Machine Learning in Economics

Module 1
Linear Regression I

Alberto Cappello

Department of Economics, Boston College

Fall 2024

Overview

Agenda:

- Simple (univariate) linear regression
- OLS estimation
- Statistical inference

Readings:

- ISLR Ch.3, section 3.1

Motivation

The simplest parametric form for the relationship between Y and X is

$$\mathbb{E}[Y|X] = f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

In most scenarios this is very far away from being realistic. Why do we still use it?

Motivation

The simplest parametric form for the relationship between Y and X is

$$\mathbb{E}[Y|X] = f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

In most scenarios this is very far away from being realistic. Why do we still use it?

- Straightforward interpretation

Motivation

The simplest parametric form for the relationship between Y and X is

$$\mathbb{E}[Y|X] = f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

In most scenarios this is very far away from being realistic. Why do we still use it?

- Straightforward interpretation
- Quick estimation on datasets of any scale

Motivation

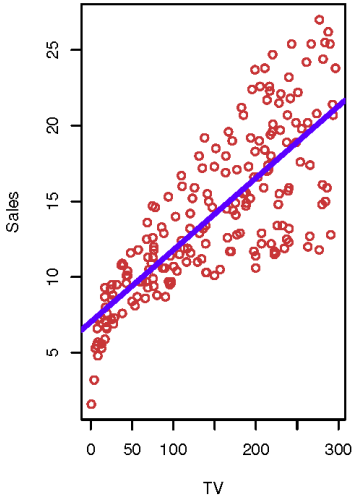
The simplest parametric form for the relationship between Y and X is

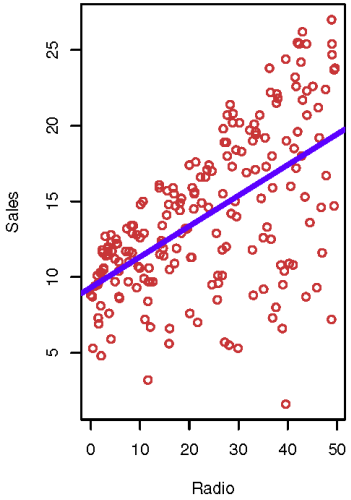
$$\mathbb{E}[Y|X] = f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

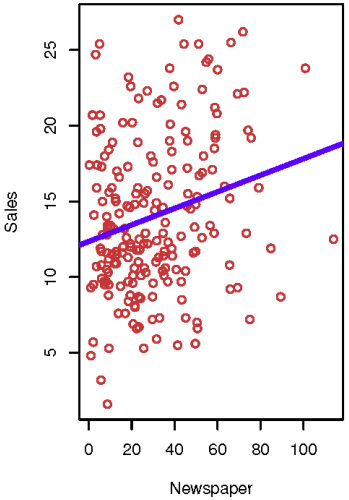
In most scenarios this is very far away from being realistic. Why do we still use it?

- Straightforward interpretation
- Quick estimation on datasets of any scale
- Well-defined statistical properties

Advertising Data







Linear Regression in Marketing

- Is there a relationship between advertising budget and sales?

Linear Regression in Marketing

- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?

Linear Regression in Marketing

- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media contribute to sales?

Linear Regression in Marketing

- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media contribute to sales?
- How accurately can we predict future sales?

Linear Regression in Marketing

- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media contribute to sales?
- How accurately can we predict future sales?
- Is the relationship linear?

Linear Regression in Marketing

- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media contribute to sales?
- How accurately can we predict future sales?
- Is the relationship linear?
- Is there synergy among the advertising media?

Simple Linear Regression

- We start with a model with only one input variable:

$$Y = \beta_0 + \beta_1 X + \epsilon$$
$$\mathbb{E}[\epsilon|X] = 0$$

where β_0 and β_1 are unknown constant parameters that represent the *intercept* and the *slope* of our regression function $f(X)$, and ϵ is the error term.

Simple Linear Regression

- We start with a model with only one input variable:

$$Y = \beta_0 + \beta_1 X + \epsilon$$
$$\mathbb{E}[\epsilon|X] = 0$$

where β_0 and β_1 are unknown constant parameters that represent the *intercept* and the *slope* of our regression function $f(X)$, and ϵ is the error term.

- Based on our data of n pairs of $\{x_i, y_i\}$, we need to come up with an estimated relationship of

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

where $\widehat{y}_i$ is our prediction of Y based on the value $X = x_i$.

Simple Linear Regression

- We start with a model with only one input variable:

$$Y = \beta_0 + \beta_1 X + \epsilon$$
$$\mathbb{E}[\epsilon|X] = 0$$

where β_0 and β_1 are unknown constant parameters that represent the *intercept* and the *slope* of our regression function $f(X)$, and ϵ is the error term.

- Based on our data of n pairs of $\{x_i, y_i\}$, we need to come up with an estimated relationship of

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

where $\widehat{y}_i$ is our prediction of Y based on the value $X = x_i$.

- What estimation method can we use?

Estimation of SLR

- The difference $e_i = y_i - \widehat{y}_i$ represents the ith *residual* or prediction error of our estimated model. Then MSE of our estimates $\widehat{\beta}_0$, $\widehat{\beta}_1$ is

$$MSE(\widehat{\beta}_0, \widehat{\beta}_1) = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2$$

Estimation of SLR

- The difference $e_i = y_i - \widehat{y}_i$ represents the ith *residual* or prediction error of our estimated model. Then MSE of our estimates $\widehat{\beta}_0$, $\widehat{\beta}_1$ is

$$MSE(\widehat{\beta}_0, \widehat{\beta}_1) = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2$$

- Classic Econometrics ditches the averaging and uses the *residual sum of squares* (RSS) as the loss function:

$$RSS(\widehat{\beta}_0, \widehat{\beta}_1) = n \cdot MSE(\widehat{\beta}_0, \widehat{\beta}_1) = \sum_{i=1}^{n} e_i^2 \to \min_{\widehat{\beta}_0, \widehat{\beta}_1}$$

Estimation of SLR

- The difference $e_i = y_i - \widehat{y}_i$ represents the ith *residual* or prediction error of our estimated model. Then MSE of our estimates $\widehat{\beta}_0$, $\widehat{\beta}_1$ is

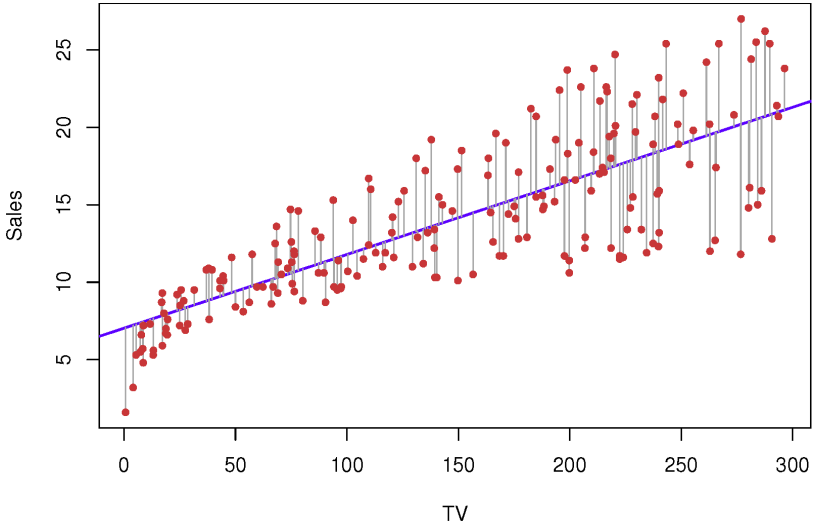
$$MSE(\widehat{\beta}_0, \widehat{\beta}_1) = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2$$

- Classic Econometrics ditches the averaging and uses the *residual sum of squares* (RSS) as the loss function:

$$RSS(\widehat{\beta}_0, \widehat{\beta}_1) = n \cdot MSE(\widehat{\beta}_0, \widehat{\beta}_1) = \sum_{i=1}^{n} e_i^2 \to \min_{\widehat{\beta}_0, \widehat{\beta}_1}$$

- The values of $\widehat{\beta}_0$, $\widehat{\beta}_1$ that minimize RSS are known as *ordinary least squares* (OLS) estimates.

Example: Advertising Data



Goodness-of-fit

- One can show that ANOVA (analysis-of-variance) decomposition of our model is

$$\underbrace{\sum(y_i - \overline{y}_i)^2}_{\text{TSS}} = \underbrace{\sum(\widehat{y}_i - \overline{y}_i)^2}_{\text{ESS}} + \underbrace{\sum e_i^2}_{\text{RSS}}$$

where *total sum of squares* TSS of y_i is partitioned into *explained sum of squares* ESS and unexplained (residual) sum of squares RSS

Goodness-of-fit

- One can show that ANOVA (analysis-of-variance) decomposition of our model is

$$\underbrace{\sum(y_i - \overline{y}_i)^2}_{\text{TSS}} = \underbrace{\sum(\widehat{y}_i - \overline{y}_i)^2}_{\text{ESS}} + \underbrace{\sum e_i^2}_{\text{RSS}}$$

  where *total sum of squares* TSS of y_i is partitioned into *explained sum of squares* ESS and unexplained (residual) sum of squares RSS
- Fraction of variation in y_i explained by our estimated model is called *R-squared*:

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

Goodness-of-fit

- One can show that ANOVA (analysis-of-variance) decomposition of our model is

$$\underbrace{\sum(y_i - \overline{y}_i)^2}_{\text{TSS}} = \underbrace{\sum(\widehat{y}_i - \overline{y}_i)^2}_{\text{ESS}} + \underbrace{\sum e_i^2}_{\text{RSS}}$$

  where *total sum of squares* TSS of y_i is partitioned into *explained sum of squares* ESS and unexplained (residual) sum of squares RSS
- Fraction of variation in y_i explained by our estimated model is called *R-squared*:

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

- The name comes from the fact that in SLR $R^2 = r^2 = \widehat{\text{Corr}}^2(x_i, y_i)$

Statistical inference

- OLS estimates have closed-form solutions:

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

where $\overline{y}$ and $\overline{x}$ are sample means of y_i and x_i

Statistical inference

- OLS estimates have closed-form solutions:

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

  where $\overline{y}$ and $\overline{x}$ are sample means of y_i and x_i
- Are β_0 and β_1 random variables? Are $\widehat{\beta}_0$ and $\widehat{\beta}_1$ random variables?

Statistical inference

- OLS estimates have closed-form solutions:

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

  where $\overline{y}$ and $\overline{x}$ are sample means of y_i and x_i
- Are β_0 and β_1 random variables? Are $\widehat{\beta}_0$ and $\widehat{\beta}_1$ random variables?
- Each sample $\{x_i, y_i\}_{i=1}^{n}$ comes from the same population, described by population regression function $f(X)$ with population parameters β_0 and β_1.

Statistical inference

- OLS estimates have closed-form solutions:

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

  where $\overline{y}$ and $\overline{x}$ are sample means of y_i and x_i
- Are β_0 and β_1 random variables? Are $\widehat{\beta}_0$ and $\widehat{\beta}_1$ random variables?
- Each sample $\{x_i, y_i\}_{i=1}^{n}$ comes from the same population, described by population regression function $f(X)$ with population parameters β_0 and β_1.
- Our sample estimates $\widehat{\beta}_0$, $\widehat{\beta}_1$ will be different for each sample we draw from population, because even with exactly same values of x_i our sample will have random values of ϵ_i as part of y_i.

Exact Statistical Inference

- Formula for $\widehat{\beta}_1$ can be rewritten as

$$\widehat{\beta_1} = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \overline{x})\epsilon_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

with conditional mean of $\widehat{\beta_1}$ given our sample being

$$\mathbb{E}\left[\widehat{\beta}_1 | X\right] = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \overline{x})\mathbb{E}\left[\epsilon_i | X\right]}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Exact Statistical Inference

- Formula for $\widehat{\beta}_1$ can be rewritten as

$$\widehat{\beta_1} = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \overline{x})\epsilon_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

with conditional mean of $\widehat{\beta_1}$ given our sample being

$$\mathbb{E}\left[\widehat{\beta}_1 | X\right] = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \overline{x})\mathbb{E}\left[\epsilon_i | X\right]}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

- Under our zero conditional mean assumption $\mathbb{E}\left[\epsilon_i | X\right] = 0$ we get

$$\mathbb{E}\left[\widehat{\beta}_1 | X\right] = \beta_1 \quad \Rightarrow \quad \mathbb{E}\left[\widehat{\beta}_1\right] = \beta_1$$

and

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} \quad \Rightarrow \quad \mathbb{E}\left[\widehat{\beta}_0\right] = \beta_0$$

Exact Statistical Inference

- Under ZCM assumption and linear parametric form of $f(X)$ OLS estimates are *unbiased* — on average across repeated samples we get the true parameters' values.

Exact Statistical Inference

- Under ZCM assumption and linear parametric form of $f(X)$ OLS estimates are *unbiased* — on average across repeated samples we get the true parameters' values.
- What about accuracy/spread across repeated sample? Since OLS estimates are unbiased, their MSE is equal to their variance:

$$Var(\widehat{\beta}_1|X) = Var\left(\beta_1 + \frac{\sum_{i=1}^{n}(x_i - \overline{x})\epsilon_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\Big|X\right) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Exact Statistical Inference

- Under ZCM assumption and linear parametric form of $f(X)$ OLS estimates are *unbiased* — on average across repeated samples we get the true parameters' values.
- What about accuracy/spread across repeated sample? Since OLS estimates are unbiased, their MSE is equal to their variance:

$$Var(\widehat{\beta}_1|X) = Var\left(\beta_1 + \frac{\sum_{i=1}^{n}(x_i - \overline{x})\epsilon_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\Big|X\right) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

- The formula above is valid only under the i.i.d. assumption for our data — i.e. that all our observations are identical independent draws from the same population. This ensures that regression errors ϵ_i are *homoscedastic* with the same constant variance $Var(\epsilon_i|X) = \sigma^2$ and *serially uncorrelated* with $\text{Cov}(\epsilon_i, \epsilon_j|X) = 0$.

Exact Statistical Inference

- Under ZCM assumption and linear parametric form of $f(X)$ OLS estimates are *unbiased* — on average across repeated samples we get the true parameters' values.
- What about accuracy/spread across repeated sample? Since OLS estimates are unbiased, their MSE is equal to their variance:

$$Var(\widehat{\beta}_1|X) = Var\left(\beta_1 + \frac{\sum_{i=1}^{n}(x_i - \overline{x})\epsilon_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\Big|X\right) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

- The formula above is valid only under the i.i.d. assumption for our data — i.e. that all our observations are identical independent draws from the same population. This ensures that regression errors ϵ_i are *homoscedastic* with the same constant variance $Var(\epsilon_i|X) = \sigma^2$ and *serially uncorrelated* with $\text{Cov}(\epsilon_i, \epsilon_j|X) = 0$.
- What do violating these assumptions cause?

Homoskedasticity vs heteroskedasticity

Figure: homoskedasticity

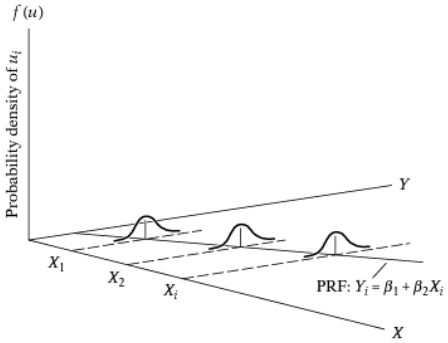
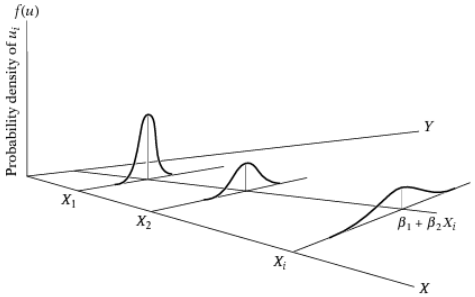


Figure: heteroskedasticity



Exact Statistical Inference

- The formula for variance of $\widehat{\beta_1}$ can be rewritten as

$$Var(\widehat{\beta}_1|X) = \frac{\sigma^2}{n \cdot \widehat{Var}(X)}$$

- It means that variation (spread) of OLS estimate $\widehat{\beta_1}$ can be measured as a ratio of noise σ^2 over signal $n \cdot \widehat{Var}(X)$.
  - Variance goes down if we have larger sample size or when X varies a lot (or both).
  - Variance goes up if unobserved error term ϵ_i has higher degree of uncertainty.

Exact Statistical Inference

- The formula for variance of $\widehat{\beta_1}$ can be rewritten as

$$Var(\widehat{\beta}_1|X) = \frac{\sigma^2}{n \cdot \widehat{Var}(X)}$$

- It means that variation (spread) of OLS estimate $\widehat{\beta_1}$ can be measured as a ratio of noise σ^2 over signal $n \cdot \widehat{Var}(X)$.
  - Variance goes down if we have larger sample size or when X varies a lot (or both).
  - Variance goes up if unobserved error term ϵ_i has higher degree of uncertainty.
- It can be shown that under $Var(\epsilon_i|X) = \sigma^2$ OLS is BLUE — Best (i.e. smallest variance) Linear Unbiased Estimator.
- In Statistics this is known as *efficiency* of an estimator, typically defined as having lower(-est) MSE.

Exact Statistical Inference

- In practice we prefer to use standard error of $\widehat{\beta_1}$ instead of variance, as the former have the same units of measurements as X.

Exact Statistical Inference

- In practice we prefer to use standard error of $\widehat{\beta_1}$ instead of variance, as the former have the same units of measurements as X.
- Since we do not know true population variance σ^2 of our error term ϵ_i, we need to estimate it using our sample OLS residuals e_i^2:

$$\widehat{\sigma}^2 = \frac{RSS}{n-2} \quad \text{and} \quad \text{SE}(\widehat{\beta_1}) = \sqrt{\frac{\widehat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

- The $(n-2)$ in denominator is called *degrees of freedom* of our regression model, as we have two equations for OLS estimates that bind together 2 out of n residuals in our model.

Exact Statistical Inference

- In small samples we cannot say anything else about the properties of OLS estimates as random variables unless we impose more assumption on what the nature of ϵ_i is.

Exact Statistical Inference

- In small samples we cannot say anything else about the properties of OLS estimates as random variables unless we impose more assumption on what the nature of ϵ_i is.
- That is why classic linear regression models assume that ϵ_i follows normal (Gaussian) distribution, which leads OLS estimates also being normal (Gaussian):

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2) \Rightarrow \widehat{\beta}_j \sim \mathcal{N}(\beta_j, Var(\widehat{\beta}_j))$$

Exact Statistical Inference

- In small samples we cannot say anything else about the properties of OLS estimates as random variables unless we impose more assumption on what the nature of ϵ_i is.
- That is why classic linear regression models assume that ϵ_i follows normal (Gaussian) distribution, which leads OLS estimates also being normal (Gaussian):

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2) \Rightarrow \widehat{\beta}_j \sim \mathcal{N}(\beta_j, Var(\widehat{\beta}_j))$$

- This allows us to compute *confidence intervals* and do *hypothesis testing* using the fact that

$$\frac{\widehat{\beta}_j - \beta_j}{SE(\widehat{\beta}_j)} \sim t_{n-2}$$

Exact Statistical Inference

- A $(1-\alpha)\%$ confidence interval for β_1 takes the form of

$$\left[\widehat{\beta_1} - t_{n-2}^{crit} \cdot \text{SE}(\widehat{\beta_1}); \widehat{\beta_1} + t_{n-2}^{crit} \cdot \text{SE}(\widehat{\beta_1})\right]$$

where t_{n-2}^{crit} is a *critical value* of t-distribution with $n-2$ degrees of freedom, equal to $(1-\alpha/2)\%$ percentile.

Exact Statistical Inference

- A $(1-\alpha)\%$ confidence interval for β_1 takes the form of

$$\left[\widehat{\beta_1} - t_{n-2}^{crit} \cdot \text{SE}(\widehat{\beta_1}); \widehat{\beta_1} + t_{n-2}^{crit} \cdot \text{SE}(\widehat{\beta_1})\right]$$

  where t_{n-2}^{crit} is a *critical value* of t-distribution with $n-2$ degrees of freedom, equal to $(1-\alpha/2)\%$ percentile.
- In repeated sampling and estimation of this confidence interval, in $(1-\alpha)\%$ of cases, the true β_1 will lie in those intervals

Exact Statistical Inference

- A $(1-\alpha)\%$ confidence interval for β_1 takes the form of

$$\left[\widehat{\beta_1} - t_{n-2}^{crit} \cdot \text{SE}(\widehat{\beta_1}); \widehat{\beta_1} + t_{n-2}^{crit} \cdot \text{SE}(\widehat{\beta_1})\right]$$

  where t_{n-2}^{crit} is a *critical value* of t-distribution with $n-2$ degrees of freedom, equal to $(1-\alpha/2)\%$ percentile.
- In repeated sampling and estimation of this confidence interval, in $(1-\alpha)\%$ of cases, the true β_1 will lie in those intervals
- For advertising data, the 95% confidence interval for β_1 in regression of `Sales` on `TV` is approximately $[0.042; 0.053]$.

Exact Statistical Inference

- The most common hypothesis test in regression analysis involves testing the *null hypothesis* of

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

versus the *alternative hypothesis* of

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

Exact Statistical Inference

- The most common hypothesis test in regression analysis involves testing the *null hypothesis* of

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

versus the *alternative hypothesis* of

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

- Mathematically, this corresponds to testing

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_A : \beta_1 \neq 0$$

since if $\beta_1 = 0$ our model reduces to $Y = \beta_0 + \epsilon$, and there is no association of X with Y.

Exact Statistical Inference

- In classic regression analysis this hypothesis is known as *significance test* — it tests for (absence of) statistically significant linear relationship between Y and X.
- To test this hypothesis we compute a *t-statistic* via

$$t = \frac{\widehat{\beta_1} - 0}{SE(\widehat{\beta_1})}$$

which under H_0 has a t-distribution with $n - 2$ degrees of freedom

Exact Statistical Inference

- In classic regression analysis this hypothesis is known as *significance test* — it tests for (absence of) statistically significant linear relationship between Y and X.
- To test this hypothesis we compute a *t-statistic* via

$$t = \frac{\widehat{\beta_1} - 0}{SE(\widehat{\beta_1})}$$

  which under H_0 has a t-distribution with $n - 2$ degrees of freedom
- Finally we either compare it to a critical value for a given *significance level* α (e.g. $t_{n-2}^{crit} \approx 2$ for $\alpha = 5\%$), or compute the *p-value* of our hypothesis — probability of observing any value equal to $|t|$ or larger.

Exact Statistical Inference

- Example using advertising data:

Variable	Coefficient	SE	t	p-value
`Intercept`	7.0325	0.4578	15.36	<0.0001
`TV`	0.0475	0.0027	17.67	<0.0001

$R^2 = 0.612 \quad \widehat{\sigma} = 3.26$

`Sales` - sales in thousands of units, `TV` - TV ad budget in thousands of $.

Exact Statistical Inference

- Example using advertising data:

Variable	Coefficient	SE	t	p-value
`Intercept`	7.0325	0.4578	15.36	<0.0001
`TV`	0.0475	0.0027	17.67	<0.0001

$$R^2 = 0.612 \quad \widehat{\sigma} = 3.26$$

`Sales` - sales in thousands of units, `TV` - TV ad budget in thousands of \$.

- Both coefficients are statistically significant on any reasonable significance level α.

Exact Statistical Inference

- Example using advertising data:

Variable	Coefficient	SE	t	p-value
`Intercept`	7.0325	0.4578	15.36	<0.0001
`TV`	0.0475	0.0027	17.67	<0.0001

$$R^2 = 0.612 \quad \widehat{\sigma} = 3.26$$

`Sales` - sales in thousands of units, `TV` - TV ad budget in thousands of \$.

- Both coefficients are statistically significant on any reasonable significance level α.
- On average and other things equal, extra \$1000 spent on TV ads is associated with extra 47 units sold across all markets.

Exact Statistical Inference

- Example using advertising data:

Variable	Coefficient	SE	t	p-value
`Intercept`	7.0325	0.4578	15.36	<0.0001
`TV`	0.0475	0.0027	17.67	<0.0001

$$R^2 = 0.612 \quad \widehat{\sigma} = 3.26$$

`Sales` - sales in thousands of units, `TV` - TV ad budget in thousands of \$.

  - Both coefficients are statistically significant on any reasonable significance level α.
  - On average and other things equal, extra \$1000 spent on TV ads is associated with extra 47 units sold across all markets.

- All these conclusions are only valid because we made the following assumption:

$$\epsilon_i | X \sim \mathcal{N}(0, \sigma^2)$$

Asymptotic (large sample) inference

- Normality of ϵ_i is a very strong assumption, which often is unrealistic or even mathematically infeasible.

Asymptotic (large sample) inference

- Normality of ϵ_i is a very strong assumption, which often is unrealistic or even mathematically infeasible.
- Good news — in large samples we can replace this assumption with asymptotic equivalent using such powerful statistical results as Law of Large Numbers (LLN) and Central Limit Theorem

Asymptotic (large sample) inference

- Normality of ϵ_i is a very strong assumption, which often is unrealistic or even mathematically infeasible.
- Good news — in large samples we can replace this assumption with asymptotic equivalent using such powerful statistical results as Law of Large Numbers (LLN) and Central Limit Theorem
- **LLN**: Let $X_1, X_2, X_3....X_n$ be i.i.d. random variables with a finite expected value $EX_i = \mu < \infty$. Then for an ϵ

$$\lim_{n \to \infty} P(|\bar{X} - \mu| \geq \epsilon) = 0 \tag{1}$$

$$plim(\bar{X}) = \mu \tag{2}$$

Asymptotic (large sample) inference

- Normality of ϵ_i is a very strong assumption, which often is unrealistic or even mathematically infeasible.
- Good news — in large samples we can replace this assumption with asymptotic equivalent using such powerful statistical results as Law of Large Numbers (LLN) and Central Limit Theorem
- **LLN**: Let $X_1, X_2, X_3....X_n$ be i.i.d. random variables with a finite expected value $EX_i = \mu < \infty$. Then for an ϵ

$$\lim_{n\to\infty} P(|\bar{X} - \mu| \geq \epsilon) = 0 \tag{1}$$

$$plim(\bar{X}) = \mu \tag{2}$$

- **CLT**: Let $X_1, X_2, X_3....X_n$ be i.i.d. random variables from the same distribution with mean μ and variance σ^2

$$\bar{X}_{n\to\infty} \sim \mathcal{N}(\mu, \sigma^2/n) \tag{3}$$

Asymptotic (large sample) inference

- While these results still require certain assumptions to hold, their key advantage is that given large enough dataset, we can obtain near exact inference without the need to do repeated sampling.

Asymptotic (large sample) inference

- While these results still require certain assumptions to hold, their key advantage is that given large enough dataset, we can obtain near exact inference without the need to do repeated sampling.
- LLN allows us to establish *consistency* of OLS estimates:

$$\text{plim}(\widehat{\beta_1}) = \text{plim}\left(\beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})\epsilon_i}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}\right) = \beta_1 + \frac{Cov(X, \epsilon)}{Var(X)} = \beta_1$$

where the last step is due to $\mathbb{E}[\epsilon|X] = 0$.

Asymptotic (large sample) inference

- While these results still require certain assumptions to hold, their key advantage is that given large enough dataset, we can obtain near exact inference without the need to do repeated sampling.
- LLN allows us to establish *consistency* of OLS estimates:

$$\text{plim}(\widehat{\beta_1}) = \text{plim}\left(\beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})\epsilon_i}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}\right) = \beta_1 + \frac{Cov(X, \epsilon)}{Var(X)} = \beta_1$$

  where the last step is due to $\mathbb{E}[\epsilon|X] = 0$.
- CLT allows us to establish *asymptotic normality* of OLS estimates:

$$\frac{\widehat{\beta_j} - \beta_j}{SE(\widehat{\beta_j})} \overset{a}{\sim} \mathcal{N}(0, 1)$$

Asymptotic (large sample) inference

- While these results still require certain assumptions to hold, their key advantage is that given large enough dataset, we can obtain near exact inference without the need to do repeated sampling.
- LLN allows us to establish *consistency* of OLS estimates:

$$\text{plim}(\widehat{\beta_1}) = \text{plim}\left(\beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})\epsilon_i}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}\right) = \beta_1 + \frac{Cov(X, \epsilon)}{Var(X)} = \beta_1$$

  where the last step is due to $\mathbb{E}[\epsilon|X] = 0$.
- CLT allows us to establish *asymptotic normality* of OLS estimates:

$$\frac{\widehat{\beta}_j - \beta_j}{SE(\widehat{\beta}_j)} \overset{a}{\sim} \mathcal{N}(0, 1)$$

- The rest of the inference (CIs, hypothesis testing) can be performed in the same exact way.