EMTH211 Statistics Section - Notes

Richard Vale

September 25, 2014

1 Introduction

These notes will cover the last twelve lectures of the course. They are partially based on handwritten notes by Dominic Lee. It is not intended that everything in these notes be examinable. Only the parts covered in class will be examinable.

2 Data

2.1

We are going to apply linear algebra to the analysis of data. The data will consist of measurements of things. For example, here is a table of pressure and temperature measurements for a boiler.

Temp (°C)	Pressure (kPa)
0	91
10	95
20	100
30	101
40	107
50	112

We could write the temperature and pressure measurements as vectors

$$\mathbf{x} = (0, 10, 20, 30, 40, 50)$$

and

$$\mathbf{y} = (91, 95, 100, 101, 107, 112).$$

We might be interested in the relation between $\mathbf{x}$ and $\mathbf{y}$. For example, does increasing the temperature cause the pressure to increase? Does increasing the pressure cause the temperature to increase? Can we predict one if we know the other? When talking about the variables without specific data values in mind, we often write them in lower case as x and y. In this example we might say $x =$ temperature, $y =$ pressure.

2.2

We might prefer to plot the (temperature, pressure) pairs as in Figure 1. This plot can help us to answer our questions. What does it suggest to you?

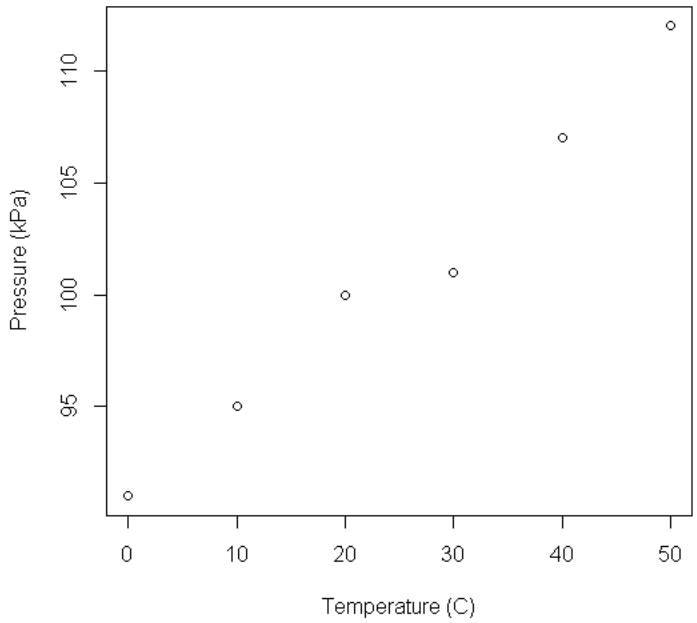

Figure 1: Running example: pressure versus temperature in a boiler.

3 Centering

3.1

If you have seen it before, you might remember that the ideal gas law says that pressure is proportional to temperature. $P \propto T$. This means that there is a constant α, a number, such that $P = \alpha T$. Can we find the value of this constant from our data? Well, $91 = \alpha.0$ and $95 = \alpha.10$ and so $\alpha = 95/10 = 91/0$. There is no suitable value of α. What went wrong?

3.2

In physics, temperature has to be measured in Kelvin, so $0°\text{C}$ should really be written $273K$ and so on for the other temperatures. If we didn't know anything about temperature, this would be very confusing. The *scale* on which our measurements are measured is a distraction. We can get rid of this distraction by centering our measurements. This is done by subtracting the mean.

3.3 The Mean

The mean of a vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is

$$\overline{\mathbf{x}} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n).$$

To centre a vector $\mathbf{x}$, you subtract its mean from each entry. The centered version of $\mathbf{x}$ is given by

$$\mathbf{x} - \overline{\mathbf{x}}\mathbf{1} = (x_1 - \overline{\mathbf{x}}, \ldots, x_n - \overline{\mathbf{x}})$$

where $\mathbf{1}$ is a vector $(1, 1, \ldots, 1)$ of all ones. The notation $\overline{\mathbf{x}}$ looks a bit like a vector, so it is better to write the mean of $\mathbf{x}$ as $\overline{x}$ to remind us that it is a scalar. For the temperature and pressure vectors we have

$$\overline{x} = \frac{1}{6}(0 + 10 + 20 + 30 + 40 + 50) = 25, \qquad \overline{y} = 303/3 = 101$$

and

$$\mathbf{x} - \overline{x}\mathbf{1} = (-25, -15, -5, 5, 15, 25)$$
$$\mathbf{y} - \overline{y}\mathbf{1} = (-10, -6, -1, 0, 6, 11).$$

3.4

The mean of a centered vector is always zero because the sum of the entries of $\mathbf{x} - \overline{x}\mathbf{1}$ is

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \overline{x}) &= \sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\overline{x} \\ &= n\overline{x} - n\overline{x} \\ &= 0.\end{aligned}$$

3.5

The mean gives us a way to split $\mathbf{R}^n$ into two orthogonal subspaces

$$V = \{\mathbf{x} : \sum_i x_i = 0\}$$

the space of all vectors which sum to zero, and

$$V^\perp = \operatorname{span}\{\mathbf{1}\}$$

the space of all vectors which are scalar multiples of $(1, 1, \ldots, 1)$. Because every $\mathbf{x}$ can be written

$$\mathbf{x} = (\mathbf{x} - \overline{x}\mathbf{1}) + \mathbf{1}$$

we have $\mathbf{R}^n = V + V^\perp$ and we can check that V and $V^\perp$ are orthogonal since for every $\mathbf{x}$ we have

$$\begin{aligned}\langle \mathbf{x} - \overline{x}\mathbf{1}, \mathbf{1}\rangle &= \langle \mathbf{x}, \mathbf{1}\rangle - \overline{x}\langle \mathbf{1}, \mathbf{1}\rangle \\ &= \sum_{i=1}^{n} x_i - \overline{x}n \\ &= 0\end{aligned}$$

where $\langle -, -\rangle$ is the dot product, so $\mathbf{R}^n = V \oplus V^\perp$.

3.6

Why will it be helpful to centre the data? Going back to the ideal gas law, suppose we know that $P \propto T$ but we cannot remember what scale should be used to measure temperature. Then we know that

$$P = \alpha(T + T_0) = \alpha T + \alpha T_0$$

for some unknown T_0. If we take the means of a collection of measurements satisfying this relationship, you can check that

$$\overline{P} = \alpha\overline{T} + \alpha T_0$$

and so

$$P - \overline{P} = \alpha T + \alpha T_0 - \alpha \overline{T} - \alpha T_0 = \alpha(T - \overline{T})$$

which is great because now we can work out α from the centered measurements without worrying about where the zero is on our temperature (or pressure) scale! (But if you now try to find α, you will find that there is still no value of α that fits, because of measurement errors.)

4 Spread

4.1 Variance and standard deviation

The mean gives us a way to measure what the typical values of the variables are. We would also like to measure how spread out the values are around the mean. This can be measured using the *variance*, which is defined by

$$\mathrm{var}(\mathbf{x}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

In vector notation, this is

$$\mathrm{var}(\mathbf{x}) = \frac{1}{n-1}||\mathbf{x} - \overline{x}\mathbf{1}||^2$$

where $||\cdot||$ is the Euclidean norm (another name for the 2–norm.)

4.2

The *standard deviation* is the square root of the variance

$$\mathrm{sd}(\mathbf{x}) = \frac{1}{\sqrt{n-1}}||\mathbf{x} - \overline{x}\mathbf{1}||.$$

The advantage of doing this is that the standard deviation is measured in the same units in which $\mathbf{x}$ is measured. So, for example, the standard deviation of our temperature vector $\mathbf{x}$ comes out to be about $18.7°\mathrm{C}$ from the following calculation:

$$\mathbf{x} = (0, 10, 20, 30, 40, 50)$$

$$\mathbf{x} - \overline{x}\mathbf{1} = (-25, -15, -5, 5, 15, 25)$$

$$\mathrm{var}(\mathbf{x}) = \frac{1}{5}(25^2 + 15^2 + 5^2 + 5^2 + 15^2 + 25^2)$$

$$\mathrm{sd}(\mathbf{x}) = \sqrt{\frac{1}{5}(1750)} \simeq 18.7$$

The standard deviation is also written as s_x and the variance as s_x^2.

4.3 Aside: Why the Euclidean norm? Why $n-1$?

The motivation behind the definition of the variance is that the squared length of the centered data is a measure of how much the data varies from its centre, and dividing by $n-1$ enables us to calculate the variance "per dimension" in the space V of Section 3.5. This all sounds plausible, but there are some natural questions. *Why* choose the 2–norm to analyse data? Why not choose some other norm, for example, the 1–norm?

4.4

Unfortunately, this question has no obvious answer. The original motivation behind the choice of the 2–norm was that it made the calculations easier. It is also a natural thing to do if we agree that the mean is the right measure of central tendency, because if $\mathbf{x}$ is a given vector then there is a unique number z which minimises

$$f(z) = \sum_{i=1}^{n} (z - x_i)^2$$

and that number must be the mean. [Exercise: prove this using calculus.] But this does not mean that we can't get other useful measures of centre and spread from other norms. In applications many different notions are indeed used. For example, MRI scans and jpeg images would not work if everybody used the 2–norm all the time! In this course, we will gloss over this point and stick to the mean and standard deviation.

4.5

Another natural question is: why the $n - 1$? This has probably cropped up in your earlier statistics classes. The usual answer is to do with getting an unbiased estimate of the variance in sampling theory. I don't really want to get into it here, but there really isn't much motivation for it. If you are analysing real data and n is so small that it makes a difference whether you use n or $n - 1$ in the denominator of the standard deviation, then you should think carefully about whether you ought to be using statistics at all.

5 Correlation

5.1

Everything we have done so far has only been applied to one variable. We want to understand the relationship between two variables. How does one vary when the other varies? Can we use one to predict the other?

5.2 Pearson correlation

If we have centered data vectors $\mathbf{x}$ and $\mathbf{y}$ then $\mathbf{y}$ is a perfect predictor of $\mathbf{x}$ when

$$\mathbf{y} = \alpha \mathbf{x}$$

for some α, just like in the case of the ideal gas law. If the ideal gas law holds, then pressure determines temperature and temperature determines pressure. In this case the vectors $\mathbf{y}$ and $\mathbf{x}$ are parallel. What about when $\mathbf{y}$ cannot be used to predict $\mathbf{x}$? This is less clear, but the opposite of being parallel is being orthogonal, so we will say that vectors are uncorrelated when they are orthogonal.

5.3

The *Pearson correlation* measures how close two data vectors are to being parallel. It is simply defined as

$$\cos(\theta)$$

where

$$\theta = \angle \mathbf{x} - \overline{x}\mathbf{1}, \mathbf{y} - \overline{y}\mathbf{1}$$

is the angle between the centered $\mathbf{x}$ and $\mathbf{y}$. Because there is a formula for the dot product $\mathbf{v} \cdot \mathbf{w} = ||\mathbf{v}||||\mathbf{w}|| \cos(\theta)$, the Pearson correlation can also be written

$$\mathrm{corr}(\mathbf{x}, \mathbf{y}) = \frac{(\mathbf{x} - \overline{x}\mathbf{1}) \cdot (\mathbf{y} - \overline{y}\mathbf{1})}{||\mathbf{x} - \overline{x}\mathbf{1}||||\mathbf{y} - \overline{y}\mathbf{1}||}$$

The Pearson correlation is also just called the correlation and is written r_{xy}.

5.4

Example: for our temperature and pressure vectors we had

$$\mathbf{x} - \overline{x}\mathbf{1} = (-25, -15, -5, 5, 15, 25)$$
$$\mathbf{y} - \overline{y}\mathbf{1} = (-10, -6, -1, 0, 6, 11)$$

and so

$$r_{xy} = \frac{(-25)(-10) + (-15)(-6) + (-5)(-1) + (5)(0) + (15)(6) + (25)(11)}{\sqrt{25^2 + 15^2 + 5^2 + 5^2 + 15^2 + 25^2}\sqrt{10^2 + 6^2 + 1^2 + 0^2 + 6^2 + 11^2}} = \frac{710}{\sqrt{1750}\sqrt{294}} \simeq 0.9898.$$

Is this big or small? If you think about the definition, $r_{xy} = \cos(\theta)$, so the maximum possible value r_{xy} can take is 1, so this is a very high correlation. The centered temperature and pressure vectors are almost parallel, as we suspected from the ideal gas law.

5.5 Covariance

The *covariance* of two vectors $\mathbf{x}$ and $\mathbf{y}$ is defined by

$$\mathrm{cov}(\mathbf{x}, \mathbf{y}) = \frac{1}{n-1}(\mathbf{x} - \overline{x}\mathbf{1}) \cdot (\mathbf{y} - \overline{y}\mathbf{1})$$

and also denoted c_{xy}. Notice that $c_{xx} = s_x^2$; the covariance of a vector with itself is the variance. The covariance measures how $\mathbf{x}$ and $\mathbf{y}$ tend to vary with each other. Note that

$$r_{xy} = \frac{c_{xy}}{s_x s_y}$$

because the $n - 1$'s cancel in the numerator and denominator. This is the formula for r_{xy} that you often find in statistics texts. It is quite painful to prove from this definition that $-1 \leq r_{xy} \leq 1$. But with our definition, it is very easy because r_{xy} is the cosine of something and cosine always takes values between -1 and 1. Here is an example of linear algebra making our lives easier!

5.6 Properties of the correlation

We can work out some important properties of Pearson correlation:

- *Symmetry.* Because $(\mathbf{x} - \overline{x}\mathbf{1}) \cdot (\mathbf{y} - \overline{y}\mathbf{1}) = (\mathbf{y} - \overline{y}\mathbf{1}) \cdot (\mathbf{x} - \overline{x}\mathbf{1})$, it follows that
$$\mathrm{corr}(\mathbf{x}, \mathbf{y}) = \mathrm{corr}(\mathbf{y}, \mathbf{x}).$$
- *Location invariance.* Adding a number to all the entries of a vector does not change the correlation of that vector with any other vector. If we write $\mathbf{x} + b$ for the vector whose entries are $x_i + b$, $1 \leq i \leq n$, then $\overline{\mathbf{x} + b} = \overline{x} + b$ and so centering $\mathbf{x} + b$ gives the same result as centering $\mathbf{x}$. Since the correlation is the angle between the centered vectors, this does not change the correlation.
- *Scale invariance.* Multiplying a vector by a *positive* scalar does not change its angle with any other vector, so
$$\mathrm{corr}(a\mathbf{x}, \mathbf{y}) = \mathrm{corr}(\mathbf{x}, \mathbf{y})$$
if $a > 0$. On the other hand, multiplying by a *negative* scalar flips the sign of the correlation because
$$\begin{aligned}\frac{(a\mathbf{x} - \overline{a\overline{x}}\mathbf{1}) \cdot (\mathbf{y} - \overline{y}\mathbf{1})}{||a\mathbf{x} - \overline{a\overline{x}}\mathbf{1}||||\mathbf{y} - \overline{y}\mathbf{1}||} &= a\frac{(\mathbf{x} - \overline{x}\mathbf{1}) \cdot (\mathbf{y} - \overline{y}\mathbf{1})}{||a(\mathbf{x} - \overline{x}\mathbf{1})||||\mathbf{y} - \overline{y}\mathbf{1}||} \\ &= a\frac{(\mathbf{x} - \overline{x}\mathbf{1}) \cdot (\mathbf{y} - \overline{y}\mathbf{1})}{|a|||\mathbf{x} - \overline{x}\mathbf{1}||||\mathbf{y} - \overline{y}\mathbf{1}||} \\ &= \frac{a}{|a|}\mathrm{corr}(\mathbf{x}, \mathbf{y})\end{aligned}$$
and $a/|a| = -1$ if $a < 0$.

These properties make sense because if correlation is to measure the extent of the association between $\mathbf{x}$ and $\mathbf{y}$, it should not matter in what units $\mathbf{x}$ and $\mathbf{y}$ are measured. Converting, for example, between cm and inches corresponds to multiplication by a constant, and this does not change the correlation.

5.7 What does correlation measure?

Correlation measures the *linear* association between two variables. If one is a linear function of the other with a positive slope, then the correlation between them will be 1. If the slope is negative, then the correlation will be -1. But variables can be related in a non-linear way. For example, suppose we push a toy car at different velocities and measure its kinetic energy. Suppose $\mathbf{v} = (-2, -1, 0, 1, 2)$ is the vector of velocities and $\mathbf{E} = (4, 1, 0, 1, 4)$ is the vector of corresponding energies. Then

$$\text{cov}(\mathbf{v}, \mathbf{E}) = \frac{1}{4}((-2)(4) + (-1)(1) + 0.0 + 1.1 + 2.4) = 0$$

and so

$$\text{corr}(\mathbf{v}, \mathbf{E}) = 0$$

and velocity and energy are uncorrelated, even though there is a perfect relationship between them: $E = v^2$. The situation is depicted in Figure 2. Correlation cannot be used to detect associations that are not linear, so you should be careful about saying that there is no relationship between variables just because they happen to be uncorrelated.

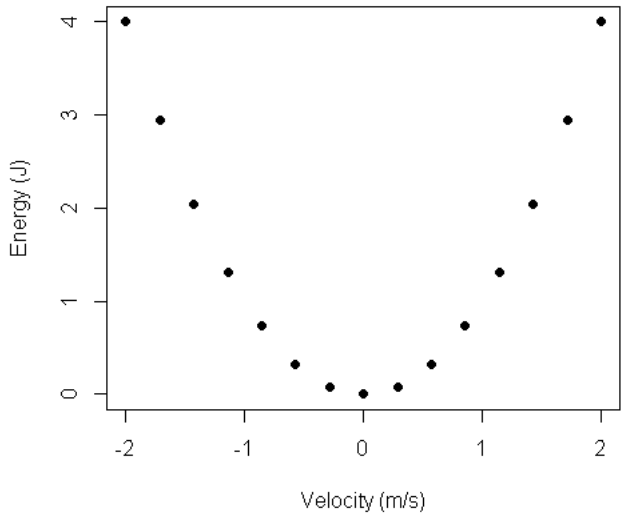

Figure 2: Correlation only measures *linear* association. Here there is a clear pattern, but the correlation is zero.

6 Exercises for week 1

1. If $\mathbf{x}$ is a vector, what is the correlation between $\mathbf{x}$ and $-\mathbf{x}$?
2. How does the *covariance* between two vectors change if one of them is multiplied by a scalar?
3. The number of birds observed at a feeder is observed to be smaller on cold days. Is the correlation between number of birds and temperature positive, negative or zero?

4. If more shipwrecks happen near the shore than further away, is the correlation between distance to the shore and number of shipwrecks positive, negative or zero?

5. If $2.5\text{cm} = 1\text{inch}$, how will the standard deviation of some measurements in cm change if we rewrite them in inches? How will the variance of the measurements change?

6. We are investigating the relationship between time and the stock price of a company. Time is measured in days since 1900. How will the correlation change if we choose to measure time in weeks since 1950 instead?

7. A person's BMI is defined by
$$BMI = \frac{\text{mass(kg)}}{(\text{height(m)})^2}$$
If we measure the mass and BMI of a number of people who are all 1.8m tall, what will be the correlation between mass and BMI?

8. If we measure the mass in pounds instead of kg, how will the correlation change?

9. If we take a group of people who all weigh 100kg and measure their BMI and height, will the correlation between height and BMI be positive, negative or zero?

10. The *Spearman correlation* between two data vectors is defined to be the Pearson correlation applied to their ranks. The rank of an entry is i if it is the i^{th} smallest. So for example, if $\mathbf{x} = (1, 2, 5)$ and $\mathbf{y} = (-5, -1, 0)$ then in $\mathbf{y}$, $\text{rank}(-5) = 1$, $\text{rank}(-1) = 2$ and $\text{rank}(0) = 3$. Similarly, the ranks of $\mathbf{x}$ are $(1, 2, 3)$ and so the Spearman correlation between $\mathbf{x}$ and $\mathbf{y}$ is $\text{corr}((1, 2, 3), (1, 2, 3)) = 1$.
    (a) What is the Pearson correlation between $(1, 2, 3)$ and $(1, 4, 9)$?
    (b) What is the Spearman correlation between $(1, 2, 3)$ and $(1, 4, 9)$?
    (c) If Pearson correlation measures linear association, what does Spearman correlation measure?

7 Simple linear regression

7.1

Correlation gives us one way to measure the strength of a linear relationship. Now we want to find an equation for the linear relationship. There is seldom an exact linear relationship, but we can find an equation which is close to exact in some sense. Finding such an equation is called linear regression. If we have just two variables x and y it is called "simple". This type of linear regression can't go too badly wrong because you can always plot y and x to see what is happening. If we have more than two variables, it is called "multiple" regression. This can go wrong in all sorts of interesting ways, which we will discuss next week.

7.2 Simple linear regression

If we have vectors $\mathbf{x}$ and $\mathbf{y}$ of measurements with n entries, we can look for a relationship of the form
$$y_i = b_0 + b_1 x_i + e_i$$
where e_i is an error term. As a vector equation, this would be written
$$\mathbf{y} = X \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} + \mathbf{e}$$
where
$$X = \begin{bmatrix} \mathbf{1} & \mathbf{x} \end{bmatrix}$$
is a $n \times 2$ matrix with columns $\mathbf{1}$ and $\mathbf{x}$ and $\mathbf{e}$ is the vector $(e_1, e_2, \ldots, e_n)$. We don't know what $\mathbf{e}$ is, but we want to make it as small as possible. To minimise the Euclidean norm of $\mathbf{e}$, we can use the *method of least squares.*

7.3 Least squares (review)

Hopefully you recall the method of least squares from earlier in the course. If not, we work through it again. The vector $(b_0, b_1)^T$ that we seek is the best approximation to $\mathbf{y}$ in the subspace spanned by the columns of X. We can find this vector by projecting $\mathbf{y}$ orthogonally onto the span of the columns of X. Let $\widehat{\mathbf{y}}$ be the orthogonal projection of $\mathbf{y}$ onto the span of the columns of $\mathbf{x}$. An orthonormal basis for the span of the columns of $\mathbf{x}$ is

$$\{\frac{\mathbf{x}-\overline{x}\mathbf{1}}{||\mathbf{x}-\overline{x}\mathbf{1}||}, \frac{\mathbf{1}}{||\mathbf{1}||}\}$$

and so

$$\widehat{\mathbf{y}} = \left(\frac{\mathbf{y}\cdot(\mathbf{x}-\overline{x}\mathbf{1})}{||\mathbf{x}-\overline{x}\mathbf{1}||}\right)\frac{(\mathbf{x}-\overline{x}\mathbf{1})}{||\mathbf{x}-\overline{x}\mathbf{1}||} + \left(\frac{\mathbf{y}\cdot\mathbf{1}}{||\mathbf{1}||}\right)\frac{\mathbf{1}}{||\mathbf{1}||}$$

Rearranging gives

$$\widehat{\mathbf{y}} = \frac{\mathbf{y}\cdot(\mathbf{x}-\overline{x}\mathbf{1})}{||\mathbf{x}-\overline{x}\mathbf{1}||^2}\mathbf{x} + \left(\frac{\mathbf{y}\cdot\mathbf{1}}{||\mathbf{1}||^2} - \overline{x}\frac{\mathbf{y}\cdot(\mathbf{x}-\overline{x}\mathbf{1})}{||\mathbf{x}-\overline{x}\mathbf{1}||^2}\right)\mathbf{1}$$

from which we can read off the values of b_0 and b_1 which give the least squares solution.

$$b_1 = \frac{\mathbf{y}\cdot(\mathbf{x}-\overline{x}\mathbf{1})}{||\mathbf{x}-\overline{x}\mathbf{1}||^2}$$

$$b_0 = \frac{\mathbf{y}\cdot\mathbf{1}}{||\mathbf{1}||^2} - \overline{x}\frac{\mathbf{y}\cdot(\mathbf{x}-\overline{x}\mathbf{1})}{||\mathbf{x}-\overline{x}\mathbf{1}||^2} = \overline{y} - b_1\overline{x}$$

Usually you should use software to calculate these, of course. For hand calculation, it is better to find b_1 first and then plug in to the formula for b_0 using $\overline{x}$ and $\overline{y}$.

7.4 Example

Let us work through this for the temperature and pressure data. We have

$$\mathbf{x} = (0, 10, 20, 30, 40, 50), \quad \mathbf{x}-\overline{x}\mathbf{1} = (-25, -15, -5, 5, 15.25)$$

$$\mathbf{y} = (91, 95, 100, 101, 107, 112), \quad \mathbf{y}\cdot(\mathbf{x}-\overline{x}\mathbf{1}) = -25(91)+-15(95)+-5(100)+5(101)+15(107)+25(112) = 710$$

$$b_1 = \frac{\mathbf{y}\cdot(\mathbf{x}-\overline{x}\mathbf{1})}{||\mathbf{x}-\overline{x}\mathbf{1}||^2} = \frac{710}{1750} \simeq 0.41$$

$$b_0 = \overline{y} - b_1\overline{x} = 101 - \frac{710}{1750}\times 25 \simeq 90.9$$

and so our least squares equation is

$$\widehat{\mathbf{y}} = 90.9 + 0.406\mathbf{x}.$$

In statistics, it is common to put a hat over a number which has been estimated from the data. So it would be more usual to write

$$\widehat{b}_0 = 90.9, \qquad \widehat{b}_1 = 0.406$$

Does our solution look sensible? We can check by tabulating the values of temperature, pressure, and predicted pressure using the equation. See Table 1. It certainly looks like a linear relationship is a good fit. If we were unaware of the ideal gas law, this could be a really useful thing to know. We can use our new equation to predict pressure from temperature, without having to do an experiment.

We can also add the line of best fit to Figure 1 to get Figure 3.

Temp (°C) x_i	Pressure (kPa) y_i	$\widehat{b}_0 + \widehat{b}_1 x_i$
0	91	90.9
10	95	94.9
20	100	99.0
30	101	103
40	107	107
50	112	111

Table 1: Temperature, pressure, and predicted pressure

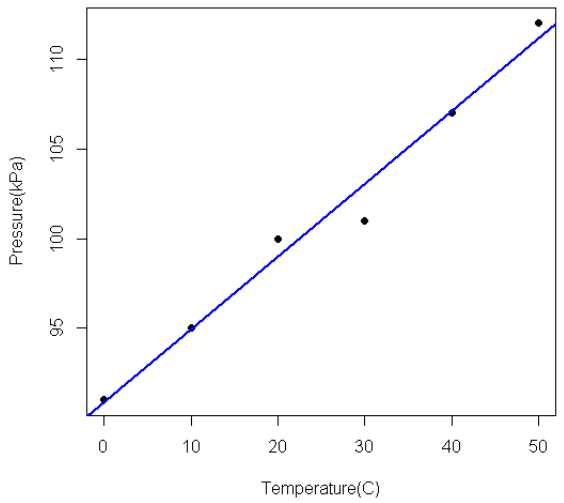

Figure 3: Running example: pressure versus temperature in a boiler.

7.5 The slope of the regression line

The slope of the regression line $\widehat{b}_1$ can be expressed in terms of the correlation. Since $\mathbf{1} \perp \mathbf{x} - \overline{x}\mathbf{1}$, we have

$$\widehat{b}_1 = \frac{\mathbf{y} \cdot (\mathbf{x} - \overline{x}\mathbf{1})}{||\mathbf{x} - \overline{x}\mathbf{1}||^2} = \frac{(\mathbf{y} - \overline{y}\mathbf{1}) \cdot (\mathbf{x} - \overline{x}\mathbf{1})}{||\mathbf{x} - \overline{x}\mathbf{1}||^2} = r_{xy}\frac{s_y}{s_x}$$

This neat result enables us to see that the slope of the regression line is just the correlation times the standard deviation of y divided by the standard deviation of x. It makes sense, because if you think about a plot of y versus x, when x gets more spread out, the slope should get shallower, but when y gets more spread out, the slope should get steeper. The slope is not defined if $s_x = 0$; in this case the regression line would be vertical.

7.6 What if we swap x and y?

Something strange immediately appears: if we swapped the roles of x and y, the slope would be different! But the regression line was supposed to describe the relationship between x and y. What is going on? The answer is that the regression line should be thought of as an equation for *predicting* y given x. It is the best one, in the sense that it minimises the sum of squared errors in the predictions *of* y. The line that minimises the sum of squared errors in the predictions of x from y is a different line in general.

7.7 Why least squares?

Again, we are faced with the question of why we should minimise the sum of *squares*. Again, the answer is really because it makes the mathematics easier. As G. Udny Yule said in 1897: "This is done solely for the convenience of analysis." He was probably the last statistics writer who was honest about this! One benefit, however, of using least squares is that it allows us to make a lot of statistical tests, if we make assumptions about how the true y values deviate from the predicted y values. There are alternatives to least squares for finding a regression line. One early writer, Boscovich, minimised the sum of absolute errors (the L^1 norm.) Galton, who coined the term *regression*, used a line which passed through the medians of the x and y–values. Shortly before the advent of computers, Tukey advocated a similar method for drawing lines of best fit by eye.

7.8 Goodness of fit

In our example, the regression line looked pretty good, in the sense of being close to the data points. We feel intuitively that it will not be so good if it is further away from the data points. How can we measure this? A natural way is to look at the vector of predicted values, the third column of Table 1 in our example, and compare it with the second column, the vector of actual values. We want these to coincide. They will coincide if and only if

$$||\widehat{\mathbf{y}} - \mathbf{y}|| = 0$$

but we don't want to take $||\widehat{\mathbf{y}} - \mathbf{y}||^2$ as a measure of goodness of fit because it increases if there are more data points and also increases if the y–values are more spread out. We can get rid of the problem of having too many dimensions by dividing by $n - 1$ and get rid of the problem of the spread of the y–values by dividing by s_y^2, so we define the measure of goodness of fit as

$$R^2 = 1 - \frac{||\widehat{\mathbf{y}} - \mathbf{y}||^2}{||\mathbf{y} - \overline{y}\mathbf{1}||^2}$$

instead. This gets bigger when the model fits better, and it always lies between 0 and 1. Indeed, because $\widehat{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ onto the plane spanned by $\mathbf{x}$ and $\mathbf{1}$, by translation we see that $\widehat{\mathbf{y}} - \overline{y}\mathbf{1}$ is the orthogonal projection of $\mathbf{y} - \overline{y}\mathbf{1}$ onto the plane spanned by $\mathbf{x}$ and $\mathbf{1}$. By the pythagorean theorem, we have

$$||\mathbf{y} - \overline{y}\mathbf{1}||^2 = ||\widehat{\mathbf{y}} - \overline{y}\mathbf{1}||^2 + ||\widehat{\mathbf{y}} - \mathbf{y}||^2.$$

Dividing by $||\widehat{\mathbf{y}} - \overline{y}\mathbf{1}||^2$ and rearranging gives

$$1 - \frac{||\widehat{\mathbf{y}} - \mathbf{y}||^2}{||\mathbf{y} - \overline{y}\mathbf{1}||^2} = \frac{||\widehat{\mathbf{y}} - \overline{y}\mathbf{1}||^2}{||\mathbf{y} - \overline{y}\mathbf{1}||^2}$$

which is certainly greater than or equal to zero and cannot exceed 1 because the projection of a vector is always shorter than the vector itself.

7.9 R-squared

The name might give us a clue about another way of writing R^2. First, notice that the mean of $\widehat{\mathbf{y}}$ is just $\overline{y}$. This is because the mean of $\widehat{b}_0 + \widehat{b}_1 x$ is $\widehat{b}_0 + \widehat{b}_1 \overline{x} = \overline{y}$. Therefore, R^2 is equal to $s^2_{\widehat{y}}/s^2_y$. But

$$s^2_{\widehat{y}} = s^2_{\widehat{b}_0 + \widehat{b}_1 x} = \widehat{b}_1^2 s_x^2 = r^2_{xy} \frac{s_y^2}{s_x^2} s_x^2$$

and so

$$R^2 = r^2_{xy}$$

which is just the square of the correlation coefficient. This immediately leads to the question: why did we define R^2 in a complicated way and why is there a capital R? The reason is because when we have more than one x–variable next week, there will be no single correlation coefficient, but the definition $R^2 = 1 - \frac{||\widehat{\mathbf{y}} - \mathbf{y}||^2}{||\mathbf{y} - \overline{y}\mathbf{1}||^2} = \frac{||\widehat{\mathbf{y}} - \overline{y}\mathbf{1}||^2}{||\mathbf{y} - \overline{y}\mathbf{1}||^2}$ will still make sense and the geometric argument which we made will still be correct.

7.10 Regression in statistics

So far, we have used regression as a way of finding an equation to predict y given x, taking into account errors in measuring y. Statisticians understand regression in quite a different way. It is important to understand the statistical way, otherwise you will not be able to understand the output of software which performs regression.

7.11

In statistics, it is assumed that there is a large population of all possible pairs of (x, y) values and the data (x_i, y_i) is a random sample from this population of size n. In the large population, the equation

$$y = b_0 + b_1 x + e$$

holds, where e is a random error term which is different for every (x, y) pair. Using least squares, an equation of the form

$$\widehat{y}_i = \widehat{b}_0 + \widehat{b}_1 x_i$$

is obtained from the data. It is desired to learn about the *true* b_0 and b_1 using the fitted $\widehat{b}_0$ and $\widehat{b}_1$.

7.12 Assumptions

To make any progress, some assumptions have to be made about the errors e. It is assumed that the errors are independent and follow a normal distribution with mean 0 and constant variance σ^2. Hopefully you have seen the normal distribution in earlier studies. If this assumption about the errors is true, then it is possible to make statements about how much $\widehat{b}_1$ deviates from b_1 by constructing a *confidence interval*.

7.13 Confidence intervals

A $100(1-\alpha)\%$ *confidence interval* for b_1 is an interval constructed from a sample of values of (x, y) in such a way that, if we took many samples and constructed an interval for each one, the proportion of such intervals which contained the true b_1 would be $(1-\alpha)$. The question which a confidence interval answers is: "How certain are we that y really depends on x, given that it either depends on x linearly or not at all, and that the assumptions of Section 7.12 are true?". If the confidence interval does not contain 0, then it is plausible that $b_1 \neq 0$, otherwise not. The idea is illustrated in Figure 4.

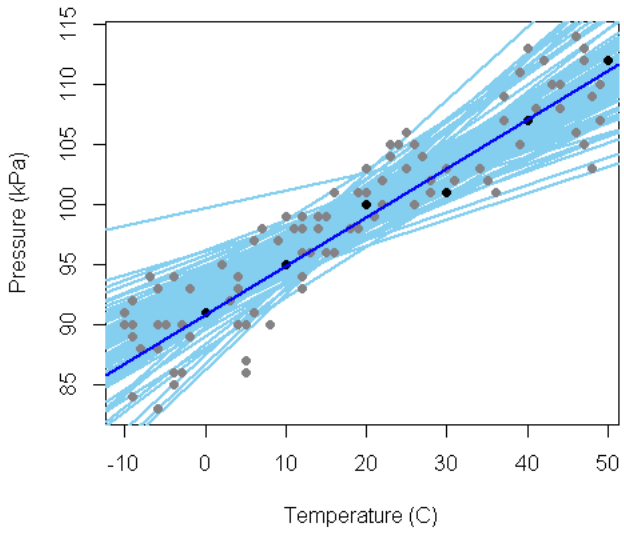

Figure 4: Every time we take a sample from the population, we get a different regression line. We can use a confidence interval to describe the range of slopes which these lines are likely to have.

7.14

It can be shown using statistics that a $100(1-\alpha)\%$ confidence interval for b_1 is

$$\widehat{b}_1 \pm t^*_{n-2}(1-\alpha/2)\mathrm{SE}(\widehat{b}_1)$$

where

$$\mathrm{SE}(\widehat{b}_1) = \sqrt{\frac{1}{n-2}\frac{||\widehat{\mathbf{y}}-\mathbf{y}||^2}{||\mathbf{x}-\overline{x}\mathbf{1}||^2}}$$

and $t^*_{n-2}(1-\alpha/2)$ is the number such that the area under the density function for the t–distribution with $n-2$ degrees of freedom and above the point $t^*_{n-2}(1-\alpha/2)$ is $1-\alpha/2$. (The t–distribution looks like the normal distribution but has fatter tails. You can get this number from Matlab by typing `tinv(1-alpha/2, n-2)`.) For example, if $\alpha = 0.95$ and $n = 8$ then you should get $t^*_{n-2}(1-\alpha/2) = 2.4469$. You can also get these numbers from look-up tables.

7.15 Prediction intervals

Before working through an example, there is another application of the t–distribution to *prediction intervals.* If x^* is a value of x, which may or may not be one of the x_i in the data, and (x^*, y^*) is a point in the population (where

y^* is some value which is unknown) then the predicted value of y^* is $\widehat{y}^* = \widehat{b}_0 + \widehat{b}_1 x^*$. A $100(1-\alpha)\%$ prediction interval for y^* is an interval such that, if the regression assumptions are true, then if we took many samples of the (x, y) and built a model and constructed a prediction interval from each one, a proportion $(1-\alpha)$ of these intervals would contain the true y^*.

7.16

Given x^*, a $100(1-\alpha)\%$ prediction interval for y^* is

$$\widehat{y}^* \pm t^*_{n-2}(1-\alpha/2)\frac{||\widehat{\mathbf{y}} - \mathbf{y}||}{\sqrt{n-2}}\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{||\mathbf{x} - \overline{x}\mathbf{1}||^2}}$$

Notice that the prediction interval becomes wider as x^* is further from $\overline{x}$. In predictive modelling in general, you should be wary about making predictions outside the range of x–values for which you have data. This is called *extrapolation*. A good example is Hooke's Law from physics (force exerted by a spring is proportional to extension) which is linear up to the elastic limit, but then breaks down.

7.17 Examples

For the temperature and pressure example again, we have $n = 6$ and

$$\mathbf{y} - \widehat{\mathbf{y}} = (91 - 90.9, 95 - 95, 100 - 99.0, 101 - 103, 107 - 107, 112 - 111) = (0.1, 0, 1, -2, 0, 1)$$

These numbers (called *errors* or *residuals*) should add up to zero. The reason why they don't in this case is rounding error. We then have

$$||\mathbf{y} - \widehat{\mathbf{y}}||^2 = 0.1^2 + 0^2 + 1^2 + 2^2 + 0^2 + 1^2 = 6.01.$$

We also have $||\mathbf{x} - \overline{x}\mathbf{1}||^2 = 1750$ and so

$$\mathrm{SE}(\widehat{b}_1) = \sqrt{\frac{1}{4}\frac{6.01}{1750}} = 0.029.$$

From software we find that $t_4(0.975) \simeq 2.78$. A 95% confidence interval for b_1 is therefore

$$\widehat{b}_1 \pm 2.78 \times 0.029 = 0.406 \pm 0.08 = (0.33, 0.49).$$

The interpretation is that we are very confident that the values of pressure change when the temperature changes; the variation of pressure with temperature which we observed in the data is probably not just a random accident. This does not mean that changing the temperature *causes* the pressure to change. They might both be caused by some common third thing.

7.18

Suppose we want to predict the pressure if the temperature is $x^* = 24°\mathrm{C}$. Then a 95% prediction interval is

$$(91 + 0.41 \times 24) \pm 2.78\sqrt{\frac{6.01}{4}}\sqrt{1 + \frac{1}{6} + (24 - 25)^2/1750} = 100.84 \pm 2.78 \times 1.249 \times 1.08 = 100.84 \pm 3.75$$

or

$$(97°\mathrm{C}, 105°\mathrm{C}).$$

The interval is pretty wide, which reflects the fact that we have a very small amount of data to go on.

8 Exercises for week 2

1. What is the slope of the regression line if we regress $\frac{y-\overline{y}}{s_y}$ on $\frac{x-\overline{x}}{s_x}$?

2. If $\mathbf{a} \perp \mathbf{v} - \mathbf{w}$ and
$$||\mathbf{w}||^2 + ||\mathbf{v} - \mathbf{w}||^2 = ||\mathbf{v}||^2$$
show that
$$||\mathbf{w} - \mathbf{a}||^2 + ||\mathbf{v} - \mathbf{w}||^2 = ||\mathbf{v} - \mathbf{a}||^2.$$
Convince yourself that this is true geometrically in $\mathbb{R}^2$ and $\mathbb{R}^3$.

3. Define an interval for a regression slope b_1 as follows. Generate a random number u between 0 and 1. If $u < 0.95$, take the interval $(-\infty, \infty)$. Otherwise take the empty interval. Explain why this is a 95% confidence interval for b_1.

4. Suppose that when we regress $y =$ height (in cm) on $x =$ foot length (in cm) for a group of people, the slope is 0.9. What is the slope if we choose to re-express foot length in mm?

5. The lengths in cm and venom strength in ppm for six spiders of a particular species are given by

length	2	2.5	2.6	4.8	5.0	5.1
venom	10	9.6	9.0	12	11	10.7

   (a) What is the R^2 if we regress venom on length?

   (b) Does venom tend to increase with length, or decrease with length?

   (c) Make a plot of the data. What do you notice?

   (d) For this species of spider, females tend to be about twice the size of males. Does venom tend to increase or decrease with length for male spiders? For female spiders?

6. Calculate $X^T X$ where X is the matrix with columns $\mathbf{1}$ and $\mathbf{x}$ where $\mathbf{1}, \mathbf{x} \in \mathbb{R}^n$.

7. Let $\mathbf{x} = (0, 1, 1)$, $\mathbf{y} = (0, -1, 1)$.

   (a) Find the least squares solution to $\mathbf{y} = b_0 + b_1\mathbf{x}$.

   (b) Show that there is more than one choice of (b_0, b_1) which minimises
$$\sum_{i=1}^{3} |y_i - (b_0 + b_1 x_i)|.$$
   (Hint: drawing a picture will probably help. In fact, there are infinitely many solutions. The uniqueness of the least squares solution is one reason for preferring least squares to other approaches.)

8. Suppose we want to find a and b in a relationship of the form $y = ax^b$, where x and y are positive quantities which we have measured, subject to some error. Since $\log(y) = \log(a) + b\log(x)$, it is common to plot $\log(y)$ versus $\log(x)$ and then look for a line of best fit. We could use least squares to do this, and get a confidence interval for b and prediction intervals for new values of x. Explain why these intervals will probably not be correct.

9 Multiple Regression

9.1

Multiple regression is also just called “linear regression". It is the generalisation of simple linear regression to the situation in which there are several x–variables and the model becomes

$$\mathbf{y} = b_0 + b_1\mathbf{x_1} + b_2\mathbf{x_2} + \cdots + b_p\mathbf{x_p} + \mathbf{e}.$$

Here, p is the number of *predictors* $\mathbf{x_i}$. Instead of a line, we seek a hyperplane which is as close as possible to containing the points $(x_1, x_2, \ldots, x_p, y)$. In the case $p = 2$, this can be visualised as a plane in $\mathbb{R}^3$.

9.2 The Normal Equations

As in the case where $p = 1$, we seek the least squares solution to the system

$$\mathbf{y} = X\mathbf{b}$$

where

$$X = \begin{bmatrix}\mathbf{1} & \mathbf{x_1} & \cdots & \mathbf{x_p}\end{bmatrix}$$

is a $n \times (p + 1)$ matrix called the *model matrix*. The least squares solution $\mathbf{b}$ satisfies

$$\widehat{\mathbf{y}} = X\mathbf{b}$$

where $\widehat{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ onto the column space of X. Since $\widehat{y}$ must satisfy

$$\widehat{\mathbf{y}} \cdot \mathbf{x}_i = \mathbf{y} \cdot \mathbf{x}_i$$

for all i, and

$$\widehat{\mathbf{y}} \cdot \mathbf{1} = \mathbf{y} \cdot \mathbf{1}$$

we must have

$$X^T\mathbf{y} = X^T\widehat{\mathbf{y}} = X^TX\mathbf{b}$$

and from this it follows that if X^TX is invertible, then

$$\mathbf{b} = (X^TX)^{-1}X^T\mathbf{y}$$

is the solution. In real life, $\mathbf{b}$ is never computed by finding the inverse of X^TX but it can be computed in more numerically stable ways, for example by using the QR factorisation. In statistics, the equation

$$X^T\widehat{\mathbf{y}} = X^TX\mathbf{b}$$

is called the *normal equations*.

9.3 R^2

The measure of goodness-of-fit R^2 is defined in the same way as for simple linear regression, but it can no longer be interpreted as the square of the correlation coefficient. It is a popular way of describing the quality of the fit, but is not necessarily useful.

9.4 Confidence intervals for b_i

A $100(1-\alpha)\%$ confidence interval for b_i is given by

$$\widehat{b}_i \pm t^*_{n-p-1}(1-\alpha/2)\mathrm{SE}(\widehat{b}_i)$$

where

$$\mathrm{SE}(\widehat{b}_i) = \sqrt{\frac{1}{n-p-1}||\widehat{\mathbf{y}}-\mathbf{y}||^2(X^TX)^{-1}_{ii}}$$

Here, $(X^TX)^{-1}_{ii}$ denotes the i^{th} diagonal entry of $(X^TX)^{-1}$. This is quite similar to the formula in the $p=1$ case except for the $n-p-1$ where we had $n-2$ before. Just like in the $p=1$ case, the confidence interval is only valid if the assumptions that the regression errors and independent, normally distributed and have constant variance are met.

9.5 *Ceteris Paribus*

The interpretation of the confidence interval is similar to the case of simple linear regression. We look for whether there is evidence that $b_i \neq 0$ and interpret this as indicating that knowing x_i tells us something about y when all other x_j are kept equal. This "all other things being equal" (often called "ceteris paribus") is very important and is often overlooked. For example, suppose y is the salary of a cricketer, x_1 is the number of years they have been playing and x_2 is the number of runs they have scored in their career. If

$$y = b_0 + b_1x_1 + b_2x_2$$

we expect y to increase as x_1 increases and also increase as x_2 increases. However, we would not expect b_1 and b_2 to both be positive. Why? If x_2 is fixed, then *among cricketers with the same number of runs*, we would expect those who have been playing longer to have lower salaries, because they took longer to score their runs. So we would actually expect $b_1 < 0, b_2 > 0$.

9.6 Pitfalls in multiple regression

The "all other things being equal" problem is not the only thing which can go wrong when interpreting the results of a multiple regression. Here are some other common ones.

- **Non-linearity.** In the $p = 1$ case, you can always plot the data and decide whether it looks curved. For $p > 2$, this is impossible, which makes it more difficult to tell whether a linear relationship fits your data well or not. This leads to the careless over-use of linear regression, both in hard and soft sciences.
- **Specification bias.** This is a name for the problem of an important variable being left out of your model. For example, a story is told by Mosteller and Tukey of a regression done to try to find out the factors which determined the success of bombing in World War II. It was found that the number of enemy fighters had a positive coefficient, i.e. the more fighters encountered, the better the performance of the bombers. This probably happened because there were more fighters when there were fewer clouds, which also made bombing more accurate. But cloud cover was not included in the model.
- **Kitchen-sink regression.** In the social sciences, it is common to throw as many variables as possible into a regression model to "control" for their effects. When irrelevant variables are included, this leads to problems in interpreting the regression coefficients and can make predictions less accurate. For example, an American data mining company tried to predict the success of teams in the 2012 Olympics by using a regression involving dozens of variables including religion and export volume. They failed completely.

On the other hand, regression can work well if done correctly. A famous example is the analysis of wine prices by the economist Orley Ashenfelter, who was able to correctly predict the quality of 1989 Bordeaux wines, even though his model was regarded with scorn by wine experts. Usually, however, linear regression is not the best approach for prediction, although it has the advantage that it is simple and the prediction equation which it produces is understandable and can be easily computed.

9.7 Instability

You might be wondering what happens when $X^T X$ is singular. Then it cannot be inverted. It is actually very rare that $X^T X$ is exactly singular, but it can be ill-conditioned and this causes various problems, sometimes known as *instability*. It turns out that $X^T X$ is singular exactly when the columns of X are linearly dependent (called multicollinearity). When the columns of X are close to being linearly dependent, the regression is unstable.

9.8

Instability is important because it can affect the values of the regression coefficients and make the results difficult to interpret. For example, if there is approximate linear dependence, say $\mathbf{x}_2 \simeq 2\mathbf{x}_1$, then we might have

$$\mathbf{y} = b_0 + b_1\mathbf{x}_1 + b_2\mathbf{x}_2 \simeq b_0 + (b_1 - 3)\mathbf{x}_1 + (b_2 + 1.5)\mathbf{x}_2 \simeq \cdots$$

and the regression coefficients are likely to be numerically unstable. They might have huge standard errors or take unreasonable values, depending on the software used. Different fields have different ways of assessing multicollinearity. A simple one, used in econometrics, is the variance inflation factor or VIF. The VIF of a variable x_i in a regression is

$$VIF(x_i) = \frac{1}{1 - R^2_{i|j}}$$

where $R^2_{i|j}$ is the R^2 from a regression of x_i on $\{x_j : j \neq i\}$. A large value of VIF is evidence of multicollinearity (why?) Alternative measures of multicollinearity are based on the determinant of $X^T X$; when $\det(X^T X)$ is close to zero, $X^T X$ is close to being singular. There is no single measure of multicollinearity which is in universal use.

10 Example

Let us do a small example in Matlab. (The calculations for multiple regression become tedious if done by hand.) The data consist of femur lengths x_1 in m, brain case volume x_2 in ml, and estimated spine length y in m, for seven stegosaurus skeletons in more-or-less complete condition. We are interested in how spine length varies with femur length and brain case volume. First we input the data:

```
x1 = [1.03, 0.99, 1.08, 1.2, 0.94, 0.97, 1.01]
x2 = [15, 15.1, 14, 13.2, 10.9, 14.1, 14.5]
y  = [9.2, 9.5, 8.4, 12, 9.3, 7.7, 10]
```

Now plot the data:

```
scatter3(x1, x2, y, 'black', 'fill')
```

and now turn on 3d rotation so you can drag the plot with the mouse:

```
rotate3d()
```

Create the model matrix `X`

```
X = [ones(7,1) x1' x2']
```

Find the least squares solution to the normal equations

```
b = linsolve(X' * X, X' * y')
```

Our regression equation is

$$y = 0.81 + 11.2x_1 - 0.21x_2.$$

We can add the plane to the plot.

```
[XX,YY] = meshgrid(0.8:0.01:1.2, 7:0.1:16);
ZZ = b(1)+b(2).*XX + b(3).*YY;
hold on
s = surf(XX,YY,ZZ)
set(s, 'facecolor', 'none');
```

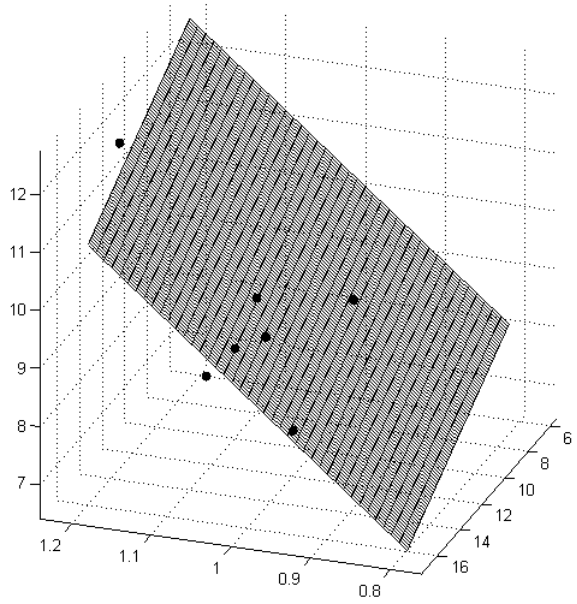

Figure 5: Least squares plane in Matlab

Let us predict the length of a stegosaurus with a 10 ml brain and a femur length of 0.8 m.

```
[1 0.8 15] * b
```

The answer is 6.6 m. We haven't covered prediction intervals for multiple regression, but just like in the one-variable case, with so few data points the prediction interval for this observation is pretty wide. We would, however, like to use our formula to get confidence intervals for the regression coefficients. The residual sum of squares $||\widehat{\mathbf{y}} - \mathbf{y}||^2$ is given by

```
RSS = sum((X*b - y').^2)
```

which is 5.26. We also need the diagonal entries of $(X^T X)^{-1}$ which we get from

```
xx = diag(inv(X' * X))
```

We compute a 95% confidence interval for b_1, which is `b(2)` in the Matlab output because Matlab starts indexing vectors at 1, so our (b_0, b_2, b_2) is Matlab's `[b(0) b(1) b(2)]`.

```
b(2) + tinv(0.975, 7-2) * [-1, 1] .* sqrt(1/(7-2-1) * RSS * xx(2))
```

The 95% confidence interval for b_1 is

$$(-2.78, 25.16)$$

which is very wide. Similarly, we use `xx(3)` to get a confidence interval for b_2 of

$$(-1, 0.63)$$

at the 95% level. Notice that taking 95% as a confidence level is quite arbitrary. This leads to the question: which confidence level would lead us to believe that y *did* vary with x_1 or x_2? This is the confidence level at which 0 lies just outside the confidence interval. Twice the corresponding α is called the *p–value*. We can calculate it as follows:

```
2 * (1 - tcdf(b(2)/sqrt(1/(7-2-1) * RSS * xx(2)), 7-2-1))
```

and get 0.108. The interpretation of this is that if the points were really random scatter and the regression assumptions were true, the probability of seeing a value of b_1 at least as extreme as observed is a little more than 1 in 10, so not very unlikely. Similarly, you can do the p–value for b_2 and should get 0.555; above 50%. This is useful information; if anything is a significant predictor of y, it is femur length and not brain case size.

But these p–values and confidence intervals are misleading if the regression is unstable, so we should check that by finding the R^2 if we regress x_1 on x_2. This is just the square of the correlation.

```
corr([x1' x2']).^2
```

The $(1, 2)$–entry of this matrix of correlations is the correlation between x_1 and x_2. It is 0.01, corresponding to a VIF of just above 1. This means that we do not have to worry about instability. Nevertheless, the regression is useless because we have no compelling evidence of a pattern in the data, as opposed to random scatter.

Note that when calculating confidence intervals and p–values we should also check the regression assumptions of independent, normally-distributed errors with constant variance. Usually this is done by making plots of the *residuals* $y_i - \widehat{y}_i$. Again, with only seven data points, this is rather pointless as there is almost no way we could discover evidence of non-normality, even if it was present.

11 Exercises for week 3

1. If we replace x_i by ax_i in a multiple regression, how does $\widehat{b}_i$ change? How does $\mathrm{SE}(\widehat{b}_i)$ change?

2. A genetics researcher has data from 20 patients in which 500 genes are present or absent and wants to regress height on the variables $x_1, x_2, \ldots, x_{500}$ where $x_i = 1$ if gene i is present and $x_i = 0$ if gene i is absent. Can this be done using linear regression? Why or why not?

3. Revisiting the spider data, we add a dummy variable which is 0 for male spiders and 1 for females.

length	2	2.5	2.6	4.8	5.0	5.1
venom	10	9.6	9.0	12	11	10.7
sex	0	0	0	1	1	1

   (a) Work out the regression coefficient for the length variable when $y =$ venom is regressed on $x_1 =$ length and $x_2 =$ sex. Does venom tend to increase or decrease with length?

   (b) Find the VIF for the length variable. (Hint: you only need to calculate the correlation between length and sex for this.)

4. For $p = 1$, check that the formula for $\mathrm{SE}(\widehat{b}_1)$ reduces to the formula from simple linear regression.

5. Let A be a subspace of $\mathbb{R}^n$, let $\mathbf{a} \in A$ and let $\mathbf{y} \in \mathbb{R}^n$. Let $\widehat{\mathbf{y}}$ be the orthogonal projection of $\mathbf{y}$ onto A. Explain why $\mathbf{y} \cdot \mathbf{a} = \widehat{\mathbf{y}} \cdot \mathbf{a}$. Hint: write $\mathbf{y} = \widehat{\mathbf{y}} + \mathbf{z}$ with $\mathbf{z} \in A^{\perp}$.

6. An economist wants to compute a regression model for predicting a country's exports using the variables $x_1 =$ 2006 GDP, $x_2 =$ 2007 GDP and $x_3 =$ 2008 GDP. What problem is this likely to run into? (Note: the problem can be overcome using modified versions of linear regression.)

7. Let y be the number of weightlifting medals won by a country in the olympics. If you regress y on a set of variables which includes the maximum weight lifted by an athlete from that country (in lb) and the maximum weight lifted by an athlete from that country (in kg), what will happen? Would your answer be different if you used software to do the regression?

8. A linear regression is used to assess 10 different mixtures of concrete. The amounts of water (kg), gypsum (kg) and ash (kg) are the predictors and the breaking strength (N) is the y–variable. A 95% confidence interval for the coefficient of gypsum is $(-0.2, 4.8)$. Carefully explain *exactly* what this interval means.

9. Sometimes it is not desirable to have an intercept in a regression.

   (a) Let $\mathbf{x}$ and $\mathbf{y}$ be vectors of data of length n. Find the value of b which minimises

   $$\sum_{i=1}^{n} (y_i - bx_i)^2.$$

   (b) Suppose we have several x–variables, x_1, …, x_p. Write down a matrix equation for the value of $\mathbf{b} = (b_1, \ldots, b_p)$ which minimises

   $$\sum_{i=1}^{n} (y_i - b_1 x_{1i} - b_2 x_{2i} - \cdots - b_p x_{pi})^2.$$