37262 Mathematical Statistics

Lecture 7

Regression

- The process of estimating the value of one or more variable based on knowledge of the values of one or more other variables is **regression**.

- The variables of which we estimate the values are **dependent** variables or **response** variables.

- The variables from which we estimate the dependent variables are **independent** variables or **predictor** variables.

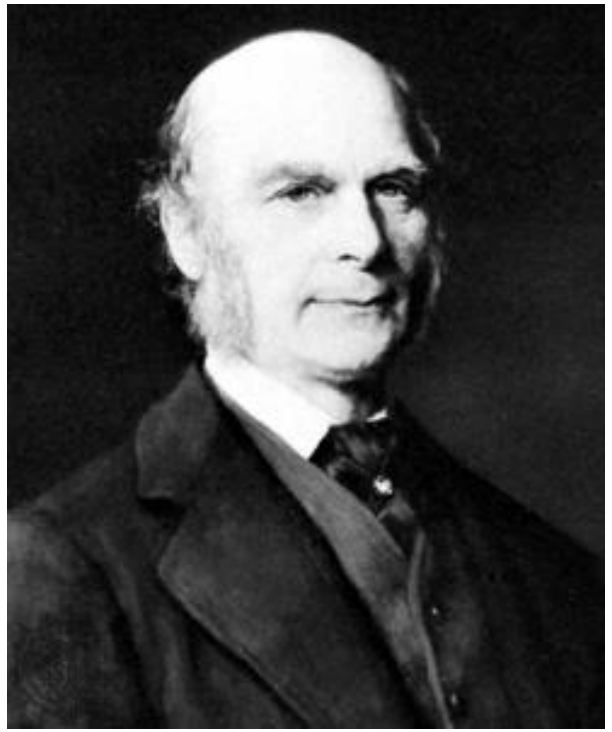
- By convention, we will usually plot a response variable on the vertical (usually *y*) axis of a graph and a predictor variable on the horizontal (usually *x*) axis.

Regression

- The term “regression” comes from the observations of Francis Galton who studied the heights of adults and the heights of their parents.

- He observed that people whose parents were taller than average were likely themselves to be taller than average, but by a smaller amount than their parents’ heights (and similarly with people whose parents were shorter than average.)

- As such, he observed that there was a connection between heights between generations but that, over generations, these deviations away from the average regressed towards the mean.



Francis Galton

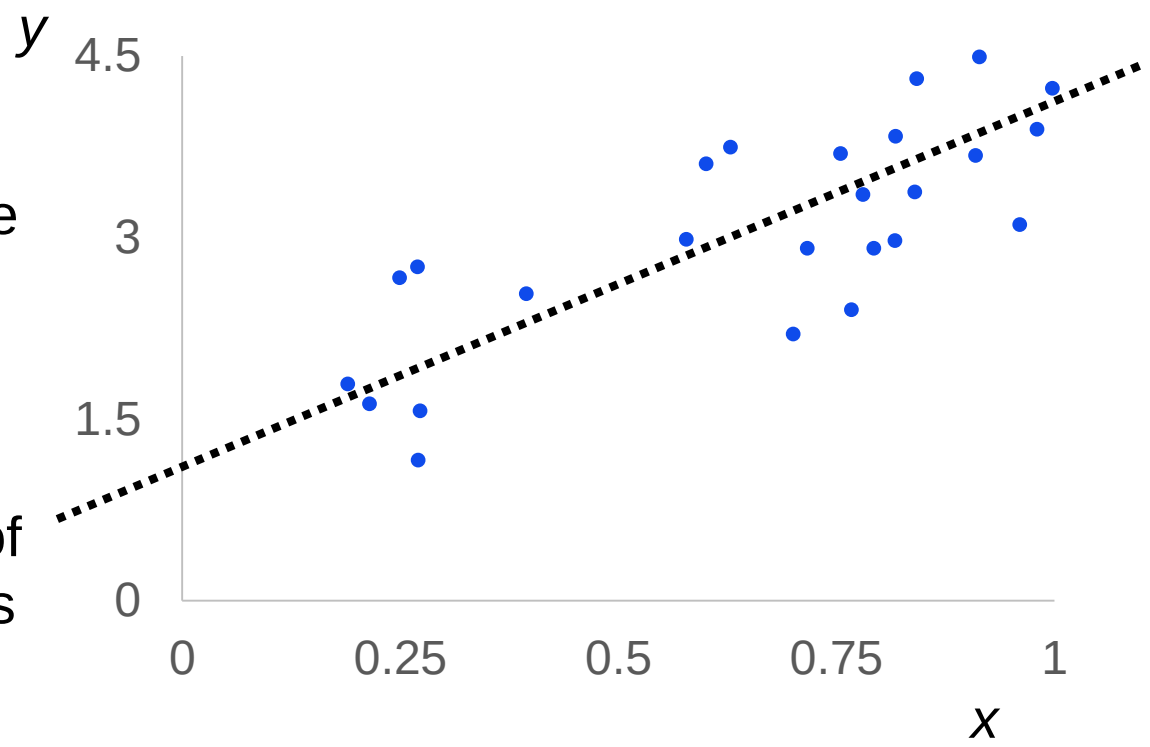
(*1822 – 1911*)

Simple Linear Regression

- The most basic regression model, often called **simple linear regression** involves estimating the relationship between one predictor variable and one response variable.

- Consider the following dataset and the problem of quantifying the relationship between observations of a predictor *x* and a response *y*.

- Here, we might visually estimate the trend to be characterised by something like the dashed line.



Simple Linear Regression

- Let $\{x_1, x_2, ..., x_n\}$ be the values of a set of predictor variables corresponding to response variables $\{y_1, y_2, ..., y_n\}$ with corresponding indexing i.e. when the predictor took the value x_1, the observed response was y_1 etc.

- Formally, we can define a simple linear regression model $y_i = \alpha + \beta x_i + \varepsilon_i$. This seeks to explain the response variable by a multiple of the predictor variable plus a fixed constant.

- The term β is sometimes called a **slope** coefficient and α is an **intercept** coefficient.

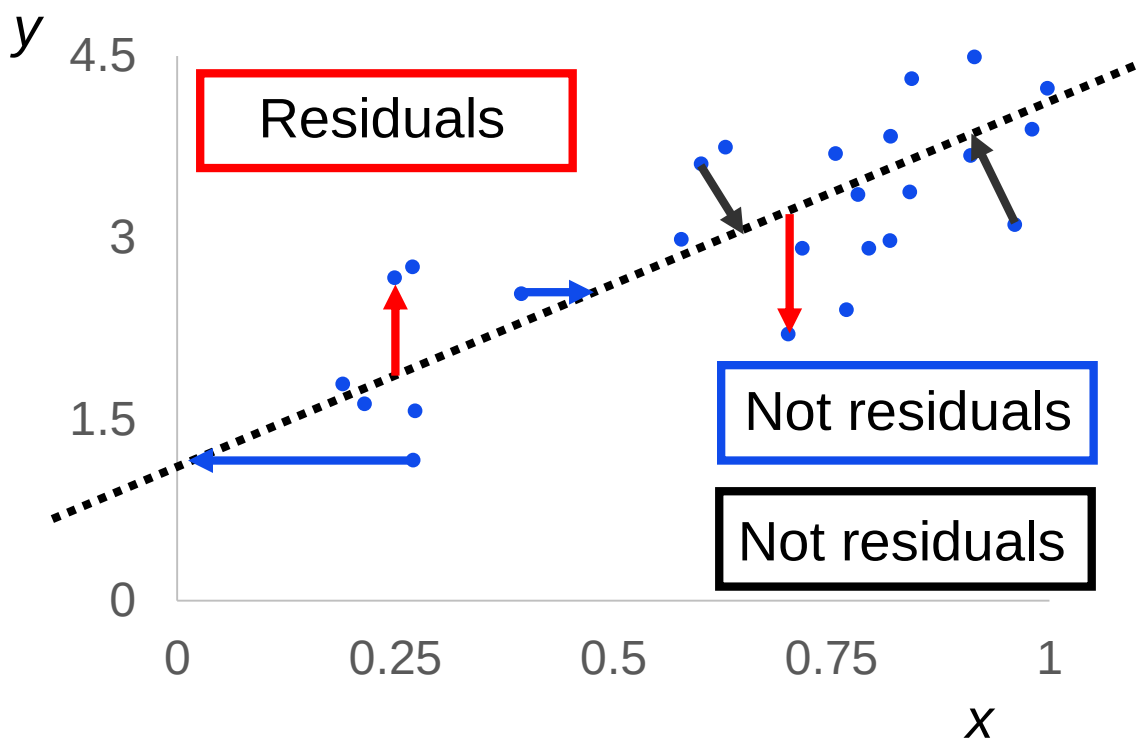
- The **residual error** term $\varepsilon_i = y_i - \alpha - \beta x_i$ measures the difference between a simple linear predictor and the observed value of the response variable.

Residuals

- We can visualise the residuals as the distances between the line used for prediction and the observed value in the dimension of the response variable.

- Mathematically, we seek a way of selecting the "best" prediction line such that the residual errors are not too large.

- Usually, we do this by first ensuring $E(\varepsilon_i) = 0$ so $\sum_{i=1}^{n} \varepsilon_i = 0$.



Ordinary Least Squares

- The most common technique for estimating the coefficients in a simple linear regression model is **ordinary least squares**.

- To do this, we choose α and β such that the sum of squared residuals is minimised.

- The squaring has two benefits
  - positive and negative residuals of the same magnitude are treated equally;
  - very large residuals are strongly penalised since a residual of 2 appears four times as large in the objective function (sum of squared residuals) as a residual of 1 does.

Ordinary Least Squares

- $\sum_{i=1}^{n} \varepsilon_i = 0$ hence $\sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0.$
- This gives $\sum_{i=1}^{n} y_i - n\alpha - \sum_{i=1}^{n} \beta x_i = 0$ so $\alpha = \frac{\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \beta x_i}{n}.$
- Our estimate of α is therefore $\hat{\alpha} = \bar{y} - \beta \bar{x}$ where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$
- We now seek $\operatorname{argmin}_{\beta} \left(\sum_{i=1}^{n} \varepsilon_i^{\,2} \right) = \operatorname{argmin}_{\beta} \left(\sum_{i=1}^{n} (y_i - \hat{\alpha} - \beta x_i)^2 \right).$

Ordinary Least Squares

- We seek $\frac{\partial}{\partial\beta}\sum_{i=1}^{n}(y_i-\hat{\alpha}-\beta x_i)^2=0 \;=\frac{\partial}{\partial\beta}\sum_{i=1}^{n}(y_i-(\bar{y}-\beta\bar{x})-\beta x_i)^2.$

- $\frac{\partial}{\partial\beta}\sum_{i=1}^{n}\left((y_i-\bar{y})-\beta(x_i-\bar{x})\right)^2=\frac{\partial}{\partial\beta}\sum_{i=1}^{n}\left((y_i-\bar{y})^2-2\beta(x_i-\bar{x})(y_i-\bar{y})+\beta^2(x_i-\bar{x})^2\right).$

- $\frac{\partial}{\partial\beta}\sum_{i=1}^{n}\left((y_i-\bar{y})^2-2\beta(x_i-\bar{x})(y_i-\bar{y})+\beta^2(x_i-\bar{x})^2\right)=\sum_{i=1}^{n}\left(-2(x_i-\bar{x})(y_i-\bar{y})+2\beta(x_i-\bar{x})^2\right)=0.$

- Solving this, we obtain $\hat{\beta}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}.$

Ordinary Least Squares

- Consider the following dataset

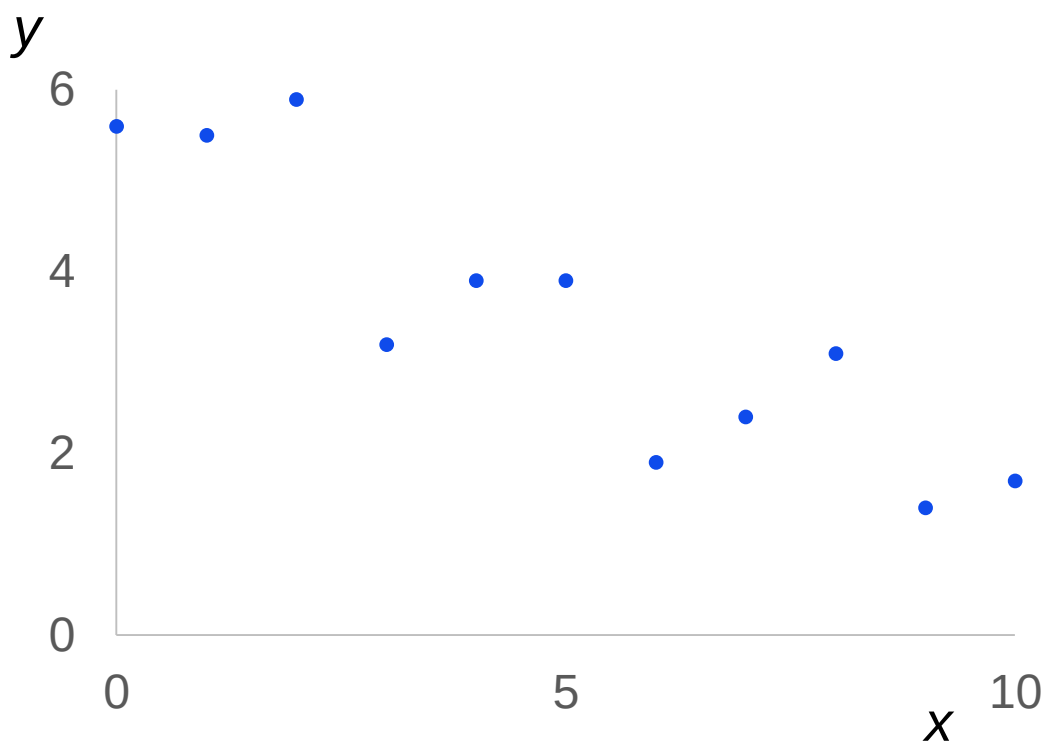
x	y
0	5.6
1	5.5
2	5.9
3	3.2
4	3.9
5	3.9
6	1.9
7	2.4
8	3.1
9	1.4
10	1.7

- We calculate

$$\sum_{i=0}^{10} x_i = 55 \text{ and } \sum_{i=0}^{10} y_i = 38.5.$$

- This gives $\bar{x} = \frac{55}{11} = 5$

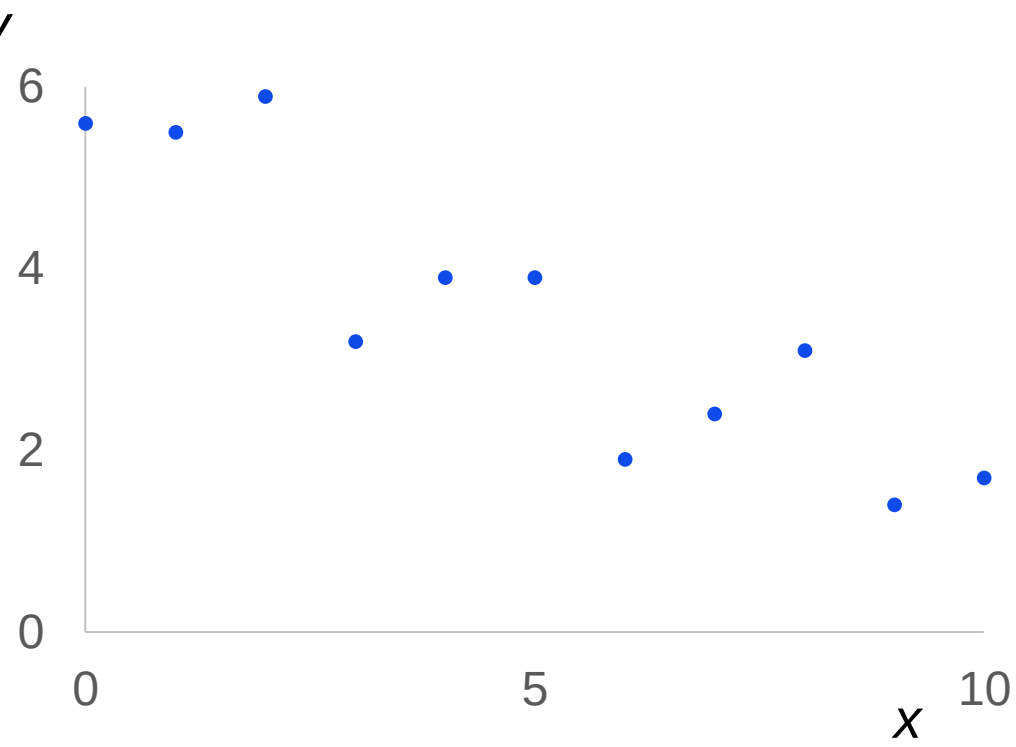
and $\bar{y} = \frac{38.5}{11} = 3.5.$



Ordinary Least Squares

- Consider the following dataset:

x	y	$x-\bar{x}$	$y-\bar{y}$
0	5.6	–5	2.1
1	5.5	–4	2
2	5.9	–3	2.4
3	3.2	–2	–0.3
4	3.9	–1	0.4
5	3.9	0	0.4
6	1.9	1	–1.6
7	2.4	2	–1.1
8	3.1	3	–0.4
9	1.4	4	–2.1
10	1.7	5	–1.8



- We calculate

$$\sum_{i=0}^{10}(x_i-\bar{x})^2 =$$

$$(-5)^2+(-4)^2+\ldots+(5)^2=110$$

and $$\sum_{i=0}^{10}(x_i-\bar{x})(y-\bar{y}) =$$

$$(-5)(-2.1)+\ldots+(5)(-1.8)=$$

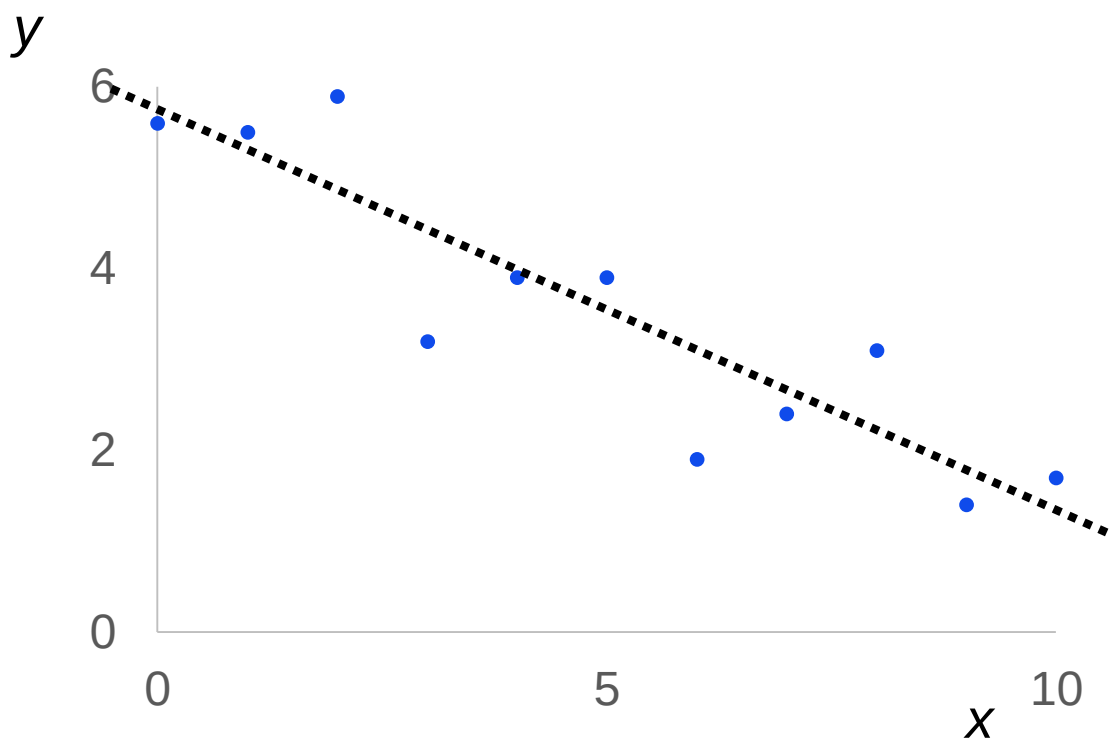
–47.9.

Ordinary Least Squares

- Putting these together, we obtain

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{-47.9}{110} \approx -0.436$$

and $\hat{\alpha} = \bar{y} - \beta\bar{x}$ hence $\hat{\alpha} = 3.5 - 5\left(\frac{-47.9}{110}\right) \approx 5.68.$



- Our simple linear regression model is therefore $y_i = 5.68 - 0.436x_i + \varepsilon_i$.

Simple Linear Regression

- Linear regression only means that the model is linear in its terms. It can be used to examine nonlinear relationships. For example, a quadratic relationship $y_i = \alpha + \beta x_i^2 + \varepsilon_i$ can be considered as a linear regression model (as plotting y against x^2 would give a straight line.)

- Note that throughout the calculation we have seen, we have made no assumptions about the distribution of the residuals.

- It is, however, common to additionally assume that the residuals are independent and identically distributed realisations of a normal variable $\varepsilon_i \sim N(0,\sigma^2)$ for some fixed σ^2.

- These additional assumptions allow for some further analyses of the regression model.

Multiple Linear Regression

- We now consider the case of having multiple predictor variables for a single response variable.

- With *n* observations of *k* predictors, we have a model $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i$.

- The same procedures can be followed as with simple linear regression i.e. setting

  $\sum_{i=1}^{n} \varepsilon_i = 0$ hence $\sum_{i=1}^{n} (y_i - \alpha - \beta_1 x_{1i} - \beta_2 x_{2i} - ... - \beta_k x_{ki}) = 0$

  and minmising $\sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} (y_i - \alpha - \beta_1 x_{1i} - \beta_2 x_{2i} - ... - \beta_k x_{ki})^2$

Multiple Linear Regression

- We can minimise $\sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} (y_i - \alpha - \beta_1 x_{1i} - \beta_2 x_{2i} - \ldots - \beta_k x_{ki})^2$ by differentiating with respect to each of $\beta_1, \beta_2, \ldots, \beta_k$ and solving simultaneous for each $\frac{\partial}{\partial \beta_j} \sum_{i=1}^{n} \varepsilon_i^{\,2} = 0.$

- This can get quite messy, but we can exploit some relatively simple matrix algebra to keep the calculation much tidier.

Multiple Linear Regression: Matrix Form

- Defining $\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, $\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}$ and $\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{pmatrix}$

  we can write the regression model as $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$.

  (Here $\boldsymbol{I}_n$ is the n-by-n identity matrix.)

- The column of 1s at the start of the data matrix $\boldsymbol{X}$ simply adds the α intercept parameter.

- Note that $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$ assumes that the residuals are all normally distributed and that they are all uncorrelated with each other.

Multiple Linear Regression: Matrix Form

- $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ so $\boldsymbol{X}^t\boldsymbol{Y} = \boldsymbol{X}^t\boldsymbol{X\beta} + \boldsymbol{X}^t\boldsymbol{\varepsilon}$.
- $\boldsymbol{X}^t\boldsymbol{Y} - \boldsymbol{X}^t\boldsymbol{\varepsilon} = \boldsymbol{X}^t\boldsymbol{X\beta}$ hence $(\boldsymbol{X}^t\boldsymbol{X})^{-1}(\boldsymbol{X}^t\boldsymbol{Y} - \boldsymbol{X}^t\boldsymbol{\varepsilon}) = \boldsymbol{\beta}$.
- This is optimised when $\boldsymbol{X}^t\boldsymbol{\varepsilon} = \boldsymbol{0}$ since the residuals are unrelated to $\boldsymbol{X}$ and each is mean zero.
- $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}(\boldsymbol{X}^t\boldsymbol{Y})$.

Multiple Linear Regression: Matrix Form

x_1	x_2	y
1	1	3
1	0	6
2	1	7
2	0	10
3	1	11

- Consider fitting a model of the form $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ to the following dataset:
- In matrix form, we define $\boldsymbol{Y} = \begin{pmatrix} 3 \\ 6 \\ 7 \\ 10 \\ 11 \end{pmatrix}$, $\boldsymbol{X} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \\ 1 & 2 & 0 \\ 1 & 3 & 1 \end{pmatrix}$ and $\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix}$.
- $\boldsymbol{X}^t\boldsymbol{Y} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 2 & 3 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 3 \\ 6 \\ 7 \\ 10 \\ 11 \end{pmatrix} = \begin{pmatrix} 37 \\ 76 \\ 21 \end{pmatrix}$ and $\boldsymbol{X}^t\boldsymbol{X} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 2 & 3 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \\ 1 & 2 & 0 \\ 1 & 3 & 1 \end{pmatrix} = \begin{pmatrix} 5 & 9 & 3 \\ 9 & 19 & 6 \\ 3 & 6 & 3 \end{pmatrix}$

Multiple Linear Regression: Matrix Form

x_1	x_2	y
1	1	3
1	0	6
2	1	7
2	0	10
3	1	11

- $\boldsymbol{X}^t\boldsymbol{X} = \begin{pmatrix} 5 & 9 & 3 \\ 9 & 19 & 6 \\ 3 & 6 & 3 \end{pmatrix}$ hence $(\boldsymbol{X}^t\boldsymbol{X})^{-1} = \frac{1}{15}\begin{pmatrix} 21 & -9 & 3 \\ -9 & 6 & -3 \\ -3 & -3 & 14 \end{pmatrix}$
- $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}(\boldsymbol{X}^t\boldsymbol{Y}) = \frac{1}{15}\begin{pmatrix} 21 & -9 & 3 \\ -9 & 6 & -3 \\ -3 & -3 & 14 \end{pmatrix}\begin{pmatrix} 37 \\ 76 \\ 21 \end{pmatrix} = \frac{1}{15}\begin{pmatrix} 30 \\ 60 \\ -45 \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \\ -3 \end{pmatrix}.$
- Our least squares estimate for the model parameters is therefore $y_i = 2 + 4x_{1i} - 3x_{2i} + \varepsilon_i$
- In this case (but not in general) we can easily verify this is the best fit, since all residuals are zero.

Generalised Least Squares

- We can also fit more complex models, such as time series models (where the observations have an inherent order in time) or multilevel models (where observations may belong to a group where there is some group-level effect.)

- This is done by fitting a model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{V}_n)$ where the off-diagonal elements of $\boldsymbol{V}_n$ capture the covariances between residuals.

- We then simply multiply by $\boldsymbol{V}_n^{-0.5}$ to obtain $\boldsymbol{V}_n^{-0.5}\boldsymbol{Y} = \boldsymbol{V}_n^{-0.5}\boldsymbol{X\beta} + \boldsymbol{V}_n^{-0.5}\boldsymbol{\varepsilon}$ which is of the form $\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}$ where $\tilde{\boldsymbol{\varepsilon}} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$.

- We can then solve as before via ordinary least squares.

Generalised Least Squares

- $\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}$ hence $\hat{\boldsymbol{\beta}} = (\tilde{\boldsymbol{X}}^t\tilde{\boldsymbol{X}})^{-1}(\tilde{\boldsymbol{X}}^t\tilde{\boldsymbol{Y}})$
- $\tilde{\boldsymbol{X}} = \boldsymbol{V}_n^{-0.5}\boldsymbol{X}$ hence $\tilde{\boldsymbol{X}}^t = \boldsymbol{X}^t\boldsymbol{V}_n^{-0.5}$. Similarly $\tilde{\boldsymbol{Y}} = \boldsymbol{V}_n^{-0.5}\boldsymbol{Y}$.
- Together, these give $\hat{\boldsymbol{\beta}} = (\tilde{\boldsymbol{X}}^t\tilde{\boldsymbol{X}})^{-1}(\tilde{\boldsymbol{X}}^t\tilde{\boldsymbol{Y}}) = (\boldsymbol{X}^t\boldsymbol{V}_n^{-0.5}\boldsymbol{V}_n^{-0.5}\boldsymbol{X})^{-1}(\boldsymbol{X}^t\boldsymbol{V}_n^{-0.5}\boldsymbol{V}_n^{-0.5}\boldsymbol{Y}) = (\boldsymbol{X}^t\boldsymbol{V}_n^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}^t\boldsymbol{V}_n^{-1}\boldsymbol{Y})$.

Coefficient of Variation (R-Squared)

- A common way to quantify how much of the variation in a response variable is explained by a regression model is the **coefficient of variation**, commonly written as ***R*-squared** or R^2. This is simply a proportion between 0 and 1.

- An *R*-squared of 1 implies that all residuals are zero and hence 100% of variation in the response variable is characterised by the model.

- An *R*-squared of 0 implies that the model captures none of the variability in the response variable.

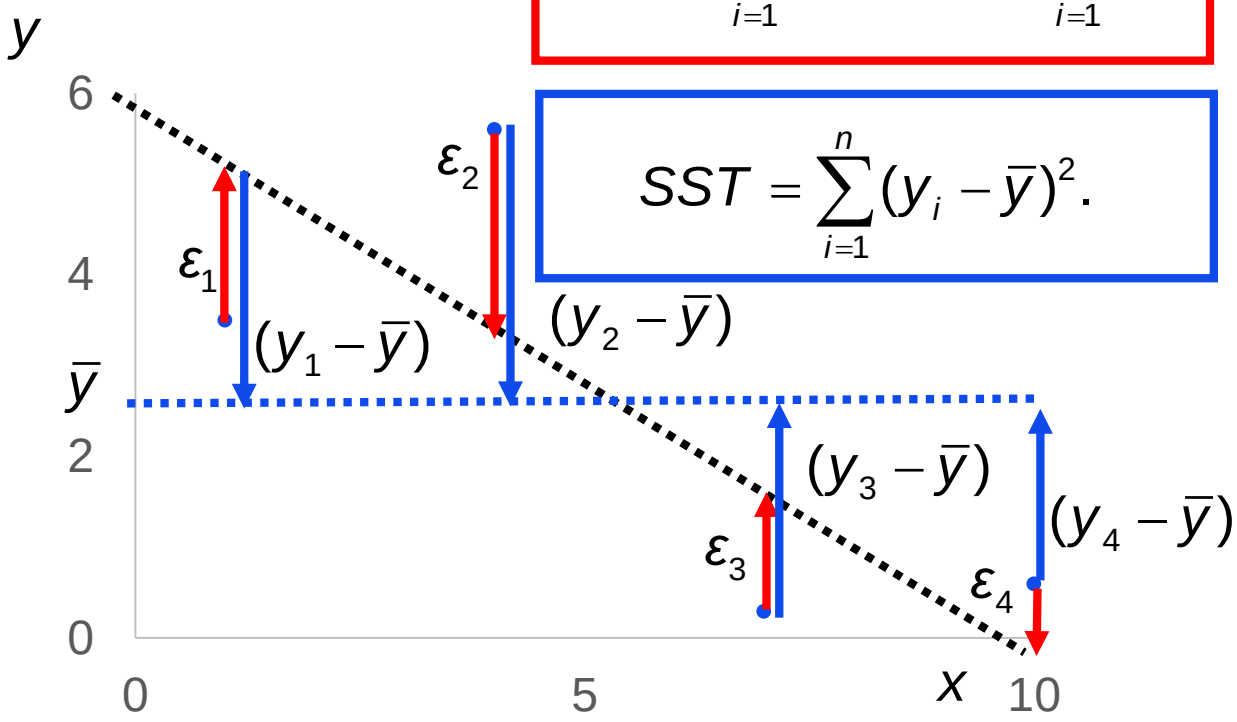
- This is a measure of how much of the variability in the response captured by a model. It does not tell us whether or not variables should be in that model. A predictor may only explain a small proportion of response variable, but still be a necessary inclusion in the final model.

Coefficient of Variation (R-Squared)

- The total squared variation (about the mean) in a response variable y is $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$.
- The sum of the squared errors around the fitted value $\hat{y}$ is $SSE = \sum_{i=1}^{n}(y_i - \hat{y})^2$.
- The sum of the squared deviation to the regression line $SSR = SST - SSE$.
- $R^2 = \frac{SSR}{SST} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y})^2 = \sum_{i=1}^{n}\varepsilon_i^2$$

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2.$$



Coefficient of Variation (R-Squared)

- Some sources define the residual error not as $SSR = SST - SSE$ but as $SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$.
- It is not instantly obvious that these definitions are equivalent.
- $$SST - SSE = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \bar{y} + \hat{y}_i - \hat{y}_i)^2 - \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n} (\hat{y}_i - \bar{y})(y_i - \hat{y}_i) - \sum_{i=1}^{n} (y_i - \hat{y})^2$$

- We therefore have that $SSR = SST - SSE$ only if $\sum_{i=1}^{n} (\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = 0$.

Coefficient of Variation (R-Squared)

- By definition, the parameters satisfy $\hat{\beta} = \text{argmin}_{\beta}\left(\sum_{i=1}^{n}(y_i - \hat{\alpha} - \beta x_i)^2\right)$ and $\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i) = 0.$
- $\frac{\partial}{\partial \beta}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \beta x_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\alpha} - \beta x_i)(-x_i) = 0$
- $\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)(-x_i) = 0$ hence $\sum_{i=1}^{n}\hat{\beta}x_i(y_i - \hat{\alpha} - \hat{\beta}x_i) = 0$
- $\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i) = 0$ gives $\sum_{i=1}^{n}(\hat{\alpha} - \bar{y})(y_i - \hat{\alpha} - \hat{\beta}x_i) = 0.$

$$\sum_{i=1}^{n}(\hat{\alpha} - \bar{y})(y_i - \hat{\alpha} - \hat{\beta}x_i) + \sum_{i=1}^{n}\hat{\beta}x_i(y_i - \hat{\alpha} - \hat{\beta}x_i) = 0 \;= \sum_{i=1}^{n}(\hat{\alpha} + \hat{\beta}x_i - \bar{y})(y_i - \hat{\alpha} - \hat{\beta}x_i)$$

$$= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = 0.$$

Assessing Model Fit

- Consider fitting a linear regression model with n observations of k predictors

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i.$$

- We wish to test the hypotheses $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$

  $H_1 : \beta_i \neq 0$ for some $i \in \{1, 2, \ldots, k\}$

- We can consider the proportion of variation in the data explained by the null model.

- $SST = \sum_{i=1}^{n} (y_i - \alpha)^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2.$

- $SSE = \sum_{i=1}^{n} (y_i - \alpha - \beta_1 x_{1i} - \beta_2 x_{2i} - \ldots - \beta_k x_{ki})^2.$

Assessing Model Fit

- Under the assumption that residuals are independent and normally distributed with equal variance σ^2

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 \sim \sigma^2 \chi^2(n-1)$$

- The variation around the fitted line is $SSE = \sum_{i=1}^{n}(y_i - \alpha - \beta_1 x_{1i} - \beta_2 x_{2i} - \ldots - \beta_k x_{ki})^2 \sim \sigma^2 \chi^2(n-k-1)$

- The variation explained by the model is therefore $SSR = (SST - SSE) \sim \sigma^2 \chi^2(k)$

- We can therefore test, under the null hypothesis, whether the proportion of variation explained by the model is consistent with the hypothesis.

Assessing Model Fit

- We know that the weighted ratio of two chi-squared variables is *F*-distributed.
- If $SSR \sim \sigma^2 \chi^2(k)$ and $SSE \sim \sigma^2 \chi^2(n-k-1)$ then $\dfrac{\left(\frac{SSR}{k}\right)}{\left(\frac{SSE}{n-k-1}\right)} \sim F(k, n-k-1)$.
- We can then reject the null hypothesis if the (weighted) proportion of variation explained by the fitted regression line is greater than the critical value from the *F*-distribution.
- Note that unlike the simple least squares calculation, we do require the assumption of normality of residuals.
- With categorical variables, this is the basis of analysis of variance (ANOVA.)