Regression

(Module 5)

Statistics (MAST20005) & Elements of Statistics (MAST90058)

Semester 2, 2022

Contents

Aims of this module

- Introduce the concept of **regression**
- Show a simple model for studying the relationship between two variables
- Discuss correlation and how it relates to regression

1 Introduction

Relationships between two variables

We have studied how to do estimation for some simple scenarios:

- iid samples from a single distribution (X_i)
- comparing iid samples from two different distributions (X_i & Y_j)
- differences between paired measurements ($X_i - Y_i$)

We now consider how to analyse bivariate data more generally, i.e. two variables, X and Y, measured at the same time, i.e. as a pair.

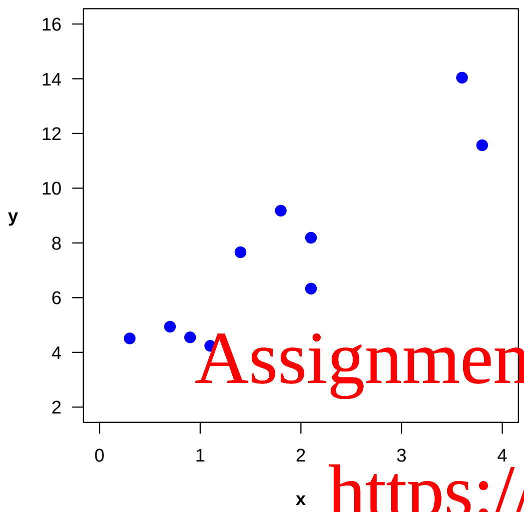
The data consist of pairs of data points, (x_i, y_i).

These can be visualised using a *scatter plot*.

Example data

x_i	y_i
1.80	9.18
1.40	7.66
2.10	6.33
0.30	4.51
3.60	14.04
0.70	4.94
1.10	4.24
2.10	8.19
0.90	4.55
3.80	11.57

$n = 10$



2 Regression

Regression

Often interested in how Y depends on X. For example, we might want to use X to predict Y.

In such a setting, we will assume that the X values are known and fixed (henceforth, x instead of X), and look at how Y varies given x.

Example: Y is a student's final mark for Statistics, and x is their mark for the prerequisite subject Probability. Does x help to predict Y?

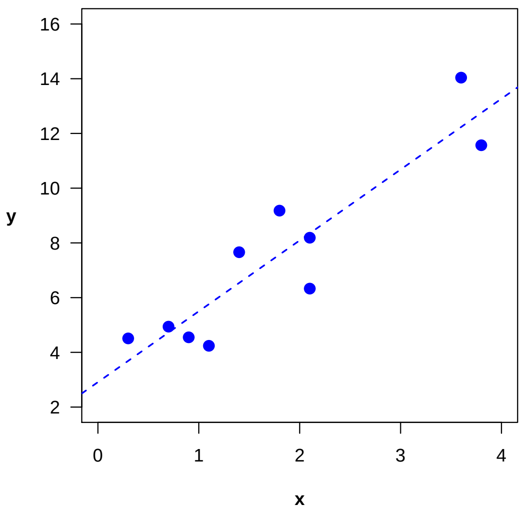
The *regression* of Y on x is the conditional mean, $\mathbb{E}(Y \mid x) = \mu(x)$.

The regression can take any form. We consider *simple linear regression*, which has the form of a **straight line**:

$$\mathbb{E}(Y \mid x) = \alpha + \beta x \quad \text{and} \quad \operatorname{var}(Y \mid x) = \sigma^2.$$

Example: simple linear regression model

$$\begin{aligned} \mathbb{E}(Y \mid x) &= \alpha + \beta x \\ \operatorname{var}(Y \mid x) &= \sigma^2 \end{aligned}$$



Terminology

- Y is called a *response* variable. Can also be called an *outcome* or *target* variable. Please do **not** call it the 'dependent' variable.
- x is called a *predictor* variable. Can also be called an *explanatory* variable. Please do **not** call it an 'independent' variable.
- $\mu(x)$ is called the *(linear) predictor function* or sometimes the *regression curve* or the *model equation*.
- The parameters in the predictor function are called *regression coefficients*.

Why 'regression'?

It is strange terminology, but it has stuck.

Refers to the idea of *'regression to the mean'*: if a variable is extreme on its first measurement, it will tend to be closer to the average on its second measurement, and vice versa.

First described by Sir Francis Galton when studying the inheritance of height between fathers and sons. In doing so, he invented the technique of simple linear regression.

Linearity

A regression model is called *linear* if it is linear in the coefficients.

It doesn't have to define a straight line!

Complex and non-linear functions of x are allowed, as long as the resulting predictor function is a linear combination (i.e. an additive function) of them, with the coefficients 'out the front'.

For example, the following are linear models:

$$\mu(x) = \alpha + \beta x + \gamma x^2$$
$$\mu(x) = \frac{\alpha}{x} + \frac{\beta}{x^2}$$
$$\mu(x) = \alpha \sin x + \beta \log x$$

The following are NOT linear models:

$$\mu(x) = \alpha \sin(\beta x)$$
$$\mu(x) = \frac{\alpha}{1 + \beta x}$$
$$\mu(x) = \alpha x^{\beta}$$

…but the last one can be re-expressed as a linear model on a log scale (by taking logs of both sides),

$$\mu^*(x) = \alpha^* + \beta \log x$$

3 Simple linear regression

Estimation goals

Back to our simple linear regression model:

$$\mathbb{E}(Y \mid x) = \alpha + \beta x \quad \text{and} \quad \operatorname{var}(Y \mid x) = \sigma^2.$$

- We wish to estimate the slope (β), the intercept (α), the variance of the errors (σ^2), their standard errors and construct confidence intervals for these quantities.
- Often want to use the fitted model to make predictions about future observations (i.e. predict Y for a new x).
- Note: the Y_i are not iid. They are independent but have different means, since they depend on x_i.
- We have not (yet) assumed any specific distribution for Y, only a conditional mean and variance.

Reparameterisation

Changing our model slightly…

Let $\alpha_0 = \alpha + \beta\bar{x}$, which gives,

$$\begin{aligned}\mathbb{E}(Y \mid x) &= \alpha + \beta x \\ &= \alpha_0 + \beta(x - \bar{x})\end{aligned}$$

Now our model is in terms of α_0 and β.

This will make calculations and proofs simpler.

3.1 Point estimation of the mean

Least squares estimation

Choose α_0 and β to minimize the sum of squared deviations:

$$H(\alpha_0, \beta) = \sum_{i=1}^{n} \left(y_i - \alpha_0 - \beta\left(x_i - \bar{x}\right)\right)^2$$

Solve this by finding the partial derivatives and setting to zero:

$$0 = \frac{\partial H(\alpha_0, \beta)}{\partial \alpha_0} = 2\sum_{i=1}^{n}[y_i - \alpha_0 - \beta(x_i - \bar{x})](-1)$$
$$0 = \frac{\partial H(\alpha_0, \beta)}{\partial \beta} = 2\sum_{i=1}^{n}[y_i - \alpha_0 - \beta(x_i - \bar{x})](-(x_i - \bar{x}))$$

These are called the *normal equations*.

Least squares estimators

Some algebra yields the *least square estimators*,

$$\hat{\alpha}_0 = \bar{Y}, \quad \hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Another expression for $\hat{\beta}$ is:

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

These are equivalent, due to the following result:

$$\sum (x_i - \bar{x})(Y_i - \bar{Y}) = \sum (x_i - \bar{x}) Y_i.$$

Can also then get an estimator for α:

$$\begin{aligned} \hat{\alpha} &= \hat{\alpha}_0 - \hat{\beta}\bar{x} \\ &= \bar{Y} - \hat{\beta}\bar{x}. \end{aligned}$$

And also an estimator for the predictor function,

$$\begin{aligned} \hat{\mu}(x) &= \hat{\alpha} + \hat{\beta} x \\ &= \hat{\alpha}_0 + \hat{\beta}(x - \bar{x}) \\ &= \bar{Y} + \hat{\beta}(x - \bar{x}). \end{aligned}$$

Ordinary least squares

This method is sometimes called *ordinary least squares* or *OLS*.

Other variants of least squares estimation exist, with different names. For example, 'weighted least squares'.

Example: least squares estimates

For our data:

$$\begin{aligned} \bar{x} &= 1.78 \\ \bar{y} &= 7.52 = \hat{\alpha}_0 \\ \hat{\alpha} &= 2.91 \\ \hat{\beta} &= 2.59 \end{aligned}$$

The fitted model equation is then:

$$\hat{\mu}(x) = 2.91 + 2.59x$$

```
> rbind(y, x)
  [,1] [,2] [,3] [,4]  [,5] [,6] [,7] [,8] [,9] [,10]
y 9.18 7.66 6.33 4.51 14.04 4.94 4.24 8.19 4.55 11.57
x 1.80 1.40 2.10 0.30  3.60 0.70 1.10 2.10 0.90  3.80

> model1 <- lm(y ~ x)
> model1

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
      2.911        2.590
```

Properties of these estimators

What do we know about these estimators?

They are all linear combinations of the Y_i,

$$\hat{\alpha}_0 = \sum_{i=1}^{n} \left(\frac{1}{n}\right) Y_i$$
$$\hat{\beta} = \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{K}\right) Y_i$$

where $K = \sum_{i=1}^{n} (x_i - \bar{x})^2$.

This allows us to easily calculate means and variances.

Means?

$$\mathbb{E}(\hat{\alpha}_0) = \mathbb{E}(\bar{Y}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(Y_i) = \frac{1}{n}\sum_{i=1}^{n} [\alpha_0 + \beta(x_i - \bar{x})] = \alpha_0$$

$$\begin{aligned}\mathbb{E}(\hat{\beta}) &= \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{K}\, \mathbb{E}(Y_i) = \frac{1}{K}\sum_{i=1}^{n} (x_i - \bar{x})(\alpha_0 + (x_i - \bar{x})\beta) \\ &= \frac{1}{K}\sum_{i=1}^{n} (x_i - \bar{x})\alpha_0 + \frac{K}{K}\beta = \beta\end{aligned}$$

This also implies, $\mathbb{E}(\hat{\alpha}) = \alpha$ and $\mathbb{E}(\hat{\mu}(x)) = \mu(x)$, and so we have that all of the estimators are unbiased.

Variances?

$$\mathrm{var}(\hat{\alpha}_0) = \mathrm{var}(\bar{Y}) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}(Y_i) = \frac{\sigma^2}{n}$$

$$\begin{aligned}\mathrm{var}(\hat{\beta}) &= \mathrm{var}\left(\sum_{i=1}^{n} \frac{(x_i - \bar{x})}{K} Y_i\right) = \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{K}\right)^2 \mathrm{var}(Y_i) \\ &= \frac{1}{K^2}\sum_{i=1}^{n} (x_i - \bar{x})^2 \mathrm{var}(Y_i) = \frac{1}{K^2}\sum_{i=1}^{n} (x_i - \bar{x})^2 \sigma^2 \\ &= \frac{1}{K^2}\sigma^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{K^2}\sigma^2 K \\ &= \frac{\sigma^2}{K}\end{aligned}$$

Similarly,

$$\begin{aligned}\mathrm{var}(\hat{\alpha}) &= \left(\frac{1}{n} + \frac{\bar{x}^2}{K}\right)\sigma^2 \\ \mathrm{cov}(\hat{\alpha}_0, \hat{\beta}) &= 0 \\ \mathrm{var}(\hat{\mu}(x)) &= \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{K}\right)\sigma^2\end{aligned}$$

Can we get their standard errors?

We need an estimate of σ^2.

3.2 Interlude: Analysis of variance

Analysis of variance: iid model

For $X_i \sim \mathrm{N}(\mu, \sigma^2)$ iid,

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2$$

Analysis of variance: regression model

$$\begin{aligned}
&\sum_{i=1}^{n}(Y_i - \alpha_0 - \beta(x_i - \bar{x}))^2 \\
&\quad = \sum_{i=1}^{n}(Y_i - \hat{\alpha}_0 - \hat{\beta}(x_i - \bar{x}) + \hat{\alpha}_0 + \hat{\beta}(x_i - \bar{x}) - \alpha_0 - \beta(x_i - \bar{x}))^2 \\
&\quad = \sum_{i=1}^{n}(Y_i - \hat{\alpha}_0 - \hat{\beta}(x_i - \bar{x}) + (\hat{\alpha}_0 - \alpha_0) + (\hat{\beta} - \beta)(x_i - \bar{x}))^2 \\
&\quad = \sum_{i=1}^{n}(Y_i - \hat{\alpha}_0 - \hat{\beta}(x_i - \bar{x}))^2 + n(\hat{\alpha}_0 - \alpha_0)^2 + K(\hat{\beta} - \beta)^2
\end{aligned}$$

Note that the cross-terms disappear. Let's see...

The cross-terms...

$$t_1 = 2\sum_{i=1}^{n}(Y_i - \hat{\alpha}_0 - \hat{\beta}(x_i - \bar{x}))(\hat{\alpha}_0 - \alpha_0)$$

$$t_2 = 2\sum_{i=1}^{n}(Y_i - \hat{\alpha}_0 - \hat{\beta}(x_i - \bar{x}))(\hat{\beta} - \beta)(x_i - \bar{x})$$

$$t_3 = 2\sum_{i=1}^{n}(x_i - \bar{x})(\hat{\beta} - \beta)(\hat{\alpha}_0 - \alpha_0)$$

Since $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ and $\sum_{i=1}^{n}(Y_i - \hat{\alpha}_0) = \sum_{i=1}^{n}(Y_i - \bar{Y}) = 0$, the first and third cross-terms are easily shown to be zero.

For the second term,

$$\begin{aligned}
\frac{t_2}{2(\hat{\beta} - \beta)} &= \sum_{i=1}^{n}(Y_i - \bar{Y})(x_i - \bar{x}) - \hat{\beta}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \sum_{i=1}^{n}(Y_i - \bar{Y})(x_i - \bar{x}) - \hat{\beta}K \\
&= \sum_{i=1}^{n} Y_i(x_i - \bar{x}) - \sum_{i=1}^{n} Y_i(x_i - \bar{x}) \\
&= 0
\end{aligned}$$

Therefore, all the cross-terms are zero.

Back to the analysis of variance formula. . .

$$\begin{aligned}
&\sum_{i=1}^{n}(Y_i-\alpha_0-\beta(x_i-\bar{x}))^2 \\
&\qquad=\sum_{i=1}^{n}(Y_i-\hat{\alpha}_0-\hat{\beta}(x_i-\bar{x}))^2+n(\hat{\alpha}_0-\alpha_0)^2+K(\hat{\beta}-\beta)^2
\end{aligned}$$

Taking expectations gives,

$$\begin{aligned}
n\sigma^2 &= \mathbb{E}(D^2)+\sigma^2+\sigma^2 \\
\Rightarrow\quad \mathbb{E}(D^2) &= (n-2)\sigma^2
\end{aligned}$$

where

$$D^2=\sum_{i=1}^{n}(Y_i-\hat{\alpha}_0-\hat{\beta}(x_i-\bar{x}))^2.$$

3.3 Point estimation of the variance

Variance estimator

Based on these results, we have an unbiased estimator of the variance,

$$\hat{\sigma}^2=\frac{1}{n-2}D^2.$$

The inferred mean for each observation is called its *fitted value*, $\hat{Y}_i=\hat{\alpha}_0+\hat{\beta}(x_i-\bar{x})$.

The deviation from each fitted value is called a *residual*, $R_i=Y_i-\hat{Y}_i$.

The variance estimator is based on the sum of squared residuals, $D^2=\sum_{i=1}^{n}R_i^2$.

Example: variance estimate

For our data:

$$\begin{aligned}
d^2 &= 16.12 \\
\hat{\sigma}^2 &= 2.015 \\
\hat{\sigma} &= 1.42
\end{aligned}$$

3.4 Standard errors of the estimates

Standard errors

We can subsitute $\hat{\sigma}^2$ into the formulae for the standard deviation of the estimators in order to calculate standard errors.

For example,

$$\begin{aligned}
\mathrm{var}(\hat{\beta}) &= \frac{\sigma^2}{K} \\
\Rightarrow\quad \mathrm{se}(\hat{\beta}) &= \frac{\hat{\sigma}}{\sqrt{K}}
\end{aligned}$$

Example: standard errors

For our data:

$$
\begin{aligned}
\operatorname{se}(\hat{\alpha}_0) &= \frac{\hat{\sigma}}{\sqrt{n}} = 0.449 \\
\operatorname{se}(\hat{\beta}) &= \frac{\hat{\sigma}}{\sqrt{K}} = 0.404 \\
\operatorname{se}(\hat{\mu}(x)) &= \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{K}} = 1.42 \times \sqrt{\frac{1}{10} + \frac{(x-1.78)^2}{12.34}}
\end{aligned}
$$

3.5 Confidence intervals

Maximum likelihood estimation

Want to also construct confidence intervals. This requires further assumptions about the population distribution.

Let's assume a normal distribution:

$$Y_i \sim \mathrm{N}(\alpha + \beta x_i, \sigma^2).$$

Alternative notation (commonly used for regression/linear models):

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad \text{where} \quad \epsilon_i \sim \mathrm{N}(0, \sigma^2).$$

Let's maximise the likelihood...

Since the Y_i's are independent, the likelihood is:

$$
\begin{aligned}
L(\alpha, \beta, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - \alpha - \beta x_i)^2}{2\sigma^2}\right\} \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{\sum_{i=1}^{n}(y_i - \alpha_0 - \beta(x_i - \bar{x}))^2}{2\sigma^2}\right\} \\
-\ln L(\alpha, \beta, \sigma^2) &= \frac{n}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha_0 - \beta(x_i - \bar{x}))^2 \\
&= \frac{n}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2} H(\alpha_0, \beta)
\end{aligned}
$$

The α_0 and β that maximise the likelihood (minimise the log-likelihood) are the same as those that minimise the sum of squares, H.

The OLS estimates are the same as the MLEs!

What about σ^2?

Differentiate by σ, set to zero, solve...

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n} D^2$$

This is biased. Prefer to use the previous, unbiased estimator,

$$\hat{\sigma}^2 = \frac{1}{n-2} D^2$$

Sampling distributions

The $Y_1, \cdots, Y_n$ are independent normally distributed random variables.

Except for $\hat{\sigma}^2$, our estimators are linear combinations of the Y_i so will also have normal distributions, with mean and variance as previously derived.

For example,

$$\hat{\beta} \sim \mathrm{N}\left(\beta, \frac{\sigma^2}{K}\right).$$

Moreover, we know $\hat{\alpha}_0$ and $\hat{\beta}$ are independent, because they are bivariate normal rvs with zero covariance.

Using the analysis of variance decomposition (from earlier), we can show that,

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}.$$

Therefore, we can define pivots for the various mean parameters. For example,

$$\frac{\hat{\beta} - \beta}{\hat{\sigma}/\sqrt{K}} \sim t_{n-2}$$

and

$$\frac{\hat{\mu}(x) - \mu(x)}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{K}}} \sim t_{n-2}$$

This allows us to construct confidence intervals.

Example: confidence itervals

For our data, a 95% CI for β is:

$$\hat{\beta} \pm c\,\frac{\hat{\sigma}}{\sqrt{K}} = 2.59 \pm 2.31 \times 0.404 = (1.66, 3.52)$$

where c is the 0.975 quantile of t_{n-2}.

A 95% CI for $\mu(3)$ is:

$$\hat{\mu}(3) \pm c \times \mathrm{se}(\hat{\mu}(3)) = 10.68 \pm 2.31 \times 0.667 = (9.14, 12.22)$$

3.6 Prediction intervals

Deriving prediction intervals

Use the same trick as we used for the simple model,

$$\begin{aligned}
Y^* &\sim \mathrm{N}\left(\mu(x^*), \sigma^2\right) \\
\hat{\mu}(x^*) &\sim \mathrm{N}\left(\mu(x^*), \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{K}\right)\sigma^2\right) \\
Y^* - \hat{\mu}(x^*) &\sim \mathrm{N}\left(0, \left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{K}\right)\sigma^2\right)
\end{aligned}$$

A 95% PI for Y^* is given by:

$$\hat{\mu}(x^*) \pm c\,\hat{\sigma}\,\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{K}}$$

Example: prediction interval

A 95% PI for Y^* corresponding to $x^* = 3$ is:

$$10.68 \pm 2.31 \times 1.42 \times \sqrt{1 + \frac{1}{10} + \frac{(3 - 1.78)^2}{12.34}} = (7.06, 14.30)$$

Much wider than the corresponding CI, as we've seen previously.

3.7 R examples

```
> model1 <- lm(y ~ x)
> summary(model1)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-2.01970 -1.05963  0.02808  1.04774  1.80580

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.9114     0.8479   3.434 0.008908 **
x             2.5897     0.4041   6.408 0.000207 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.419 on 8 degrees of freedom
Multiple R-squared:  0.8369,Adjusted R-squared:  0.8166
F-statistic: 41.06 on 1 and 8 DF,  p-value: 0.0002074

> # Confidence intervals for mean parameters
> confint(model1)
                2.5 %   97.5 %
(Intercept) 0.9560629 4.866703
x           1.6577220 3.521623

> # Data to use for prediction.
> data2 <- data.frame(x = 3)

> # Confidence interval for mu(3).
> predict(model1, newdata = data2, interval = "confidence")
      fit      lwr      upr
1 10.6804 9.142823 12.21798

> # Prediction interval for y when x = 3.
> predict(model1, newdata = data2, interval = "prediction")
      fit   lwr     upr
1 10.6804 7.064 14.2968
```

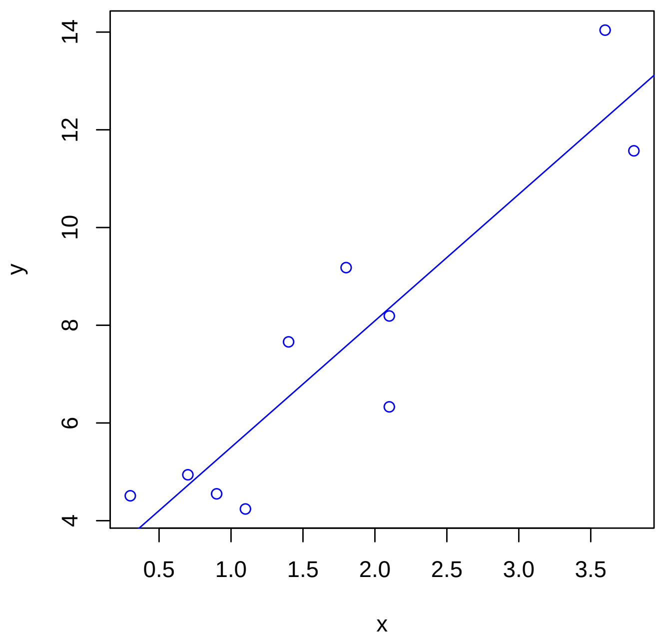

R example explained

- The `lm` (linear model) command fits the model.
- `model1` is an object that contains all the results of the regression needed for later calculations.
- `summary(model1)` acts on `model1` and summarizes the regression.
- `predict` can calculate CIs and PIs.
- R provides more detail than we need at the moment. Much of the output relates to hypothesis testing that we will get to later.

Plot data and fitted model

```
> plot(x, y, col = "blue")
> abline(model1, col = "blue")
```

The command `abline(model1)` adds the fitted line to a plot.

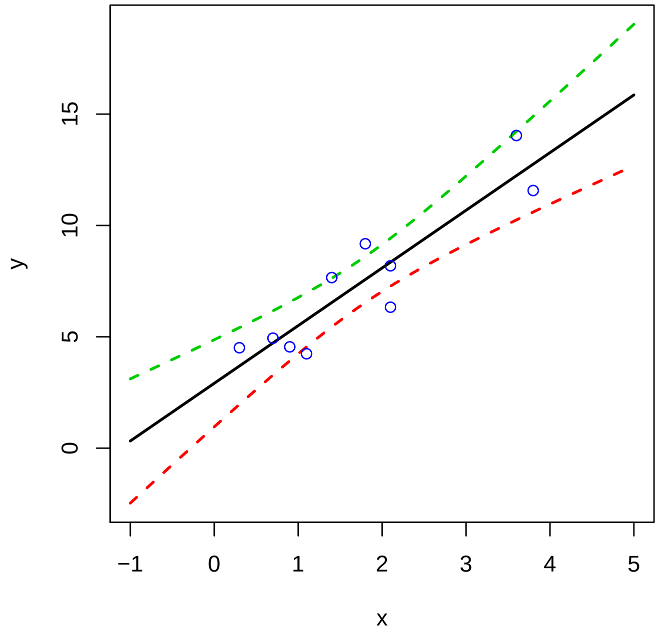


Fitted values and CIs for their means

```
> predict(model1, interval = "confidence")
         fit       lwr       upr
1   7.572793  6.537531  8.608056
2   6.536924  5.442924  7.630925
3   8.349695  7.272496  9.426895
4   3.688285  1.963799  5.412771
5  12.234204 10.247160 14.221248
6   4.724154  3.280382  6.167925
7   5.760023  4.546338  6.973707
8   8.349695  7.272496  9.426895
9   5.242088  3.921478  6.562699
10 12.752138 10.603796 14.900481
```

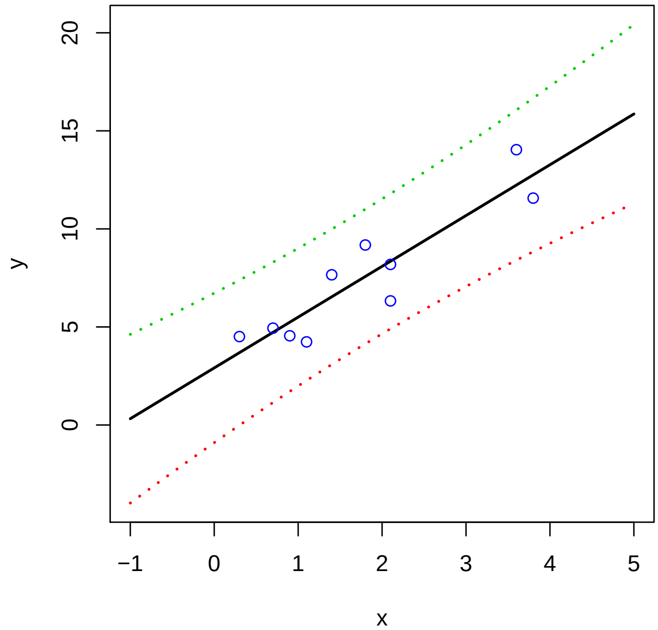

Confidence band for the mean

```
> data3 <- data.frame(x = seq(-1, 5, 0.05))
> y.conf <- predict(model1, data3, interval = "confidence")
> head(cbind(data3, y.conf))
      x       fit       lwr      upr
1 -1.00 0.3217104 -2.468232 3.111653
2 -0.95 0.4511941 -2.295531 3.197919
3 -0.90 0.5806777 -2.122943 3.284298
4 -0.85 0.7101613 -1.950472 3.370794
5 -0.80 0.8396449 -1.778124 3.457414
6 -0.75 0.9691286 -1.605906 3.544164

> matplot(data3$x, y.conf, type = "l", lty = c(1, 2, 2),
+         lwd = 2, xlab = "x", ylab = "y")
> points(x, y, col = "blue")
```



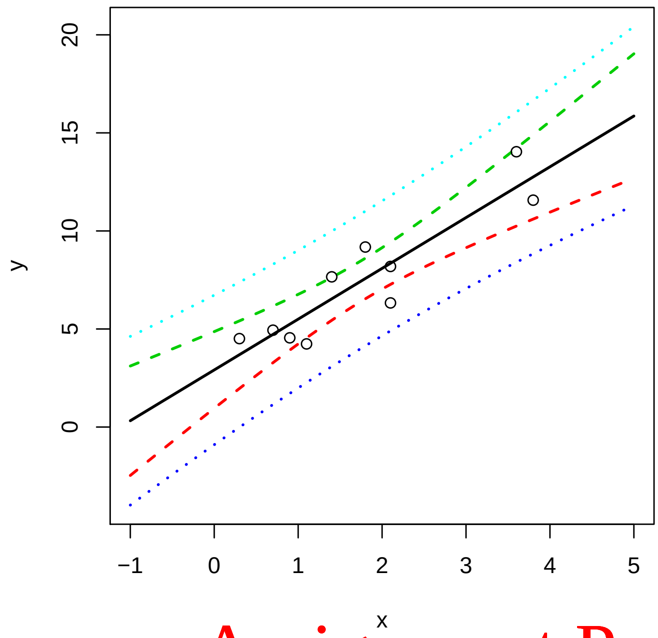
Prediction bands for new observations

```
> y.pred <- predict(model1, data3, interval = "prediction")
> head(cbind(data3, y.pred))
      x       fit       lwr      upr
1 -1.00 0.3217104 -3.979218 4.622639
2 -0.95 0.4511941 -3.821827 4.724215
3 -0.90 0.5806777 -3.664763 4.826119
4 -0.85 0.7101613 -3.508034 4.928357
5 -0.80 0.8396449 -3.351646 5.030936
6 -0.75 0.9691286 -3.195606 5.133863

> matplot(data3$x, y.pred, type = "l", lty = c(1, 3, 3),
+         lwd = 2, xlab = "x", ylab = "y")
> points(x, y, col = "blue")
```

Both bands plotted together

```
> matplot(data3$x, cbind(y.conf, y.pred[, -1]),
+         type = "l", lwd = 2, lty = c(1, 2, 2, 3, 3),
+         xlab = "x", ylab = "y")
> points(x, y)
```



3.8 Model checking

Checking our assumptions

What modelling assumptions have we made?

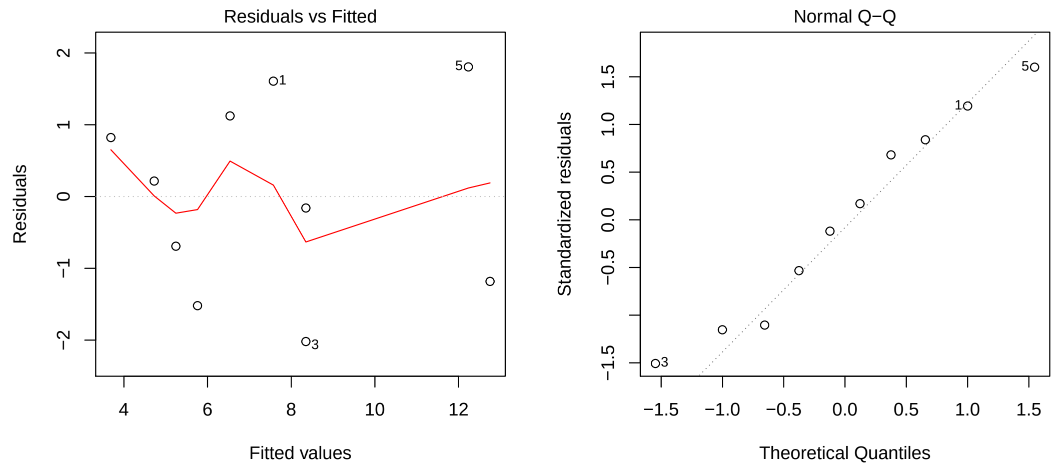
- Linear model for the mean
- Equal variances for all observations (*homoscedasticity*)
- Normally distributed residuals

Ways to check these:

- Plot the data and fitted model together (done!)
- Plot residuals vs fitted values
- QQ plot of the residuals

In R, the last two of these are very easy to do:

```
> plot(model1, 1:2)
```



4 Further regression models

Multiple regression

- What if we have more than one predictor?
- Observe $x_{i1}, \ldots, x_{ik}$ as well as y_i (for each i)
- Can fit a *multiple regression model*:

$$\mathbb{E}(Y \mid x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- This is linear in the coefficients, so is still a linear model

- Fit by method of least squares by minimising:

$$H = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik})^2$$

- Take partial derivatives, etc., and solve for $\beta_0, \ldots, \beta_k$.
- The subject Linear Statistical Models (MAST30025) looks into these types of models in much more detail.

Two-sample problem

- The two-sample problem can be expressed as a linear model!
- Sample $Y_1, \ldots, Y_n \sim \mathrm{N}(\mu_1, \sigma^2)$ and $Y_{n+1}, \ldots, Y_{n+m} \sim \mathrm{N}(\mu_2, \sigma^2)$.
- Define *indicator variables* (x_{i1}, x_{i2}) where $(x_{i1}, x_{i2}) = (1, 0)$ for $i = 1, \ldots, n$ and $(x_{i1}, x_{i2}) = (0, 1)$ for $i = n + 1, \ldots, n + m$.
- Observed data: $(y_i,\ x_{i1},\ x_{i2})$
- Then $Y_1, \ldots, Y_n$ each have mean $1 \times \beta_1 + 0 \times \beta_2 = \mu_1$ and $Y_{n+1}, \ldots, Y_{n+m}$ each have mean $0 \times \beta_1 + 1 \times \beta_2 = \mu_2$.
- This is in the form a multiple regression model.
- The *general linear model* unifies many different types of models together into a common framework. The subject MAST30025 covers this in more detail.

5 Correlation

5.1 Definitions

Correlation coefficient

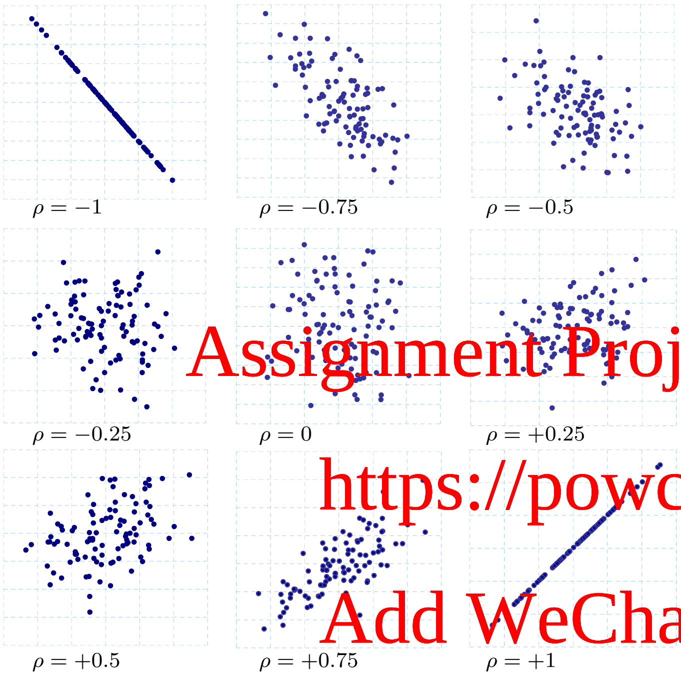
(Revision) for two rvs X and Y, the *correlation coefficient*, or simply the *correlation*, is defined as:

$$\rho = \rho_{XY} = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var} X \operatorname{var} Y}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

This is a quantitative measure of the strength of relationship, or *association*, between X and Y.

We will now consider inference on ρ, based on an iid sample of pairs (X_i, Y_i).

Note: unlike in regression, X is now considered as a random variable.



5.2 Point estimation

Sample covariance

To estimate $\operatorname{cov}(X,Y)$ we use the *sample covariance*:

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}\right)$$

You can check that this is unbiased, $\mathbb{E}(S_{XY}) = \sigma_{XY} = \operatorname{cov}(X,Y)$.

Sample correlation coefficient

To estimate ρ we use the *sample correlation coefficient* (also known as *Pearson's correlation coefficient*):

$$R = R_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

You can check that $|R| \leqslant 1$, just like $|\rho| \leqslant 1$.

This gives a point estimate of ρ.

For further results, we make some more assumptions...

5.3 Relationship to regression

Bivariate normal

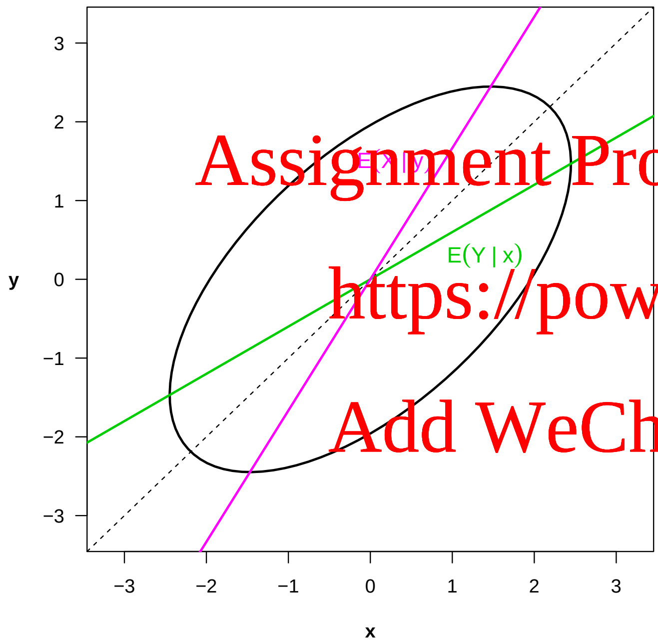
Assume X and Y have correlation ρ and follow a bivariate normal distribution,

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathrm{N}_2\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}\right)$$

In this case, the regressions are linear,

$$\begin{aligned} \mathbb{E}(X \mid Y = y) &= \mu_X + \frac{\rho\sigma_X}{\sigma_Y}(y - \mu_Y) = \alpha' + \beta' y \\ \mathbb{E}(Y \mid X = x) &= \mu_Y + \frac{\rho\sigma_Y}{\sigma_X}(x - \mu_X) = \alpha + \beta x \end{aligned}$$

Note: $\beta' \neq 1/\beta$



Variance explained

An alternative analysis of variance decomposition:

$$\begin{aligned} \sum (Y_i - \bar{Y})^2 &= \sum (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2 + \hat{\beta}^2 \sum (x_i - \bar{x})^2 \\ &= (1 - R^2) \sum (Y_i - \bar{Y})^2 + R^2 \sum (Y_i - \bar{Y})^2 \end{aligned}$$

This implies that R^2 is the proportion of the variation in Y 'explained' by x.

In this usage, R^2 is called the *coefficient of determination*.

Remarks

- For simple linear regression, the coefficient of determination is the same as the square of the sample correlation, with both being denoted by R^2.

- Also, the proportion of Y explained by x is the same as the proportion of X explained by y. Both are equal to R^2, which is a symmetric expression of both X and Y.
- For more complex models, the coefficient of determination is more complicated: it needs to be calculated using all predictor variables together.

5.4 Confidence interval

Approximate sampling distribution

Define:

$$g(r) = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

This function has a standard name, $g(r) = \operatorname{artanh}(r)$, and so does it's inverse, $g^{-1}(r) = \tanh(r)$. The function $g(r)$ is also known as the *Fisher transformation.*

The following is a widely used approximation:

$$g(R) \approx \mathrm{N}\left(g(\rho), \frac{1}{n-3}\right)$$

We can use this to construct approximate confidence intervals.

Example: correlation

For our data:

$$r = 0.91$$
$$r^2 = 0.84$$

An approximate 95% CI for $g(\rho)$ is:

$$g(r) \pm \frac{c}{\sqrt{n-3}} = 1.56 \pm 1.96 \times 0.378 = (0.819, 2.30)$$

where $c = \Phi^{-1}(1-\alpha/2)$. Transforming this to an approximate 95% CI for ρ:

$$(\tanh(0.819), \tanh(2.30)) = (0.67, 0.98)$$

5.5 R example

```
> cor(x, y)
[1] 0.9148421

> cor(x, y)^2
[1] 0.836936

> cor.test(x, y)

Pearson's product-moment correlation

data:  x and y
t = 6.4078, df = 8, p-value = 0.0002074
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6726924 0.9799873
sample estimates:
      cor
0.9148421
```

```
> model1 <- lm(y ~ x)
> summary(model1)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-2.01970 -1.05963  0.02808  1.04774  1.80580

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.9114     0.8479   3.434 0.008908 **
x             2.5897     0.4041   6.408 0.000207 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.419 on 8 degrees of freedom
Multiple R-squared:  0.8369,Adjusted R-squared:  0.8166
F-statistic: 41.06 on 1 and 8 DF,  p-value: 0.0002074
```