EDUR 8131
Chat 11

Notes 8b Multiple Regression
Topics 1.1 to 1.3 cover in Chat 10.

1. Notes 8b Multiple Regression (resume with 1.6 Inference in Regression)

Summary of content covered in Chat 10 (Topics 1 to 5)

1.1 Regression Equation

Y = b0 + b1 X1 + b2 X2 + e
Y’ = b0 + b1 X1 + b2 X2 (prediction equation, used to obtain predicted value of Y)

1.2 Literal Interpretation of Coefficients and Predicted Values

See Box Office sales example below.

1.3 Predicted Values vs. Expected Change

See Box Office sales example below.

1.4 Obtaining Regression Estimates

Box Office Sales

http://www.bwgriffin.com/gsu/courses/edur8131/data/movie_sales.sav

> [Copy and paste into News Item in folio if folks have trouble downloading; Expression Web _target built into link above so new folder opens]

Contains the following variables

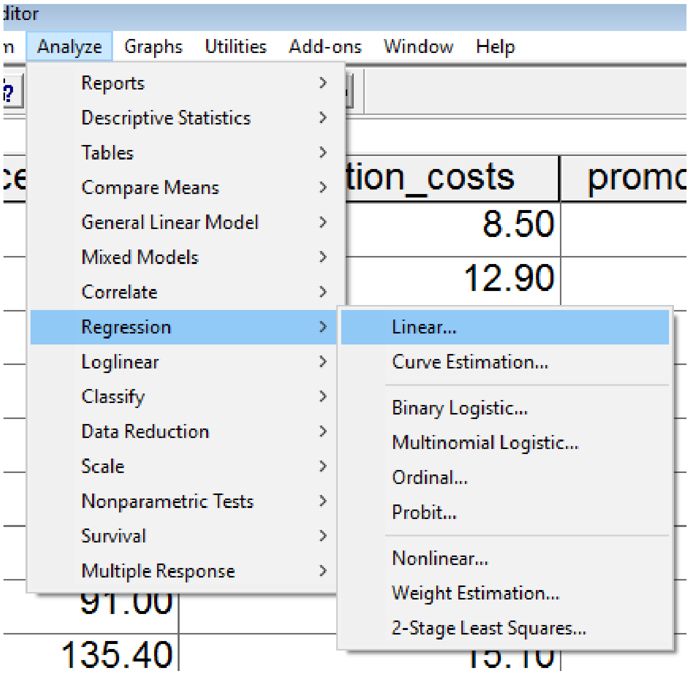
DV = Box office sales (in millions of dollars)
IV or Predictors =

- Production costs (in millions of dollars)
- Promotion costs (in millions of dollars)
- Book sales (in millions of books sold prior to movie)

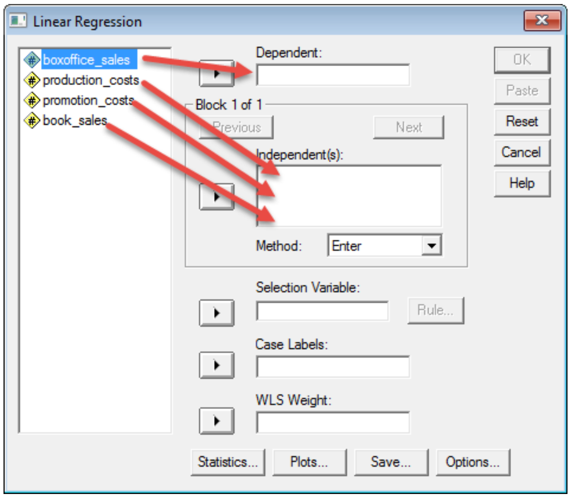
(a) What is the regression equation and coefficient estimates for the above data?

> Y’ = b0 + b1 (X1) + b2 (X2) + b3 (X3)
> Sales’ = b0 + b1 (production costs) + b2 (promotion costs) + b3 (book sales)

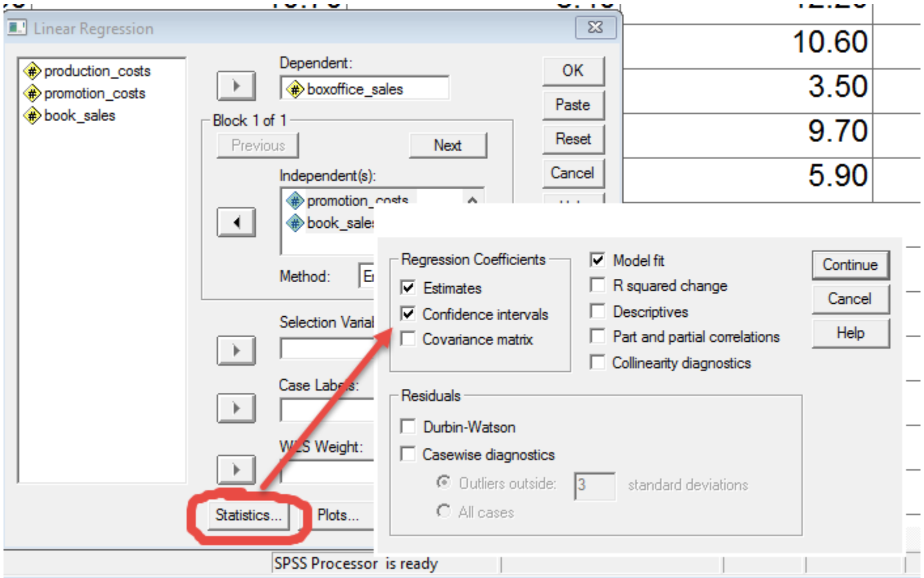
To run SPSS regression, follow these steps:



Move IV and DV to appropriate boxes



Then, to get confidence intervals, click on Statistics, then CI.



Dependent: boxoffice_sales

Independent(s): promotion_costs, book_sales

Regression Coefficients
☑ Estimates
☑ Confidence intervals
☐ Covariance matrix

☑ Model fit
☐ R squared change
☐ Descriptives
☐ Part and partial correlations
☐ Collinearity diagnostics

Residuals
☐ Durbin-Watson
☐ Casewise diagnostics
● Outliers outside: 3 standard deviations
○ All cases

Then click continue and ok to run.

Post value of intercept obtained; this will help ensure we all are obtaining that same regression results.

SPSS Results

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	7.676	6.760		1.135	.299	-8.866	24.218
	production_costs	3.662	1.118	.421	3.276	.017	.927	6.397
	promotion_costs	7.621	1.657	.559	4.598	.004	3.566	11.676
	book_sales	.828	.539	.127	1.536	.175	-.491	2.148

a. Dependent Variable: boxoffice_sales

So regression equation with coefficient estimates inserted would be:

BO Sales' = 7.676 + 3.662 (production $) + 7.621 (promotion $) + 0.828 (Book millions)

(b) Literal Interpretation of Each Coefficient

b0 = 7.676:

The predicted box office sales is $7.676 million when production costs are $0, promotion costs are $0, and when book sales are 0

b1 = 3.662 (production costs millions):

For each 1 million dollar increase in production costs, one would expect box office sales to increase by 3.662 millions of dollars controlling for promotion costs and book sales.

b2 = 7.621 (promotion costs millions):

For each 1 million dollar increase in promotion costs, one would expect box office sales to increase by 7.621 millions of dollars controlling for production costs and book sales.

b3 = 0.828 (book sales in millions):
For addition million books sold, one would expect box office sales to increase by 0.828 millions of dollars controlling for promotion and production costs.

Summary of literal interpretation for slope:

The slope tells us how much change can be expected on the DV (box office sales) for a one unit increase in the IV (e.g., production costs) controlling for the other predictors.

General interpretation for slopes (b1, b2, and b3)

Each of the production costs, promotion costs, and book sales is positively associated with box office sales; the greater production costs, promotion costs, or book sales, the greater is box office sales.

BO Sales' = 7.676 + 3.662 (production $) + 7.621 (promotion $) + 0.828 (Book millions)

For each of the following values of the predictors, what are the predicted box office sales?

Example 1

production cost = $5 million
promotion cost = $10 million
book sales = 1.75 million

Predicted box office sales would be:

Sales' = 7.676 + 3.662*(5) + 7.621*(10) + 0.828*(1.75)
= $103.645 million

Example 2

BO Sales' = 7.676 + 3.662 (production $) + 7.621 (promotion $) + 0.828 (Book millions)

production cost = $15 million
promotion cost = $828,103
book sales = 93,511

First, one must convert promotion costs and book sales into units of millions
Promotion costs in millions of dollars = 828,103 / 1,000,000 = .828103
Book sales in millions = 93,511 / 1,000,000 = .093511

With these numbers, what value do you obtain for predicted Box Office sales?

Sales’ = 7.676 + 3.662*(15) + 7.621*(.828103) + 0.828*(.093511)
= $68.994 millions

Example 3

BO Sales’ = 7.676 + 3.662 (production $) + 7.621 (promotion $) + 0.828 (Book millions)

production cost = $64,000
promotion cost = $6,283
book sales = 0

Sales’ = 7.676 + 3.662*(.064000) + 7.621*(.006283) + 0.828*(0)
= $7.958 millions

(e) Expected Change

Regression equation

Sales’ = 7.676 + 3.662 (production $) + 7.621 (promotion $) + 0.828 (Book millions)

Example 1

If book sales increases by 50,000 units, what is the expected change in box office sales?

Convert book sales to millions of units = 50000/1000000 = .05
b3*(change in IV) = 0.828*(.050000) = $0.0414 millions of dollars (or $41,400)

Example 2

Sales’ = 7.676 + 3.662 (production $) + 7.621 (promotion $) + 0.828 (Book millions)

If promotion costs were to be increased by $345,000, what is the expected change in box office sales?

b2*(change in IV) = 7.621*(.345) = $2.629 millions

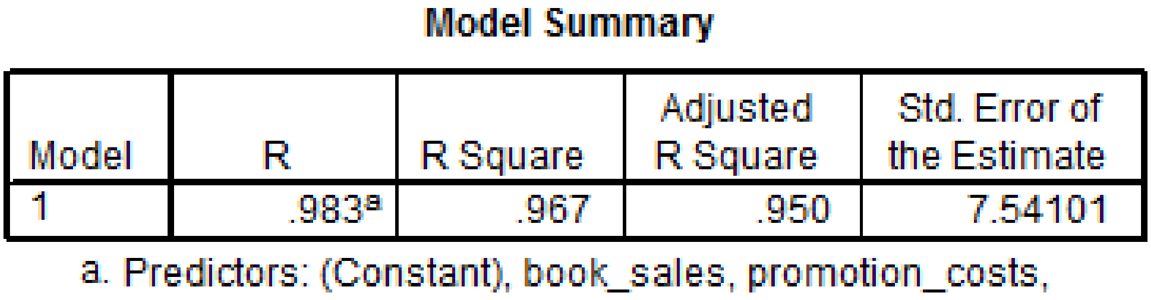
Example 3

If production costs were decreased by $785,000, what is the expected change in box office sales?

b1*(change in IV) = 3.662*(-.785) = -$2.874 millions

1.5 Model Fit

SPSS Results



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.983[a]	.967	.950	7.54101

a. Predictors: (Constant), book_sales, promotion_costs,

(a) Multiple R – how obtained (what information used to calculate) and what does it mean?

R = .983

Remember that R is the Pearson correlation between Y and predicted Y'

So the correlation between observed box office sales and predicted box office sales (using the regression equation) is .983

[Demonstrate in Excel using regression equation]

(b) Multiple R^2 – how obtained and what does it mean, what is interpretation of R^2?

Multiple R squared = R^2 = .967 = R^2 (i.e., .983^2)

Proportion reduction in prediction error, or proportion of variance in DV (box office sales) that can be predicted by the regression equation.

(c) Adjusted R^2 – what does it mean?

Adjusted R^2 = .95
Has the same interpretation as R^2 given above.

SEE = standard deviation of residuals (except divided by df2 = n-k-1)

1.6 Inference in Regression

(a) Overall model fit (i.e., does the regression model, as a whole, help predict the DV?)

Ho: $R^2 = 0.00$
H_1: $R^2 \neq 0.00$

For Movie Data R^2 = .967 (extremely high value)

This null is assessed via the ANOVA F-ratio.

If significant (i.e., calculated F is larger than critical F, or $p \leq \alpha$), then overall model predicts more variance in DV than would be expected by chance alone.

SPSS Results for Movie Data

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9932.463	3	3310.821	58.221	.000[b]
	Residual	341.201	6	56.867		
	Total	10273.664	9			

a. Dependent Variable: boxoffice_sales

b. Predictors: (Constant), book_sales, promotion_costs, production_costs

Mean Square Residual = variance of residuals (sometimes called variance error)

Question

If we were able to predict perfectly the DV with our regression model, what would be the value of the Mean Square Residual (also called Mean Square Error or MSE)?

Question

Is this F ratio significant at the .05 level?

F = 58.221
p-value = .000

Recall the decision rule for hypothesis testing

If $p \leq \alpha$ reject Ho; if $p > \alpha$ fail to reject Ho.

$\alpha = .05$
p = 0.000

Answer

If $0.000 \leq .05$ reject Ho; if $p > \alpha$ fail to reject Ho

Reject Ho: $R^2 = 0.00$

Yes, significant. This result tells us something in our regression model is able to predict more variance in the DV (movie box office sales) than would be expected by chance alone.

If the overall model is judged to predict move variance in the DV than would be expected by random variation—random chance—then the next question is which variables or predictors are helpful in predicting the DV?

(b) Regression Coefficients

Ho: $b_1 = 0.00$
H_1: $b_1 \neq 0.00$

Test each individually with ratio of ***regression coefficient*** to ***standard error of regression coefficient***:

b_1 / se b_1 = t ratio

If t-ratio is significant, then this suggests the regression coefficient is sufficiently different from 0.00 (as specified by the null) that we believe the difference from 0.00 is greater than would be expected by chance.

SPSS Results

Coefficients[a]

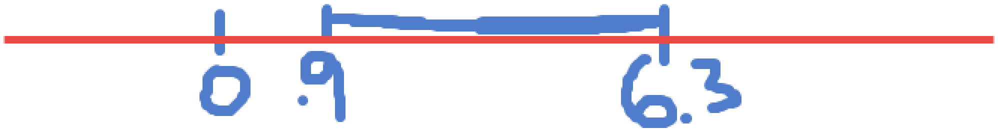
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	7.676	6.760		1.135	.299	-8.866	24.218
	production_costs	3.662	1.118	.421	3.276	.017	.927	6.397
	promotion_costs	7.621	1.657	.559	4.598	.004	3.566	11.676
	book_sales	.828	.539	.127	1.536	.175	-.491	2.148

a. Dependent Variable: boxoffice_sales

b1: production costs – is this significant at .05 level? Note that "Sig." is the p-value for the t-test in regression in the table above.

P = .017, which is less than .05, so reject Ho

Remember, you can check the accuracy of your inference with p-values by examining the CI; if 0 is not within the CI, reject Ho. For example, with production costs, the interval is .9 to 6.3, note 0 is not with this interval (see number line below).,



Interpretation – we believe the true partial slope relating production costs to box office sales, controlling for promotion costs and book sales, is somewhere between .9 and 6.3.

b2: promotions – is this significant at .05 level?

SPSS Results

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)	7.676	6.760		1.135	.299	-8.866	24.218
	production_costs	3.662	1.118	.421	3.276	.017	.927	6.397
	promotion_costs	7.621	1.657	.559	4.598	.004	3.566	11.676
	book_sales	.828	.539	.127	1.536	.175	-.491	2.148

a. Dependent Variable: boxoffice_sales

P = .004, so reject Ho

b3: book sales – is significant at .05 level?

P = .175, so fail to reject

We could also use the confidence interval for hypothesis testing, how?

Like with the two group t-test, if 0.00 lies within the CI, fail to reject Ho.

1.7 Confidence Intervals for Regression Coefficients

Confidence intervals provide a range of values that could represent the population parameter. The range is within a specified level of confidence ($1 – \alpha$), so there is an upper and lower limit to the interval. The interval is conceptually identical to confidence intervals found in t-tests for mean differences.

Formula for CI in Regression

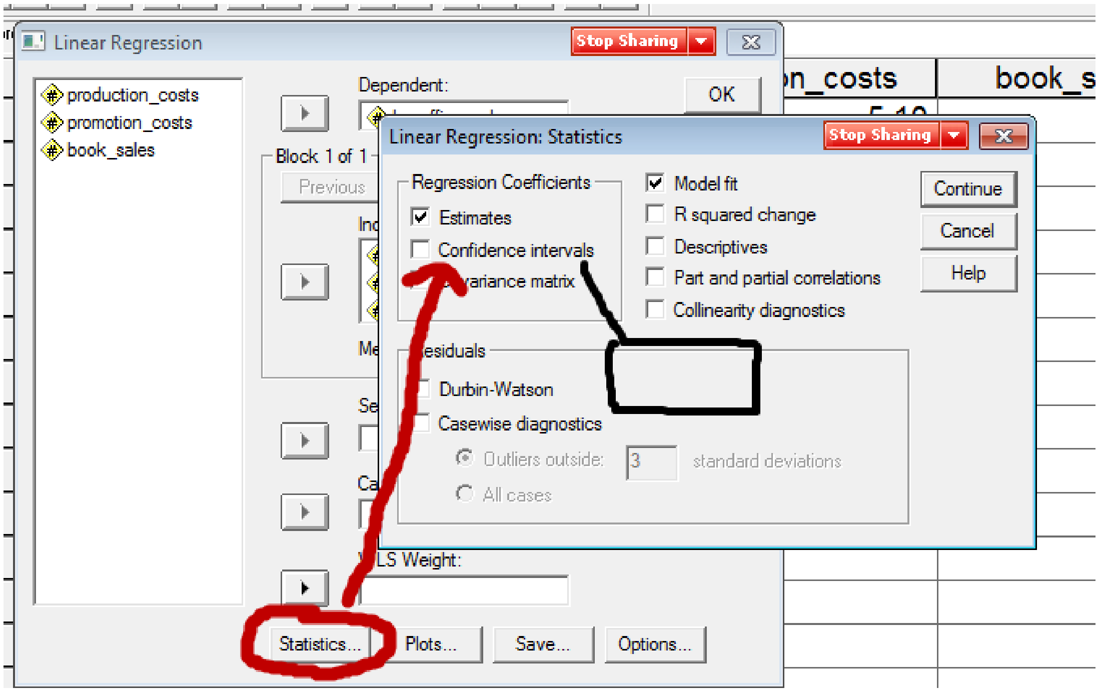
$b_i \pm$ (critical t for specified α and df)*(standard error of b_i)

where b_i is one of the regression coefficients found in a given regression equation.

Example

What is the 95% confidence interval for b1 (coefficient for production costs in millions)?

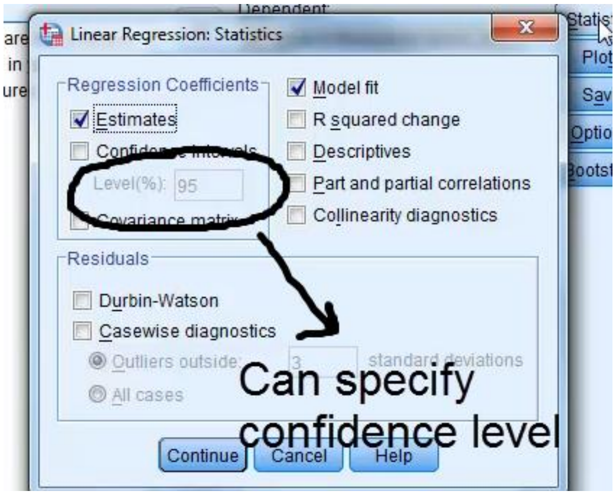
Calculate by hand and also use SPSS



Regression Coefficients
- ☑ Estimates
- □ Confidence intervals
- □ Covariance matrix

- ☑ Model fit
- □ R squared change
- □ Descriptives
- □ Part and partial correlations
- □ Collinearity diagnostics

Residuals
- □ Durbin-Watson
- □ Casewise diagnostics
  - ● Outliers outside: 3 standard deviations
  - ○ All cases



Regression Coefficients
- ☑ Estimates
- □ Confidence intervals
  - Level(%): 95
- □ Covariance matrix

- ☑ Model fit
- □ R squared change
- □ Descriptives
- □ Part and partial correlations
- □ Collinearity diagnostics

Residuals
- □ Durbin-Watson
- □ Casewise diagnostics
  - ● Outliers outside: 3 standard deviations
  - ○ All cases

Can specify confidence level

SPSS Results

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)	7.676	6.760		1.135	.299	-8.866	24.218
	production_costs	3.662	1.118	.421	3.276	.017	.927	6.397
	promotion_costs	7.621	1.657	.559	4.598	.004	3.566	11.676
	book_sales	.828	.539	.127	1.536	.175	-.491	2.148

a. Dependent Variable: boxoffice_sales

b_1 = 3.662

se b_1 = 1.118

α = .05

df = n-k-1

(where n is total sample size, and k is number of regression coefficients estimated excluding the intercept b0)

Critical t = find this using critical t-ratio table (see course web site in t-test area)

Question

What are the df for this regression?

Answer

- n = 10
- k = 3 (b1, b2, and b3)
- so n-k-1 = 10 – 3 – 1 = 6

Also can find df as the residual df in the ANOVA summary table

SPSS Results for Movie Data

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9932.463	3	3310.821	58.221	.000[b]
	Residual	341.201	6	56.867		
	Total	10273.664	9			

a. Dependent Variable: boxoffice_sales

b. Predictors: (Constant), book_sales, promotion_costs, production_costs

b_i ± (critical t for specified α and df)*(standard error of b_i)

Upper Limit = b_i + (critical t for specified α and df)*(standard error of b_i)

= 3.662 + (2.45) *(1.118)

= 3.662 + (2.45)*(1.118)

= 3.662 + (2.7391)

= 6.4011

Lower Limit = b_i - (critical t for specified α and df)*(standard error of b_i)

= 3.662 - (2.45) *(1.118)

= 3.662 - (2.45)*(1.118)

= 3.662 - (2.7391)

= 0.9229

Compare the above with SPSS generated CI.

Ours 95% CI= 0.9229 to 6.4011
SPSS 95% CI= 0.927 to 6.397

Interpretation

We can be 95% confidence (or whatever level of confidence used) that the true population slope for production costs may be small as 0.9229 or as large as 6.4011; i.e., we can be 95% confidence the population slope lies between 0.9229 and 6.4011.

1.8 APA Style Presentation with Box Office Sales

Important

Remember, do NOT use correlations reported by SPSS regression command because SPSS reports the 1-tail p-values; we want the 2-tailed p-values, so use bivariate correlation command for the correlations.

Table 1: Descriptive Statistics and Correlations for Box Office Sales Data

	Correlations			
Variable	1	2	3	4
1. Box Office Sales	---			
2. Production Costs	.917*	---		
3. Promotion Costs	.930*	.790*	---	
4. Book Sales	.475	.429	.299	---
Mean	85.24	8.74	4.90	9.92
SD	33.79	3.89	2.48	5.17

Note: All variables reported in millions of dollars except for Book Sales. n = 10
* p < .05

Table 2: Regression of Box Office Sales on Production Costs, Promotion Costs, and Book Sales

Variable	b	se b	95% CI	t
Production Costs	3.66	1.12	0.93, 6.40	3.28*
Promotion Costs	7.62	1.66	3.57, 11.68	4.60*
Book Sales	0.83	0.54	-0.49, 2.15	1.54
Intercept	7.68	6.76	-8.87, 24.22	1.14

Note: R^2 = .97, adj. $\underline{R}^2$ = .95, F = 58.22*, df = 3,6; n = 10
*p < .05.

Based upon the regression results, both production and promotion costs are statistically associated with box office sales; book sales, however, is not related to box office sales. As the regression results show, both production and promotion costs positively predict box office sales: as production or promotion costs increase, box office sales are predicted to increase as well. Book sales does not appear to predict box office sales once production and promotion costs are taken into account.

Note about Test 3 – How to determine whether to use correlation or regression

Correlation answers simple question of whether variables are **related**.

Correlation, as we have learned it, **will not address**

- **control** of variables (**partialing** effects of one IV while controlling for a second IV)
- **prediction**, ability to predict the DV from values on the IVs since no prediction equation (regression equation) is present
- also, with correlation one need not have specific IV and DV, but with regression, there must be at least one DV.

Regression addresses both concerns that correlation cannot handle.