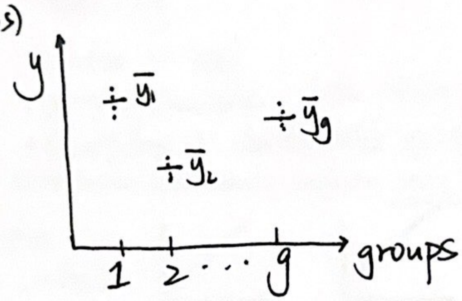
Simple Linear Regression (Ch. 3 and 12)

Regression vs. ANOVA

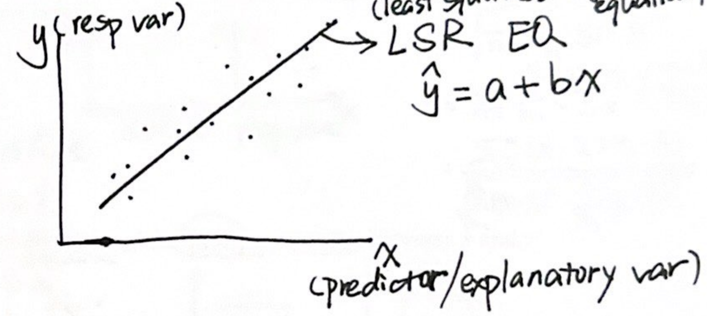
ANOVA compares means of several groups

resp = Quantitative
predictor = Categorical
(groups)



REGRESSION predicts average y for a particular x

Both x and y are quantitative



Least Squares Regression (LSR) Method:

- Find the "best fitting" line through $\boxed{\text{a set of } (x,y) \text{ points.}}$ → our data
- The regression line will minimize the sum of the **squared** vertical "distances" from points to the line.

goal: minimize $\Sigma(y_i - \hat{y}_i)^2$

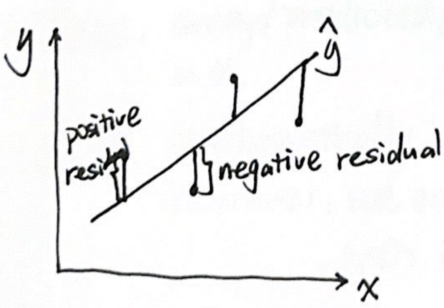
squared residual

- The sum of the vertical "distances" has to be zero.

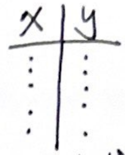
$\Sigma(y_i - \hat{y}_i) = 0$

The **residuals** are the vertical distances from the points to the line

residual = observed y − predicted y = $y_i - \hat{y}_i$



△can only compute residuals for observed data points.



Basics of Regression

- **Collect data** (x,y), both quantitative.

n observations
(means n x-values AND n y-values because we observe both for each subject)

summary statistics

mean	$\bar{x}$	$\bar{y}$
s.d.	S_x	S_y
corr.	r	

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

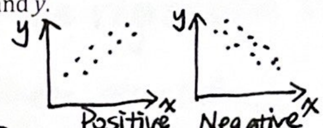
$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

$S_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$

$S_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}$

$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{S_x}\right)\left(\frac{y_i-\bar{y}}{S_y}\right) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \cdot \sum_{i=1}^{n}(y_i-\bar{y})^2}}$

- **Scatterplot**. Determine the relationship between x and y.



- Pos / Neg
- Linear or not
- strong / weak – how close to a line
- Outliers ? Influential outliers ?

- **Correlation coefficient** r: measures strength and direction of the linear association between x and y

$-1 \le r \le 1$ (no units)

The strength of the association:

$r<0.2$	very weak
$0.2 \sim 0.4$	weak
$0.4 \sim 0.6$	moderate
$0.6 \sim 0.8$	strong
$r>0.8$	very strong

Note: These are rather arbitrary limits, and the context of the results should be considered.

- **Compute LSR Equation:**

$\hat{y} = a + bx$

$b(\text{slope}) = r\frac{S_y}{S_x}$

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

The value of y when $x=0$.

- **Interpretation:**

- Slope: average/predicted/expected change in y for a one-unit change in x.
- y-int: Mathematically it is the average value of y when $x=0$. However, we only interpret it if $x=0$ makes sense AND is close to values of x observed.

EXTRAPOLATION:
Using the regression equation to predict for value of x far from data observed.

- **Coefficient of determination** R^2

$R^2 = (r)^2 =$ Percentage of variability in y explained by the regression on x.

Example: Suppose we collect data on UF students where x = height in inches, y = weight in pounds. Suppose the least squares regression equation is $\hat{y} = -250 + 6x$ and $r = 0.7$.

- Interpret the correlation and R^2. (linear association)

→ Corr $r = .7$ strong, positive correlation btwn ht and wt.

→ $R^2 = (.7)^2 = .49$ Interpretation: 49% of variability in wt is explained by the regression on ht.

What about when $R^2 = 0.37 / 0.40$?
$|r| > 0.6$ in this case ⇒ strong linear association.

- Interpret the slope and the intercept (if appropriate).

→ Slope (6): On average, we expect 6 extra pounds for each extra inch of height

→ y-int (−250): Mathematically, it is the value of $\hat{y}$ when $x = 0$.
But we DO NOT interpret it because $x = 0$" tall is impossible AND very far from hts of college students.

- Predict the (average) weight for someone whose height is 5'9".

5'9" = 69" so $x = 69$ inches ⇒ $\hat{y} = -250 + 6 \times 69 = 164$ pounds
(1' = 12")

Now, suppose we have one person in the data set with ht = 69" and wt = 160 pounds. Then we can find the residual 160 − 164 = −4 pounds. Hence, that person weighs 4 pounds less than the prediction.

- There was one person in the data set with height of 69" and weight of 160 pounds. Find their residual.

- Would we predict the weight for someone who is 2 ft tall (e.g. a small infant)?

NO. Too far from heights of college students
It would be ~~exp~~ extrapolation!