Linear Model

1. $Y = \beta_0 + \beta_1 \log x + \beta_2 x^2 + \epsilon \rightarrow$ A linear model

   $x = \frac{\beta_1 X}{\beta_0 + X} + \epsilon$ is a non linear.

   The 2$^{\text{nd}}$ one can't be expressed as a linear combination of known functions of X.

2. We are only dealing with models, which is not necessarily the truth.

   > "All models are wrong but some models are useful" – Box (1976)

   Not even looking for correct or true.

3. We do not claim a causal relation between X & Y.

   i.e. Trying to predict Y based on X.

   Not claiming X causes Y or Y causes X.

4. **Interpretation of the coefficients.**

   Univariate ($p = 1$) regression

   $Y = \beta_0 + \beta_1 X + \epsilon$

   - $\beta_0 \rightarrow$ expected value of Y when $X = 0$.
   - $\beta_1 \rightarrow$ avg change in Y for unit change in X.

   Horse Power (X_2) in car weights (X)

   For every value of X_2 fixed, Y is also a function of X_1.

   Multiple Regression

   $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon, \quad \beta_1 > 0.$

   If we disregard X_2, then the overall pattern of Y on X_1 is increasing.

   In decreasing $Y = \beta_0 + \beta_1 X_1 + \epsilon$, $\beta_1 < 0$.

   In multiple regression, the interpretation of β_1 is the avg change in Y for unit change in X_1, when X_2 is held constant.

   In general the regression coefficient measures the change in response for unit change in the corresponding predictor when all other predictors are held constant.

Collinearity

A linear dependence b/w 2 or more predictors/columns of the design matrix.

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

e.g. $x_1 = 2x - 7$

$x_1 = x_2 + 2x_3$

At the population level makes the coeff $(\beta_1, \beta_2, \beta_3)$ ill defined.

$y = x_1 + x_2 + \epsilon \quad (1, 1, 1)$
$y = 3 - 2x_2 + \epsilon \quad (0, 3)$
$y = 1.5x_1 + \epsilon \quad (1.5, 0)$

Infinitely many β's represent same plane.

At the estimate level The OLS estimate is $\hat{\beta} = (X^T X)^{-1} X^T Y$. If $x_1 = 2x_2$ then $(X^T X)$ is not full rank. As $\det(X^T X) = 0 \rightarrow$ then $(X^T X)^{-1}$ can't be defined.

Let $\det(X^T X)$ not be zero, but very small, then $(X^T X)^{-1}$ becomes unstable conceptually. This is called near collinearity & we should try to avoid. We want to remove some of the columns. But which ones?

- If the relation involves only 2 elements, then a pairwise plot e.g., $(x_1, x_2), (x_1, x_3), (x_2, x_3)$, will reflect an exact straight line in one of the plots. Then drop any one of the variables involved in that plot.
- Even a very high correlation b/w (x_1, x_3) dictates that we drop one of them.
- If $x_1 = x_2 + 2x_3 \rightarrow$ This can't be detected from a particular plot. One must check the determinant of sub-matrices. For predictors e.g., (X_1, X_2, X_3, X_4), analyze $(X^T X)^{-1}$.

Each diagonal element of the inverse is related to the determinant of the minor of the original matrix after removing that particular row and column.

→ The columns **not** involved in the linear relation will have small numbers in the corresponding diagonal of the inverse matrix.

- Retain those.

→ From the subset that is not retained, remove elements one by one & check if the overall determinant stabilizes.

→ **Collinearity is:** Linear or approximate linear relationship between the predictor variables.

→ **Why is it a problem?** The estimate $\hat{\beta}$ is unstable, i.e., the variance $(X^T X)^{-1}\sigma^2$ is very large.

→ **How to detect?** $\det(X^T X) \approx 0$. Also can be done by using a pairs plot.

L-6 Another Problem of Multiple Regression

Interaction

The nature of a relationship between y & X_1 depends on the value of X_2.

Our assumption is that every variable makes a distinct additive contribution to the response. The interaction model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

Start with a bigger (with interaction) model. Test the hypothesis whether the interaction term is zero. If the hypothesis $(H_0 : \beta_3 = 0)$ is not rejected, then we can go ahead with the simpler model (with no interaction).

Normal Equations & the Geometry of Least-Squares

The model is $Y = X\beta + \epsilon$. We find the coefficients by minimizing $\sum_{i=1}^{n} \epsilon_i^2 = ||\epsilon||^2$.

Geometrically, $Y \in R^n$, and the fitted values $X\beta$ form a plane (or subspace) in that n-dimensional space. The distance is minimized when we drop a perpendicular/normal from the point Y to the plane.

The normal equations are derived from this condition:

- The residual vector ϵ must be orthogonal to the column space of X: $X^T\epsilon = \mathbf{0}$.
- Substituting $\epsilon = Y - X\beta$, we get: $X^T(Y - X\beta) = \mathbf{0}$.
- This gives the final form: $X^T Y = X^T X\beta$.

Inference on Linear Regression

The linear model is $Y = X\beta + \epsilon$, with the standard assumptions $E[\epsilon] = \mathbf{0}$ and $\text{Var}(\epsilon) = \sigma^2 I_n$.

We now add a new assumption: ϵ follows a Normal distribution.

$$\epsilon \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 I_n)$$

This implies $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$ and $\text{Var}(\epsilon_i) = \sigma^2$. It follows that:

$$Y \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$$

Likelihood Function

The likelihood function $L(\beta, \sigma^2|\mathbf{Y})$ is the probability density of the observed data, viewed as a function of the parameters. Given that $\epsilon = Y - X\beta$:

$$\begin{aligned} L(\beta, \sigma^2|Y_1, \dots, Y_n) &= \frac{1}{(2\pi)^{n/2}|\sigma^2 I_n|^{1/2}} \exp\left\{-\frac{1}{2}(Y - X\beta)^T(\sigma^2 I_n)^{-1}(Y - X\beta)\right\} \\ &= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta)\right\} \end{aligned}$$

Log-Likelihood

It's often easier to work with the natural logarithm of the likelihood:

$$l(\beta, \sigma^2|\mathbf{Y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta)$$

MLE for β

To find the Maximum Likelihood Estimator (MLE) for β, we need to maximize the log-likelihood function. Notice that the only term involving β is $-(Y - X\beta)^T(Y - X\beta)$. Maximizing the log-likelihood is therefore equivalent to **minimizing** the sum of squared residuals, $(Y - X\beta)^T(Y - X\beta)$.

This is precisely the objective of the method of least squares. Therefore, under the normality assumption, the MLE for β is identical to the least squares estimator:

$$\hat{\beta}_{MLE} = \hat{\beta}_{LS} = (X^T X)^{-1} X^T Y$$

MLE for σ^2

To find the MLE for σ^2, we take the partial derivative of the log-likelihood with respect to σ^2 and set it to zero, substituting $\hat{\beta}$ for β:

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(Y - X\hat{\beta})^T(Y - X\hat{\beta}) \overset{\text{set}}{=} 0$$

Solving for σ^2 yields the MLE:

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}(Y - X\hat{\beta})^T(Y - X\hat{\beta}) = \frac{\text{SSR}}{n}$$

Note that this estimator is biased. The unbiased estimator for σ^2 that we typically use is $\hat{\sigma}^2 = \frac{\text{SSR}}{n-(p+1)}$.

To find $E[\textbf{SSR}]$

First, we express the residual vector, $Y - X\hat{\beta}$, in terms of the error vector ϵ.

$$\begin{aligned} \hat{\beta} &= (X^TX)^{-1}X^TY \\ Y &= X\beta + \epsilon \end{aligned}$$

Substituting these into the residual expression:

$$\begin{aligned} Y - X\hat{\beta} &= Y - X(X^TX)^{-1}X^TY \\ &= (I - X(X^TX)^{-1}X^T)Y \\ &= (I - X(X^TX)^{-1}X^T)(X\beta + \epsilon) \\ &= (X\beta - X(X^TX)^{-1}X^TX\beta) + (I - X(X^TX)^{-1}X^T)\epsilon \\ &= (X\beta - X\beta) + (I - X(X^TX)^{-1}X^T)\epsilon \\ \implies Y - X\hat{\beta} &= [I - X(X^TX)^{-1}X^T]\epsilon \end{aligned}$$

(a) Let's define the matrix $A = I - X(X^TX)^{-1}X^T$. It is symmetric ($A^T = A$) and idempotent ($A^2 = A$).

(b) Given the property that for a random vector $\mathbf{v} \sim \mathcal{N}(\mu, \Sigma)$, any linear transformation is $A\mathbf{v} \sim \mathcal{N}(A\mu, A\Sigma A^T)$.

Since $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$, the distribution of the residual vector $A\epsilon$ is:

$$Y - X\hat{\beta} \sim \mathcal{N}(\mathbf{0}, \sigma^2 AA^T) = \mathcal{N}(\mathbf{0}, \sigma^2 A)$$

This simplification works because A is symmetric ($A^T = A$) and idempotent ($A^2 = A$), so $AA^T = A^2 = A$.

(c) Now, we use a key theorem regarding quadratic forms of normal variables:

> **Theorem:** If a matrix A is idempotent and symmetric with rank K, and a random vector $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \sigma^2 A)$, then the quadratic form:
>
> $$\frac{1}{\sigma^2}\mathbf{U}^T\mathbf{U} \sim \chi^2_K$$
>
> where $K = \text{Rank}(A)$.

In our case, we let $U = Y - X\hat{\beta}$. It follows that:

$$\frac{1}{\sigma^2}U^T U = \frac{1}{\sigma^2}(Y - X\hat{\beta})^T(Y - X\hat{\beta}) = \frac{\text{SSR}}{\sigma^2} \sim \chi^2_K$$

(d) We know that the expected value of a chi-squared random variable is its degrees of freedom: $E[\chi^2_K] = K$.

This allows us to find the expected value of the Sum of Squared Residuals (SSR):

$$E\left[\frac{\text{SSR}}{\sigma^2}\right] = K \implies E[\text{SSR}] = \sigma^2 K$$

(e) The final step is to find K, the rank of matrix A. For an idempotent matrix, the rank is equal to its trace.

$$\begin{aligned} K = \text{Rank}(A) &= \text{trace}(A) = \text{tr}(I_n - X(X^TX)^{-1}X^T) \\ &= \text{tr}(I_n) - \text{tr}(X(X^TX)^{-1}X^T) \\ &= \text{tr}(I_n) - \text{tr}((X^TX)^{-1}X^TX) \quad \text{(using the cyclic property of trace)} \\ &= n - \text{tr}(I_{p+1}) \quad (\text{since } X^TX \text{ is } (p+1)\times(p+1)) \\ &= n - (p+1) \end{aligned}$$

Therefore, we arrive at the result:

$$E[\text{SSR}] = \sigma^2(n - (p+1))$$

Results to be discussed in next class

(a) The sampling distribution of the coefficient estimator is:

$$\hat{\beta} \sim \mathcal{N}_{p+1}(\beta, \sigma^2(X^TX)^{-1})$$

Furthermore, $\hat{\beta}$ and the variance estimator σ^2 are independent, where $s^2 = \frac{\text{RSS}}{n-(p+1)}$.

For a single coefficient $\hat{\beta}_i$, the marginal distribution is $\hat{\beta}_i \sim \mathcal{N}(\beta_i, \sigma^2((X^T X)^{-1})_{ii})$. The pivotal quantity for forming confidence intervals and conducting t-tests is:

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{s^2((X^T X)^{-1})_{ii}}} \sim t_{n-(p+1)}$$

(b) The scaled residual sum of squares follows a chi-squared distribution:

$$\frac{\text{SSR}}{\sigma^2} = \frac{(n-(p+1))s^2}{\sigma^2} \sim \chi^2_{n-(p+1)} \quad \text{(under normality of } \epsilon)$$

(c) Recall the estimators for the variance:

- $\hat{\sigma}^2_{MLE} = \frac{1}{n}\text{SSR}$ (Maximum Likelihood Estimator)
- $\hat{\sigma}^2_{unbaised} = \frac{1}{n-(p+1)}\text{SSR}$ (Unbiased Estimator)

(d) **Derivation of the Distribution of $\hat{\beta}$**

We start with the estimator (which is both the least squares and max likelihood solution) and the model:

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

Substituting the model into the estimator equation:

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T (X\beta + \epsilon) \\ &= (X^T X)^{-1} X^T X \beta + (X^T X)^{-1} X^T \epsilon \\ &= I\beta + (X^T X)^{-1} X^T \epsilon \\ &= \beta + A\epsilon, \quad \text{where } A = (X^T X)^{-1} X^T. \end{aligned}$$

Since $\hat{\beta}$ is a linear transformation of the normally distributed vector $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$, $\hat{\beta}$ is also normally distributed with mean $E[\hat{\beta}] = \beta$ and variance $\text{Var}(A\epsilon) = A(\sigma^2 I_n)A^T = \sigma^2 AA^T$.

$$AA^T = (X^T X)^{-1} X^T X ((X^T X)^{-1})^T = (X^T X)^{-1}$$

Thus, the distribution is $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1})$.

(e) **Special Case: Simple Linear Regression ($p = 1$)**

For the model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, the design matrix leads to:

$$(X^T X)^{-1} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}^{-1} = \frac{1}{n\sum x_i^2 - (\sum x_i)^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}$$

The denominator can be simplified to $n\sum(x_i - \bar{x})^2$. The variance of $\hat{\beta}_1$ depends on the $(2,2)$ entry of the matrix above:

Confidence Intervals (CI) and Hypothesis Tests for Regression Coefficients

We start with the sampling distribution of the intercept estimator, $\hat{\beta}_0$:

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma^2 \frac{\sum x_i^2}{n\sum(x_i - \bar{x})^2}\right)$$

To get the Standard Error (S.E.), we replace the unknown population variance σ^2 with its sample estimate, s^2.

Inference for the Slope (β_1)

Let's test if the slope is significantly different from zero.

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_a : \beta_1 \neq 0$$

If σ were known, our test statistic would be:

$$Z = \frac{\hat{\beta}_1 - 0}{\sigma/\sqrt{\sum(x_i - \bar{x})^2}} \sim \mathcal{N}(0,1)$$

Since σ is unknown, this is not a usable statistic. We must estimate it with s, where $s^2 = \frac{1}{n-2}\sum(y_i - \hat{y}_i)^2$. If we can show normal and chi squared are independent we have a t-distribution for our statistic.

$$t = \frac{\hat{\beta}_1 - 0}{s/\sqrt{\sum(x_i - \bar{x})^2}} \sim t_{n-2}$$

This result stems from the fact that $\frac{(n-2)SSR}{\sigma^2} \sim \chi^2_{n-2}$ and the t-distribution is a ratio of a standard normal and the square root of a scaled, independent chi-squared variable.

Inference for the Intercept (β_0)

We can perform a similar test for the intercept.

$$H_0 : \beta_0 = \beta_{0,\text{null}} \quad \text{vs.} \quad H_a : \beta_0 \neq \beta_{0,\text{null}}$$

The corresponding t-statistic is:

$$t = \frac{\hat{\beta}_0 - \beta_{0,\text{null}}}{s\sqrt{\frac{\sum x_i^2}{n\sum(x_i - \bar{x})^2}}} \sim t_{n-2}$$

Confidence Interval for β_0

A $(1-\alpha)\%$ confidence interval for β_0 is constructed by inverting the t-test:

$$\hat{\beta}_0 \pm t_{\alpha/2,n-2} \cdot s\sqrt{\frac{\sum x_i^2}{n\sum(x_i - \bar{x})^2}}$$

Confidence Interval for β_1

A $(1-\alpha)\%$ confidence interval for β_1 is :

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \cdot s\sqrt{\frac{1}{\sum(x_i-\bar{x})^2}}$$

$l = lm(y \sim x)$
summary(l)

Proof of Independence: $\hat{\beta} \perp s^2$

A critical result in linear regression is proving that the coefficient estimator, $\hat{\beta}$, is statistically independent of the variance estimator, s^2. This property is what allows us to form valid t-statistics for inference.

Since we define $s^2 = \frac{1}{n-(p+1)}U^T U$, where U is the residual vector, the proof simplifies to showing that $\hat{\beta}$ and U are independent.

The Proof

We can express both $\hat{\beta}$ and U as linear transformations of the same error vector, ϵ:

$$\hat{\beta} = \beta + \underbrace{(X^T X)^{-1} X^T}_{A} \epsilon$$
$$U = \underbrace{[I - X(X^T X)^{-1} X^T]}_{B} \epsilon$$

We can analyze their joint behavior by stacking them into a single vector:

$$\begin{pmatrix} \hat{\beta} \\ U \end{pmatrix} = \begin{pmatrix} \beta \\ \mathbf{0} \end{pmatrix} + \begin{pmatrix} (X^T X)^{-1} X^T \\ I - X(X^T X)^{-1} X^T \end{pmatrix} \epsilon$$

Since this stacked vector is a linear transformation of ϵ (which is normally distributed), the vector itself is jointly multivariate normal. A key property of the multivariate normal distribution is that if the covariance between two sub-vectors is zero, then the vectors are independent.

Our goal is to show that $\mathrm{Cov}(\hat{\beta}, U) = \mathbf{0}$. The covariance between the transformations is calculated as $A \cdot \mathrm{Var}(\epsilon) \cdot B^T$.

Given that $\text{Var}(\epsilon) = \Sigma = \sigma^2 I_n$, we just need to show that the matrix product $AB^T = \mathbf{0}$.

$$\begin{aligned}
AB^T &= \left((X^TX)^{-1}X^T\right)\left(I - X(X^TX)^{-1}X^T\right)^T \\
&= (X^TX)^{-1}X^T(I - X(X^TX)^{-1}X^T) && \text{(since the matrix B is symmetric)} \\
&= (X^TX)^{-1}X^T - (X^TX)^{-1}\underbrace{X^TX(X^TX)^{-1}}_{I}X^T \\
&= (X^TX)^{-1}X^T - (X^TX)^{-1}X^T \\
&= \mathbf{0}
\end{aligned}$$

Since the covariance matrix is the zero matrix, $\hat{\beta}$ and U are independent. This confirms that $\hat{\beta}$ and s^2 are also independent.

Summary for Inference

This independence is the foundation for all standard inference in linear models. The key results are:

- The coefficient estimator follows a Normal distribution: $\hat{\beta} \sim \mathcal{N}$
- The variance estimator follows a (scaled) Chi-squared distribution: $s^2 \propto \chi^2$
- Crucially, they are independent: $\hat{\beta} \perp s^2$

This allows for the construction of element-wise confidence intervals and hypothesis tests for each individual coefficient, β_i.

ANOVA (Analysis of Variance)

The total variability in the response can be partitioned into the variability explained by the regression model and the unexplained (residual) variability. This is summarized in an ANOVA table.

ANOVA Table for Simple Linear Regression

Source of Variation	df	SS (Sum of Squares)	MS (Mean Square)	F-statist
Regression (SSReg)	1	$\sum(\hat{y}_i - \bar{y})^2$	SSReg/1	$\frac{\text{MSReg}}{\text{MSR}}$
Residual/Error (SSR)	$n-2$	$\sum(y_i - \hat{y}_i)^2$	SSR/$(n-2)$	
Total (SST)	$n-1$	$\sum(y_i - \bar{y})^2$		

F-Test for Overall Model Significance

The F-test is used to determine if at least one predictor variable in the model is significantly related to the response variable.

- **Hypotheses:** For a general multiple regression model, the hypotheses are:

  $$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

  $$H_a : \text{At least one } \beta_j \neq 0$$

  We reject the null hypothesis (H_0) for large values of the F-statistic.
- **Theoretical Basis (Cochran's Theorem):** The distribution of the F-statistic relies on the properties of the sums of squares.
  - Under the null hypothesis, the scaled regression sum of squares follows a chi-squared distribution: $\frac{\text{SSReg}}{\sigma^2} \sim \chi^2_p$.
  - The scaled residual sum of squares always follows a chi-squared distribution: $\frac{\text{SSR}}{\sigma^2} \sim \chi^2_{n-(p+1)}$.
  - A key result is that SSReg and SSR are independent random variables.

Matrix Formulation of Sums of Squares

In matrix form, the linear model is written as $Y = X\beta + \epsilon$.

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

The sums of squares can be represented as quadratic forms of the vector Y. Their distributions rely on the assumption that the errors are normally and independently distributed, $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$.

- SST $= \sum(y_i - \bar{y})^2 \implies \frac{\text{SST}}{\sigma^2} \sim \chi^2_{n-1}$
- SSR $= \sum(y_i - \hat{y}_i)^2 \implies \frac{\text{SSR}}{\sigma^2} \sim \chi^2_{n-(p+1)}$

The total sum of squares can be decomposed into components using symmetric and idempotent matrices ($B^{(i)}$):

$$\epsilon^T \epsilon = \epsilon^T B^{(1)} \epsilon + \epsilon^T B^{(2)} \epsilon + \epsilon^T B^{(3)} \epsilon$$

Matrix Decomposition of Sums of Squares

The partitioning of the total sum of squares in ANOVA can be formally shown using a set of symmetric, idempotent projection matrices. Let's define:

- $B^{(1)} = \mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T$ (Corresponds to the correction for the mean)
- $B^{(2)} = I - X(X^T X)^{-1} X^T$ (The residual-maker matrix, for SSR)

- $B^{(3)} = X(X^TX)^{-1}X^T - \mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T$ (The regression matrix, for SSReg)

For such idempotent matrices, the rank is equal to the trace.

- $\text{rank}(B^{(1)}) = \text{tr}(B^{(1)}) = 1$
- $\text{rank}(B^{(2)}) = \text{tr}(I) - \text{tr}(X(X^TX)^{-1}X^T) = n - (p+1)$
- $\text{rank}(B^{(3)}) = \text{tr}(X(X^TX)^{-1}X^T) - \text{tr}(B^{(1)}) = (p+1) - 1 = p$

Notice that the ranks (degrees of freedom) sum to the total number of observations: $1 + (n - (p+1)) + p = n$.

Cochran's Theorem Application

By Cochran's Theorem, if we express the total sum of squares of the errors, $\epsilon^T\epsilon$, using these matrices, we find that:

i. The resulting quadratic forms, $\epsilon^T B^{(i)}\epsilon$, are distributed as scaled chi-squared variables: $\frac{1}{\sigma^2}\epsilon^T B^{(i)}\epsilon \sim \chi^2_{\text{rank}(B^{(i)})}$.

ii. These quadratic forms (which correspond to SSReg and SSR) are mutually independent.

This theorem provides the theoretical foundation for the degrees of freedom used in the ANOVA table and the independence required to form a valid F-statistic.

Comparing the F-test and t-tests

The overall F-test evaluates the null hypothesis that all regression coefficients (excluding the intercept) are simultaneously equal to zero.

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

- For testing $\beta = 0$, we can use even t-test or ANOVA test. These give same result when p=1. for $p \geq 2$, these give different results. In multiple ($p \geq 2$) regression , the order in which the variable are added make a difference. The F test has to be interpreted as the contribution of one predictor given (conditional) on the effect of all predictors being taken into account. The t-test looks at the marginal effect of each predictor.

Prediction and Confidence Intervals in Linear Regression

Model Setup

Let our observed data be $(x_1, y_1), \ldots, (x_n, y_n)$. The parametric model is:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \text{where } \epsilon_i \sim \text{i.i.d. } \mathcal{N}(0, \sigma^2)$$

The fitted value (our point estimate) for the response at a new point x_0 is:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

In matrix notation, this is written as $\hat{y}_0 = \mathbf{x}_0^T \hat{\beta}$, where the vector for the new observation is $\mathbf{x}_0^T = \begin{pmatrix} 1 & x_0 \end{pmatrix}$.

1. Confidence Interval for the Mean Response ($E[y|x_0]$)

This interval provides a range for the average value of y for all subjects with a predictor value of x_0.

- **Expected Value:** The estimator $\hat{y}_0$ is an unbiased estimator of the true mean response:

$$E[\hat{y}_0] = E[\mathbf{x}_0^T \hat{\beta}] = \mathbf{x}_0^T E[\hat{\beta}] = \mathbf{x}_0^T \beta = E[y_0]$$

- **Variance:** The variance of this estimator depends on how far x_0 is from the center of the data:

$$\text{Var}(\hat{y}_0) = \text{Var}(\mathbf{x}_0^T \hat{\beta}) = \mathbf{x}_0^T \text{Var}(\hat{\beta}) \mathbf{x}_0 = \sigma^2 \mathbf{x}_0^T (X^T X)^{-1} \mathbf{x}_0$$

- **Distribution:** The estimator is normally distributed:

$$\hat{\beta}_1 x_0 \sim \mathcal{N}\left(\mathbf{x}_0^T \beta, \sigma^2 \mathbf{x}_0^T (X^T X)^{-1} \mathbf{x}_0\right)$$

Since σ^2 is unknown, we use s^2 and the t-distribution. The $(1-\alpha)\%$ confidence interval is:

$$\hat{\beta}_0 \pm t_{\alpha/2, n-2} \cdot s \sqrt{\mathbf{x}_0^T (X^T X)^{-1} \mathbf{x}_0}$$