STAT 331: Applied Linear Regression

Professor Kun Liang
LaTeXer Iris Jiang

Spring 2020

Contents

1. Introduction to Regression

1.1 What is regression

Definition 1.1 — Regression analysis.
Regression analysis is a statistical methodology that models the functional relationship between a response variable y and one or more explanatory variables $x_1, x_2, \ldots, x_p$.

A typical regression model is:

$$y = f(x_1, x_2, \ldots, x_p) + \epsilon$$

- y: dependent variable or response variable
- $x_1, x_2, \ldots, x_p$: covariates, explanatory variables, independent variables, or predictors
- ϵ: random error term

Regression models can be used to:

- Identify important predictors
- Estimate regression coefficients
- Estimate the response for given values of predictors
- Predict of future values of response

In STAT 331, we focus on the simplest form of regression: linear models

$$\begin{aligned} y &= f(x_1, x_2, \ldots, x_p) + \epsilon \\ &= \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon \end{aligned}$$

where the β's are the regression parameters(coefficients).
Linear in the parameter (not predictor). Linear model is the basic building block of more complicated models

R We refer to the model as linear in the parameters β's ($\frac{\partial f}{\partial \beta_i}$ do not depend on the parameters)

Are the following models linear?

(1) $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$
(2) $f(x) = \beta_0 + \beta_1 e^{\beta_2 x}$
(3) $f(x_1, x_2) = \beta_0 + \beta_1 x_1 x_2$

(1) This is a linear model. The predictor is x, this is not a linear model on the predictor but we define the linear model as to parameter, $\beta_0, \beta_1, \beta_2$ in this case.
(2) This is not a linear model. If taking derivative to β_1, the result involves β_2.
(3) This is a linear model.

1.2 Why linear model?

- Linear model is easy to implement and interpret
- All functions can be approximated locally by a linear function
- The simplest starting model to fit

1.3 Sample vs. population

Definition 1.2 — Sample.
A **sample** is the collection of units (people, animals, cities, whatever you study) that is actually measure or surveyed.

Definition 1.3 — Population.
The **population** is the large group of unites we are interested in, from which the sample was selected.

R We assume the data we have a representative sample (random sample) from a larger population

2. Simple Linear Regression (SLR)

2.1 Population model

$$y = \beta_0 + \beta_1 x + \epsilon$$

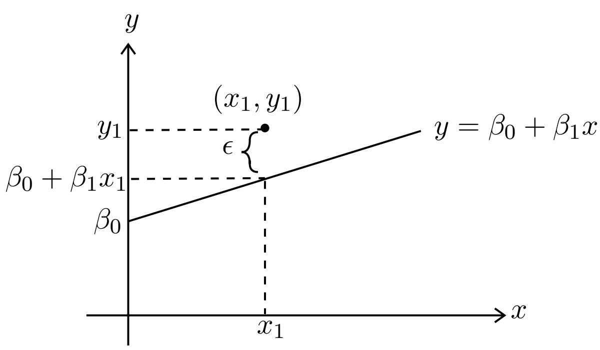
- y: response
- β_0, β_1: regression Coefficients
- x: predictor
- ϵ: random error
- $\beta_0 + \beta_1 x$: systematic (deterministic) part

Observed sample: suppose we have n pairs of observations (x_i, y_i), $i = 1, \ldots, n$. Then

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- x_i: fixed and known (for this course)
- β: fixed and unknown
- ϵ_i: random and unknown
- y_i: random and known

"known" means we can observe them



- β_0: intercept
- β_1: slope

2.2 Assumptions

(1) $E(\epsilon_i) = 0$
(2) $\epsilon_1, \ldots, \epsilon_n$ are statistically independent
(3) Constant variance: $\text{Var}(\epsilon_i) = \sigma^2 \implies \text{Var}(y_i) = \sigma^2$
The randomness of y_i comes from ϵ_i
(4) ϵ_i is normally distributed. $\epsilon_i \sim N(0, \sigma^2)$ and $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

Note 2.1 Assumption 1 to 3 are called Gauss-Markov assumptions.

Assumption 4 is stronger than all 3 assumptions combined.
1 to 3 are useful is you want to point estimate β.
4 is useful for further results.
There is no guarantee that the assumptions are correct. We will talk about model diagnostic and checking these assumptions.

2.3 Least Square Estimation (LSE)

2.3.1 Task

Given the sample observation (x_i, y_i), $i = 1, \ldots, n$, estimate (β_0, β_1) as $(\hat{\beta}_0, \hat{\beta}_1)$ such that the values of

$$r_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x = y_i - \hat{y}_i$$

are "small".

- r_i: residual
- y_i: fitted value

We define discrepancy function

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n} r_i^2$$

The reason we use square: the least square method provides an elegant solution; ϵ's follow normal distribution the Least Squared Estimation has the equivalence with Maximum Likelihood estimation.

2.3.2 Goal and Derivation

Minimize $S(\beta_0, \beta_1)$, i.e. solve$\begin{cases} \frac{\partial S(\beta_0,\beta_1)}{\partial \beta_0} = 0 \\ \frac{\partial S(\beta_0,\beta_1)}{\partial \beta_1} = 0 \end{cases}$.

$$\begin{cases} \frac{\partial S(\beta_0,\beta_1)}{\partial \beta_0} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-1) = 0 \\ \frac{\partial S(\beta_0,\beta_1)}{\partial \beta_1} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-x_i) = 0 \end{cases}$$

This is called Normal Equation

$$\implies \begin{cases} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 & \ldots(1) \\ \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)(x_i) = 0 & \ldots(2) \end{cases}$$

$$(1) \implies n\hat{\beta}_0 = \sum_{i=1}^n y_i - \hat{\beta}_1 \sum_{i=1}^n x_i \text{ or } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \text{ where } \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i \text{ and } \bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$$

$$(2) \implies \sum_{i=1}^n x_i y_i - \hat{\beta}_0 \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 = 0$$

we can replace $\hat{\beta}_0$ from previous results.

$$\implies \sum_{i=1}^n x_i y_i - \bar{y}\sum_{i=1}^n x_i - \hat{\beta}_1 \bar{x} \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 = 0$$

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

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \text{ where } S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) \text{ and } S_{xx} = \sum (x_i - \bar{x})^2$$

$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ is the fractions of the sample covariance between x and y over the sampel variance of x

Thus, the solution is

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \end{cases}$$

S_{xx} purely comes from x so is considered fixed, however S_{xy} is the joint quantity between x and y, so it is considered random. Both are functions of y.

2.3.3 The properties of $\hat{\beta}_0$ and $\hat{\beta}_1$

Property 2.1 — The properties of $\hat{\beta}_0$ and $\hat{\beta}_1$.

(1) LSEs are unbiased, i.e. $E(\hat{\beta}_0) = \beta_0$ and $E(\hat{\beta}_1) = \beta_1$

(2) $\text{Var}\left(\hat{\beta}_0\right) = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}})$

$\text{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{s_{xx}}$

(3) $\text{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) = -\frac{\sigma^2 \bar{x}}{s_{xx}}$

Proof. Express $\hat{\beta}_1$ as a linear combination of y_i's.

$$\begin{aligned} \hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{S_{xx}} \\ &\quad (\text{Since } \sum_{i=1}^n (x_i - \bar{x})\bar{y} = \bar{y}\sum_{i=1}^n (x_i - \bar{x}) = 0) \\ &= \frac{\sum (x_i - \bar{x}) y_i}{S_{xx}} \\ &= \sum \frac{x_i - \bar{x}}{S_{xx}} y_i \\ &\quad (\frac{x_i - \bar{x}}{S_{xx}} \text{ can be seen as some constant } c_i) \end{aligned}$$

The constant c_i's have a few properties.

(i) $\sum c_i = \sum \frac{x_i - \bar{x}}{S_{xx}} = 0$

(ii) $\sum c_i x_i = \frac{1}{S_{xx}} \sum x_i (x_i - \bar{x}) = \frac{1}{S_{xx}} \left(\sum x_i^2 - \bar{x} \sum x_i\right) = 1$

(iii) $\sum c_i^2 = \sum \frac{(x_i-\bar{x})^2}{S_{xx}} = \frac{1}{S_{xx}}$

$$E(\hat{\beta}_1) = E\left(\sum c_i y_i\right) = \sum c_i E(y_i)$$

Recall $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, and $\beta_0 + \beta_1 x_i$ is a fix constant; ϵ_i is the only random part, and we have $E(\epsilon_i) = 0$

$$\begin{aligned}
E(\hat{\beta}_1) &= \sum c_i E(y_i) = \beta_0 \underbrace{\sum c_i}_{=0} + \beta_1 \underbrace{\sum c_i x_i}_{=1} = \beta_1 \\
\mathrm{Var}\left(\hat{\beta}_1\right) &= \mathrm{Var}\left(\sum c_i y_i\right) \text{ (y_i's are independent)} \\
&= \sum c_i^2 \mathrm{Var}(y_i) = \sigma^2 \sum c_i^2 \\
&= \frac{\sigma^2}{S_{xx}} \\
E(\hat{\beta}_0) &= E(\bar{y} - \bar{x}\hat{\beta}_1) = \frac{1}{n}\sum E(y_i) - \bar{x}E(\hat{\beta}_1) \\
&= \frac{1}{n}\sum(\beta_0 + \beta_1 x_i) - \bar{x}\beta_1 \\
&= \beta_0 + \beta_1 \frac{1}{n}\sum x_i - \beta_1 \bar{x} \\
&= \beta_0
\end{aligned}$$

Note that

$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1 = \sum(\frac{1}{n}y_i - \bar{x}c_i y_i) = \sum(\frac{1}{n} - c_i\bar{x})y_i$$

$\frac{1}{n} - c_i\bar{x}$ is constant and let's call it k_i.

$$\begin{aligned}
\mathrm{Var}\left(\hat{\beta}_0\right) &= \mathrm{Var}\left(\sum k_i y_i\right) = \sum k_i^2 \mathrm{Var}(y_i) \\
&= \sigma^2 \sum k_i^2 = \sigma^2 \sum(\frac{1}{n} - c_i\bar{x})^2 \\
&= \sigma^2 \sum(\frac{1}{n^2} - \frac{2}{n}c_i\bar{x} + c_i^2\bar{x}^2) \\
&\quad \text{(Recall } \sum c_i = 0 \text{ and } \sum c_i^2 = \frac{1}{S_{xx}}) \\
&= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) \\
\mathrm{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) &= \mathrm{Cov}\left(\sum_{i=1}^n k_i y_i, \sum_{j=1}^n c_j y_j\right) = \sum_{i=1}^n \sum_{j=1}^n k_i c_i \mathrm{Cov}(y_i, y_j) \\
&\quad \left(\text{Note } \mathrm{Cov}(y_i, y_j) = \begin{cases} \sigma^2 & i = j \\ 0 & i \neq j \end{cases}\right) \\
&= \sigma^2 \sum_i k_i c_i = \sigma^s \sum(\frac{1}{n} - c_i\bar{x})c_i \\
&= \sigma^2\left(\frac{1}{n}\sum c_i - \bar{x}\sum c_i^2\right) \\
&= -\frac{\sigma^2\bar{x}}{S_{xx}}
\end{aligned}$$

■

2.3.4 Properties of the residual r_i

Recall

$$r_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1$$

Property 2.2 — Properties of the residual r_i.
Under LS fit

(1) $\sum_{i=1}^{n} r_i = 0$
(It comes from $\frac{\partial S(\beta_0,\beta_1)}{\partial \beta_0} = 0$)
(2) $\sum_{i=1}^{n} r_i x_i = 0$
(It comes from $\frac{\partial S(\beta_0,\beta_1)}{\partial \beta_1} = 0$)
(3) $r_i \hat{y}_i = 0$
(It comes from $\sum r_i(\hat{\beta}_0 + \hat{\beta}_1 x_i) = \hat{\beta}_0 \sum r_i + \hat{\beta}_1 \sum r_i x_i = 0$)
(4) The point $(\bar{x}, \bar{y})$ is always on the fitted regression line
(It comes from $\hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y}$)

The first property shows the residual vector and the 1 vector are perpendicular since the inner product equals to 0.
The second property shows the residual vector and the $\vec{x}$ are perpendicular since the inner product equals to 0
The third property shows the residual vector and the fitted value $\vec{y}$ are perpendicular since the inner product equals to 0

2.3.5 The Estimator of σ^2

Notice that $\begin{cases} \epsilon_i = y_i - (\beta_0 + \beta_1 x_i) \\ r_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \end{cases}$

Recall that $\epsilon_i \sim N(0, \sigma^2) \implies E(\epsilon_i^2) = E^2(\epsilon_i) + \mathrm{Var}(\epsilon_i) = \sigma^2$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i$

Instead $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} r_i$, however $E(\hat{\sigma}^2) \neq \sigma^2$

If we define $s^2 = \frac{1}{n-2}\sum_{i=1}^{n} r_i^2$, then $E(s^2) = \sigma^2$.
Intuitively, we need to estimate β_0 and β_1, and we only have $n-2$ degrees of freedom (d.f.) left to estimate σ^2.
Think about $(r_1, \ldots, r_n)$, $\sum r_i = 0$ and $\sum r_i x_i = 0$

2.3.6 Confidence Interval and Hypothesis Testing

Results: $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$
Recall $\hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i$, $E(\hat{\beta}_1) = \beta_1$ and $\mathrm{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{S_{xx}}$
If σ^2 is known,

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1)$$

Replace σ^2 with s^2,

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{s^2}{S_{xx}}}} \sim t_{n-2}$$

d.f. is $n-2$ since we have to estimate 2 variables. note the numerator and denominator are independent, and $\hat{\beta}_1$ is normally distributed and $\sqrt{\frac{s^2}{S_{xx}}}$ can be viewed as a square root of a *chi* distribution, hence the whole thing follows a T distribution.
$\sqrt{\frac{s^2}{S_{xx}}}$ is the standard error of $\hat{\beta}_1 - \beta_1$
A $100(1-\alpha)\%$ confidence interval (C.I.) for β_1 is:

$$\Pr(-t_{n-2,\frac{\alpha}{2}} < \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{s^2}{S_{xx}}}} < t_{n-2,\frac{\alpha}{2}}) = 1-\alpha$$

$t_{n-2,\frac{\alpha}{2}}$ denotes the upper $\frac{\alpha}{2}$ quantile of t_{n-2}

$$\Pr(\hat{\beta}_1 - t_{n-2,\frac{\alpha}{2}}\text{SE}(\hat{\beta}_1) < \beta_1 < \hat{\beta}_1 + t_{n-2,\frac{\alpha}{2}}\text{SE}(\hat{\beta}_1)) = 1-\alpha$$

Thus, the C.I. is $[\hat{\beta}_1 - t_{n-2,\frac{\alpha}{2}}\text{SE}(\hat{\beta}_1), \hat{\beta}_1 + t_{n-2,\frac{\alpha}{2}}\text{SE}(\hat{\beta}_1)]$ or $\hat{\beta}_1 \pm t_{n-2,\frac{\alpha}{2}}\text{SE}(\hat{\beta}_1)$

2.3.7 Hypothesis Testing

$H_0 : \beta_1 = \beta_2^*$ versus $H_a : \beta_1 \neq \beta_1^*$
Under H_0,

$$t = \frac{\hat{\beta}_1 - \beta_1^*}{\text{SE}(\hat{\beta}_1)} \sim t_{n-2}$$

Under H_0, t should follow a standard t distribution, however under H_a, t might follows a "nonsense" t distribution (a shift to the left or right to the standard t distribution).
If $|t| = \left|\frac{\hat{\beta}_1 - \beta_1^*}{\text{SE}(\hat{\beta}_1)}\right| \geq t_{n-2,\frac{\alpha}{2}}$, we reject H_0 at the significance level α.
Alternatively, we compute the p-value (the probability we observe a test statistic that is as extreme are more extreme than the observed one):

$$p = \Pr(|T| \geq |t|), \text{ where } T \sim t_{n-2}$$

and reject H_0 if $p \leq \alpha$

- T: random variable follows the reference distribution under H_0
- t: observed statistic

Typically, $H_0 : \beta_1 = 0$ is often of interest

2.4 Prediction

Inference of $\mu_0 = \beta_0 + \beta_1 x_0$ for some predictor value x_0

$$y_0 = \beta_0 + \beta_1 x_0 + \epsilon_0$$

$$E(\epsilon_0) = 0;\ E(y_0) = \beta_0 + \beta_1 x_0 = \mu_0$$

To estimate μ_0, we compute $\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.
Recall

$$\hat{\beta}_1 = \sum c_i y_i$$

where $c_i = \frac{x_i - \bar{x}}{S_{xx}}$ and $S_{xx} = \sum (x_i - \bar{x})^2$.

$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1 = \sum k_i y_i$$

where $k_i = \frac{1}{n} - c_i\bar{x} = \frac{1}{n} - \frac{\bar{x}(x_i-\bar{x})}{S_{xx}}$
It is easy to show that

$$\hat{\mu}_0 = \sum d_i y_i$$

where $d_i = \frac{1}{n} + \frac{(x_0-\bar{x})(x_i-\bar{x})}{S_{xx}}$.
$\hat{\mu}_0$ can be written as a linear combination of y_i's
Note that when $x_0 = \bar{x}$, $d_i = \frac{1}{n}$ for all i, then

$$\hat{\mu}_0 = \bar{y} = \hat{\beta}_0 + \hat{\beta}_1\bar{x}$$

If both x_0 and x_i are on the same side of $\bar{x}$ you get up-weighted, if one of them is below $\bar{x}$ and the other is above $\bar{x}$, thwn you get down-weighted.

$$E(\hat{\mu}_0) = E(\hat{\beta}_0 + \hat{\beta}_1 x_0) = E(\hat{\beta}_0) + x_0 E(\hat{\beta}_1) = \beta_0 + x_0\beta_1 = \mu_0$$

since $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators.

$$\mathrm{Var}(\hat{\mu}_0) = \mathrm{Var}\left(\sum d_i y_i\right) = \sum d_i^2 \mathrm{Var}(y_i) = [\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}]\sigma^2$$

The variance can be reduced by increasing the sample size n.

2.4.1 Prediction of future values

Question: What is your best guess of the value of y given that $x = x_p$?

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \to ((x_1, y_1), \cdots, (x_n, y_n)) \to (x_p, y_p)$$

Model: $y_p = \beta_0 + \beta_1 x_p + \epsilon_p$.
Prediction: $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_p$

Results of $\hat{y}_p$

(1) $E(y_p - \hat{y}_p) = 0$
which means $\hat{y}_p$ is an unbiased prediction.
Note: we should not write $E(\hat{y}_p) = y_p$ because y_p is not a constant (it is random).

(2) $\mathrm{Var}(y_p - \hat{y}_p) = [1 + \frac{1}{n} + \frac{(x_p-\bar{x})^2}{S_{xx}}]\sigma^2$
because

$$y_p - \hat{y}_p = \underbrace{(\beta_0 + \beta_1 x_p + \epsilon_p)}_{\mu_p} - \underbrace{(\hat{\beta}_0 + \hat{\beta}_1 x_p)}_{\hat{\mu}_p} = (\mu_p - \hat{\mu}_p) + \epsilon_p$$

$$\mathrm{Var}(y_p - \hat{y}_p) = \mathrm{Var}(\hat{\mu}_p) + \mathrm{Var}(\epsilon_p) = [\frac{1}{n} + \frac{(x_p-\bar{x})^2}{S_{xx}}]\sigma^2 + \sigma^2$$

(3) $\frac{y_p - \hat{y}_p}{SE(y_p-\hat{y}_p)} \sim t_{n-2}$
where $SE(y_p - \hat{y}_p) = \sqrt{[1 + \frac{1}{n} + \frac{(x_p-\bar{x})^2}{S_{xx}}]s^2}$, where $s = \frac{\sum(y_i-\bar{y})^2}{n-2}$
Thus $100(1-\alpha)\%$ prediction interval for y_p is

$$\hat{y}_p \pm t_{n-1,\frac{\alpha}{2}} SE(y_p - \hat{y}_p)$$

The prediction interval is trying to capture the variation of a future observation, and the confidence interval for the fitted value is trying to capture the variation in the mean value. Hence the prediction interval will be larger since there is an additional σ^2

2.5 Analysis of Variance (ANOVA) for Testing $H_0 : \beta_1 = 0$

The total variation among y_i's is measured by the total sum of squares (SST) defined as

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

Note that $y_i - \bar{y} = y_i - \hat{y}_i + \hat{y}_i - \bar{y}$ Now

$$\begin{aligned} SST &= \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \\ &= \sum_{i=1}^{n} r_i^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\ &= \text{sum of Squares of errors (SSE)} + \text{sum of squares of regression (SSR)} \end{aligned}$$

SSE: The residual the cannot be explained by your model.
SSR: how much your model can explain

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

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum_{i=1}^{n} r_i(\hat{y}_i - \bar{y}) \\ &= \sum_{i=1}^{n} r_i\hat{y}_i - \sum_{i=1}^{n} r_i\bar{y} \\ &\quad \text{recall} \sum r_i = 0, \ \sum r_i x_i = 0, \ \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \\ &\quad \text{hence } \sum r_i\hat{y}_i = 0 \\ &= 0 \end{aligned}$$

If $H_0 : \beta_1 = 0$ is true, $y_i = \beta_0 + \epsilon_i$.
SSR should be relatively "small" comparing to SSE.
Under $H_0 : \beta_1 = 0$, we have the following results:

(1) $\frac{\text{SSR}}{\sigma^2} \sim \chi^2_1$
(2) $\frac{\text{SST}}{\sigma^2} \sim \chi^2_{n-1}$
(3) $\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-2}$ and it is independent of SSR

Proof. (1)

$$\begin{aligned} \text{SSR} &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\ &= \sum(\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2 \\ &= \sum(\bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1 x_i - \bar{y})^2 \\ &= \hat{\beta}_1^2 \sum(x_i - \bar{x})^2 \\ &= \hat{\beta}_1^2 S_{xx} \end{aligned}$$

Recall that $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$, hence $(\frac{\hat{\beta}_1-\beta_1}{\frac{\sigma}{\sqrt{S_{xx}}}})^2 \sim \chi_1^2$.
Under $H_0 : \beta_1 = 0$,

$$\frac{\hat{\beta}_1^2 S_{xx}}{\sigma^2} \sim \chi_1^2$$

Thus, $\frac{\text{SSR}}{\sigma^2} \sim \chi_1^2$

(2) $y_1, \ldots, y_n \overset{i.i.d}{\sim} N(\beta_0, \sigma^2)$, i.e. $\frac{y_i-\beta_0}{\sigma} \sim N(0,1)$.
Thus, $\sum \frac{(y_i-\beta_0)^2}{\sigma^2} \sim \chi_n^2$. Therefore, $(\frac{\bar{y}-\beta_0}{\frac{\sigma}{\sqrt{n}}})^2 \sim \chi_1^2$ because $\bar{y} \sim N(\beta_0, \frac{\sigma^2}{n})$.

$$\begin{aligned}
\text{SST} &= \sum_{i=1}^{n}(y_i-\bar{y})^2 \\
&= \sum[(y_i-\beta_0)-(\bar{y}-\beta_0)]^2 \\
&= \sum(y_i-\beta_0)^2 - n(\bar{y}-\beta_0)^2 \\
\iff \frac{\sum(y_i-\beta_0)^2}{\sigma^2} &= \frac{\text{SST}}{\sigma^2} + (\frac{\bar{y}-\beta_0}{\frac{\sigma}{\sqrt{n}}})^2
\end{aligned}$$

For $Q_1 = \frac{\text{SST}}{\sigma^2}$ where $\text{SST} = \sum_{i=1}^{n}(y_i-\bar{y})^2$, there are only $n-1$ degree of freedom, because $\sum(y_i-\bar{y}) = 0$, i.e. $d_1 = n-1$
For $Q_2 = (\frac{\bar{y}-\beta_0}{\frac{\sigma}{\sqrt{n}}})^2$, we only have $\bar{y}$ and $d_2 = 1$.

Theorem 2.3 — (Simplified) Cochran's Theorem.
Suppose $U_i \overset{i.i.d}{\sim} N(0,1)$, $i = 1, \cdots, n$ and $\sum_{i=1}^{n} U_i^2 = Q_1 + Q_2$.
Let d_1 and d_2 be the degree of freedom of Q_1 and Q_2, which are the number of linearly independent linear combinations of y_i's in Q_1 and Q_2.
If $d_1 + d_2 = n$, Q_1 and Q_2 are independent, and furthermore, $Q_1 \sim \chi_{d_1}^2$ and $Q_2 \sim \chi_{d_2}^2$.

By Cochran's Theorem (2.3), Q_1 and Q_2 are independent and $\frac{\text{SST}}{\sigma^2} \sim \chi_{n-1}^2$.

(3) $\frac{\text{SSE}}{\sigma^2} \sim \chi_{n-2}^2$ and it is independent of SSR
We have $\frac{\text{SST}}{\sigma^2} = \frac{\text{SSE}}{\sigma^2} + \frac{\text{SSR}}{\sigma^2}$. Previously we have $\frac{\text{SST}}{\sigma^2} \sim \chi_{n-1}^2$ and $\frac{\text{SSR}}{\sigma^2} \sim \chi_1^2$.
The degree of freedom of $\text{SSE} = \sum(y_i-\hat{y}_i)^2$ is $n-2$.
(r_i subjects to 2 linear constraints: $\sum r_i = 0$ and $\sum r_i x_i = 0$)
By Cochran's Theorem (2.3), $\frac{\text{SSE}}{\sigma^2} \sim \chi_{n-2}^2$ and is independent of SSR.

■

F-satistic

$$F = \frac{\frac{\text{SSR}}{\sigma^2}/1}{\frac{\text{SSR}}{\sigma^2}/(n-2)} = \frac{\text{SSR}}{\text{SSE}/(n-2)} = \frac{\text{Mean Squares of Regression (MSR)}}{\text{Mean Squares of Error (MSE)}} \sim F(1, n-2)$$

Test $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$.
We reject H_0 at the α level if

$$F > F_\alpha(1, n-2) \text{ (Upper } \alpha \text{ quantile)}$$

Recall that

$$\frac{\hat{\beta}_1 - \beta_1^*}{\text{se}(\hat{\beta}_1)} \sim t_{n-2}$$

can test $H_0 : \beta_1 = \beta_1^*$
In SLR, testing $H_0 : \beta_1 = 0$ using t-test and F-test are equivalent. If $X \sim t_{n-2},\ X^2 \sim F(1, n-2)$

ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F
Regression	SSR	1	MSR= $\frac{\text{SSR}}{1}$	
Residual	SSE	$n-2$	MSE= $\frac{\text{SSE}}{n-2}$	$\frac{\text{MSR}}{\text{MSE}}$
Total	SST	$n-1$		

Coefficient of Determination: $R^2 = \frac{\text{SSR}}{\text{SST}}$.
In SLR,

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\hat{\beta}_1 S_{xx}}{S_{yy}} = \frac{(\frac{S_{xy}}{S_{xx}})^2 S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx}S_{yy}}$$

Recall that the sample correlation coefficient $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$ and thus $R^2 = r^2$

3. Multiple Linear Regression (MLR)

3.1 Review of Random Vector and Matrix Algebra

3.1.1 Noations

- Capital letters are matrices or vectors: A, X, Σ
- Lower letters are scalars: a, x, σ
- Underlined lower cases are vectors $\underset{\sim}{a}, \underset{\sim}{x}$
- All vectors are column vectors by default
- The transpose of any matrix A is A^T (occasionally A')

Random Vector

Let $Y = (y_1, y_2, \cdots, y_n)^T$ be a vector of random variables with

- $E(y_i) = \mu_i$
- $\text{Var}(y_i) = \sigma_i^2$
- $\text{Cov}(y_i, y_j) = \sigma_{ij}$

$$E(Y) = (\mu_1, \cdots, \mu_n)$$

$$\text{Var}(Y) = E([Y - E(Y)][Y - E(Y)]^T) = (\sigma_{ij}) \text{ (covariance matrix)}$$

Basic Properties

Suppose $A = (a_{ij})_{m\times n}$, $\underset{\sim}{b} = (b_1, \ldots, b_m)^T$, and $\underset{\sim}{c} = (c_1, \ldots, c_n)^T$ are matrix and vectors of constants

- $E(AY + \underset{\sim}{b}) = AE(Y) + \underset{\sim}{b}$
- $\text{Var}\left(Y + \underset{\sim}{c}\right) = \text{Var}(Y)$
- $\text{Var}\left(AY + \underset{\sim}{b}\right) = A\text{Var}(Y)A^T$

Quadratic Form

Suppose $A = (a_{ij})_{n\times n}$ is symmetric, i.e. $a_{ij} = a_{ji}$, $\forall i, j$. Then

$$f = Y^T AY = \sum_i \sum_j a_{ij} y_i y_j$$

is called a **quadratic form**.

Differentiation over Linear and Quadratic Forms

- $f = f(Y) = f(y_1, \ldots, y_n)$; $\frac{df}{dY} = (\frac{\partial f}{\partial y_1}, \ldots, \frac{\partial f}{\partial y_n})$
- $f = \underset{\sim}{c}^T Y = \sum_{i=1}^n c_i y_i$; $\frac{df}{dY} = \underset{\sim}{c}^T$
- $f = Y^T AT$, where $A = (a_{ij})_{n \times n}$ is symmetric, $\frac{df}{dY} = 2Y^T A^T$

Trace

For a matrix, the **trace** is defined as

$$tr(A_{m\times m}) = \sum_{i=1}^{m} a_{ii}$$

Note that $tr(BC) = tr(CB)$

Rank

The **rank** of a matrix denoted $rank(A)$ is the maximum number of *linearly independent* columns (or rows) of A.
Note that vectors $Y_1, \ldots, Y_n$ are linearly independent if and only if

$$c_1 Y_1 + \cdots + c_n Y_n = 0$$

implies $c_1 = \cdots c_n = 0$

Eigenvector and Eigenvalue

A non-zero vector $\underset{\sim}{v_i}$ is an **eigenvector** of $A_{m\times m}$ if

$$A\underset{\sim}{v_i} = \lambda_i \underset{\sim}{v_i};\ i = 1, 2, \ldots, m$$

where λ_i is the corresponding i-th **eigenvalue**

Orthogonality

- Two vectors X and Y are orthogonal if $Y^T X = 0$
- A square matrix A is an orthogonal matrix if

$$A^T A = AA^T = I$$

Idempotent

A matrix A is **idempotent** if $AA = A$
Results:

(1) If A is idempotent, then all its eigenvalues are either 0 or 1.
(2) If A is idempotent and symmetric, there exists an orthogonal matrix P such that $A = P\Lambda P^T$ where

$$\Lambda = \begin{bmatrix} 1 & \cdots & 0 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 & 0 & 0 \\ 0 & \cdots & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

and $tr(A) = rank(A) = tr(\Lambda)$ which is equivalent to the number of eigenvalues being 1.

3.2 Multiple Linear Regression (MLR)

3.2.1 Model

'For $i = 1, \cdots, n$

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

where

- $x_{i1}, \ldots, x_{ip}$ are p fixed known predictor variable
- $\beta_0, \ldots, \beta_p$ are fixed but unknown regression parameters
- ϵ_i is random and unknown error
- y_i is random but observable response

3.2.2 Assumptions

(1) $E(\epsilon_i) = 0$
(2) $\mathrm{Var}(\epsilon_i) = \sigma^2$
(3) $\epsilon_1, \ldots, \epsilon_n$ are independent
(4) $\epsilon_1, \ldots, \epsilon_n \overset{i.i.d}{\sim} N(0, \sigma^2) \implies y_i \sim N(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, 0)$

3.2.3 Regression Coefficients $\beta_1, \ldots, \beta_p$

β_j: the average amount of increase (or decrease) in response when the j-th predictor x_j increase (or decrease) by one unit while holding all other predictors fixed/constant
$H_0 : \beta_j = 0$ means x_j is not (linearly) related to y, given all other predictors in the model.

Matrix Form Representation

$$\underbrace{\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}}_{Y} = \underbrace{\begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_0 \end{bmatrix}}_{\underset{\sim}{\beta}} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \vdots \\ \beta_0 \end{bmatrix}$$

In short,

$$Y = x\underset{\sim}{\beta} + \underset{\sim}{\epsilon}$$

and

$$\underset{\sim}{\epsilon} \sim MVN(\underset{\sim}{0},\ \sigma^2 I)$$

$$Y \sim MVN(X\underset{\sim}{\beta},\ \sigma^2 I)$$

Note 3.1 $Y \sim MVN(\underset{\sim}{\mu},\ \Sigma) \implies$ The probability density function of Y is

$$f(Y) = \left[\frac{1}{2\pi}\right]^{\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(Y - \underset{\sim}{\mu})^T \Sigma^{-1} (Y - \underset{\sim}{\mu})\}$$

where $|\Sigma|$ is the determinant of Σ

Property 3.1 If $Y \sim MVN(\underset{\sim}{\mu}, \Sigma)$ and $Z = AY$, then $Z \sim MVN(A\underset{\sim}{\mu},\ A\Sigma A^T)$

3.3 Least Squares Estimate (LSE) of $\underset{\sim}{\beta} = (\beta_0, \dots, \beta_p)^T$

Minimize

$$\begin{aligned} S(\underset{\sim}{\beta}) &= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_p x_{ip})^2 \\ &= (Y - X\underset{\sim}{\beta})^T (Y - X\underset{\sim}{\beta}) \\ &= Y^T Y - Y^T X \underset{\sim}{\beta} - \underset{\sim}{\beta}^T X^T Y + \underset{\sim}{\beta}^T X^T X \underset{\sim}{\beta} \end{aligned}$$

Note that $Y^T X \underset{\sim}{\beta} = \underset{\sim}{\beta}^T X^T Y$ is a scalar,

$$S(\underset{\sim}{\beta}) = Y^T Y - 2Y^T X \underset{\sim}{\beta} + \underset{\sim}{\beta}^T X^T X \underset{\sim}{\beta}$$

Recall $\frac{d \underset{\sim}{c}^T Y}{dY} = \underset{\sim}{c}^T$ and $\frac{dY^T AY}{dY} = 2Y^T A^T$

$$\begin{aligned} \frac{dS(\underset{\sim}{\beta})}{d\underset{\sim}{\beta}} &= -2Y^T X + 2\underset{\sim}{\beta}^T X^T X = 0 \\ \implies & \hat{\underset{\sim}{\beta}}^T X^T X = Y^T \\ & X^T X \hat{\underset{\sim}{\beta}} = X^T Y \\ & \hat{\underset{\sim}{\beta}} = (X^T X)^{-1} x^t y \end{aligned}$$

Note $X^T X$ needs to be full rank (rank p+1) to be invertible.

3.3.1 Properties of LSE $\hat{\underset{\sim}{\beta}}$

(1) $\hat{\underset{\sim}{\beta}}$ is unbiased

(2) $\text{Var}\left(\hat{\underset{\sim}{\beta}}\right) = \sigma^2 (X^T X)^{-1}$

(3) Fitted values $\hat{Y} = X\hat{\underset{\sim}{\beta}} = X(X^T X)^{-1} X^T Y = HY$

- H is idempotent and symmetric
- H is typically called the "hat" matrix because it produces $\hat{Y}$

Check:

$$\begin{aligned} HH &= X(X^T X)^{-1} X (X^T X)^{-1} X^T = X(X^T X)^{-1} X^T = H \\ H^T &= [X(X^T X)^{-1} X^T]^T = X(X^T X)^{-1} X^T = H \end{aligned}$$

(4) Residuals $\underset{\sim}{r} = Y - \hat{Y} = Y - HY = (I - H)Y$

- The matrix $I - H$ is idepotgent and symmetric
- $\sum_{i=1}^{n} r_i = 0$

  $X^T \underset{\sim}{r} = \underset{\sim}{0}$

  $Y^T \underset{\sim}{r} = 0$

$X = [\underset{\sim}{x}_0, \underset{\sim}{x}_1, \dots, \underset{\sim}{x}_p]$

(5) $E(\underset{\sim}{r}) = 0$

(6) $\text{Var}\left(\underset{\sim}{r}\right) = \sigma^2(I - H)$

Proof.

(1)

$$\begin{aligned} E(\underset{\sim}{\hat{\beta}}) &= E[(X^TX)^{-1X^TY}]) \\ &= (X^TX)^{-1}X^TE(Y) \\ &= (X^TX)^{-1}X^TX\underset{\sim}{\beta} \\ &= \underset{\sim}{\beta} \end{aligned}$$

(2)

$$\begin{aligned} \text{Var}\left(\underset{\sim}{\hat{\beta}}\right) &= \text{Var}\big([(X^TX)^{-1}X^TY]\big) \\ &= (X^TX)^{-1}X^T\text{Var}(Y)X[(X^TX)^{-1}]^T \end{aligned}$$

Note that $[(X^TX)^{-1}]^T = [(X^TX)^T]^{-1} = (X^TX)^{-1}$

$$\begin{aligned} &= (X^TX)^{-1}X^T\sigma^2IX(X^TX)^{-1} \\ &= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} \\ &= \sigma^2X^T \end{aligned}$$

(4)

$$\begin{aligned} X^T\underset{\sim}{r} &= X^T(Y - \hat{Y}) \\ &= X^TY - X^THY \\ &= X^TY - X^TX(X^TX)^{-1}X^TY \\ &= X^TY - X^TY \\ &= \underset{\sim}{0} \\ \hat{Y}\underset{\sim}{r} &= (X\underset{\sim}{\hat{\beta}})^T\underset{\sim}{r} \\ &= \underset{\sim}{\hat{\beta}}^TX^T\underset{\sim}{r} \\ &= 0 \end{aligned}$$

■