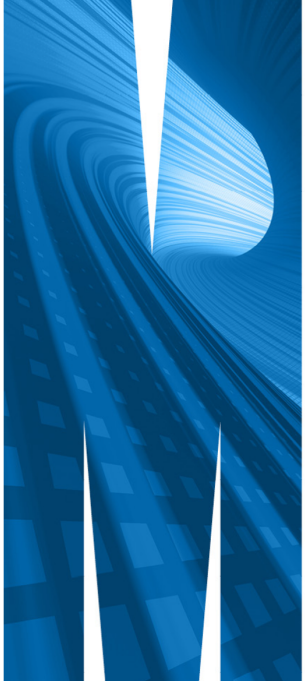
MONASH University

Regression models

Statistical Thinking (ETC2420 / ETC5242)

Week 9, Semester 2, 2025



Outline

Outline

Learning goals for Week 9

- Understand and recognise different goals of statistical modelling
- Review simple linear regression
- Understand the least squares approach to fitting a simple linear regression model
- Apply the statistical concepts learnt thus far (e.g., sampling distributions) to a simple linear regression model
- Calculate confidence intervals and carry out hypothesis tests using a simple linear regresison model
- Use a fitted regression model for prediction
- Assess the adequacy of simple linear regression model

Outline

Relationships between two variables

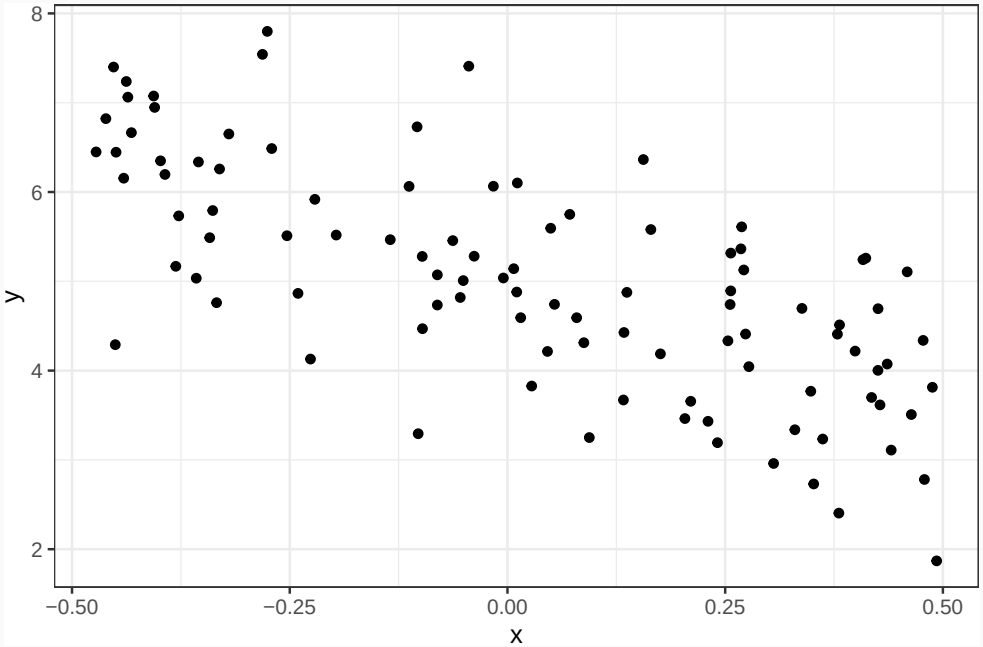
We have studied how to do estimation for some simple scenarios:

- 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**.


8
6
4
2
y
−0.50
−0.25
0.00
0.25
0.50
x

Regression model

- Often interested in how Y depends on X.
- For example, we might want to use X to predict Y.
- 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 the price of a house and x is the floor area. Does x help to predict Y?
- The **regression** of Y on x is the conditional mean:

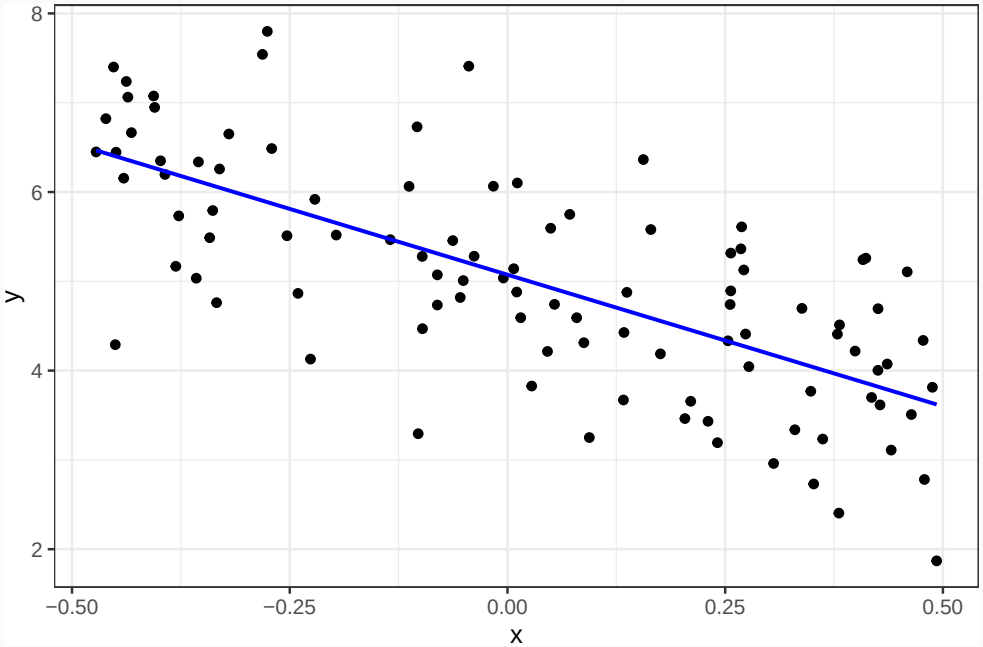
$$\mathbb{E}(Y \mid x) = \mu(x)$$

- The regression can take any form. Can consider arbitrary functions for $\mu(x)$.
- We consider **simple linear regression**, which has the form of a **straight line**:

$$\mathbb{E}(Y \mid x) = \beta_0 + \beta_1 x$$

- We also assume constant variance:

$$\mathrm{var}(Y \mid x) = \sigma^2$$


8
6
4
2
y
−0.50
−0.25
0.00
0.25
0.50
x

Simple linear regression

- Explains how the response variable, y, changes (linearly) in relation to an explanatory variable, x, on average.
- Usually have a sample and use i to index the variables: x_i and y_i.
- Different ways to write the model:

$$\mathbb{E}(Y_i \mid x_i) = \beta_0 + \beta_1 x_i$$
$$Y_i = \beta_0 + \beta_1 x_i + E_i$$
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- Can omit the subscript i if referring to a generic pair of values, (x, Y).
- The regression line is an average, it "balances out" the points above and below the line.

Terminology

- Y is called a **response** variable. Can also be called an **outcome** or **target** variable. Do NOT call it the 'dependent' variable.
- x is called a **predictor** variable. Can also be called an **explanatory** variable. Do NOT call it an 'independent' variable.
- $\mu(x)$ is called the **(linear) predictor function** or sometimes the **regression line** or **regression curve** or the **model equation**.
- The parameters in the predictor function (β_0 and β_1) 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 (in an algebraic sense).

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 (an additive function) of them, with the coefficients 'out the front'.

General form:

$$\mu(x) = \beta_1 f_1(x) + \cdots + \beta_k f_k(x)$$

Linearity examples

The following are linear models:

$$\mu(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

$$\mu(x) = \frac{\beta_1}{x} + \frac{\beta_2}{x^2}$$

$$\mu(x) = \beta_1 \sin x + \beta_2 \log x$$

The following are NOT linear models:

$$\mu(x) = \beta_0 \sin(\beta_1 x)$$

$$\mu(x) = \frac{\beta_0}{1 + \beta_1 x}$$

$$\mu(x) = \beta_0 x^{\beta_1}$$

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

$$\mu^*(x) = \beta_0^* + \beta_1 \log x$$

What is “simple”?

A simple linear regression is the **simplest** model equation that is **linear** and includes **one explanatory variable**.

$$\mu(x) = \beta_0 + \beta_1 x$$

Can you think of a *simpler* model?

Outline

Estimation goals

Simple linear regression model:

$$\mathbb{E}(Y \mid x) = \beta_0 + \beta_1 x \quad \text{and} \quad \operatorname{var}(Y \mid x) = \sigma^2$$

- We wish to estimate the slope (β_1), the intercept (β_0), 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.

Least squares estimation

- Define the sum of squared deviations:

$$H(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

- Find β_0 and β_1 that minimises this sum. (Can do this using partial derivatives.)
- This gives the **least squares estimators**:

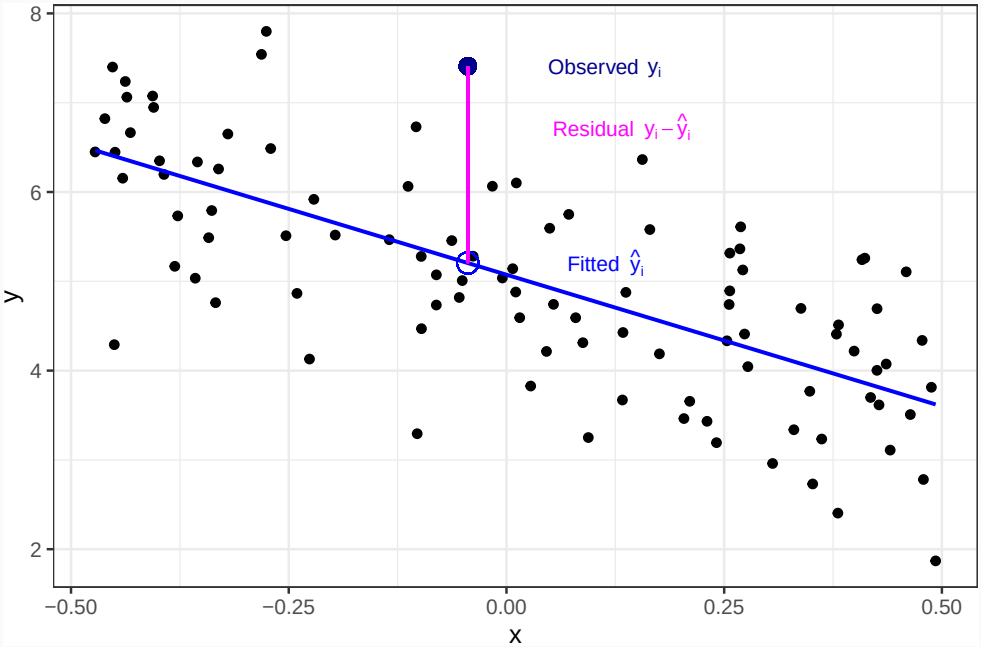
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) Y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

- This method is often called **ordinary least squares** (OLS).
- Gives a “line of best fit”.
- The **fitted values** (or **predicted values**) are those that lie on the regression line:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- The deviations from the line are called **residuals**:

$$E_i = Y_i - \hat{Y}_i$$


Observed y_i
Residual $y_i - \hat{y}_i$
Fitted $\hat{y}_i$
y
x
8
6
4
2
−0.50
−0.25
0.00
0.25
0.50

Parameter interpretation

- Fitted line: $\hat{\mu}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$
- $\hat{\beta}_0$ is the **intercept** of the fitted line with y-axis
- $\hat{\beta}_1$ is the **slope** of the fitted line

Does the point $(\bar{x}, \bar{y})$ lie on the fitted line?

R example

Peek at the data:

```
head(df)
```

```
# A tibble: 6 x 2
         x     y
     <dbl> <dbl>
1 -0.271    6.49
2 -0.338    5.79
3  0.256    4.74
4  0.488    3.81
5 -0.450    4.29
6 -0.00470  5.04
```

Fit the model:

```
lm(y ~ x, data = df)
```

```
Call:
lm(formula = y ~ x, data = df)

Coefficients:
(Intercept)            x
      5.074       -2.948
```

Properties of these estimators

- All of these estimators are **unbiased**.

$$\mathbb{E}(\hat{\beta}_0) = \beta_0 \qquad \mathbb{E}(\hat{\beta}_1) = \beta_1 \qquad \mathbb{E}(\hat{\mu}(x)) = \mu(x)$$

- Their variances are:

$$\begin{aligned}
\operatorname{var}(\hat{\beta}_1) &= \frac{1}{K}\sigma^2 \\
\operatorname{var}(\hat{\beta}_0) &= \left(\frac{1}{n} + \frac{\bar{x}^2}{K}\right)\sigma^2 \\
\operatorname{var}(\hat{\mu}(x^*)) &= \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{K}\right)\sigma^2
\end{aligned}$$

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

Quantifying uncertainty, first steps

Variance estimator

The following estimator of σ^2 is unbiased:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} E_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Note the similarity to the sample variance:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

Standard errors

- Plug in $\hat{\sigma}$ to get standard errors.
- For example:

$$\mathrm{se}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{K}}$$

Coefficient of determination (R^2)

Define the following sums of squares:

$$\text{SSR} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \quad \text{Regression sum of squares}$$

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \quad \text{Error sum of squares}$$

$$\text{SST} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \quad \text{Total sum of squares}$$

Define:

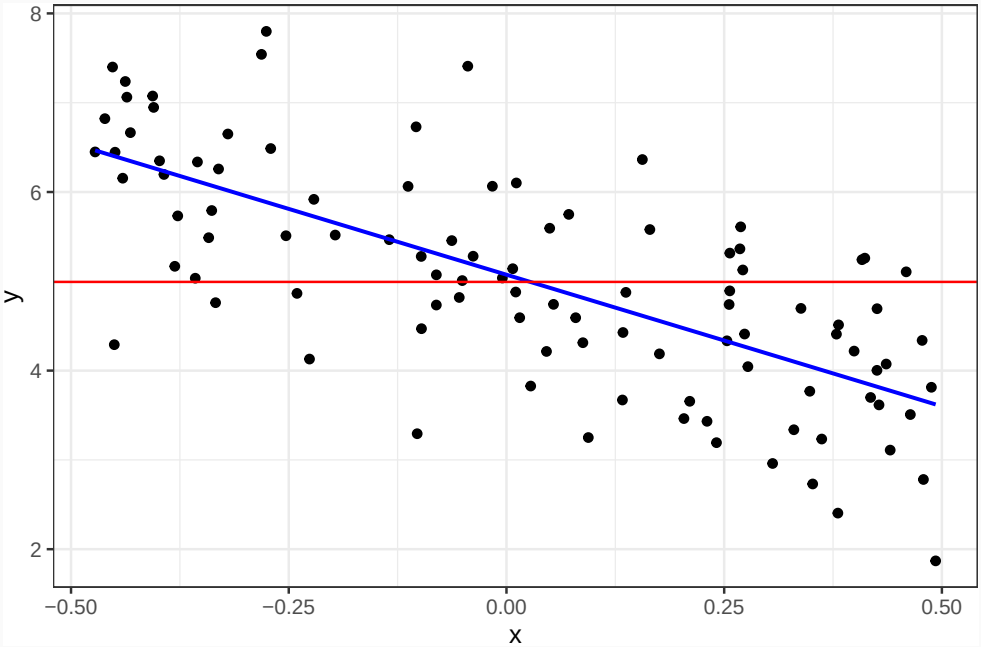
$$R^2 = \frac{\text{SSR}}{\text{SST}}$$

R^2 is called the **coefficient of determination**.

It quantifies the **proportion of variation** of the response variables (Y_i's) that is **explained** by the regression model.

Note: we can show that $\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}$, which means $0 \leqslant R^2 \leqslant 1$.

(Next slide:) R^2 gives us an idea of how much better our predictions of Y would be if we used the fitted model instead of just the mean $\bar{Y}$.


8
6
4
2
y
−0.50
−0.25
0.00
0.25
0.50
x

R example

```r
m1 <- lm(y ~ x, data = df)
glance(m1)
```

```
# A tibble: 1 x 12
  r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>
1     0.504         0.499 0.894      99.5 1.39e-16     1  -130.  265.  273.
# i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```r
tidy(m1)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)     5.07    0.0898      56.5  1.32e-76
2 x              -2.95    0.296       -9.97 1.39e-16
```

Outline

More assumptions required

- So far, have **not** assumed any specific probability distribution.
- We want to be able to calculate confidence intervals and hypothesis tests.
- This requires further assumptions.
- Let's assume a **normal distribution**:

$$Y_i \sim \mathrm{N}(\beta_0 + \beta_1 x_i, \sigma^2).$$

- Alternative notation (more common for regression models):

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \text{where} \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2).$$

Maximum likelihood estimation

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

$$
\begin{aligned}
L(\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^{n} f(Y_i = y_i \mid x_i) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right\} \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left\{ -\frac{\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right\} \\
-2 \ln L(\beta_0, \beta_1, \sigma^2) &= n \ln(2\pi\sigma^2) + \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \\
&= n \ln(2\pi\sigma^2) + \frac{1}{\sigma^2} H(\beta_0, \beta_1)
\end{aligned}
$$

The β_0 and β_1 that maximise the likelihood (minimise the log-likelihood) are the same as those that minimise the sum of squared deviations, $H(\beta_0, \beta_1)$.

$\Rightarrow$ The OLS estimates are the same as the MLEs!

What about σ^2?

Differentiate by σ, set to zero, solve…

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} E_i^2$$

This is biased.

We prefer to use the previous, unbiased estimator,

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} E_i^2$$

Sampling distributions

Under the normal distribution assumption, we can derive the sampling distributions of all of our estimators.

The ones that are the most useful are for the slope, β_1, and the predictor function, $\mu(x)$:

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{K}} \sim \mathrm{t}_{n-2}$$

and

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

This allows us to construct confidence intervals and perform hypothesis tests. . . . in practice, R does all the calculations for us!

Examples: confidence intervals

```
confint(m1)
```

```
               2.5 %    97.5 %
(Intercept)  4.89570  5.251963
x           -3.53417 -2.361084
```

```
df2 <- tibble(x = 0.2)
predict(m1, newdata = df2, interval = "confidence")
```

```
       fit      lwr      upr
1 4.484306 4.280027 4.688586
```

Examples: confidence intervals

The 95% CI for β_1 is:

$$\hat{\beta}_1 \pm c \times \mathrm{se}(\hat{\beta}_1) = -2.95 \pm 1.98 \times 0.296$$
$$= (-3.54, -2.36)$$

The 95% CI for $\mu(0.2)$ is:

$$\hat{\mu}(0.2) \pm c \times \mathrm{se}(\hat{\mu}(0.2)) = 4.48 \pm 1.98 \times \mathrm{se}(\hat{\mu}(0.2))$$
$$= (4.28, 4.69)$$

In both cases, $c = 1.98$ is the 0.975 quantile of t_{98}.

Examples: hypothesis tests

```
tidy(m1)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)     5.07    0.0898      56.5  1.32e-76
2 x              -2.95    0.296       -9.97 1.39e-16
```

Examples: hypothesis tests

The R output on the previous slide shows the results of the following tests:

$$H_0 : \beta_0 = 0 \quad \text{vs} \quad H_1 : \beta_0 \neq 0$$

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0$$

Outline

Predicting a future value

Given a new predictor value x^*, how can we predict the corresponding response value, Y^*?

A **point prediction** is given directly from the fitted regression line:

$$\hat{Y}^* = \hat{\mu}(x^*) = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

How certain are we of this prediction?

Can we get an interval estimate for the predicted value?

Prediction intervals

We can show the following distribution relates the point predictor, $\hat{\mu}(x^*)$, and the true future value, Y^*,

$$\frac{\hat{\mu}(x^*) - Y^*}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{K}}} \sim \mathrm{t}_{n-2}$$

This allows us to construct the following interval estimate for Y^*:

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

This is known as a **prediction interval** (PI).

Notes:

- A prediction interval (PI) is similar to a confidence interval (CI), but is for estimating a random quantity, Y^*, rather than a fixed quantity, such as $\mu(x^*)$.
- The extra "1+" term will make a PI much wider than a corresponding CI.

Example: prediction intervals

```
df2 <- tibble(x = c(0.1, 0.15, 0.2))
df2
```

```
# A tibble: 3 x 1
      x
  <dbl>
1  0.1
2  0.15
3  0.2
```

```
predict(m1, newdata = df2, interval = "prediction")
```

```
       fit      lwr      upr
1 4.779069 2.995628 6.562510
2 4.631688 2.847305 6.416070
3 4.484306 2.698501 6.270112
```

These are much wider than the corresponding CIs!

Prediction intervals vs confidence intervals

- Confidence intervals are for **fixed** quantities, such as parameters.
- Prediction intervals are for **random** quantities, such as future or unobserved values.
- Same (frequentist) interpretation of probability: a “95% CI” or “95% PI” means that 95% of such intervals should contain the target value, under hypothetical repeated sampling.
- A prediction interval could also be called a “predictive confidence interval”.

Prediction intervals vs confidence intervals

For regression models

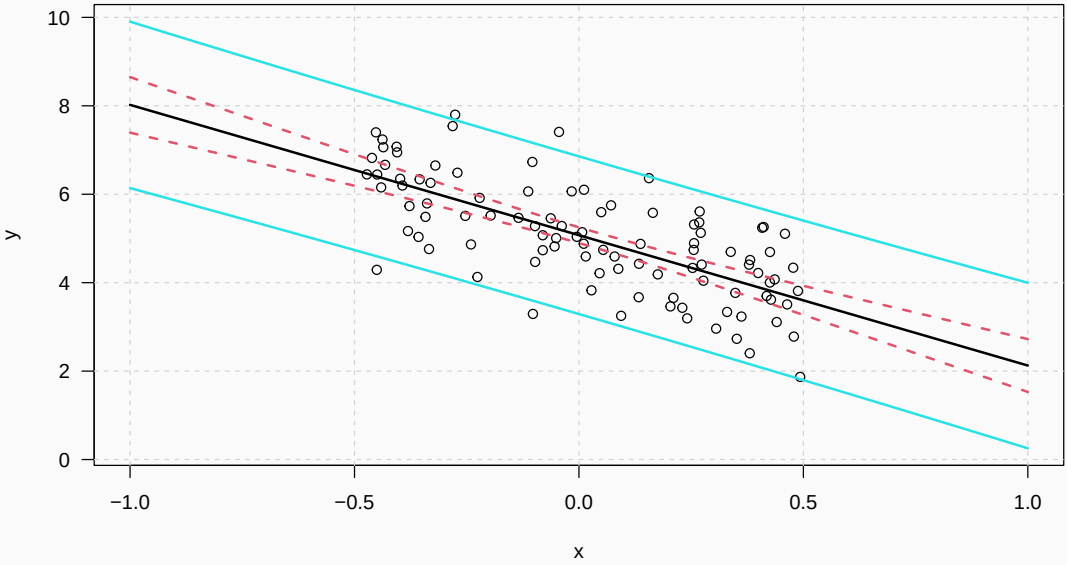
We might be interested in two related quantities:

- $\mu(x^*)$, which is a fixed value
- Y^*, which is random variable with mean $\mu(x^*)$

The point estimates/predictions are the same: $\hat{Y}^* = \hat{\mu}(x^*) = \hat{\beta}_0 + \hat{\beta}_1 x^*$.

But the PI for Y^* will be much wider than the CI for $\mu(x^*)$.

Prediction intervals vs confidence intervals



Reverse prediction

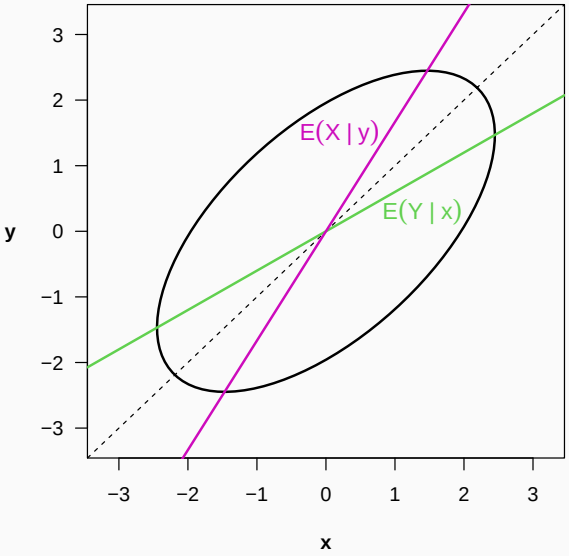
Suppose we:

- Predict X using Y, using a regression model
- Predict Y using X, using a regression model

Is one model the “inverse” of the other?

Are we using the same “line of best fit”?

Regression relationships are *not* symmetric



Outline

The three goals of statistical modelling

Description:

- Summarise or represent data structure in a compact manner.
- Provide a simple quantitative summary of certain features of the world.
- Describe associations (not necessarily causal) between variables.

Prediction:

- Predict new or future observations.

“Explanation” (causal inference):

- Infer the effect of **interventions**.
- Predict potential observations after an intervention is applied.
- Test and assess causal explanations of the world.

Description

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Interested in:

- Whether the relationship between Y and X is close enough to a straight line for this to be an acceptable simplification of reality.
- The value of β_1 as describing how Y changes on average as X changes (but remembering that it is not necessarily causal).

Prediction

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Interested in:

- Accurately predicting Y for any given X.
- Don't care about the regression coefficients, only care about prediction accuracy.
- It doesn't matter how "bad" the model fits, e.g. if the true relationship is far from a straight line, as long as the prediction accuracy is acceptable.

Explanation / Causal inference

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Interested in:

- Predictions of Y under interventions, described by different values of X
- The value of β_1, as describing an average causal effect of changes in X
- These interpretations are only valid if the model is correctly specified and is used with an appropriate dataset that admits a causal interpretation. For example, data from a randomised controlled trial.

This type of modelling is more advanced, and requires more care and training, than for description or prediction. We mainly discuss the other two uses in this unit.

Outline

What is a “good” model?

Depends on your goals

- Description: Does it describe the data or population adequately well?
- Prediction: Does it give accurate predictions?
- Explanation / causal inference: Does it accurately estimate the causal effect?

We focus mainly on the **description** goal for now…

Checking our assumptions

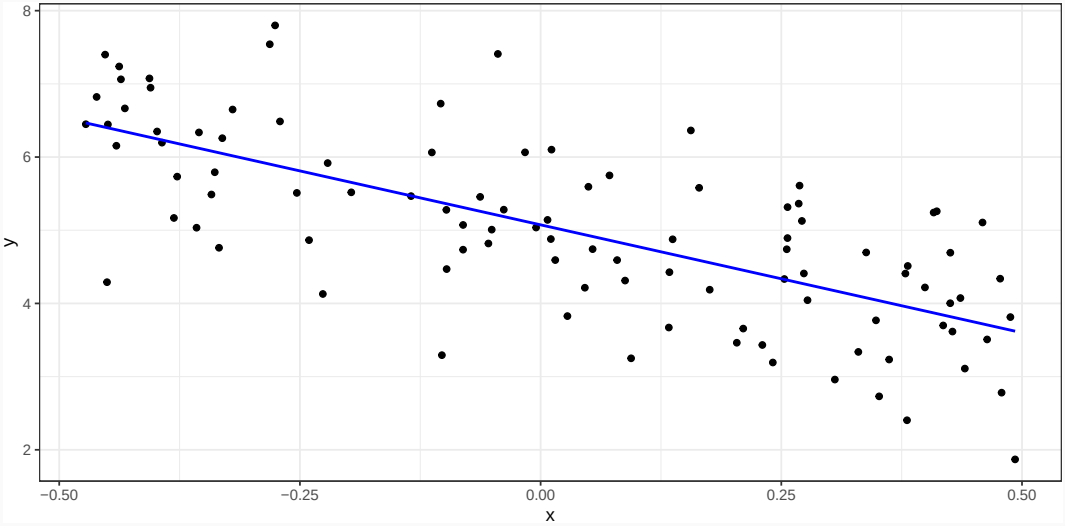
What modelling assumptions have we made?

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

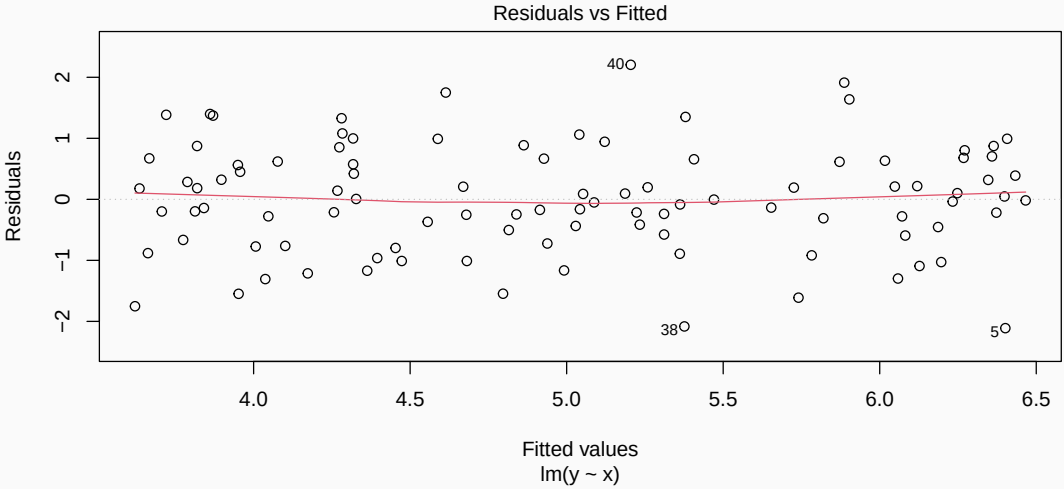
Ways to check these:

- Plot the data and fitted model together
- Plot residuals vs fitted values
- QQ plot of the residuals

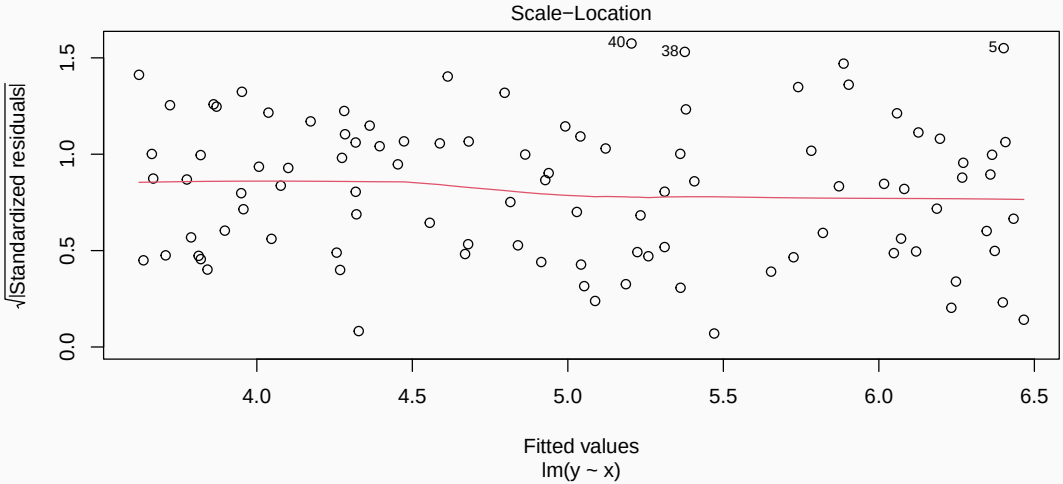
Data and fitted model



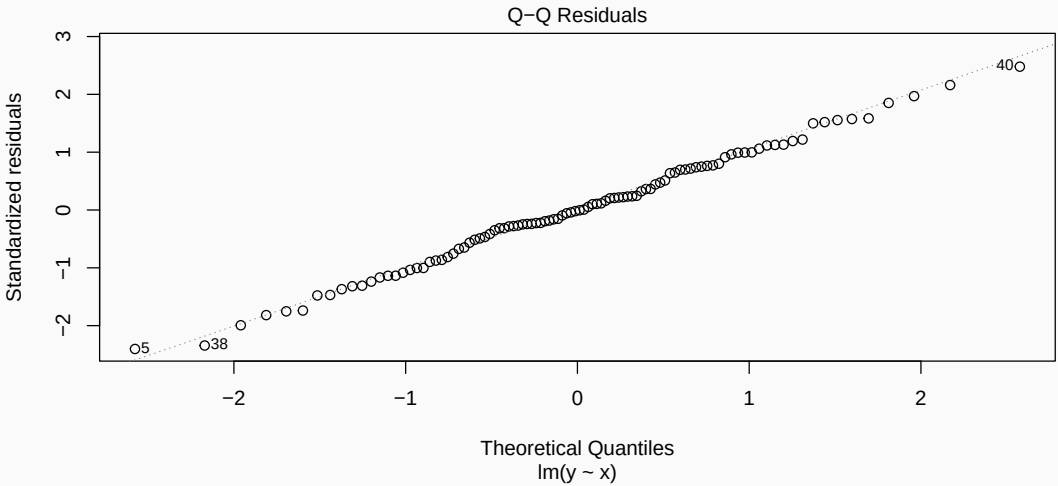
Residuals vs fitted values



Residuals vs fitted values (alternate version)



QQ plot of residuals



Diagnostics in R

To draw the previous 3 plots:

```
m1 <- lm(y ~ x, data = df)
plot(m1, 1:3)
```

More info on the following help page:

```
help("plot.lm")
```

Transformations

If the model is a poor fit, one thing to consider is whether a **transformations of** y, such as taking a logarithm, might improve it.

- Shift values first, then take logarithm to avoid log of a negative number
- Other transformations are possible (e.g. power transform y^c or y^{-c})
- The linear regression just needs to be linear in parameters (β's)
- We can do anything to x and/or y to capture non-linear patterns

Assessing model fit for prediction

Briefly for now (we will cover in more detail in later weeks):

- Use a separate dataset where you can measure the accuracy of your predictions.
- This is known as a **test dataset** or **validation dataset**. (The data used to fit the model is known as the **training dataset**.)
- If possible, use data that is collected differently, separately, or in a different context to the data used for fitting the model. This is known as **external validation** and will show how **generalisable** the model is.
- If you don't have a separate dataset, can split your own data into two parts, or use **cross-validation**.

Further reading

PIs and CIs:

- The difference between prediction intervals and confidence intervals (blog post)

Goals of statistical modelling (somewhat technical):

- Shmueli (2010), To Explain or to Predict?
- Hernán et al. (2019), A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks.