Module 3: Linear Regression

Lecture 4
Jan 19th, 2023

$\begin{pmatrix} X \\ Y \end{pmatrix}$ $f(x) = E(Y \mid X = x)$

Simple Linear Regression $Y = f(X) + \epsilon = \beta_0 + \beta_1 X + \epsilon$

$\boxed{Y = \boxed{\beta_0 + \beta_1 X} + \varepsilon}$

$= E(\beta_0 + \beta_1 X + \varepsilon \mid X = x)$

$= \beta_0 + \beta_1 x + E(\varepsilon)$

$= \beta_0 + \beta_1 x$

Residual sum of squares (RSS)

$\approx \quad \hat{\beta}_1 \pm 2 \cdot \boxed{\mathrm{SE}(\hat{\beta}_1)}$

Confidence interval.

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

where $\sigma^2 = \mathrm{Var}(\epsilon)$ $\qquad \hat{\sigma}^2 = \frac{\mathrm{RSS}}{n-2}$

ε ?

$\boxed{\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2}$

$\beta_1 \quad \hat{\beta}_1$

2. Sigma

Topics

MICHIGAN STATE UNIVERSITY

- Confidence interval, hypothesis test, and p-value for coefficient estimates
- Residual standard error (RSE)
- R squared
- Setup for multiple linear regression

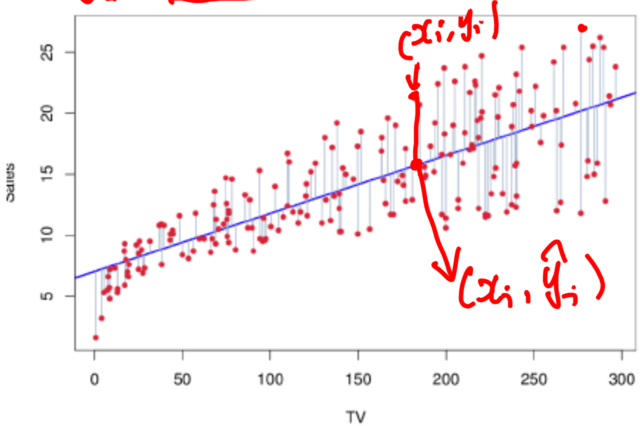
Set up

MICHIGAN STATE UNIVERSITY

$$Y = f(X) + \epsilon = \beta_0 + \beta_1 X + \epsilon$$

metric ?

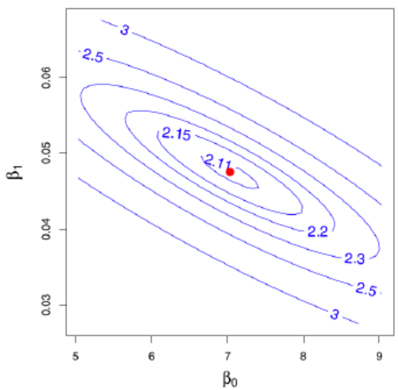
- Given $(x_1, y_1), \cdots, (x_n, y_n)$
- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be prediction for Y on ith value of X.
- $e_i = y_i - \hat{y}_i$ is the ith residual

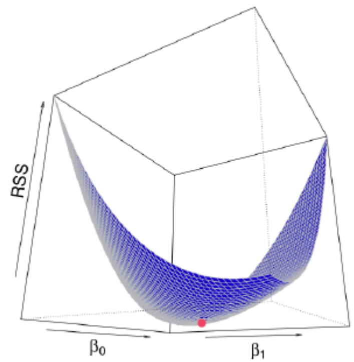


Ordinary Least Squares Regression

MICHIGAN STATE UNIVERSITY

OLS





Residual sum of squares RSS is

$$RSS = e_1^2 + \cdots + e_n^2$$
$$= \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Least squares criterion

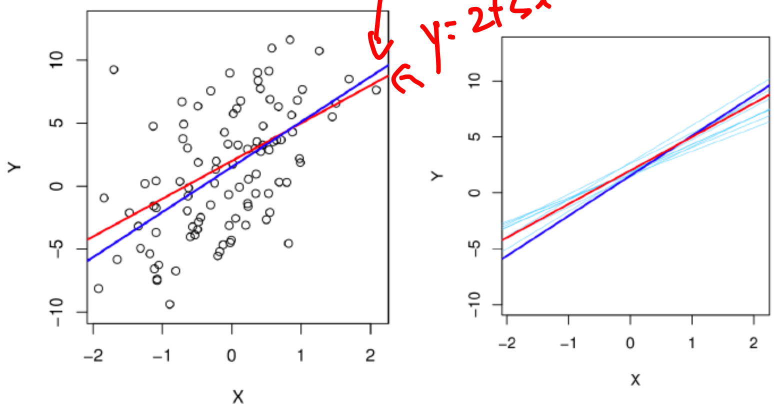
Find β_0 and β_1 that minimize the RSS.

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

Ordinary Least Squares

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

$y = 2+3x$



- 100 data points drawn from $Y = 2 + 3X + \varepsilon$
- ε drawn from normal distribution with mean 0
- Red line is true relationship, blue is least squares estimate
- Repeat this 10 times and plot all the found lines (in variations of blue)
- The resulting models are slightly different but are all around the red true relationship

Variance of OLS estimates

MICHIGAN STATE UNIVERSITY

- Variance of linear regression estimates:

$$\mathrm{SE}(\hat\beta_0) = \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]$$

$$\mathrm{SE}(\hat\beta_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

where $\sigma^2 = \mathrm{Var}(\varepsilon)$

- Residual standard error is an estimate of σ

$$RSE = \sqrt{RSS/(n-2)}$$

$Y = \beta_0 + \beta_1 X + \varepsilon$

$\Downarrow$

$\varepsilon = Y - \beta_0 - \beta_1 X$

$\hat\varepsilon_i = y_i - \hat\beta_0 - \hat\beta_1 x_i$

$\vdots$

$\hat\varepsilon_n$

$\hat\sigma^2 = ?$

$\mathrm{Var}(\varepsilon) = E\left((\varepsilon - E\varepsilon)^2\right)$

$\approx \frac{1}{n}\left(\left(\hat\varepsilon_i - \frac{\sum \hat\varepsilon_i}{n}\right)^2\right)$

$\frac{\sum \hat\varepsilon_i}{n} = 0$

$= \frac{1}{n}(\hat\varepsilon_i^2) = \frac{1}{n}(y_i - \hat\beta_0 - \hat\beta_1 x_i)^2$

$\hat\sigma^2 = \frac{RSS}{n-2}$

Confidence Interval

That is, there is approximately a 95% chance that the interval

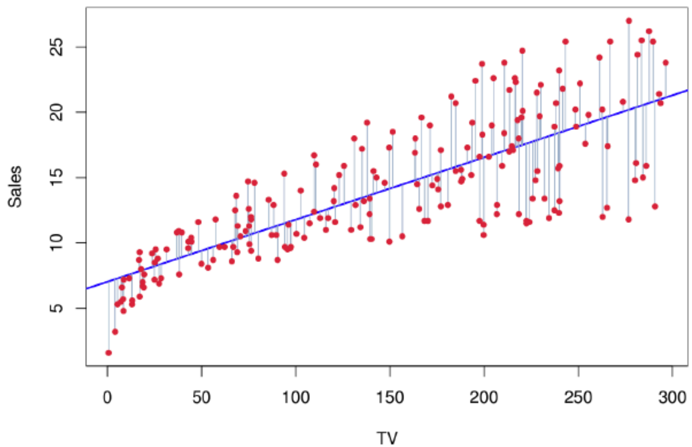
$$\approx \quad \left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\right]$$

not ideal.

will contain the true value of β_1 (under a scenario where we got repeated samples like the present sample)

For the advertising data, the 95% confidence interval for β_1 is $[0.042, 0.053]$

Confidence Interval: Ad data



For the advertising data set, the 95% CIs are:

- β_1 :: $[0.042, 0.053]$
  - First line through [0,8] and [300,20.6]
  - Second line through [0,7] and [300,21.9]
- β_0 :: $[6.130, 7.935]$

Hypothesis Testing

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

- Standard errors can also be used to perform hypothesis tests on the coefficients. The most common hypothesis test involves testing the null hypothesis of

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

versus the *alternative hypothesis*

$$H_A: \quad \text{There is some relationship between } X \text{ and } Y.$$

$Y = \beta_0 + \beta_1 X + \varepsilon$

$\text{if } \beta_1 = 0$

$\boxed{Y = \beta_0 + \varepsilon}$

Hypothesis Testing

MICHIGAN STATE UNIVERSITY

- Standard errors can also be used to perform hypothesis tests on the coefficients. The most common hypothesis test involves testing the null hypothesis of

H_0 : There is no relationship between X and Y

versus the *alternative hypothesis*

H_A : There is some relationship between X and Y.

- Mathematically, it is

$$H_0 : \beta_1 = 0$$

versus

$$H_A : \beta_1 \neq 0,$$

since if $\beta_1 = 0$ then the model reduces to $Y = \beta_0 + \epsilon$, and X is not associated with Y.

Hypothesis Testing

- Mathematically, it is

$$H_0 : \beta_1 = 0$$

versus

$$H_A : \beta_1 \neq 0,$$

since if $\beta_1 = 0$ then the model reduces to $Y = \beta_0 + \epsilon$, and X is not associated with Y.

- We have $\hat{\beta}_1$ from data and want to test whether it is far from 0. But how far?

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

Hypothesis Testing

MICHIGAN STATE UNIVERSITY

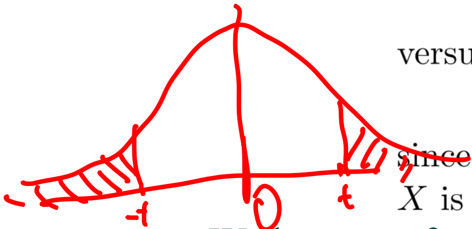
- Mathematically, it is

$$H_0 : \beta_1 = 0$$

versus

$$H_A : \beta_1 \neq 0,$$

since if $\beta_1 = 0$ then the model reduces to $Y = \beta_0 + \epsilon$, and X is not associated with Y.



- We have $\hat{\beta}_1$ from data and want to test whether it is far from 0. But how far?

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

$$\hat{\sigma}^2 = \frac{RSS}{n-2}$$

- This will have a t-distribution with *n-2* degrees of freedom, assuming $\beta_1 = 0$

- Using statistical software, it is easy to compute the probability of observing any value equal to $|t|$ or larger. We call this probability the p-value.

- We will specify an alpha value (0.05) before Hypothesis testing. If p-value less than the alpha value, we will reject the null hypothesis.

Test Statistic and p-value

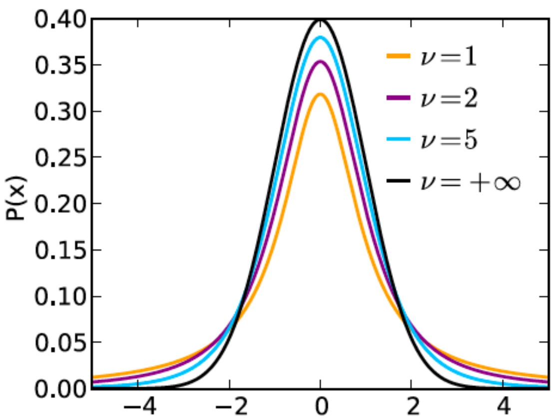
$Y = \beta_0 + \beta_1 X + \boxed{\varepsilon}$

$\varepsilon \sim N(0, \sigma^2)$

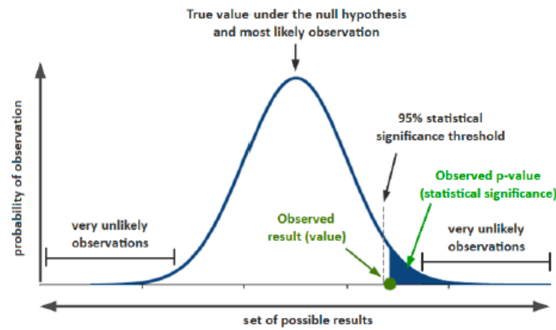
Test statistic:

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

t-distribution with $n-2$ degrees of freedom



Draw me:



Results for the Advertising Data

MICHIGAN STATE UNIVERSITY

	Coefficient	Std. Error	t-statistic	p-value
β_0 Intercept	7.0325	0.4578	15.36	< 0.0001
TV β_1	0.0475	0.0027	17.67	< 0.0001

0.05

- Since p-value < 0.05, we reject the null hypothesis and conclude that TV is related to sale.

Assessing the Accuracy of the

MICHIGAN STATE UNIVERSITY

- Residual Standard Error

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

- estimate of the standard deviation of ε
- Avg amount that the response will deviate from the true regression line
- avg amount response will deviate from the true regression line

Assessing the Accuracy of the

- Residual Standard Error

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

- R-square: fraction of variance explained by the linear model

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

← variance can be explained by your model

where $\mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the *total sum of squares*.

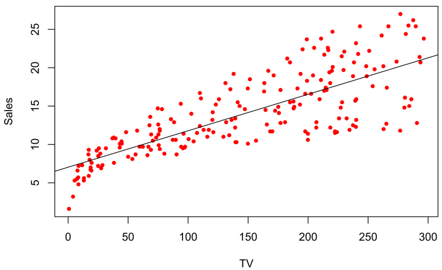
where $\mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total sum of squares

↑ nothing about X

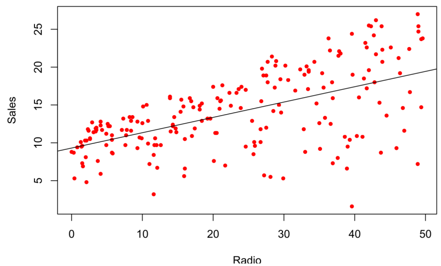
$$\mathrm{RSS} = \sum \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

$$= \sum \varepsilon_i^2$$

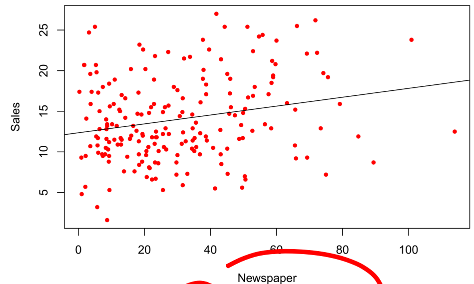
Advertisement Data



$$R^2 = 0.61$$



$$R^2 = 0.33$$



$$R^2 = 0.05$$

Correlation Coefficient

MICHIGAN STATE UNIVERSITY

- Measures dependence between two random variables X and Y

$$r = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

- Correlation coefficient r is between [-1, 1]
  - 0: Variables are not linearly related
  - 1: Variables are perfectly related (same)
  - -1: Variables are negatively related (different)

$$\boxed{R^2 = r^2}$$