Linear Regression

Professor Jie Peng, PhD

Department of Statistics

University of California, Davis

ANOVA: F Tests

Expected Values of SS and MS

Under simple regression model:

- Expected values of SS:

$$E(SSE) = (n-2)\sigma^2, \quad E(SSR) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

- Expected values of MS:

$$E(MSE) = \sigma^2, \qquad E(MSR) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

- Note that $E(MSR) \geq E(MSE)$ and " $=$ " holds iff $\beta_1 = 0$.

Optional Reading: Sampling Distributions of SS

Under the Normal error model:

- $SSE \sim \sigma^2 \chi^2_{(n-2)}$
- SSE and SSR are independent.

F Test for Linear Association between X and Y

- $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$
- F ratio: $F^* = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/(n-2)}$
- Null distribution of F^*: $F^* \underset{H_0:\beta_1=0}{\sim} F_{1,n-2}$.
- Decision rule at the significance level α:
  - Critical value approach:

    $$\text{reject } H_0 \text{ if} \quad F^* > F(1-\alpha; 1, n-2),$$

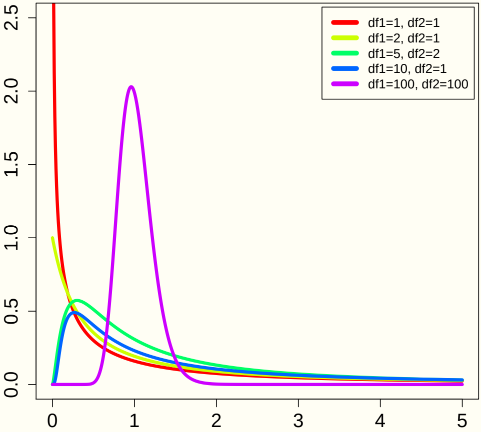
    where $F(1-\alpha; 1, n-2)$ is the $(1-\alpha)$100th percentile of the $F_{1,n-2}$ distribution.
  - P-value approach: reject H_0 if p-value$< \alpha$ where p-value$= P(F_{1,n-2} > F^*)$.

Optional Reading: Definition of F Distributions

If $Z_1 \sim \chi^2_{(df_1)}$, $Z_2 \sim \chi^2_{(df_2)}$ and Z_1, Z_2 are independent, then $\frac{Z_1/df_1}{Z_2/df_2} \sim F_{df_1,df_2}$.

F Distributions

Figure: F distributions: probability density function



Relationship between F Tests and T Tests

In simple linear regression, the F-test is equivalent to the **two-sided** t-test for testing $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$. This is because:

- $F^* = (T^*)^2$
- $F(1 - \alpha; 1, n - 2) = t^2(1 - \alpha/2; n - 2)$

ANOVA Table for Simple Regression

Source of Variation	SS	d.f.	MS=SS/d.f.	F^*
Regression	$SSR = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$	1	$MSR = SSR/1$	MSR/MSE
Error	$SSE = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$	$n-2$	$MSE = SSE/(n-2)$	
Total	$SSTO = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$	$n-1$		

Reading: Heights

Source of Variation	SS	d.f.	MS=SS/d.f.	F^*
Regression	$SSR = 1234$	1	$MSR = 1234$	245
Error	$SSE = 4659$	926	$MSE = 5.03$	
Total	$SSTO = 5893$	927		

- Test whether there is a linear association between parent's height and child's height at significance level $\alpha = 0.01$.
- $F(0.99; 1, 926) = 6.66 < F^* = 245$, so reject $H_0 : \beta_1 = 0$ and conclude that there is a significant linear association between parent's height and child's height.

Coefficient of Determination

Coefficient of Determination R^2

R^2 is a descriptive measure for **linear association** between X and Y:

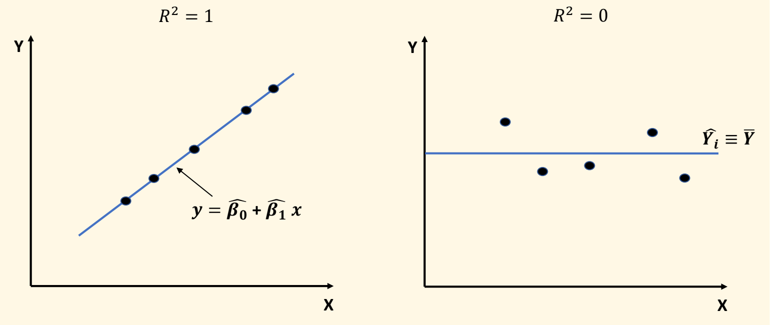
$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}.$$

- Heights: $R^2 = \frac{1234}{5893} = 0.209$. So 20% of variation in child's height may be explained by the variation in parent's height.

Properties of R^2

- $0 \leq R^2 \leq 1$.
  - In simple linear regression, $R^2 = r_{Xy}^2$.
- If all observations fall on one straight line, then $R^2 = 1$.
  - X accounts for all variation in the observations.
- If the fitted regression line is horizontal, i.e., $\hat{\beta}_1 = 0$, then $R^2 = 0$.
  - X is of no use in explaining variation in the observations.
  - There is no evidence of linear association between X and Y in the data.

Figure:



$R^2 = 1$
Y
X
$y = \widehat{\beta_0} + \widehat{\beta_1}\, x$
$R^2 = 0$
Y
X
$\widehat{Y}_i \equiv \overline{Y}$

Caution when Interpreting R^2

Is any of the following statements true?

- *"A large R^2 means that the estimated regression line must be a good fit of the data."*
- *"A near zero R^2 means that X and Y are not related."*

Figure:

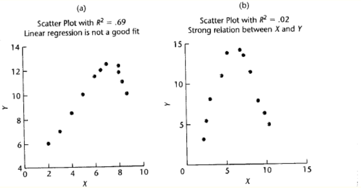


FIGURE 2.9 Illustrations of Two Misunderstandings about Coefficient of Determination.

Ref.: Applied Linear Statistical Models by Kutner, Nachtsheim, Neter and Li

If the relationship between X and Y is indeed linear, is any of the following statements true?

- *"A large R^2 means that there must be a (statistically) significant linear association between X and Y."*
- *"A near zero R^2 means that there is no (statistically) significant linear association between X and Y."*

Model Diagnostics:

Overview

Assumptions of the Normal Error Model

- **Linearity** of the regression relation
- **Constant variance** of the error terms
- **Independence** of the error terms
- **Normality** of the error terms

Consequences of Model Departures

- With regard to regression relation:
  - **Nonlinearity** of the regression relation: E.g.

    $Y_i = \beta_0 + \beta_1 X_i^2 + \varepsilon_i$
  - **Omission of important predictor variable(s)**: E.g.

    $Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \varepsilon_i$
- These departures usually have serious consequences such as biased estimators.

- With regard to error distribution:
  - **Nonconstant variance (a.k.a. heteroscedasticity)** or **Nonindependence** $\Longrightarrow$ invalid variance estimation $\Longrightarrow$ invalid inference (e.g., inflated type-I error, under-coverage of CIs)
  - **Nonnormality**: small departures – not serious due to CLT; major departures – could be serious for small data sets
- **Outliers**: could be serious for small data sets

Diagnostic Tools: Residual Plots

We use various residual plots to detect departures from the model.

- Examine regression relation and error variance:
  - residuals vs. fitted values
  - residuals vs. X variable(s)
  - residuals vs. omitted X variable(s)
- Examine error distribution:
  - Normality: normal probability plot (Q-Q plot) of residuals
  - Independence: sequence plot of residuals

Remedial Measures

Mild departures often do not need to be fixed. For more serious departures:

- Fix regression relation: transformation of the response variable and/or transformation(s) of the X variable(s);
- Fix error distribution: transformation of the response variable;

Model Diagnostics:

Nonlinearity Detection

Detection of Nonlinearity

We can use residuals vs. fitted values plot or residuals vs. X values plot:

- If the plot shows a clear nonlinear pattern, then it is an indication of possible nonlinearity in the regression relation. This is because the nonlinearity unaccounted for by the regression function of the model would be left in the residuals.
- If the plot shows a clear linear pattern, then it is an indication of possible omission of an important X variable. This is because the relation unaccounted for by the X-variables of the model would be left in the residuals.

Simulation Experiment

- Data: 30 cases with $X \sim N(100, 16^2)$, $\varepsilon \sim N(0, 10^2)$,

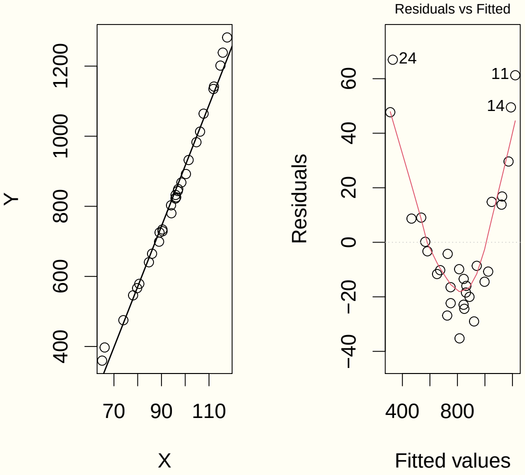
$$Y_i = 5 - X_i + 0.1X_i^2 + \varepsilon_i, \quad i = 1, \cdots, 30$$

- Fitted model: simple linear regression

Coefficients	Estimate	Std. Error	t value	$Pr(>\|t\|)$
Intercept	-811.8518	35.2767	-23.01	<2e-16 ***
X	17.2787	0.3695	46.76	<2e-16 ***

$\sqrt{MSE} = 27.6$, $R^2 = 0.9874$

Figure: Left: scatter plot; Right: residual vs. fitted value



Model Diagnostics: Unequal Variance Detection

Unequal Variance

- Sometimes variance increases (or decreases) with X-value. E.g., in financial data, the volume of transactions often has a role in the volatility of market.
- Data may come from different strata with different variability. E.g., measuring instruments with different precision may have been used to obtain the observations.

Detection of Nonconstancy in Variance

We can use the residuals vs. fitted values plot:

- If the plot shows an unequal spread of the residuals along the horizontal axis (i.e., the fitted values), then this is an indication of unequal variance.

Simulation Experiment

- Data: 100 cases with $X_i = \frac{i}{10}$, $u_i \sim N(0,1)$,

$$Y_i = 2 + 3X_i + \sigma(X_i) \cdot u_i, \quad i = 1, \cdots, 100,$$

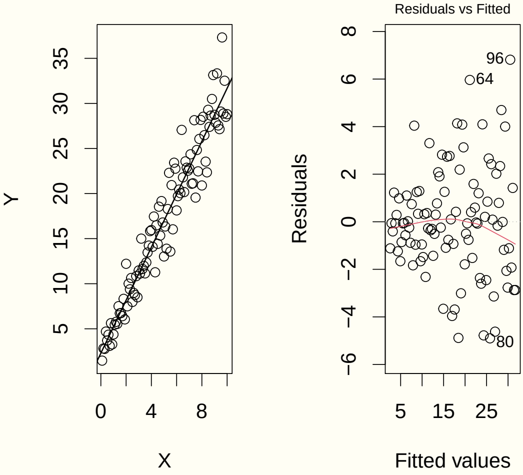
where $\sigma^2(x) = \exp(1 + 0.1x)$. Note that the error term $\varepsilon_i = \sigma(X_i) \cdot u_i$ has variance $\sigma^2(X_i)$ increasing with the X-value.

- Fitted model: simple linear regression

Coefficients	Estimate	Std. Error	t value	$Pr(>\|t\|)$
Intercept	2.29130	0.46689	4.908	3.67e-06 ***
X	2.93869	0.08027	36.612	< 2e-16 ***

$\sqrt{MSE} = 2.317$, $R^2 = 0.9319$.

Figure: Left: scatter plot; Right: residual vs. fitted value



Model Diagnostics:

Non-normality Detection

Detection of Non-normality

We can use *Normal probability plot (a.k.a. Normal Q-Q plot)* of residuals:

- If the residuals are (approximately) normally distributed, then the points on the Q-Q plot should be (nearly) on a straight line.
- Departures from the straight line pattern could indicate **skewed** (non-symmetry) or **heavy-tailed** (more probability mass on tails than a Normal distribution) distributions.
- Other types of departures (e.g., nonlinearity) may affect the distribution of the residuals, thus it is better to examine these assumptions before checking normality.

Reading: Q-Q Plot

Q-Q stands for quantile-quantile. Q-Q plot is a graphical tool to compare the empirical distribution of a sample with a reference distribution (by default: the reference is $N(0, 1)$).

- $e_{(k)}$'s – the *sample quantiles or empirical quantiles*: the kth smallest data in the sample;
- $z_{(k)}$'s – the *theoretical quantiles* under the reference distribution;
- Q-Q plot is simply the scatter plot of $e_{(k)}$'s vs. $z_{(k)}$'s
- A (nearly) straight line pattern indicates that the sample is likely from the reference distribution.

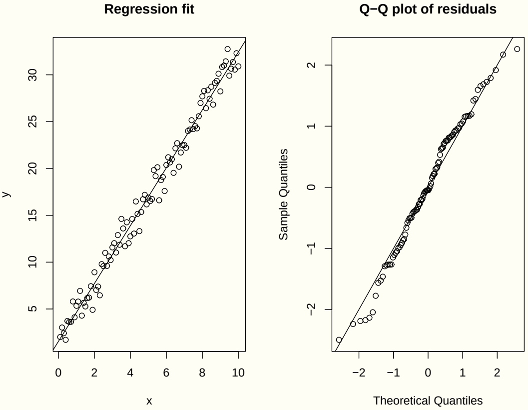
Case i	X_i	Y_i	$\widehat{Y}_i$	e_i
1	0.22	1.79	2.33	-0.54
2	3.55	5.66	5.90	-0.23
3	1.86	3.34	4.09	-0.75
4	3.29	5.83	5.62	0.22
5	1.25	4.74	3.43	1.31

$e_{(2)}$, the second smallest residual, is -0.54 and its corresponding theoretical quantile under Normality (with continuity correction) is:

$$z_{(2)} = \sqrt{MSE} \times Z\left((2 - 0.375)/(5 + 0.25)\right)$$

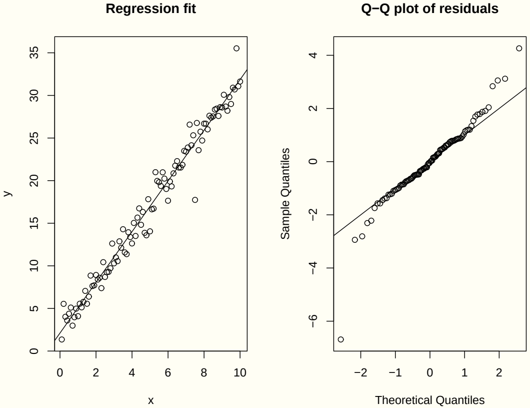
$$= \sqrt{0.8905} \times Z(0.31) = 0.944 \times (-0.497) = -0.469.$$

Error distribution: Normal$(0, 1)$



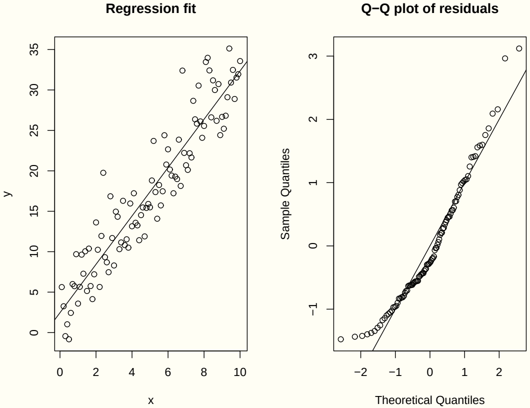
Normal Q-Q plot shows a straight line pattern.

Error distribution: $t_{(5)}$ – symmetrical but heavy-tailed



Normal Q-Q plot shows more probability mass on both tails compared to a Normal distribution.

Error distribution: centered $\chi^2_{(5)}$ – right-skewed



Normal Q-Q plot shows more probability mass on the right tail and

less probability mass on the left tail compared to a Normal

distribution.

Remedial Measures:

Transformations

Transformation of X

Transformation of X may be used to “linearize” a nonlinear relationship. Some commonly used transformations are:

- Increasing and concave downward: $X' = \log X$ or $X' = \sqrt{X}$
- Increasing and concave upward: $X' = X^2$ or $X' = \exp(X)$
- Decreasing and concave upward: $X' = 1/X$ or $X' = \exp(-X)$

Transformation of Y

Transformation of Y may be used to fix error distribution such as unequal variance or non-normality, which often appear together. Commonly used transformations are

- $Y' = \sqrt{Y}$
- $Y' = \log Y$
- $Y' = 1/Y$

- Sometimes, we may need to add a constant to the transformation, e.g. $X' = 1/(c + X)$, to avoid negative or near zero values.
- A simultaneous transformation of both X and Y might be needed to maintain a linear relationship.
- The **Box-Cox** procedure is often used to choose the appropriate power transformation Y^λ for the response variable.

  Notes: (i) $\lambda \approx 0$ *corresponds to the logarithm transformation:* $\log(Y)$; *(ii) to apply the Box-Cox procedure, the Y values must be all positive.*

Optional Reading: Box-Cox Procedure Details

- For each $\lambda \in R$, define the transformed observations as

$$Y_i^* = \begin{cases} K_1 \frac{Y_i^\lambda - 1}{\lambda}, & if, \quad \lambda \neq 0 \\ K_2 \log(Y_i), & if, \quad \lambda = 0 \end{cases}, \quad K_2 = (\prod_{j=1}^n Y_j)^{1/n}, \ K_1 = 1/K_2^{\lambda - 1}$$

- For each λ, fit a regression model on the transformed data Y^* and derive $SSE(\lambda)$ (or maximum loglikelihood).
- Find the λ that minimizes SSE (or maximizes maximum loglikelihood) and apply the corresponding power transformation.

Simple Regression: Matrix Form

Simple Linear Regression in Matrix Form

The regression equations:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \cdots n$$

can be expressed in a compact matrix form:

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times 2}{\mathbf{X}} \underset{2\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\epsilon}}$$

- **Response vector Y** and **error vector** : $n \times 1$ column vectors

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{bmatrix}, \qquad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_i \\ \vdots \\ \epsilon_n \end{bmatrix}$$

- **Design matrix**: $n \times 2$ matrix:

$$\mathbf{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_i \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}$$

- **Coefficient vector**: 2×1 column vector:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

The model assumptions:

$$E(\epsilon_i) = 0, \quad Var(\epsilon_i) = \sigma^2, \quad \text{for all } i = 1, \cdots, n$$

$$Cov(\epsilon_i, \epsilon_j) = 0, \quad \text{for all } i \neq j$$

can be expressed in matrix form:

$$\mathbf{E}\{\boldsymbol{\epsilon}\} = \mathbf{0}_n, \quad \boldsymbol{\sigma}^{\mathbf{2}}\{\boldsymbol{\epsilon}\} = \sigma^2 \mathbf{I}_n.$$

Mean of the error vector:

$$\mathbf{E}\{\boldsymbol{\epsilon}\} := \begin{bmatrix} E(\epsilon_1) \\ E(\epsilon_2) \\ \vdots \\ E(\epsilon_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{0}_n,$$

where $\mathbf{0}_n$ is the $n \times 1$ zero vector.

Variance-covariance matrix of the error vector:

$$\boldsymbol{\sigma^2}\{\boldsymbol{\epsilon}\}: \quad = \begin{bmatrix} Var(\epsilon_1) & Cov(\epsilon_1, \epsilon_2) & \cdots & Cov(\epsilon_1, \epsilon_n) \\ Cov(\epsilon_2, \epsilon_1) & Var(\epsilon_2) & \cdots & Cov(\epsilon_2, \epsilon_n) \\ \vdots & \vdots & \vdots & \vdots \\ Cov(\epsilon_n, \epsilon_1) & Cov(\epsilon_n, \epsilon_2) & \cdots & Var(\epsilon_n) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}_n,$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix.

Mean response vector: $n \times 1$ column vector:

$$\mathbf{E}\{\mathbf{Y}\} = \begin{bmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_i) \\ \vdots \\ E(Y_n) \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \vdots \\ \beta_0 + \beta_1 X_i \\ \vdots \\ \beta_0 + \beta_1 X_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_i \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \mathbf{X}\boldsymbol{\beta}.$$

Summary

The simple regression model can be expressed as:

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times 2}{\mathbf{X}}\,\underset{2\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\epsilon}}$$

- $\boldsymbol{\epsilon}$ is a random vector with $\mathbf{E}\{\boldsymbol{\epsilon}\} = \mathbf{0}_n,\ \boldsymbol{\sigma}^{\mathbf{2}}\{\boldsymbol{\epsilon}\} = \sigma^2\mathbf{I}_n$.
- Normal error model: $\boldsymbol{\epsilon} \sim \mathrm{Normal}_n(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$.

In terms of the response vector:

$$\mathbf{E}\{\mathbf{Y}\} = \mathbf{X}\boldsymbol{\beta}, \quad \boldsymbol{\sigma}^{\mathbf{2}}\{\mathbf{Y}\} = \sigma^2\mathbf{I}_n.$$

Least Squares Estimation: Matrix Form

Least Squares Estimation in Matrix Form

Least squares criterion:

$$Q(b_0, b_1) = \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_i))^2$$

can be expressed in matrix form : $\mathbf{b} = (b_0, b_1)^T$

$$Q(\mathbf{b}) = (\mathbf{Y} - \mathbf{Xb})' (\mathbf{Y} - \mathbf{Xb}) = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{Xb} + \mathbf{b}'\mathbf{X}'\mathbf{Xb}.$$

LS estimators:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{bmatrix} \overline{Y} - \hat{\beta}_1 \overline{X} \\ \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} \end{bmatrix},$$

provided that X_is are <u>not</u> all equal.

- Note that $\hat{\boldsymbol{\beta}}$ is a linear transformation of the observations vector **Y**.

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{bmatrix}, \quad \mathbf{X}'\mathbf{Y} = \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_i Y_i \end{bmatrix}.$$

When

$$D := n\sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i\right)^2 = n\sum_{i=1}^n (X_i - \overline{X})^2 \neq 0$$

$$\begin{aligned} (\mathbf{X}'\mathbf{X})^{-1} &= \begin{bmatrix} \frac{\sum_{i=1}^n X_i^2}{n\sum_{i=1}^n (X_i - \overline{X})^2} & -\frac{\sum_{i=1}^n X_i}{n\sum_{i=1}^n (X_i - \overline{X})^2} \\ -\frac{\sum_{i=1}^n X_i}{n\sum_{i=1}^n (X_i - \overline{X})^2} & \frac{n}{n\sum_{i=1}^n (X_i - \overline{X})^2} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2} & -\frac{\overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2} \\ -\frac{\overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2} & \frac{1}{\sum_{i=1}^n (X_i - \overline{X})^2} \end{bmatrix}. \end{aligned}$$

Reading: Deriving LS Estimator

- Differentiate $Q(\cdot)$ with respect to $\mathbf{b}$: $\frac{\partial}{\partial \mathbf{b}} Q = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\mathbf{b}$.
- Set the gradient to zero $\Longrightarrow$ *normal equation*:

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}.$$

- Multiply both sides by $(\mathbf{X}'\mathbf{X})^{-1}$:

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

- The left hand side becomes $\mathbf{I}_2\mathbf{b} = \mathbf{b}$, and the right hand side is the solution.

Fitted Values and Residuals: Matrix Form

Fitted Values and Residuals

- Fitted values vector: $n \times 1$ column vector:

$$\widehat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y},$$

  where $\mathbf{H} := \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is called the **hat matrix**.

- Residuals vector: $n \times 1$ column vector:

$$\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y}.$$

- Note that, fitted values vector $\widehat{\mathbf{Y}}$ and residuals vector $\mathbf{e}$ are linear transformations of the observations vector $\mathbf{Y}$.

Hat Matrix

The hat matrix

$$\underset{n\times n}{\mathbf{H}} := \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

and the matrix

$$\mathbf{I}_n - \mathbf{H} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

are $n \times n$ **projection matrices**. Meaning that they satisfy:

- **Symmetric**: $\mathbf{H}' = \mathbf{H}, \quad (\mathbf{I}_n - \mathbf{H})' = \mathbf{I}_n - \mathbf{H}$
- **Idempotent**: $\mathbf{H}^2 := \mathbf{H}\mathbf{H} = \mathbf{H}, \quad (\mathbf{I}_n - \mathbf{H})^2 = \mathbf{I}_n - \mathbf{H}$.

Moreover, $rank(\mathbf{H}) = 2$, $rank(\mathbf{I}_n - \mathbf{H}) = n - 2$ (provided that X_is are not all equal).

LS Estimation: Mean and Variance

Review: Linear Transformations of Random Vector

If $\mathbf{Z}$ is an $r \times 1$ random vector, and $\mathbf{A}$ is an $s \times r$ non-random matrix,

then

$$\underset{s\times 1}{\mathbf{W}} = \underset{s\times r}{\mathbf{A}}\ \underset{r\times 1}{\mathbf{Z}}$$

is an $s \times 1$ random vector with

$$\begin{aligned} \mathbf{E}\{\mathbf{W}\} &= \mathbf{E}\{\mathbf{AZ}\} = \mathbf{AE}\{\mathbf{Z}\} \\ \sigma^2\{\mathbf{W}\} &= \sigma^2\{\mathbf{AZ}\} = \mathbf{A}\sigma^2\{\mathbf{Z}\}\mathbf{A}' \end{aligned}$$

If further $\mathbf{B}$ is a $t \times r$ non-random matrix, then

$$Cov(\mathbf{AZ}, \mathbf{BZ}) = \mathbf{A}\sigma^2\{\mathbf{Z}\}\mathbf{B}'$$

LS Estimation: Expectations

- LS estimator is unbiased:

$$\mathbf{E}\{\hat{\beta}\} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{E}\{\mathbf{Y}\} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta = \beta$$

- Expectation of the fitted values:

$$\mathbf{E}\{\widehat{\mathbf{Y}}\} = \mathbf{E}\{\mathbf{X}\hat{\beta}\} = \mathbf{X}\mathbf{E}\{\hat{\beta}\} = \mathbf{X}\beta = \mathbf{E}\{\mathbf{Y}\}$$

- Expectation of the residuals:

$$\mathbf{E}\{\mathbf{e}\} = \mathbf{E}\{\mathbf{Y} - \widehat{\mathbf{Y}}\} = \mathbf{E}\{\mathbf{Y}\} - \mathbf{E}\{\widehat{\mathbf{Y}}\} = \mathbf{0}_n$$

LS Estimation: Variance-Covariance Matrices

Variance-covariance of the LS estimator:

$$
\begin{aligned}
\boldsymbol{\sigma}^{\mathbf{2}}\{\hat{\boldsymbol{\beta}}\} &= \boldsymbol{\sigma}^{\mathbf{2}}\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\} = \left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\boldsymbol{\sigma}^{\mathbf{2}}\{\mathbf{Y}\}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)' \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 \begin{bmatrix} \frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2} & -\frac{\overline{X}}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \\ -\frac{\overline{X}}{\sum_{i=1}^{n}(X_i - \overline{X})^2} & \frac{1}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \end{bmatrix}
\end{aligned}
$$

What is the covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$? What happens if $\overline{X} = 0$?

- Variance-covariance of the fitted values:

$$\boldsymbol{\sigma}^{\mathbf{2}}\{\widehat{\mathbf{Y}}\} = \mathbf{H}\boldsymbol{\sigma}^{\mathbf{2}}\{\mathbf{Y}\}\mathbf{H}' = \sigma^2\mathbf{H}$$

- Variance-covariance of the residuals:

$$\boldsymbol{\sigma}^{\mathbf{2}}\{\mathbf{e}\} = (\mathbf{I}_n - \mathbf{H})\boldsymbol{\sigma}^{\mathbf{2}}\{\mathbf{Y}\}(\mathbf{I}_n - \mathbf{H})' = \sigma^2(\mathbf{I}_n - \mathbf{H})$$

Are residuals uncorrelated? Do they have the same variance? How about the fitted values? What are the covariances between the residuals and fitted values?