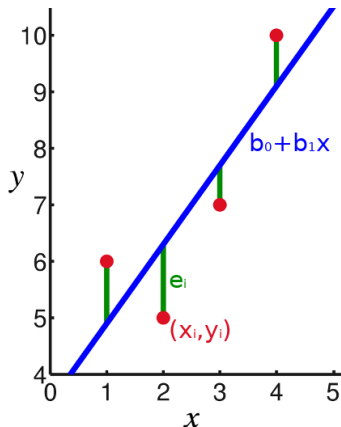


Covariance and correlation give some idea of direction and strength of linear dependencies. **Simple linear regression** tries to give an explicit linear function that explains one random variable Y by another random variable X

Here E are random variations in Y that cannot be explained by X .

- Random variables are denoted by capital letters, e.g. X, Y, E .
- Individual observations are denoted by lower-case letters often with an index, e.g. x_i, y_i, e_i .
- Samples are given by vectors, e.g. $\mathbf{x}, \mathbf{y}, \mathbf{e}$.

•



For random variables X, Y, E

$$Y = b_0 + b_1X + E$$

For samples x_j, y_j, e_j

$$y_i = b_0 + b_1 x_i + e_i, \quad i = 1, 2, \dots, n$$

Applications

- b_0 and b_1 can contain interesting information about the data generating process, e.g. constant in the ideal gas law.
- Predict new values of Y given potential values of X , e.g. predict electricity consumption for the next day.

Notation

- () () () ()

Citizenship

Journal of Management Inquiry

1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025, 2026, 2027, 2028, 2029, 2030, 2031, 2032, 2033, 2034, 2035, 2036, 2037, 2038, 2039, 2040, 2041, 2042, 2043, 2044, 2045, 2046, 2047, 2048, 2049, 2050, 2051, 2052, 2053, 2054, 2055, 2056, 2057, 2058, 2059, 2060, 2061, 2062, 2063, 2064, 2065, 2066, 2067, 2068, 2069, 2070, 2071, 2072, 2073, 2074, 2075, 2076, 2077, 2078, 2079, 2080, 2081, 2082, 2083, 2084, 2085, 2086, 2087, 2088, 2089, 2090, 2091, 2092, 2093, 2094, 2095, 2096, 2097, 2098, 2099, 2100, 2101, 2102, 2103, 2104, 2105, 2106, 2107, 2108, 2109, 2110, 2111, 2112, 2113, 2114, 2115, 2116, 2117, 2118, 2119, 2120, 2121, 2122, 2123, 2124, 2125, 2126, 2127, 2128, 2129, 2130, 2131, 2132, 2133, 2134, 2135, 2136, 2137, 2138, 2139, 2140, 2141, 2142, 2143, 2144, 2145, 2146, 2147, 2148, 2149, 2150, 2151, 2152, 2153, 2154, 2155, 2156, 2157, 2158, 2159, 2160, 2161, 2162, 2163, 2164, 2165, 2166, 2167, 2168, 2169, 2170, 2171, 2172, 2173, 2174, 2175, 2176, 2177, 2178, 2179, 2180, 2181, 2182, 2183, 2184, 2185, 2186, 2187, 2188, 2189, 2190, 2191, 2192, 2193, 2194, 2195, 2196, 2197, 2198, 2199, 2200, 2201, 2202, 2203, 2204, 2205, 2206, 2207, 2208, 2209, 2210, 2211, 2212, 2213, 2214, 2215, 2216, 2217, 2218, 2219, 2220, 2221, 2222, 2223, 2224, 2225, 2226, 2227, 2228, 2229, 2230, 2231, 2232, 2233, 2234, 2235, 2236, 2237, 2238, 2239, 2240, 2241, 2242, 2243, 2244, 2245, 2246, 2247, 2248, 2249, 2250, 2251, 2252, 2253, 2254, 2255, 2256, 2257, 2258, 2259, 2260, 2261, 2262, 2263, 2264, 2265, 2266, 2267, 2268, 2269, 2270, 2271, 2272, 2273, 2274, 2275, 2276, 2277, 2278, 2279, 2280, 2281, 2282, 2283, 2284, 2285, 2286, 2287, 2288, 2289, 2290, 2291, 2292, 2293, 2294, 2295, 2296, 2297, 2298, 2299, 2300, 2301, 2302, 2303, 2304, 2305, 2306, 2307, 2308, 2309, 2310, 2311, 2312, 2313, 2314, 2315, 2316, 2317, 2318, 2319, 2320, 2321, 2322, 2323, 2324, 2325, 2326, 2327, 2328, 2329, 2330, 2331, 2332, 2333, 2334, 2335, 2336, 2337, 2338, 2339, 2340, 2341, 2342, 2343, 2344, 2345, 2346, 2347, 2348, 2349, 2350, 2351, 2352, 2353, 2354, 2355, 2356, 2357, 2358, 2359, 2360, 2361, 2362, 2363, 2364, 2365, 2366, 2367, 2368, 2369, 2370, 2371, 2372, 2373, 2374, 2375, 2376, 2377, 2378, 2379, 2380, 2381, 2382, 2383, 2384, 2385, 2386, 2387, 2388, 2389, 2390, 2391, 2392, 2393, 2394, 2395, 2396, 2397, 2398, 2399, 2400, 2401, 2402, 2403, 2404, 2405, 2406, 2407, 2408, 2409, 2410, 2411, 2412, 2413, 2414, 2415, 2416, 2417, 2418, 2419, 2420, 2421, 2422, 2423, 2424, 2425, 2426, 2427, 2428, 2429, 2430, 2431, 2432, 2433, 2434, 2435, 2436, 2437, 2438, 2439, 2440, 2441, 2442, 2443, 2444, 2445, 2446, 2447, 2448, 2449, 2450, 2451, 2452, 2453, 2454, 2455, 2456, 2457, 2458, 2459, 2460, 2461, 2462, 2463, 2464, 2465, 2466, 2467, 2468, 2469, 2470, 2471, 2472, 2473, 2474, 2475, 2476, 2477, 2478, 2479, 2480, 2481, 2482, 2483, 2484, 2485, 2486, 2487, 2488, 2489, 2490, 2491, 2492, 2493, 2494, 2495, 2496, 2497, 2498, 2499, 2500, 2501, 2502, 2503, 2504, 2505, 2506, 2507, 2508, 2509, 2510, 2511, 2512, 2513, 2514, 2515, 2516, 2517, 2518, 2519, 2520, 2521, 2522, 2523, 2524, 2525, 2526, 2527, 2528, 2529, 2530, 2531, 2532, 2533, 2534, 2535, 2536, 2537, 2538, 2539, 2540, 2541, 2542, 2543, 2544, 2545, 2546, 2547, 2548, 2549, 2550, 2551, 2552, 2553, 2554, 2555, 2556, 2557, 2558, 2559, 2560, 2561, 2562, 2563, 2564, 2565, 2566, 2567, 2568, 2569, 2570, 2571, 2572, 2573, 2574, 2575, 2576, 2577, 2578, 2579, 2580, 2581, 2582, 2583, 2584, 2585, 2586, 2587, 2588, 2589, 2590, 2591, 2592, 2593, 2594, 2595, 2596, 2597, 2598, 2599, 2600, 2601, 2602, 2603, 2604, 2605, 2606, 2607, 2608, 2609, 2610, 2611, 2612, 2613, 2614, 2615, 2616, 2617, 2618, 2619, 2620, 2621, 2622, 2623, 2624, 2625, 2626, 2627, 2628, 2629, 2630, 2631, 2632, 2633, 2634, 2635, 2636, 2637, 2638, 2639, 2640, 2641, 2642, 2643, 2644, 2645, 2646, 2647, 2648, 2649, 2650, 2651, 2652, 2653, 2654, 2655, 2656, 2657, 2658, 2659, 2660, 2661, 2662, 2663, 2664, 2665, 2666, 2667, 2668, 2669, 2670, 2671, 2672, 2673, 2674, 2675, 2676, 2677, 2678, 2679, 2680, 26

$$\mathbf{X} = [\mathbf{1} \ \mathbf{x}] = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

1. 2. 3. 4. 5. 6. 7. 8. 9. 10. 11. 12. 13. 14. 15. 16. 17. 18. 19. 20. 21. 22. 23. 24. 25. 26. 27. 28. 29. 30. 31. 32. 33. 34. 35. 36. 37. 38. 39. 40. 41. 42. 43. 44. 45. 46. 47. 48. 49. 50. 51. 52. 53. 54. 55. 56. 57. 58. 59. 60. 61. 62. 63. 64. 65. 66. 67. 68. 69. 70. 71. 72. 73. 74. 75. 76. 77. 78. 79. 80. 81. 82. 83. 84. 85. 86. 87. 88. 89. 90. 91. 92. 93. 94. 95. 96. 97. 98. 99. 100. 101. 102. 103. 104. 105. 106. 107. 108. 109. 110. 111. 112. 113. 114. 115. 116. 117. 118. 119. 120. 121. 122. 123. 124. 125. 126. 127. 128. 129. 130. 131. 132. 133. 134. 135. 136. 137. 138. 139. 140. 141. 142. 143. 144. 145. 146. 147. 148. 149. 150. 151. 152. 153. 154. 155. 156. 157. 158. 159. 160. 161. 162. 163. 164. 165. 166. 167. 168. 169. 170. 171. 172. 173. 174. 175. 176. 177. 178. 179. 180. 181. 182. 183. 184. 185. 186. 187. 188. 189. 190. 191. 192. 193. 194. 195. 196. 197. 198. 199. 200. 201. 202. 203. 204. 205. 206. 207. 208. 209. 210. 211. 212. 213. 214. 215. 216. 217. 218. 219. 220. 221. 222. 223. 224. 225. 226. 227. 228. 229. 230. 231. 232. 233. 234. 235. 236. 237. 238. 239. 240. 241. 242. 243. 244. 245. 246. 247. 248. 249. 250. 251. 252. 253. 254. 255. 256. 257. 258. 259. 260. 261. 262. 263. 264. 265. 266. 267. 268. 269. 270. 271. 272. 273. 274. 275. 276. 277. 278. 279. 280. 281. 282. 283. 284. 285. 286. 287. 288. 289. 290. 291. 292. 293. 294. 295. 296. 297. 298. 299. 300. 301. 302. 303. 304. 305. 306. 307. 308. 309. 310. 311. 312. 313. 314. 315. 316. 317. 318. 319. 320. 321. 322. 323. 324. 325. 326. 327. 328. 329. 330. 331. 332. 333. 334. 335. 336. 337. 338. 339. 340. 341. 342. 343. 344. 345. 346. 347. 348. 349. 350. 351. 352. 353. 354. 355. 356. 357. 358. 359. 360. 361. 362. 363. 364. 365. 366. 367. 368. 369. 370. 371. 372. 373. 374. 375. 376. 377. 378. 379. 380. 381. 382. 383. 384. 385. 386. 387. 388. 389. 390. 391. 392. 393. 394. 395. 396. 397. 398. 399. 400. 401. 402. 403. 404. 405. 406. 407. 408. 409. 410. 411. 412. 413. 414. 415. 416. 417. 418. 419. 420. 421. 422. 423. 424. 425. 426. 427. 428. 429. 430. 431. 432. 433. 434. 435. 436. 437. 438. 439. 440. 441. 442. 443. 444. 445. 446. 447. 448. 449. 450. 451. 452. 453. 454. 455. 456. 457. 458. 459. 460. 461. 462. 463. 464. 465. 466. 467. 468. 469. 470. 471. 472. 473. 474. 475. 476. 477. 478. 479. 480. 481. 482. 483. 484. 485. 486. 487. 488. 489. 490. 491. 492. 493. 494. 495. 496. 497. 498. 499. 500. 501. 502. 503. 504. 505. 506. 507. 508. 509. 510. 511. 512. 513. 514. 515. 516. 517. 518. 519. 520. 521. 522. 523. 524. 525. 526. 527. 528. 529. 530. 531. 532. 533. 534. 535. 536. 537. 538. 539. 540. 541. 542. 543. 544. 545. 546. 547. 548. 549. 550. 551. 552. 553. 554. 555. 556. 557. 558. 559. 560. 561. 562. 563. 564. 565. 566. 567. 568. 569. 570. 571. 572. 573. 574. 575. 576. 577. 578. 579. 580. 581. 582. 583. 584. 585. 586. 587. 588. 589. 590. 591. 592. 593. 594. 595. 596. 597. 598. 599. 600. 601. 602. 603. 604. 605. 606. 607. 608. 609. 610. 611. 612. 613. 614. 615. 616. 617. 618. 619. 620. 621. 622. 623. 624. 625. 626. 627. 628. 629. 630. 631. 632. 633. 634. 635. 636. 637. 638. 639. 640. 641. 642. 643. 644. 645. 646. 647. 648. 649. 650. 651. 652. 653. 654. 655. 656. 657. 658. 659. 660. 661. 662. 663. 664. 665. 666. 667. 668. 669. 670. 671. 672. 673. 674. 675. 676. 677. 678. 679. 680. 681. 682. 683. 684. 685. 686. 687. 688. 689. 690. 691. 692. 693. 694. 695. 696. 697. 698. 699. 700. 701. 702. 703. 704. 705. 706. 707. 708. 709. 710. 711. 712. 713. 714. 715. 716. 717. 718. 719. 720. 721. 722. 723. 724. 725. 726. 727. 728. 729. 730. 731. 732. 733. 734. 735. 736. 737. 738. 739. 740. 741. 742. 743. 744. 745. 746. 747. 748. 749. 750. 751. 752. 753. 754. 755. 756. 757. 758. 759. 760. 761. 762. 763. 764. 765. 766. 767. 768. 769. 770. 771. 772. 773. 774. 775. 776. 777. 778. 779. 780. 781. 782. 783. 784. 785. 786. 787. 788. 789. 790. 791. 792. 793. 794. 795. 796. 797. 798. 799. 800. 801. 802. 803. 804. 805. 806. 807. 808. 809. 810. 811. 812. 813. 814. 815. 816. 817. 818. 819. 820. 821. 822. 823. 824. 825. 826. 827. 828. 829. 830. 831. 832. 833. 834. 835. 836. 837. 838. 839. 840. 84

$\hat{b}_0, \quad \hat{b}_1, \quad \hat{\mathbf{e}}$

100

1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025, 2026, 2027, 2028, 2029, 2030, 2031, 2032, 2033, 2034, 2035, 2036, 2037, 2038, 2039, 2040, 2041, 2042, 2043, 2044, 2045, 2046, 2047, 2048, 2049, 2050, 2051, 2052, 2053, 2054, 2055, 2056, 2057, 2058, 2059, 2060, 2061, 2062, 2063, 2064, 2065, 2066, 2067, 2068, 2069, 2070, 2071, 2072, 2073, 2074, 2075, 2076, 2077, 2078, 2079, 2080, 2081, 2082, 2083, 2084, 2085, 2086, 2087, 2088, 2089, 2090, 2091, 2092, 2093, 2094, 2095, 2096, 2097, 2098, 2099, 2100, 2101, 2102, 2103, 2104, 2105, 2106, 2107, 2108, 2109, 2110, 2111, 2112, 2113, 2114, 2115, 2116, 2117, 2118, 2119, 2120, 2121, 2122, 2123, 2124, 2125, 2126, 2127, 2128, 2129, 2130, 2131, 2132, 2133, 2134, 2135, 2136, 2137, 2138, 2139, 2140, 2141, 2142, 2143, 2144, 2145, 2146, 2147, 2148, 2149, 2150, 2151, 2152, 2153, 2154, 2155, 2156, 2157, 2158, 2159, 2160, 2161, 2162, 2163, 2164, 2165, 2166, 2167, 2168, 2169, 2170, 2171, 2172, 2173, 2174, 2175, 2176, 2177, 2178, 2179, 2180, 2181, 2182, 2183, 2184, 2185, 2186, 2187, 2188, 2189, 2190, 2191, 2192, 2193, 2194, 2195, 2196, 2197, 2198, 2199, 2200, 2201, 2202, 2203, 2204, 2205, 2206, 2207, 2208, 2209, 2210, 2211, 2212, 2213, 2214, 2215, 2216, 2217, 2218, 2219, 2220, 2221, 2222, 2223, 2224, 2225, 2226, 2227, 2228, 2229, 2230, 2231, 2232, 2233, 2234, 2235, 2236, 2237, 2238, 2239, 2240, 2241, 2242, 2243, 2244, 2245, 2246, 2247, 2248, 2249, 2250, 2251, 2252, 2253, 2254, 2255, 2256, 2257, 2258, 2259, 2260, 2261, 2262, 2263, 2264, 2265, 2266, 2267, 2268, 2269, 2270, 2271, 2272, 2273, 2274, 2275, 2276, 2277, 2278, 2279, 2280, 2281, 2282, 2283, 2284, 2285, 2286, 2287, 2288, 2289, 2290, 2291, 2292, 2293, 2294, 2295, 2296, 2297, 2298, 2299, 2300, 2301, 2302, 2303, 2304, 2305, 2306, 2307, 2308, 2309, 2310, 2311, 2312, 2313, 2314, 2315, 2316, 2317, 2318, 2319, 2320, 2321, 2322, 2323, 2324, 2325, 2326, 2327, 2328, 2329, 2330, 2331, 2332, 2333, 2334, 2335, 2336, 2337, 2338, 2339, 2340, 2341, 2342, 2343, 2344, 2345, 2346, 2347, 2348, 2349, 2350, 2351, 2352, 2353, 2354, 2355, 2356, 2357, 2358, 2359, 2360, 2361, 2362, 2363, 2364, 2365, 2366, 2367, 2368, 2369, 2370, 2371, 2372, 2373, 2374, 2375, 2376, 2377, 2378, 2379, 2380, 2381, 2382, 2383, 2384, 2385, 2386, 2387, 2388, 2389, 2390, 2391, 2392, 2393, 2394, 2395, 2396, 2397, 2398, 2399, 2400, 2401, 2402, 2403, 2404, 2405, 2406, 2407, 2408, 2409, 2410, 2411, 2412, 2413, 2414, 2415, 2416, 2417, 2418, 2419, 2420, 2421, 2422, 2423, 2424, 2425, 2426, 2427, 2428, 2429, 2430, 2431, 2432, 2433, 2434, 2435, 2436, 2437, 2438, 2439, 2440, 2441, 2442, 2443, 2444, 2445, 2446, 2447, 2448, 2449, 2450, 2451, 2452, 2453, 2454, 2455, 2456, 2457, 2458, 2459, 2460, 2461, 2462, 2463, 2464, 2465, 2466, 2467, 2468, 2469, 2470, 2471, 2472, 2473, 2474, 2475, 2476, 2477, 2478, 2479, 2480, 2481, 2482, 2483, 2484, 2485, 2486, 2487, 2488, 2489, 2490, 2491, 2492, 2493, 2494, 2495, 2496, 2497, 2498, 2499, 2500, 2501, 2502, 2503, 2504, 2505, 2506, 2507, 2508, 2509, 2510, 2511, 2512, 2513, 2514, 2515, 2516, 2517, 2518, 2519, 2520, 2521, 2522, 2523, 2524, 2525, 2526, 2527, 2528, 2529, 2530, 2531, 2532, 2533, 2534, 2535, 2536, 2537, 2538, 2539, 2540, 2541, 2542, 2543, 2544, 2545, 2546, 2547, 2548, 2549, 2550, 2551, 2552, 2553, 2554, 2555, 2556, 2557, 2558, 2559, 2560, 2561, 2562, 2563, 2564, 2565, 2566, 2567, 2568, 2569, 2570, 2571, 2572, 2573, 2574, 2575, 2576, 2577, 2578, 2579, 2580, 2581, 2582, 2583, 2584, 2585, 2586, 2587, 2588, 2589, 2590, 2591, 2592, 2593, 2594, 2595, 2596, 2597, 2598, 2599, 2600, 2601, 2602, 2603, 2604, 2605, 2606, 2607, 2608, 2609, 2610, 2611, 2612, 2613, 2614, 2615, 2616, 2617, 2618, 2619, 2620, 2621, 2622, 2623, 2624, 2625, 2626, 2627, 2628, 2629, 2630, 2631, 2632, 2633, 2634, 2635, 2636, 2637, 2638, 2639, 2640, 2641, 2642, 2643, 2644, 2645, 2646, 2647, 2648, 2649, 2650, 2651, 2652, 2653, 2654, 2655, 2656, 2657, 2658, 2659, 2660, 2661, 2662, 2663, 2664, 2665, 2666, 2667, 2668, 2669, 2670, 2671, 2672, 2673, 2674, 2675, 2676, 2677, 2678, 26

- The values $\mathbf{e} = (e_1, e_2, \dots, e_n)^\top$ describe by how much \mathbf{y} differs from the linear relation $b_0 + b_1\mathbf{x}$.
- Estimate $\hat{\mathbf{e}}$, \hat{b}_0 , and \hat{b}_1 in a way that keeps $\hat{\mathbf{e}}$ as small as possible. We will minimise

$$\|\hat{\mathbf{e}}\|^2 = \sum_{i=1}^n \hat{e}_i^2 = \sum_{i=1}^n (y_i - \hat{b}_0 - \hat{b}_1 x_i)^2.$$

- This will make $\bar{\hat{\mathbf{e}}} = 0$ and minimise $var(\hat{\mathbf{e}})$.

Question

Because minimising $\sum_{i=1}^n \hat{e}_i^2$ leads to a least squares problem and this is "easy" to compute.

Since

$$\mathbf{e} = \mathbf{y} - \mathbf{X} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} \quad \text{with} \quad \mathbf{X} = [\mathbf{1} \ \mathbf{x}],$$

we are looking for the least squares solution to

$$\mathbf{y} = \mathbf{X} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$$

This can be found by projecting \mathbf{y} on an orthonormal basis for the span of the columns of \mathbf{X} .

Estimation

Least squares

The two vectors

$$\frac{\mathbf{1}}{\|\mathbf{1}\|}, \quad \frac{\tilde{\mathbf{x}}}{\|\tilde{\mathbf{x}}\|} = \frac{\mathbf{x} - \bar{x}\mathbf{1}}{\|\mathbf{x} - \bar{x}\mathbf{1}\|}$$

are orthonormal and have the same span as the columns of $\mathbf{X} = [\mathbf{1} \ \mathbf{x}]$. Let $\hat{\mathbf{y}}$ be the orthogonal projection onto this basis.

$$\begin{aligned} \hat{\mathbf{y}} &= \frac{\mathbf{y} \cdot (\mathbf{x} - \bar{x}\mathbf{1})}{\|\mathbf{x} - \bar{x}\mathbf{1}\|} \frac{\mathbf{x} - \bar{x}\mathbf{1}}{\|\mathbf{x} - \bar{x}\mathbf{1}\|} + \frac{\mathbf{y} \cdot \mathbf{1}}{\|\mathbf{1}\|} \frac{\mathbf{1}}{\|\mathbf{1}\|} \\ &= \frac{\mathbf{y} \cdot (\mathbf{x} - \bar{x}\mathbf{1})}{\|\mathbf{x} - \bar{x}\mathbf{1}\|^2} \mathbf{x} + \left(\frac{\mathbf{y} \cdot \mathbf{1}}{\|\mathbf{1}\|^2} - \bar{x} \frac{\mathbf{y} \cdot (\mathbf{x} - \bar{x}\mathbf{1})}{\|\mathbf{x} - \bar{x}\mathbf{1}\|^2} \right) \mathbf{1} \\ &= \underbrace{\frac{\mathbf{y} \cdot \tilde{\mathbf{x}}}{\|\tilde{\mathbf{x}}\|^2}}_{\hat{b}_1} \mathbf{x} + \underbrace{\left(\bar{y} - \bar{x} \frac{\mathbf{y} \cdot \tilde{\mathbf{x}}}{\|\tilde{\mathbf{x}}\|^2} \right)}_{\bar{y} - \hat{b}_1 \bar{x} = \hat{b}_0} \mathbf{1} \end{aligned}$$

Estimation

Result of least squares

$$\hat{\mathbf{y}} = \hat{b}_0 \mathbf{1} + \hat{b}_1 \mathbf{x} \quad \text{or}$$

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i \quad i = 1, 2, \dots, n$$

$$\mathbf{y} = \hat{b}_0 \mathbf{1} + \hat{b}_1 \mathbf{x} + \hat{\mathbf{e}} \quad \text{or}$$

$$y_i = \hat{b}_0 + \hat{b}_1 x_i + \hat{e}_i \quad i = 1, 2, \dots, n$$

with

$$\bullet \hat{b}_1 = \frac{\mathbf{y} \cdot \tilde{\mathbf{x}}}{\|\tilde{\mathbf{x}}\|^2} = \frac{\tilde{\mathbf{y}} \cdot \tilde{\mathbf{x}}}{\|\tilde{\mathbf{x}}\|^2} = \text{corr}(\mathbf{x}, \mathbf{y}) \frac{sd(\mathbf{y})}{sd(\mathbf{x})}$$

$$\bullet \hat{b}_0 = \bar{\mathbf{y}} - \hat{b}_1 \bar{\mathbf{x}}$$

$$\bullet \hat{e}_i = y_i - \hat{y}_i$$

$$\bullet \hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}}$$

A scatter plot illustrating linear regression. The x-axis is labeled x and ranges from 0 to 5. The y-axis is labeled y and ranges from 4 to 10. A solid blue line represents the fitted regression line. The equation for the line is shown as $\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$. Five data points are plotted as red dots. For each point, a vertical green line segment connects the point to the regression line, representing the residual. One point is specifically labeled (x_i, y_i) in red, and its corresponding residual is labeled \hat{e}_i in green.

- \hat{b}_0, \hat{b}_1 - ordinary least squares (OLS) estimators
- \hat{y}_i - fitted values
- $\hat{\mathbf{y}}$ - vector of fitted values
- \hat{e}_i - residuals
- $\hat{\mathbf{e}}$ - vector of residuals
- $\hat{y} = \hat{b}_0 + \hat{b}_1 x$, with $x \in \mathbb{R}$
- regression line.

The regression coefficients have a straight forward interpretation.

- For every increase of x by one unit a change of y by \hat{b}_1 units can be expected on average.
- If $x = 0$, then on average $y = \hat{b}_0$.

Estimation

Properties

- $\widehat{\mathbf{e}} = 0$
- $\widehat{\mathbf{y}} = \bar{\mathbf{y}}$
- $\text{cov}(\mathbf{x}, \widehat{\mathbf{e}}) = 0$, i.e. $\widetilde{\mathbf{x}} \perp \widehat{\mathbf{e}}$
- $\text{cov}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = 0$, i.e. $\widetilde{\widehat{\mathbf{y}}} \perp \widehat{\mathbf{e}}$

Hence, the residuals of the OLS estimator ...

- contain no information that could be explained by a linear function of \mathbf{x} .
- are on average 0.
- have the smallest variance possible.

Estimation

Example Temperature and pressure

$$\tilde{\mathbf{x}} = (-25, -15, -5, 5, 15, 25)^\top$$

$$\mathbf{y} = (91, 95, 100, 101, 107, 112)^\top$$

$$\bar{\mathbf{y}} = 101 \quad \text{and} \quad \bar{\mathbf{x}} = 25$$

$$\begin{aligned} \mathbf{y} \cdot \tilde{\mathbf{x}} = \tilde{\mathbf{y}} \cdot \tilde{\mathbf{x}} &= \left((-25)91 + (-15)95 + (-5)100 \right. \\ &\quad \left. + (5)101 + (15)107 + (25)112 \right) \\ &= 710 \end{aligned}$$

$$\|\tilde{\mathbf{x}}\|^2 = 1750$$

$$\hat{b}_1 = \frac{\mathbf{y} \cdot \tilde{\mathbf{x}}}{\|\tilde{\mathbf{x}}\|^2} = \frac{710}{1750} \approx 0.406$$

$$\hat{b}_0 = \bar{\mathbf{y}} - \hat{b}_1 \bar{\mathbf{x}} = 101 - \frac{710}{1750} 25 \approx 90.9.$$

Hence, $\hat{y} = 90.9 + 0.406x$.

Estimation

Example

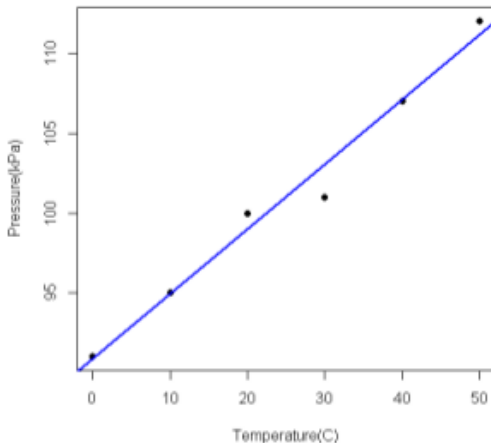


Figure 3: Running example: pressure versus temperature in a boiler.

Estimation

Example

Fitted values

| Temp (°C) x_i | Pressure (kPa) y_i | $\hat{b}_0 + \hat{b}_1 x_i$ |
|-----------------|----------------------|-----------------------------|
| 0 | 91 | 90.9 |
| 10 | 95 | 94.9 |
| 20 | 100 | 99.0 |
| 30 | 101 | 103 |
| 40 | 107 | 107 |
| 50 | 112 | 111 |

Table 1: Temperature, pressure, and predicted pressure

Interpretation

- The regression line is $\hat{y} = 90.9 + 0.406x$.
- For every increase (decrease) of temperature by one degree Celsius an increase (decrease) of pressure by $b_1 = 0.406$ kPa can be expected.
- The pressure that we can expect at 0°C is $b_0 = 90.9$ kPa.

2.3 Goodness-of-fit

Notation

- **Total sum of squares (TSS)**

$$TSS = \|\tilde{\mathbf{y}}\|^2 = \sum_{i=1}^n (y_i - \bar{y})^2$$

This quantifies all variation in the sample \mathbf{y} .

- **Explained sum of squares (ESS)**

$$ESS = \|\hat{\tilde{\mathbf{y}}}\|^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

The variation in \mathbf{y} that is explained by the regression line.

- **Residual sum of squares (RSS)**

$$RSS = \|\hat{\mathbf{e}}\|^2 = \sum_{i=1}^n \hat{e}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

The variation in \mathbf{y} that is not explained by the regression line.

Goodness-of-fit

We want to measure how good the regression line fits the data by taking the ratio of explained and total sum of squares.

A large value indicates good fit, a small value indicates a less good fit.

Definition The **R-squared** of the regression is

$$R^2 = \frac{ESS}{TSS}$$

Remember $\tilde{\mathbf{y}} \perp \hat{\mathbf{e}}$

Pythagorean theorem

$$\|\mathbf{y} - \bar{y}\mathbf{1}\|^2 = \|\bar{y}\mathbf{1} - \hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{y}} - \mathbf{y}\|^2, \quad \text{i.e.} \quad \|\tilde{\mathbf{y}}\|^2 = \|\hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{e}}\|^2$$

Hence, $TSS = ESS + RSS$ and

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}.$$

Goodness-of-fit

Properties of the R-squared

- $0 \leq R^2 \leq 1$
- If $R^2 = 1$, the fit is perfect. All data points are on the regression line and all residuals equal 0.
- If $R^2 = 0$, the regression line does not fit the data and is useless.
- If R^2 is close to 1, it is very likely that the regression found an actual relation in the data and will probably work well for prediction.
- If R^2 is small, it is unclear whether we found an actual relation or not. Prediction will very likely not work well.

2.4 Confidence intervals

Remember

We assume the regression equation with the true b_0 and b_1 for the random variables X and Y

$$Y = b_0 + b_1X + E.$$

With samples \mathbf{x} and \mathbf{y} we estimate \hat{b}_0 and \hat{b}_1 such that

$$\mathbf{y} = \hat{b}_0\mathbf{1} + \hat{b}_1\mathbf{x} + \hat{\mathbf{e}}.$$

Estimation errors

- Every time we collect data for the random variables X and Y , we will get different data vectors \mathbf{x} and \mathbf{y} and different estimates \hat{b}_0 and \hat{b}_1 .
- The estimators \hat{b}_0 and \hat{b}_1 can be understood as random variables.
- The estimation errors $\hat{b}_0 - b_0$ and $\hat{b}_1 - b_1$ will be random.
- How much can we trust \hat{b}_0 and \hat{b}_1 ?

Confidence intervals

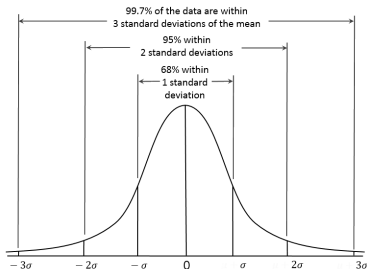
We will answer this question only for b_1 and only under the following assumption.

Assumption

The relation $Y = b_0 + b_1X + E$ holds with $E \sim \mathcal{N}(0, \sigma^2)$, i.e. normally distributed with mean 0 and variance σ^2 .

Interpretation

- When Y is on average equal to $b_0 + b_1X$, E will have mean 0.
- E has the probability density $f_E(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$



68% chance that $E \in [-\sigma, \sigma]$

95% chance that
 $E \in [-1.96\sigma, 1.96\sigma]$

99.7% chance that $E \in [-3\sigma, 3\sigma]$

Confidence intervals

Definition

A $100(1 - \alpha)\%$ **confidence interval** for b_1 is an interval $[c, d]$ constructed from the data vectors \mathbf{x} , \mathbf{y} such that the chance for b_1 to be in $[c, d]$ is $100(1 - \alpha)\%$.

In other words

If we collect a lot of different samples and construct a confidence interval for each sample

$$\mathbf{x}_1, \mathbf{y}_1 \rightsquigarrow [c_1, d_1]$$

$$\mathbf{x}_2, \mathbf{y}_2 \rightsquigarrow [c_2, d_2]$$

$$\vdots$$

then $b_1 \in [c_j, d_j]$ can be expected in $100(1 - \alpha)\%$ of the cases.

Remark

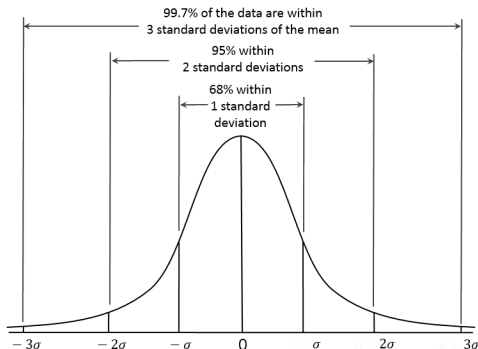
Typically, we choose $\alpha = 0.05$ or $\alpha = 0.1$ and construct 95% or a 90% confidence interval respectively.

Confidence intervals

Example

Let \hat{u} be an estimator for some value u with $\hat{u} \sim \mathcal{N}(u, \sigma^2)$.

- Hence, with a 95% chance $|\hat{u} - u| \leq 1.96\sigma$.
- Thus, with a 95% chance $u \in [\hat{u} - 1.96\sigma, \hat{u} + 1.96\sigma]$.
- $[\hat{u} - 1.96\sigma, \hat{u} + 1.96\sigma]$ is a 95% confidence interval.



Confidence intervals

t-distribution

Confidence intervals for b_1 work similarly, except that we cannot work with the normal distribution.

- Instead we have to work with the so called ***t*-distribution** with $n - 2$ **degrees of freedom**.
- Replace 1.96 by the so called **critical value** $t_{1-\frac{\alpha}{2}}(n-2)$.
- The critical value can be looked up in a table or use the Matlab command `tinv(1- α /2,n-2)`.

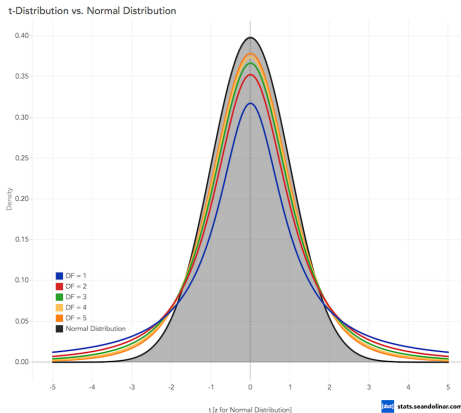
Theorem With probability $100(1 - \alpha)\%$

$$b_1 \in \left[\hat{b}_1 - t_{1-\frac{\alpha}{2}}(n-2)se(\hat{b}_1), \hat{b}_1 + t_{1-\frac{\alpha}{2}}(n-2)se(\hat{b}_1) \right]$$

where $se(\hat{b}_1) = \sqrt{\frac{1}{n-2} \frac{RSS}{\|\tilde{\mathbf{x}}\|^2}}$ is the so called **standard error**.

Confidence intervals

t -distribution



- It is similar to the normal distribution but has heavier tails.
- Often called **Student's t -distribution**.
- First published 1908 by William Sealy Gosset under the pseudonym "Student" while working for the Guinness Brewery.

Confidence intervals

t-distribution

| cum. prob
one-tail | $t_{.50}$ | $t_{.25}$ | $t_{.20}$ | $t_{.15}$ | $t_{.10}$ | $t_{.05}$ | $t_{.025}$ | $t_{.01}$ | $t_{.005}$ | $t_{.001}$ | $t_{.0005}$ |
|-----------------------|------------------|-----------|-----------|-----------|-----------|-----------|------------|-----------|------------|------------|-------------|
| two-tails | 1.00 | 0.50 | 0.40 | 0.30 | 0.20 | 0.10 | 0.05 | 0.02 | 0.01 | 0.002 | 0.001 |
| df | | | | | | | | | | | |
| 1 | 0.000 | 1.000 | 1.376 | 1.963 | 3.078 | 6.314 | 12.71 | 31.82 | 63.66 | 318.31 | 636.62 |
| 2 | 0.000 | 0.816 | 1.061 | 1.386 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 22.327 | 31.599 |
| 3 | 0.000 | 0.765 | 0.978 | 1.250 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 10.215 | 12.924 |
| 4 | 0.000 | 0.741 | 0.941 | 1.190 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 7.173 | 8.610 |
| 5 | 0.000 | 0.727 | 0.920 | 1.156 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 5.893 | 6.869 |
| 6 | 0.000 | 0.718 | 0.906 | 1.134 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 5.208 | 5.959 |
| 7 | 0.000 | 0.711 | 0.896 | 1.119 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 4.785 | 5.408 |
| 8 | 0.000 | 0.706 | 0.889 | 1.108 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 4.501 | 5.041 |
| 9 | 0.000 | 0.703 | 0.883 | 1.100 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 4.297 | 4.781 |
| 10 | 0.000 | 0.700 | 0.879 | 1.093 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 4.144 | 4.587 |
| 11 | 0.000 | 0.697 | 0.876 | 1.088 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 4.025 | 4.437 |
| 12 | 0.000 | 0.695 | 0.873 | 1.083 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 3.930 | 4.318 |
| 13 | 0.000 | 0.694 | 0.870 | 1.079 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 3.852 | 4.221 |
| 14 | 0.000 | 0.692 | 0.868 | 1.076 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 3.787 | 4.140 |
| 15 | 0.000 | 0.691 | 0.866 | 1.074 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 3.733 | 4.073 |
| 16 | 0.000 | 0.690 | 0.865 | 1.071 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 3.686 | 4.015 |
| 17 | 0.000 | 0.689 | 0.863 | 1.069 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 3.646 | 3.965 |
| 18 | 0.000 | 0.688 | 0.862 | 1.067 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 3.610 | 3.922 |
| 19 | 0.000 | 0.688 | 0.861 | 1.066 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.579 | 3.883 |
| 20 | 0.000 | 0.687 | 0.860 | 1.064 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.552 | 3.850 |
| 21 | 0.000 | 0.686 | 0.859 | 1.063 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 3.527 | 3.819 |
| 22 | 0.000 | 0.686 | 0.858 | 1.061 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 3.505 | 3.792 |
| 23 | 0.000 | 0.685 | 0.858 | 1.060 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 | 3.485 | 3.768 |
| 24 | 0.000 | 0.685 | 0.857 | 1.059 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 | 3.467 | 3.745 |
| 25 | 0.000 | 0.684 | 0.856 | 1.058 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 | 3.450 | 3.725 |
| 26 | 0.000 | 0.684 | 0.856 | 1.058 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 | 3.435 | 3.707 |
| 27 | 0.000 | 0.684 | 0.855 | 1.057 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 | 3.421 | 3.690 |
| 28 | 0.000 | 0.683 | 0.855 | 1.056 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 | 3.408 | 3.674 |
| 29 | 0.000 | 0.683 | 0.854 | 1.055 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 | 3.396 | 3.659 |
| 30 | 0.000 | 0.683 | 0.854 | 1.055 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 | 3.385 | 3.646 |
| 40 | 0.000 | 0.681 | 0.851 | 1.050 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 | 3.307 | 3.551 |
| 60 | 0.000 | 0.679 | 0.848 | 1.045 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 | 3.232 | 3.460 |
| 80 | 0.000 | 0.678 | 0.846 | 1.043 | 1.292 | 1.664 | 1.990 | 2.374 | 2.639 | 3.195 | 3.416 |
| 100 | 0.000 | 0.677 | 0.845 | 1.042 | 1.290 | 1.660 | 1.984 | 2.364 | 2.626 | 3.174 | 3.390 |
| 1000 | 0.000 | 0.675 | 0.842 | 1.037 | 1.282 | 1.646 | 1.962 | 2.330 | 2.581 | 3.098 | 3.300 |
| Z | 0.000 | 0.674 | 0.842 | 1.036 | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 | 3.090 | 3.291 |
| | 0% | 50% | 60% | 70% | 80% | 90% | 95% | 98% | 99% | 99.8% | 99.9% |
| | Confidence Level | | | | | | | | | | |

Notice

Rows - degree of freedoms ($n - 2$)

Columns - confidence level

Look for $t_{1-\frac{\alpha}{2}}$ in the first row.

Or look for α in the row “two-tails”.

Or look for the confidence % in the last row.

Confidence intervals

Interpretation

- A small confidence interval indicates that the estimator \hat{b}_1 is quite precise.
- A wide confidence interval indicates that the estimator \hat{b}_1 is less reliable.
- If 0 is not in the confidence interval, we have strong evidence that $b_1 \neq 0$ and that X has an impact on Y .
We say b_1 is **significant**.
- If we collect more data and increase the sample size n , the confidence interval becomes tighter.

Remember

$$b_1 \in \left[\hat{b}_1 - t_{1-\frac{\alpha}{2}}(n-2)se(\hat{b}_1), \hat{b}_1 + t_{1-\frac{\alpha}{2}}(n-2)se(\hat{b}_1) \right]$$

$$\text{with } se(\hat{b}_1) = \sqrt{\frac{1}{n-2} \frac{RSS}{\|\tilde{\mathbf{x}}\|^2}}$$

Confidence intervals

Example 95% confidence interval for temperature and pressure

$$\hat{b}_0 = 90.9, \hat{b}_1 = 0.406$$

$$\hat{y} = 90.9 + 0.406x$$

$$\mathbf{y} = (91, 95, 100, 101, 107, 112)^\top$$

$$\hat{\mathbf{e}} =$$

$$\begin{aligned} & (91 - 90.9, 95 - 95, 100 - 99, 101 - 103.1, 107 - 107.1, 112 - 111.2)^\top \\ &= (0.1, 0, 1, -2.1, -0.1, 0.8)^\top \end{aligned}$$

$$RSS = 0.1^2 + 0^2 + 1^2 + 2.1^2 + 0.1^2 + 0.8^2 = 6.07$$

Hence, the standard error is

$$se(\hat{b}_1) = \sqrt{\frac{1}{6-2} \frac{6.07}{1750}} \approx 0.029.$$

Confidence intervals

Example 95% confidence interval for temperature and pressure

We have $n - 2 = 6 - 2 = 4$ degrees of freedom and

$$1 - \frac{\alpha}{2} = 1 - \frac{0.05}{2} = 0.975.$$

$$t_{0.975}(4) \approx 2.78$$

$$t_{0.975}(4)se(\hat{b}_1) \approx 2.78 \times 0.029 \approx 0.081$$

This gives the 95% confidence interval for b_1

$$[0.406 - 0.081, 0.406 + 0.081] = [0.325, 0.487].$$

We can see that b_1 is significant.

2.5 t-test

Statisticians like to describe the statement that b_1 is significant in a different way. It is common to derive this statement from a statistical **hypothesis test**. Let us understand statistical testing step by step.

Hypothesis

- A statistical test is a tool that gives evidence **against** a hypothesis, which is called the **null hypothesis** H_0 .
- If you want evidence that b_1 is significant, you need to start with the opposite statement.

$$H_0 : b_1 = 0$$

- Sometime the **alternative hypothesis** H_1 is introduced as well. Although, it is not necessary to state it explicitly because it is always the opposite of H_0 .

$$H_1 : b_1 \neq 0$$

***t*-test**

Test statistic

- We want to find evidence against H_0 . However, we assume for a moment that H_0 is true.
- The next step is to find an estimator of which we know the distribution whenever H_0 is true. We know that

$$\frac{\hat{b}_1}{se(\hat{b}_1)} \sim t \quad \text{with } n - 2 \text{ degrees of freedom}$$

when $b_1 = 0$. This estimator is called **test statistic**.

- If $\hat{b}_1/se(\hat{b}_1)$ takes a value that is very unlikely for the t -distribution, we have evidence that H_0 was probably not true in the first place.

t-test

Critical values

- Critical values have the property that there is only a $100\alpha\%$ chance that

$$\hat{b}_1/se(\hat{b}_1) < -t_{1-\frac{\alpha}{2}}(n-2) \quad \text{or} \quad t_{1-\frac{\alpha}{2}}(n-2) < \hat{b}_1/se(\hat{b}_1)$$

when $\frac{\hat{b}_1}{se(\hat{b}_1)} \sim t$ with $n-2$ degrees of freedom.

- Hence, when

$$|\hat{b}_1/se(\hat{b}_1)| > t_{1-\frac{\alpha}{2}}(n-2)$$

we **reject** the null-hypothesis on the $100(1-\alpha)\%$ **confidence level**.

- If

$$|\hat{b}_1/se(\hat{b}_1)| \leq t_{1-\frac{\alpha}{2}}(n-2),$$

we **fail to reject** the null-hypothesis. We have neither evidence for nor against the null-hypothesis!

Note

- The test described above is called **two sided t -test**.
- The test rejects H_0 when 0 is not in the $100(1 - \alpha)\%$ confidence interval and fails to reject when 0 is in the interval.
- We cannot rule out a null-hypothesis with 100% certainty. We have to live with the fact that we can only reject on a certain confidence level, e.g. 90% or 95%.
- α is called the **significance level** of the test.

One sided t -test

- _____

0 1 = 1 1 1

- 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99

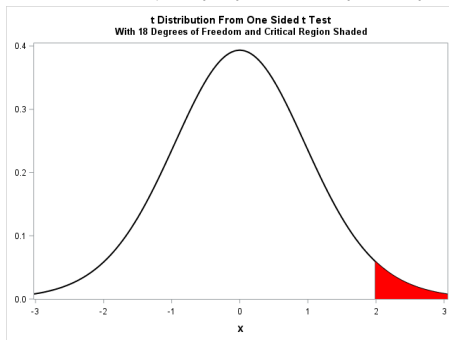
1. *Journal of the American Medical Association*, 1997; 277: 1001-1005.

- Fail to reject if $-t_{1-\alpha}(n-2) \leq \hat{b}_1/se(\hat{b}_1)$.

One sided t -test

- $$H_0 : b_1 \leq 0.$$

- Fail to reject if $\hat{b}_1/se(\hat{b}_1) \leq t_{1-\alpha}(n-2)$.



t-test

Example Temperature and pressure $\hat{y} = 90.9 + 0.406x$

Can we confirm that b_1 is positive on the 99% confidence level?

$$H_0 : b_1 \leq 0$$

The critical value is $t_{0.99}(4) = 3.747$.

Compute the test statistic:

$$\begin{aligned}\hat{b}_1 &= 0.406 \\ se(\hat{b}_1) &\approx 0.029 \\ \frac{\hat{b}_1}{se(\hat{b}_1)} &\approx \frac{0.406}{0.029} \approx 14.\end{aligned}$$

Since $14 > 3.747$ the null hypothesis can be rejected. We have strong evidence that $b_1 > 0$.

2.6 Prediction intervals

In many applications \hat{b}_0 and \hat{b}_1 are estimated and used to predict Y for a new value x^* of X by the corresponding fitted value \hat{y}^*

$$\hat{y}^* = \hat{b}_0 + \hat{b}_1 x^*.$$

How precise is this prediction compared to the actual value y^* ?

Theorem

A $100(1 - \alpha)\%$ confidence interval for y^* is $[\hat{y}^* - \tau, \hat{y}^* + \tau]$ with

$$\tau = t_{1-\frac{\alpha}{2}}(n-2) \sqrt{\frac{RSS}{n-2}} \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\|\tilde{\mathbf{x}}\|^2}}.$$

Notation

This interval is called **prediction interval**.

$$\hat{y}^* \pm t_{1-\frac{\alpha}{2}}(n-2) \sqrt{\frac{RSS}{n-2}} \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\|\tilde{\mathbf{x}}\|^2}}$$

100

Prediction intervals are

- wide for small n and if x^* is far away from the mean \bar{x} .
- small for large n and if x^* is close to \bar{x} .

Interpolation works much better than extrapolation. Be careful if x^* is outside the range of the data \mathbf{x} .

Prediction intervals

$$\hat{y}^* \pm t_{1-\frac{\alpha}{2}}(n-2) \sqrt{\frac{RSS}{n-2}} \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\|\tilde{\mathbf{x}}\|^2}}$$

Example Temperature and pressure $\hat{y} = 90.9 + 0.406x$

Let $x^* = 24^\circ\text{C}$. What is a 95% prediction interval?

The critical value is $t_{0.975}(4) = 2.78$. The prediction interval is given by

$$\begin{aligned} 90.9 + 0.406 \times 24 \pm 2.78 \sqrt{\frac{6.07}{4}} \sqrt{1 + \frac{1}{6} + \frac{(24 - 25)^2}{1750}} \\ \approx 100.84 \pm 3.77 \end{aligned}$$

or equivalently $[97.14, 104.54]$.