Stat 849: ggplot2 graphics

Douglas Bates

University of Wisconsin - Madison
and R Development Core Team
<Douglas.Bates@R-project.org>

Sept 08, 2010

Outline

Outline

Outline

ggplot2

The `pima` data set from the `faraway` package

Univariate summary plots

Bivariate plots

Simple regression or ancova lines

Ancova

Outline

ggplot2

The `pima` data set from the `faraway` package

Univariate summary plots

Bivariate plots

Simple regression or ancova lines

Ancova

Outline

ggplot2

The `pima` data set from the `faraway` package

Univariate summary plots

Bivariate plots

Simple regression or ancova lines

Ancova

Outline

ggplot2

The `pima` data set from the `faraway` package

Univariate summary plots

Bivariate plots

Simple regression or ancova lines

Ancova

Outline

The ggplot2 graphics package

- Another advanced graphics package for *R* is `ggplot2` by Hadley Wickham (a recent Iowa State Stats Ph.D., now at Rice).
- His book is listed as one of the references on the course web site.
- The core chapter introducing the basic function called `qplot` can be obtained from the URL in the links section on the course web site.
- I will use data from the `faraway` package to accompany Julian Faraway's freely available book "Practical Regression and Anova using R" to illustrate the use of `qplot`.

Outline

Examining the pima data

```
> library(faraway)
> str(pima)

'data.frame': 768 obs. of  9 variables:
 $ pregnant : int  6 1 8 1 0 5 3 10 2 8 ...
 $ glucose  : int  148 85 183 89 137 116 78 115 197 125 ...
 $ diastolic: int  72 66 64 66 40 74 50 0 70 96 ...
 $ triceps  : int  35 29 0 23 35 0 32 0 45 0 ...
 $ insulin  : int  0 0 0 94 168 0 88 0 543 0 ...
 $ bmi      : num  33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...
 $ diabetes : num  0.627 0.351 0.672 0.167 2.288 ...
 $ age      : int  50 31 32 21 33 30 26 29 53 54 ...
 $ test     : int  1 0 1 0 1 0 1 0 1 1 ...

> head(pima)

  pregnant glucose diastolic triceps insulin  bmi diabetes age test
1        6     148        72      35       0 33.6    0.627  50    1
2        1      85        66      29       0 26.6    0.351  31    0
3        8     183        64       0       0 23.3    0.672  32    1
4        1      89        66      23      94 28.1    0.167  21    0
5        0     137        40      35     168 43.1    2.288  33    1
6        5     116        74       0       0 25.6    0.201  30    0
```

Recoding the missing data

- As Faraway indicates, several of the values of variables that cannot reasonably be zero are recorded as zero.
- A bit of research shows that these are missing data values. Also the `test` variable is a factor, not numeric.

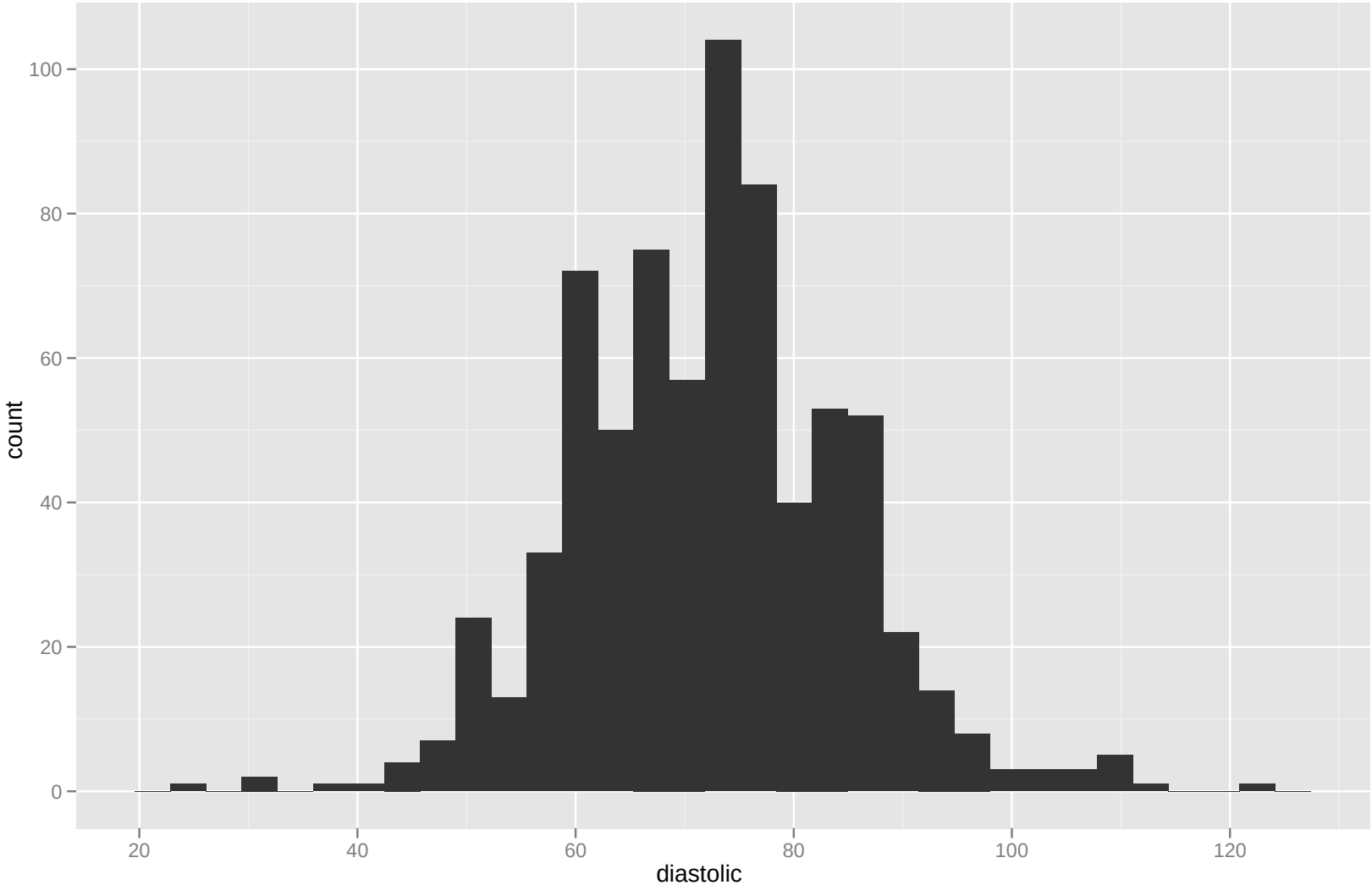
```
> pima <- within(pima, {
+     diastolic[diastolic == 0] <- glucose[glucose ==
+         0] <- triceps[triceps == 0] <- insulin[insulin ==
+         0] <- bmi[bmi == 0] <- NA
+     test <- factor(test, labels = c("negative", "positive"))
+ })
> head(pima, 3)
```

```
  pregnant glucose diastolic triceps insulin  bmi diabetes age
1        6     148        72      35      NA 33.6    0.627  50
2        1      85        66      29      NA 26.6    0.351  31
3        8     183        64      NA      NA 23.3    0.672  32
      test
1 positive
2 negative
3 positive
```

Outline

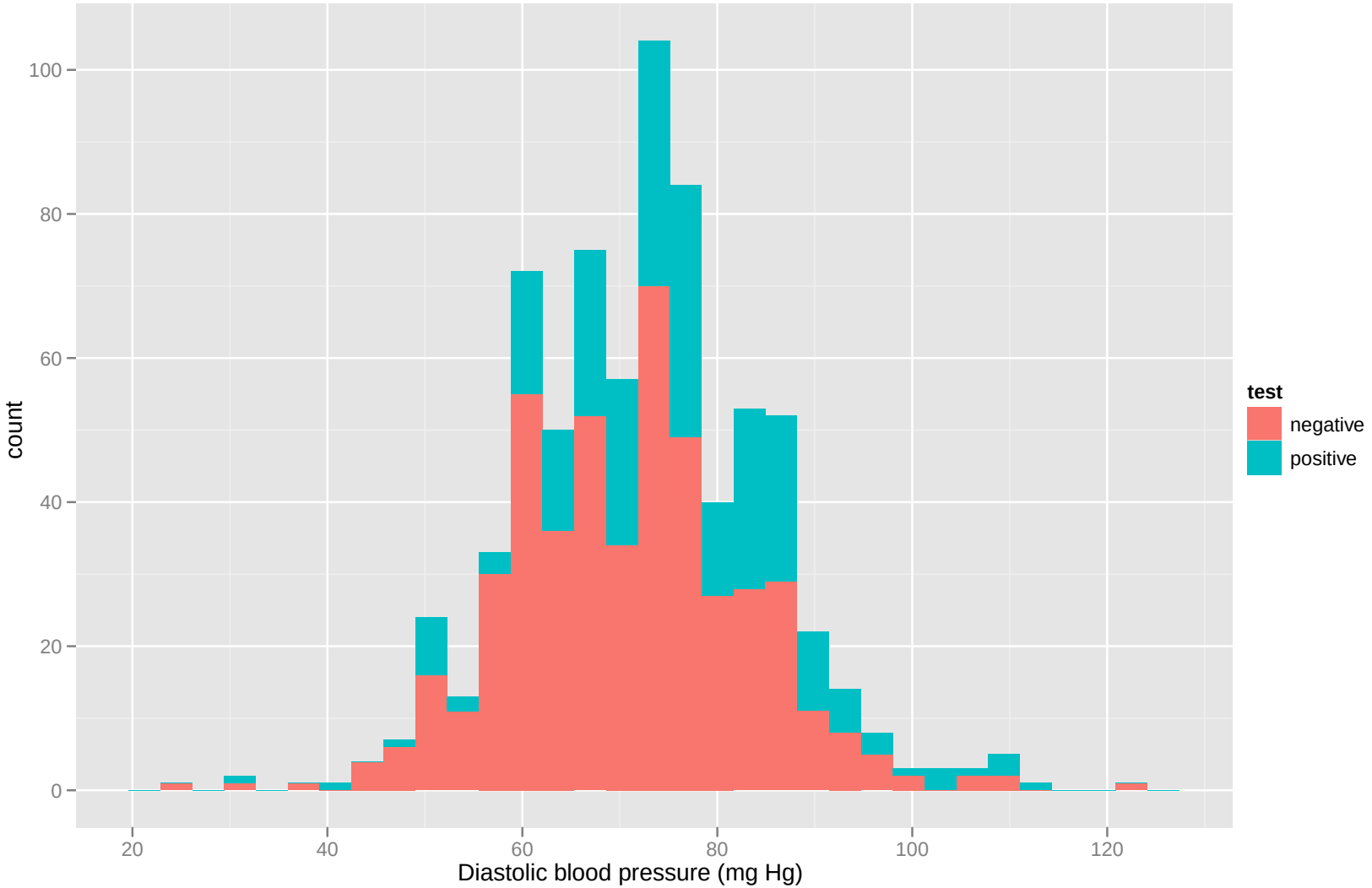
Histogram of diastolic blood pressure

```
> qplot(diastolic, data = pima, geom = "histogram")
```



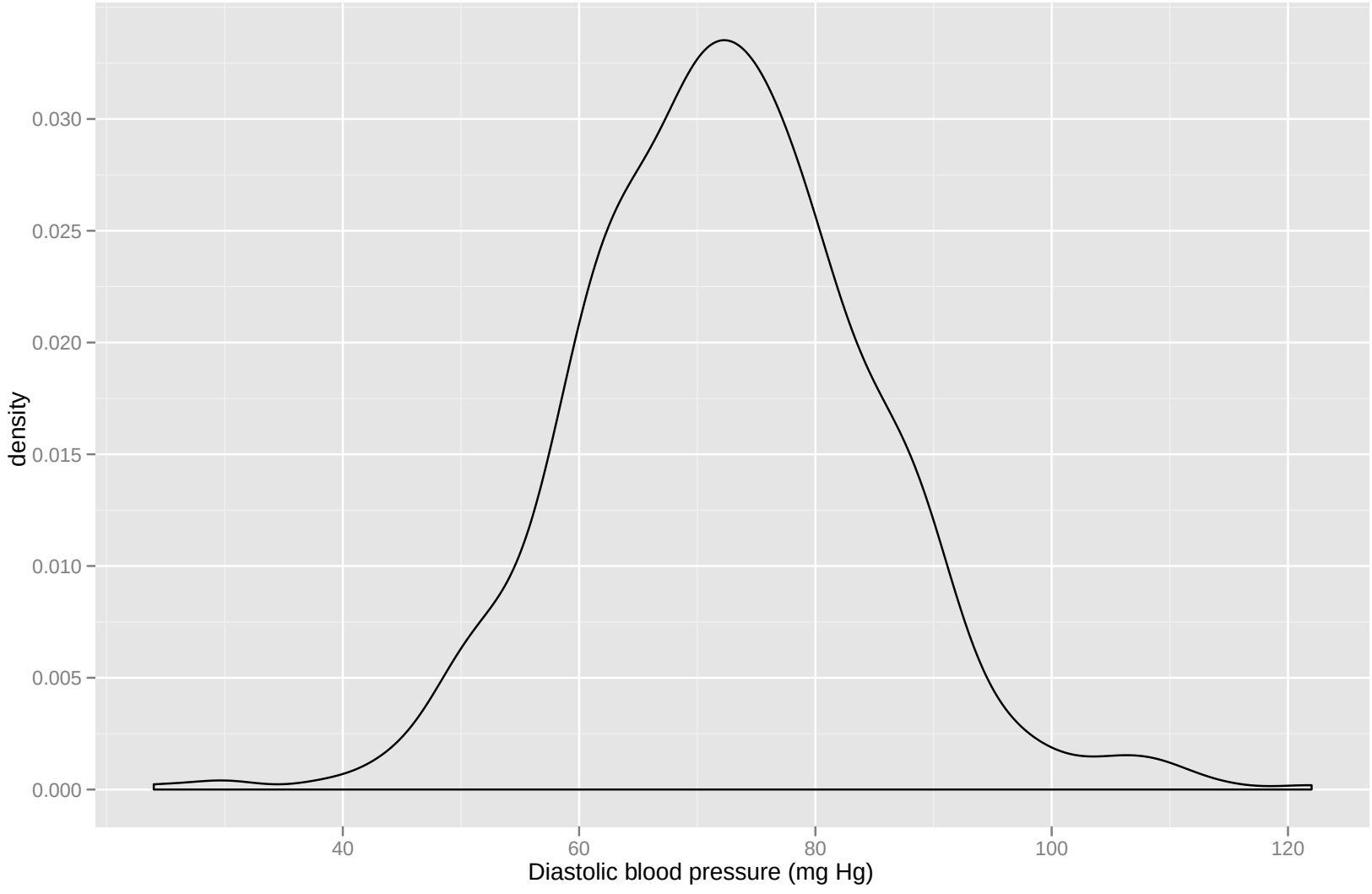
Histogram of diastolic bp by test

```
> qplot(diastolic, data = pima, geom = "histogram", fill = test)
```



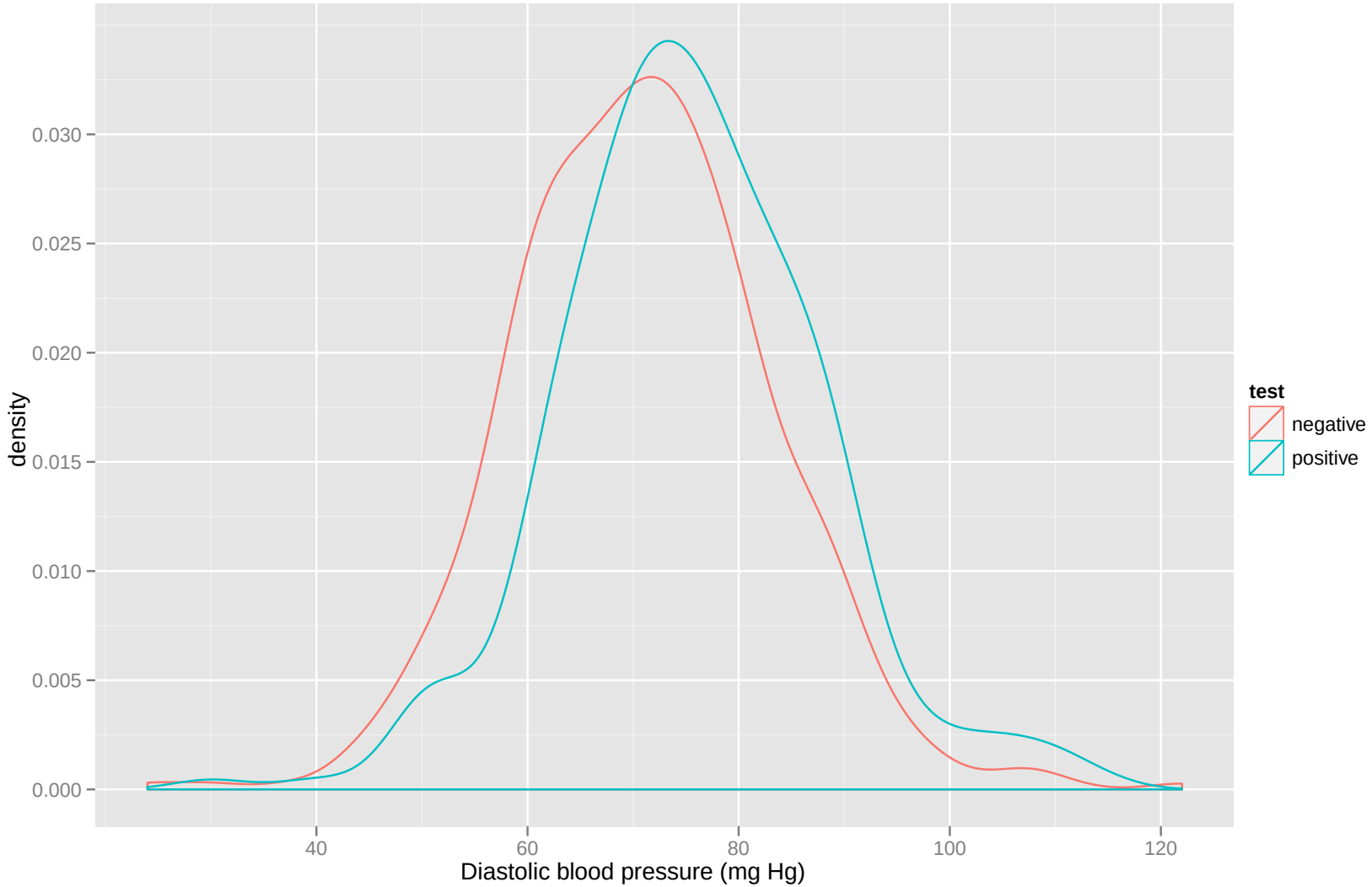
Empirical density plot

```
> qplot(diastolic, data = pima, geom = "density")
```



Empirical density of diastolic by test

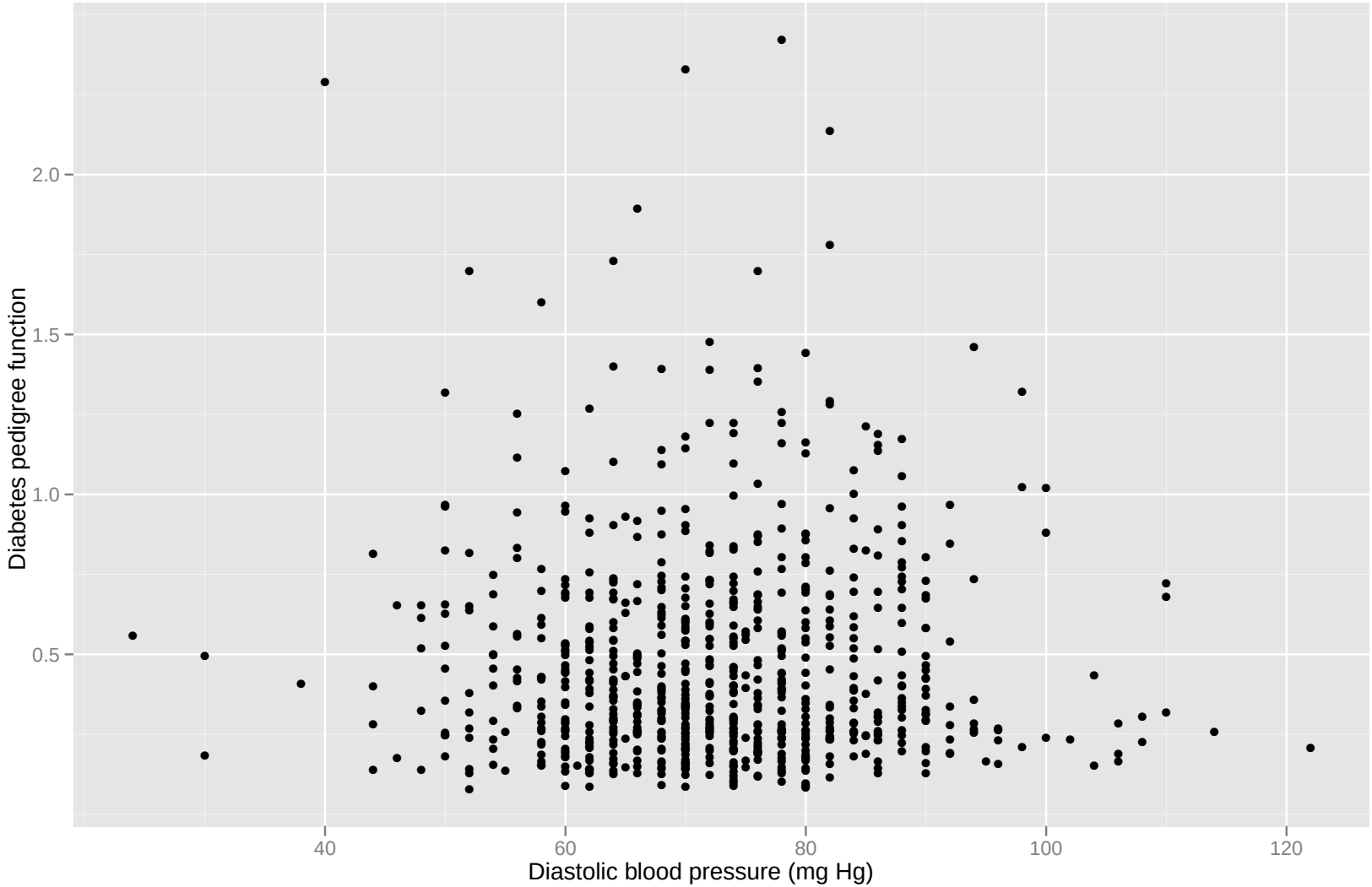
```
> qplot(diastolic, data = pima, geom = "density", color = test)
```



Outline

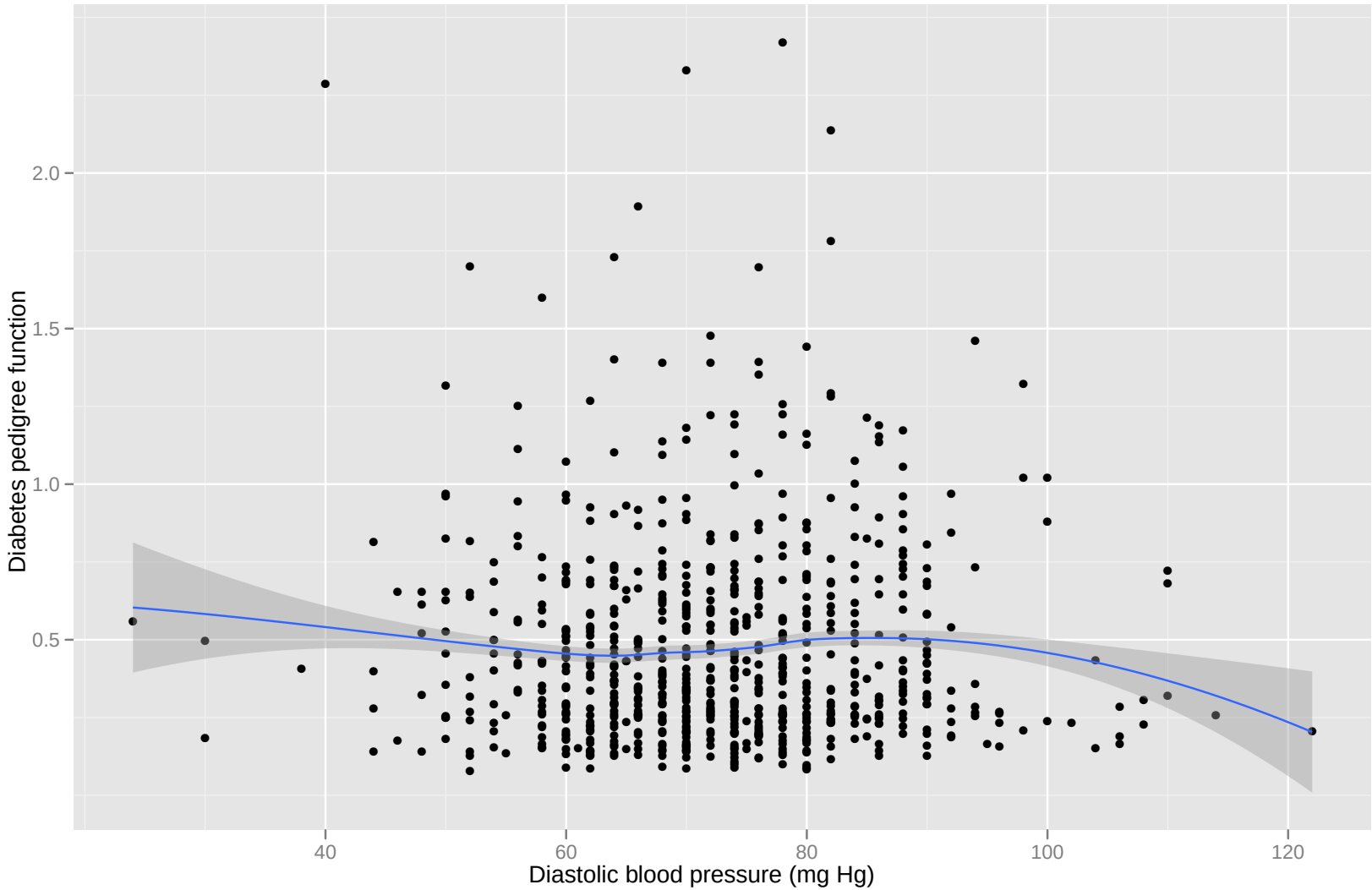
Simple scatterplot, c.f. Fig. 1.2a, p. 13

```
> qplot(diastolic, diabetes, data = pima, xlab = ...)
```



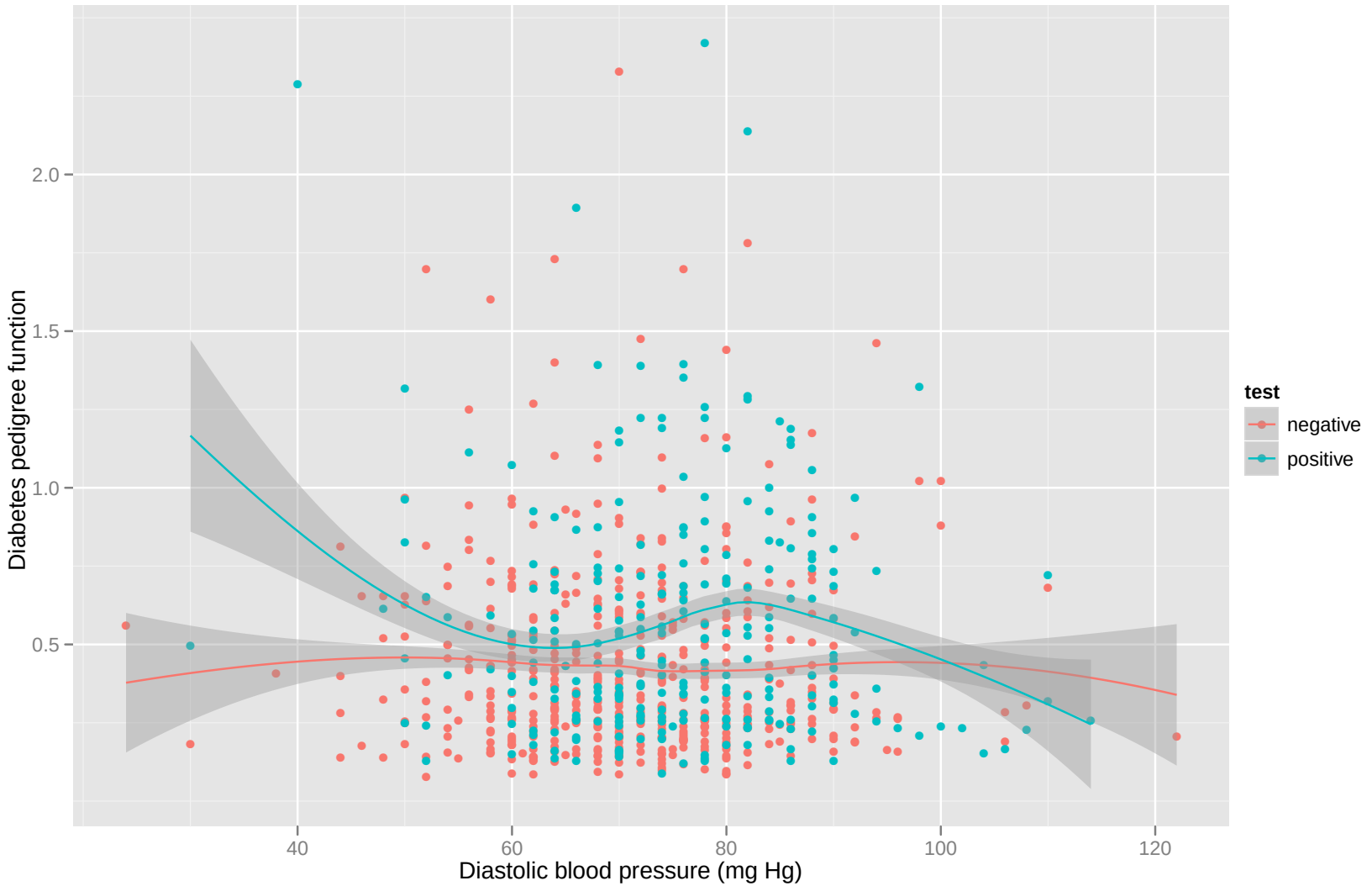
Adding a scatterplot smoother

```
> qplot(diastolic, diabetes, data = pima, geom = c("point",
+     "smooth"))
```



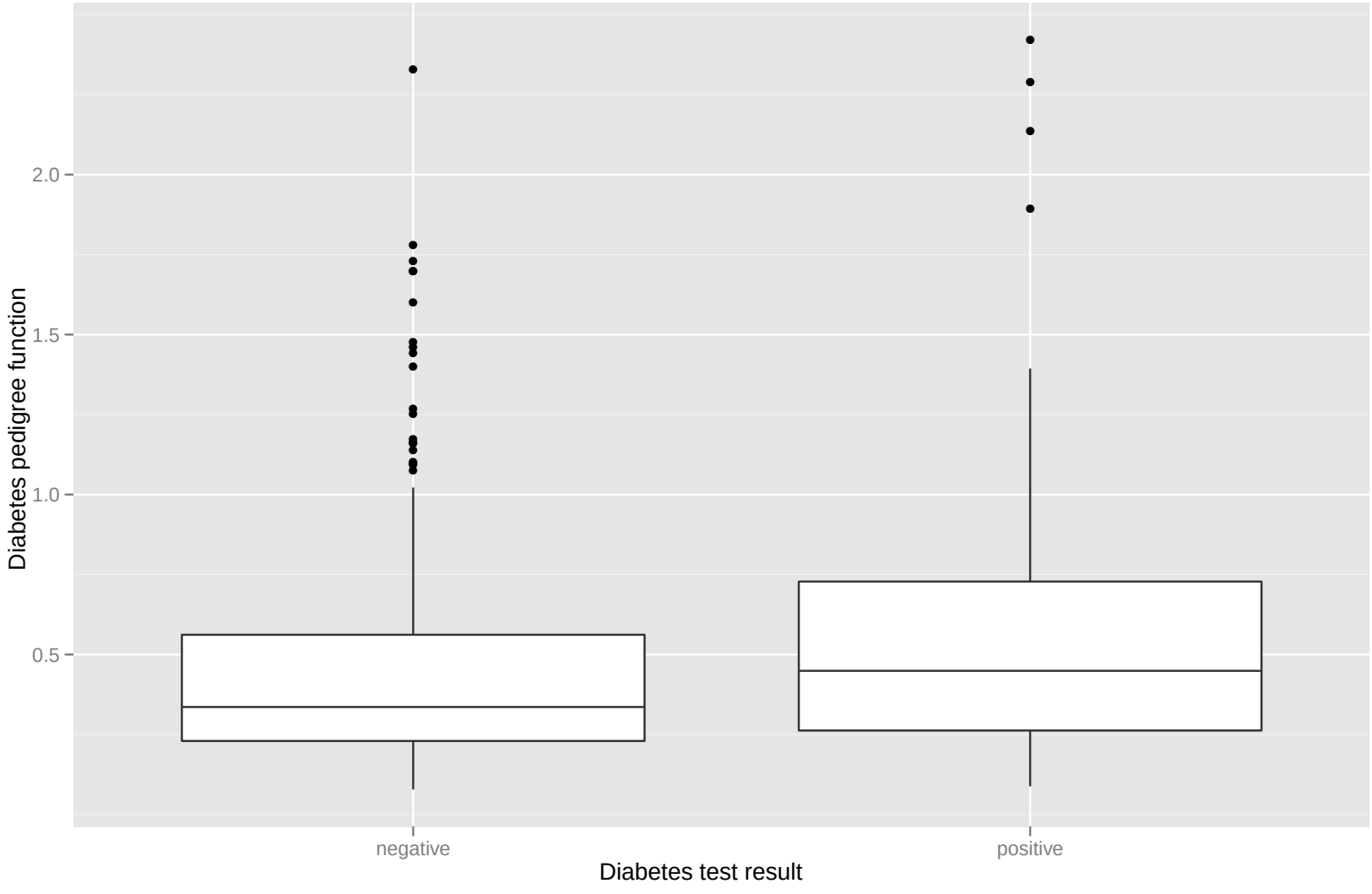
Multiple smoothers by group

```
> qplot(diastolic, diabetes, data = pima, geom = c("point",
+     "smooth"), color = test)
```



Comparative boxplots - apparently only vertical

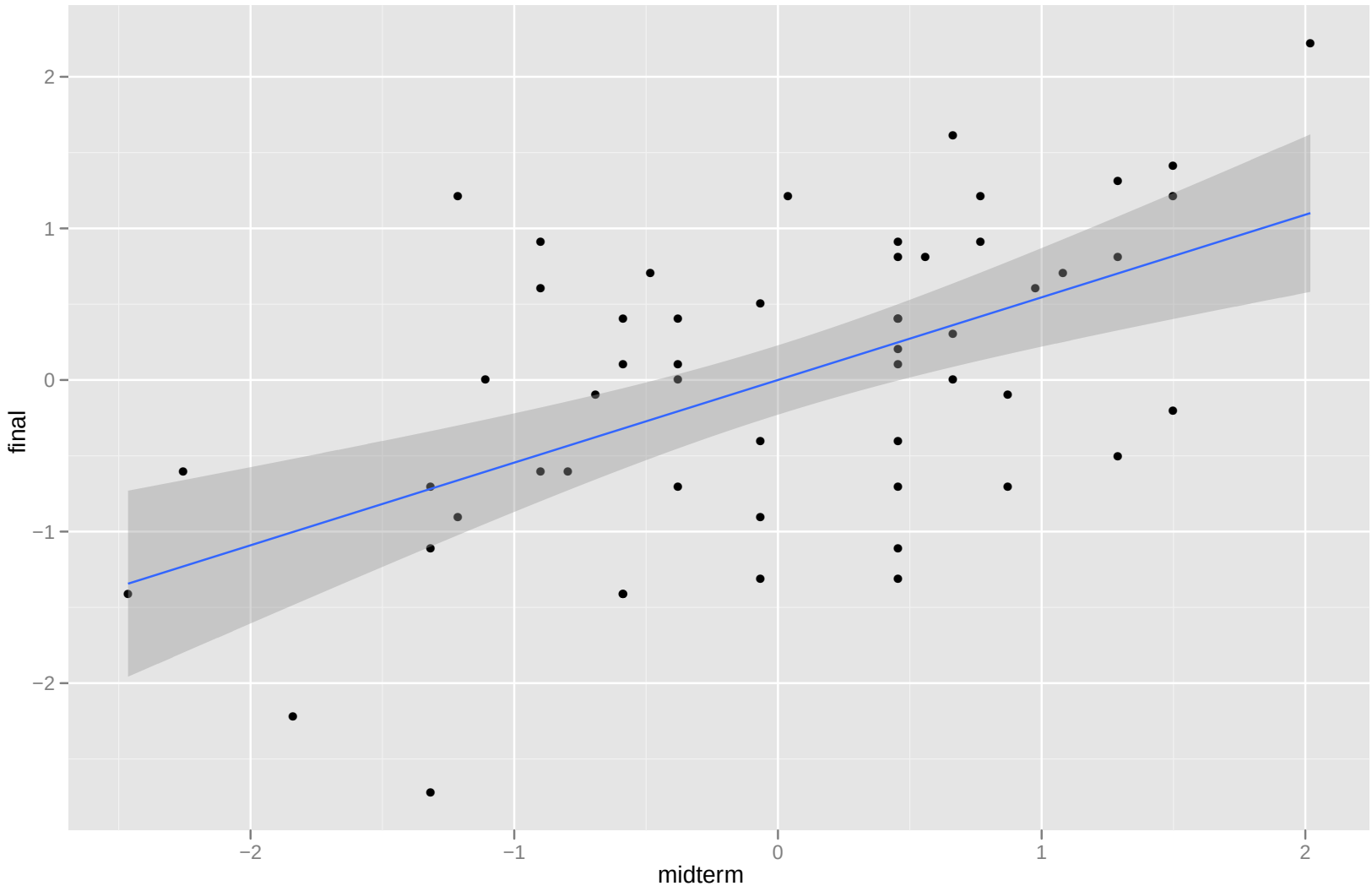
```
> qplot(test, diabetes, data = pima, geom = c("boxplot"))
```



Outline

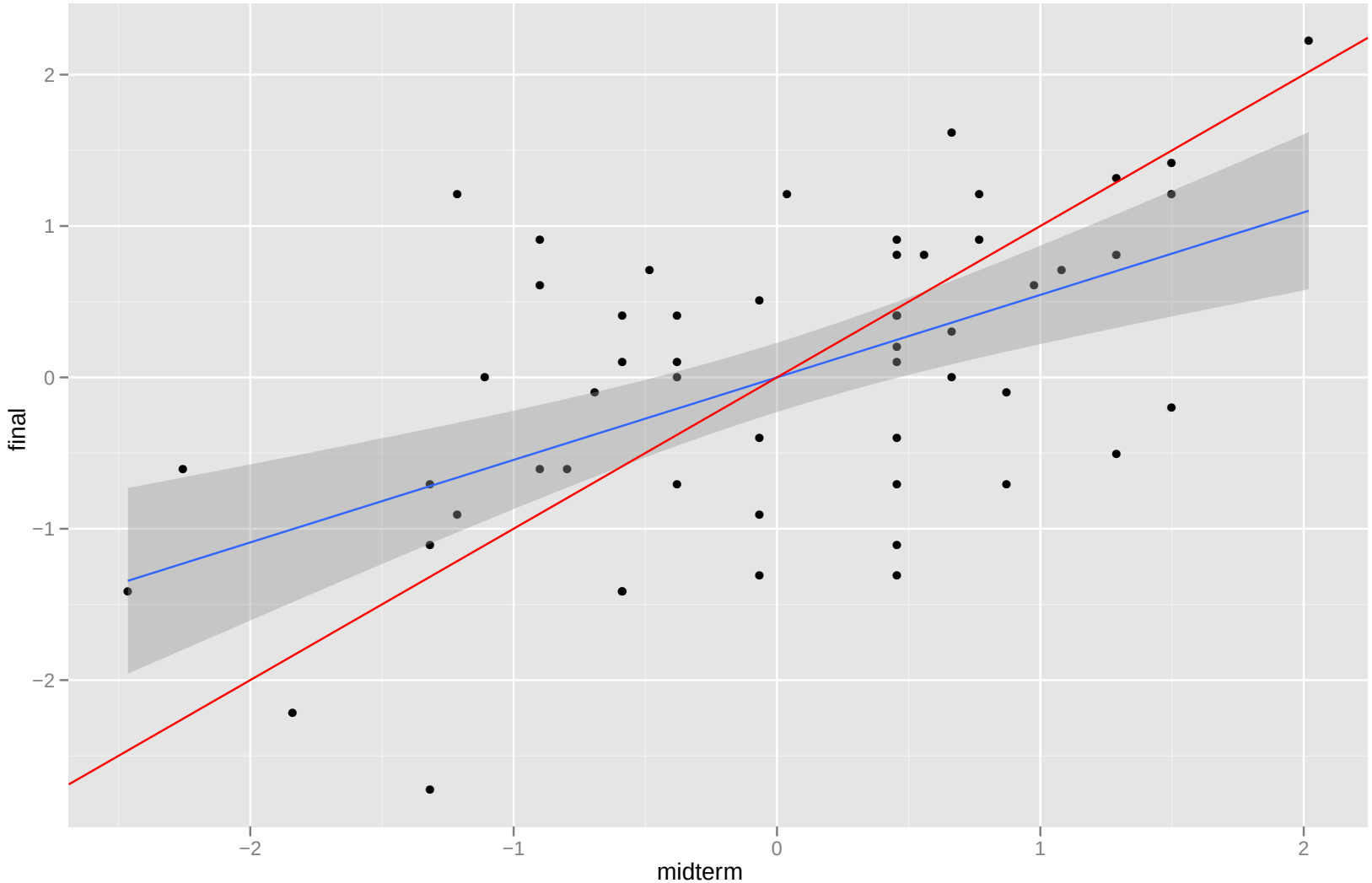
Adding a simple linear regression line - c.f. Fig. 1.3, p. 14

```
> (p <- qplot(midterm, final, data = stat500, geom = c("point",
+     "smooth"), method = "lm"))
```



Adding a reference line - c.f. Fig. 1.3, p. 14

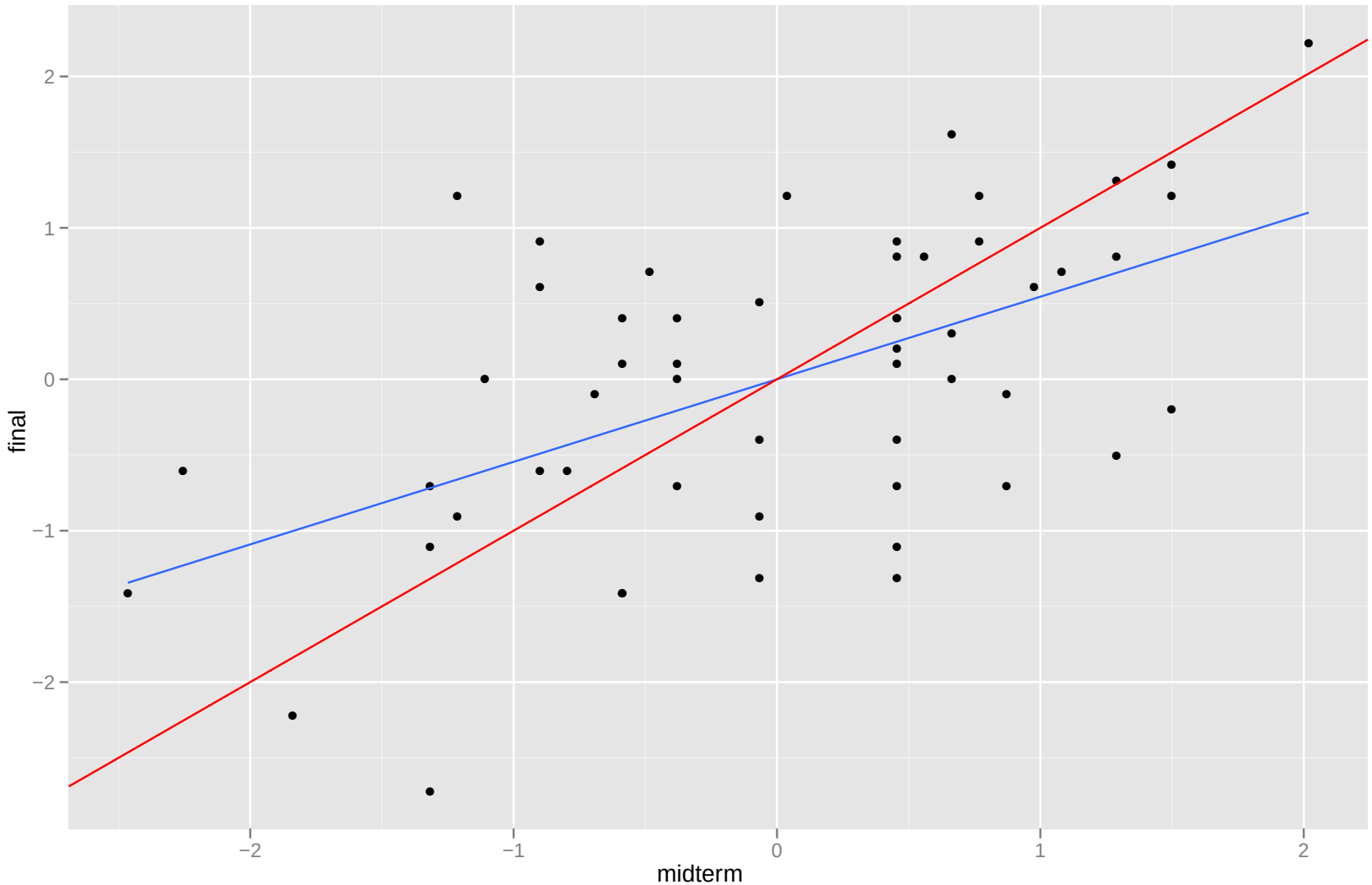
```
> p + geom_abline(intercept = 0, slope = 1, color = "red")
```



Suppressing the confidence band

It happens that the defaults are `intercept=0` and `slope=1`

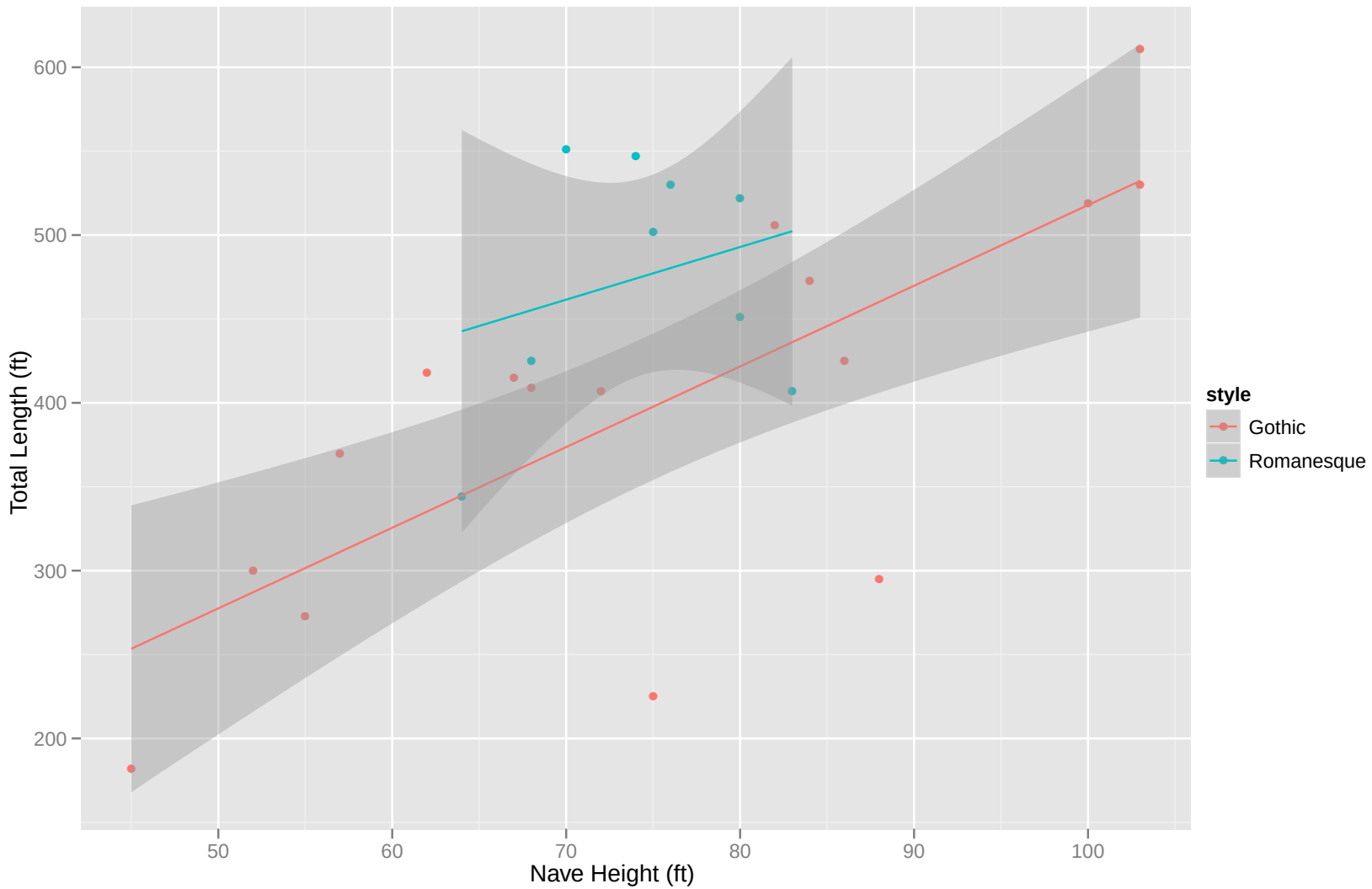
```
> (p <- qplot(midterm, final, data = stat500, geom = c("point",
+     "smooth"), method = "lm", se = FALSE) + geom_abline(color
```



Outline

Plotting multiple groups and lines, c.f. Fig. 15.2, p. 163

```
> levels(cathedral$style) <- c("Gothic", "Romanesque")
> qplot(x, y, data = cathedral, geom = c("point", "smooth"),
+     method = "lm", color = style, xlab = ...)
```



Plotting multiple groups in separate panels

```
> qplot(x, y, data = cathedral, geom = c("point", "smooth"),
+       method = "lm", facets = . ~ style, xlab = ...)
```