Regression Models
Lecture I: Review of Simple Linear Regression & Motivating Multiple Regression

DT9002: Postgraduate Certificate in Applied Statistics

Dr Joe Condon

School of Mathematical Sciences
Technological University Dublin

Regression Models

Regression models look at the relationship between a response variable(s) and one or more predictors.

They are used everywhere in science, medicine, social science, business and economics. [even humanities sometimes!]

Some examples:

- Life expectancy given medical and lifestyle factors
- Consumer sentiment and interest rates
- Yield from a chemical reaction and amount of catalysts used
- Annual salary and education
- Probability of defaulting on a bank loan based on socio-economic factors
- Response to a drug and dose used
- Many, many more...

Purpose of Regression?

To answer/address some or all of the following:

- Is there a relationship between the response and the predictors?
- What is the nature of the relationship, e.g. straight line, curved line, strong weak etc.
- Which predictors (all/some) from a set are related to the response?
- Can I make meaningful predictions of the response using its relationship with the predictors?

Modelling cycle

- Prepare data.
- Summary statistics/plots.
- Propose/formulate and fit a model.
- Assess evidence for that models' existence (hypothesis testing).
- Consider alternative models (model building).
- Examine model adequacy - diagnose model failings/shortcomings.
- Fit alternative models and cycle through until a final working model is identified.
- Predict from model if that is the goal of the analysis. Assess model performance in prediction.

History?

Around 1795 Carl Friedrich Gauss probably developed least-squares regression to solve astronomical problems.

Certainly by 1805-1810 both Gauss and Legendre were publishing least squares (regression) papers.

The normal probability distribution is also one of his discoveries - hence it is also called the Gaussian distribution.

There are many geniuses in maths - Gauss is up there with the greatest of them.

A defining feature of the regression model is that the relationship between the response and the predictors has a stochastic (random) element - i.e. the relationship is not fully deterministic.

We will initially look at a sub-class of regression models - linear models. Then we will expand our repertoire to include an important class of non-linear models - generalised linear models.

Regression models are applied to datasets consisting of a number of variables (response and predictors) measured against an experimental (or sample) unit. This collection of variables recorded on an experimental unit is called an ‘observation’.

Example: LDL Data

i	Weight (kg)	Cholestoral (mg/dL)
1	100	160
2	105	150
3	90	120
4	80	90
5	80	110
6	85	130
7	87	110
8	92	140
9	90	130
10	95	140
11	93	120
12	85	120
13	85	110
14	70	100
15	85	100

We use the following convention:
Each observation is given an index i,
i =1=first observation,
i =2=second observation, etc.
Each observation has a response variable y_i. In this case y_i is the LDL level for each child

Each observation has at least one predictor variable. These are denoted $x_{i1}, x_{i2}, x_{i3}, \ldots$

If there is is only one predictor variable, then we denote it as x_i.

Example: Dose-response data

An experiment was performed to test the effect of a steroid on the activity levels of rats. Eleven rats were used in the study, different doses of the steroid were administered and their activity levels were recorded over a period.

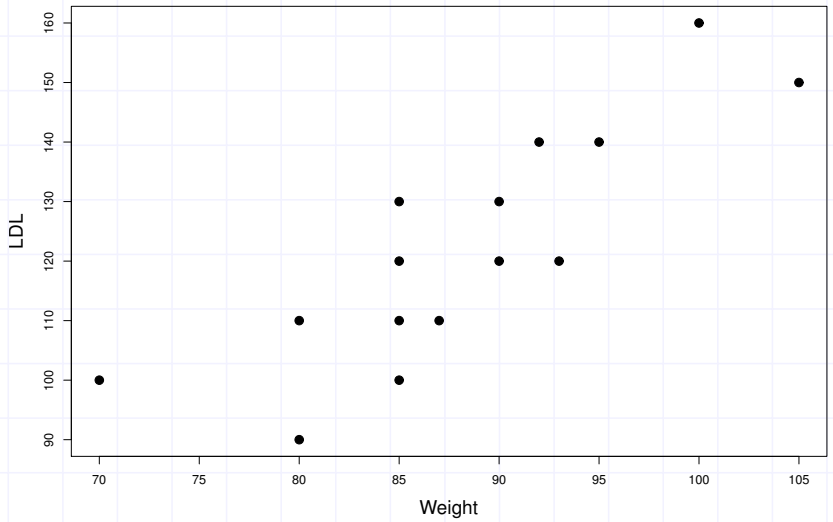
Dose	Activity
10.45	991
12.57	1233
15.62	1229
25.98	1684
30.52	1862
34.06	1919
41.17	2082
50.78	1776
61.01	1528
71.76	881
79.20	1101

Example: Bread-wrapper data

Seal Strength	Sealing Temp.	% polyethylene
6.6	225	0.5
6.9	285	0.5
7.9	225	0.5
6.1	285	0.5
9.2	225	1.7
6.8	285	1.7
10.4	225	1.7
7.3	285	1.7
9.8	204.5	1.1
5	305.5	1.1
6.9	255	1.1
6.3	255	1.1
4	255	0.09
8.6	255	2.11
10.1	255	1.1
9.9	255	1.1
12.2	255	1.1
9.7	255	1.1
9.7	255	1.1
9.6	255	1.1

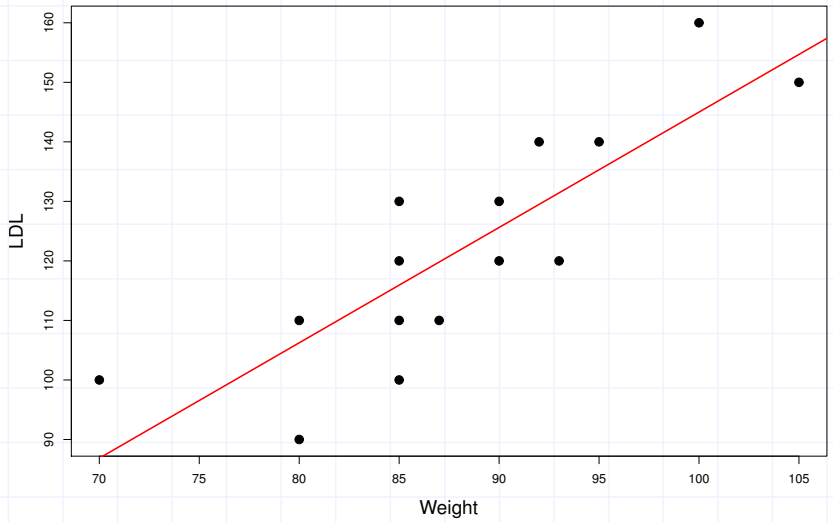
Motivating Examples

LDL Data: The relationship between the weight of obese children and LDL ('bad') cholesterol.



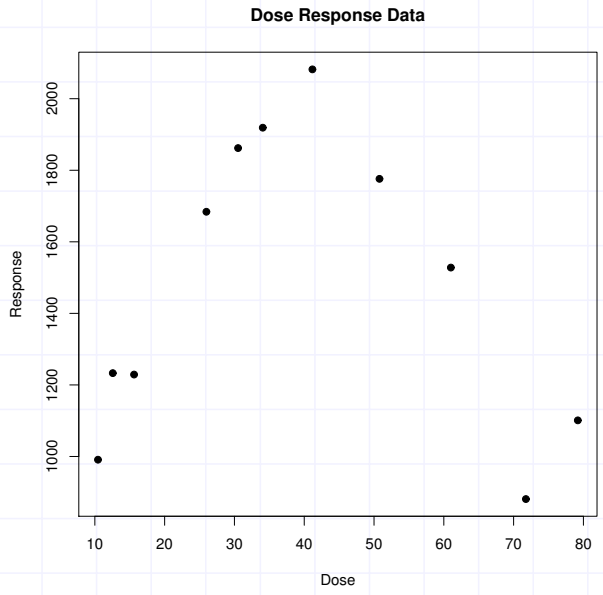
A straight line model seems appropriate here, i.e.

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



This is called a ‘simple linear regression model’.

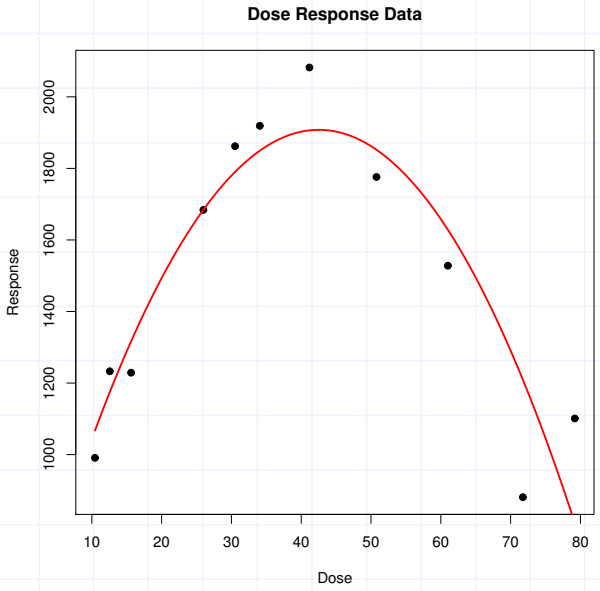
Dose-response Data: The relationship between the dose of a steroid given to rats and their activity levels.



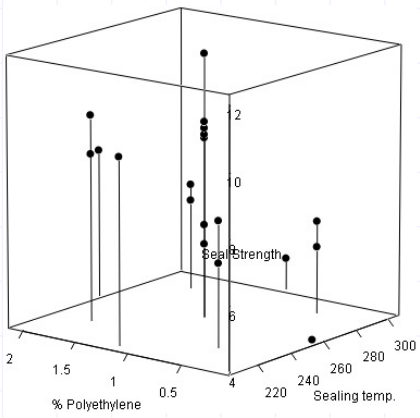
A non straight-line (quadratic) model may be appropriate here, i.e.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

where y_i is the activity level for rat i and x_i is the dose of steroid they received.



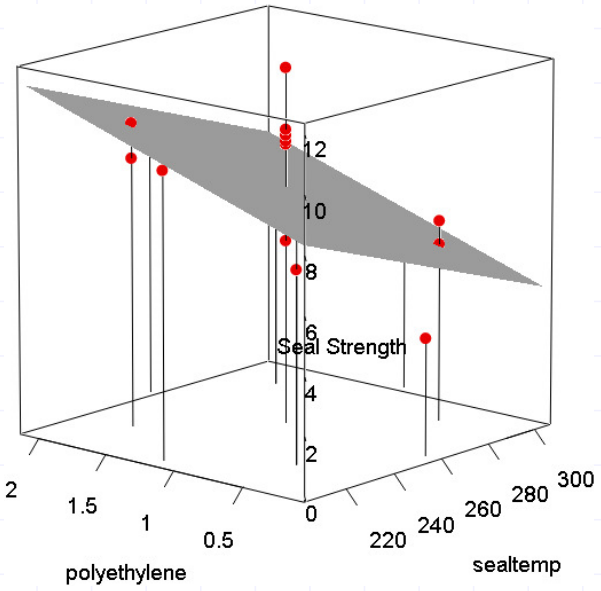
Example: The Bread-Wrapper Data: The relationship between the strength of the wrapping on a loaf of bread, and the sealing temperature t which the stock (i.e. glue) was applied and the % polyethylene in the stock.



A planar response surface model may be appropriate here, i.e.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

where y_i is the strength for wrapper i, x_{i1} is the sealing temperature and x_{i2} is the % polyethylene used in the stock.



For the three models considered here so far, i.e.,

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i \\ y_i &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \\ y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \end{aligned}$$

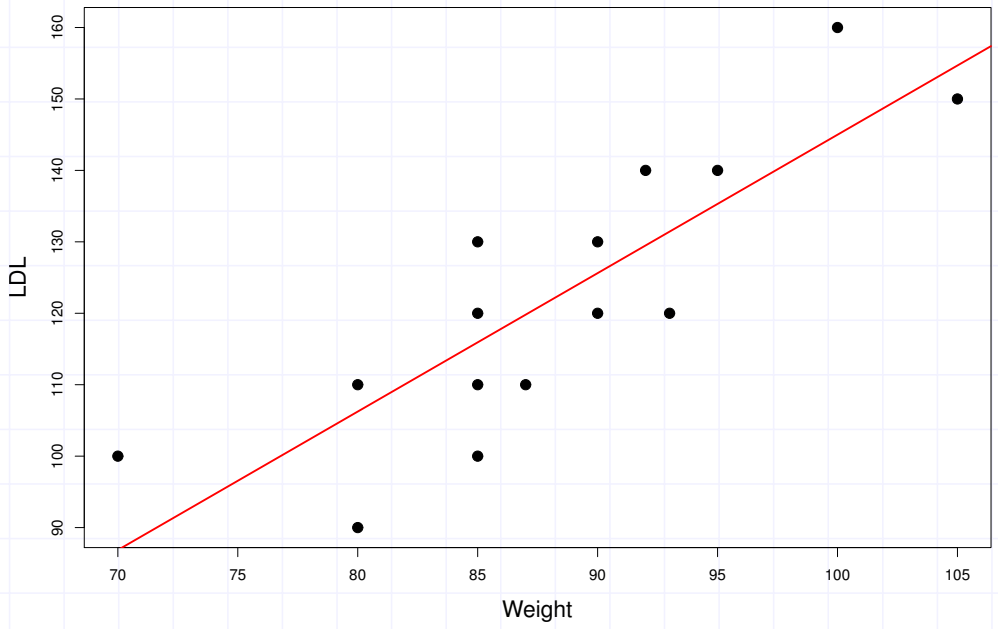
NB. The unknown regression parameters, the β's, enter each of the models as linear coefficients, which is to say that they multiply a predictor/predictors but not each other, nor are they raised to a power etc.

The predictors are not required to be linear however.

LDL Data

- A medical researcher is investigating the relationship between weight in obese children and LDL Cholesterol in their blood stream.
- A central question: does increasing weight predict an increase in LDL Cholesterol levels in obese children? Do these data supply any reasonable evidence of such a relationship?
- They randomly recruit 15 obese children and measure (i) their weight (kg) and (ii) their blood LDL level (mg/dL).
- They the draw a scatterplot of the results.

The data are shown in the plots below.



The line is called the 'regression line'.

- We are modelling the relationship as a straight line.
- What straight line do we choose from the infinity of lines available?
- There are numerous choice, but for linear statistical models we choose a particular mathematical criterion for finding the line, called the criterion of **least squares**.
- The criterion of least squares has some 'nice' statistical properties - it is also very mathematically convenient. However, it is the statistical properties that we focus on in modern times - the mathematically convenience of it is somewhat moot in an era of ubiquitous computing.

The model for the straight line is:

$$\text{LDL} = \beta_0 + \beta_1(\text{weight}) + \varepsilon.$$

Where
β_0 is the intercept
β_1 is the slope
ε is the error term - i.e. the distance of the observed data point to the line.

We say that we are *regressing* LDL on Weight.

More generally, letting y = LDL and x = Weight, we are regressing y on x.

So the general simple linear regression model becomes:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (1)$$

NB. y_i is also called the response or dependent variable, and x_i the predictor or independent variable.

For the time being we assume the following:

1. Both x_i and y_i variables are on the continuous level of measurement.
2. The x_i variables are not random variables.
3. The y_i variables are random variables - consisting of both a non-random (deterministic/structural) and random components.

Least Squares Criterion

The idea is to choose a line that is simultaneously close to all points by finding the line that minimises the sum of squared vertical distances from the observed data points to the line.

Mathematically this is: measure the squared vertical distances from the points to the line - so need to add up this distance over all points. For a given point y_i we get:

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

Squaring and writing in terms of the distance we get,

$$\varepsilon_i^2 = (y_i - \beta_0 - \beta_1 x_i)^2$$

and finally summing over the n data points we get an overall measure of squared distance, i.e. the Sum of Squares

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

The task now is to find the values of β_0, β_1 such that the objective function Q in equation (2) is minimised.

NB: The objective function Q is called **least squares criterion**.

To do this solve the following equations simultaneously.

$$(1)\ \frac{\partial Q}{\partial \beta_0} = 0 \qquad (2)\ \frac{\partial Q}{\partial \beta_1} = 0$$

The solutions for the two parameters (i.e. β's) are,

$$\hat{\beta}_1 = S_{xy}/S_{xx} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where

$$S_{xy} = \sum_{i=1}^{n} xy - \frac{\sum_{i=1}^{n} x \sum_{i=1}^{n} y}{n} \qquad S_{xx} = \sum_{i=1}^{n} x^2 - \frac{(\sum_{i=1}^{n} x)^2}{n}$$

The hats are used to indicate that these are estimates (in fact MLEs - more on this later).

We can see know that this is mathematically convenient - maths gives us relatively easy equations and formulae as solutions.

Since we have minimised the squared distance, this method is called the Method of Least Squares or simply Least Squares (LS). It can be shown that under fairly general conditions these estimates are the unique minimisers of (2).

Solving for the LS estimates for the LDL data we get using R,

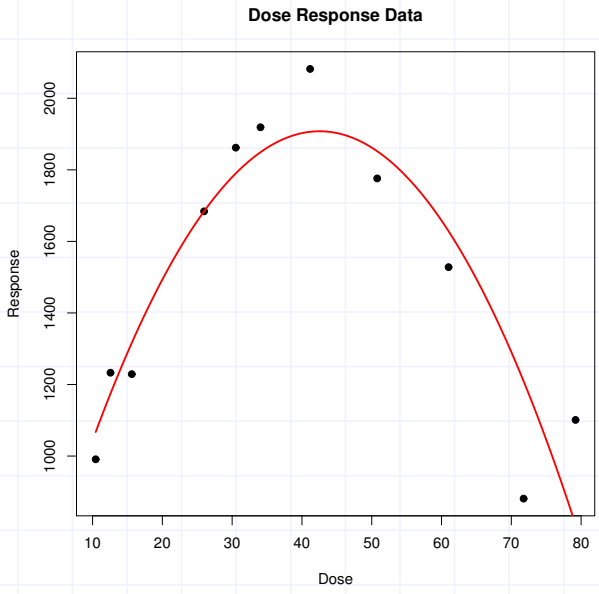
```
> ldl = read.table("ldldata.txt",header=T,sep=',')
> attr(ldl,'names')=c("Weight","LDL")
> fit_ldl=lm(LDL~Weight,data=ldl)
> summary(fit_ldl)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -48.7814    30.8577  -1.581    0.138
Weight        1.9378     0.3486   5.559 9.25e-05
```

How are the slope and intercept interpreted?

Dose-Response Data

An experiment was performed to test the effect of a steroid on the activity levels of rats. Eleven rats were used in the study, different doses of the steroid were administered and their activity levels were recorded over a period.

Dose	Activity
10.45	991
12.57	1233
15.62	1229
25.98	1684
30.52	1862
34.06	1919
41.17	2082
50.78	1776
61.01	1528
71.76	881
79.20	1101



Mathematically, a line that goes smoothly up and then down (or vice versa) could be a quadratic line.

The quadratic model is:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

where y_i is the activity level for rat i and x_i is the dose of steroid they received.

How do we find the estimates for β_0, β_1 and β_2?

We use the LS method again.

Define the objective function:

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2)^2$$

NB. This is a function of three unknowns.

Take the partial derivative WRT each unknown and set each equal to zero. Then solve the three equations simultaneously.

$$(1)\ \frac{\partial Q}{\partial \beta_0} = 0 \qquad (2)\ \frac{\partial Q}{\partial \beta_1} = 0 \qquad (3)\ \frac{\partial Q}{\partial \beta_2} = 0$$

This is entirely feasible - but we can be more mathematically efficient - see 'Matrix Formulation of LS Model' in the appendix.

Formulae exists for this case as well (see Appendix) but no-longer simple ones. We rely on software like R, SAS, SPSS, etc. for fitting such models.

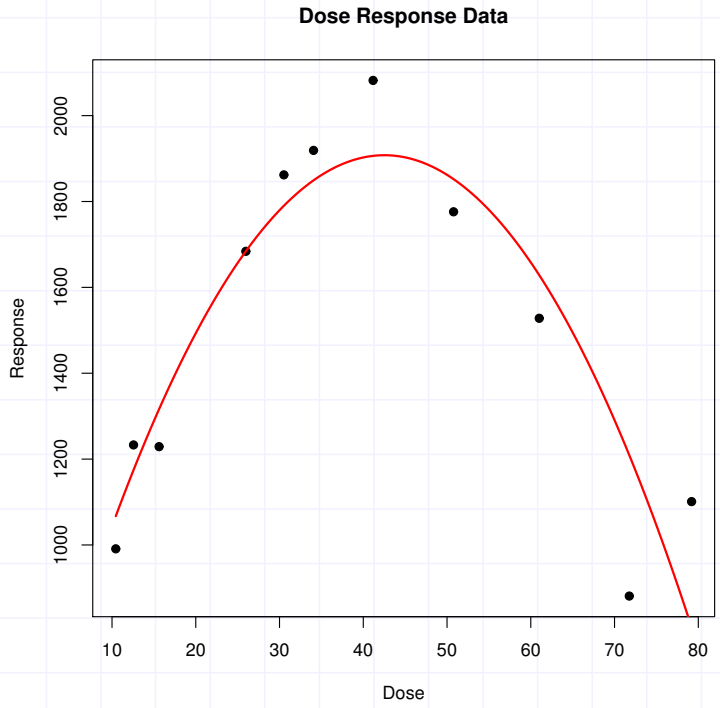
Using R we might try the following:

```
> dr = read.table("doseresponse.txt",header=T,sep=' ')
> fit_dr<- lm(activity~dose+I(dose^2),data=dr)
> summary(fit_dr)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 430.0759   205.3052   2.095 0.069496 .
dose         69.5021    11.2468   6.180 0.000265 ***
I(dose^2)    -0.8172     0.1257  -6.502 0.000188 ***
---
Residual standard error: 182.7 on 8 degrees of freedom
Multiple R-squared:  0.8428,  Adjusted R-squared:  0.8035
F-statistic: 21.45 on 2 and 8 DF,  p-value: 0.0006106
```

So, the LS equation for the quadratic model to these data is:

$$y = 430.0754 + 69.5021x - 0.8172x^2$$

If you plot this function over the scatterplot of the data - then you get the following curved line.



Bread-wrapper Data

We can use a similar approach in fitting a model to the Bread-wrapper data.

Seal Strength	Sealing Temp.	% polyethylene
6.6	225	0.5
6.9	285	0.5
7.9	225	0.5
6.1	285	0.5
9.2	225	1.7
6.8	285	1.7
10.4	225	1.7
7.3	285	1.7
9.8	204.5	1.1
5	305.5	1.1
6.9	255	1.1
6.3	255	1.1
4	255	0.09
8.6	255	2.11
10.1	255	1.1
9.9	255	1.1
12.2	255	1.1
9.7	255	1.1
9.7	255	1.1
9.6	255	1.1

The model we are fitting is:

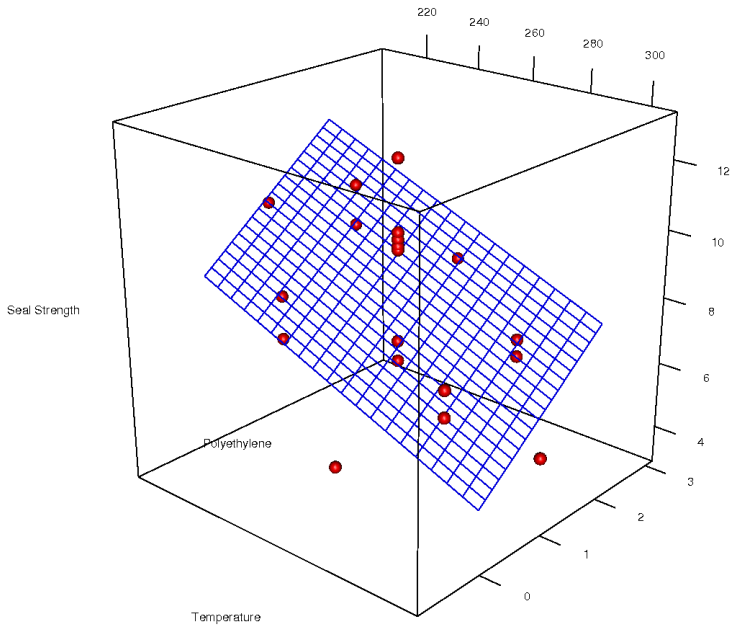
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

where x_{i1} is the sealing temperature for observation i and x_{i2} is the corresponding % of polyethylene used.

Mathematically, such an equation defines a plane in 3 dimensions.

Once again we get the solution using R and the LS criterion:

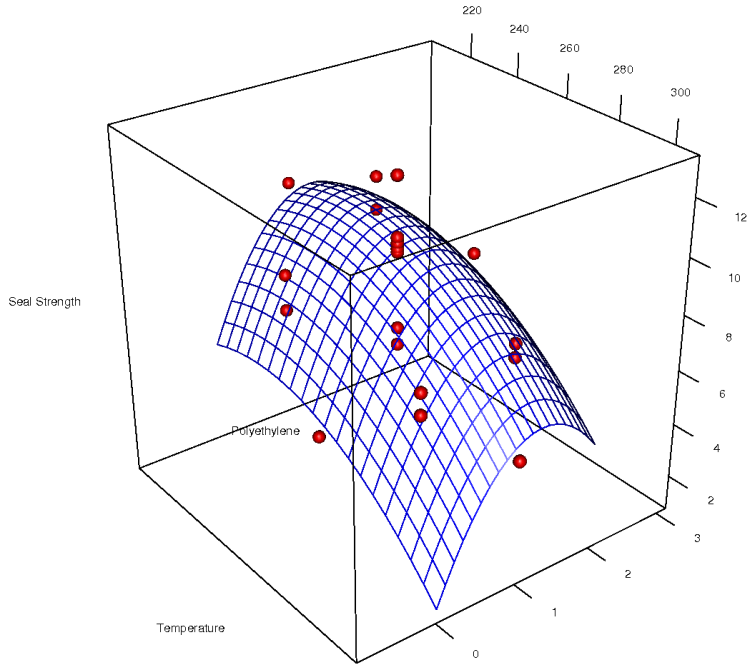
```
1 bw = read.table("breadwrapper.txt",header=T,sep='')
2 fit_bw=lm(Seal_Strength~sealtemp+polyethylene,data=bw)
3 summary(fit_bw)
```



We could also fit a quadratic (i.e. curved) surface in 3 dimensions :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \beta_3 x_{i2} + \beta_4 x_{i2}^2 + \varepsilon_i$$

```
1 fit2_bw=lm(Seal_Strength~sealtemp+I(sealtemp^2)+polyethylene
      +I(polythylene^2),data=bw)
2 summary(fit2_bw)
```



Back to the LDL model

Statistics of Least Squares

What have we got at this stage?

- Some idea that the relationship between the weight and LDL might be approx. 1 to 2 (i.e. the slope is nearly 2).
- Since this is an experiment to find something out, there presumably are hypotheses to be tested. For example, here the hypothesis might be that there is no linear relationship between weight and LDL level. Or even that the relationship between weight and LDL is linear with intercept = 0, and slope = 1?
- Look at the scatterplot again - visually there is a relationship but it is not perfect - i.e. all the points are not an a straight line.

- There are also different results for LDL given for the same weight in different children - so they are not the same. However, this is just one small sample of obese children - the same experiment using a different sample of obese children may give different results.

So, perhaps a more reasonable question is that weight and LDL are linearly related on average.
This is where statistics comes in.

$$\begin{aligned} E[y] &= \beta_0 + \beta_1 x + E[\varepsilon] \\ &= \beta_0 + \beta_1 x \end{aligned} \tag{3}$$

- This equates to saying that the 'true' average relationship is linear (i.e. in exactly on the LS line).

- So what is the rest? = is random noise (or experimental error) etc.
  Importantly, therefore, each of the individual distances (i.e. the ε's or **ERROR**s) are random variables (RVs) with expectation zero, and some variance σ^2.
- This is equivalent to saying that y is a RV with mean $= \beta_0 + \beta_1 x$ and some variance σ^2.
- Note also, that we are assuming that the X's are fixed and known - i.e. they are constants and are NOT RVs.

Lets say you are appraising this research - perhaps the first question is as follows;

Does this experiment supply any reasonable evidence that the weight of a child in linearly related (on average) to the LDL level in their blood stream?

The classical regression model assumption are that the y's are independent normal RV's and have the same variance σ^2, but with differing means given by $\beta_0 + \beta_1 x_i$, i.e.:

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2) \qquad \Rightarrow \varepsilon_i \sim N(0, \sigma^2) \tag{4}$$

Note, β_1 is not a random variable but it's estimate $\hat{\beta}_1$ is.

We can show the following (under fairly general conditions);

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

This immediately gives us a method for hypothesis testing and confidence interval estimation, e.g.

$$\begin{aligned} H_0 : \beta_1 &= \beta_1^0, \qquad e.g.\ \beta_1 = 0 \\ H_a : \beta_1 &\neq \beta_1^0, \qquad e.g.\ \beta_1 \neq 0 \end{aligned}$$

$$z = \frac{\hat{\beta}_1 - \beta_1^0}{\sqrt{\sigma^2 / S_{xx}}} \qquad \Rightarrow z \sim N(0,1) \quad [under H_0 :]$$

But, we don't know the value of σ^2! Therefore we have to estimate it and this has theoretical consequences - i.e. we can't use a simple z-test!

Estimating the variance

The model says that the errors are *i.i.d.* normal RVs with mean zero.

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

We estimate errors using the residuals:

$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

and use the observed sample variance of the residuals as an estimate for σ^2.

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

Note: the denominator here is $n-2$ as opposed to the usual $n-1$. This is because two degrees of freedom are lost to the two estimated parameters, one for the slope and one for the intercept. using the correct degrees of freedom makes s^2 an unbiased estimator.

R gives the following summary:

```
> summary(fit_ldl)

Call:
lm(formula = LDL ~ Weight, data = ldl)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -48.7814    30.8577  -1.581    0.138
Weight        1.9378     0.3486   5.559 9.25e-05 ***
---

Residual standard error: 11.13 on 13 degrees of freedom
Multiple R-squared:  0.7039,  Adjusted R-squared:  0.6811
F-statistic:  30.9 on 1 and 13 DF,  p-value: 9.248e-05
```

So, $s^2 = 11.13^2 = 123.9$

Inference on the Slope

What happens when you use s^2 in place of σ^2 in a z-test?

Solution is from William Sealy Gosset (1908) aka Student working in Jame's Gate.

$$H_0 : \beta_1 = \beta_1^0$$

$$t = \frac{\hat{\beta}_1 - \beta_1^0}{\sqrt{s^2/S_{xx}}} \sim t_{(n-2)} \quad (5)$$

Where $t_{(n-2)}$ is Student's t-distribution with $n - 2$ degrees of freedom.

Using R we get the following from the `lm(...)` function:

```
> ldl = read.table("ldldata.txt",header=T,sep=',')
> attr(ldl,'names')=c("Weight","LDL")
> fit_ldl=lm(LDL~Weight,data=ldl)
> summary(fit_ldl)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -48.7814    30.8577   -1.581    0.138
Weight        1.9378     0.3486    5.559 9.25e-05
```

What are the conclusions about the hypothesis test concerning the slope?

How would a rejection region approach have been applied here?

CI for slope

Confidence Intervals (CIs) for the slope can also be derived from the t distribution and the variance for the $\hat{\beta}_1$. A $(1-\alpha)\%$ CI for the slope is given by,

$$\hat{\beta}_1 \quad \pm \quad t_{1-\alpha/2,n-2}\sqrt{\frac{s^2}{S_{xx}}} \tag{6}$$

```
> confint(fit_ldl)
                    2.5 %     97.5 %
(Intercept) -115.445298 17.882537
Weight         1.184652  2.690871
```


Inference on the Intercept

Conducting statistical inference on the intercept is only occasionally of interest.

Nevertheless, here are the relevant formulae:

Hypothesis testing:

$$H_0 : \beta_0 \quad = \quad \beta_0^0$$

$$t \quad = \quad \frac{\hat{\beta}_0 - \beta_0^0}{s\sqrt{(1/n) + (\bar{x}^2/S_{xx})}} \sim t_{(n-2)} \tag{7}$$

Confidence intervals:

$$\hat{\beta}_0 \quad \pm \quad t_{1-\alpha/2,n-2}\sqrt{s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)} \tag{8}$$

Exercise

Given the following data answer the questions below:

x	1	2	3	4	5
y	5	7	9	8	17

1. Find the equation of the LS line to these data by hand.
2. Test the null hypothesis that there is no simple linear relationship between the x and y - also by hand.
3. Confirm you answer using R .

Regression & Distributional Assumptions

People tend to get caught up in these - and admittedly they are not unimportant.

But, there are generally alternatives/work-arounds when we suspect the assumptions are being violated.

E.G. Let's assume that for the LDL data that our $H_0 : \beta_1 = 0$ is true. That means that response is unrelated to predictor (ldl level to weight). If this as the case, then ldl level can be viewed as a random allocation with respect to weight.

So, what if we did a random allocation of ldl level to weight and then fit a regression model and see if the slope from that fit is similar to the one we calculated from the original data.

If the slopes are similar we might conclude that the slope calculated from the original data is consistent with our null hypothesis.

If the slopes are very different we might conclude that the null hypothesis is not well supported by the data and therefore reject it.

Try the following code:

```
## ldl randomisation models
## set random seed for reproducible random number stream
set.seed(2987887)
## simple random permutation of ldl values
ldl_random=sample(ldl$LDL)
## fit model to randomly allocated responses
fit_random=lm(ldl_random~ldl$Weight)
## extract slope from model fit
coef(fit_random)[2]
```

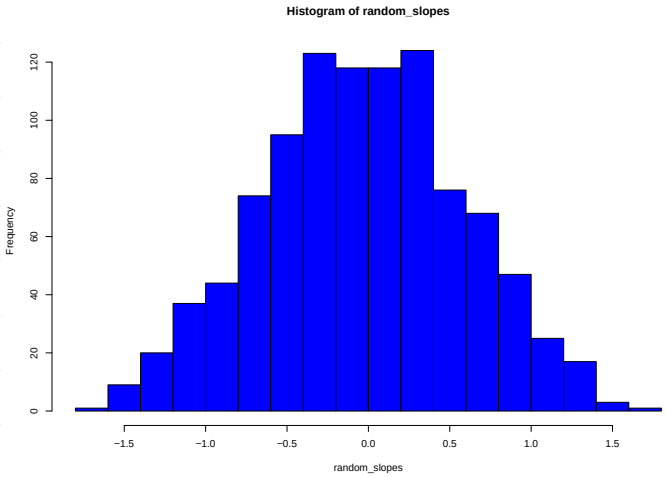
```
> coef(fit_random)[2]
ldl$Weight
 -0.719796
```


The one result we have is not very close to 1.94, but maybe this is because the random allocation is an unusual one.

Solution, repeat the random allocation say 1,000 time and calculate what proportion of the results are outside the range ± 1.94 - we can treat this proportion as a p-value based on randomisation!

```
### ldl randomisation models
## set random seed for reproducible random number stream
set.seed(2987887)
## create a vector called random_slopes to store the 1,000
    results
random_slopes=NA
## repeat the random allocation independently 1,000 times
    using a 'for' loop
for(i in 1:1000){ ## start of 'for' loop on this line
  ## simple random permutation of ldl values
  ldl_random=sample(ldl$LDL)
  ## fit model to randomly allocated responses
  fit_random=lm(ldl_random~ldl$Weight)
  ## extract slope from model fit
  random_slopes[i]=coef(fit_random)[2]
} ## end of 'for' loop on this line
summary(random_slopes)
```

And the summary statistics for the random slopes are:




```
> summary(random_slopes)
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-1.867000 -0.445200 -0.003923 -0.007243  0.427600  1.859000
```

Giving a randomisation p-value of zero.

Appendix

Matrix Formulation of LS Model

We could re-formulate the dose-response model using matrix algebra.

$$\overbrace{\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \\ y_{11} \end{pmatrix}}^{Y} = \overbrace{\begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ 1 & x_4 & x_4^2 \\ 1 & x_5 & x_5^2 \\ 1 & x_6 & x_6^2 \\ 1 & x_7 & x_7^2 \\ 1 & x_8 & x_8^2 \\ 1 & x_9 & x_9^2 \\ 1 & x_{10} & x_{10}^2 \\ 1 & x_{11} & x_{11}^2 \end{pmatrix}}^{X} \overbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}}^{\beta} + \overbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \end{pmatrix}}^{\varepsilon}$$

Using the actual data we get:

$$\overbrace{\begin{pmatrix} 991 \\ 1233 \\ 1229 \\ 1684 \\ 1862 \\ 1919 \\ 2082 \\ 1776 \\ 1528 \\ 881 \\ 1101 \end{pmatrix}}^{Y} = \overbrace{\begin{pmatrix} 1 & 10.45 & 109.20 \\ 1 & 12.57 & 158.00 \\ 1 & 15.62 & 243.98 \\ 1 & 25.98 & 674.96 \\ 1 & 30.52 & 931.47 \\ 1 & 34.06 & 1160.08 \\ 1 & 41.17 & 1694.97 \\ 1 & 50.78 & 2578.61 \\ 1 & 61.01 & 3722.22 \\ 1 & 71.76 & 5149.50 \\ 1 & 79.20 & 6272.64 \end{pmatrix}}^{X} \overbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}}^{\beta} + \overbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \end{pmatrix}}^{\varepsilon}$$

and with the vectors Y, β, ε and the matrix X so defined, we can write very succinctly:

$$Y = X\beta + \varepsilon \tag{9}$$

Using these, and being careful about our matrix/vector dimensions and multiplication we get:

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta) \tag{10}$$

NB: using A' to indicate the transpose of A. Also, I'm using uppercase for both vectors and matrices.

Now, we need to minimise a scalar Q with respect to the elements of the vector β.

Derivative of Q WRT to (elements of) a vector

- In reality what we are doing is taking the partial derivatives WRT to each element in the vector and stacking them in a column vector.
- The specialised notation used here consists of the following: (a) treat the vector as though it were a scalar and apply the usual rules of taking derivatives, but (b) keep a careful account of the dimensions of all the vectors and matrices in the resulting expressions to ensure they are correct/comfortable.

$$\begin{aligned} Q &= (Y - X\beta)'(Y - X\beta) \\ \Rightarrow \quad \frac{\partial Q}{\partial \beta} &= -2X'Y + 2X'X\beta = 0 \\ \Rightarrow \quad \hat{\beta} &= (X'X)^{-1}X'Y \end{aligned}$$

Applying the formula to the dose-response data we get the following:

$$\overbrace{\begin{pmatrix} 1.2632141 & -0.0639283 & 0.0006518 \\ -0.0639284 & 0.0037908 & -0.0000414 \\ 0.0006518 & -0.0000414 & 0.0000005 \end{pmatrix}}^{(X'X)^{-1}} \overbrace{\begin{pmatrix} 16286 \\ 630536 \\ 30939083 \end{pmatrix}}^{X'Y} = \overbrace{\begin{pmatrix} 430.0754 \\ 69.5021 \\ -0.8172 \end{pmatrix}}^{\hat{\beta}}$$

So, the LS equation for the quadratic model to these data is:

$$y = 430.0754 + 69.5021x - 0.8172x^2$$

What happens if you decide, that a cubic model might be better than quadratic model for the dose-response data?

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i$$

Now we have four parameters to estimate. Define the following:

$$\overbrace{\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \\ y_{11} \end{pmatrix}}^{Y} = \overbrace{\begin{pmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ 1 & x_3 & x_3^2 & x_3^3 \\ 1 & x_4 & x_4^2 & x_4^3 \\ 1 & x_5 & x_5^2 & x_5^3 \\ 1 & x_6 & x_6^2 & x_6^3 \\ 1 & x_7 & x_7^2 & x_7^3 \\ 1 & x_8 & x_8^2 & x_8^3 \\ 1 & x_9 & x_9^2 & x_9^3 \\ 1 & x_{10} & x_{10}^2 & x_{10}^3 \\ 1 & x_{11} & x_{11}^2 & x_{11}^3 \end{pmatrix}}^{X} \overbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}}^{\beta} + \overbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \end{pmatrix}}^{\varepsilon}$$

Notice that the equations can be written exactly the same as before, i.e.:

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

$$\begin{aligned} \Rightarrow Q &= \varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta) \\ \Rightarrow & \frac{\partial Q}{\partial \beta} = -2X'Y + 2X'X\beta = 0 \\ \Rightarrow & \hat{\beta} = (X'X)^{-1}X'Y \end{aligned}$$

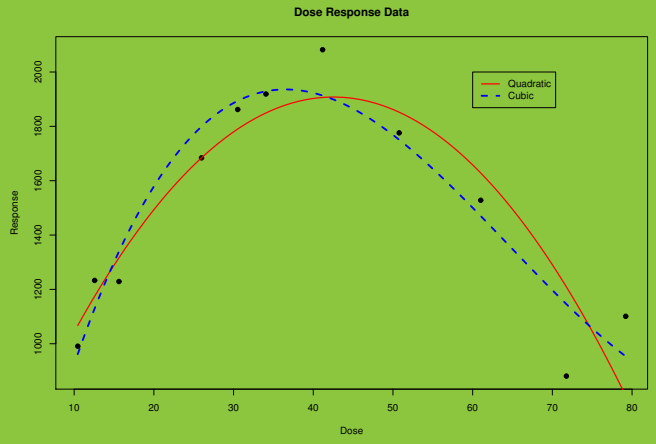
The only thing that has changed are the dimensions of the X matrix and the β vector (one extra column and row respectively).

The model from these specifications for the X matrix and β vector is:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i$$

Performing the matrix algebra we get:

$$\hat{\beta} = \begin{pmatrix} -157.87 \\ 132.06 \\ -2.52 \\ 0.013 \end{pmatrix}$$



Bread-wrapper Data

We can use exactly the same matrix algebra in fitting a model to the Bread-wrapper data.

Seal Strength	Sealing Temp.	% polyethylene
6.6	225	0.5
6.9	285	0.5
7.9	225	0.5
6.1	285	0.5
9.2	225	1.7
6.8	285	1.7
10.4	225	1.7
7.3	285	1.7
9.8	204.5	1.1
5	305.5	1.1
6.9	255	1.1
6.3	255	1.1
4	255	0.09
8.6	255	2.11
10.1	255	1.1
9.9	255	1.1
12.2	255	1.1
9.7	255	1.1
9.7	255	1.1
9.6	255	1.1

The model we are fitting is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

where x_{i1} is the sealing temperature for observation i and x_{i2} is the corresponding % of polyethylene used.

This is a plane in 3 dimensions.

$$\overbrace{\begin{pmatrix} 6.6 \\ 6.9 \\ 7.9 \\ 6.1 \\ 9.2 \\ 6.8 \\ 10.4 \\ 7.3 \\ 9.8 \\ 5 \\ 6.9 \\ 6.3 \\ 4 \\ 8.6 \\ 10.1 \\ 9.9 \\ 12.2 \\ 9.7 \\ 9.7 \\ 9.6 \end{pmatrix}}^{Y} = \overbrace{\begin{pmatrix} 1 & 225 & 0.5 \\ 1 & 285 & 0.5 \\ 1 & 225 & 0.5 \\ 1 & 285 & 0.5 \\ 1 & 225 & 1.7 \\ 1 & 285 & 1.7 \\ 1 & 225 & 1.7 \\ 1 & 285 & 1.7 \\ 1 & 204.5 & 1.1 \\ 1 & 305.5 & 1.1 \\ 1 & 255 & 1.1 \\ 1 & 255 & 1.1 \\ 1 & 255 & 0.09 \\ 1 & 255 & 2.11 \\ 1 & 255 & 1.1 \\ 1 & 255 & 1.1 \\ 1 & 255 & 1.1 \\ 1 & 255 & 1.1 \\ 1 & 255 & 1.1 \\ 1 & 255 & 1.1 \end{pmatrix}}^{X} \overbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}}^{\beta} + \overbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{15} \\ \varepsilon_{16} \\ \varepsilon_{17} \\ \varepsilon_{18} \\ \varepsilon_{19} \\ \varepsilon_{20} \end{pmatrix}}^{\varepsilon}$$

Once again we get the solution using the LS criterion:

$$\hat{\beta} = (X'X)^{-1}X'Y = \begin{pmatrix} 15.658 \\ -0.037 \\ 1.700 \end{pmatrix}$$