Lecture 22

Linear Regression and Correlation

Today's Updates / Reminders

- PRWA (Article) grades are published. Two scores… one out of 80 and one out of 20.
  - I have set up a regrade request form. In most cases, peers are a mix of generous and harsh. You likely earned credit for things you missed, didn't earn credit for things you included. In most cases, this balances out. The regrade will throw out all peer reviews and my grade will be the sole determination. You score can increase, but it can also decrease. The score change only applies to the submission part of the grade (the grade you earned for your grading is unchanged… the part out of 20). Regrade requests must be received by Friday, 4/24.
- HW 9 is due the end of the day, TODAY.
- Lab 9 is open and due the end of the day, last day of class.
- Homework 10 is open and due the end of the day, last day of class.
- I will hold an optional exam review on Sunday, May 3, SAS1102, 2-3pm. Final exam review and key have been published.
- ClassEval open. Since this is just my second year at NC State, please help me get better by filling one out.

Inference for Regression

Chapter 12

Example: Osteoporosis is a condition where bones become weak. Exercise is one way to produce strong bones and prevent osteoporosis. Since we use our dominant arm more than our non-dominant arm, we expect, when measuring bone strength, to find stronger bones in the dominant arm. The table below compares the bone strength in the arms (in $cm^4/1000$) of 15 young men.

ID	Nondominant	Dominant	ID	Nondominant	Dominant
1	15.7	16.3	9	15.9	20.1
2	25.2	26.9	10	13.7	18.7
3	17.9	18.7	11	17.7	18.7
4	19.1	22.0	12	15.5	15.2
5	12.0	14.8	13	14.4	16.2
6	20.0	19.8	14	14.1	15.0
7	12.3	13.1	15	12.3	12.9
8	14.4	17.5			

Applet for Correlation and Regression

arm_bone_strength.xlsx
Google Sheets

Equation of the Least-Squares Regression Line:

Explanatory variable x $\rightarrow \bar{x}, s_x$

Response variable y $\rightarrow \bar{y}, s_y$

Then, the equation for the least-squares regression line is:

$\hat{y} = b_0 + b_1 x$ $\hat{y}$: predicted value of y

where $b_1 = r\frac{s_y}{s_x}$ and $b_0 = \bar{y} - b_1 \bar{x}$

b_1 = slope

b_0 = intercept

Alternate Notation:

$s_{xx} = \sum(x_i - \bar{x})^2$

$s_{yy} = \sum(y_i - \bar{y})^2$

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

$s_{xx} = 174.497$

$s_{yy} = 186.689$

$s_{xy} = 163.325$

$$b_1 = \frac{s_{xy}}{s_{xx}} = \frac{163.325}{174.497} = 0.936$$

Note: The slope and intercept of the least square line depend on the units of measurement!

From previous example:

$$\bar{x} = 16.01 \;;\; s_x = 3.53$$
$$\bar{y} = 17.73 \;;\; s_y = 3.65$$
$$r = 0.905$$

$$\hat{y} = 2.7386 + 0.936\,x$$

x	y	ŷ
15.7	16.3	17.4338
25.2	26.9	26.3258
17.9	18.7	19.493
19.1	22	20.6162
12	14.8	13.9706
20	19.8	21.4586
12.3	13.1	14.2514
14.4	17.5	16.217
15.9	20.1	17.621
13.7	18.7	15.5618
17.7	18.7	19.3058
15.5	15.2	17.2466
14.4	16.2	16.217
14.1	15	15.9362
12.3	12.9	14.2514

We can calculate the standard deviation of ŷ :

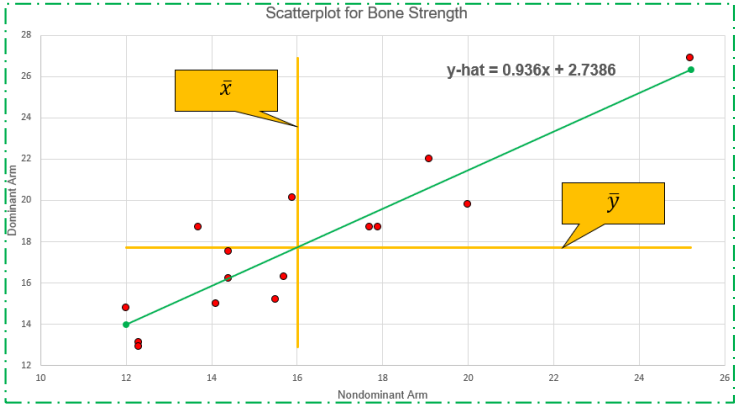
$$s_{\hat{y}} = 3.30$$

$$r^2 = \frac{s_{\hat{y}}^2}{s_y^2} = \frac{(3.30)^2}{(3.65)^2} = \frac{10.89}{13.32} = 0.82$$

$$r = \sqrt{0.82} = 0.905 \rightarrow \text{same as before}$$

82% of the variation is explained by the least-squares regression

Alternate Concept of b_1



Nondomi	Dominant		xi - xbar	(xi-xbar)^2	yi-ybar	(yi-ybar)^2	(xi-xbar)(yi-ybar)
15.7	16.3		-0.313	0.098	-1.427	2.035	0.447
25.2	26.9		9.187	84.395	9.173	84.150	84.272
17.9	18.7		1.887	3.560	0.973	0.947	1.836
19.1	22		3.087	9.528	4.273	18.261	13.190
12	14.8		-4.013	16.107	-2.927	8.565	11.746
20	19.8		3.987	15.894	2.073	4.299	8.266
12.3	13.1		-3.713	13.789	-4.627	21.406	17.180
14.4	17.5		-1.613	2.603	-0.227	0.051	0.366
15.9	20.1		-0.113	0.013	2.373	5.633	-0.269
13.7	18.7		-2.313	5.352	0.973	0.947	-2.252
17.7	18.7		1.687	2.845	0.973	0.947	1.642
15.5	15.2		-0.513	0.264	-2.527	6.384	1.297
14.4	16.2		-1.613	2.603	-1.527	2.331	2.463
14.1	15		-1.913	3.661	-2.727	7.435	5.217
12.3	12.9		-3.713	13.789	-4.827	23.297	17.923
16.01333	17.72667	SUM	0	174.497	0	186.689	163.325

s_{xx} s_{yy} s_{xy}

Alternate Notation:

$s_{xx} = \sum(x_i - \bar{x})^2$

$s_{yy} = \sum(y_i - \bar{y})^2$

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

$s_{xx} = 174.497$

$s_{yy} = 186.689$

$s_{xy} = 163.325$

$$b_1 = \frac{s_{xy}}{s_{xx}} = \frac{163.325}{174.497} = 0.936$$

- s_{xx} represents how far, in the x-direction, each value is from $\bar{x}$. What would happen if we did $x_i - \bar{x}$ and summed that? It would sum to zero!
- s_{yy} represents how far, in the y-direction, each value is from $\bar{y}$.
- s_{xy} represents how much the points tend to fall in the upper right and lower left (positive s_{xy}, positive correlation) or vice versa (negative s_{xy}, negative correlation).

Residuals

Residual: **is the difference between an observed value of the response and the value predicted by the regression line.**

$$\text{residual} = \text{observed } y - \text{predicted } y$$

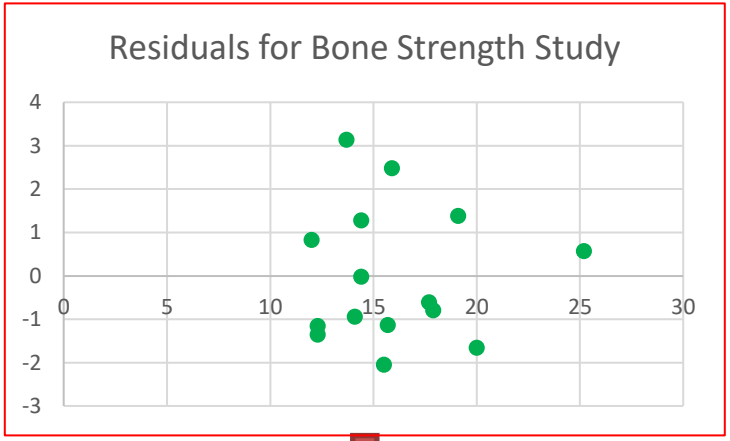
$$e = y - \hat{y}$$

Residual plot: **is a scatterplot of the residuals against the explanatory variable x. They help us assess the fit of a regression line.**

Example: Bone strength study

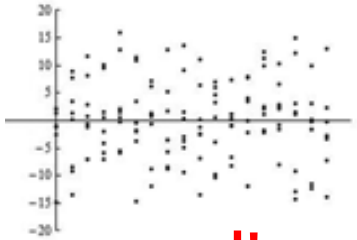
x	y	ŷ	Residual y-ŷ
15.7	16.3	17.4338	-1.1338
25.2	26.9	26.3258	0.5742
17.9	18.7	19.493	-0.793
19.1	22	20.6162	1.3838
12	14.8	13.9706	0.8294
20	19.8	21.4586	-1.6586
12.3	13.1	14.2514	-1.1514
14.4	17.5	16.217	1.283
15.9	20.1	17.621	2.479
13.7	18.7	15.5618	3.1382
17.7	18.7	19.3058	-0.6058
15.5	15.2	17.2466	-2.0466
14.4	16.2	16.217	-0.017
14.1	15	15.9362	-0.9362
12.3	12.9	14.2514	-1.3514

What do you think the sum of the residuals will be?



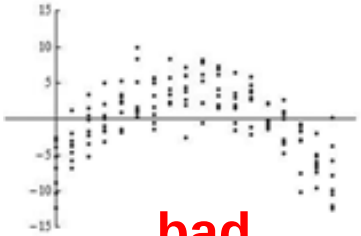
If the regression line catches the overall pattern of the data→ the residual plot should be a uniform horizontal band of points centered at zero, with no visible pattern or functional dependence, i.e., **residual values should be equally and randomly spaced around the horizontal axis.**

Residual Plot A



good!

Residual Plot B

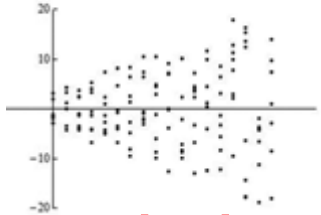


bad

Residual Plot C

good!

Residual Plot D



bad

More on Residuals and Error

- As stated earlier, $e = y - \hat{y}$ and call this a residual, or error.
- We can find sum of squares of error, or SSE using:

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

For the arm bone strength study, SSE = 33.822 and n=15 Calculate and interpret s_e.

On average, the y's are about 1.6 from the regression line.

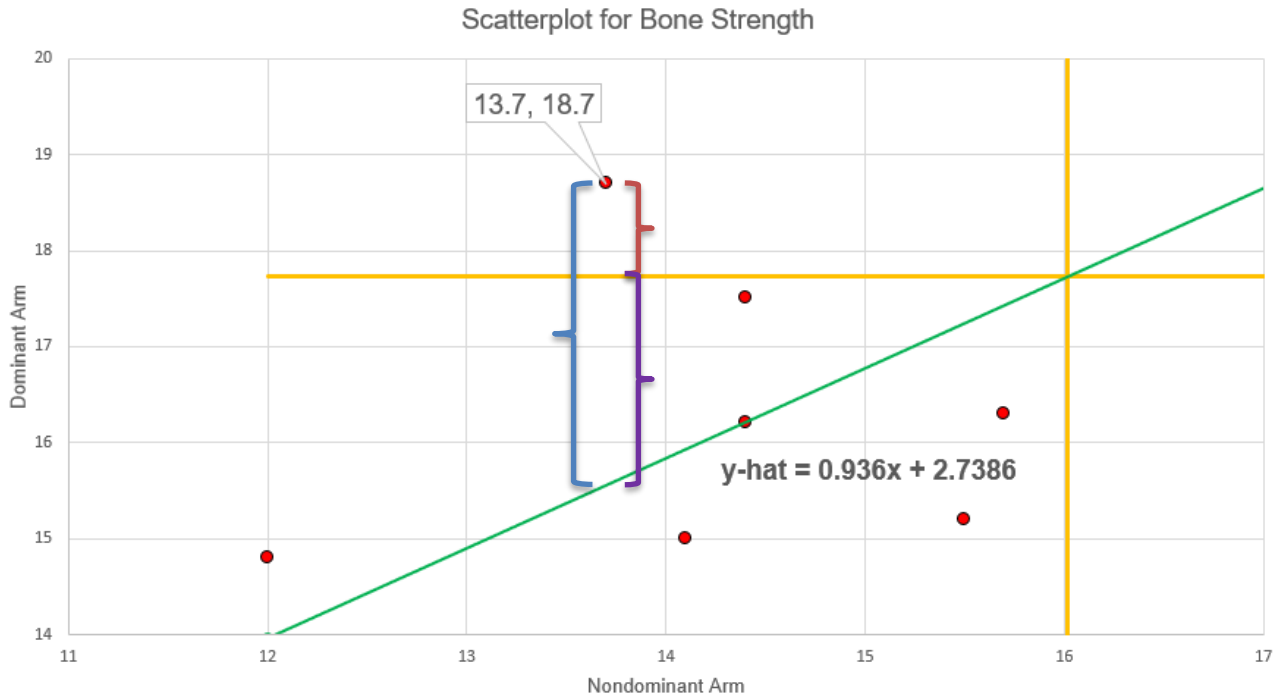
- When we find the "least squares" regression line, **this** is the value that we are minimizing.
- If we find the variance of e, we can refer to this as a variation in error, which we call **mean square error (MSE)**. And s_e is the standard deviation of the error.

$$s_e = \hat{\sigma}_e = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{SSE}{n-2}} = \sqrt{MSE}$$

Applet for Correlation and Regression

- s_e represents the typical amount by which an observation deviates vertically from the regression. It describes, on average, how far off the predicted $\hat{y}$ is to the observed y.

Least-squares regression line for the Bone Strength study:



Rewriting $e_i = y_i - \hat{y}_i$, we can say $y_i = \hat{y}_i + e_i$

If we subtract $\bar{y}$ on both sides, we get:

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + e_i$$

We can sum these squares:

$$\underbrace{\sum (y_i - \bar{y})^2}_{S_{yy}} = \sum (\hat{y}_i - \bar{y})^2 + \underbrace{\sum e_i^2}_{SSE}$$

Now using our formulas from the last slide, we can say:

Total Sum of Squares = Model Sum of Squares + Error Sum of Squares

SST = SSM + SSE

And the definition of R^2 is the fraction of variation in the values that is explained by the model… so $R^2 = \frac{SSM}{SST}$

From slide 6 earlier

$$r^2 = \frac{\text{variance of predicted values } \hat{y}}{\text{variance of observed values } y} = \frac{s_{\hat{y}}^2}{s_y^2}$$

DATA = FIT + RESIDUAL

R^2 = coefficient of determination

- Let's talk more about R^2 and how we understand it.
  - You can find the correlation coefficient, r. Then square that.

$$r = 0.905 \rightarrow\rightarrow r^2 = 0.905^2 = 0.82$$

  - From the last lecture (and first slides from today), you can find it by:

$$R^2 = \frac{s_{\hat{y}}^2}{s_y^2} = \frac{variance\ of\ \hat{y}}{variance\ of\ y} = \frac{10.9191}{13.33495} = 0.82$$

  - And now, SST = SSM + SSE leads us to quantify it as:

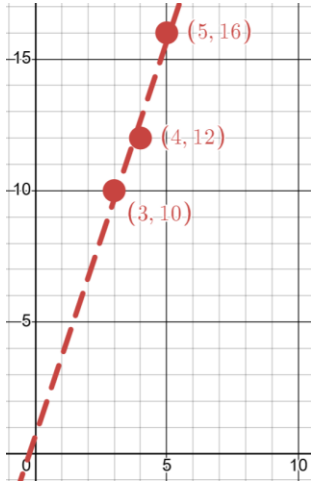
$$R^2 = \frac{SSM}{SST} = \frac{152.8674}{186.6893} = 0.82$$

- The last two describe the definition best! How much TOTAL variation in dominant arm strength is there? And how much variation can be accounted for using our best-fit model, $\hat{y}$?

Example

- You're stranded all alone without any technology (think UNC home football game in the 4th quarter… I mean, nobody there… at all!)
- You have the data set below and you know the line of best fit. But you need to find SSE and SSM so you can find R^2 for this model.
- Line of best fit is $\hat{y} = 3x + 2/3$

Points	$\hat{y}$	e	e^2	$\hat{y} - \bar{y}$	$(\hat{y} - \bar{y})^2$
(3,10)					
(4,12)					
(5,16)					

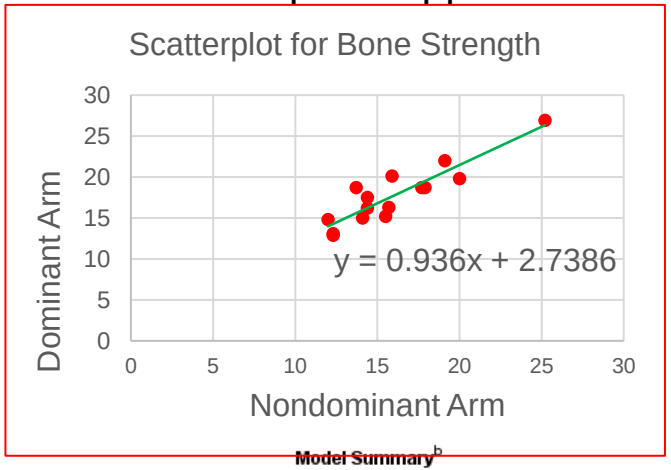


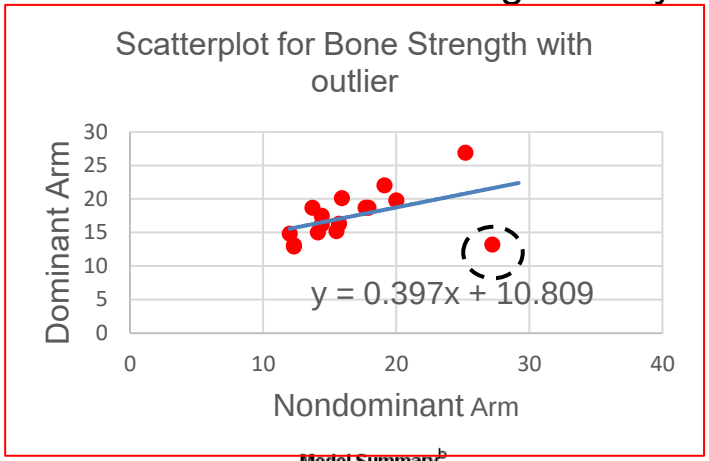
Effects of outliers:

Outlier: **an observation that lies outside the overall pattern of the observations.**

Influential outlier: **The act of removing it changes the result of the calculations considerably.**

Example: Suppose we add an outlier in the bone strength study





Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.905[a]	.819	.805	1.61298

a. Predictors: (Constant), Nondominant

b. Dependent Variable: Dominant

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.473[a]	.223	.168	3.37963

a. Predictors: (Constant), nondomin

b. Dependent Variable: domin

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t
1	(Constant)	2.739	1.999		1.
	Nondominant	.936	.122	.905	7.

a. Dependent Variable: Dominant

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t
1	(Constant)	10.809	3.413		3.
	nondomin	.397	.198	.473	2.

This outlier is influential!

Linear Regression → two types of inference or significance test:

- **Significance test for the population correlation**
- **Significance test for the slope of the regression line in the population**

All the tests that we'll consider in linear regression are going to be two-sided!!

Inference for Correlation

Finding and Testing ρ

Inference for the population correlation

So far we talked about the correlation between x & y in the sample → **r**
What can we infer about the value of the correlation for the population?
The question we want to answer is:
Are dominant arm bone strength, and non-dominant arm bone strength independent in the population?

ρ = correlation of the population. It is the correlation between x and y measured on every member of the population.
When **ρ = 0** → there is no linear association between the variables in the population (i.e. x & y are not correlated in the population) → **x and y are independent in the population.**
The null hypothesis to test is that the population correlation is zero.

Significance test for a zero population correlation is characterized by:

Null hypothesis $H_0 : \rho = 0$ (x and y are independent, i.e. not correlated)

Alternative hypothesis: $H_a : \rho \neq 0$ (x and y are not independent, i.e they are correlated)

Test-statistic for the correlation: $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$;

where n = sample size, and r = sample correlation

It has a t(n-2) distribution (df = n-2).

P-val = 2 P(T >| t |) (2-tail)

If P-val $\leq \alpha$, we reject the null. There is enough evidence to conclude that x & y are correlated (i.e. not independent) in the population.

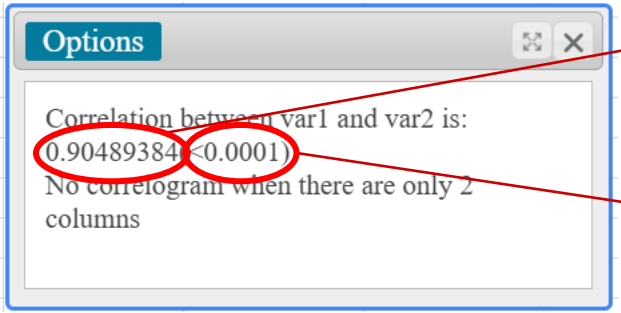
If P-val > α, we fail to reject the null. We don't have enough evidence to say that x & y are correlated (i.e. not independent) in the population.

To perform this test with StatCrunch we need to look at the correlations output!

To get the correlations output from StatCrunch, follow these steps:

1. The two quantitative variables (explanatory and response) are in two separate columns. Both are numeric variables.
2. Stat → Summary Stat → Correlation
3. Hold ctrl and click both variables
4. Under Display, click "Two-Sided P-value", then click Compute.

Let's calculate t.



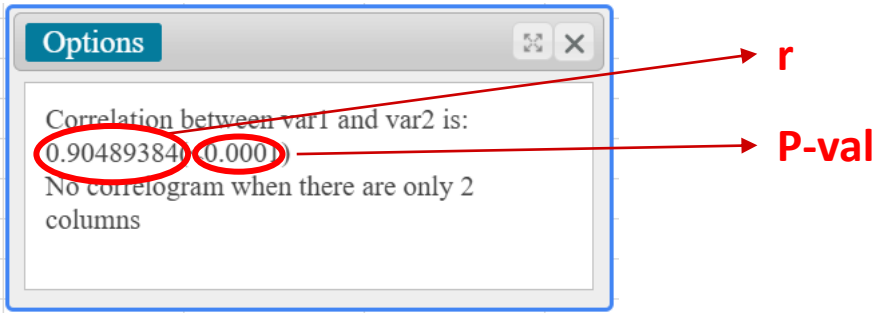
r = estimate for the population correlation

P-val for a 2 sided H_a!!

Add another column of data and try it… sneak peak at multiple regression!

H_0: $\rho = 0$ (X and Y are not correlated, i.e. they are independent)
H_a: $\rho \neq 0$ (X and Y are correlated, i.e. they are not independent)

P-value = 0.000 <0.05→ Reject Ho
At 5% significance level, we have evidence that dominant and non-dominant arm bone strength **are correlated** (**i.e. they are not independent) in the population.**



Inference for Regression

Finding and Testing β

Inference for the Regression Line

Model's assumptions:

1. In linear regression the explanatory variable x is quantitative and can take many values, so we can think of the different values of x as defining different subpopulations. For each subpopulation x_i we consider a SRS consisting of different values of y → the responses corresponding to different values of x are independent.
2. For each value of x, the observed values of the response variable y are assumed to be Normally distributed with a mean that depends on x → $\mu_y(x)$
3. All those means lie on a line when plotted against x:

$$\mu_y = \mu_y(x) = \beta_0 + \beta_1 x$$ → Population's regression line

4. The standard deviations of all the Normal distributions corresponding to the response y are the same and equal to a certain value σ → the standard deviation is assumed to be the same (constant).

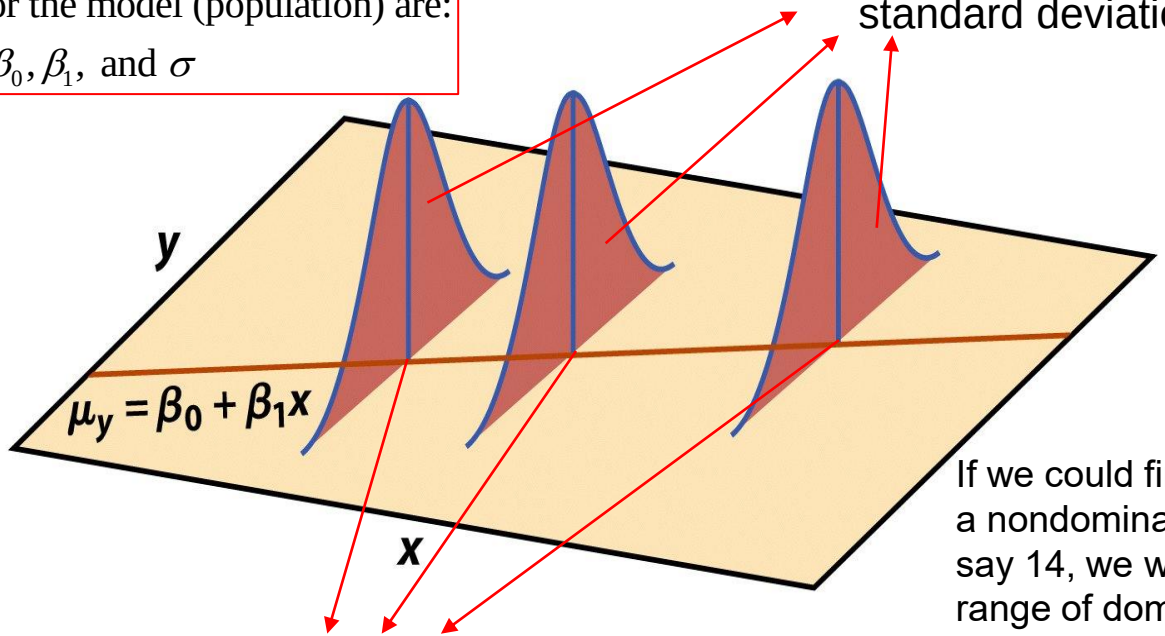
$$\sigma_y = \sigma_y(x) = \sigma$$

To visualize this (from the textbook):

The parameters for the model (population) are:

$\beta_0, \beta_1,$ and σ

All y have Normal distributions with the same spread → same standard deviation σ



The centers of the distributions lie on a line given by the equation $\mu_y(x) = \beta_0 + \beta_1 x$

If we could find 100 people with a nondominant arm strength of, say 14, we would get a wide range of dominant arm strengths. In fact, we should get a normal curve distribution of y's for any given x.

Checking for assumptions: Residuals and Normality Plots

What kind of plot will let us see if the model assumptions are met?

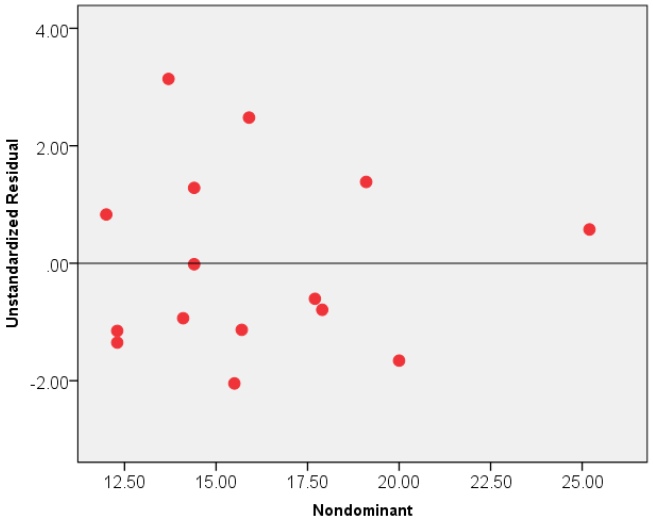
- **Scatterplot (y vs. x)**: Let us check if $\mu_y=\beta_0+\beta_1 x$. If we see the data are aligned in the scatterplot (linear pattern) → assumption 3 is established.

- **Residual plot ((y-ŷ) vs. x):** Let us check if the values of y for different values of x are independent, i.e. **there is no pattern or trend** (assumption 1). Let us see if the we could have the same σ (for example, if the **spread of data around 0 line is similar or uniform** for all values of x, that is a check for assumption 4)

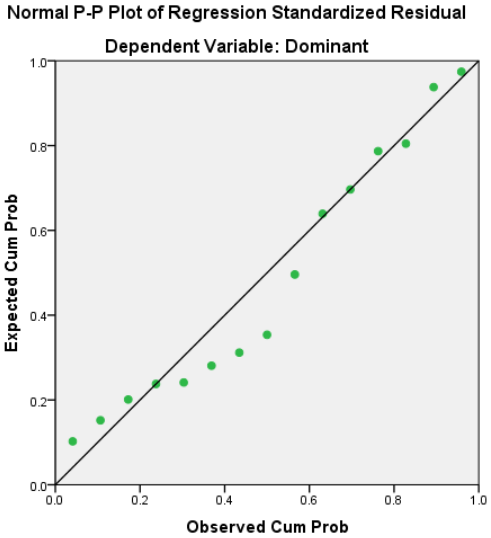
- **Normal probability plot for the residuals:** For assumption 2: y Normally distributed. The residuals are also Normally distributed with mean equal zero. If the Normal probability plot for the residuals shows a linear behavior → Normality.

For the Bone Strength study:

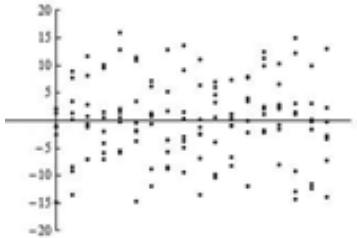
Residual plot ((y-ŷ) vs. x):



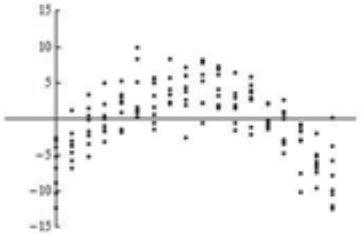
Normal probability plot for the residuals



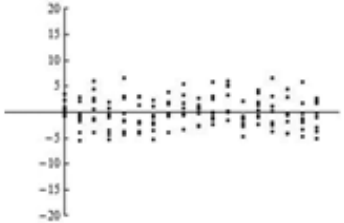
Residual Plot A



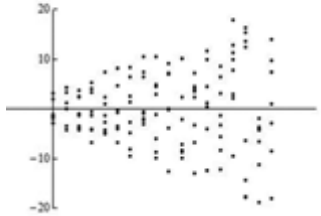
Residual Plot B



Residual Plot C

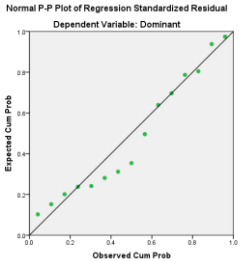


Residual Plot D



Constant Variance, Linearity, Normality

- For regression analyses, remember that we are making several big assumptions:
  1) by doing linear regression (as opposed to quadratic, polynomial, exponential, etc.), we expect the points to lie close to a straight line.
  2) we are assuming the spread of the corresponding y-values are also equally balanced.
  3) we expect the distribution of the x's and y's to be normal
  4) Finally, we also assume that the responses (y) corresponding to different values of x to be independent.
- For **linearity** and **constant variance**, we look at the **residuals plot**.
- For **normality**, we must look at a normal probability plot (see example to the right).

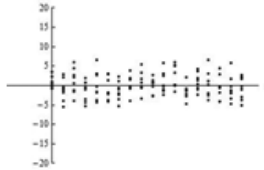


Linearity

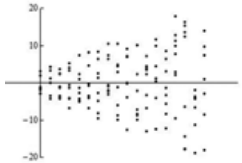
- Linearity can be determined from a residual plot. (You can also determine this from a scatterplot… the most basic question, “Do the points on the x-y scatterplot appear to go in a linear direction?”)
- On a residual plot, to be linear, the residuals must stay balanced about the horizontal axis. Essentially, a line of best fit for the residuals must have a slope of zero. Remember, when r=0, the slope of any best fit line is also zero.
- Keep in mind that you can be linear, yet not have constant variance. All the plots to the right are examples of linearity but may or may not have constant variance (see D). The key aspect is that the points lie balanced on the horizontal axis.

Residual Plot A

Residual Plot C



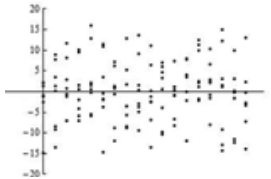
Residual Plot D



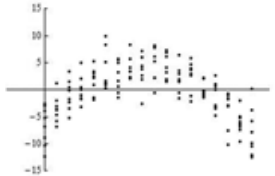
Constant Variance

- Constant variance can be determined from a residual plot. There is no easy way to determine constant variance from a scatterplot.
- Recall that variance is related to the standard deviation by:
  - $Variance = (Standard\ Deviation)^2$
  - $Standard\ Deviation = \sqrt{Variance}$
- When we say variance is constant, we are saying the spread is constant. The reason we must have constant variance is because we assume there is one σ for any response y.
- Keep in mind that you can have constant variance, yet not be linear. All the plots to the right are examples of constant variance but may or may not have linearity (see B). The key aspect is that the width or spread of the points must be uniform.

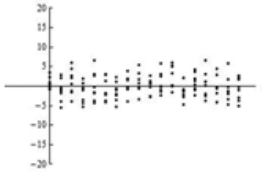
Residual Plot A



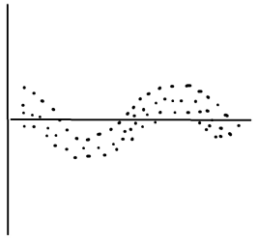
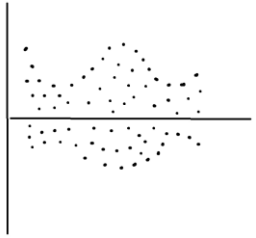
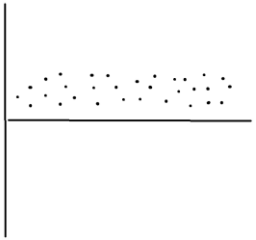
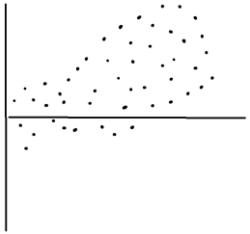
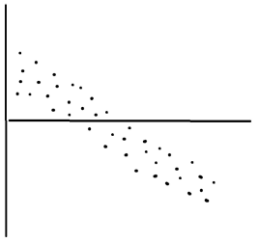
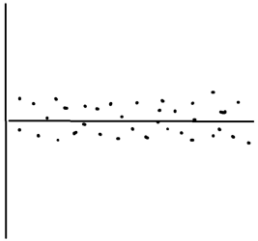
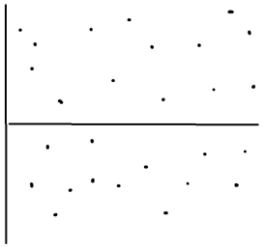
Residual Plot B



Residual Plot C



Examples and Counter Examples

LINEARITY – points evenly distributed on both sides of the horizontal axis

CONSTANT VARIANCE – points evenly spread about whatever trend may be occurring (linear or otherwise)

Assuming the conditions for the regression model are met, we can perform inference for the slope β_1 of the population regression line. We will use software for this!!

The question we want to answer is:

Is the regression coefficient non zero in the population?

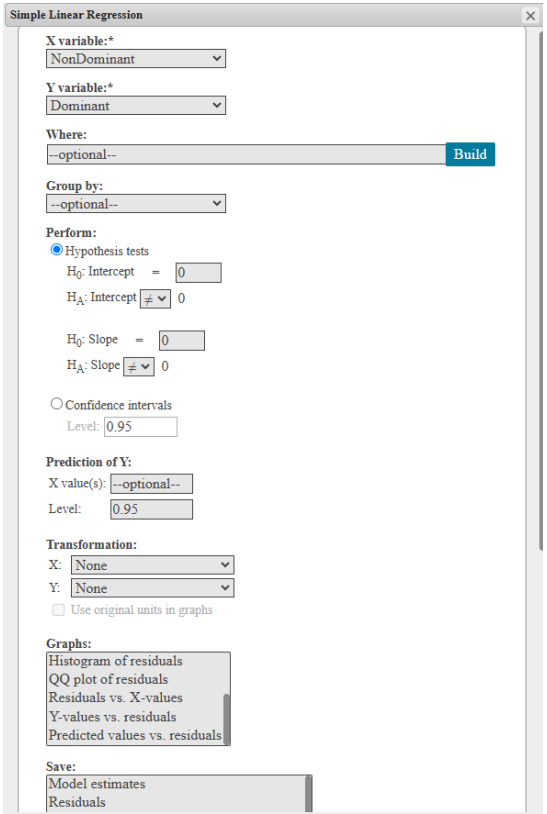
Or, do we have a linear relationship in the population?

$$H_0: \beta_1 = 0 \qquad H_a: \beta_1 \neq 0$$

→ **2-sided test**

Let's go look at StatCrunch with our data again.

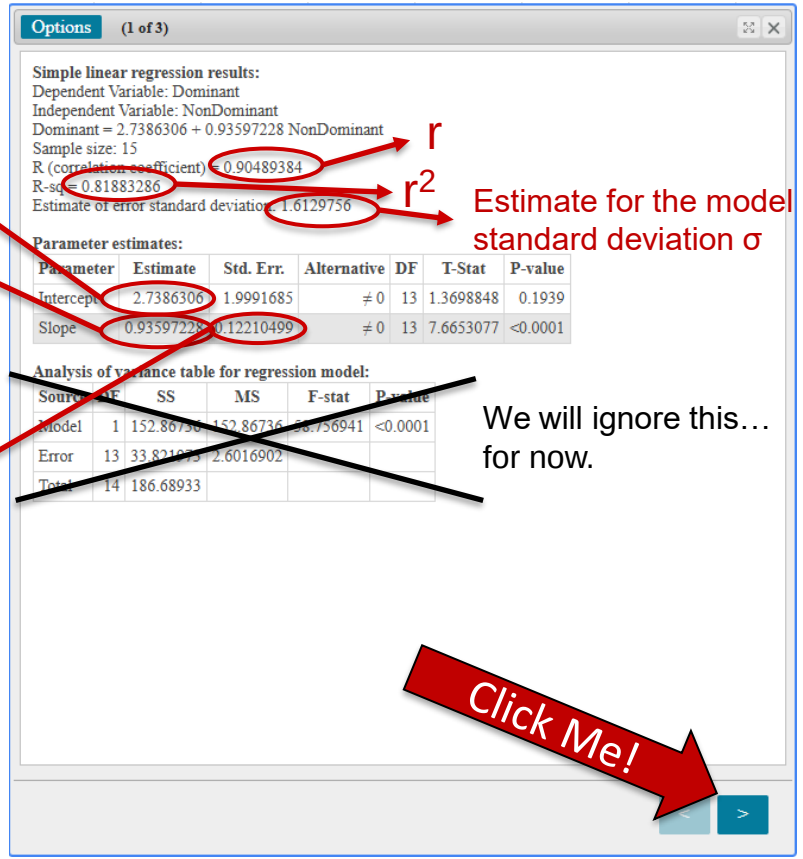
StatCrunch



b_0

b_1

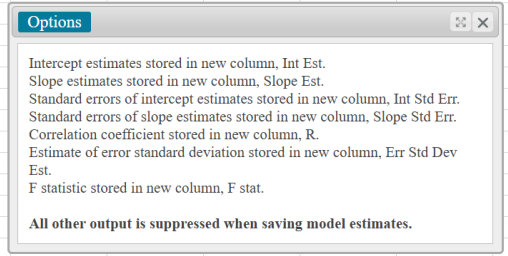
$$t_{b_1} = \frac{b_1}{SE_{b_1}}$$



Other Helpful Options

- You can click “Model Estimates” under the Save dialog box and get much of the same output, which gets dumped back on the spreadsheet view alongside the data.
- Selecting “Predicted Values” will give you $\hat{y}$ values in a column. Likewise, selecting residuals will give you all the $y - \hat{y}$, in a neatly calculated list.
- We also need scatterplots, residual plots, and normal probability plots, all easily found in the Graphs dialog box:
  - Select Fitted Line Plot (with mean interval), QQ plot of residuals, residuals and x-values

Int Est	Slope Est	Int Std Err	Slope Std Err	R	Err Std Dev Est	F stat
2.7386306	0.93597228	1.9991685	0.12210499	0.90489384	1.6129756	58.756941



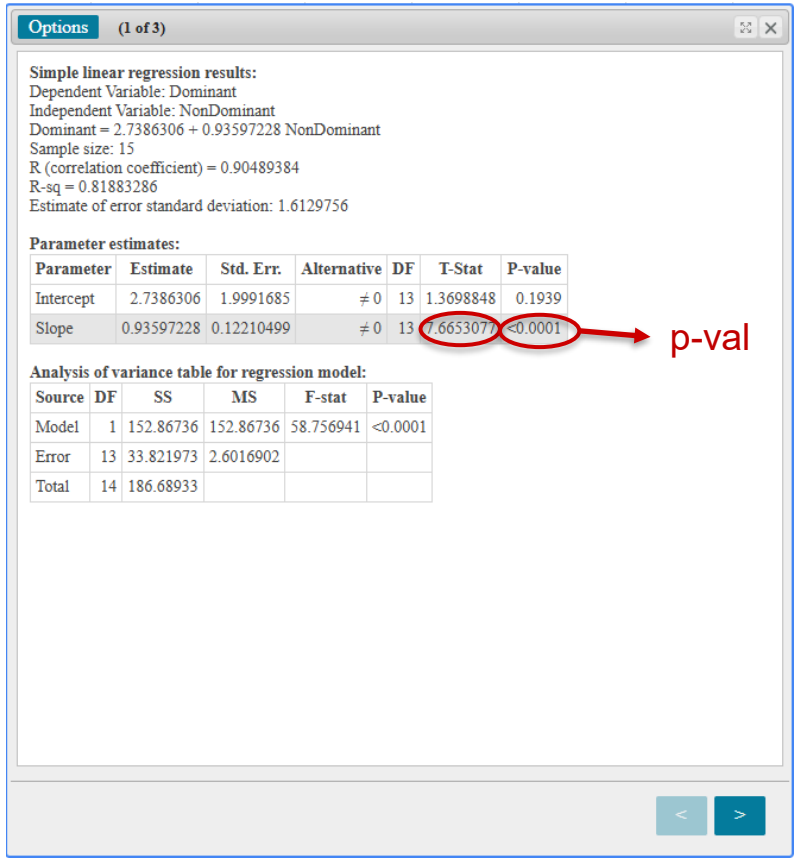
NonDominant	Dominant	Predicted Value	Residuals
15.7	16.3	17.433395	-1.1333954
25.2	26.9	26.325132	0.574868
17.9	18.7	19.492534	-0.79253437
19.1	22	20.615701	1.3842989
12	14.8	13.970298	0.82970208
20	19.8	21.458076	-1.6580762
12.3	13.1	14.25109	-1.1510896
14.4	17.5	16.216631	1.2833686
15.9	20.1	17.62059	2.4794102
13.7	18.7	15.561451	3.1385492
17.7	18.7	19.30534	-0.60533991
15.5	15.2	17.246201	-2.0462009
14.4	16.2	16.216631	-0.016631391
14.1	15	15.93584	-0.93583971
12.3	12.9	14.25109	-1.3510896

Significance test for the Slope:

Is the regression coefficient non zero in the population?

Or, do we have a linear relationship in the population?

$H_0: \beta_1 = 0$
$H_a: \beta_1 \neq 0$
From the StatCrunch output
$t = 7.665$, $P-value = 0.000 < 0.05 = \alpha \rightarrow$ reject H_0
At 5% significance level, we have evidence that the slope in the model (or the population slope) is different from zero.



When doing inference - Which hypothesis test do you use?

For Simple Linear Regression the *t*-statistics and *P*-values are identical for either symbol.

Use	*If the words are:*
β_1	Slope, regression coefficient
ρ	Correlation, independence
Either β_1 or ρ	linear relationship

Output Calculations

Simple linear regression results:
Dependent Variable: Dominant
Independent Variable: NonDominant
Dominant = 2.7386306 + 0.93597228 NonDominant
Sample size: 15
R (correlation coefficient) = 0.90489384
R-sq = 0.81883286
Estimate of error standard deviation: 1.6129756

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	2.7386306	1.9991685	≠ 0	13	1.3698848	0.1939
Slope	0.93597228	0.12210499	≠ 0	13	7.6653077	<0.0001

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	152.86736	152.86736	58.756941	<0.0001
Error	13	33.821973	2.6016902		
Total	14	186.68933			

- R^2 =
- Regression standard error, s =
- r =
- t_{b1} =
- MSE =
- And as we know from one-way ANOVA, F-stat comes from MSM/MSE = 152.86736/2.60169 = 58.757

Confidence Interval for b_1 (Slope)

- If we did this arm study 100 times, we will get some variation in the slope.
- In fact, we can use some of our old favorite inference techniques!
- Confidence interval for β:

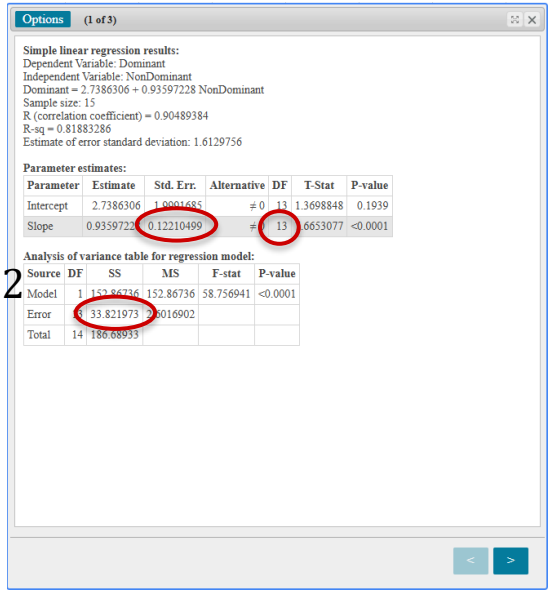
$$point\ estimate\ \pm confidence\ level\ *standard\ error$$

- So in this case,

$$b_1 \pm t_{df} \cdot SE_{b_1}$$

- Recall that t_{df} in this case is n – 2.
- $SE_{b_1} = \sqrt{\frac{1}{n-2} \cdot \frac{\sum(y_i-\hat{y}_i)^2}{\sum(x_i-\bar{x})^2}} = \sqrt{\frac{SSE}{(n-2)\cdot s_{xx}}} = \frac{s_e}{\sqrt{s_{xx}}} = \sqrt{\frac{33.822}{13\cdot 177.497}} = 0.122$
- So for df = 13 and a 95% CI, we find $t_{\alpha/2} = 2.160$ and thus:

  $0.936 \pm 2.16*0.122 = (0.672, 1.200) = CI\ for\ b_1$


Options (1 of 3)

Simple linear regression results:
Dependent Variable: Dominant
Independent Variable: NonDominant
Dominant = 2.7386306 + 0.93597228 NonDominant
Sample size: 15
R (correlation coefficient) = 0.90489384
R-sq = 0.81883286
Estimate of error standard deviation: 1.6129756

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	2.7386306	1.0991685	≠ 0	13	1.3698848	0.1939
Slope	0.9359722	0.12210499	≠ 0	13	7.6653077	<0.0001

Analysis of variance table for regression model:

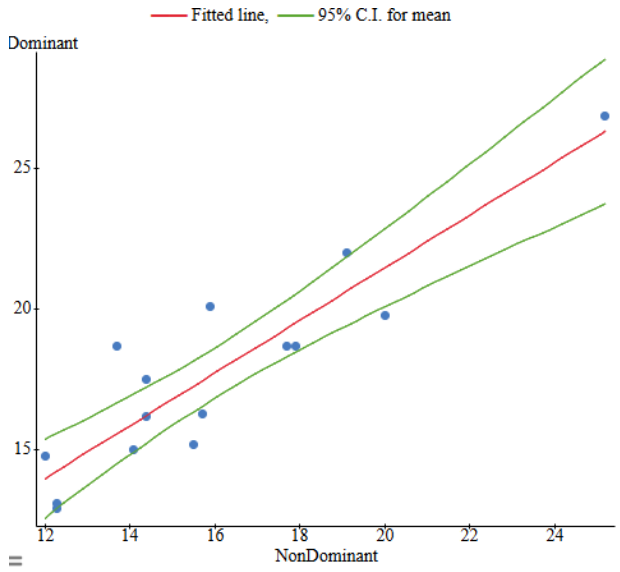
Source	DF	SS	MS	F-stat	P-value
Model	1	152.86736	152.86736	58.756941	<0.0001
Error	13	33.821973	2.6016902		
Total	14	186.68933			

Confidence Interval for Mean Response

- We can also calculate a confidence interval for the population mean μ_y of all responses y when x takes the value x* (within the range of the data tested).
- The confidence interval for the mean response μ_y at a given x^* of x is:

$$\hat{\mu}_y \pm t_{df} SE_{\hat{\mu}}$$

- A separate confidence interval could be calculated for μ_y along all the values that x takes. Graphically, the series of confidence intervals is shown as a continuous curve on either side of ŷ.
- In StatCrunch, choose Graphs, Fitted Line Plot – with mean interval



Confidence Interval for Mean Response

- How to find $SE_{\hat{\mu}}$.

MSE

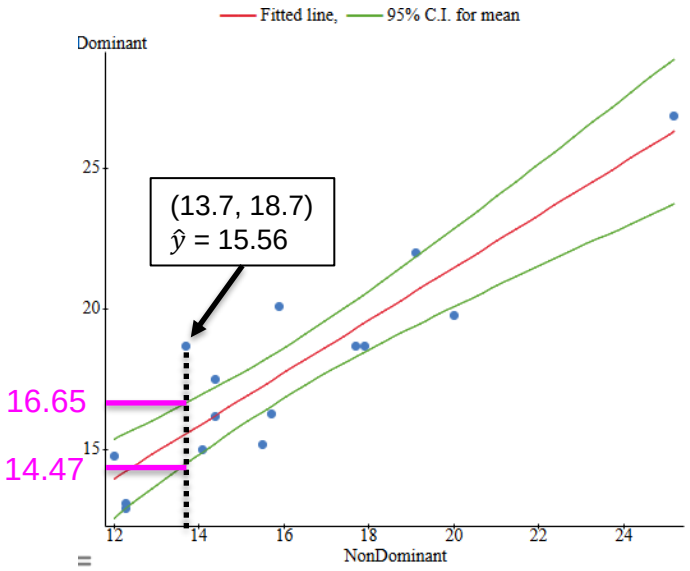
$$SE_{\hat{\mu}} = \sqrt{s_e^2 \cdot \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{s_{xx}}\right]}$$

- For example, for the measurement (13.7, 18.7), let's find the CI for 13.7.

$$15.56 \pm 2.16\sqrt{2.6017 \cdot \left[\frac{1}{15} + \frac{(13.7 - 16.013)^2}{174.497}\right]}$$

$$= (14.47, 16.65)$$

- Notice how the CI is NOT the same width for all x. Why not?
- In StatCrunch, click Save, Standard Error for Mean Response



Prediction Intervals

- Since there is variability involved in using a model created from sample data, a prediction interval is better than a single prediction (namely, if we collect data from different samples, the values of the y variable will probably be different for the different samples).
- One use of regression is for **predicting** the value of y at some value of x within the range of the data tested. Reliable predictions require statistical inference.
- To estimate an *individual* response y for a given value of x, we use a **prediction interval**.
- If we randomly sampled many times, many different values of y would be obtained for a particular x following N(0, σ) around the mean response μ_y.
- The prediction interval for a single observation on y when x takes the value x^* is:

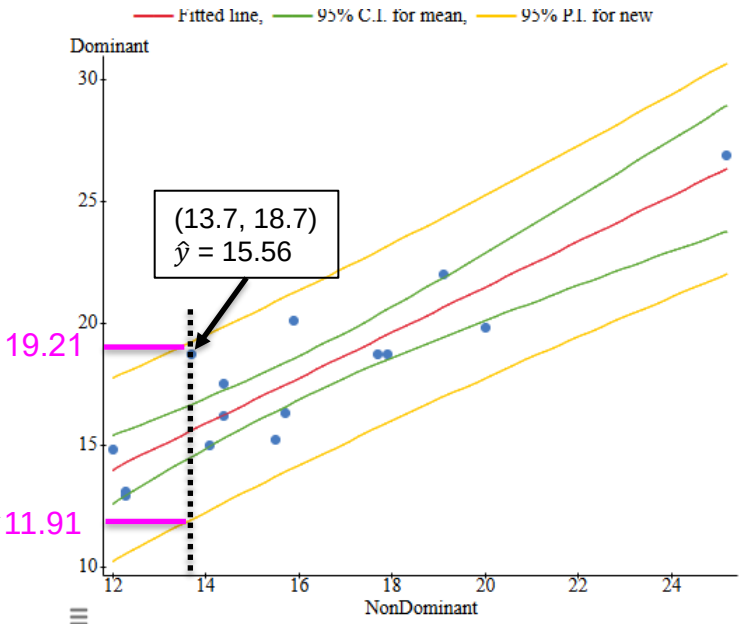
$$\hat{y} \pm t_{df} SE_{\hat{y}}$$

- We'll use StatCrunch for this.

Prediction Interval

- Accounts for :
  - ✓ Error in estimating β_0 and β_1
  - ✓ Uncertainty about the value of ***y*** being predicted.

- The prediction interval accounts for error in estimating β_0 and β_1 as well as uncertainty about the value of y being predicted.
- Graphically, the series of prediction intervals is shown as a continuous curve on either side of $\hat{y}$. These prediction intervals are wider than the corresponding confidence intervals for μ_y.
- To find $SE_{\hat{y}} = \sqrt{s_e^2 \cdot \left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{s_{xx}}\right]}$
- Notice this is almost identical to $SE_{\hat{\mu}}$ but the extra 1+ term is what increases the variability
- In StatCrunch, type a value for x^* like 13.7 in the field Prediction of Y:



Predicted values:

X value	Pred. Y	s.e.(Pred. y)	95% C.I. for mean	95% P.I. for new
13.7	15.561451	0.50322465	(14.4743, 16.648602)	(11.911179, 19.211723)

Example

- Moondollars is a coffee franchise that has many coffee stores all over the world. MoonDollars coffee is investigating a linear regression model to predict the annual income for a store based on the population of the city where the store is located. In the proposed regression model, annual store income is the response variable and population of the city is the explanatory variable.
- A random sample of 20 stores is selected and measurements are observed.
- Remember the trick? Right click, Copy Link and import from web in StatCrunch. Here is the data… moondollars.csv

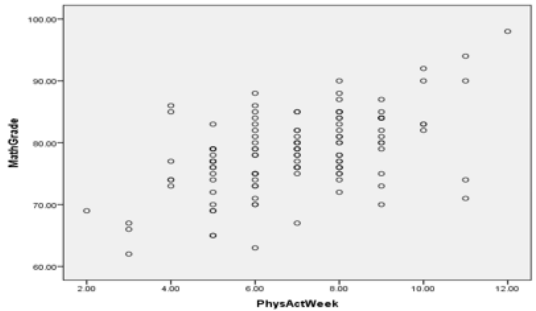
Analysis

- Let's complete all of the following:
  - Create a scatterplot
  - Find r, R^2, and the line of best fit
  - Predict the income of a store in a city of 2,120,000 and find the residual
  - Predict the income of a store in Mexico City (9.2 million). Is this valid?
  - Create a residual plot and check for linearity and constant variance
  - Do a hypothesis test at α=0.05 for correlation… $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ and StatCrunch
  - Do a hypothesis test at α=0.05 for regression.
  - Create a 95% CI for b_1, a 95% CI for mean response for 2,120, and a 95% PI for 2,120.

Is there evidence of a positive correlation between weekly time dedicated to physical activities (hours/week) and math grades for high school students? Can the math grades of high school students be predicted based on the weekly time dedicated to physical activities?
Below are the SPSS outputs of a study based on a sample of 120 high school students. For **Questions 42-44** in the next page choose the correct answer and write the corresponding letter on the line when needed.

Correlations

		PhysActWeek	MathGrade
PhysActWeek	Pearson Correlation	1	.520
	Sig. (2-tailed)		.000
	N	120	120
MathGrade	Pearson Correlation	.520	1
	Sig. (2-tailed)	.000	
	N	120	120



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.520[a]	.271	.264	5.49677

a. Predictors: (Constant), PhysActWeek

b. Dependent Variable: MathGrade

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	66.456	1.886		35.240	.000	62.722	70.190
	PhysActWeek	1.733	.262	.520	6.616	.000	1.214	2.251

a. Dependent Variable: MathGrade

42. (2 points) Provide the value of the percentage of variability in the data that is explained by the regression model. Write as a percentage with one decimal place.

Correlations

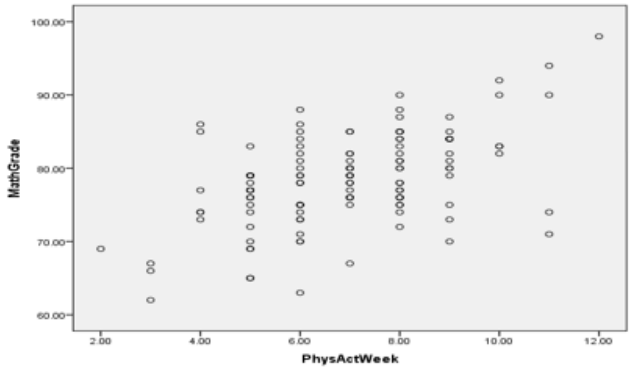
		PhysActWeek	MathGrade
PhysActWeek	Pearson Correlation	1	.520
	Sig. (2-tailed)		.000
	N	120	120
MathGrade	Pearson Correlation	.520	1
	Sig. (2-tailed)	.000	
	N	120	120

_____ 43. (2 points) Based on the output provided, we can conclude that there is evidence of a non-zero correlation between math grades and weekly time dedicated to physical activities for all high school students.

A. Because the correlation for this data set is 0.52.
B. Because the sign of the correlation for this data set is positive.
C. Because the P-value for the correlation between those two variables is 0.0.
D. Because the sample size is large enough.

_____ 44. (2 points) Using the regression equation from the SPSS outputs, if a high school student does fifteen hours of physical activities per week, her/his math grade would be 92.451.

A. This is a valid prediction as it is based on the regression equation.
B. This is an extrapolation.
C. This is an outlier.



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.520[a]	.271	.264	5.49677

a. Predictors: (Constant), PhysActWeek

b. Dependent Variable: MathGrade

4. From the coefficients output, at 5% significance level, we conclude that:

A. The population slope is positive because the lower and upper bound of the confidence interval for the slope are positive.
B. The population slope is positive because the B coefficient for PhysActWeek is 1.733.
C. The population slope is different from zero because the confidence interval for the slope doesn't contain the zero value.
D. We cannot reach any conclusion about the population slope from the coefficients output.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.520[a]	.271	.264	5.49677

a. Predictors: (Constant), PhysActWeek

b. Dependent Variable: MathGrade

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	66.456	1.886		35.240	.000	62.722	70.190
	PhysActWeek	1.733	.262	.520	6.616	.000	1.214	2.251

a. Dependent Variable: MathGrade