Outline

1) Linear models – simple linear regression

2) Model formulation

3) Linear least squares

4) Maximum likelihood estimators

Simple Linear Regression

Mitsubishi Example : Consider the case of a small business owner named Ingrid who wants to purchase a fleet of Mitsubishi Sigmas. To reduce expenditure she decides to purchase second-hand Sigmas and wants to estimate how much they will cost.

A member of Ingrid's accounts department looks through the classifieds of the local newspaper and comes up with the following data on age and price of a set of 39 Mitsubishi Sigma cars (source: Exploring Statistics with Minitab, P. Martin, L. Roberts, R. Pierce, Nelson 1994)

Can Ingrid use the above data to work out how much she will expect to pay for the cars?

Mitsubishi Example

Download the dataset (**mitsub.txt**) from Moodle and load it in RStudio.

A first step might be to look at some summary statistics:

```
> summary(mitsub)
      age             price
 Min.   : 6.00   Min.   : 450
 1st Qu.:10.00   1st Qu.:2850
 Median :12.00   Median :3500
 Mean   :11.79   Mean   :3625
 3rd Qu.:14.00   3rd Qu.:4350
 Max.   :15.00   Max.   :8999
```

Some standard statistical calculations allow Ingrid to predict with 95% certainty that a randomly chosen second hand sigma will cost between \$668 and \$6625. Use the **quantile** function to find these values.

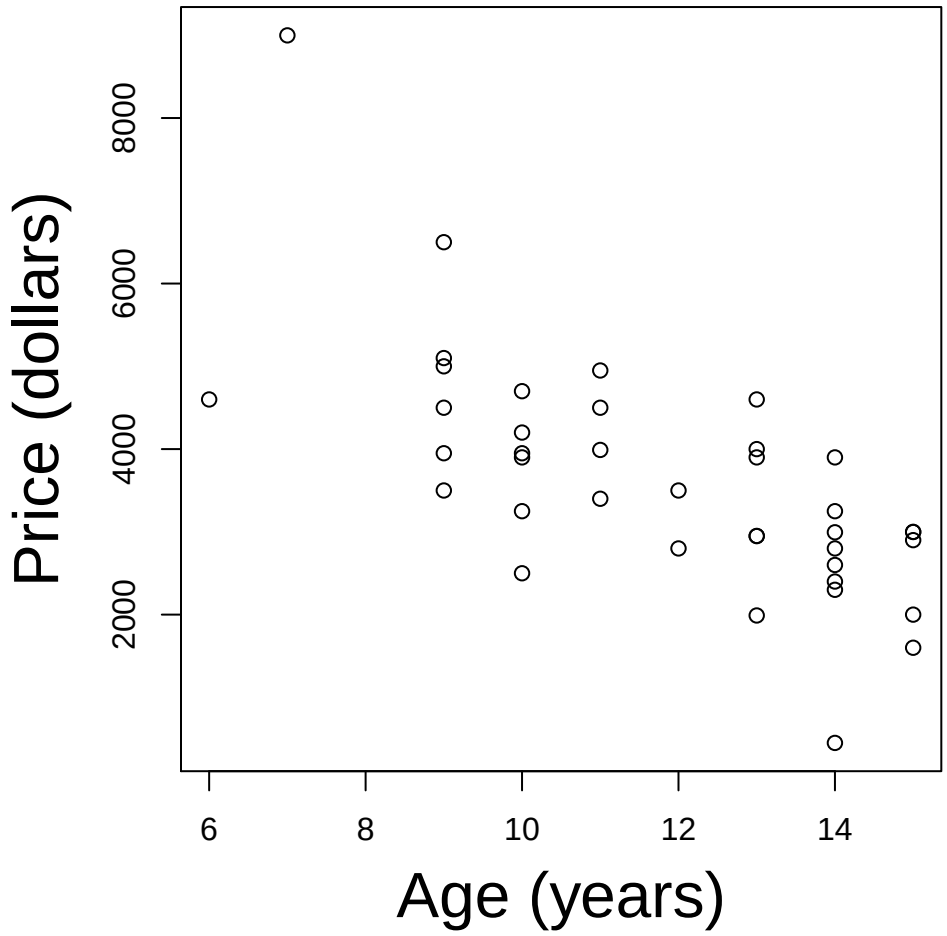
Mitsubishi Example

This analysis gives Ingrid a vague idea about how much she will need to spend to build up its fleet, but the range of the interval is probably too large to be of much practical use

How can she improve her prediction of price? One possibility is to try to make use of the additional data that we have available on age of the cars. Since **age** and **price** are likely to be correlated in some way we might be able to take advantage of this correlation to get better predictions.

The **scatterplot** of **price** versus **age** shown in the figure below confirms the unsurprising fact that prices tend to decrease as **age** increases. Q: write the code to reproduce this plot.

Mitsubishi Example



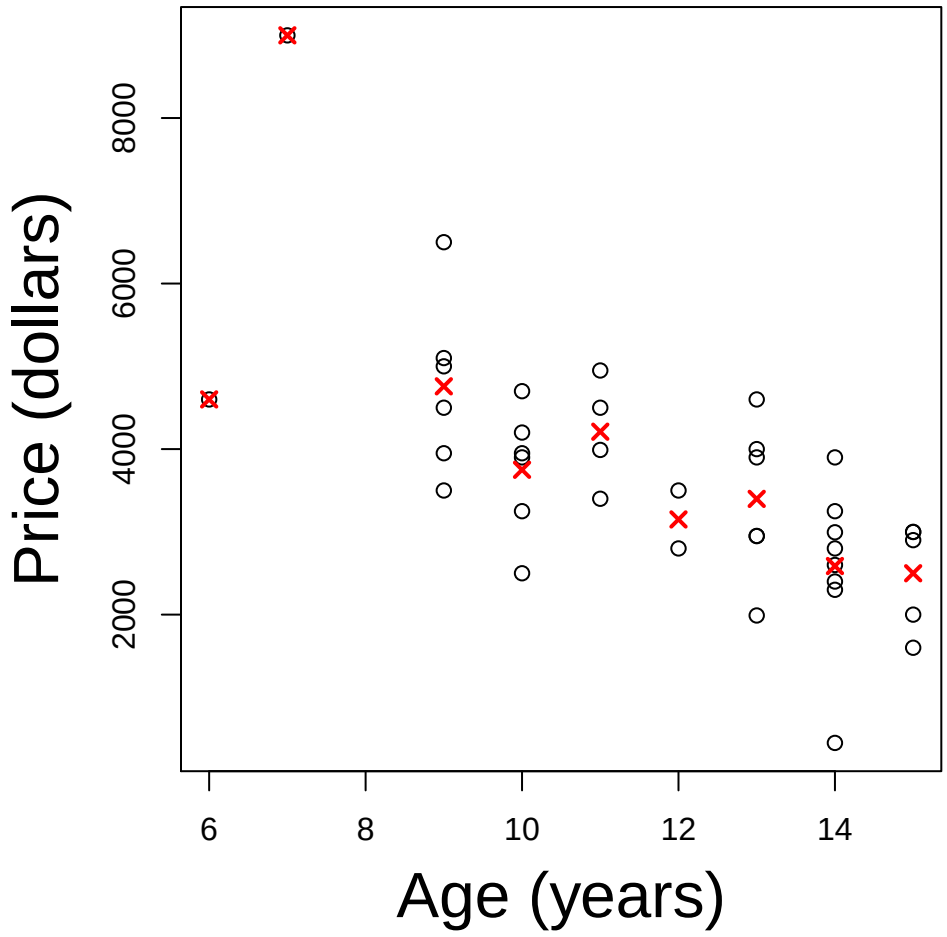
Mitsubishi Example

The age of a car is almost always known. Suppose that we want to predict the price of a randomly chosen 10 year old car. Then we can focus on the sample of prices corresponding to cars with an age of 10 years. Graphically, this corresponds to those observations within the vertical strip at `age=10`.

Within this strip the average price is \$3750. If we now take a different vertical strip corresponding to say, 14 year old cars then we get a different average price: \$2587. Q: Find these values using R.

If this process is repeated at several values of the age variable then we get the following figure - a whole suite of mean values, conditional on the value of age. Q: Write the R code to obtain the following plot.

Mitsubishi Example



Mitsubishi Example

If we make the strips finer and finer then we get what is known as a regression curve. This is defined to be the mean value of price for a given value of age.

One could think about repeating the calculations performed above based on the sample of prices of 10 year old cars to obtain a new prediction interval. However, there are only five such observations, and therefore we would not expect much accuracy. The situation is even worse for other car ages. There are only 2 cars aged 12 years old, and no cars in the sample are 8 years old. How do we use the data to predict prices of cars with these ages?

Mitsubishi Example

The usual approach is to model the regression curve. This means that we make some assumptions about what the regression curve look like. The curve in the above plot is approximately linear, so it may be reasonable to postulate the *simple linear regression model* for these data and use this to predict price from age.

Then the fitted linear regression model can be shown to be

$$\hat{\text{price}} = 8452 - 409\ \textbf{age}$$

Model formulation

- In the Mitsubishi example, we say that **age** is a **predictor** for **price**. The variable **age** is called the **predictor variable** and **price** is the **response variable**.

- The usual generic notation for simple regression is

$$x = \text{ predictor variable}$$

$$y = \text{ response variable}$$

- The prediction equation is

$$\hat{y} = \beta_0 + \beta_1 x$$

Model formulation

- Therefore, if y is any particular observation with corresponding x-value equal to x then we can write

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

where the *error* ϵ is given by

$$\epsilon = y - \hat{y} = y - (\beta_0 + \beta_1 x)$$

- Since we assume that $\beta_0 + \beta_1 x$ is the average of the ys in the vertical strip about x, we will have that

$$E(\epsilon) = 0$$

- Also $Var(y)$ in the vertical strip is equal to σ^2 so we have

$$Var(\epsilon) = \sigma^2$$

- Finally because the observations in the vertical strips are normally distributed, we have

$$\epsilon \sim N(0, \sigma^2)$$

Model formulation

- Thus we can write the simple linear regression model as

$$y = \beta_0 + \beta_1 x + \epsilon, \quad \epsilon \sim N(0, \sigma^2).$$

- Suppose we have n observations, $(x_1, y_1), \ldots, (x_n, y_n)$ from this model, then

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \ldots, n$$

- We assume further that the observations are collected independently of each other so that the ϵ_i are independent. This means that a knowledge of one of the e_i does not tell us anything about the value of the other e_is

Linear Least Squares

- In order to fit a straight line to a plot of points (x_i, y_i), where $i = 1, \ldots, n$, the slope and intercept of the line

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

must be found from the data in some manner

- In order to fit a p-th order polynomial, $p+1$ coefficients must be determined

- Other functional forms besides linear and polynomial ones may be fit to data, and in order to do so, parameters associated with those forms must be determined

Linear Least Squares

- The most common method for determining the parameters in curve-fitting problems is the method of least squares

- The idea is to minimize the sum of squared deviations of the predicted, or *fitted* values, (given by the curve) from the actual observations.

- Applying the method of **least squares**, we choose the slope and intercept of the straight line to minimize

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

- Note that β_0 and β_1 are chosen to minimize the sum of squared vertical deviations, or prediction errors

- To find β_0 and β_1, we calculate

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i)$$

- Setting these partial derivatives equal to zero, we have that

the minimizers $\hat{\beta}_0$ and $\hat{\beta}_1$ satisfy

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

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

- Solving for $\hat{\beta}_0$ and $\hat{\beta}_1$, we obtain

$$\hat{\beta}_0 = \frac{(\sum_{i=1}^{n} x_i^2)(\sum_{i=1}^{n} y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} x_i y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

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

Linear Least Squares

- Functional forms more complicated than straight lines are often fit to data, for example, when there are more than one predictor variables available, we may fit a line of the form

$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

where β_i could be estimated from data

- Or we may be able to fit functions of the following form to decay curves

$$f(t) = Ae^{-\alpha t} + Be^{-\beta t}$$

where the function f is linear in the parameters A and B and nonlinear in the parameters α and β, from data of the form $(y_i, t_i), i = 1, \ldots, n$

- When the function to be fitted is linear in the unknown parameters, the minimisation is relatively straightforward, since calculating partial derivatives and setting them equal to zero produces a set of simultaneous linear equations that can be solved in closed form. This special case is known as **linear least squares**

Linear Least Squares

- If the function to be fit is not linear in the unknown parameters, a system of nonlinear equations must be solved to find the coefficients. Typically, the solution cannot be found in closed form, so an iterative procedure must be used

- For our purposes, the general formulation of the linear least squares problem is as follows: a function of the form

$$\begin{aligned} & f(x_1, x_2, \ldots, x_{p-1}) \\ = \; & \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{p-1} x_{p-1} \end{aligned}$$

involving p unknown parameters, $\beta_0, \beta_1, \ldots, \beta_{p-1}$ is to be fit to n data points

$$\begin{gathered} y_1, x_{11}, x_{12}, \ldots, x_{1,p-1} \\ y_2, x_{21}, x_{22}, \ldots, x_{2,p-1} \\ \vdots \\ y_n, x_{n1}, x_{n2}, \ldots, x_{n,p-1} \end{gathered}$$

- The function $f(x)$ is called the **linear regression** of y on x
- We will always assume that $p < n$, that is, there are fewer unknown parameters than observations
- Fitting a straight line clearly follows this format

- A quadratic can be fit in this way by setting $x_1 = x$ and $x_2 = x^2$

- Many functions that are not initially linear in the unknowns can be put into linear form by means of a suitable transformation

Simple Linear Regression

- Recall the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \ldots, n$$

where ϵ_i are independent random variables with

$$E(\epsilon_i) = 0, Var(\epsilon_i) = \sigma^2$$

the x_is are assumed to be fixed

- Under the simple linear regression model, the least squares estimates are unbiased:

$$E(\hat{\beta}_j) = \beta_j, j = 0, 1.$$

Proof: In lecture.

Simple Linear Regression

- Note that here the proof does not depend on the assumption that ϵ_i are independent and have the same variance, only on the assumptions that the errors are additive and $E(\epsilon_i) = 0$

- From the simple linear regression model, $Var(y_i) = \sigma^2$ and $Cov(y_i, y_j) = 0$, where $i \neq j$, this makes the computation of the variances of the $\hat{\beta}_j$s straightforward.

- Under the assumptions of the simple linear regression model

$$Var(\hat{\beta}_0) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$Var(\hat{\beta}_1) = \frac{n\sigma^2}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\sigma^2 \sum_{i=1}^n x_i}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$

Proof: In lecture.

- We see that the variances of the slope and intercept depend on the x_i and on the error variances, σ^2

- The x_is are known, therefore to estimate the variance of slope and intercept, we only need to estimate σ^2

Simple Linear Regression

- Since in the simple linear regression model, σ^2 is just the expected squared deviation of th y_is from the line $\beta_0 + \beta_1 x_i$, it is natural to base an estimate of σ^2 on the average squared deviations of the data about the fitted line, we define the **residual sum of squares (RSS)** to be

$$RSS = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

and

$$S^2 = \frac{RSS}{n-2}$$

is an unbiased estimator of σ^2

Simple Linear Regression

- The divisor $n-2$ is used rather than n because two parameters have been estimated from the data, giving $n-2$ degrees of freedom

- The variances of $\hat{\beta}_0$ and $\hat{\beta}_1$ are thus estimated by replacing σ^2 by S^2, yielding estimates that we will denote $s^2_{\hat{\beta}_0}$ and $s^2_{\hat{\beta}_1}$.

- If the errors ϵ_i are independent normal random variables, then the estimated slope and intercept, being linear combinations of independent random variables, are normally distributed as well

Simple Linear Regression

- More generally, if the ϵ_i are independent and the x_i satisfy certain assumptions, a version of the central limit theorem implies that, for large n, the estimated slope and intercept are approximately normally distributed

- The normality assumption, or its approximation makes possible the construction of confidence intervals and hypothesis tests. It can be shown that

$$\frac{\hat{\beta}_i - \beta_i}{s_{\hat{\beta}_i}} \sim t_{n-2}$$

  which implies that the t distribution can be used for confidence intervals and hypothesis tests.

Mitsubishi Example:

For the Mitsubishi price/age data, using the R commands

```
> # Question:
> # Write down the code
> # Hint: use 'lm' function
```

gives that the estimates of the coefficients are

$$\hat{\beta}_0 = 8451.591 \quad \hat{\beta}_1 = -409.2175$$

and the residual variance is

$$\hat{\sigma}^2 = 1045^2$$

The R output is

```
> summary(lm(price ~ age))

Call:
lm(formula = price ~ age)

Residuals:
    Min      1Q  Median      3Q     Max
-2272.5  -504.8  -122.5   568.2  3411.9

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8451.59     840.05  10.061 3.89e-12 ***
age          -409.22      69.79  -5.863 9.62e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1045 on 37 degrees of freedom
Multiple R-Squared: 0.4816,     Adjusted R-squared: 0.4676
F-statistic: 34.38 on 1 and 37 DF,  p-value: 9.618e-07
```

Mitsubishi Example:

Hence the fitted line is

$$\hat{\text{price}} = 8451.59 - 409.22 \text{ age}.$$

The estimated price of new Mitsubishi Sigma cars (**age** =0) is \$8451.59

The estimated depreciation rate of Mitsubishi Sigma cars is \$409.22 per year

The standard error of the intercept, $s_{\hat{\beta}_0} = 840.05$. A 95% confidence interval for the intercept, β_0 based on the t distribution with 37 df is

$$\hat{\beta}_0 \pm t_{37}(0.025) s_{\hat{\beta}_0} \Longrightarrow (6748.897, 10153.10).$$

Mitsubishi Example:

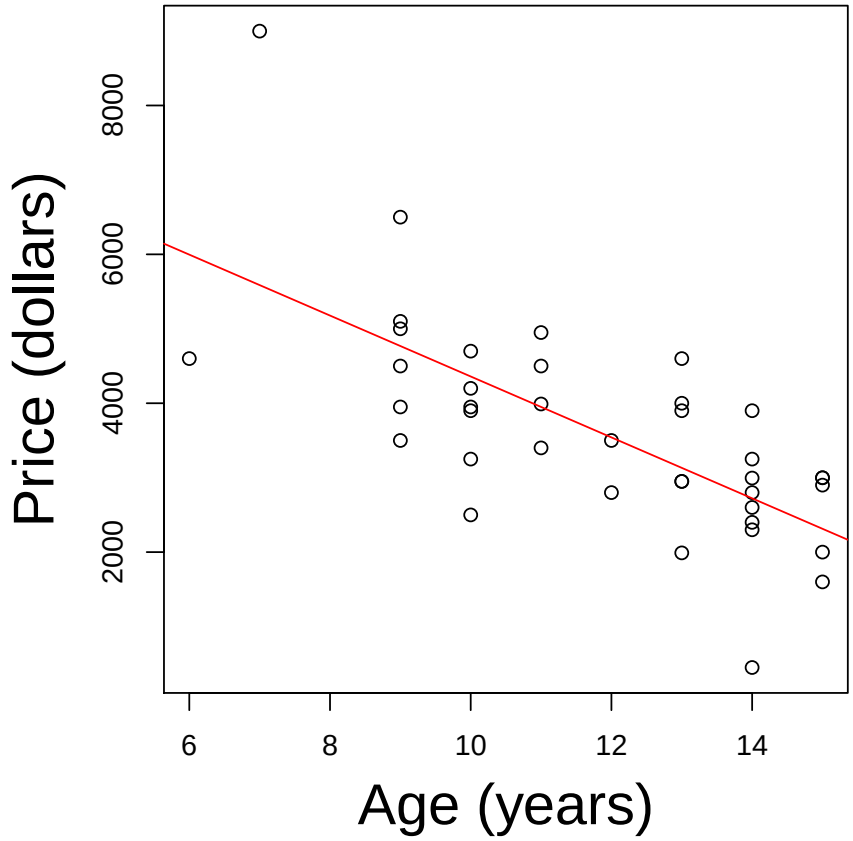
Similarly, a 95% confidence interval for the slope β_1, is

$$\hat{\beta}_1 \pm t_{37}(0.025) s_{\hat{\beta}_1} \Longrightarrow (-550.628, -267.8120).$$

To test the null hypothesis $H_0 : \beta_0 = 0$, we would use the t statistic $\hat{\beta}_0 / s_{\hat{\beta}_0} = 10.061$. The hypothesis would be rejected at significance level $\alpha = 0.05$, there is strong evidence that the intercept is non-zero.

Q: Write the code that returns the predicted price of a 10 years old car (Hint: use the `predict` function)

Q: Write the code that produces the graph below (Hint: use the **`abline`** function)



Maximum likelihood estimators in simple linear regression model

MLEs of β_0, β_1 and σ^2 under normality assumptions?

Maximizing the likelihood is equivalent to maximizing the log-likelihood: if $L_1 < L_2$ are values of the likelihood function for two different points in parameter space, $\log(L_1) < \log(L_2)$.

Likelihood for single observation y_i:

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right)$$

Independent errors, full likelihood:

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right)$$

$$= \quad (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right).$$

Log likelihood:

$$-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

$$= \quad -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

Regardless of σ^2, the above is maximized with respect to β_0 and

β_1 by minimizing

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

$\Longrightarrow$ Least squares and maximum likelihood estimators of β_0, β_1 coincide

Maximum likelihood estimator of error variance

Maximum likelihood estimator of σ^2?

$$\frac{\partial}{\partial \sigma^2}\left(-\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right)$$

$$= \quad -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

Set to zero, write $\hat{\sigma}^2$ for MLE, substitute $\hat{\beta}_0$ and $\hat{\beta}_1$ for β_0 and β_1, which gives

$$\hat{\sigma}^2 \quad = \quad \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$