# QS101: Introduction to Quantitative Methods in Social Science

## Week 18: Linear Regression

Dr. Florian Reiche
Teaching Fellow in Quantitative Methods
Course Director BA Politics and Sociology
Deputy Director of Student Experience and Progression

February 27, 2015

Formatting

Plotting the Regression Line

Linear Regression with STATA

# Formatting

# Designing Tables . . .

- . . . requires you to think about what the reader knows and wants to know

This section draws heavily on Stimson, James A. (n.d.) Professional Writing in Political Science - A Highly opinionated Essay, available at `http://www.unc.edu/~jstimson/Working_Papers_files/Writing.pdf`

# Designing Tables . . .

- . . . requires you to think about what the reader knows and wants to know
- Tables should always be composed so that a reader can pick one up and understand its content, **without having read the text**.

This section draws heavily on Stimson, James A. (n.d.) Professional Writing in Political Science - A Highly opinionated Essay, available at `http://www.unc.edu/~jstimson/Working_Papers_files/Writing.pdf`

# Designing Tables . . .

- . . . requires you to think about what the reader knows and wants to know
- Tables should always be composed so that a reader can pick one up and understand its content, **without having read the text**.
- Professional type-setting practice in recent years has moved toward simplicity and away from extensive use of highlighting

This section draws heavily on Stimson, James A. (n.d.) Professional Writing in Political Science - A Highly opinionated Essay, available at `http://www.unc.edu/~jstimson/Working_Papers_files/Writing.pdf`

# Captions

- The reader is asking, "Why am I looking at these numbers?", and the title should answer that question.

# Captions

- The reader is asking, "Why am I looking at these numbers?", and the title should answer that question.
- Note: Do not make it too fancy, it needs to be in **plain English**

# The Stub

- The stub is the leftmost column in which you name the indicators for which coefficients will be presented.

# The Stub

- The stub is the leftmost column in which you name the indicators for which coefficients will be presented.
- Abbreviate nothing.

# The Stub

- The stub is the leftmost column in which you name the indicators for which coefficients will be presented.
- Abbreviate nothing.
- Never ever ever use computer variable names to stand for concepts (FU-BAR . . . )

# The Stub

- The stub is the leftmost column in which you name the indicators for which coefficients will be presented.
- Abbreviate nothing.
- Never ever ever use computer variable names to stand for concepts (FU-BAR . . . )
- Since interpreting an unstandardized coefficient requires us to know about the measurement of the indicators, then more information is better than less (e.g. "Income in $ thousands" instead of "Income")

# Asterisks

- Professional Consensus: varying number according to significance level

# Asterisks

- Professional Consensus: varying number according to significance level
- Basically up to you, as it is often more confusing than clarifying

# Figures

- Figures are called "Figures", not graphs and not charts, and have captions *beneath* the picture.

# Figures

- Figures are called "Figures", not graphs and not charts, and have captions *beneath* the picture.
- Coloured figures are not acceptable

# Figures

- Figures are called "Figures", not graphs and not charts, and have captions *beneath* the picture.
- Coloured figures are not acceptable
- Make sure to define and distinguish lines

# Figures

- Figures are called "Figures", not graphs and not charts, and have captions *beneath* the picture.
- Coloured figures are not acceptable
- Make sure to define and distinguish lines
- 50 shades of grey . . .

# Figures

- Figures are called "Figures", not graphs and not charts, and have captions *beneath* the picture.
- Coloured figures are not acceptable
- Make sure to define and distinguish lines
- 50 shades of grey . . .
- Do not refer to particular colours

Sample Word Document

# Plotting the Regression Line

# binscatter

- There is a very helpful command for this, install:

# binscatter

- There is a very helpful command for this, install:
- **`ssc install binscatter`**

# `binscatter`

- There is a very helpful command for this, install:
- **`ssc install binscatter`**
- If you need help with this package, type: **`help binscatter`**

# `binscatter` (contd.)

- Example: `scatter c_fimngrs_dv c_tvhours`

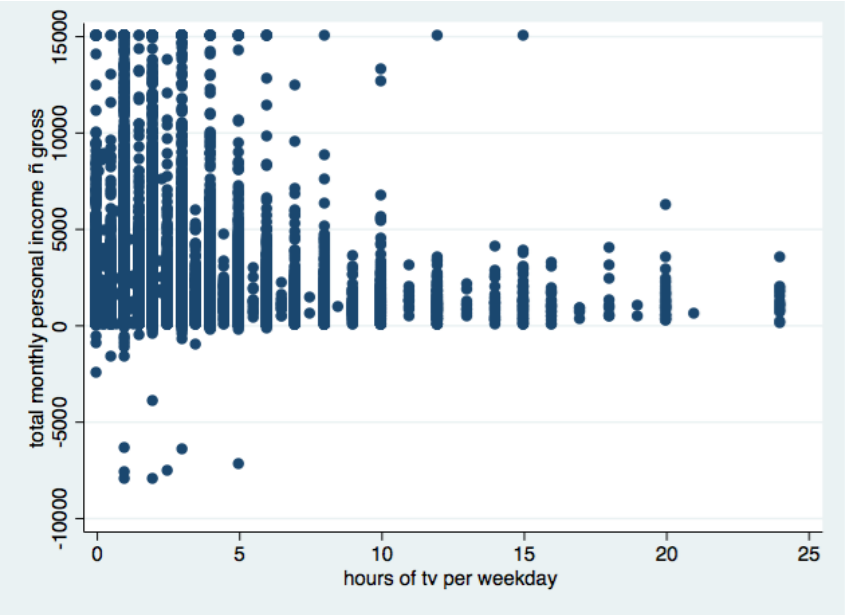


Figure: Scatter Plot: personal monthly income (gross), and hours of tv watched daily

# `binscatter` (contd.)

- Example: **`binscatter c_fimngrs_dv c_tvhours`**

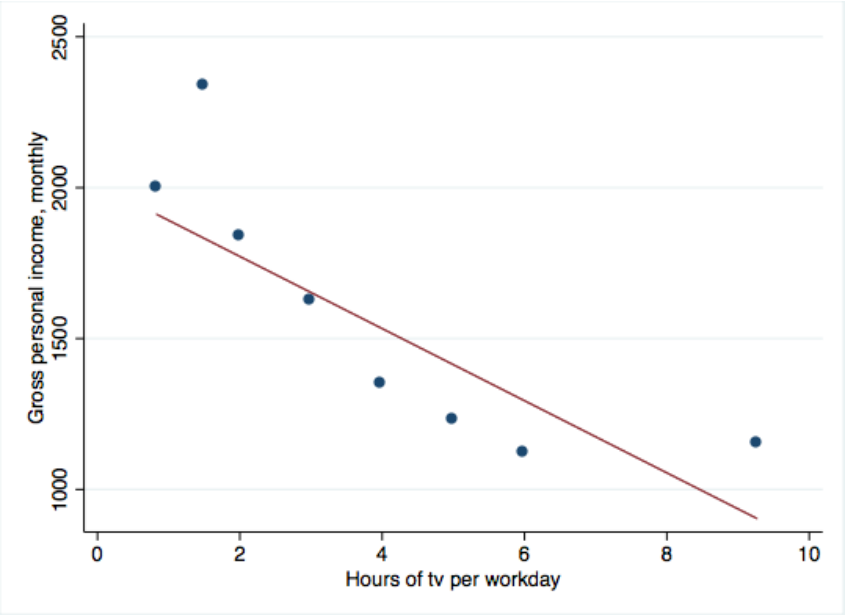


Figure: Linear Relationship between gross personal monthly income and tv hours watched daily

# Task

- We need to stick to continuous variables for now for our analysis.

# Task

- We need to stick to continuous variables for now for our analysis.
- These are:

# Variable Selection for Today

- We need to stick to continuous variables for now for our analysis.
- These are:

| Variable Name | Label | Wave |
|---|---|---|
| jbhrs | no. of hours normally worked per week | ABC |
| jbttwt | minutes spent travelling to work | ABC |
| netuse | frequency of using the internet | ABC |
| nch14 | resp number of children under 15 resp is responsible for | ABC |
| nnatch | number of biological children in household | ABC |
| nadoptch | number of adoptive children in household | ABC |
| nmar | number of marriages | A - - |
| volfreq | frequency of volunteering | - B - |
| volhrs | hours spent volunteering in last 4 weeks | - B - |
| charfreq | frequency donated to charity | - B - |
| charam | amount given to charity last 12 months | - B - |
| howlng | hours per week on housework | - B - |
| tvhours | hours of tv per weekday | - - C |

# Task

- Perform some analysis with regards to how these variables are related to personal income

# Task

- Perform some analysis with regards to how these variables are related to personal income
- Use the **`binscatter`** command to do so

# Task

- Perform some analysis with regards to how these variables are related to personal income
- Use the **`binscatter`** command to do so
- Compare your results to correlation coefficients (use **`pwcorr`**)

# Task

- Perform some analysis with regards to how these variables are related to personal income
- Use the **`binscatter`** command to do so
- Compare your results to correlation coefficients (use **`pwcorr`**)
- Make sure to recode the variables as it is necessary

# Linear Regression with STATA

# The Basic Command in Stata

- **regress** *depvar indepvar*

# The Basic Command in Stata

- **regress** *depvar indepvar*
- Example: **regress c_fimngrs_dv c_tvhours**

# Example

```
. regress c_fimngrs_dv c_tvhours
```

| Source | SS | df | MS |
|---|---|---|---|
| Model | 2.8824e+09 | 1 | 2.8824e+09 |
| Residual | 1.2168e+11 | 45650 | 2665466.72 |
| Total | 1.2456e+11 | 45651 | 2728548.88 |

| | |
|---|---|
| Number of obs | = 45652 |
| F( 1, 45650) | = 1081.40 |
| Prob > F | = 0.0000 |
| R-squared | = 0.0231 |
| Adj R-squared | = 0.0231 |
| Root MSE | = 1632.6 |

| c_fimngrs_dv | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| c_tvhours | -119.5112 | 3.63426 | -32.88 | 0.000 | -126.6344 | -112.388 |
| _cons | 2011.039 | 13.34122 | 150.74 | 0.000 | 1984.89 | 2037.188 |

Figure: Relationship between personal gross monthly income and tv hours watched daily

# Interpretation of Stata Output

- Top left-hand is an ANOVA

# Interpretation of Stata Output

- Top left-hand is an ANOVA
- The F-Value and the p-value are reported in the section on the top right

# Interpretation of Stata Output

- Top left-hand is an ANOVA
- The F-Value and the p-value are reported in the section on the top right
- What do these tell us here?

# Interpretation of Stata Output (contd.)

- The first column lists the outcome variable *c_fimngrs_dv*

# Interpretation of Stata Output (contd.)

- The first column lists the outcome variable *c_fimngrs_dv*
- This is followed by the independent variable *c_tvhours*, and the constant *_cons*

# Interpretation of Stata Output (contd.)

- The first column lists the outcome variable *c_fimngrs_dv*
- This is followed by the independent variable *c_tvhours*, and the constant *_cons*
- The equation could be written as:

# Interpretation of Stata Output (contd.)

- The first column lists the outcome variable *c_fimngrs_dv*
- This is followed by the independent variable *c_tvhours*, and the constant *_cons*
- The equation could be written as:

Gross, personal monthly income = 2011.039 - 119.5112 (hours)

# Using the Equation

Gross, personal monthly income = 2011.039 - 119.5112 (hours)

- A person not watching any telly would earn £2011.04

# Using the Equation

Gross, personal monthly income = 2011.039 - 119.5112 (hours)

- A person not watching any telly would earn £2011.04
- A person watching 1 hour of telly would earn $2011.039 - 119.5112 \times 1 = 1891.53$

# Using the Equation

Gross, personal monthly income = 2011.039 - 119.5112 (hours)

- A person not watching any telly would earn £2011.04
- A person watching 1 hour of telly would earn $2011.039 - 119.5112 \times 1 = 1891.53$
- A person watching 2 hours of telly would earn $2011.039 - 119.5112 \times 2 = 2011.039 - 239.02 = 1772.02$

# Significance

- `Std. Err.` gives you the standard error

# Significance

- `Std. Err.` gives you the standard error
- $t$ is calculated by dividing the regression coefficient by its standard error

# Significance

- `Std. Err.` gives you the standard error
- $t$ is calculated by dividing the regression coefficient by its standard error
- Test of significance will be covered next week

# Confidence Intervals

- This is the range within which the true slope is contained with 95% certainty

# Confidence Intervals

- This is the range within which the true slope is contained with 95% certainty
- If it does not contain zero, the slope will be statistically significant

# Graphical Depiction

- We can show our regression line in graphical form with and without confidence intervals

# Graphical Depiction

- We can show our regression line in graphical form with and without confidence intervals
- Type: **`twoway (lfitci c_fimngrs_dv c_tvhours)`**

# Output

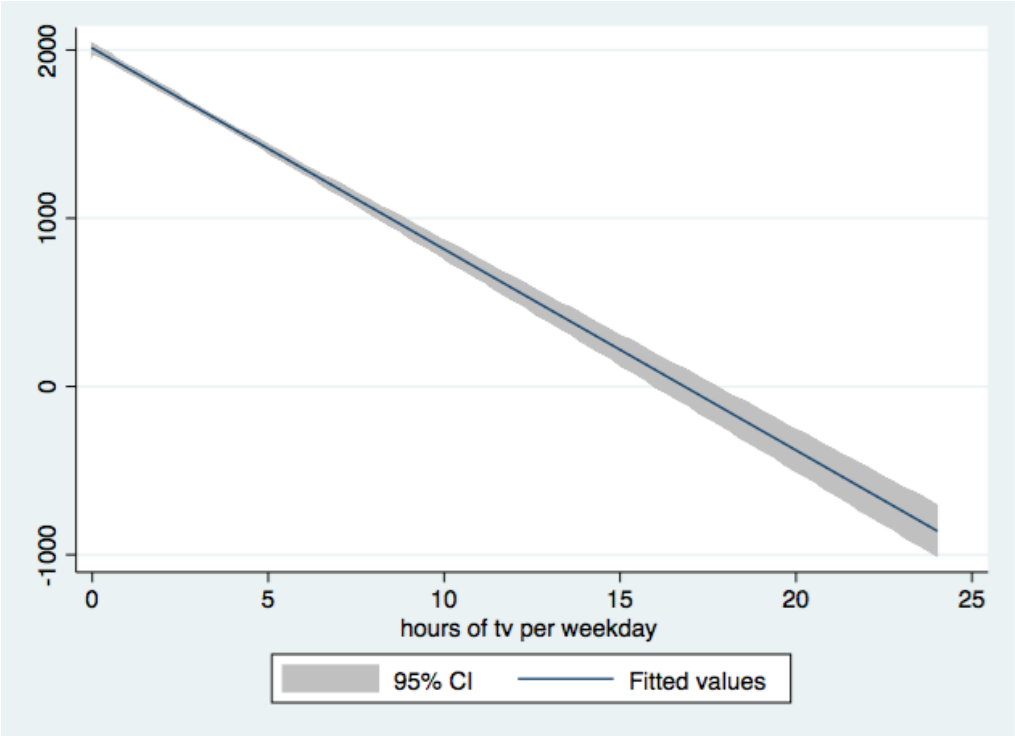


Figure: Predicted Relationship between gross personal monthly income and tv hours watched daily

# Your Turn

- Perform 2 individual linear regression analyses, choosing one predictor from the list each
- Interpret the output
- Compose at least one regression prediction plot