Indian Statistical Institute

Multivariate Data and Regression

Statistics - III

Prepared by:

Utkarsh Bharadwaj

Fall 2026, New Delhi

Table of Contents

1 Simple Linear Regression

Note

This is moreover of a revision of already covered topics from 1st semester along with some additional topics. To view Statistics-I notes from 1st semester, kindly click here

1.1 Motivation

We study regression for the following reasons

1. To predict the value of Y given a number of response variables $X_1, \ldots, X_p$. In SLR, we typically only have exactly one *predictor variable* (X) and one *response variable* (Y)
2. To study the impact of a predictor variable on the response variable. This, as well as the above mentioned point is done by *Estimation of Parameters*
3. To conduct test of hypothesis on the parameters

1.2 SLR Model

We'll be discussing the model for Simple Linear Regression here

The Model

Consider an array of data points $(x_1, y_1), \ldots, (x_n, y_n)$ plotted on a 2D graph. We need to find the line of best fit through the graph. Consider the model below

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

where β_0 and β_1 are the unknown parameters to be estimated. Here Y is the **output/response** variable (*dependent*) and X is the **input/predictor** variable (*independent*); ϵ is the unobservable error component. We call $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ the **population model**, and its data-driven counterpart

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

the **estimated (fitted) line**.

Note

It is important to note that β_0 and β_1 are NOT *random variables*. Thus they do not follow any probability distribution. They are just unknown constants. They must be treated like any other parameter from a probability distribution

Assumption

We operate under the following assumption in SLR

1. There exists a linear model between the predictor and response variable
2. $Cov(\epsilon_i, \epsilon_j) = \begin{cases} 0 & \text{if } i \neq j \\ \sigma^2 & \text{if } i = j \end{cases}$ Thus, all ϵ_i's are independent with a fixed unknown constant as their variance
3. $E[\epsilon_i] = 0$ and $Var(\epsilon_i) = \sigma^2$ for all $i = 1, \ldots, n$
4. x_i's maybe fixed or else we condition on them

Note

Assumption of *Normality* for ϵ is not required

Since ϵ_i is random, Y_i is also random (even though β_0, β_1 are fixed constants). This gives us

$$E[Y_i] = \beta_0 + \beta_1 x_i, \qquad Var(Y_i) = \sigma^2$$

Parameters of the Model

The unknown parameters that we are trying to estimate in SLR are β_0, β_1 and σ^2. Everything else (hypothesis tests, prediction, diagnostics) is built on top of estimating these three quantities.

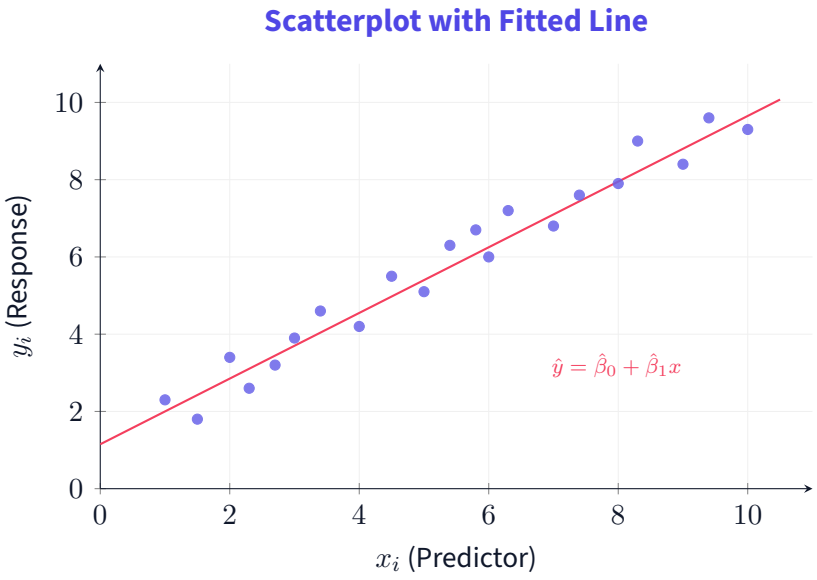
1.3 The Scatterplot

Why Start with a Scatterplot?

Before fitting any line, the very first step is to plot y_i against x_i for all i. This is a purely visual check: does a straight line look like a reasonable summary of the relationship, at least approximately? If the cloud of points is curved, or has no discernible trend at all, fitting a straight line would be misleading.

Example: Infant Mortality vs Female Literacy

Plotting infant mortality rate (y) against female literacy rate (x) across regions, the points broadly slope downward and fall roughly along a line. This visual pattern is what justifies attempting a linear model in the first place.



1.4 Goal of Regression Analysis

The Three-Step Cycle

Once we accept that a linear model is visually reasonable, the analysis proceeds in the following stages

1. **Estimation (Fitting):** Estimate the parameters $(\beta_0, \beta_1, \sigma^2)$ from data. This can be followed up by hypothesis tests on the parameters and by prediction of y at new values of x
2. **Diagnostics:** Once fitted, check whether the model assumptions (linearity, independence, homoscedasticity, etc.) are actually satisfied by the data. This is done by examining the *residuals*
3. **Iteration:** If assumptions are violated, we modify the model (transform x or y, add terms, etc.) and refit. This cycle repeats until we arrive at a model that reasonably satisfies the assumptions

Note

This is not a one-shot procedure , regression modelling is inherently iterative. Fit $\rightarrow$ Diagnose $\rightarrow$ Refine $\rightarrow$ Repeat.

1.5 Least Squares Estimation

Motivating the Least Squares Criterion

Before jumping to the regression setup, consider a simpler question: given observations $y_1, \ldots, y_n$, what single number a best predicts Y? If "best" means minimizing the sum of squared prediction errors, we solve

$$\arg\min_a \sum_{i=1}^{n} (y_i - a)^2$$

Differentiating with respect to a and setting to zero gives $a = \bar{y}$. So $\bar{y}$ is the least-squares-optimal constant predictor of y, and it minimizes the sum of squared distances from all the observations.

Regression extends exactly this idea: instead of predicting y by a single constant, we predict it by a *line* $\beta_0 + \beta_1 x_i$, and we choose $(\hat{\beta}_0, \hat{\beta}_1)$ to minimize the sum of squared vertical distances between the observed y_i and the fitted value on the line

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

An Alternative Criterion

Least squares is not the only option. One could instead minimize the sum of *absolute* deviations

$$\min_{\beta_0, \beta_1} \sum_i |y_i - \beta_0 - \beta_1 x_i|$$

This is called *Least Absolute Deviation (LAD)* regression. It is more robust to outliers than least squares but does not admit a clean closed-form solution and is computationally harder (requires linear programming). We stick to least squares.

Summation Lemmas , Building Blocks for the Derivation

Two summation identities recur constantly in everything that follows, so it is worth recording them before differentiating anything. First

$$\sum_{i=1}^{n} (x_i - \bar{x}) = \sum x_i - \sum \bar{x} = n\bar{x} - n\bar{x} = 0$$

Second, define $S_{xx} = \sum (x_i - \bar{x})^2$. Expanding the square, one of the two cross terms involves the constant $\bar{x}$ multiplied by the identity above, so it drops out, leaving

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum (x_i - \bar{x}) x_i$$

Analogously we will use $S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y}) = \sum(x_i - \bar{x})y_i$ and, later, $S_{yy} = \sum(y_i - \bar{y})^2 = \sum(y_i - \bar{y})y_i$. These three quantities recur throughout the rest of this chapter.

```
# Deriving the Least Squares Estimators
```

Set $Q(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$. We minimize Q jointly over β_0, β_1 by setting both partial derivatives to zero.

Step 1 , Partial derivative w.r.t. β_0:

$$\frac{\partial Q}{\partial \beta_0} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-1) = 0$$

$$\implies -2\left(\sum y_i - n\beta_0 - \beta_1 \sum x_i\right) = 0$$

$$\implies n\beta_0 = \sum y_i - \beta_1 \sum x_i \quad \implies \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \text{(Eq. 1)}$$

Step 2 , Partial derivative w.r.t. β_1:

$$\frac{\partial Q}{\partial \beta_1} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-x_i) = 0$$

$$\implies \sum x_i y_i - \beta_0 \sum x_i - \beta_1 \sum x_i^2 = 0 \qquad \text{(Eq. 2)}$$

Step 3 , Solve the system: Substitute Eq. 1 into Eq. 2

$$\sum x_i y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) \sum x_i - \hat{\beta}_1 \sum x_i^2 = 0$$

$$\implies \sum x_i y_i - n\bar{x}\bar{y} = \hat{\beta}_1 \left(\sum x_i^2 - n\bar{x}^2\right)$$

$$\implies \hat{\beta}_1 = \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2}$$

By the Summation Lemmas above, $\sum x_i y_i - n\bar{x}\bar{y} = \sum(x_i - \bar{x})(y_i - \bar{y}) = S_{xy}$ and $\sum x_i^2 - n\bar{x}^2 = \sum(x_i - \bar{x})^2 = S_{xx}$, so the numerator and denominator collapse to exactly these centered sums, giving the clean final form

$$\boxed{\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}}$$

Least Squares Estimator , Final Form

$$\boxed{\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}} \qquad \boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

Geometrically: we are shifting and tilting the line so as to minimize the sum of squared *vertical* distances of the points in the scatterplot from the line.

1.6 Expectation and Variance of the Parameter Estimates

Setup: The Linear Constant k_i

To simplify all the moment calculations that follow, define a constant k_i that depends only on the (fixed) x-values

$$k_i = \frac{x_i - \bar{x}}{\sum(x_i - \bar{x})^2} = \frac{x_i - \bar{x}}{S_{xx}}$$

This lets us rewrite the slope estimator as a *linear combination of the Y_i's*

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x_i - \bar{x})(Y_i - \bar{Y})}{S_{xx}} = \frac{\sum(x_i - \bar{x})Y_i}{S_{xx}} = \sum_{i=1}^{n} k_i Y_i$$

(the $\bar{Y}\sum(x_i - \bar{x})$ piece drops out because $\sum(x_i - \bar{x}) = 0$). Writing $\hat{\beta}_1$ this way is what makes its expectation and variance easy to compute directly from the properties of Y_i.

```
# Three Properties of k_i
```

1. $\sum k_i = \dfrac{\sum(x_i - \bar{x})}{S_{xx}} = \dfrac{0}{S_{xx}} = 0$
2. $\sum k_i x_i = \dfrac{\sum(x_i - \bar{x})x_i}{S_{xx}} = \dfrac{S_{xx}}{S_{xx}} = 1$
3. $\sum k_i^2 = \sum \left(\dfrac{x_i - \bar{x}}{S_{xx}}\right)^2 = \dfrac{1}{S_{xx}^2}\sum(x_i - \bar{x})^2 = \dfrac{S_{xx}}{S_{xx}^2} = \dfrac{1}{S_{xx}}$

These three identities are all that is needed to derive every moment of $\hat{\beta}_0, \hat{\beta}_1$ below.

Unbiasedness of $\hat{\beta}_1$

$$\begin{aligned} E[\hat{\beta}_1] &= E\left[\sum k_i Y_i\right] = \sum k_i \, E[Y_i] \qquad \text{(the } k_i\text{'s are fixed constants)} \\ &= \sum k_i(\beta_0 + \beta_1 x_i) \\ &= \beta_0 \underbrace{\sum k_i}_{=0} + \beta_1 \underbrace{\sum k_i x_i}_{=1} \\ &= \beta_1 \end{aligned}$$

So $\hat{\beta}_1$ is an **unbiased** estimator of β_1.

Unbiasedness of $\hat{\beta}_0$

Using $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$

$$\begin{aligned} E[\hat{\beta}_0] &= E[\bar{Y}] - \bar{x}\, E[\hat{\beta}_1] \\ &= (\beta_0 + \beta_1 \bar{x}) - \bar{x}(\beta_1) \\ &= \beta_0 \end{aligned}$$

So $\hat{\beta}_0$ is also **unbiased**. Neither result used normality of ϵ_i , only $E[\epsilon_i] = 0$.

Variance of $\hat{\beta}_1$

Since $\hat{\beta}_1 = \sum k_i Y_i$ and the Y_i's are independent

$$Var(\hat{\beta}_1) = Var\left(\sum k_i Y_i\right) = \sum k_i^2 \, Var(Y_i) + \underbrace{\sum\sum Cov(Y_i, Y_j)}_{=0 \;\forall\; i \neq j}$$

$$\begin{aligned} &= \sigma^2 \sum k_i^2 \\ &= \sigma^2 \cdot \frac{1}{S_{xx}} \qquad \text{(Property 3)} \\ &= \frac{\sigma^2}{S_{xx}} \end{aligned}$$

Side Derivation: $Cov(\bar{Y}, \hat{\beta}_1) = 0$

This intermediate result is needed to find $Var(\hat{\beta}_0)$ below. Recall $\hat{\beta}_1 = \sum_j k_j Y_j$ and

$\bar{Y} = \frac{1}{n}\sum_i Y_i$

$$
\begin{aligned}
Cov(\bar{Y}, \hat{\beta}_1) &= Cov\left(\frac{1}{n}\sum_{i=1}^{n} Y_i,\ \sum_{j=1}^{n} k_j Y_j\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} k_i\, Var(Y_i) + \underbrace{\sum\sum Cov(Y_i, Y_j)}_{=0\ \forall\ i\neq j} \\
&= \frac{1}{n}\sum_{i=1}^{n} k_i\, \sigma^2 \\
&= \frac{\sigma^2}{n}\underbrace{\sum_{i=1}^{n} k_i}_{=0} \qquad \text{(Property 1)} \\
&= 0
\end{aligned}
$$

Variance of $\hat{\beta}_0$

Using $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$

$$
\begin{aligned}
Var(\hat{\beta}_0) &= Var(\bar{Y}) + \bar{x}^2\, Var(\hat{\beta}_1) - 2\bar{x}\ \underbrace{Cov(\bar{Y}, \hat{\beta}_1)}_{=0} \\
&= \frac{\sigma^2}{n} + \bar{x}^2 \cdot \frac{\sigma^2}{S_{xx}} \\
&= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) \\
&= \frac{\sigma^2(S_{xx} + n\bar{x}^2)}{n\, S_{xx}} = \frac{\sigma^2 \sum x_i^2}{n\, S_{xx}}
\end{aligned}
$$

(the last step uses $S_{xx} = \sum x_i^2 - n\bar{x}^2$, so $S_{xx} + n\bar{x}^2 = \sum x_i^2$).

Covariance of $\hat{\beta}_0$ and $\hat{\beta}_1$

$$
\begin{aligned}
Cov(\hat{\beta}_0, \hat{\beta}_1) &= Cov(\bar{Y} - \hat{\beta}_1\bar{x},\ \hat{\beta}_1) \\
&= \underbrace{Cov(\bar{Y}, \hat{\beta}_1)}_{=0} - \bar{x}\, Cov(\hat{\beta}_1, \hat{\beta}_1) \\
&= -\bar{x}\, Var(\hat{\beta}_1) \\
&= -\frac{\bar{x}\,\sigma^2}{S_{xx}}
\end{aligned}
$$

Quick Reference: Sampling Properties of $\hat{\beta}_0, \hat{\beta}_1$

$$E[\hat{\beta}_1] = \beta_1 \quad \text{(unbiased)}, \qquad Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

$$E[\hat{\beta}_0] = \beta_0 \quad \text{(unbiased)}, \qquad Var(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) = \frac{\sigma^2 \sum x_i^2}{nS_{xx}}$$

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}\,\sigma^2}{S_{xx}}$$

Note

Both $\hat{\beta}_0$ and $\hat{\beta}_1$ are *unbiased* estimators of β_0, β_1 regardless of normality , this only requires $E[\epsilon_i] = 0$. Normality of ϵ_i is what upgrades these moment results into exact sampling *distributions* (needed for t-tests and confidence intervals on β_0, β_1).

1.7 The Asymmetric Role of Predictor and Response

This section discusses our choice of Regression Line

x on y vs y on x

A subtle but important point:

1. Regression does *not* minimize perpendicular distance from the points to the line (that would be Total Least Squares / orthogonal regression)
2. We treat x and y *asymmetrically*: we minimize the *vertical* distance (in y) of the points from the line, treating x as given/fixed
3. If instead we regress x on y, i.e., fit the model

   $$x = \alpha_0 + \alpha_1 y + \epsilon$$

   then by the same least-squares logic applied with roles swapped

   $$\hat{\alpha}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(y_i - \bar{y})^2}, \qquad \hat{\alpha}_0 = \bar{x} - \hat{\alpha}_1\bar{y}$$

   This minimizes *horizontal* distance instead of vertical distance , a genuinely different line in general!

Two Different Lines

The regression of y on x and the regression of x on y are **not** the same line (unless $r = \pm 1$). Regressing y on x answers "what y do I expect for a given x?" while regressing x on y answers the reverse question. Always be clear about which variable is the response.

Types of Regression, by Choice of Offset

More generally, there are three natural lines one could fit to the same scatterplot, depending on which kind of distance is minimized

1. **Vertical Offsets (OLS):** Minimize $\sum_i (y_i - \hat{y}_i)^2$. Assumes the error lies entirely in Y , this is exactly what ordinary least squares does, and is the line we derived above
2. **Horizontal Offsets:** Minimize $\sum_i (x_i - \hat{x}_i)^2$. Assumes the error lies entirely in X , this is precisely the regression of x on y derived above, with $\hat{\alpha}_1, \hat{\alpha}_0$
3. **Perpendicular Offsets:** Minimize the Euclidean (perpendicular) distance from each point to the line. Assumes error in *both* X and Y , this is *orthogonal regression / Total Least Squares*, and is a genuinely different line from either of the above

OLS uses vertical offsets specifically because, under the standard SLR assumptions, X is treated as fixed/known (e.g., a chosen dosage or time point), so *all* measurement error is assumed to live in the response Y , there is simply no error in X to account for.

Common Point, Related Slopes

Both fitted lines pass through the point of means $(\bar{x}, \bar{y})$

$$y - \bar{y} = \hat{\beta}_1 (x - \bar{x}), \qquad x - \bar{x} = \hat{\alpha}_1 (y - \bar{y})$$

Multiplying the two slopes together gives a beautifully clean identity

$$\hat{\beta}_1 \, \hat{\alpha}_1 = r^2$$

where r is the sample correlation coefficient between x and y. Since $r^2 \leq 1$, the two lines can only coincide when $|r| = 1$, i.e., when the points fall exactly on a line.

1.8 Regression to the Mean

Standardized Form of the Regression Line

Rewrite the fitted line as $y - \bar{y} = \hat{\beta}_1(x - \bar{x})$. Standardizing both x and y (dividing deviations by their respective standard deviations s_x, s_y) gives

$$\frac{y - \bar{y}}{s_y} = r \cdot \frac{x - \bar{x}}{s_x}$$

Since $|r| \leq 1$, a one standard-deviation change in (standardized) x predicts *less than* one standard deviation change in (standardized) y , unless $|r| = 1$. In other words, predictions for y are always pulled back *toward the mean* relative to how extreme x was.

Example: Galton's Height Data

Francis Galton observed that tall fathers tend to have sons who are tall, but on average *not as tall as the father* , their height "regresses" back toward the population mean height. Symmetrically, sons of very short fathers tend to be short but not as short, on average, as the father. This is precisely the phenomenon described by $|\hat{\beta}_1| < 1$ in standardized units, and is in fact the historical origin of the term "regression."

A second classic example: individuals who score exceptionally well on one test of a pair of similar tests tend to score somewhat less exceptionally, on average, on the second test , and vice versa for very low scorers.

Not a Causal Effect

Regression to the mean is a statistical artefact of imperfect correlation ($|r| < 1$), not evidence of an active "dampening" causal mechanism. It is easy to mistake this for a real effect (e.g., "rookie of the year" players often underperform in their second season , this is regression to the mean, not necessarily a decline in ability).

MLE vs Least Squares

If we additionally assume $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, then the least squares estimators $\hat{\beta}_0, \hat{\beta}_1$ coincide exactly with the *Maximum Likelihood Estimators* of β_0, β_1. However, the least squares criterion itself is more general than MLE , it needs **no** distributional assumption on ϵ_i to be defined or to be a sensible criterion. Normality is only invoked later, when we want to do exact hypothesis tests / confidence intervals.

1.9 Model Evaluation

Is x Actually Useful?

If we ignored x entirely and predicted every y_i by $\bar{y}$, the total prediction error is captured by the **Total Sum of Squares**

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

After fitting the regression line, the leftover, unexplained error is the **Residual Sum of Squares**

$$SS_{Res} = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Intuitively, if the regression is doing anything useful at all, it should explain away some of the variation that $\bar{y}$ alone could not , i.e., we expect $SS_{Res} < SST$. The bigger the gap $(SST - SS_{Res})$, the more useful the model.

`# Proof that` $SST \geq SS_{Res}$ `(Decomposition of Variance)`

Recall $S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y})$ and $S_{xx} = \sum(x_i - \bar{x})^2$ from the Summation Lemmas earlier, and additionally define $S_{yy} = \sum(y_i - \bar{y})^2 = SST$. Recall also $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$ and $\hat{\beta}_1 = S_{xy}/S_{xx}$. Substitute into SS_{Res}

$$\begin{aligned}
SS_{Res} &= \sum \left(y_i - (\bar{y} - \hat{\beta}_1\bar{x}) - \hat{\beta}_1 x_i\right)^2 = \sum \left[(y_i - \bar{y}) - \hat{\beta}_1(x_i - \bar{x})\right]^2 \\
&= \sum (y_i - \bar{y})^2 - 2\hat{\beta}_1 \sum (x_i - \bar{x})(y_i - \bar{y}) + \hat{\beta}_1^2 \sum (x_i - \bar{x})^2 \\
&= S_{yy} - 2\hat{\beta}_1 S_{xy} + \hat{\beta}_1^2 S_{xx} \\
&= S_{yy} - 2\frac{S_{xy}}{S_{xx}} S_{xy} + \left(\frac{S_{xy}}{S_{xx}}\right)^2 S_{xx} \\
&= S_{yy} - \frac{S_{xy}^2}{S_{xx}}
\end{aligned}$$

So $SS_{Res} = SST - \dfrac{S_{xy}^2}{S_{xx}}$. Since $\dfrac{S_{xy}^2}{S_{xx}} \geq 0$, it follows immediately that $SS_{Res} \leq SST$. ■

Regression Sum of Squares

The amount of variation *explained away* by the regression is called the **Regression Sum of Squares**

$$SS_{Reg} = SST - SS_{Res} = \frac{S_{xy}^2}{S_{xx}}$$

This gives the fundamental decomposition of total variability

$$SST = SS_{Reg} + SS_{Res}$$

i.e., *Total variation = Variation explained by regression + Variation left unexplained*.

Coefficient of Determination (R^2)

The proportion of total variation in y that is explained by the linear regression on x is

$$R^2 = \frac{SS_{Reg}}{SST} = \frac{S_{xy}^2/S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx}S_{yy}} = r^2$$

where r is the sample correlation between x and y. So R^2 is nothing but the square of the correlation coefficient in SLR (this identity generalizes, in modified form, to multiple regression too).

Note

- $0 \leq R^2 \leq 1$, with larger R^2 indicating a better linear fit
- R^2 measures *proportion of variability of y explained by regression on x* , it says nothing about causation, and nothing about whether the linear form itself is correct

ANOVA Table for Regression

All these sums of squares are conventionally organized into an **Analysis of Variance (ANOVA)** table

Source	SS	df	MS	F
Regression	SS_{Reg}	1	$MS_{Reg} = \frac{SS_{Reg}}{1}$	$F = \frac{MS_{Reg}}{MS_{Res}}$
Residual	SS_{Res}	$n-2$	$MS_{Res} = \frac{SS_{Res}}{n-2}$	
Total	SST	$n-1$		

Note

- Under normality of ϵ_i, each sum of squares (suitably scaled) follows a χ^2 distribution with the stated degrees of freedom
- The ratio of two independent χ^2 random variables, each divided by its own df, follows an F-distribution. Hence $F = MS_{Reg}/MS_{Res}$ has an $F_{1,n-2}$ distribution under H_0
- This is exactly what lets us build a formal hypothesis test out of the ANOVA decomposition

Hypothesis Test for Significance of Regression

We test

$$H_0 : \beta_1 = 0 \quad \text{(regression is not significant)}$$
$$H_a : \beta_1 \neq 0 \quad \text{(regression is significant)}$$

Under H_0, $F = MS_{Reg}/MS_{Res} \sim F_{1,n-2}$. A neat way to rewrite the F-statistic is directly in terms of R^2

$$F = \frac{MS_{Reg}}{MS_{Res}} = \frac{SS_{Reg}/1}{SS_{Res}/(n-2)} = \frac{R^2 \cdot SST}{(1-R^2)SST/(n-2)} = \frac{R^2}{1-R^2}(n-2)$$

Reading the Test

- A significant F tells us the model has significantly reduced variation in the data, i.e., x is genuinely useful in predicting y
- Ideally we want *both* a significant F-statistic *and* a large R^2
- In practice, it is common to see a significant F with only a moderate R^2 (say 0.5,0.6). This means x is statistically useful, but a substantial amount of variation in y remains unexplained , significance and explanatory power are related but distinct notions

1.10 Estimating σ^2

Residuals and the Unbiased Estimator

Define the fitted value $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ and the residual

$$\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

An unbiased estimator of σ^2 is

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

Why $n-2$?

We divide by $n-2$, not n, because *two* degrees of freedom have already been "used up" in estimating β_0 and β_1 from the data. This is exactly analogous to the simpler case: for the constant model $y_i = a + \epsilon_i$ (only 1 parameter estimated), the unbiased variance estimator is $\frac{1}{n-1}\sum(y_i - \bar{y})^2$, dividing by $n-1$ because one degree of freedom ($\bar{y}$) was used. Each estimated parameter costs one degree of freedom.

 # Proof: $\hat{\sigma}^2 = MS_{Res}$ is an Unbiased Estimator of σ^2

We show $E[SS_{Res}] = (n-2)\sigma^2$; dividing through by $n-2$ immediately gives $E[MS_{Res}] = \sigma^2$.

Step 1 - Expectation of SST. Using the model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, we can write

$$Y_i - \bar{Y} = \beta_1(x_i - \bar{x}) + (\epsilon_i - \bar{\epsilon})$$

Squaring and summing over i, the cross term $2\beta_1\sum(x_i - \bar{x})(\epsilon_i - \bar{\epsilon})$ vanishes *in expectation* (since $E[\epsilon_i - \bar{\epsilon}] = 0$), leaving

$$SST = \sum(Y_i - \bar{Y})^2 \implies E[SST] = \beta_1^2\sum(x_i - \bar{x})^2 + E\left[\sum(\epsilon_i - \bar{\epsilon})^2\right]$$

$$= \beta_1^2 S_{xx} + E\left[\sum(\epsilon_i - \bar{\epsilon})^2\right]$$

Since the ϵ_i's are i.i.d. with variance σ^2, the quantity $\sum(\epsilon_i - \bar{\epsilon})^2$ is exactly the numerator of a sample-variance calculation on the ϵ_i's, so it has expectation $(n-1)\sigma^2$ (one degree of freedom is used up estimating $\bar{\epsilon}$). Hence

$$\boxed{E[SST] = \beta_1^2 S_{xx} + (n-1)\sigma^2}$$

Step 2 - Expectation of SS_{Reg}. Recall $SS_{Reg} = \hat{\beta}_1^2 S_{xx}$, so

$$E[SS_{Reg}] = S_{xx}\,E[\hat{\beta}_1^2] = S_{xx}\left(Var(\hat{\beta}_1) + \left(E[\hat{\beta}_1]\right)^2\right)$$

using the identity $E[W^2] = Var(W) + (E[W])^2$ for any random variable W. Substituting $Var(\hat{\beta}_1) = \sigma^2/S_{xx}$ and $E[\hat{\beta}_1] = \beta_1$ (both proved in the previous section)

$$\boxed{E[SS_{Reg}] = S_{xx}\left(\frac{\sigma^2}{S_{xx}} + \beta_1^2\right) = \sigma^2 + \beta_1^2 S_{xx}}$$

Step 3 - Expectation of SS_{Res}. Since $SS_{Res} = SST - SS_{Reg}$ (the ANOVA identity)

$$\begin{aligned} E[SS_{Res}] = E[SST] - E[SS_{Reg}] &= \left[\beta_1^2 S_{xx} + (n-1)\sigma^2\right] - \left[\sigma^2 + \beta_1^2 S_{xx}\right] \\ &= (n-1)\sigma^2 - \sigma^2 = (n-2)\sigma^2 \end{aligned}$$

Conclusion.

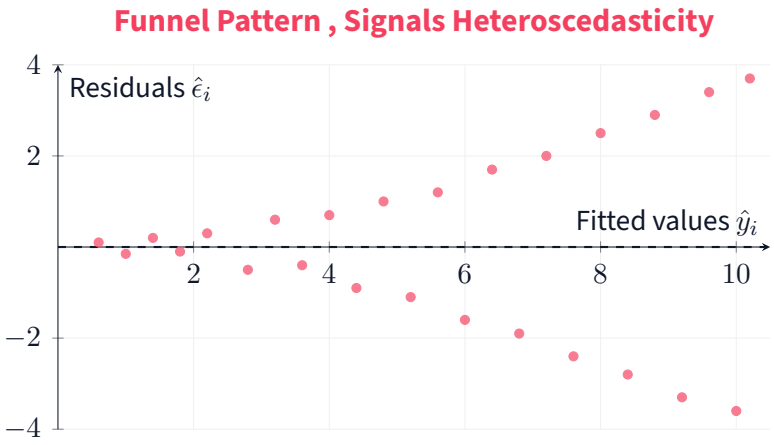
$$\boxed{E[\hat{\sigma}^2] = E\left[\frac{SS_{Res}}{n-2}\right] = \frac{(n-2)\sigma^2}{n-2} = \sigma^2} \quad \blacksquare$$

1.11 Diagnostics for Simple Linear Regression

Checking the Assumptions

Once the model is fit, we must go back and check whether the assumptions we started with actually hold, using the residuals $\hat{\epsilon}_i$ as our diagnostic tool

1. **Linearity:** Plot $\hat{\epsilon}_i$ against x_i (or against fitted values $\hat{y}_i$). A systematic pattern (e.g., a curve) suggests the true relationship is not linear
2. **Independence of errors:** ϵ_i's should be independent of each other. Hard to check directly from a single sample; a specific and important violation is *serial dependence*, common in time-series data (residuals near each other in time tend to be correlated)
3. **Independence of errors and predictor:** ϵ_i should show no pattern when plotted against x_i
4. **Zero mean of errors:** $E(\epsilon_i) = 0$
5. **Constant variance (Homoscedasticity):** $Var(\epsilon_i) = \sigma^2$ for all i. Plot $\hat{\epsilon}_i$ against $\hat{y}_i$; a *funnel shape* (spread increasing or decreasing systematically) signals non-constant variance, i.e., *heteroscedasticity*
6. **Normality of errors:** If normality of ϵ_i is assumed (needed for exact tests), check using a **QQ-plot** of the residuals $\hat{\epsilon}_i$
7. **Leverage:** Identify points that have unusually high influence on the fitted line , leverage measures how much the regression line shifts due to the presence of a single point



Note

Compare this against an *ideal* residual plot: a structureless, constant-width band of points scattered around the horizontal zero line, with no funnel, no curve, and no trend. That is exactly what we hope to see if all assumptions hold.

1.12 Bivariate Normal Distribution

Reference for this section: Sec 2.11 of Kutner et al., Applied Linear Statistical Models

Why Do We Need This?

So far x_i has been treated as *fixed* (non-random). This is realistic in some settings , e.g., a demand curve, where we deliberately set the price X and observe demand Y. But in many real situations we cannot fix X at all.

Example: Age and Blood Pressure

Consider studying the effect of age on blood pressure by recording (Age, BP) for patients as they walk into a clinic. Here $X = \text{Age}$ is clearly *random*, not something the experimenter chooses. Does the linear regression framework still make sense if we condition on X?

The answer is **yes** , provided (X, Y) jointly follow a **Bivariate Normal** distribution. In that case, the conditional distribution of Y given $X = x$ is itself Normal

$$Y \mid X = x \;\sim\; N(\beta_0 + \beta_1 x,\ \sigma^2)$$

and this can be derived directly from the joint bivariate normal density.

Deriving the Conditional Distribution

The conditional density is, by definition, the ratio of joint to marginal density

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

Let

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right), \qquad X \sim N(\mu_1, \sigma_1^2)$$

The joint density is

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right]\right\}$$

and the marginal density of X is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\}$$

Write $u = x - \mu_1$, $v = y - \mu_2$ for brevity, so the two exponents above become $-\frac{1}{2(1-\rho^2)}\left[\frac{u^2}{\sigma_1^2} - 2\rho\frac{uv}{\sigma_1\sigma_2} + \frac{v^2}{\sigma_2^2}\right]$ and $-\frac{u^2}{2\sigma_1^2}$ respectively.

Step 1 , Divide the prefactors.

$$\frac{1/\left(2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}\right)}{1/\left(\sqrt{2\pi}\,\sigma_1\right)} = \frac{\sqrt{2\pi}\,\sigma_1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} = \frac{1}{\sqrt{2\pi}\,\sigma_2\sqrt{1-\rho^2}}$$

Step 2 , Subtract the exponents. Dividing densities means subtracting exponents; collect the u^2-only terms first

$$-\frac{u^2}{2\sigma_1^2(1-\rho^2)} - \left(-\frac{u^2}{2\sigma_1^2}\right) = \frac{u^2}{2\sigma_1^2}\left[1 - \frac{1}{1-\rho^2}\right] = \frac{u^2}{2\sigma_1^2}\cdot\frac{(1-\rho^2)-1}{1-\rho^2}$$

$$= \frac{u^2}{2\sigma_1^2}\cdot\frac{-\rho^2}{1-\rho^2} = -\frac{\rho^2u^2}{2\sigma_1^2(1-\rho^2)}$$

Combined with the remaining uv and v^2 terms carried over unchanged from the joint exponent, the full exponent of the ratio is

$$-\frac{\rho^2u^2}{2\sigma_1^2(1-\rho^2)} + \frac{\rho\,uv}{\sigma_1\sigma_2(1-\rho^2)} - \frac{v^2}{2\sigma_2^2(1-\rho^2)}$$

Factor $-\dfrac{1}{2\sigma_2^2(1-\rho^2)}$ out of all three terms (multiply each term by $-2\sigma_2^2(1-\rho^2)$ and check it reproduces the original coefficients)

$$= -\frac{1}{2\sigma_2^2(1-\rho^2)}\left[v^2 - 2\rho\frac{\sigma_2}{\sigma_1}uv + \rho^2\frac{\sigma_2^2}{\sigma_1^2}u^2\right]$$

Step 3 , Complete the square in v. The bracketed quadratic is a perfect square of the form $a^2 - 2ab + b^2 = (a-b)^2$ with $a = v$, $b = \rho\frac{\sigma_2}{\sigma_1}u$

$$v^2 - 2\rho\frac{\sigma_2}{\sigma_1}uv + \rho^2\frac{\sigma_2^2}{\sigma_1^2}u^2 = \left(v - \rho\frac{\sigma_2}{\sigma_1}u\right)^2$$

so the exponent collapses to the single squared term

$$-\frac{1}{2\sigma_2^2(1-\rho^2)}\left(v - \rho\frac{\sigma_2}{\sigma_1}u\right)^2$$

Step 4 , Substitute back $u = x - \mu_1$, $v = y - \mu_2$. Combining the prefactor from Step 1 with the exponent from Step 3

$$f_{Y|X=x}(y) = \frac{1}{\sqrt{2\pi}\,\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2\sigma_2^2(1-\rho^2)}\left[y - \left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1)\right)\right]^2\right\}$$

Punchline

This is exactly the density of a Normal distribution

$$Y \mid X = x \;\sim\; N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1),\; \sigma_2^2(1-\rho^2)\right)$$

which is precisely the simple linear regression model $Y = \beta_0 + \beta_1 x + \epsilon$ with $\epsilon \sim N\big(0, \sigma_2^2(1-\rho^2)\big)$, once we identify

$$\beta_1 = \rho\frac{\sigma_2}{\sigma_1}, \qquad \beta_0 = \mu_2 - \beta_1\mu_1, \qquad \sigma^2 = Var(\epsilon) = \sigma_2^2(1-\rho^2)$$

Population vs Sample Quantities

Here β_0, β_1 are the *population* parameters of the bivariate normal model, and they exactly mirror the least squares formulas we derived earlier from data

$$\hat{\beta}_{1,LS} = \frac{S_{xy}}{S_{xx}} \;\longleftrightarrow\; \beta_1 = \rho\frac{\sigma_2}{\sigma_1}, \qquad \hat{\beta}_{0,LS} = \bar{y} - \hat{\beta}_1\bar{x} \;\longleftrightarrow\; \beta_0 = \mu_2 - \beta_1\mu_1$$

There is also a direct population-level analogue of R^2

$$1 - R^2 = 1 - \frac{SS_{Reg}}{SST} = \frac{SS_{Res}}{SST} \quad\longleftrightarrow\quad 1 - \rho^2 = \frac{\sigma^2}{\sigma_2^2}$$

$$\text{where} \quad \sigma^2 = Var(\epsilon), \qquad \sigma_2^2 = Var(Y)$$

Summary

When $(X, Y) \sim N_2(\cdot)$, the simple linear regression model holds automatically , the conditional mean of Y given $X = x$ is linear in x. All the SLR parameters are *population versions* of the least squares estimates. Equivalently: **the least squares estimators are the method-of-moments estimators of the parameters of the bivariate normal model.**

1.13 Matrix Notation

Writing SLR in Matrix Form

The model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ for $i = 1, \ldots, n$ can be stacked into a single matrix equation $y = X\beta + \epsilon$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1}, \quad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}_{n\times 2}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}_{2\times 1}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}_{n\times 1}$$

Note

The first column of X being all 1's is exactly what generates the intercept β_0 in every row of $X\beta$. This matrix formulation is what generalizes seamlessly to *multiple* regression , we would simply add more columns to X, one per predictor.

```
# Normal Equations in Matrix Form
```

Minimizing $\sum_i (y_i - \beta_0 - \beta_1 x_i)^2$ w.r.t. β_0, β_1 (as before) gives the two normal equations

$$\sum (y_i - \beta_0 - \beta_1 x_i) = 0 \implies n\beta_0 + \left(\sum x_i\right)\beta_1 = \sum y_i$$

$$\sum x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \implies \left(\sum x_i\right)\beta_0 + \left(\sum x_i^2\right)\beta_1 = \sum x_i y_i$$

These two scalar equations are exactly equivalent to the single matrix equation

$$X^T X\beta = X^T y$$

Solving (assuming X^TX is invertible, i.e., x_i's are not all identical)

$$\boxed{\hat{\beta} = (X^TX)^{-1}X^Ty}$$

Residuals and Sums of Squares, in Matrix Form

The residual vector and the two key sums of squares can be written compactly as

$$\hat{\epsilon} = y - \hat{y} = y - X\hat{\beta} = y - X(X^TX)^{-1}X^Ty$$

$$SS_{Res} = \hat{\epsilon}^T\hat{\epsilon} = y^Ty - y^TX(X^TX)^{-1}X^Ty$$

$$SST = y^Ty - \frac{1}{n^2}y^T\mathbf{1}\mathbf{1}^Ty$$

where $\mathbf{1}$ is the $n \times 1$ vector of all ones. This matrix machinery is what makes the extension from Simple to *Multiple* Linear Regression almost immediate , every formula above carries over unchanged, just with X now having $p + 1$ columns instead of 2.

2 Multiple Linear Regression

2.1 Introduction and Motivation

The Setup

Multiple Linear Regression expands our framework to handle a single continuous response variable, denoted as y, using multiple predictors, denoted as $x_1, \ldots, x_p$. Each observation in our dataset can be represented as a tuple $(y_i, x_{i1}, \ldots, x_{ip})$ for $i = 1, \ldots, n$.

Example: Weather Prediction

Suppose we want to predict rainfall on a given day i. The variables would be defined as follows:

- $i \to$ Specific day.
- $y_i \to$ Total Rainfall on day i (Response).
- $x_{i1} \to$ Total Temperature at 6 AM on day i (Predictor 1).
- $x_{i2} \to$ Relative Humidity at 8 AM on day i (Predictor 2).
- $x_{ip} \to$ Other contributing factors (Predictor p).

2.2 The Linear Model and Assumptions

The Model

For n observations, the multiple linear regression model is formulated as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad \forall\, i = 1, \ldots, n$$

Assumptions

We operate under the following assumptions for Multiple Linear Regression:

1. The linear model holds true.
2. ϵ_i are independent of each other.
3. ϵ_i are independent of the predictors $x_{i1}, \ldots, x_{ip}$.

4. $E[\epsilon_i] = 0$ and $Var[\epsilon_i] = \sigma^2$ (where ϵ_i are continuous random variables).

2.3 Matrix Notation

Writing the Model in Matrix Form

To handle multiple predictors efficiently, we stack our equations into matrices and vectors:

$$\underline{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1} \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}_{(p+1)\times 1} \qquad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}_{n\times 1}$$

$$X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}_{n\times(p+1)}$$

The overall model can now be written compactly as

$$Y = X\beta + \epsilon \quad \Longrightarrow \quad \epsilon = Y - X\beta$$

2.4 Least Squares Estimation

Least Squares Criterion

The goal is to minimize the sum of squares of errors, $\sum_{i=1}^{n} \epsilon_i^2$.
Using matrix notation, the sum of squared errors is

$$\sum \epsilon_i^2 = \underline{\epsilon}^T \underline{\epsilon} = (Y - X\beta)^T (Y - X\beta)$$

Expanding this expression gives the objective function $S(\beta)$:

$$\begin{aligned} S(\beta) &= Y^T Y - \beta^T X^T Y - Y^T X\beta + \beta^T (X^T X)\beta \\ &= Y^T Y - 2Y^T X\beta + \beta^T (X^T X)\beta \end{aligned}$$

```
# Deriving the Normal Equations
```

We need to take the derivative of $S(\beta)$ with respect to β and set it equal to 0 to find the normal equations.

$$\frac{\partial S(\beta)}{\partial \beta} = \begin{pmatrix} \frac{\partial S}{\partial \beta_0} \\ \frac{\partial S}{\partial \beta_1} \\ \vdots \\ \frac{\partial S}{\partial \beta_p} \end{pmatrix} = 0$$

Step 1 - Derivative of the linear term: Let A be a $1 \times (p+1)$ matrix, where $A = -2Y^T X$. We know that $\frac{\partial}{\partial \beta}(A\beta) = A^T$. Therefore:

$$\frac{\partial}{\partial \beta}(-2Y^T X\beta) = (-2Y^T X)^T = -2X^T Y$$

Step 2 - Derivative of the quadratic term: Consider the quadratic term with a symmetric matrix $B_{(p+1)\times(p+1)}$. The derivative of a quadratic form is $\frac{\partial(\beta^T B\beta)}{\partial \beta} = 2B\beta$. Since $X^T X$ is symmetric, we get:

$$\frac{\partial(\beta^T X^T X\beta)}{\partial \beta} = 2X^T X\beta$$

Step 3 - Solving for $\hat{\beta}$: Setting the full derivative to zero yields the Normal Equations:

$$-2X^T Y + 2X^T X\beta = 0 \quad \Longrightarrow \quad X^T X\beta = X^T Y$$

Multiplying both sides by the inverse of $(X^T X)$ gives the least squares estimate of β:

$$\boxed{\hat{\beta} = (X^T X)^{-1} X^T Y}$$

Dimensionality Condition

Condition: $n > p+1$. In regression, there must be more data points than predictor dimensions. Otherwise, if $n = p+1$, we can achieve a perfect fit with all $\hat{\epsilon}_i = 0$, leading to a trivial least squares solution $\hat{\beta} = X^T Y$. All formulation of the least squares or best line relies on having more points than dimensions.

2.5 Expectation, Variance, and Covariance of Estimators

Properties of $\hat{\beta}$

Using the rules of expectation $E[AX+b] = AE[X]+b$ and variance $Var(AX+b) = AVar(X)A^T$, we can find the properties of our estimator matrix.

```
# Unbiasedness and Variance
```

Expectation:

$$\begin{aligned} E[\hat{\beta}] &= \beta + (X^T X)^{-1} X^T E[\epsilon] \\ &= \beta \end{aligned}$$

Variance:

$$Var(\hat{\beta}) = (X^T X)^{-1} X^T Var(\epsilon) X (X^T X)^{-1}$$

If $Var(\epsilon) = \sigma^2 I$, then this simplifies to:

$$\boxed{Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}}$$

Variance Properties Reminder

If $Ax = \sum_{j=1}^{K} a_{ij} x_j$, then $Var(Ax + By) = a^2 Var(x) + b^2 Var(y) + 2abCov(x, y)$. For a random vector X and a matrix A: $Var(AX) = A\,Var(X) A^T$.

2.6 Diagnostics and Residuals

Verifying Model Assumptions

The fitted values are defined as $\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$. The residuals are given by $\hat{\epsilon}_i = y_i - \hat{y}_i$.
Assumptions need to be verified for the errors, which we do in practice by analyzing the residuals:

- Plot residuals against fitted values of each predictor to verify linear model dependence on x_i.
- Plot squared residuals against fitted values of each predictor to check for homoscedasticity (verifying that variance is the same).
- Use a Normal Q-Q plot if normality is assumed to check for outliers, influential observations, and leverage.

2.7 Sum of Squares Decomposition

Sum of Squares in Matrix Form

The total variation and residual errors can be decomposed as follows:

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

$$SS_{Res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\epsilon_i^2$$

$$= Y^TY - 2Y^TX\hat{\beta} + \hat{\beta}^T(X^TX)\hat{\beta}$$

Substituting $\hat{\beta} = (X^TX)^{-1}X^TY$, we can rewrite the Residual Sum of Squares cleanly as:

$$\boxed{SS_{Res} = Y^T\big(I - X(X^TX)^{-1}X^T\big)Y}$$

Degrees of Freedom and Evaluation Metrics

Degrees of Freedom:

- Regression (SS_{Reg}): p.
- Residual (SS_{Res}): $n - (p + 1)$.
- Total (SST): $n - 1$.

Key Metrics:

- $R^2 = 1 - \frac{SS_{Res}}{SST}$.
- Estimated variance: $\hat{\sigma}^2 = \frac{SS_{Res}}{n-(p+1)}$.
- Adjusted $R^2 = 1 - \frac{SS_{Res}/(n-(p+1))}{SST/(n-1)}$.