Regression (Chapter 11)

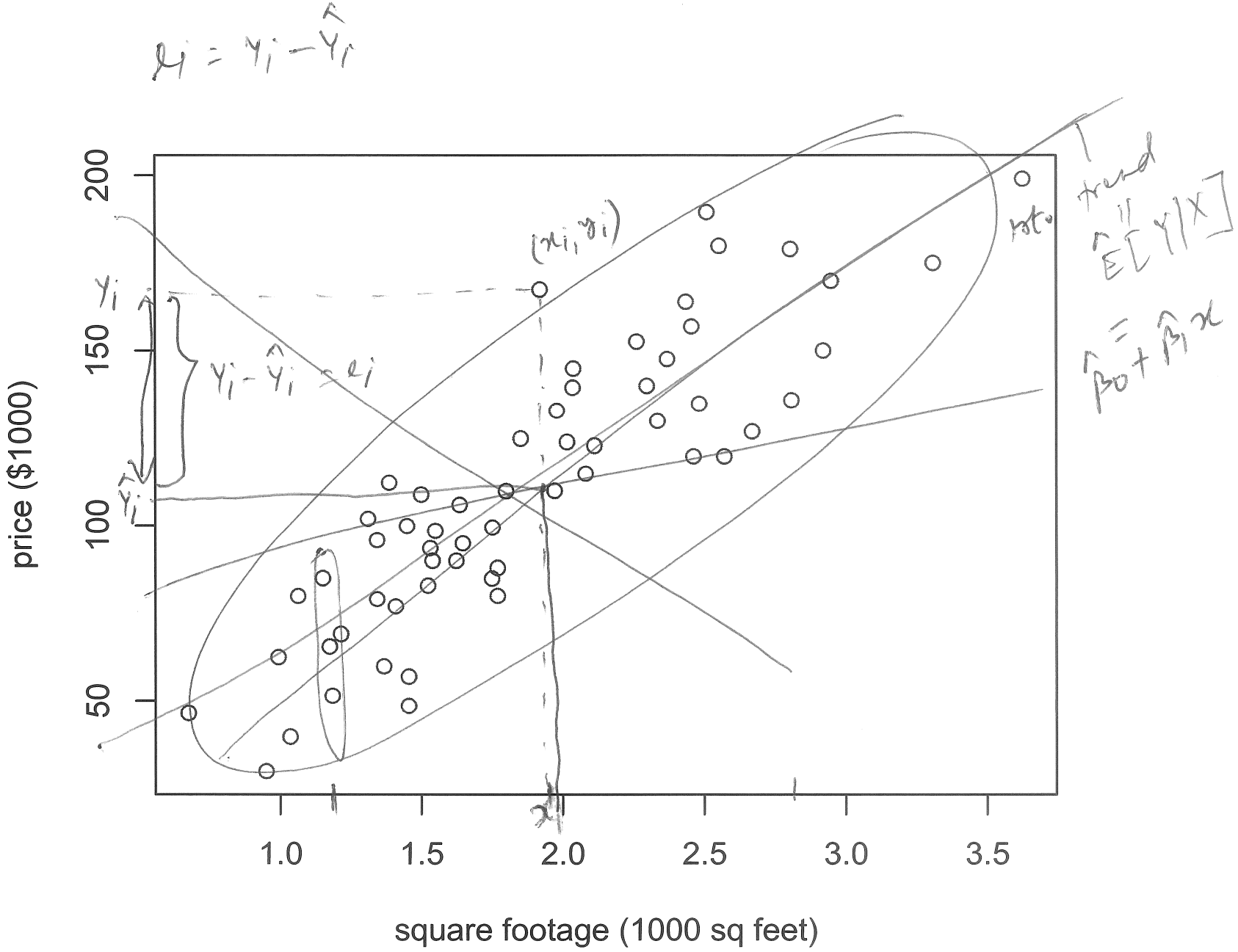
as opposed to qualitative or categorical.

Setup: Have data on two *quantitative* variables — X and Y — on a sample of n subjects. (relationship)

Q: Is there any association between X and Y? What kind?

Scatterplot: Data: (Y_i, X_i), $i = 1, 2, \ldots, n$. Coming from ith subject.

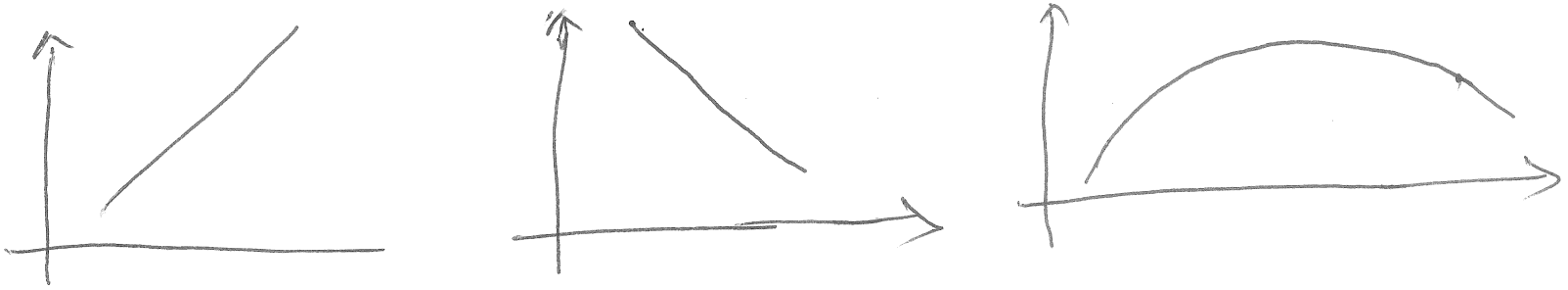
- Plot y against x
- Look for the **trend** in the plot — a smooth curve that shows how the average value of Y changes with x } $\therefore E[Y|X]$
- Trend may be linear or non-linear
- If there is a trend, then the two variables are associated. In this case, x may be used to predict y
- Trend may be strong or weak. It is strong if the points are tightly clustered around the trend (small scatter)
- No trend: No association — i.e., the variables are independent, and x is not helpful for predicting y.
- Predicted response: $\hat{Y} = \hat{E}[Y|X]$


$e_i = y_i - \hat{y}_i$
price ($1000)
square footage (1000 sq feet)
200
150
100
50
1.0
1.5
2.0
2.5
3.0
3.5
(x_i, y_i)
y_i
$\hat{y}_i$
$y_i - \hat{y}_i = e_i$
trend
$\hat{E}[Y|X]$
$= \hat{\beta}_0 + \hat{\beta}_1 x$

look at the "trend"

Overall pattern in a scatterplot

Form:



Direction: (Only for linear trend)

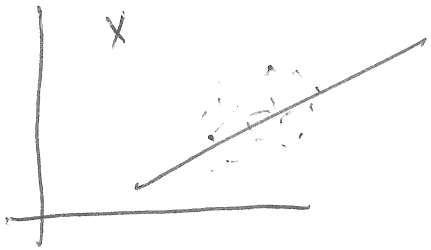
- ⊕ ve or ⊖ ve

Strength - assess the scatter of the points: around the trend.

small scatter ⇒ relationship b/w y and x is strong

large " ⇒ " " " " is weak.

Outliers - observations that don't follow overall pattern:



Focus on linear trend

Correlation Coefficient

Population correlation: A measure of **linear** relationship between X and Y. It is defined as,

$$\rho = \frac{\text{cov}(X, Y)}{\text{sd}(X)\text{sd}(Y)},$$

where $\text{cov}(X, Y) = E\{(X - E(X))(Y - E(Y))\}$ is covariance between X and Y.

Sample correlation: Estimator of ρ. It is defined as

$$r = \frac{S_{xy}}{S_x S_y},$$

where

$$S_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{x})(Y_i - \bar{y}) \quad \text{— sample covariance}$$

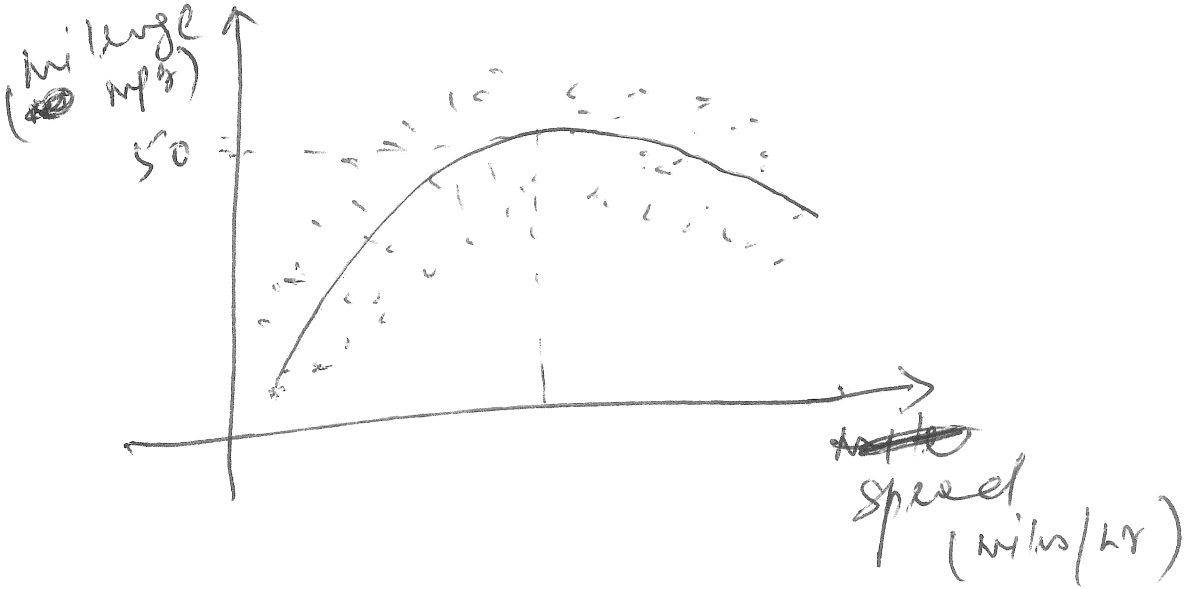
$$S_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{x})^2$$

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

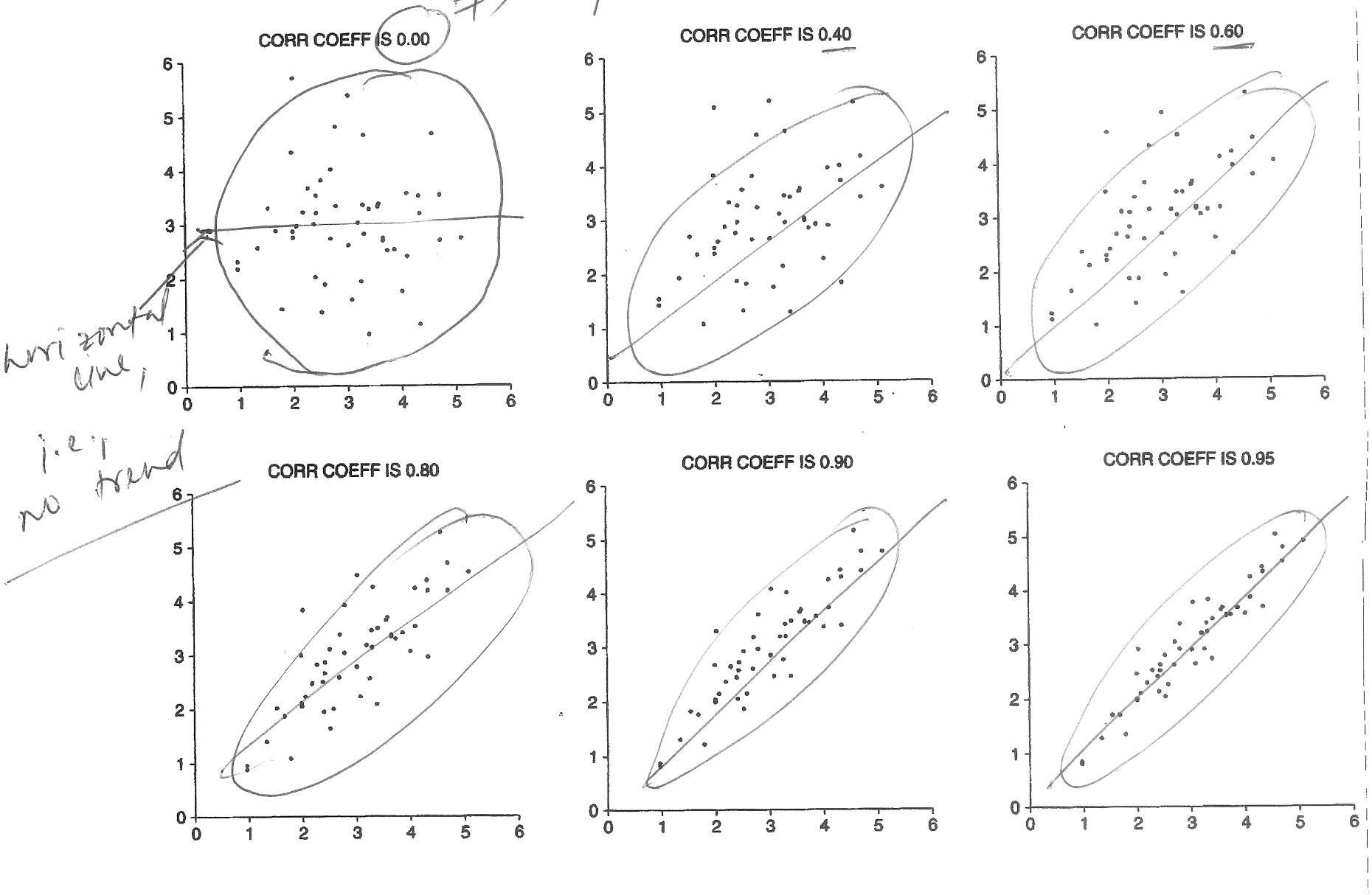
Properties of ρ and r:

- Range between -1 to 1
- Sign tells us direction of the linear relationship
- Absolute value tells us strength of the linear relationship
- Perfect correlation: $|\rho| = 1$ or $|r| = 1 \Rightarrow Y = a + bX$
- Unit free
- **No change** if X and Y are interchanged or if X is replaced by $aX + b$ and/or Y is replaced by $cY + d$, where a and c have the same sign. The sign will reverse if a and c have different signs.
- Zero correlation: **No linear** relationship. But there may be non-linear relationship.
- Independent X and Y: zero correlation, but the converse may not be true

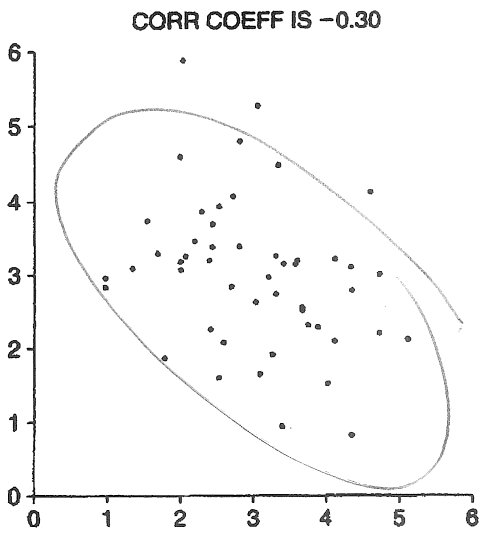
Example non-linear relationship:

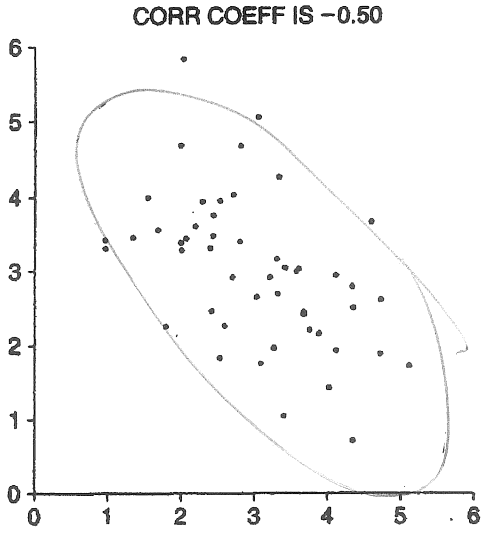


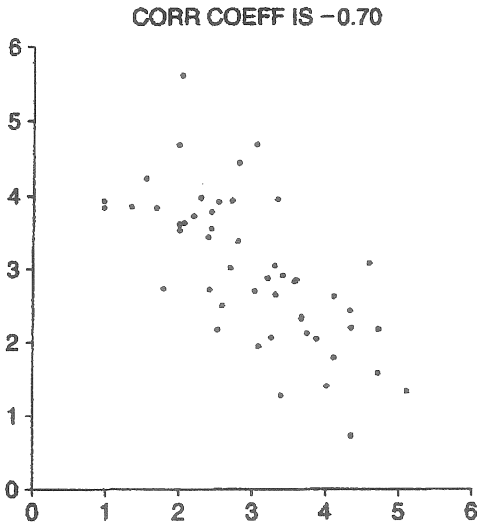
Examples of Positive Correlation

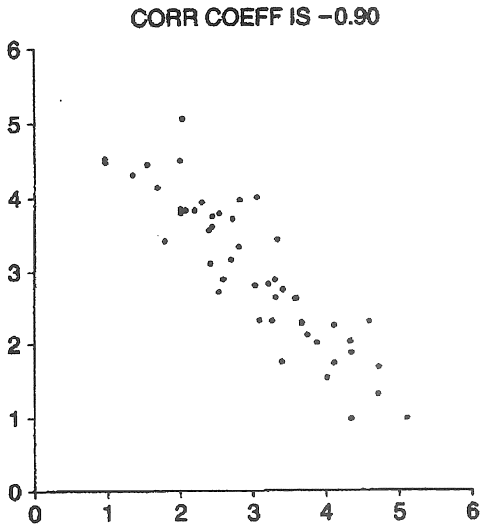


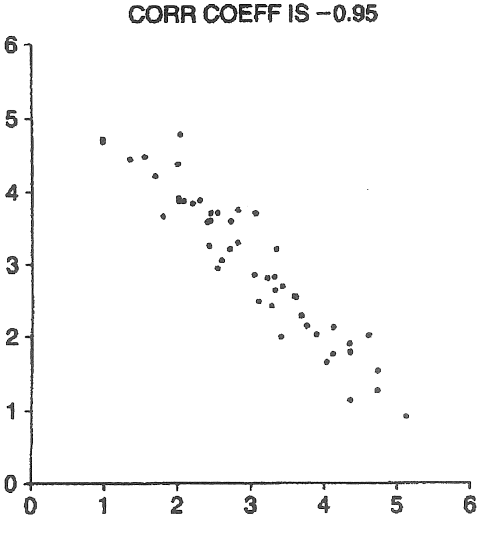
Examples of Negative Correlation

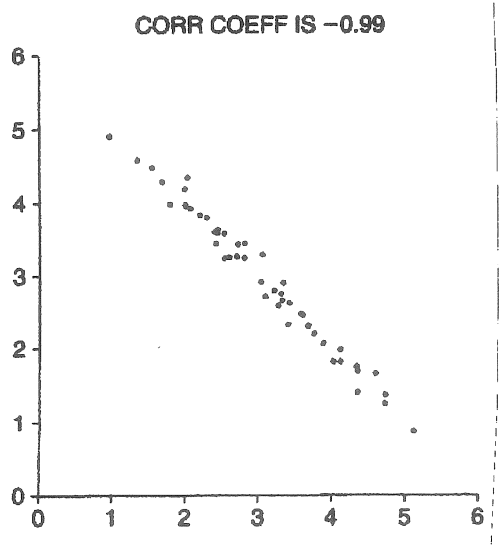








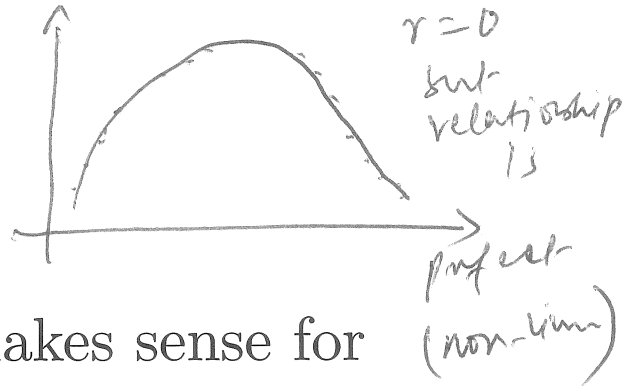




Caution: Non-linear relationships

Ex: Scatter plot of speed and mileage (miles per gallon) of an automobile.

More non-linear patterns: Any ~~things~~ trend that is not linear is non-linear



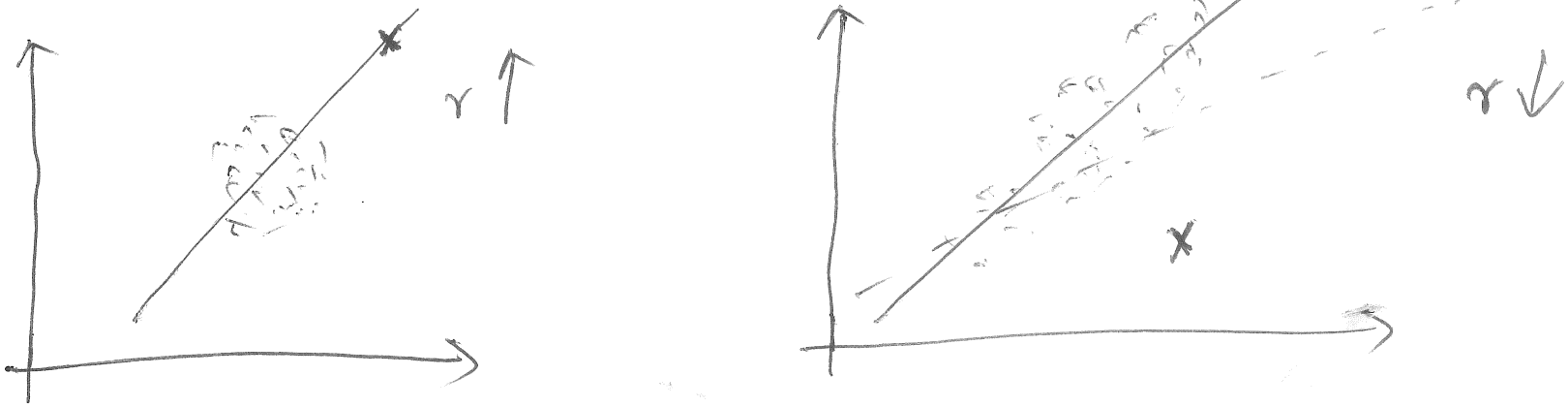
You can always compute r — but it doesn't makes sense for curve patters.

Lesson: Correlation only measures the strength of **linear** association — i.e., how close the points are to a straight line.

(case of linear trend)

Caution: Outliers

The position of an outlier relative to the rest of the ("cloud") of points determines how it affects r. Outliers may decrease or increase the value of r.



Note: Just knowing the value of r will give no information about whether outliers are present. That's why it is important to look at scatterplots.

Simple Linear Regression

Setup: Have data $(X_i, Y_i), i = 1, \ldots, n$, on two quantitative variables X & Y. Their scatterplot shows a linear relationship. Need an equation that would allow us to predict Y from X.

Response variable (Y): variable to be predicted (or modeled), aka, *dependent variable*.

Predictor (X): variable used to predict Y, aka, *independent or explanatory variable or covariate*.

Regression model: A function that models mean response — $E(Y|X = x)$ — as a function of x

Simple linear regression: $E(Y|X = x) = \beta_0 + \beta_1 x$

- Assumes mean response changes *linearly* with x
- β_0 : intercept — $E(Y|X = 0)$
- β_1 : slope — rate of change of mean response. It represents the change in mean when x increases by 1 unit.

- The regression coefficients are estimated from data.
- Let $(\hat{\beta}_0, \hat{\beta}_1)$ = estimator of (β_0, β_1).

Observed response: Y_i when $X = x_i$, $i = 1, \ldots, n$.

Fitted response: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$

- *Estimated mean response* when $X = x$
- *Response predicted* by the regression line
- $(\hat{Y}, x)$ falls on the regression line
- Fitted values: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, $i = 1, \ldots, n$.

Residuals: $e_i = Y_i - \hat{Y}_i$, $i = 1, \ldots, n$.

- Vertical distance between observed and predicted Y's
- Error in prediction
- Large residuals: observed and fitted Ys are too far

Least squares method for estimating coefficients: Find $(\hat{\beta}_0, \hat{\beta}_1)$ that minimize the *sum of squares of residuals*

measures how close the line is to data.

our focus.

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$

- Results in the **line of best fit** — the line is such that the fitted Ys are "closest" to the observed Ys
- Fitted regression line: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- Other criteria possible, e.g., minimizing $\sum_{i=1}^{n} |e_i|$, but the resulting estimates don't have simple expressions

valid.

(but note that $\sum_{i=1}^{n} e_i$ is not a valid)

To minimize $\sum_{i=1}^n e_i^2$ wrt (β_0, β_1), solve the **normal equations**

$$\frac{\partial \sum_{i=1}^n e_i^2}{\partial \beta_0} = 0, \quad \frac{\partial \sum_{i=1}^n e_i^2}{\partial \beta_1} = 0,$$

resulting in the **least squares estimates**

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}, \quad \hat{\beta}_1 = r S_y / S_x,$$

where r is **sample correlation**, and S_x and S_y are **standard deviations** of x and y samples, respectively.

Recall that:

$$S_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \overline{x})^2, \quad S_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \overline{y})^2,$$

$$r = \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{S_x S_y}$$

The fitted regression line

Fitted regression line: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x.$ Plugging-in $\hat{\beta}_0$ and $\hat{\beta}_1$,

$$\hat{Y} = (\bar{Y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 x = \bar{Y} + \hat{\beta}_1 (x - \bar{x})$$

$$= \bar{Y} + r \frac{S_y}{S_x} (x - \bar{x}) \Rightarrow \frac{(\hat{Y} - \bar{Y})}{S_y} = r \frac{(x - \bar{x})}{S_x}$$

implying that

$$\frac{\hat{Y} - \overline{Y}}{S_y} = r \frac{(x - \overline{x})}{S_x}.$$

- If x is 1 SD away from its mean $\overline{x}$, $\hat{Y}$ is r SD away from its mean $\overline{Y}$. Since $|r| \leq 1$, this means $\hat{Y}$ is **closer** to $\overline{Y}$ (in units of SD) than x is to $\overline{x}$ — **regression toward mean**.
- The fitted line passes through the points $(\overline{x}, \overline{y})$.
- The sign of slope $\hat{\beta}_1$ is same as the sign of r.
- The sum of residuals, $\sum_{i=1}^{n} e_i =$
- The average of fitted values, $(1/n) \sum_{i=1}^{n} \hat{Y}_i =$