      Source |       SS       df       MS              Number of obs =  199119
-------------+------------------------------           F(  1,199117) = 4286.23
       Model |  1042.28362     1  1042.28362           Prob > F      =  0.0000
    Residual |  48419.3766199117  .243170481           R-squared     =  0.0211
-------------+------------------------------           Adj R-squared =  0.0211
       Total |  49461.6602199118  .248403762           Root MSE      =  .49312

------------------------------------------------------------------------------
       swish |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     2.dist2 |  -.1447232   .0022106   -65.47   0.000    -.1490558   -.1403905
       _cons |   .5310807   .0015489   342.89   0.000      .528045    .5341165
------------------------------------------------------------------------------
```

And graph the predicted values…

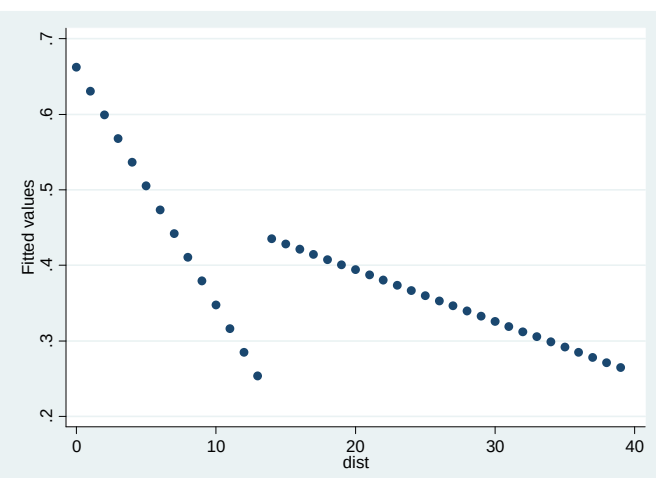
Exploring Functional Forms: *NBA Shots*



Hmmm… not sure about fitting that linear function in the first regression. Let's allow for different slopes and intercepts for the two median-tiles:

```
. gen ddist2 = dist*(dist2==2)

. reg swish dist i.dist2 ddist2

      Source |       SS       df       MS              Number of obs =  199119
-------------+------------------------------           F(  3,199115) = 3400.71
       Model |  2410.76556     3  803.588519           Prob > F      =  0.0000
    Residual |  47050.8946199115  .236300101           R-squared     =  0.0487
-------------+------------------------------           Adj R-squared =  0.0487
       Total |  49461.6602199118  .248403762           Root MSE      =  .48611

------------------------------------------------------------------------------
       swish |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dist |  -.0314502   .0004243   -74.12   0.000    -.0322818   -.0306186
     2.dist2 |  -.1318809   .0088172   -14.96   0.000    -.1491623   -.1145994
      ddist2 |   .0246311   .0005801    42.46   0.000     .0234942     .025768
       _cons |   .6623326   .0023381   283.28   0.000     .6577499    .6669152
------------------------------------------------------------------------------
```

Here are the predicted values:



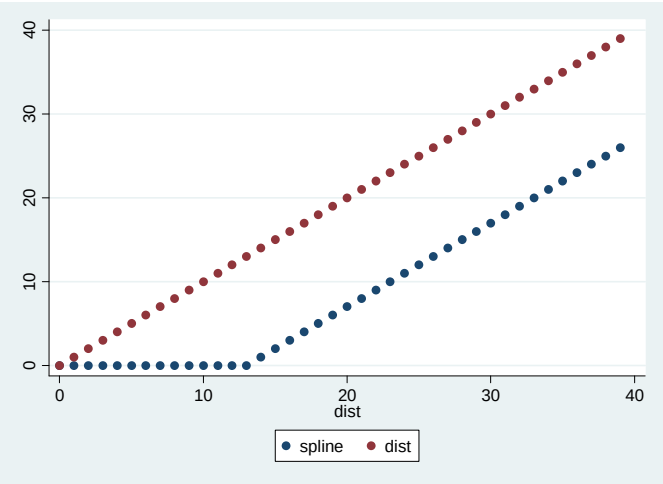
Right! So if you step back from the 13' jumper and go for the 14' shot, your probability of success almost doubles. I don't think so.

B. Linear Splines: Tieing predicted values together

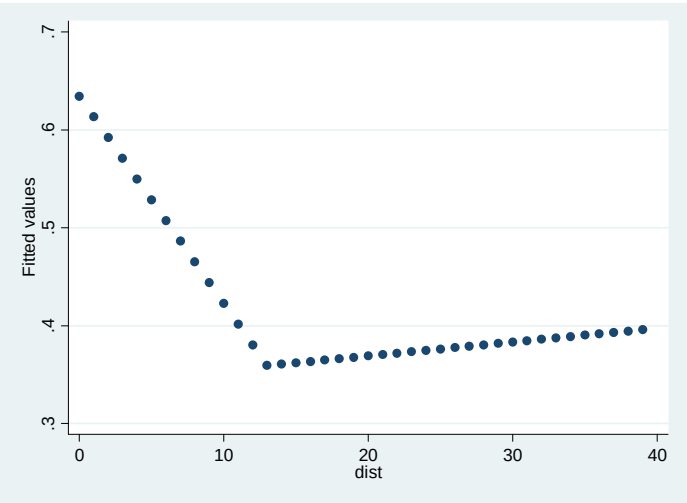
So bring on Linear Splines! … which allow for different slopes and intercepts, but tie the two predicted values together so there is no discontinuity.

```
. gen spline= max(dist-13,0)
```

Here's what the spline looks like (compared to dist):




```
. reg swish dist spline

      Source |       SS       df       MS              Number of obs =  199119
-------------+------------------------------           F(  2,199116) = 4464.79
       Model |  2122.95613     2  1061.47807           Prob > F      =  0.0000
    Residual |  47338.7041199116   .23774435           R-squared     =  0.0429
-------------+------------------------------           Adj R-squared =  0.0429
       Total |  49461.6602199118  .248403762           Root MSE      =  .48759

------------------------------------------------------------------------------
       swish |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dist |  -.0211727   .0003064   -69.10   0.000    -.0217733   -.0205722
      spline |    .022581   .0005789    39.01   0.000     .0214465    .0237156
       _cons |   .6345183   .0022048   287.79   0.000      .630197    .6388397
------------------------------------------------------------------------------
```



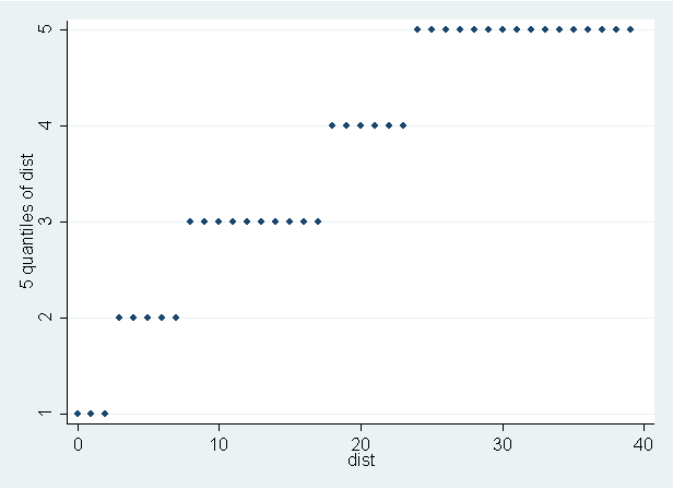
Well at least the predicted values tie together now. But what about that slope for the second median-tile?

Maybe we need to look at more quantile groups.. bring on the quintile (5 equally sized groups) dummies?

C. Generate Quantile Dummies II: Quintile Dummies

Percentile dummies for 0%-20%, -40%, -60%, -80% and %100.

```
. xtile dist5=dist, n(5)
```

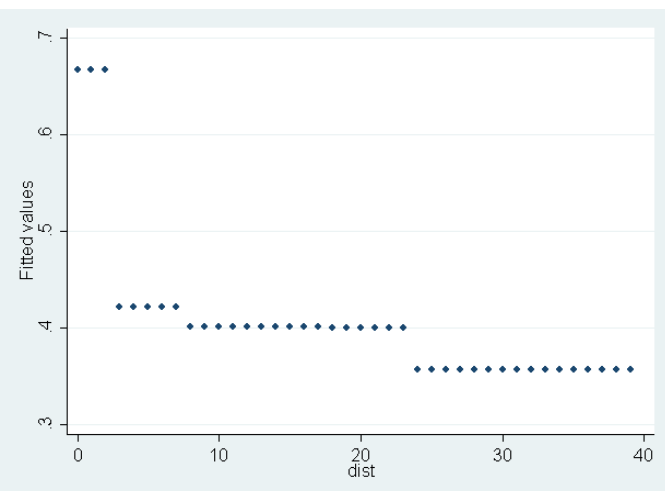



```
. tab dist5

5 quantiles |
    of dist |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |     47,972       24.09       24.09
          2 |     32,992       16.57       40.66
          3 |     42,273       21.23       61.89
          4 |     39,464       19.82       81.71
          5 |     36,418       18.29      100.00
------------+-----------------------------------
      Total |    199,119      100.00
```

Exploring Functional Forms: *NBA Shots*

```
. mean dist, over(dist5)
. bysort dist5: summ dist

    dist5    |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        1    |     47972    1.321521    .6779594          0          2
        2    |     32992    4.535615    1.452283          3          7
        3    |     42273    13.12587    3.092234          8         17
        4    |     39464    20.26449    1.774758         18         23
        5    |     36418    25.29115     1.40181         24         39

. reg swish i.dist5

      Source |       SS       df       MS              Number of obs =  199119
-------------+------------------------------           F(  4,199114) = 2986.11
       Model |  2799.18944     4  699.797361           Prob > F      =  0.0000
    Residual |  46662.4708199114  .234350527           R-squared     =  0.0566
-------------+------------------------------           Adj R-squared =  0.0566
       Total |  49461.6602199118  .248403762           Root MSE      =   .4841

------------------------------------------------------------------------------
       swish |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       dist5 |
          2  |  -.2460069   .0034624   -71.05   0.000    -.2527932   -.2392206
          3  |  -.2666731   .0032294   -82.58   0.000    -.2730026   -.2603436
          4  |  -.2681974   .0032899   -81.52   0.000    -.2746455   -.2617493
          5  |  -.3111654   .0033645   -92.48   0.000    -.3177598   -.3045709
             |
       _cons |   .6674727   .0022102   301.99   0.000     .6631407    .6718047
------------------------------------------------------------------------------
```

And the predicted values?



Exploring Functional Forms: *NBA Shots*

Aside: Alternative approach using xi to generate _I* variables:

```
. xi i.dist5
i.dist5           _Idist5_1-5         (naturally coded; _Idist5_1 omitted)

. reg swish dist _Idist*

      Source |       SS       df       MS              Number of obs =  199119
-------------+------------------------------           F(  5,199113) = 2410.85
       Model |  2823.46886     5  564.693772           Prob > F      =  0.0000
    Residual |  46638.1913199113  .234229766           R-squared     =  0.0571
-------------+------------------------------           Adj R-squared =  0.0571
       Total |  49461.6602199118  .248403762           Root MSE      =  .48397

------------------------------------------------------------------------------
       swish |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dist |  -.0059246   .0005819   -10.18   0.000    -.0070652   -.0047841
   _Idist5_2 |  -.2269645   .0039345   -57.69   0.000    -.2346761    -.219253
   _Idist5_3 |  -.1967365   .0075901   -25.92   0.000    -.2116129   -.1818601
   _Idist5_4 |   -.155967   .0115035   -13.56   0.000    -.1785137   -.1334204
   _Idist5_5 |  -.1691539   .0143483   -11.79   0.000    -.1972761   -.1410316
       _cons |   .6753022   .0023397   288.63   0.000     .6707165    .6798879
------------------------------------------------------------------------------
```

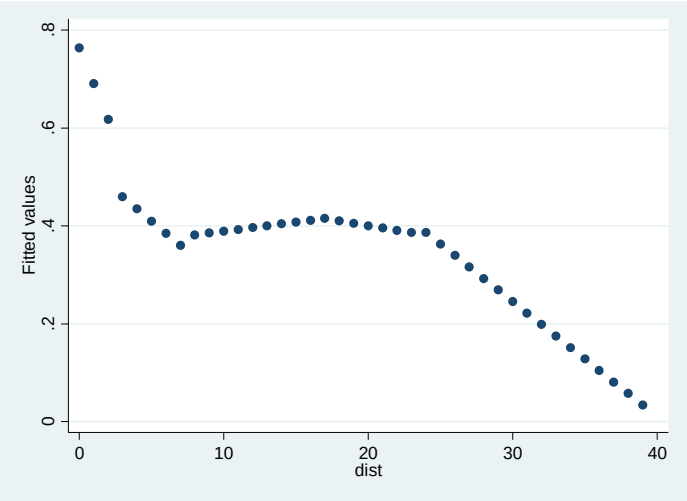
Back to Business!

Let's allow for different slopes and intercepts for the different quintiles:

```
. reg swish c.dist##i.dist5

      Source |       SS       df       MS              Number of obs =  199119
-------------+------------------------------           F(  9,199109) = 1432.30
       Model |  3007.52556     9  334.169506           Prob > F      =  0.0000
    Residual |  46454.1346199109   .23331007           R-squared     =  0.0608
-------------+------------------------------           Adj R-squared =  0.0608
       Total |  49461.6602199118  .248403762           Root MSE      =  .48302

------------------------------------------------------------------------------
       swish |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dist |  -.0728427   .0032529   -22.39   0.000    -.0792184   -.0664671
             |
       dist5 |
          2  |  -.2291294   .0099696   -22.98   0.000    -.2486696   -.2095893
          3  |   -.412145   .0113274   -36.38   0.000    -.4343464   -.3899436
          4  |  -.2684954   .0282851    -9.49   0.000    -.3239334   -.2130573
          5  |   .1873447   .0459907     4.07   0.000      .097204    .2774854
             |
dist5#c.dist |
          2  |   .0478978   .0037329    12.83   0.000     .0405814    .0552142
          3  |   .0765917   .0033405    22.93   0.000     .0700445    .0831389
          4  |   .0681071   .0035297    19.30   0.000      .061189    .0750251
          5  |   .0493257   .0037205    13.26   0.000     .0420337    .0566176
             |
       _cons |   .7637359   .0048315   158.07   0.000     .7542663    .7732054
------------------------------------------------------------------------------
```

And the predicted values?

Exploring Functional Forms: *NBA Shots*

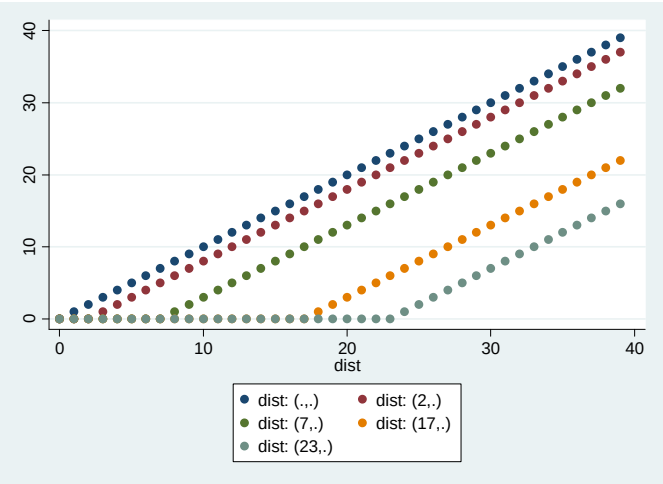


The discontinuities aren't as bad as with the median-tiles… but they are there. So let's use the linear splines to tie the predicted values together.

Use the mkspline command in Stata:

```
. mkspline spline 5 = dist, pctile marginal
```

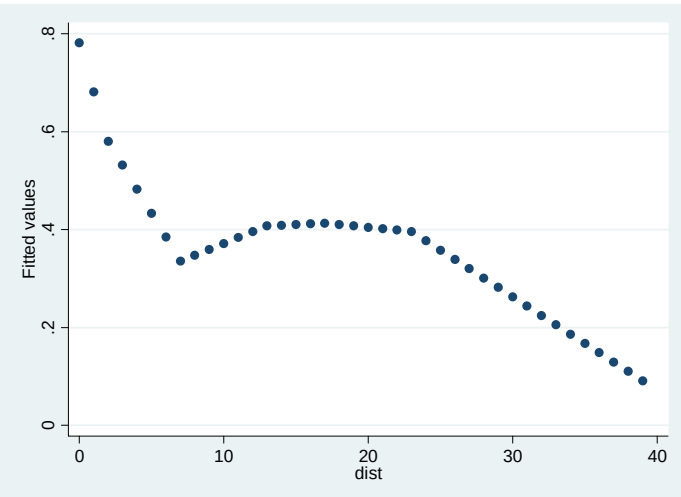
Here's what the splines look like… and yes they are all of the form max(dist-x, 0), where the x's take on different values as above.



And the regression results?

Exploring Functional Forms: *NBA Shots*

```
. reg swish spline*

      Source |       SS       df       MS              Number of obs =  199119
-------------+------------------------------           F(  5,199113) = 2456.52
       Model |  2873.83789     5  574.767577           Prob > F      =  0.0000
    Residual |  46587.8223199113  .233976799           R-squared     =  0.0581
-------------+------------------------------           Adj R-squared =  0.0581
       Total |  49461.6602199118  .248403762           Root MSE      =  .48371

------------------------------------------------------------------------------
       swish |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     spline1 |  -.1014631   .0030204   -33.59   0.000     -.107383   -.0955432
     spline2 |   .0542955   .0035577    15.26   0.000     .0473224    .0612685
     spline3 |   .0548282   .0013497    40.62   0.000     .0521829    .0574736
     spline4 |  -.0119232   .0011474   -10.39   0.000    -.0141721   -.0096743
     spline5 |  -.0142517   .0019425    -7.34   0.000     -.018059   -.0104445
       _cons |   .7821701   .0047741   163.84   0.000     .7728129    .7915273
------------------------------------------------------------------------------
```



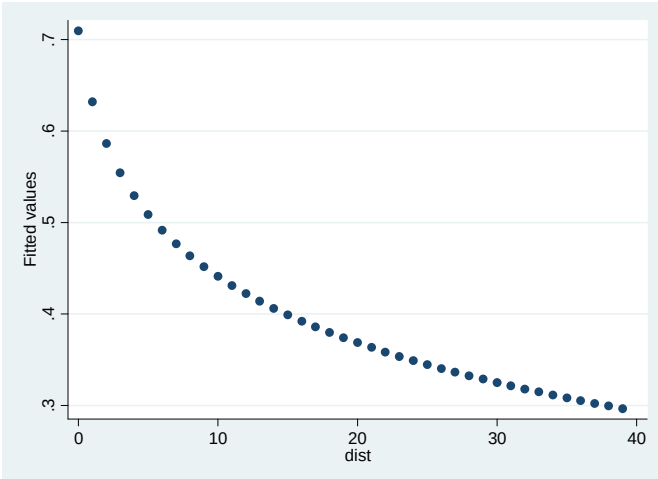
Better!... but do we really believe that shot success rates increase in the third quintile?

D. It's always a good idea to explore logarithms!

Maybe taking ln's will provide a better fit. So generate ln(dist) and estimate the model (I actually compute ln(dist+1) to avoid the undefined ln(.)'s at 0).

```
. gen lndist=ln(dist+1)
. reg swish lndist

      Source |       SS       df       MS              Number of obs =  199119
-------------+------------------------------           F(  1,199117) =10245.14
       Model |  2420.40774     1  2420.40774           Prob > F      =  0.0000
    Residual |  47041.2525199117  .236249303           R-squared     =  0.0489
-------------+------------------------------           Adj R-squared =  0.0489
       Total |  49461.6602199118  .248403762           Root MSE      =  .48605

------------------------------------------------------------------------------
       swish |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lndist |  -.1120073   .0011066  -101.22   0.000    -.1141762   -.1098384
       _cons |   .7095249   .0026949   263.29   0.000      .704243    .7148067
------------------------------------------------------------------------------
```

Exploring Functional Forms: *NBA Shots*



Not bad… we no longer have that increasing success rates in the third quintile… but do we really believe that 30% success rate at 38 feet? Time for more flexible functional forms!

E. Bring on the Polynomials!

```
gen d1 = dist
gen d2 = dist^2
gen d3 = dist^3
gen d4 = dist^4
gen d5 = dist^5
```

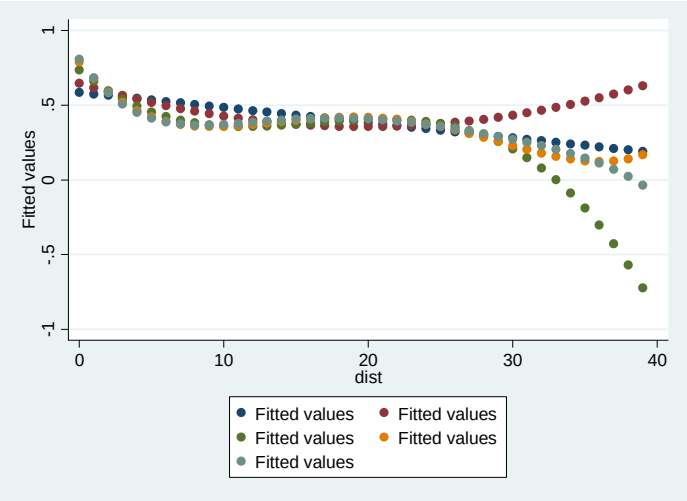
And regress swish on 1st through 5th order polynomials:

	(1) swish	(2) swish	(3) swish	(4) swish	(5) swish
d1	-0.0101***	-0.0293***	-0.0805***	-0.122***	-0.142***
	(-85.74)	(-59.76)	(-72.28)	(-64.93)	(-45.28)
d2		0.000740***	0.00539***	0.0115***	0.0162***
		(40.25)	(58.09)	(47.48)	(25.80)
d3			-0.000110***	-0.000413***	-0.000807***
			(-51.13)	(-36.48)	(-16.32)
d4				0.00000483***	0.0000182***
				(27.26)	(11.07)
d5					-0.000000157***
					(-8.18)
_cons	0.587***	0.647***	0.737***	0.790***	0.810***
	(318.81)	(273.60)	(250.75)	(224.33)	(190.69)
N	199119	199119	199119	199119	199119
R-sq	0.036	0.043	0.056	0.059	0.060
adj. R-sq	0.036	0.043	0.056	0.059	0.060
F	7351.6	4515.6	3921.2	3137.7	2524.4
rmse	0.489	0.487	0.484	0.483	0.483

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

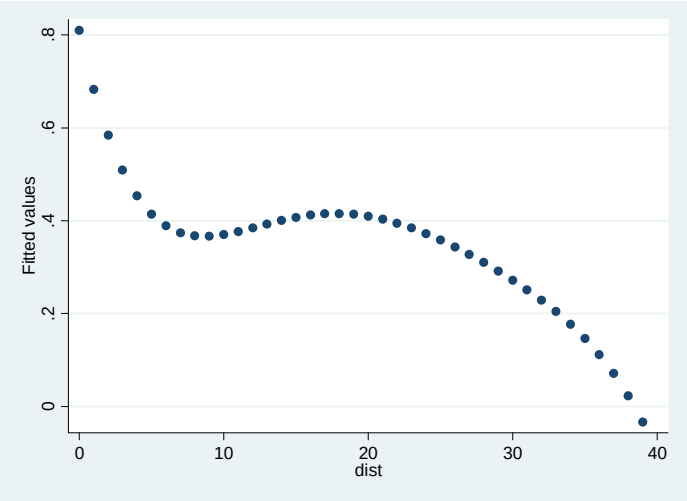
Exploring Functional Forms: *NBA Shots*

And the predicted values?



Nothing like a negative predicted probability of success!

What about the 5th order polynomial?



Better… but we still see success rates increasing between about 8 and 18 feet. Maybe that's what happens?

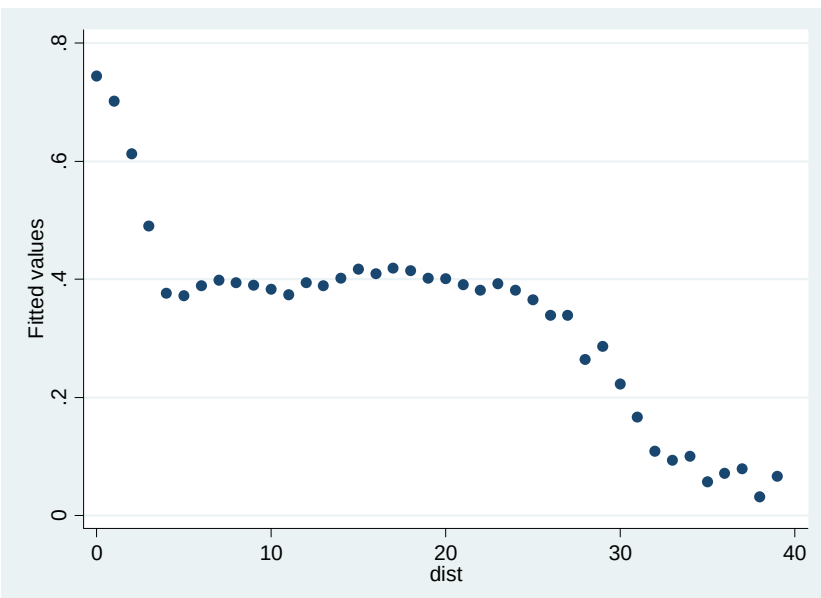
F. Bring on *Fixed Effects* ... dummies for each distance!

Since we have so much data, why are we worrying about functional forms at all? Why not just regress swish on distance dummies (*fixed effects*)?

```
. reg swish i.dist

      Source |       SS       df       MS              Number of obs =  199119
-------------+------------------------------           F( 39,199079) =  337.39
       Model |  3066.54516    39  78.6293631           Prob > F      =  0.0000
    Residual |   46395.115199079  .233048765           R-squared     =  0.0620
-------------+------------------------------           Adj R-squared =  0.0618
       Total |  49461.6602199118  .248403762           Root MSE      =  .48275

------------------------------------------------------------------------------
       swish |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dist |
          1  |  -.0432072   .0071661    -6.03   0.000    -.0572526   -.0291618
          2  |  -.1318084   .0071569   -18.42   0.000    -.1458357   -.1177811
          3  |  -.2540993   .0077454   -32.81   0.000    -.2692801   -.2389185
          4  |   -.368612   .0088803   -41.51   0.000    -.3860171   -.3512069
          5  |  -.3724521   .0091155   -40.86   0.000    -.3903182    -.354586
```

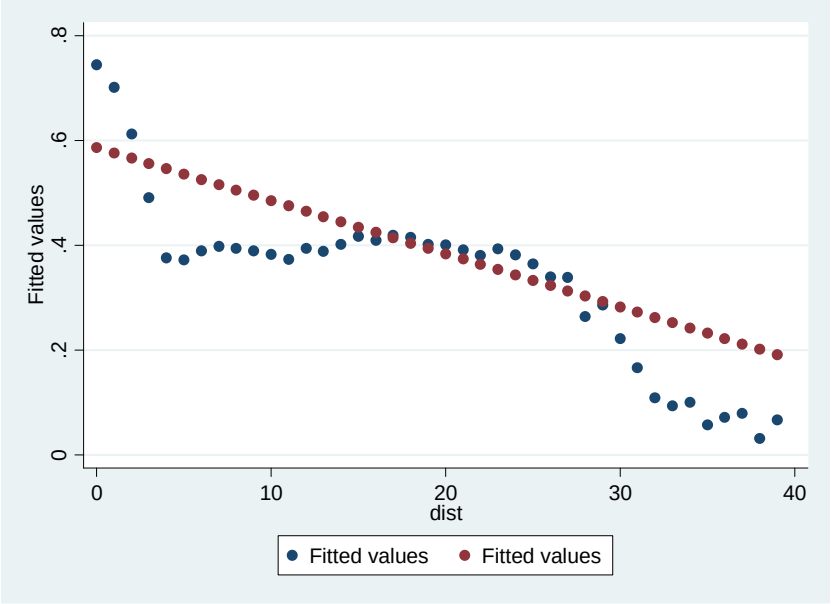
skip some lines

```
         39  |  -.6779811   .0722433    -9.38   0.000    -.8195763    -.536386
             |
       _cons |   .7446478   .0063432   117.39   0.000     .7322153    .7570803
------------------------------------------------------------------------------
```



Coming full circle, let's compare these predictions to what we first saw in the simple OLS model regressing shot success on distance:

Exploring Functional Forms: ***NBA Shots***



Hmmm… Maybe it is worth exploring function forms after all!