EAS 508 Week 3: Linear Regression

Jianzhen Liu, PhD

University at Buffalo

Why Linear regression

- Linear regression is a simple approach to supervised learning. It assumes that the dependence of Y on $X_1, X_2, \cdots, X_p$ is linear.
- True regression functions are never linear!
- Although it may seem overly simplistic, linear regression is extremely useful both conceptually and practically.

Steps in Regression Analysis

1. Statement of the problem
2. Using regression for:
   - Diagnostic
   - Predictive,
   - or Prescriptive analytics
3. Selection of potentially relevant response and explanatory variables
4. Data collection
   - Internal data external data,
   - purchased data,
   - experiments, etc.

Steps in Regression Analysis (Continued)

5. Choice of fitting method:
   - Ordinary least squares (OLS),
   - Generalized least squares,
   - Maximum likelihood,
   - Etc.
6. Model fitting
7. Model validation (diagnostics)
8. Refine the model & iterate from step 3
9. Use of the mode

Business Examples

Option	Description
1. Used car price	odometer reading, age of car, condition
2. Sales	Madvertisement spending
3. Time taken to repair a product	experience of technician in years
4. Product added to shopping cart?	ratings, price
5. Starting salary of new employee	work experience, years of education
6. Sale price of house	square feet, # of bedrooms, location
7. Will customer default?	credit balance, income, age
8. Will customer churn?	Will customer churn?

Quiz (True/False)

- Could a variable, say price, be either a dependent or an independent variable?
  - **True**
  - **False**
- A variable that takes binary values (pass/fail or true/false) cannot be a dependent variable.
  - **True**
  - **False**

Answer (True/False)

- Could a variable, say price, be either a dependent or an independent variable?
  - Answer: **True**. Depends on the purpose of your model; see where price appears in examples #1 and #4 in the previous
- A variable that takes binary values (pass/fail or true/false) cannot be a dependent variable.
  - Answer: **FALSE**. We do use 0/1 dependent variables in logistic regression models; #7 in the previous slide is one example.

Simple Linear Regression

Our goal is to find the "best" line that describes a linear relationship; that is, find (β_0, β_1) where

$$Y = \beta_0 + \beta_1 x + \epsilon.$$

Equivalently, estimating:

- β_0 **Intercept**
- β_1 **Slope**, also known as coefficients or parameters

ϵ is the (random) deviance of the data from the linear model

Simple linear regression using a single predictor X

Given some estimates β_0 and β_1 for the model coefficients,

we predict future sales using

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

Equivalently, estimating:

- $\hat{\beta}_0$ is an estimate of **Intercept** β_0
- $\hat{\beta}_1$ is an estimiate of **Slope** β_1

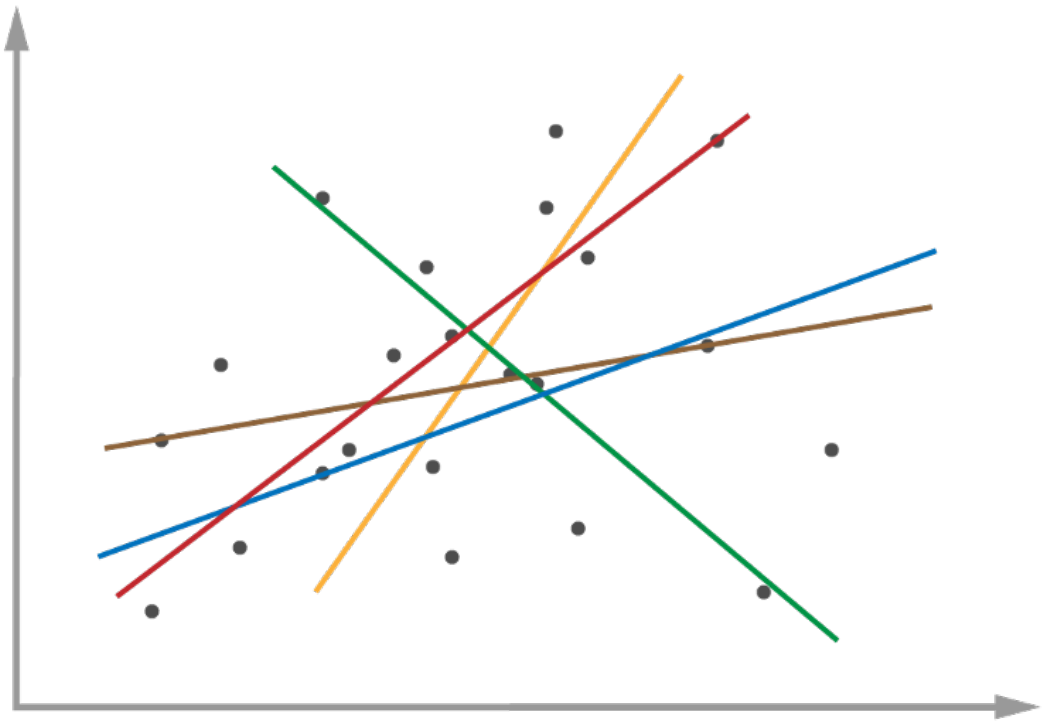
where $\hat{y}$ indicates a prediction of Y on the basis of $X = x$.

The hat ˆ symbol denotes an estimated value.

Simple Linear Regression: Which line?

Which line to choose?

- The line that fits the data "best" where 'best" is in reference to a given criterion.



How to find the best line?

Look Into Linear Regression

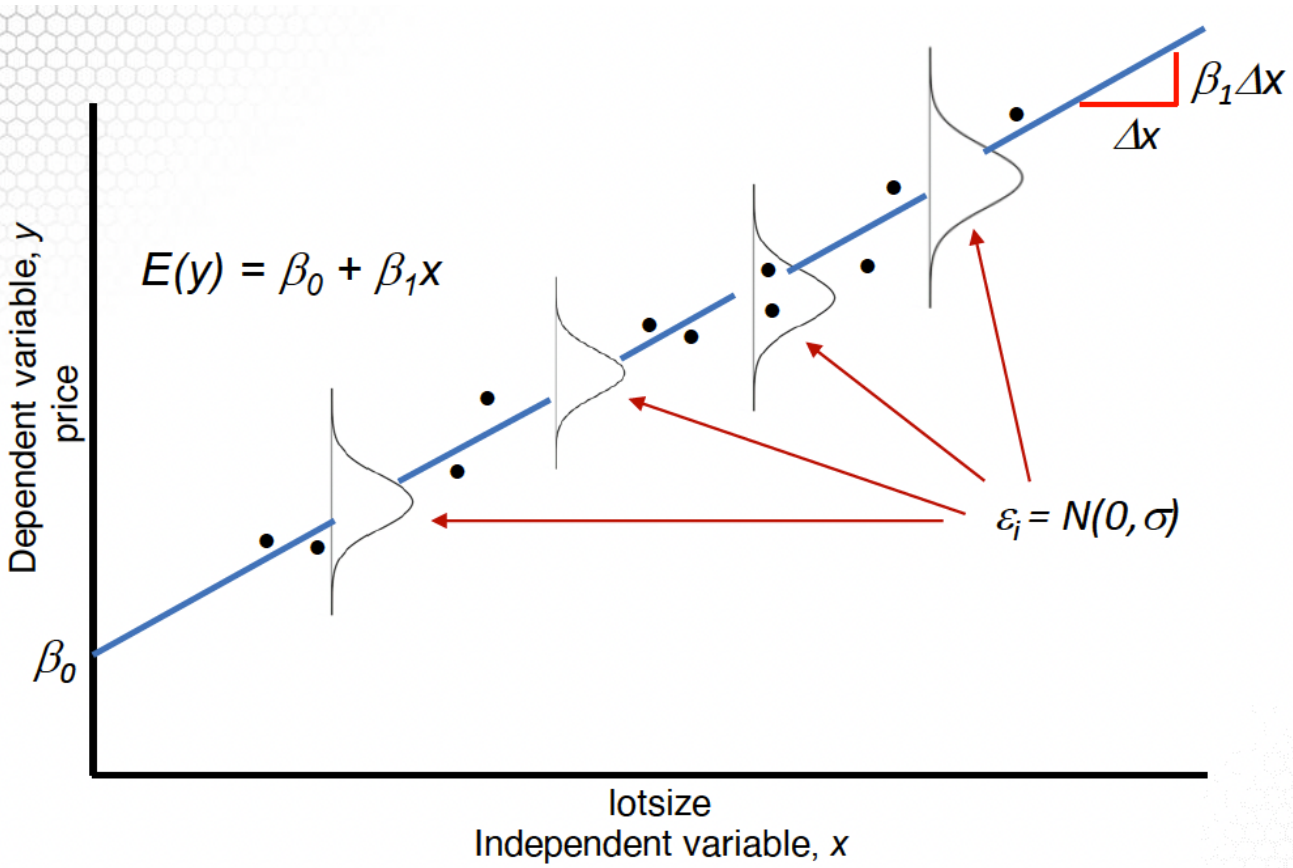
- We observe the data in the Housing dataset (which is a sample)
- We want to build a model for the population

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

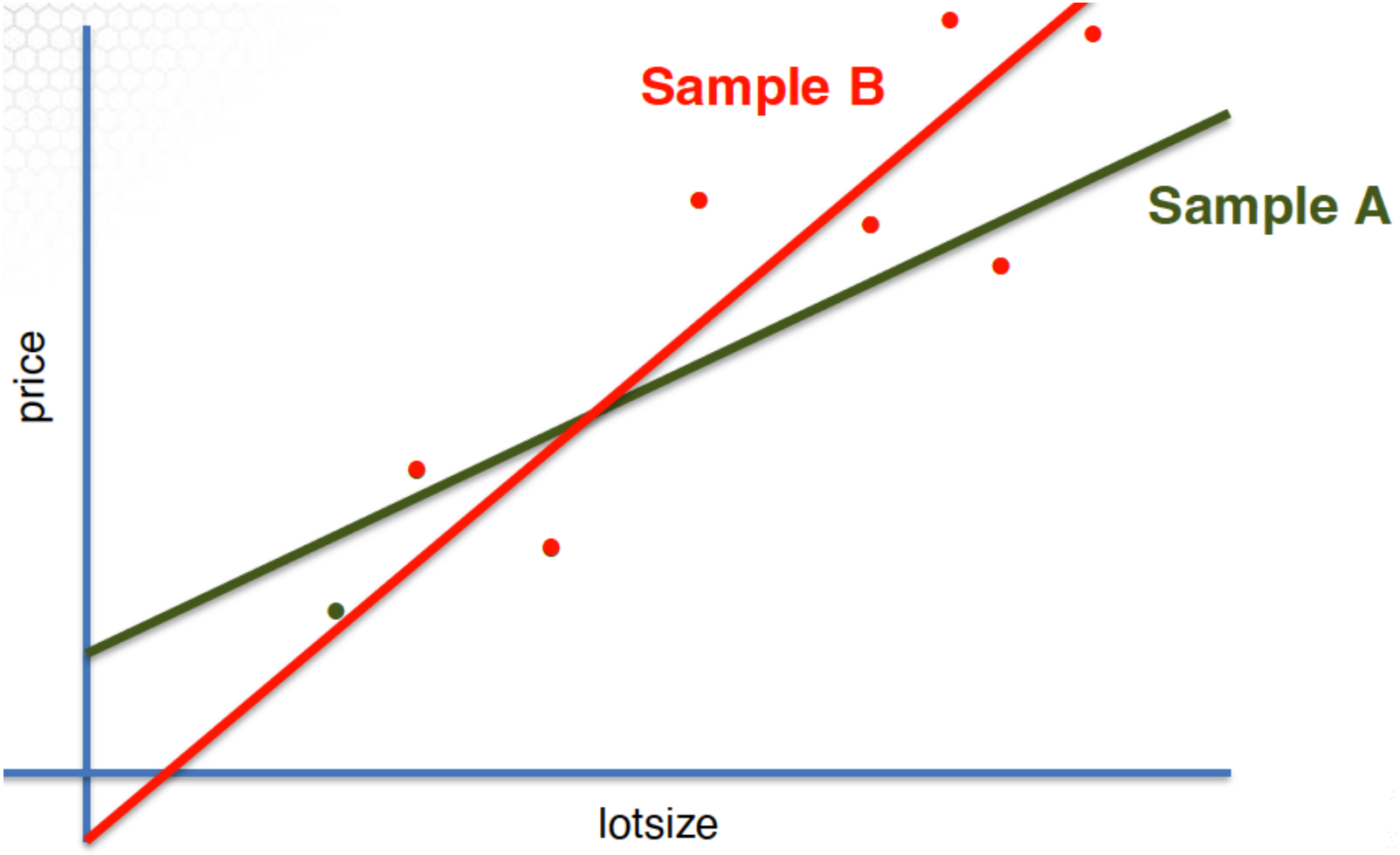
- ϵ_i are independent and identically distributed (i.i.d .) random variables, which are normally distributed with mean 0 and constant standard deviation σ
- However, we do not know β_0, β_1 or σ, so we need to estimate them based on the sample in the Housing dataset
- Using this sample, we are going to build a model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

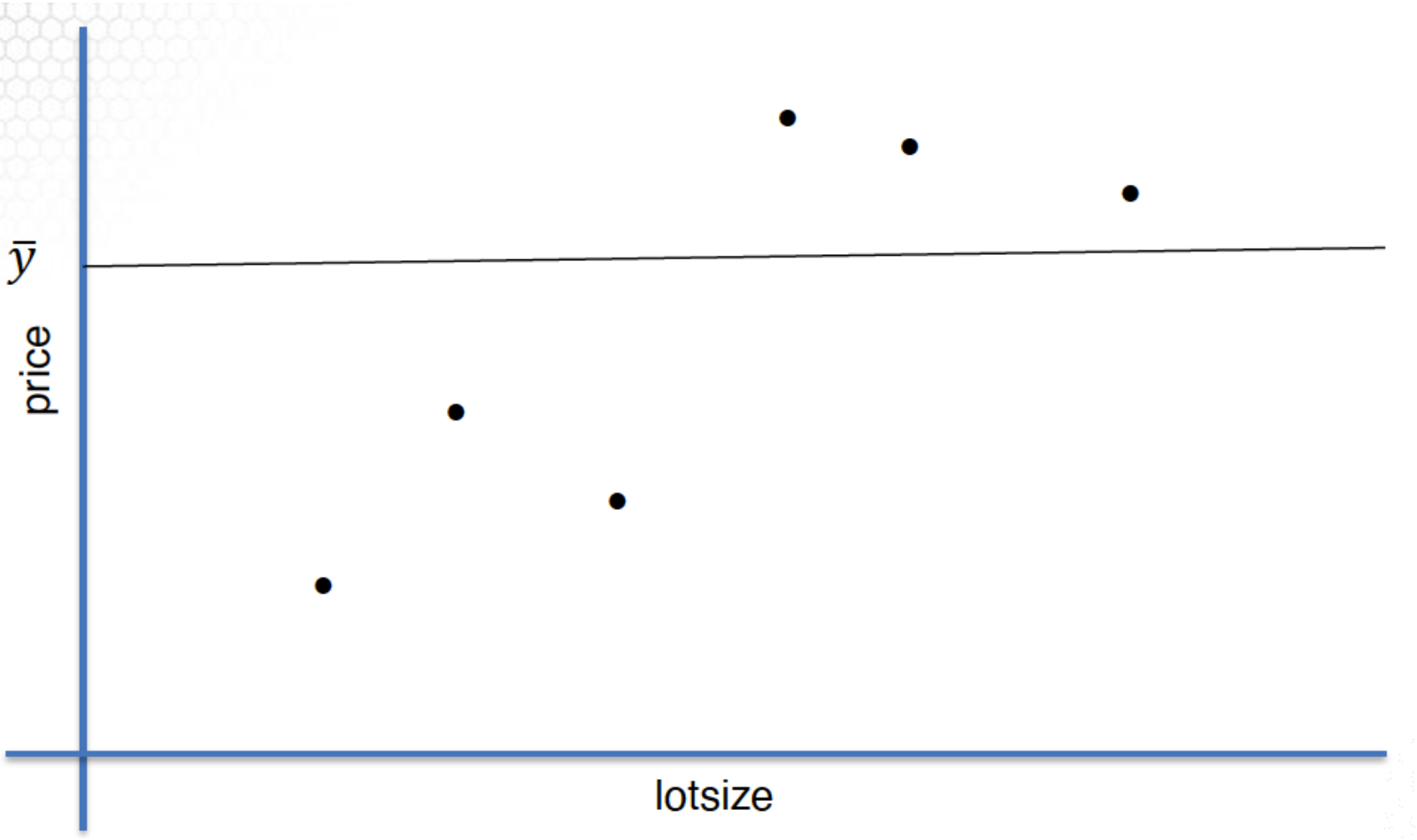
Polulation Model



Estimates of Slope and Intercept Depend on the Sample Being Used

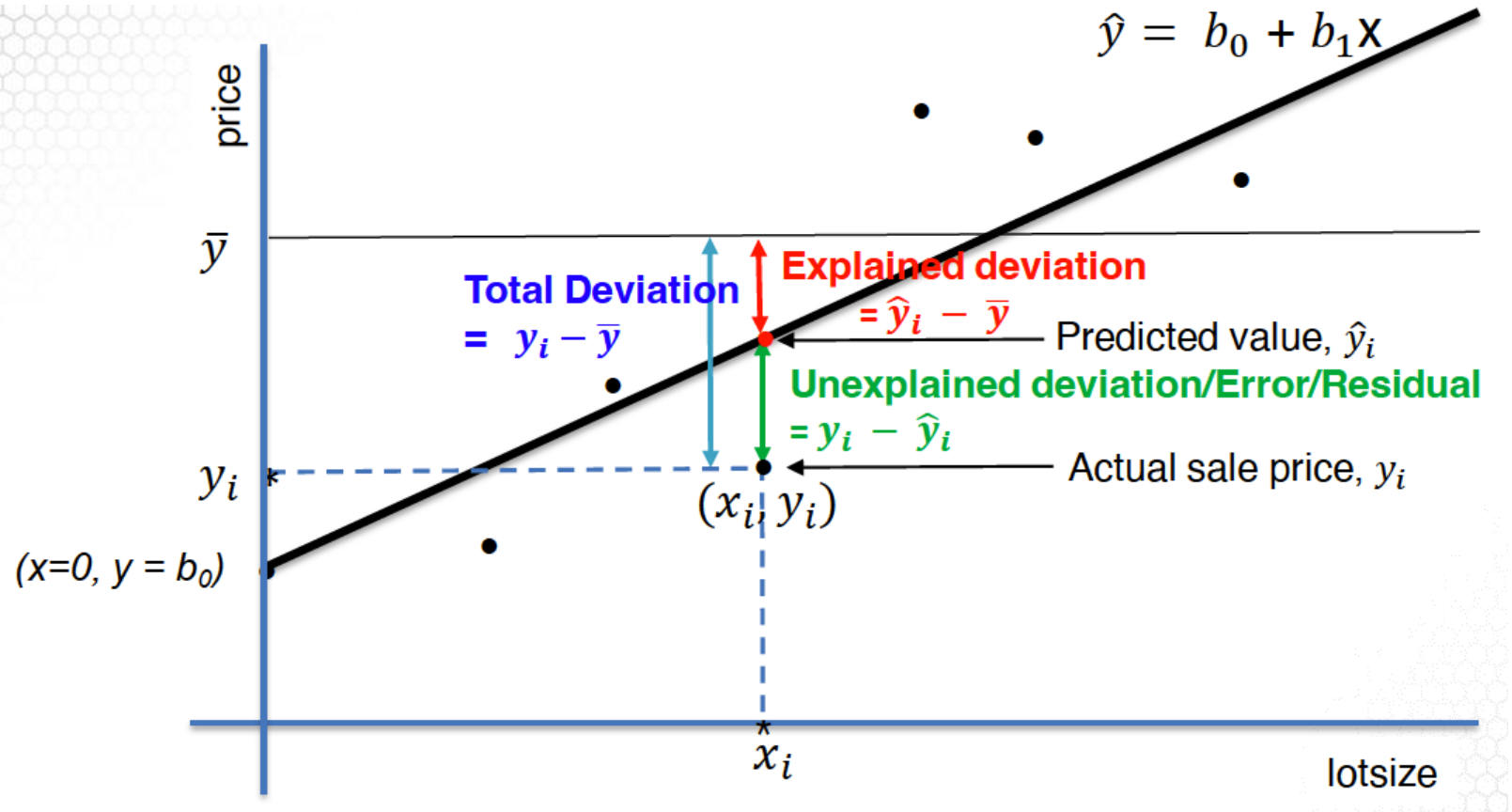


Use Ordinary Least Squares (OLS) to Fit the Line



Use Ordinary Least Squares (OLS) to Fit the Line

Total Deviation = Explained Deviation + Unexplained Deviation



Estimation of the parameters by least squares

- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X. Then $e_i = y_i - \hat{y}_i$ represents the ith **residual**, also called **error**.
- We define the residual sum of squares (RSS) as

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

or equivalently,

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

- **Remark:** residual sum of squares (RSS) is also called sum of squares errors, **SSE**.

Summing the Deviations

$\sum_i (y_i - \bar{y})^2$	$=$	$\sum_i (y_i - \hat{y}_i)^2$	$+$	$\sum_i (\bar{y} - \hat{y}_i)^2$
TSS	$=$	**RSS**	$+$	**ESS**
Total Sum of Squares	$=$	Residual Sum of Squares	$+$	Explained Sum of Squares

- Note that there are other names for the equation:

SST	$=$	SSE	$+$	SSR
Total Sum of Squares	$=$	Sum of Squares Errors	$+$	Sum of Squares Regression

Regression Output R^2 and Adjusted R^2

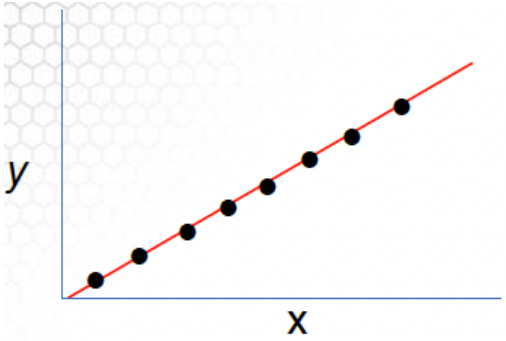
Coefficient of determination R^2 - A measure of the overall strength of the relationship between the dependent variable Y and independent variables X

- $R^2 = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS} = \frac{\text{Explained deviation (ESS)}/}{\text{Total Deviation (TSS)}}$
- $R^2 \rightarrow$ how much of the variation in Y (from the mean) has been explained

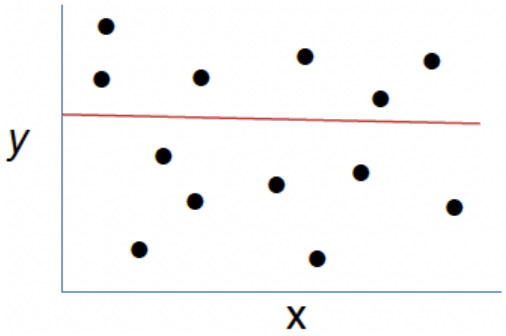
Adjusted R^2

- Adding a penalty for the number of independent variables (p)
- Adjusted $R^2 = 1 - \frac{\frac{RSS}{(n-p-1)}}{\frac{TSS}{(n-1)}} = SSE \sim F$

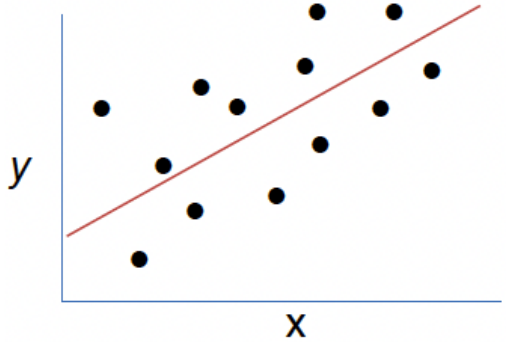
Understand R^2



$R^2 = 1$,
X accounts for all Y variation



$R^2 = 0$,
X accounts for none of the Y variation



$R^2 = 0.75$,
X accounts for most of the Y variation

#

Quiz - $R^2 = 0$, implies that X values account for all of the variation in the Y values

Quiz Answer

- $R^2 = 0$, implies that X values account for all of the variation in the Y values

  `True or False`

- R^2 can take any value from - infinity to + infinity

  `True or False`

Quiz Answer

- $R^2 = 0$, implies that X values account for all of the variation in the Y values

Answer:

FALSE . R 2 = 0 implies that X values account for none of the variation in the Y values

- R^2 can take any value from - infinity to + infinity

Answer:

False. It can only take on values between 0 and 1.

Model Estimation: Approach

$\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$	$\rightarrow$	To estimate (β_0, β_1), we find values that minize RSS (SSE)

The values that minimize **RSS** (SSE):

- $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
  - where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ are the sample means.
  - S_{xy} is the covariance of x, y and S_{xx} is the vairance of x.

Model Estimation: Approach (con'd)

Begin with the minimization problem:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2$$

To solve, take the first order derivatives of the function to be minimized and equate to 0:

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2 = 0$$
$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2 = 0$$

1. Result into a system of linear equation in β_0 and β_1
2. Solve using linear algebra
3. Solutions to the system are $\hat{\beta}_0$ and $\hat{\beta}_1$

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

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

Some Definitions

The following terms are known as sums of squares.

$$S_{xx} = \sum_i (x_i - \bar{x})^2 = \sum_i x_i^2 - n\bar{x}^2$$

$$S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i x_i y_i - n\bar{x}\bar{y} \quad \textit{Why?}$$

$$S_{yy} = \sum_i (y_i - \bar{y})^2 = \sum_i y_i^2 - n\bar{y}^2$$

Fitted Values and Residuals

Given the estimates of β_0 and β_1, we define:

- Fitted values: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- Residuals: $e_i = \hat{\epsilon}_i = y_i - \hat{y}_i$
- Mean squared error: Estimator for σ^2

$$MSE = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{RSS}{n-2}$$

Variance Sampling Distribution

- Recall that $\epsilon = (y_i - (\beta_0 + \beta_1 x_i))$
- $\hat{\epsilon}$ is replaced by $\hat{\epsilon}_i = (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))$
- We lose two degrees of freedom because

$$\beta_0 \leftarrow \hat{\beta}_0$$
$$\beta_1 \leftarrow \hat{\beta}_1$$

Thus, assuming that

$$\epsilon_i \sim N(0, \sigma^2) \rightarrow \hat{\sigma}^2 = MSE \sim \chi^2_{n-2},$$

This is called the sampling distribution of $\hat{\sigma}^2$

Model Parameter Interpretation

Commonly interested in the behavior of β_1

- A positive value of β_1 is consistent with a direct relationship between x and y; e.g., higher values of height are associated with higher values of weight, or lower values of revenue are associated with lower values of profit;
- A negative value of β_1 is consistent with an inverse relationship between x and y; e.g., higher price of a product is associated with lower demand, or a lower inflation rate is associated with a higher savings rate;
- A close-to-zero value of β_1 means that there is not a significant association between x and y.

Model Estimate Interpretation

The Least Squares estimated coefficients have specific interpretations:

- $\hat{\beta}_1$ is the estimated expected change in the response variable associated with one unit of change in the predicting variable;
- $\hat{\beta}_0$ is the estimated expected value of the response variable when the predicting variable equals zero.

Regression Estimators: Properties

For the slope parameter β_1, we can show that

- $E[\hat{\beta}_1] = \beta_1$, so $\hat{\beta}_1$ is unbiased.
- $SE[\hat{\beta}_1]^2 = Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Regression Estimators: Properties

Furthermore, $\hat{\beta}_1$ is a linear combination of $\{Y_1, \cdots, Y_n\}$. If we assume that $\epsilon_i \sim Normal(0, \sigma^2)$, then β_1 is also distributed as

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2})$$

Regression Estimators: Properties

We do not know σ^2. We can replace it by MSE, but then the sampling distribution becomes the t-distribution with $n-2$ df.

$$\hat{\sigma} = MSE = \frac{\sum_i \hat{\epsilon}_i^2}{n-2} \sim \chi^2_{n-2}$$

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MSE}{S_{xx}}}} \sim t_{n-2}$$

Inference for Slope Parameter

Given the sampling distribution of $\hat{\beta}_1$, we can derive confidence intervals and perform hypothesis testing for β_1:

$$\hat{\beta}_1 - (t_{\frac{\alpha}{2},n-2}\sqrt{\frac{MSE}{S_{xx}}}), \hat{\beta}_1 - (t_{\frac{\alpha}{2},n-2}\sqrt{\frac{MSE}{S_{xx}}})$$

Confidence Interval Derivation

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MSE}{S_{xx}}}} \sim t_{n-2}$$

$1 - \alpha$ considence interval:

$\hat{\beta}_1$	$\pm$	$t_{\frac{\alpha}{2},n-2}$	$\sqrt{\frac{MSE}{S_{xx}}}$
$\uparrow$		$\uparrow$	
Estimate of β_1		t-critical point	Standard Deviation/error or $\hat{\beta}_1$
		Sampling distribution of $\hat{\beta}_1$ is t_{n-2}	Variance$[\hat{\beta}_1] = \frac{\sigma^2}{S_{xx}}$ replace σ^2 with MSE

Testing Significance of Regression

One way we can test statistical significance is to use the t-test for

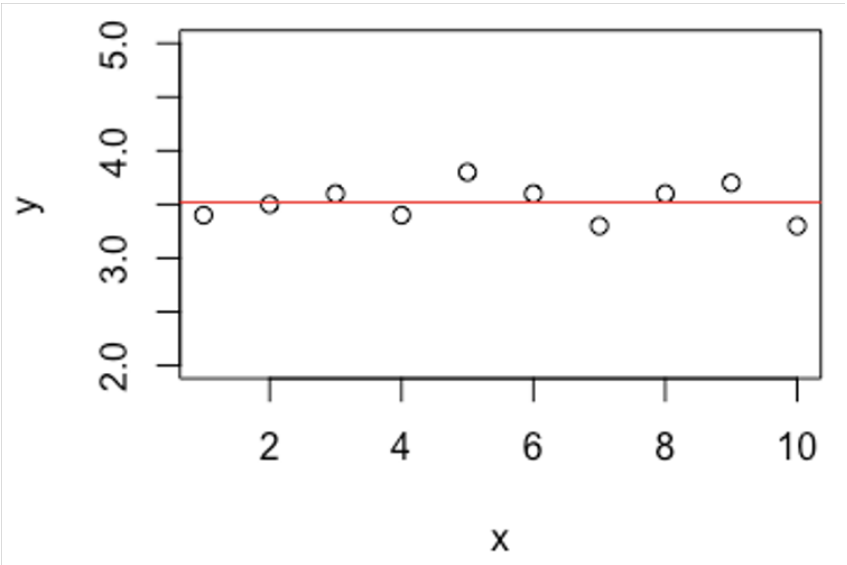
$$H_0 : \beta_1 = 0 \; H_a : \beta_1 \neq 0$$

$$t - value = \frac{\hat{\beta}_1 - 0}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} = \frac{\beta_1 \hat{S}_{xx}}{\hat{\sigma}}$$

We reject H_0 if $|t - value|$ is large. If the null hypothesis is rejected, we interpret this as β_1 being **statistically significant**.

Testing Significance of Regression

- A regression model is useful if the relationship between y and x is significant.
- If there is no relationship, then $\beta_1 = 0$. It means we cannot express y in terms of x.



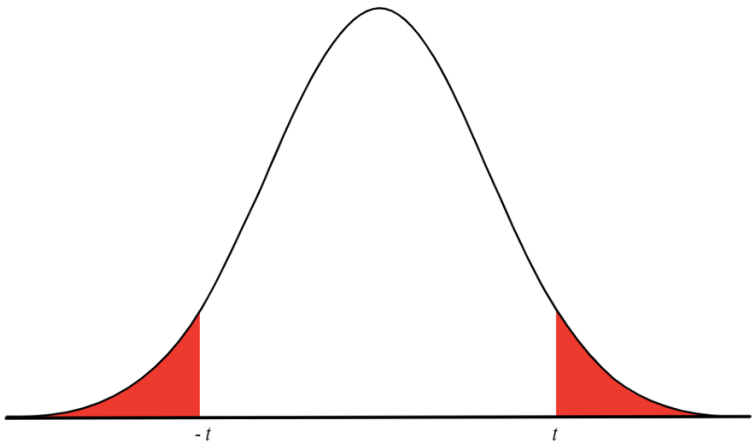
Testing Significance of Regression

Hypothesis: $H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq= 0$

- Test statistic $t - value = \frac{\hat{\beta}_1 - 0}{\sqrt{\frac{MSE}{S_{xx}}}} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \sim t_{n-2}$ where SE means 'standard error' and

Testing Significance of Regression



2 x the area to the right of $|t|$

Critical Region:

$|t - value| > t_{\frac{\alpha}{2}(n-2)} \rightarrow$ Reject H_0

P-value: $p = 2 \times P(t_{n-2} > |t_0|) < \alpha \rightarrow$ Reject H_0

Inference for Intercept Parameter

Now, let's see the intercept

- $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$
- $\rightarrow E[\hat{\beta}_0] = E[\bar{Y}] - E[\hat{\beta}_1]\bar{x} = \beta_0$
- $Var(\hat{\beta}_0) = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})$
- Confidence interval (replace σ^2 with MSE):

$$(\hat{\beta}_0 - t_{\frac{\alpha}{2},n-2}\sqrt{MSE(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}, \hat{\beta}_0 + t_{\frac{\alpha}{2},n-2}\sqrt{MSE(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})})$$

Example:

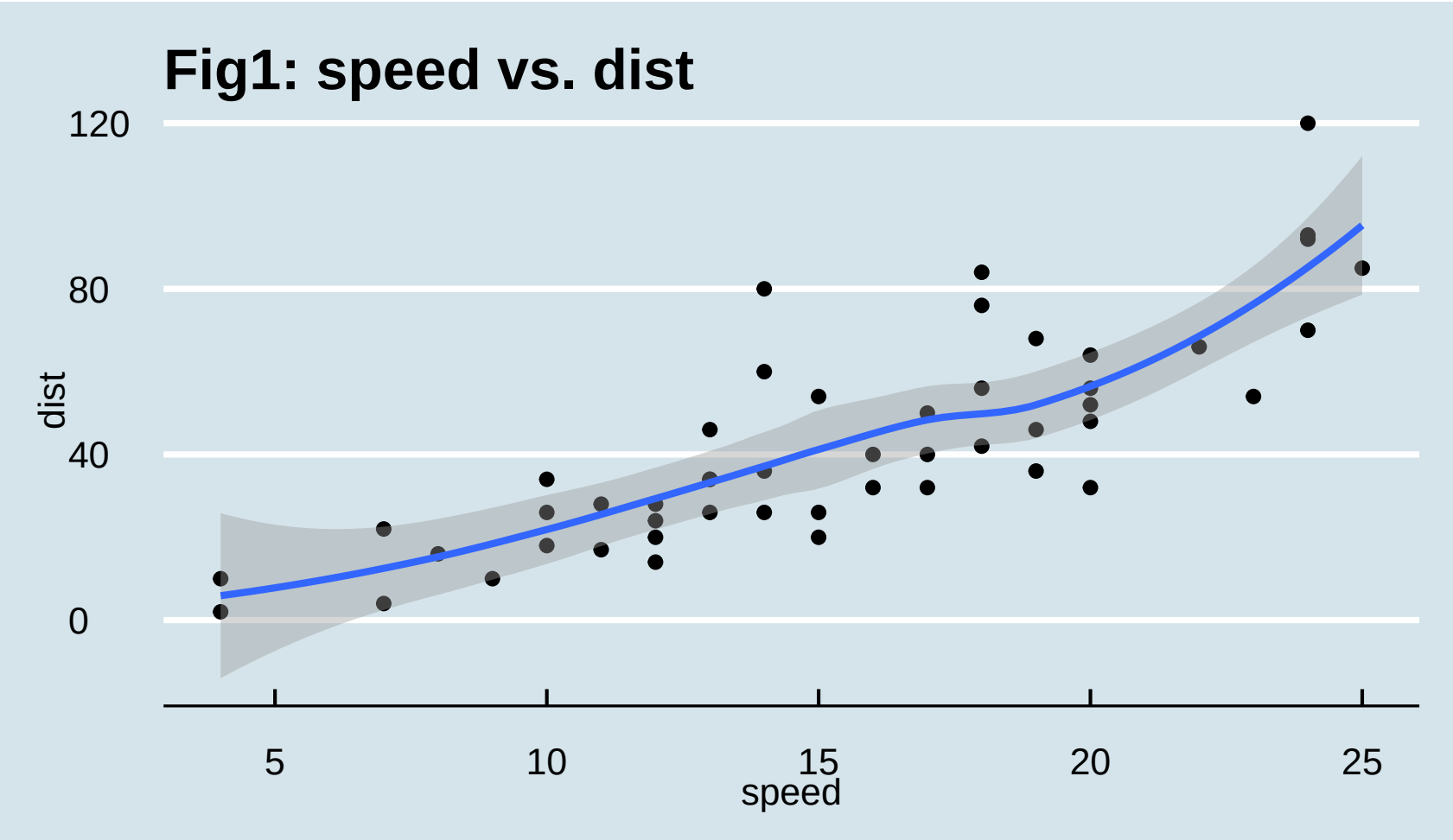
```
head(cars)
```

```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
## 4     7   22
## 5     8   16
## 6     9   10
```

Example

```
## Warning: package 'ggthemes' was built under R version 4.
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y
```



Example

```
model<-lm(dist~speed,data=cars)
summary(model)
```

Call: lm(formula = dist ~ speed, data = cars)

Residuals: Min 1Q Median 3Q Max -29.069 -9.525 -2.272 9.215 43.201

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791 6.7584 -2.601 0.0123 *
speed 3.9324 0.4155 9.464 1.49e-12 *** — Signif. codes: 0 ‘’ ***0.001*** ’’ *0.01* ’’ 0.05 ‘.’ 0.1 ’ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom Multiple R-squared: 0.6511, Adjusted R-squared: 0.6438 F-statistic: 89.57 on 1 and 48 DF, p-value: 1.49e-12

Example

```
confint(model,level=0.99)
```

```
                0.5 %     99.5 %
(Intercept) -35.706610 0.5484205
speed         2.817919 5.0468988
```

Example

```
new_speed<-data.frame(speed=21)
predict(model,newdata=new_speed, interval='confidence',leve
```

```
    fit       lwr       upr
1 65.00149 59.65934 70.34364
```

Multiple Linear Regression: Notation

Notation	Meaning
$i = 1, 2, \cdots, n$	i refers to the i th observation or record in a data set of records (typically a sample of the population)
$X_1 = 11 \quad X_{2,1} \quad \cdots \quad X_{p1}$ $X_2 = 12 \quad X_{2,2} \quad \cdots \quad X_{p2}$ $X_n = 1n \quad X_{2,n} \quad \cdots \quad X_{pn}$	n observations of the p explanatory variables
$y_1, y_2, \cdots, y_n$	n observations of the dependent variable
$\beta_0, \beta_1, \cdots, \beta_p$	Parameters of the regression line for the entire population
$\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$	Pestimates of $\beta_0, \beta_1, \cdots, \beta_p$

Multiple Linear Regression, with p Explanatory Variables

- Predict for Y at X_i

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \hat{\beta}_1 X_{2,i} + \cdots + \hat{\beta}_1 X_{p,i}$$

- Residual:

$e_i = y_i - \hat{y}_i$

Our goal is finding $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$ to minimize the residual sum of squares

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \hat{\beta}_1 X_{2,i} + \cdots + \hat{\beta}_1 X_{p,i})^2)$$

A Case Study: Using R to Estimate a Linear Model

Using the Housing Dataset in the Ecdat package in R

```
## Loading required package: Ecfun

##
## Attaching package: 'Ecfun'

## The following object is masked from 'package:base':
##
##     sign

##
## Attaching package: 'Ecdat'

## The following object is masked from 'package:datasets':
##
##     Orange
```