ST3131 Simple Regression: Inference

Semester 2 2023/2024

If printing, do DOUBLE-SIDED, each side TWO slides.

Introduction

Let x and y be variables on n subjects, with $s_x, s_y > 0$.

- NEW: Assume y_i is a realisation of random variable Y_i, $i = 1, \ldots, n$, with

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where

1. β_0 and β_1 are parameters, i.e., unknown constants.
2. $\varepsilon_1, \ldots, \varepsilon_n$ are independent $N(0, \sigma^2)$ RV's, σ a parameter.

This is a statistical model, called a simple regression model.

- The goal is inference: (i) to estimate the parameters and (ii) to test hypotheses about the parameters.

Topics

(I) Simple regression model

(II) Parameter estimation

(III) Confidence intervals

(IV) Hypothesis tests

(V) Simulating a simple regression model

(I) Hooke's Law

Suspend weight x_i on a metal rod, and measure the corresponding length y_i, for $i = 1, \ldots, n$, with $s_x > 0$.

- Hooke's Law says that provided the weight is not too large, y is a linear function of x. Due to measurement errors, we have

$$y_i \approx \beta_0 + \beta_1 x_i, \qquad 1 \leq i \leq n$$

  β_0: natural length of the rod, β_1: characteristic of metal,

- High school inference: draw scatter diagram, then draw the "best" straight line to estimate slope β_1 and y-intercept β_0.

- Uncertainties can be gauged from wriggling the best line.

(I) Parameters and measurement errors

- β_0 and β_1 are parameters: constants of unknown values. If known, no need to estimate.
- Imagine trying to measure $\beta_0 + \beta_1 x_i$, but with error ϵ_i. I.e.,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad 1 \leq i \leq n$$

  $\epsilon_1, \ldots, \epsilon_n$ are called measurement errors.
- Can β_0 and β_1 be determined?

(I) Simple regression model

Let x and y be data variable of length n. The simple regression model says

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad 1 \leq i \leq n$$

β_0 and β_1 are parameters;
$\epsilon_1, \ldots, \epsilon_n$ are realisations of $N(0, \sigma^2)$, σ is a parameter.

- This is a statistical model: a random mechanism for generating data (y given x here).
- We will be able to estimate β_0, β_1, and quantify the likely errors in the estimates, by estimating σ.
- All computations are in terms of y and x, but the interpretation depends on the model assumption: that ϵ's come from some normal RV with expectation 0.

(I) Mechanics of model

Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent $N(0, \sigma^2)$ RV's. Thus the model assumes that y_i is a realisation of

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- Let the values of β_0, β_1, σ be given. For each i, generate a realisation ϵ_i from ε_i. Then a realisations of Y_i is determined:

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

  y_i is revealed, but not ϵ_i. This gives a data set.

- Repeating the previous step gives another data set, with the same x values, but likely different y values. The process can be simulated in a computer, to generate many data sets.

- In practice, the analyst has only one data set to learn what β_0, β_1, σ might be.

(I) Consequences of model (1)

According to the model,

- For each i,

$$Y_i \sim \mathrm{N}(\beta_0 + \beta_1 x_i, \sigma^2)$$

  The i-th true response $\mathrm{E}(Y_i) = \beta_0 + \beta_1 x_i$ is only indirectly observed through y_i. Since the variance is constant, the scatter diagram of (x, y) should be homoschedastic.

- Since $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed (IID), a plot of $\epsilon_1, \ldots, \epsilon_n$ against $1, \ldots, n$ should show no systematic pattern.

- Can $Y_1, \ldots, Y_n$ be IID?

(I) Consequences of model (2)

Define

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Calculate

- $\mathrm{E}(\overline{Y})$
- $\mathrm{var}(\overline{Y})$
- $\mathrm{SD}(\overline{Y})$

(I) Setting up a model

State the simple regression model for

1. Pearson data: x = father's height, y = son's height.

2. Steam plant data: y = amount of steam, x = average temperature.

(I) Connection to real problems

- Does β_1 have a causal interpretation?
- How real are the assumption that $\epsilon_1, \ldots, \epsilon_n$ are realisations of IID normal RV's with expectation 0?

Consider
(i) Hooke's Law
(ii) Pearson data
(iii) Steam plant data: x, y as before

Generally, a model should not be taken too seriously unless it has been verified.

Topics

(I) Simple regression model

(II) Parameter estimation

(III) Confidence intervals

(IV) Hypothesis tests

(V) Simulating a simple regression model

(II) Estimating β_0 and β_1

In the equations

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad 1 \leq i \leq n$$

we know $y_1, \ldots, y_n, x_1, \ldots, x_n$, but not $\beta_0, \beta_1, \epsilon_1, \ldots, \epsilon_n$. The parameters β_0 and β_1 cannot be determined.

Gauss, Legendre, etc: let

$$S(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

The values (b_0, b_1) that minimise S are the least square (LS) estimates of (β_0, β_1).

Can you write down the LS estimates?

(II) LS estimates

We denote the LS estimates by $\hat{\beta}_0$ and $\hat{\beta}_1$. The "hat" is a reminder that these are estimates of β_0 and β_1.

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{ns_x^2}, \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Report $\hat{\beta}_1$ and $\hat{\beta}_0$ to 2 decimal places, for data from

1. Pearson
2. Steam plant, same x and y as before

(II) LS estimators

How good are the LS estimates? Is there any systematic error, or bias? How large is the random error?

In order to answer these questions, we need to see the LS estimates $\hat{\beta}_1$ and $\hat{\beta}_0$ are realisations of two RV's called estimators. The LS estimators are denoted by the same symbols:

$$\hat{\beta}_1 =$$

$$\hat{\beta}_0 =$$

This is an exception to the rule of using capital and small letters for RV and realisation.

(II) Predicted values and errors

- Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the LS estimates. The predicted or fitted value at x_i and the prediction error or residual are

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \qquad y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

- Now let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the LS estimators. The random predicted value and random residual are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \qquad Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

While the hatted symbols do double duties, let them be distinguished by the accompanying small or capital letters.

(II) Distribution of $(\hat{\beta}_0, \hat{\beta}_1)$ (1)

- How to get a realisation of $(\hat{\beta}_0, \hat{\beta}_1)$? This is driven by the simple regression model.
- Plotting many realisations on the plane gives a visualisation of the distribution. Will it look like an ellipse? Will the correlation be positive, negative, or 0?
- How to simulate the distribution in a computer?

(II) Distribution of $(\hat{\beta}_0, \hat{\beta}_1)$ (2)

Simulation requires specific values of β_0, β_1 and σ. A theoretical approach covers all possible cases.

Observe that both $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear functions of $Y_1, \ldots, Y_n$ (T1Q4). Hence there is a $2 \times n$ matrix A such that

$$AY = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \hat{\beta}$$

What can we conclude from slide 1.51?

The expectation and variance of $\hat{\beta}$ can be calculated more directly.

(II) Expectations

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{ns_x^2}$$

- Can we calculate $\mathrm{E}(\hat{\beta}_1)$ directly? Or is there a simpler way?

- $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{x}$. Calculate $\mathrm{E}(\hat{\beta}_0)$.

(II) $\text{var}(\hat{\beta}_1)$

(II) $\text{cov}(\overline{Y}, \hat{\beta}_1)$, $\text{var}(\hat{\beta}_0)$ and $\text{cov}(\hat{\beta}_0, \hat{\beta}_1)$

(II) $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased

- For any fixed values of β_0, β_1, $\sigma > 0$,

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \frac{\sigma^2}{n s_x^2} \begin{bmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \right)$$

- Since $\mathrm{E}(\hat{\beta}_0) = \beta_0$ and $\mathrm{E}(\hat{\beta}_1) = \beta_1$, the LS estimators are unbiased: no systematic error.
- Generate many realisations of $(\hat{\beta}_0, \hat{\beta}_1)$. In the scatter diagram, roughly what percentage of points will be within one $\mathrm{SD}(\hat{\beta}_1)$ in the vertical direction of the point (β_0, β_1)?

(II) Standard error (SE)

- For the Pearson data set, β_1 is estimated as 0.51. Since it is unbiased, the error

$$0.51 - \beta_1$$

  is due to random fluctuations, in the error RV's ε's.

- Can the error be calculated? Can it be estimated?

- The size of the error can be quantified by $\mathrm{SD}(\hat{\beta}_1)$. This is called the standard error.

$$\mathrm{SE} = \frac{\sigma}{\sqrt{n}} \frac{1}{s_x}$$

  But this has to be estimated.

(II) Estimating σ^2

$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where ϵ_i is a realisation of ε_i.

$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$, where e_i is the residual.

- In Pearson data set, $e_i = y_i - 33.89 - 0.51 x_i \approx \epsilon_i$. By the Law of Large Numbers, as $n \to \infty$, $\frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2 \to \sigma^2$, so it is reasonable to estimate σ^2 with

$$\frac{1}{n}\sum_{i=1}^{n} e_i^2 \approx 5.93$$

  Note $e_i = y_i - \hat{y}_i$.

- But this estimate is too small, since the LS estimates by definition minimise the sum of squares of prediction errors.

(II) Joint distribution of $(\hat{\beta}_0, \hat{\beta}_1)$ and $Y - \hat{Y}$

- Fact: Under the model,

$$(1) \qquad \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \frac{\sigma^2}{n s_x^2} \begin{bmatrix} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{bmatrix} \right)$$

$$(2) \qquad \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sigma^2} \sim \chi^2_{n-2}$$

(3) $(\hat{\beta}_0, \hat{\beta}_1)$ and $Y - \hat{Y}$ are independent.

- (1) has been seen before. The proofs of (2) and (3) use a change of coordinate technique from linear algebra, skipped in this course.

(II) Unbiased estimate of σ^2

- Fact: From (2), an unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

  For Pearson data set, this works out to be 5.94, slightly larger than before.

- Statisticians prefer unbiased estimators, even though $\hat{\sigma}$ is biased for σ. The square root of both 5.93 and 5.94 are about 2.44.

(II) Degree of freedom

- We say Y has n degrees of freedom, since it has n independent RV's.
- (2) implies $Y - \hat{Y}$ has $n-2$ degrees of freedom. The reason it is less: the normal equations

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i) = 0, \qquad \sum_{i=1}^{n}(Y_i - \hat{Y}_i)x_i = 0$$

impose two constraints on the random residuals. Hence, if we know $n-2$ of the residuals, then the last two are determined.

(II) SE for LS estimates

We complete the estimation of β_0 and β_1 for the Pearson data set.

- The LS estimate of β_0 is 33.89. The SE is

$$\mathrm{SD}(\hat{\beta}_0) = \frac{\sigma}{\sqrt{n}} \frac{\sqrt{\overline{x^2}}}{s_x} \approx 1.83$$

  We write $\beta_0 \approx 33.89 \pm 1.83$.

- The LS estimate of β_1 is 0.51. The SE is

$$\mathrm{SD}(\hat{\beta}_1) = \frac{\sigma}{\sqrt{n}} \frac{1}{s_x} \approx 0.03$$

  $\beta_1 \approx 0.51 \pm 0.03$.

(II) Linear functions of β_0 and β_1

Let

$$\theta = c_0\beta_0 + c_1\beta_1$$

where c_0 and c_1 are constants.

How to estimate θ, and compute an approximate SE?

(II) The importance of statistical models

The simple regression model says

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

where $\varepsilon_1, \ldots, \varepsilon_n$ are IID N(0,σ^2) RV's.

- The model postulates parameters β_0, β_1, and a random mechanism for producing $y_1, \ldots, y_n$. These are needed to talk about estimation and SE.
- Without a model, there is no parameter to estimate. This was the case in descriptive simple regression. The intercept 33.89 and gradient 0.51 are purely descriptive, not estimates of parameters.
- That the calculations can be done easily might encourage an erroneous belief that "estimate" and "SE" are meaningful without a statistical model.
- Statistical inference is not just a bunch of algorithms.

Topics

(I) Simple regression model

(II) Parameter estimation

(III) Confidence intervals

(IV) Hypothesis tests

(V) Simulating a simple regression model

(III) A t distribution from $\hat{\beta}_1$

Show that

- $$\frac{\hat{\beta}_1 - \beta_1}{\sigma/(\sqrt{n}s_x)} \sim \mathsf{N}(0,1)$$
- $$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$$

Deduce that

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/(\sqrt{n}s_x)} \sim t_{n-2}$$

(III) Upper quantile

Let $Z \sim \mathrm{N}(0,1)$. For $0 < p < 1$, let z_p be such that

$$\Pr(Z > z_p) = p$$

By sketching a normal curve, see that

- $z_p > 0$ if $p < 0.5$
- $z_p < 0$ if $p > 0.5$
- $z_{1-p} = -z_p$

Similarly, let $t_{p,n}$ be such that $\Pr(t_n > t_{p,n}) = p$.

(III) Random interval for β_1

Let $\alpha > 0$ be smaller than 0.5. Under the simple regression model,

$$\Pr\left(-t_{\frac{\alpha}{2},n-2}\frac{\hat{\sigma}}{\sqrt{n}s_x} \leq \hat{\beta}_1 - \beta_1 \leq t_{\frac{\alpha}{2},n-2}\frac{\hat{\sigma}}{\sqrt{n}s_x}\right) = 1-\alpha$$

implying

$$\Pr\left(\hat{\beta}_1 - t_{\frac{\alpha}{2},n-2}\frac{\hat{\sigma}}{\sqrt{n}s_x} \leq \beta_1 \leq \hat{\beta}_1 + t_{\frac{\alpha}{2},n-2}\frac{\hat{\sigma}}{\sqrt{n}s_x}\right) = 1-\alpha$$

In words, the random interval

$$\left(\hat{\beta}_1 - t_{\frac{\alpha}{2},n-2}\frac{\hat{\sigma}}{\sqrt{n}s_x}, \hat{\beta}_1 + t_{\frac{\alpha}{2},n-2}\frac{\hat{\sigma}}{\sqrt{n}s_x}\right)$$

covers β_1 with probability $1-\alpha$.

(III) Confidence interval for β_1

- Any realisation of the random interval

$$\left(\hat{\beta}_1 - t_{\frac{\alpha}{2},n-2}\frac{\hat{\sigma}}{\sqrt{n}s_x}, \hat{\beta}_1 + t_{\frac{\alpha}{2},n-2}\frac{\hat{\sigma}}{\sqrt{n}s_x}\right)$$

  is a $(1-\alpha)$-confidence interval (CI) for β_1.

- Pearson data set: $n = 1078$, let $\alpha = 0.05$. $t_{\frac{\alpha}{2},n-2} \approx 1.96$. $1.96 \times 0.03 \approx 0.05$. A 95%-CI for β_1 is

$$(0.51 - 0.05, 0.51 + 0.05) \approx (0.46, 0.57)$$

  We say

  "We are 95% confident that β_1 lies between 0.46 and 0.57."

(III) Frequency interpretation of a CI

- For the Pearson data, $\Pr(0.46 \leq \beta_1 \leq 0.57) = 0.95$ is not sensible. There is no random variable here. β_1 is a constant, which is either in (0.46,0.57) or not in there.
- Imagine generating many data sets from the model, using the same values of β_0, β_1, σ, and $x_1, \ldots, x_n$. For each data set, compute a 95%-CI for β_1. Then around 95% of these intervals contain β_1.
- Even with many CI's, we will not know exactly which contain β_1 and which do not. All that we are assured of is the percentage of CI's that do so. Hence the name "frequency interpretation".

(III) Linear functions of β_0 and β_1

Let

$$\theta = c_0\beta_0 + c_1\beta_1$$

where c_0 and c_1 are constants.

How to construct a CI for θ?

(III) More on CI

- The previous conclusion about a parameter, in the form of estimate $\pm$ SE, is a point estimation.
- The CI is an interval estimation. Generally, the $(1-\alpha)$-CI on a parameter has the form

$$(\text{estimate} - \kappa\,\text{SE}, \text{estimate} + \kappa\,\text{SE})$$

  where κ depends on α and the distribution of the estimator.
- As $n \to \infty$, $t_n \to$ N(0,1). Thus, for large n, a κ based on a t-distribution can be approximated by one based on the standard normal distribution. In the Pearson data set, $z_{\frac{\alpha}{2}} \approx 1.96$, virtually identical with $t_{\frac{\alpha}{2},n-2}$.

Topics

(I) Simple regression model

(II) Parameter estimation

(III) Confidence intervals

(IV) Hypothesis tests

(V) Simulating a simple regression model

(IV) Is $\beta_1 = 0$?

Suppose $\beta_1 = 0$.

- Is it useful to use x to predict y?
- Generate y under the model. How would a scatter diagram of x and y look like?
- What is $\Pr(\hat{\beta}_1 = 0)$?

Even if $\beta_1 = 0$, the data will show some trend, by chance. This makes it challenging to answer "Yes" or "No" to the question. Statisticians focus on finding evidence to say "No".

(IV) Testing the hypothesis $\beta_1 = 0$

- The null and alternative hypotheses are

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0$$

- Basic idea: if the estimate of β_1 is far away from 0, then we are more skeptical of H_0, to the extent of rejecting it. Conversely, if it is not too far away from 0, then we do not reject H_0.
- The distance from 0 is measured relative to the SE. In Pearson data set, $\hat{\beta}_1 = 0.51$, SE is about 0.03, so the test statistic is about $0.51/0.03 \approx 19$.

(IV) The t statistic

The t statistics for testing H_0 is the random variable

$$t = \frac{\hat{\beta}_1}{\hat{\sigma}/(\sqrt{n}s_x)}$$

Suppose H_0 is true.

- What is the distribution of t?
- For large n, what is roughly the distribution of t?

Is 19 an extreme realisation of t, if H_0 is true?

(IV) P value

- Instead of acting on a test statistic like 19, statisticians like to use the P value: the probability that the random test statistic is as extreme as, or more extreme than, the observed test statistic. In this case,

$$P = \Pr(|t_{1076}| \geq 19) \approx 0$$

  Since P is so small, we feel strongly about rejecting H_0, and conclude that β_1 is not 0.

- Frequency interpretation: Imagine generating many data sets, assuming $\beta_1 = 0$. Out of the corresponding t statistics, we will see none that is more than 19 in absolute value. This means that our observed statistic of 19 is way too large, to the extent of making $\beta_1 = 0$ not believable.

(IV) More on t test

- The test is two-tailed, since H_1 says $\beta_1 \neq 0$. If instead H_1 says $\beta_1 > 0$, then the test is one-tailed, and the P value is half as before: $\Pr(t_{1076} \geq 19)$.
- P value is not the probability that H_0 is true. H_0 is either true or false, and it is not sensible to talk about this probability.
- How small should P be to reject H_0. Practitioners like to use 0.05, or 0.01, but there really is not good answer. Like there is no good answer to: What temperature in C is too hot? The smaller the P value, the more unlikely that H_0 is true.

(IV) Duality between CI and test

Consider a test using the t statistics for

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0$$

of size α, i.e., it rejects H_0 if $P < \alpha$.

We can do this test by just looking at the $(1 - \alpha)$-CI.

- In Pearson data, a 95%-CI for β_1 is (0.46, 0.57). Since the interval does not contain 0, we conclude that H_0 is rejected by the t-test of size 0.05, i.e., we conclude that $P < 0.05$.
- Conversely, if the 95%-CI contains 0, then H_0 is not rejected at level 0.05.
- More generally, based on the CI (0.46, 0.57) for β_1, we know that $H_0 : \beta_1 = 0.50$ is not rejected at level 0.05.

(IV) Revisiting ANOVA

- The analysis of variance says

$$\text{var}(y) = \text{var}(\hat{y}) + \text{var}(e)$$

where the residuals are $e = y - \hat{y}$.

- Multiplying n throughout gives

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

(IV) Sums of squares

Denote

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 \quad : \quad \text{Total Sum of Squares (SST)}$$

$$\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 \quad : \quad \text{Regression Sum of Squares (SSR)}$$

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad : \quad \text{Error Sum of Squares (SSE)}$$

The analysis of variance is

$$\text{SST} = \text{SSR} + \text{SSE}$$

For Pearson, SST $\approx$ 8541.63, SSR $\approx$ 2145.35, SSE $\approx$ 6396.28.

(IV) Random SS

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Always,

- Always, $\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-2}$.
- SSR and SSE are independent.

If $\beta_1 = 0$, $\frac{\text{SSR}}{\sigma^2} \sim \chi^2_1$. Hence under H_0,

$$F = \frac{\text{SSR}}{\text{SSE}/(n-2)} \sim F_{1,n-2}$$

(IV) F test of $\beta_1 = 0$

Besides t, here is another way to test

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0$$

- Let f be the realisation of F, i.e.,

$$f = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-2)} = \frac{(n-2)r^2}{1-r^2}$$

- Larger f is stronger evidence against H_0. The P value is

$$\Pr(F_{1,n-2} \geq f)$$

- In Pearson, $f \approx 360.90$. $P \approx 0$.

Actually, the t and F statistics are related: $t^2 = F$.

(IV) t vs F

Two ways to test

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0$$

1. Under H_0,
$$t = \frac{\hat{\beta}_1}{\hat{\sigma}/(\sqrt{n}s_x)} \sim t_{n-2}$$
2. Under H_0,
$$F = \frac{\text{SSR}}{\text{SSE}/(n-2)} \sim F_{1,n-2}$$

They are equivalent.
For test statistics, $F = t^2$.
For distributions, $F_{1,n-2} = t^2_{n-2}$.

(IV) Revisiting `lm`

`regr = lm(y ~ x)`

What do we recognise from the following outputs?

```
summary(regr)
```

```
anova(regr)
```

Topics

(I) Simple regression model

(II) Parameter estimation

(III) Confidence intervals

(IV) Hypothesis tests

(V) Simulating a simple regression model

Simulating RV

- A random variable (RV) is a random mechanism for generating numbers. The realisations are unpredictable individually, but when aggregated, exhibit regularity, as described by Law of Large Numbers.
- Simulation: a non-random mechanism that imitates the realisations of an RV. A computer is a great for simulation, relying on pseudo-random number generators, which are largely deterministic.
- Computer simulations are not realisations of an RV. But they are rather good for studying an RV.

Simulating normal RV

- The following simulates 1000 realisations from N(0,1).

```
sim1 = rnorm(1000)
```

Say what you expect to see from

```
mean(sim1)
sd(sim1)
hist(sim1)
```

then run them.

- Repeat the above, for these two lines.

```
sim2 = rnorm(1000, 2, 3)
sim3 = rnorm(1000, 2, 3) + rnorm(1000, 1, 4)
```

Simulating a regression model (1)

- Let x be as in Pearson data set.

  Set $\beta_0 = 33.89$, $\beta_1 = 0.51$, $\sigma = 2.44$.

- Simulate the son's height from

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

  where $\varepsilon_1, \ldots, \varepsilon_n$ are independent $\mathrm{N}(0, \sigma^2)$ RV's.

- Denote the simulated values by $y_1^*, \ldots, y_n^*$. What can you say about $\mathrm{cor}(x, y^*)$?

- Let the y-intercept and slope of the regression line of y^* on x be $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$. What can you say about them?

Simulating a regression model (2)

Repeat the simulation 1000 times, so that you have 1000 realisations of the RV's $(\hat{\beta}_0^*, \hat{\beta}_1^*)$.

- Make a histogram of the realisations of $\hat{\beta}_0^*$.
- Make a histogram of the realisations of $\hat{\beta}_1^*$.
- Make a scatter diagram for the realisations of $(\hat{\beta}_0^*, \hat{\beta}_1^*)$.

Simulating a regression model (3)

- What is the theoretical distribution of $\hat{\beta}_0^*$? Make a normal quantile plot for the realisations.
- What is the theoretical distribution of $\hat{\beta}_1^*$? Make a normal quantile plot for the realisations.
- What is the theoretical correlation between $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$? Compare with the correlation between the realisations.

Conclusion (1)

- Let $x_1, \ldots, x_n$ be given, with $s_x > 0$. The simple regression model assumes data $y_1, \ldots, y_n$ have been generated from the RV's

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

  where β_0, β_1 and σ are parameters, and $\varepsilon_1, \ldots, \varepsilon_n$ are independent $\mathrm{N}(0, \sigma^2)$ RV's. The realisation of ε_i is ϵ_i, so that $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.

- β_0 and β_1 are estimated by $\hat{\beta}_0$ and $\hat{\beta}_1$, which are numerically the y-intercept and slope of the regression line of y on x.

- Under the model, the estimates are realisations of RV's called estimators, which support the estimation of SE's, construction of CI's, and testing of hypotheses.

Conclusion (2)

- The i-th fitted value is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, and the residual is $e_i = y_i - \hat{y}_i$. An unbiased estimate of σ^2 is

$$\frac{1}{n-2} \sum_{i=1}^{n} e_i^2$$

- The model is a mental construction imposed on the data (x, y) for the purpose of inference. All computations can be done with no regard for the model, but the interpretation depends on the model providing reasonable fit to the data.
- If the model seems inadequate, then we should be cautious about the inference results. However, the descriptive techniques from the previous lecture can always be used.

Computation summary

Basic quantities: $\overline{x}$, $\overline{y}$, s_x^2, s_y^2, s_{xy}.

- Data summary $\overline{x}$, $\overline{y}$, s_x, s_y, $r = \frac{s_{xy}}{s_x s_y}$.
- LS estimates $\hat{\beta}_1 = \frac{r\, s_y}{s_x} = \frac{s_{xy}}{s_x^2}$, $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$.
- $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, $e = y - \hat{y}$. Unbiased estimate of σ^2: $\hat{\sigma}^2 = \frac{n}{n-2}(1 - r^2)s_y^2$.
- ANOVA table:

SS	Definition	Computation	df
SSR	$\sum_{i=1}^n (\hat{y}_i - \overline{y})^2$	$nr^2 s_y^2$	1
SSE	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	$n(1 - r^2)s_y^2$	$n - 2$
SST	$\sum_{i=1}^n (y_i - \overline{y})^2$	ns_y^2	$n - 1$