**Multiple Regression: examples, how to handle insignificant x’s, how to compare models**

In multiple regression, the population slopes are estimated using all of the x’s at once. The procedure takes into account each x’s relationship with y, while simultaneously accounting for each x’s relationship to all the other x’s. Here’s what I mean, conceptually speaking:

The upshot of all of this is that a couple of our interpretations have to change to acknowledge the fact that the other x’s are being accounted for. Other than that, things stay pretty much the same.

**Example: Showtime Theaters**

An analyst at Showtime Theaters has been asked to analyze the relationship between weekly on-line advertising ($1000s), weekly television advertising ($1000s), and weekly gross revenue ($1000s)

* What is the Dependent Variable? (in other words, what is being explained?)
* What are the Independent Variables? (in other words, what variables might explain changes in the dependent variable?
* Note: all of the variables are expressed in thousands of dollars. Regression calculates the estimated regression equation in the same units as the variables. This means that you have to consider the units when interpreting the slopes, when plugging x values into the ERE to make predictions, and when interpreting predicted values. The equation will expect x values in the same units as the data, and will produce y values in the same units as the data.
* 15 weeks are randomly selected. Weekly on-line advertising ($1000s), weekly television advertising ($1000s), and weekly gross revenue ($1000s) are measured.
* This analysis will use an significance level

Here is the dataset:

|  |  |  |  |
| --- | --- | --- | --- |
| Week  (i) | Weekly Revenue | TV Advertising | On-line Advertising |
| 1 | 97.92 | 5 | 1.5 |
| 2 | 95.94 | 4 | 1.5 |
| 3 | 93.8 | 2.5 | 2.5 |
| 4 | 95.31 | 3 | 3.3 |
| 5 | 94.62 | 3.5 | 2 |
| 6 | 94.63 | 2.5 | 4.2 |
| 7 | 94.64 | 3.2 | 2 |
| 8 | 93.47 | 2.2 | 4 |
| 9 | 97.73 | 4.6 | 3 |
| 10 | 96.3 | 4.25 | 2.6 |
| 11 | 95.6 | 3.9 | 1.7 |
| 12 | 94.46 | 3.3 | 1.8 |
| 13 | 93.01 | 2.7 | 2.25 |
| 14 | 93.07 | 2.75 | 1.2 |
| 15 | 95.25 | 3.4 | 3.1 |

Now we need to scatterplot y versus each x to look for non-linear patterns.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |
| Multiple R | 0.9745 |  |  |  |  |  |
| R Square | 0.9497 |  |  |  |  |  |
| Adjusted R Square | 0.9413 |  | NOTE: all numbers have been rounded to | | | |
| Standard Error | 0.3618 |  | four decimal places. | | | |
| Observations | 15 |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |
| Regression | 2 | 29.6320 | 14.8160 | 113.1856 | 1.63E-08 |  |
| Residual | 12 | 1.5704 | 0.1309 |  |  |  |
| Total | 14 | 31.2024 |  |  |  |  |
|  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 99%* | *Upper 99%* |
| Intercept | 87.0304 | 0.6208 | 140.1907 | 1.16E-20 | 85.1346 | 88.9263 |
| TV advertising | 1.9365 | 0.1292 | 14.9884 | 3.92E-09 | 1.5420 | 2.3311 |
| On-line advertising | 0.5980 | 0.1163 | 5.1419 | 0.000244 | 0.2426 | 0.9534 |

What is the Regression Model we are estimating here?

What is the estimated regression equation (ERE)?

Do the hypothesis test to determine whether there is a statistically significant relationship between TV Advertising and Revenue.

Do the hypothesis test to determine whether there is a statistically significant relationship between On-line Advertising and Revenue.

Do the hypothesis test to determine whether *the overall regression model* is statistically significant (See **Ch 15: Handout #2**)

Now that we have confirmed this regression reflects real relationships in the population, we can go ahead and start using the results.

**Remember what the ERE was?**

**What is the predicted Gross Weekly Revenue if $3750 is spent on TV Advertising and $2300 is spent on On-line Advertising?**

* Note: the problem gives the spending in $s, but the Estimated Regression Equation is calculated using data in $1000s. So you must convert the $s to $1000s:
* plug the x-values into the ERE:
* **Interpretation:** For all weeks in which TV Advertising is $3750 and On-line Advertising is $2300, Gross Weekly Revenue is predicted to be $95,670 on average.

**What is the interpretation of the slope coefficient on TV advertising?**

* + NOTE: for all interpretations, we should convert the $1000s to $s. If the units were already in $s, we would leave them alone, so here we multiply 1.9365 x $1000 = $1936.50.
  + **Interpretation:** For every $1000 increase in TV Advertising, Gross Weekly Revenue is expected to increase by $1936.50 on average, holding On-line Advertising constant.

**What is the interpretation of the slope coefficient on On-line Advertising?** *(multiply 0.5980 x $1000 to convert to dollars)*

* + **Interpretation:** For every $1000 increase in On-line Advertising, Gross Weekly Revenue is expected to increase by $598 on average, holding TV Advertising constant.
  + Uh-oh

**Interpret the Coefficient of Determination,**

* + First, multiply
  + **Interpretation:** 94.97% of the variability in Gross Weekly Revenue is explained by TV Advertising and On-line Advertising

**Interpret the Standard Error of the Estimate,**

* + Again, it helps communicate this interpretation if you convert , which is in $1000s, to $s:
  + **Interpretation:** If we used the Estimated Regression Equation to predict Gross Weekly Revenue, our average error would be $361.80. Not bad!

There ends our discussion of the Showtime Theaters regression. Moving on to other important issues…

**What about insignificant Independent Variables?**

* So far, we have only dealt with models where all the IVs are significant. This is NOT always the case.
* If you do not reject the null hypothesis in a t-test for a slope coefficient, then you CANNOT conclude that the population slope on that x is different from zero. Therefore, you CANNOT conclude that the IV for that slope is related to the DV
  + This may be a substantively interesting result. Was it an IV that you were certain would help explain the DV? You may need to investigate why it does not in this case.
  + However, if you are using the model for prediction, then you should drop the insignificant IV from the analysis and re-estimate the regression model including only the significant IVs.

Let’s look at an example:

Cruise Ship Ratings

* *Condé Nast Traveler* magazine conducts an annual Reader’s Choice Survey in which they ask readers to rate small cruise ships on several criteria: Itineraries/Schedule, Shore Excursions, and Food/Dining. The readers also give an overall quality rating to each ship.
  + The variable values represent the percentage of readers who rated the cruise ship excellent or very good on each criterion.
* Research question: which (or what combination) of the three criteria explain the overall rating?
* What is the DV? What are the IVs?
* Use an α = 0.05 significance level

Here is the dataset:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ShipID | Ship | Overall Score | Itineraries/  Schedule | Shore Excursions | Food/Dining |
| 1 | Seabourn Odyssey | 94.4 | 94.6 | 90.9 | 97.8 |
| 2 | Seabourn Pride | 93 | 96.7 | 84.2 | 96.7 |
| 3 | National Geographic Endeavor | 92.9 | 100 | 100 | 88.5 |
| 4 | Seabourn Sojourn | 91.3 | 88.6 | 94.8 | 97.1 |
| 5 | Paul Gauguin | 90.5 | 95.1 | 87.9 | 91.2 |
| 6 | Seabourn Legend | 90.3 | 92.5 | 82.1 | 98.8 |
| 7 | Seabourn Spirit | 90.2 | 96 | 86.3 | 92 |
| 8 | Silver Explorer | 89.9 | 92.6 | 92.6 | 88.9 |
| 9 | Silver Spirit | 89.4 | 94.7 | 85.9 | 90.8 |
| 10 | Seven Seas Navigator | 89.2 | 90.6 | 83.3 | 90.5 |
| 11 | Silver Whisperer | 89.2 | 90.9 | 82 | 88.6 |
| 12 | National Geographic Explorer | 89.1 | 93.1 | 93.1 | 89.7 |
| 13 | Silver Cloud | 88.7 | 92.6 | 78.3 | 91.3 |
| 14 | Celebrity Xpedition | 87.2 | 93.1 | 91.7 | 73.6 |
| 15 | Silver Shadow | 87.2 | 91 | 75 | 89.7 |
| 16 | Silver Wind | 86.6 | 94.4 | 78.1 | 91.6 |
| 17 | SeaDream II | 86.2 | 95.5 | 77.4 | 90.9 |
| 18 | Wind Star | 86.1 | 94.9 | 76.5 | 91.5 |
| 19 | Wind Surf | 86.1 | 92.1 | 72.3 | 89.3 |
| 20 | Wind Spirit | 85.2 | 93.5 | 77.4 | 91.9 |

Here is the regression output:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |
| Multiple R | 0.865922 |  |  |  |  |  |
| R Square | 0.749821 |  |  |  |  |  |
| Adjusted R Square | 0.702912 |  |  |  |  |  |
| Standard Error | 1.387747 |  |  |  |  |  |
| Observations | 20 |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |
| Regression | 3 | 92.35202 | 30.78401 | 15.9847 | 4.52E-05 |  |
| Residual | 16 | 30.81348 | 1.925843 |  |  |  |
| Total | 19 | 123.1655 |  |  |  |  |
|  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| Intercept | 35.61838 | 13.23083 | 2.692075 | 0.01603 | 7.570276 | 63.66648 |
| Itineraries/Schedule | 0.110454 | 0.129662 | 0.851859 | 0.406863 | -0.16442 | 0.385325 |
| Shore Excursions | 0.244537 | 0.043357 | 5.64005 | 3.69E-05 | 0.152624 | 0.336451 |
| Food/Dining | 0.247357 | 0.062117 | 3.982137 | 0.001072 | 0.115675 | 0.379038 |

**Is this model statistically significant overall?**

(remember, we are using α = 0.05)

* The F test statistic has p-value = 0.0000452
  + 0.0000452 ≤ 0.05, so we would reject in the F test. YES, the overall model is statistically significant.

**Going from the bottom of the Coefficients Table up, is each IV statistically significant?**

* Food/Dining: p-value = 0.001072
  + 0.001072 ≤ 0.05, so we would reject in the t-test for . YES, Food/Dining has a statistically significant relationship with Overall Score.
* Shore Excursions: p-value = 0.0000369
  + 0.0000369 ≤ 0.05, so we would reject in the t-test for . YES, Shore Excursions has a statistically significant relationship with Overall Score.
* What about Itineraries/Schedules? p-value = 0.406863
  + Uh-oh! 0.406863 > 0.05, so we would **not** reject in the t-test for . NO, Itineraries/Schedule does NOT have statistically significant relationship with Overall Score.

We assume this model is being used for prediction purposes, so we should drop the Itineraries/Schedule variable by going into Excel and re-running the regression with only Food/Dining and Shore Excursions included in the x variable range. Here are the results of that:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | | |  | |  | |  | |  | |  |
|  |  |  | |  | |  | |  | |  | |
| *Regression Statistics* | | |  | |  | |  | |  | |  |
| Multiple R | 0.859345 |  | |  | |  | |  | |  | |
| R Square | 0.738474 |  | |  | |  | |  | |  | |
| Adjusted R Square | 0.707706 |  | |  | |  | |  | |  | |
| Standard Error | 1.376504 |  | |  | |  | |  | |  | |
| Observations | 20 |  | |  | |  | |  | |  | |
|  |  |  | |  | |  | |  | |  | |
| ANOVA |  |  | |  | |  | |  | |  | |
|  | *df* | *SS* | | *MS* | | *F* | | *Significance F* | |  | |
| Regression | 2 | 90.95451 | | 45.47725 | | 24.00154 | | 1.12E-05 | |  | |
| Residual | 17 | 32.21099 | | 1.894764 | |  | |  | |  | |
| Total | 19 | 123.1655 | |  | |  | |  | |  | |
|  |  |  | |  | |  | |  | |  | |
|  | *Coefficients* | *Standard Error* | | *t Stat* | | *P-value* | | *Lower 95%* | | *Upper 95%* | |
| Intercept | 45.17796 | 6.951848 | | 6.498698 | | 5.46E-06 | | 30.51084 | | 59.84508 | |
| Shore Excursions | 0.252892 | 0.041891 | | 6.036882 | | 1.33E-05 | | 0.16451 | | 0.341275 | |
| Food/Dining | 0.248189 | 0.061606 | | 4.028667 | | 0.000871 | | 0.118212 | | 0.378166 | |

Now all the variables are statistically significant, and see how the slopes changed? Even though Itineraries/Schedule was not statistically significant, it was still messing up the estimates for the other slopes.

**Model Comparisons:**

Frequently when using regression, we end up with more than one model for a particular dependent variable. We need some criteria to choose between models in that situation.

*for model comparisons:*

* When comparing two models that explain the same DV, it is tempting to choose the one with the higher Coefficient of Determination (), but this would **NOT** be the right thing to do. increases if you add additional IVs, even if the new variables do not add any substantive explanatory power to the model.
* , , is modified to account for the number of IVs in the model. It only increases if the added variables improve the model more than would be expected to happen by chance. Therefore, is a useful statistic to compare two models. The one with the **HIGHER** is **BETTER**. (NOTE: is NOT interpretable like is.)

*Standard Error of the Estimate, , for model comparisons:*

* Reminder: the Standard Error of the Estimate, , is the average error we would make when using the estimated regression equation to predict the DV.
* Two models with the same DV measured in the same units can be compared using s. The one with the **LOWER** **s** gives more accurate predictions, and therefore is the **BETTER** model.

*Illustration of this:*

Here is a summary table of some statistics from the two cruise ship regressions (one with all three IVs, one with only the two significant IVs):

NOTICE: is higher in Model 1, even though Model 1 contained an insignificant IV. This is because increases with each additional IV, even if the IV does not add ANY explanatory power to the model AT ALL.

This is why is NOT a good statistic to use when comparing two models.

|  |  |  |
| --- | --- | --- |
|  | **Model 1:**  **Three IVs** | **Model 2:**  **Two IVs** |
| Coefficient of Determination | 0.7498 | 0.7385 |
| Adjusted | 0.7029 | 0.7077 |
| Standard Error of the Estimate (s) | 1.3877 | 1.3765 |

* is useful for model comparisons. It is not affected by additional IVs like is. **HIGHER values of are BETTER**, so **Model 2** is better than Model 1
* The Standard Error of the Estimate (*s*) gives the average error made when using the regression equation to predict the DV. The lower the *s*, the more accurate the predictions are. When comparing two models, **LOWER values of the Standard Error of the Estimate are BETTER**. So, **Model 2** is better than Model 1