The Stata cheat sheet for regular expressions (regex)

Core functions

```
[jkl]    = j or k or l
[^jkl]   = everything except j,k, or l
[j|l]    = j or l
[a-z]    = all lowercase letters
[a-zA-Z] = all lower and uppercase letters
[0-9]    = find all numbers

\d all numbers                    \D all non numbers
\w all alphanumeric chars.        \W all non-alphanumeric chars.
\s all spaces                     \S all non spaces

()       sub-expressions or tokens
[]       exact matches or negations
{}       define ranges
```

Special characters that require the \ escape function

```
[ \ ^ $ . | ? * + ( ) { }
```

Quantifiers and anchors

```
^    matches the beginning of a string
$    matches the end of a string
.    matches any character
|    the separator for or. Same as in Stata
?    matches zero or one instance
*    matches zero or more instances
+    matches one or more instances

^x   starts with x
z&   ends with z
\b   word boundary
```

Stata commands (v. 14+)

```
help string functions
help Unicode locale
help set locale_functions
help tockenize

ustregexm   Match a pattern
ustrregexs  Sub-expression (token) of a matched pattern
ustrregexra Replace a pattern
```

Greedy versus Possessive matching

Pattern	Greedy	Reluctant	Possessive
0 or 1	?	??	?+
0 or more	*	*?	*+
1 or more	+	+?	++
y times	{y}	{y}?	{y}+
>=y times	{y,}	{y,}?	{y,}+
>=y and <=z	{y,z}	{y,z}?	{y,z}+

Sample Stata syntax

```
gen var = ustrregexs(0) if ustrregexm("My email address is other-name123@dmail.com", "\b[a-zA-Z]+[_|\-|\.]?[a-zA-Z0-9]+@[a-zA-Z]+\.[com|net]+\b")
```