Ch. 11 - Simple Regression

Ercan Karadas

New York University

ercan@nyu.edu

December 10, 2015

Overview of Linear Models

An equation can be fit to show the best linear relationship between two variables:

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

where

- Y is the dependent variable and
- X is the independent variable
- β_0 is the Y-intercept
- β_1 is the slope

Least Squares Regression

- Estimates for coefficients β_0 and β_1 are found using a Least Squares Regression technique.
- The least-squares regression line, based on sample data, is

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

where b_1 is the slope of the line and b_0 is the y-intercept term:

$$b_1 = \frac{Cov(x, y)}{s_x^2} = r\frac{s_y}{s_x} \qquad b_0 = \overline{y} - b_1\overline{x}$$

- 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

- The relationship between X and Y is described by a linear function
- Changes in Y are assumed to be influenced 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_i is a random error term.

Simple Linear Regression Model

The population regression model:

Dependent Variable

Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

$$\boxed{y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}}$$

Linear Regression Assumptions

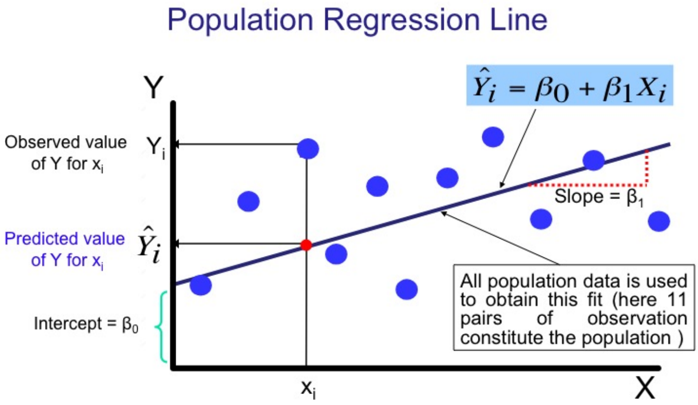
- The true relationship form is linear (Y is a linear function of X, plus random error)
- The error terms, ε_i are independent of the x values
- The error terms ε_i are random variables with mean 0 and constant variance, σ^2:

  $$E(\varepsilon_i) = 0 \qquad \text{and} \qquad Var(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2$$

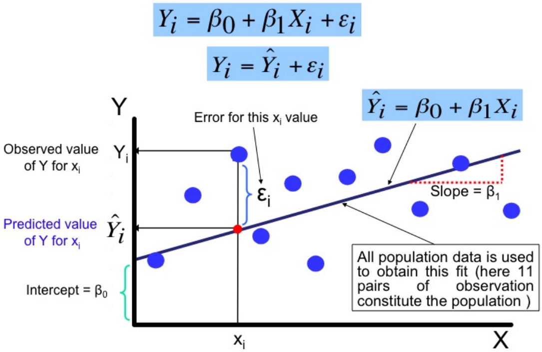
  (the uniform variance property is called homoscedasticity)
- The random error terms, ε_i, are not correlated with one another, so that

  $$E(\varepsilon_i \varepsilon_j) = 0, \quad \text{for all } i \neq j$$

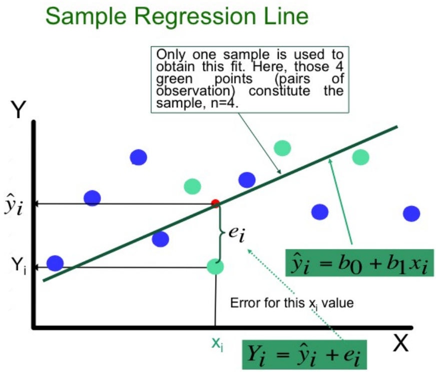
- If we had all population data (blue spots below) we could compute the β_0 and β_1
- And this would be the best model we could hope to work with...



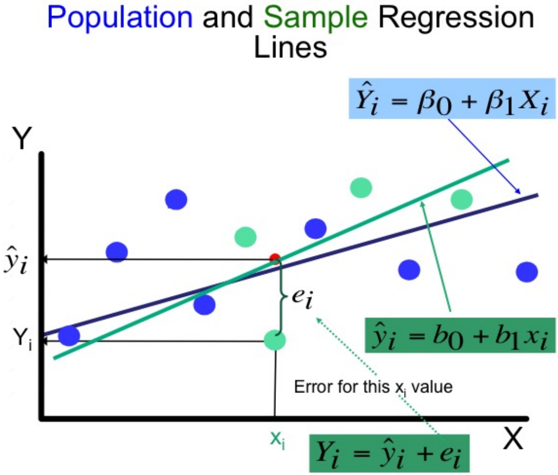
- Even in this case our population regression line, $\hat{Y}_i$, would not match all the observations, Y_i.
- For each X_i, the model prediction, $\hat{Y}_i$, contains some error, ε_i, and therefore we have $Y_i = \hat{Y}_i + \varepsilon_i$
- But this is due to the linear structure of the model that we imposed.
- A much more serious problem is that in practice we never have the population data!



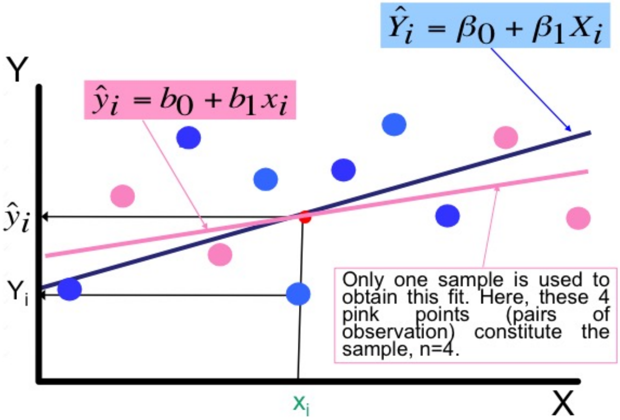
- In another words, in practice, we do not observe all of the blue points in the figure.
- Instead, suppose that we decided to work with a sample of size 4 and a random sample turned out to be the green points in the figure below
- A linear fit just taking into account these four points is also depicted in the figure



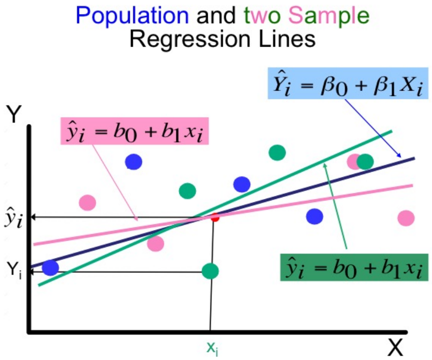
- Note that the population regression line $\hat{Y}_i$, and the sample regression line $\hat{y}_i$ are different.



Population and (another) Sample Regression Lines



- The population regression line is unique, i.e. β_0 and β_1 are fixed, but unobserved
- On the other hand, we can compute the coefficients, b_0 and b_1, in the sample regression line, but they can change from sample to sample
- Therefore, in the regression analysis, we will study
  - based on the sample results, b_0 and b_1, what we can say about the unobserved population parameters β_0 and β_1
  - and potential problems that we might encounter



Simple Linear Regression Equation

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

Estimated (or predicted) y value for observation i

Estimate of the regression intercept

Estimate of the regression slope

Value of x for observation i

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

The individual random error terms e_i have a mean of zero

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

Least Squares Coefficient 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 residuals (errors), SSE:

$$\begin{aligned}\min \text{ SSE} &= \min \sum_{i=1}^{n} e_i^2 \\ &= \min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \min \sum_{i=1}^{n} \big(y_i - (b_0 + b_1 x_i)\big)^2\end{aligned}$$

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

Least Squares Coefficient Estimators

- The slope coefficient estimator is

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{Cov(x, y)}{s_x^2} = r\frac{s_y}{s_x}$$

  where $s_x^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2/(n-1)$ is the sample variance for X, similarly for s_y^2, and r is the correlation coefficient between X and Y.
- And the intercept estimator is

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

Computer Computation of Regression Coefficients

The coefficients b_0 and b_1 , and other regression results in this chapter, will be found using a computer because

- Hand calculations are tedious
- Statistical routines are built into Excel
- Other statistical analysis software can be used
- You are not expected to remember the formulas for b_0 and b_1, but developing an intuitive understanding for the expressions is important and expected.
- For example, why $Cov(x, y)$ appears in the numerator and s_x^2 in the denominator of the expression for b_1?
- The regression line always goes through the mean $(\overline{x}, \overline{y})$, why?
- Also, note that in order to compute b_0, first we need to compute b_1.

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, otherwise in most cases it can be interpreted as average effect of factors other than x.
- 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

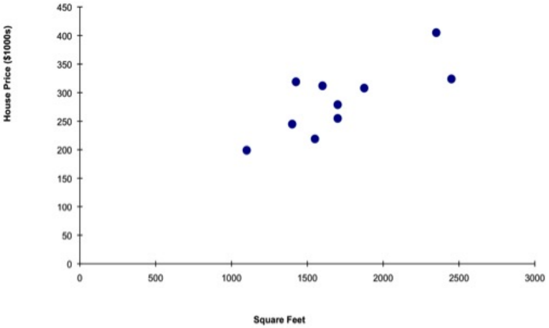
- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)

- 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

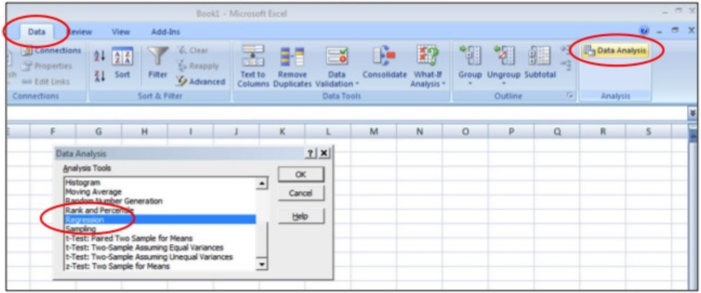
House Price in $1000s (Y)	Square Feet (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

House price model: scatter plot



Regression Using Excel

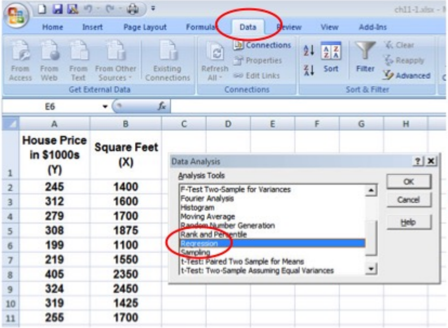
- Excel will be used to generate the coefficients and measures of goodness of fit for regression
  - Data / Data Analysis / Regression



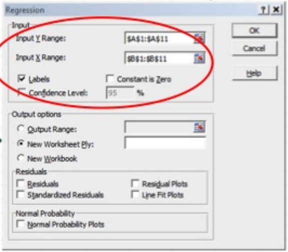
Regression Using Excel

(continued)

- Data / Data Analysis / Regression



Provide desired input:



Excel Output

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.762113713					
5	R Square	0.580817312					
6	Adjusted R Square	0.528419476					
7	Standard Error	41.33032365					
8	Observations	10					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	18934.9348	18934.9348	11.0848	0.01039	
13	Residual	8	13665.5652	1708.1957			
14	Total	9	32600.5				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	98.24833	58.03348	1.69296	0.12892	-35.57711	232.07377
18	Square Feet (X)	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Excel Output

(continued)

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:

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

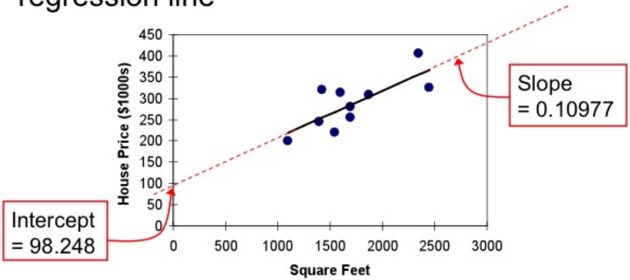
ANOVA

	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



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

Interpretation of the Intercept, b_0

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

- 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)
  - Here, no houses had 0 square feet, so $b_0 = 98.24833$ just indicates that, for houses within the range of sizes observed, $98,248.33 is the portion of the house price not explained by square feet

Interpretation of the Slope Coefficient, b_1

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

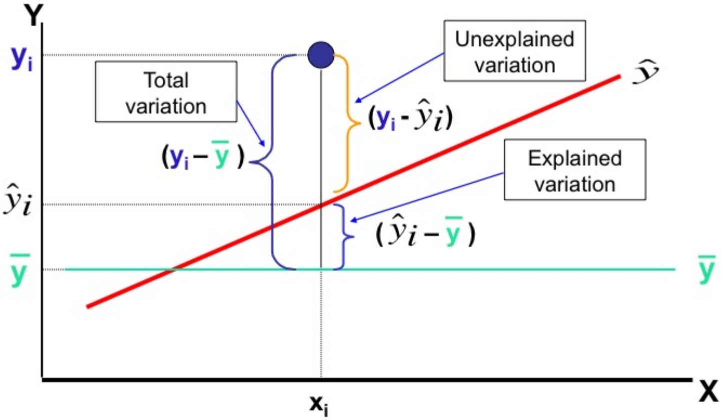
- b_1 measures the estimated change in the average value of Y as a result of a one-unit change in X
  - Here, $\boxed{b_1 = .10977}$ tells us that the average value of a house increases by .10977($1000) = $109.77, on average, for each additional one square foot of size

Explanatory Power of a Linear Regression Equation / Analysis of Variance

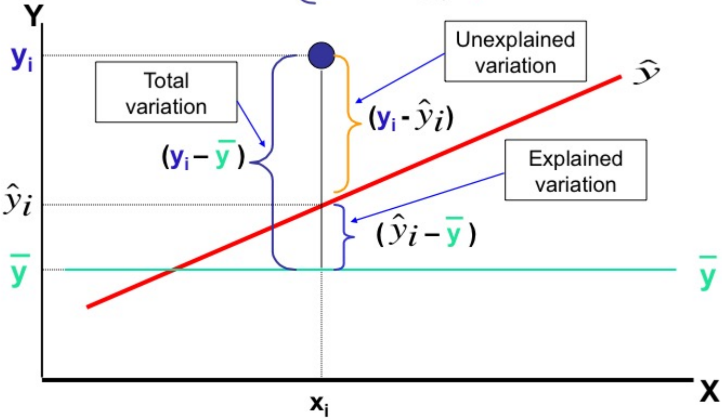
Question: How can we asses performance of the house model?
A good model should give a satisfactory answer to the following question: What is your best guess for the price of a randomly chosen house in the city?

- **Without a model.** Since we already have the price data for n houses in the city, $\{y_1, \ldots, y_n\}$, our best would be $\overline{y}$.
- If it turns out that actual price of the house is y_i, our (prediction) error would be
  $\Longrightarrow y_i - \overline{y} \quad$ (crude prediction error)
- **With a model.** Our prediction would be $\hat{y}_i = b_0 + b_1 x_i$, where x_i is the size of the house that is assumed to be known.
- And therefore our prediction error would be
  $\Longrightarrow y_i - \hat{y}_i \quad$ (sophisticated prediction error)
- Therefore, we can asses performance of the model by looking at how much the sophisticated error improves the initial crude prediction error.

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \quad \text{(error decomposition)}$$



$$SST = \sum(y_i - \bar{y})^2 = \begin{cases} SSE = \sum(y_i - \hat{y}_i)^2 \\ SSR = \sum(\hat{y}_i - \bar{y})^2 \end{cases}$$



- Total variation is made up of two parts:

$$SST = SSR + SSE$$

Total Sum of Squares	Regression Sum of Squares	Error (residual) 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

Analysis of Variance

- 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

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

$$R^2 = \frac{SSR}{SST} = \frac{\text{Regression Sum of Squares}}{\text{Total Sum of Squares}}$$

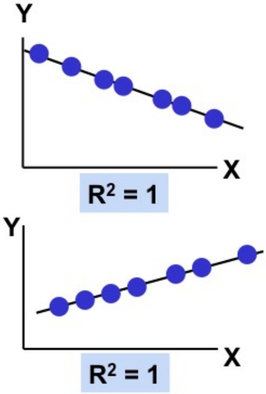
- Note that

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

- For a linear regression model with only one independent variable, we also have

$$r^2 = R^2$$

Examples of Approximate R^2 Values

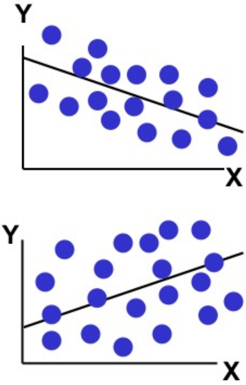


$R^2 = 1$

Perfect linear relationship between X and Y:

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

Examples of Approximate R^2 Values

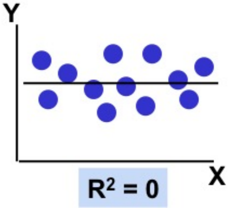


$0 < R^2 < 1$

Weaker linear relationships between X and Y:

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

Examples of Approximate R^2 Values



$R^2 = 0$

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)

Excel Output

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

Statistical Inference: Hypothesis Tests and C. I.

Let us start with a summary of what we have done so far:

- We can think of the population regression line, $\hat{Y}_i = \beta_0 + \beta_1 X_i$, as the best fit we could potentially obtain.
- However, since we do not have the whole population data we do not know β_0 and β_1 in practice.
- Therefore, we work with a sample instead, and we can compute the corresponding sample regression coefficients, b_0 and b_1, as estimates of the population regression coefficients β_0 and β_1:

$$b_0 \longrightarrow \beta_0 \quad \text{and} \quad b_1 \longrightarrow \beta_1$$

- However, as we change the sample, fitted sample regression line changes as well (compare the coefficients b_0 and b_1 for the green and pink samples).
- Depending the sample, sometimes we over-predict the population slope ($b_1 > \beta_1$, green sample) or under-predict ($b_1 < \beta_1$, pink sample).

- The situation is quite similar to the case when tried to make inference about the unobserved population mean, μ, based on the sample mean, $\overline{X}$: $\quad \overline{X} \longrightarrow \mu$
- Just like $\overline{X}$, here b_0 **and** b_1 **are random variables**: they are formulas/expressions in terms of the sample data. Therefore, we can talk about the distributions of b_0 and b_1, their expected values, and variances.
- We were able to construct confidence interval for μ and do hypothesis testing based on $\overline{X}$ only after we worked out the distribution of $\overline{X}$. For example, when the population is normal, we said

$$\overline{X} \sim N(\mu, \sigma^2/n)$$

- Therefore, in order to able to make some inference for β_0 and β_1, we need to determine how b_0 and b_1 are distributed. For example,

$$E(b_1) =?, \quad Var(b_1) =?, \quad b_1 \sim N(\cdot, \cdot) \quad ?$$

- I will focus on the distribution of b_1, the other case is similar.

- The distribution of b_1 critically depends on our assumptions regarding the population error terms ε_i's
- Recall that we made the following assumptions on ε_i's:

1.	$E(\varepsilon_i) = 0$	(errors cancel out on average)
2.	$Var(\varepsilon_i) = \sigma^2$	(errors have the same variance)
3.	$E(\varepsilon_i \varepsilon_j) = 0$	(errors are independent)

- Together these three (actually just 1 and 3 are enough) assumptions ensures that b_1 will be equal to β_1 on average, i.e.

$$E(b_1) = \beta_1$$

- What about $Var(b_1)$?
- $var(b_1) = \sigma^2_{b_1} = \sum_{i=1}^{n} (b_{1,i} - \overline{b}_1)^2/(n-1)$
- $\sigma^2_{b_1}$ depends on two things:
  - $Var(\varepsilon_i) = \sigma^2$
  - s^2_X, variance of x_i's.

Dependence on $Var(\varepsilon_i) = \sigma^2$

There is another way to think about ε_i terms:

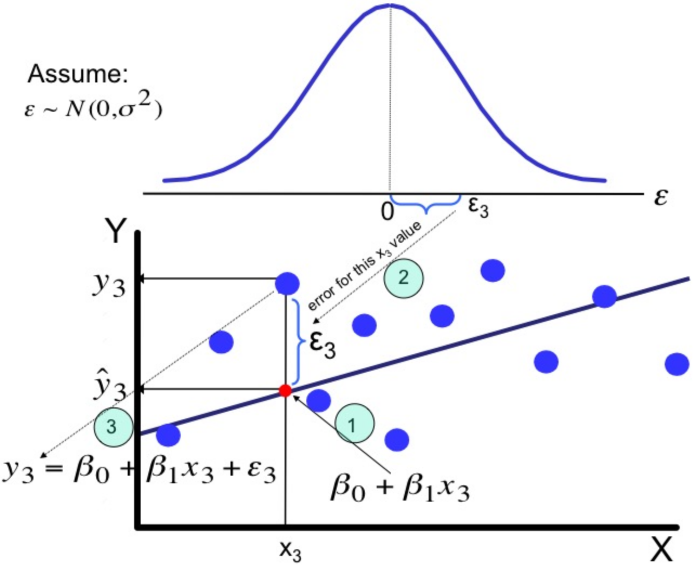
- Suppose that actually the relationship between X and Y is perfectly linear as given by the population regression line

$$y_i = \beta_0 + \beta_1 x_i$$

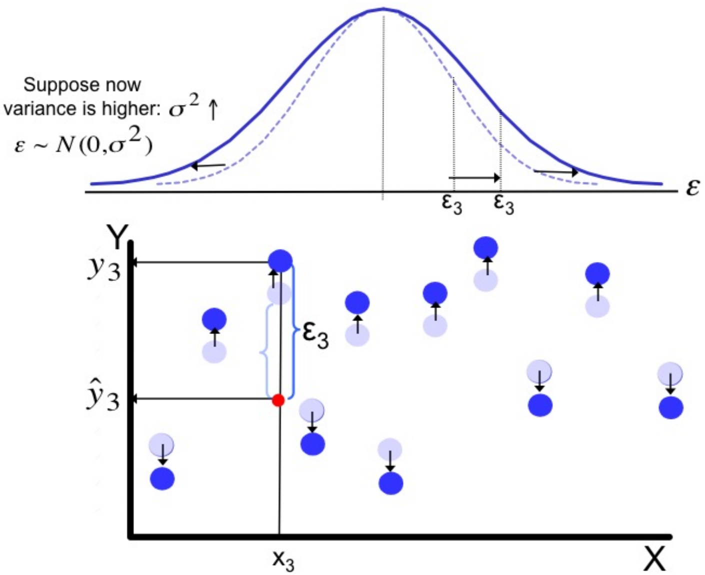
- However, we are unable to measure y_i with exact precision, instead what we observe is actually $y_i + \varepsilon_i$
- When we pass ε_i to the other side, we obtain the population regression model that we have been working with

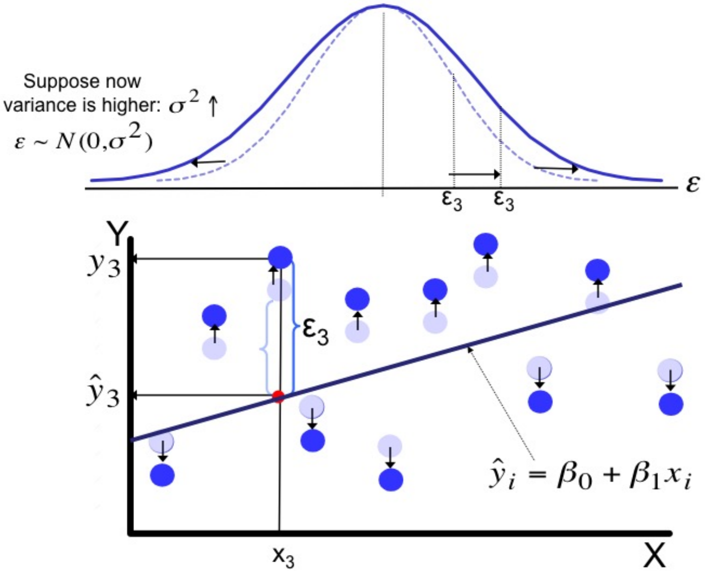
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

  (here writing $-\varepsilon_i$ or ε does not make any difference since $E(\varepsilon_i) = 0$)
- Therefore, this interpretation allows us think as if: for a given x_i, first find $\beta_0 + \beta_1 x_i$, then add a random measurement term ε_i to obtain the true y_i as $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

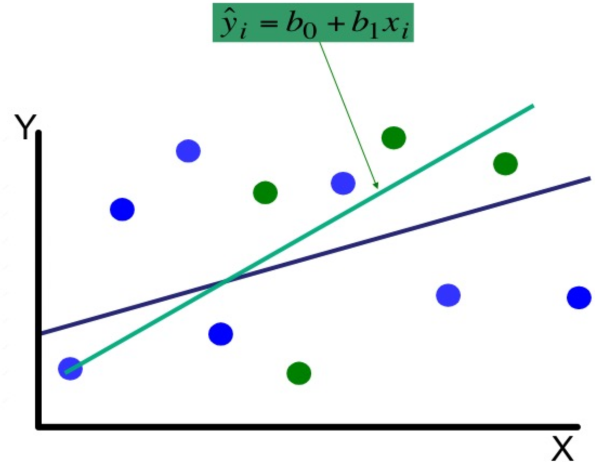

Assume:
$\varepsilon \sim N(0,\sigma^2)$
0
ε_3
ε
Y
error for this x_3 value
2
y_3
ε_3
$\hat{y}_3$
3
1
$y_3 = \beta_0 + \beta_1 x_3 + \varepsilon_3$
$\beta_0 + \beta_1 x_3$
x_3
X

Main Question: How does a change in the variance of measurement error, σ^2, effect $\sigma^2_{b_1}$?

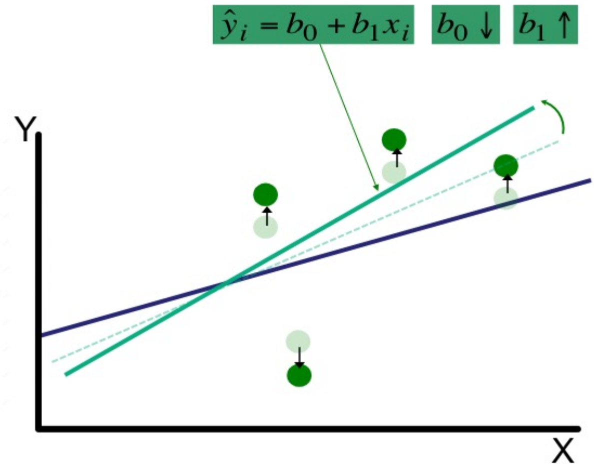



Suppose now variance is higher: $\sigma^2 \uparrow$
$\varepsilon \sim N(0,\sigma^2)$
ε
ε_3
ε_3
Y
y_3
ε_3
$\hat{y}_3$
$\hat{y}_i = \beta_0 + \beta_1 x_i$
x_3
X

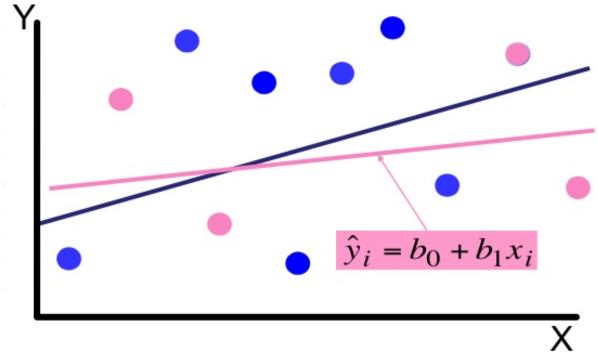
- The sample regression line fit for the new green sample shown below



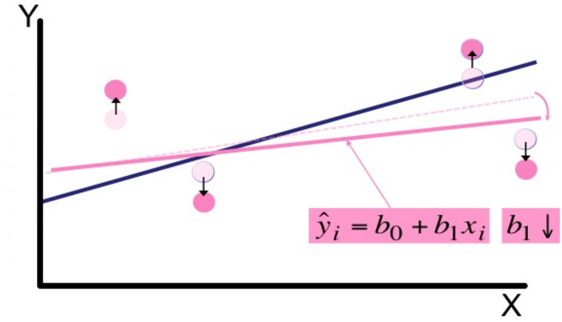
- The sample regression lines fit for the new and old green samples shown together below
- Note that it is steeper $\Longrightarrow$ b_1 is larger compare to the previous case (before increasing σ^2)



- The sample regression line fit for the new pink sample shown below

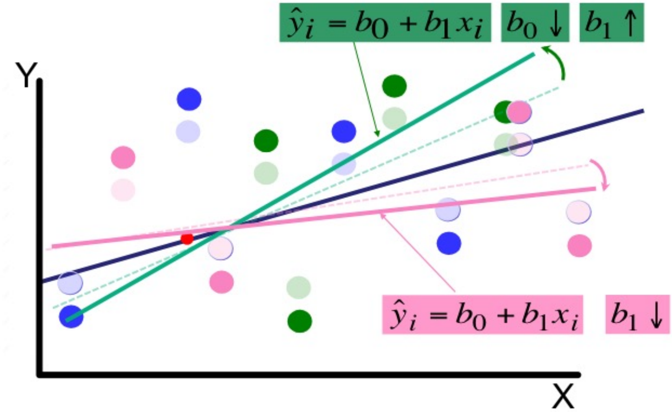


- The sample regression lines fit for the new and old pink samples shown together below
- Note that it is flatter $\Longrightarrow$ b_1 is smaller compare to the previous case (before increasing σ^2)



Result 1: The higher the variance of the population error terms the higher the variance of the slope coefficient of the sample regression:

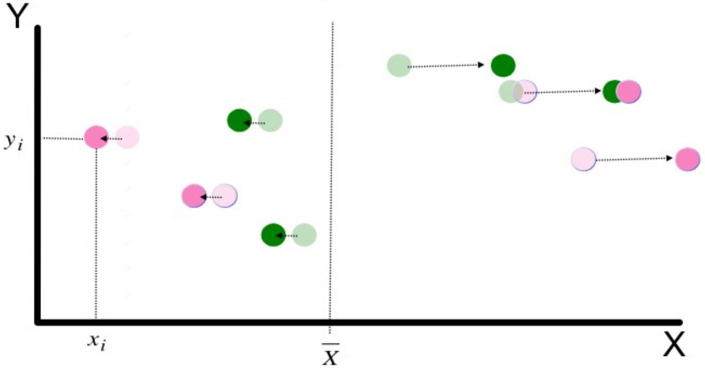
$$\sigma^2 \uparrow \quad \Longrightarrow \quad \sigma^2_{b_1} \uparrow$$



Dependence on s_X^2

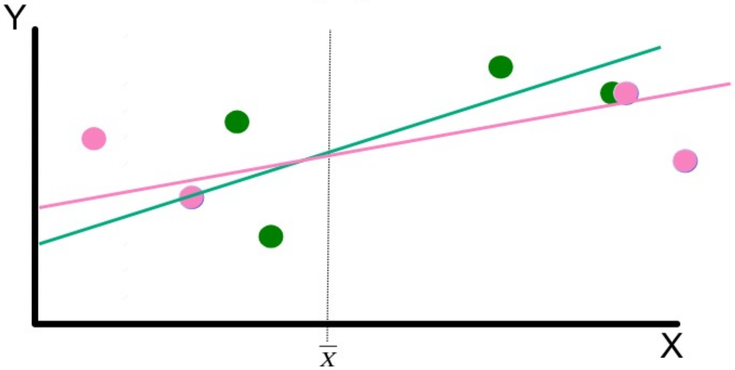
Suppose that the variance of the independent variable is now higher :

$$s_X^2 = \frac{\sum\left(x_i - \overline{X}\right)^2}{n-1} \uparrow$$



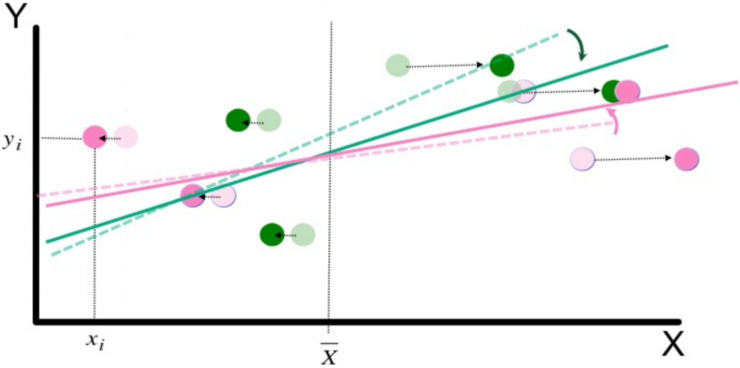
Suppose now that the variance of the independent variable is higher :

$$s_X^2 = \frac{\sum (x_i - \overline{X})^2}{n-1} \uparrow$$



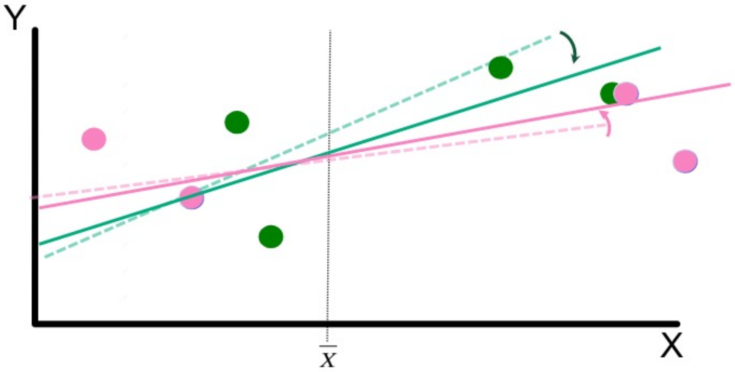
Suppose now that the variance of the independent variable is higher :

$$s_X^2 = \frac{\sum \left(x_i - \overline{X}\right)^2}{n-1} \uparrow$$



Result 2: The higher the variance of the independent variable, the higher the variance of the slope coefficient of the sample regression:

$$s_X^2 \uparrow \implies \sigma_{b_1}^2 \downarrow$$



- In summary, we have seen that

$$\sigma^2 \uparrow \Longrightarrow \sigma^2_{b_1} \uparrow \quad \text{and} \quad s^2_X \uparrow \Longrightarrow \sigma^2_{b_1} \downarrow$$

- In fact, it can be shown that the exact formula for $\sigma^2_{b_1}$ is

$$\sigma^2_{b_1} = \frac{\sigma^2}{(n-1)s^2_X}$$

- Finally, we have found an expression for $\sigma^2_{b_1}$!
- But unfortunately, it contains an unobserved population parameter, σ^2, the variance of population error terms.
- Now, all we need to do is to find a good estimator for σ^2.
- Since σ^2 is the variance of the population error terms ε_i, we should first try the variance of sample error terms e_i:

$$s^2_e = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SSE}{n-2} \qquad \text{(some sources also use } \hat{\sigma}^2\text{)}$$

(Division by $n-2$ instead of $n-1$ is because two parameters, b_0 and b_1, need to be estimated.)

- s_e is called the standard error of the estimate

- We shall use s_e^2 as an estimator for σ^2 to obtain an estimator for $\sigma_{b_1}^2$ as

$$s_{b_1}^2 = \frac{s_e^2}{(n-1)s_X^2}$$

- This is the estimator for the variance of the regression slope coefficient b_1.
- From a particular sample data, this estimator $s_{b_1}^2$ (s_{b_1}) gives us an estimate of the variance (standard deviation) of the regression slope term.

Inference with b_1: summary

Inference with b_1

- $E(b_1) = \beta_1$
- $s_{b_1}^2 \to \sigma_{b_1}^2 = Var(b_1)$
- $\frac{b_1 - \beta_1}{s_{b_1}} = t_{n-2}$

  when either
  - $\varepsilon_i \sim N(0, \cdot)$
  - n is large enough

Inference with $\overline{X}$

- $E(\overline{X}) = \mu$
- $s^2/n \to \sigma^2/n = Var(\overline{X})$
- $\frac{\overline{X} - \mu}{s/\sqrt{n}} = t_{n-1}$

  when either
  - $X_i \sim N(0, \cdot)$
  - n is large enough

Excel Output

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

$$\boxed{s_e = 41.33032}$$

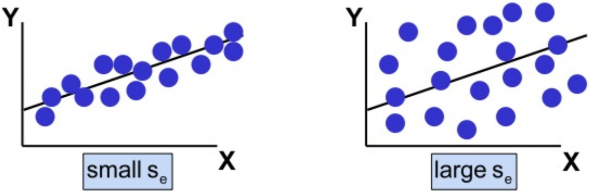
ANOVA

	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

Excel Output

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

$$s_{b_1} = 0.03297$$

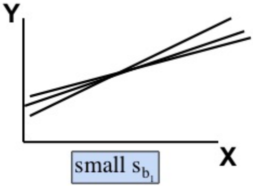
ANOVA

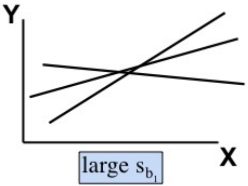
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	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

- We can use t-test for claims regarding the population slope in the same way we did in chapter 9.
- For example, suppose we would like to test whether $\beta_1 = \beta_1^*$:
  - Null and alternative hypotheses:
    - $H_0 : \beta_1 = \beta_1^*$
    - $H_0 : \beta_1 \neq \beta_1^*$
  - Test statistic:
    - $t_{n-2} = \frac{b_1 - \beta_1^*}{s_{b_1}} \qquad [df = n-2]$
  - Decision Rule:
    - Reject if $|t_{n-2}| > t_{n-2,\alpha/2}$
- Of course, we can address one-sided tests too:

$$H_0 : \beta_1 \geq \beta_1^* \quad \text{or} \quad H_0 : \beta_1 \leq \beta_1^*$$

- In these cases, the test statistic remains the same but the decision rule will have to be modified as we discussed in Ch. 9.

A special t-Test: Test of Significance

- For $\beta_1^* = 0$, t-test is called the significance test for the slope coefficient, and H_0 and H_1 have a natural meaning:
  - Null and alternative hypotheses:
    - $H_0 : \beta_1 = 0$: no linear relationship between X and Y.
    - $H_1 : \beta_1 \neq 0$: a linear relationship does exist between X and Y.
  - Test statistic:
    - $t_{n-2} = \frac{b_1 - 0}{s_{b_1}} = \frac{b_1}{s_{b_1}}, \qquad [df = n-2]$
  - Decision Rule:
    - Reject if $|t_{n-2}| > t_{n-2,\alpha/2}$
- Note: t-stats that are presented in the regression output by default are **only** for the case $H_0 : \beta_1 = 0$, i.e. they are test statistics for the significance test. If you would like to test something else as we discussed in the previous slide, you need to compute the corresponding t-statistic first.

Inference about the Slope: t Test

(continued)

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 significantly 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

$$t_8 = \frac{b_1 - 0}{s_{b_1}} = \frac{0.10977 - 0}{0.03297} = 3.32938$$

Inferences about the Slope: t Test Example

(continued)

Test Statistic: $t_8 = 3.329$

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

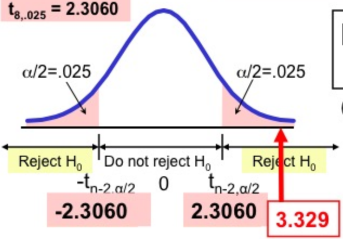
From Excel output:

b_1 $\quad s_{b_1}$ $\quad t$

	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

(continued)

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

(continued)

	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

Hypothesis Test for Population Slope Using the F Distribution

- F test statistic:

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

where

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

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

k: the number of independent variables in the regression model
n: sample size

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

Hypothesis Test for Population Slope Using the F Distribution

- An alternate test for the hypothesis that the slope is zero:
  $H_0 : \beta_1 = 0$
  $H_1 : \beta_1 \neq 0$
- Use the F statistic:

$$F = \frac{MSR}{MSE} = \frac{SSR}{s_e^2}$$

- The decision rule is

$$\text{reject } H_0 \text{ if } F \geq F_{1,n-2,\alpha}$$

Excel Output

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

(continued)

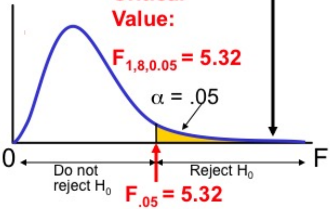
$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

$\alpha = .05$

$\underbrace{df_1 = 1 \qquad df_2 = 8}$

Critical Value:



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

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

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

- (Example Cont'd) Predict the price for a house with 2000 square feet:

$$\begin{aligned} \text{House}\hat{\ }\text{Price} &= 98.25 + 0.1098(sq.ft) \\ &= 98.25 + 0.1098(2000) \\ &= 317.85 \end{aligned}$$

- 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