Linear Regression-2

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 <u>*residual sum of squares*</u> *(*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

Assessing the Accuracy of the Coefficient Estimates

Standard Errors associated with coefficients

95% confidence interval associated with coefficients

Simple Linear Regression

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

Assessing the Accuracy of the Coefficient Estimates

Standard Errors associated with coefficients

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

Where $\sigma^2 = \mathrm{var}(\epsilon)$ and it is assumed that ϵ_i are uncorrelated and has same variance σ^2

95% confidence interval associated with coefficients

$$\hat{\beta}_0 \pm 1.96\, SE(\hat{\beta}_0) \qquad \hat{\beta}_1 \pm 1.96\, SE(\hat{\beta}_1)$$

Simple Linear Regression

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

Assessing the Accuracy of the Coefficient Estimates

Standard Errors associated with coefficients

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

Where $\sigma^2 = var(\epsilon)$ and it is assumed that ϵ_i are uncorrelated and has same variance σ^2

95% confidence interval associated with coefficients

$\hat{\beta}_0 \pm 1.96\, SE(\hat{\beta}_0)$ $\qquad$ $\hat{\beta}_1 \pm 1.96\, SE(\hat{\beta}_1)$

Simple Linear Regression

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

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

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

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

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

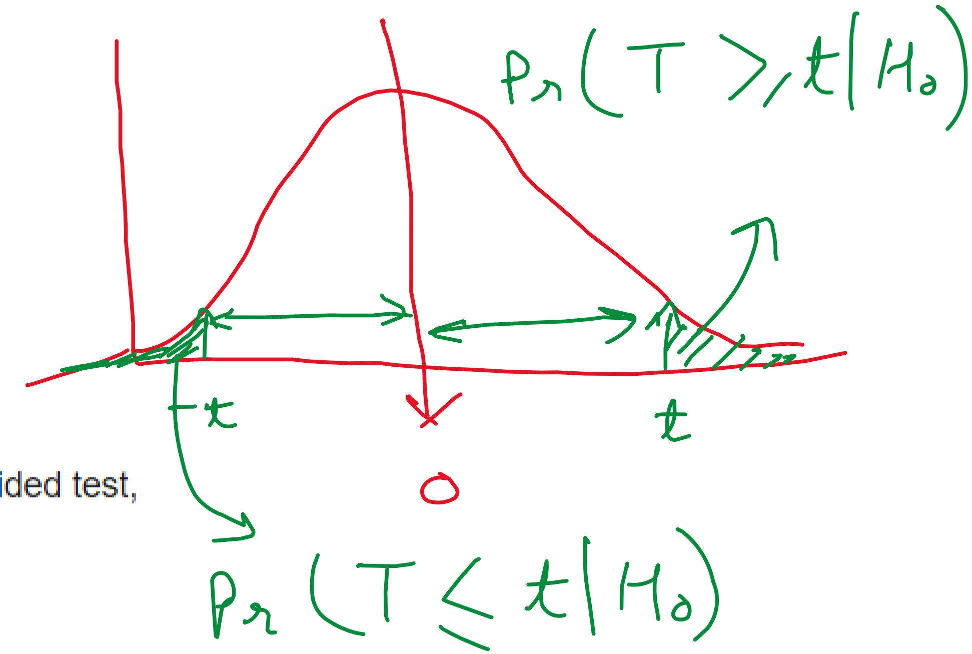
p-value is defined as

- $\Pr(T \geq t|H)$ 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,

Simple Linear Regression

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

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



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,

95% Confidence

$\alpha = 1 - C = 0.05$

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.

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,

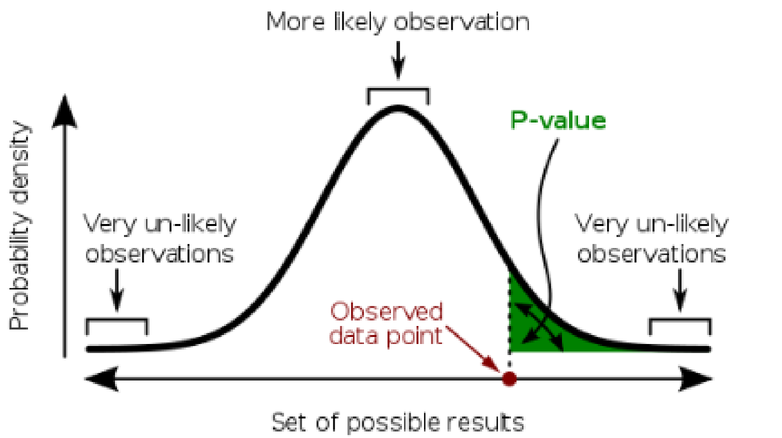
Notice that just by replacing T by $-T$ one converts a test based on extremely large values to a test based on extremely small values; ar by replacing T by $|T|$ one gets a test with *p*-value

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

Simple Linear Regression

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,



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

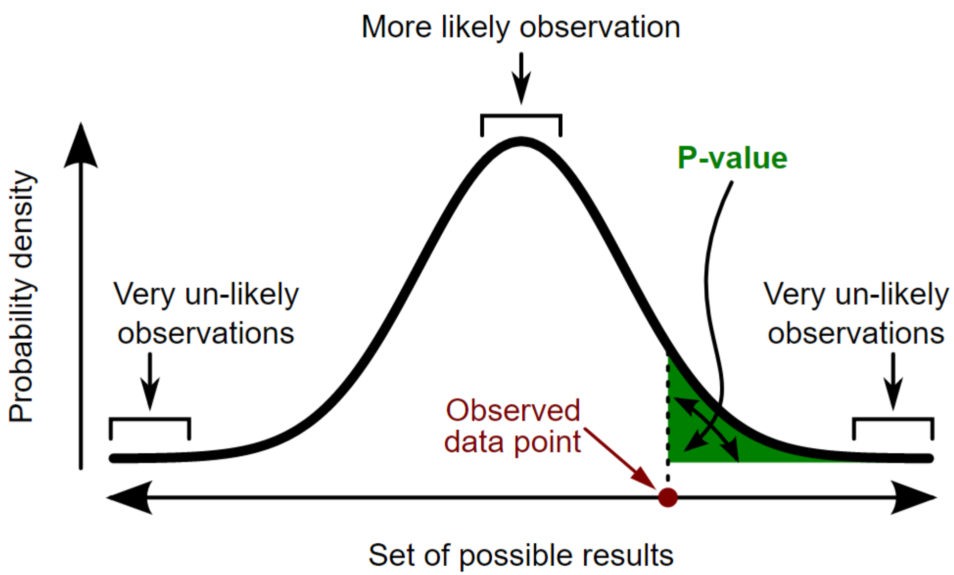
Notice that just by replacing T by $-T$ one converts a test based on extremely large values to a test based on extremely small values; and by replacing T by $|T|$ one gets a test with *p*-value

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

Simple Linear Regression

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

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



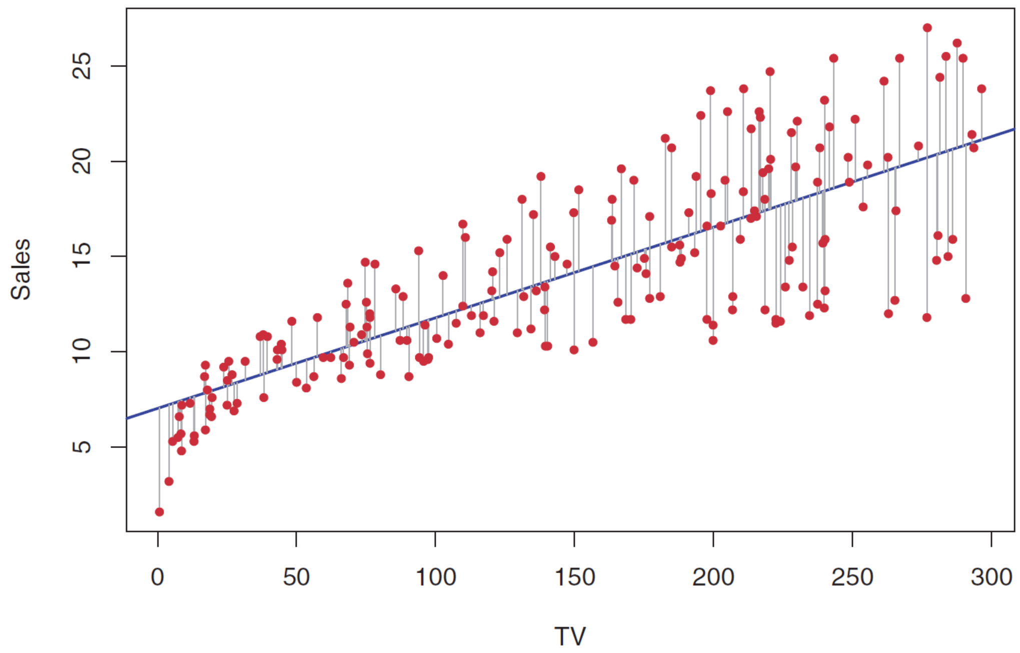
A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

For a p-dimensional X vector
t-dist will be of
$n-1-p$ dof

Simple Linear Regression

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

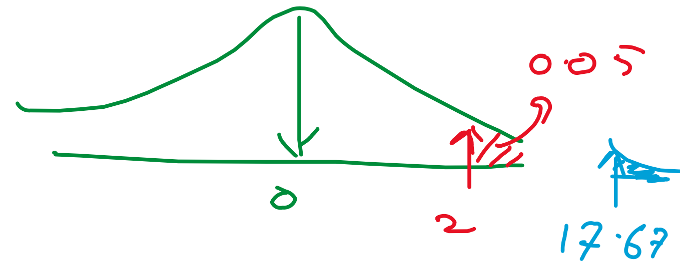
- So a small p-value means that there is a small chance that your results are random. Thus, they are not random. So we can infer that there is an association between the predictor and the response (i.e we *reject the null hypothesis)*



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.*

C	0.95	0.99
α	0.05	0.01
t^*	2	2.75

P-value



$\sqrt{\operatorname{var} \hat{\beta}_0}$

$\approx Pr(T \geq |t| \mid H_0)$

	Coefficient	Std. error	t-statistic	p-value
Intercept → $\hat{\beta}_0$	7.0325	0.4578	15.36	< 0.0001
TV → $\hat{\beta}_1$	0.0475	0.0027	17.67	< 0.0001

$\sqrt{\operatorname{var} \hat{\beta}_1}$

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).

SE of a mean of a RV

Standard error $= \sqrt{Var(\mu)}$

$= \dfrac{\sigma}{\sqrt{n}}$

$SE(\hat{\beta}_0) = ?$ $SE(\hat{\beta}_1) = ?$

SE of a mean of a RV

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

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

$$\sigma^2 = Var(\epsilon) = \text{Residual Std error}^2 = (RSE)^2$$

$$\hat{RSE} = \sqrt{\frac{RSS}{n-2}}$$

$$\beta_1 = \hat{\beta}_1 \pm \underset{\substack{\uparrow \\ \alpha}}{2}\, \hat{SE}(\hat{\beta}_1)$$

$$\beta_0 = \hat{\beta}_0 \pm \underset{\substack{\uparrow \\ \alpha}}{2}\, \hat{SE}\, \hat{\beta}_0)$$

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

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

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