# Segmented Linear Regression

## (Piecewise, Changepoint, Broken-Line Regression)

BIOS 6618

CU Anschutz

1. Motivating Example

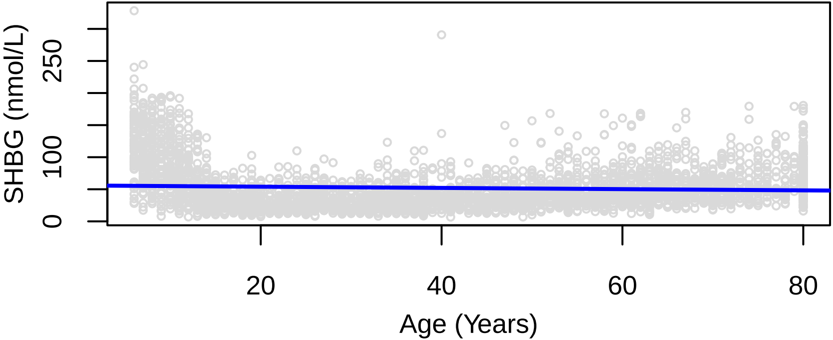
2. Introduction to Segmented Regression

3. Example in R

# Motivating Example

# Motivation

We are interested in the association between sex hormone binding globulin (SHBG) and age in years for males 6-80 years old from the NHANES 2015-2016 cycle ($n = 3390$ with SHBG)*. The scatterplot suggests there is some sort of non-linear trend:

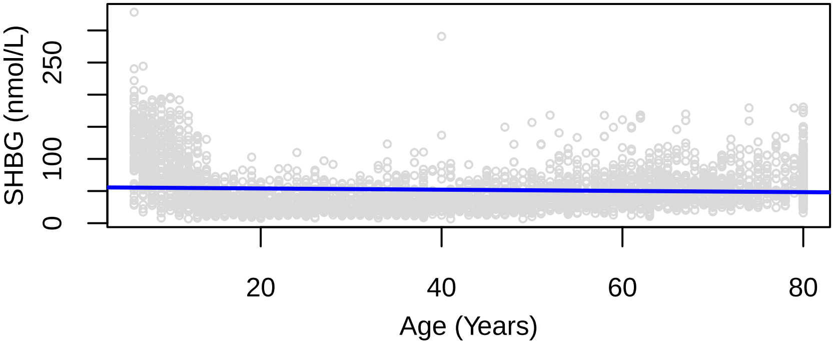


*Note: NHANES uses a complex, multistage, probability sampling design to mirror the US population, but will ignore this for our example. Also, all individuals 80+ are given an age of 80.

# Linear Regression

One modeling strategy from class would be simple linear regression:

```
mod_lm <- lm(SHBG~RIDAGEYR, data=dat)
```

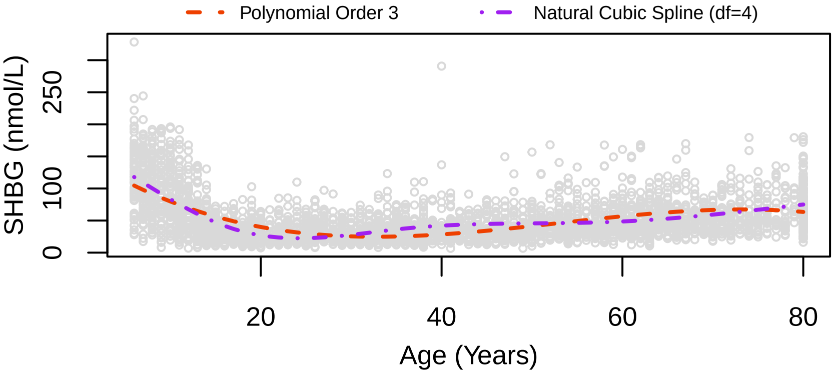


We can see that our SLR does not fit the data very well.

# Regression with Polynomials or Splines

Another modeling strategy may be including polynomial or spline terms in our linear regression:

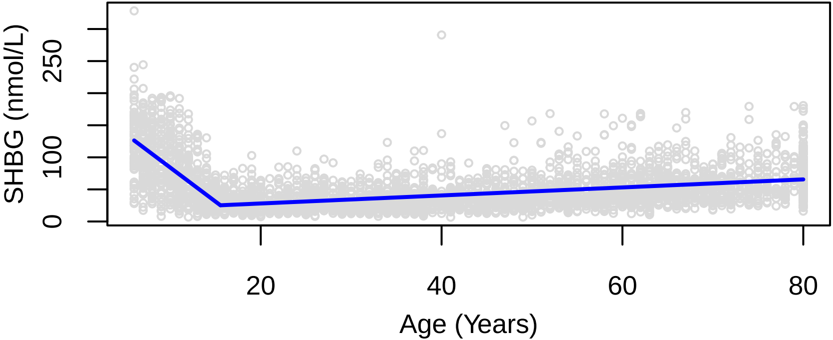
```
mod_poly3 <- lm(SHBG~poly(RIDAGEYR,3, raw=T), data=dat) # Order 3
mod_ns <- lm(SHBG ~ splines::ns(RIDAGEYR, df=4), data=dat)
```



We see better fits than the SLR, but the $\hat{\beta}$'s are challenging to interpret.

# Segmented Regression

One approach that may be more interpretable is conducting a *segmented regression* analysis (also known as piecewise, broken-line, or changepoint regression model):



# Introduction to Segmented Regression

# Segmented Regression

A segmented regression evaluates the relationship between your response ($Y$) and the explanatory variables ($X$) based on fitting piecewise linear regressions that allow changes in the model at a *breakpoint/changepoint* $\psi$.

Depending on the context, a segmented model could have multiple breakpoints (e.g., $\psi_1, \psi_2$).

The breakpoint can be provided based on context (e.g., time an intervention occurred, average age of puberty, etc.) or estimated from the data (e.g., using algorithms).

# Extension from Linear Regression

The segmented regression model is an extension of a linear regression model. Consider a linear regression model with one predictor $X$, then our segmented regression model would be:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 (X_i - \psi)_+ + \epsilon_i,\ \epsilon \sim N(0, \sigma^2_{Y|X})$$

where

- $(X_i - \psi)_+ = (X_i - \psi) \times I(X_i > \psi)$ indicates if an observation $X_i$ is above the breakpoint (with $I(\cdot)$ being the indicator function)
- $\epsilon_i$ is our traditional error term from a linear regression model
- $\beta_0$ is our intercept for the first segment
- $\beta_1$ is the slope for $X$ in our first segment
- $\beta_2$ is the difference-in-slope for segment 2 compared to segment 1

# Extension from Linear Regression

We can rewrite the mean response, $E(Y)$:

$$E(Y) = \beta_0 + \beta_1 X_i + \beta_2 (X_i - \psi)_+$$

to identify the linear regressions within each segment. For instance, segment 1's linear regression is:

$$E(Y) = \beta_0 + \beta_1 X_i$$

Then, for values of $X_i > \psi$, segment 2's linear regression is:

$$E(Y) = \beta_0^* + (\beta_1 + \beta_2) X_i$$

where $\beta_0^*$ is estimated based on the change in slope and location of $\psi$.

# Hypothesis Testing for Breakpoints

We may be interested in testing if a breakpoint is warranted, or if a more parsimonious model (e.g., simple linear regression without a breakpoint) could be used.

If the breakpoint does not exist, the difference-in-slopes parameter must be zero, indicating a potential hypothesis test is: $H_0: \ \beta_2 = 0$.

However, the validity conditions for standard statistical tests (e.g., Wald) are not satisfied, with *p*-values being underestimated.

Therefore, a different approach is needed...

# Hypothesis Testing for Breakpoints: Davies' Test

One approach for evaluating breakpoints is *Davies' test*. It assumes there are $K$ fixed, ordered values of breakpoints $\psi_1 < \psi_2 < \cdots < \psi_K$ spanning the range of $X$ and our test statistics $\{S(\psi_k)\}_k$ have standard normal distributions for fixed $\psi_k$ $(k = 1, ..., K)$.[1]

Davies provides an upper bound for a one-sided p-value of:

$$p \approx \Phi(-M) + V \exp(-M^2/2)(8\pi)^{-1/2}$$

where $M = \max(S(\psi_k))$ and $V = \sum_k(|S(\psi_k) - S(\psi_{k-1})|)$ is the total variation of $\{S(\psi_k)\}_k$.

As an upper bound, Davies overestimates and is slightly conservative.

Usually, $5 \leq K \leq 10$ or the quantiles of $X$.

[1] Muggeo, V. M. (2008). Segmented: an R package to fit regression models with broken-line relationships. *R news*, 8(1), 20-25.

# Example in R

# R Code

Let’s revisit our NHANES example and apply a segmented regression.

We will examine:

- General code to fit/evaluate
- Testing if the change in slopes before/after the breakpoint is significant
- Changing the number of breakpoints

First we will load the `segmented` package:

```r
library(segmented)
```

# Example: Segmented Regression with 1 Breakpoint

```
mod_lm <- lm(SHBG~RIDAGEYR, data=dat) # First fit a lm/glm object
os <- segmented(mod_lm) # Fit segmented regression
summary(os) # See summary
```

```
##
##  ***Regression Model with Segmented Relationship(s)***
##
## Call:
## segmented.lm(obj = mod_lm)
##
## Estimated Break-Point(s):
##                   Est. St.Err
## psi1.RIDAGEYR 15.541  0.202
##
## Meaningful coefficients of the linear terms:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 189.8303     3.8113   49.81   <2e-16 ***
## RIDAGEYR    -10.5888     0.3497  -30.28   <2e-16 ***
## U1.RIDAGEYR  11.2143     0.3508   31.97       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.48 on 3386 degrees of freedom
## Multiple R-Squared: 0.4055,  Adjusted R-squared: 0.405
```

# Example: Davies' Test

Let's evaluate if the difference-in-slopes is significantly different from 0 (i.e., do we need a breakpoint):

```r
# Notice we provide the lm() model:
davies.test(mod_lm, k=10)
```

```
##
##  Davies' test for a change in the slope
##
## data:  formula = SHBG ~ RIDAGEYR ,   method = lm
## model = gaussian , link = identity
## segmented variable = RIDAGEYR
## 'best' at = 14.222, n.points = 8, p-value < 2.2e-16
## alternative hypothesis: two.sided
```

We see that $p < 0.001$, so we reject $H_0$ that $\beta_2 = 0$ and conclude that there is a significant difference-in-slopes (i.e., a breakpoint at 15.541 years does lead to a significant change before and after this age in SHBG).

# Example: Confidence Interval Around Breakpoint Location

The segmented packages allows us to easily estimate a CI around the breakpoint location:

```r
confint(os)
```

```
##                   Est. CI(95%).low CI(95%).up
## psi1.RIDAGEYR 15.5412     15.1457    15.9367
```

The CI can be interpreted as being 95% confident that the true breakpoint falls between 15.1457 and 15.9367.

Note, given our larger sample size the 95% CI is fairly narrow. In smaller samples there may be more uncertainty around the “true” breakpoint location.

# Example: SLR in Each Segment

Next, let's estimate the fitted regression within each segment:

```
intercept(os) # intercept in each segment
```

```
## $RIDAGEYR
##                Est.
## intercept1 189.830
## intercept2  15.547
```

```
slope(os) # slopes of age (X) in each segment
```

```
## $RIDAGEYR
##              Est.  St.Err. t value CI(95%).l CI(95%).u
## slope1 -10.58900 0.349660 -30.283 -11.27400   -9.9032
## slope2   0.62551 0.027737  22.551   0.57113    0.6799
```

We see that before 15.5 years, the fitted regression is:

$$\hat{Y}_{I(X\leq 15.5)} = 189.830 + -10.589X_{age}$$

And after 15.5 years it is:

$$\hat{Y}_{I(X>15.5)} = 15.547 + 0.626X_{age}$$

# Example: SLR in Each Segment

We can also evaluate statistical significance by examining the CIs:

```
## $RIDAGEYR
##              Est.  St.Err. t value CI(95%).l CI(95%).u
## slope1 -10.58900 0.349660 -30.283 -11.27400   -9.9032
## slope2   0.62551 0.027737  22.551   0.57113    0.6799
```

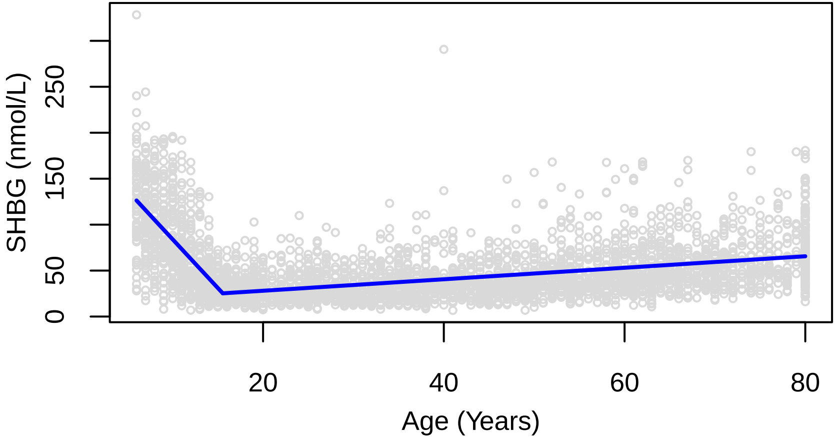
For those 6 to 15.5 years, there is a significant decrease in average SHBG of 10.59 nmol/L (95% CI: -11.27 to -9.90 nmol/L) for every one year increase in age.

For those 15.5 to 80 years, there is a significant increase in average SHBG of 0.63 nmol/L (95% CI: 0.57 to 0.68 nmol/L) for every one year increase in age.

Note, `slope()` estimates are based on a limiting Gaussian distribution and the approximation does not support estimating a *p*-value from the *t*-value. See `?slope` help documentation for more information.

# Example: Plotting Segmented Results

```r
par( mgp=c(2,1,0) )
plot(x=dat$RIDAGEYR, y=dat$SHBG, xlab='Age (Years)', ylab='SHBG (nmol/L)',
     col='gray85', cex.lab=0.8, cex.axis=0.8, cex=0.5)
plot(os, add=T, col='blue', lwd=2)
```



# Example: More Than 1 Breakpoint?

We can also fit models with more than 1 breakpoint or evaluate if a model could benefit from more breakpoints.

First, we can evaluate if our current segmented regression would benefit from an additional breakpoint by providing the os segmented regression object and using `davies.test()`:

```
davies.test(os)
```

```
##
##  Davies' test for a change in the slope
##
## data:  formula = SHBG ~ RIDAGEYR + U1.RIDAGEYR ,   method = segmented.lm
## model = gaussian , link = identity
## segmented variable = RIDAGEYR
## 'best' at = 47.111, n.points = 8, p-value = 4.6e-08
## alternative hypothesis: two.sided
```

We see from this output that a model with one more breakpoint may result in a better fit by using the `npsi` argument:

```
os2 <- segmented(mod_lm, npsi=2) # two breakpoints
```

# Example: More Than 1 Breakpoint?

Would a model with 3 breakpoints be better than one with 2?

```
davies.test(os2)
```

```
##
##  Davies' test for a change in the slope
##
## data:  formula = SHBG ~ RIDAGEYR + U1.RIDAGEYR + U2.RIDAGEYR ,   method = segment
## model = gaussian , link = identity
## segmented variable = RIDAGEYR
## 'best' at = 71.778, n.points = 8, p-value = 0.04781
## alternative hypothesis: two.sided
```

Since $p$=0.048 $<$ 0.05, we may wish to add another breakpoint:

```
os3 <- segmented(mod_lm, npsi=3) # three breakpoints
davies.test(os3)$p.value # 4th not warranted
```

```
## [1] 0.8729932
```

A 4th breakpoint does not appear to be warranted given $p = 0.873 > 0.05$.

# Example: AIC/BIC to Select Breakpoints

We can also compare models with model selection criterion (e.g., BIC):

```
# AIC
c('1 Breakpoint'=AIC(os), '2 Breakpoints'=AIC(os2), '3 Breakpoints' = AIC(os3))
```

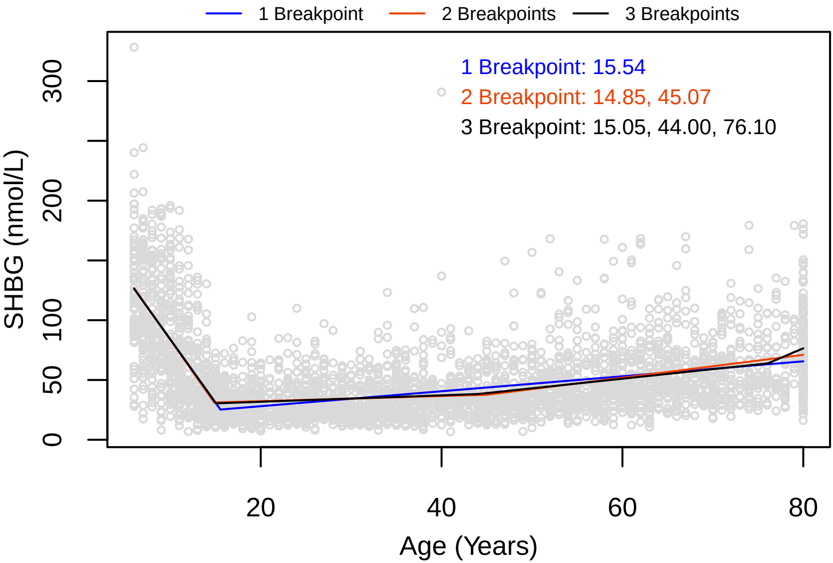
```
##  1 Breakpoint 2 Breakpoints 3 Breakpoints
##      32092.79      32055.64      32049.56
```

```
# BIC
c('1 Breakpoint'=BIC(os), '2 Breakpoints'=BIC(os2), '3 Breakpoints' = BIC(os3))
```

```
##  1 Breakpoint 2 Breakpoints 3 Breakpoints
##      32123.44      32098.54      32104.72
```

The model with 2 breakpoints minimizes our BIC, whereas the model with 3 breakpoints minimizes AIC. Depending on our context and desire for a parsimonious (i.e., simpler) model, we could use any of our models with 1, 2, or 3 breakpoints.

# Example: Plot with 1/2/3 Breakpoints



# Segmented Regression Summary

Segmented regression models help us address non-linear trends by fitting separate (linear) piecewise regressions. They may be especially useful for problems where identifying or testing a changepoint is the primary research question.

There are also alternatives to using segmented regression that we discuss in other lectures, including polynomial regression or regression models with splines for continuous predictors.