Econometric Methods

Module 1
Simple Linear Regression I

Alberto Cappello

Department of Economics, Boston College

Spring 2025

Motivating Example

- Is there a relationship between advertising budget and sales?

Motivating Example

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

Motivating Example

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

Motivating Example

- 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?

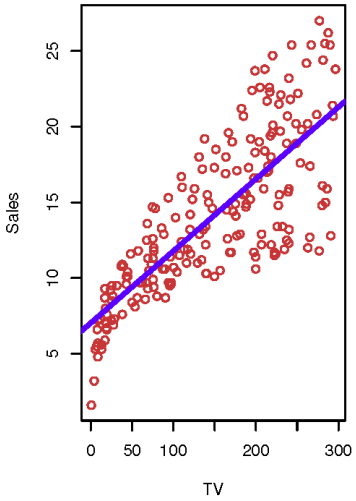
Motivating Example

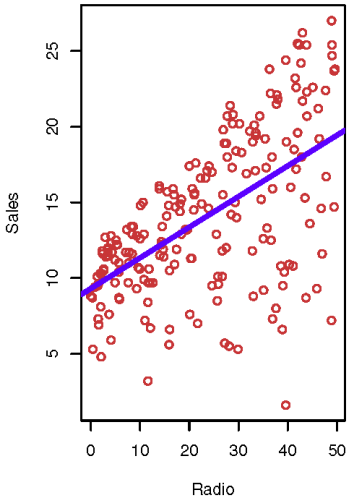
- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media contribute to sales?
- How accurately can we predict future sales?
- Is the relationship linear?

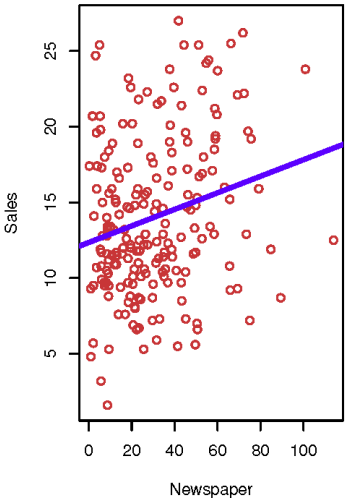
Motivating Example

- 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?

Advertising Data







Motivating Example

How can we model the relationship between Sales and Advertisement? Let

- Y = number of unit of a good/service sold.
- X = advertisement budget in thousands of dollars.

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

$$\mathbb{E}[Y|X] = f(X) = \beta_0 + \beta_1 X$$

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

Motivating Example

How can we model the relationship between Sales and Advertisement? Let

- Y = number of unit of a good/service sold.
- X = advertisement budget in thousands of dollars.

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

$$\mathbb{E}[Y|X] = f(X) = \beta_0 + \beta_1 X$$

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

- Straightforward interpretation

Motivating Example

How can we model the relationship between Sales and Advertisement? Let

- Y = number of unit of a good/service sold.
- X = advertisement budget in thousands of dollars.

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

$$\mathbb{E}[Y|X] = f(X) = \beta_0 + \beta_1 X$$

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

Motivating Example

How can we model the relationship between Sales and Advertisement? Let

- Y = number of unit of a good/service sold.
- X = advertisement budget in thousands of dollars.

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

$$\mathbb{E}[Y|X] = f(X) = \beta_0 + \beta_1 X$$

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

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? $\rightarrow$ Ordinary Least Squares

Ordinary Least Squares (OLS)

- The difference $e_i = y_i - \widehat{y}_i$ represents the ith *residual* or prediction error of our estimated model.
- We want to find the value of $(\widehat{\beta}_0, \widehat{\beta}_1)$ that minimize the Residual Sum of Squares (RSS)

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

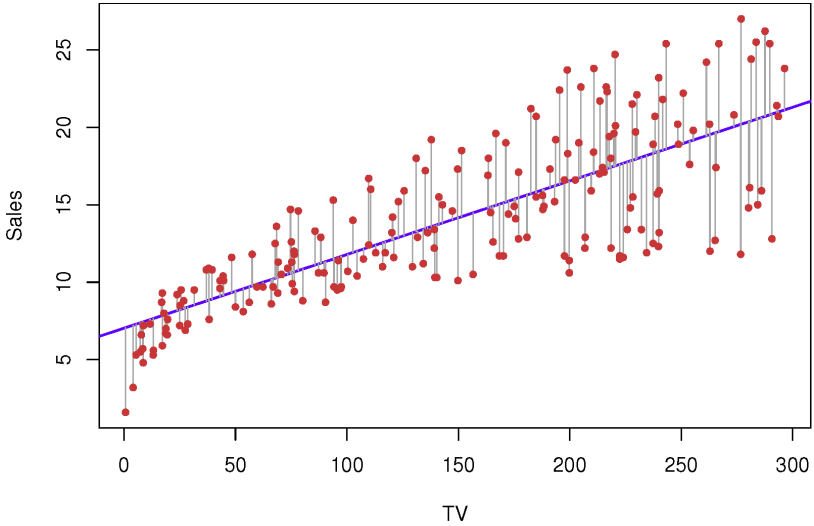
Ordinary Least Squares (OLS)

- The difference $e_i = y_i - \widehat{y}_i$ represents the ith *residual* or prediction error of our estimated model.
- We want to find the value of $(\widehat{\beta}_0, \widehat{\beta}_1)$ that minimize the Residual Sum of Squares (RSS)

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

- 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

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

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

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

  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

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

  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

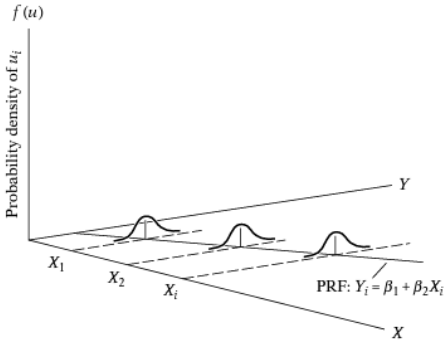
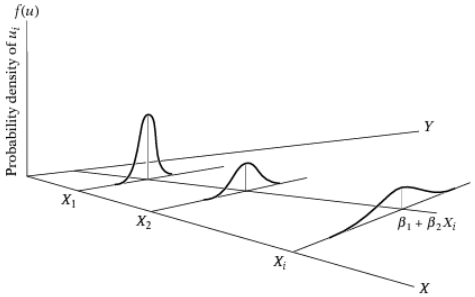


Figure: heteroskedasticity



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.

Extensions of Linear Model

- The main issue with linear model is that all variables have fixed marginal effects:

$$\beta_1 = \frac{\partial \mathbb{E}[Y|X]}{\partial X} = \frac{\mathbb{E}[\Delta Y|X]}{\Delta X}$$

- This is totally unrealistic in many cases — e.g. effect of years of education, standardized test scores, number of kids, etc.
- There are three most common ways to change that and yet retain the useful simplicity of a linear model: interactions, non-linear transformations and higher order polynomials.

Polynomials

- The most straightforward way to make marginal effects vary with values of X is to add powers of corresponding regressors:

$$wage_i = \beta_0 + \beta_1\, exper_i + \beta_2\, exper_i^2 + \epsilon_i$$

- Marginal effects become variable:

$$\frac{\mathbb{E}[\Delta wage|\ldots]}{\Delta exper} = \beta_1 + 2\beta_2\, exper$$

  but β_1 no longer have meaningful interpretation for most cases.
- Notice that $\hat{wage}_i = \hat{\beta}_0 + \hat{\beta}_1\, exper_i + \hat{\beta}_2\, exper_i^2$ is parabola shaped (not a line).
- Higher order polynomial regression are very good at predicting, but mostly useless for inference (more on this in later chapters).

Qualitative predictors

- OLS is a mathematical algorithm that find optimal solutions over a space of $p+1$ numerical variables $Y, X_1, X_2, \ldots$
- But not every observable feature/predictor has a natural numerical scale to be measured along.

Qualitative predictors

- OLS is a mathematical algorithm that find optimal solutions over a space of $p+1$ numerical variables $Y, X_1, X_2, \ldots$
- But not every observable feature/predictor has a natural numerical scale to be measured along.
- *Qualitative* or *factor* variables such as gender, race, city district or education major clearly are important in many statistical applications, but all of them lack numerical scale.
- The solution is to split the data for every such variable into non-overlapping groups and assign a binary 0/1 variable to identify every group.

Qualitative predictors

- The simplest case is when our qualitative predictor has only two possible values in the data, e.g. having a college degree:

$$college_i = \begin{cases} 0, \text{person i does not have a college degree} \\ 1, \text{person i has a college degree} \end{cases}$$

Qualitative predictors

- The simplest case is when our qualitative predictor has only two possible values in the data, e.g. having a college degree:

$$college_i = \begin{cases} 0, \text{person i does not have a college degree} \\ 1, \text{person i has a college degree} \end{cases}$$

- Then a wage equation can take the form of

$$wage_i = \beta_0 + \beta_1 exper_i + \beta_2 college_i + \epsilon_i$$

or

$$wage_i = \begin{cases} \beta_0 + \beta_1 exper_i + & \epsilon_i, \quad \text{no college degree} \\ \beta_0 + \beta_1 exper_i + \beta_2 + & \epsilon_i, \quad \text{college degree} \end{cases}$$

Qualitative predictors

- The simplest case is when our qualitative predictor has only two possible values in the data, e.g. having a college degree:

$$college_i = \begin{cases} 0, \text{person i does not have a college degree} \\ 1, \text{person i has a college degree} \end{cases}$$

- Then a wage equation can take the form of

$$wage_i = \beta_0 + \beta_1 exper_i + \beta_2 college_i + \epsilon_i$$

or

$$wage_i = \begin{cases} \beta_0 + \beta_1 exper_i + & \epsilon_i, \quad \text{no college degree} \\ \beta_0 + \beta_1 exper_i + \beta_2 + & \epsilon_i, \quad \text{college degree} \end{cases}$$

- In this case β_2 gains a very special interpretation: it stands for a ceteris paribus fixed difference in average wage for college educated vs non-college educated workers.

Qualitative predictors

- If we have k possible values (groups), we need to create $k-1$ *dummy variables* that take values 0 and 1 to differentiate between $k-1$ groups and a baseline group.
- Each individual dummy variable will show the fixed difference between one of the groups and the baseline group. The difference between two dummies will show the fixed difference between those two groups only.

Qualitative predictors

- Suppose our qualitative predictor takes three possible values in the data, e.g. major $\in$ economics, maths, physics

$$economics_i = \begin{cases} 0, \text{person i does not have an economics degree} \\ 1, \text{person i has an economics degree} \end{cases}$$

$$math_i = \begin{cases} 0, \text{person i does not have a math degree} \\ 1, \text{person i has a math degree} \end{cases}$$

Qualitative predictors

- Suppose our qualitative predictor takes three possible values in the data, e.g. major $\in$ economics, maths, physics

$$economics_i = \begin{cases} 0, \text{person i does not have an economics degree} \\ 1, \text{person i has an economics degree} \end{cases}$$

$$math_i = \begin{cases} 0, \text{person i does not have a math degree} \\ 1, \text{person i has a math degree} \end{cases}$$

- Then a wage equation is

$$wage_i = \beta_0 + \beta_1 exper_i + \beta_2 economics_i + \beta_3 math_i + \epsilon_i$$

or

$$wage_i = \begin{cases} \beta_0 + \beta_1 exper_i + & \epsilon_i, & \text{neither econ nor math degree} \\ \beta_0 + \beta_1 exper_i + \beta_2 + & \epsilon_i, & \text{economics degree} \\ \beta_0 + \beta_1 exper_i + & \beta_3 + \epsilon_i, & \text{math degree} \end{cases}$$

Interactions

- In our previous analysis of the advertising data, we assumed that the effect on `sales` of increasing one advertising medium is independent of the amount spent on the other media.
- For example, in the model

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon$$

the average effect on `sales` of a one-unit increase in `TV` is always β_1, regardless of the amount spent on `radio`.

Interactions

- But suppose that spending money on radio advertising actually increases the effectiveness of TV advertising, so that the slope term for `TV` should increase as `radio` increases.
- In this situation, given a fixed budget of \$100 000, spending half on `radio` and half on `TV` may increase sales more than allocating the entire amount to either `TV` or to `radio`.
- In marketing, this is known as a *synergy effect*, and in statistics it is referred to as an *interaction effect*.
- We can capture this using the model

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{TV} \times \texttt{radio} + \epsilon$$

Interactions

- But suppose that spending money on radio advertising actually increases the effectiveness of TV advertising, so that the slope term for `TV` should increase as `radio` increases.
- In this situation, given a fixed budget of \$100 000, spending half on `radio` and half on `TV` may increase sales more than allocating the entire amount to either `TV` or to `radio`.
- In marketing, this is known as a *synergy effect*, and in statistics it is referred to as an *interaction effect*.
- We can capture this using the model

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{TV} \times \texttt{radio} + \epsilon$$

- the average effect on `sales` of a one-unit increase in `TV` is $(\beta_1 + \beta_3 \times \texttt{radio})$, which depends on `radio` spend

Interactions

- A more interesting way to add interactions is by interacting usual numerical variables with dummy (qualitative) variables:

$$wage_i = \beta_0 + \beta_1\, exper_i + \beta_2\, college_i + \beta_3\, exper_i \times college_i + \epsilon_i$$

Interactions

- A more interesting way to add interactions is by interacting usual numerical variables with dummy (qualitative) variables:

$$wage_i = \beta_0 + \beta_1\, exper_i + \beta_2\, college_i + \beta_3\, exper_i \times college_i + \epsilon_i$$

- The model above allows college graduates to have not only fixed difference in wages, but also a different return to experience:

$$wage_i = \begin{cases} \beta_0 + & \beta_1 exper_i + & \epsilon_i, & \text{no college degree} \\ \beta_0 + \beta_2 + & (\beta_1 + \beta_3) exper_i + & \epsilon_i, & \text{college degree} \end{cases}$$

Non-Linear Transformations

- Another way to achieve non-constant marginal effects is to use non-linear transformation of regressors, e.g. natural logs:

$$\log(wage)_i = \beta_0 + \beta_1 \log(exper)_i + \beta_2 college_i + \epsilon_i$$

Non-Linear Transformations

- Another way to achieve non-constant marginal effects is to use non-linear transformation of regressors, e.g. natural logs:

$$\log(wage)_i = \beta_0 + \beta_1 \log(exper)_i + \beta_2 college_i + \epsilon_i$$

- The model still remains linear in parameters (betas), so OLS estimation proceeds as usual. However, marginal effect of years of experience on wage is no longer the same for every extra year of experience:

$$\beta_1 = \frac{\mathbb{E}[\Delta \log(wage)|\ldots]}{\Delta \log(exper)} \approx \frac{\mathbb{E}[\%\Delta wage|\ldots]}{\%\Delta exper}$$

- The latter fraction is known as *elasticity* and plays a very important role in Economics.