Assessing the scale of continuous model covariates

HL Chapter 4 – part 1

Continuous variables

Continuous model covariates

- Continuous model covariates are assumed to be linear in the logit
- Example: x=age, y=adverse birth outcome
  - Increase in age of, say, 5 years has the same effect on the logit no matter what age we start at
  - The effect of a 5 year age increase on the logit is the same among 14 year olds and among 60 year olds
  - This is biologically incorrect

Continuous model covariates

- Solution:
  - Categorize the continuous variable using biologically meaningful cutpoints; or
  - Assess the scale of the variable and transform, if necessary

Example

- Hypothetical data
- Dependent variable=*y*
- Independent variable=*x*

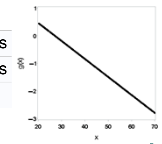
SAS

```
libname sdat 'C:\ERHS642';
/*
data sdat.scale_example; set scale_example; run;
proc contents data=sdat.scale_example; run;
*/

data scale_example; set sdat.scale_example; run;
```

Conclusions for linear *x*?

Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
x	10.0000	0.522	0.442	0.615



- For every 10 point increase in *x*, the probability of *y* decreases by approximately 50%
- An increase in *x* has the same effect on the logit no matter what *x* we start at

SAS

```
* x modeled using linear scale *;

proc logistic descending data=scale_example;
  model y=x/clodds=wald;
  units x=10;
  output out=pdat xbeta=g p=pihat;
run;

proc sort data=pdat; by x; run;
```

SAS

```
axis1 minor=none label=(f=swiss h=2.5 'x');
axis2 minor=none label=(f=swiss h=2.5 a=90 'g(x)');
goptions FTEXT=swissb HTEXT=2.0 HSIZE=6 in
VSIZE=6 in;
symbol1 c=black v=dot;
axis1 minor=none label=(f=swiss h=2.5 'x');
axis2 minor=none label=(f=swiss h=2.5 a=90 'g(x)');
proc gplot data=pdat;
        plot g*x/overlay haxis=axis1 vaxis=axis2;
run; quit;
```

Is this conclusion correct?

Assessing scale - Approaches

- Splines 
- Categorizing 
- Fractional polynomials 

Splines

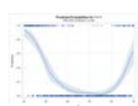
Splines

- Effect of *x* on the logit is not forced to be linear
- Effect of *x* on the logit is allowed to “follow” the data

Splines – Number of knots

- Knots = connection points
- How many?
  - 3-5
  - More knots means more flexibility
  - More variables are needed to model more knots

Few connection points (knots)



Many connection points (knots)

Splines – Spacing of knots

- Could space evenly
- Could use quartiles
- Could use percentiles shown in HL Table 4.1

Splines – Selecting connections

- Select connections between knots
  - Constant connection
  - Linear connection
  - Cubic connection
  - Others

Example

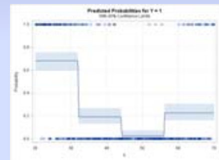
- Let’s choose 3 knots based on quartiles of *x*:
  - 32, 44 and 56 (Could choose different knots)
- Let’s try three different connections

SAS

```
proc univariate data=scale_example; var x; run;
```

Example: Constant connection/SAS

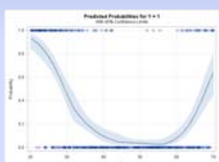
```
proc logistic descending data=scale_example;
 effect xs=spline(x/knotmethod=list(32 44 56)
                          basis=tpf(noint) degree=0);
 model y=xs;
 effectplot;
run;
```



Same as a design variable plot

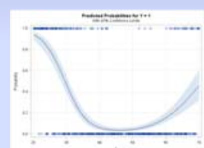
Example: Linear connection/SAS

```
proc logistic descending data=scale_example;
 effect xs=spline(x/knotmethod=list(32 44 56)
                          basis=tpf(noint) degree=1);
 model y=xs;
 effectplot;
run;
```



Example: Cubic connection/SAS

```
proc logistic descending data=scale_example;
 effect xs=spline(x/knotmethod=list(32 44 56)
                          basis=tpf(noint) naturalcubic);
 model y=xs;
 effectplot;
run;
```



Recall conclusions for linear *x*

- In this example, using *x* linear we concluded
  - For every 10 point increase in *x*, the probability of *y* decreases by approximately 50%
  - An increase in *x* has the same effect on the logit no matter what *x* we start at

Conclusions for splines

- In this example, using splines we conclude
  - A 10 point increase in *x* results in a sharp decrease in the probability of *y* for younger ages
  - A 10 point increase in *x* has little effect on the probability of *y* for ages in the middle range
  - A 10 point increase in *x* results in a sharp increase in the probability of *y* for older ages
  - The effect of an increase in *x* on the logit depends on what *x* we start at

Linear x vs. splines

- In this example, using x linear we drew the wrong conclusions

Splines pros and cons

- Splines pros
  - Easy to use
  - Quick method to check for non-linearity in the logit
  - Can compare different splines and select "the best" model, i.e. the model with the smallest deviance

Splines pros and cons

- Splines cons
  - Too many choices (knots, connections)
  - Still, many possible knots and connections are not tested
  - Must check for statistical significance! The shape of the plot may just be noise.
  - Very difficult to obtain interpretable ORs; therefore, finding the best spline model is not very useful in practice

Categorizing

Categorizing continuous variables

- Based on the spline plots it may be possible to establish cutpoints
- Alternatively, quartiles can be used
- For resulting categorical variables with more than 2 categories, design variables must be used in the model

Conclusions for categorical x

- In this example, let's use the "spline knots" (i.e. the quartiles of x) as cutpoints

Contrast	OR	Confidence Limits		P-value
32-44 vs. 20-32	0.1093	0.0607	0.1971	<.0001

- Compared to those with x between 20 and 32
  - The probability of y is decreased about 90% among those with x between 32 and 44

Conclusions for categorical *x*

Contrast	OR	Confidence Limits		P-value
44-56 vs. 20-32	0.0117	0.00349	0.0390	<.0001

- Compared to those with *x* between 20 and 32
  - The probability of *y* is decreased about 99% among those with *x* between 44 and 56

Conclusions for categorical *x*

Contrast	OR	Confidence Limits		P-value
56-70 vs. 20-32	0.1365	0.0781	0.2386	<.0001

- Compared to those with *x* between 20 and 32
  - The probability of *y* is decreased about 85% among those with *x* between 56 and 70

Conclusions for categorical *x*

Contrast	OR	Confidence Limits		P-value
44-56 vs. 32-44	0.1067	0.0311	0.3659	0.0004

- Compared to those with *x* between 32 and 44
  - The probability of *y* is decreased about 90% among those with *x* between 44 and 56

Conclusions for categorical *x*

Contrast	OR	Confidence Limits		P-value
56-70 vs. 44-56	11.6959	3.4606	39.5292	<.0001

- Compared to those with *x* between 44 and 56
  - The probability of *y* is increased almost 12fold among those with *x* between 56 and 70

Conclusions for categorical *x* Summary

- As *x* increases, the probability of *y* drops sharply
- However, between *x*=32 and *x*=56, the decrease in the probability of *y* tapers off
- Between *x*=56 and *x*=70, the probability of *y* increases slightly
- After *x*=70, the probability of *y* increases sharply
- This agrees with the spline plots

Categorizing pros and cons

- Categorizing pros
  - Easy to do
  - If meaningful cutpoints are chosen, results and conclusions are meaningful
- Categorizing cons
  - Increases the number of variables in the model
  - May result in misclassification

SAS: Change data step as follows

```
data scale_example; set sdat.scale_example;

  * categorize x for design variable plot          *;
  * category boundaries from proc univariate below *;
        if 20<=x<32 then x_c=1;
  else if 32<=x<44 then x_c=2;
  else if 44<=x<56 then x_c=3;
  else if 56<=x<70 then x_c=4;
run;
```

SAS: Then…

```
proc logistic descending data=scale_example;
  class x_c/param=ref ref=first;
  model y=x_c;
  contrast '32-44 vs. 20-32' x_c  1  0 0/estimate=exp;
  contrast '44-56 vs. 20-32' x_c  0  1 0/estimate=exp;
  contrast '56-70 vs. 20-32' x_c  0  0 1/estimate=exp;
  contrast '44-56 vs. 32-44' x_c -1  1 0/estimate=exp;
  contrast '56-70 vs. 44-56' x_c  0 -1 1/estimate=exp;
run;
```

fp method

Fractional polynomial (fp) procedure

- Model the continuous variable using many different scales (e.g., linear, quadratic, cubic, log-transformed, etc.)
- Compare the different models and select "the best" model, i.e. the model with the smallest deviance
- Transform the continuous variable accordingly

fp procedure: "One power" transformations

- $x^{-2} = {}^{1}/_{x^2}$
- $x^{-1} = {}^{1}/_{x}$
- $x^{-0.5} = {}^{1}/_{\sqrt{x}}$
- $x^{0} = \ln(x)$
- $x^{0.5} = \sqrt{x}$
- x^{2}
- x^{3}

fp procedure: "Two power" transformations

- x^{p_1} <u>and</u> x^{p_2} if $p_1 \neq p_2$
  - Example: For p1=2 and p2=3, use x^2 and x^3
- x^{p_1} <u>and</u> $x^{p_2}\ln(x)$ if $p_1 = p_2$
  - Example: For p1=2 and p2=2, use x^2 and $x^2\ln(x)$

Example: Results from fp procedure

Dev_linear	e_fp1	Dev_fp1	e1_fp2	e2_fp2	Dev_fp2	p_lin_fp1	p_lin_fp2	p_fp1_fp2
521.007	-2	452.668	2	2	386.868	0.0000	0.0000	0.0000

Transformation	Deviance	P-values for comparisons	
Linear	521.007		
		0.0000	0.0000
$x^{-2} = {}^{1}/_{x^2}$	452.668		
		0.0000	
x^2 and $x^2 \ln(x)$	386.868		

Interpretation of fp results

- ${}^{1}/_{x^2}$ better than linear (p<0.0001)
- x^2 and $x^2 \ln(x)$ better than linear (p<0.0001)
- x^2 and $x^2 \ln(x)$ better than ${}^{1}/_{x^2}$ (p<0.0001)

Best fp-"one power" transformed *x* (${}^{1}/_{x^2}$): Logit difference

- Let's use a 10 year increase in age
- Logit difference

$$g(x+10) - g(x)$$

$$= \left\{\beta_0 + \beta_1 \left(\frac{1}{(x+10)^2}\right)\right\} - \left\{\beta_0 + \beta_1 \left(\frac{1}{(x)^2}\right)\right\}$$

$$= \beta_1 \left(\frac{1}{(x+10)^2}\right) - \beta_1 \left(\frac{1}{(x)^2}\right)$$

$$= \beta_1 \left(\frac{1}{(x+10)^2} - \frac{1}{x^2}\right)$$

Best fp-"one power" transformed *x* (${}^{1}/_{x^2}$): Logit difference

At

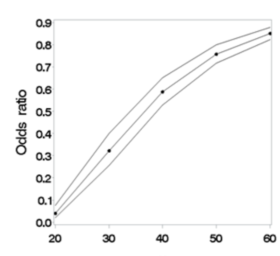
- $x = 20: g(30) - g(20) = \beta_1 \left(\frac{1}{30^2} - \frac{1}{20^2}\right) = -0.00139\beta_1$
- $x = 30: g(40) - g(30) = \beta_1 \left(\frac{1}{40^2} - \frac{1}{30^2}\right) = -0.00049\beta_1$
- $x = 40: g(50) - g(40) = \beta_1 \left(\frac{1}{50^2} - \frac{1}{40^2}\right) = -0.00023\beta_1$
- $x = 50: g(60) - g(50) = \beta_1 \left(\frac{1}{60^2} - \frac{1}{50^2}\right) = -0.00012\beta_1$
- $x = 60: g(70) - g(60) = \beta_1 \left(\frac{1}{70^2} - \frac{1}{60^2}\right) = -0.00007\beta_1$

Best fp-"one power" transformed *x* (${}^{1}/_{x^2}$): Contrasts and ORs

x	SAS Contrasts (${}^{1}/_{x^2} = \boldsymbol{fp1}$)	OR	Confidence limits		P-value
30 vs. 20	$fp1\ -0.00139$	0.0399	0.0212	0.0748	<.0001
40 vs. 30	$fp1\ -0.00049$	0.3211	0.2572	0.4009	<.0001
50 vs. 40	$fp1\ -0.00023$	0.5867	0.5287	0.6512	<.0001
60 vs. 50	$fp1\ -0.00012$	0.7572	0.7171	0.7995	<.0001
70 vs. 60	$fp1\ -0.00007$	0.8502	0.8237	0.8776	<.0001

The effect on y of a 10 point increase in *x* depends on *x*

Best fp-"one power" transformed *x* (${}^{1}/_{x^2}$): ORs



The effect on y of a 10 point increase in *x* depends on *x*

Best fp-“one power” transformed x ($^1/_{x^2}$): Conclusions

- A 10 point increase in *x* decreases the risk of *y* by
  - ≈ 95% at x=20
  - ≈ 70% at x=30
  - ≈ 40% at x=40
  - ≈ 25% at x=50
  - ≈ 15% at x=60
- Based on the spline plots we know that the last two interpretations are incorrect

Best fp-“two power” transformed x (x^2 and $x^2\ln(x)$): Logit difference 1

- Let’s use a 10 year increase in age
- Logit difference

$$g(x+10)-g(x)$$

$$=\{\beta_0+\beta_1(x+10)^2+\beta_2(x+10)^2\ln(x+10)\}$$
$$-\{\beta_0+\beta_1x^2+\beta_2x^2\ln(x)\}$$

$$=\beta_1\{(x+10)^2-x^2\}+\beta_2\{(x+10)^2\ln(x+10)-x^2\ln(x)\}$$

Best fp-“two power” transformed x (x^2 and $x^2\ln(x)$): Logit difference 1

At

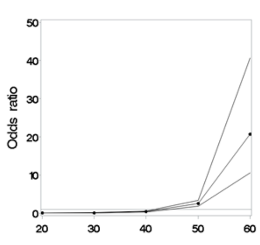
- $x=20: g(30)-g(20)$
  $=\beta_1\{30^2-20^2\}+\beta_2\{30^2\ln(30)-20^2\ln(20)\}$
  $=500\beta_1+1862.8\beta_2$
- Etc.
- $x=60: g(70)-g(60)=$
  $\beta_1\{70^2-60^2\}+\beta_2\{70^2\ln(70)-60^2\ln(60)\}=$
  $1300\beta_1+6078\beta_2$

Best fp-“two power” transformed x (x^2 and $x^2\ln(x)$): Contrasts and ORs 1

x	SAS Contrasts ($x^2=fp2a$, $\ln(x)=fp2b$)	OR	Confidence limits		P-value
30 vs. 20	fp2a 500 fp2b 1862.8	0.0465	0.0269	0.0802	<.0001
40 vs. 30	fp2a 700 fp2b 2841.1	0.1089	0.0740	0.1604	<.0001
50 vs. 40	fp2a 900 fp2b 3877.9	0.4301	0.3587	0.5158	<.0001
60 vs. 50	fp2a 1100 fp2b 4959.6	2.5346	1.8776	3.4216	<.0001
70 vs. 60	fp2a 1300 fp2b 6078.0	20.7175	10.5528	40.6730	<.0001

The effect on *y* of a 10 point increase in *x* depends on *x*

Best fp-“two power” transformed x (x^2 and $x^2\ln(x)$): ORs 1



The effect on y of a 10 point increase in *x* depends on *x*

Best fp-“two power” transformed x (x^2 and $x^2\ln(x)$): Conclusions 1

- A 10 point increase in *x* decreases the risk of *y* by
  - ≈ 95% at x=20
  - ≈ 90% at x=30
  - ≈ 55% at x=40
- A 10 point increase in *x* increases the risk of *y*
  - ≈ 2.5fold at x=50
  - ≈ 21fold at x=60

Best fp-“two power” transformed x (x^2 and $x^2\ln(x)$): Logit difference 2

- Let’s compare different ages to age 20
- Logit difference

$$g(x) - g(20)$$

$$= \{\beta_0 + \beta_1 x^2 + \beta_2 x^2 \ln(x)\} - \{\beta_0 + \beta_1 20^2 + \beta_2 20^2 \ln(20)\}$$

$$= \beta_1\{x^2 - 20^2\} + \beta_2\{x^2 \ln(x) - 20^2 \ln(20)\}$$

Best fp-“two power” transformed x (x^2 and $x^2\ln(x)$): Logit difference 2

At

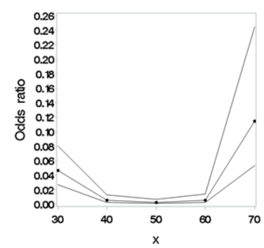
- $x = 30: g(30) - g(20)$
$$= \beta_1\{30^2 - 20^2\} + \beta_2\{30^2 \ln(30) - 20^2 \ln(20)\}$$
$$= 500\beta_1 + 1862.8\beta_2$$
- Etc.
- $x = 70: g(70) - g(20)$
$$= \beta_1\{70^2 - 20^2\} + \beta_2\{70^2 \ln(70) - 20^2 \ln(20)\}$$
$$= 4500\beta_1 + 19{,}619.3\beta_2$$

Best fp-“two power” transformed x (x^2 and $x^2\ln(x)$): Contrasts and ORs 2

x	SAS Contrasts ($x^2 = fp2a$, $ln(x) = fp2b$)	OR	Confidence limits		P-value
30 vs. 20	fp2a 500 fp2b 1862.8	0.04650	0.02690	0.0802	<.0001
40 vs. 20	fp2a 1200 fp2b 4703.9	0.00506	0.00200	0.0128	<.0001
50 vs. 20	fp2a 2100 fp2b 8581.8	0.00218	0.00075	0.00635	<.0001
60 vs. 20	fp2a 3200 fp2b 13541.3	0.00552	0.00216	0.0141	<.0001
70 vs. 20	fp2a 4500 fp2b 19619.3	0.11430	0.05340	0.2444	<.0001

The effect on *y* of a 10 point increase in *x* depends on *x*

Best fp-“two power” transformed x (x^2 and $x^2\ln(x)$): ORs 2



The effect on y of a 10 point increase in *x* depends on *x*

Best fp-“two power” transformed x (x^2 and $x^2\ln(x)$): Conclusions 2

- Compared to subjects with *x*=20 the probability of *y* is decreased by
  - ≈ 95.0% for subjects with *x*=20
  - ≈ 99.5% for subjects with *x*=30
  - ≈ 99.8% for subjects with *x*=40
  - ≈ 99.5% for subjects with *x*=50
  - ≈ 90.0% for subjects with *x*=60

fp procedure pros and cons

- Pros
  - Automated
  - Lots of possible transformations tested

fp procedure pros and cons

- Cons
  - Best transformation may be very complex and hard to interpret/explain to lay persons
  - Based on statistical significance
  - Many possible transformations are not tested
  - Values of the variable of interest must be > 0 (transformations include division by the variable and the natural log (ln) of the variable)

SAS fp macro

```
** Macro for fp assessment **;
%macro fp1(dset,y,var,lb,p1);
 %do %until(&p1=7);
 %put ***** &p1 *****;
 ODS output FitStatistics = mfs;
 data fpdat; set &dset; if &var>&lb; pc=&p1/2;
  if pc ne 0 then F1=&var**pc;
  else if pc = 0 then F1=log(&var);
 run;
```

Placeholders for data set name (dset), outcome variable (y) and variable being tested for scale (var). Ignore lb and p1.

F1 represents the variable being tested for scale

SAS fp macro, cont.

```
proc logistic descending data=fpdat;
 *______________________________;
 model &y=F1; run;
*_______________________________;

data mfs; set mfs; if criterion='-2 Log L';
 drop Criterion InterceptOnly; run;
proc append data=mfs base=tres; run;
proc datasets; delete fpdat mfs; run;
quit;
```

F1 represents the variable being tested for scale

SAS fp macro, cont.

```
%let p1=%eval(&p1+1);
%end;
%mend fp1;
*________________________________;
%fp1(scale_example,y,x,0,-4);
*________________________________;
```

Enter actual names of dataset (here, scale_example), outcome variable (here, *y*) and variable being tested for scale (here, *x*).
Don't change 0 or -4.

SAS fp macro, cont.

```
data pvals; do p1=-4 to 6; output; end; run;
data pvals; set pvals; p1=p1/2; run;
data tres; merge pvals tres; if p1 in (-1.5, 1.5, 2.5) then
delete; run;
proc sort data=tres; by InterceptAndCovariates; run;
data tres; set tres; if _N_=1 or p1=1; run;
```

SAS fp macro, cont.

```
%macro fp2(dset,y,var,lb,p1,p2);
 %do %until(&p1=7);
  %do %until(&p2=7);
 %put ***** &p1 &p2 *****;
 ODS output FitStatistics = mfs;
```

Placeholders for data set name (dset), outcome variable (y), variable being tested for scale (var). Ignore lb, p1 and p2.

SAS fp macro, cont.

```
data fpdat; set &dset;
 if &var>&lb; pc1=&p1/2; pc2=&p2/2;
 if pc1 ne 0 then F1=&var**pc1;
 else if pc1 = 0 then F1=log(&var);
 if pc1 ne pc2 then do;
  if pc2 ne 0 then F2=&var**pc2;
  else if pc2 = 0 then F2=log(&var); end;
 if pc1=pc2 then F2=F1*log(&var);
run;
```

F1 and F2 represent the variable being tested for scale

SAS fp macro, cont.

```
proc logistic descending data=fpdat;
*______________________;
 model &y=F1 F2;
*______________________;
run;
data mfs; set mfs; if criterion='-2 Log L';
drop Criterion InterceptOnly; run;
 proc append data=mfs base=tres2; run;
 proc datasets; delete fpdat mfs; run;
 quit;
```

F1 and F2 represent the variable being tested for scale

SAS fp macro, cont.

```
%let p2=%eval(&p2+1);
 %end;
 %let p2=%eval(-4);
 %let p1=%eval(&p1+1);
 %end;
%mend fp2;
*__________________________;
%fp2(scale_example,y,x,0,-4,-4);
*__________________________;
```

Enter actual names of dataset (here, scale_example), outcome variable (here, *y*) and variable being tested for scale (here, *x*). Don't change 0 or -4.

SAS fp macro, cont.

```
data pvals2; do p1=-4 to 6; do p2=-4 to 6; output;end;
end; run;
data pvals2; set pvals2; p1=p1/2; p2=p2/2; run;
data tres2; merge pvals2 tres2;
 if p1 in (-1.5, 1.5, 2.5) or p2 in (-1.5, 1.5, 2.5) then
delete; run;
proc sort data=tres2; by InterceptAndCovariates; run;
data tres2; set tres2; if _N_=1; run;
```

SAS fp macro, cont.

```
data comb; set tres tres2; run;
data c1; set comb; if p1=1 and p2=.;
 rename InterceptAndCovariates=Dev_linear;
 drop p1 p2; run;
data c2; set comb; if p1 ne 1 and p2=.;
 rename InterceptAndCovariates=Dev_fp1;
 rename p1=e_fp1; drop p2; run;
data c3; set comb; if p2 ne .;
  rename InterceptAndCovariates=Dev_fp2;
  rename p1=e1_fp2; rename p2=e2_fp2; run;
```

SAS fp macro, cont.

```
data c;
 merge c1 c2 c3;
 diff_lin_fp1=Dev_linear-Dev_fp1;
 diff_lin_fp2=Dev_linear-Dev_fp2;
 diff_fp1_fp2=Dev_fp1-Dev_fp2;

 p_lin_fp1=1-probchi(diff_lin_fp1,1);
 p_lin_fp2=1-probchi(diff_lin_fp2,3);
 p_fp1_fp2=1-probchi(diff_fp1_fp2,2);
run;
```

SAS fp macro, cont.

```
proc print noobs data=c;
  var Dev_linear e_fp1 Dev_fp1 e1_fp2 e2_fp2 Dev_fp2
                                        p_lin_fp1 p_lin_fp2 p_fp1_fp2;
  format p_lin_fp1 p_lin_fp2 p_fp1_fp2 6.4;
run;

proc datasets;
  delete tres tres2 pvals pvals2 comb c c1 c2 c3;
run; quit;
* End macro for fp assessment *;
```

SAS fp, cont.

- After establishing the best transformations, continue as follows:

SAS: Add transformations to data step

```
data scale_example; set sdat.scale_example;
  * categorize x for design variable plot         *;
  * category boundaries from proc univariate below *;
      if 20<=x<32 then x_c=1;
  else if 32<=x<44 then x_c=2;
  else if 44<=x<56 then x_c=3;
  else if 56<=x<70 then x_c=4;
 * transformations suggested by fp procedure *;
  fp1=1/(x**2);
  fp2a=x**2;
  fp2b=x**2*log(x);
run;
```

SAS: Contrasts – one power

```
* x modeled using one power fp transformation *;
proc logistic descending data=scale_example;
  model y=fp1;
  contrast '30 vs. 20' fp1 -0.00139/estimate=exp;
  contrast '40 vs. 30' fp1 -0.00049/estimate=exp;
  contrast '50 vs. 40' fp1 -0.00023/estimate=exp;
  contrast '60 vs. 50' fp1 -0.00012/estimate=exp;
  contrast '70 vs. 60' fp1 -0.00007/estimate=exp;
run;
```

SAS: OR plot – one power

```
* Creation of odds ratio plot for 10 point increase in x
modeled using one power fp transformation *;
data fp1plot;  input x OR CIL CIU;
cards;
20 0.0399    0.0212    0.0748
30 0.3211    0.2572    0.4009
40 0.5867    0.5287    0.6512
50 0.7572    0.7171    0.7995
60 0.8502    0.8237    0.8776
run;
```

SAS: OR plot – one power

```
axis1 minor=none label=(f=swiss h=2.5 'x');
axis2 minor=none label=(f=swiss h=2.5 a=90 'Odds ratio');
goptions FTEXT=swissb HTEXT=2.0 HSIZE=6 in
VSIZE=6 in;
symbol1 c=black i=join;
symbol2 c=black i=join;
Symbol3 c=black i=join;
proc gplot data=fp1plot;
        plot (OR CIL CIU)*x/overlay haxis=axis1
vaxis=axis2 vref=1;
run; quit;
```

SAS: Contrasts 1 – two powers

```
* x modeled using two power fp transformation, contrasts
for 10 point increase in x *;
proc logistic descending data=scale_example;
 model y=fp2a fp2b;
 contrast '30 vs. 20' fp2a  500 fp2b 1862.8/estimate=exp;
 contrast '40 vs. 30' fp2a  700 fp2b 2841.1/estimate=exp;
 contrast '50 vs. 40' fp2a  900 fp2b 3877.9/estimate=exp;
 contrast '60 vs. 50' fp2a 1100 fp2b 4959.6/estimate=exp;
 contrast '70 vs. 60' fp2a 1300 fp2b 6078.0/estimate=exp;
run;
```

SAS: OR plot 1 – two powers

```
* Creation of odds ratio plot for 10 point increase in x
modeled using two power fp transformation *;
data fp2plot;  input x OR CIL CIU;
cards;
20 0.0465    0.0269     0.0802
30 0.1089    0.0740     0.1604
40 0.4301    0.3587     0.5158
50 2.5346    1.8776     3.4216
60 20.7175   10.5528    40.673
run;
```

SAS: OR plot 1 – two powers

```
proc gplot data=fp2plot;
        plot (OR CIL CIU)*x/overlay haxis=axis1
vaxis=axis2 vref=1;
run; quit;
```

SAS: Contrasts 2 – two powers

```
* x modeled using two power fp transformation, contrasts
for comparison to x=20 *;
proc logistic descending data=scale_example;
 model y=fp2a fp2b;
 contrast '30 vs. 20' fp2a  500 fp2b  1862.8/estimate=exp;
 contrast '40 vs. 20' fp2a 1200 fp2b  4703.9/estimate=exp;
 contrast '50 vs. 20' fp2a 2100 fp2b  8581.8/estimate=exp;
 contrast '60 vs. 20' fp2a 3200 fp2b 13541.3/estimate=exp;
 contrast '70 vs. 20' fp2a 4500 fp2b 19619.3/estimate=exp;
run;
```

SAS: OR plot 2 – two powers

```
* Creation of odds ratio plot for comparison to x=20 with x
modeled using two power fp transformation *;
data fp2plot;  input x OR CIL CIU;
cards;
30 0.0465    0.0269     0.0802
40 0.00506   0.0020     0.0128
50 0.00218   0.000748   0.00635
60 0.00552   0.00216    0.0141
70 0.1143    0.0534     0.2444
run;
```

SAS: OR plot 2 – two powers

```
proc gplot data=fp2plot;
        plot (OR CIL CIU)*x/overlay haxis=axis1
vaxis=axis2 vref=1;
run; quit;
```

Modeling variables with many zeros

Modeling variables with many zeros

Example: Number of cigarettes smoked per day

- Continuous variable among smokers
- 0 for all non-smokers

How should this variable be modeled?

Ideas

Idea 1:

- Categorize the variable with many zeros
- There may be biologically meaningful cutpoints
- Otherwise, median or quartiles of the non-zero part of the variable can be used as cutpoints

Idea 2:

- Dichotomize the variable with many zeros (here, number of cigarettes smoked per day)
- 0=non-smoker and 1=smoker
- Use dichotomous <u>and</u> continuous variable

Idea 2, Example

- lc = lung cancer
- cigs = number of cigarettes smoked per day
- smo = smoking status

smo=0 (non-smoker) if cigs=0
smo=1 (smoker) if cigs>0

Proportion with cigs=0

- Proc freq: Almost 30% of observations have cigs=0
- Cigs is a variable with many zeros

smo	Frequency	Percent
0	180	29.13

SAS:
```
data cigs; set sdat.cigs; rename lung_cancer=lc; run;
proc freq data=cigs; tables smo; run;
```

- Must assess the scale of the non-zero part of cigs

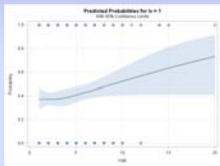
Scale of the non-zero part: Splines

- Choose 3 knots based on quartiles of *cigs*:
  - 2, 3 and 5
- Choose natural cubic connection

Scale of the non-zero part: Splines

```
proc logistic descending data=cigs;  where cigs>0;
  effect xs=spline(cigs/knotmethod=list(2 3 5)
                          basis=tpf(noint) naturalcubic);
  model lc=xs;
  effectplot;
run;
```

Looks fairly linear



Scale of the non-zero part

- In this example, we can keep *cigs* linear
- If the spline suggests the necessity to transform the non-zero part of cigs
  - Categorize *cigs*; or
  - Use fp method for the non-zero part of cigs

Contrast 1 for variables with many zeros

- Logit: $g(smo, cigs) = \beta_0 + \beta_1 smo + \beta_2 cigs$
- Smoking *c* cigarettes vs. non-smoking

$$g(smo = 1, cigs = c) - g(smo = 0, cigs = 0)$$
$$= (\beta_0 + \beta_1(1) + \beta_2(c)) - (\beta_0 + \beta_1(0) + \beta_2(0)) = \beta_1 + \beta_2 c$$

c	SAS Contrasts	OR	Confidence limits		P-value
5	$smo\ 1\ cigs\ 5$	6.5562	3.8696	11.1082	<.0001
10	$smo\ 1\ cigs\ 10$	9.3373	4.8687	17.9072	<.0001
15	$smo\ 1\ cigs\ 15$	13.298	5.4996	32.1544	<.0001

Contrast 1 for variables with many zeros

- Smokers of 5 cigs per day are more than 6 times as likely to get lung cancer as non-smokers
- Smokers of 10 cigs per day are more than 9 times as likely to get lung cancer as non-smokers
- Smokers of 15 cigs per day are more than 13 times as likely to get lung cancer as non-smokers

Contrast 2 for variables with many zeros

- Logit: $g(smo, cigs) = \beta_0 + \beta_1 smo + \beta_2 cigs$
- Smoking *x+c* cigarettes vs. smoking x cigarettes

$$g(smo = 1, cigs = x + c) - g(smo = 1, cigs = x)$$
$$= (\beta_0 + \beta_1(1) + \beta_2(x + c)) - (\beta_0 + \beta_1(1) + \beta_2(x)) = c\beta_2$$

c	SAS Contrasts	OR	Confidence limits		P-value
5	$smo\ 0\ cigs\ 5$	1.4242	1.0304	1.9684	0.0322
10	$smo\ 0\ cigs\ 10$	2.0283	1.0617	3.8748	0.0322

Contrast 2 for variables with many zeros

- A 5 cigarette increase among smokers increases the lung cancer risk by about 40%
- A 10 cigarette increase among smokers doubles the lung cancer risk

SAS: Contrasts

```
proc univariate data=cigs; where smo=1; var cigs; run;

proc logistic descending data=cigs;
  model lung_cancer=smo cigs;
  contrast ' 5 cigs vs. ns' smo 1 cigs  5/estimate=exp;
  contrast '10 cigs vs. ns ' smo 1 cigs 10/estimate=exp;
  contrast '15 cigs vs. ns ' smo 1 cigs 15/estimate=exp;
  contrast '5 cig increase, smo' smo 0 cigs  5/estimate=exp;
  contrast '10 cig increase, smo' smo 0 cigs 10/estimate=exp;
run;
```