**How to Interpret Stats Output**

Let’s illustrate how to interpret crosstabs and regressions using a dataset of Texas school districts (I downloaded the data from <https://rptsvr1.tea.texas.gov/perfreport/snapshot/2018/index.html>). The following variables are analyzed:

* charter: a dummy variable where 0s indicate traditional school districts and 1s indicate charter school districts
* passrate: the percentage of students in the district who meet grade level standards in all subjects on the state’s standardized exams (STAAR)
* tsal: the average teacher salary in dollars
* tturn: the percentage of teachers who turned over from last year (did not return to teach this year)

**Crosstabs**

Whenever I’m looking at crosstabs, I focus on the percentages/proportions (the 2nd number listed in each cell, right below the number of observations) and first figure out whether the columns or the rows add up to 100% (which I can easily find in the “total” row/column).

* If columns add to 100%, I will focus on one row at a time and figure out what type of characteristic that row represents; we’ll call this characteristic X. Then, I interpret each percentage (P) in that row as “P% of group A have characteristic X,” substituting in the column identifier for group A.
* If the rows add up to 100%, I do the exact opposite. I focus on one column at a time, and call the characteristics represented by that *column* characteristic X. Then, I still interpret each percentage (P) in that row as “P% of group A have characteristic X,” substituting in whatever the *row* indicates for group A.

Now an example from **Stata**:

. tab grade charter, col

+-------------------+

| Key |

|-------------------|

| frequency |

| column percentage |

+-------------------+

| charter

grade | 0 1 | Total

-----------+----------------------+----------

A | 121 32 | 153

| 16.29 36.78 | 18.43

-----------+----------------------+----------

B | 334 22 | 356

| 44.95 25.29 | 42.89

-----------+----------------------+----------

C | 232 15 | 247

| 31.22 17.24 | 29.76

-----------+----------------------+----------

D | 46 11 | 57

| 6.19 12.64 | 6.87

-----------+----------------------+----------

F | 10 7 | 17

| 1.35 8.05 | 2.05

-----------+----------------------+----------

Total | 743 87 | 830

| 100.00 100.00 | 100.00

Since the columns sum to 100%, I’ll focus on one of the rows:

. tab grade charter, col

+-------------------+

| Key |

|-------------------|

| frequency |

| column percentage |

+-------------------+

| charter

grade | 0 1 | Total

-----------+----------------------+----------

A | 121 32 | 153

| 16.29 36.78 | 18.43

-----------+----------------------+----------

B | 334 22 | 356

| 44.95 25.29 | 42.89

-----------+----------------------+----------

C | 232 15 | 247

| 31.22 17.24 | 29.76

-----------+----------------------+----------

D | 46 11 | 57

| 6.19 12.64 | 6.87

-----------+----------------------+----------

F | 10 7 | 17

| 1.35 8.05 | 2.05

-----------+----------------------+----------

Total | 743 87 | 830

| 100.00 100.00 | 100.00

This row is labeled as having districts that were given an A rating (by the Texas Education Agency), so **an A rating** is our characteristic X. The first column indicates **16.29**%, and it shows observations where charter =0, so Group A is **traditional school districts** (non-charters). Thus, I say “**16.29**% of **traditional school districts** have **an A rating**” (remember, the formula is “P% of group A have characteristic X”). By comparison, moving over to the 2nd column (charter school districts), I see that “**36.78**% of **charter school districts** have **an A rating**.” Therefore, charter school districts are more likely than traditional school districts to get the top grade.

If we move down to the bottom row in the table, we can see that “1.35% of traditional school districts have an F rating” while “8.05% of charter school districts have an F rating.” So charter school districts are also overrepresented among the worst schools.

The output in **R** looks rather similar, except it shows proportions rather than percentages:

> CrossTable(data$grade, data$charter, prop.r=FALSE, prop.c=TRUE, prop.t=FALSE, prop.chisq=FALSE)

Cell Contents

|-------------------------|

| N |

| N / Col Total |

|-------------------------|

Total Observations in Table: 830

| data$charter

data$grade | 0 | 1 | Row Total |

-------------|-----------|-----------|-----------|

A | 121 | 32 | 153 |

| 0.163 | 0.368 | |

-------------|-----------|-----------|-----------|

B | 334 | 22 | 356 |

| 0.450 | 0.253 | |

-------------|-----------|-----------|-----------|

C | 232 | 15 | 247 |

| 0.312 | 0.172 | |

-------------|-----------|-----------|-----------|

D | 46 | 11 | 57 |

| 0.062 | 0.126 | |

-------------|-----------|-----------|-----------|

F | 10 | 7 | 17 |

| 0.013 | 0.080 | |

-------------|-----------|-----------|-----------|

Column Total | 743 | 87 | 830 |

| 0.895 | 0.105 | |

-------------|-----------|-----------|-----------|

With crosstabs, you can always switch the columns and rows while still showing the same info. In **Stata**:

. tab charter grade, row

+----------------+

| Key |

|----------------|

| frequency |

| row percentage |

+----------------+

| grade

charter | A B C D F | Total

-----------+-------------------------------------------------------+----------

0 | 121 334 232 46 10 | 743

| 16.29 44.95 31.22 6.19 1.35 | 100.00

-----------+-------------------------------------------------------+----------

1 | 32 22 15 11 7 | 87

| 36.78 25.29 17.24 12.64 8.05 | 100.00

-----------+-------------------------------------------------------+----------

Total | 153 356 247 57 17 | 830

| 18.43 42.89 29.76 6.87 2.05 | 100.00

This table gives us the same percentages as before, but the orientation has switched. Rows total to 100%, so we begin by focusing on one column

. tab charter grade, row

+----------------+

| Key |

|----------------|

| frequency |

| row percentage |

+----------------+

| grade

charter | A B C D F | Total

-----------+-------------------------------------------------------+----------

0 | 121 334 232 46 10 | 743

| 16.29 44.95 31.22 6.19 1.35 | 100.00

-----------+-------------------------------------------------------+----------

1 | 32 22 15 11 7 | 87

| 36.78 25.29 17.24 12.64 8.05 | 100.00

-----------+-------------------------------------------------------+----------

Total | 153 356 247 57 17 | 830

| 18.43 42.89 29.76 6.87 2.05 | 100.00

Focusing on the first column, we again see that “16.29% of traditional school districts have an A rating” while “36.78% of charter school districts have an A rating.”

Again, **R** looks very similar for crosstabs:

> CrossTable(data$charter, data$grade, prop.r=TRUE, prop.c=FALSE, prop.t=FALSE, prop.chisq=FALSE)

Cell Contents

|-------------------------|

| N |

| N / Row Total |

|-------------------------|

Total Observations in Table: 830

| data$grade

data$charter | A | B | C | D | F | Row Total |

-------------|-----------|-----------|-----------|-----------|-----------|-----------|

0 | 121 | 334 | 232 | 46 | 10 | 743 |

| 0.163 | 0.450 | 0.312 | 0.062 | 0.013 | 0.895 |

-------------|-----------|-----------|-----------|-----------|-----------|-----------|

1 | 32 | 22 | 15 | 11 | 7 | 87 |

| 0.368 | 0.253 | 0.172 | 0.126 | 0.080 | 0.105 |

-------------|-----------|-----------|-----------|-----------|-----------|-----------|

Column Total | 153 | 356 | 247 | 57 | 17 | 830 |

-------------|-----------|-----------|-----------|-----------|-----------|-----------|

Now, let’s see what happens if the percentages are calculated by rows instead of columns in the original crosstab. This time, I’ll just show **Stata**:

. tab grade charter, row

+----------------+

| Key |

|----------------|

| frequency |

| row percentage |

+----------------+

| charter

grade | 0 1 | Total

-----------+----------------------+----------

A | 121 32 | 153

| 79.08 20.92 | 100.00

-----------+----------------------+----------

B | 334 22 | 356

| 93.82 6.18 | 100.00

-----------+----------------------+----------

C | 232 15 | 247

| 93.93 6.07 | 100.00

-----------+----------------------+----------

D | 46 11 | 57

| 80.70 19.30 | 100.00

-----------+----------------------+----------

F | 10 7 | 17

| 58.82 41.18 | 100.00

-----------+----------------------+----------

Total | 743 87 | 830

| 89.52 10.48 | 100.00

Since the rows total to 100%, we focus on one column to start. Now, our interpretation is flipped from before. This column is labeled as having districts that are traditional districts (charter =0), so **being a traditional district** is our characteristic X. The first row indicates **79.08**%, and it shows observations that are A-rated, so Group A is **A-rated districts**. Thus, I say “**79.08**% of **A-rated districts** are **traditional districts**” (I changed the grammar a bit but still followed the basic formula is “P% of group A have characteristic X”). Moving down to the 2nd row (B-rated districts), I see that “**93.82**% of **B-rated districts** are **traditional districts**.” Therefore, B-rated districts are more likely to be traditional districts than A-rated districts. Conversely, A-rated districts are more likely to be charter districts than B-rated districts. We could, of course, continue down the rows considering other ratings that districts get.

**Regression**

Now, let’s look at some regressions. We want to set up a model where we predict teacher turnover based on salaries. Before learning our results, we can write our regression equation as:

In **Stata**:

. reg tturn tsal

Source | SS df MS Number of obs = 1,192

-------------+---------------------------------- F(1, 1190) = 80.74

Model | 9456.63937 1 9456.63937 Prob > F = 0.0000

Residual | 139371.613 1,190 117.119003 R-squared = 0.0635

-------------+---------------------------------- Adj R-squared = 0.0628

Total | 148828.253 1,191 124.96075 Root MSE = 10.822

------------------------------------------------------------------------------

tturn | Coef. Std. Err. t P>|t| [95% Conf. Interval]

-------------+----------------------------------------------------------------

tsal | -.000536 .0000597 -8.99 0.000 -.000653 -.000419

\_cons | 46.89468 2.859738 16.40 0.000 41.28399 52.50538

------------------------------------------------------------------------------

What we care about for now is the two coefficients, which appear in the column labeled “Coef.” Using the coefficients, we can write out the regression equation as:

The dependent variable (tturn) is always listed just to the left of “Coef.” in the Stata output. Notice that in the equation we multiply the independent variable (tsal) by its coefficient, and the constant (\_cons) is not multiplied by anything.

In **R**, the output will look like this:

> summary(lm(tturn ~ tsal, data=data))

Call:

lm(formula = tturn ~ tsal, data = data)

Residuals:

Min 1Q Median 3Q Max

-25.233 -6.591 -1.718 4.669 88.484

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 46.89468505 2.85973817 16.398 <2e-16 \*\*\*

tsal -0.00053600 0.00005965 -8.986 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.82 on 1190 degrees of freedom

(8 observations deleted due to missingness)

Multiple R-squared: 0.06354, Adjusted R-squared: 0.06275

F-statistic: 80.74 on 1 and 1190 DF, p-value: < 2.2e-16

Note that another name for the *constant* is the *intercept*, and this R output relies on the latter term. R also lists the intercept first. Because the order in which we list items we are adding does not matter (e.g., 2+3 is equivalent to 3+2), either of the following are a valid way to write the estimated regression equation:

Using the equation we just created, we can calculate the predicted value of teacher turnover when teacher salaries are $50,000:

Thus, we project turnover to be 21.9% if the average teacher salary is $50,000.

Now, if we want to interpret the slope coefficient *b* (-0.0005), we use the formula: “A one-unit increase in the **independent variable** predicts a ***b****-*unit change in the **dependent variable** (change will be an increase if ***b*** is positive, or a decrease if ***b*** is negative).”

So, a one-dollar increase in **average teacher salaries** predicts a **0.0005 percentage-point decrease** in the **teacher turnover rate**. Note that the projected change is tiny because changing salaries by $1 is a tiny change. A more meaningful interpretation might be that a $1000 increase in salary is associate with a 0.5 percentage-point decrease in teacher turnover.

Let’s practice with a couple more regressions:

**Stata**:

. reg passrate tsal

Source | SS df MS Number of obs = 1,184

-------------+---------------------------------- F(1, 1182) = 74.36

Model | 10494.2938 1 10494.2938 Prob > F = 0.0000

Residual | 166819.855 1,182 141.133549 R-squared = 0.0592

-------------+---------------------------------- Adj R-squared = 0.0584

Total | 177314.149 1,183 149.885164 Root MSE = 11.88

------------------------------------------------------------------------------

passrate | Coef. Std. Err. t P>|t| [95% Conf. Interval]

-------------+----------------------------------------------------------------

tsal | .0005821 .0000675 8.62 0.000 .0004496 .0007145

\_cons | 16.61448 3.2339 5.14 0.000 10.26965 22.9593

------------------------------------------------------------------------------

**R**:

> summary(lm(passrate ~ tsal, data=data))

Call:

lm(formula = passrate ~ tsal, data = data)

Residuals:

Min 1Q Median 3Q Max

-35.122 -7.961 -0.722 7.171 43.412

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 16.6144751 3.2339000 5.138 0.000000325 \*\*\*

tsal 0.0005820 0.0000675 8.623 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.88 on 1182 degrees of freedom

(16 observations deleted due to missingness)

Multiple R-squared: 0.05918, Adjusted R-squared: 0.05839

F-statistic: 74.36 on 1 and 1182 DF, p-value: < 2.2e-16

The predicted student pass rate when teacher salaries are $50,000 is:

We project the student pass rate to be 46.6% if the average teacher salary is $50,000.

Finally, we would say that a one-dollar increase in average teacher salaries predicts a 0.0006 percentage-point *increase* in the student pass rate. Again, the projected change is tiny because changing salaries by $1 is a tiny change.

If we have multiple independent variables, we can add subscripts to our coefficients (e.g., *b*1) when writing out an initial regression equation:

**Stata**:

. reg passrate tsal tturn

Source | SS df MS Number of obs = 1,181

-------------+---------------------------------- F(2, 1178) = 140.95

Model | 34067.0699 2 17033.535 Prob > F = 0.0000

Residual | 142357.802 1,178 120.847031 R-squared = 0.1931

-------------+---------------------------------- Adj R-squared = 0.1917

Total | 176424.872 1,180 149.512604 Root MSE = 10.993

------------------------------------------------------------------------------

passrate | Coef. Std. Err. t P>|t| [95% Conf. Interval]

-------------+----------------------------------------------------------------

tsal | .0003554 .0000645 5.51 0.000 .0002289 .0004818

tturn | -.4210002 .0300135 -14.03 0.000 -.4798862 -.3621143

\_cons | 36.37731 3.301514 11.02 0.000 29.8998 42.85481

------------------------------------------------------------------------------

**R**:

> summary(lm(passrate ~ tsal + tturn, data=data))

Call:

lm(formula = passrate ~ tsal + tturn, data = data)

Residuals:

Min 1Q Median 3Q Max

-36.258 -7.514 -1.101 6.370 39.981

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 36.37730646 3.30151427 11.018 < 2e-16 \*\*\*

tsal 0.00035536 0.00006446 5.513 0.0000000434 \*\*\*

tturn -0.42100021 0.03001353 -14.027 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.99 on 1178 degrees of freedom

(19 observations deleted due to missingness)

Multiple R-squared: 0.1931, Adjusted R-squared: 0.1917

F-statistic: 141 on 2 and 1178 DF, p-value: < 2.2e-16

The predicted student pass rate when the average salary is $50,000 and the turnover rate is 10%:

Thus, we project the student pass rate to be 50.2% if the average salary is $50,000 and the turnover rate is 10%.

Finally, we would say that a $1000 increase in teacher salary predicts a .36 (.00036 × 1000) percentage-point increase in the student pass rate. And a one percentage-point increase in teacher turnover predicts a 0.46 percentage-point *decrease* in the student pass rate.