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.

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 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)

<table>
<thead>
<tr>
<th></th>
<th>pregnant</th>
<th>glucose</th>
<th>diastolic</th>
<th>triceps</th>
<th>insulin</th>
<th>bmi</th>
<th>diabetes</th>
<th>age</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>6</td>
<td>148</td>
<td>72</td>
<td>35</td>
<td>0</td>
<td>33.6</td>
<td>50</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>85</td>
<td>66</td>
<td>29</td>
<td>0</td>
<td>26.6</td>
<td>0.351</td>
<td>31</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>8</td>
<td>137</td>
<td>0</td>
<td>0</td>
<td>23.3</td>
<td>0.672</td>
<td>32</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>89</td>
<td>66</td>
<td>23</td>
<td>94</td>
<td>28.1</td>
<td>0.167</td>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>35</td>
<td>168</td>
<td>43.1</td>
<td>2.288</td>
<td>33</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>5</td>
<td>116</td>
<td>74</td>
<td>0</td>
<td>25.6</td>
<td>0.201</td>
<td>30</td>
<td>0</td>
</tr>
</tbody>
</table>

Outline

ggplot2

The pima data set from the faraway package

Univariate summary plots

Bivariate plots

Simple regression or ancova lines

Ancova
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.

```r
> 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)

 preg  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
```

Histogram of diastolic blood pressure

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

Histogram of diastolic bp by test

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

Empirical density plot

```r
> qplot(diastolic, data = pima, geom = "density")
```
Empirical density of diastolic by test

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

![Empirical density of diastolic by test](image)

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

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

![Simple scatterplot](image)

Adding a scatterplot smoother

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

![Adding a scatterplot smoother](image)

Multiple smoothers by group

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

![Multiple smoothers by group](image)
Comparative boxplots - apparently only vertical

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

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

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

Plotting multiple groups in separate panels

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