Section 3: Simple Linear Regression

Jared S. Murray

Regression: General Introduction

- Regression analysis is the most widely used statistical tool for understanding relationships among variables
- It provides a conceptually simple method for investigating functional relationships between one or more factors and an outcome of interest
- The relationship is expressed in the form of an equation or a model connecting the response or dependent variable and one or more explanatory or predictor variable

Why?

Straight prediction questions:

- For who much will my house sell?
- How many runs per game will the Red Sox score this year?
- Will this person like that movie? (Netflix rating system)

Explanation and understanding:

- What is the impact of MBA on income?
- How does the returns of a mutual fund relates to the market?
- Does Walmart discriminates against women regarding salaries?

1st Example: Predicting House Prices

Problem:

- Predict market price based on observed characteristics

Solution:

- Look at property sales data where we know the price and some observed characteristics.
- Build a decision rule that predicts price as a function of the observed characteristics.

Predicting House Prices

What characteristics do we use?

We have to define the variables of interest and develop a specific quantitative measure of these variables

- Many factors or variables affect the price of a house
  - size
  - number of baths
  - garage, air conditioning, etc
  - neighborhood

Predicting House Prices

To keep things super simple, let's focus only on size.

The value that we seek to predict is called the dependent (or output) variable, and we denote this:

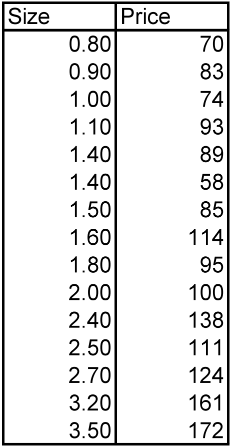
- $Y =$ price of house (e.g. thousands of dollars)

The variable that we use to guide prediction is the explanatory (or input) variable, and this is labelled

- $X =$ size of house (e.g. thousands of square feet)

Predicting House Prices

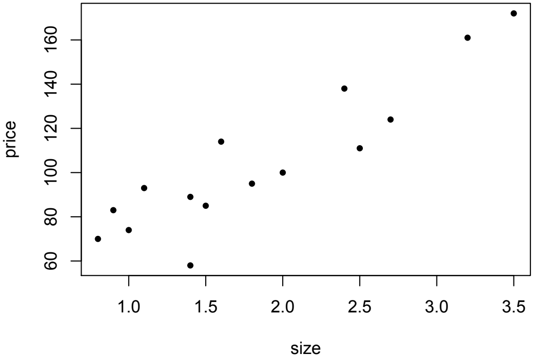
What does this data look like?



Size	Price
0.80	70
0.90	83
1.00	74
1.10	93
1.40	89
1.40	58
1.50	85
1.60	114
1.80	95
2.00	100
2.40	138
2.50	111
2.70	124
3.20	161
3.50	172

Predicting House Prices

It is much more useful to look at a scatterplot



In other words, view the data as points in the $X \times Y$ plane.

Regression Model

$Y =$ response or outcome variable

$X =$ explanatory or input variables

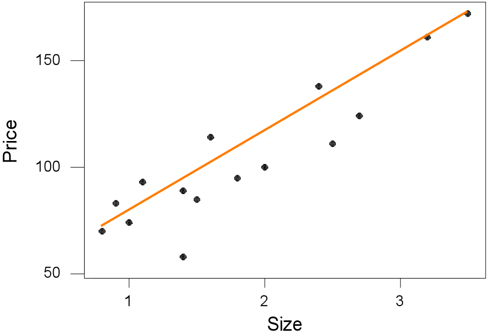
A linear relationship is written

$$Y = b_0 + b_1 X + e$$

Linear Prediction

Appears to be a linear relationship between price and size:

As size goes up, price goes up.



The line shown was fit by the "eyeball" method.

Linear Prediction

Recall that the equation of a line is:

$$Y = b_0 + b_1 X + e$$

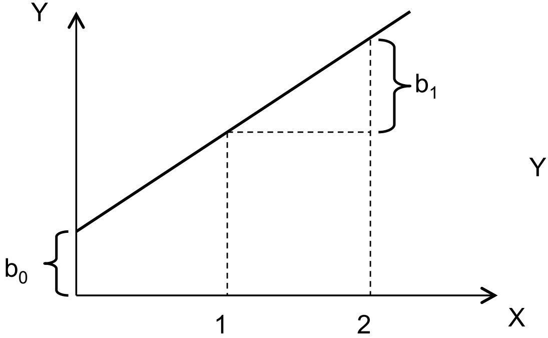
Where b_0 is the intercept and b_1 is the slope.

The intercept value is in units of Y (\$1,000).

The slope is in units of Y *per* units of X (\$1,000/1,000 sq ft).

The residual e is in units of Y (\$1,000).

Linear Prediction



$Y = b_0 + b_1X$

Our "eyeball" line has $b_0 = 35$, $b_1 = 40$.

Linear Prediction

We can now predict the price of a house when we know only the size; just read the value off the line that we've drawn.

For example, given a house with of size $X = 2.2$.

Predicted price $\hat{Y} = 35 + 40(2.2) = 123$.

Note: Conversion from 1,000 sq ft to \$1,000 is done for us by the slope coefficient (b_1)

Linear Prediction

Can we do better than the eyeball method?

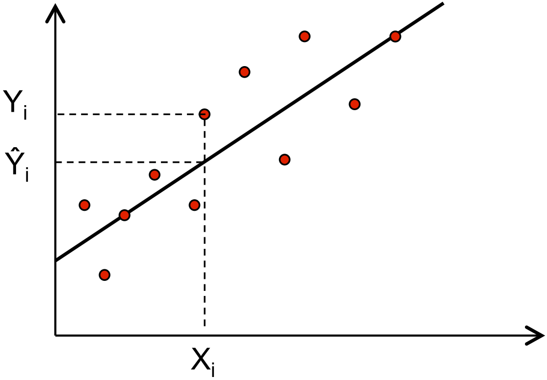
We desire a strategy for estimating the slope and intercept parameters in the model $\hat{Y} = b_0 + b_1 X$

A reasonable way to fit a line is to minimize the amount by which the fitted value differs from the actual value.

This amount is called the residual.

Linear Prediction

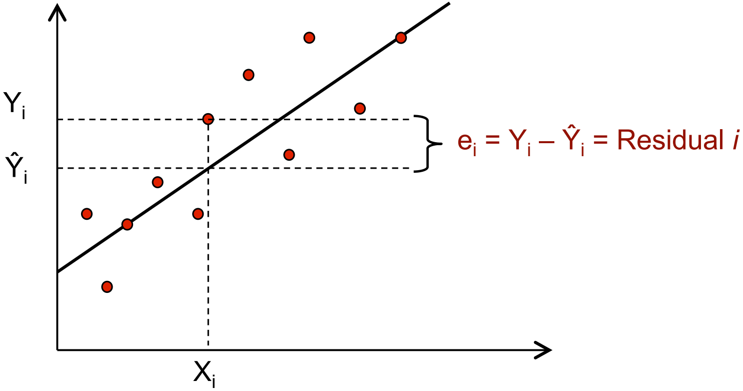
What is the "fitted value"?



The dots are the observed values and the line represents our fitted values given by $\hat{Y}_i = b_0 + b_1 X_1$.

Linear Prediction

What is the “residual”’ for the *i*th observation’?



We can write $Y_i = \hat{Y}_i + (Y_i - \hat{Y}_i) = \hat{Y}_i + e_i$.

Least Squares

Ideally we want to minimize the size of all residuals:

- If they were all zero we would have a perfect line.
- Trade-off between moving closer to some points and at the same time moving away from other points.

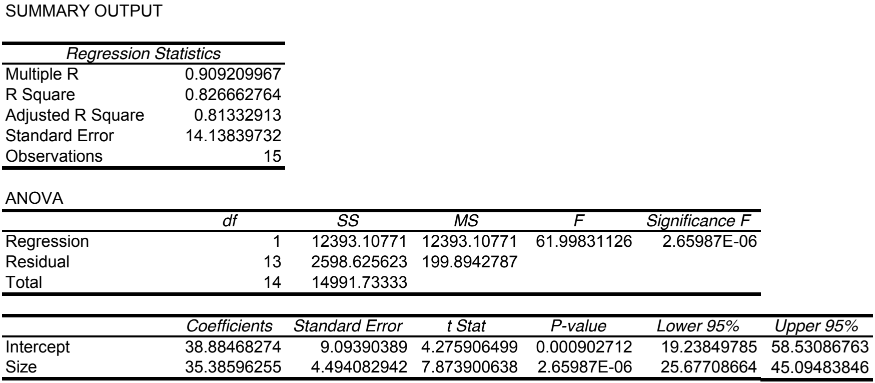
The line fitting process:

- Give weights to all of the residuals.
- Minimize the "total" of residuals to get best fit.

Least Squares chooses b_0 and b_1 to minimize $\sum_{i=1}^{N} e_i^2$

$$\sum_{i=1}^{N} e_i^2 = e_1^2 + e_2^2 + \cdots + e_N^2 = (Y_1 - \hat{Y}_1)^2 + (Y_2 - \hat{Y}_2)^2 + \cdots + (Y_N - \hat{Y}_N)^2$$

Least Squares – Excel Output



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.909209967
R Square	0.826662764
Adjusted R Square	0.81332913
Standard Error	14.13839732
Observations	15

ANOVA

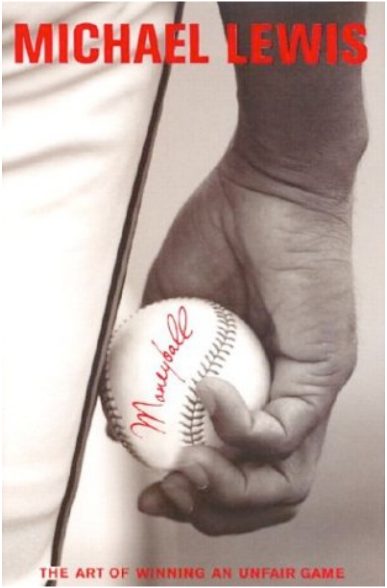
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	12393.10771	12393.10771	61.99831126	2.65987E-06
Residual	13	2598.625623	199.8942787		
Total	14	14991.73333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	38.88468274	9.09390389	4.275906499	0.000902712	19.23849785	58.53086763
Size	35.38596255	4.494082942	7.873900638	2.65987E-06	25.67708664	45.09483846

2nd Example: Offensive Performance in Baseball

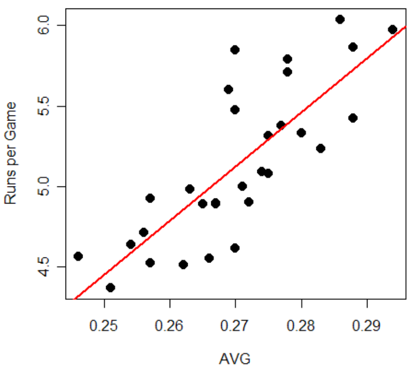
1. Problems:
   - Evaluate/compare traditional measures of offensive performance
   - Help evaluate the worth of a player
2. Solutions:
   - Compare *prediction rules* that forecast runs as a function of either AVG (batting average), SLG (slugging percentage) or OBP (on base percentage)

2nd Example: Offensive Performance in Baseball



Baseball Data – Using AVG

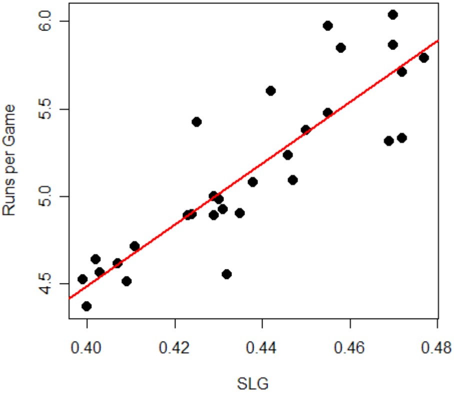
Each observation corresponds to a team in MLB. Each quantity is the average over a season.



- Y = runs per game; X = AVG (average)

LS fit: Runs/Game = -3.93 + 33.57 AVG

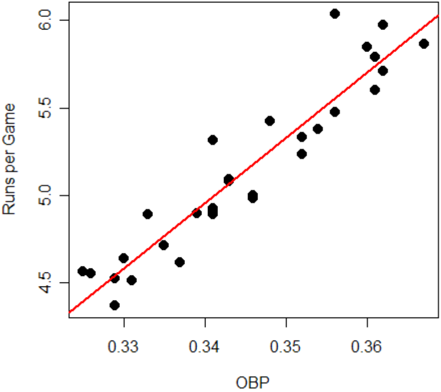
Baseball Data – Using SLG



- Y = runs per game
- X = SLG (slugging percentage)

LS fit: Runs/Game = -2.52 + 17.54 SLG

Baseball Data – Using OBP



- Y = runs per game
- X = OBP (on base percentage)

LS fit: Runs/Game = -7.78 + 37.46 OBP

Baseball Data

- What is the best prediction rule?
- Let's compare the predictive ability of each model using the average squared error

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N} e_i^2} = \sqrt{\frac{\sum_{i=1}^{N}\left(\widehat{Runs_i} - Runs_i\right)^2}{N}}$$

Place your Money on OBP!!!

	Root Average Squared Error
AVG	0.29
SLG	0.23
OBP	0.16

Estimation of Error Variance

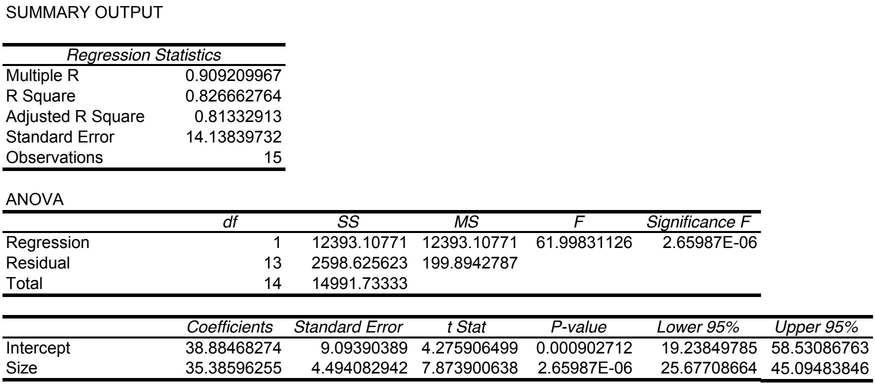
We can quantify the variability around the line by computing the variance of the residuals:

$$s = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} e_i^2} = \sqrt{\frac{\sum (Y_i - \hat{Y}_i)^2}{n-2}}$$

This is measured in the same units as Y. It's also called the regression standard error.

(Don't worry about this $n-2$ yet!).

Back to the House Data



SUMMARY OUTPUT

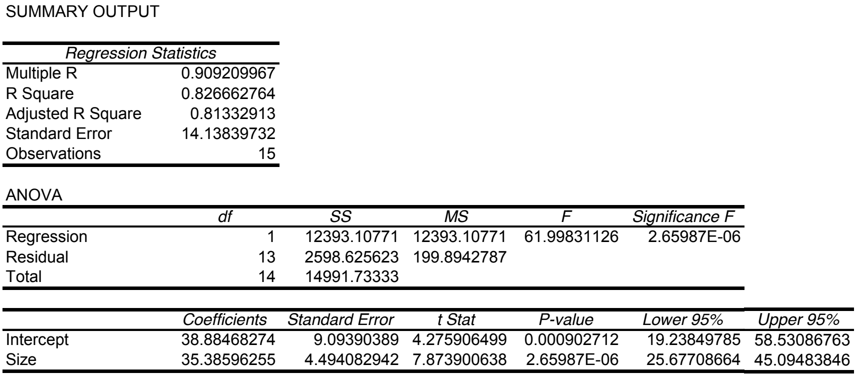
Regression Statistics	
Multiple R	0.909209967
R Square	0.826662764
Adjusted R Square	0.81332913
Standard Error	14.13839732
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	12393.10771	12393.10771	61.99831126	2.65987E-06
Residual	13	2598.625623	199.8942787		
Total	14	14991.73333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	38.88468274	9.09390389	4.275906499	0.000902712	19.23849785	58.53086763
Size	35.38596255	4.494082942	7.873900638	2.65987E-06	25.67708664	45.09483846

Back to the House Data



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.909209967
R Square	0.826662764
Adjusted R Square	0.81332913
Standard Error	14.13839732
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	12393.10771	12393.10771	61.99831126	2.65987E-06
Residual	13	2598.625623	199.8942787		
Total	14	14991.73333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	38.88468274	9.09390389	4.275906499	0.000902712	19.23849785	58.53086763
Size	35.38596255	4.494082942	7.873900638	2.65987E-06	25.67708664	45.09483846

$$R^2 = 0.82$$

The Simple Linear Regression Model

The power of statistical inference comes from the ability to make precise statements about the accuracy of the forecasts.

In order to do this we must invest in a probability model.

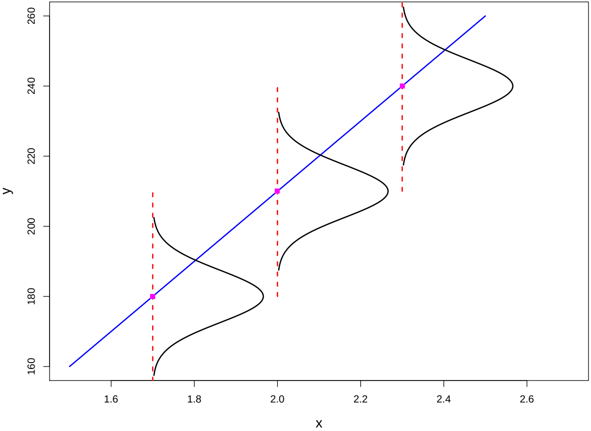
Simple Linear Regression Model: $Y = \beta_0 + \beta_1 X + \varepsilon$

$$\varepsilon \sim \mathrm{N}(0, \sigma^2)$$

- $\beta_0 + \beta_1 X$ represents the "true line"; The part of Y that depends on X.
- The error term ε is independent "idosyncratic noise"; The part of Y not associated with X.

The Simple Linear Regression Model

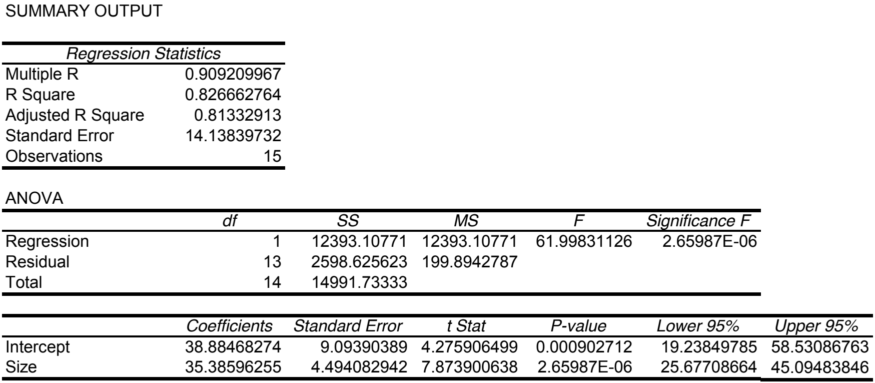
$$Y = \beta_0 + \beta_1 X + \varepsilon$$



The conditional distribution for Y given X is Normal:

$$Y|X = x \sim \mathrm{N}(\beta_0 + \beta_1 x, \sigma^2).$$

Back to the House Data



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.909209967
R Square	0.826662764
Adjusted R Square	0.81332913
Standard Error	14.13839732
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	12393.10771	12393.10771	61.99831126	2.65987E-06
Residual	13	2598.625623	199.8942787		
Total	14	14991.73333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	38.88468274	9.09390389	4.275906499	0.000902712	19.23849785	58.53086763
Size	35.38596255	4.494082942	7.873900638	2.65987E-06	25.67708664	45.09483846

Sampling Distribution of Least Squares Estimates

How much do our estimates depend on the particular random sample that we happen to observe? Imagine:

- Randomly draw different samples of the same size.
- For each sample, compute the estimates b_0, b_1, and s.

If the estimates don't vary much from sample to sample, then it doesn't matter which sample you happen to observe.

If the estimates do vary a lot, then it matters which sample you happen to observe.

Sampling Distribution of b_1

The sampling distribution of b_1 describes how estimator $b_1 = \hat{\beta}_1$ varies over different samples with the X values fixed.

It turns out that b_1 is normally distributed (approximately): $b_1 \sim N(\beta_1, s_{b_1}^2)$.

- b_1 is unbiased: $E[b_1] = \beta_1$.
- s_{b_1} is the standard error of b_1. In general, the standard error is the standard deviation of an estimate. It determines how close b_1 is to β_1.
- This is a number directly available from the regression output.

Confidence Intervals

Since $b_1 \sim N(\beta_1, s_{b_1}^2)$, Thus:

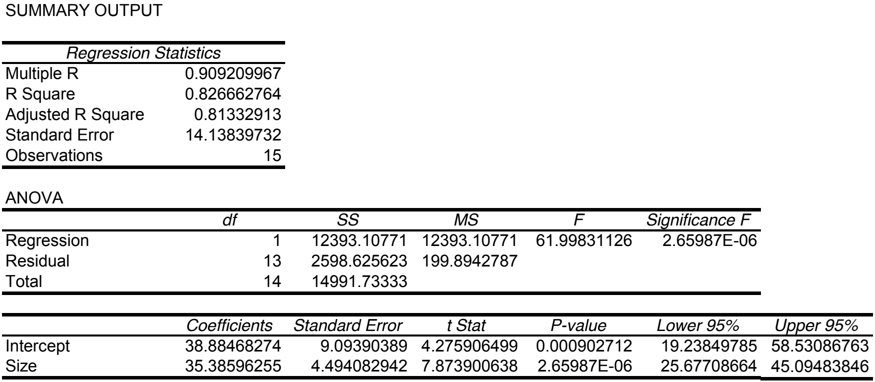
- 68% Confidence Interval: $b_1 \pm 1 \times s_{b_1}$
- 95% Confidence Interval: $b_1 \pm 2 \times s_{b_1}$
- 99% Confidence Interval: $b_1 \pm 3 \times s_{b_1}$

Same thing for b_0

- 95% Confidence Interval: $b_0 \pm 2 \times s_{b_0}$

The confidence interval provides you with a set of plausible values for the parameters

Back to the House Data



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.909209967
R Square	0.826662764
Adjusted R Square	0.81332913
Standard Error	14.13839732
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	12393.10771	12393.10771	61.99831126	2.65987E-06
Residual	13	2598.625623	199.8942787		
Total	14	14991.73333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	38.88468274	9.09390389	4.275906499	0.000902712	19.23849785	58.53086763
Size	35.38596255	4.494082942	7.873900638	2.65987E-06	25.67708664	45.09483846

$$[b_1 - 2 \times s_{b_1}; b_1 + 2 \times s_{b_1}] \approx [25.67; 45.09]$$

Testing

Suppose we want to assess whether or not β_1 equals a proposed value β_1^0. This is called hypothesis testing.

Formally we test the null hypothesis:

$H_0: \beta_1 = \beta_1^0$

vs. the alternative

$H_1: \beta_1 \neq \beta_1^0$

Testing

That are 2 ways we can think about testing:

1. Building a test statistic... the t-stat,

$$t = \frac{b_1 - \beta_1^0}{s_{b_1}}$$

   This quantity measures how many standard deviations the estimate (b_1) from the proposed value (β_1^0).

   If the absolute value of t is greater than 2, we need to worry (why?)... we reject the hypothesis.

Testing

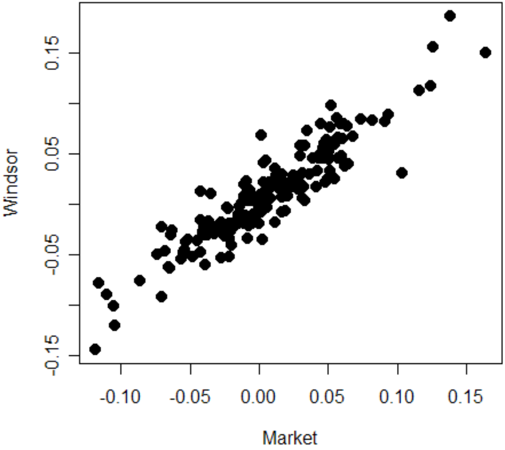
2. Looking at the confidence interval. If the proposed value is outside the confidence interval you reject the hypothesis.

   Notice that this is equivalent to the t-stat. An absolute value for t greater than 2 implies that the proposed value is outside the confidence interval... therefore reject.

   This is my preferred approach for the testing problem. You can't go wrong by using the confidence interval!

Example: Mutual Funds

Let's investigate the performance of the Windsor Fund, an aggressive large cap fund by Vanguard...



The plot shows monthly returns for Windsor vs. the S&P500

Example: Mutual Funds

Consider a CAPM regression for the Windsor mutual fund.

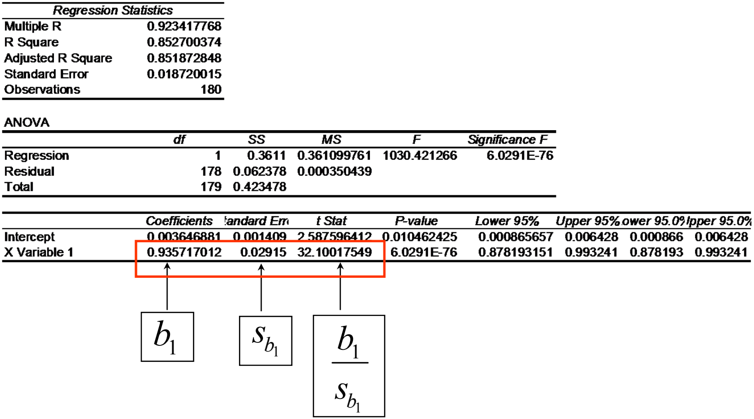
$$r_w = \beta_0 + \beta_1 r_{sp500} + \epsilon$$

Let's first test $\beta_1 = 0$

$H_0: \ \beta_1 = 0$. Is the Windsor fund related to the market?

$H_1: \ \beta_1 \neq 0$

Example: Mutual Funds



Regression Statistics	
Multiple R	0.923417768
R Square	0.852700374
Adjusted R Square	0.851872848
Standard Error	0.018720015
Observations	180

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.3611	0.361099761	1030.421266	6.0291E-76
Residual	178	0.062378	0.000350439		
Total	179	0.423478			

	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	0.003646881	0.001409	2.587596412	0.010462425	0.000865657	0.006428	0.000866	0.006428
X Variable 1	0.935717012	0.02915	32.10017549	6.0291E-76	0.878193151	0.993241	0.878193	0.993241

b_1 s_{b_1} $\frac{b_1}{s_{b_1}}$

- $t = 32.10...$ reject $\beta_1 = 0$!!
- the 95% confidence interval is $[0.87; 0.99]$... again, reject!!

Example: Mutual Funds

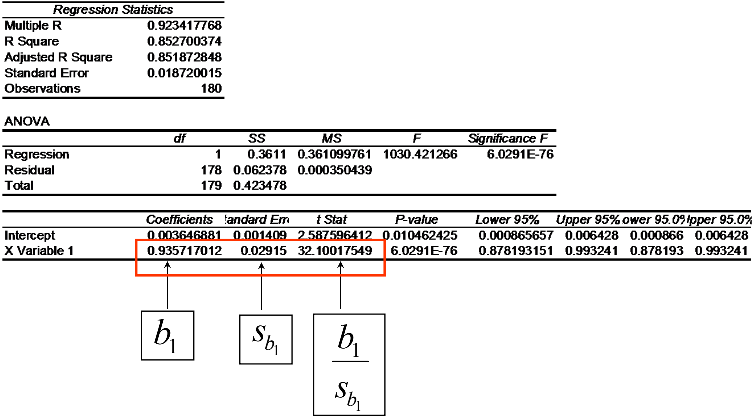
Now let's test $\beta_1 = 1$. What does that mean?

$H_0 :$ $\beta_1 = 1$ Windsor is as risky as the market.

$H_1 :$ $\beta_1 \neq 1$ and Windsor softens or exaggerates market moves.

We are asking whether or not Windsor moves in a different way than the market (e.g., is it more conservative?).

Example: Mutual Funds



Regression Statistics	
Multiple R	0.923417768
R Square	0.852700374
Adjusted R Square	0.851872848
Standard Error	0.018720015
Observations	180

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.3611	0.361099761	1030.421266	6.0291E-76
Residual	178	0.062378	0.000350439		
Total	179	0.423478			

	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*lpper 95.0%*
Intercept	0.003646881	0.001409	2.587596412	0.010462425	0.000865657	0.006428	0.000866	0.006428
X Variable 1	0.935717012	0.02915	32.10017549	6.0291E-76	0.878193151	0.993241	0.878193	0.993241

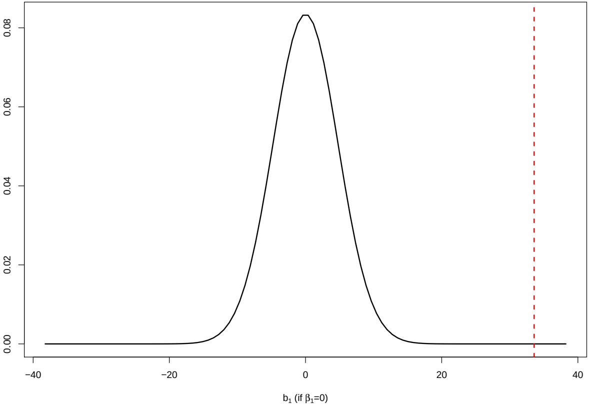
- $t = \frac{b_1 - 1}{s_{b_1}} = \frac{-0.0643}{0.0291} = -2.205$... reject.
- the 95% confidence interval is $[0.87; 0.99]$... again, reject, but...

P-values

- The *p*-value provides a measure of how weird your estimate is **if** the null hypothesis is true
- Small p-values are evidence against the null hypothesis
- In the AVG vs. R/G example... $H_0 : \beta_1 = 0$. How weird is our estimate of $b_1 = 33.57$?
- Remember: $b_1 \sim N(\beta_1, s^2_{b_1})$... If the null was true $(\beta_1 = 0)$, $b_1 \sim N(0, s^2_{b_1})$

P-values

- Where is 33.57 in the picture below?

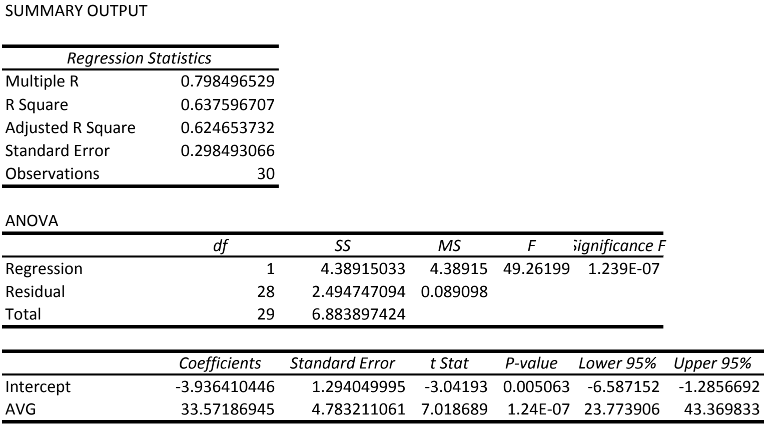


The *p*-value is the probability of seeing b_1 equal or greater than 33.57 in absolute terms. Here, *p*-value=0.000000124!!

Small p-value = bad null

P-values

- $H_0 : \beta_1 = 0$... p-value $=$ 1.24E-07... reject!



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.798496529
R Square	0.637596707
Adjusted R Square	0.624653732
Standard Error	0.298493066
Observations	30

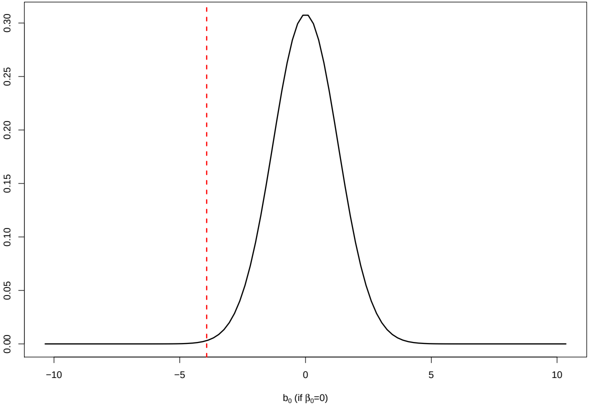
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	4.38915033	4.38915	49.26199	1.239E-07
Residual	28	2.494747094	0.089098		
Total	29	6.883897424			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-3.936410446	1.294049995	-3.04193	0.005063	-6.587152	-1.2856692
AVG	33.57186945	4.783211061	7.018689	1.24E-07	23.773906	43.369833

P-values

- How about $H_0 : \beta_0 = 0$? How weird is $b_0 = -3.936$?

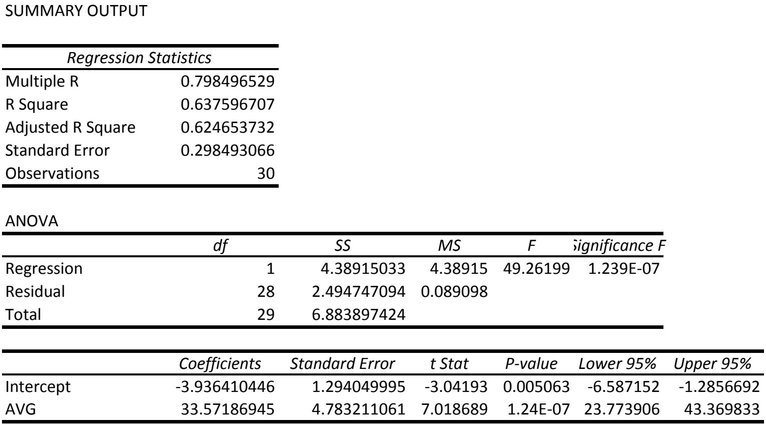


The *p*-value (the probability of seeing b_0 equal or greater than -3.936 in absolute terms) is 0.005.

Small p-value = bad null

P-values

- $H_0 : \beta_0 = 0$... p-value $= 0.005$... we still reject, but not with the same strength.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.798496529
R Square	0.637596707
Adjusted R Square	0.624653732
Standard Error	0.298493066
Observations	30

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	1	4.38915033	4.38915	49.26199	1.239E-07
Residual	28	2.494747094	0.089098		
Total	29	6.883897424			

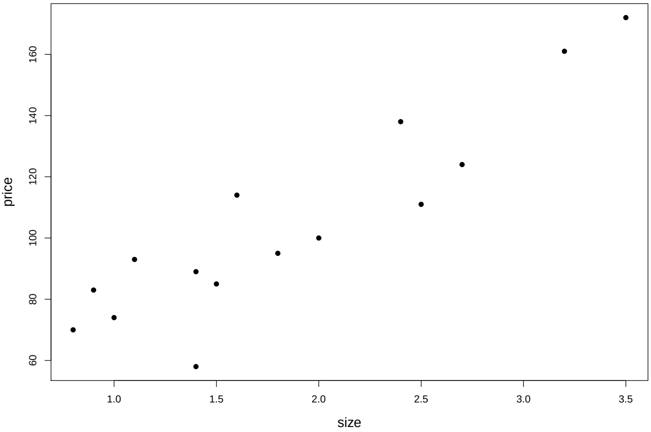
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-3.936410446	1.294049995	-3.04193	0.005063	-6.587152	-1.2856692
AVG	33.57186945	4.783211061	7.018689	1.24E-07	23.773906	43.369833

Testing – Summary

- Large t or small p-value mean the same thing...
- p-value < 0.05 is equivalent to a t-stat > 2 in absolute value
- Small p-value means something weird happen if the null hypothesis was true...
- Bottom line, small p-value $\rightarrow$ REJECT! Large t $\rightarrow$ REJECT!
- But remember, always look at the confidence interveal!

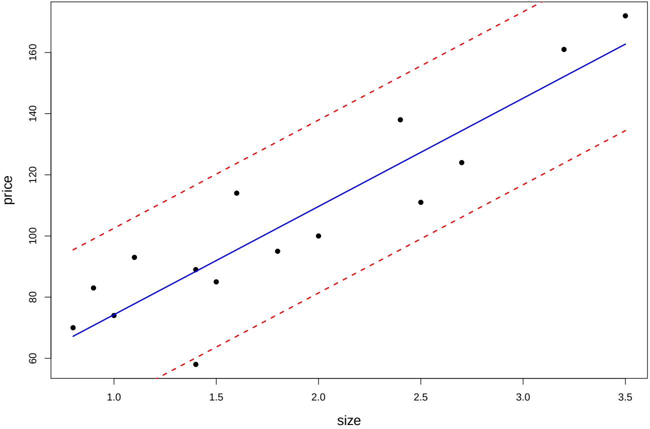
House Data – one more time!

- $R^2 = 82\%$
- Great R^2, we are happy using this model to predict house prices, right?



House Data – one more time!

- But, $s = 14$ leading to a predictive interval width of about US$60,000!! How do you feel about the model now?
- As a practical matter, s is a much more relevant quantity than R^2. Once again, *intervals* are your friend!



Dummy Variables

Imagine you are a trial lawyer and you want to file a suit against a company for salary discrimination... you gather the following data...

```
 Gender    Salary
1    Male   32.0
2  Female   39.1
3  Female   33.2
4  Female   30.6
5    Male   29.0
... ... ...
208 Female  30.0
```

Dummy Variables

You want to relate salary(Y) to gender(X)... how can we do that?

Gender is an example of a categorical variable. The variable gender separates our data into 2 groups or categories. The question we want to answer is: *"how is your salary related to which group you belong to..."*

Could we think about additional examples of categories potentially associated with salary?

Dummy Variables

We can use regression to answer these question but we need to recode the categorical variable into a dummy variable

```
 Gender    Salary  Sex
1     Male  32.00   1
2   Female  39.10   0
3   Female  33.20   0
4   Female  30.60   0
5     Male  29.00   1
... ... ...
208  Female  30.00  0
```

Note: In Excel you can create the dummy variable using the formula:

=IF(Gender="Male",1,0)

Dummy Variables

Now you can present the following model in court:

$$Salary_i = \beta_0 + \beta_1 Sex_i + \epsilon_i$$

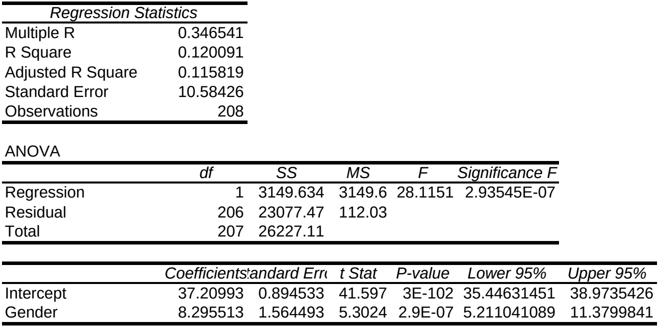
How do you interpret β_1?

$$\begin{aligned} E[Salary|Sex = 0] &= \beta_0 \\ E[Salary|Sex = 1] &= \beta_0 + \beta_1 \end{aligned}$$

β_1 is the male/female difference

Dummy Variables

$$Salary_i = \beta_0 + \beta_1 Sex_i + \epsilon_i$$



Regression Statistics	
Multiple R	0.346541
R Square	0.120091
Adjusted R Square	0.115819
Standard Error	10.58426
Observations	208

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3149.634	3149.6	28.1151	2.93545E-07
Residual	206	23077.47	112.03		
Total	207	26227.11			

	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	37.20993	0.894533	41.597	3E-102	35.44631451	38.9735426
Gender	8.295513	1.564493	5.3024	2.9E-07	5.211041089	11.3799841

$\hat{\beta}_1 = b_1 = 8.29...$ on average, a male makes approximately \$8,300 more than a female in this firm.

How should the plaintiff's lawyer use the confidence interval in his presentation?