Centering

Some researchers transform $X$ and/or $Y$ to center them.

$$X_{\text{centered}} = X_i - \overline{X}$$

$$Y_{\text{centered}} = Y_i - \overline{Y}$$

Note that $\overline{X}_{\text{centered}} = \emptyset$

$\overline{Y}_{\text{centered}} = \emptyset$

[Some researchers simple regard this as more user/reader-friendly: they see $\emptyset$ as a natural neutral point for their variable.]

If we center $X$, then the line will cross the $Y$ axis. (and, of course, $x=0=\overline{x}$) will be on line.

If we center $Y$, then the line will cross the $X$ axis.

If we center both, the line will go through the origin.

The slope doesn't change - only the intercept (which usually isn't of primary interest)
Standardized Estimates

Recall centering - some authors subtract $x$ from $\overline{x}$ and/or $y$ from $\overline{y}$. This has no effect on $b$ - it just shifts $a$ (it's as if you are moving the line around a 2-dimensional space w/out altering the slope.)

"Standardized estimates" involve a different sort of transformation. We can standardize $x$ and $y$

$$X_{\text{stand}} = \frac{X_i - \overline{X}}{S_x}$$

$$Y_{\text{stand}} = \frac{Y_i - \overline{Y}}{S_y}$$

$X_{\text{stand}}$ is number of standard deviation units $X_i$ is from $\overline{X}$. $Y_{\text{stand}}$ is number of standard deviation units $Y_i$ is from $\overline{Y}$. 
Standardized variables (that is, $X_{\text{stand}}$ and $Y_{\text{stand}}$) are really unit-free (but we talk about them in terms of 'standard deviation units').

We can run an OLS regression....

$$Y_{\text{stand}} = \alpha + \beta X_{\text{stand}} + \varepsilon_i$$

estimated by

$$Y_{\text{stand}} = a + b^* X_{\text{stand}} + \varepsilon_i$$

$b^*$ is the amount of change in $Y$ (in standard deviation units) associated w/a one-standard deviation-unit change in $X$.

Researchers often use standardized estimates in multivariate regression to compare the effects of independent variables. For example,

$$Y_{\text{stand}} = a + b_{1} X_{1\text{stand}} + b_{2} X_{2\text{stand}} + \varepsilon$$
both \( x_{1\text{stand}} \) and \( x_{2\text{stand}} \) are standardized measured in standard deviation units. So, some researchers believe that \( b_{1}^{*} \) and \( b_{2}^{*} \) are directly comparable. 

\[ b_{1}^{*} \text{ and } b_{2}^{*} \text{ are "standardized coefficients", commonly called beta-weights, which is sometimes shortened to 'beta's' - not to be confused w/ the population slope } \beta. \]

Note

\[
\text{Standardized estimate} = \text{beta weight} = \frac{b}{\frac{S_{x}}{S_{y}}} \]  

(bivariate case)

\[ \text{unstandardized coefficient} \]

also recall in bivariate case:

\[
\text{corr}(x, y) = b \cdot \frac{S_{x}}{S_{y}}
\]

So, in the bivariate case,

\[
\text{beta weight} = \text{corr}(x, y)
\]
So, beta weights have the same limitations as correlations.

* Many scholars argue that we can't compare beta-weights across variables, even in a single equation. Beta-weights are based on standard deviations of $X$ and $Y$ — which vary across independent variables.

* So, it might be more useful to compare (for instance) one year of education w/ $1,000 of income when predicting donations to interest groups —

$$\hat{Y} = a + b_1 x_1 + b_2 x_2$$

You can compare the change in $Y$ caused by going from low-income to high-income with the change in $Y$ associated w/having a college degree (or beyond). You don't need standardized estimates.
* Standardized estimates lack units — though we speak casually about "standard deviation units" — those really aren’t units, in the same sense that "years" or "miles" or "$1,000 dollar" are.

* Standardized estimates can’t be compared across samples.

(see Gary King article for more discussion)
Regression Forced Through Origin

Suppose we believe that when $X$ is $\emptyset$, $Y$ should be $\emptyset$.

OLS/SPSS (and other packages) will allow you to force the line through the origin. Why would one do that? (Advantages?)

* Might be more compatible w/theory
e.g.

$$X = \# \text{ of accidents}$$
$$Y = \text{money lost due to accidents}$$

when $X = \emptyset$, $Y$ should be $\emptyset$.

* If compatible w/theory, model is more parsimonious if you don't estimate $\alpha$ - less collinearity. (Particularly helpful if you have a small sample size.)

Disadvantages

* Throwing away information. You can test whether $\alpha \neq \emptyset$

$$Y = a + bX$$

$\Leftrightarrow$ if $\alpha$ is significantly different from $\emptyset$, you may have a problem.
It could be a sign of mismeasurement. Have you left a variable out? That could bias your estimates, including your intercept estimate.

* If $\emptyset$ is not part of the range of $X$, you are extrapolating — danger!!

$Y = \text{money spent on education}$

$X = \text{number of children in school}$

Once there's just one child, the amount of money is fairly substantial.

If you forced the line through the origin, you wouldn't be estimating the true slope.

The relationship is non-linear — a big jump between $\emptyset$ children and 1 child, then a smaller effect as you add children, but consistent.
but presumably it doesn't make sense to think about the case w/ 0 children. It's not only not in your sample - it's not even part of the population to which you're generalizing. There are no schools with no children, no countries with 0 people, etc.

(see graphs)

* also - the $R^2$ loses meaning - it can be negative or greater than 1. In a very real sense, you are no longer fitting a line based on minimizing the sum of the squared errors.
Functional Transformations of $X$ (non-linear relationships)

Analogy to standardization -
we transformed our variables to standardize them.

$$X_{stand} = \frac{X_i - \overline{X}}{S_x} \quad \text{and} \quad Y_{stand} = \frac{Y_i - \overline{Y}}{S_y}$$

and computed a new regression equation

$$Y_{stand} = a^* + b^* X_{stand} + e^*_i$$

Where $\beta$- weight

What we’re about to do now also involves transformations (but doesn’t involve standard deviations of $X$ and $Y$ - so we won’t have the same problems associated w/standardized estimates.)

Previously, we thought of our data as

$$\{y_1, y_2, \ldots, y_n\} \quad \text{and} \quad \{x_1, x_2, \ldots, x_n\}$$

But more generally we can think of it as
\{y_1, y_2, \ldots, y_n\}$ and $\{f(x_1), f(x_2), \ldots, f(x_n)\}$

$f(x)$ is just a function of $x$ ....

It might mean

$f(x) = x$

or

$f(x) = x^2$

etc.

We transform $X$ (into some function of $x$) when we expect non-linear relationships.

If we observe a relationship between two variables $y$ and $X$ such as

and we can find a function $f(x)$ that exactly duplicates this relationship
then the plot of \( Y \) and \( f(x) \) will be a straight line.

\[
\begin{align*}
Y \quad &\quad f(x) \\
\end{align*}
\]

So, we rerun our regression - no longer using \( X \), but using \( f(x) \).

That is, we're **not** estimating

\[
y = \alpha + \beta x + \epsilon
\]

but rather

\[
y = \alpha + \beta f(x) + \epsilon
\]

[and \( f(x) \) might be \( x^2, \sqrt{x} \)]

etc.

Let's think about an example. Many relationships that we discuss in the social sciences look like this \( \downarrow \)
In other words, \( X \) has a very pronounced effect on \( Y \) at lower levels of \( X \). Population and income are two examples of independent variables that often have this relationship to dependent variables.

For instance

\[
Y_{X} = \text{dollars spent on luxury items}
\]

\[X = \text{income}\]

As income changes in $10,000 increments, there's likely a more marked effect on spending if one goes from $10,000 to $30,000 (3 unit change) than if there's an increase from $5,000,000 income.
to $5,030,000. The change in absolute terms is the same - 3 of those $10,000 units - but that change likely makes a bigger difference at lower levels of income.

Another example of the same relationship is population. Adding 1,000 residents to Grand Isle will have a bigger effect on all sorts of socio-political phenomena - demand than adding 1,000 residents to NYC.

One common function $f(x)$ that captures this relationship is the log

\[ \log(x) \]

So, in our syntax for OLS, we compute a new variable that = \( \log \) of \( x \)

\[ \text{In SPSS, it's } \log x = \ln(x). \]

or any name you choose
Then you rerun your regression analysis with the new, transformed variable $f(x)$. That is,

$$ Y = a + b \log(x) + e $$

Note $b$ is no longer the change in $Y$ associated with a one-unit change in $X$. It's the change in $Y$ associated w/a one-unit change in log of $X$.

Another common relationship: $X^2$

Here, this is useful if we expect change to increase across $X$—that is, maybe more pronounced changes at higher values of $X$.

A common way to handle expected curvilinear relationships is to incorporate $X$ and $X^2$ into the equation.
How, in a paper, should transformations be interpreted?

Well, a single variable transformation such as log of $x$ or $x^2$ is fairly well known—so, for instance, for log of $x$:

"there are more substantial/ marked/ pronounced changes in $y$ associated w/income change for low-income individuals"

(or, associated w/population increases in rural areas," etc.)

But including two variables (e.g. $x$, $x^2$) or a less well known transformation can be more complicated

But you can still explain to your reader the association or relationship between $x$ and $y$.

For instance, $y = 3 + 2x^2 + e_i$
\[ Y = 5 + 2x^2 + e \]

Pick a few \( x \)'s and demonstrate the relationship b/w \( x \) and \( y \)

<table>
<thead>
<tr>
<th>As ( x ) changes from</th>
<th>to</th>
<th>( \hat{y} ) changes from</th>
<th>to</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 ( \text{2 units} )</td>
<td>3</td>
<td>7</td>
<td>23</td>
</tr>
<tr>
<td>4 ( \text{2 units} )</td>
<td>6</td>
<td>37</td>
<td>77</td>
</tr>
<tr>
<td>7 ( \text{2 units} )</td>
<td>9</td>
<td>103</td>
<td>167</td>
</tr>
<tr>
<td>10 ( \text{2 units} )</td>
<td>12</td>
<td>205</td>
<td>293</td>
</tr>
</tbody>
</table>

A change in \( \hat{y} \) of 16 \( \leftarrow \) smaller change

40

44

88 \( \leftarrow \) bigger change

These sorts of tables are really useful to help a researcher understand his/her results. (and can easily be done in excel)

Note \( \Rightarrow \) of course, with multivariate regression, things are more complicated. But you can still compute \( \hat{y} \) - just alter the value of one independent variable, as above, and hold everything else at its mean.]
In an actual paper, you might just list hypothetical X's, and then \( \hat{y} \)’s (predicted y’s) associated w/ those X’s (holding other variables at their mean). [we’ll do more of this in Problem Set 2]

Or, you can present your results graphically [see the State Court paper I posted for an example— that is logit, not OLS, but it’s the same principle.]

How do you know if using a transformation helped produce a better fit of the data? Compare \( R^2 \)!

\[ R^2 \] from \( \hat{y} = a + b \times \) versus \[ R^2 \] from \( \hat{y} = a + b \cdot f(x) \)

Remember, \( R^2 \) is a measure of linear fit of data! Higher means better fit of data to