Regression Analysis

Fall 2021

Department of Mathematics
Hong Kong Baptist University

Simple linear regression with normal errors

Model:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \ i = 1, \ldots, n,$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d with $\mathcal{N}(0, \sigma^2)$ distribution.

Least Squares Estimator (LSE):

$$\hat{\beta}_1 = \frac{C_{xY}}{C_{xx}}, \ \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x},$$

where

$$\begin{aligned} \bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i, \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \\ C_{xY} &= \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}), \ C_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2. \end{aligned}$$

Note

$$
\begin{aligned}
C_{xY} &= \sum_{i=1}^{n}(x_i Y_i - x_i \bar{Y} - \bar{x} Y_i + \bar{x}\bar{Y}) \\
&= \sum_{i=1}^{n}(x_i - \bar{x}) Y_i + \sum_{i=1}^{n}(-x_i + \bar{x})\bar{Y} \\
&= \sum_{i=1}^{n}(x_i - \bar{x}) Y_i - n\bar{x}\bar{Y} + n\bar{x}\bar{Y} \\
&= \sum_{i=1}^{n}(x_i - \bar{x}) Y_i \, .
\end{aligned}
$$

Similarly,

$$\begin{aligned} C_{xx} &= \sum_{i=1}^{n}(x_i^2 - 2\bar{x}x_i + \bar{x}^2) = \sum_{i=1}^{n} x_i x_i - n\bar{x}\bar{x} \\ &= \sum_{i=1}^{n} x_i x_i - \bar{x}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n}(x_i - \bar{x})x_i \,. \end{aligned}$$

It follows from the previous results that

$$\hat{\beta}_1 = \frac{C_{xY}}{C_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{C_{xx}} \equiv \sum_{i=1}^{n} k_i Y_i \,,$$

where $k_i = \frac{x_i - \bar{x}}{C_{xx}}, i = 1, \ldots, n.$

Therefore,

$$
\begin{aligned}
\mathbb{E}(\hat{\beta}_1) &= \mathbb{E}\Big(\sum_{i=1}^{n} k_i Y_i\Big) = \sum_{i=1}^{n} k_i \mathbb{E}(Y_i) = \sum_{i=1}^{n} k_i(\beta_0 + \beta_1 x_i) \\
&= \sum_{i=1}^{n} \frac{x_i - \bar{x}}{C_{xx}} \beta_0 + \sum_{i=1}^{n} \frac{(x_i - \bar{x})x_i}{C_{xx}} \beta_1 \\
&= \beta_1 \,, \\
\mathbf{var}(\hat{\beta}_1) &= \mathbf{var}\Big(\sum_{i=1}^{n} k_i Y_i\Big) = \sum_{i=1}^{n} k_i^2 \mathbf{var}(Y_i) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{C_{xx}^2} \sigma^2 \\
&= \frac{\sigma^2}{C_{xx}} \,,
\end{aligned}
$$

and

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

For the estimator of σ^2,

$$\hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2},$$

we can show that

$$\frac{n-2}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{n-2}\,,$$

and $\hat{\sigma}^2$ is independent of $\hat{\beta}_1$. Therefore

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2/C_{xx}}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/C_{xx}}} \Bigg/ \sqrt{\frac{n-2}{\sigma^2}\hat{\sigma}^2/(n-2)} \;\sim\; t_{n-2}\,.$$

Hypothesis Testing and Confidence Intervals

$H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$
Test statistic: $T = \dfrac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / C_{xx}}}$.
Reject H_0 if $|t| > t_{n-2,1-\alpha/2}$, where t is observed value of T.

$$
\begin{aligned}
1-\alpha &= P\left(-t_{n-2,1-\alpha/2} \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2 / C_{xx}}} \leq t_{n-2,1-\alpha/2} \right) \\
&= P\left(\hat{\beta}_1 - t_{n-2,1-\alpha/2} \frac{\hat{\sigma}}{\sqrt{C_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{n-2,1-\alpha/2} \frac{\hat{\sigma}}{\sqrt{C_{xx}}} \right)
\end{aligned}
$$

Level $1-\alpha$ confidence interval for β_1 is

$$
\hat{\beta}_1 \pm t_{n-2,1-\alpha/2} \frac{\hat{\sigma}}{\sqrt{C_{xx}}}.
$$

For $\hat{\beta}_0$, we have

$$\begin{aligned}
\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{x} = \frac{\sum_{i=1}^n Y_i}{n} - \sum_{i=1}^n k_i Y_i \bar{x} \\
&= \sum_{i=1}^n (\frac{1}{n} - k_i \bar{x}) Y_i \\
&\equiv \sum_{i=1}^n c_i Y_i \, .
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbb{E}(\hat{\beta}_0) &= \mathbb{E}(\bar{Y} - \hat{\beta}_1 \bar{x}) = \mathbb{E}\Big(\frac{\sum_{i=1}^n Y_i}{n}\Big) - \beta_1 \bar{x} \\
&= \frac{\sum_{i=1}^n (\beta_0 + \beta_1 x_i)}{n} - \beta_1 \bar{x} \\
&= \frac{\sum_{i=1}^n \beta_0}{n} + \beta_1 \frac{\sum_{i=1}^n x_i}{n} - \beta_1 \bar{x} \\
&= \beta_0 \, ,
\end{aligned}$$

$$
\begin{aligned}
\mathbf{var}(\hat{\beta}_0) &= \mathbf{var}\Big(\bar{Y} - \hat{\beta}_1\bar{x}\Big) = \mathbf{var}\big(\bar{Y}\big) + \bar{x}^2\mathbf{var}(\hat{\beta}_1) - 2\mathbf{cov}(\bar{Y}, \hat{\beta}_1\bar{x}) \\
&= \frac{\sigma^2}{n} + \bar{x}^2\frac{\sigma^2}{C_{xx}} - 2\bar{x}\mathbf{cov}(\sum_{i=1}^{n} Y_i/n, \sum_{i=1}^{n} k_i Y_i) \\
&= \frac{\sigma^2}{n} + \bar{x}^2\frac{\sigma^2}{C_{xx}} - 2\bar{x}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{cov}(Y_i, k_j Y_j) \\
&= \frac{\sigma^2}{n} + \bar{x}^2\frac{\sigma^2}{C_{xx}} - 2\bar{x}\frac{1}{n}\sum_{i=1}^{n}\mathbf{cov}(Y_i, k_i Y_i) \\
&= \frac{\sigma^2}{n} + \bar{x}^2\frac{\sigma^2}{C_{xx}} - 2\bar{x}\frac{1}{n}\sum_{i=1}^{n} k_i\sigma^2 \\
&= \frac{\sigma^2}{n} + \bar{x}^2\frac{\sigma^2}{C_{xx}} - 2\bar{x}\frac{1}{n}\sum_{i=1}^{n}\frac{x_i - \bar{x}}{C_{xx}}\sigma^2 \\
&= \sigma^2\Big(\frac{1}{n} + \frac{\bar{x}^2}{C_{xx}}\Big),
\end{aligned}
$$

and

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

In addition, we can show that $\hat{\beta}_0$ and $\hat{\sigma}^2$ are independent, and so we have

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{C_{xx}}}} \sim t_{n-2},$$

which allows us to construct hypothesis testing and confidence intervals for β_0.

Estimation of mean response at $x = x_0$

For a given value of $x = x_0$,

$$Y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0\,, \varepsilon_0 \sim \mathcal{N}(0, \sigma^2) \text{ independent of } \varepsilon_1, \ldots, \varepsilon_n,$$

an estimator for the mean response $\mathbb{E}(Y_0)$ at $x = x_0$ is the value of the least squares regression line at $x = x_0$: $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$. Then

$$\begin{aligned}
\mathbb{E}(\hat{Y}_0) &= \mathbb{E}(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \beta_0 + \beta_1 x_0\,, \\
\mathbf{var}(\hat{Y}_0) &= \mathbf{var}\Big(\hat{\beta}_0 + \hat{\beta}_1 x_0\Big) = \mathbf{var}(\hat{\beta}_0) + \mathbf{var}(\hat{\beta}_1) x_0^2 + 2x_0 \mathbf{cov}(\hat{\beta}_0, \hat{\beta}_1) \\
&= \sigma^2\Big(\frac{1}{n} + \frac{\bar{x}^2}{C_{xx}}\Big) + x_0^2 \frac{\sigma^2}{C_{xx}} + 2x_0 \mathbf{cov}(\bar{Y} - \hat{\beta}_1 \bar{x}, \hat{\beta}_1) \\
&= \sigma^2\Big(\frac{1}{n} + \frac{\bar{x}^2}{C_{xx}}\Big) + x_0^2 \frac{\sigma^2}{C_{xx}} + 2x_0 \mathbf{cov}(\bar{Y}, \hat{\beta}_1) - 2x_0 \bar{x} \mathbf{var}(\hat{\beta}_1) \\
&= \sigma^2\Big(\frac{1}{n} + \frac{\bar{x}^2}{C_{xx}}\Big) + x_0^2 \frac{\sigma^2}{C_{xx}} + 0 - 2x_0 \bar{x} \frac{\sigma^2}{C_{xx}} \\
&= \sigma^2\Big[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{C_{xx}}\Big]\,.
\end{aligned}$$

Thus,

$$\hat{Y}_0 \sim \mathcal{N}\left(\beta_0 + \beta_1 x_0, \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{C_{xx}}\right]\right),$$

and

$$\frac{\hat{Y}_0 - (\beta_0 + \beta_1 x_0)}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{C_{xx}}}} \sim t_{n-2}.$$

This leads to the following level-$(1-\alpha)$ confidence interval for the mean response at $x = x_0$:

$$\hat{Y}_0 \pm t_{n-2;1-\alpha/2} \cdot \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{C_{xx}}}$$

Prediction for a future response at $x = x_0$

As a prediction for a future response Y_0 at $x = x_0$,

$$Y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0 \,,$$

where $\varepsilon_0 \sim \mathcal{N}(0, \sigma^2)$ is independent of $\varepsilon_1, \ldots, \varepsilon_n$, $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ is again unbiased i.e. $\mathbb{E}(\hat{Y}_0 - Y_0) = 0$, and its mean squared prediction error is

$$\begin{aligned}
\mathbb{E}[(\hat{Y}_0 - Y_0)^2] &= \mathbb{E}[\hat{Y}_0 - \mathbb{E}(\hat{Y}_0) + \mathbb{E}(\hat{Y}_0) - Y_0]^2 \\
&= \mathbf{var}(\hat{Y}_0) - 2\mathbf{cov}(\hat{Y}_0, Y_0) + \mathbf{var}(Y_0) \\
&= \mathbf{var}(\hat{Y}_0) + \mathbf{var}(Y_0) \\
&= \sigma^2 \Big[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{C_{xx}} \Big] .
\end{aligned}$$

This leads to a level-$(1-\alpha)$ prediction interval for Y_0:

$$\hat{Y}_0 \pm t_{n-2;1-\alpha/2} \cdot \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{C_{xx}}} .$$

Note that the standard error of the prediction of an individual is always larger than that of the estimation of the mean, but both of them will take their smallest values at $x_0 = \overline{x}$.

Extrapolation (estimation of the mean response or prediction of an individual response for values of x that fall outside the range of the values of x in the sample) may lead to large error because (i) the standard error is large, and more importantly (ii) we use the regression model to describe only the relationship between x and Y for values of x that fall in the observed range.

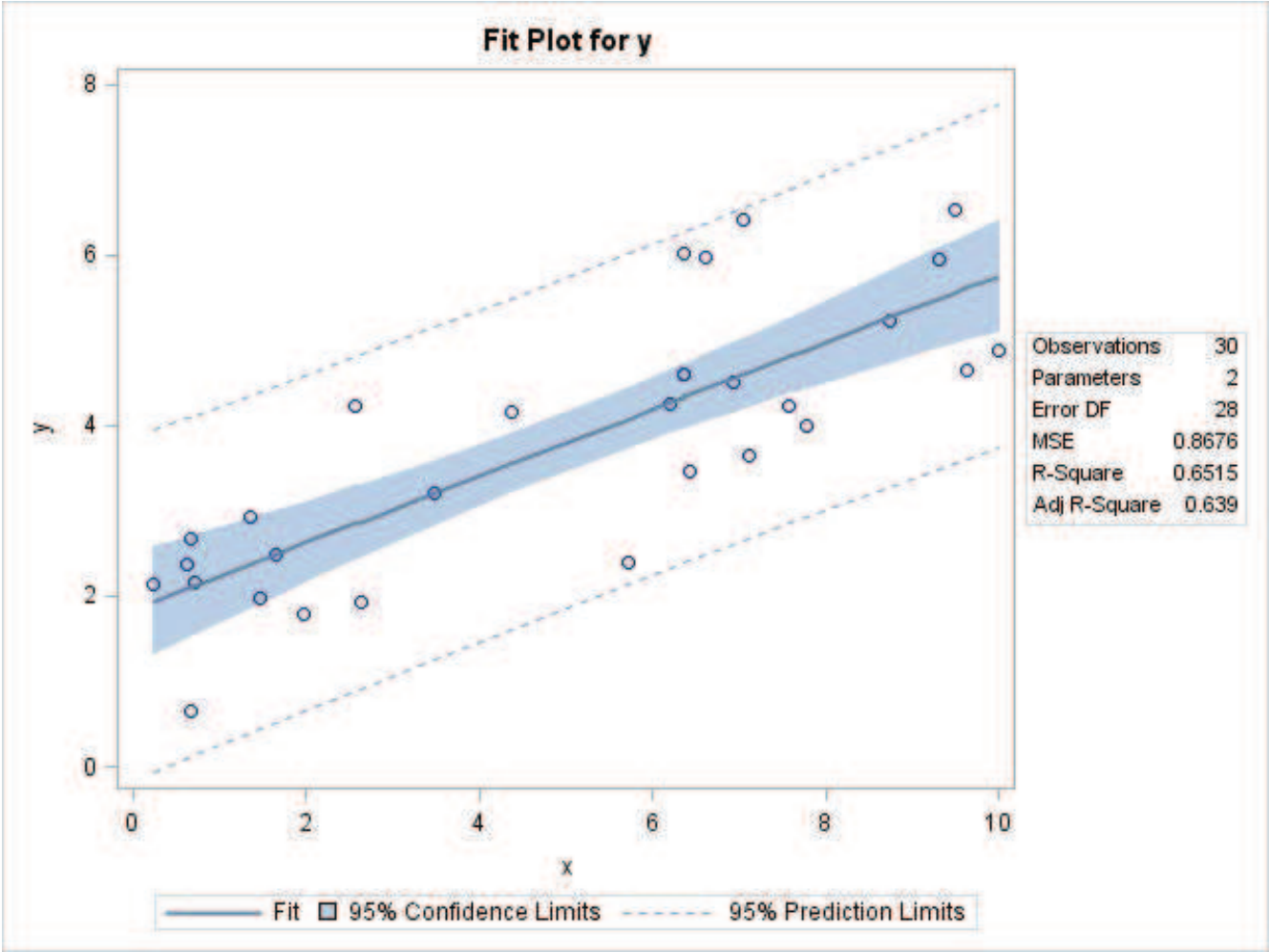

Figure: The default output of the confidence envelopes together with the regression line using the statement `model y=x` under `proc reg`.

Analysis of Variance

Define

$$SST = \text{total sum of squares} \equiv \sum (Y_i - \overline{Y})^2 \qquad (df = n-1),$$

$$SSR = \text{regression sum of squares} \equiv \sum (\hat{Y}_i - \overline{\hat{Y}})^2 \qquad (df = 1),$$

$$SSE = \text{error sum of squares} \equiv \sum (Y_i - \hat{Y}_i)^2 = \sum e_i^2 \qquad (df = n-2).$$

Note

$$\overline{\hat{Y}} = \frac{\sum \hat{Y}_i}{n} = \frac{\sum (\hat{\beta}_0 + \hat{\beta}_1 x_i)}{n} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x} = \overline{Y},$$

and so $SSR = \sum (\hat{Y}_i - \overline{\hat{Y}})^2 = \sum (\hat{Y}_i - \overline{Y})^2$.

The regression sum of squares SSR is the variability in $Y_1, \ldots, Y_n$ accounted for by the regression model, and it has only $df = 1$ coming from the slope parameter because the variation in $\hat{Y}_1, \ldots, \hat{Y}_n$ is fixed once the slope of the line is fixed (the intercept is determined by the slope and $\overline{Y}$).

The error sum of squares SSE is the variability in $Y_1, \ldots, Y_n$ not accounted for by the regression model, and it has $df = n - 2$ because though the variation comes from all n observed data, it loses two degrees of freedom as we estimated two parameters in order to get all the $\hat{Y}_i$.

Now, we break the total variability into two components:

$$\underbrace{\sum (Y_i - \overline{Y})^2}_{SST} = \sum (Y_i - \hat{Y}_i + \hat{Y}_i - \overline{Y})^2$$

$$= \underbrace{\sum (\hat{Y}_i - \overline{Y})^2}_{SSR} + \underbrace{\sum (Y_i - \hat{Y}_i)^2}_{SSE} + 2 \underbrace{\sum (\hat{Y}_i - \overline{Y})(Y_i - \hat{Y}_i)}_{= \underbrace{\sum e_i \hat{Y}_i}_{=0} - \overline{Y} \underbrace{\sum e_i}_{=0}}.$$

Hence, we have:

$$SST = SSR + SSE. \tag{1}$$

It is then obvious that the sum of the degrees of freedom of SSR and SSE should be equal to the degrees of freedom of SST.

$$
\begin{aligned}
& \sum \varepsilon_i^2 \\
= & \sum (Y_i - \beta_0 - \beta_1 x_i)^2 \\
= & \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i + \hat{\beta}_0 + \hat{\beta}_1 x_i - \overline{Y} + \overline{Y} - \beta_0 - \beta_1 x_i)^2 \\
= & \sum (e_i + \overline{Y} - \hat{\beta}_1 \overline{x} + \hat{\beta}_1 x_i - \overline{Y} + \overline{Y} - \beta_0 - \beta_1 x_i + \beta_1 \overline{x} - \beta_1 \overline{x})^2 \\
= & \sum [e_i + (\overline{Y} - \beta_0 - \beta_1 \overline{x}) + (\hat{\beta}_1 - \beta_1)(x_i - \overline{x})]^2 \\
= & \sum e_i^2 + n(\overline{Y} - \beta_0 - \beta_1 \overline{x})^2 + (\hat{\beta}_1 - \beta_1)^2 \sum (x_i - \overline{x})^2 ,
\end{aligned}
$$

where the third equality holds because $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}$ and the last equality holds because $\sum e_i = 0$ and $\sum e_i x_i = 0$.

Note that $\frac{\sum \varepsilon_i^2}{\sigma^2} \sim \chi_n^2$, and we can show that $\frac{\sum e_i^2}{\sigma^2}$, $\frac{n(\overline{Y} - \beta_0 - \beta_1 \overline{x})^2}{\sigma^2} \sim \chi_1^2$ and $\frac{(\hat{\beta}_1 - \beta_1)^2 \sum (x_i - \overline{x})^2}{\sigma^2} \sim \chi_1^2$ are independent.

Therefore, we have $\frac{\sum e_i^2}{\sigma^2} \sim \chi_{n-2}^2$.

In addition, we have

$$\begin{aligned}\sum \varepsilon_i^2 &= \sum e_i^2 + n(\overline{Y} - \beta_0 - \beta_1 \overline{x})^2 + \hat{\beta}_1^2 \sum (x_i - \overline{x})^2 \\ &\quad + \beta_1(\beta_1 - 2\hat{\beta}_1) \sum (x_i - \overline{x})^2 \\ &= SST + n(\overline{Y} - \beta_0 - \beta_1 \overline{x})^2 + \beta_1(\beta_1 - 2\hat{\beta}_1) \sum (x_i - \overline{x})^2 ,\end{aligned}$$

and so SST has $n - 1$ degrees of freedom.

Coefficient of Determination

Moreover, from equation (1), we have

$$1 = \underbrace{\frac{SSR}{SST}}_{:=R^2} + \frac{SSE}{SST},$$

where $0 \leq R^2 \leq 1$, called the *coefficient of determination*, is interpreted as the proportion of the sum of squares of deviations of the Y values about their mean that can be attributed to a linear relationship between Y and x.

- If $R^2 = 1$, then $SSE = 0$ and all points are lying on the regression line;
- if $R^2 = 0$, then no variability in Y_i is explained by the regression line, meaning that the regression line is flat so that it does not explain any variability in Y_i;
- if $R^2 = 0.8$, then the regression line accounts for 80% of the total variability of Y_i around their mean.

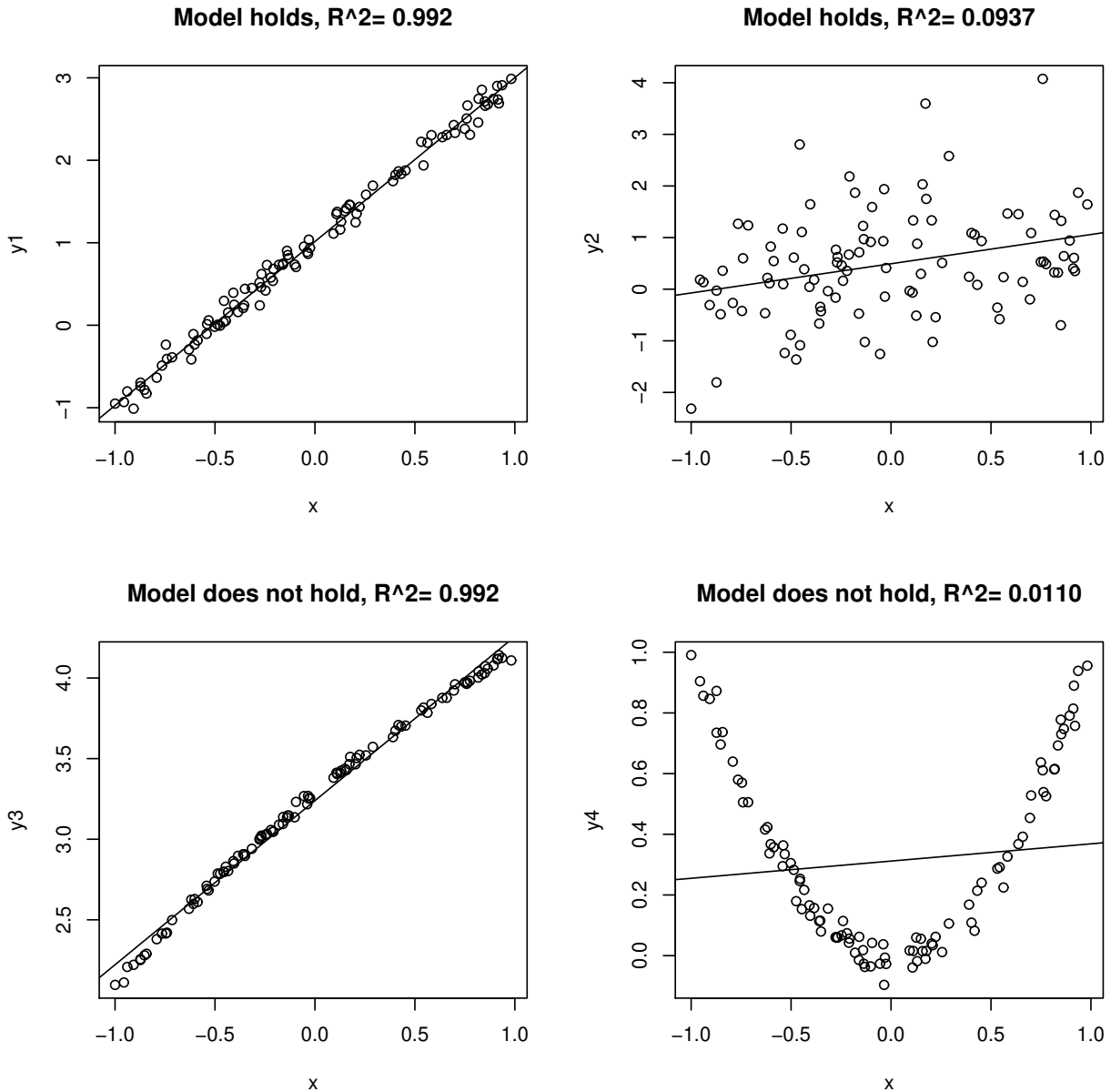

Figure: Value of R^2 in different situations.

Mean Square

Recall that each sum of squares has its degrees of freedom. The ratio of the sum of squares to its degrees of freedom is called the mean square. In particular,

$$MSE := \frac{SSE}{n-2} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2} = \hat{\sigma}^2,$$

which is an estimator for the variance σ^2 of ε; the alphabet M in *MSE* stands for *Mean* (and we may read *MSE* as *mean squared error*).

Note that $(n-2)\hat{\sigma}^2/\sigma^2 = SSE/\sigma^2 \sim \chi^2_{n-2}$, and the mean of the χ^2-distribution is its degrees of freedom. Hence, $\mathbb{E}((n-2)\hat{\sigma}^2/\sigma^2) = n-2 \Rightarrow \mathbb{E}(\hat{\sigma}^2) = \sigma^2$.

In general, the denominator of an *MS* (mean square) is the *df* of *SS* (sum of squares) because we should divide an *SS* by the number of free terms that have been summed up in order to calculate its mean.

By the same token, we denote the ratio of SSR to its degrees of freedom by MSR:

$$MSR := \frac{SSR}{k}$$

(stands for *regression mean square*), where k is the number of independent variables in the model. Here we consider just one independent variable, i.e. we have $k = 1$ and so $MSR = SSR$. Since $\hat{Y}_i - \overline{Y} = \hat{\beta}_1(x_i - \overline{x})$, we have

$$MSR := \frac{SSR}{1} = \sum(\hat{Y}_i - \overline{Y})^2 = \sum \hat{\beta}_1^2(x_i - \overline{x})^2 = \hat{\beta}_1^2 C_{xx}.$$

Then

$$\begin{aligned}
\mathbb{E}(MSR) = \mathbb{E}(\hat{\beta}_1^2 C_{xx}) = C_{xx} \cdot \mathbb{E}(\hat{\beta}_1^2) &= C_{xx}\left\{\mathbf{var}(\hat{\beta}_1) + [\mathbb{E}(\hat{\beta}_1)]^2\right\} \\
&= C_{xx}\left\{\frac{\sigma^2}{C_{xx}} + \beta_1^2\right\} \\
&= \sigma^2 + \beta_1^2 C_{xx}.
\end{aligned}$$

Thus, if $\beta_1 = 0$ and so $Y = \beta_0 + \varepsilon$ only, then $\mathbb{E}(MSR) = \sigma^2$, meaning that MSR is also an estimator of σ^2 if Y is just a constant plus a random error.

That is to say, for the same σ^2, we have two different estimators, namely, $MSE = \hat{\sigma}^2$ and, if $Y = \beta_0 + \varepsilon$, $MSR = \hat{\beta}_1^2 C_{xx}$.

Hence, if $\beta_1 = 0$, the ratio MSR/MSE should be close to 1.

However, if the ratio MSR/MSE is large, then MSR contains not only the variation from the random errors but also some extra variation caused by the variation in the expected values of the responses $Y_1, \ldots, Y_n$, meaning that β_1 is not zero.

ANOVA F-test

$H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$
Because under the null hypothesis that $\beta_1 = 0$, we have $SSR/\sigma^2 \sim \chi^2_1$ and so the test statistic is

$$F := \frac{MSR}{MSE} = \frac{\frac{SSR}{\sigma^2}/1}{\frac{SSE}{\sigma^2}/(n-1-1)} \sim F_{1,n-1-1},$$

giving us the rejection region $\{F \geq F_{1,n-2;1-\alpha}\}$, where $F_{1,n-2;1-\alpha}$ is the $(1-\alpha)$-quantile of the $F_{1,n-2}$ distribution.

This is a one-sided rejection region because a small value of F means no extra variation from the independent variables, which means there is no evidence against the null hypothesis. This test for $\mathrm{H}_0 : \beta_1 = 0$ is called *ANOVA* F-test.

ANOVA Table

ANOVA table

Source of variation	Sum of Squares	degrees of freedom	Mean square	F-ratio	p-value
Model	SSR	1	$MSR = \frac{SSR}{1}$	$\frac{MSR}{MSE}$	$\Pr(F_{1,n-1-1} \geq F\text{-ratio})$
Error	SSE	$n-1-1$	$MSE = \frac{SSE}{n-1-1}$		
Total	SST	$n-1$			

For the simple linear regression this ANOVA F-test is equivalent to the t-test for testing $\mathrm{H}_0 : \beta_1 = 0$, because

$$t_{df}^2 = \frac{z^2}{(\chi_{df}^2/df)} = \frac{\chi_1^2/1}{(\chi_{df}^2/df)} = F_{1,df}$$

and for testing $\mathrm{H}_0 : \beta_1 = 0$, since $\mathbf{var}(\hat{\beta}_1) = \sigma^2/C_{xx}$, we use the following form as the test statistic for the t-test:

$$T = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/C_{xx}}} = (\text{sign of } \hat{\beta}_1) \times \sqrt{\frac{\hat{\beta}_1^2 C_{xx}}{MSE}} = (\text{sign of } \hat{\beta}_1) \times \sqrt{\frac{MSR}{MSE}}.$$

Thus, the t-test is equivalent to the F-test here. Note that the t-statistic has the same sign as $\hat{\beta}_1$, and so the t-test allows us to have one-sided alternative, while the F-test only allows us to have two-sided alternative. (When we move to models with $k > 1$ independent variables, the ANOVA F-test is a test for all $\beta_1, \ldots, \beta_k$ while the t-test is a test for an individual β_i, and so these two tests are not equivalent in general. That is why we have two tests.)

Connection with Correlation Analysis

Let us consider the algebraic form of the coefficient of determination R^2 in more details. It is defined by

$$R^2 := \frac{SSR}{SST}.$$

In the above discussion, we already noticed that $\hat{Y}_i - \overline{Y} = \hat{\beta}_1(x_i - \overline{x})$, which implies that

$$SSR := \sum(\hat{Y}_i - \overline{Y})^2 = \hat{\beta}_1^2 \sum(x_i - \overline{x})^2 = \hat{\beta}_1^2 C_{xx},$$

and because $SST := \sum(Y_i - \overline{Y})^2 =: C_{YY}$, we have

$$R^2 = \frac{\hat{\beta}_1^2 C_{xx}}{C_{YY}} = \frac{\left(\frac{C_{xY}}{C_{xx}}\right)^2 C_{xx}}{C_{YY}} = \frac{C_{xY}^2}{C_{xx} C_{YY}}.$$

For the simple linear regression, we consider deterministic predictor x.

Suppose we have paired data $\{(X_i, Y_i)\}$ where both X_i and Y_i are random. Then we are talking about *correlation analysis*, in which we use a quantity describing the relationship between the two random variables X and Y, namely, the correlation coefficient, which is defined by

$$\rho := \frac{\mathbf{cov}(X, Y)}{\sqrt{\mathbf{var}(X)\mathbf{var}(Y)}},$$

where $\mathbf{cov}(X, Y) = \mathbb{E}\{(X - \mu_X)(Y - \mu_Y)\}$.

The sample covariance and the sample variances are

$$\widehat{\mathbf{cov}(X,Y)} = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{n-1},$$

$$\widehat{\mathbf{var}(X)} = \frac{\sum(X_i - \overline{X})^2}{n-1}, \quad \widehat{\mathbf{var}(Y)} = \frac{\sum(Y_i - \overline{Y})^2}{n-1}$$

and hence the sample correlation coefficient, denoted by r, is

$$r := \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2 \sum(Y_i - \overline{Y})^2}}.$$

Algebraically, we can see that

$$R^2 = r^2,$$

but note that r can be positive or negative and its sign is the same as that of the slope of the regression line. Thus,

$$r = \hat{\beta}_1 \times \sqrt{\frac{C_{xx}}{C_{YY}}} = (\text{sign of } \hat{\beta}_1) \times \sqrt{R^2}.$$

In particular, suppose (X, Y) follows the bivariate normal distribution. In MATH2206 you learned that if $\rho = 0$, then

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2},$$

which can be used to test the null hypothesis $\mathrm{H}_0\colon \rho = 0$. This t-distribution of the test statistic actually comes from the F-distribution of the ANOVA, which can be seen by noting the fact that algebraically,

$$1 - r^2 = \frac{SSE}{SST} \quad \text{and} \quad r^2 = \frac{SSR}{SST},$$

and so the ANOVA F-statistic in this case can be re-written as:

$$\begin{aligned} F &= \frac{MSR}{MSE} = \frac{SSR/1}{SSE/(n-2)} = \frac{SSR \cdot (n-2)}{SSE} \\ &= \frac{SSR \cdot (n-2)}{SST \cdot (1-r^2)} = \frac{(SSR/SST) \cdot (n-2)}{(1-r^2)} = \frac{r^2(n-2)}{(1-r^2)}. \end{aligned}$$

Take the square root of each side with the corresponding sign given to the right-hand side will lead to the t-statistic.

In this context, we prefer t-test to F-test because the former (which has either the plus or minus sign) allows us to have one-sided test.

However, this t-distribution result is not true when $\rho \neq 0$. That is, we are not able to construct confidence intervals for ρ.

To address this point, we used the Fisher Z-transform:

$$Z := \tanh^{-1} r = \frac{1}{2} \log \frac{1+r}{1-r} \overset{\text{approx.}}{\sim} \mathcal{N}\big(\tanh^{-1}(\rho), \frac{1}{n-3}\big),$$

where $\tanh(\cdot)$ is the hyperbolic tangent.

The Fisher transform allows us to construct confidence intervals and also allows us to test the null hypothesis that $\mathrm{H}_0 : \rho = \rho_0$.

Note that although R^2 (the coefficient of determination) and r^2 (the square of the sample correlation coefficient) are algebraically the same, they are different notions.

The coefficient of determination tells you how well your regression line (where the independent variable x is deterministic) can explain the variability in $Y_1, \ldots, Y_n$, while the correlation coefficient is a measure of the strength of the linear relationship between two random variables.

There is a true but unknown population correlation coefficient ρ, but we do not have a population coefficient of determination.

Also, note that high correlation does not imply causality.

Simple linear regression without intercept

In some particular applications (e.g. size–weight relationship), we may know in advance that the regression line must pass through the origin so that

$$Y_i = \beta_1 x_i + \varepsilon_i,\ i = 1, \ldots, n,$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. with $\mathcal{N}(0, \sigma^2)$ distribution.

The mathematical treatment of such a regression is almost the same as that of the regression with an unknown intercept β_0, except that now we have only one unknown parameter to estimate. In particular, the least squares estimator for $\hat{\beta}_1$ is

$$\hat{\beta}_1 = \frac{\sum x_i Y_i}{\sum x_i^2};$$

it is unbiased with variance $\mathbf{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum x_i^2}$, and

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum x_i^2}\right).$$

The least squares fitted valued for the mean response $\mathbb{E}(Y_0)$ at $x = x_0$, where

$$Y_0 = \beta_1 x_0 + \varepsilon_0,\ \varepsilon_0 \sim \mathcal{N}(0, \sigma^2) \text{ independent of } \varepsilon_1, \ldots, \varepsilon_n,$$

is $\hat{Y}_0 = \hat{\beta}_1 x_0$; it is unbiased for $\mathbb{E}(Y_0)$ and

$$\mathbf{var}(\hat{Y}_0) = x_0^2 \mathbf{var}(\hat{\beta}_1) = \sigma^2 \frac{x_0^2}{\sum x_i^2}\,.$$

As a prediction for a future response Y_0, $\hat{Y}_0$ has mean squared prediction error

$$\mathbb{E}[(Y_0 - \hat{Y}_0)^2] = \sigma^2 \left(1 + \frac{x_0^2}{\sum x_i^2}\right).$$

And,

$$\hat{\sigma}^2 = \frac{SSE}{n-1}, \quad \text{where } SSE = \sum Y_i^2 - \hat{\beta}_1 \sum x_i Y_i.$$

These resemble the corresponding ones in the ordinary simple linear regression, except that the sample means $\overline{x}$ and $\overline{Y}$ are now gone, and the value of the degrees of freedom of $\hat{\sigma}^2$ is $n-1$.

Sum of squares:

$$
\begin{aligned}
SST &:= \sum Y_i^2, \\
SSR &:= \sum \hat{Y}_i^2 = \sum \hat{\beta}_1^2 x_i^2, \\
SSE &:= \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - \hat{\beta}_1 x_i)^2 = \sum Y_i^2 - \hat{\beta}_1 \sum x_i Y_i.
\end{aligned}
$$

$$
\begin{aligned}
\sum Y_i^2 &= \sum (Y_i - \hat{\beta}_1 x_i + \hat{\beta}_1 x_i)^2 \\
&= \sum (Y_i - \hat{\beta}_1 x_i)^2 + 2 \sum (Y_i - \hat{\beta}_1 x_i) \hat{\beta}_1 x_i + \sum \hat{\beta}_1^2 x_i^2 \\
&= \sum (Y_i - \hat{\beta}_1 x_i)^2 + \hat{\beta}_1^2 \sum x_i^2
\end{aligned}
$$

The degrees of freedom of SSE is $n - 1$.

The SST is now called `Uncorrected Total Sum of Squares` because it is just $\sum Y_i^2$, i.e. the data are not centered by subtracting the sample mean from them (i.e. are uncorrected), and hence its degrees of freedom is n.