BIOS 6312: Modern Regression Analysis

Andrew J. Spieker, Ph.D.

Assistant Professor of Biostatistics
Vanderbilt University Medical Center

Set 2 supplementary slides for `R` enthusiasts

Version: 01/25/2021

Examples for Set 2

Examples for R enthusiasts:

- Regression of A1c on continuous age (Slide 135)
- Diagnostic plots for age and A1c (Slide 162)

Examples for Set 2

Examples for R enthusiasts:

- **Regression of A1c on continuous age (Slide 135)**
- Diagnostic plots for age and A1c (Slide 162)

Analysis of A1c and age

Reading in the REACH data:

- Read in data:

```
reach.data <- read.csv("reach.csv",
                       header = TRUE,
                       stringsAsFactors = FALSE)
```

Analysis of A1c and age

Scatterplot:

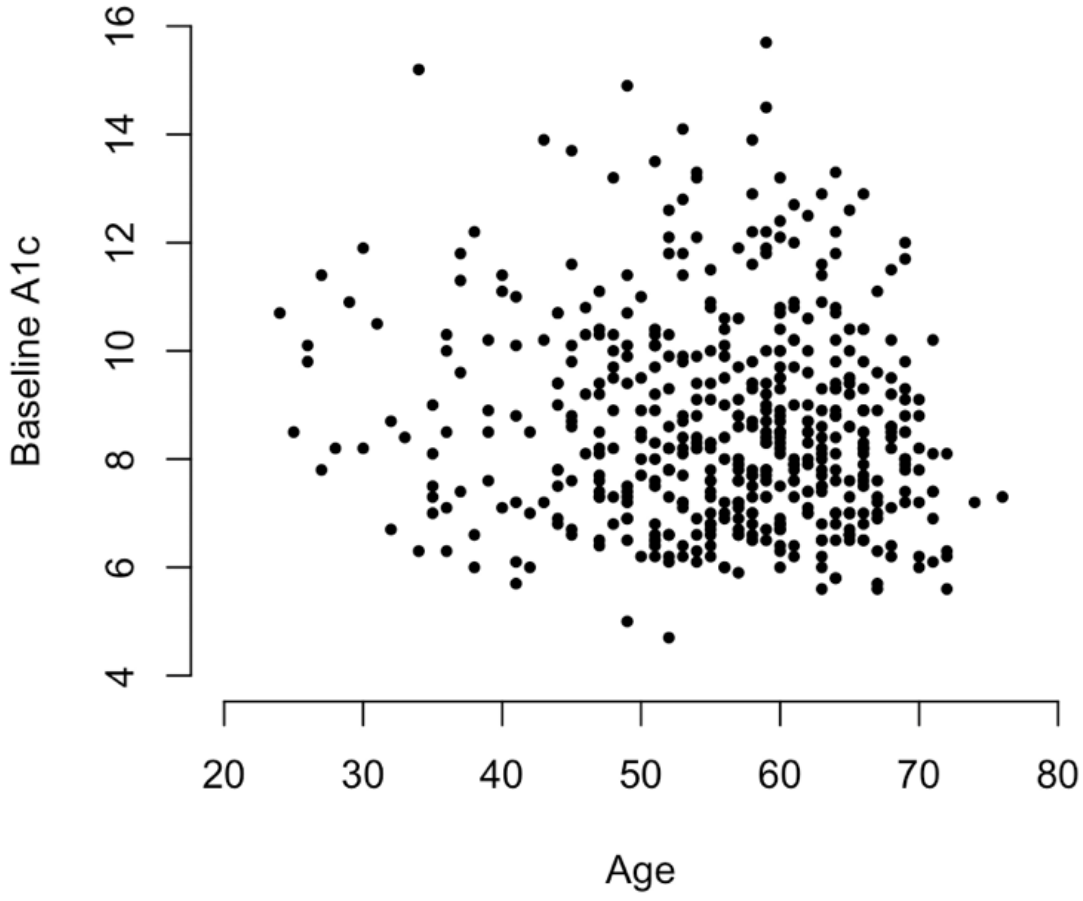
- Without a LOWESS smoother.

```
plot(reach.data$age, reach.data$a1c.0,
     xlim = c(20, 80), ylim = c(4, 16),
     xlab = "Age", ylab = "Baseline A1c",
     cex = 0.8, pch = 20,
     frame.plot = FALSE)
```

- Highly customizable.

Analysis of A1c and age

Scatterplot:



Analysis of A1c and age

Scatterplot:

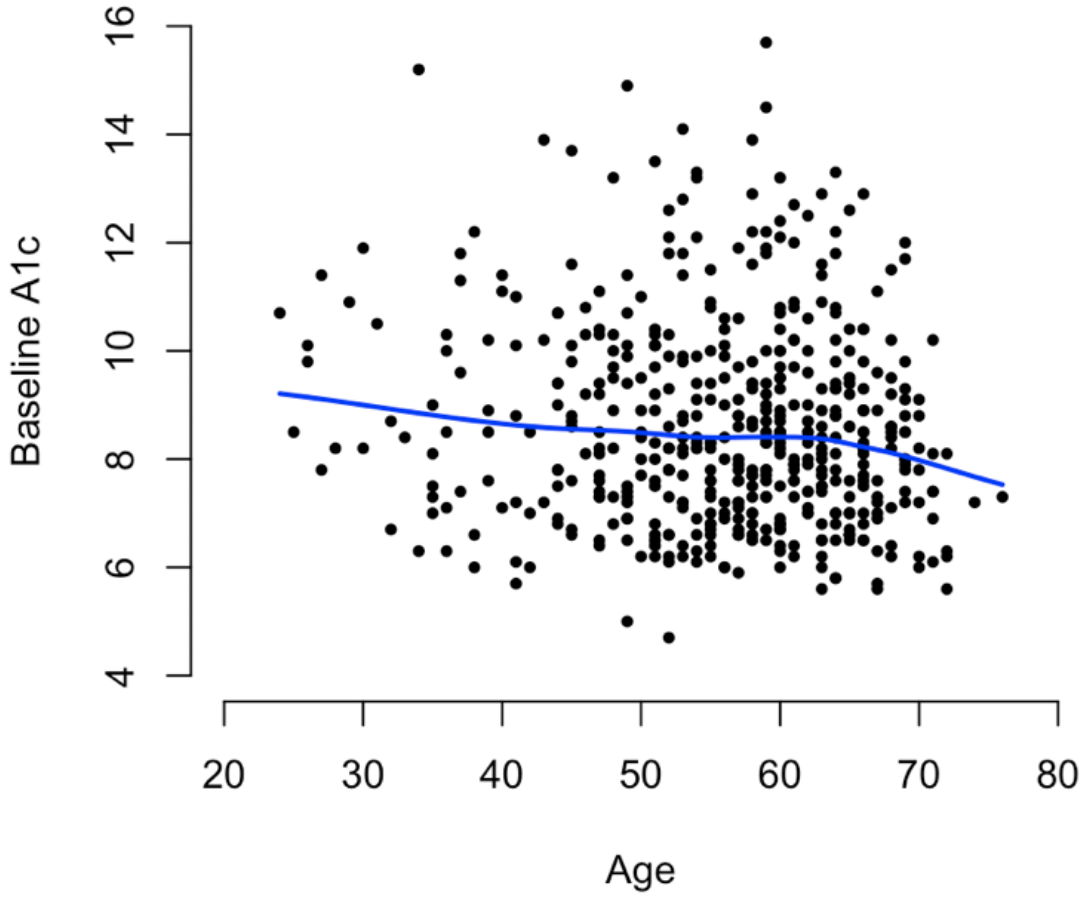
- With a LOWESS smoother.

```
scatter.smooth(reach.data$age, reach.data$a1c.0,
               xlim = c(20, 80), ylim = c(4, 16),
               xlab = "Age", ylab = "Baseline A1c",
               cex = 0.8, pch = 20,
               lpars = list(lwd = 2, col = "blue"),
               frame.plot = FALSE)
```

- Option `lpars` contains options specific to the smoothing ling.

Analysis of A1c and age

Scatterplot: With LOWESS smoother



Analysis of A1c and age

Regression fit:

- Fit regression model and print summary of results.

```
regr.a1c <- lm(a1c.0 ~ age, data = reach.data)
summary(regr.a1c)
```

Analysis of A1c and age

Regression fit:

- Print summary of results

```
Call:
lm(formula = a1c.0 ~ age, data = reach.data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0097 -1.4270 -0.2879  1.1100  7.1426

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.840478   0.490875  20.047   <2e-16 ***
age         -0.021747   0.008641  -2.517   0.0122 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.883 on 493 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.01268,Adjusted R-squared:  0.01068
F-statistic: 6.333 on 1 and 493 DF,  p-value: 0.01217
```

- Be warned: output is based on non-robust standard errors!

Analysis of A1c and age

Extracting robust standard errors:

- Need to install (and load) the `sandwich` package.

```
## Load library
library("sandwich")

## Huber-White variance
robust.var <- vcovHC(regr.a1c, type = "HC1")
> robust.var
             (Intercept)           age
(Intercept)  0.234640771 -3.986932e-03
age         -0.003986932  6.985083e-05

## Print standard errors for coefficients of interest
> sqrt(diag(robust.var))
(Intercept)         age
0.484397327 0.008357681
```

- Agrees with Stata output (Slide 139).

Examples for Set 2

Examples for R enthusiasts:

- *Regression of A1c on continuous age (Slide 135)*
- **Diagnostic plots for age and A1c (Slide 162)**

Diagnostic plots: age and A1c

Fitted/predicted values:

- This is a continuation of the previous example.
- Extract fitted values from regression fit:

```
fitted <- regr.a1c$fitted.values
```

Diagnostic plots: age and A1c

Studentized residuals:

- Studentized residuals not readily available.

```
## Residuals from regression model
resid <- regr.a1c$residuals

## Estimate error variance
sigma.hat <- sd(regr.a1c$residuals)

## Create hat matrix
dsn.X <- cbind(1, regr.a1c$model$age)
H <- dsn.X %*% solve(t(dsn.X) %*% dsn.X) %*% t(dsn.X)

## Diagonal entries (leverage)
lvg <- diag(H)

## Create studentized residuals
st.resid <- resid/(sigma.hat * sqrt(1 - lvg))
```

Diagnostic plots: age and A1c

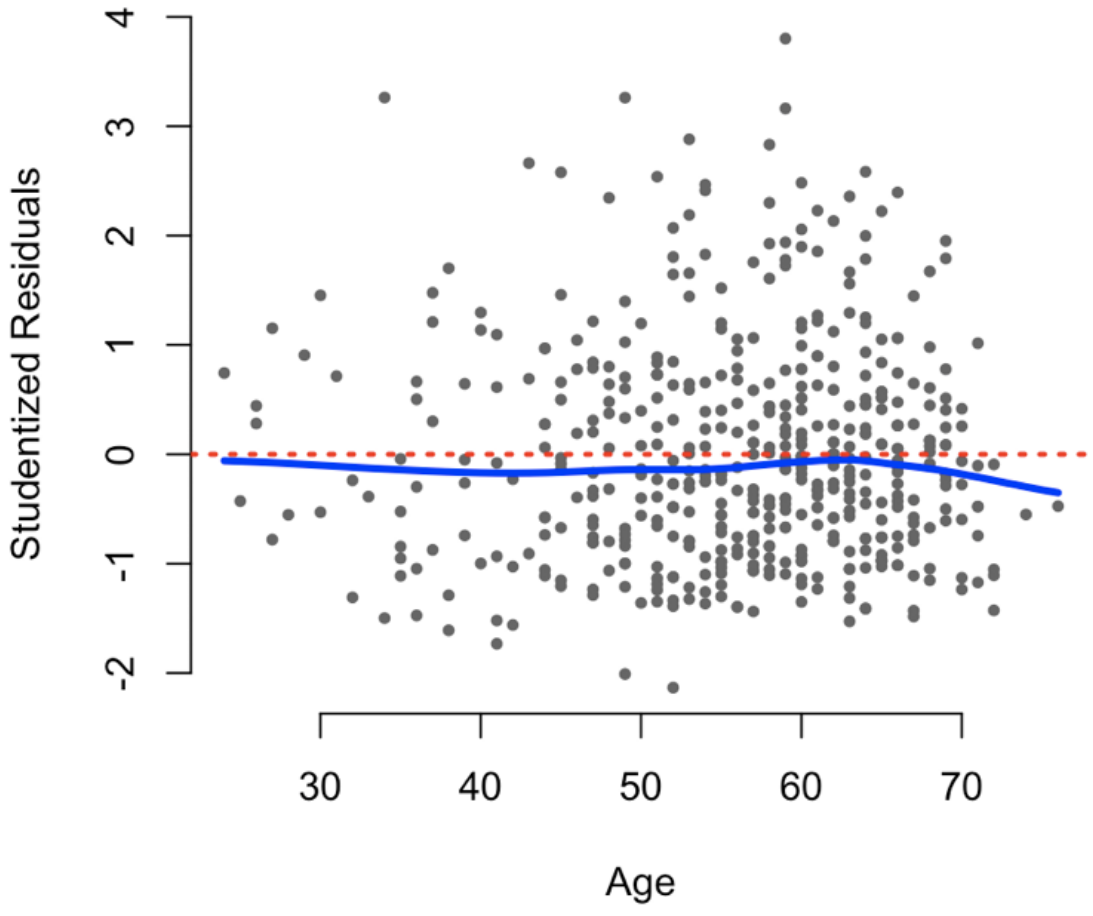
Residual-versus-predictor plot:

- Include LOWESS and an indicator of the x-axis.

```
scatter.smooth(regr.a1c$model$age, st.resid,
     xlab = "Age",
     ylab = "Studentized Residuals",
     cex = 0.8, pch = 20, col = "gray40",
     lpars = list(lwd = 3, col = "blue"),
     frame.plot = FALSE)
abline(0,0, lty = 3, lwd = 2, col = "red")
```

Diagnostic plots: age and A1c

Residual-versus-predictor plot:



Diagnostic plots: age and A1c

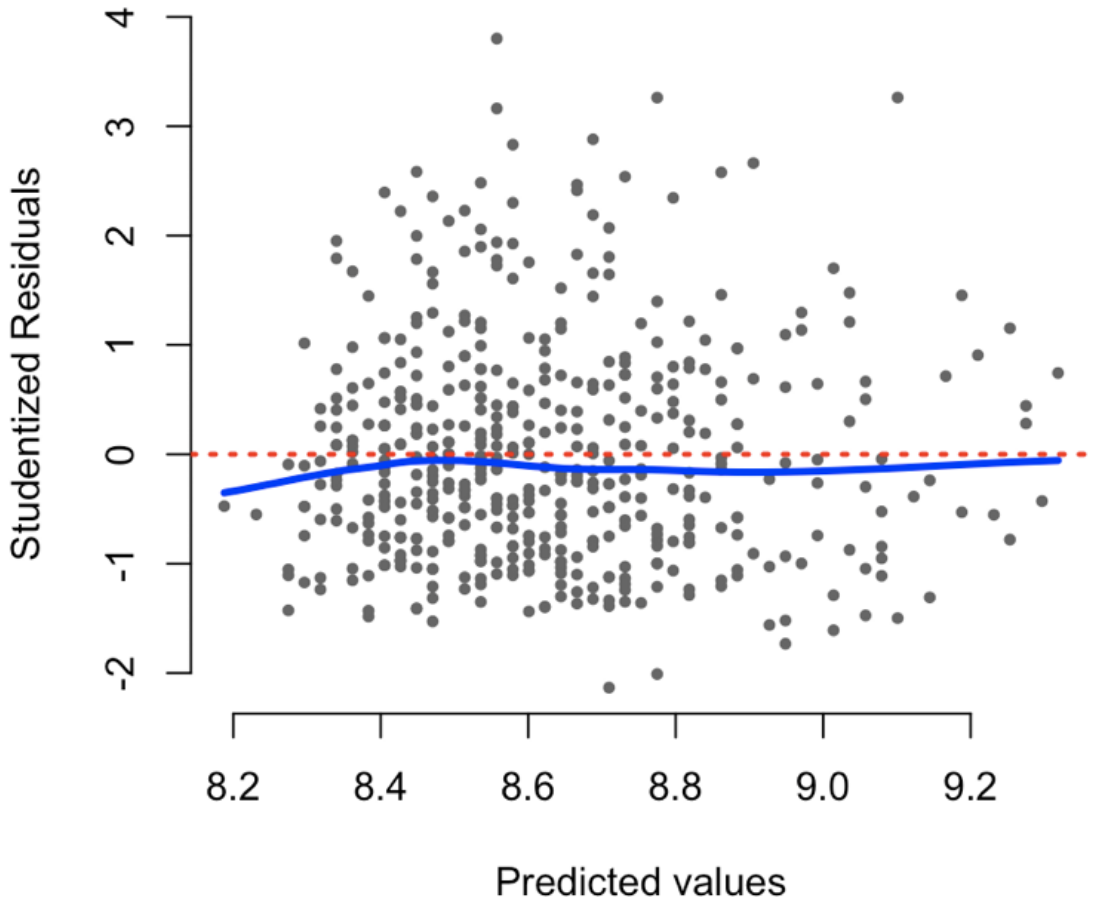
Residual-versus-fitted plot:

- Include LOWESS and an indicator of the x-axis.

```
scatter.smooth(fitted, st.resid,
               xlab = "Predicted values",
               ylab = "Studentized Residuals",
               cex = 0.8, pch = 20, col = "gray40",
               lpars = list(lwd = 3, col = "blue"),
               frame.plot = FALSE)
abline(0,0, lty = 3, lwd = 2, col = "red")
```

Diagnostic plots: age and A1c

Residual-versus-fitted plot:



Diagnostic plots: age and A1c

Quantile-quantile plot:

- Include reference line.

```
qqnorm(st.resid, frame = FALSE,
       cex = 0.8, pch = 20, col = "gray40",
       xlim = c(-4,4), ylim = c(-4, 4))
qqline(st.resid, lwd = 1.5)
```

Diagnostic plots: age and A1c

Quantile-quantile plot: