Correlation & Regression, III

9.07

4/6/2004

Outline

- Relationship between correlation and regression, along with notes on the correlation coefficient
- Effect size, and the meaning of r
- Other kinds of correlation coefficients
- Confidence intervals on the parameters of correlation and regression

Review

- Linear regression refers to fitting a best fit line $y=a+bx$ to the bivariate data (x, y), where
  $a = m_y - bm_x$
  $b = \text{cov}(x, y)/s_x^2 = ss_{xy}/ss_{xx}$
- Correlation, r, is a measure of the strength and direction (positive vs. negative) of the relationship between x and y.
  $r = \text{cov}(x, y)/(s_x s_y)$
  (There are various other computational formulas, too.)

Relationship between r and regression

- $r = \text{cov}(x, y)/(s_x s_y)$
- In regression, the slope, $b = \text{cov}(x, y)/s_x^2$
- So we could also write $b = r\cdot(s_y/s_x)$
- This means $b = r$ when $s_x = s_y$

Notes on the correlation coefficient, r

1. The correlation coefficient is the slope (b) of the regression line when both the X and Y variables have been converted to z-scores, i.e. when $s_x = s_y = 1$.
   Or more generally, when $s_x = s_y$.

   For a given s_x and s_y, the larger the size of the correlation coefficient, the steeper the slope.

Notes on the correlation coefficient, r

2. The correlation coefficient is invariant under linear transformations of x and/or y.
- (r is the average of $z_x z_y$, and z_x and z_y are invariant to linear transformations of x and/or y)

Invariance of r to linear transformations of x and y

- A linear change in scale of either x or y will not change r.
- E.G. converting height to meters and weight to kilograms will not change r.
- This is just the sort of nice behavior we'd like from a measure of the strength of the relationship.
  - If you can predict height in inches from weight in lbs, you can just as well predict height in meters from weight in kilograms.

How do correlations (=r) and regression differ?

- While in regression the emphasis is on predicting one variable from the other, in correlation the emphasis is on the degree to which a linear model may describe the relationship between two variables.
- The regression equation depends upon which variable we choose as the explanatory variable, and which as the variable we wish to predict.
- The correlation equation is symmetric with respect to x and y – switch them and r stays the same.

Correlation is symmetric wrt x & y, but regression is not

$a = m_y - bm_x$
$b = \text{cov}(x, y)/s_x^2$
$r = \text{cov}(x, y)/(s_x\, s_y)$

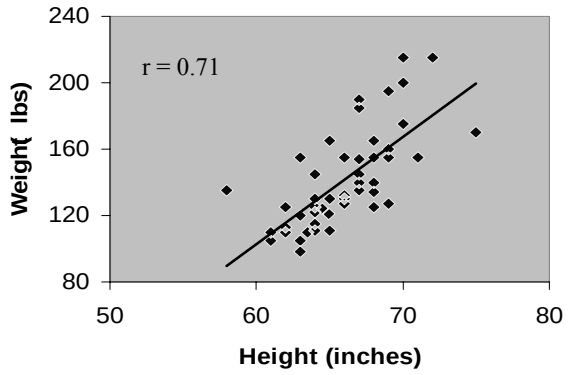
$\xrightarrow{x \leftrightarrow y}$

$a = m_x - bm_y$
$b = \text{cov}(x, y)/s_y^2$
$r = \text{cov}(x, y)/(s_x\, s_y)$

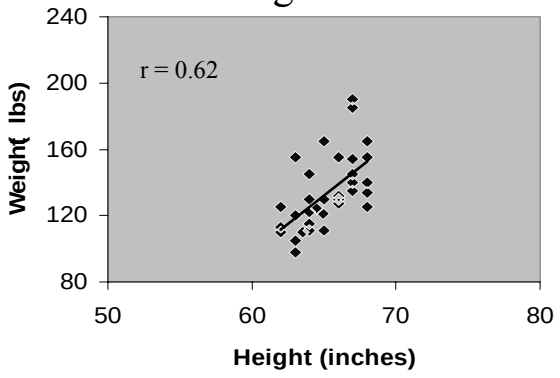
To look out for, when calculating r:

- In regression, we had to watch out for outliers and extreme points, because they could have an undue influence on the results.
- In correlation, the key thing to be careful of is not to artificially limit the range of your data, as this can lead to inaccurate estimates of the strength of the relationship (as well as give poor linear fits in regression)
  - Often it gives an underestimate of r, though not always

Correlation over a normal range

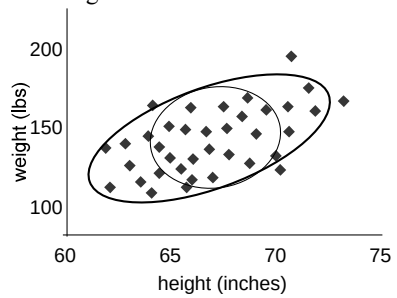


Correlation over a narrow range of heights



Correlation over a limited range

- A limited range will often (though not always) lead to an underestimate of the strength of the association between the two variables



Outline

- Relationship between correlation and regression, along with notes on the correlation coefficient
- Effect size, and the meaning of r
- Other kinds of correlation coefficients
- Confidence intervals on the parameters of correlation and regression

The meaning of r

- We've already talked about r indicating both whether the relationship between x and y is positive or negative, and the strength of the relationship
- The correlation coefficient, r, also has meaning as a measure of *effect size*

Effect size

- When we talked about effect size before, it was in the context of a two-sample hypothesis test for a difference in the mean.
- If there were a significant difference, we decided it was likely there was a real systematic difference between the two samples.
- Measures of effect size attempt to get at how big is this systematic effect, in an attempt to begin to answer the question "how important is it?"

Effect size & regression

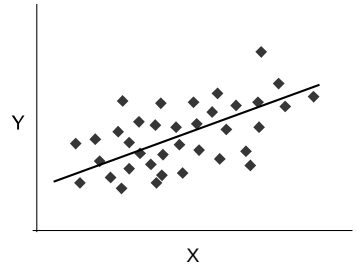
- In the case of linear regression, the *systematic effect* refers to the linear relationship between x and y
- A measure of effect size should get at how important (how strong) this relationship is
  - The fact that we're talking about strength of relationship should be a hint that effect size will have something to do with r

The meaning of r and effect size

- When we talked about two-sample tests, one particularly useful measure of effect size was the *proportion of the variance in y accounted for by knowing x*
- (You might want to review this, to see the similarity to the development on the following slides)
- The reasoning went something like this, where here it's been adapted to the case of linear regression:

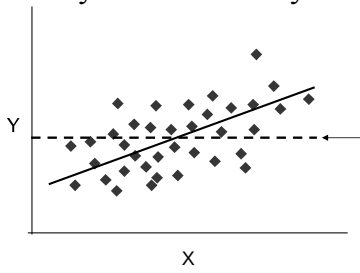
Predicting the value of y

- If x is correlated with y, then the situation might look like this:



Predicting the value of y

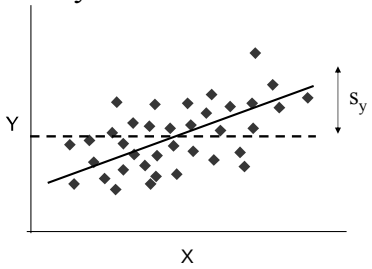
- Suppose I pick a random individual from this scatter plot, don't tell you which, and ask you to estimate y for that individual.



It would be hard to guess! The best you could probably hope for is to guess the mean of all the y values (at least your error would be 0 on average)

How far off would your guess be?

- The variance about the mean y score, s_y^2, gives a measure of your uncertainty about the y scores.

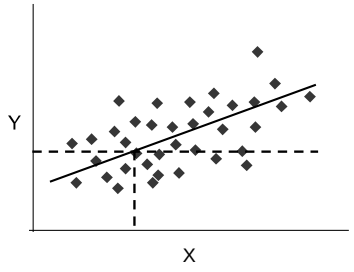


Predicting the value of y when you know x

- Now suppose that I told you the value of x, and again asked you to predict y.
- This would be somewhat easier, because you could use regression to predict a good guess for y, given x.

Predicting the value of y when you know x

- Your best guess is y', the predicted value of y given x.
- (Recall that regression attempts to fit the best fit line through the average y for each x. So the best guess is still a mean, but it's the mean y given x.)



How far off would your guess be, now?

- The variance about the mean score *for that value of x,* gives a measure of your uncertainty.
- Under the assumption of homoscedasticity, that measure of uncertainty is $s_{y'}^2$, where $s_{y'}$ is the rms error = $\mathrm{sqrt}(\Sigma(y_i - y_i')^2/N)$

The strength of the relationship between x and y

- is reflected in the extent to which knowing x reduces your uncertainty about y.
- Reduction in uncertainty = $s_y^2 - s_{y'}^2$
- Relative reduction in uncertainty: $(s_y^2 - s_{y'}^2)/ s_y^2$
- This is the *proportion of variance in y accounted for by x.*

(total variation – variation left over)/(total variation)
= (variation accounted for)/(total variation)

Unpacking the equation for the proportion of variance accounted for, $(s_y^2 - s_{y'}^2)/ s_y^2$

First, unpack $s_{y'}^2$:

$$s_{y'}^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - y_i')^2$$

where $y' = a + bx_i = (m_y - bm_x) + bx_i$

$$= m_y + \left(\frac{\text{cov}(x,y)}{s_x^2}\right)(x_i - m_x)$$

Unpacking the equation for the proportion of variance accounted for, $(s_y^2 - s_{y'}^2)/ s_y^2$

$$s_{y'}^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - y_i')^2 = \frac{1}{N}\sum_{i=1}^{N}((y_i - m_y) - \left(\frac{\text{cov}(x,y)}{s_x^2}\right)(x_i - m_x))^2$$

$$= \frac{1}{N}\sum_{i=1}^{N}(y_i - m_y)^2 + \left(\frac{\text{cov}(x,y)}{s_x^2}\right)^2 \frac{1}{N}\sum_{i=1}^{N}(x_i - m_x)^2$$

$$- 2\left(\frac{\text{cov}(x,y)}{s_x^2}\right)\sum_{i=1}^{N}(x_i - m_x)(y_i - m_y)$$

$$= s_y^2 + \frac{\text{cov}(x,y)^2}{s_x^2} - 2\frac{\text{cov}(x,y)^2}{s_x^2} = s_y^2 - \frac{\text{cov}(x,y)^2}{s_x^2}$$

Unpacking the equation for the proportion of variance accounted for, $(s_y^2 - s_{y'}^2)/ s_y^2$

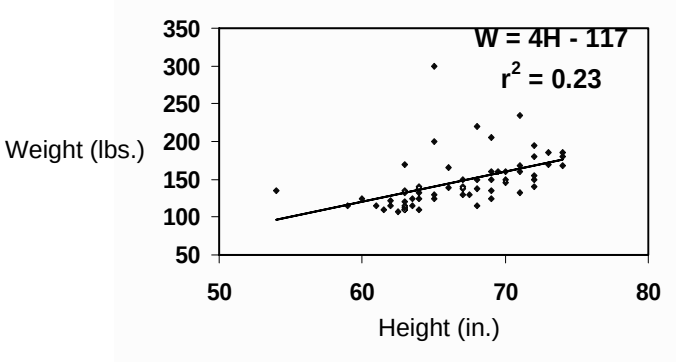
$$(s_y^2 - s_{y'}^2)/s_y^2 = \left(s_y^2 - s_y^2 + \frac{\text{cov}(x,y)^2}{s_x^2}\right) \Big/ s_y^2$$

$$= \frac{\text{cov}(x,y)^2}{s_x^2 s_y^2} = r^2 \;\; !!$$

r^2

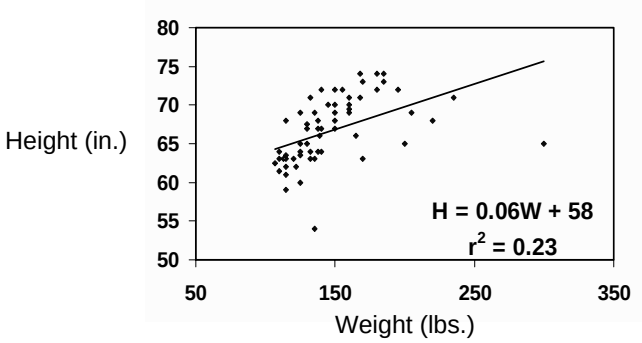
- The squared correlation coefficient (r^2) is the proportion of variance in y that can be accounted for by knowing x.
- Conversely, since r is symmetric with respect to x and y, r^2 it is the proportion of variance in x that can be accounted for by knowing y.
- r^2 is a measure of the size of the effect described by the linear relationship

Weight as a Function of Height



The linear relationship accounts for 23% of the variation in the data.

Height as a Function of Weight



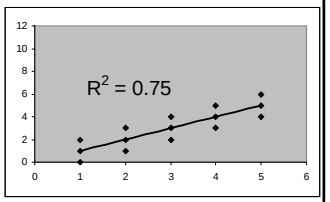
Again, accounting for 23% of the variability in the data.

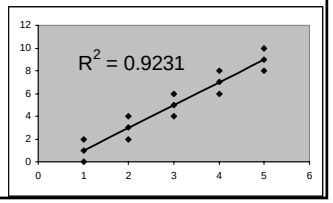
Behavior and meaning of the correlation coefficient

- Before, we talked about r as if it were a measure of the spread about the regression line, but this isn't quite true.
  - If you keep the spread about the regression line the same, but increase the slope of the line, r^2 increases
  - The correlation coefficient for zero slope will be 0 regardless of the amount of scatter about the line

Behavior and meaning of the correlation coefficient

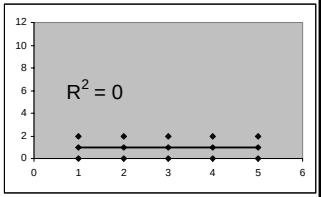
- Best way to think of the correlation coefficient: r^2 is % of variance in Y accounted for by the regression of Y on X.
  - As you increase the slope of the regression line, the total variance to be explained goes up.
  - Therefore the unexplained variance (the spread about the regression line) goes down *relative to the total variance.*
  - Therefore, r^2 increases..





Behavior and meaning of the correlation coefficient

- Best way to think of the correlation coefficient: r^2 is % of variance in Y accounted for by the regression of Y on X.
  - If the slope of the regression line is 0, then 0% of the variability in the data is accounted for by the linear relationship. So $r^2 = 0$



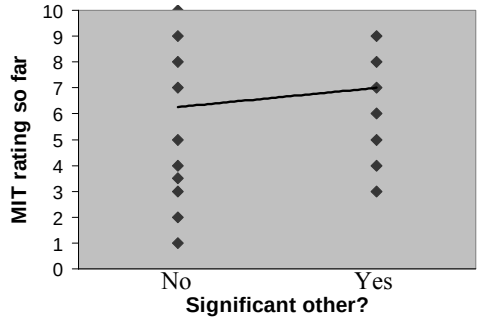
Other correlation coefficients

- This percent of variance accounted for is a useful concept.
- We talked about it before in talking about effect size.
- One thing it lets us do is compare effect sizes across very different kinds of experiments.
- Different kinds of experiments -> different correlation coefficients

r_{pb}: The point-biserial correlation coefficient

- This was the correlation coefficient we used when measuring effect size for a two-sample test.
- This is used when one of your variables takes on only two possible values (a *dichotomous* variable)
- In the two-sample case, the two possible value corresponded to the two experimental groups or conditions you wished to compare.
  - E.G. are men significantly taller than women?

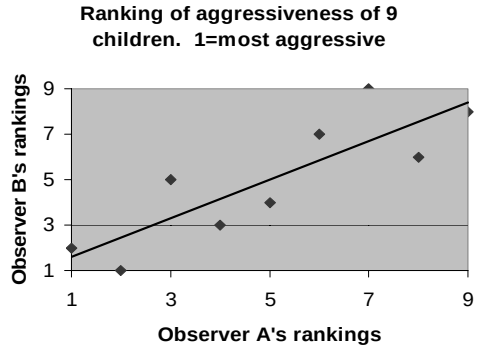
The "correlation"/"regression" associated with r_{pb}



r_s: The Spearman rank-order correlation coefficient

- Describes the linear relationship between two variables when measured by ranked scores.
- E.G. instead of using the actual height of a person, we might look at the rank of their height – are they the tallest in the group? The 2nd tallest? Compare this with their ranking in weight.
- Often used in behavioral research because a variable is difficult to measure quantitatively. May be used to compare observer A's ranking (e.g. of the aggressiveness of each child) to observer B's ranking.

Scatter plot for ranking data



Three main types of correlation coefficients: summary

- Pearson product-moment correlation coefficient
  - Standard correlation coefficient, r
  - Used for typical quantitative variables
- Point-biserial correlation coefficient
  - r_{pb}
  - Used when one variable is dichotomous
- Spearman rank-order correlation coefficient
  - r_s
  - Used when data is ordinal (ranked) (1st, 2nd, 3rd, …)

Outline

- Relationship between correlation and regression, along with notes on the correlation coefficient
- Effect size, and the meaning of r
- Other kinds of correlation coefficients
- Confidence intervals on the parameters of correlation and regression

Confidence intervals and hypothesis testing on a, b, y', and r

Best fit line for the data (the sample)

- With regression analysis, we have found line $y = a + bx$ that best fits our data.
- Our model for the data was that it was approximated by a linear relationship plus normally-distributed random error, e, i.e.

  $y = a + bx + e$

Best fit line for the population

- As you might imagine, we can think of our points (x_i, y_i) as samples from a population
- Then our best fit line is our estimate of the true best fit line for the population
- The regression model for the population as a whole:

  $y = \alpha + \beta x + \varepsilon$
- Where α and β are the parameters we want to estimate, and ε is the random error
  - Assume that for all x, the errors, ε are independent, normal, of equal σ, and mean $\mu=0$

Confidence intervals for α and β

- a and b are unbiased estimators for α and β, given our sample
- Different samples yield different regression lines.
- These lines are distributed around $y = \alpha+\beta x+\varepsilon$
- How are a and b distributed around α and β? If we know this, we can construct confidence intervals and do hypothesis testing.

An estimator, s_e for $\sigma(\varepsilon)$

- The first thing we need is an estimator for the standard deviation of the error (the spread) about the true regression line
- A decent guess would be our rms error, $\text{rms} = \text{sqrt}(\Sigma(y_i - y_i')^2/N) = s_{y'}$
- We will use a modification of this estimate:

An estimator, s_e for $\sigma(\varepsilon)$

- $s_e = \text{sqrt}(\Sigma(y_i - y_i')^2/(N-2))$
- Why n-2? We used up two degrees of freedom in calculating a and b, leaving n-2 independent pieces of information to estimate $\sigma(\varepsilon)$
- An alternative computational formula:
  $s_e = \text{sqrt}((ss_{yy} - b \cdot ss_{xy})/(n-2))$

Recall some notation

- $ss_{xx} = \Sigma(x_i - m_x)^2$
  $ss_{yy} = \Sigma(y_i - m_y)^2$
  $ss_{xy} = \Sigma(x_i - m_x)(y_i - m_y)$

Confidence intervals for α and β have the usual form

- $\alpha = a \pm t_{crit}\ SE(a)$
- $\beta = b \pm t_{crit}\ SE(b)$

  Where we use the t-distribution with N-2 degrees of freedom, for the same reason as given in the last slide.
- Why are a and b distributed according to a t-distribution?
  - This is complicated, and is offered without proof. Take it on faith (or, I suppose, try it out in MATLAB).
- So, we just need to know what SE(a) and SE(b) are…

SE(a) and SE(b) look rather unfamiliar

- $SE(a) = s_e \cdot \text{sqrt}(1/N + m_x^2/ss_{xx})$
- $SE(b) = s_e/\text{sqrt}(ss_{xx})$
- Just take these equations on faith, too.

SE(a) and SE(b)

- But some intuition: where does this ss_{xx} come from?
  - Usually we had a 1/sqrt(N) in our SE equations; now we have a $1/\text{sqrt}(ss_{xx})$
  - Like N, ss_{xx} increases as we add more data points
  - ss_{xx} also takes into account the spread of the x data
- Why care about the spread of the x data?
  - If all data points had the same x value, we'd be unjustified in drawing any conclusions about α or β, or in making predictions for any other value of x
  - If $ss_{xx} = 0$, our confidence intervals would be infinitely wide, to reflect that uncertainty

See picture on board

So, you now know all you need to get confidence intervals for α and β

- Just plug into the equation from 3 slides ago
- What about hypothesis testing?
  - Just as before when we talked about confidence intervals and hypothesis testing, we can turn our confidence interval equation into the equation for hypothesis testing
  - E.g. is the population slope greater than 0?

Testing whether the slope is greater than 0

- H_0: $\beta=0$, H_a: $\beta>0$
- $t_{obt} = b/SE(b)$
- Compare t_{obt} to t_{crit} at the desired level of significance (for, in this case, a one-tailed test). Look up t_{crit} in your t-tables with N-2 degrees of freedom.

What about confidence intervals for the mean response y' at $x=x_o$?

- The confidence interval for $y'=a+bx_o$ is
  $\alpha+\beta x_o = a+bx_o \pm t_{crit}\ SE(y')$
- Where $SE(y') =$
  $s_e \cdot \text{sqrt}(1/N + (x_o - m_x)^2/ss_{xx})$
- Note the $(x_o - m_x)$ term
  - The regression line always passes through (m_x, m_y)
  - If you "wiggle" the regression line (because you're unsure of a and b), it makes more of a difference the farther you are from the mean.

See picture on board

Versus confidence intervals on an individual's response y_{new}, given x_{new}

- Previous slide had confidence intervals for the predicted *mean* response for $x=x_o$
- You can also find confidence intervals for an *individual's* predicted y value, y_{new}, given their x value, x_{new}
- As you might imagine, $SE(y_{new})$ is bigger than $SE(y')$
  - First there's variability in the predicted mean (given by $SE(y')$)
  - Then, on top of that, there's the variability of y about the mean

See picture on board

$SE(y_{new})$

- These variances add, so
  $SE(y_{new})^2 = SE(y')^2 + s_e^2$

$SE(y_{new}) = s_e \cdot \text{sqrt}(1/N + (x_{new} - m_x)^2/ss_{xx} + 1)$

Homework notes

- Problems 5 and 9 both sound like they could be asking for either confidence intervals on y', or on y_{new}. They are somewhat ambiguously worded
- Assume that problem 5 is asking for confidence intervals on y', and problem 9 is asking for confidence intervals on y_{new}

Hypothesis testing on r

- We often want to know if there is a significant correlation between x and y
- For this, we want to test whether the population parameter, ρ, is significantly different from 0
- It turns out that for the null hypothesis H_0: $\rho=0$, $r \cdot \text{sqrt}(N-2)/\text{sqrt}(1-r^2)$ has a t distribution, with N-2 degrees of freedom
- So, just compare $t_{obt} = r \cdot \text{sqrt}(N-2)/\text{sqrt}(1-r^2)$ with t_{crit} from your t-tables.
- Don't use this test to test any other hypothesis, H_0: $\rho=\rho_0$. For that you need a different test statistic.

Examples

Previous example: predicting weight from height

x_i	y_i	(x_i-m_x)	(y_i-m_y)	$(x_i-m_x)^2$	$(y_i-m_y)^2$	$(x_i-m_x)(y_i-m_y)$
60	84	-8	-56	64	3136	448
62	95	-6	-45	36	2025	270
64	140	-4	0	16	0	0
66	155	-2	15	4	225	-30
68	119	0	-21	0	441	0
70	175	2	35	4	1225	70
72	145	4	5	16	25	20
74	197	6	57	36	3249	342
76	150	8	10	64	100	80
Sum=612	1260			ss_{xx}=240	ss_{yy}=10426	ss_{xy}=1200
m_x=68	m_y=140					

$b = ss_{xy}/ss_{xx} = 1200/240 = 5$; $a = m_y - bm_x = 140-5(68) = -200$

Confidence intervals on α and β

- $s_e = \text{sqrt}((ss_{yy} - b \cdot ss_{xy})/(n-2)) = \text{sqrt}((10426 - (5)(1200))/7) = 25.15$
- $SE(a) = s_e \cdot \text{sqrt}(1/N + m_x^2/ss_{xx}) = 25.15 \cdot \text{sqrt}(1/9 + 68^2/240) = 110.71$
- $SE(b) = s_e/\text{sqrt}(ss_{xx}) = 25.15/\text{sqrt}(240) = 1.62$
- t_{crit} for 95% confidence = 2.36, so
  $\alpha = -200 \pm (2.36)\ (110.71)$
  $\beta = 5 \pm (2.36)\ (1.62) = 5 \pm 3.82$
  - It looks like β is significantly different from 0

Confidence intervals on y' and y_{new}, for x = 76"

- $\alpha+\beta x_o = a+bx_o \pm t_{crit}\ SE(y')$
- $SE(y') = s_e \cdot \text{sqrt}(1/N + (x_o - m_x)^2/ss_{xx}) = (25.15)\ \text{sqrt}(1/9 + (76\text{-}68)^2/240) = (25.15)\ (0.61) = 15.46$
- So $\alpha+\beta x_o = -200 + 5(76) \pm (2.36)\ SE(y') = 180 \pm 36.49$ lbs
- $SE(y_{new}) = (25.15)\ \text{sqrt}(1+1/N+ (x_{new} - m_x)^2/ss_{xx}) = 29.52$
- So $y_{new} = 180 \pm (2.36)(29.52) \approx 180 \pm 70$ lbs

Testing the hypothesis that ρ≠0

- $r = \text{cov}(x, y)/(s_x\ s_y) = (ss_{xy}/N)/\text{sqrt}((ss_{xx}/N\ ss_{yy}/N)) = 1200/\text{sqrt}((240)(10426)) = 0.76$
- $t_{obt} = r \cdot \text{sqrt}(N-2)/\text{sqrt}(1-r^2) = 0.76\ \text{sqrt}(7)/\text{sqrt}(1\text{-}0.58) \approx 3.10$
- $t_{crit} = 2.36$, so ρ is significantly different from 0

A last note

- The test for β≠0 is actually the same as the test for ρ≠0.
- This should make sense to you, since if ρ≠0, you would also expect the slope to be ≠0
- Want an extra point added to your midterm score? Prove that the two tests are the same.
  - In general, not just for this example
  - Submit it with your next homework