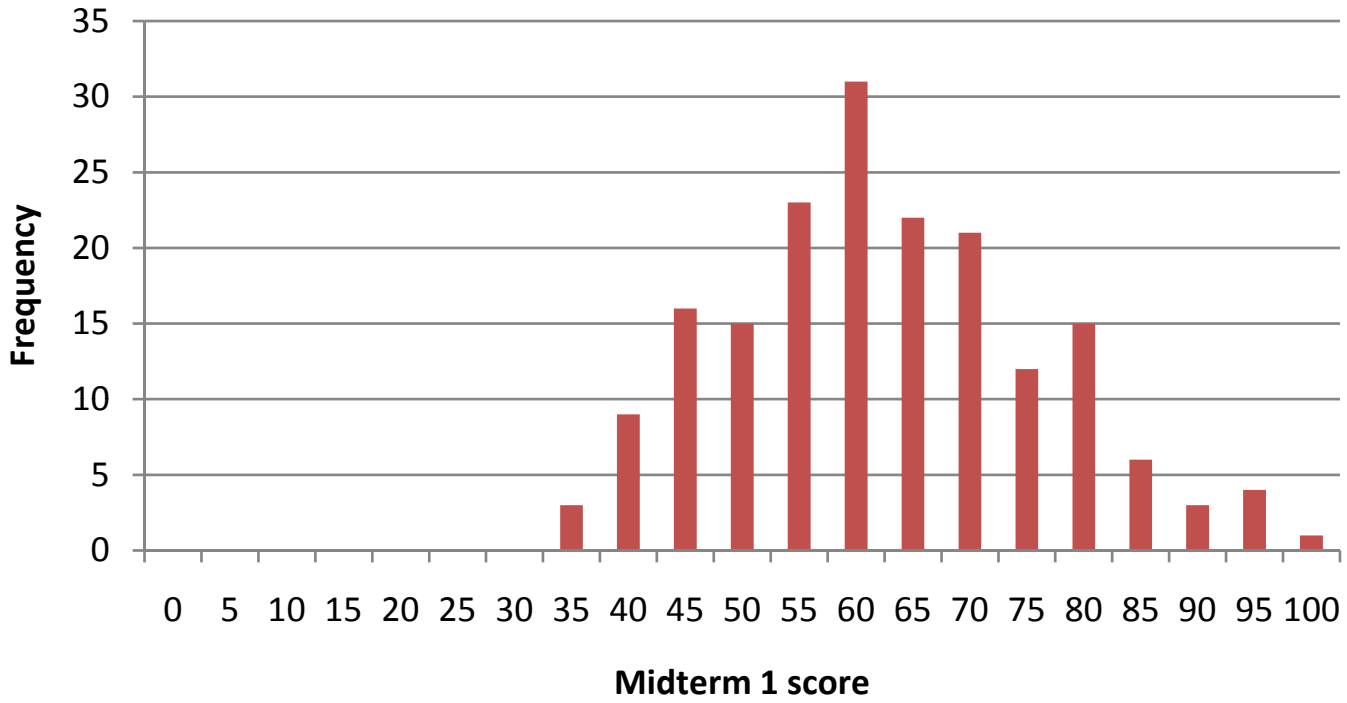
Announcements

- Grades for the first midterm are posted, solutions to the midterm are on Smartsite
- The mean was a 60.6, the median was a 60
- A rough guide to letter grades is on Smartsite (the actual curve will be set at the end of the quarter)
- Don't forget to work on Problem Set 3

Midterm 1 Grade Distribution



Reviewing the Regression Line

$$\hat{y}_i = b_1 + b_2 x_i$$

- $\hat{y}_i$: predicted value for Y for individual i
- x_i: observed value of X for individual i
- b_1: intercept (predicted value of Y when X equals 0)
- b_2: slope (predicted ΔY for a one unit increase in X)

Reviewing the Regression Line

- Recall that the residual is:

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

- We wanted to choose b_1 and b_2 to minimize the average of the squared residuals:

$$\min_{b_1, b_2} \sum (y_i - \hat{y}_i)^2$$

- Replacing $\hat{y}$ with the equation for the regression line makes this:

$$\min_{b_1, b_2} \sum (y_i - b_1 - b_2 x_i)^2$$

Reviewing the Regression Line

- If you work through the math, you come up with the following two equations giving b_1 and b_2:

$$b_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$b_1 = \bar{y} - b_2\bar{x}$$

- Notice that the first equation looks very similar to our variance and covariance formulas, we can rewrite b_2 as:

$$b_2 = \frac{s_{xy}}{s_{xx}} = r_{xy}\sqrt{\frac{s_{yy}}{s_{xx}}}$$

Calculating the Regression Line

- To calculate b_2 and b_1 yourself:
  1. Calculate the covariance of X and Y using the covariance function in Excel
  2. Calculate the variance of X using the variance function in Excel
  3. Calculate b_2 by dividing the covariance of X and Y by the variance of X
  4. Calculate b_1 by subtracting $\bar{x}$ times the b_2 you just found from $\bar{y}$ ($\bar{x}$ and $\bar{y}$ can be calculated with the average function in Excel)
- To have Excel calculate b_2 and b_1, use 'Regression' from the 'Data Analysis' choices

Assessing How Good the Fit Is

- We found the best fit for the regression line (according to our definition)
- This doesn't mean that we have a perfect fit; many data points will not be on the line
- We would like to know just how good the fit is, how well does the line fit the data?
- To answer this, we can use either the **standard error of the regression** or the **R-squared**

The Standard Error of the Regression

- Think back to the residuals: $y_i - \hat{y}_i$
- One way to check how good the fit is is to see how big the residuals are on average
- This is what the standard error of the regression does:

$$s_e^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- The smaller the standard error of the regression is, the closer the fitted values are to the actual data for y

The R-Squared

- The standard error of the regression depends on the units that Y is measured in
- The R^2 provides a standardized measure of how good the fit is
- The idea behind the R^2 is to determine how much of the observed variation in y can be explained by the regression on x
- To do this, we need to measure the total variation in y and the amount of the variation that isn't explained by the regression
- These two measures are the **total sum of squares** and the **error (or residual) sum of squares**, respectively

The R-Squared

- The total sum of squares:

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- The error sum of squares:

$$ESS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- The R-squared:

$$R^2 = 1 - \frac{ESS}{TSS}$$

The R-Squared

- The R^2 will always be between 0 and 1
- An R^2 of 1 means a perfect fit, x perfectly predicts y
- An R^2 of 0 means no fit, variation in x can't explain any of the variation in y
- One interpretation of the R^2 value is that it is the percentage of the variation in y explained by variation in x
- With a little algebra, you can show that R^2 is the square of r_{xy}
- The higher the correlation of two variables, the greater the R^2 will be

Regressing Weight on Height

Regression Statistics	
Multiple R	0.532681203
R Square	0.283749264
Adjusted R Square	0.282871505
Standard Error	29.49983204
Observations	818

SUMMARY OUTPUT: Weight as dependent variable

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	281318.8979	281318.8979	323.2658446	3.84342E-61
Residual	816	710115.9139	870.2400905		
Total	817	991434.8117			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-165.605738	18.65570156	-8.87695044	4.30095E-18	-202.224555	-128.986921
height	4.968722683	0.276353423	17.97959523	3.84342E-61	4.426275353	5.511170013

Assessing the R-squared

- In general, we'd like R^2 to be large but a low R^2 doesn't necessarily mean we have nothing of interest
- R^2 will tend to be high when:
  - Looking at certain time series data in economics
  - Looking at data from controlled experiments (especially in the physical sciences)
  - When the outcome is only dependent on a handful of observable variables
- R^2 will tend to be low when:
  - Looking at certain cross-sectional data in economics (especially wages, employment outcomes, productivity, etc.)
  - Looking at data where there are important but unobservable variables
  - Looking at poorly measured data

An Example of a Low R-Squared

Regression Statistics	
Multiple R	0.402129
R Square	0.161708
Adjusted R Square	0.128176
Standard Error	85.63869
Observations	27

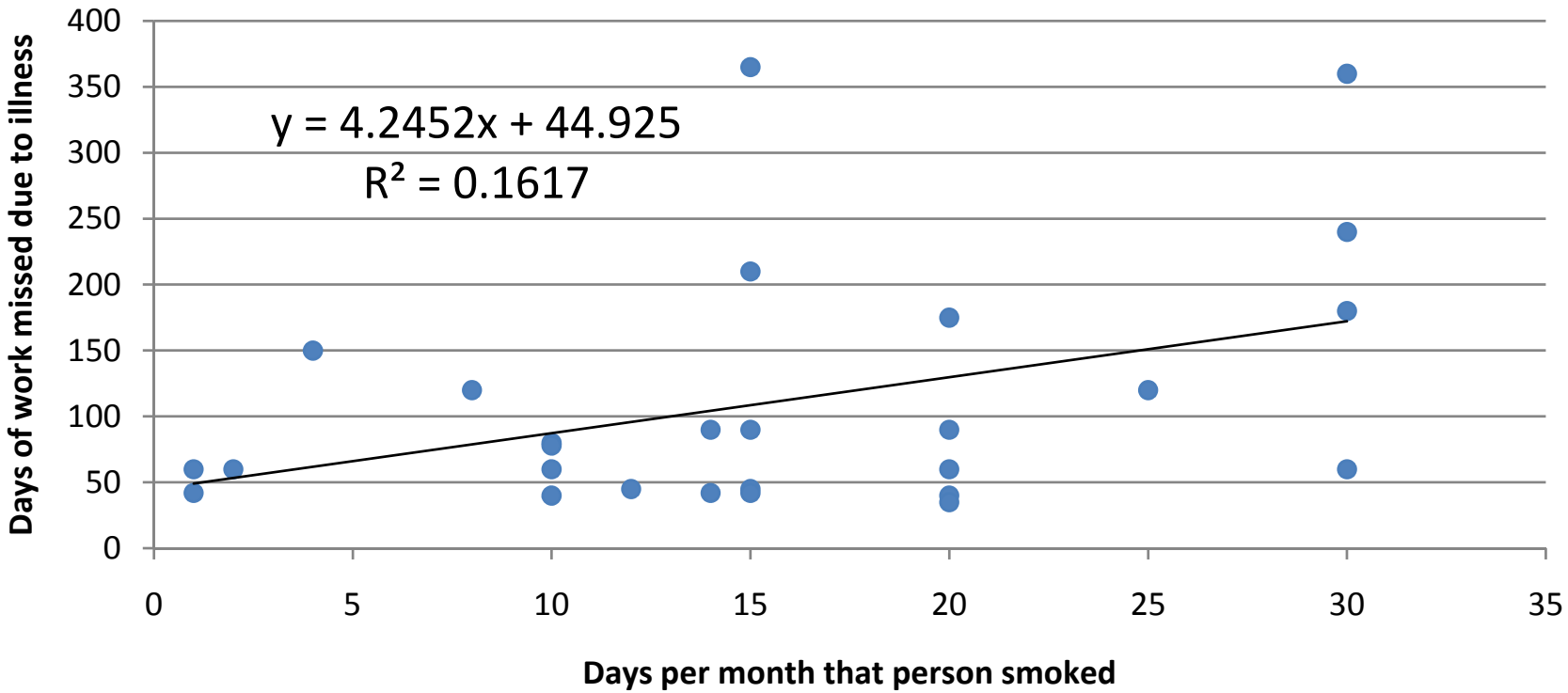
Summary Output: Lost works days in past year

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	1	35368.39	35368.39	4.822534	0.037585
Residual	25	183349.6	7333.985		
Total	26	218718			

	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Upper 95.0%*
Intercept	44.92542	34.04049	1.319764	0.198872	-25.18229	115.0331	-25.18229	115.0331
Days smoked per month	4.245225	1.933139	2.196027	0.037585	0.263851	8.2266	0.263851	8.2266

An Example of a Low R-Squared



An Example of a High R-Squared

Regression Statistics	
Multiple R	0.972705
R Square	0.946155
Adjusted R	0.946006
Standard E	2.042726
Observatio	363

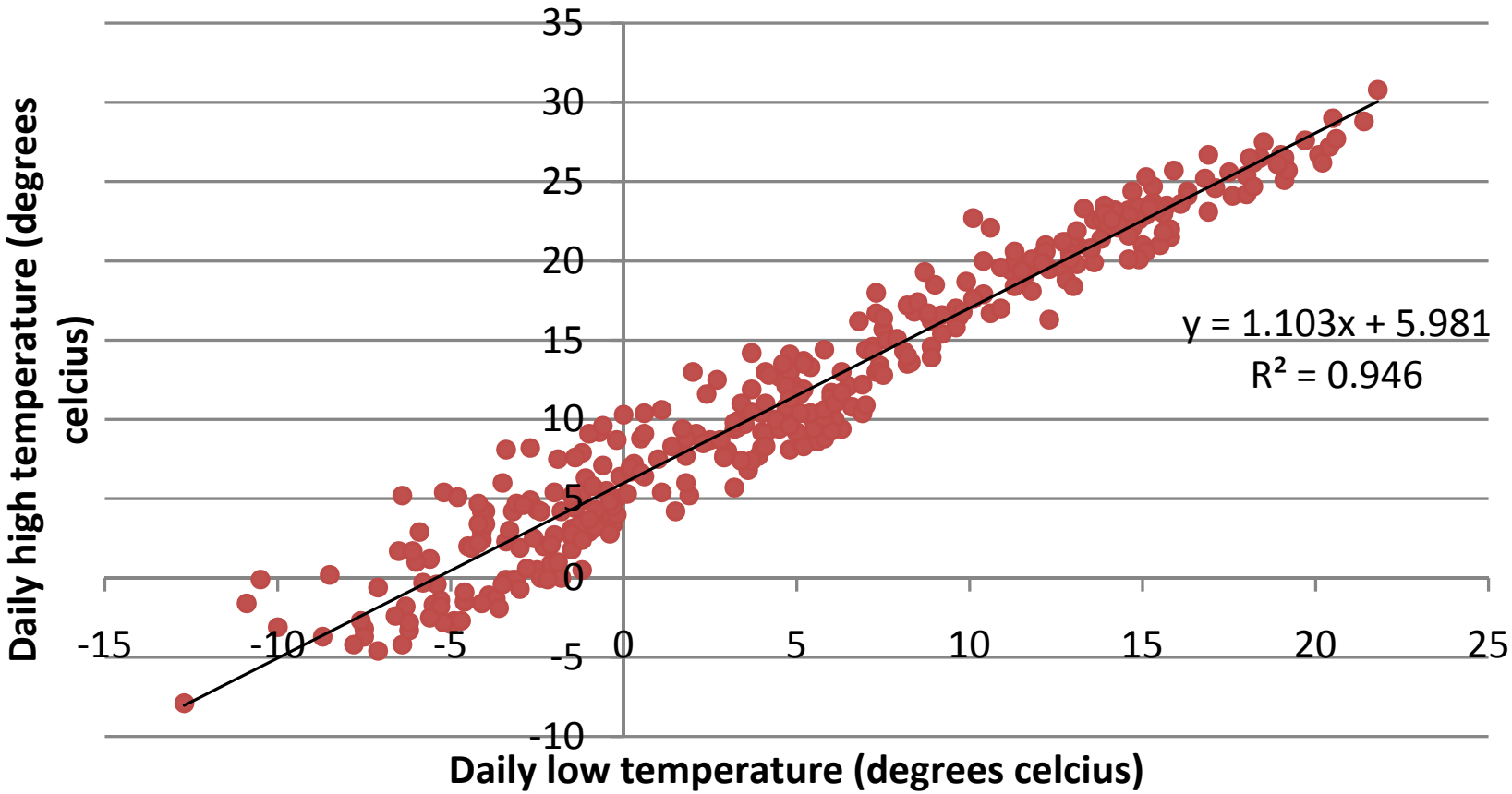
Summary output: Daily high temperature

ANOVA

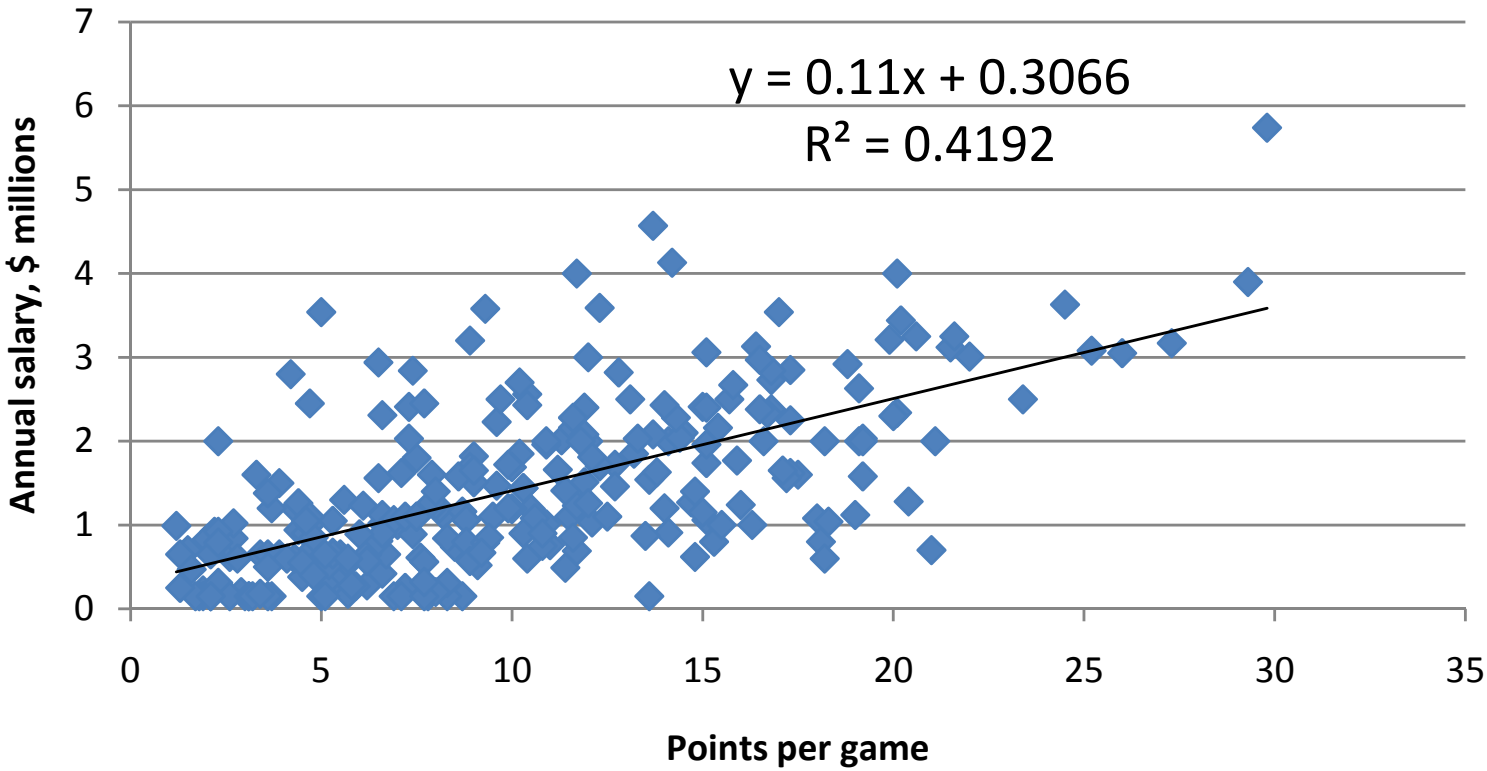
	df	*SS*	*MS*	*F*	*ignificance F*
Regression	1	26469.32	26469.32	6343.409	4E-231
Residual	361	1506.355	4.172728		
Total	362	27975.67			

	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Jpper 95.0%*
Intercept	5.981077	0.129348	46.24032	9.7E-154	5.726707	6.235446	5.726707	6.235446
Low tempe	1.103883	0.01386	79.64552	4E-231	1.076627	1.13114	1.076627	1.13114

An Example of a High R-Squared



Recapping the Regression Line



Recapping the Regression Line

SUMMARY OUTPUT: ln(salary) regressed on points per game

Regression Statistics	
R Square	0.373151498
Observations	272

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	78.69467035	78.69467	160.72608
Residual	270	132.1973423	0.48962	
Total	271	210.8920127		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-0.885888855	0.08479455	-10.44747	1.114E-21
points	0.091561535	0.007222206	12.67778	3.268E-29

From Regression to Statistical Inference

- Our coefficients and R^2 values tell us a lot about what is going on in our sample
- But to make inferences about the population, we need to do a little more work
- Just like we used the sample mean and the sample standard deviation to make inferences about the population, we'll the estimated coefficients and their standard errors to make inferences about the relationship between X and Y for the population

From Regression to Statistical Inference

- We're really interested in the relationship between X and Y at the population level
- We will assume that this relationship is linear:

$$Y = \beta_1 + \beta 2X + \varepsilon$$

- We call $\beta_1 + \beta 2X$ the **population line**
- ε is the **error term** (similar to the residual but a population concept)

From Regression to Statistical Inference

$$Y = \beta_1 + \beta_2 X + \varepsilon$$

- We want to figure out what β_1 and β_2 are based on our estimates of b_1 and b_2
- We can use statistical inference similar to what we used for the population mean (trying to infer the value of μ based on our observed $\bar{x}$)
- First we need to make a few assumptions about the relationship between X and Y and in particular about the distribution of ε

Population Assumptions

We are going to make the following set of **population assumptions**:

1. The population model is $y = \beta_1 + \beta_2 X + \varepsilon$
2. The error ε has mean zero and is unrelated with the regressor x
3. The errors for different observations have constant variance, σ_ε^2
4. The errors for different observations are unrelated
5. The errors are normally distributed: $\varepsilon \sim N(0, \sigma_\varepsilon^2)$

Population Assumptions

- Assumptions 2 through 5 imply that the errors are independently and identically normally distributed ($\varepsilon \sim N(0, \sigma_\varepsilon^2)$)
- This plus the first assumption tell us that observations of y will be independently and identically distributed:

$$y \sim N(\beta_1 + \beta_2 x, \sigma_\varepsilon^2)$$

- Think back to univariate statistical inference, we had $\bar{x} \sim N(\mu, \frac{\sigma_x^2}{n})$ and wanted to figure out μ
- Now we have $y \sim N(\beta_1 + \beta_2 x, \sigma_\varepsilon^2)$ and want to figure out β_1 and β_2

Properties of the Regression Coefficients

- Recall our formulas for b_1 and b_2:

$$b_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$b_1 = \bar{y} - b_2\bar{x}$$

- b_1 and b_2 are functions of our observations of x_i and y_i
- This means that b_1 and b_2 are random variables
- With a bunch of algebra and our population assumptions, we can derive the distributions of these two random variables

The Distribution of the Slope Coefficient

- First, the expected value of b_2 is β_2

$$E(b_2) = \beta_2$$

- This means that b_2 is an unbiased estimator of β_2
- This is a good thing, it says that on average our slope will be equal to the true β_2 for the population relationship

The Distribution of the Slope Coefficient

- Second, the standard deviation of b_2, also called the **standard error of** b_2 is:

$$s_{b_2} = \sqrt{\frac{s_e^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

- The s_e^2 in this is equation is an estimate of σ_ε^2 and is calculated as:

$$s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- The standard error will get very small as n gets very large, meaning that b_2 is a consistent estimator of β_2

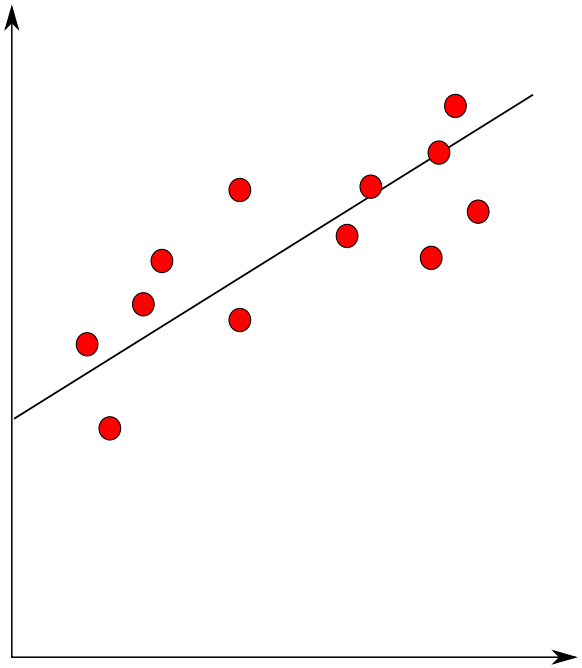
The Distribution of the Slope Coefficient

- Third, the distribution of b_2 is given by the following test statistic:
$$T = \frac{b_2 - \beta_2}{s_{b_2}}$$
- This test statistic is t distributed with $(n-2)$ degrees of freedom
- We will use this test statistic to do our statistical inference

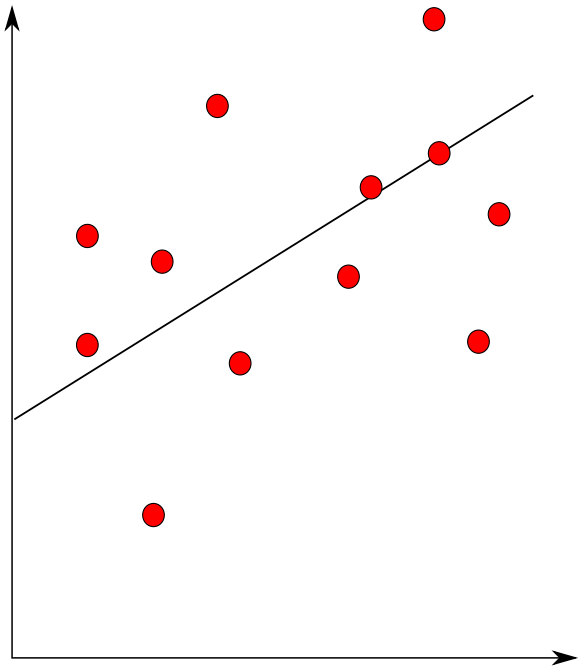
The Distribution of the Slope Coefficient

- Now we can see what sorts of things will affect the precision of our estimate of the slope coefficient
- Anything that makes the standard error of b_2 smaller will make our estimate more precise
- The standard error will be smaller if:
  - The data are closer to the regression line
  - The sample size is large
  - The spread in the x_i values is larger

The Precision of b_2

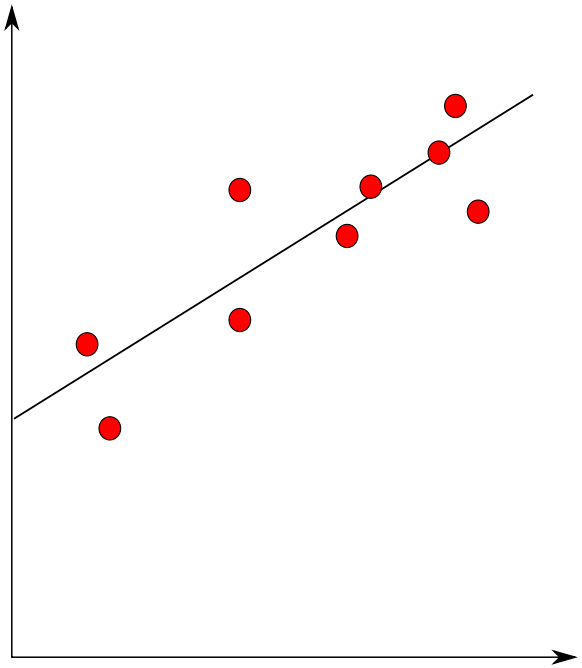


smaller standard error

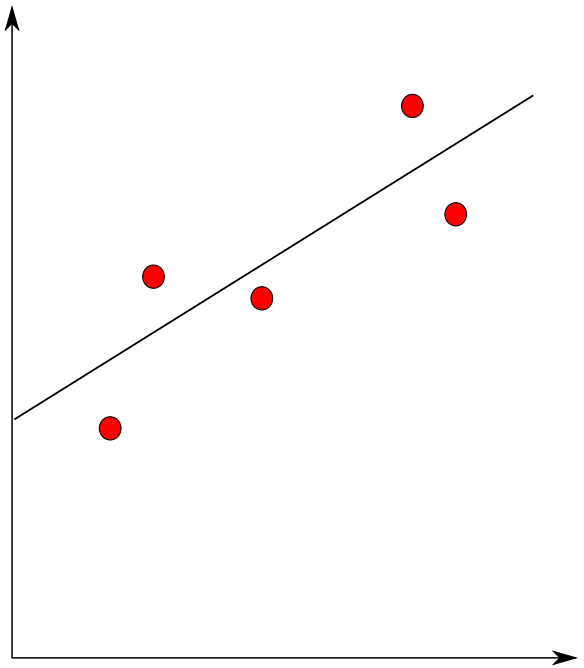


larger standard error

The Precision of b_2

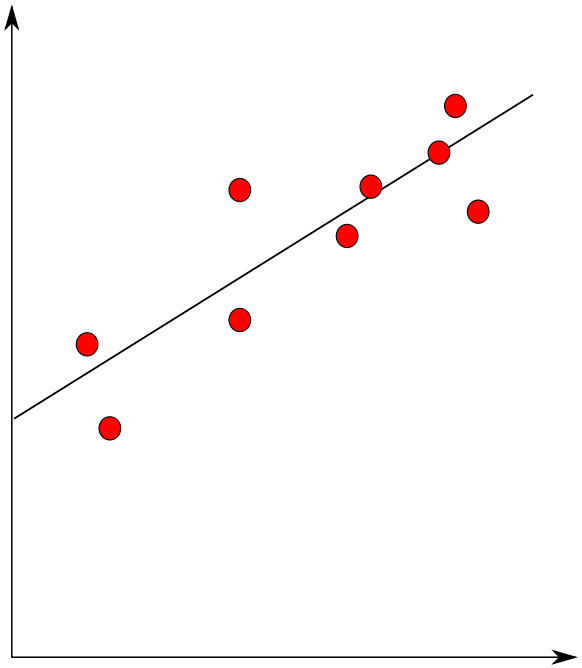


smaller standard error

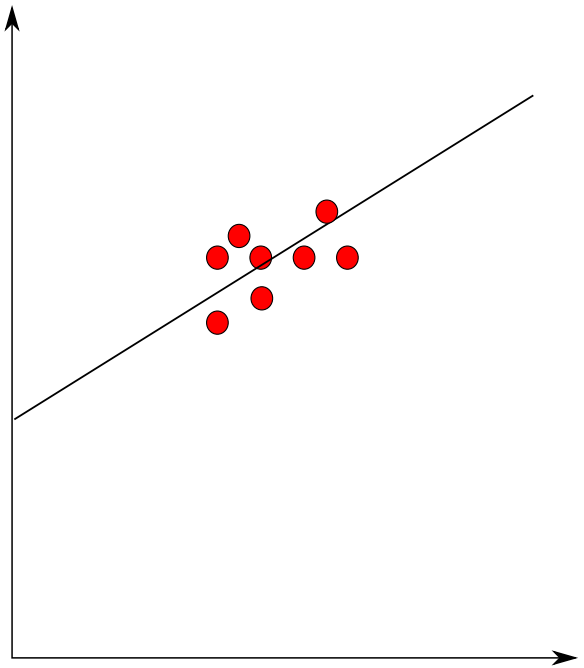


larger standard error

The Precision of b_2



smaller standard error



larger standard error

The Distribution of the Intercept

Given our population assumptions:

- $E(b_1) = \beta_1$ (so b_1 is an unbiased estimator)
- The standard error of b_1 is:

$$\frac{s_e \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

- The test statistic $T = \frac{b_1 - \beta_1}{s_{b_1}}$ is t distributed with $(n-2)$ degrees of freedom