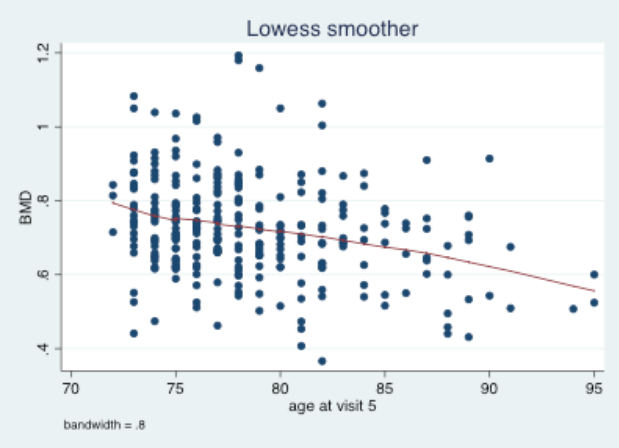
Comments on Biostatistics 208 Lab #6 –2/15/18

Linearity

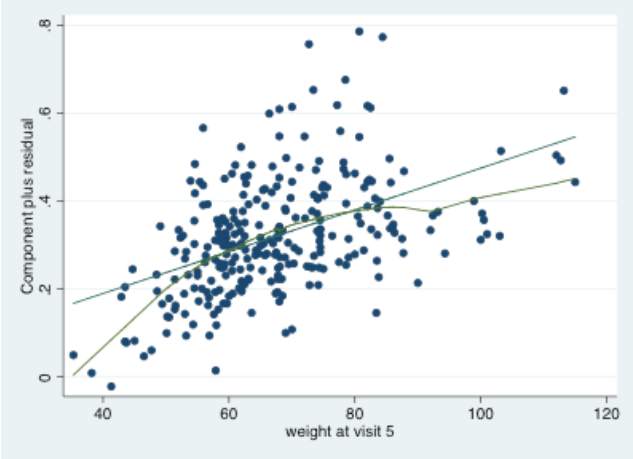
Question 1. Does there appear to be any departure from linearity in the relationship between BMD and age?



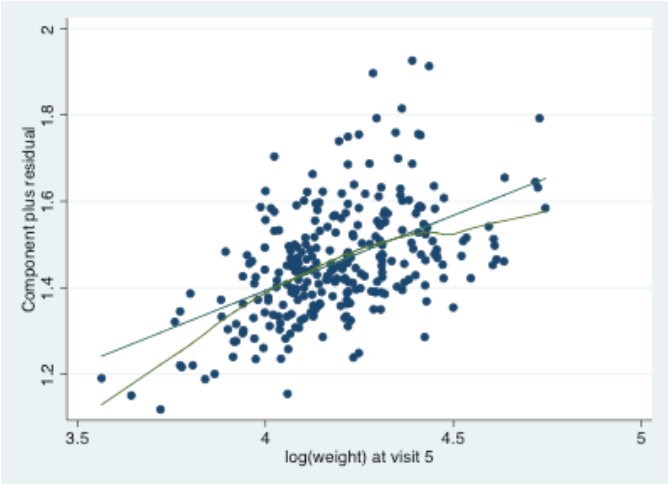
The slight S-curvature is probably not substantial enough to raise concern.

Question 2. Does the CPR plot for log-weight look better? If you think that the log-transformed predictor is a worthwhile improvement, how would you interpret the `nlcom` result?

Below is the CPR plot for the model with untransformed weight. The downward curvature in the CPR plot for weight suggests that a log-transformation might help.



The CPR plot for the log-transform result is shown next.



Log transformation does not really fix the problem, though it is reduced.

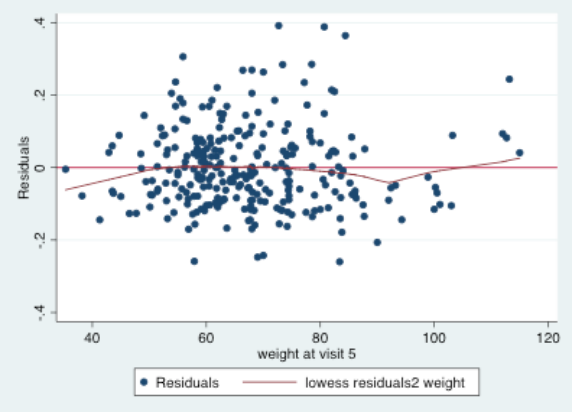
If we did decide to use this transformation, the `nlcom` result shows a 10% increase in weight is associated with an increase of 0.03 gm/cm2.

```
. nlcom _b[lweight]*log(1.1)

       _nl_1:  _b[lweight]*log(1.1)

------------------------------------------------------------------------------
         bmd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |   .0331986   .0034726     9.56   0.000     .0263923    .0400049
------------------------------------------------------------------------------
```

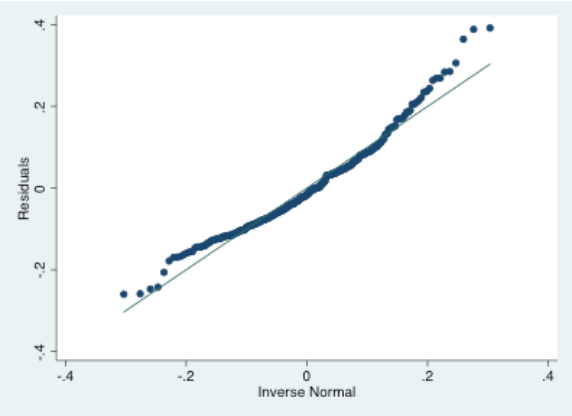
Question 3. Does the RVP plot for the quadratic model look better?



The quadratic term does a better job of addressing the problem with nonlinearity. The drawback is that the quadratic model is more complicated to present (i.e. the effect of weight now involves two coefficients).

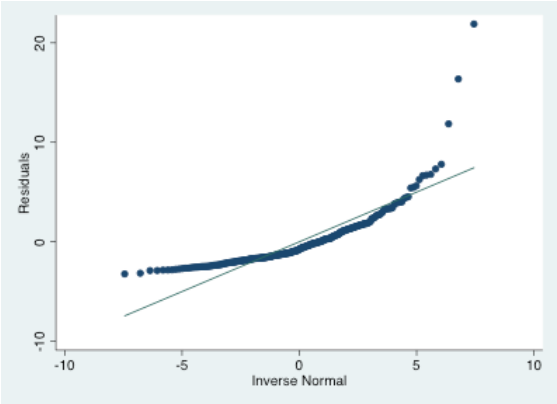
Normality of residuals

Question 4. Is there anything to worry about here, given that this is a moderately large dataset?



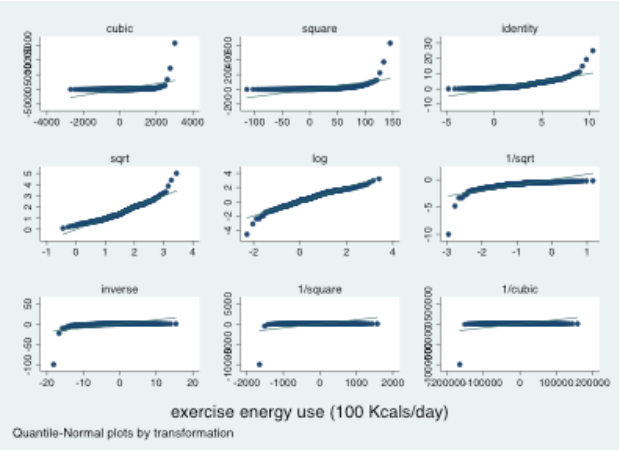
This looks pretty reassuring, although the residuals are slightly right-skewed. It is conventional to analyze BMD without transformation.

Question 5. What do you think about these residuals? Would your judgment be different if we had the full SOF dataset, with over 5,000 observations?

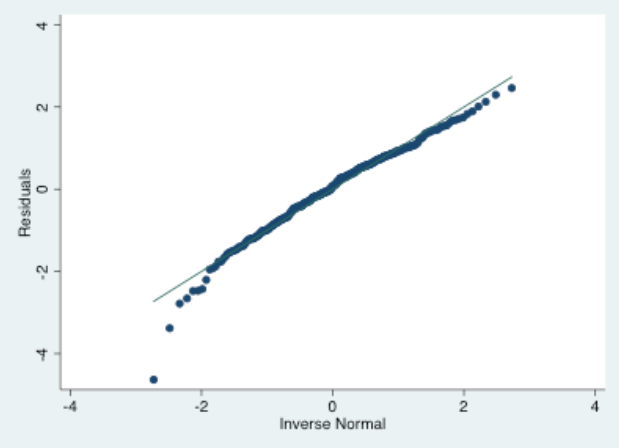


In a dataset of this moderate size, it would be prudent to question the validity of inferences from a conventional regression analysis of an outcome that is so right-skewed; another worry is that the three outliers evident might be influential if those observations also have high leverage. The departure from Normality would be less worrisome in a very large dataset. However, even then, it might be preferable to transform, for the sake of face validity and to maintain efficiency for weaker predictors.

Question 6. Does log transformation of EEU adequately address the departure from normality?



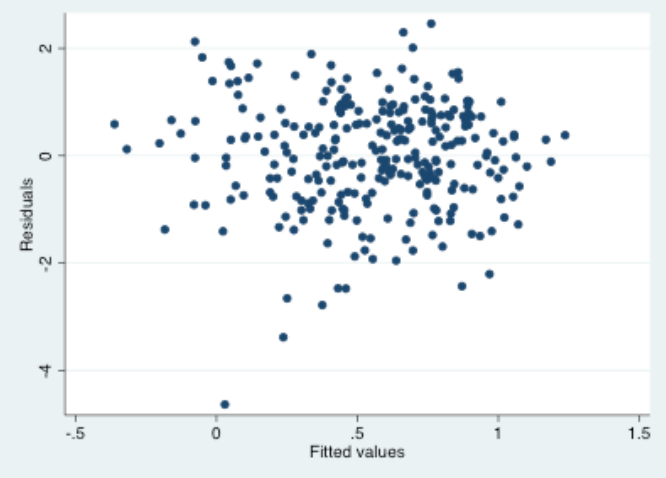
The `qladder` plot suggests that the log transformation would be a reasonable choice.

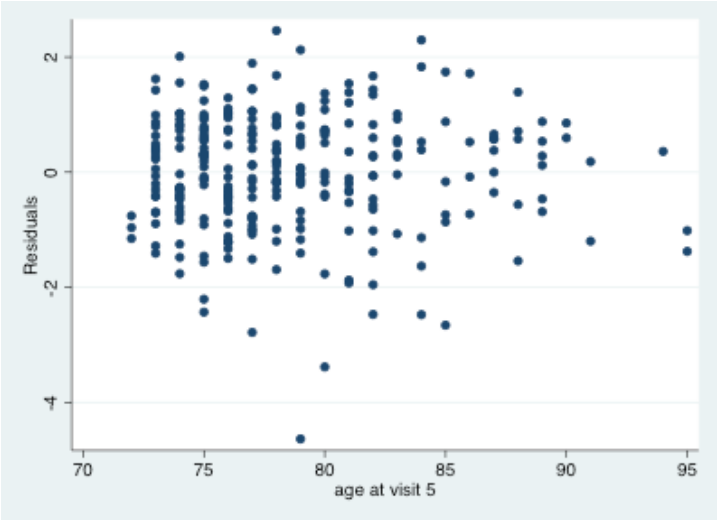


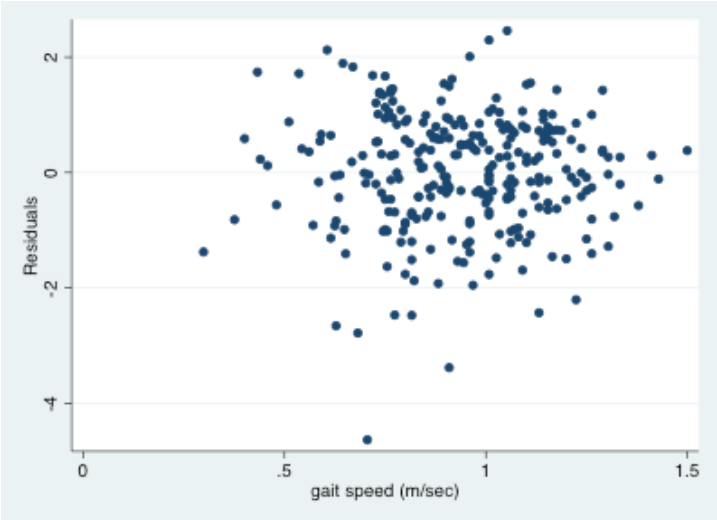
Despite the outlier on the left, the QQ-plot of the residuals looks OK.

Constant variance

Question 7. Do you find any evidence of "heteroskedasticity" (i.e. non-constant variance)?







These all look reassuring – no pronounced funnel shapes in any of them. (We do need to discount some noise in these plots). Also, the table below shows that the SD of the residuals is almost constant across levels of the binary predictor `poorhlth`. Note that less than a quarter of the sample reports being in poor or fair health, so non-constant variance would have been a concern in this case if we had found it, because the subgroups are of very unequal size.

```
. table poorhlth, c(n ln_eeuresid sd ln_eeuresid)
-------------------------------------
poor      |
health by |
self-repo |
rt        |  N(ln_eeu~d)  sd(ln_eeu~d)
----------+--------------------------
       no |           217      .9795251
      yes |            60      1.144814
-------------------------------------
```

Question 8. Are your conclusions affected? How might you report the results of this sensitivity analysis?

Using the rough cutoff of 0.2, as in the lecture, we identify seven somewhat influential observations:

```
. list bmd age lweight _dfbeta_1 if abs(_dfbeta_1) > 0.2 & ~missing(_dfbeta_1)
     +----------------------------------+
     |  bmd   age    lweight   _dfbeta_1 |
     |----------------------------------|
140. | .458    88   4.234107   -.2716317 |
151. | .914    90   4.023564    .3781947 |
200. |  .91    87   4.219508    .2188265 |
     +----------------------------------+

. list bmd age lweight _dfbeta_2 if abs(_dfbeta_2) > 0.2 & ~missing(_dfbeta_2)
     +----------------------------------+
     |   bmd   age    lweight   _dfbet~2 |
     |----------------------------------|
 94. |  1.18    78   4.435567   .2476206 |
108. |  .366    82   3.720862   .2242683 |
231. |  .441    73   3.642836   .2110856 |
262. | 1.193    78   4.390738   .2162125 |
     +----------------------------------+
```

These observations are influential because they have outlying values of BMD as well as age or weight, respectively. However, in the regression results, the coefficient estimates as well as the *p*-values and confidence intervals are qualitatively unchanged after omitting these seven observations. So I would not report the results without them; at most I would say that we checked and things looked OK.

```
. * results with complete data
. reg bmd age lweight
      Source |       SS       df       MS              Number of obs =     278
-------------+------------------------------           F(  2,   275) =   61.64
       Model |  1.61096241     2  .805481207           Prob > F      =  0.0000
    Residual |  3.59338142   275  .013066842           R-squared     =  0.3095
-------------+------------------------------           Adj R-squared =  0.3045
       Total |  5.20434383   277  .018788245           Root MSE      =  .11431

------------------------------------------------------------------------------
         bmd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0047296   .0015137    -3.12   0.002    -.0077094   -.0017497
     lweight |   .3483217   .0364352     9.56   0.000     .2765944     .420049
       _cons |  -.3640786   .2156111    -1.69   0.092    -.7885367    .0603794
------------------------------------------------------------------------------


. * results omitting potentially influential points
. reg bmd age lweight if abs(_dfbeta_1) <= .2 & abs(_dfbeta_2) <= 0.2
      Source |       SS       df       MS              Number of obs =     271
-------------+------------------------------           F(  2,   268) =   60.60
       Model |  1.37868504     2  .689342521           Prob > F      =  0.0000
    Residual |   3.0485926   268  .011375346           R-squared     =  0.3114
-------------+------------------------------           Adj R-squared =  0.3063
       Total |  4.42727765   270  .016397325           Root MSE      =  .10666

------------------------------------------------------------------------------
         bmd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0055429   .0014588    -3.80   0.000     -.008415   -.0026708
     lweight |   .3168664   .0352548     8.99   0.000     .2474548    .3862781
       _cons |  -.1709832   .2094359    -0.82   0.415    -.5833321    .2413656
------------------------------------------------------------------------------
```

Covariate overlap

Question 9. How would you characterize the degree of overlap in the two estrogen groups? Is the overlap good enough to proceed as usual?

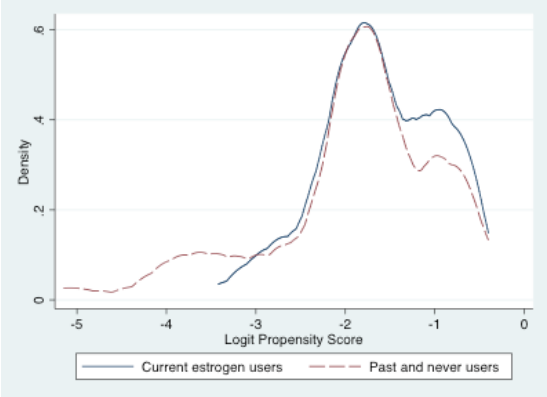
The summary statistics characterizing the groups defined by `ht_current` did not show any gross differences. In the overall check using propensity scores, here is what we found:

```
. logistic ht_current agesp* calsupp

Logistic regression                                     Number of obs   =        278
                                                        LR chi2(5)      =      19.18
                                                        Prob > chi2     =     0.0018
Log likelihood = -121.39847                             Pseudo R2       =     0.0732

------------------------------------------------------------------------------
  ht_current | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agesp1 |   .6542074   .2239392    -1.24   0.215     .3344598    1.279638
      agesp2 |   744.7702   4079.485     1.21   0.227     .0162016    3.42e+07
      agesp3 |   3.17e-07   3.97e-06    -1.20   0.231     7.18e-18    14012.11
      agesp4 |   39901.15   395380.3     1.07   0.285     .0001467    1.09e+13
     calsupp |   2.638898   .8679393     2.95   0.003     1.385034    5.027877
       _cons |   7.19e+12   1.82e+14     1.17   0.241     2.33e-09    2.22e+34
------------------------------------------------------------------------------

. predict logit_pscore, xb
```



The coefficients for the spline variables are not only un-interpretable, but alarming in this case; but this is in fact not a problem. The density plot shows that there is little overlap between the two groups of participants at the left end of the range of the logit propensity scores. To address this, a useful sensitivity analysis would be to rerun the analysis restricted to the region of good overlap: that is, omitting observations with logit propensity scores less than -3. We would also need to figure out who those participants with very low propensity scores are.

Optional: Linear and restricted cubic splines

Question 11. 1. Is there evidence for linear trend and/or departure from it?

```
. * test for linear  trend across categories
. contrast q(1).bmicat

Contrasts of marginal linear predictions
Margins      : asbalanced

------------------------------------------------
             |         df           F        P>F
-------------+----------------------------------
      bmicat |          1       39.46     0.0000
             |
 Denominator |        270
------------------------------------------------
-------------------------------------------------------------
             |   Contrast   Std. Err.     [95% Conf. Interval]
-------------+-----------------------------------------------
      bmicat |
   (linear)  |   .1020557   .0162455      .0700717    .1340396
-------------------------------------------------------------

. * test for departure from linearity
. contrast q(2/4).bmicat

Contrasts of marginal linear predictions
Margins      : asbalanced

------------------------------------------------
             |         df           F        P>F
-------------+----------------------------------
      bmicat |
(quadratic)  |          1        4.65     0.0320
    (cubic)  |          1        0.06     0.7993
  (quartic)  |          1        3.85     0.0509
      Joint  |          3        1.97     0.1186
             |
 Denominator |        270
------------------------------------------------
-------------------------------------------------------------
             |   Contrast   Std. Err.     [95% Conf. Interval]
-------------+-----------------------------------------------
      bmicat |
(quadratic)  |  -.0296468   .0137502      -.056718   -.0025756
    (cubic)  |    .002508   .0098528     -.0168901    .0219061
  (quartic)  |  -.0124878   .0063676     -.0250242    .0000486
-------------------------------------------------------------
```

The first contrast result shows that there is very stong evidence for linear trend across categories ($p < 0.00005$). However, the joint test in the second contrast result reveals little evidence ($p = 0.1186$) for departure from linearity.

Question 10. Is there evidence for changes in the linear spline slopes across regions of the plot?

```
. * linear spline fit
. mkspline bmi1 18.5 bmi2 25 bmi3 30 bmi4 35 bmi5 = bmi
. xi: reg bmd bmi1-bmi5 age i.estrogen

i.estrogen        _Iestrogen_0-2      (naturally coded; _Iestrogen_0 omitted)

      Source |       SS       df       MS              Number of obs =     278
-------------+------------------------------           F(  8,   269) =   16.65
       Model |    1.723868     8    .2154835           Prob > F      =  0.0000
    Residual |  3.48047583   269  .012938572           R-squared     =  0.3312
-------------+------------------------------           Adj R-squared =  0.3113
       Total |  5.20434383   277  .018788245           Root MSE      =  .11375
```

```
------------------------------------------------------------------------------
         bmd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bmi1 |   .0479726   .0288431     1.66   0.097    -.0088143    .1047595
        bmi2 |   .0142805   .0058679     2.43   0.016     .0027277    .0258333
        bmi3 |    .018217   .0051934     3.51   0.001     .0079922    .0284418
        bmi4 |   .0014476   .0067401     0.21   0.830    -.0118225    .0147178
        bmi5 |   .0115979   .0073531     1.58   0.116    -.0028792    .0260749
         age |  -.0039816   .0015351    -2.59   0.010     -.007004   -.0009592
_Iestrogen_1 |   .0427419   .0165707     2.58   0.010     .0101172    .0753666
_Iestrogen_2 |   .0735367   .0190421     3.86   0.000     .0360462    .1110273
       _cons |   .0061371   .5363318     0.01   0.991    -1.049805    1.062079
------------------------------------------------------------------------------

. * test for departure from linear trend
. testparm bmi*, equal

 ( 1)  - bmi1 + bmi2 = 0
 ( 2)  - bmi1 + bmi3 = 0
 ( 3)  - bmi1 + bmi4 = 0
 ( 4)  - bmi1 + bmi5 = 0

       F(  4,   269) =    1.69
            Prob > F =    0.1524
```

There is not much evidence ($p = 0.1524$) for heterogeneity across the slopes.

Question 11. Is there evidence for non-linearity in the cubic spline fit?

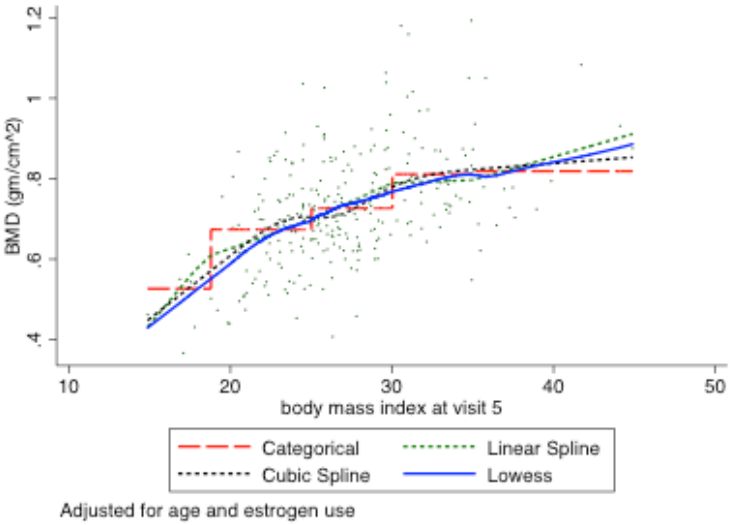
```
. * restricted cubic splines
. xi: reg bmd bmisp* age i.estrogen
i.estrogen        _Iestrogen_0-2      (naturally coded; _Iestrogen_0 omitted)

      Source |       SS       df       MS              Number of obs =     278
-------------+------------------------------           F(  7,   270) =   19.68
       Model |  1.75848253     7   .25121179           Prob > F      =  0.0000
    Residual |  3.44586131   270  .012762449           R-squared     =  0.3379
-------------+------------------------------           Adj R-squared =  0.3207
       Total |  5.20434383   277  .018788245           Root MSE      =  .11297

------------------------------------------------------------------------------
         bmd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      bmisp1 |   .0314653   .0076344     4.12   0.000     .0164347    .0464959
      bmisp2 |  -.1377472    .058356    -2.36   0.019    -.2526379   -.0228566
      bmisp3 |   .6172104   .2652112     2.33   0.021     .0950654    1.139355
      bmisp4 |  -.7514937    .328523    -2.29   0.023    -1.398286   -.1047012
         age |  -.0038331   .0015263    -2.51   0.013     -.006838   -.0008281
_Iestrogen_1 |   .0361593   .0162605     2.22   0.027     .0041458    .0681729
_Iestrogen_2 |   .0732025   .0189018     3.87   0.000     .0359889    .1104161
       _cons |   .2581214   .2166109     1.19   0.234    -.1683398    .6845826
------------------------------------------------------------------------------
. * test for overall BMI effect
. testparm bmisp*
 ( 1)  bmisp1 = 0
 ( 2)  bmisp2 = 0
 ( 3)  bmisp3 = 0
 ( 4)  bmisp4 = 0
     F(  4,   270) =   22.07
          Prob > F =    0.0000
. * test for departure from linearity
. testparm bmisp2 bmisp3 bmisp4
 ( 1)  bmisp2 = 0
 ( 2)  bmisp3 = 0
 ( 3)  bmisp4 = 0
       F(  3,   270) =    3.19
            Prob > F =    0.0242
```

The `testparm` results show that there is strong evidence for an overall BMI effect, and some evidence ($p = 0.024$) for departure from linearity in the cubic spline fit.

Question 12. How do the linear and cubic spline fits compare to the categorical and lowess fits?



Both spline fits do a good job of following the fully nonparametric lowess fit, and are more realistic than the categorical fit, especially the cubic spline fit. Explaining them would be harder, however, and would probably require a plot. In contrast the categorical fit is easily summarized in a table.