Regression and Linear Models (37252)

Lecture 2 - Simple Linear Regression I

Lecturer: Joanna Wang
Notes adopted from Scott Alexander

School of Mathematical and Physical Sciences, UTS

Autumn 2024

Acknowledgement

These notes are based, in part, on earlier versions prepared by Dr Ed Lidums and Prof. James Brown.

2. Lecture outline

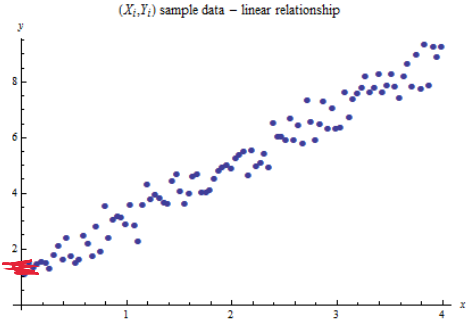
Topics:

- fitting lines to data
  - model setup
  - method of least squares
  - data transformations
- model assumptions
- model parameters and estimates
- statistical properties of estimates
  - distributions
  - T-tests
- statistical properties of model
  - prediction of $\mathbb{E}[Y|x]$
  - prediction of $Y|x$
- human calculator example
- R example

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

2. Fitting lines to data

Consider the problem of constructing a model describing the **sample** data (X_i, Y_i), $i \in \{1, \ldots, n\}$, displayed in the following scatter plot.



$y = mx + b$

In this toy example, the sample data has been generated using the rule

$$Y_i = 1 + 2X_i + \epsilon_i$$

with $n = 100$, $X_i = \frac{4}{n} i$ and $\epsilon_i \sim \mathrm{N}(0, \frac{1}{2})$.

2. Fitting lines to data – model setup

You immediately notice a strong, although imperfect, **linear** relationship in the data and wonder whether the underlying **population** might be described by

direction positive

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

which is the equation of a straight line with intercept β_0, slope β_1 that is disturbed by some RV ϵ. This of course makes Y a random variable also.

Assuming that $\mathbb{E}[\epsilon] = 0$, you decide to try and fit the model

estimate

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

$\bar{x}$ to estimate μ

by finding estimates $\hat{\beta}_0 \approx \beta_0$ and $\hat{\beta}_1 \approx \beta_1$.

The quality of the resulting model

$$\begin{aligned} \hat{Y}|x &:= \hat{\beta}_0 + \hat{\beta}_1 x \\ &\approx \mathbb{E}[Y|x] \end{aligned} \tag{2}$$

will be determined by the quality of the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

2. Fitting lines to data – method of least squares

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

This is because the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are known in **closed-form**, which means they can be calculated **exactly**.

Later in the course we will study another technique in regression ("logistic" regression) where the model parameters can only be approximated.

2. Fitting lines to data – method of least squares

The method of least squares is based on the idea of finding the estimated coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ such that error in the approximation

obs. – estimated

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

is minimised in some way.

This error term $\hat{e}$ is an estimate of the RV

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

$\hat{Y}_i$

assumed in the population model underlying the sample data.

The least squares model is the model that minimises the **sum squared errors** over the sample data; i.e.

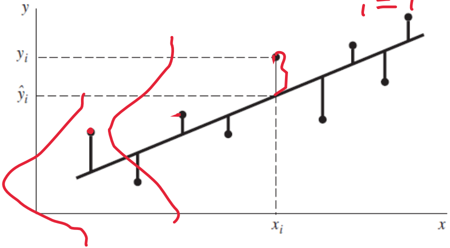
$Y_i - (\beta_0 + \beta_1 X_i)$

$$\begin{aligned}\min_{(\beta_0,\beta_1)} SSE(\beta_0,\beta_1) &= \min_{(\beta_0,\beta_1)} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \\ &= \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2. \qquad (3)\end{aligned}$$

2. Fitting lines to data – method of least squares

The residual $\hat{\epsilon}_i$ associated with the i-th data point (X_i, Y_i) is the distance between the sample data value Y_i and the value $\hat{Y}_i$ determined by the regression model line.

$$\sum_{i=1}^{9} (Y_i - \hat{y}_i)^2$$



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

2. Fitting lines to data – method of least squares

Assuming that the minimum of SSE exists, we can find $\hat{\beta}_0$ and $\hat{\beta}_1$ by differentiating SSE with respect to both β_0 and β_1, setting the equations to zero and solving.

That is, $\hat{\beta}_0$ will be found as the value such that

$$\frac{\partial}{\partial \beta_0} SSE(\beta_0, \beta_1)\Big|_{\beta_0 = \hat{\beta}_0} = 0 \tag{4}$$

and $\hat{\beta}_1$ such that

$$\frac{\partial}{\partial \beta_1} SSE(\beta_0, \beta_1)\Big|_{\beta_1 = \hat{\beta}_1} = 0. \tag{5}$$

The resulting regression line

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

is the straight line that minimises total SSE associated with the sample data.

2. Fitting lines to data – method of least squares

After performing the differentiation, the **least squares equations** (4) and (5) become

$$\partial/\partial\beta_0 = \quad \sum_{i=1}^{n} Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} X_i = 0 \tag{6}$$

and

$$\partial/\partial\beta_1 = \quad \sum_{i=1}^{n} X_i Y_i - \hat{\beta}_0 \sum_{i=1}^{n} X_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 = 0 \tag{7}$$

respectively, which after solving provide the least squares coefficients

$$\mathrm{var}(X) = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \qquad \frac{S_{xy}}{S_{xx}}$$

and

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}.$$

2. Fitting lines to data – method of least squares

These are often re-expressed as

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} \tag{8}$$

and

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \tag{9}$$

where

$$S_{XY} = \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}) \tag{10}$$

and

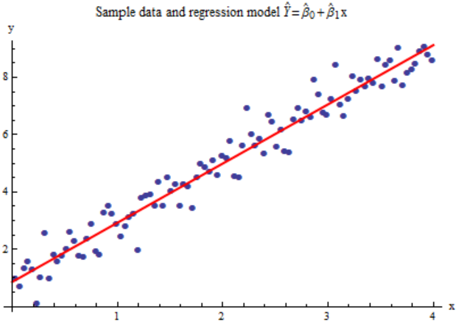
$$S_{XX} = \sum_{i=1}^{n} (X_i - \overline{X})^2. \tag{11}$$

The least squares problem is solved.

2. Fitting lines to data – method of least squares

The following graph shows the least squares line fitted to the sample data.

$y = 1 + 2x$

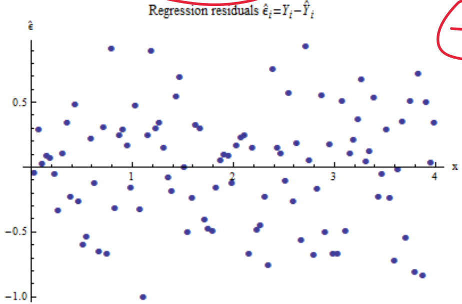


The least squares coefficients are $\hat{\beta}_0 = 1.07678$ and $\hat{\beta}_1 = 1.96832$.

2. Fitting lines to data – method of least squares

Of primary importance are the residuals, displayed in the plot below.

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

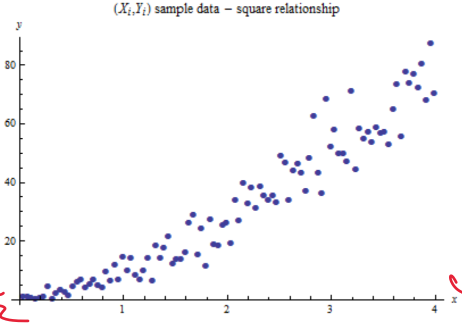


In building regression models, much of our effort will be spent analyzing the residuals for compliance with assumptions (more soon).

2. Fitting lines to data – data transformations

What if the data does not seem linear?

Y



X^2 $\sqrt{Y}$ X

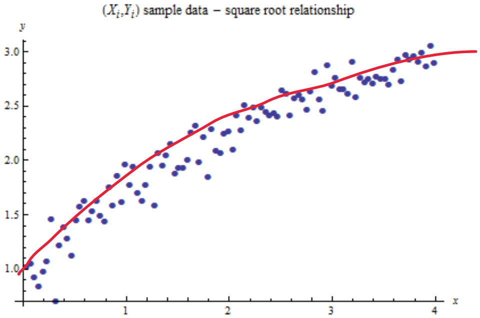
Scatter plot showing square relationship

In this case we can attempt to fit the (linear) 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 take negative values.

Here is another example showing a square root relationship.



Scatter plot showing square root relationship

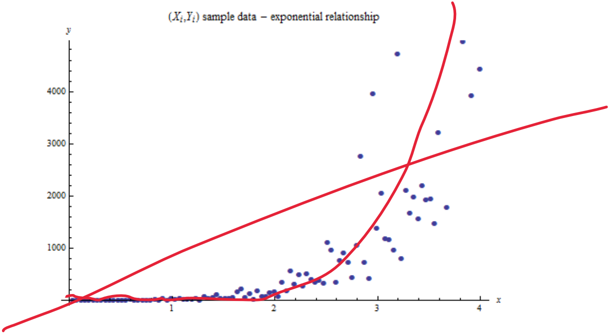
In this case we can attempt to fit the model

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

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

2. Fitting lines to data – data transformations

Another example showing an exponential relationship.



Scatter plot showing exponential relationship

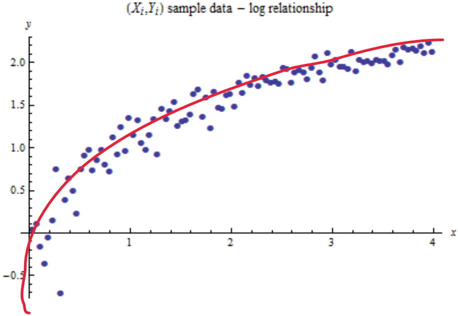
In this case we can attempt to fit the model

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

using the first alternative if the Y_i sample data take negative values.

2. Fitting lines to data – data transformations

Another example showing a log relationship.



Scatter plot showing log relationship

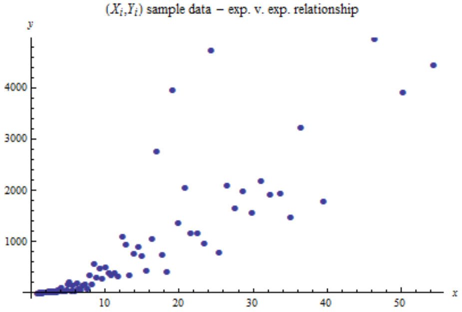
In this case we can attempt to fit the model

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

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

2. Fitting lines to data – data transformations

Sometimes we need to transform both X_i and Y_i.



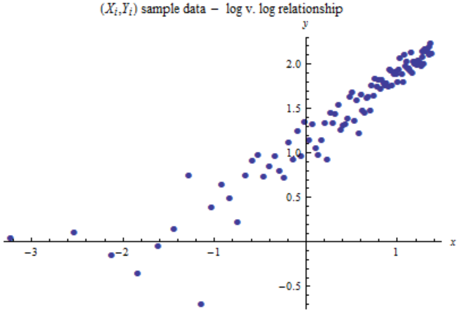
Scatter plot showing exp. v. exp. relationship

In this case we can attempt to fit the model

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

watching for the possibility that X_i or the Y_i sample data take negative values.

A final example.



Scatter plot showing log v. log relationship

In this case we can attempt to fit the model

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

2. Model assumptions

Although we have solved the least squares problem and found the regression line, we still have many questions to answer.

To answer these questions requires developing certain statistical tools, the validity of which depend on the following assumptions:

1. the underlying population model is given

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

2. the RV ϵ has $\mathbb{E}[\epsilon] = 0$ and $\mathrm{var}(\epsilon) = \sigma^2$ for all x
3. the sample data errors $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ and independent so that $Y_i|X_i \sim \mathrm{N}(\beta_0 + \beta_1 X_i, \sigma^2)$ and independent.

constant variance assumption.

normality assumption

independence assumption.

i.i.d

2. Model parameters and estimates

The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of the population parameters β_0 and β_1 are themselves RVs – construct a least squares model on a different set of sample data and the estimates will change.

Under the assumptions just stated, these estimates possess some nice statistical properties:

1. by the Gauss-Markov theorem, these are the **minimum variance linear unbiased estimators** of the population parameters β_0 and β_1.
2. they are the **maximum likelihood estimators (MLE)**.

The first property means that $\mathbb{E}[\hat{\beta}_0] = \beta_0$ and $\mathbb{E}[\hat{\beta}_1] = \beta_1$ (unbiased) and that $\text{var}(\hat{\beta}_0)$ and $\text{var}(\hat{\beta}_1)$ will be smaller than for any other possible estimates $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$.

The second property means, in a "hand waving" sort of way, that they are the estimates that make it most "likely" to observe the sample data (X_i, Y_i).

2. Model parameters and estimates

There remains one other population parameter to estimate, the variance σ^2 of the RV ϵ.

$\sum (Y_i - \hat{Y}_i)^2$

One candidate is $\frac{SSE(\hat{\beta}_0, \hat{\beta}_1)}{n}$, with SSE given by (3). We used SSE to derive our least squares line, and it turns out that this is the MLE of σ^2.

However it is a **biased** estimate in that $\mathbb{E}[\frac{SSE(\hat{\beta}_0, \hat{\beta}_1)}{n}] \neq \sigma^2$.

It turns out that an **unbiased** version of this estimate can be constructed by taking into account the **degrees of freedom** lost in finding the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. This estimate

$$S^2 = \frac{SSE(\hat{\beta}_0, \hat{\beta}_1)}{n-2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n}\hat{\epsilon}_i^2}{n-2} \tag{12}$$

degrees of freedom

does satisfy $\mathbb{E}[S^2] = \sigma^2$.

$Y = \beta_0 + \beta_1 X$

2. Statistical properties of estimates – distributions

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

It can be shown that $\hat{\beta}_0 \sim \mathrm{N}(\beta_0, \sigma^2_{\hat{\beta}_0})$ with

$$\sigma^2_{\hat{\beta}_0} = \sigma^2\Big(\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}\Big).$$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

It inherits its normality from the sample errors ϵ_i, which are assumed to be $N(0, \sigma^2)$. It is unbiased and so has mean β_0, which we stated before.

Similarly, $\hat{\beta}_1 \sim \mathrm{N}(\beta_1, \sigma^2_{\hat{\beta}_1})$ with

$$\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{S_{XX}}.$$

We see that both $\sigma_{\hat{\beta}_0}$ and $\sigma_{\hat{\beta}_1}$ depend on σ – if we don't know the latter we don't know the former.

2. Statistical properties of estimates – distributions

The **standardised** versions of the estimates

$$Z_{\hat{\beta}_0} = \frac{\hat{\beta}_0 - \beta_0}{\sigma_{\hat{\beta}_0}} \quad \sim N(0,1) \tag{13}$$

and

$$Z_{\hat{\beta}_1} = \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \quad \sim N(0,1) \tag{14}$$

are both $N(0,1)$, i.e. normally-distributed with zero mean and variance of one.

We can use these RVs as the basis of test statistics in Z-tests and to establish confidence intervals.

2. Statistical properties of estimates – distributions

However, in practice we will never know the value of σ.

Instead the sample standard deviation S is used as an approximation of σ.

This gives the unbiased estimates of $\sigma_{\hat{\beta}_0}$

$$S_{\hat{\beta}_0} = S\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}}. \tag{15}$$

and of $\sigma_{\hat{\beta}_1}$

$$S_{\hat{\beta}_1} = \frac{S}{\sqrt{S_{XX}}} \tag{16}$$

with σ described in (12).

2. Statistical properties of estimates – distributions

These estimates of $\sigma_{\hat{\beta}_0}$ and $\sigma_{\hat{\beta}_1}$ provide the alternative test statistics

$$T_{\hat{\beta}_0} = \frac{\hat{\beta}_0 - \beta_0}{S_{\hat{\beta}_0}} \tag{17}$$

and

$$T_{\hat{\beta}_1} = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}}, \tag{18}$$

which are both Student's T-distributed with $n-2$ degrees of freedom.

These RVs can be used as **test statistics** in **T-tests**.

$H_0 : \beta_0 = \beta_0^*$

$H_0 : \beta_1 = \beta_1^*$

T-test hypotheses

The null hypothesis is

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

with β_j^* some hypothesised level of β_j we hope to exclude.

The alternative hypothesis can be any of

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

R output: *R reports the results of T-tests with* $\beta_j^* = 0$.

$H_0 : \beta_0 = 0$

$H_0 : \beta_1 = 0$

2. Statistical properties of estimates – T-tests

Test statistic

The test statistic

$$t^*_{\hat{\beta}_j} = \frac{\hat{\beta}_j - \beta^*_j}{S_{\hat{\beta}_j}} \tag{19}$$

$t_{1-\alpha/2}$

depends on the sample data, and is a **realisation** of the appropriate RV described in (17)-(18).

Rejection of null hypothesis using reject/retain region

H_0 is rejected in favour of H_A at **significance level** α, $0 < \alpha < 1/2$, if

$$t^*_{\hat{\beta}_j} > t_{1-\alpha} \text{ (upper-tail test)}$$

$$|t^*_{\hat{\beta}_j}| > t_{1-\alpha/2} \text{ (two-tail test)}$$

$$t^*_{\hat{\beta}_j} < t_\alpha \text{ (lower-tail test)}$$

t_α

where t_θ is the **quantile** satisfying

$t_{1-\alpha}$

$$\text{Prob}(T_{\hat{\beta}_j} > t_\theta) = \theta.$$

2. Statistical properties of estimates – T-tests

$\alpha = 0.05$

2

Rejection of null hypothesis using p-value

Equivalently, H_0 is rejected if

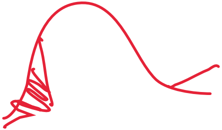
$$p < \alpha$$

where the **p-value**

$$p = \text{Prob}(T_{\hat{\beta}_j} > t^*_{\hat{\beta}_j}) \text{ (\textbf{upper-tail test})}$$

$$p = 2 \times \text{Prob}(T_{\hat{\beta}_j} > |t^*_{\hat{\beta}_j}|) \text{ (\textbf{two-tail test})}$$

$$p = \text{Prob}(T_{\hat{\beta}_j} < t^*_{\hat{\beta}_j}) \text{ (\textbf{lower-tail test})}.$$



The null hypothesis H_0 is **retained** in any other case.

~~Accept H_0~~

2. Statistical properties of estimates – T-tests

3 **Rejection of null hypothesis using CI**

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

$$\hat{\beta}_j - S_{\hat{\beta}_j} t_{1-\alpha} \leq \beta_j < \infty \text{ (\textbf{upper-tail test})}$$

$$\hat{\beta}_j - S_{\hat{\beta}_j} t_{1-\alpha/2} \leq \beta_j \leq \hat{\beta}_j + S_{\hat{\beta}_j} t_{1-\alpha/2} \text{ (\textbf{two-tail test})}$$

$$-\infty < \beta_j \leq \hat{\beta}_j + S_{\hat{\beta}_j} t_{1-\alpha} \text{ (\textbf{lower-tail test})}.$$

The null hypothesis H_0 is **retained** in any other case.

Interpretation of special case

When the null hypothesis $\beta_j = 0$ is rejected, we can say the **predictor $\boldsymbol{x_j}$ is statistically-significant** (at significance level α).

2. Statistical properties of model – prediction of $\mathbb{E}[Y|x]$

Recall we set out to model

$$\begin{aligned} \mathbb{E}[Y|x] &= \beta_0 + \beta_1 x \\ &\approx \hat{\beta}_0 + \hat{\beta}_1 x \\ &= \hat{Y}|x, \end{aligned}$$

with the last being our least squares regression model.

We know this model is unbiased as

$$\mathbb{E}[\hat{Y}|x] = \mathbb{E}[\hat{\beta}_0] + \mathbb{E}[\hat{\beta}_1]x = \beta_0 + \beta_1 x = \mathbb{E}[Y|x]$$

which follows from the unbiased nature of the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ established earlier.

But $\hat{Y}|x$ is still a RV, so we desire confidence intervals for the values it takes, or the **predictions** it makes of $\mathbb{E}[Y|x]$.

2. Statistical properties of model – prediction of $\mathbb{E}[Y|x]$

Without going into details, we can establish a $100(1-\alpha)\%$ confidence interval for the model's **prediction** of $\mathbb{E}[Y|x]$ as

$$\hat{Y}|x \pm z_{1-\alpha/2} \times \sigma \sqrt{\frac{1}{n} + \frac{(x-\overline{X})^2}{S_{XX}}} \tag{20}$$

or if σ is unknown as

$$\hat{Y}|x \pm t_{1-\alpha/2} \times S \sqrt{\frac{1}{n} + \frac{(x-\overline{X})^2}{S_{XX}}} \tag{21}$$

where the quantiles $z_{1-\alpha/2}$ and $t_{1-\alpha/2}$ are associated with the $\mathrm{N}(0,1)$ distribution and Students' T distribution with $n-2$ degrees of freedom respectively.

Again, for the values of α we are interested in, the Student's T-based CI will be wider than the $\mathrm{N}(0,1)$-based version. This is due to the fatter tails of the Student's-T distribution.

Finally, CIs for the model's **prediction** of $Y|x$ are

$$\hat{Y}|x \pm z_{1-\alpha/2} \times \sigma \sqrt{1 + \frac{1}{n} + \frac{(x - \overline{X})^2}{S_{XX}}} \tag{22}$$

or if σ is unknown

$$\hat{Y}|x \pm t_{1-\alpha/2} \times S \sqrt{1 + \frac{1}{n} + \frac{(x - \overline{X})^2}{S_{XX}}}. \tag{23}$$

Eg $BP = \beta_0 + \beta_1 age$

$\hat{BP}$ for a single person with age 60 (23)

$\hat{BP}$ for the average of a group of ppl with age 60 (21)

2. Human calculator example

OK, that's a lot of equations. However, we won't be doing the calculations by hand - OK, maybe just one example.

Consider the following data recording the age (X_i) and blood pressure (Y_i) of four individuals.

i	X_i	Y_i
1	39	144
2	47	220
3	45	138
4	47	145

Blood pressure data

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

The independent variable in this case is age (X_i) and the dependent variable is blood pressure (Y_i). (Why?)

2. Human calculator example

Our first step would normally be to plot the data in the hope of spotting a recognisable relationship (linear in the case of simple regression), but with only four data points there isn't much to see.

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

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

and look to build the model

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

to approximate

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

2. Human calculator example

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

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

$\beta_1 = \frac{Sxy}{Sxx}$

and of the Y_i data (average blood pressure)

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

Sxx Sxy

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

2. Human calculator example

From this table we can read off the figures $S_{XX} = 43$ and $S_{XY} = 189.5$.

From (8) we have

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{189.5}{43} \approx 4.41$$

and from (9)

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \approx 161.75 - 4.41 \times 44.5 \approx -34.36. \tag{24}$$

So our least squares model is

$$\hat{Y}|x = -34.36 + 4.41x$$

or in alternative notation

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

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.

$Y_i - \hat{Y}_i$

Now let's see how well this fits the data.

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

$-34.36 + 4.41 \times 39$

Prediction and residual data

Not too well, which is hardly surprising given the amount of data we had.

2. R example

Consider the data set newdata.csv, available in the Week 3 folder on Canvas.

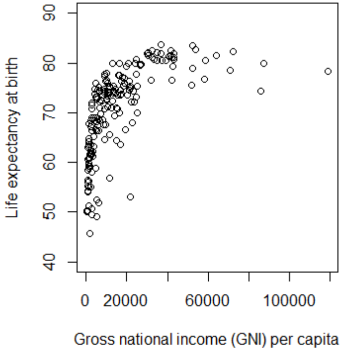
The variables of interest are *LifeExp* and *GNI*. We are going to build a model that lets us predict *LifeExp* from *GNI*.

The independent variable in this case is *GNI* (our X_i values) and the dependent variable is *LifeExp* (our Y_i values).

We would like to build a linear model, so the first thing we do is see if a linear relationship can be found.

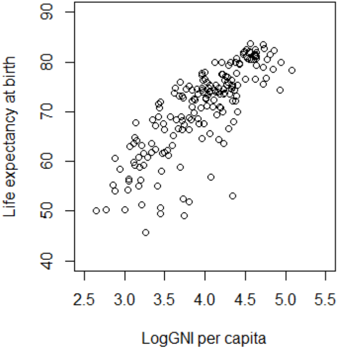
We do so with a scatter plot.

2. R example



We see that no linear relationship is apparent. We perform a log transform of *GNI*, defining the new variable *LogGNI* in the process and create a scatter plot using the new variable.

2. R example



We now see a reasonable linear relationship and decide to build our model of *LifeExp* against this new variable *LogGNI*.

2. R example

So we assume an underlying reality of

$$\mathit{LifeExp}|\mathit{LogGNI} = \beta_0 + \beta_1 \times \mathit{LogGNI} + \epsilon$$

and look to build the model

$$\widehat{\mathit{LifeExp}}|\mathit{LogGNI} = \hat{\beta}_0 + \hat{\beta}_1 \times \mathit{LogGNI}$$

as an approximation of

$$\mathbb{E}[\mathit{LifeExp}|\mathit{LogGNI}] = \beta_0 + \beta_1 \times \mathit{LogGNI}.$$

2. R example

To fit a simple linear regression model in R, we use the **lm** command.

```
> mod1<-lm(newdata$Life_exp ~ newdata$LogGNI)
> summary(mod1)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     17.2798     2.9577   5.842 2.28e-08 ***
newdata$LogGNI  13.4659     0.7408  18.177  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
```

The least squares parameter estimates are $\hat{\beta}_0 = 17.280$ and $\hat{\beta}_1 = 13.466$, resulting in the fitted model

$$\widehat{LifeExp} = 17.280 + 13.466 \times LogGNI.$$

2. R example

As part of its output 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

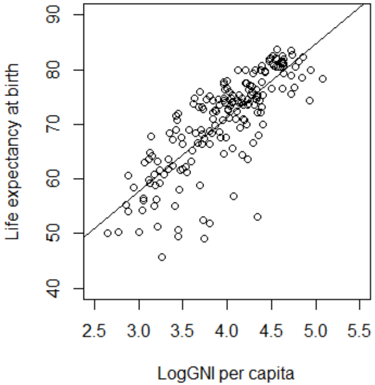
$$H_0\colon\ \beta_1 = 0$$
$$H_A\colon\ \beta_1 \neq 0.$$

The p-values associated with both of these tests are extremely small so both null hypotheses can be rejected at some significance levels $\alpha < 0.0005$.

2. R example

To aid a visual inspection, R will plot the fitted model against sample data using

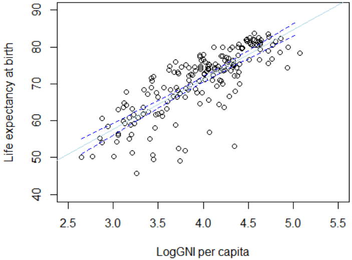
```
> abline(mod1)
```



2. R example

... and can add 95% confidence bounds for the model's prediction of $\mathbb{E}[LifeExp|LogGNI]$ ...

```
> newx <- seq(min(newdata$LogGNI), max(newdata$LogGNI), by=0.05)
> conf_interval <- predict(mod1, newdata=data.frame(LogGNI=newx),
 interval="confidence", level = 0.95)
> lines(newx, conf_interval[,2], col="blue", lty=2)
> lines(newx, conf_interval[,3], col="blue", lty=2)
```

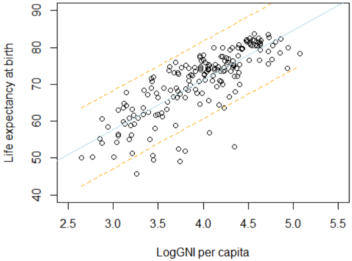


Regression Line and 95% CI for $\mathbb{E}[LifeExp|LogGNI]$

2. R example

... and can add 95% confidence bounds for the model's prediction of $LifeExp|LogGNI$.

```
pred_interval <- predict(mod1, newdata=data.frame(LogGNI=newx),
interval="prediction", level = 0.95)
lines(newx, pred_interval[,2], col="orange", lty=2)
lines(newx, pred_interval[,3], col="orange", lty=2)
```



Regression Line and 95% CI for $LifeExp|LogGNI$

2. 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 *LogGNI* as a new variable to the model?

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

We have **interpolated** within the range of the X_i sample data – can we **extrapolate** outside of this range?

We will start to answer some of these questions next week.

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.