Tutorial 6: Linear Regression

Rob Nicholls – nicholls@mrc-lmb.cam.ac.uk

MRC LMB Statistics Course 2014

Contents

1 Introduction to Simple Linear Regression 1
2 Parameter Estimation and Model Utility 3
3 Modelling Non-Linear Relationships 7
4 Multiple Regression 9
5 Choosing Between Models 11
6 Estimation of Parameters for the Simple Linear Regression Model . . 12

1 Introduction to Simple Linear Regression

It is often of interest to model linear relationships between variables, such as when wanting to know whether two variables are correlated, or when wanting to be able to predict a response given knowledge of a number of independent variables. Such questions can be addressed by considering a *parametric regression model*:

$$Y = f(X; \theta) + \varepsilon$$

in which the *response variable* Y is regressed against the *independent variables* X, given knowledge of some parameters θ and an error model ε.

The most simple form of such a model, in which there is only one independent variable, is referred to as *simple linear regression*. Such a model may be expressed:

$$Y = \alpha + \beta X + \varepsilon$$

where X is the *regressor* (also called the *predictor* or *independent variable*), Y is the *response* (also called the *dependent variable*), α and β are parameters that describe the relationship between X and Y, and the term ε represents the *error model* (the errors are also referred to as *residuals*). In simple linear regression, it is assumed that the residuals follow a Normal distribution, specifically:

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

for some parameter σ.

Such a model is useful for investigating linear relationships between variables.

It is often the case that we observe a series of n observations of the response $(y_1, \ldots, y_n)$ and predictor $(x_1, \ldots, x_n)$ variables, in which case the simple linear regression model be may expressed:

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad\qquad i = 1, \ldots, n$$

for some parameters α and β.

Here, the values of y_i and x_i are known (observed), the values of the parameters α and β are chosen, and the residuals ε_i are determined as:

$$\varepsilon_i = y_i - \alpha - \beta x_i$$

Note that we could select any values for α and β, and we would be able to select values for ε_i in order to satisfy this equation.

We want to find find values of the parameters α and β such that the errors ε_i are co-minimised. Doing so would allow α and β to be meaningful, allowing the model to be used to predict the value of the response variable Y, utilising knowledge of the regressor X.

Note that if all data points were to lie on a straight line then all errors ε_i would be zero. The errors represent random effects that we cannot account for, given the available information.

We want to use the regression model to predict the value of the response variable Y, given a particular value x of the independent variable X. Denoting the estimate of the response by $\hat{y}$, we can write:

$$\hat{y} = \hat{\alpha} + \hat{\beta} x$$

where $\hat{\alpha}$ and $\hat{\beta}$ are estimates of the parameters α and β, respectively. This essentially says that if we denote the mean of Y by $\hat{y} = E(Y|x; \hat{\alpha}, \hat{\beta})$, know the value x of the independent variable X, and assume that $\alpha = \hat{\alpha}$ and $\beta = \hat{\beta}$, then:

$$\begin{aligned}
\hat{y} &= E(Y|x; \hat{\alpha}, \hat{\beta}) \\
&= E(\alpha + \beta X + \varepsilon|x; \hat{\alpha}, \hat{\beta}) \\
&= E(\alpha|x; \hat{\alpha}, \hat{\beta}) + E(\beta X|x; \hat{\alpha}, \hat{\beta}) + E(\varepsilon|x; \hat{\alpha}, \hat{\beta}) \\
&= \hat{\alpha} + \hat{\beta} x
\end{aligned}$$

since $E(\varepsilon) = 0$, due to the assumption that $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

In order to model such a system, we need to be able to achieve estimates $\hat{\alpha}$ and $\hat{\beta}$ of the parameters α and β, respectively. However, in order to test hypotheses regarding the parameters, we must make distributional assumptions regarding the error model. Specifically:

- Residuals are Normally distributed: $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, for some parameter σ.
- Residuals are independent, i.e. $Cov(\varepsilon_i \varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0$ for all $i \neq j$.
- The expected value of the residuals is independent of the predictor variable x, i.e. $E(\varepsilon|x) = 0$ for all x.

- The variance of the residuals is independent of the predictor variable x, i.e. $Var(\varepsilon|x)$ is constant for all x. This is called *homoscedasticity*.

Consequently, we want to find optimal parameter estimates $\hat{\alpha}$ and $\hat{\beta}$ such that the errors ε_i are co-minimised, whilst satisfying the above conditions/assumptions. Of course, this is only possible if there is a linear relationship between the variables X and Y, satisfying the above conditions.

Were we to find suitable parameter estimates $\hat{\alpha}$ and $\hat{\beta}$, we would be able to predict values of the response variable corresponding to each of the observed values (x_i) of the predictor variable. Denoted by $\hat{y}_i = E(Y|X = x_i; \hat{\alpha}, \hat{\beta})$, these values are referred to as the *fitted values*, which may be calculated as:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

and the *residual errors* are given by the difference between the actual and predicted values of the response:

$$\varepsilon_i = y_i - \hat{y}_i$$

The distribution of y_i given parameter estimates $\hat{\alpha}$ and $\hat{\beta}$ is thus given by:

$$y_i \sim \mathcal{N}(\hat{y}_i, \sigma^2)$$

2 Parameter Estimation and Model Utility

Linear models can be created in R using the `lm` function, which estimates parameters for a linear model, and tests the significance of model terms as well as overall model utility.

Begin by attempting to fit a linear model to describe the relationship between two independent random Normal samples:

```
x = rnorm(100)
y = rnorm(100)
m1 = lm(y~x)
summary(m1)
```

Here, we create a linear model object `m1`, and then view useful information about the model using the `summary` function. The formula $y \sim x$, which is read as "regress y on x", corresponds to the linear model:

$$y = \alpha + \beta x + \varepsilon$$

Clearly, in this case the parameters α and β should not be significantly different from zero, since we know that the variables x and y are independent. This should be reflected in the p-values corresponding to the t-tests, which are performed to test whether the parameters are significantly different from zero – this information is displayed in the model summary table. In this case, both p-values should be greater than 0.05, indicating insignificance. Note that t-tests are used because the estimators can be considered approximately Normally distributed, providing the number of observations is large (according to the Central Limit Theorem).

The p-value corresponding to an F-test is reported at the bottom of the summary, corresponding to the overall utility of the model. This essentially tests whether the

model has any predictive power. In this case, the p-value should be greater than 0.05, since the model has no predictive power.

Another quantity of interest reported by the summary is the Multiple R^2 value (and the Adjusted R^2 value) which represents the correlation between dependent and independent variables. This is also called the *coefficient of determination*, and is equal to the square of the Pearson product-moment correlation coefficient (for simple linear regression – i.e. only one independent variable).

Now consider fitting a linear model in the presence of a true linear correlation between the variables:

```
x = rnorm(100)
y = x + rnorm(100)
m1 = lm(y∼x)
summary(m1)
```

Here, we know that there is a positive correlation between the variables, and we also know that $\beta = 1$. Indeed, we should find that the p-value corresponding to the x term now indicates significance, since β is significantly different from zero. Furthermore, the p-value corresponding to the overall F-test should also be significant, indicating that the model has predictive power.

Note that the significance of p-values corresponding to the t-tests is illustrated, as indicated in the summary output:

```
Signif.  codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

For example, p-values between 0.001 and 0.01 will be highlighted using two asterisks.

The inclusion of independent variables as regressors is only justified if the corresponding parameter estimates are significantly non-zero, i.e. there is a significant relationship between the independent and response variables. Otherwise, the model would not be meaningful and appropriate, and may suffer from overfitting. Consequently, in order to be suitably *parsimonious*, it is always desirable to only include terms that are identified as significant.

In this particular case, since the intercept term is not significant, we could exclude the parameter α from the model by specifying to force a zero intercept:

```
m1 = lm(y∼0+x)
summary(m1)
```

Now consider fitting a linear model in the presence of a true linear correlation between the variables, with a non-zero intercept:

```
x = rnorm(100)
y = 1 + x + rnorm(100)
m1 = lm(y∼x)
summary(m1)
```

In this case, we should see that both terms are significant.

Diagnostic Plots

Whenever creating a linear model, it is not sufficient to test overall utility and significance of terms – it is important to also consider whether the model assumptions are reasonable, or whether there are serious violations that would imply that the model is invalid.

In R, plotting a linear model object will result in four diagnostic plots being displayed, which help to explore relationships between the model and the observed data. These allow manual visual testing for Normality, independence, homoscedasticity and outliers. These can be plotted using the `plot` function:

```
plot(m1)
```

The four diagnostic plots displayed are:

1. *Plot of Residuals vs Fitted Values.* Residuals and fitted values should not be related, and should be independent, thus there should be no pattern. The variance of the residuals σ^2 should be constant (i.e. independent of the predictor variable), so the residuals should lie within a horizontal band of constant vertical width. Note also that the residuals will always be centred on zero, since $E(\varepsilon) = 0$. This plot also allows easy visual identification of potential outliers.

2. *Normal Q-Q Plot of the Residuals.* Allows visual testing of the assumption that the errors are Normally distributed. Non-Normality indicates assumptions are violated, which would imply that the model is inappropriate.

3. *Scale-Location Plot.* This plots:

$$\sqrt{\left|\frac{\varepsilon_i}{\sqrt{\hat{\sigma}^2}}\right|} = \sqrt{|\text{standardised residuals}|} \qquad \text{vs} \qquad \text{fitted values}$$

   which sometimes makes non-constant variance more noticeable. Again, this plot should exhibit no pattern.

4. *Index Plot of Cook's Distance.* Cook's distance is a combination of the magnitude of the residual and the leverage of the observation, thus is a measure of how influential a particular data point is, i.e. how much effect it has on the regression. A data point has high *leverage* if its x-value is extreme – a point with high leverage has the potential to be *influential*. For the sake of exploring sensitivity, it is often worth investigating the effects on the model parameter estimates of removing highly influential data points.

Task 1:

The inbuilt R dataset *faithful* pertains to the waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA.

1. Create a simple linear regression model that models the eruption duration using waiting time as the independent variable, storing the model in the variable `m1`. Look at the summary of the model.

   (a) What are the values of the estimates of the intercept and coefficient of 'waiting'?

   (b) What is the adjusted R^2 value?

   (c) Does the model have significant utility?

   (d) Are neither, one, or both of the parameters significantly different from zero?

   (e) Can you conclude that there is a linear relationship between the two variables?

2. Plot the eruption duration against waiting time. Is there anything noticeable about the data?

3. Draw the regression line corresponding to the model `m1` onto the plot (i.e. plot the fitted values – extracted using the `fitted` function – against the observed waiting times, displayed as lines). Based on this graphical representation, does the model seem reasonable?

4. Generate the four diagnostic plots corresponding to the model `m1`. Discuss the appropriateness of the model `m1` for describing the relationship between eruption duration and waiting time.

Task 2:

The inbuilt R dataset *trees* provides measurements of the girth, height and volume of timber in 31 felled black cherry trees.

1. It may be hypothesised that Height depends on Girth.

   (a) If creating a model to test this hypothesis, what is the dependent (response) variable, and what is the independent (predictor) variable?

   (b) Create a model to test this hypothesis, assuming a linear relationship between Height and Girth. Call this model `m1`. Does this model have significant utility?

   (c) What is the slope, i.e. what is the value of the parameter that describes the relationship between Height and Girth?

2. Suppose that we know that the height of a particular tree is 80ft.

   (a) Can we use the above model `m1` to estimate the Girth corresponding to the 80ft tree? If not, why not?

   (b) Create a simple linear regression model suitable for the purpose of estimating Girth given knowledge of tree Height. Call this model `m2`. Does this model have significant utility? Are all parameters significant?

   (c) What is the slope, i.e. what is the value of the parameter that describes the relationship between Girth and Height? Compare this with the slope from model `m1`. What can you conclude?

3. Note that the intercept in model `m2` is not significant.

(a) Create a third model `m3` that is similar to `m2`, but does not include the intercept parameter (this can be achieved by including an additive '`0`' term on the right hand side of the regression formula – for further information type `?formula`. Be aware that forcing a zero intercept will cause the R^2 values to no longer be useful/comparable).

(b) Plot tree Height against Girth, and display lines corresponding to the three regression models: `m1`, `m2` and `m3` (i.e. plot the fitted values – extracted using the `fitted` function – against the observed values of the independent variable). Why are these lines different?

4. Now suppose we want to model tree Volume.

(a) Create two models – one that models Volume using Girth as the independent variable, and one that models Volume using Height as the independent variable. From looking at the model summaries, which of these two models do you consider better?

(b) Consider the better of the two models. Create diagnostic plots corresponding to this model. What do you notice?

(c) From looking at the diagnostic plots, we can see that one observation is particularly influential, having a large residual given the leverage. Regenerate the model excluding this one influential observation. From looking at the summary, how does removal of this single observation affect the estimation of model parameters?

Task 3:

Consider the inbuilt R dataset *anscombe*. This dataset contains four x-y datasets, contained in the columns: ($x1$,$y1$), ($x2$,$y2$), ($x3$,$y3$) and ($x4$,$y4$).

1. For each of the four datasets, calculate the correlation between the x and y variables. What do you conclude?

2. For each of the four datasets, create a linear model that regresses y on x. Look at the summaries corresponding to these models. What do you conclude?

3. For each of the four datasets, create a plot of y against x. What do you conclude?

3 Modelling Non-Linear Relationships

It should be noted that a linear model is linear in the parameters. Consequently, linear regression is not limited to dealing with models of the form:

$$Y = \alpha + \beta X + \varepsilon$$

and actually extends to models of the more general form:

$$f(Y) = \alpha + \beta g(X) + \varepsilon$$

for some functions $f()$ and $g()$.

For example, the model:

$$\log(Y) = \alpha + \beta\sqrt{X} + \varepsilon$$

is also called a simple linear regression model, since it is linear in the parameters, and there is only one independent variable (X).

It is often necessary to transform data in order to satisfy the assumptions behind the linear model. This is often done in order to:

1. Linearise the data, i.e. make the relationship between variables more linear;
2. Stabilise the variance σ^2 of the model residuals, so that σ^2 does not depend on the independent variable(s).

Common transformations include the natural logarithm, and power transformations of the form X^n (for some $n \in \mathbb{R}$), noting that if $n < 1$ then large x-values will be proportionally reduced relative to smaller ones. Importantly, any transformations applied to the data must be monotonic, so that there is no information loss (i.e. the original data must be recoverable).

Transformations to linearise the data usually involve transforming the independent variable X, whilst variance stabilising transformations usually involve transforming the response variable Y (and possibly the independent variable X also).

Note that power relationships (that intercept the origin) may be explored, identified, and modelled by log-transforming both the independent and response variables and fitting a linear model.

Task 4:

Consider the inbuilt R dataset *cars*, which contains data regarding the speed of cars and the distances taken to stop.

1. Plot speed against dist (distance). Does the relationship between these variables seem linear?
2. Create a linear model for the relationship between speed and dist. Does this model have significant utility? Do the diagnostic plots identify any behaviour suggesting that the regression assumptions are violated?
3. Now plot speed against the square root of dist. Does this relationship seem more linear? Create a linear model for the relationship between speed and the square root of dist. Is this model better or worse than the previous model?
4. Now plot the (natural) logarithm of speed against the logarithm of dist. Does this relationship seem linear? Create a linear model for the relationship between the logarithm of speed against the logarithm of dist. Is this model better or worse than the previous model?
5. What do the parameter estimates (and standard errors) imply about the suitability of the model of speed against the square root of dist?
6. Focussing on the log-log plot, utilise the log-log model in order to add the line of fitted values – extracted using the `fitted` function – to the plot, showing the linear relationship between the transformed data.
7. Now regenerate the original plot of speed against dist. Using the `lines` function, add a curve to the plot corresponding to the fitted values of the log-log model.

8. Add a second curve to the plot, this one corresponding to the fitted values of the model of speed against the square root of dist. Are these curves similar?

Task 5:

Consider the inbuilt R dataset *Indometh*, which contains data on the pharmacokinetics of indometacin.

1. Plot time versus conc (concentration). What is the nature of the relationship between time and conc?

2. Apply monotonic transformations to the data so that a simple linear regression model can be used to model the relationship (ensure both linearity and stabilised variance, within reason). Create a plot of the transformed data, to confirm that the relationship seems linear.

3. After creating the linear model, inspect the diagnostic plots to ensure that the assumptions are not violated (too much). Are there any outliers with large influence? What are the parameter estimates? Are both terms significant?

4. Add the line of fitted values – extracted using the `fitted` function – to the plot showing the linear relationship between the transformed data.

5. Now regenerate the original plot of time versus conc (i.e. the untransformed data). Using the `lines` function, add a curve to the plot corresponding to the fitted values of the model.

4 Multiple Regression

Multiple regression is the extension of *simple linear regression* to include multiple independent variables. This more general form may be expressed:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

or alternatively:

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \varepsilon$$

or alternatively:

$$Y = \mathbf{B}^{\mathrm{T}} \mathbf{X} + \varepsilon$$

where

$$\mathbf{B} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} \qquad \text{and} \qquad \mathbf{X} = \begin{bmatrix} 1 \\ X_1 \\ \vdots \\ X_n \end{bmatrix}$$

Multiple regression models can be created in R similarly to simple linear regression models. For example,

```
x1 = rnorm(100)
x2 = rnorm(100)
y = 1 + x1 + x2*x2 + rnorm(100)
```

```
m1 = lm(y∼x1+x2)
summary(m1)
```

Note which terms are significant. Now try the model:

```
m1 = lm(y∼x1+x2*x2)
summary(m1)
```

Note that it did not produce the desired effect. In order to do this, we need to create a new variable before attempting to regress y against $x1$ and $x2^2$, e.g.:

```
x2sq = x2*x2
m1 = lm(y∼x1+x2sq)
summary(m1)
```

Note that both terms should now be significant.

We can also try including an interaction term between the independent variables, in order to detect any co-dependency:

```
m1 = lm(y∼x1+x2sq+x1:x2sq)
summary(m1)
```

or equivalently:

```
m1 = lm(y∼x1*x2sq)
summary(m1)
```

In this case, the interaction term should not be significant.

Note that the reported p-value corresponding to the F-test corresponds to the hypothesis that all parameters equal zero, although it doesn't test whether all predictor variables are required/necessary/appropriate. Also, the reported R^2 value measures how well the model predicts the observed values of the response.

Task 6:

Recall the inbuilt R dataset *trees*, which provides measurements of the girth, height and volume of timber in 31 felled black cherry trees. Suppose we want to model Volume using the Height and Girth variables.

1. Create one simple linear regression model that regresses Volume on Height, and another that regresses Volume on Girth. Which of these two models is better?

2. Now create a multiple regression model that regresses Volume on both Height and Girth. Is this model better than the previous two models? Are all model terms significant? From looking at the diagnostic plots, does it appear that this is an appropriate model? Are there any peculiarities that indicate that the modelling assumptions are invalid?

3. Now create a multiple regression model that regresses Volume on Height and Girth, also including a term representing the interaction between Height and Girth. Is this model better than the previous models? Are all model terms

significant? From looking at the diagnostic plots, does it appear that this is an appropriate model? Are there any peculiarities that indicate that the modelling assumptions are invalid?

4. Now create a multiple regression model that regresses log(Volume) on log(Height) and log(Girth). Is this model better than the previous models? Are all model terms significant? From looking at the diagnostic plots, does it appear that this is an appropriate model? What can we conclude from this?

5. Now create a multiple regression model that regresses log(Volume) on log(Height) and log(Girth), also including a term representing the interaction between log(Height) and log(Girth). Is this model better than the previous models? Are the model terms significant? What can we conclude from this?

6. Which of the above models do you think is the most appropriate for modelling Volume? Plot Volume against Girth. Add lines to the plot illustrating the fitted values – extracted using the `fitted` function – of the model you selected as being most appropriate.

Task 7:

The inbuilt R dataset *Puromycin* contains data regarding the reaction velocity versus substrate concentration in an enzymatic reaction involving untreated cells or cells treated with Puromycin.

1. Plot conc (concentration) against rate. What is the nature of the relationship between conc and rate?

2. Find a transformation that linearises the data and stabilises the variance, making it possible to use linear regression. Create the corresponding linear regression model. Are all terms significant?

3. Add the `state` term to the model. What type of variable is this? Is the inclusion of this term appropriate?

4. Now add a term representing the interaction between rate and state. Are all terms significant? What can you conclude?

5. Given this information, create the regression model you believe to be the most appropriate for modelling conc. Regenerate the plot of conc against rate. Draw curves corresponding to the fitted values of the final model onto this plot – note that two separate curves should be drawn, corresponding to the two levels of state.

5 Choosing Between Models

In multiple regression, it is often the case that there are various acceptable models, in which case it is necessary to choose between them. The number of models to consider can be very large – note that if there are k independent variables then there are 2^k possible models. Since the number of possible models to test can be large, it is often necessary to adopt a strategy for trialling different models, given a number of independent variables.

A common strategy is *stepwise regression*, which involves adding or removing terms each step, converging on an 'optimal' solution. There are two main forms of stepwise regression:

1. *Forward Selection* – start from the simplest model, and add terms one at a time until the fit cannot be improved;

2. *Backward Elimination* – start from the most complex model, and remove terms one at a time until further removal makes the fit (too much) worse.

Note that these different strategies can lead to different models being selected in some cases – they do not guarantee that the global best model will be found.

At each step of the stepwise regression procedure, we must decide whether or not to add/remove a term, or whether to stop. Note that adding more terms will always increase the R^2 value, since adding more parameters will always allow an improved fit. However, if terms are included that are no appropriate (i.e. have little/no predictive power) then the model will suffer from overfitting.

Akaike's Information Criterion (AIC) is often used to make such decisions. This score rewards models that better fit the data, whilst penalising those that use many parameters. The best model is deemed to be the one with the lowest AIC (although note that other information criteria exist).

Task 8:

The inbuilt R dataset *swiss* contains standardized fertility measure and socio-economic indicators for 47 French-speaking provinces of Switzerland.

1. Create a linear model regressing Fertility on all other variables, using the command:
   `m1 = lm(Fertility~.,data=swiss)`

   Are all terms significant?

2. Use the `step` function – i.e. use the command `step(m1)` – to perform backward elimination stepwise regression, in order to automatically remove inappropriate terms. Which term(s) were removed? What is Akaike's Information Criterion (AIC) corresponding to the final model? Are all terms in the resulting model significant?

Task 9:

The inbuilt R dataset *attitude* contains data from a survey of clerical employees.

1. Create a linear model regressing `rating` on complaints, and store the model in a variable called `m1`.

2. Use the `step` function to perform forward selection stepwise regression, in order to automatically add appropriate terms, using the command:
   `m2 = step(m1,.~.+privileges+learning+raises+critical+advance)`

   Which term(s) were added? What is Akaike's Information Criterion (AIC) corresponding to the final model? Are all terms in the resulting model significant?

6 Estimation of Parameters for the Simple Linear Regression Model

In this section, we will consider how the parameters $\hat{\alpha}$ and $\hat{\beta}$ can be estimated in the case of simple linear regression.

We want to find the optimal parameter estimates $\hat{\alpha}$ and $\hat{\beta}$ that result in the model optimally fitting the observed data. Specifically, we want to maximise the likelihood of observing these particular parameter values, given the data. This can be written:

$$\mathcal{L}(\alpha, \beta | x_1, \ldots, x_n, y_1, \ldots, y_n) \rightarrow \max$$

However, we known that:

$$\begin{aligned}
\mathcal{L}(\alpha, \beta | x_1, \ldots, x_n, y_1, \ldots, y_n) &= \prod_{i=1}^{n} f(y_i | x_i; \alpha, \beta) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y_i - E(y_i))^2}{2\sigma^2}} \\
&= \sum_{i=1}^{n} \left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_i - E(y_i))^2}{2\sigma^2} \right) \\
&= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - E(y_i))^2
\end{aligned}$$

Consequently,

$$\mathcal{L}(\alpha, \beta | x_1, \ldots, x_n, y_1, \ldots, y_n) \rightarrow \max$$

happens when:

$$S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \rightarrow \min$$

This is equivalent to:

$$S = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2 \rightarrow \min$$

or, noting that $\varepsilon_i = y_i - \hat{y}_i$:

$$S = \sum_{i=1}^{n} \varepsilon_i^2 \rightarrow \min$$

So the optimal estimates of α and β are achieved when the *residual sum of squares* (S) is minimised. Consequently, the *maximum likelihood estimates* and *least squares estimates* of α and β are equivalent (due to model form and distributional assumptions).

This optimisation problem may be solved by equating the differentials of S with respect to $\hat{\alpha}$ and $\hat{\beta}$ to zero, and simultaneously solving them thus eliminating $\hat{\alpha}$ and finding $\hat{\beta}$, and subsequently deducing $\hat{\alpha}$.

As an aside, note that $\hat{\alpha}$ and $\hat{\beta}$ are themselves random variables, and thus have distributions. Consequently, so does the fitted value $\hat{y}$.

Task 10:

Starting from the equation:

$$S = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2 \rightarrow \min$$

derive the formula for $\hat{\beta}$, and subsequently $\hat{\alpha}$.