Notes on Statistical Methods

10th Windsor Conference (Rethinking Thermal Comfort) 15 April 2018

Jane Galbraith & Rex Galbraith

Department of Statistical Science, University College London

Introduction . . . 2
Summary . . . 3

Normal distributions **4**
Univariate normal: $X \sim \mathrm{N}(\mu, \sigma^2)$ . . . 5
Measurements of span for 1200 men . . . 6
Bivariate normal: $(X, Y) \sim \mathrm{BVN}(\mu, \Sigma)$ . . . 7
Bivariate data: length and head circumference for 382 baby boys . . . 8
Here are some ellipses of constant probability density . . . 9
Here they are again with the data superimposed . . . 10
All marginal and conditional distributions are Normal. . . . 11

The method of least squares **12**
Univariate data . . . 13
Bivariate data . . . 14
Here are the three lines for the boys data . . . 15
What do these lines tell us and what are they for? . . . 16
Here are the three lines with the bivariate normal probability ellipses . . . 17
A problem of scales . . . 18
What happens to the three lines? . . . 19
We see that: . . . 20
Normal distributions and least squares . . . 21

Linear regression **22**
Example: birth weight and gestational age for 382 baby boys . . . 23
Computer output from R software . . . 24
What is the fitted model? . . . 25
Data with fitted mean line . . . 26
Hypothesis tests . . . 27
Diagnostic plots from R output . . . 28
What use is R^2 ? . . . 29
What use is R^2 ? (example) . . . 30
What use is R^2 ? (example continued) . . . 31
What use is R^2 indeed! . . . 32
Within- and between-group slopes . . . 33
Regression dilution — what is it? . . . 34
Regression dilution — does it matter? . . . 35
Regression dilution — what can we do about it? . . . 36

Ordered Categorical Data **37**
Ordinal outcome variables . . . 38
Example: ASHRAE temperature votes . . . 39
Another version with box plots and summary statistics . . . 40
Another plot with a regression line added . . . 41
Another plot with some mean ASHRAE scores added . . . 42
. . . now with a lowess line . . . 43
. . . and now all together . . . 44
The same data re-plotted by House and Room type . . . 45
Proportions and cumulative proportions . . . 46
Plotting ordinal data . . . 47
Binary outcome variables — logistic and probit regression . . . 48
Fitted logistic regression curve . . . 49
Probit regression models . . . 50
Ordinal regression models . . . 51
Fitted proportional odds regression curves . . . 52
Fitted proportions voting “comfortable” at each temperature . . . 53
Another version, but using a *non*-proportional odds model . . . 54

Looking at several variables **55**
Example: air quality data . . . 56
Times series plots with lowess smoothing . . . 57
Pairwise scatter plots . . . 58
Coplot of Ozone against Temp given Wind and Solar.R grouped . . . 59

Introduction

- Who are we?
- General themes:
  - Context and purpose — what are we trying to do?
  - Looking at data
  - Using statistical models
  - Questions (and some answers)

2 / 59

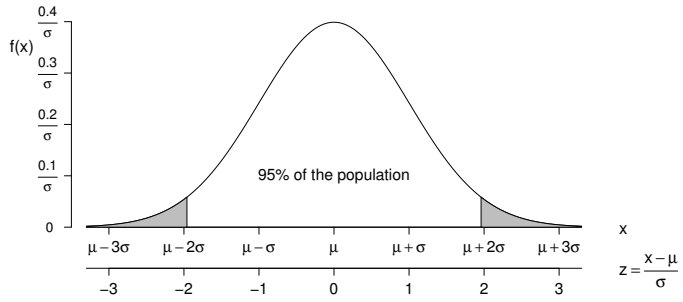
Summary

These slides deal with some basic statistical concepts that are relevant to current thermal comfort research. Topics include:

- Normal distributions (univariate and bivariate)
- The method of least squares
- Linear regression
  - Fitted model and meaning
  - Computer output, diagnostics, note on R^2
  - Regression dilution
- Ordered categorical data
  - Summarising and plotting
  - Regression models
- Looking at several variables

3 / 59

Univariate normal: $X \sim \mathrm{N}(\mu, \sigma^2)$



The **probability density function** of X is

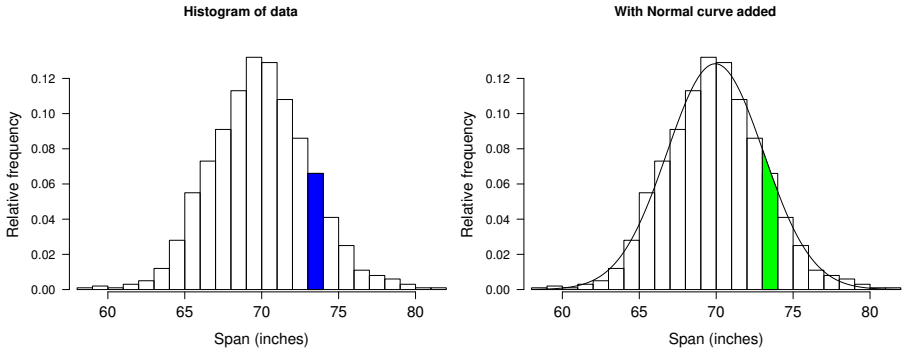
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, \quad -\infty < x < \infty$$

There are two parameters: μ and σ,
which are the **mean** and **standard deviation** of X.

From these, the probability of any event (concerning X) and be calculated

5 / 59

Measurements of span for 1200 men



The shaded area denotes the proportion of men with span between 73 and 74 inches

If we have data that we know are from a normal distribution, the summary statistics: sample size, mean and standard deviation tell us all we need to know.

Here they are: $n = 1200$ $\bar{x} = 69.94$ $s_x = 3.14$

6 / 59

Bivariate normal: $(X, Y) \sim \text{BVN}(\mu, \Sigma)$

Joint probability density function of X and Y:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)} Q\right\}$$

where

$$Q = \left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2$$

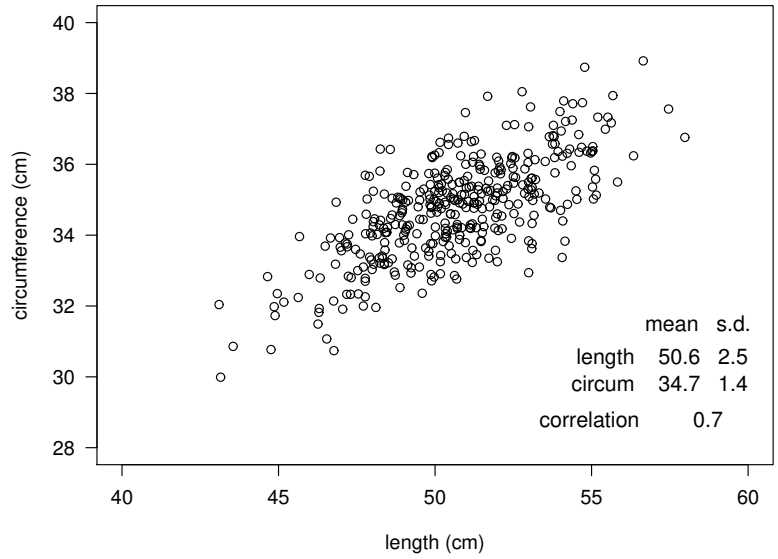
There are five parameters $\mu_x,\ \sigma_x,\ \mu_y,\ \sigma_y$ and ρ
which are the **means** and **standard deviations** of X and Y and their **correlation coefficient**.

From these, any probabilities (concerning X and Y) can be calculated.

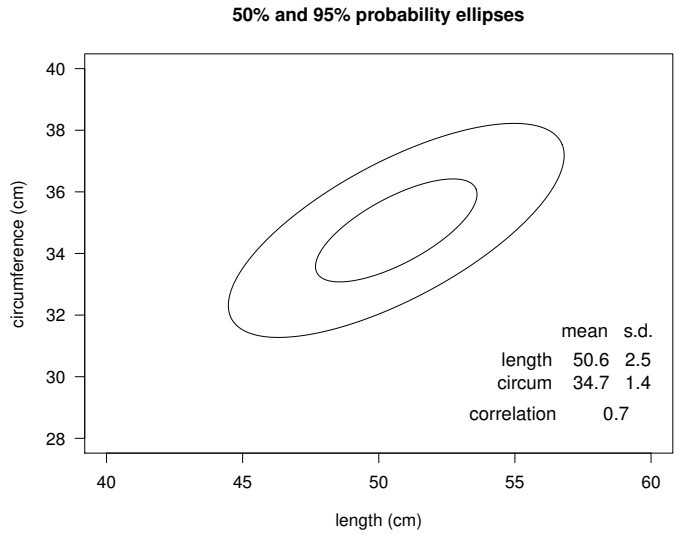
Contours of equal probability density are ellipses in the (x, y) plane.

Bivariate data: length and head circumference for 382 baby boys

Scatter plot and summary statistics:



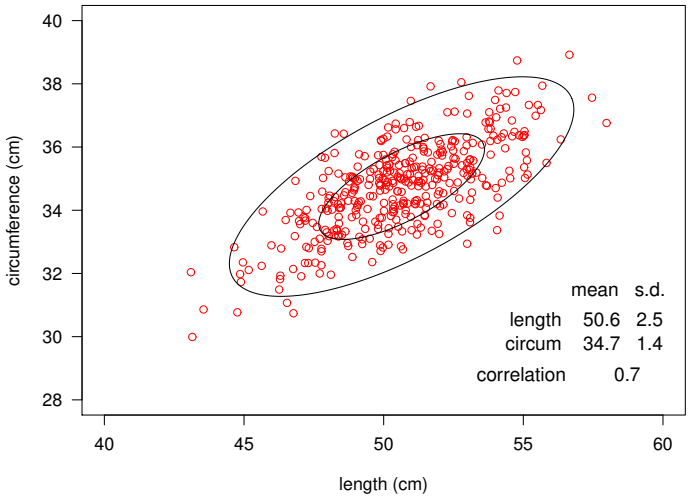
Here are some ellipses of constant probability density



— using parameter values estimated from the boys data

9 / 59

Here they are again with the data superimposed



10 / 59

All marginal and conditional distributions are Normal

Variable	Mean	Standard Deviation
X	μ_x	σ_x
Y	μ_y	σ_y
Y given x	$\mu_y + \rho \dfrac{\sigma_y}{\sigma_x}(x - \mu_x)$	$\sigma_y\sqrt{1-\rho^2}$
X given y	$\mu_x + \rho \dfrac{\sigma_x}{\sigma_y}(y - \mu_y)$	$\sigma_x\sqrt{1-\rho^2}$

And all regressions are **straight lines**.

The **mean** of Y given x is a straight line (as a function of x)
— it goes through (μ_x, μ_y) and has slope $\rho\,\sigma_y/\sigma_x$

The **standard deviation** of Y given x does not depend on x

Similarly for X given y

11 / 59

The method of least squares

12 / 59

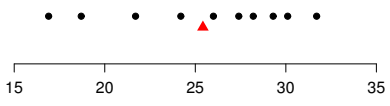
Univariate data

Question: Given n measurements $x_1, x_2, \ldots, x_n$, what value of a is closest to them in the sense of **least squares**?

i.e., what value of a minimises $(x_1 - a)^2 + (x_2 - a)^2 + \cdots + (x_n - a)^2$?

Answer: the mean value

$$a = (x_1 + x_2 + \ldots + x_n)/n$$



This is pure geometry.

There is no statistical model or distribution required.

13 / 59

Bivariate data

Question: Given n bivariate measurements $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, what values of a and b give the straight line $y = a + bx$ that is closest to them in the sense of least squares?

Answer: It depends **what** least squares you mean. For example:

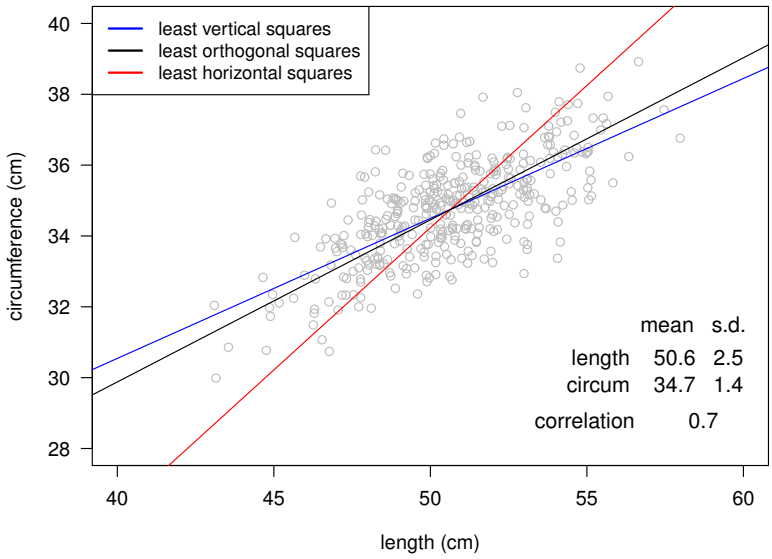
(a) Minimise $(y_1 - a - bx_1)^2 + (y_2 - a - bx_2)^2 + \cdots + (y_n - a - bx_n)^2$
then $b = r_{xy}s_y/s_x$ and $a = \bar{y} - b\bar{x}$

(b) Minimise $(x_1 - a - by_1)^2 + (x_2 - a - by_2)^2 + \cdots + (x_n - a - by_n)^2$
then $b = s_x/r_{xy}s_y$ and $a = \bar{y} - b\bar{x}$

(c) Minimise the sum of squares of the perpendicular distances from (x_i, y_i) to the line.
then $b = D + \sqrt{1 + D^2}$ where $D = \dfrac{s_y^2 - s_x^2}{2r_{xy}s_xs_y}$ and $a = \bar{y} - b\bar{x}$

Again, this is pure geometry, requiring no statistical model.

Here are the three lines for the boys data



What do these lines tell us and what are they for?

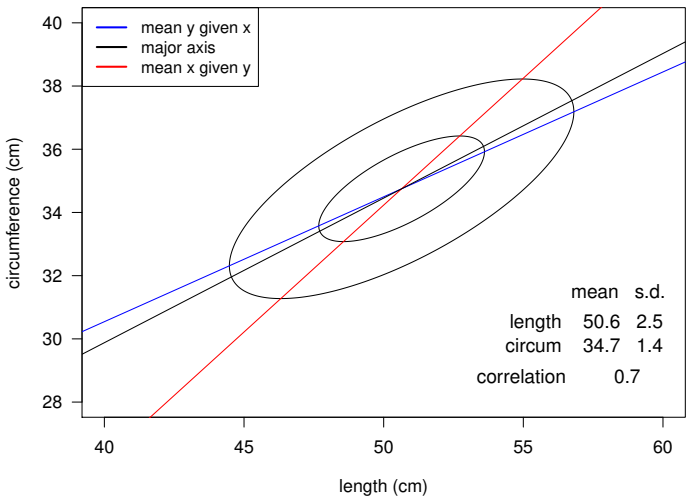
It transpires that

(a) shows the average y for a given x (as a function of x)

(b) shows the average x for a given y

(c) shows the direction of greatest variance of all linear combinations of x and y. It is called the **first principal component**

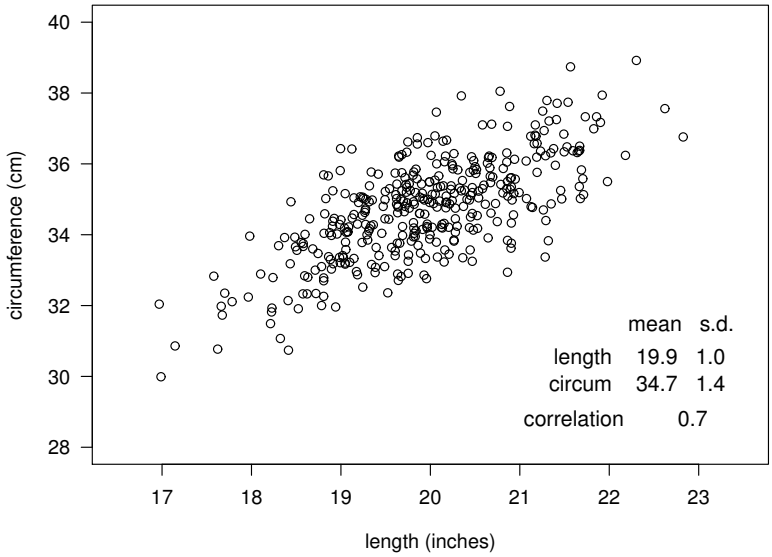
Here are the three lines with the bivariate normal probability ellipses



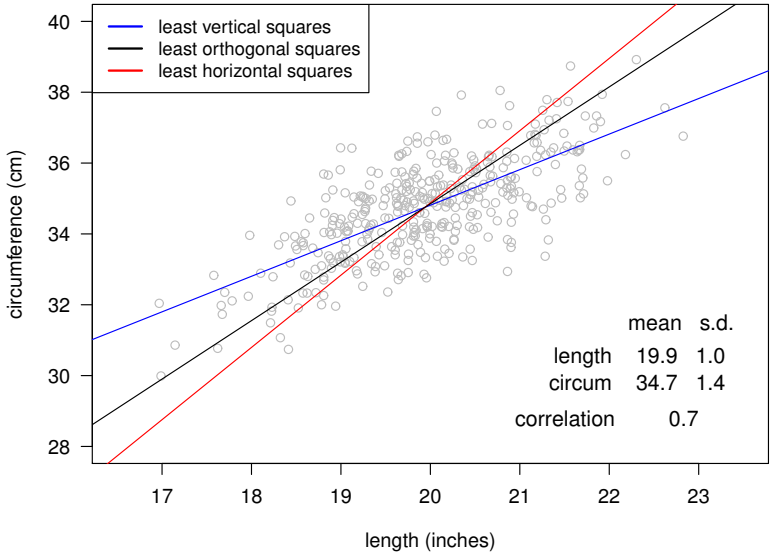
The **least orthogonal squares** (black) line, or **first principal component**, is also the direction of the **major axis** of the ellipses.

A problem of scales

Suppose we change the units of measurement of x or y. Here are the boys data again but with length now measured in inches!



What happens to the three lines?



We see that:

For cases **(a)** and **(b)** the least squares line is the same as before

— but expressed in the new units.

But **not so** for case **(c).** We get an essentially different line!

The **principal components** depend on the scales of measurement of the variables.

For this reason, in **principal components analysis** the variables are often standardised to have unit standard deivation before applying the method. But this affects the interpretation.

Normal distributions and least squares

When measurements come from a **Normal** distribution (univariate or multivariate) we can use the method of **maximum likelihood** to estimate the population parameters. These estimates turn out to be the **sample** means, standard deviations and and pairwise correlation coefficients.

Furthermore, for **location parameters** in linear models, the maximum likelihood estimates are the same as the **least squares** estimates.

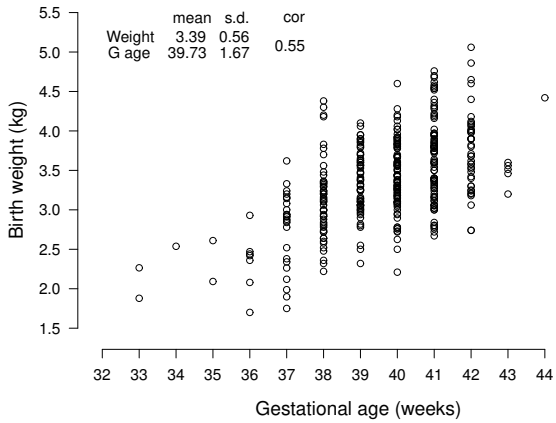
For example

- For univariate observations from a Normal population, the sample mean is the maximum likelihood estimate of the population mean
- For bivariate data where y_i is from a normal population with mean $\alpha + \beta x_i$ and standard deviation σ then the maximum likelihood estimates of α and β are the same as the least squares estimates in example (a) above.

Dispersion parameters are usually estimated from the residual scatter.

Example: birth weight and gestational age for 382 baby boys

Scatter plot and summary statistics:



Fit a linear regression model with

$y =$ birth weight and

$x =$ gestational age $- 40$ weeks.

23 / 59

Computer output from R software

```
Call: lm(formula = Bwt ~ g.age)

Residuals:
     Min       1Q   Median       3Q      Max
-1.23069 -0.31311 -0.00296  0.28962  1.30962

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.44069    0.02453  140.24   <2e-16 ***
g.age        0.18515    0.01449   12.78   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4732 on 380 degrees of freedom
Multiple R-squared:  0.3005,    Adjusted R-squared:  0.2987
F-statistic: 163.2 on 1 and 380 DF,  p-value: < 2.2e-16
```

24 / 59

What is the fitted model?

This is given by the **two coefficients** and the **residual standard deviation**

The **intercept** formally estimates the mean y when $x = 0$, in this case, the mean birth weight when gestational age is 40 weeks.[a]

The **slope** estimates the change in average y when x increases by 1 unit, in this case the change in mean birth weight when gestational age increases by 1 week.

The **residual standard deviation** estimates the standard deviation of y when x is fixed, in this case the standard devation of birth weights for boys with the same gestational age.

Birth weights of babies with gestational age $x + 40$ weeks have estimated

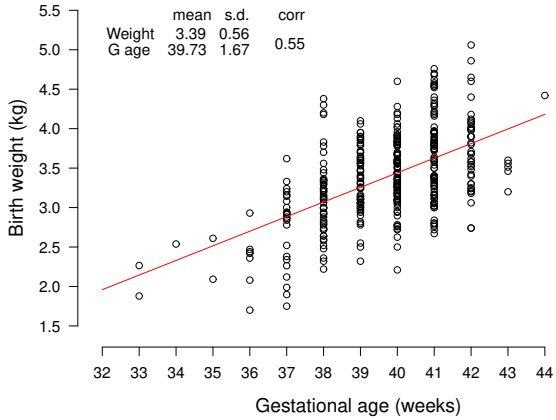
mean $3.440 + 0.185\,x$ kg and standard deviation 0.473 kg.

Don't forget to quote the residual standard deviation! We would also normally present the sample size, standard errors, residual degrees of freedom, and other diagnostics where appropriate.

[a] Often the intercept term is not physically meaningful by itself (e.g., if $x =$ gestational age then $x = 0$ is not possible in practice) but it is when combined with other parameters.

25 / 59

Data with fitted mean line



Compare the residual standard deviation 0.47 kg
with the marginal standard deviation 0.56 kg

— controlling for gestational age has reduced the scatter about the mean

— formally the marginal standard deviation of y (0.56 kg)
is the residual standard devation for the "null" model

26 / 59

Hypothesis tests

The output gives 3 P-values, relating to:

- `(intercept)`, testing the null hypothesis that when $x = 0$ the mean y equals 0

  — usually this is of no interest

- `g.age`, testing the null hypothesis that the (true) regression slope is 0

  — this is of interest if that null hypothsis is

- `F-statistic`, testing the null hypothesis that the mean y does not depend on x (this is called the "null" model)

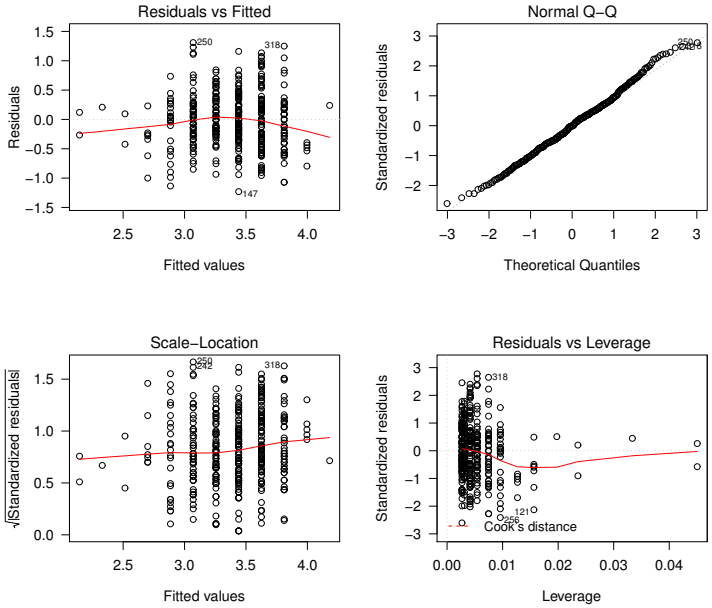
  — with just one explanatory variable, this is equivalent to the previous test, and the P-value is the same (note $12.78^2 = 163.3$).

  — it is also equivalent to testing the hypothesis that the (true) correlation coefficient is zero.

As always, their relevance depends on the context.

With more than one explanatory variable, P-values for individual coefficients are harder to interpret because they are conditional on other terms in the regression model. The null model says that the mean y does not depend on *any* of the x variables.

Diagnostic plots from R output



What use is R^2 ?

Here (with just one explanatory variable) R^2 is the square of the correlation coefficient between x and y (note: $0.55^2 = 0.30$).

Formally, R^2 measures the proportion of the variance of y that is "explained" by x. This interpretation extends to regression with several explanatory variables.

But:

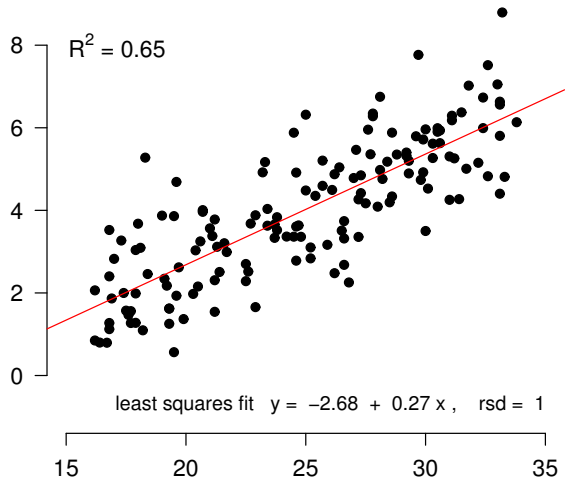
- variance **not** explained is more important than variance explained
- knowing the **amount** of variation not explained is more informative than knowing the **proportion** of it
- the **residual standard deviation** tells us explicitly how much variation in y is left over after accounting for x. **We need to know this!**

What use is R^2 ? (example)

Here are simulated values of $Y|x$ from Normal distributions with

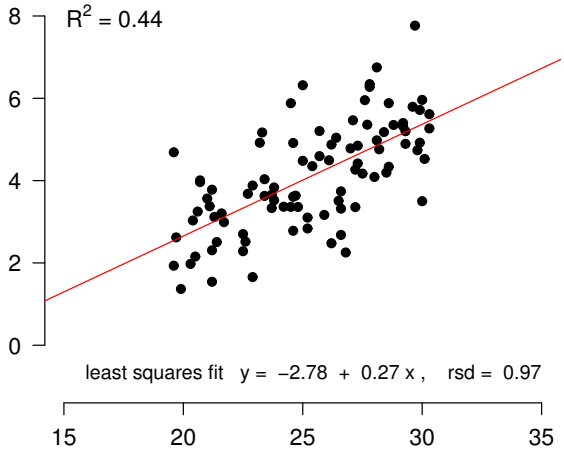
$$\mathsf{E}(Y|x) = -2.25 + 0.25x \qquad \text{and} \qquad \mathsf{var}(Y|x) = 1$$

plus the fitted least squares line, where x varies between 16 and 34



What use is R^2 ? (example continued)

Here are the results using just the data where x is between 19 and 31:



The fitted line and residual standard deviation are practically the same, but R^2 has reduced from 0.65 to 0.44.

31 / 59

What use is R^2 indeed!

In many applications, R^2 is of little use.

Often, R^2 differs between samples just because x varies by different amounts.

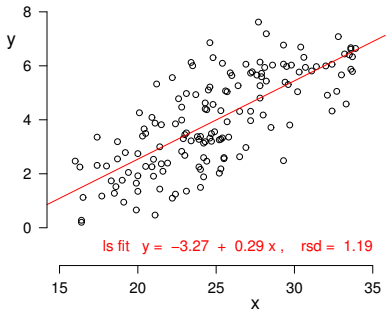
A higher value of R^2 does not mean that the fitted linear regression is “better”.

On the other hand, a smaller residual standard deviation does mean that y varies less when x is fixed.

32 / 59

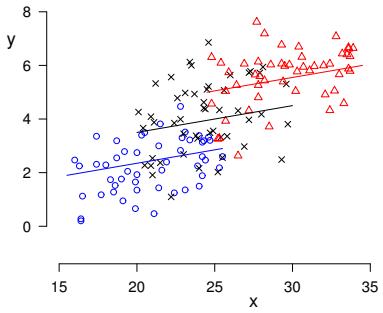
Within- and between-group slopes

Here is a scatter plot of some bivariate data showing a linear regression relationship with a slope of about 0.29.



In fact the data are from three groups with a common within-group slope of 0.1 but with different means.

Care is needed when interpreting regression slopes for data that are pooled across groups.



33 / 59

Regression dilution — what is it?

Suppose we wish to fit a straight line $\text{mean}(y) = \alpha + \beta x$
to describe how the mean of an outcome variable y
depends on a predictor (or explanatory) variable x.

But we cannot observe x precisely. Instead we observe $w = x + e$
where e is a random error with mean 0.

Then the slope of the regression of y on the observed w does not equal β.

- the regression of y on x has slope $\beta = \dfrac{\text{cov}(x, y)}{\text{var}(x)}$

- the regression of y on w has slope $\dfrac{\text{cov}(w, y)}{\text{var}(w)} = \dfrac{\text{cov}(x + e, y)}{\text{var}(x + e)}$

  if e is independent of x and y this equals $\dfrac{\text{cov}(x, y)}{\text{var}(x) + \text{var}(e)} = \dfrac{\beta}{1 + \text{var}(e)/\text{var}(x)}$

34 / 59

Regression dilution — does it matter?

Yes – if we want to estimate or test hypotheses about the parameter β

No – if we want to predict or estimate the mean of y
– or if we want to use w as an operational measurement

Example from occupational epidemiology: does cumulative exposure to carbon black in factories cause respiratory morbidity?

$y =$ measured lung function (FEV1)
$x =$ exposure to carbon black over time
$w =$ measurements of inhalable carbon black made on several days

We might want to test the hypothesis that $\beta = 0$.

We can't measure x directly, and regression of y on w might fail to reveal an important effect.

Regression dilution — what can we do about it?

Approaches:

1. Do nothing (this has advantages)
2. Get better measurents of x
3. Use the regression of y on w and apply a correction factor to the estimated slope
4. Look at the literature on "errors in variables" models.

Without further data, errors in variables models are not identifiable, so assumptions about the measurement error variances are needed.

Ordinal outcome variables

Examples: Severity of disease (absent, mild, moderate, severe),

Temperature sensitivity vote (At present I feel: cold, cool, slightly cool, neutral, slightly warm, warm, hot)

Nature: Each individual (unit) is assigned to one of several ordered categories. The difference between adjacent categories does not necessarily have the same meaning at different points on the scale.

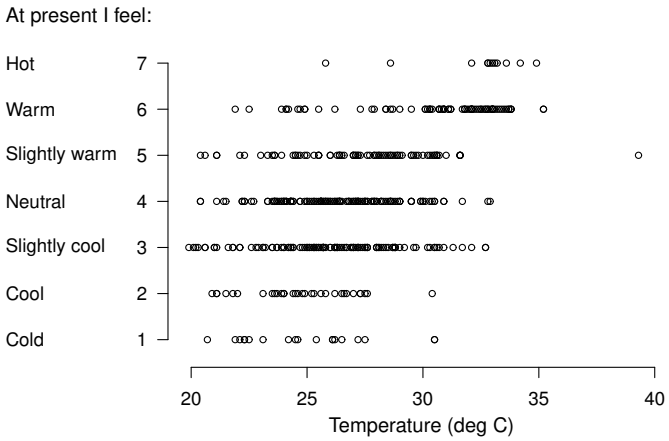
Approaches:

1. Create quantitative data by assigning a numerical score to each category.
2. Imagine we have coarsely grouped data from an underlying continuous distribution; again use methods for quantitative data.
3. Reduce to binary variables by merging categories and use methods for binary data.
4. Treat directly as ordinal — e.g., using proportional odds or cumulative link models.

38 / 59

Example: ASHRAE temperature votes

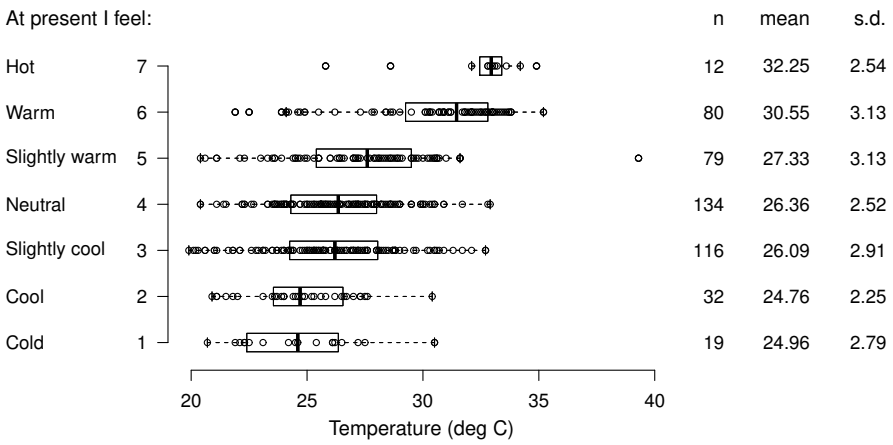
Illustrative data showing 472 temperature sensation votes (ASHRAE scale) plotted against actual temperatures in °C:



How does ASHRAE score depend on temperature?

39 / 59

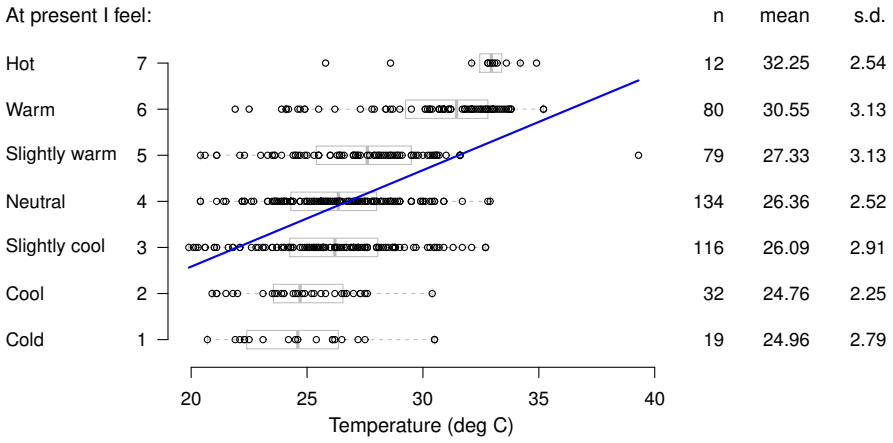
Another version with box plots and summary statistics



This shows how temperature varies for each ASHRAE score.

40 / 59

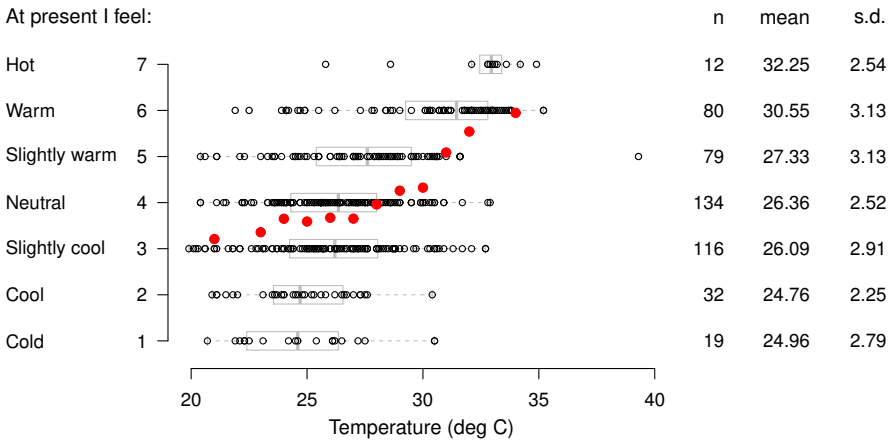
Another plot with a regression line added



Blue line: least squares fit $y = -1.60 + 0.21x$ residual s.d. $= 1.20$

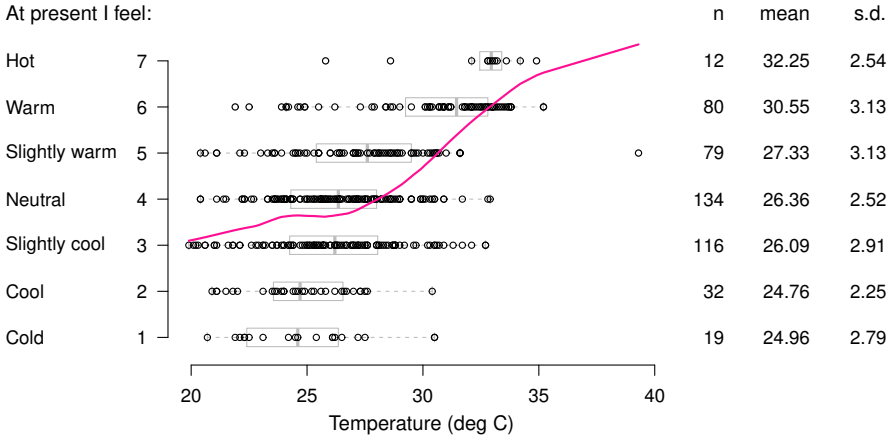
41 / 59

Another plot with some mean ASHRAE scores added



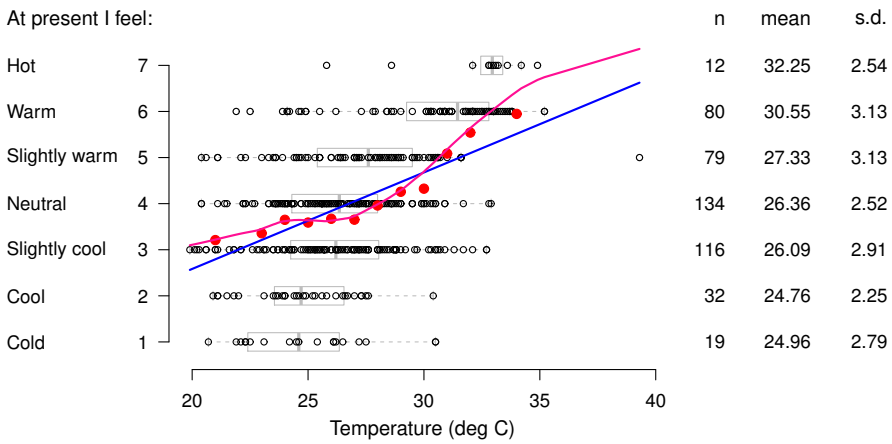
Red dots show the mean ASHRAE score *(regarded as a quantitative variable)*
for temperatures in each 1°C interval,
with tails grouped below 23.5°C and above 32.5°C

. . . now with a lowess line



Pink line: lowess with smoothing span `f = 0.35`

. . . and now all together



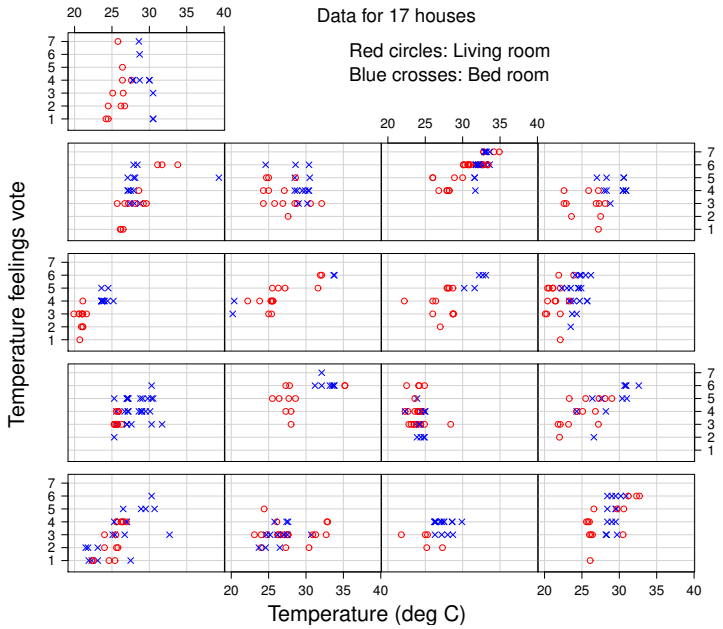
Pink line: lowess with smoothing span `f = 0.35`

Blue line: least squares fit $y = -1.60 + 0.21x$ residual s.d. $= 1.20$

— a straight line looks like a poor description

44 / 59

The same data re-plotted by House and Room type



Look for patterns between houses and rooms

Features seen from the pooled data might be due to differences between houses or rooms . . .

. . . or to other uncotrolled factors

45 / 59

Proportions and cumulative proportions

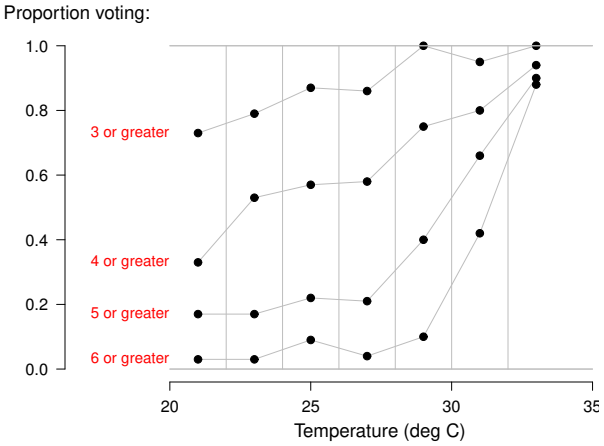
Temperature	Proportion voting:					
range °C	$1, 2$	3	4	5	$6, 7$	n
≤ 22	0.27	0.40	0.17	0.13	0.03	30
22 – 24	0.21	0.26	0.36	0.14	0.03	58
24 – 26	0.13	0.30	0.36	0.13	0.09	101
26 – 28	0.14	0.28	0.36	0.18	0.04	107
28 – 30	0.00	0.25	0.34	0.30	0.10	67
30 – 32	0.05	0.15	0.14	0.24	0.42	59
> 32	0.00	0.06	0.04	0.02	0.88	50

	≥ 1	≥ 3	≥ 4	≥ 5	≥ 6
≤ 22	1.00	0.73	0.33	0.17	0.03
22 – 24	1.00	0.79	0.53	0.17	0.03
24 – 26	1.00	0.87	0.57	0.22	0.09
26 – 28	1.00	0.86	0.58	0.21	0.04
28 – 30	1.00	1.00	0.75	0.40	0.10
30 – 32	1.00	0.95	0.80	0.66	0.42
> 32	1.00	1.00	0.94	0.90	0.88

46 / 59

Plotting ordinal data

When there are more then 2 ordered categories, proportions for intermediate categories can be hard to interpret. It is more logical to plot **cumulative** proportions— e.g., as given in the lower table above.



This shows the proportion voting y or greater for y = 3, 4, 5 and 6 for each $2°$C temperature interval.

Vertical distances between the curves give the proportions in the upper table.

47 / 59

Binary outcome variables — logistic and probit regression

Define an outcome variable y which equals 1 when the ASHRAE vote is 5, 6 or 7 (i.e., "I feel slightly warm or warmer") and 0 when the preference vote is below 5 (i.e., "I feel neutral or cooler"). Let p denote the probability that $y = 1$ when the temperature is x.

A logistic regression equation is of the form

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta x$$

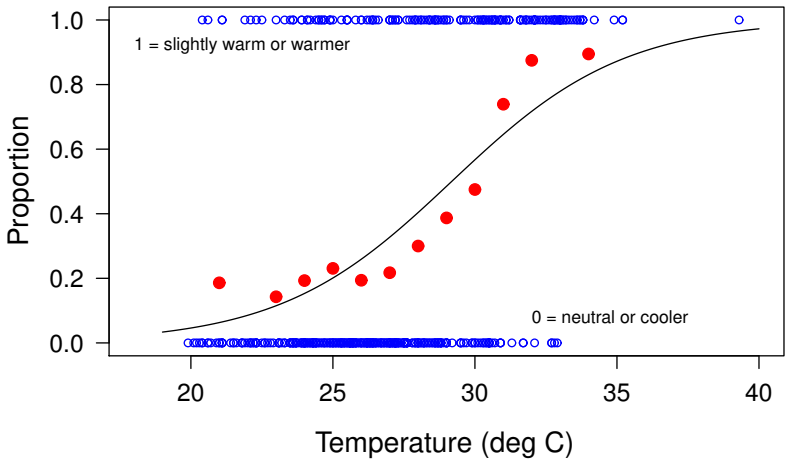
That is, the log odds of *slightly warm or warmer* changes with temperature in a straight line with intercept α and slope β.

This equation may be written to express p as a function of x:

$$p = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}$$

which is not a straight line, but an S-shaped (or reverse S-shaped) curve.

Fitted logistic regression curve



Blue dots denote observed responses

Red dots denote observed proportions of 1s in 1°C intervals

Solid black line is the fitted logistic curve with $\alpha = -9.63$ and $\beta = 0.33$

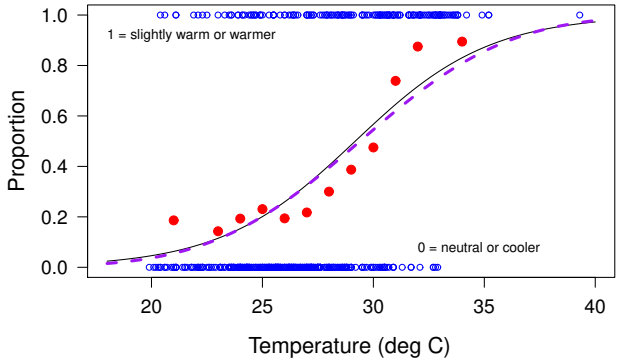
Probit regression models

These are very similar to logistic regression models. The model equations are of the form

$$\Phi^{-1}(p) = \gamma + \delta x \qquad \text{and} \qquad p = \Phi(\gamma + \delta x)$$

where $\Phi(z)$ is the standard Normal distribution function.

In practice there is very little difference between logistic and probit models in terms of how well they fit data; the main differences are in the interpretation of the parameters. Here is the previous graph again, with the logistic curve in black and a fitted probit curve (purple broken line), with $\gamma = -5.59$ and $\delta = 0.19$.



50 / 59

Ordinal regression models

Let y denote the ASHRAE score, which is an ordinal variable taking values $1, 2, 3, 4, 5, 6$ or 7.

Let q_k denote the probability that $y \geq k$, for $k = 1, 2, \ldots, 7$.

For example, q_5 is the probability that the individual feels sightly warm or warmer. And of course $q_1 = 1$.

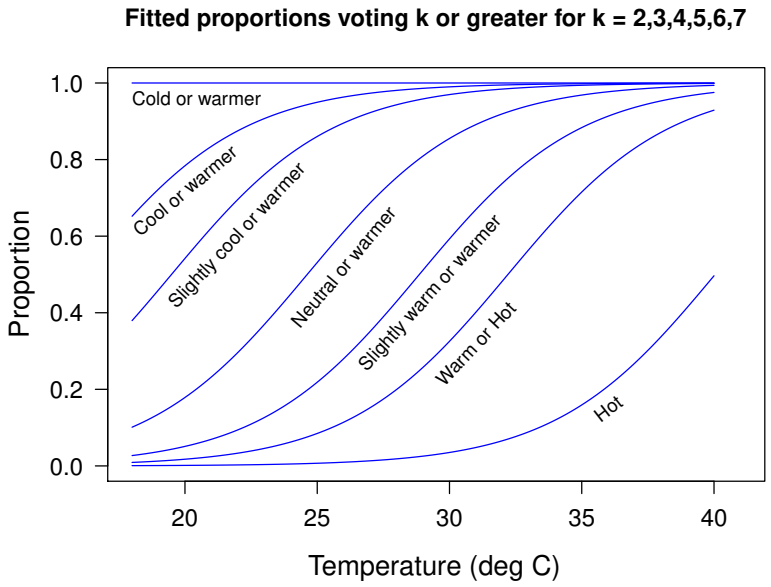
An ordinal regression model will specify how q_k depends on explanatory variables such as actual temperature, for each value of k.

Often a *proportional odds* model is used, where the coefficient β is the same for every value of k.

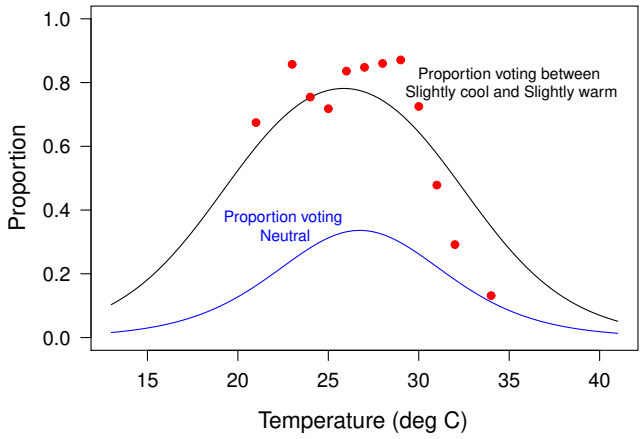
These models are an extension of models for binary responses (when y takes only two values).

51 / 59

Fitted proportional odds regression curves



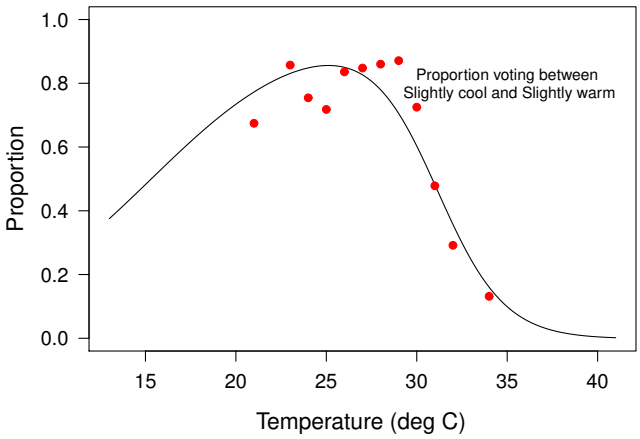
Fitted proportions voting “comfortable” at each temperature



The blue curve here is the vertical distance between the curves “Slightly warm or warmer” and “Neutral or warmer” in the previous figure.

Likewise, the black curve is the vertical distance between the curves “Warm or Hot” and “Slightly cool or warmer” in the previous figure. Red dots are observed proportions.

Another version, but using a *non*-proportional odds model



The black curve is the vertical distance between the fitted logistic curves for “Warm or Hot” and “Slightly cool or warmer”, but now with different intercepts *and* slopes. Red dots are observed proportions.

Looking at several variables

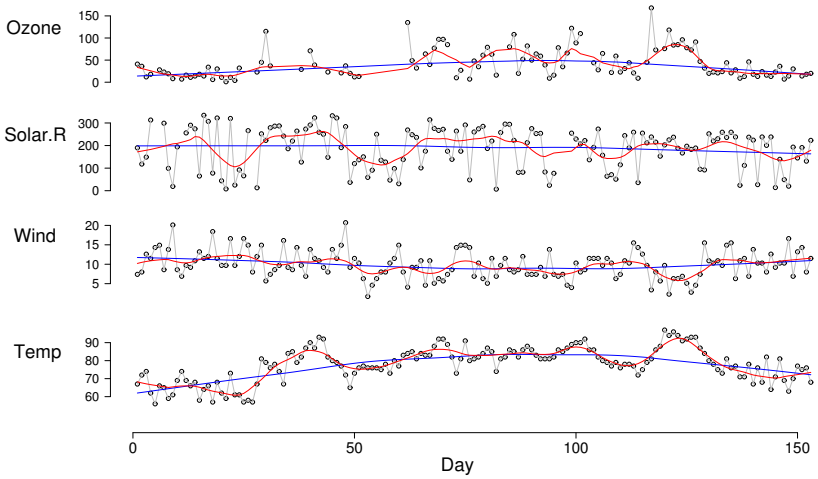
Example: air quality data

	Ozone	Solar.R	Wind	Temp	Month	Day
1	41	190	7.4	67	5	1
2	36	118	8.0	72	5	2
3	12	149	12.6	74	5	3
4	18	313	11.5	62	5	4
5	NA	NA	14.3	56	5	5
6	28	NA	14.9	66	5	6
7	23	299	8.6	65	5	7
...						
151	14	191	14.3	75	9	28
152	18	131	8.0	76	9	29
153	20	223	11.5	68	9	30

Daily readings of the following air quality values for May 1, 1973 (a Tuesday) to September 30, 1973.

- 'Ozone': Mean ozone in parts per billion from 1300 to 1500 hours at Roosevelt Island
- 'Solar.R': Solar radiation in Langleys in the frequency band 4000-7700 Angstroms from 0800 to 1200 hours at Central Park
- 'Wind': Average wind speed in miles per hour at 0700 and 1000 hours at La Guardia Airport
- 'Temp': Maximum daily temperature in degrees Fahrenheit at La Guardia Airport.

Time series plots with lowess smoothing

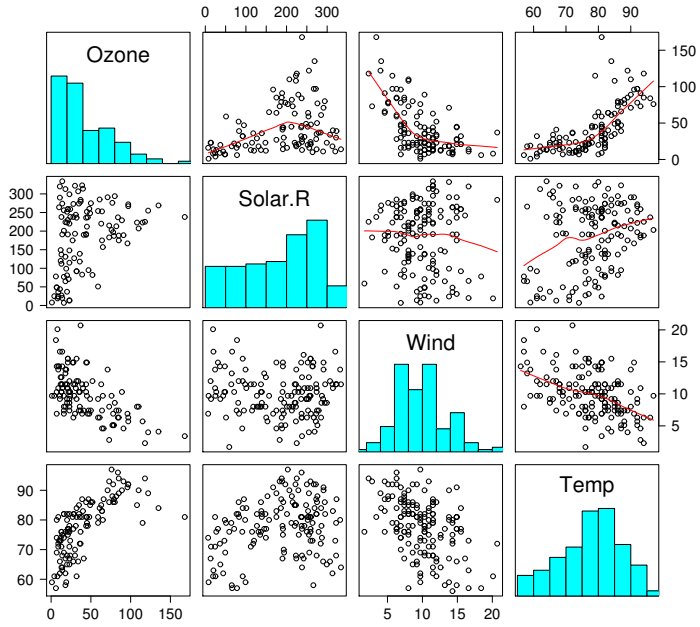


Blue lines use the default $f = 2/3$, red lines use $f = 0.1$

f is the 'smoother span' — the proportion of points in the plot that influence the smooth at each value. Larger values give more smoothness.

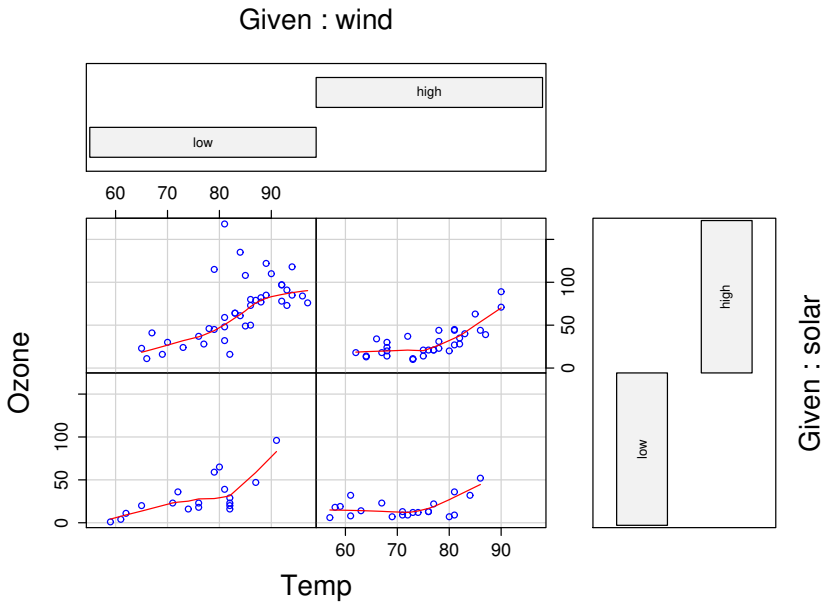
57 / 59

Pairwise scatter plots



58 / 59

Coplot of Ozone against Temp given Wind and Solar.R grouped



59 / 59