Understanding Data and Statistical Design (60117)

Chapter 5

Simple linear regression I

Subject Coordinator: Stephen Woodcock
Lecture notes: Scott Alexander

School of Mathematical and Physical Sciences, UTS

Autumn 2024

Chapter outline

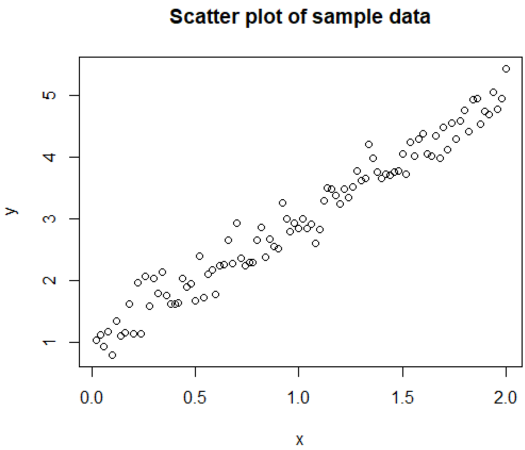
Topics:

- fitting lines to data
  - introductory example
  - model setup
  - method of least squares
  - introductory example
  - data transformations
- regression model
  - assumptions
  - properties of estimators
- coefficient T-test
  - running the test
  - two tail example
- using the model
  - prediction interval of $\mathbb{E}[Y]$
  - prediction interval of Y
- human calculator example
- R example

See Chapters 1 and 2 of Draper and Smith (1998).

Fitting lines to data – introductory example

As an introductory example, consider a sample of data consisting of 100 observations of (x, y) pairs, with a plot of this shown below.



Fitting lines to data – model setup

Suppose we wish to fit a model to this sample data – what sort of model should we choose?

To answer this question we need to decide what the nature of the relationship is between x and y.

If we were to draw a curve that "best fit" this data, what would the curve look like?

The plot suggests a straight line, as there is no sign of curvature, be it positive or negative.

But a straight line does not completely describe the data – there appears to be a **linear relationship** between x and y, but one that is disturbed by some **noise** in the data.

Fitting lines to data – model setup

The plot suggests the relationship between **predictor** x and **response** y could be described as

$$y = \beta_0 + \beta_1 x + \epsilon.$$

This is the equation of a straight line with **intercept** β_0 and **slope** β_1, disturbed by an observation of some RV ϵ which we call the **noise** or **error** term.

Our sample is one of many possible samples and we suppose it has been drawn from a population described by

$$Y = \beta_0 + \beta_1 x + \epsilon. \tag{1}$$

Notation

- β_0 and β_1 are unknown constants
- x is non-random
- y is an observation of the RV Y
- ϵ is an observation if used in the context of y and a RV if used in the context of Y

Fitting lines to data – model setup

Using the sample data, we calculate **estimates** $\hat{\beta}_0$ and $\hat{\beta}_1$ which define the **fitted regression model**

mean the estimation of the the sample

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Our sample is one of many, so we suppose our fitted regression model is an observation of the **population regression model**

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Assuming $\mathbb{E}[\epsilon] = 0$, the fitted model is used to **estimate** the **mean** or

$$\mathbb{E}[Y] = \beta_0 + \beta_1 x.$$

Notation

- $\hat{y}$ is an observation of the RV $\hat{Y}$
- $\hat{\beta}_0$ and $\hat{\beta}_1$ are observations if used in the context of $\hat{y}$ and RVs if used in the context of $\hat{Y}$

Fitting lines to data – model setup

The **statistical model** for simple linear regression is defined via the random sample

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

for $i \in \{1, \ldots, n\}$ where

- Y_i is the i-th response
- β_0 is the intercept coefficient
- β_1 is the slope coefficient
- x_i is the i-th predictor value
- ϵ_i is the i-th noise or error term.

The **sample data** is an observation of the random sample

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

for $i \in \{1, \ldots, n\}$.

Fitting lines to data – method of least squares

How should we calculate the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ so that our fitted regression model provides the **line of best fit** to the sample data?

There are many methods that are suited to this situation, but the one that is most widely used is the **method of least squares**.

This method provides estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ that are known in **closed form**, which means they can be **calculated exactly**.

It turns out that estimates calculated in this way are equal to those calculated using **maximum likelihood estimation**, another common statistical technique.

Fitting lines to data – method of least squares

The method of least squares is based on the idea of finding **estimators** $\hat{\beta}_0$ and $\hat{\beta}_1$ such that **residual**

$$\begin{aligned}\hat{\epsilon} &= Y - \hat{Y} \\ &= Y - \hat{\beta}_0 - \hat{\beta}_1 x\end{aligned}$$

is minimised in some way.

This residual RV $\hat{\epsilon}$ is an **estimator** of the noise RV

$$\begin{aligned}\epsilon &= Y - \mathbb{E}[Y] \\ &= Y - \beta_0 - \beta_1 x\end{aligned}$$

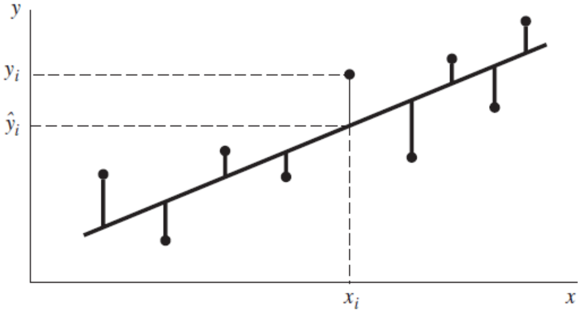
assumed in the population model from which the sample data is drawn.

Fitting lines to data – method of least squares

For a given sample of data, the residual $\hat{\epsilon}_i$ associated with the i-th data point (x_i, y_i) is the vertical distance between the observation y_i and the estimate $\hat{y}_i$ determined by the regression line, i.e.

$$\hat{\epsilon}_i = y_i - \hat{y}_i.$$

An example is shown below.



Regression line and residual. Source: Wackerly et al. (2008) page 569

Fitting lines to data – method of least squares

The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are the values that β_0 and β_1 would take to minimise the **sum square error**

$$sse(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n} \epsilon_i^2.$$

We can write this problem mathematically as

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{(\beta_0, \beta_1)}{\operatorname{argmin}}\, sse(\beta_0, \beta_1)$$

and solve using techniques of calculus.

Fitting lines to data – method of least squares

In practice we don't need to do this ourselves, as R will perform all calculations for us.

However, we will outline the solution for those interested (ignore if not).

Through differentiation we define the **normal equations**

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \big|_{\beta_0 = \hat{\beta}_0} = 0$$

and

$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \big|_{\beta_1 = \hat{\beta}_1} = 0.$$

Fitting lines to data – method of least squares

After performing the differentiation, the normal equations become

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0$$

and

$$\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0.$$

The solution of these is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{2}$$

and

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}, \tag{3}$$

where $\overline{x}$ and $\overline{y}$ are sample means of the x_i and y_i data respectively.

Fitting lines to data – method of least squares

It is common for the notation

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}}$$

to be used where

$$s_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \begin{cases} < 0 \to -\beta_1 \\ > 0 \to +\beta_1 \end{cases}$$

and

$$s_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Fitting lines to data – method of least squares

There remains one other population parameter to find an estimator for, the variance σ^2 of the noise RV ϵ.

It turns out that an **unbiased** estimate for σ^2 is given by

$$s^2 = \frac{sse(\hat{\beta}_0, \hat{\beta}_1)}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2} = \frac{\sum_{i=1}^{n}\hat{\epsilon}_i^2}{n-2}.$$

df

how many pieces of independed variable can vary, how many parameters we estimate

The least squares problem is solved.

Fitting lines to data – introductory example

Returning to the introductory example, below is the output produced by R when fitting the model (see R code file on Canvas).

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

Residuals:
     Min       1Q   Median       3Q      Max
-0.53381 -0.17184 -0.00689  0.14039  0.55751

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.99466    0.04784   20.79   <2e-16 ***
x            1.98904    0.04112   48.37   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2374 on 98 degrees of freedom
Multiple R-squared:  0.9598,    Adjusted R-squared:  0.9594
F-statistic:  2339 on 1 and 98 DF,  p-value: < 2.2e-16
```

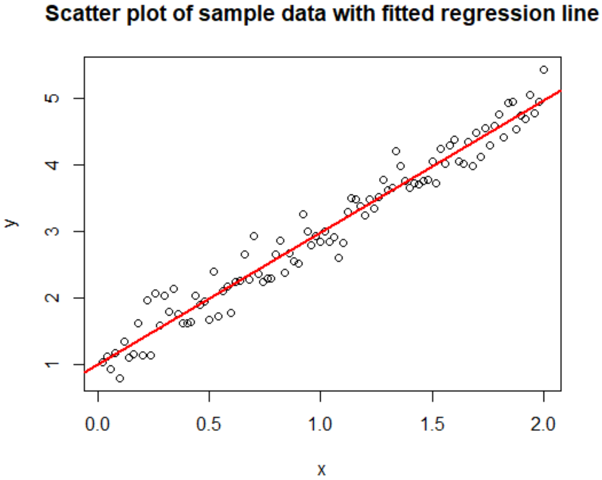
The fitted regression line equation is

$$\hat{y} = 0.99466 + 1.98904x$$

and $s = 0.2374$, which is the model's estimate of σ.

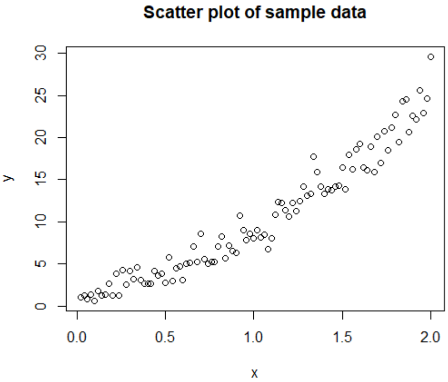
Fitting lines to data – introductory example

The following plot shows the least squares regression line fitted to the sample data from the introductory example.



Fitting lines to data – data transformations

What if the data is not linear but displays a squared relationship?



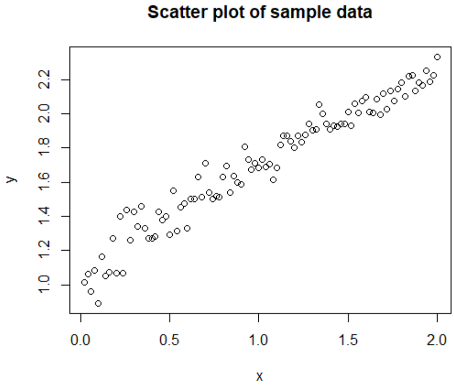
In this case we can attempt to fit the model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^2 \quad \text{or} \quad \sqrt{\hat{y}} = \hat{\beta}_0 + \hat{\beta}_1 x$$

using the first alternative if the y_i sample data takes negative values.

Fitting lines to data – data transformations

Here is another example showing a square root relationship.



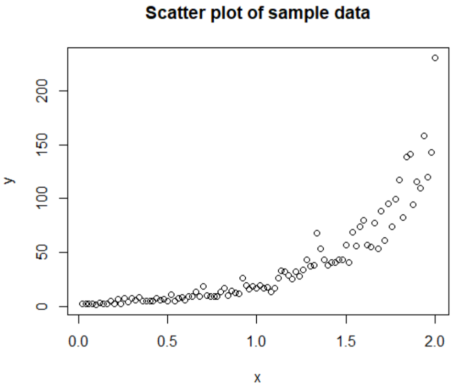
In this case we can attempt to fit the model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \sqrt{x} \quad \text{or} \quad \hat{y}^2 = \hat{\beta}_0 + \hat{\beta}_1 x$$

using the second alternative if the x_i sample data takes negative values.

Fitting lines to data – data transformations

Another example showing an exponential relationship.



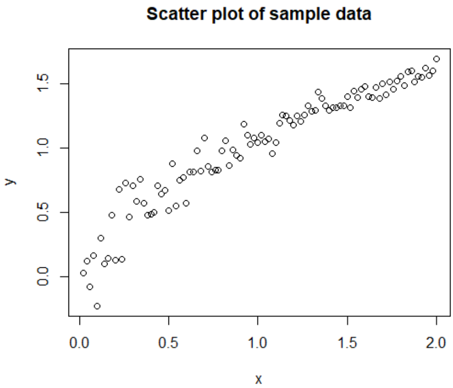
In this case we can attempt to fit the model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 e^x \quad \text{or} \quad \log(\hat{y}) = \hat{\beta}_0 + \hat{\beta}_1 x$$

using the first alternative if the y_i sample data takes negative values.

Fitting lines to data – data transformations

Another example showing a log relationship.



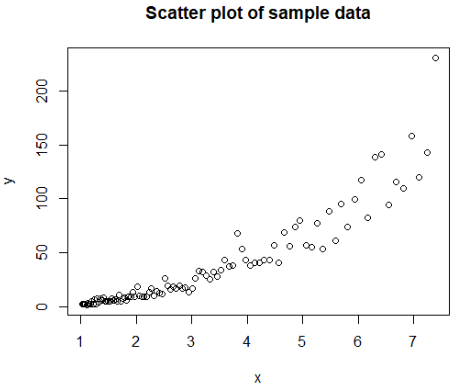
In this case we can attempt to fit the model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \log(x) \quad \text{or} \quad e^{\hat{y}} = \hat{\beta}_0 + \hat{\beta}_1 x$$

using the second alternative if the x_i sample data take negative values.

Fitting lines to data – data transformations

Sometimes we need to transform both variables.



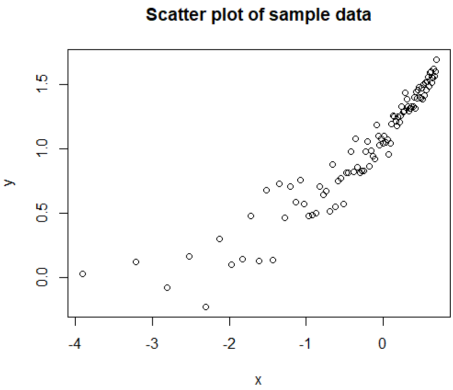
In this case we can attempt to fit the model

$$\log(\hat{y}) = \hat{\beta}_0 + \hat{\beta}_1 \log(x)$$

watching out if x_i or y_i sample data take negative values.

Fitting lines to data – data transformations

A final example.



In this case we can attempt to fit the model

$$e^{\hat{y}} = \hat{\beta}_0 + \hat{\beta}_1 e^{x}.$$

Regression model – assumptions

Although fitting the regression model using least squares requires no assumptions about the nature of the data, to go further and develop tools to analyse the fitted model does.

We make the assumptions:

- $\epsilon_i \sim N(0, \sigma)$, i.e. normally distributed with $\mathbb{E}[\epsilon_i] = 0$ and $\text{var}(\epsilon_i) = \sigma^2$
- ϵ_i are all independent from each other.

These assumptions can be re-stated in terms of the response variable as $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma)$ and independent.

In summary, the assumptions are:

1. normality
2. constant variance
3. independence.

Regression model – properties of estimators

We have the **estimators**

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x} \quad \text{and} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Under the assumptions, it can be shown that

$$\hat{\beta}_0 \sim N(\beta_0, \sigma_{\hat{\beta}_0}) \quad \text{and} \quad \hat{\beta}_1 \sim N(\beta_1, \sigma_{\hat{\beta}_1}) \tag{4}$$

where

$$\sigma^2_{\hat{\beta}_0} = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{s_{xx}} \right) \quad \text{and} \quad \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{s_{xx}}.$$

Both of these **estimators are unbiased**, which means

$$\mathbb{E}[\hat{\beta}_0] = \beta_0 \quad \text{and} \quad \mathbb{E}[\hat{\beta}_1] = \beta_1.$$

That is, the **means of the estimators equal what they are estimating**.

Regression model – properties of estimators

In practice we will never know the value of σ^2.

In its place we use the unbiased estimator

$$S^2 = \frac{SSE(\hat{\beta}_0, \hat{\beta}_1)}{n-2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2} = \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{n-2}.$$

Unbiased estimators of $\sigma^2_{\hat{\beta}_0}$ and $\sigma^2_{\hat{\beta}_1}$ are

$$S^2_{\hat{\beta}_0} = S^2\left(\frac{1}{n} + \frac{\overline{x}^2}{s_{xx}}\right) \quad \text{and} \quad S^2_{\hat{\beta}_1} = \frac{S^2}{s_{xx}}$$

respectively.

Regression model – properties of estimators

We can now define the RVs

$$T_{\hat{\beta}_0} = \frac{\hat{\beta}_0 - \beta_0}{S_{\hat{\beta}_0}} \quad \text{and} \quad T_{\hat{\beta}_1} = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}}. \tag{5}$$

Under the assumptions, these are both Student's T-distributed with $n - 2$ degrees of freedom.

We will use them as **test statistics** in ***T*-tests** to test hypothesised values of the unknown parameters β_0 and β_1.

Coefficient T-test – running the test

Hypotheses

The **null hypothesis** for this test is

$$H_0\colon\ \beta_j = \beta_j^*, \quad j \in \{0, 1\},$$

while the **alternative hypothesis** may be any of

$$H_A\colon\ \beta_j < \beta_j^* \text{ (lower tail test)}$$
$$H_A\colon\ \beta_j \neq \beta_j^* \text{ (two tail test)}$$
$$H_A\colon\ \beta_j > \beta_j^* \text{ (upper tail test)}$$

where β_j^* is the hypothesised value of β_j.

Test statistic

The test statistic is calculated from the sample data as

$$t^*_{\hat{\beta}_j} = \frac{\hat{\beta}_j - \beta_j^*}{s_{\hat{\beta}_j}}.$$

Under H_0, $t^*_{\hat{\beta}_j}$ is an observation of the appropriate $T(n-2)$ RV in (5).

Coefficient T-test – running the test

Test decision – lower tail test
H_0 is rejected at significance level $0 < \alpha < 1$ if

$$t^*_{\hat{\beta}_j} < t_\alpha,$$

where the quantile t_α is from $T(n-2)$ distribution.

Equivalently, H_0 is rejected if β_j^* falls outside the $100(1-\alpha)\%$ **lower tail confidence interval (CI)** for μ given by

$$-\infty < \beta_j \leq \hat{\beta}_j + s_{\hat{\beta}_j} t_{1-\alpha}$$

or if the p-value

$$p = \text{Prob}(T < t^*_{\hat{\beta}_j}) < \alpha$$

where $T \sim T(n-2)$.

The null hypothesis H_0 is retained in any other case.

Coefficient T-test – running the test

Test decision – two tail test
H_0 is rejected at significance level $0 < \alpha < 1$ if

$$|t^*_{\hat{\beta}_j}| > t_{1-\alpha/2},$$

where the quantile $t_{1-\alpha/2}$ is from $T(n-2)$ distribution.

Equivalently, H_0 is rejected if β^*_j falls outside the $100(1-\alpha)\%$ **two tail CI** for μ given by

$$\hat{\beta}_j - s_{\hat{\beta}_j} t_{1-\alpha/2} \leq \beta_j \leq \hat{\beta}_j + s_{\hat{\beta}_j} t_{1-\alpha/2}$$

or if the p-value

$$p = 2 \times \text{Prob}(T > |t^*_{\hat{\beta}_j}|) < \alpha$$

where $T \sim T(n-2)$.

The null hypothesis H_0 is retained in any other case.

Coefficient T-test – running the test

Test decision – upper tail test
H_0 is rejected at significance level $0 < \alpha < 1$ if

$$t^*_{\hat{\beta}_j} > t_{1-\alpha},$$

where the quantile $t_{1-\alpha}$ is from $T(n-2)$ distribution.

Equivalently, H_0 is rejected if β^*_j falls outside the $100(1-\alpha)\%$ **upper tail CI** for μ given by

$$\hat{\beta}_j - s_{\hat{\beta}_j} t_{1-\alpha} \leq \beta_j < \infty$$

or if the p-value

$$p = \text{Prob}(T > t^*_{\hat{\beta}_j}) < \alpha$$

where $T \sim T(n-2)$.

The null hypothesis H_0 is retained in any other case.

Coefficient T-test – two tail example

As part of its output, R provides details for two tail T-tests with alternative hypotheses $\beta_0 \neq 0$ and $\beta_1 \neq 0$.

The test of $\beta_1 \neq 0$ is particularly important because if this conclusion cannot be drawn, then the population model is

$$Y = \beta_0 + \epsilon,$$

i.e a constant plus noise with the predictor x not even appearing.

If we can conclude that $\beta_1 \neq 0$ then we have shown that a significant relationship exists between Y and x and can claim that the **fitted regression model is significant**.

Let's document this test for our introductory example.

Coefficient T-test – two tail example

Hypotheses

$$H_0\colon\ \beta_1 = 0$$
$$H_A\colon\ \beta_1 \neq 0$$

R output

Below is the output produced by R including two tail CIs for the coefficients (see accompanying R code file).

$$\frac{Est - H_0}{s}$$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.99466    0.04784   20.79   <2e-16 ***
x            1.98904    0.04112   48.37   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2374 on 98 degrees of freedom
Multiple R-squared:  0.9598,    Adjusted R-squared:  0.9594
F-statistic:  2339 on 1 and 98 DF,  p-value: < 2.2e-16

                2.5 %   97.5 %
(Intercept) 0.8997179 1.089595
x           1.9074331 2.070648
```

Coefficient T-test – two tail example

Test decision – using rejection region
The rejection region is defined by the 0.975 quantile from $T(98)$ distribution, which R calculates as

$$t_{0.975} = 1.984467.$$

The absolute value of the test statistic reported by R

$$|t^*_{\hat{\beta}_j}| = 48.37 > 1.984467 = t_{0.975}.$$

Accordingly, the null hypothesis H_0 is rejected at significance level $\alpha = 0.05$.

Coefficient T-test – two tail example

Test decision – using CI
With the hypothesised value $\beta_1 = 0$ outside the 95% two tail CI, reported by R as $[1.9074331, 2.070648]$, H_0 is rejected at significance level $\alpha = 0.05$.

Test decision – using p-value
With the reported p-value satisfying

$$p < 2 \times 10^{-16} < 0.05 = \alpha,$$

H_0 is rejected at significance level $\alpha = 0.05$.

Conclusion
The regression is significant (there is a significant relationship between predictor and response).

Using the model

Once we have a significant regression, we can use it for prediction.

Suppose x^* is a new value of the predictor that was not in the original sample data to which the model was fitted.

We want to be able to predict the response Y^* to this new value x^*, which are related according to our population model

$$Y^* = \beta_0 + \beta_1 x^* + \epsilon.$$

Using our fitted model, we can estimate this as

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*.$$

We can calculate **confidence intervals for such predictions**.

Using the model – prediction interval of $\mathbb{E}[Y]$

First we place bounds on $\mathbb{E}[Y^*]$.

The fitted model's $100(1-\alpha)\%$ **prediction interval** for $\mathbb{E}[Y^*]$ can be calculated as

$$\hat{y}(x^*) \pm t_{1-\alpha/2} \times s\sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{s_{xx}}}$$

where the quantile $t_{1-\alpha/2}$ is from Students' T-distribution with $n-2$ degrees of freedom.

This is also referred to as the **mean prediction interval**.

Note that this is just a confidence interval by another name.

Using the model – prediction interval of Y

We can also place bounds on Y^*.

The fitted model's $100(1-\alpha)\%$ **prediction interval** for Y^* can be calculated as as

$$\hat{y}(x^*) \pm t_{1-\alpha/2} \times s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{s_{xx}}}$$

Interval bigger, estimate less precise

where the quantile $t_{1-\alpha/2}$ is from Students' T-distribution with $n-2$ degrees of freedom.

This is also referred to as the **individual prediction interval**.

Note that this is just a confidence interval by another name.

Human calculator example

Although in practice R will make most calculations for us, let's illustrate one example by preforming the calculations by hand.

Consider the following data recording the age (x_i) and blood pressure (y_i) of four individuals, with sample data displayed below.

i	x_i	y_i
1	39	144
2	47	220
3	45	138
4	47	145

We are going to build a model that allows us to predict blood pressure from age.

The independent variable in this case represents age and the dependent variable represents blood pressure. (Why not the other way around?)

Human calculator example

Our first step would normally be to plot the data, but with only four data points there isn't much to see.

We suppose that the true population relationship between age and blood pressure is

$$Y = \beta_0 + \beta_1 x + \epsilon$$

and look to build the fitted regression model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

to estimate

$$\mathbb{E}[Y] = \beta_0 + \beta_1 x.$$

Human calculator example

First we calculate the sample average of the x_i data (average age)

$$\overline{x} = \frac{39 + 47 + 45 + 47}{4} = 44.5$$

and of the y_i data (average blood pressure)

$$\overline{y} = \frac{144 + 220 + 138 + 145}{4} = 161.75.$$

Theses sample averages are then used to construct the following table.

i	x_i	y_i	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})^2$	$(x_i - \overline{x})(y_i - \overline{y})$
1	39	144	-5.5	-17.75	30.25	97.625
2	47	220	2.5	58.25	6.25	145.625
3	45	138	0.5	-23.75	0.25	-11.875
4	47	145	2.5	-16.75	6.25	-41.875
					43.00	189.500

Human calculator example

From this table we can read off the figures $s_{xx} = 43$ and $s_{xy} = 189.5$.

From (2) we have

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}} = \frac{189.5}{43} \approx 4.41$$

and from (3)

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 161.75 - 4.41 \times 44.5 \approx -34.36.$$

So our least squares model fitted to the sample data is

$$\hat{y} = -34.36 + 4.41x.$$

Obviously, this is not a terribly sophisticated model – for one, it predicts negative blood pressure up until 7.8 years of age.

Be careful extrapolating model outside range of sample data.

Human calculator example

We can now calculate the regression predictions on the sample data and associated residuals, with the results displayed below.

i	x_i	y_i	$\hat{y}_i$	$\hat{\epsilon}_i$
1	39	144	137.63	6.37
2	47	220	172.91	47.09
3	45	138	164.09	-26.09
4	47	145	172.91	-27.91

Obviously, with a sample of only four observations we can't expect much from the fitted model.

R example

Now an example using R.

We are going to build a model that lets us predict average life expectancy from per capita gross national income.

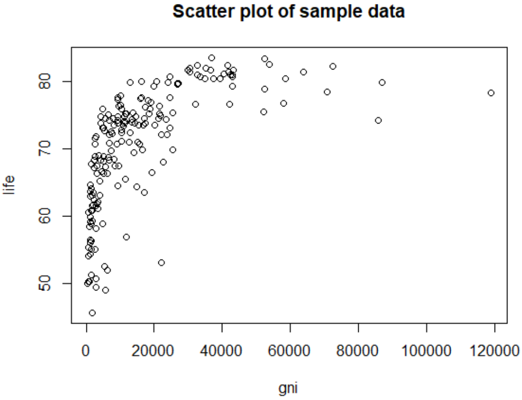
Our sample data contains observations for $n = 187$ countries, with the variables summarised in the table below (data in life.data.csv on Canvas).

Name	**Type**	**Description**
life	response	life expectancy (years)
gni	predictor	per capita gross national income (USD)

We will fit a simple linear regression model, so the first thing we do is see if a linear relationship can be found.

R example

Below is a scatter plot of the sample data $(gni_i, life_i)$.



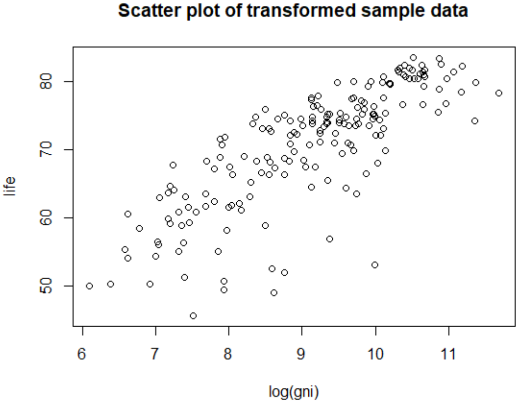
We see that no linear relationship is apparent.

R example

Now we consider the transformed predictor

$$gniLog = \log(gni)$$

and create a scatter plot of the transformed sample data $(gniLog_i, life_i)$.



Now we now see a reasonable linear relationship.

R example

So we assume the population has the form

$$LIFE = \beta_0 + \beta_1 \times gniLog + \epsilon$$

and look to build the fitted regression model

$$\widehat{life} = \hat{\beta}_0 + \hat{\beta}_1 \times gniLog$$

to estimate

$$\mathbb{E}[LIFE] = \beta_0 + \beta_1 \times gniLog.$$

R example

Using R we obtain the following summary of the fitted model.

```
Call:
lm(formula = life ~ gniLog, data = life.data)

Residuals:
    Min      1Q  Median      3Q     Max
-22.647  -2.267   1.020   3.354   8.938

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.2798     2.9577   5.842 2.28e-08 ***
gniLog        5.8482     0.3217  18.177  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.306 on 185 degrees of freedom
Multiple R-squared:  0.6411,    Adjusted R-squared:  0.6391
F-statistic: 330.4 on 1 and 185 DF,  p-value: < 2.2e-16
```

We we can also obtain confidence intervals on the parameters.

```
                2.5 %    97.5 %
(Intercept) 11.444555 23.114985
gniLog       5.213424  6.482883
```

The least squares parameter estimates are $\hat{\beta}_0 = 17.2798$ and $\hat{\beta}_1 = 5.8482$, resulting in the fitted model

$$\widehat{life} = 17.2798 + 5.8482 \times gniLog.$$

R example

As part of the output above, R reports the p-values associated with T-tests on the parameters β_0 and β_1.

The hypotheses for the tests on β_0 are

$$H_0\colon\ \beta_0 = 0$$
$$H_A\colon\ \beta_0 \neq 0$$

and for β_1 are

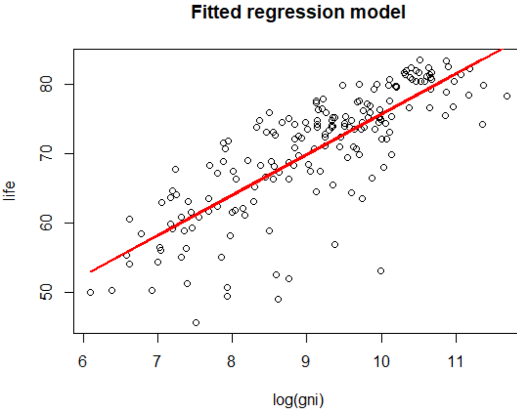
$$H_0\colon\ \beta_1 = 0$$
$$H_A\colon\ \beta_1 \neq 0.$$

The p-values associated with these tests are both well below our usual significance level of $\alpha = 0.05$.

So both null hypotheses can be rejected and we conclude that both β_0 and β_1 are different from zero. We can also see this from the absence of zero in the parameter CIs.

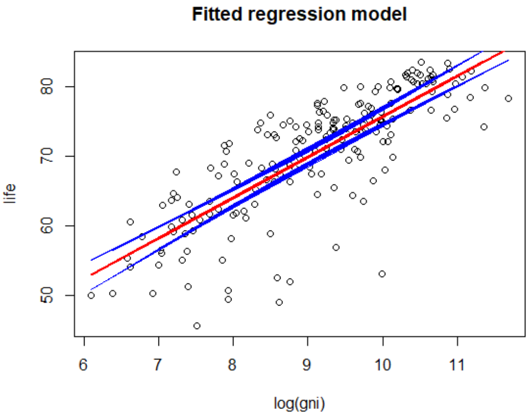
R example

Below is a plot of the fitted regression model against the sample data ...



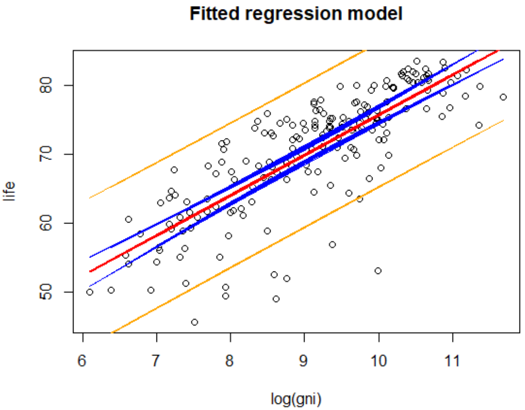
R example

... to which we can add the 95% prediction interval for $\mathbb{E}[LIFE]$ ...



R example

... to which we can add the 95% prediction interval for *LIFE*.



R example

Of course, there are other questions to answer.

Are the assumptions, on which the statistical tests are built, valid?

How well does the model fit the data?

Is there some non-linear component that can be captured by adding some function of *gniLog* as a new variable to the model?

Are there variables other than *gniLog* that we should consider adding to the model?

We will show how to go about answering these sort of questions in following chapters.

References I

Draper, N. R. and Smith, H. (1998). *Applied regression analysis*. Wiley-Interscience, Somerset, US.

Wackerly, D., Mendenhall, W., and Scheaffer, R. L. (2008). *Mathematical Statistics with Applications*. Thomson Brooks/Cole, Belmont, CA, 7 edition edition.