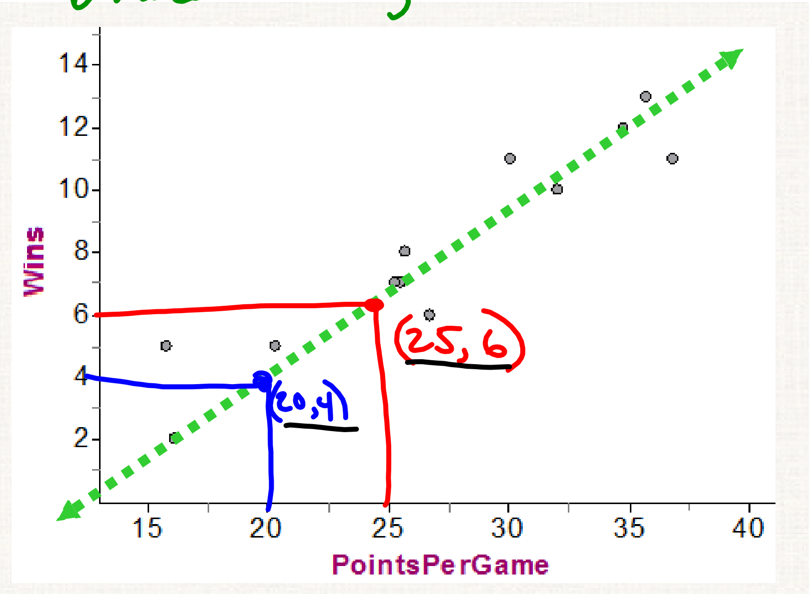
3.2 Least squares Regression

Regression line: line that describes how a response variable, y, changes as an explanatory variable, x, changes. It is used to predict y from x.

Draw a regression line



A regression line has the form $\hat{y} = a + bx$ $(\hat{y} = b_0 + b_1 x)$

- $\hat{y}$ is the predicted value of the response variable for a given explanatory variable, x.
- b is the slope, which is the amount of change we predict in the response variable for every 1 unit change in the explanatory variable.
- a is the y-intercept, which is the predicted value of y when x=0.

Ex: Write the equation of the regression line for the SEC football example. Interpret the slope & y-intercept.

$\boxed{\hat{y} = -4 + .4x}$

$\widehat{win} = -4 + .4(\text{points scored})$

(25, 6) $b = .4$

x y

$6 = a + .4(25)$

$6 = a + 10$

$-10 \quad -10$

$a = -4$

a = when you score zero points, we predict -4 wins.

b = We predict .4 wins for every point per game scored.

Extrapolation is the use of regression for prediction far outside the interval of values of the explanatory variable (x) used to obtain the regression line.
Beware extrapolating data, often the values will be inaccurate.

Check your understanding, p. 168

1. 40. We predict a rat will gain 40g/week.

2. 100. At birth, we predict they will weigh 100g.

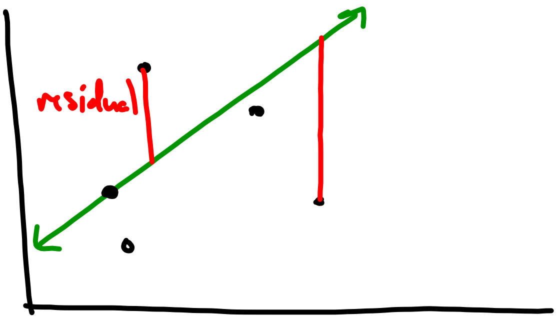
3. $\widehat{\text{weight}} = 100 + 40(16)$
$= 740\text{ g.}$

4. $\widehat{\text{weight}} = 100 + 40(104)$
$= 4260\text{ g}$
$\approx 9.4\text{ lb}$

<u>Residuals</u>: difference in y between an actual value (point on scatterplot) and a predicted value (point on the regression line).

residual = <u>a</u>ctual y − <u>p</u>redicted y = $y - \hat{y}$

Ex:



Ex: Find the residual for 32 points per game (<u>Georgia</u>) in context.

$$\text{residual} = y - \hat{y}$$
$$= 10 - 8.8$$
$$= 1.2$$

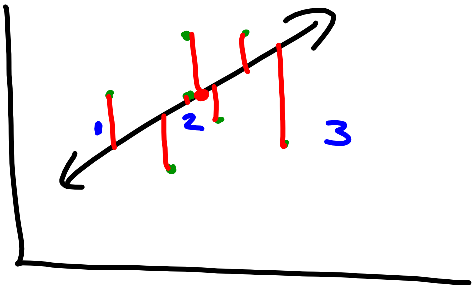
$$\widehat{\text{wins}} = -4 + .4(\text{points per game})$$
$$= -4 + .4(32)$$
$$= 8.8$$

The predicted number of wins was underestimated by 1.2 wins.

Least Squares Regression Line (LSRL)

line that makes the squared residuals as small as possible.

Ex:



p.170 Activity

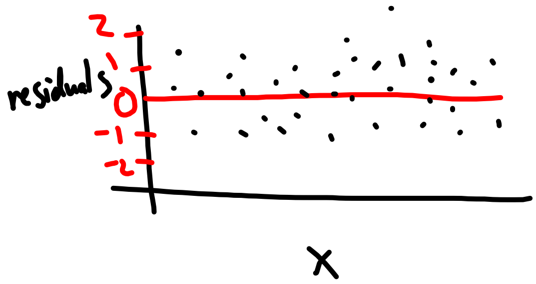
- an outlier "pulls" the line towards it. The further away from the middle the outlier is, the more effect it will have on the LSRL.
- Always goes through $(\bar{x}, \bar{y})$

p. 172, check your understanding

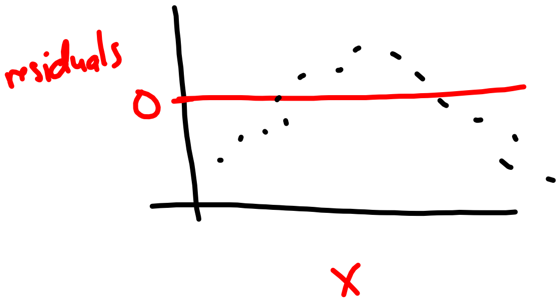
Residual Plot: scatterplot of residuals against the explanatory variable, x.

This helps us to assess whether a linear model is appropriate.

- A residual plot shows a linear model is appropriate if it is well scattered.



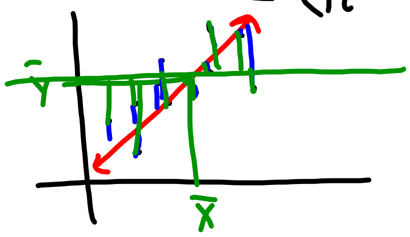
- A residual shows a linear model is not appropriate if it shows a distinct pattern.



Ex: make a residual plot for SEC football data

Coefficient of Determination, r^2, is the fraction of the variation in the values of y accounted for by the LSRL of y on x.

$$r^2 = 1 - \frac{\Sigma(\text{residuals})^2}{\Sigma(y_i - \bar{y})^2} = 1 - \frac{\text{error from LSRL y values}}{\text{error from the mean of y values}}$$



$\bar{y} = 8.08$

In context, r^2 is interpreted: "___% of the variation in (y-variable) is accounted for by the linear model relating (y-variable) to (x-variable)"

Ex: SEC football example.

88% of the variation in number of wins is accounted for by the linear model relating wins to points per game.

Calculating an LSRL

slope $= b = r \cdot \frac{s_y}{s_x}$

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

Properties of Correlation + Regression

1. The distinction of variables as explanatory and response is important regression.
2. correlation + regression should only be used for straight line relationships.
3. correlation and regression are not resistant to outliers. We call an observation an influential point if removing it would drastically change the result of a calculation
4. An association between an explanatory + response variable, even if very strong, is not by itself good evidence changes in x cause changes in y.

association ≠ causation