Data Analysis

Simple (univariate) linear regression

National Research University Higher School of Economics
Master's Program "Big Data Systems"

Fall 2018

Correlation Analysis

- The population correlation coefficient is denoted ρ (the Greek letter rho)
- The sample correlation coefficient is

$$r = \frac{s_{xy}}{s_x s_y}$$

where

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Hypothesis Test for Correlation

- To test the null hypothesis of no linear association,

$$H_0 : \rho = 0$$

the test statistic follows the Student's t distribution with (n – 2) degrees of freedom:

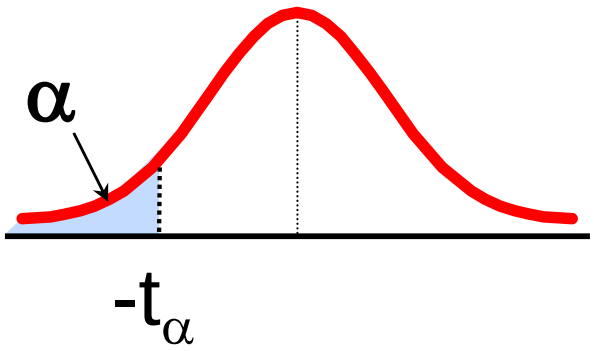
$$t = \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}}$$

Decision Rules

Hypothesis Test for Correlation

Lower-tail test:

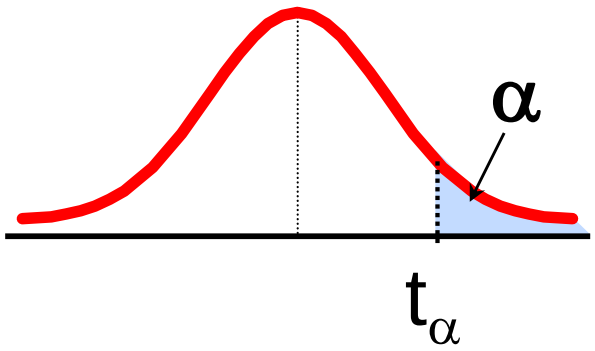
$H_0: \rho \geq 0$
$H_1: \rho < 0$



Reject H_0 if $t < -t_{n-2,\,\alpha}$

Upper-tail test:

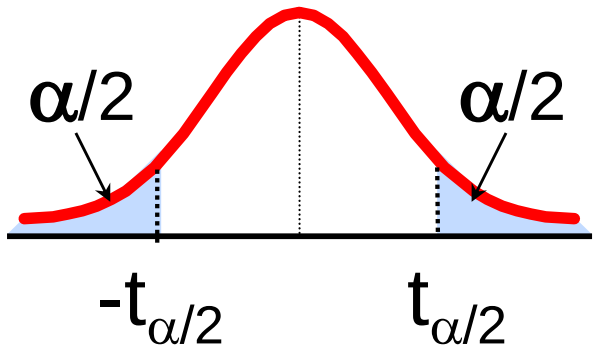
$H_0: \rho \leq 0$
$H_1: \rho > 0$



Reject H_0 if $t > t_{n-2,\,\alpha}$

Two-tail test:

$H_0: \rho = 0$
$H_1: \rho \neq 0$



Reject H_0 if $t < -t_{n-2,\,\alpha/2}$ or $t > t_{n-2,\,\alpha/2}$

Where $t = \dfrac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}}$ has $n - 2$ d.f.

Introduction to Regression Analysis

- Regression analysis is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable

Dependent variable: the variable we wish to explain
(also called the endogenous variable)

Independent variable: the variable used to explain the dependent variable
(also called the exogenous variable)

Linear Regression Model

- Regression analysis is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable

Dependent variable: the variable we wish to explain
(also called the endogenous variable)

Independent variable: the variable used to explain the dependent variable
(also called the exogenous variable)

Simple Linear Regression Model

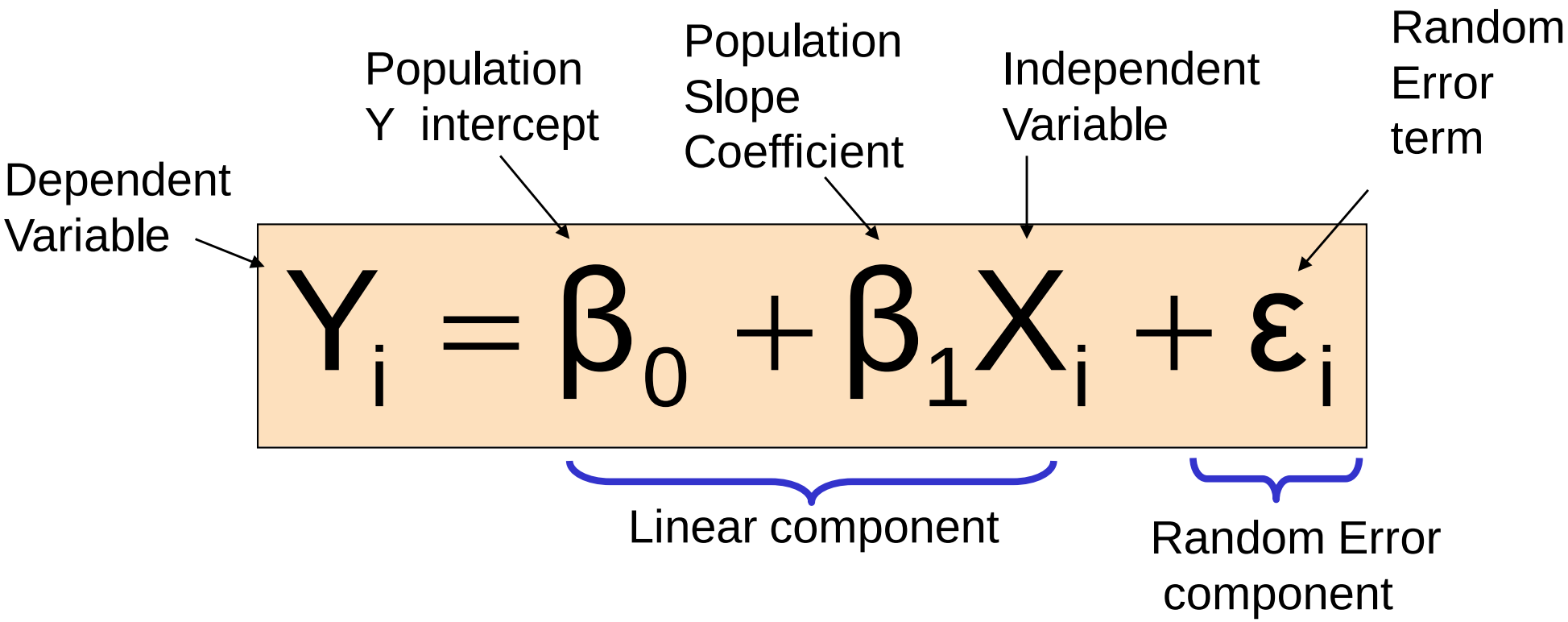
- The relationship between X and Y is described by a linear function
- Changes in Y are assumed to be caused by changes in X
- Linear regression population equation model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

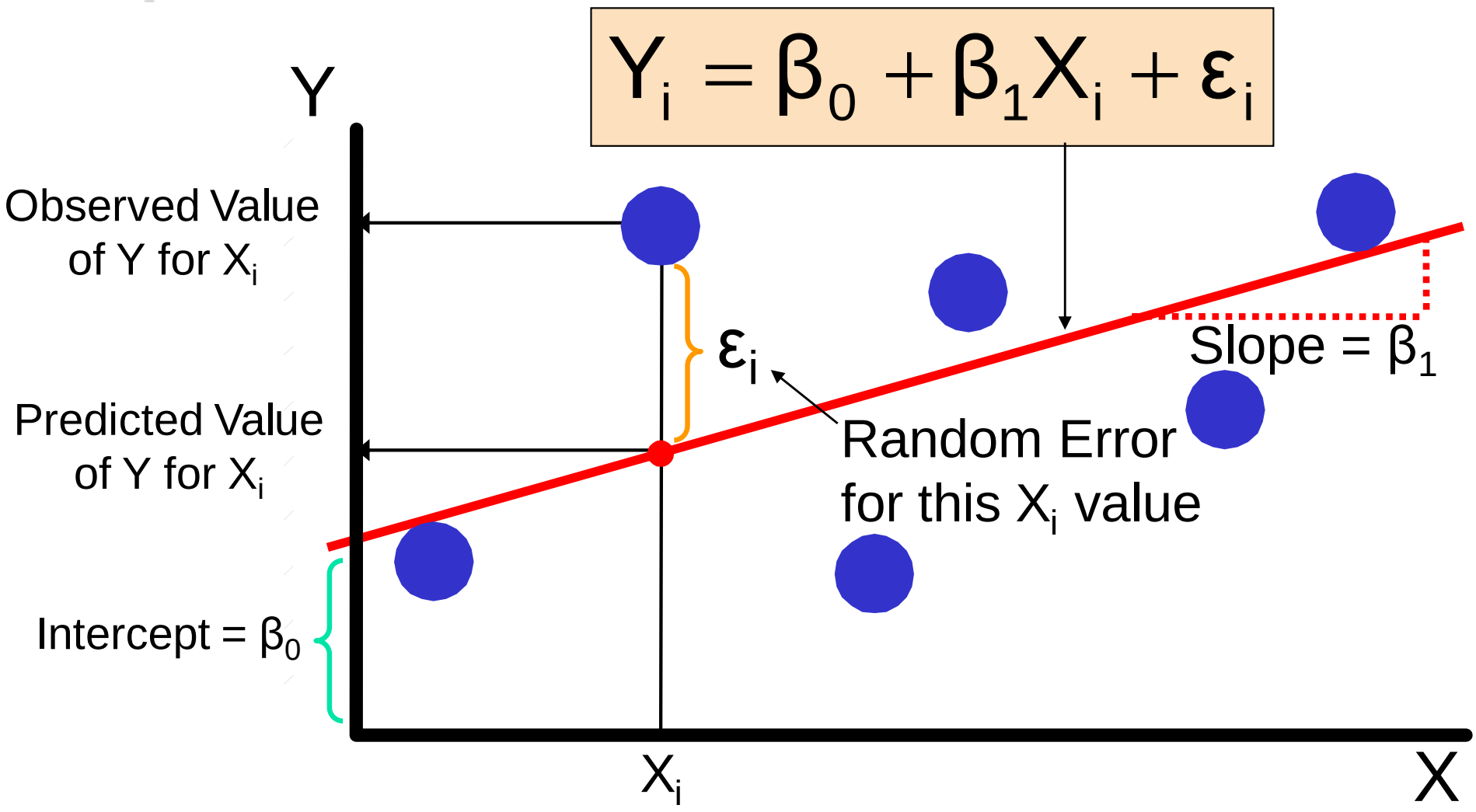
- Where β_0 and β_1 are the population model coefficients and ε is a random error term.

Simple Linear Regression Model (2)

The population regression model:

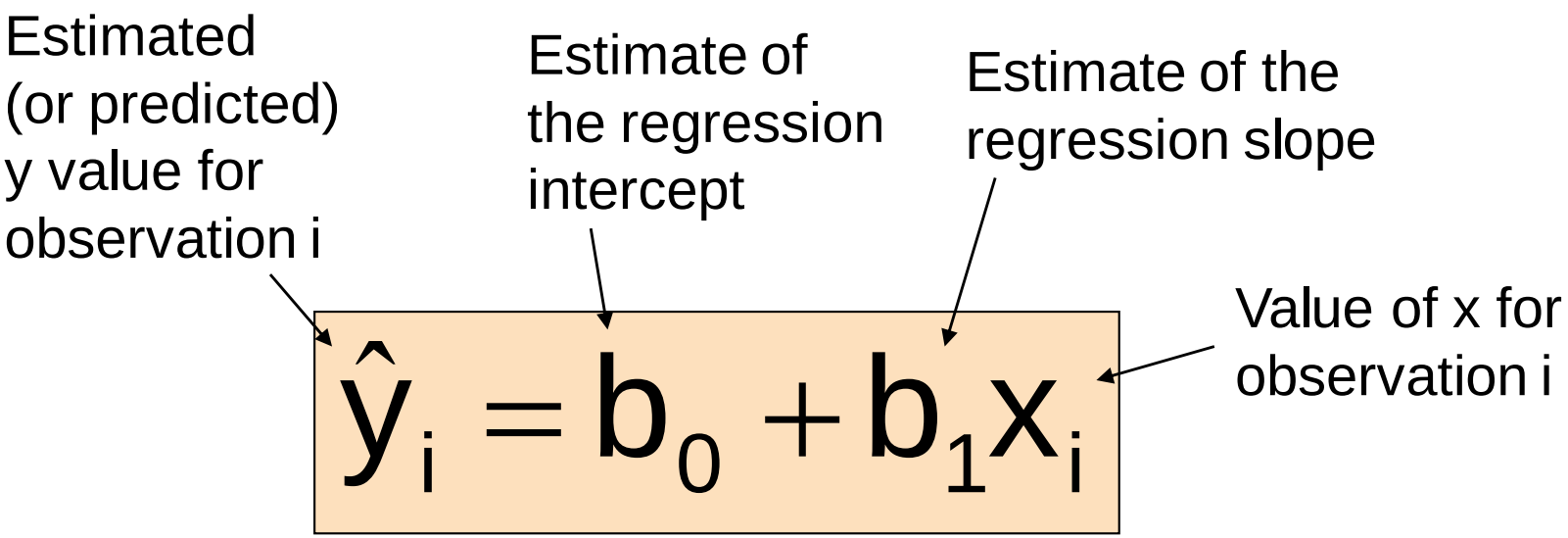


Simple Linear Regression Model (3)



Simple Linear Regression Equation

The simple linear regression equation provides an estimate of the population regression line



The individual random error terms e_i have a mean of zero

$$e_i = (y_i - \hat{y}_i) = y_i - (b_0 + b_1 x_i)$$

Simple linear regression: things to remember

- Get the summary statistics of the regression model
- Interpret coefficient of determination R^2
- Regression slope, β_1:
  - Assess the standard error of β_1
  - Perform and interpret the t-test for non-zero β_1. Confirm that analysis by examining confidence interval for β_1
  - As an additional check, do the F-test (the same purpose)
- Predictions:
  - Interpolation or extrapolation (be cautious)
  - Prediction for mean
  - Prediction for a single value
- Analysis of residuals = model checking. Important!

Least Squares Estimators

- b_0 and b_1 are obtained by finding the values of b_0 and b_1 that minimize the sum of the squared differences between y and $\hat{y}$:

$$\begin{aligned} \min \ SSE &= \min \sum e_i^2 \\ &= \min \sum (y_i - \hat{y}_i)^2 \\ &= \min \sum [y_i - (b_0 + b_1 x_i)]^2 \end{aligned}$$

Differential calculus is used to obtain the coefficient estimators b_0 and b_1 that minimize SSE

Least Squares Estimators (2)

- The slope coefficient estimator is

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = r_{xy}\frac{s_Y}{s_X}$$

- And the constant or y-intercept is

$$b_0 = \bar{y} - b_1\bar{x}$$

- The regression line always goes through the mean $\bar{x}$, $\bar{y}$

Finding the Least Squares Equation

- The coefficients b_0 and b_1, and other regression results in this chapter, will be found using a computer
  - Hand calculations are tedious
  - Statistical routines are built into Excel
  - Other statistical analysis software can be used

Linear Regression Model Assumptions

- The true relationship form is linear (Y is a linear function of X, plus random error)
- The error terms, ε_i are independent of the x values
- The error terms are random variables with mean 0 and constant variance, σ^2

  (the constant variance property is called homoscedasticity)

$$E[\varepsilon_i] = 0 \quad \text{and} \quad E[\varepsilon_i^2] = \sigma^2 \quad \text{for } (i = 1, \ldots, n)$$

- The random error terms, ε_i, are not correlated with one another, so that

$$E[\varepsilon_i \varepsilon_j] = 0 \quad \text{for all } i \neq j$$

Interpretation of the Slope and the Intercept

- b_0 is the estimated average value of y when the value of x is zero (if $x = 0$ is in the range of observed x values)
- b_1 is the estimated change in the average value of y as a result of a one-unit change in x

Simple Linear Regression Example

- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)

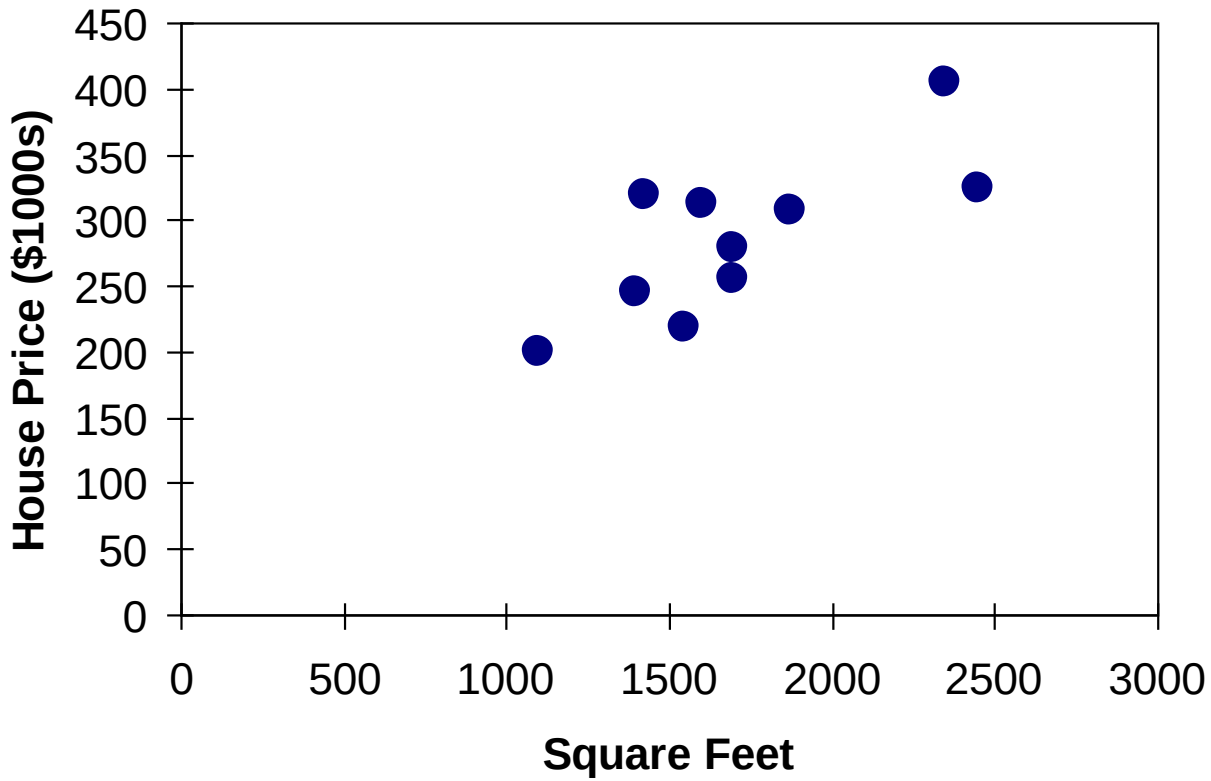
- A random sample of 10 houses is selected
  - Dependent variable (Y) = house price in $1000s
  - Independent variable (X) = square feet

Sample Data for House Price Model

House Price in $1000s (Y)	Square Feet (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Graphical Presentation

- House price model: scatter plot



Regression Table

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

The regression equation is:

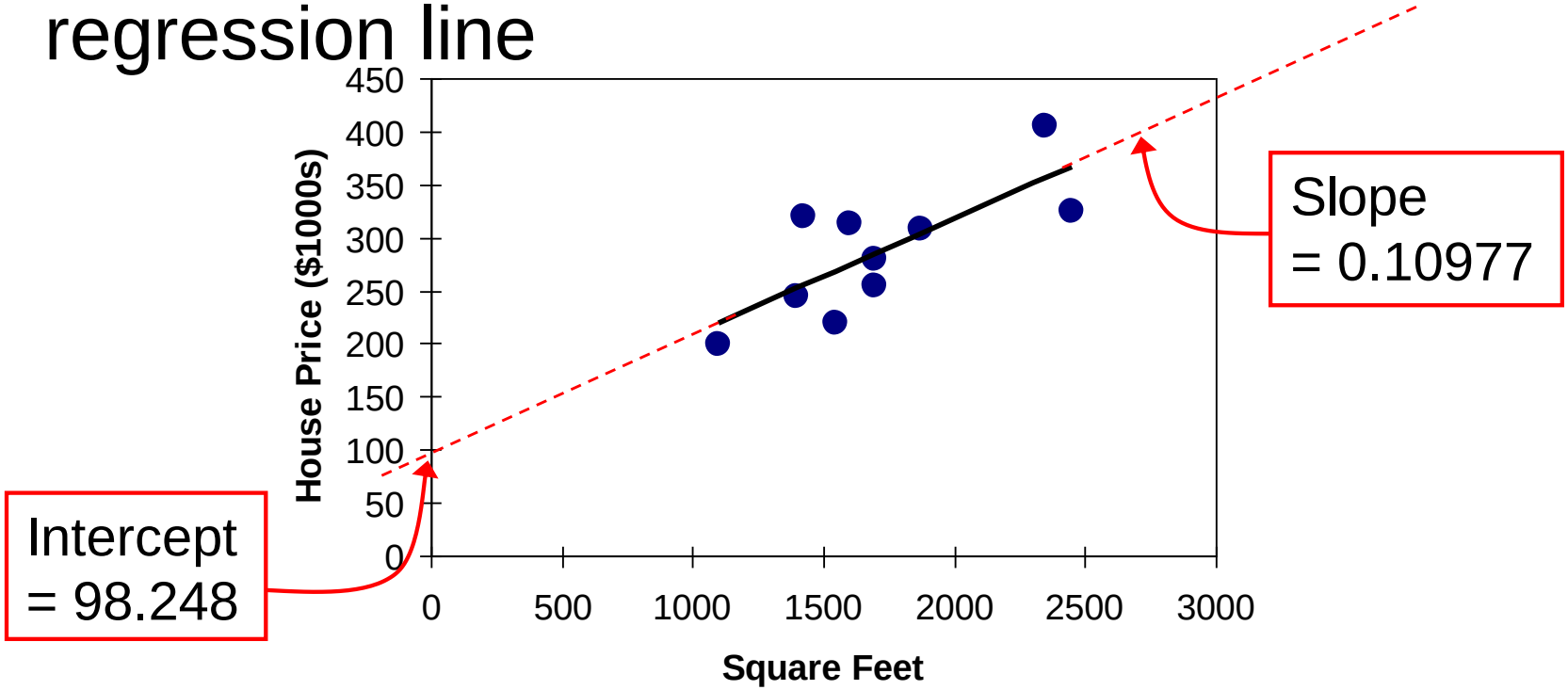
$$\widehat{\text{house price}} = 98.24833 + 0.10977\ (\text{square feet})$$

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Graphical Presentation

- House price model: scatter plot and regression line



$$\widehat{\text{house price}} = 98.24833 + 0.10977\ (\text{square feet})$$

Measures of Variation

- Total variation is made up of two parts:

$$SST = SSR + SSE$$

Total Sum of Squares	Regression Sum of Squares	Error Sum of Squares
$SST = \sum (y_i - \bar{y})^2$	$SSR = \sum (\hat{y}_i - \bar{y})^2$	$SSE = \sum (y_i - \hat{y}_i)^2$

where:

$\bar{y}$ = Average value of the dependent variable

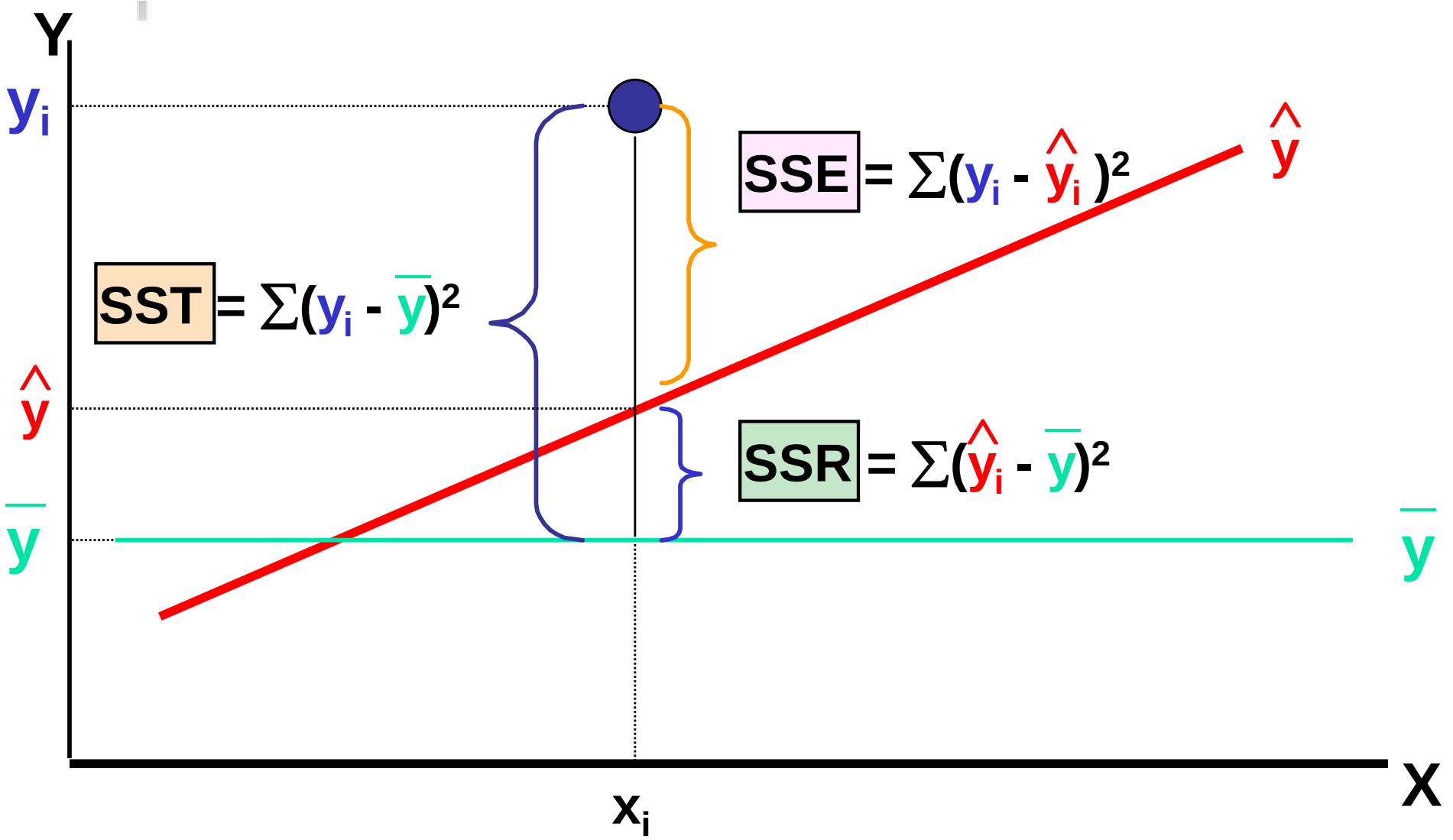
y_i = Observed values of the dependent variable

$\hat{y}_i$ = Predicted value of y for the given x_i value

Measures of Variation (2)

- SST = total sum of squares
  - Measures the variation of the y_i values around their mean, $\bar{y}$
- SSR = regression sum of squares
  - Explained variation attributable to the linear relationship between x and y
- SSE = error sum of squares
  - Variation attributable to factors other than the linear relationship between x and y

Measures of Variation (3)



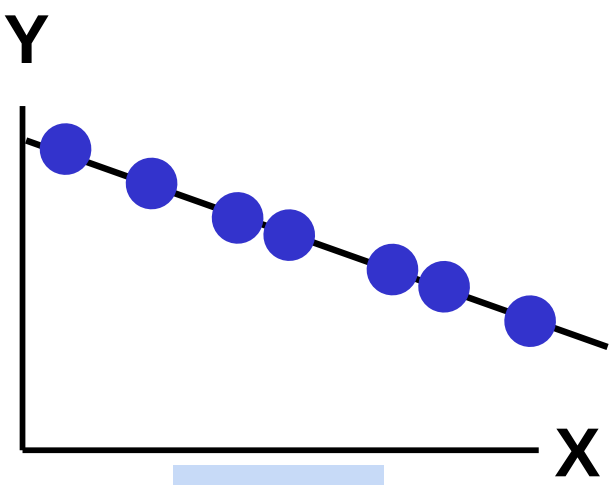
Coefficient of Determination, R^2

- The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable
- The coefficient of determination is also called R-squared and is denoted as R^2

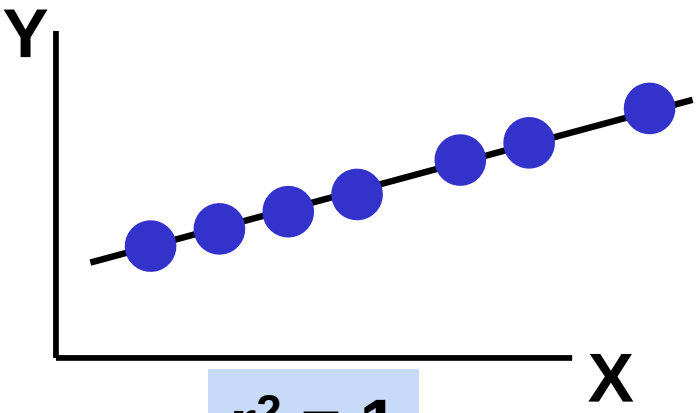
$$R^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

note: $0 \leq R^2 \leq 1$

Examples of R^2 Values



$r^2 = 1$



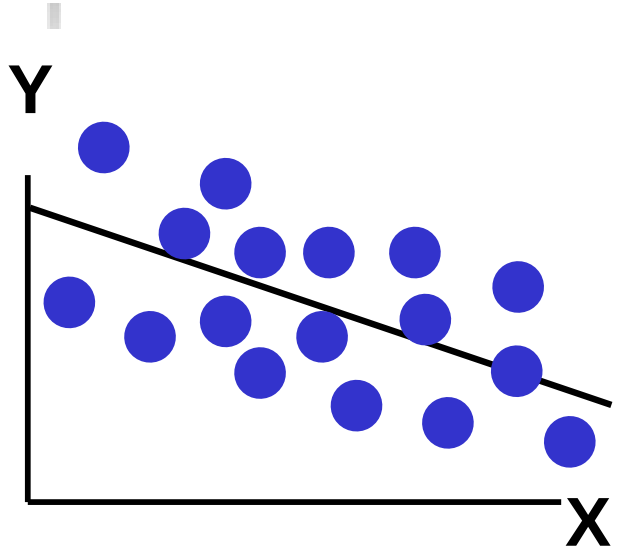
$r^2 = 1$

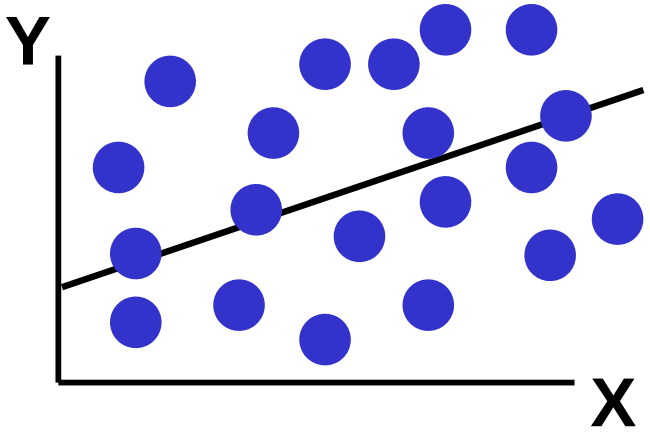
$r^2 = 1$

Perfect linear relationship between X and Y:

100% of the variation in Y is explained by variation in X

Examples of R^2 Values (2)



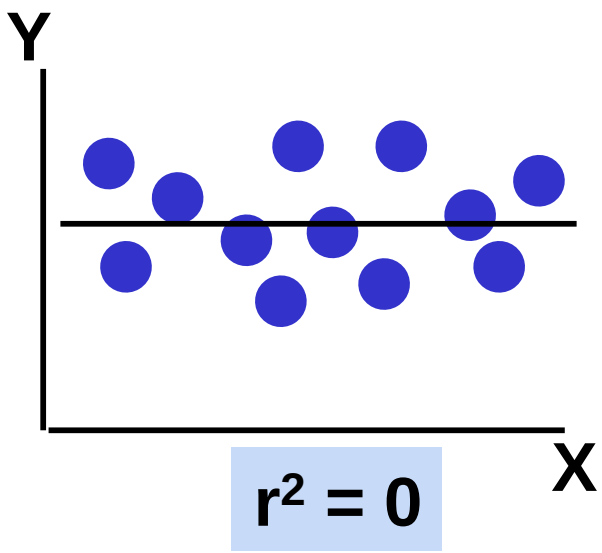


$0 < r^2 < 1$

Weaker linear relationships between X and Y:

Some but not all of the variation in Y is explained by variation in X

Examples of R^2 Values (3)



$r^2 = 0$

No linear relationship between X and Y:

The value of Y does not depend on X. (None of the variation in Y is explained by variation in X)

Regression Table

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$R^2 = \frac{SSR}{SST} = \frac{18934.9348}{32600.5000} = 0.58082$$

58.08% of the variation in house prices is explained by variation in square feet

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Correlation coefficient and R^2

- The coefficient of determination, R^2, for a simple regression is equal to the simple correlation squared

$$R^2 = r_{xy}^2$$

Estimation of Model Error Variance

- An estimator for the variance of the population model error is

$$\hat{\sigma}^2 = s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SSE}{n-2}$$

- Division by n – 2 instead of n – 1 is because the simple regression model uses two estimated parameters, b_0 and b_1, instead of one

$s_e = \sqrt{s_e^2}$ is called the standard error of the estimate

Regression Table

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

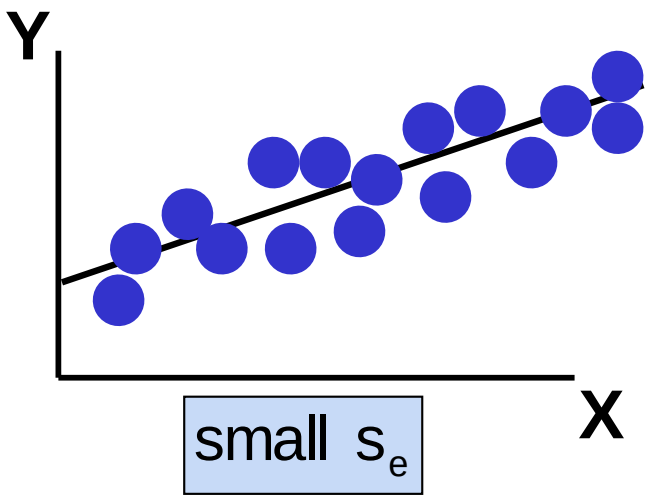
$$s_e = 41.33032$$

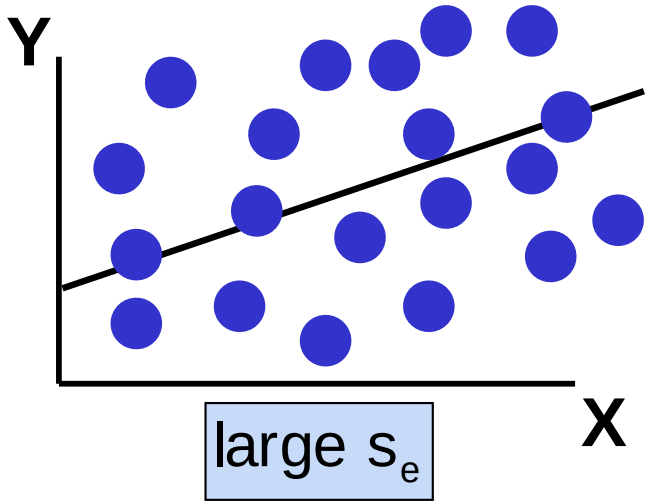
ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Comparing Standard Errors

- s_e is a measure of the variation of observed y values from the regression line





The magnitude of s_e should always be judged relative to the size of the y values in the sample data

i.e., s_e = \$41.33K is moderately small relative to house prices in the \$200 - \$300K range

Inferences About the Regression Model

- The variance of the regression slope coefficient (b_1) is estimated by

$$s_{b_1}^2 = \frac{s_e^2}{\sum (x_i - \bar{x})^2} = \frac{s_e^2}{(n-1)s_x^2}$$

where:

s_{b_1} = Estimate of the standard error of the least squares slope

$$s_e = \sqrt{\frac{SSE}{n-2}}$$ = Standard error of the estimate

Regression Table

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

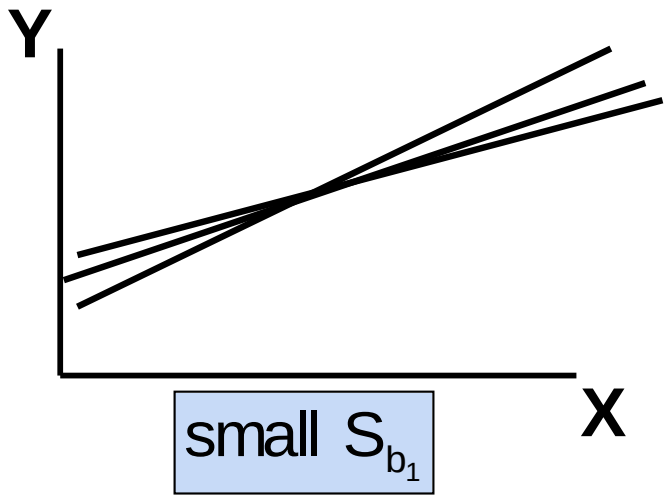
$$S_{b_1} = 0.03297$$

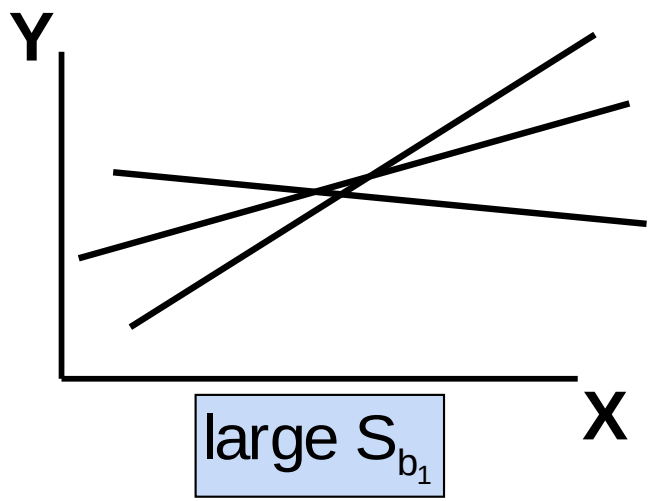
ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Comparing Standard Errors of the Slope

S_{b_1} is a measure of the variation in the slope of regression lines from different possible samples





Inference about the Slope: t Test

- t test for a population slope
  - Is there a linear relationship between X and Y?
- Null and alternative hypotheses

H_0: $\beta_1 = 0$ (no linear relationship)
H_1: $\beta_1 \neq 0$ (linear relationship does exist)

- Test statistic

$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

$$d.f. = n - 2$$

where:

b_1 = regression slope coefficient

β_1 = hypothesized slope

s_{b1} = standard error of the slope

Inference about the Slope: t Test (2)

House Price in $1000s (y)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Estimated Regression Equation:

$$\widehat{\text{house price}} = 98.25 + 0.1098\ (\text{sq.ft.})$$

The slope of this model is 0.1098

Does square footage of the house affect its sales price?

Inferences about the Slope: t Test Example

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

From Excel output:

b_1 S_{b_1}

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{0.10977 - 0}{0.03297} = 3.32938$$

Inferences about the Slope: t Test Example (2)

Test Statistic: **t = 3.329**

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

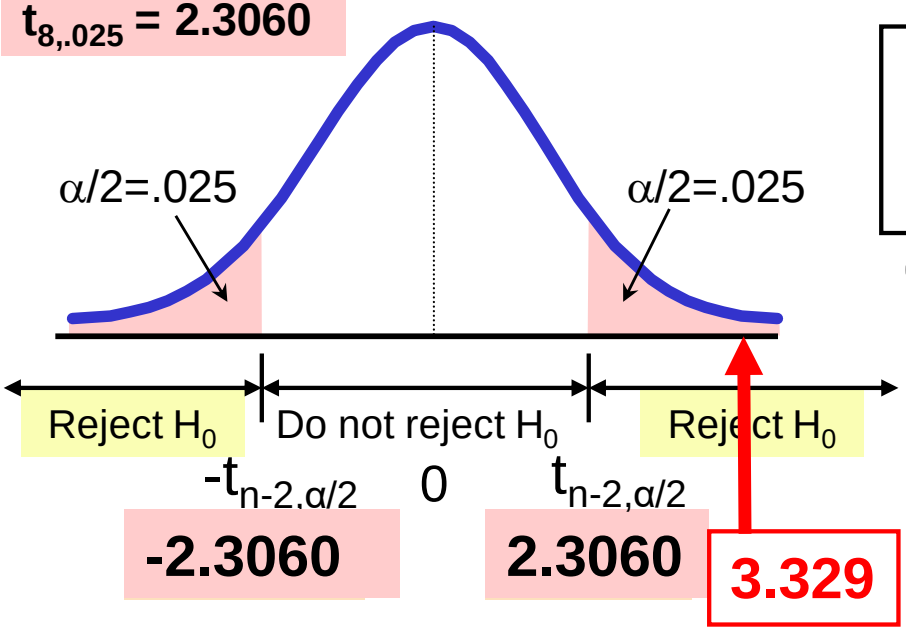
From Excel output:

b_1 S_{b_1} t

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	98.24833	58.03348	1.69296	0.12892
Square Feet	0.10977	0.03297	3.32938	0.01039

d.f. = 10-2 = 8

$t_{8,.025} = 2.3060$



Decision:

Reject H_0

Conclusion:

There is sufficient evidence that square footage affects house price

Inferences about the Slope: t Test Example (3)

P-value = **0.01039**

H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$

From Excel output:

P-value

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**

This is a two-tail test, so the p-value is

$P(t > 3.329)+P(t < -3.329) = 0.01039$

(for 8 d.f.)

Decision: P-value < α so Reject H_0

Conclusion:

There is sufficient evidence that square footage affects house price

Confidence Interval Estimate for the Slope

Confidence Interval Estimate of the Slope:

$$b_1 - t_{n-2,\alpha/2} s_{b_1} < \beta_1 < b_1 + t_{n-2,\alpha/2} s_{b_1}$$

d.f. = n - 2

Excel Printout for House Prices:

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**	**-35.57720**	**232.07386**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**	**0.03374**	**0.18580**

At 95% level of confidence, the confidence interval for the slope is (0.0337, 0.1858)

Confidence Interval Estimate for the Slope (2)

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**	**-35.57720**	**232.07386**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**	**0.03374**	**0.18580**

Since the units of the house price variable is $1000s, we are 95% confident that the average impact on sales price is between $33.70 and $185.80 per square foot of house size

This 95% confidence interval does not include 0.

Conclusion: There is a significant relationship between house price and square feet at the .05 level of significance

F-Test for Significance

- F Test statistic:

$$F = \frac{MSR}{MSE}$$

where

$$MSR = \frac{SSR}{k}$$

$$MSE = \frac{SSE}{n-k-1}$$

where F follows an F distribution with k numerator and $(n-k-1)$ denominator degrees of freedom

(k = the number of independent variables in the regression model)

Regression Table

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$F = \frac{MSR}{MSE} = \frac{18934.9348}{1708.1957} = 11.0848$$

With 1 and 8 degrees of freedom

P-value for the F-Test

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

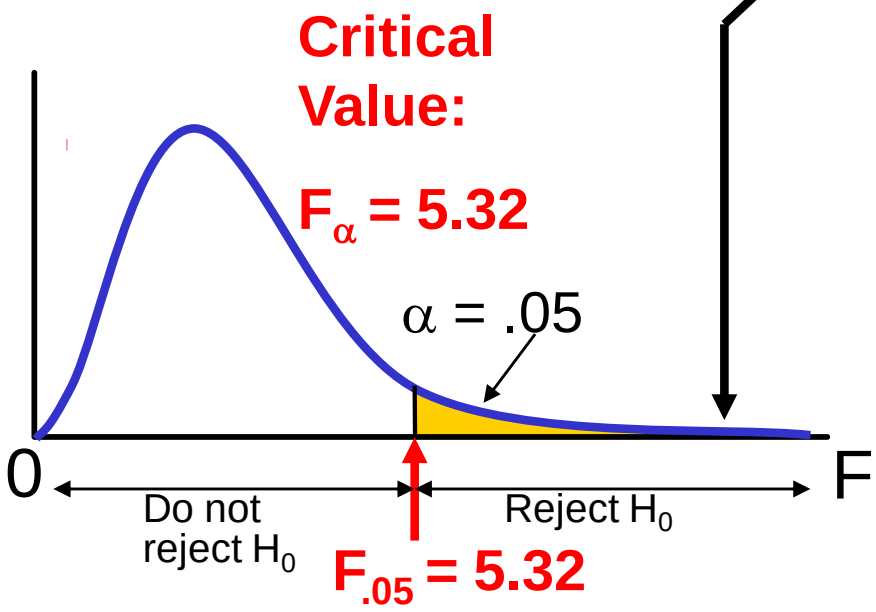
F-Test for Significance (2)

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

$\alpha = .05$

$df_1 = 1 \qquad df_2 = 8$



Test Statistic:

$$F = \frac{MSR}{MSE} = 11.08$$

Decision:

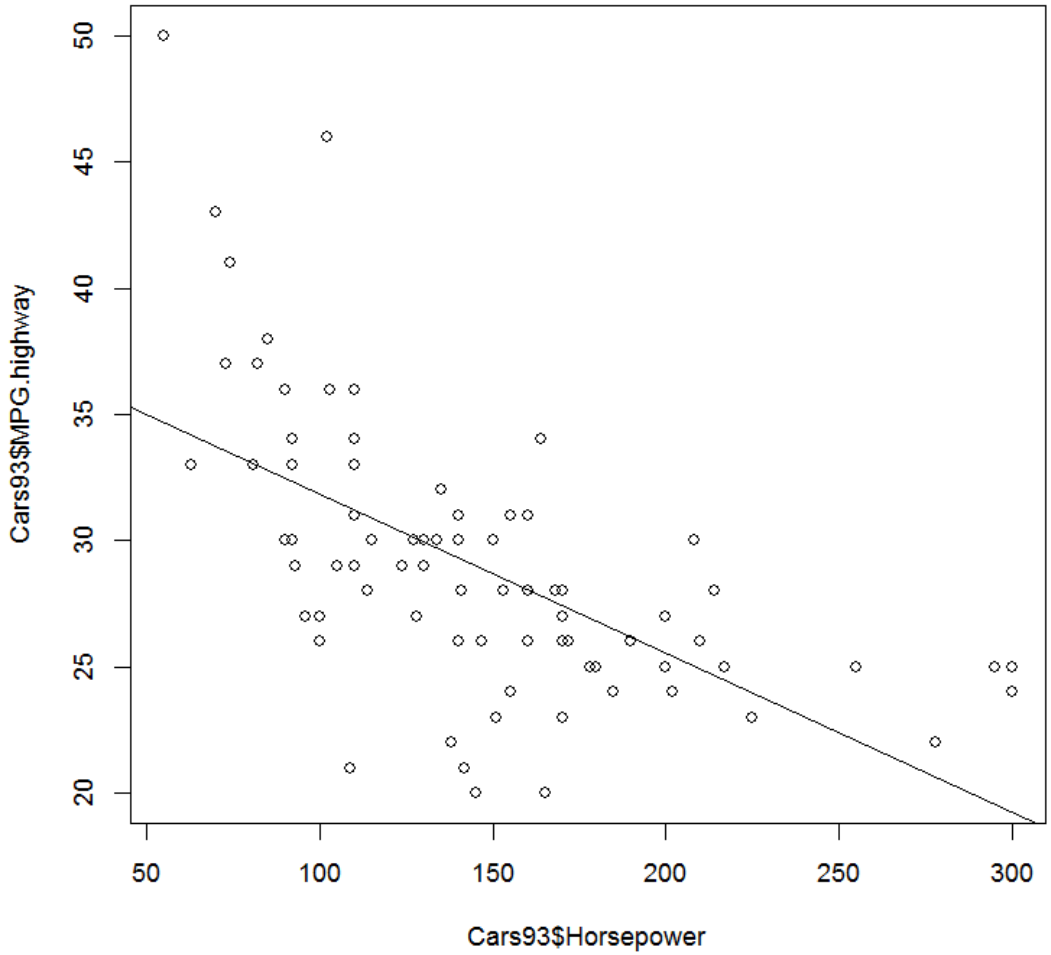
Reject H_0 at $\alpha = 0.05$

Conclusion:

There is sufficient evidence that house size affects selling price

Regression in R

```
> plot(Cars93$MPG.highway ~ Cars93$Horsepower)
> abline(da)
```



Regression in R (2)

Table 10.1 Extractor functions for the result of lm ()

summary ()	returns summary information about the regression
plot ()	makes diagnostic plots
coef()	returns the coefficients
residuals ()	returns the residuals (can be abbreviated resid())
fitted ()	returns the residuals $\hat{y}_i$
deviance()	returns RSS
predict ()	performs predictions
anova ()	finds various sums of squares
AIC ()	is used for model selection

Regression Table in R (3)

```
> da=lm(Cars93$MPG.highway ~ Cars93$Horsepower)
> summary(da)
Call:
lm(formula = Cars93$MPG.highway   Cars93$Horsepower)
Residuals:
Min      1Q     Median      3Q     Max
-10.2808 -2.2178 -0.1763 1.6727 15.3161
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        38.149884 1.282045 29.757 < 2e-16 ***
Cars93$Horsepower -0.063019 0.008381 -7.519 3.74e-11 ***
Residual standard error:  4.21 on 91 degrees of freedom
Multiple R-squared:  0.3832, Adjusted R-squared:  0.3764
F-statistic:  56.54 on 1 and 91 DF, p-value:  3.744e-11
```

- Confidence interval for β_1:

```
> betahatl=-0.063019 # read from summary
> SE=0.008381 # read from summary
> tstar=qt(1-0.05/2,df= n-2)
> c(betahatl-tstar*SE, betahatl+tstar*SE)
[1] -0.07966683 -0.04637117
```

Prediction

- The regression equation can be used to predict a value for y, given a particular x

- For a specified value, x_{n+1} , the predicted value is

$$\hat{y}_{n+1} = b_0 + b_1 x_{n+1}$$

Predictions Using Regression Analysis

Predict the price for a house with 2000 square feet:

$$\widehat{\text{house price}} = 98.25 + 0.1098\ (\text{sq.ft.})$$

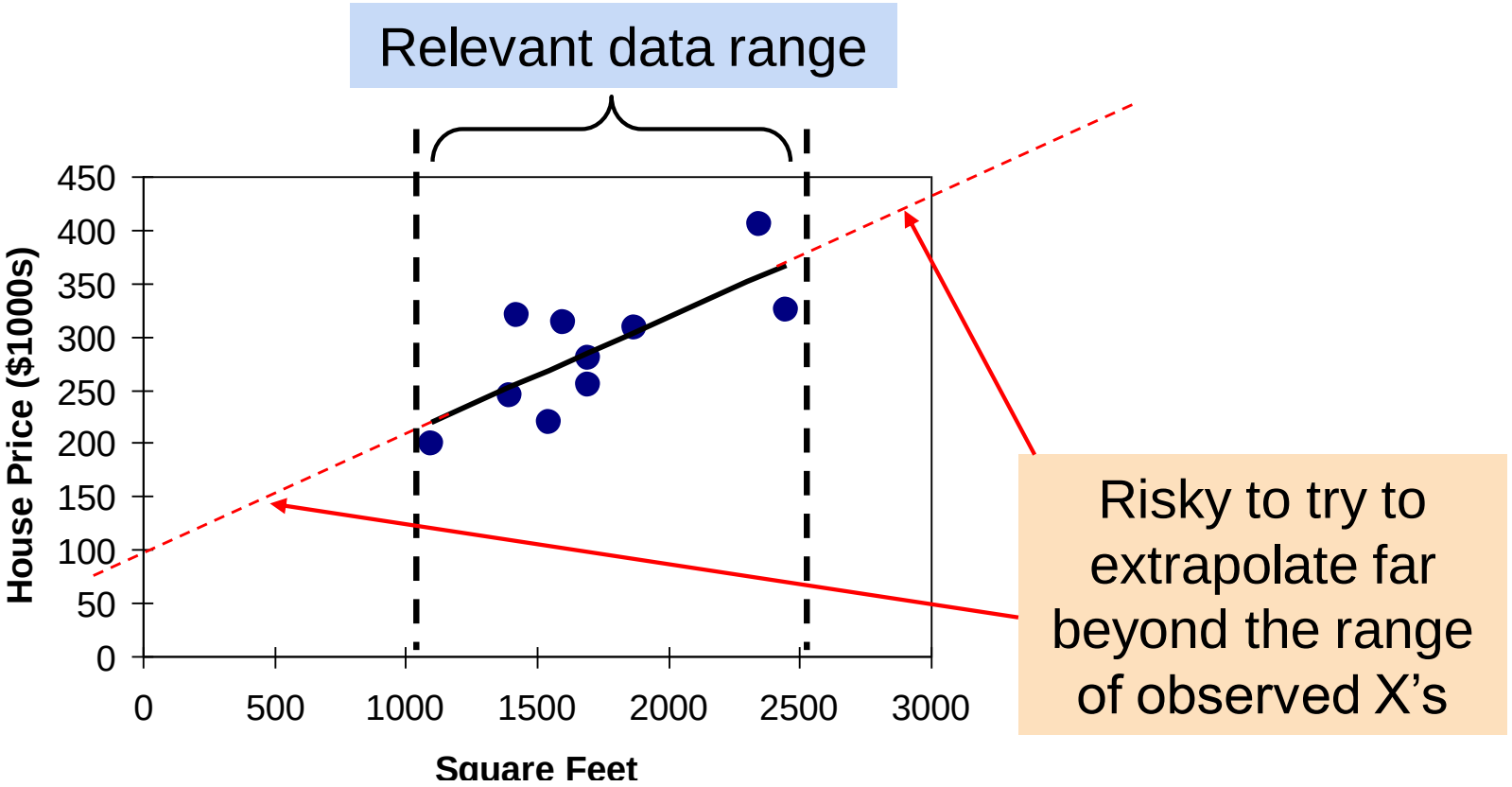
$$= 98.25 + 0.1098(200\,0)$$

$$= 317.85$$

The predicted price for a house with 2000 square feet is 317.85($1,000s) = $317,850

Relevant Data Range

- When using a regression model for prediction, only predict within the relevant range of data



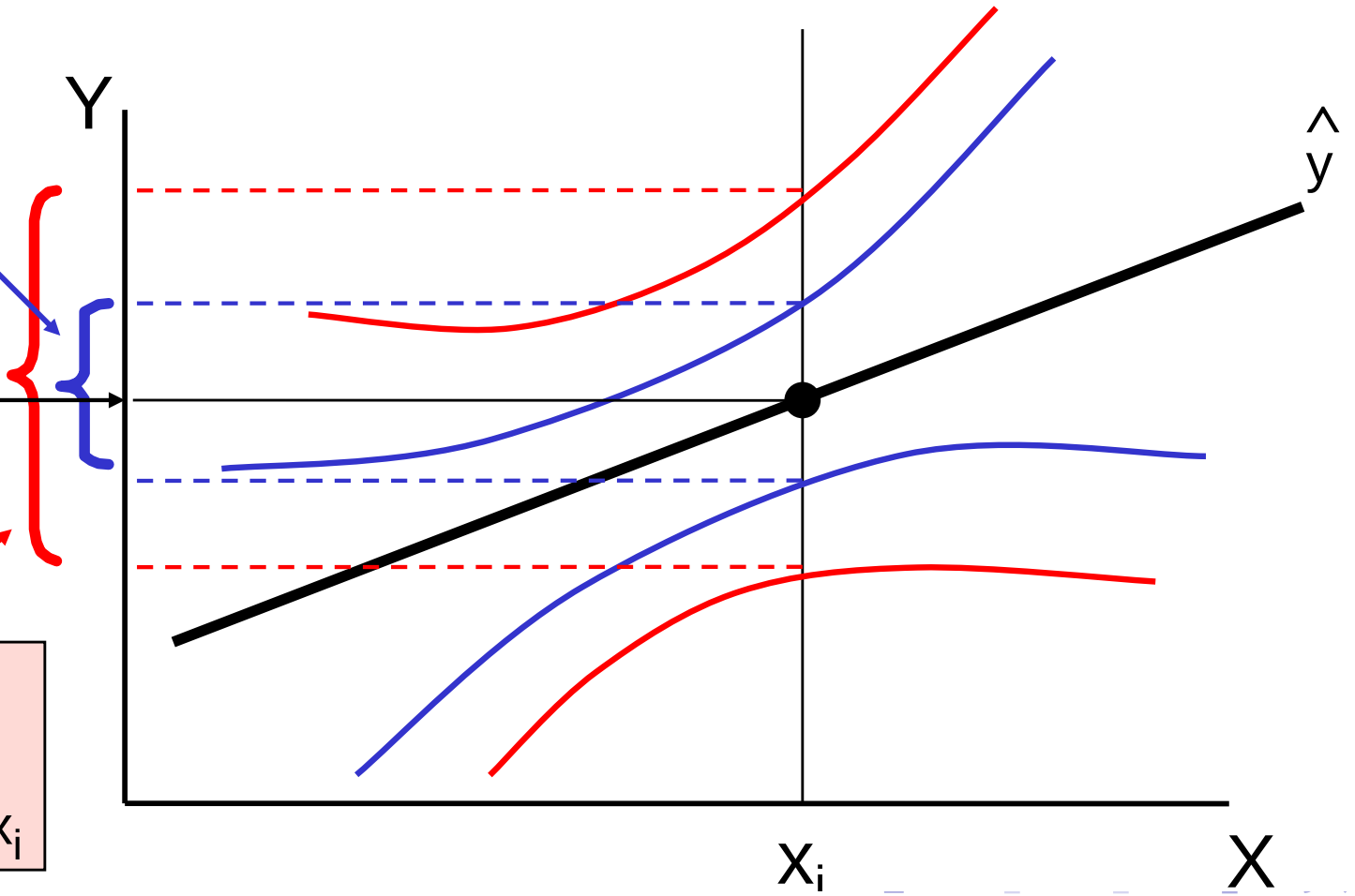
Estimating Mean Values and Predicting Individual Values

Goal: Form intervals around y to express uncertainty about the value of y for a given x_i

Confidence Interval for the expected value of y, given x_i

$\hat{y} = b_0 + b_1 x_i$

Prediction Interval for an single observed y, given x_i



Prediction in R: interpolating or extrapolating

```
> y=Cars93$MPG.highway;x=Cars93$Horsepower
> predict(lm(y~x),new=data.frame(x=123.2))
1
30.38597
```

Confidence Interval for the Average Y, given X

Confidence interval estimate for the **expected value of y** given a particular x_i

Confidence interval for $E(Y_{n+1} \mid X_{n+1})$:

$$\hat{y}_{n+1} \pm t_{n-2,\alpha/2} s_e \sqrt{\left[\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right]}$$

Notice that the formula involves the term $(x_{n+1} - \bar{x})^2$

so the size of interval varies according to the distance x_{n+1} is from the mean, $\bar{x}$

Prediction Interval for an Individual Y, given X

Confidence interval estimate for an **actual observed value of y** given a particular x_i

Confidence interval for $\hat{y}_{n+1}$:

$$\hat{y}_{n+1} \pm t_{n-2,\alpha/2} s_e \sqrt{\left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right]}$$

This extra term adds to the interval width to reflect the added uncertainty for an individual case

Estimation of Mean Values: Example

Confidence Interval Estimate for $E(Y_{n+1}|X_{n+1})$

Find the 95% confidence interval for the mean price of 2,000 square-foot houses

Predicted Price $\hat{y}_i = 317.85$ ($1,000s)

$$\hat{y}_{n+1} \pm t_{n-2,\alpha/2} s_e \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}} = 317.85 \pm 37.12$$

The confidence interval endpoints are 280.66 and 354.90, or from $280,660 to $354,900

Prediction Interval for Mean Y, given X in R

```
> pred.res = predict(res, interval="confidence")
```

Estimation of Individual Values: Example

Confidence Interval Estimate for $\hat{y}_{n+1}$

Find the 95% confidence interval for an individual house with 2,000 square feet

Predicted Price $\hat{y}_i = 317.85$ ($1,000s)

$$\hat{y}_{n+1} \pm t_{n-1,\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum (X_i - \overline{X})^2}} = 317.85 \pm 102.28$$

The confidence interval endpoints are 215.50 and 420.07, or from $215,500 to $420,070

Prediction Interval for an Individual Y, given X in R

```
> pred.res = predict(res, int = "pred")
> pred.res
fit lwr upr
1 193.2 185.0 201.4
2 189.4 181.3 197.5
...
```

Plotting the prediction interval:

```
> age.sort = sort(unique(age))
> pred.res = predict(res.mhr, newdata = data.frame(age =
age.sort), int="pred")
> pred.res[,2]
1 2 3 4 5 6 7 8 9
185.0 181.3 177.6 173.9 170.1 166.3 162.4 158.5 154.6
> plot(mhr ~ age); abline(res)
> lines(age.sort,pred.res[,2] , lty=2) # lower curve
> lines(age.sort,pred.res[,3], lty=2) # upper curve
```

Assumptions of Regression

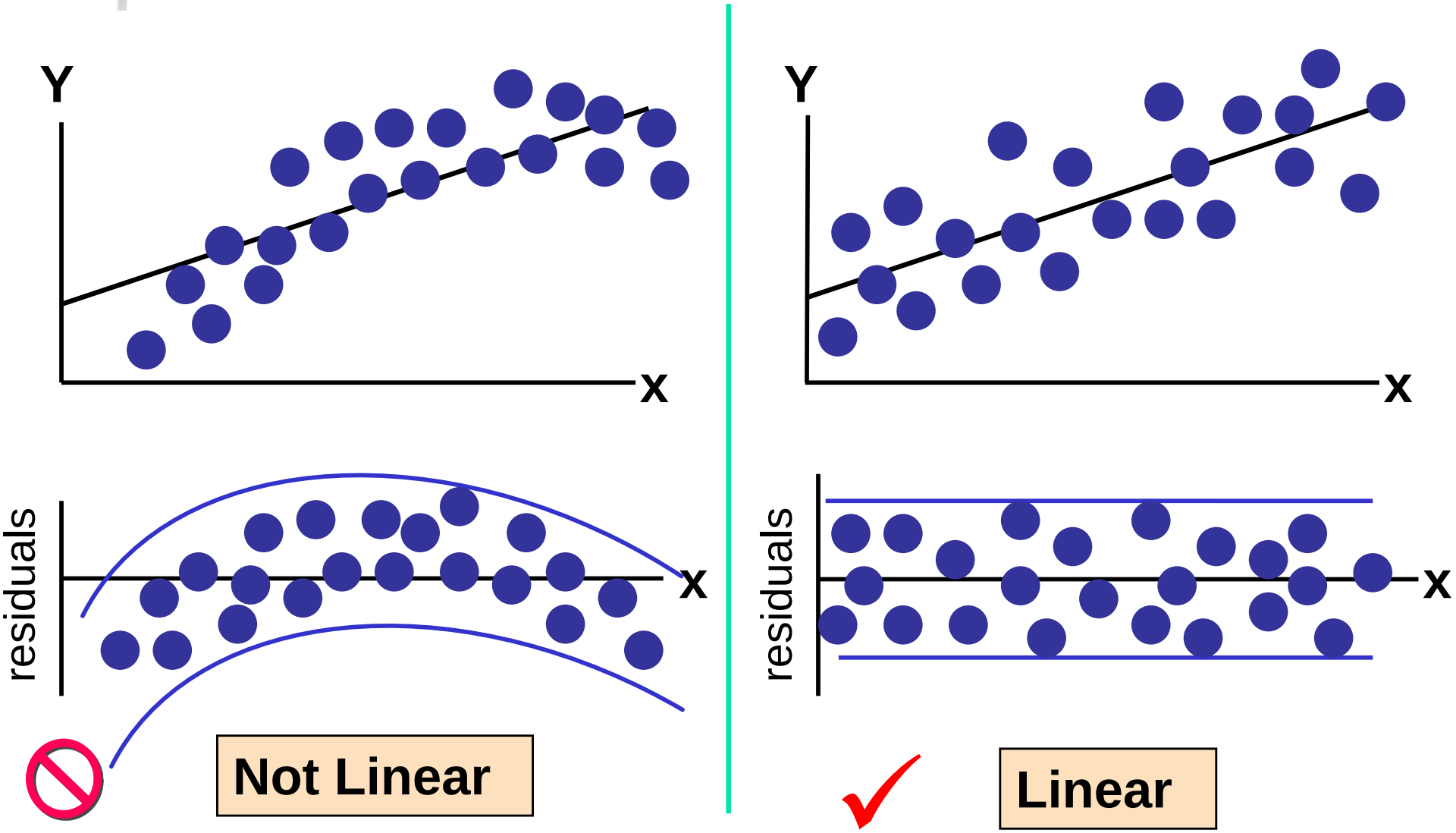
- Normality of Error
  - Error values (ε) are normally distributed for any given value of X
- Homoscedasticity
  - The probability distribution of the errors has constant variance
- Independence of Errors
  - Error values are statistically independent

Residual Analysis

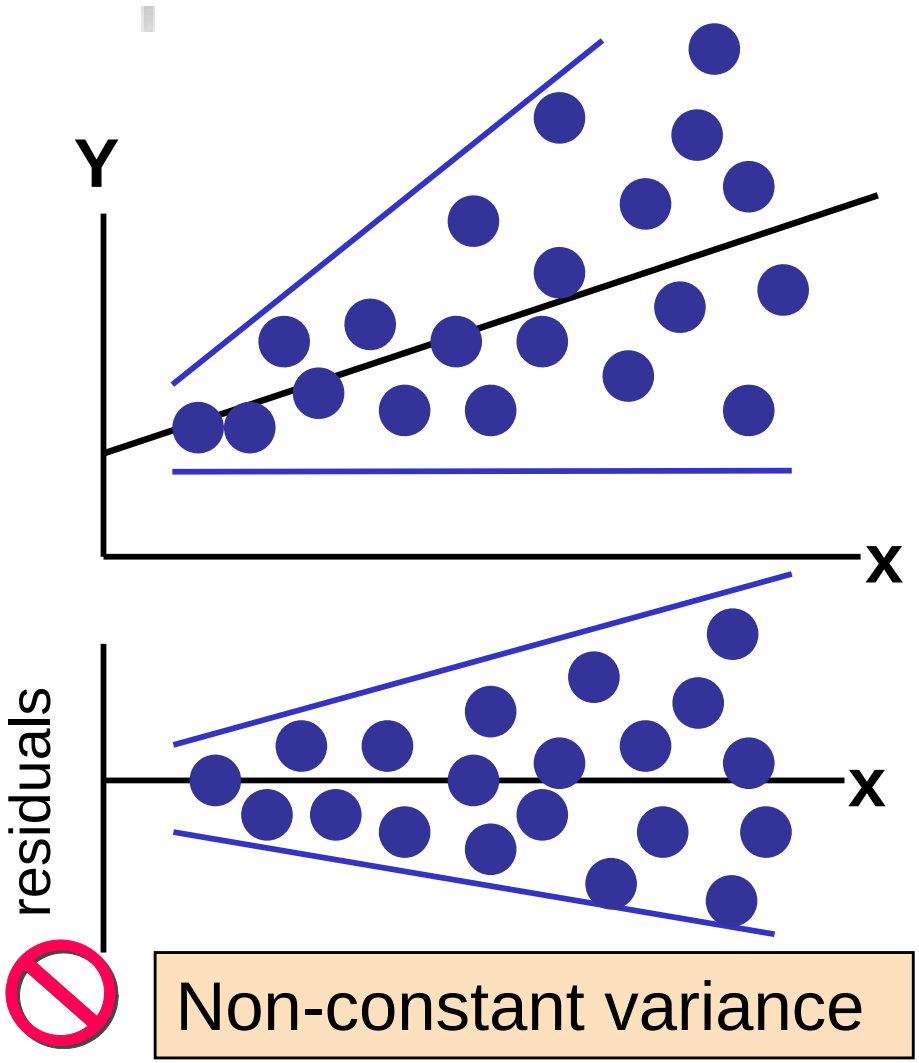
$$e_i = y_i - \hat{y}_i$$

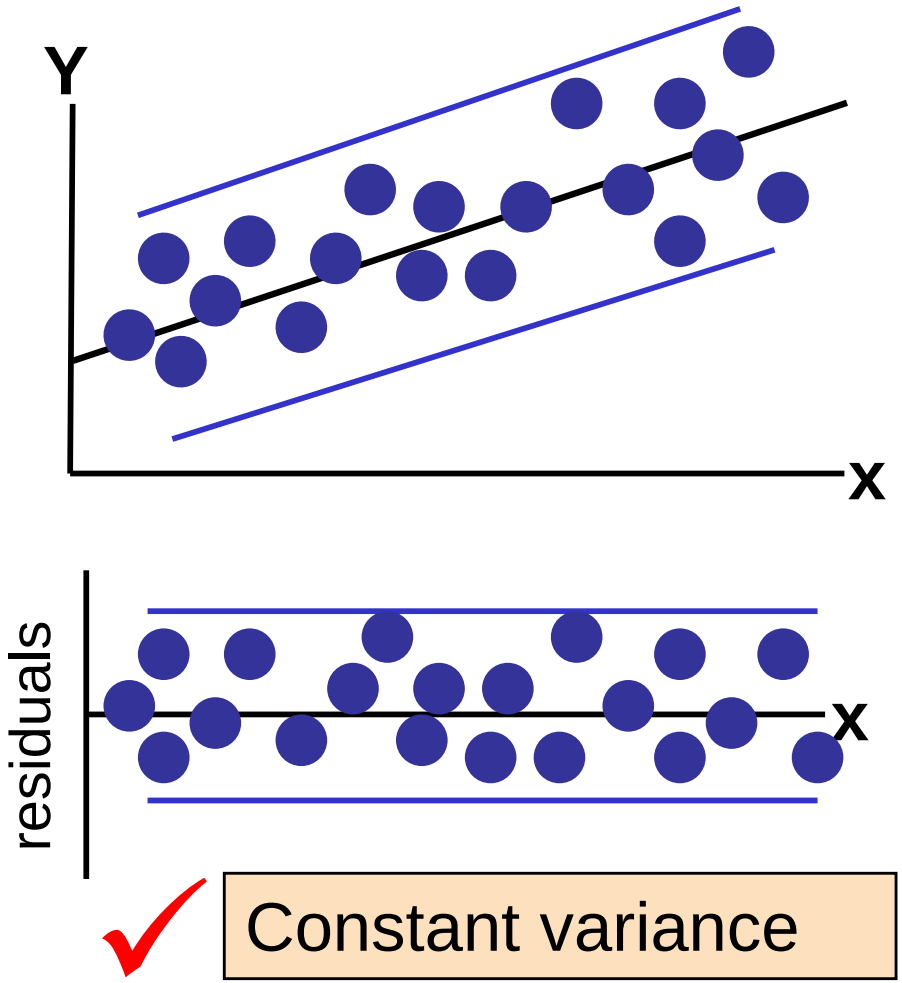
- The residual for observation i, e_i, is the difference between its observed and predicted value
- Check the assumptions of regression by examining the residuals
  - Examine for linearity assumption
  - Examine for constant variance for all levels of X (homoscedasticity)
  - Evaluate normal distribution assumption
  - Evaluate independence assumption
- Graphical Analysis of Residuals
  - Can plot residuals vs. X

Residual Analysis for Linearity



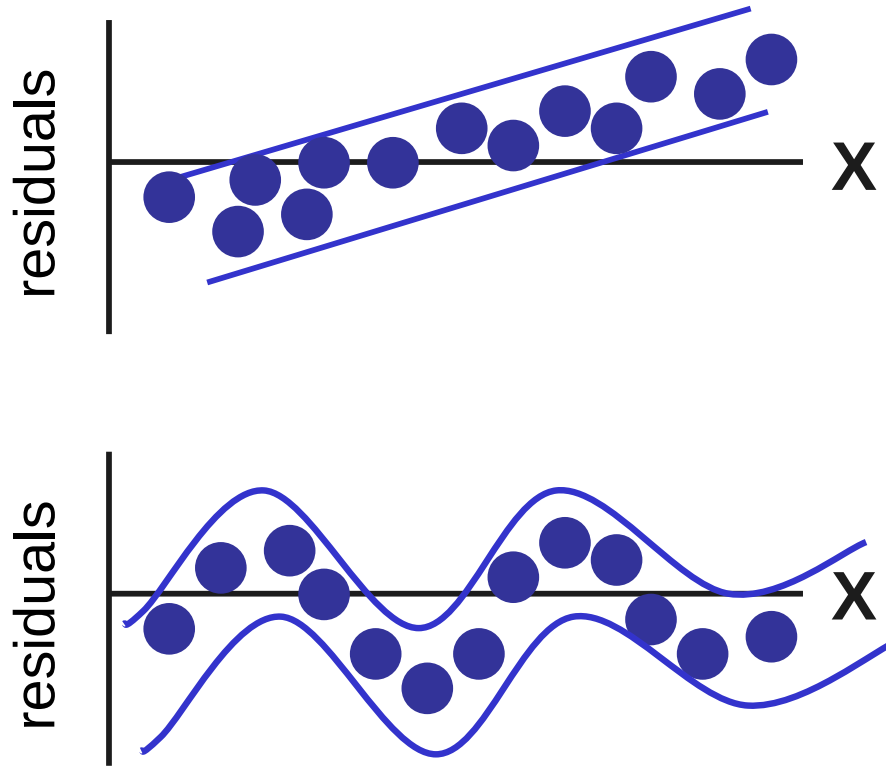
Residual Analysis for Homoscedasticity

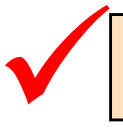




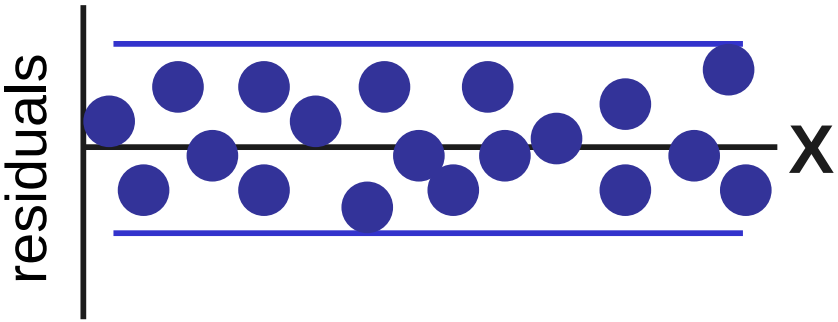
Residual Analysis for Independence

Not Independent



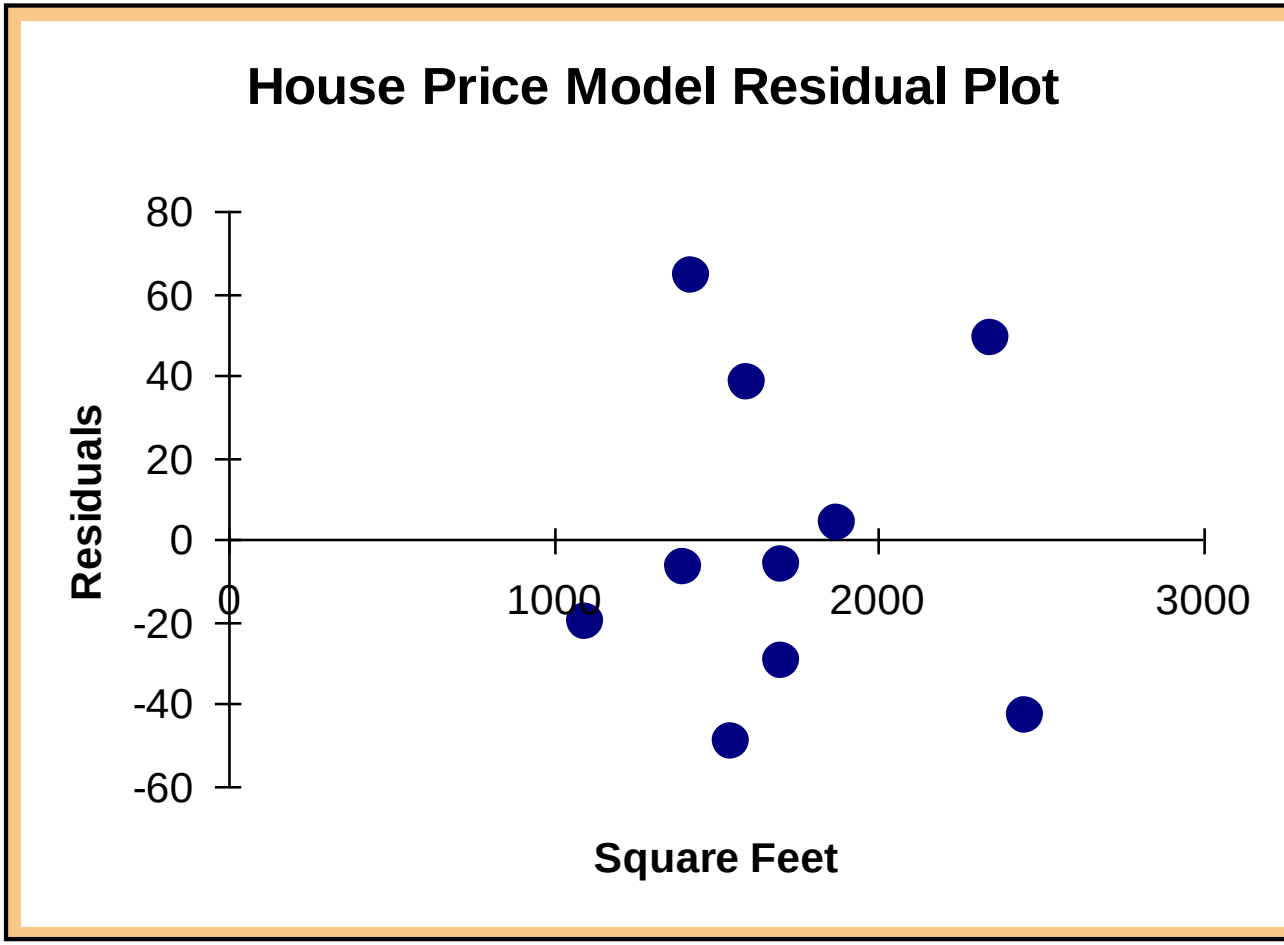


Independent



Residual Output

RESIDUAL OUTPUT		
	Predicted House Price	*Residuals*
1	251.92316	-6.923162
2	273.87671	38.12329
3	284.85348	-5.853484
4	304.06284	3.937162
5	218.99284	-19.99284
6	268.38832	-49.38832
7	356.20251	48.79749
8	367.17929	-43.17929
9	254.6674	64.33264
10	284.85348	-29.85348



Does not appear to violate any regression assumptions

Acknowledgments

Slides are partially adapted from those accompanying Newbold's textbook

References

See Chapters 1-8 of [1] for more.

[1] J. Verzani.
Using R for Introductory Statistics, Second Edition.
Chapman & Hall/CRC The R Series. Taylor & Francis, 2014.