Regression Analysis - Intro

If we determine that two variables $X$ and $Y$ are correlated, we can use the value of one variable, $X$, to "predict" the corresponding value of the other variable, $Y$.

**EX:** Use height ($X$) to predict weight ($Y$)

Use education level ($X$) to predict income ($Y$)
We use a "regression equation" to make predictions

\[ y = b_0 + b_1 x \]

\[ \uparrow \]
\[ y \text{-intercept} \]

\[ \text{slope (regression coefficient)} \]

Book uses:

\[ y = b + m x \]
Computing a Regression Equation

Given a set of bivariate data \((X, Y)\) we can compute \(b_0\) and \(b_1\) as:

\[
\begin{align*}
    b_1 &= \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}} \\
    b_0 &= \frac{SS(xy)}{SS(x)}
\end{align*}
\]
and
\[ b_0 = \overline{Y} - b_1 \overline{X} \]

Ex: (cont. from ch 3.2 notes)
\[ X = \text{hours of TV per day} \]
\[ Y = \text{cholesterol level} \]

From Sec 3.2 we computed
\[ SS(xy) = \sum xy - \frac{(\sum x)(\sum y)}{n} \]
\[ = 90.4 \]
\[ SS(x) = \sum x^2 - \frac{(\sum x)^2}{n} \]
\[ = 4.97 \]
So \[ b_1 = \frac{SS(xy)}{SS(x)} = \frac{90.44}{4.97} = 18.2 \]

Next \[ \bar{X} = \frac{\sum X}{n} = \frac{18.5}{8} \]
\[ = 2.31 \]
\[ \bar{Y} = \frac{\sum Y}{n} = \frac{1597}{8} \]
\[ = 199.63 \]

So \[ b_0 = \bar{Y} - b_1 \bar{X} \]
\[ = 199.63 - (18.2)(2.31) \]
\[ = 199.63 - 42.04 \]
\[ = 157.59 \]

So our regr. eq. is

\[ y = b_0 + b_1 x \]

\[ y = 157.6 + 18.2 x \]
Plotting the Regression Line on the Scatter Plot

Ex: (cont.)

\[ y = 157.6 + 18.2 \times x \]
Notes:
1) Regr. line is the "line of best fit" in the sense that this line is the line which is closest to all data points simultaneously.

2) The regr. line \( y = b_0 + b_1 x \) passes thru \((\bar{x}, \bar{y})\).

3) Don't predict \( y \)-values for \( x \)'s outside of the range of the data.
Predicting Values of $y$

Ex: (cont.)

Predict the cholesterol level ($y$) for a person who watches 2 hours TV a day ($x$).

For $x=2$ the predicted $y$ is

\[ \hat{y} = 157.6 + 18.2 \times 2 \]

\[ \hat{y} = 157.6 + 18.2 (2) \]

\[ \hat{y} = 194.0 \]
Interpretation:
The predicted $y$-value is the mean cholesterol level of all persons who watch 2 hours TV a day.
Residuals

For each observed y-value, we can compute a residual - the difference between the observed y-value and the predicted y-value.

residual = y - \hat{y}

Ex: (cont.)
For x = 2.5 the observed y = 200 and \hat{y} = 157.6 + 18.2 x
= 157.6 + 18.2 (2.5)
= 203.1

Thus, the residual for this obs.
residual = y - \hat{y} \\
= 200 - 203.1 \\
= -3.1

Note: A residual may be positive or negative (or zero)

Interpretation: The observed y-value for x = 2.5 is 3.1 units less than the predicted y-value for x = 2.5