Stat 13, Tue 5/29/12.
1. Drawing the reg. line.
2. Making predictions.
3. Interpreting b and r.
4. RMS residual.
5. r^2.
6. Residual plots.

Final exam is Thur, 6/7, in class.
Hw7 is due Tue, 6/5, and is from the handout, which is from "An Introduction to the Practice of Statistics" 3rd ed. by Moore and McCabe.
Problems 2.20, 2.30, 2.38, 2.44, 2.46, and 2.102.
On 2.20, you don't need to make the scatterplot on the computer, and it doesn't have to look perfect. On 2.102, the last sentence should be "In particular, suggest some other variables that may be confounded with heavy TV viewing and may contribute to poor grades."

1. Drawing the regression line.

A couple more useful facts about the least squares regression line:

(a) the reg. line always goes through ($\bar{x}$,$\bar{y}$).

(b) the sum of the residuals always = 0.

Fact (a) is useful for plotting the regression line.

Put one point at ($\bar{x}$, $\bar{y}$).

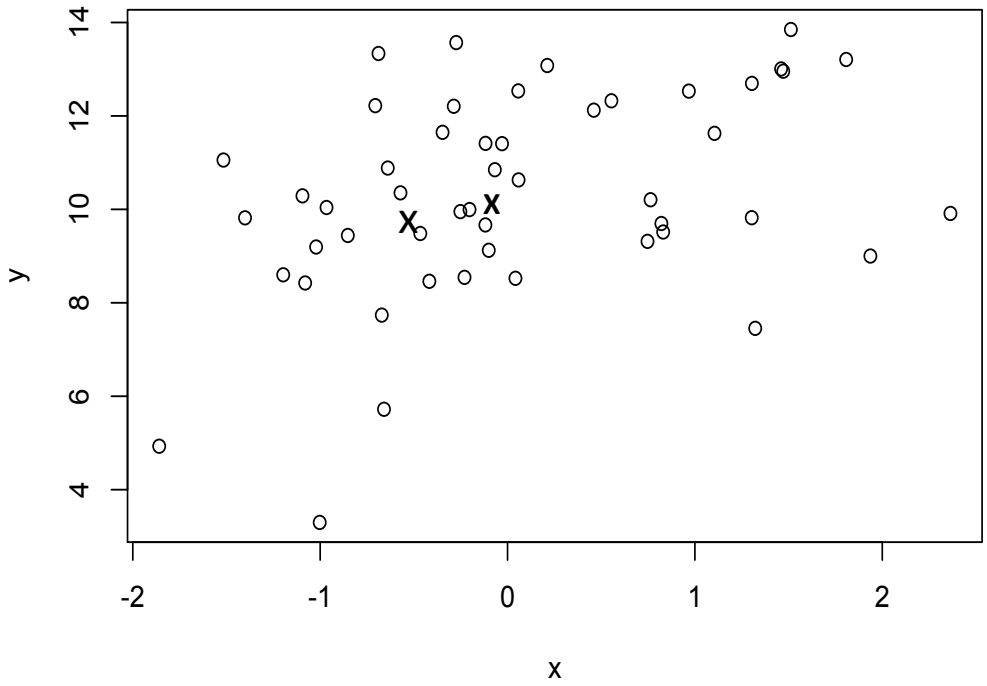
Choose some number c so that $\bar{x}$ + c is on the graph,

and put another pt. at

($\bar{x}$ + c, $\bar{y}$ + bc),

where b is the regression slope.

Connect the 2 points.



1. Drawing the regression line.

A couple more useful facts about the least squares regression line:

(a) the reg. line always goes through ($\bar{x}$,$\bar{y}$).

(b) the sum of the residuals always = 0.

Fact (a) is useful for plotting the regression line.

Put one point at ($\bar{x}$, $\bar{y}$).

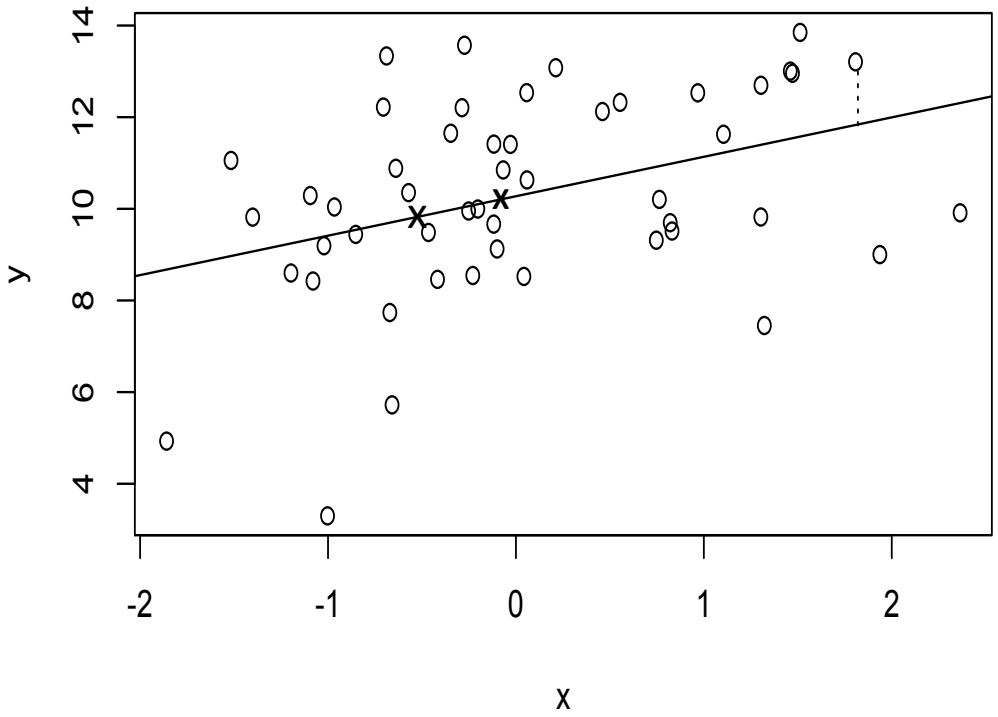
Choose some number c so that $\bar{x}$ + c is on the graph,

and put another pt. at

($\bar{x}$ + c, $\bar{y}$ + bc),

where b is the regression slope.

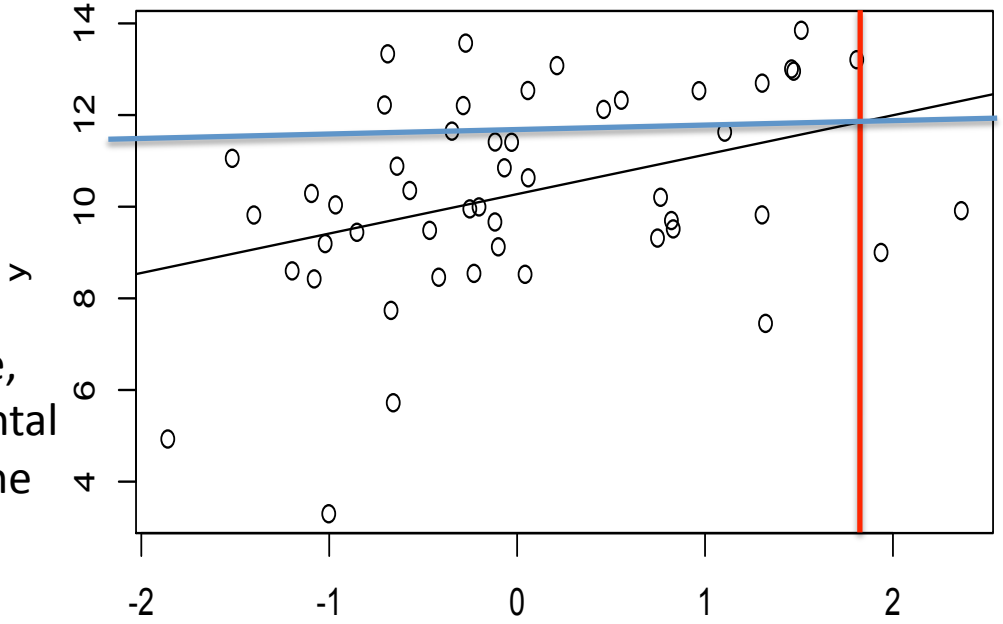
Connect the 2 points.



2. Making predictions.

Given that the regression line is $\hat{y} = a + b\,X_i$, where you've found *a* and *b*, to make a prediction of Y for a given value of X, just plug X in for X_i in the above equation. Graphically, you can also make this prediction using the *up and over line*, i.e. by drawing the corresponding vertical line on the scatterplot, seeing where it intersects the regression line, and then drawing the corresponding horizontal line and seeing where it intersects the y axis.

For instance, on this scatterplot, to make a prediction for X = 1.8, draw the red vertical line X = 1.8, see where it intersects the reg. line, and then draw a horizontal line from there to see the corresponding y value. Here, $\hat{y}$ = 11.5.



Note that, for pure prediction, you don't really care about confounding factors. If it works, it works. Confounding is a problem when trying to infer causation though. e.g. ice cream sales and crime.

3. Interpreting b and r.

The correlation, r, tells you the strength of the linear relationship between X and Y. If the points fall exactly on a line sloping up, then r is exactly 1, and if the line slopes down, then r = -1.

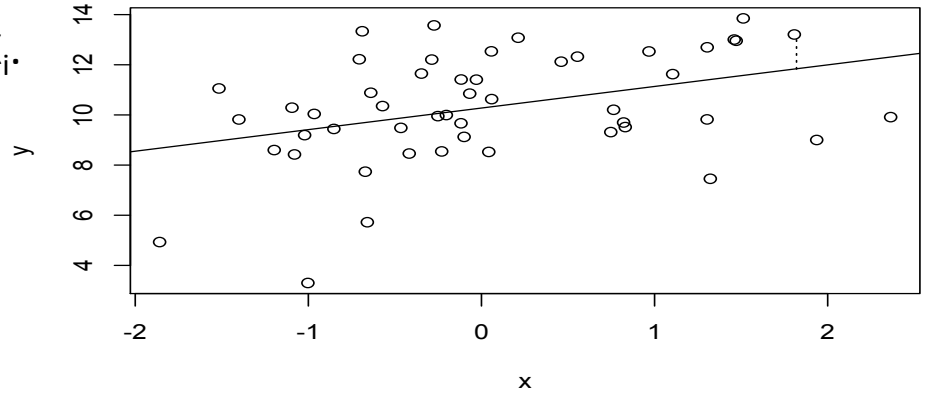
The slope, b, of the regression line tells you how much your predicted y value, $\hat{y}$, would increase if you increase X by one unit. Note that neither tells you anything about causation, especially if the data are observational rather than experimental. But if you are interested in prediction rather than in causation, b is extremely important.

To say that the relationship between X and Y is causal would mean that if you increase X by one unit, and everything else is held constant, then you actually increase that person's Y value by 1 unit.

The regression line is $\hat{y} = a + b\,X_i$.

$b = rs_y/s_x$.

$a = \bar{y} - b\bar{x}$.



Now we will talk about ways to assess how well the regression line fits.

4. RMS residual.

RMS means *root mean square*. The RMS of a bunch of numbers is found by taking those numbers, squaring them, taking the mean of those squares, and then taking the square root of that mean.
e.g. RMS{-1,2,-3,4,10} = $\sqrt{}$ mean $\{(-1)^2, 2^2, (-3)^2, 4^2, 10^2\}$ = $\sqrt{}$ mean{1,4,9,16,100} = $\sqrt{}$ 26 ~ 5.10.

The RMS of some numbers indicates the *typical size* (or abs. value) of the numbers. For instance, σ is the RMS of the deviations from μ.

Thus the RMS of the residuals, $Y_i - \hat{y}_i$, tells us the typical size of these residuals, which indicates how far off our regression predictions would typically be. Smaller RMS residual indicates better fit.

It turns out that the RMS of the residuals = $s_y\sqrt{(1-r^2)}$. Very simple!

Note that the sum of the residuals = 0. Therefore,
SD of residuals = RMS of (residuals - mean residual) = RMS of residuals.

Suppose the RMS residual = 10. Then the +/- for your reg. prediction is 10.

Another way to assess how well the regression line fits.

5. r^2.

r^2 is simply the correlation squared, and it indicates the proportion of the variation in Y that is *explained* by the regression line. Larger r^2 indicates better fit.

The justification for that sentence stems from the fact that, as mentioned on the previous slide, SD of residuals = RMS of residuals = $s_y\sqrt{(1-r^2)}$.
Therefore, squaring both sides,
$Var(e_i) = Var(Y_i)\ (1-r^2)$.
$= Var(Y_i) - r^2\ Var(Y_i)$.
So, $r^2\ Var(Y_i) = Var(Y_i) - Var(e_i)$.
$r^2 = [Var(Y_i) - Var(e_i)] / Var(Y_i)$.

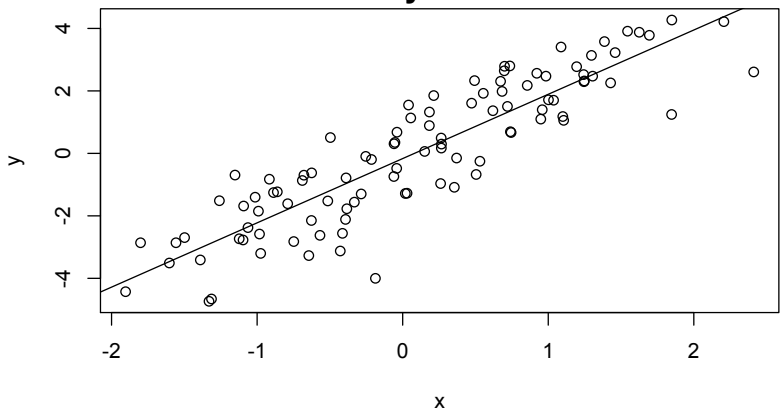
$Var(Y_i)$ is the amount of variation in the observed y values, and $Var(e_i)$ is the variation in the residuals, i.e. the variation left over after fitting the line by regression. So, $Var(Y_i) - Var(e_i)$ is the variation *explained by* the regression line, and this divided by $Var(Y_i)$ is the proportion of variation explained by the reg. line.

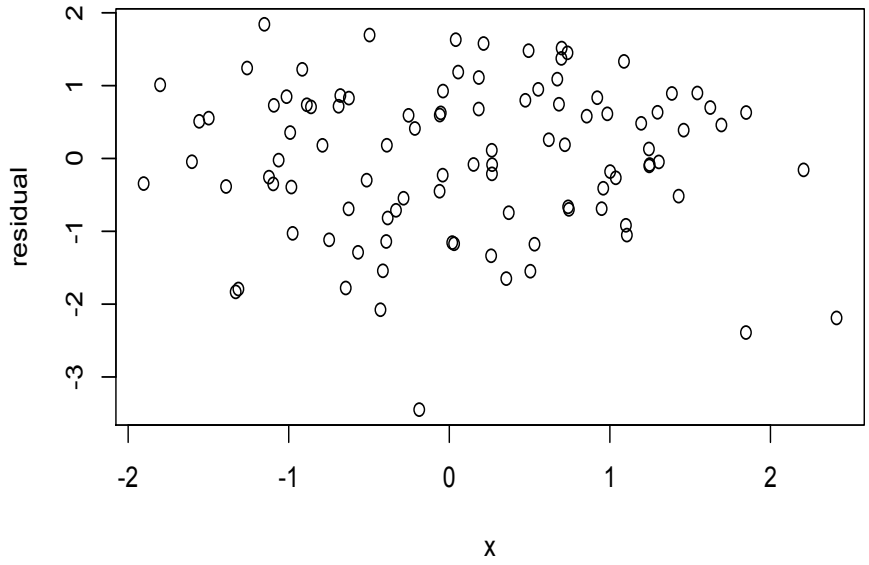
e.g. if r = -0.8. Then $r^2 = 0.64$, so we'd say that 64% of the variation in Y is explained by the regression line, and 36% is left over, or residual variation, after fitting the regression line. If instead r = 0.10, then $r^2 = 0.01$, so only 1% of the variation in Y is explained by the regression line.

Another way to assess how well the regression line fits.

6. Residual plots.
For each observed point (X_i, Y_i), plot X_i on the x axis and e_i on the y axis.
$e_i = Y_i - \hat{y}_i$.

From the residual plot, you can eyeball the typical size of the residuals, and you can also see potential:
* outliers
* curvature
* heteroskedasticity.



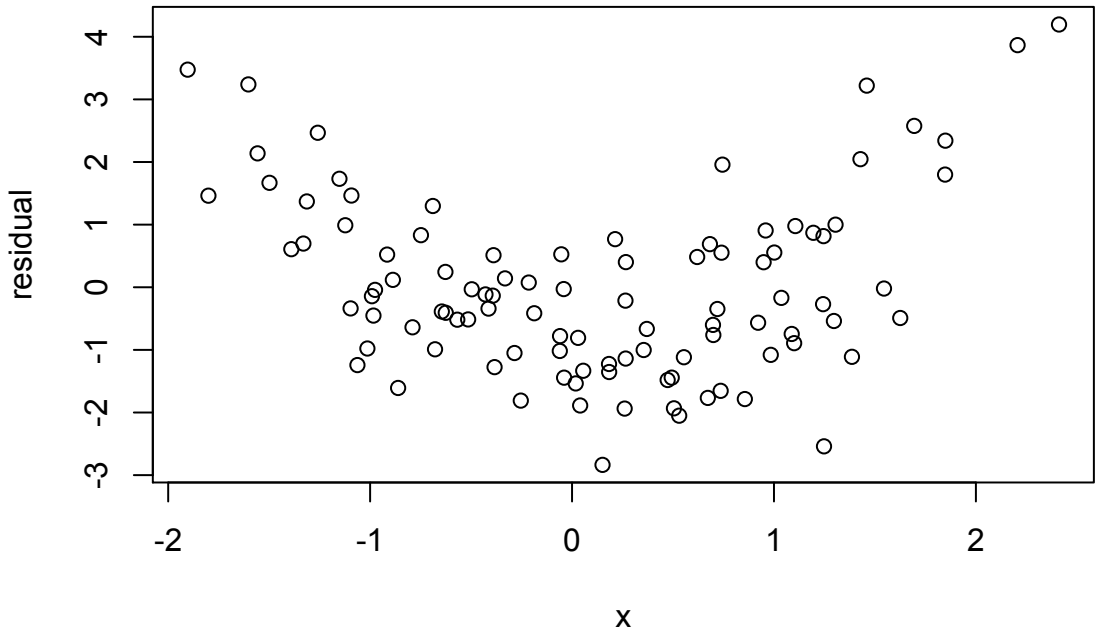


Outliers may deserve further attention, or reg. predictions may be good for some observations but lousy for others. The prediction $\hat{y}$ +/- RMS residual may be a lousy summary, if you're typically off by very little and occasionally off by a ton.

Another way to assess how well the regression line fits.

6. Residual plots.
For each observed point (X_i, Y_i), plot X_i on the x axis and e_i on the y axis.
$e_i = Y_i - \hat{y}_i$.

From the residual plot, you can eyeball the typical size of the residuals, and you can also see potential:
* outliers
* curvature
* heteroskedasticity.

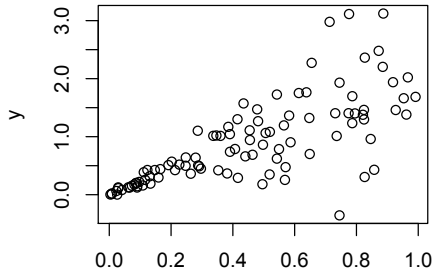


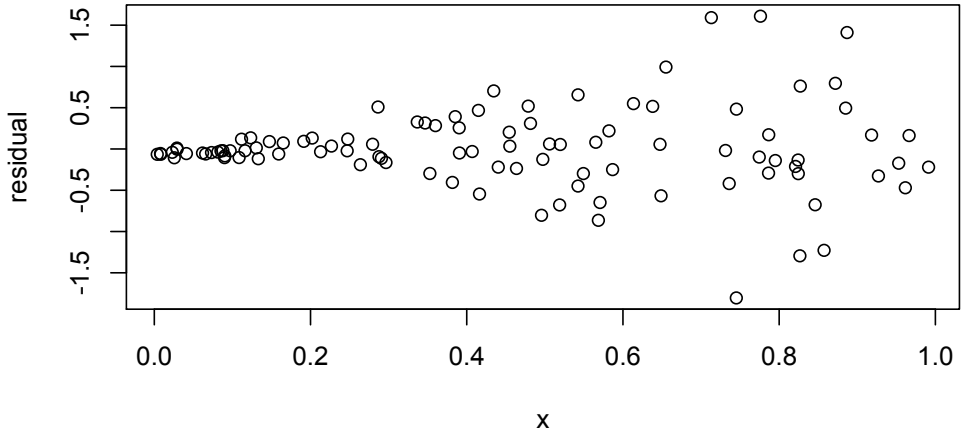
When curvature is present, the regression line may still be the best fitting line, but it does not seem to predict optimally: we could do better with a curve, rather than a line.

Another way to assess how well the regression line fits.

6. Residual plots.
For each observed point (X_i, Y_i), plot X_i on the x axis and e_i on the y axis.
$e_i = Y_i - \hat{y}_i$.

From the residual plot, you can eyeball the typical size of the residuals, and you can also see potential:
* outliers
* curvature
* heteroskedasticity.





Heteroskedasticity means non-constant variation. Homoskedasticity means constant variation. When we use the RMS residual for a +/- for our predictions, we're assuming homoskedasticity. If the data are heteroskedastic, then the RMS residual will be a lousy summary of how much we're off by. e.g. in this example, we know that when x is near 0, we're off by very little, and when x is large, we are often off by a lot.