STA 371G: Statistics and Modeling

Simple Linear Regression: Sampling Distributions and Confidence Intervals for Regression Parameters

Mingyuan Zhou
McCombs School of Business
The University of Texas at Austin

http://mingyuanzhou.github.io/teaching

Estimation for the SLR Model

SLR assumes every observation in the dataset was generated by the model:

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

This is a model for the conditional distribution of Y given X.

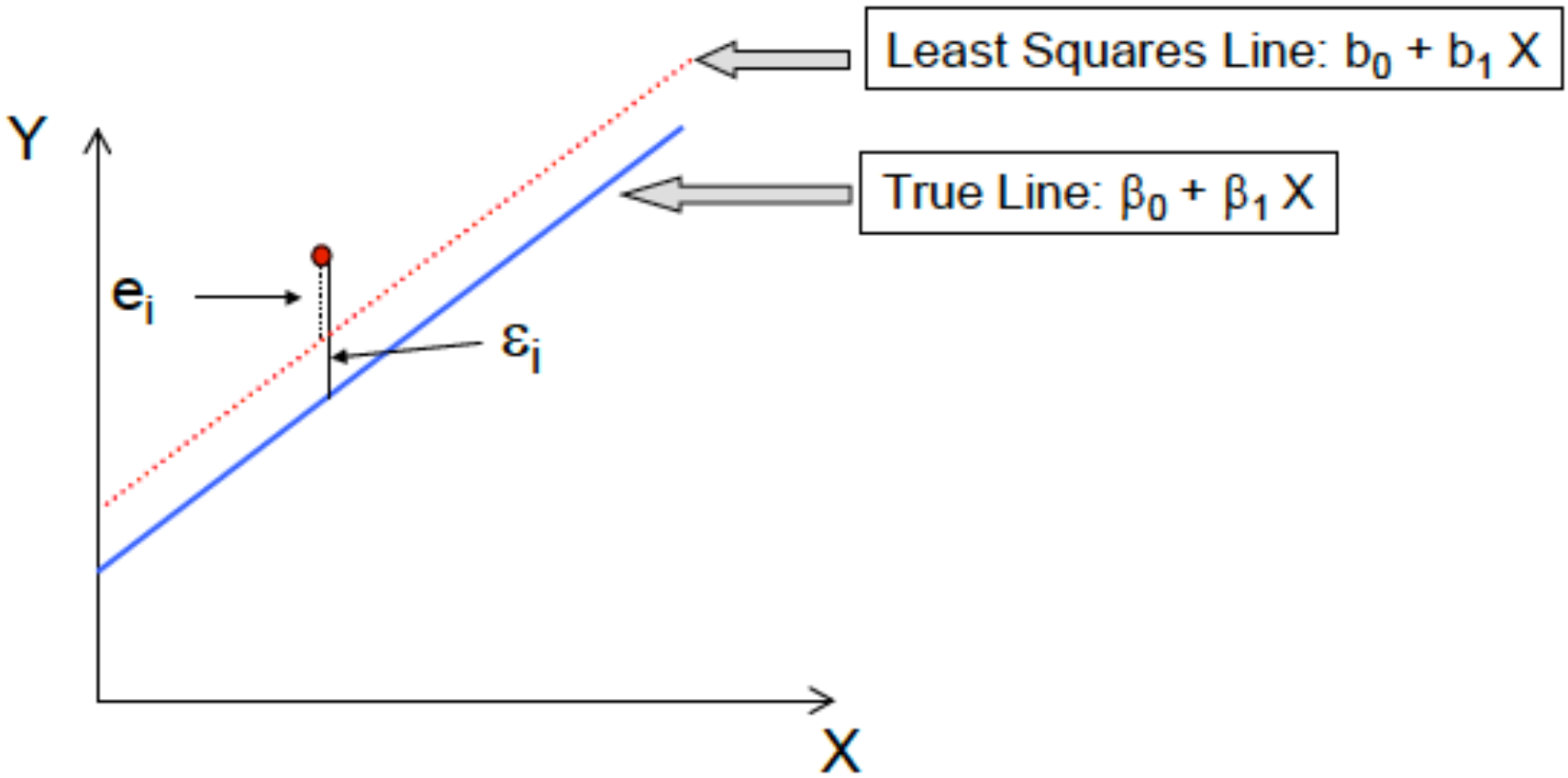
We use Least Squares *to estimate* β_0 and β_1:

$$\hat{\beta}_1 = b_1 = r_{xy} \times \frac{s_y}{s_x}$$

$$\hat{\beta}_0 = b_0 = \bar{y} - b_1 \bar{x}$$

Estimation for the SLR Model

NOTE!!: β_0 **is not** b_0, β_1 **is not** b_1 **and** ε_i **is not e**



Degrees of Freedom (Optional)

Degrees of Freedom is the number of times you get to observe useful information about the variance you're trying to estimate.

For example, to estimate the population variance σ^2 from a random sample, since $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, we've only had $n-1$ chances for deviation from the sample mean, and we estimate σ^2 with $s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)$.

In SLR, since $\sum_{i=1}^{n} e_i = 0$ and $\sum_{i=1}^{n}(x_i - \bar{x})e_i = 0$, we only get $n-2$ real observations of variability $\Rightarrow DoF = n-2$.

In regression with p coefficients (e.g., $p = 2$ in SLR), we only get $n-p$ real observations of variability $\Rightarrow DoF = n-p$.

Estimation of Error Variance

In SLR, we estimate σ^2 with:

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{SSE}{n-2}$$

(2 is the number of regression coefficients; i.e. 2 for β_0 and β_1).

We have $n-2$ degrees of freedom because 2 have been "used up" in the estimation of b_0 and b_1.

We usually use $s = \sqrt{SSE/(n-2)}$, in the same units as Y. It's also called the regression standard error.

Estimation of Error Variance

Where is s in the 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

s

ANOVA

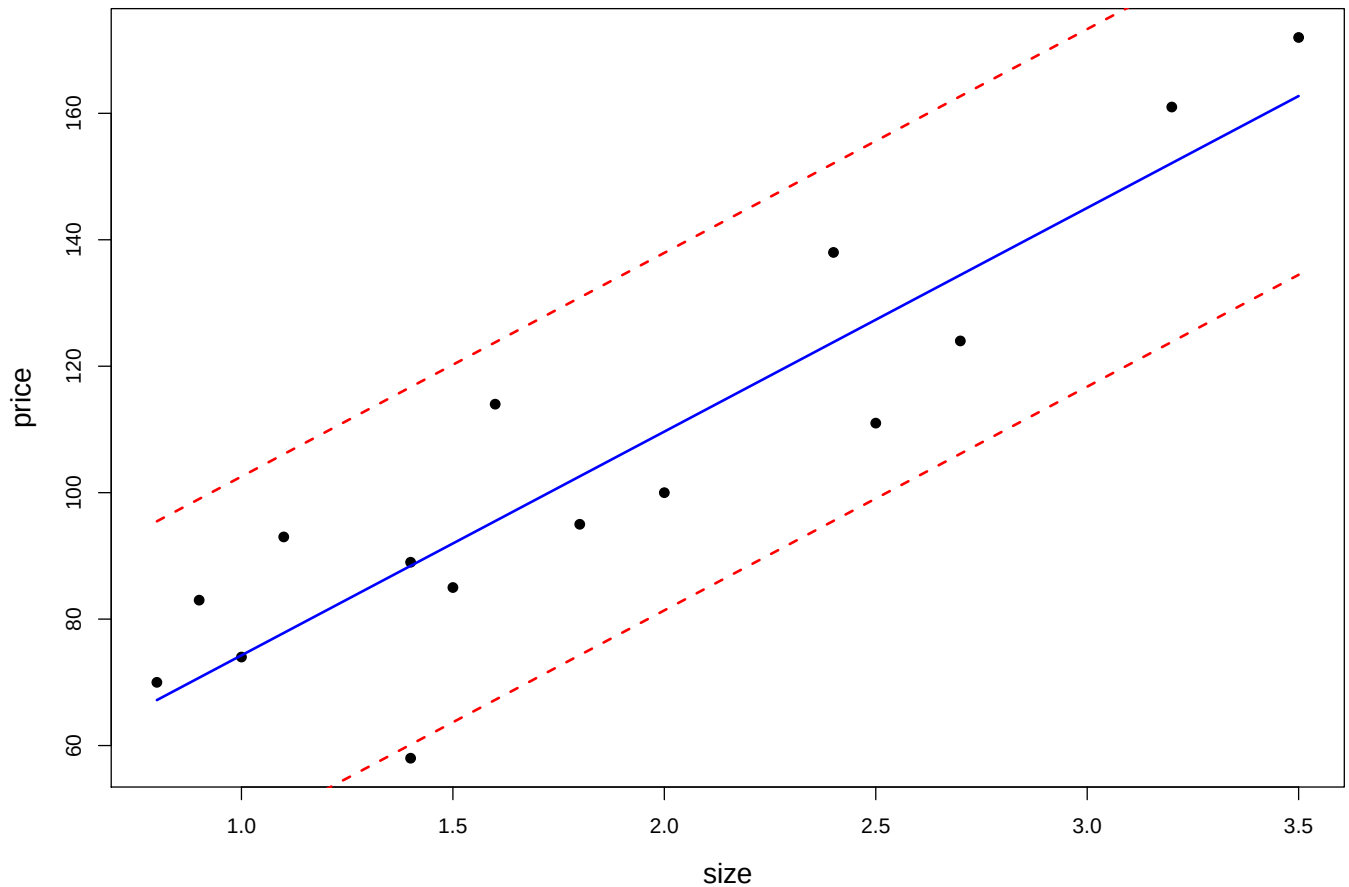
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	12393.10771	12393.10771	61.99831126	2.65987E-06
Residual	13	2598.625623	199.8942787		
Total	14	14991.73333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	38.88468274	9.09390389	4.275906499	0.000902712	19.23849785	58.53086763	19.23849785	58.53086763
X Variable 1	35.38596255	4.494082942	7.873900638	2.65987E-06	25.67708664	45.09483846	25.67708664	45.09483846

Remember that whenever you see "standard error" read it as estimated standard deviation: σ is the standard deviation.

One Picture Summary of SLR

- The plot below has the house data, the fitted regression line $(b_0 + b_1 X)$ and $\pm 2 * s$...
- From this picture, what can you tell me about β_0, β_1 and σ^2? How about b_0, b_1 and s^2?



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.

The Importance of Understanding Variation

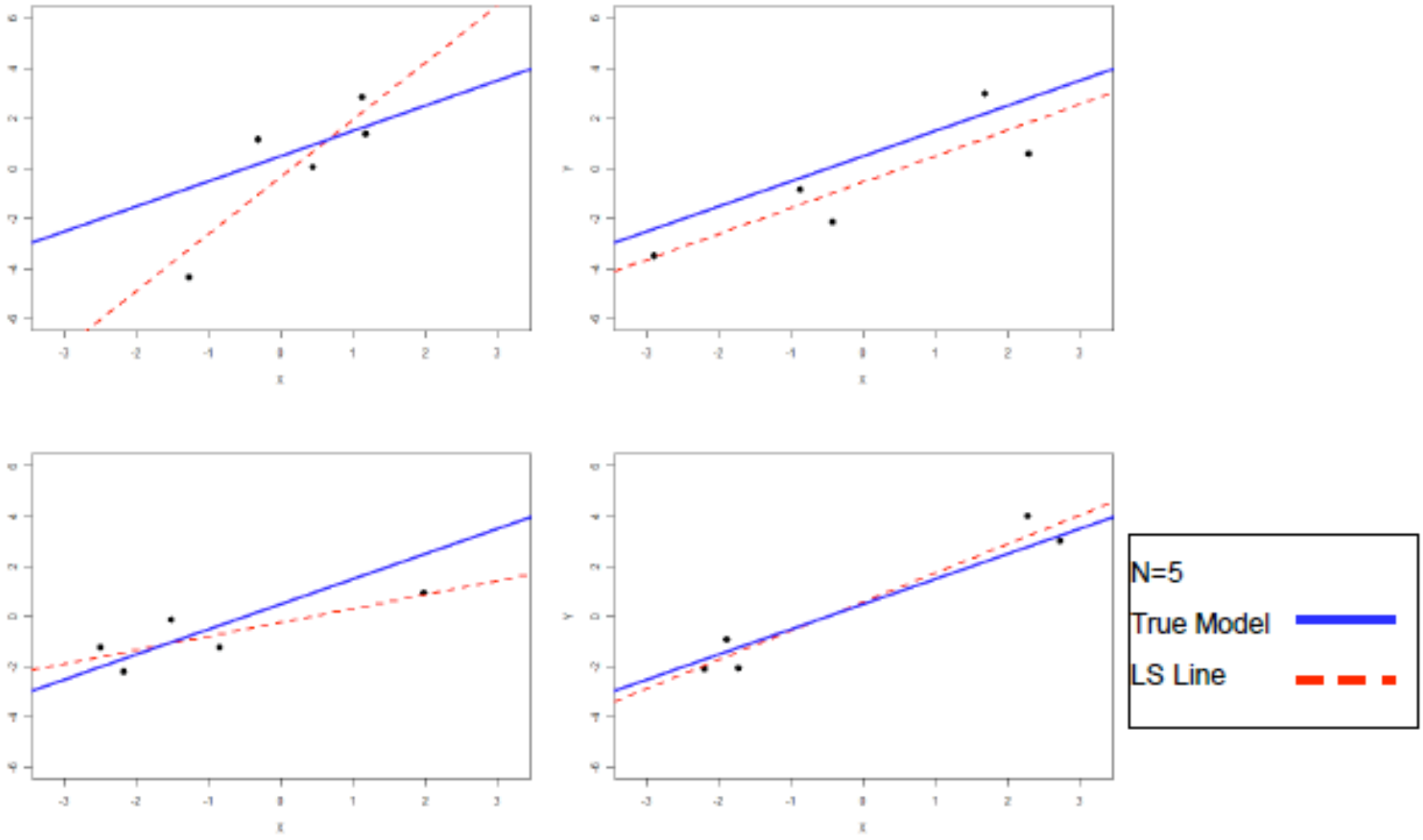
When estimating a quantity, it is vital to develop a notion of the precision of the estimation; for example:

- estimate the slope of the regression line
- estimate the value of a house given its size
- estimate the expected return on a portfolio
- estimate the value of a brand name
- estimate the damages from patent infringement

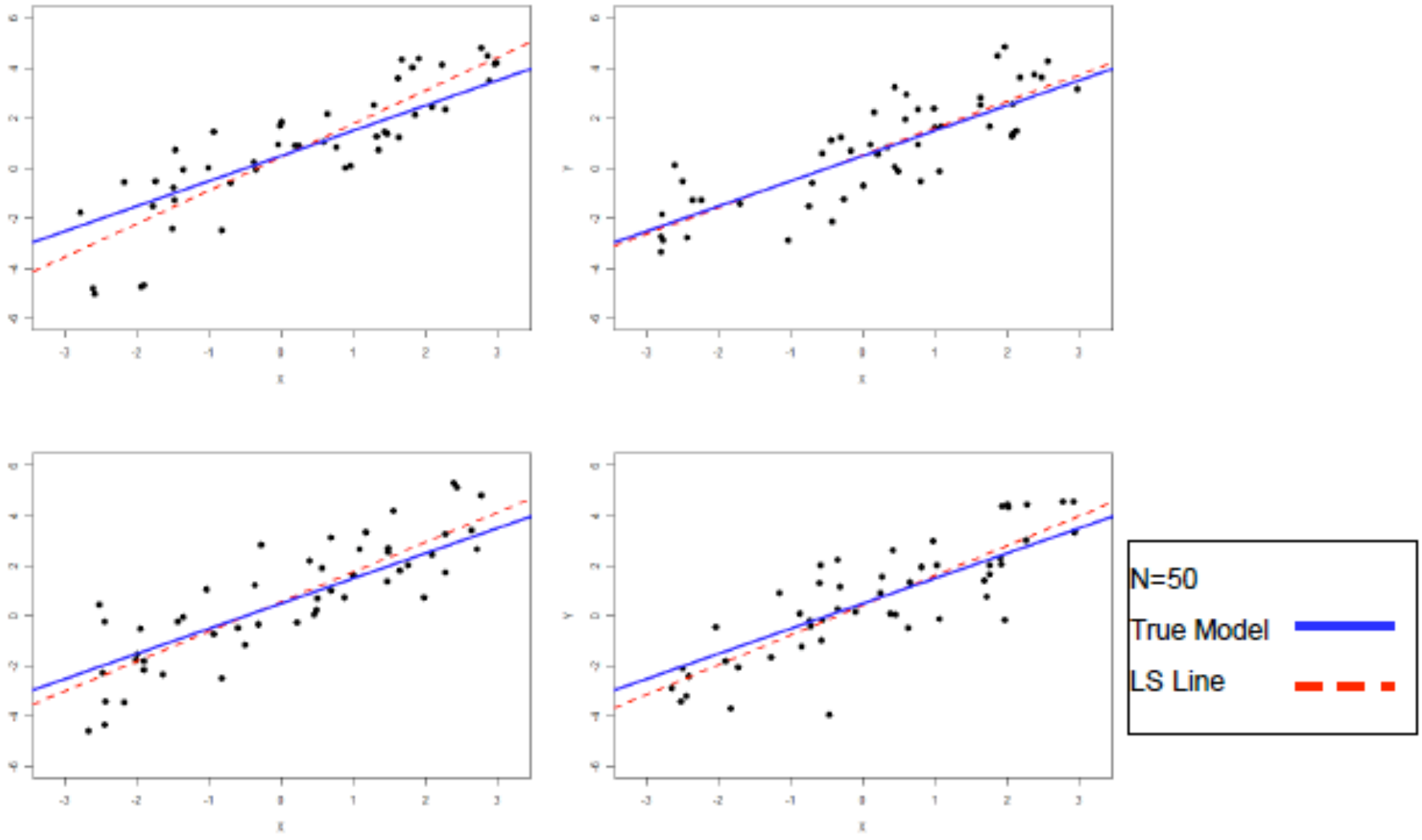
Why is this important?

We are making decisions based on estimates, and these may be very sensitive to the accuracy of the estimates!

Sampling Distribution of Least Squares Estimates



Sampling Distribution of Least Squares Estimates



Sampling Distribution of Least Squares Estimates

LS lines are much closer to the true line when $n = 50$.

For $n = 5$, some lines are close, others aren't:

we need to get "lucky" if the sample size is small

Sampling Distribution of Least Squares Estimates

- The least squares estimates of regression coefficients are:

$$b_1 = r_{xy}\frac{s_y}{s_x} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad b_0 = \bar{y} - b_1\bar{x}$$

- The model assumes that $y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ \epsilon_i \sim \mathcal{N}(0, \sigma^2)$, thus

$$\bar{y} = \beta_0 + \beta_1\bar{x} + \bar{\epsilon}$$

$$y_i - \bar{y} = \beta_1(x_i - \bar{x}) + (\epsilon_i - \bar{\epsilon})$$

Sampling Distribution (Optional)

Assuming we know the variance σ^2:

- One may show that b_1 is normal distributed, with

$$E(b_1) = \beta_1, \quad \sigma_{b_1}^2 = Var(b_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$b_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

- b_0 is normal distributed, with

$$E(b_0) = \beta_0, \quad \sigma_{b_0}^2 = Var(b_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

$$b_0 \sim \mathcal{N}\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)\right)$$

Normal and Student's t (Optional)

Recall what *Student* discovered:

If $\theta \sim N(\mu, \sigma^2)$, but you estimate $\sigma^2 \approx s^2$ based on $n - p$ degrees of freedom, then $\theta \sim t_{n-p}(\mu, s^2)$.

For example:

- $\bar{Y} \sim t_{n-1}(\mu, s_y^2/n)$.
- $b_0 \sim t_{n-2}\left(\beta_0, s_{b_0}^2\right)$ and $b_1 \sim t_{n-2}\left(\beta_1, s_{b_1}^2\right)$

The t distribution is just a fat-tailed version of the normal. As $n - p \longrightarrow \infty$, our tails get skinny and the t becomes normal.

Standardized Normal and Student's t (Optional)

We'll also usually standardize things:

$$\frac{b_j - \beta_j}{\sigma_{b_j}} \sim N(0,1) \Longrightarrow \frac{b_j - \beta_j}{s_{b_j}} \sim t_{n-2}(0,1)$$

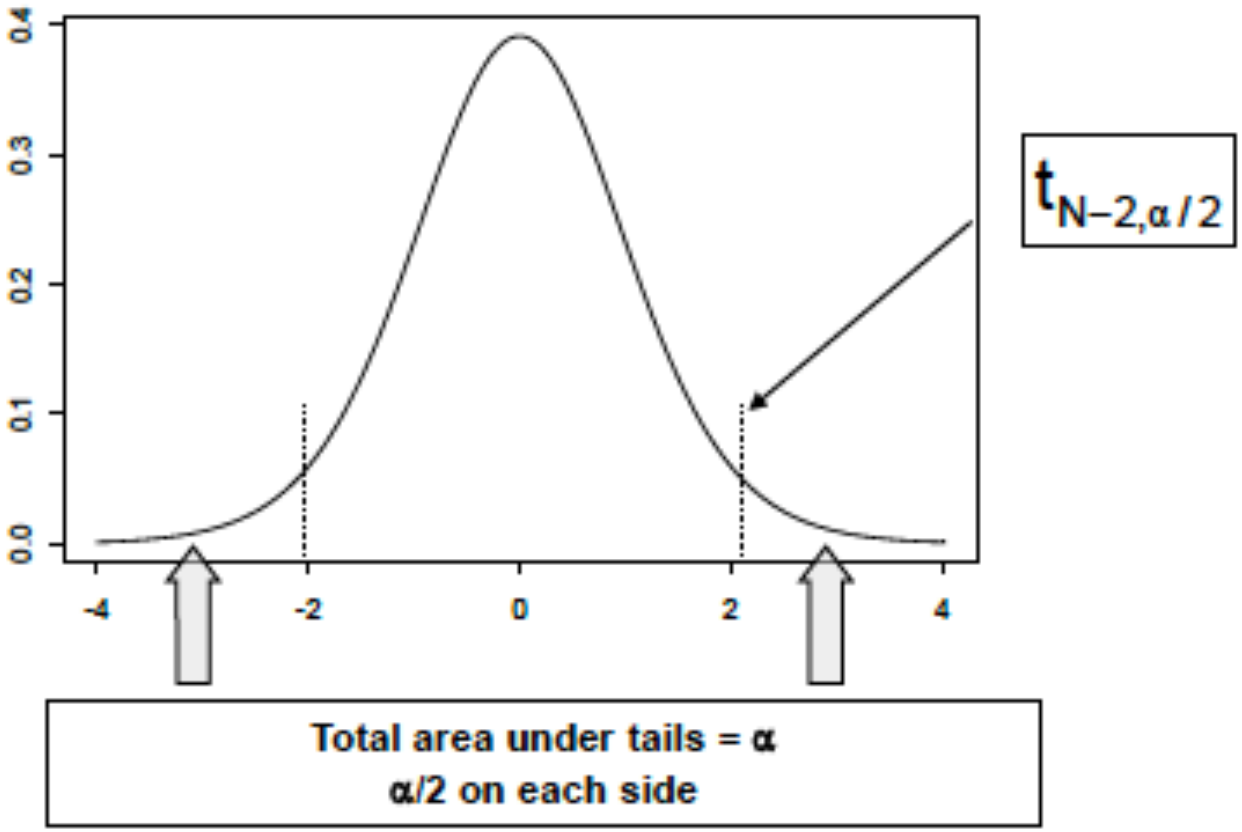
We use $Z \sim N(0,1)$ and $Z_{n-p} \sim t_{n-p}(0,1)$ to represent standard random variables.

Notice that the t and normal distributions depend upon assumed values for β_j: this forms the basis for confidence intervals, hypothesis testing, and p-values.

Testing and Confidence Intervals (Optional)

Suppose Z_{n-p} is distributed $t_{n-p}(0,1)$. A centered interval is

$$\mathrm{P}(-t_{n-p,\alpha/2} < Z_{n-p} < t_{n-p,\alpha/2}) = 1-\alpha$$



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.
- s_{b_1} is directly available from the regression output.

Sampling Distribution of b_1

Can we intuit what should be in the formula for s_{b_1}?

- How should s figure in the formula?
- What about n?
- Anything else?

$$s_{b_1}^2 = \frac{s^2}{\sum (X_i - \bar{X})^2} = \frac{s^2}{(n-1)s_x^2}$$

Three Factors:

sample size (n), error variance (s^2), and X-spread (s_x).

Sampling Distribution of b_0

The intercept is also normal and unbiased: $b_0 \sim N(\beta_0, s^2_{b_0})$.

$$s^2_{b_0} = \mathrm{var}(b_0) = s^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s^2_x} \right)$$

What is the intuition here?

Estimated Variance

We estimate variation with "sample standard deviations":

$$s_{b_1} = \sqrt{\frac{s^2}{(n-1)s_x^2}} \quad s_{b_0} = \sqrt{s^2\left(\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_x^2}\right)}$$

Recall that $s = \sqrt{\sum e_i^2/(n-2)}$ is the estimator for $\sigma = \sigma_\varepsilon$.
Hence, $s_{b_1} = \hat{\sigma}_{b_1}$ and $s_{b_0} = \hat{\sigma}_{b_0}$ are estimated coefficient sd's.

A high level of info/precision/accuracy means small s_b values.

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: $b_1 \pm 3 \times s_{b_1}$

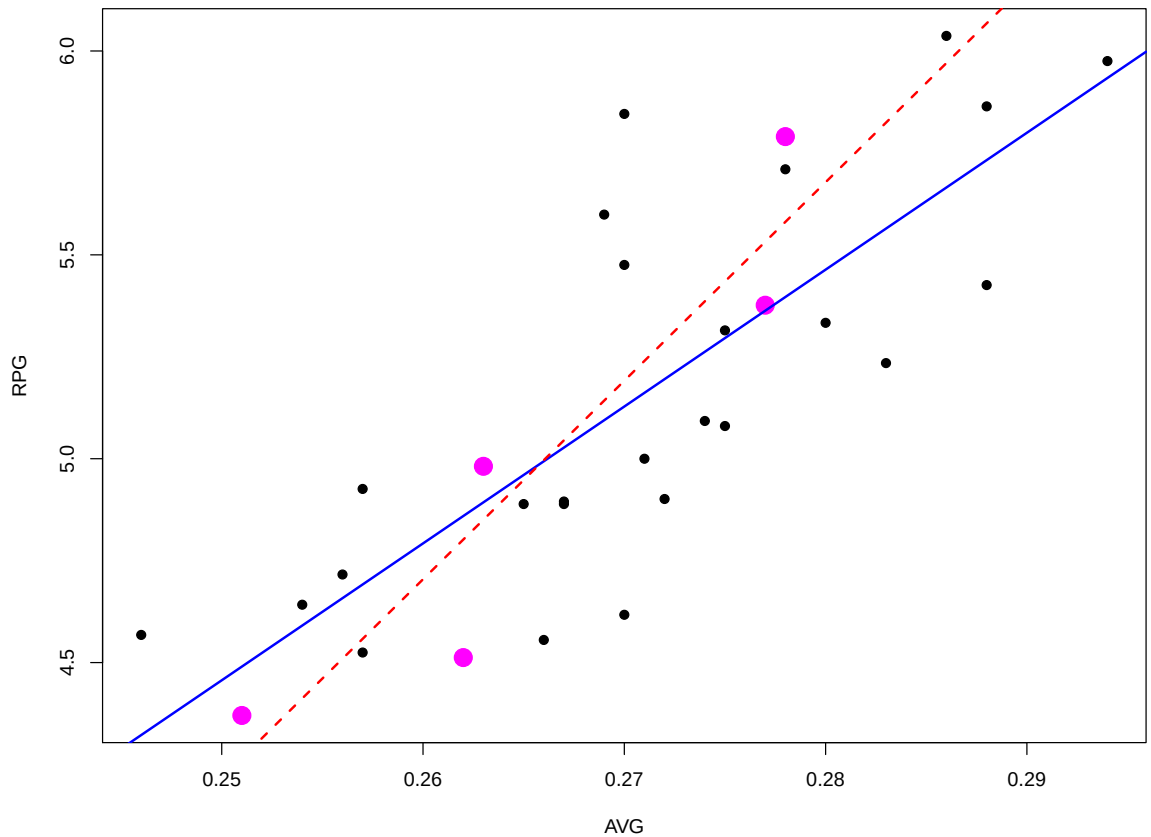
Same thing for b_0

- 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

Example: Runs per Game and AVG

- blue line: all points
- red line: only purple points
- Which slope is closer to the true one? How much closer?



Example: Runs per Game and AVG

Regression with all points

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

$$s_{b_1} = 4.78$$

Example: Runs per Game and AVG

Regression with subsample

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.933601392
R Square	0.87161156
Adjusted R Square	0.828815413
Standard Error	0.244815842
Observations	5

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1.220667405	1.220667	20.36659	0.0203329
Residual	3	0.17980439	0.059935		
Total	4	1.400471795			

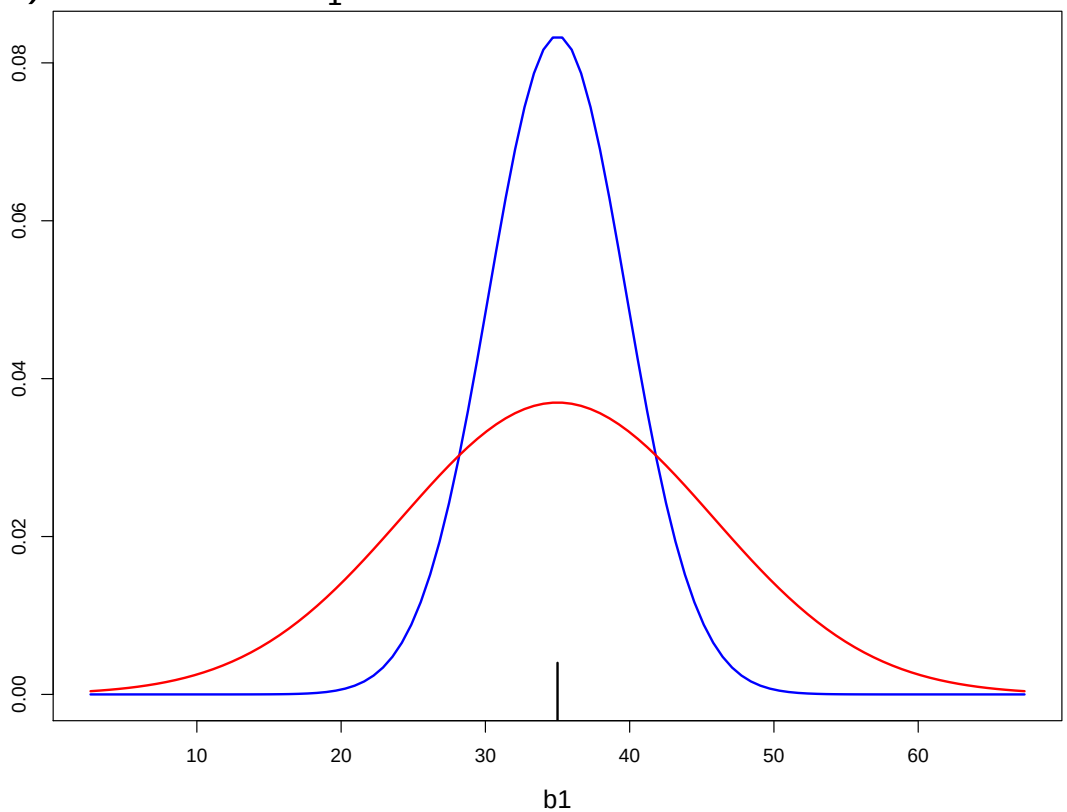
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-7.956288201	2.874375987	-2.76801	0.069684	-17.10384	1.191259
AVG	48.69444328	10.78997028	4.512936	0.020333	14.355942	83.03294

$$s_{b_1} = 10.78$$

Example: Runs per Game and AVG

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

- Suppose $\beta_1 = 35$
- blue line: $N(35, 4.78^2)$; red line: $N(35, 10.78^2)$
- $(b_1 - \beta_1) \approx \pm 2 \times s_{b_1}$



Example: Runs per Game and AVG

Regression with all points

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

$$[b_1 - 2 \times s_{b_1}; b_1 + 2 \times s_{b_1}] \approx [23.77; 43.36]$$

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 errors the estimate (b_1) is 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 null hypothesis.

Testing

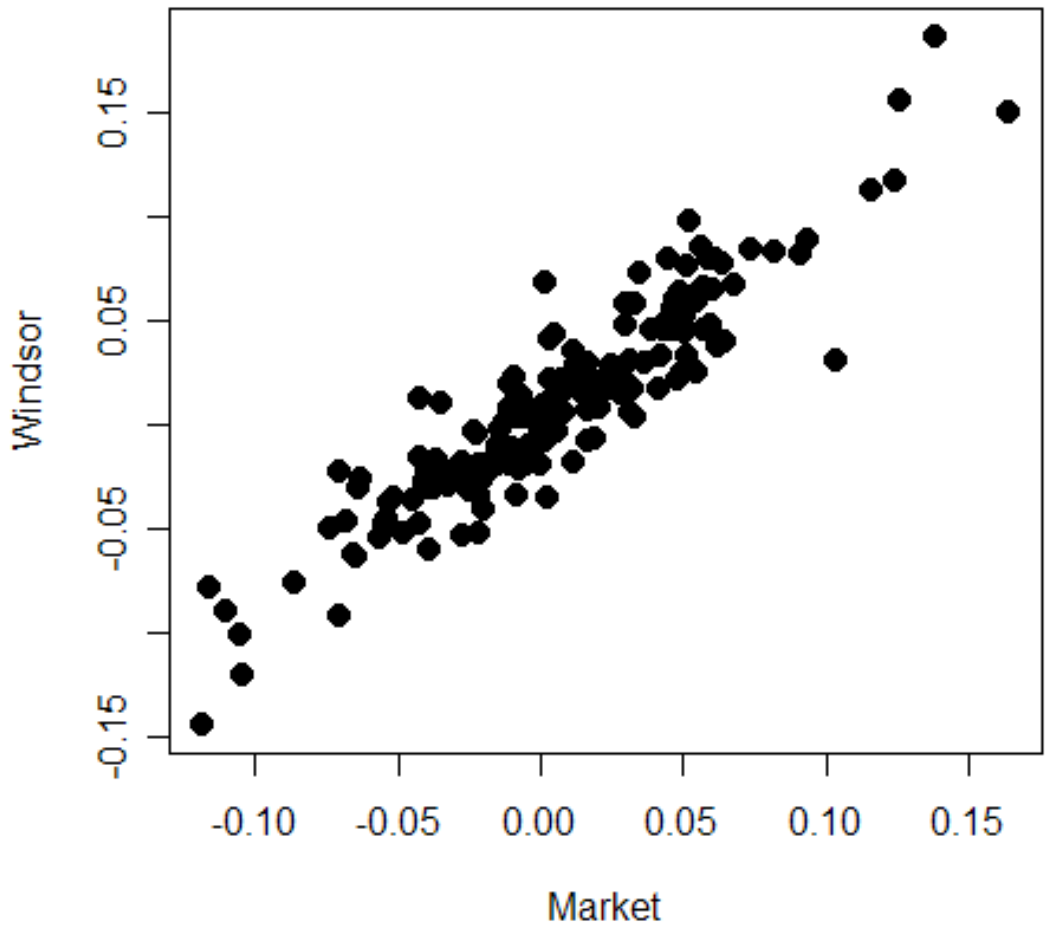
2. Looking at the confidence interval. If the proposed value is outside the confidence interval you reject the null hypothesis.

   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%*	*Upper 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 $\quad$ s_{b_1} $\quad$ $\dfrac{b_1}{s_{b_1}}$

- $t = 32.10...$ 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 Er*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Ipper 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 $\quad s_{b_1}$ $\quad \dfrac{b_1}{s_{b_1}}$

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

Testing – Why I like Conf. Int.

- Suppose in testing $H_0 : \beta_1 = 1$ you got a t-stat of 6 and the confidence interval was

$$[1.00001, 1.00002]$$

Do you reject $H_0 : \beta_1 = 1$? Could you justify that to you boss? Probably not! (why?)

Testing – Why I like Conf. Int.

- Now, suppose in testing $H_0 : \beta_1 = 1$ you got a t-stat of -0.02 and the confidence interval was

$$[-100, 100]$$

Do you accept $H_0 : \beta_1 = 1$? Could you justify that to you boss? Probably not! (why?)

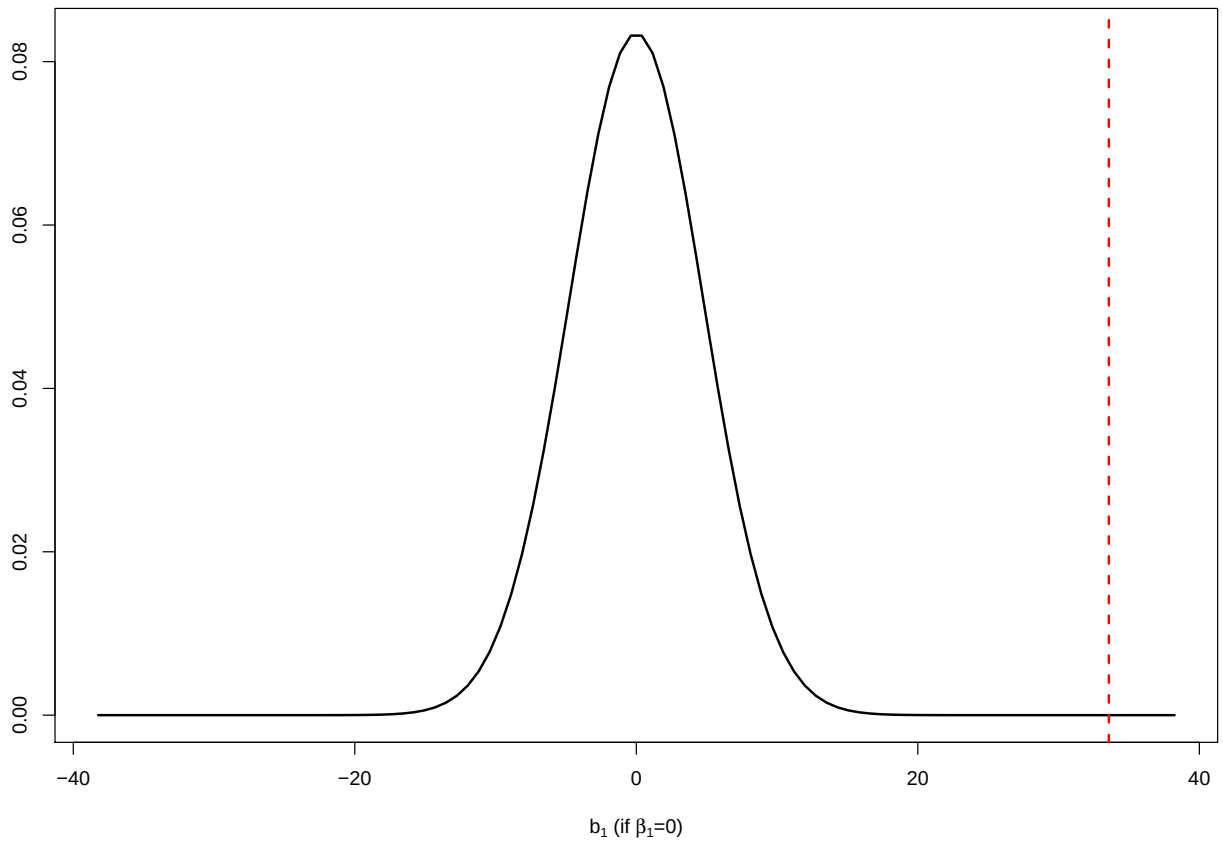
The Confidence Interval is your best friend when it comes to testing!!

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: $b_1 \sim N(\beta_1, s_{b_1}^2)$... If the null was true ($\beta_1 = 0$), $b_1 \sim N(0, s_{b_1}^2)$

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 $= 1.24\text{E-}07$... reject!

SUMMARY OUTPUT

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

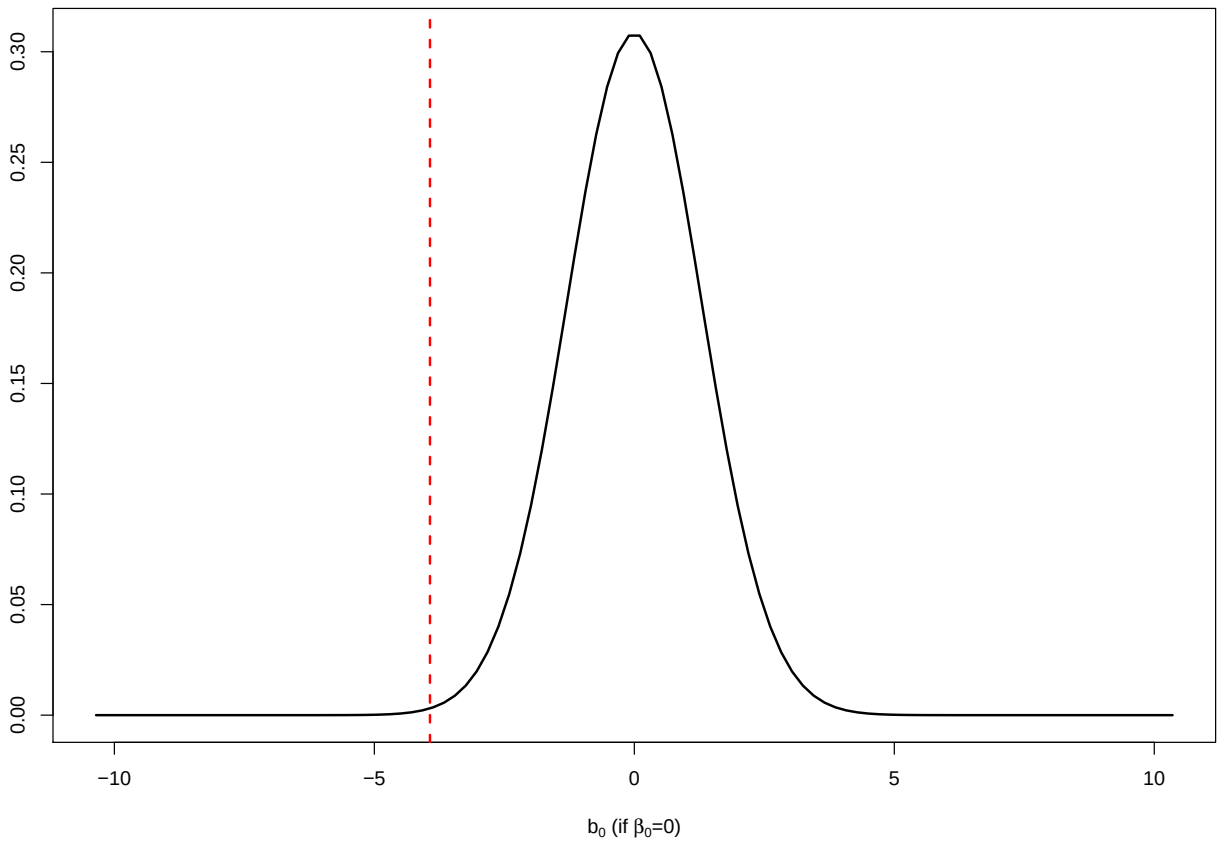
ANOVA

	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_1 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*	*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

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!

Forecasting

The conditional forecasting problem: Given covariate X_f and sample data $\{X_i, Y_i\}_{i=1}^n$, predict the "future" observation y_f.

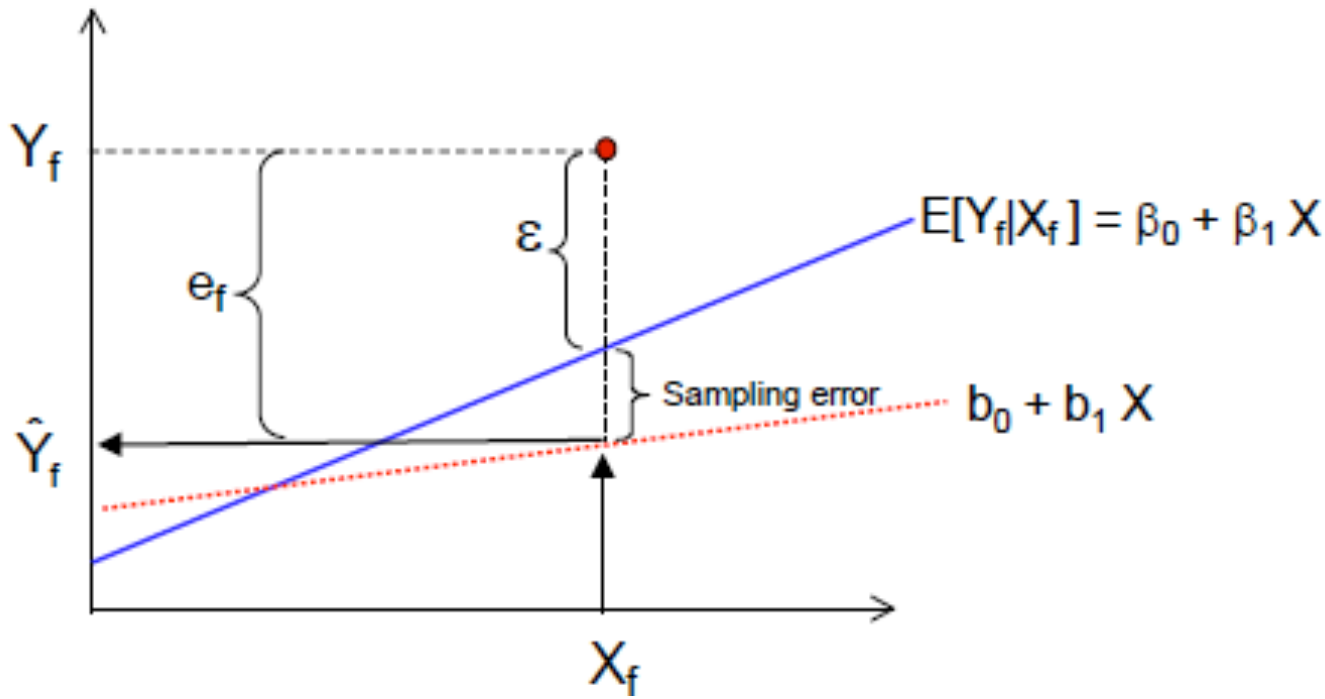
The solution is to use our LS fitted value: $\hat{Y}_f = b_0 + b_1 X_f$.

This is the easy bit. The hard (and very important!) part of forecasting is assessing uncertainty about our predictions.

Forecasting

If we use $\hat{Y}_f$, our prediction error is

$$\begin{aligned} e_f = Y_f - \hat{Y}_f &= Y_f - b_0 - b_1 X_f \\ &= (\beta_0 + \beta_1 X_f + \epsilon) - (b_0 + b_1 X_f) \\ &= (\beta_0 - b_0) + (\beta_1 - b_1) X_f + \epsilon \end{aligned}$$



Forecasting

The most commonly used approach is to assume that $\beta_0 \approx b_0$, $\beta_1 \approx b_1$ and $\sigma \approx s$... in this case, the error is just ϵ hence the 95% plug-in prediction interval is:

$$b_0 + b_1 X_f \pm 2 \times s$$

It's called "plug-in" because we just plug-in the estimates (b_0, b_1 and s) for the unknown parameters (β_0, β_1 and σ).

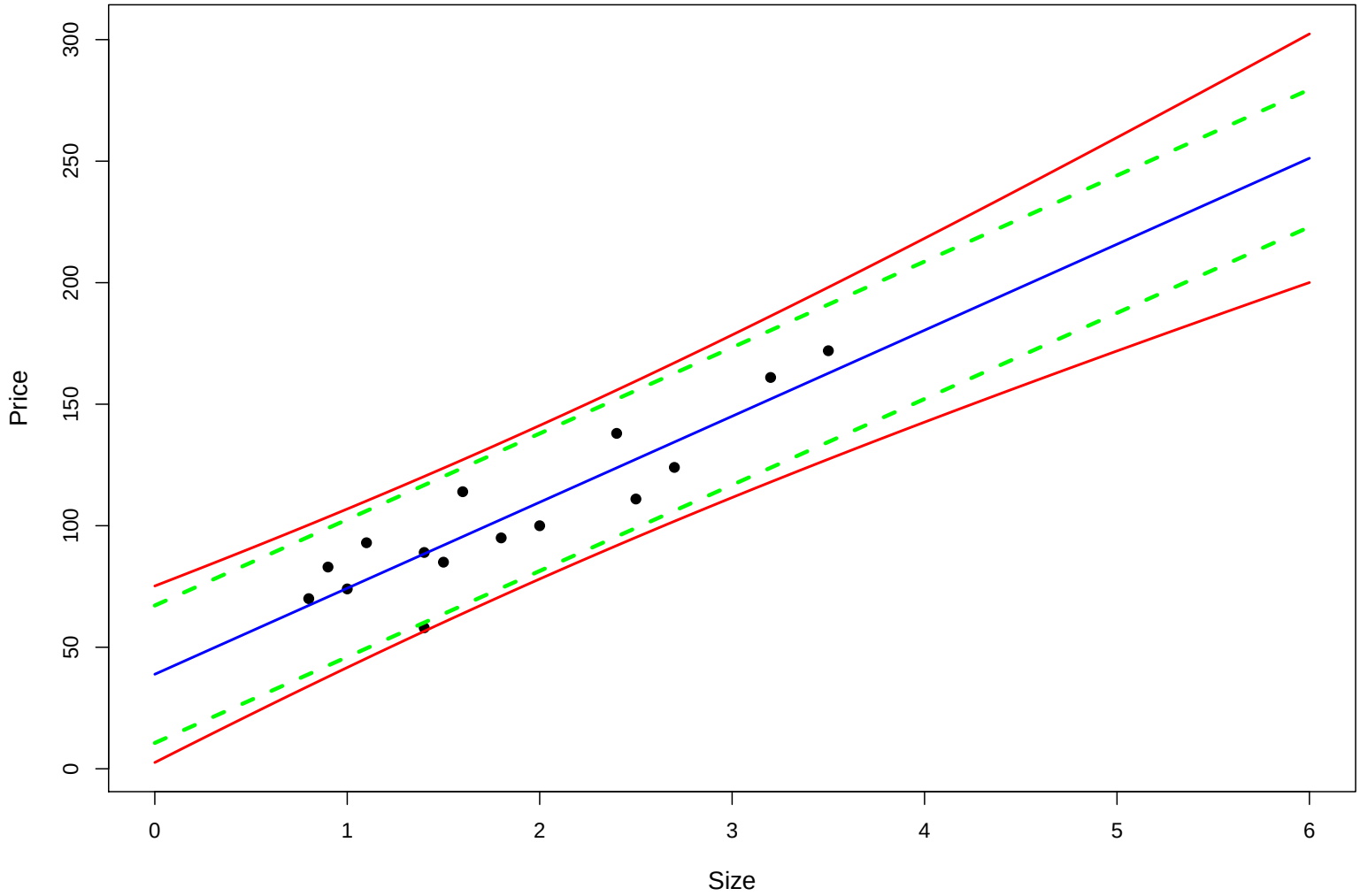
Forecasting

Just remember that you are uncertain about b_0 and b_1! As a practical matter if the confidence intervals are big you should be careful!! Some statistical software will give you a larger (and correct) predictive interval.

A large predictive error variance (high uncertainty) comes from

- Large s (i.e., large ε's).
- Small n (not enough data).
- Small s_x (not enough observed spread in covariates).
- Large difference between X_f and $\bar{X}$.

Forecasting



- Red lines: prediction intervals
- Green lines: "plug-in" prediction intervals

The Importance of Considering and Reporting Uncertainty

In 1997 the Red River flooded Grand Forks, ND overtopping its levees with a 54-feet crest. 75% of the homes in the city were damaged or destroyed!

It was predicted that the rain and the spring melt would lead to a 49-feet crest of the river. The levees were 51-feet high.

The Water Services of North Dakota had explicitly avoided communicating the uncertainty in their forecasts as they were afraid the public would loose confidence in their abilities to predict such events.

The Importance of Considering and Reporting Uncertainty

It turns out the prediction interval for the flood was $49ft \pm 9ft$ leading to a 35% probability of the levees being overtopped!!

Should we take the point prediction ($49ft$) or the interval as an input for a decision problem?

In general, the distribution of potential outcomes are very relevant to help us make a decision

The Importance of Considering and Reporting Uncertainty

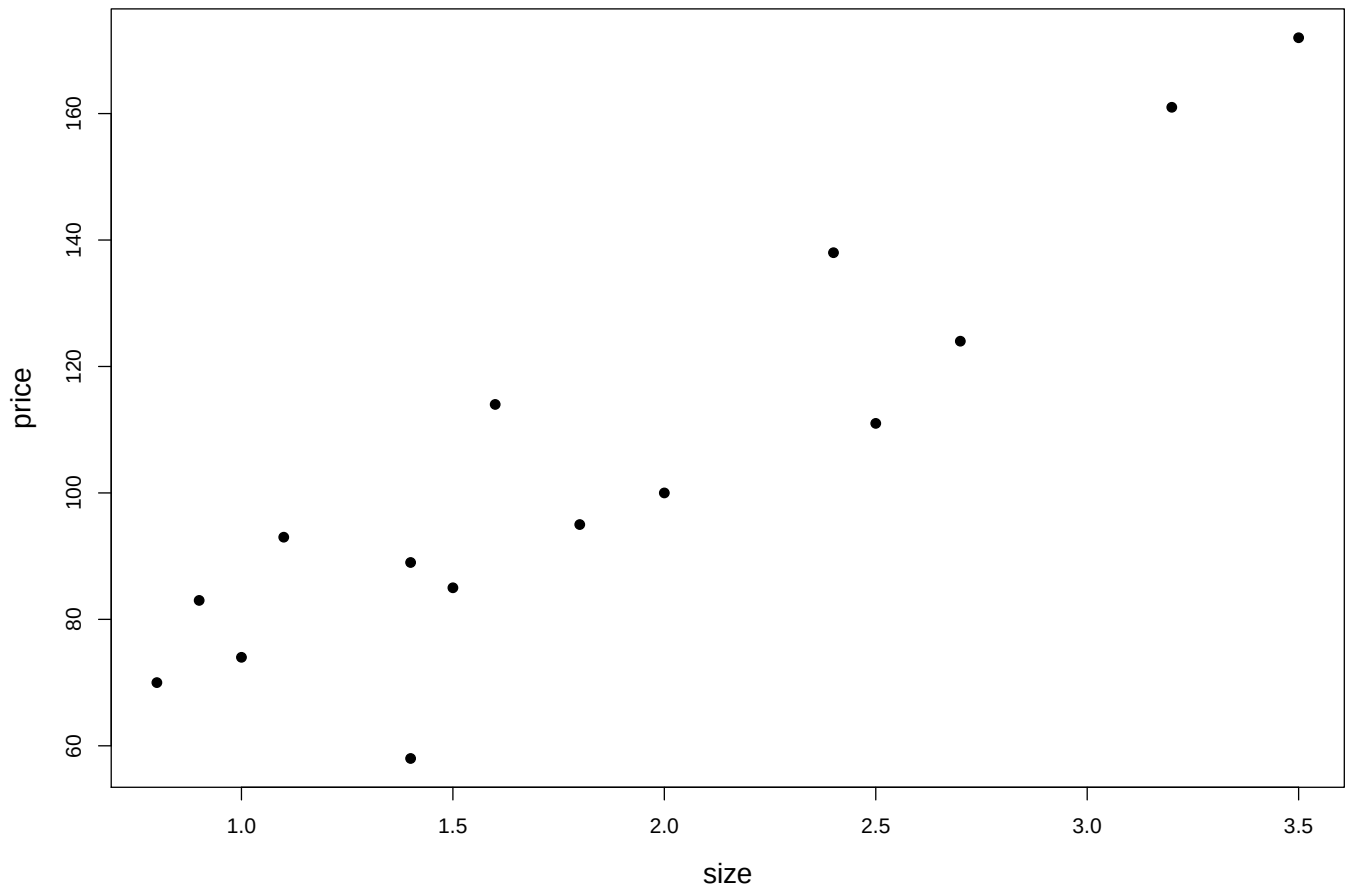
The answer seems obvious in this example (and it is!)... however, you see these things happening all the time as people tend to underplay uncertainty in many situations!

"Why do people not give intervals? Because they are embarrassed!"
Jan Hatzius, Goldman Sachs economists talking about economic forecasts...

Don't make this mistake! Intervals are your friend and will lead to better decisions!

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!