Linear Regression (I)

Zhijian He

2021-06-08

Background

High on the list of problems that experimenters most frequently need to deal with is **the determination of the relationships that exist among the various components of a complex system**. If those relationships are sufficiently understood, there is a good possibility that the system's output can be effectively modeled, maybe even controlled.

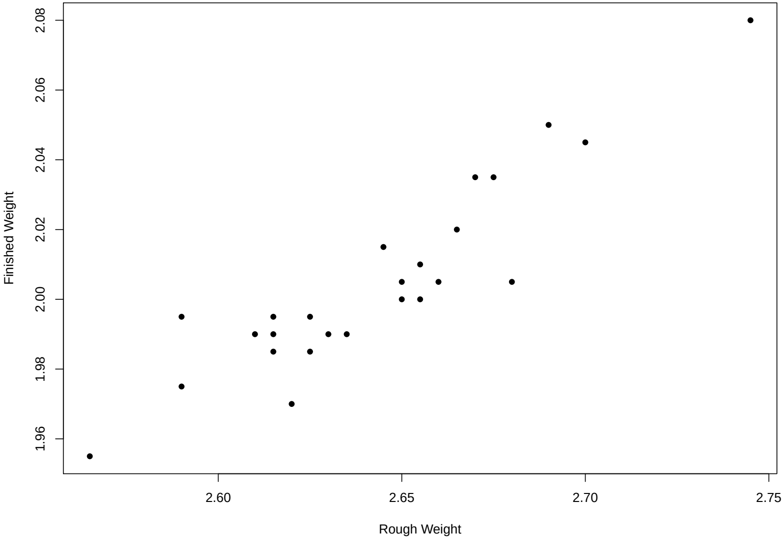
Input	X
Output	Y

- What is the relationship between X and Y? How to model it?
- If you have data from the system $(x_i, y_i), i = 1, 2, \ldots, n$, how to fit your model?
- If the relationship is well-understood, what would you do?

Case study 1

A manufacturer of air conditioning units is having assembly problems due to the failure of a connecting rod to meet finished-weight specifications. Too many rods are being completely tooled, then rejected as overweight. To reduce that cost, the company's quality-control department wants to quantify the relationship between the weight of the **finished rod**, y, and that of the **rough casting**, x. Castings likely to produce rods that are too heavy can then be discarded before undergoing the final (and costly) tooling process.

Graphed data



Simple Linear Models

The system:

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

The data: (x_i, y_i), $i = 1, \ldots, n$

The linear model is given by

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

- ϵ_i are random (need some assumptions) and unobservable
- x_i are **fixed** (*independent/predictor* variable)
- y_i are random (*dependent/response* variable)
- β_0 is the *intercept*
- β_1 is the *slope*

To-do-lists

1. Point estimation: β_0 and β_1
2. Confidence interval: β_0 and β_1
3. Hypothesis testing of $H_0 : \beta_i = 0 \ vs. \ H_1 : \beta_i \neq 0$
4. Prediction: Given x, how to predict y?
5. Control y: Under the constraint on y, what should x be?

Choose β_0, β_1 to minimize

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

The minimizers $\hat{\beta}_0, \hat{\beta}_1$ are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y}) x_i}{\sum_{i=1}^{n} (x_i - \bar{x}) x_i}, \ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

`Regression function`: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

Some useful notations

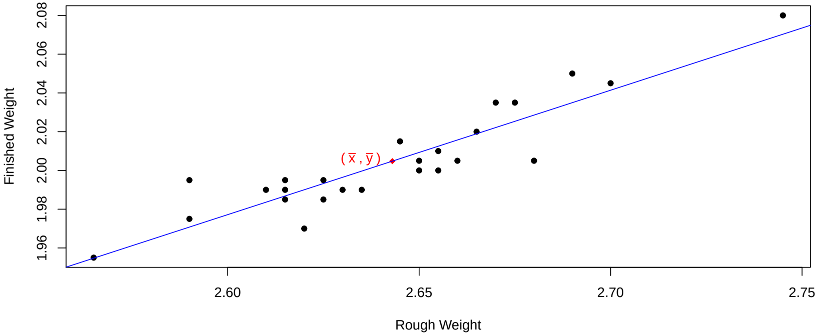
$$\ell_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

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

$$\ell_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})} = \frac{\ell_{xy}}{\ell_{xx}} = \frac{1}{\ell_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})y_i$$

The results for Case study 1



- The least squares estimates are

$$\hat{\beta}_1 = \frac{\ell_{xy}}{\ell_{xx}} = \frac{0.023565}{0.0367} = 0.642, \ \ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 0.308.$$

- The regression function is $\hat{y} = 0.308 + 0.642x$.

`Assumption A1`: $E[\epsilon_i] = 0, i = 1, \ldots, n.$

`Theorem 1`: Under Assumption A1, $\hat{\beta}_0, \hat{\beta}_1$ are unbiased estimators for β_0, β_1, respectively.

`Assumption A2`: $Cov(\epsilon_i, \epsilon_j) = \sigma^2 1\{i = j\}$.

`Theorem 2`: Under Assumption A2, we have

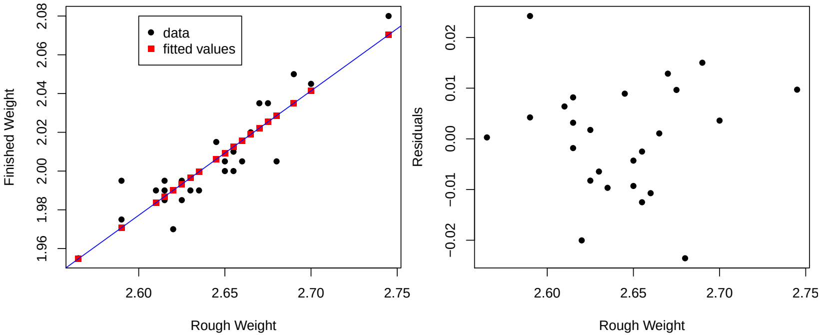
$$Var[\hat{\beta}_0] = \left(\frac{1}{n} + \frac{\bar{x}^2}{\ell_{xx}}\right)\sigma^2,\ Var[\hat{\beta}_1] = \frac{\sigma^2}{\ell_{xx}}$$

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

Note that: $Cov(\bar{y}, \hat{\beta}_1) = 0$.

Fitted values and residuals

- fitted values: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- residuals: $\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$
- the sum of squared errors (SSE): $S_e^2 = \sum_{i=1}^n \hat{\epsilon}_i^2$




```
## SSE =  0.002942958
```

Estimation of σ^2

- Residual standard error σ
- Residual variance σ^2

`Theorem 3`: Let

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

Under Assumptions A1 and A2, we have $E[\hat{\sigma}^2] = \sigma^2$. I.e., $\hat{\sigma}^2$ is an unbiased estimate of σ^2.

Normal distributions assumption

`Assumption B`: $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2), i = 1, \ldots, n.$

`Theorem 4`: Under Assumption B, we have

(1). $\hat{\beta}_0 \sim N(\beta_0, (\frac{1}{n} + \frac{\bar{x}^2}{\ell_{xx}})\sigma^2)$

(2). $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\ell_{xx}})$

(3). $\frac{(n-2)\hat{\sigma}^2}{\sigma^2} = \frac{S_e^2}{\sigma^2} \sim \chi^2(n-2)$

(4). $\hat{\sigma}^2$ is independent of $(\hat{\beta}_0, \hat{\beta}_1)$.

Standard error of the estimators:

- $se(\hat{\beta}_0) := \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\ell_{xx}}}\hat{\sigma}$
- $se(\hat{\beta}_1) := \sqrt{1/\ell_{xx}}\hat{\sigma}$

Inferences about β_1

- In the more realistic setting of unknown σ, using Theorem 4 gives
$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{\ell_{xx}}} = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \sim t(n-2).$$
The $1-\alpha$ confidence interval for β_1 is
$$\hat{\beta}_1 \pm t_{1-\alpha/2}(n-2)se(\hat{\beta}_1).$$
- For testing
$$H_0 : \beta_1 = \beta_1^* \ vs. \ H_1 : \beta_1 \neq \beta_1^*,$$
we chosse the test statistic as
$$T_1 = \frac{\hat{\beta}_1 - \beta_1^*}{se(\hat{\beta}_1)}.$$
We reject H_0 if $|T_1| > t_{1-\alpha/2}(n-2)$.

Inferences about β_0

Similarly, for drawing inferences about β_0, we can use

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}\sqrt{1/n + \bar{x}^2/\ell_{xx}}} = \frac{\hat{\beta}_0 - \beta_0}{se(\hat{\beta}_0)} \sim t(n-2).$$

- The $1-\alpha$ confidence interval for β_0 is

$$\hat{\beta}_0 \pm t_{1-\alpha/2}(n-2)se(\hat{\beta}_0).$$

- For testing

$$H_0 : \beta_0 = \beta_0^* \ vs. \ H_1 : \beta_0 \neq \beta_0^*,$$

  we chosse the test statistic as

$$T_2 = \frac{\hat{\beta}_0 - \beta_0^*}{se(\hat{\beta}_0)}.$$

  We reject H_0 if $|T_2| > t_{1-\alpha/2}(n-2)$.

Inferences about σ^2

Note that

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} = \frac{S_e^2}{\sigma^2} \sim \chi^2(n-2).$$

The $1-\alpha$ confidence interval for σ^2 is

$$\left[\frac{(n-2)\hat{\sigma}^2}{\chi^2_{1-\alpha/2}(n-2)}, \frac{(n-2)\hat{\sigma}^2}{\chi^2_{\alpha/2}(n-2)}\right]$$

or,

$$\left[\frac{S_e^2}{\chi^2_{1-\alpha/2}(n-2)}, \frac{S_e^2}{\chi^2_{\alpha/2}(n-2)}\right].$$

R code for handling Case study 1

```
rough_weight = c(2.745, 2.700, 2.690, 2.680, 2.675,
2.670, 2.665, 2.660, 2.655, 2.655, 2.650, 2.650,
2.645, 2.635, 2.630, 2.625, 2.625, 2.620, 2.615,
2.615, 2.615, 2.610, 2.590, 2.590, 2.565)
finished_weight = c(2.080, 2.045, 2.050, 2.005, 2.035,
2.035, 2.020, 2.005, 2.010, 2.000, 2.000, 2.005, 2.015,
1.990, 1.990, 1.995, 1.985, 1.970, 1.985, 1.990, 1.995,
1.990, 1.975, 1.995, 1.955)
lm.rod = lm(finished_weight~rough_weight)
summary(lm.rod) #output the results
```

Summary report

```
##
## Call:
## lm(formula = finished_weight ~ rough_weight)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.023558 -0.008242  0.001074  0.008179  0.024231
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.30773    0.15608   1.972   0.0608 .
## rough_weight  0.64210    0.05905  10.874 1.54e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01131 on 23 degrees of freedom
## Multiple R-squared:  0.8372, Adjusted R-squared:  0.8301
## F-statistic: 118.3 on 1 and 23 DF,  p-value: 1.536e-10
```

LSE vs. MLE

Consider the model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ \ \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), i = 1, \dots, n.$$

- look at the relationship between the LSE and MLE for β_0 and β_1
- work out the MLE for σ^2 and compare it with $\hat{\sigma}^2$

The answers:

$$\hat{\beta}_j^{LSE} = \hat{\beta}_j^{MLE},\ \ j = 0, 1$$

$$\hat{\sigma}^2_{MLE} = \frac{S_e^2}{n} = \frac{(n-2)\hat{\sigma}^2}{n}$$

Question: for the two estimates of σ^2, which one is better?

Prediction

The model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ \ \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), i = 1, \ldots, n.$$

For given **a new** x_{n+1}, we would like to draw inferences about the future observation y_{n+1}, where

$$y_{n+1} = \beta_0 + \beta_1 x_{n+1} + \epsilon_{n+1}.$$

- confidence interval (CI) for the future expected value $E[y_{n+1}] = \beta_0 + \beta_1 x_{n+1}$
- prediction interval (PI) for the future observation y_{n+1}

Drawing Inferences about the future expected value

A natural unbiased estimate is $\hat{y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}$.

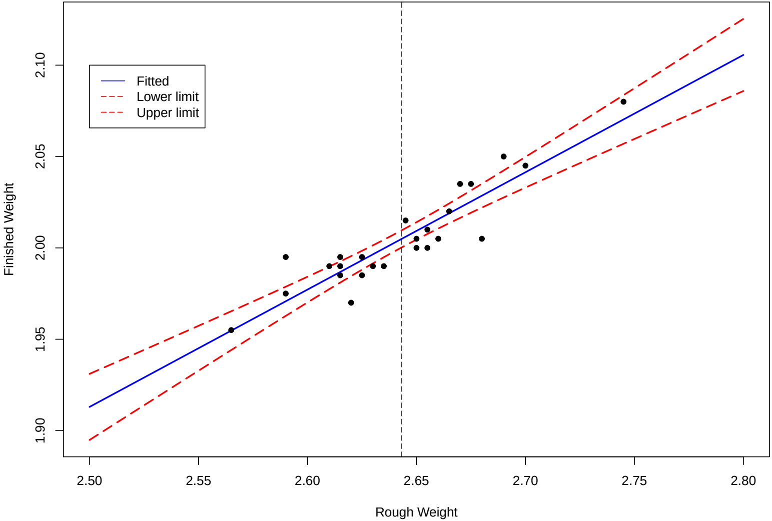
`Theorem 5`: Suppose Assumption B is satisfied. Then we have

$$\frac{\hat{y}_{n+1} - E[y_{n+1}]}{\sigma\sqrt{\frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\ell_{xx}}}} \sim N(0,1).$$

For unknown σ, we replace σ with $\hat{\sigma}$ to arrive at $t(n-2)$ distribution. The $1-\alpha$ CI for $E[y_{n+1}] = \beta_0 + \beta_1 x_{n+1}$ is given by

$$\hat{y}_{n+1} \pm t_{1-\alpha/2}(n-2)\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\ell_{xx}}}.$$

Case study 1: Confidence interval



Drawing Inferences about the future value

`Definition`: A **prediction interval (PI)** is a range of numbers that contains y_{n+1} with a specified probability.

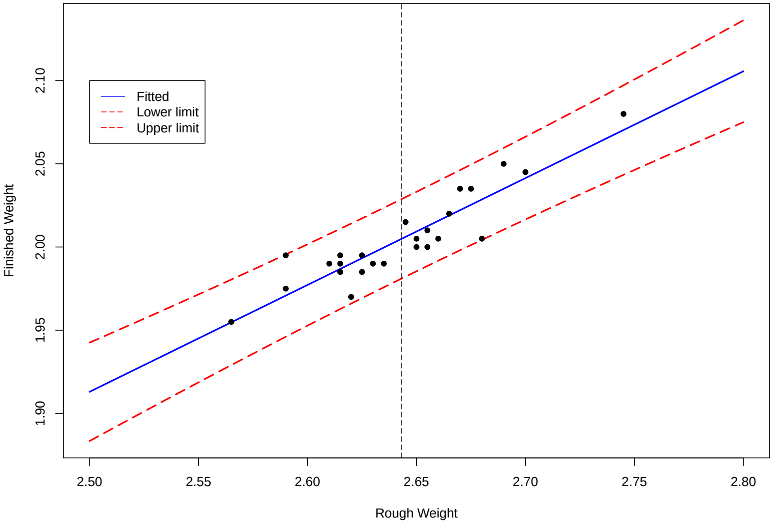
`Theorem 6`: Suppose Assumption B is satisfied. Let $y_{n+1} = \beta_0 + \beta_1 x_{n+1} + \epsilon_{n+1}$, where $\epsilon_{n+1} \sim N(0, \sigma^2)$ is independent of ϵ_i's. Then

$$\frac{\hat{y}_{n+1} - y_{n+1}}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\ell_{xx}}}} \sim N(0,1).$$

The $1-\alpha$ PI for y_{n+1} is given by

$$\hat{y}_{n+1} \pm t_{1-\alpha/2}(n-2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\ell_{xx}}}.$$

Case study 1: Prediction interval



How to control the future observation?

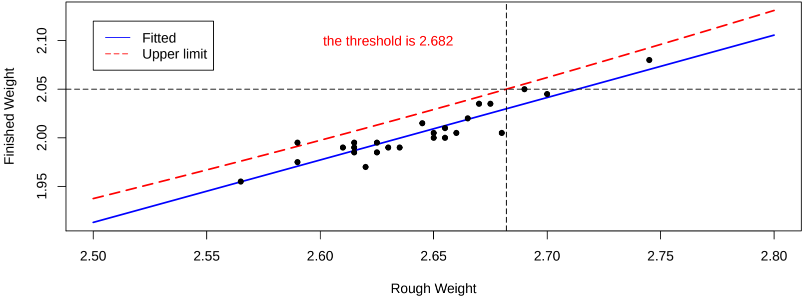
Consider case study 1 again. Castings likely to produce rods that are too heavy can then be discarded before undergoing the final (and costly) tooling process. The company's quality-control department wants to produce the rod y_{n+1} with weights no larger than 2.05 with probability no less than 0.95. How to choose the rough casting?

Now we want $y_{n+1} \leq y_0 = 2.05$ with probability no less than $1-\alpha$. I.e., $P(y_{n+1} \leq y_0) \geq 1-\alpha$. How to choose x_{n+1}?

Recall that

$$\frac{y_{n+1} - \hat{y}_{n+1}}{\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_{n+1}-\bar{x})^2}{\ell_{xx}}}} \sim t(n-2).$$

How to control the future observation?



$$\hat{\beta}_0 + \hat{\beta}_1 x_{n+1} + t_{1-\alpha}(n-2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\ell_{xx}}} \le y_0$$

Multiple linear regression

Consider a model of the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i,\ \ i = 1, \ldots, n.$$

In the matrix form:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

$$Y = X\beta + \epsilon$$

- the matrix X is called the **design matrix**

Least squares estimation (LSE)

Find β to minimize

$$Q(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_{p-1}x_{i,p-1})^2$$
$$= ||Y - X\beta||^2 = (Y - X\beta)^\top (Y - X\beta)$$
$$= Y^\top Y - 2Y^\top X\beta + \beta^\top X^\top X\beta.$$

If we differentiate Q with respect to each β_i and set the derivatives equal to zero, we see that the minimizers $\hat{\beta}_0, \ldots, \hat{\beta}_{p-1}$ satisfy

$$\frac{\partial Q}{\partial \beta_i} = -2(Y^\top X)_i + 2(X^\top X)_{i\cdot}\hat{\beta} = 0.$$

Normal equations

We thus arrive at the so-called **normal equations**:

$$X^\top X\hat{\beta} = X^\top Y$$

If the matrix $X^\top X$ is **nonsingular**, the formal solution is

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y.$$

`Lemma 1`: The matrix $X^\top X$ is nonsingular if and only if $\mathrm{rank}(X) = p$.

NOTE: In what follows, we assume that $\mathrm{rank}(X) = p < n$. If $p > n$, it belongs to the field of high-dimensional statistics.

Expected values and variances

`Assumption A`: Assume that $E[\epsilon] = 0$ and $Var[\epsilon] = \sigma^2 I_n$.

`Theorem 7`: Suppose that Assumption A is satisfied and $\mathrm{rank}(X) = p < n$, we have

(1). $E[\hat{\beta}] = \beta$,

(2). $Var[\hat{\beta}] = \sigma^2 (X^\top X)^{-1}$.

Estimation of σ^2

Definition:

- **The fitted values**: $\hat{Y} = X\hat{\beta}$
- **The vector of residuals**: $\hat{\epsilon} = Y - \hat{Y}$
- **The sum of squared errors (SSE)**: $S_e^2 = Q(\hat{\beta}) = ||Y - \hat{Y}||^2 = ||\hat{\epsilon}||^2$

Note that

$$\hat{Y} = X\hat{\beta} = X(X^\top X)^{-1}X^\top Y =: PY$$

- **The projection matrix**: $P = X(X^\top X)^{-1}X^\top$

The vector of residuals is then $\hat{\epsilon} = (I_n - P)Y$.

The projection matrix

Two useful properties of P are given in the following lemma.

The projection matrix:

$$P = X(X^\top X)^{-1}X^\top$$

`Lemma 2`: Let P be defined as before. Then

$$P = P^\top = P^2$$

$$I_n - P = (I_n - P)^\top = (I_n - P)^2.$$

The sum of squared residuals is then

$$S_e^2 := ||\hat{\epsilon}||^2 = Y^\top (I_n - P)^\top (I_n - P)Y = Y^\top (I_n - P)Y.$$

Estimation of σ^2

`Theorem 8`: Suppose that Assumption A is satisfied and $\text{rank}(X) = p < n$,

$$\hat{\sigma}^2 = \frac{S_e^2}{n-p}$$

is an unbiased estimate of σ^2.

Normal distribution

`Assumption B`: Assume that $\epsilon \sim N(0, \sigma^2 I_n)$.

`Theorem 9`: Suppose that Assumption B is satisfied and $\mathrm{rank}(X) = p < n$, we have

(1). $\hat{\beta} \sim N(\beta, \sigma^2(X^\top X)^{-1})$,

(2). $\frac{(n-p)\hat{\sigma}^2}{\sigma^2} = \frac{S_e^2}{\sigma^2} \sim \chi^2(n-p)$,

(3). $\hat{\epsilon}$ is independent of $\hat{Y}$,

(4). S_e^2 (or equivalently $\hat{\sigma}^2$) is independent of $\hat{\beta}$.

Confidence intervals for β_i

Let $C = (X^\top X)^{-1}$ with entries c_{ij}. By Theorem 9, we have

$$\frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{c_{ii}}} \sim N(0,1),$$

$$\frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}\sqrt{c_{ii}}} = \frac{\hat{\beta}_i - \beta_i}{se(\hat{\beta}_i)} \sim t(n-p).$$

- **standard error**: $se(\hat{\beta}_i) = \hat{\sigma}\sqrt{c_{ii}}$

If σ^2 is unknown, for each β_i, the $100(1-\alpha)\%$ CI is

$$\hat{\beta}_i \pm t_{1-\alpha/2}(n-p) se(\hat{\beta}_i).$$

Hypothesis tests on β_i

Consider the test

$$H_0 : \beta_i = \beta_i^* \ vs. \ H_1 : \beta_i \neq \beta_i^*.$$

The test statistic is

$$T = \frac{\hat{\beta}_i - \beta_i^*}{se(\hat{\beta}_i)}.$$

The rejection region is

$$W = \{|T| > t_{1-\alpha/2}(n-p)\}.$$

NOTE: We are particularly interested in the case of $\beta_i^* = 0$.

Consider the hypothesis test:

$$H_0 : \beta_1 = \cdots = \beta_{p-1} = 0 \ vs. \ H_1 : \beta_{i^*} \neq 0 \text{ for some } i^* \geq 1.$$

`Definition`:

- **The total sum of squares (SST)**: $S_T^2 = \sum_{i=1}^n (y_i - \bar{Y})^2$
- **The sum of squares due to regression (SSR)**: $S_R^2 = \sum_{i=1}^n (\hat{y}_i - \bar{Y})^2$
- **The sum of squared errors (SSE)**: $S_e^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$

The relationship is

$$S_T^2 = S_R^2 + S_e^2.$$

The GLR test

The likelihood function for Y is given by

$$L(\beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} e^{-\frac{||Y - X\beta||^2}{2\sigma^2}}.$$

The likelihood ratio is then given by

$$\lambda = \frac{\sup_{\theta \in \Theta} L(\beta, \sigma^2)}{\sup_{\theta \in \Theta_0} L(\beta, \sigma^2)} = \left(\frac{S_T^2}{S_e^2}\right)^{n/2} = \left(1 + \frac{S_R^2}{S_e^2}\right)^{n/2}.$$

F-test

`Theorem 10`: Suppose that Assumption B is satisfied and $\mathrm{rank}(X) = p < n$, we have

(1). $S_R^2, S_e^2, \bar{Y}$ are independent, and

(2). if the null $H_0 : \beta_1 = \cdots = \beta_{p-1} = 0$ is true,

$$S_R^2/\sigma^2 \sim \chi^2(p-1),$$

$$F = \frac{S_R^2/(p-1)}{S_e^2/(n-p)} \overset{H_0}{\sim} F(p-1, n-p).$$

We take F as the test statistic. The rejection region is

$$W = \{F > F_{1-\alpha}(p-1, n-p)\}.$$

More general tests

$$H_0 : \beta_i = 0 \text{ for all } i \in I \ vs. \ H_1 : \beta_{i^*} \neq 0 \text{ for some } i^* \in I.$$

$$H_0 : H\beta = 0 \ vs. \ H_1 : H\beta \neq 0.$$

See Pages 178-192 of our textbook for details.

Coefficient of determination (R-squared):

$$R^2 = \frac{S_R^2}{S_T^2} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{S_e^2}{S_T^2}.$$

- a crude measure of the strength of a relationship that has been fit by least squares.
- the proportion of the variability of the dependent variable that can be explained by the independent variables.

Adjusted R-squared:

$$\tilde{R}^2 = 1 - \frac{S_e^2/(n-p)}{S_T^2/(n-1)} = 1 - \frac{n-1}{n-p} \times \frac{S_e^2}{S_T^2} < R^2.$$

It is easy to see that

$$F = \frac{S_T^2 R^2/(p-1)}{S_T^2(1-R^2)/(n-p)} = \frac{R^2/(p-1)}{(1-R^2)/(n-p)}.$$

R-squared vs. correlation coefficient

For the simple linear model $p = 2$, we have

$$S_R^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{\ell_{xy}^2}{\ell_{xx}}.$$

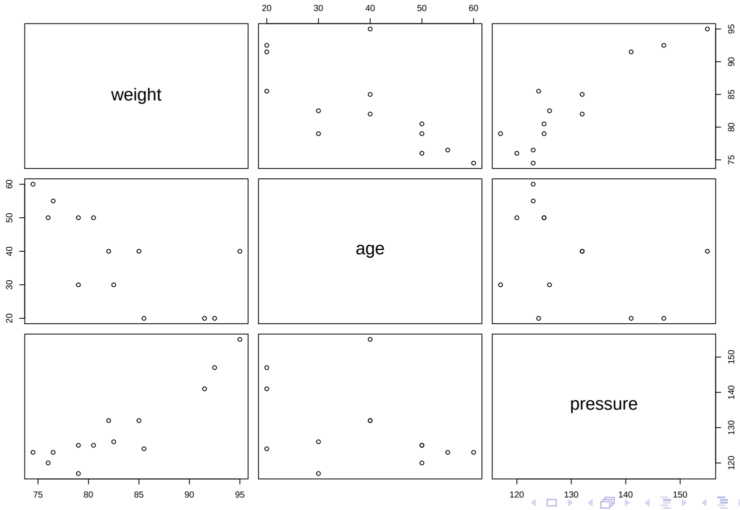
This gives

$$R^2 = \frac{\ell_{xy}^2}{\ell_{xx}\ell_{yy}} = \rho^2,$$

where the **correlation coefficient** between x_i and y_i is

$$\rho = \frac{\ell_{xy}}{\sqrt{\ell_{xx}\ell_{yy}}} = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2}}.$$

It is found that the systolic pressure is linked to the weight and the age. We now have the following data.



Summary report

```
##
## Call:
## lm(formula = pressure ~ weight + age, data = blood)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0404 -1.0183  0.4640  0.6908  4.3274
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -62.96336   16.99976  -3.704 0.004083 **
## weight        2.13656    0.17534  12.185 2.53e-07 ***
## age           0.40022    0.08321   4.810 0.000713 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.854 on 10 degrees of freedom
## Multiple R-squared:  0.9461, Adjusted R-squared:  0.9354
## F-statistic: 87.84 on 2 and 10 DF,  p-value: 4.531e-07
```

The regression function

$$\hat{y} = -62.96336 + 2.13656x_1 + 0.40022x_2$$

- $R^2 = 0.9461$
- the estimated covariance matrix $\hat{\sigma}^2(X^\top X)^{-1}$ is

	intercept	weight	age
intercept	288.991861	-2.9499280	-1.1174334
weight	-2.949928	0.0307450	0.0102176
age	-1.117433	0.0102176	0.0069243

Confidence interval for $E[y_{n+1}]$

Consider

$$y_{n+1} = \beta_0 + \beta_1 x_{n+1,1} + \cdots + \beta_{p-1} x_{n+1,p-1} + \epsilon_{n+1}.$$

Under Assumption B, $y_{n+1} = v^\top \beta + \epsilon_{n+1} \sim N(v^\top \beta, \sigma^2)$, where $v = (1, x_{n+1,1}, x_{n+1,2}, \dots, x_{n+1,p-1})^\top$. An unbiased estimate of the expected value of $E[y_{n+1}] = v^\top \beta$ is the fitted value

$$\hat{y}_{n+1} = v^\top \hat{\beta} \sim N(v^\top \beta, \sigma^2 v^\top (X^\top X)^{-1} v).$$

The $100(1-\alpha)\%$ CI for $E[y_{n+1}]$ is

$$\hat{y}_{n+1} \pm t_{1-\alpha/2}(n-p)\hat{\sigma}\sqrt{v^\top (X^\top X)^{-1} v}.$$

Prediction interval for y_{n+1}

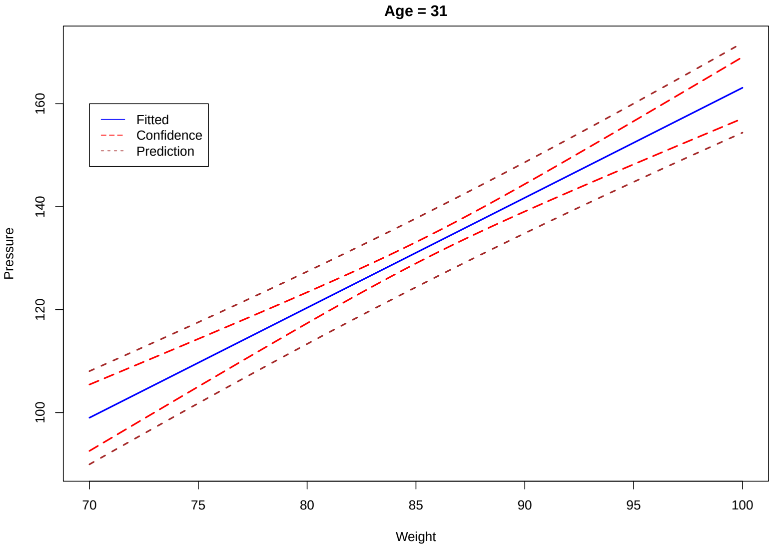
Similarly,

$$\frac{y_{n+1} - \hat{y}_{n+1}}{\hat{\sigma}\sqrt{1 + v^\top (X^\top X)^{-1} v}} \sim t(n-p).$$

The $100(1-\alpha)\%$ prediction interval for y is

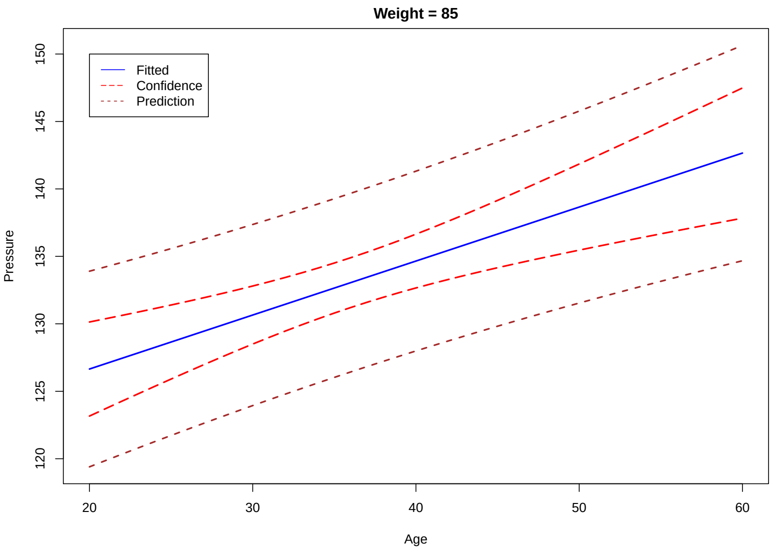
$$\hat{y}_{n+1} \pm t_{1-\alpha/2}(n-p)\hat{\sigma}\sqrt{1 + v^\top (X^\top X)^{-1} v}.$$

Case study 2



Age = 31
Fitted
Confidence
Prediction
160
140
120
100
Pressure
70
75
80
85
90
95
100
Weight

Case study 2


Weight = 85
Fitted
Confidence
Prediction
Pressure
120
125
130
135
140
145
150
20
30
40
50
60
Age

Inherently Linear models:

$$f(y) = \beta_0 + \beta_1 g_1(x_1, \dots, x_{p-1}) + \dots$$
$$+\beta_{k-1} g_{k-1}(x_1, \dots, x_{p-1}) + \epsilon$$

Let $y^* = f(y),\ x_i^* = g_i(x_1, \dots, x_{p-1})$. The transformed model is linear

$$y^* = \beta_0 + \beta_1 x_1^* + \cdots + \beta_{k-1} x_{k-1}^* + \epsilon.$$

Examples

- Polynomial models:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_{p-1} x^p + \epsilon$$

- Interaction models:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \beta_3 x_1 x_2 + \epsilon$$

- Multiplicative models:

$$y = \gamma_1 X_1^{\gamma_2} X_2^{\gamma_3} \epsilon^*$$

- Exponential models:

$$y = \exp\{\beta_0 + \beta_1 x_1 + \beta_2 x_2\} + \epsilon^*$$

- Reciprocal models:

$$y = \frac{1}{\beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_{p-1} x^p + \epsilon}$$

- Semilog models:

$$y = \beta_0 + \beta_1 \log(x) + \epsilon$$

- Logit models:

$$\log\left(\frac{y}{1-y}\right) = \beta_0 + \beta_1 x + \epsilon$$

- Probit models: $\Phi^{-1}(y) = \beta_0 + \beta_1 x + \epsilon$, where Φ is the CDF of $N(0,1)$.

- The response variables: $y_i \in \{0, 1\}$
- The predictor variables: $x_{i1}, \ldots, x_{i,p-1}$
- $y_i \sim B(1, p(x_{i1}, \ldots, x_{i,p-1}))$

How to model the probability of success

$$\mathbb{P}(y_i = 1 | x_{i1}, \ldots, x_{i,p-1}) = p(x_{i1}, \ldots, x_{i,p-1})?$$

The **logit transformation** is defined as

$$\psi(p) := \ln \frac{p}{1-p} : (0, 1) \mapsto \mathbb{R}.$$

The **inverse logit transformation**

$$\psi^{-1}(t) = \frac{e^t}{1 + e^t}.$$

Logit (逻辑斯谛) regression

$$\psi(p(x_{i1},\dots,x_{i,p-1})) = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{j,p-1}$$

OR

$$p(x_{i1},\dots,x_{i,p-1}) = \psi^{-1}\left(\beta_0 + \sum_{j=1}^{p-1} \beta_j x_{j,p-1}\right)$$

That is

$$\mathbb{P}(y_i = 1 | x_{i1},\dots,x_{i,p-1}) = \frac{\exp\left\{\beta_0 + \sum_{j=1}^{p-1} \beta_j x_{j,p-1}\right\}}{1+\exp\left\{\beta_0 + \sum_{j=1}^{p-1} \beta_j x_{j,p-1}\right\}}.$$

How to estimate β_i?

See Pages 215–219 of our textbook.

MLE

The likelihood function:

$$L(\beta) = f(y_1, \dots, y_n) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i},$$

where

$$p_i := \mathbb{P}(y_i = 1 | x_{i1}, \dots, x_{i,p-1}) = \frac{\exp\left\{\beta_0 + \sum_{j=1}^{p-1} \beta_j x_{j,p-1}\right\}}{1 + \exp\left\{\beta_0 + \sum_{j=1}^{p-1} \beta_j x_{j,p-1}\right\}}.$$

The log-likelihood function:

$$\begin{aligned}\ell(\beta) = \ln L(\beta) &= \sum_{i=1}^{n} y_i \ln p_i + \sum_{i=1}^{n} (1-y_i) \ln(1-p_i) \\ &= \sum_{i=1}^{n} y_i \left(\beta_0 + \sum_{j=1}^{p-1} \beta_j x_{j,p-1}\right) - \sum_{i=1}^{n} \ln \left(\beta_0 + \sum_{j=1}^{p-1} \beta_j x_{j,p-1}\right).\end{aligned}$$

Probit regressian

The **Probit transformation** is defined as

$$\psi(p) := \Phi^{-1}(p) : (0,1) \mapsto \mathbb{R},$$

where $\Phi(x)$ is the CDF of $N(0,1)$.

Application: Titanic data

Data: `https://www.kaggle.com/c/titanic/overview`

- **Survived**: Passenger survival indicator (1 if survived)
- **Pclass**: Passenger class
- **Sex**: Sex of the passenger
- **Age**: Age of the passenger
- **SibSp**: Number of siblings/spouses aboard
- **Parch**: Number of parents/children aboard

Survived	Pclass	Sex	Age	SibSp	Parch
0	3	male	22	1	0
1	1	female	38	1	0
1	3	female	26	0	0
1	1	female	35	1	0
0	3	male	35	0	0
0	3	male	NA	0	0

R code for Logit regession

```r
reg = glm(Survived~., family=binomial(link = "logit"),data=data)
summary(reg)
```

```
##
## Call:
## glm(formula = Survived ~ ., family = binomial(link = "logit"),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7563  -0.6498  -0.3840   0.6222   2.4561
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.619681   0.546711  10.279  < 2e-16 ***
## Pclass      -1.316003   0.140868  -9.342  < 2e-16 ***
## Sexmale     -2.637379   0.219402 -12.021  < 2e-16 ***
## Age         -0.044451   0.008159  -5.448  5.1e-08 ***
## SibSp       -0.364586   0.126493  -2.882  0.00395 **
## Parch       -0.037142   0.119602  -0.311  0.75614
## ---
```

R code for Logit regession (improved)

```
reg = glm(Survived~as.factor(Pclass)+Sex+Age+SibSp,
          family=binomial(link = "logit"),data=data)
summary(reg)
```

```
##
## Call:
## glm(formula = Survived ~ as.factor(Pclass) + Sex + Age + SibSp,
##     family = binomial(link = "logit"), data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7876  -0.6417  -0.3864   0.6261   2.4539
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         4.334201   0.450700   9.617  < 2e-16 ***
## as.factor(Pclass)2 -1.414360   0.284727  -4.967 6.78e-07 ***
## as.factor(Pclass)3 -2.652618   0.285832  -9.280  < 2e-16 ***
## Sexmale            -2.627679   0.214771 -12.235  < 2e-16 ***
## Age                -0.044760   0.008225  -5.442 5.27e-08 ***
## SibSp              -0.380190   0.121516  -3.129  0.00176 **
```

R code for Probit regession

```
reg = glm(Survived~as.factor(Pclass)+Sex+Age+SibSp,
          family=binomial(link = "probit"),data=data)
summary(reg)
```

```
##
## Call:
## glm(formula = Survived ~ as.factor(Pclass) + Sex + Age + SibSp,
##     family = binomial(link = "probit"), data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9426  -0.6536  -0.3772   0.6372   2.4731
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         2.496516   0.246885  10.112  < 2e-16 ***
## as.factor(Pclass)2 -0.815986   0.162947  -5.008 5.51e-07 ***
## as.factor(Pclass)3 -1.495787   0.158124  -9.460  < 2e-16 ***
## Sexmale            -1.547109   0.119421 -12.955  < 2e-16 ***
## Age                -0.025079   0.004638  -5.407 6.39e-08 ***
## SibSp              -0.225302   0.068818  -3.274  0.00106 **
```