Linear Regression-3

Prof. Asim Tewari
IIT Bombay

Simple Linear Regression

It assumes that there is approximately a linear relationship between *X* and *Y*

$$Y \approx \beta_0 + \beta_1 X \quad \text{or} \quad Y = \beta_0 + \beta_1 X + \epsilon.$$

β0 and β1 are intercept slope known as the model coefficients or parameters

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

Hat symbol, ˆ , to denote the estimated value for an unknown parameter or coefficient

Simple Linear Regression

Estimating the Coefficients

- Least squares approach

The least squares approach chooses parameters to minimize the *residual sum of squares* (RSS)

$e_i = y_i - \hat{y}_i$ represents i_{th} residual

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

Simple Linear Regression

Estimating the Coefficients

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$

Simple Linear Regression

$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ / $Var(\epsilon_i) = \sigma^2$

Assessing the Accuracy of the Coefficient Estimates

Standard Errors associated with coefficients

$$\left(SE(\hat{\beta}_1)\right)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$Var(\hat{\beta}_1)$

$$\left(SE(\hat{\beta}_0)\right)^2 = \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$Var(\hat{\beta}_0)$

95% confidence interval associated with coefficients

$$\beta_0 = \hat{\beta}_0 \pm 1.96\, SE(\hat{\beta}_0) \; ; \; \beta_1 = \hat{\beta}_1 \pm 1.96\, SE(\hat{\beta}_1)$$

CLT with 95% C.I

Simple Linear Regression

$$\boxed{y = f(x, \beta_0, \beta_1, \ldots \beta_j) + \epsilon}$$

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

- zero mean
- σ^2 var
- Uncorrelated

Population

X	$x_1, x_2 \ldots \ldots$	$- \cdot \infty$
Y	$y_1, y_2 \ldots \ldots$	$- \cdot\cdot \infty$

Sample is of size n

$$\left.\begin{matrix} x_1, x_2 \ldots\ldots x_n \\ y_1, y_2 \ldots\ldots y_n \end{matrix}\right\} \Rightarrow \text{Find}$$

β_0, β_1

$\hat{\beta}_0, \hat{\beta}_1$

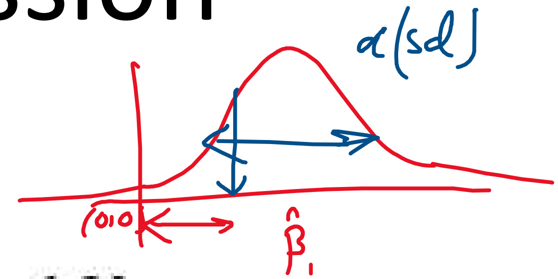
Assessing the accuracy of fit

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

Q1. How close is $\hat{\beta}_0$ to β_0 and $\hat{\beta}_1$ to β_1

Simple Linear Regression

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



Hypothesis tests on the coefficients

$$H_0 : \text{There is no relationship between } X \text{ and } Y$$

versus the *alternative hypothesis*

$$H_a : \text{There is some relationship between } X \text{ and } Y$$

Mathematically, this corresponds to testing

$$H_0 : \beta_1 = 0 \qquad \text{versus} \qquad H_a : \beta_1 \neq 0$$

For this we calculate t statistics which measures the number of standard deviations that $\hat{\beta}_1$ is away from 0.

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)},$$

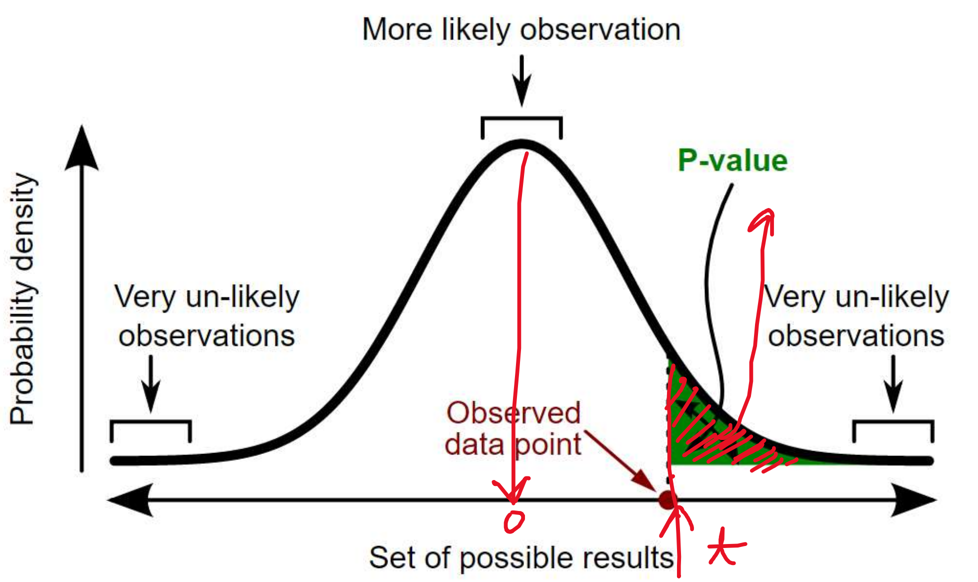
Simple Linear Regression

P-Value is the probability of observing any value equal to |t| or larger for a t-distribution with n−2 degrees of freedom

Sample of size n → $\hat{\beta}_0$, $\hat{\beta}_1$, $SE(\hat{\beta}_1)$

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)},$$ t

t



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

Simple Linear Regression

Pr (observation | hypothesis) ≠ Pr (hypothesis | observation)

The probability of observing a result given that some hypothesis is true is *not equivalent* to the probability that a hypothesis is true given that some result has been observed.

Sample of size n → $\hat{\beta}_0, \hat{\beta}_1$ → $t = \dfrac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$ → compare with std. t-dist of $n-2$ degrees of freedom

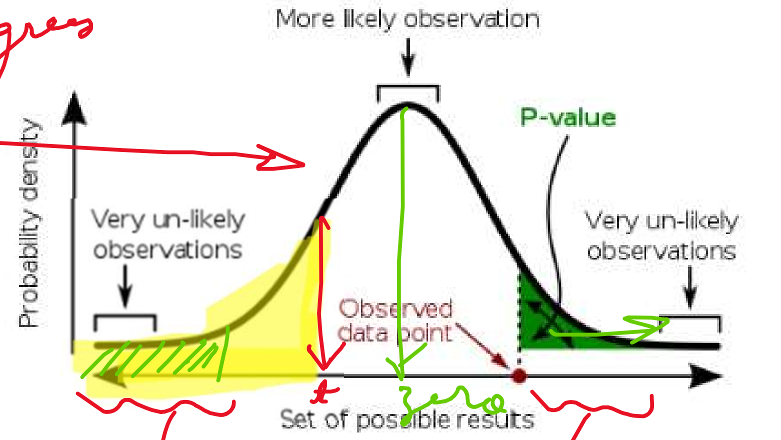
Simple Linear Regression

$$H_0 : \beta_1 = 0 \qquad\qquad H_a : \beta_1 \neq 0$$

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)},$$

$\nu = n - 2$ degrees

t-dist



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

p-value is defined as

- $\Pr(T \geq t|H)$ for a one-sided (right tail) test,
- $\Pr(T \leq t|H)$ for a one-sided (left tail) test,
- $2\min\{\Pr(T \leq t|H), \Pr(T \geq t|H)\}$ for a two-sided test,

Notice that just by replacing T by $-T$ one converts a test based on extremely large values to a test based on extremely small values; a by replacing T by $|T|$ one gets a test with *p*-value

- $\Pr(T \leq -|t||H) + \Pr(T \geq +|t||H)\}$.

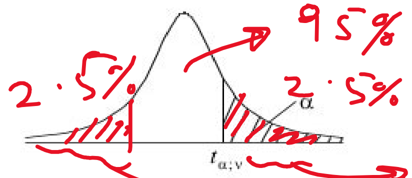
Simple Linear Regression

- The p-value represents the chance your results could be random (i.e. happened by chance).

- So a small p-value means that there is a small chance that your results are random. Thus, they are not random. So we can infer that there is an association between the predictor and the response (i.e we *reject the null hypothesis)*

Table of the Student's *t*-distribution

The table gives the values of $t_{\alpha;\nu}$ where $\Pr(T_\nu > t_{\alpha;\nu}) = \alpha$, with ν degrees of freedom



5%

2 5%

ν \ α	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	6.314	12.076	31.821	63.657	318.310	636.620
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

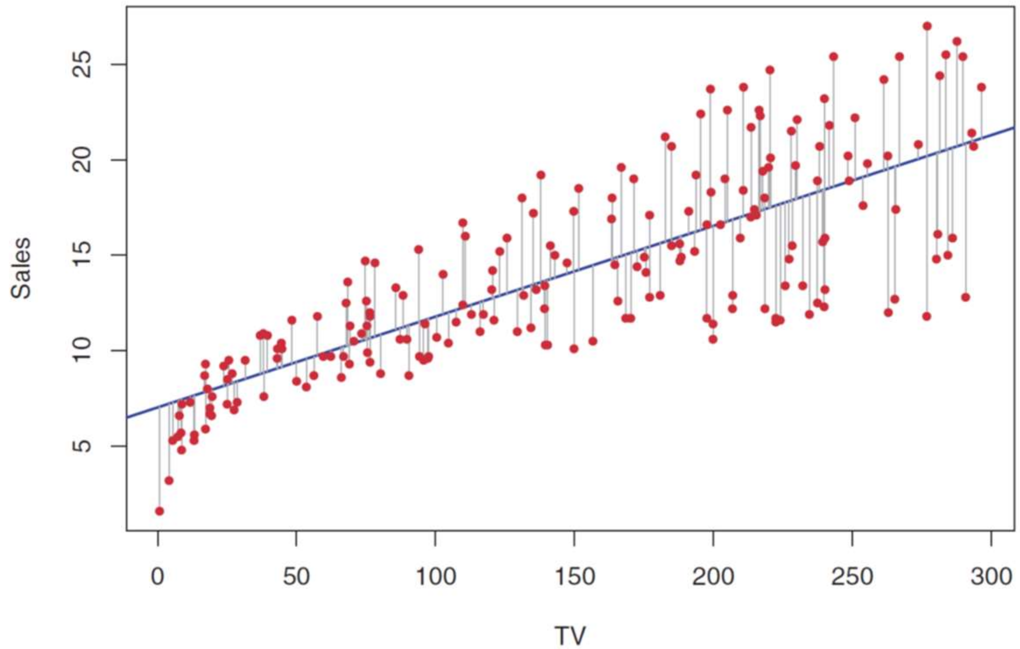
$22 - 2 = 20$

$20 \to \infty$

$\pm 2.086(SE)$

$\boxed{\pm 1.96(SE)}$

1.96



For the Advertising *data, the least squares fit for the regression of* sales *onto* TV *is shown. The fit is found by minimizing the sum of squared errors. Each grey line segment represents an error, and the fit makes a compromise by averaging their squares. In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.*

P-value

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

For the Advertising data, coefficients of the least squares model for the regression of number of units sold on TV advertising budget. An increase of $1,000 *in the TV advertising budget is associated with an increase in sales by around 50 units (Recall that the sales variable is in thousands of units, and the TV variable is in thousands of dollars).*

Simple Linear Regression

Assessing the Accuracy of the Model

Residual Standard Error (RSE)

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

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

R^2 Statistic: The RSE provides an absolute measure, R^2 provides a relative measure

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \quad \text{where } \mathrm{TSS} = \sum (y_i - \bar{y})^2$$

$$R = \mathrm{Cor}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$