VE401 Recitation 7

Simple Linear Regression

Example: We want to do physical experiment to measure the gravitational acceleration $g \approx 9.8$ m/s^2. We have several weights of 10g each. We use springs to measure their gravitational forces:

In[584]:=
```
X = {10, 20, 30, 40, 50, 60, 70, 80};
```

In[•]:=
```
Residual = RandomVariate[NormalDistribution[0, 0.05], Length[X]]
```

Out[•]=
```
{-0.0115325, 0.0157878, 0.0294766,
 -0.0597483, 0.0173925, 0.0349411, 0.0903353, 0.122338}
```

In[585]:=
```
Y = 0.0098 * X + Residual
```

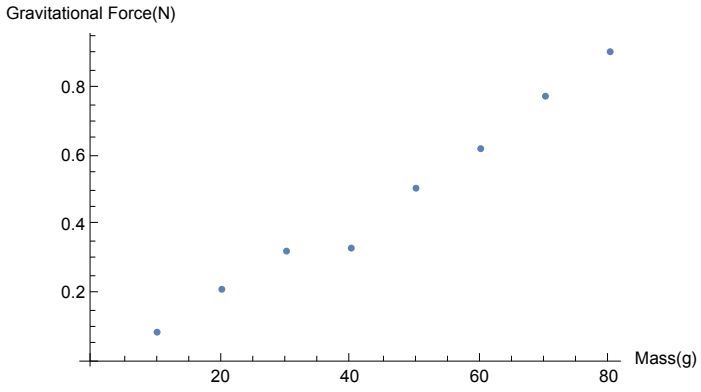
Out[•]=
```
{0.0864675, 0.211788, 0.323477, 0.332252, 0.507393, 0.622941, 0.776335, 0.906338}
```

In[844]:=
```
Data = Transpose[{X, Y}]
```

Out[•]=
```
{{10, 0.0864675}, {20, 0.211788}, {30, 0.323477}, {40, 0.332252},
 {50, 0.507393}, {60, 0.622941}, {70, 0.776335}, {80, 0.906338}}
```

In[587]:=
```
ListPlot[Data, AxesLabel → {"Mass(g)", "Gravitational Force(N)"}]
```

Out[•]=



Setting and Assumptions

Setting: we have

- An independent variable X, or predictor variable or regressor,
- A dependent variable Y, or a response variable.

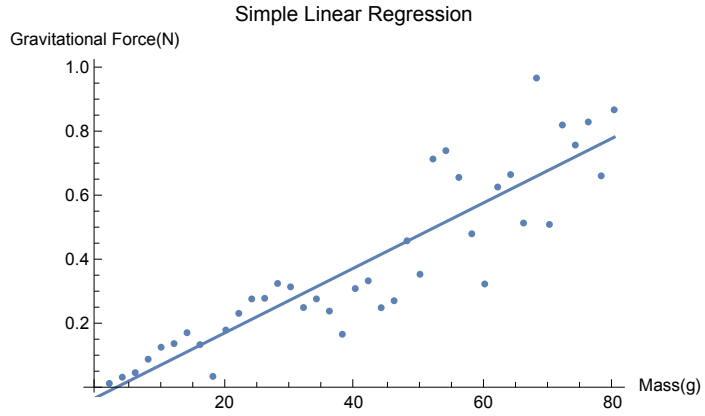
Assumptions:

- Y is approximately linearly dependent on X, $Y_i = \beta_0 + \beta_1 x_i + E_i$ for some β_0, $\beta_1 \in \mathbb{R}$ for all x.

- The error term E_i follows a normal distribution with mean 0 and variance σ^2. This means that $Y|x$ is also normal with variance σ^2 and mean $\mu_{Y|x} = \beta_0 + \beta_1 x$.

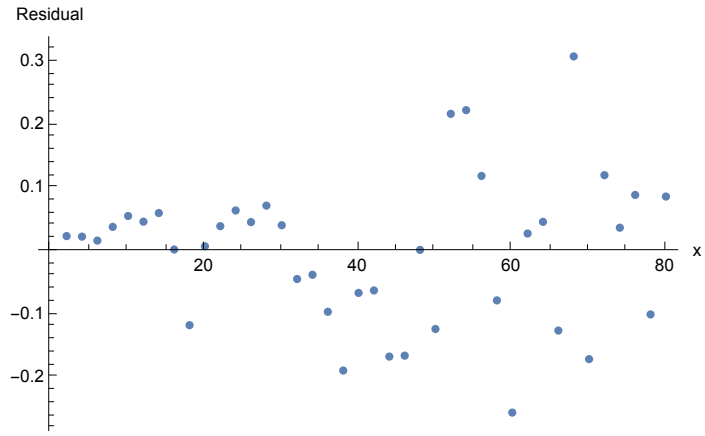
If my spring has more variance when the mass is large, the data may look like:

Out[•]=



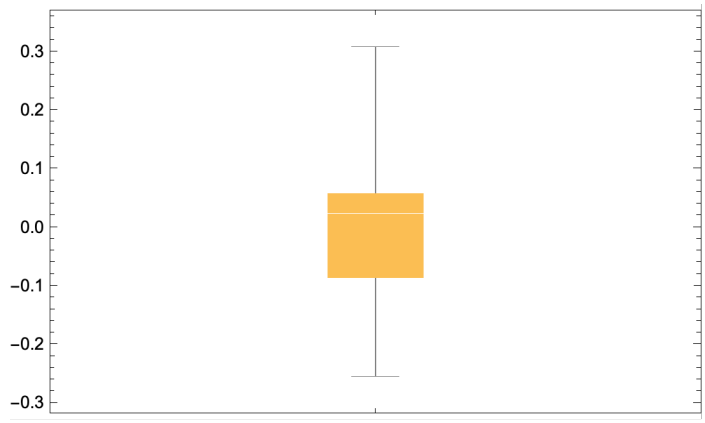
In[•]:=
```
ListPlot[Transpose[{Xalt, lmAlt["FitResiduals"]}],
 AxesLabel → {"x", "Residual"}]
```

Out[•]=



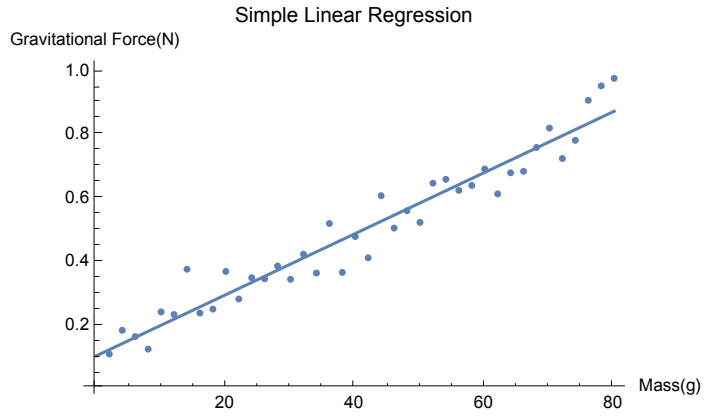
In[•]:=
```
BoxWhiskerChart[lmAlt["FitResiduals"]]
```

Out[•]=



If the spring has wrong labels, the data may look like:

Out[•]=



In[•]:= **lmMov**

Out[•]= FittedModel[0.10274 + 0.00963947 m]

- $Y \mid x_i$ and $Y \mid x_j$ are independent if $x_i \neq x_j$.

Least-Squares Estimation

Intuition: we want to find b_0 and b_1 such that $\text{SS}_E := \sum_{i=1}^n (y_i - (b_0 + b_1 x_i))^2 = S_{yy} - b_1 S_{xy}$ is minimized.

Mathematical representation:

$$b_1 = \frac{n \sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \qquad \left(\text{point estimate of } \beta_1 = \frac{\text{Cov}[X,Y]}{\text{Var}[X]}\right)$$

$$b_0 = \frac{1}{n}\sum_{i=1}^n y_i - b_1 \cdot \frac{1}{n}\sum_{i=1}^n x_i = \bar{y} - b_1 \bar{x} \qquad \left(\text{point estimate of } \beta_0 = \text{E}[Y] - \beta_1 \text{E}[X]\right)$$

where

$$S_{xx} := \sum_{i=1}^n x_i^2 - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 \qquad (\text{estimates } (n-1)\,\text{Var}[X])$$

$$S_{yy} := \sum_{i=1}^n y_i^2 - \frac{1}{n}\left(\sum_{i=1}^n y_i\right)^2 = \sum_{i=1}^n (y_i - \bar{y})^2 \qquad (\text{estimates } (n-1)\,\text{Var}[Y])$$

$$S_{xy} := \sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i) = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \qquad (\text{estimates } (n-1)\,\text{Cov}[X, Y])$$

In[•]:= **Data**

Out[•]= {{10, 0.0864675}, {20, 0.211788}, {30, 0.323477}, {40, 0.332252}, {50, 0.507393}, {60, 0.622941}, {70, 0.776335}, {80, 0.906338}}

Example: For our data, we can calculate using the statistics function of Casio calculators

Calculate S_{xx}, S_{yy} and S_{xy}.

We have

In[•]:=
```
Covariance[X, Y] * (8 - 1)
```

Out[•]=
```
48.1768
```

$$S_{xx} := \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = 20\,400 - \frac{1}{8}(360)^2 = 4200$$

$$S_{yy} := \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 = 2.337 - \frac{1}{8}(3.767)^2 = 0.5632$$

$$S_{xy} := \sum_{i=1}^{n} x_i\, y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) = 217.69 - \frac{1}{8}\,360 \times 3.767 = 48.177$$

Calculate the estimation of β_0 and β_1 of the model.

We have

$$b_1 = \frac{n\sum_{i=1}^{n} x_i\, y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{8\times 217.69 - 360\times 3.77}{8\times 20\,400 - (360)^2} = 0.0114$$

$$b_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - b_1 \cdot \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{y} - b_1\,\overline{x} = 0.47 - 0.0114 \times 45 = -0.043$$

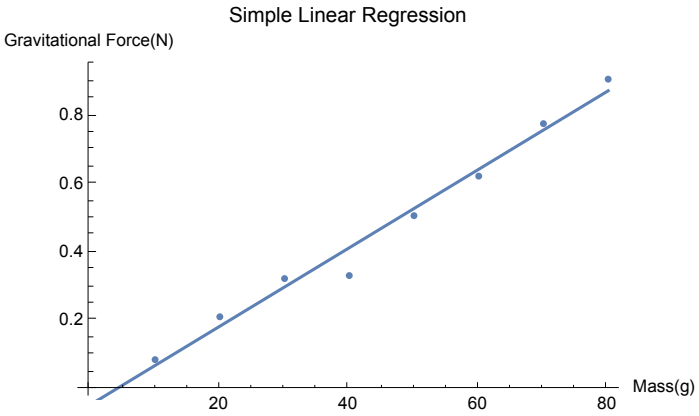
which is fairly close to the true value of $\beta_1 = 0.0098$, $\beta_0 = 0$.

In[•]:=
```
lm = LinearModelFit[Data, m, m]
```

Out[•]=
```
FittedModel[ -0.0453065 + 0.0114707 m ]
```

In[•]:=
```
Show[ListPlot[Data], Plot[lm[x], {x, 0, 80}],
 PlotLabel → "Simple Linear Regression",
 AxesLabel → {"Mass(g)", "Gravitational Force(N)"}]
```

Out[•]=



The $SS_E = S_{yy} - b_1\, S_{xy} = 0.5632 - 0.01147 \times 48.177 = 0.0106$.

```
In[•]:= Total[lm["FitResiduals"]^2]
Out[•]= 0.010609
```

Distribution of the Least Squares Estimator

Theorem:

Estimator	Distribution	Mean	Variance
B_1 (estimator of β_1)	Normal	β_1	$\frac{\sigma^2}{S_{xx}}$
B_0 (estimator of β_0)	Normal	β_0	$\sigma^2 \frac{\sum_{i=1}^n x_i}{n S_{xx}}$

Important property: $\sum_{i=1}^n (x_i - \overline{x}) = 0$, $\sum_{i=1}^n (x_i - \overline{x})^2 = \sum_{i=1}^n (x_i - \overline{x}) x_i + \overline{x} \sum_{i=1}^n (x_i - \overline{x}) = \sum_{i=1}^n (x_i - \overline{x}) x_i$.

Proof: For B_1, we have

$$\begin{aligned} B_1 &= \frac{\sum_{i=1}^n (x_i-\overline{x})\left(Y_i-\overline{Y}\right)}{\sum_{i=1}^n (x_i-\overline{x})^2} \\ &= \frac{\sum_{i=1}^n (x_i-\overline{x})\, Y_i - \overline{Y} \sum_{i=1}^n (x_i-\overline{x})}{\sum_{i=1}^n (x_i-\overline{x})^2} \\ &= \frac{\sum_{i=1}^n (x_i-\overline{x})\, Y_i}{\sum_{i=1}^n (x_i-\overline{x})^2} = \sum_{i=1}^n \frac{(x_i-\overline{x})}{\sum_{j=1}^n (x_j-\overline{x})^2} Y_i = : \sum_{i=1}^n c_i\, Y_i \end{aligned}$$

Therefore B_1 is a linear combination of i.i.d. normal distributed RV, so B_1 is normally distributed. furthermore,

$$\begin{aligned} B_1 &= \frac{\sum_{i=1}^n (x_i-\overline{x})\,(\beta_0+\beta_1 x_i+E_i)}{\sum_{i=1}^n (x_i-\overline{x})^2} \\ &= \beta_1 \frac{\sum_{i=1}^n (x_i-\overline{x})\, x_i}{\sum_{i=1}^n (x_i-\overline{x})^2} + \frac{\sum_{i=1}^n (x_i-\overline{x})\, E_i}{\sum_{i=1}^n (x_i-\overline{x})^2} \\ &= \beta_1 + \frac{\sum_{i=1}^n (x_i-\overline{x})\, E_i}{\sum_{i=1}^n (x_i-\overline{x})^2} \end{aligned}$$

Therefore the expectation is given by

$$\begin{aligned} \mathrm{E}[B_1] &= \beta_1 + \mathrm{E}\left[\frac{\sum_{i=1}^n (x_i-\overline{x})\, E_i}{\sum_{i=1}^n (x_i-\overline{x})^2}\right] \\ &= \beta_1 + \frac{\sum_{i=1}^n (x_i-\overline{x})\, \mathrm{E}[E_i]}{\sum_{i=1}^n (x_i-\overline{x})^2} \\ &= \boxed{\beta_1} \end{aligned}$$

and

$$\begin{aligned}\text{Var}[B_1] &= 0 + \text{Var}\left[\frac{\sum_{i=1}^n (x_i-\bar{x})\,E_i}{\sum_{i=1}^n (x_i-\bar{x})^2}\right] \\ &= \sum_{i=1}^n c_i^2\,\text{Var}[E_i] \\ &= \sum_{i=1}^n \left(\frac{(x_i-\bar{x})}{\sum_{j=1}^n (x_j-\bar{x})^2}\right)^2 \sigma^2 \\ &= \frac{\sum_{i=1}^n (x_i-\bar{x})^2}{\left(\sum_{j=1}^n (x_j-\bar{x})^2\right)^2}\,\sigma^2 \\ &= \frac{\sigma^2}{\sum_{j=1}^n (x_j-\bar{x})^2} = \boxed{\frac{\sigma^2}{S_{xx}}}\end{aligned}$$

Verify the same theorem for $B_0 = \overline{Y} - B_1\,\bar{x}$.

For B_0, since $B_0 = \overline{Y} - B_1\,\bar{x}$ and $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i = \frac{1}{n}\sum_{i=1}^n (\beta_0 + \beta_1 x_i + E_i) = \beta_0 + \beta_1\,\frac{1}{n}\sum_{i=1}^n x_i + \frac{1}{n}\sum_{i=1}^n E_i$ we have

$$\begin{aligned}\text{E}[B_0] &= \text{E}[\overline{Y}] - \bar{x}\,\text{E}[B_1] \\ &= \beta_0 + \beta_1\,\bar{x} - \beta_1\,\bar{x} \\ &= \boxed{\beta_0}\end{aligned}$$

and

$$\begin{aligned}\text{Var}[B_0] &= \text{Var}[\overline{Y}] + (\bar{x})^2\,\text{Var}[B_1] \\ &= \frac{\sigma^2}{n} + \frac{\sigma^2\left(\frac{1}{n}\sum_{i=1}^n x_i\right)^2}{S_{xx}} \\ &= \sigma^2\left(\frac{1}{n} + \frac{\left(\frac{1}{n}\sum_{i=1}^n x_i\right)^2}{\sum_{i=1}^n x_i^2 - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)^2}\right) \\ &= \sigma^2\left[\frac{1}{n} + \left(-\frac{1}{n} + \frac{\frac{1}{n}\sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i^2 - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)^2}\right)\right] \\ &= \boxed{\sigma^2\,\frac{\sum_{i=1}^n x_i}{n\,S_{xx}}}\end{aligned}$$

If the mean of error is not 0, are the estimators B_0, B_1 still unbiased?

They will no longer be unbiased.

If the variance of error is not constant, but the mean is 0, are the estimators B_0, B_1 still unbiased?

In this case they are still unbiased.

Least Squares Estimator for the Variance

The variance σ^2 of $Y \mid x$ can be estimated by

$$S^2 := \frac{\text{SS}_E}{n-2} = \frac{1}{n-2}\sum_{i=1}^n (y_i - (b_0 + b_1\,x_i))^2$$

And

$$\frac{(n-2)\,S^2}{\sigma^2} = \frac{\mathrm{SS}_E}{\sigma^2} \approx \chi^2_{n-2}$$

Using this we can calculate the $100\,(1-\alpha)\,\%$ confidence interval

$$B_1 \pm t_{\alpha/2,n-2}\,\frac{S}{\sqrt{S_{xx}}} \qquad B_0 \pm t_{\alpha/2,n-2}\,\frac{S\sqrt{\sum_{i=1}^n x_i}}{\sqrt{n\,S_{xx}}}$$

Calculate the confidence interval for the parameter of previous model.

We can calculate $S^2 = \mathrm{SS}_E/(n-2) = 0.106/6 = 0.0176$.

```
In[•]:= lm["EstimatedVariance"]
Out[•]= 0.00176817
In[•]:= lm["ParameterConfidenceIntervals", ConfidenceLevel → 0.95]
Out[•]= {{-0.125479, 0.0348662}, {0.00988302, 0.0130583}}
```

Test for Significance of Regression

Testing Parameter	Null Hypothesis	Test Statistics
β_1	$H_0 : \beta_1 = 0$	$T_{n-2} = \frac{B_1}{S/\sqrt{S_{xx}}} = \frac{R}{\sqrt{1-R^2}}\sqrt{n-2}$

We reject H_0 if $T_{n-2} > t_{\alpha/2,n-2}$.

Test the significance of regression of our previous model.

```
In[•]:= lm["ParameterTable"]
```

	Estimate	Standard Error	t-Statistic	P-Value
1	−0.0453065	0.0327648	−1.38278	0.215995
x	0.0114707	0.00064884	17.6787	2.10336×10^{-6}

We conclude that there is NO evidence that the intercept is not 0, and there is evidence that the slope is not 0.

Distribution of Estimated Mean (Predictor)

Estimator	Distribution	Mean	Variance
$\hat{Y}\,\vert\,x$ (estimator of $\mu_{Y\vert x}$)	Normal	$\mu_{Y\vert x}$	$\sigma\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}$
$\hat{Y}\,\vert\,x - Y\,\vert\,x$	Normal	0	$\left(1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}\right)\sigma^2$

The $100\,(1-\alpha)\,\%$ ***confidence interval*** for the conditional mean is

$$\hat{\mu}_{Y|x} \pm t_{\alpha/2,n-2}\, S \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$$

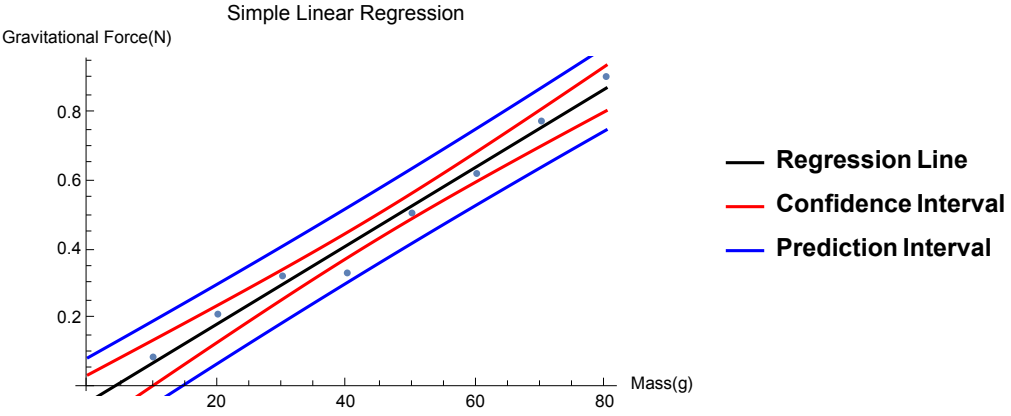
With $100\,(1-\alpha)\,\%$ chance, the conditional mean $\mu_{Y|x}$ will lie in this interval.

The $100\,(1-\alpha)\,\%$ ***prediction interval*** for the observed value is

$$\hat{\mu}_{Y|x} \pm t_{\alpha/2,n-2}\, S \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$$

With $100\,(1-\alpha)\,\%$ chance, the newly observed value $Y \mid x_{\text{new}}$ will lie in this interval.

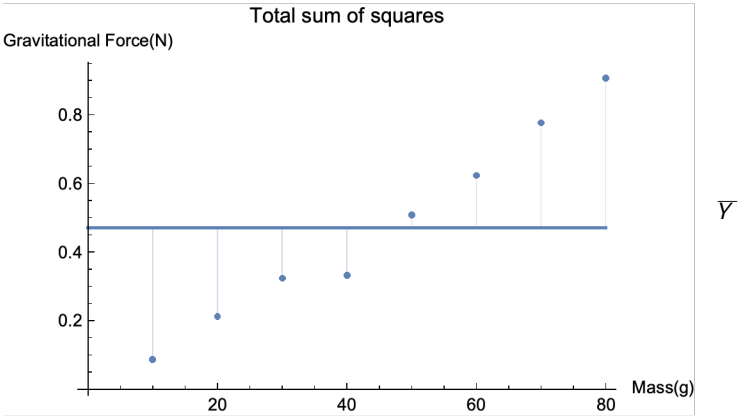
In[•]:=
```
conf = lm["MeanPredictionBands", ConfidenceLevel → 0.95];
pred = lm["SinglePredictionBands", ConfidenceLevel → 0.95];
Show[ListPlot[Data],
 Plot[{lm[x], conf, pred}, {x, 0, 80}, PlotStyle → {Black, Red, Red, Blue, Blue},
  PlotLegends → Automatic], PlotLabel → "Simple Linear Regression",
 AxesLabel → {"Mass(g)", "Gravitational Force(N)"}]
```



R^2

Intuition: We want to see how much variance is explained by our model.

Mathematical representation: The ***total sum of squares*** $\text{SS}_T = S_{yy} = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$, representing the variance of Y.

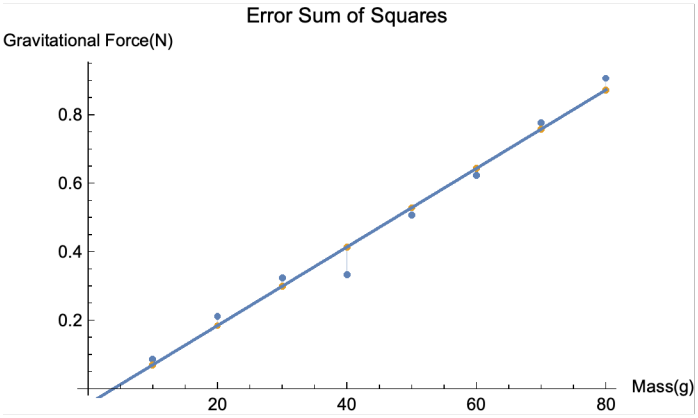
Out[•]=


$\overline{Y}$

The ***error sum of squares*** $\text{SS}_E = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 x))^2$, representing the variance of Y that

remains after applying the model.

Out[•]=



The ***correlation of determination*** is

$$R^2 = \frac{SS_T - SS_E}{SS_T} = \frac{S_{xy}^2}{S_{xx}\, S_{yy}}$$

Note. R^2 is the square of the estimator of ρ_{XY}.

Calculate R^2 of our model.

The R^2 value of our model is $\frac{S_{xy}^2}{S_{xx}\, S_{yy}} = \frac{48.177^2}{4200 \times 0.5632} = 0.9812$

In[•]:=
```
lm["RSquared"]
```

Out[•]=
```
0.981164
```

Testing for Lack of Fit

Testing Parameter	Null Hypothesis	Test Statistics
$SS_{E,\text{lf}}$	H_0 : the linear regression model is appropriate (the error due to lack of fit is small)	$F_{k-2,n-k} = \frac{SS_{E,\text{lf}}/(k-2)}{SS_{E,\text{pe}}/(n-k)}$

We reject H_0 if $F_{k-2,n-k} > f_{\alpha,k-2,n-k}$.

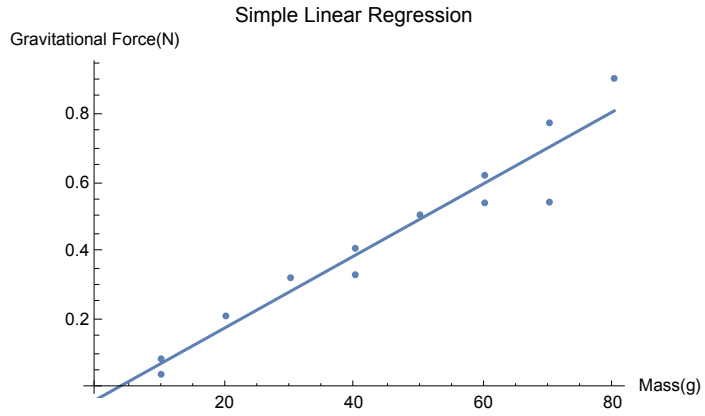
Example: We perform 4 more experiments to measure $SS_{E,\text{pe}}$,

Out[•]=
```
{{10, 0.0864675}, {20, 0.211788}, {30, 0.323477}, {40, 0.332252},
 {50, 0.507393}, {60, 0.622941}, {70, 0.776335}, {80, 0.906338},
 {10, 0.0412471}, {40, 0.409575}, {60, 0.54236}, {70, 0.544538}}
```

Steps:

- Perform linear regression on the set of data,

Out[•]=



- Identify that there are k distinct x values. In this case $k = 8$.

In[•]:=
```
lmNew
```

Out[•]=
```
FittedModel[ -0.0324704 + 0.0105451 x ]
```

- Calculate for the i^{th} group ($i = 1, \ldots, k$) of repeated x the ***internal sum of squares***, $\sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_i\right)^2$.

$$\text{For } n = 10,\ \sum_{j=1}^{2} \left(Y_{1,j} - \overline{Y}_1\right)^2 = 0.001$$

$$\text{For } n = 40,\ \sum_{j=1}^{2} \left(Y_{2,j} - \overline{Y}_2\right)^2 = 0.003$$

$$\text{For } n = 60,\ \sum_{j=1}^{2} \left(Y_{3,j} - \overline{Y}_3\right)^2 = 0.003$$

$$\text{For } n = 70,\ \sum_{j=1}^{2} \left(Y_{4,j} - \overline{Y}_4\right)^2 = 0.027$$

- Sum them up to get ***error sum of squares due to pure error***. $\text{SS}_{E,\text{pe}} / \sigma^2$ follows a chi-squared distribution with $n - k = 12 - 8 = 4$ degrees of freedom.

$$\text{SS}_{E,\text{pe}} = 0.001 + 0.003 + 0.003 + 0.027 = 0.034$$

- The sum of square error is calculated by

$$\text{SS}_E = S_{yy} - b_1 S_{xy} = 0.052$$

In[•]:=
```
Total[lmNew["FitResiduals"]^2]
```

Out[•]=
```
0.0515848
```

- The error due to lack of fit is

$$\begin{aligned} \text{SS}_{E,\text{lf}} &= \text{SS}_E - \text{SS}_{E,\text{pe}} \\ &= 0.052 - 0.034 \\ &= 0.018 \end{aligned}$$

$\text{SS}_{E,\text{lf}} / \sigma^2$ follows $k - 2 = 6$ degree of freedom.

- Calculate F statistics,

In[234]:=
```
1 - CDF[FRatioDistribution[6, 4], (0.018 / 6) / (0.034 / 4)]
```

Out[234]=
```
0.8771642274430168
```

There is no evidence that the linear model is not appropriate here.

Multiple Linear Regression

Settings and Assumptions

To discuss the model

$$Y\,|\,x = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + E$$

We gives the same assumptions, and define

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},\ \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{\mathrm{p}1} \\ 1 & x_{12} & & \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1\,n} & & x_{\mathrm{p}n} \end{pmatrix},\ \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix},\ \boldsymbol{E} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

We have $\boldsymbol{Y} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{E}$.

Least Squares Estimation

To minimize $\mathrm{SS}_E = (Y - X\,b)^T\,(Y - X\,b)$, we have $b = \left(X^T\,X\right)^{-1}\,X^T\,Y$.

Example: I want to check whether gravitational force is related to the square of mass, so I fit my data to the model $y = b_0 + b_1\,x + b_2\,x^2$. I calculate

In[823]:=
```
y = Transpose[Data][[2]];
x = Transpose[Table[Function[x, x^k] /@ Transpose[Data][[1]], {k, 0, 2}]];
{MatrixForm[x], MatrixForm[y]}
```

Out[•]=
$$\left\{\begin{pmatrix} 1 & 10 & 100 \\ 1 & 20 & 400 \\ 1 & 30 & 900 \\ 1 & 40 & 1600 \\ 1 & 50 & 2500 \\ 1 & 60 & 3600 \\ 1 & 70 & 4900 \\ 1 & 80 & 6400 \end{pmatrix},\ \begin{pmatrix} 0.0864675 \\ 0.211788 \\ 0.323477 \\ 0.332252 \\ 0.507393 \\ 0.622941 \\ 0.776335 \\ 0.906338 \end{pmatrix}\right\}$$

In[827]:=
```
b = Inverse[Transpose[x].x].Transpose[x].y;
MatrixForm[b]
```

Out[•]//MatrixForm=
$$\begin{pmatrix} 0.0350763 \\ 0.00664771 \\ 0.0000535885 \end{pmatrix}$$

So the result become $y = 0.035 + 0.0066\,x + 0.00005\,x^2$.

In[841]:=
```
lmQuadratic = LinearModelFit[{x, y}]
```

Out[•]=
```
FittedModel[ 0.0350763 #1 + 0.00664771 #2 + 0.0000535885 #3 ]
```

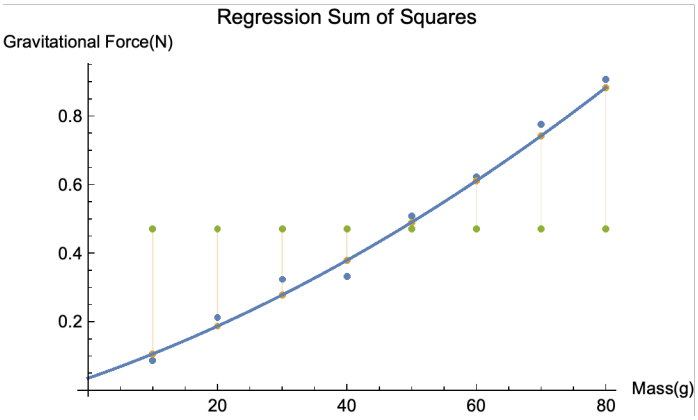
Error Analysis

We define an ***orthogonal projection*** matrix

$$P = \frac{1}{n}\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & & \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & & 1 \end{pmatrix}$$

which can map a vector in $\mathbb{R}^n$ to the mean of its entries. It has the nice property that $P^2 = P$, $P^T = P$.

Out[•]=



We can then write $\mathrm{SS}_T = \langle (\mathbb{1}_n - P)\, Y, (\mathbb{1}_n - P)\, Y\rangle = \langle Y, (\mathbb{1}_n - P)\, Y\rangle$. Furthermore, we define the hat matrix $H = X\left(X^T X\right)^{-1} X^T$, then $\hat{Y} = X\, b = X\left(X^T X\right)^{-1} X^T\, Y = H\, Y$.

$$\begin{aligned} \mathrm{SS}_T &= \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 = \langle Y, (\mathbb{1}_n - P)\, Y\rangle \\ &= \langle Y, (\mathbb{1}_n - H)\, Y\rangle + \langle Y, (H - P)\, Y\rangle \\ &= \mathrm{SS}_E + \mathrm{SS}_R \\ &= \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 + \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2 \end{aligned}$$

where SS_R is the ***regression sum of squares***. Then the coefficient of multiple determination is

$$R^2 = \frac{\mathrm{SS}_T - \mathrm{SS}_E}{\mathrm{SS}_T} = \frac{\mathrm{SS}_R}{\mathrm{SS}_T}$$

In[1019]:=
```
lm["RSquared"]
```

Out[•]=
```
0.981164
```

In[842]:=
```
lmQuadratic["RSquared"]
```

Out[•]=
```
0.98973
```

Distribution of Sum of Squares Error

- SS_E / σ^2 follows a chi-squared distribution with $n - p - 1$ degrees of freedom.
- If $\beta_1 = \beta_2 = \cdots = \beta_p = 0$, then SS_R / σ^2 follows a chi-squared distribution with p degrees of freedom.
- SS_R and SS_E are independent.
- The estimator for variance $S^2 = \frac{SS_E}{n-p-1}$ is unbiased.

```
In[•]:= lm["EstimatedVariance"]
Out[•]= 0.00176817
In[•]:= lmQuadratic["EstimatedVariance"]
Out[•]= 0.00115691
```

F-test for significance of Regression

Testing Parameter	Null Hypothesis	Test Statistics
$\beta_1, \beta_2, \ldots, \beta_p$	$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$	$F_{p,n-p-1} = \frac{SS_R/p}{SS_E/(n-p-1)}$

We reject H_0 if $F_{p,n-p-1} > f_{\alpha,p,n-p-1}$.

R Demonstration

The R language is powerful and easy to use when doing regression analysis. See the VE401R-C7_R_Demo folder.