Regression (Chapter 11)

Setup: Have data on two *quantitative* variables — X and Y — on a sample of n subjects. (x_i, y_i), $i = 1, 2, \ldots, n$.

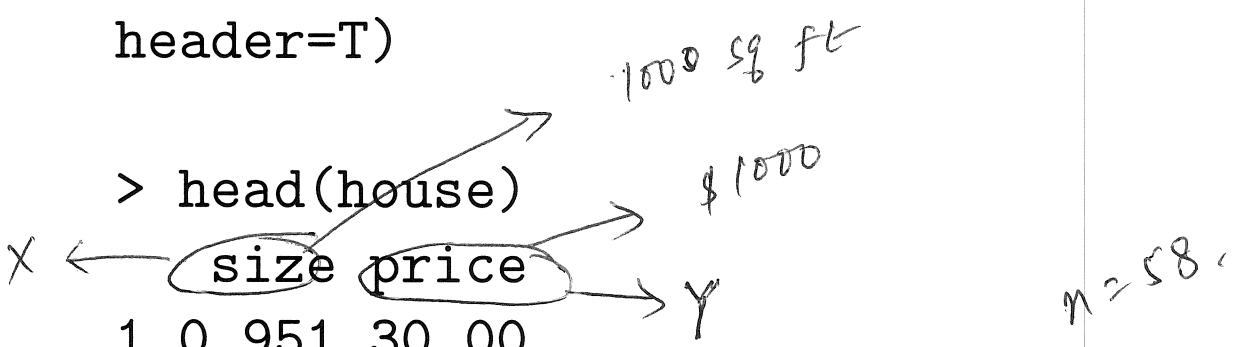
Q: Is there any association between X and Y? What kind?

Scatterplot:

- 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
- 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.

Example: House price data

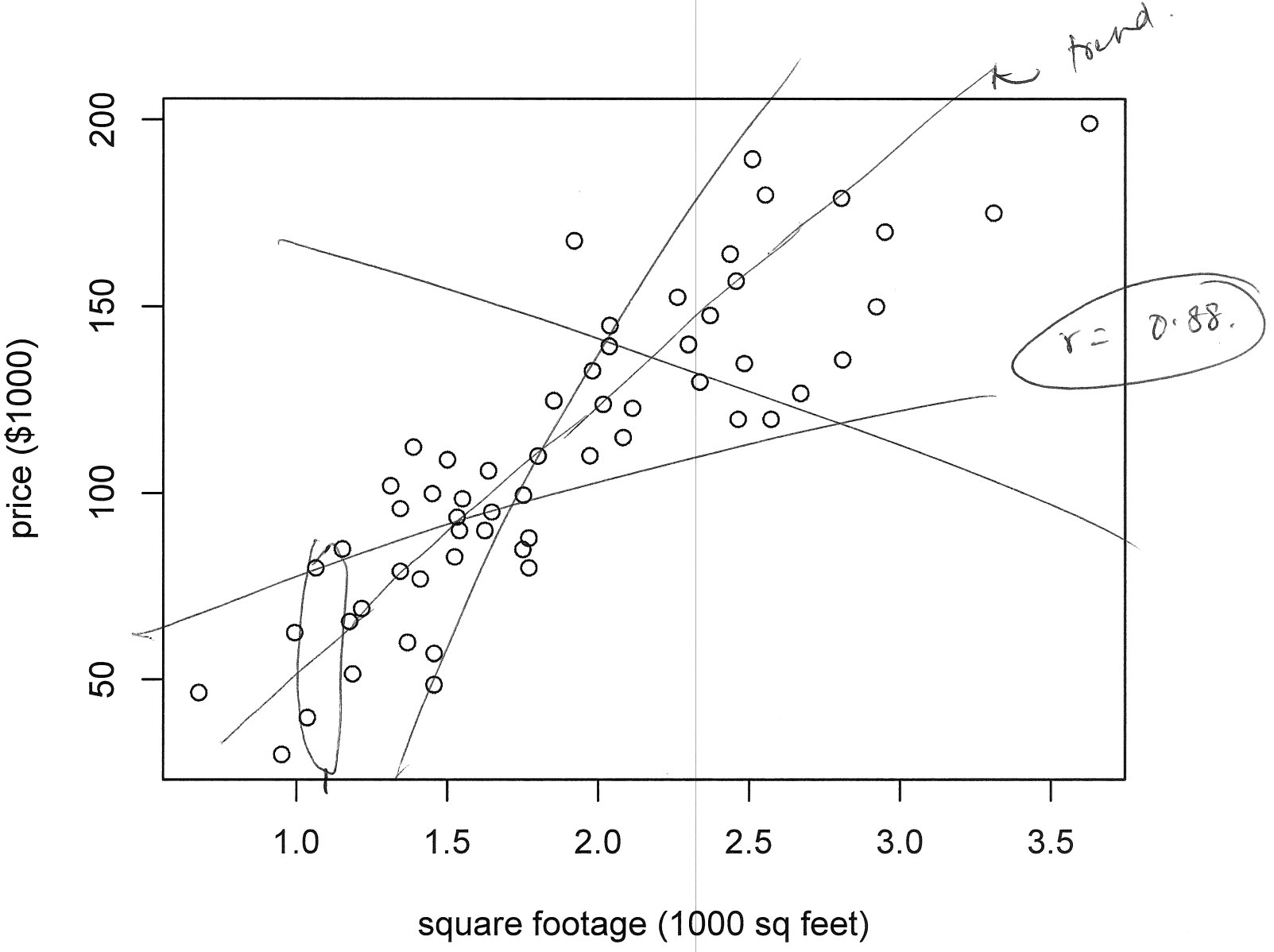
```
house <- read.table(file="house_price.txt", sep=",",
header=T)
```




```
> head(house)
  size price
1 0.951 30.00
2 1.036 39.90
3 0.676 46.50
4 1.456 48.60
5 1.186 51.50
6 1.456 56.99
>
```

```
> str(house)
'data.frame': 58 obs. of  2 variables:
 $ size : num  0.951 1.036 0.676 1.456 1.186 ...
```

```
 $ price: num  30 39.9 46.5 48.6 51.5 ...
>

# Make a scatterplot

plot(house$size, house$price,
xlab="square footage (1000 sq feet)",
ylab="price ($1000)")
```

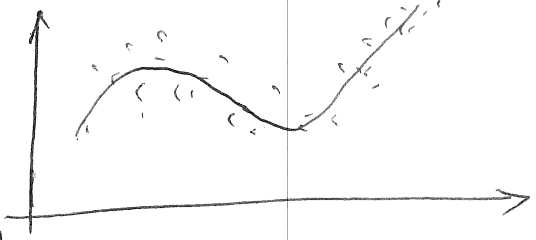

price ($1000)
50
100
150
200
1.0
1.5
2.0
2.5
3.0
3.5
square footage (1000 sq feet)
r found.
r = 0.88.

'Look at the trend'

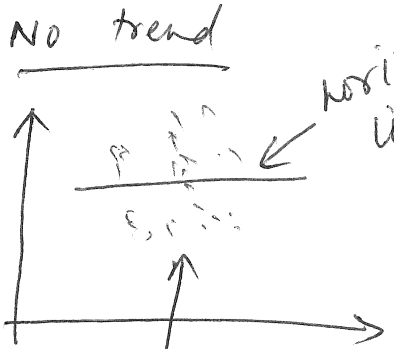
Overall pattern in a scatterplot

Form: Linear trend

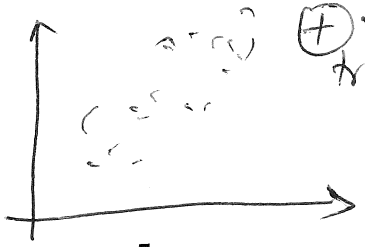
Non-linear trend



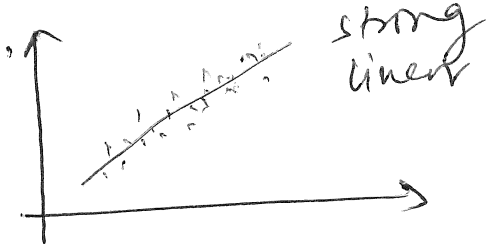
No trend

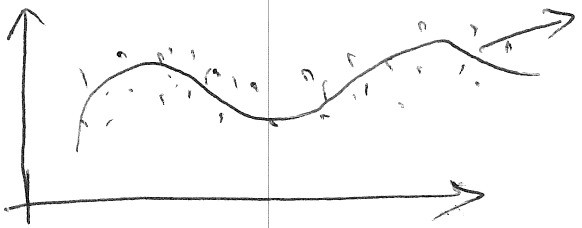


Direction: (Linear trend)

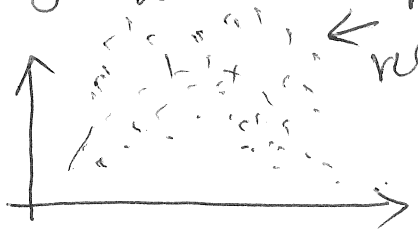


Strength - assess the scatter of the points:

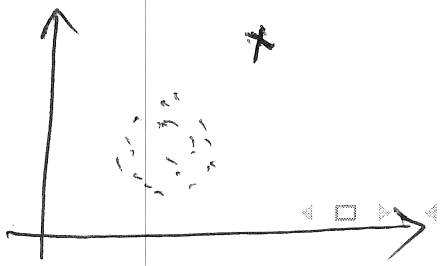




strong non-linear relationship



Outliers - observations that don't follow overall pattern:



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

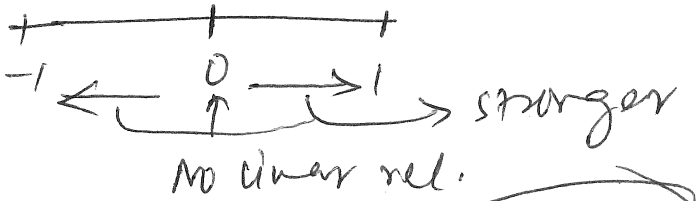
$$r = \frac{S_{xy}}{S_x S_y},$$

where $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$

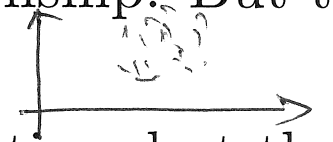
$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

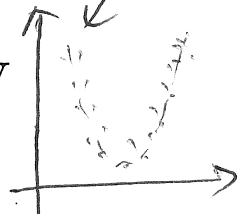
Properties of ρ and r:

- Range between -1 to 1
- Sign tells us direction of the linear trend (+ve: upwards slope; −ve: downwards slope.)
- Absolute value tells us strength
- Perfect correlation: $|r|$ or $|\rho| = 1$ ($Y = aX + b$)
- Unit free



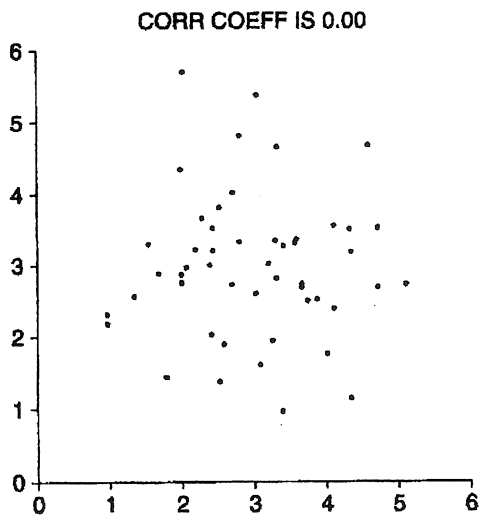
- **No change** if X and Y are interchanged or if X is replaced by $aX + b$ and/or Y is replaced by $cY + d$, where a and c have the same sign. The sign will reverse if a and c have different signs. (verify this)
- Zero correlation: **No linear** relationship. But there may be non-linear relationship.

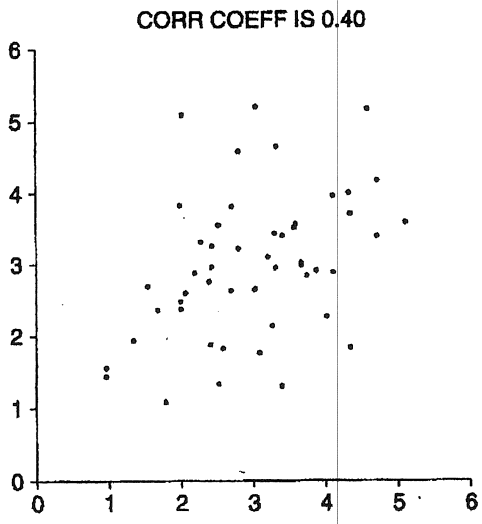


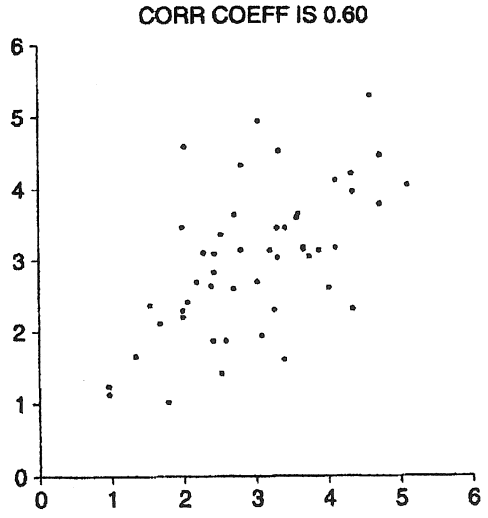


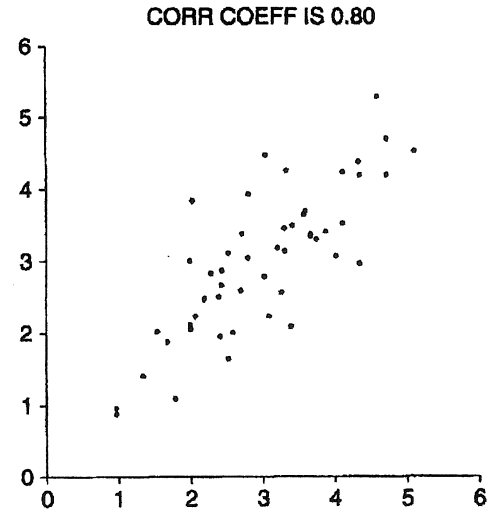
- Independent X and Y: zero correlation, but the converse may not be true

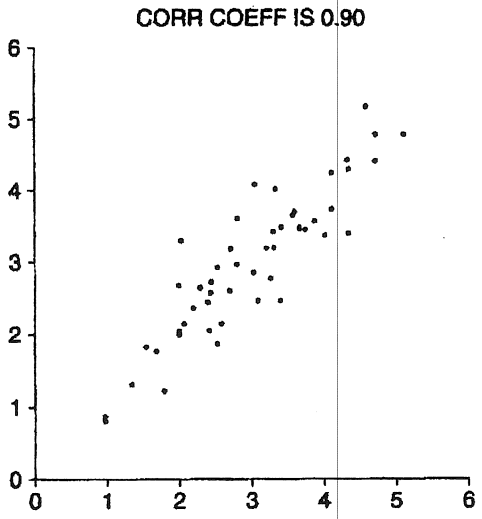
Examples of Positive Correlation

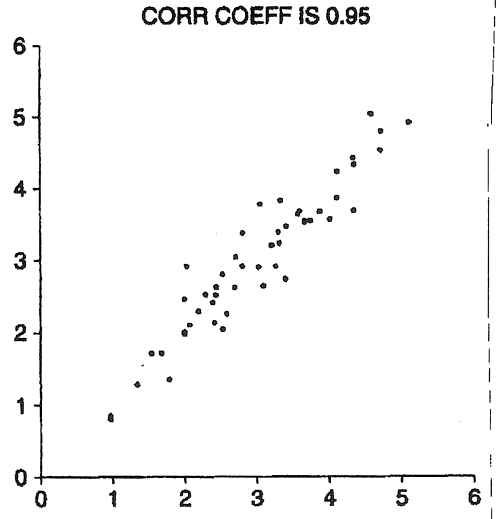




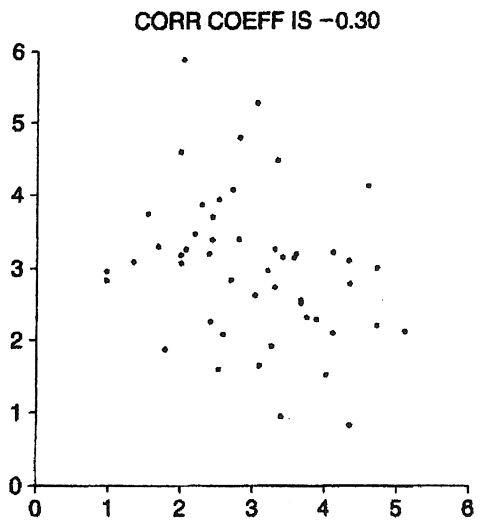


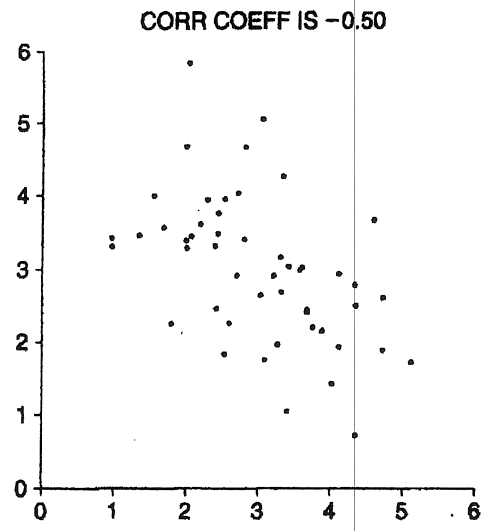


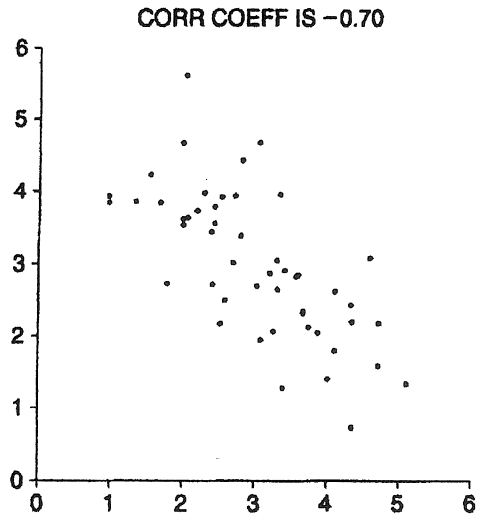


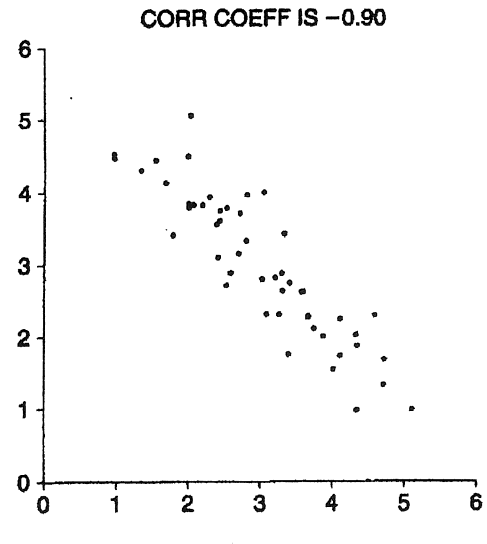


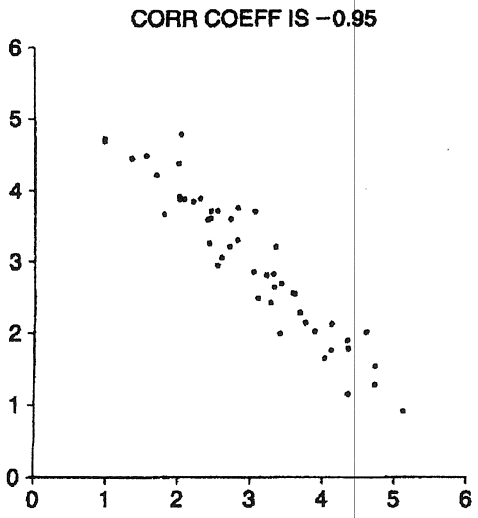
Examples of Negative Correlation

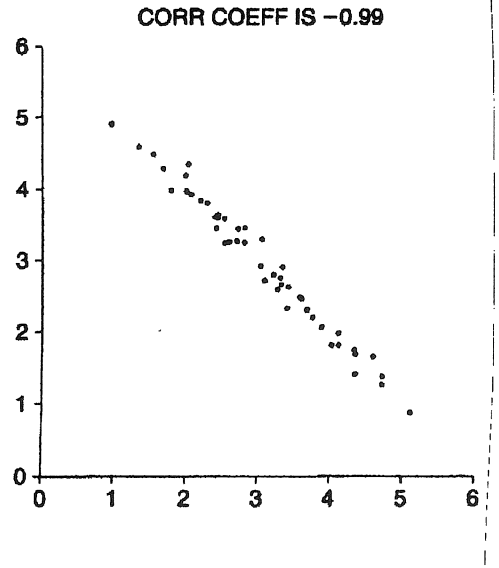






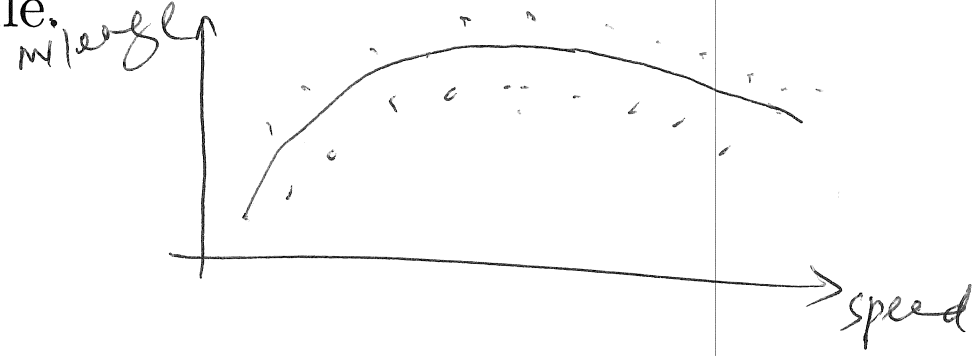




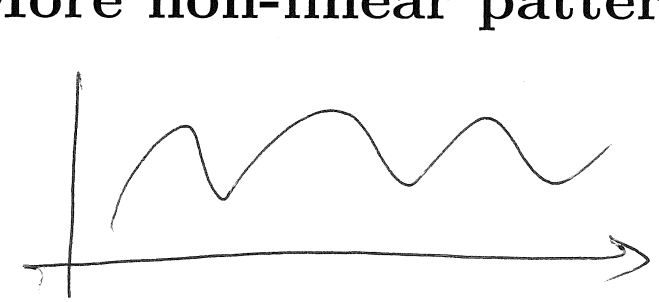


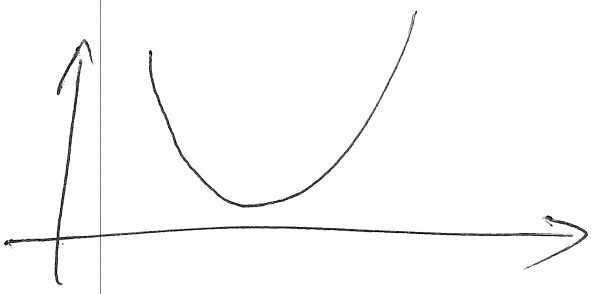
Caution: Non-linear relationships

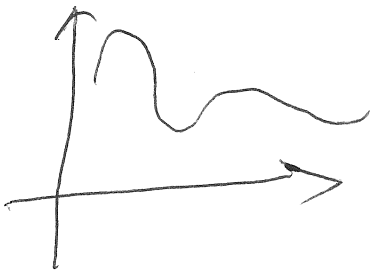
Ex: Scatter plot of speed and mileage (miles per gallon) of an automobile.



More non-linear patterns:





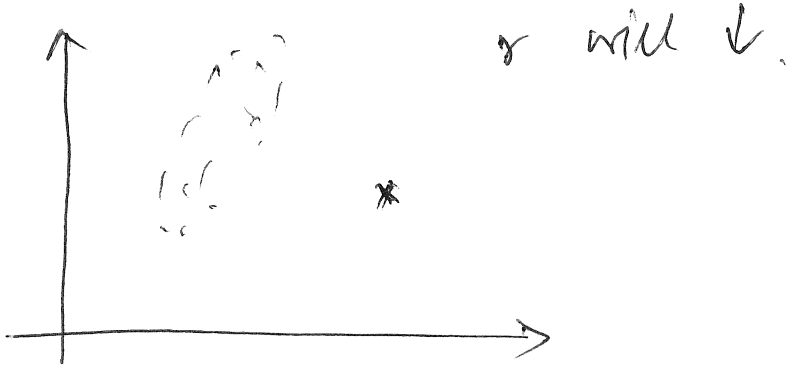


You can always compute r — but it doesn't makes sense for curve patters.

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

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

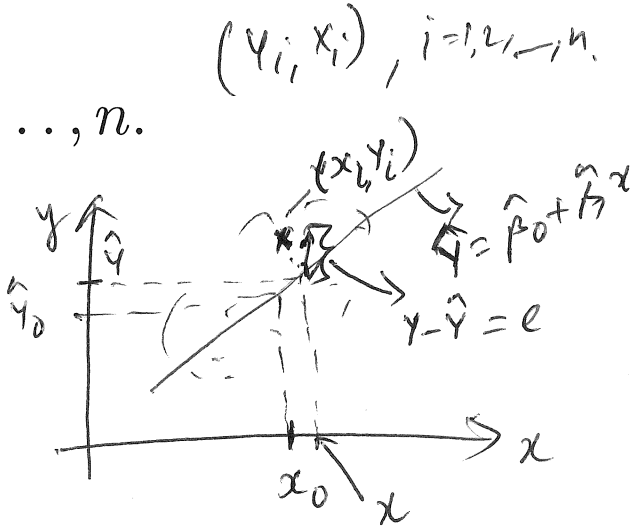
Simple linear regression: $E(Y|X = x) = \beta_0 + \beta_1 x$

- 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*

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

- 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

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

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

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

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:

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

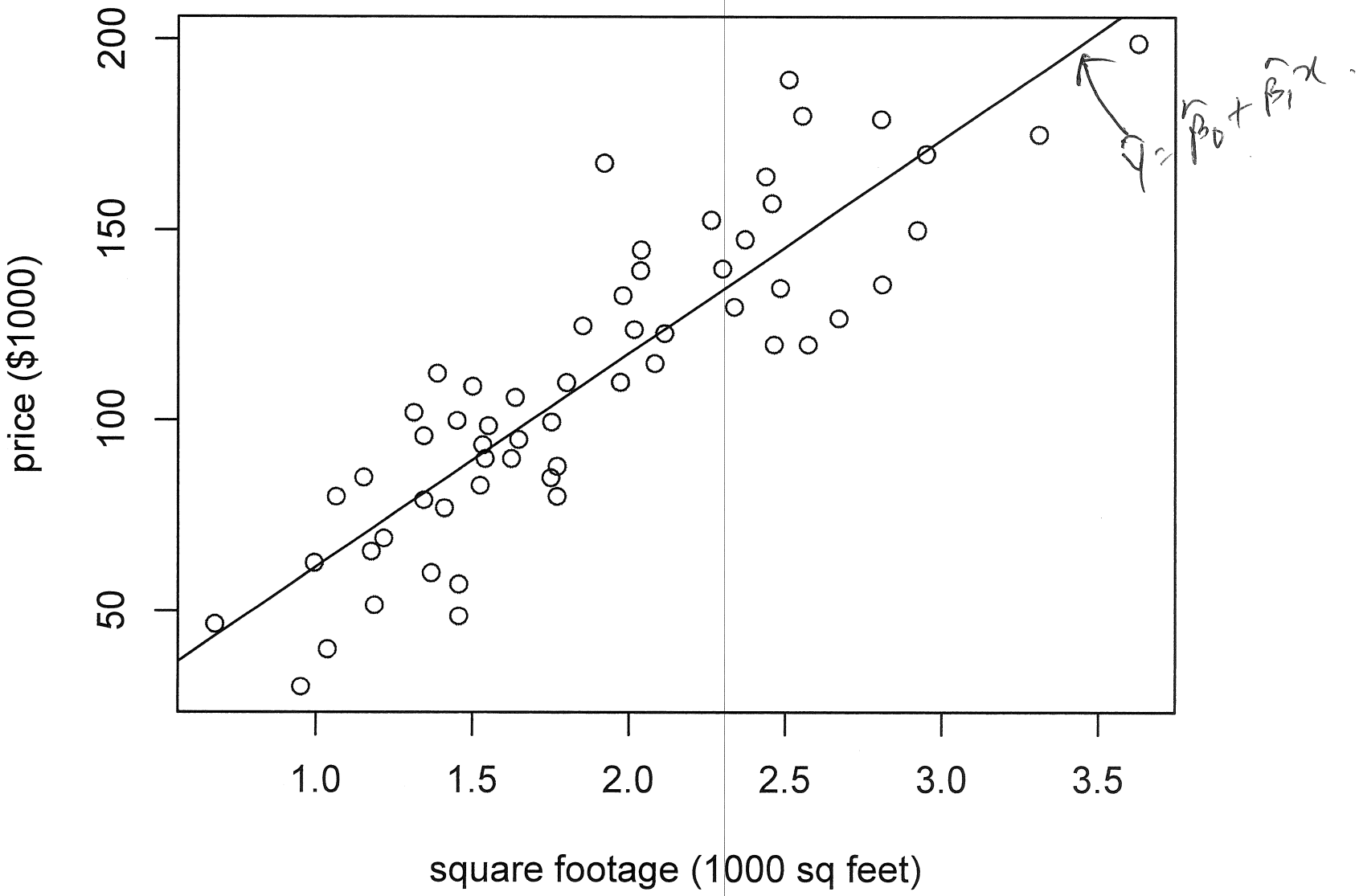
$$r = \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{S_x S_y}$$

Ex: Let's get the fitted line for the house price data and add it to the scatterplot.

```
x <- house$size
y <- house$price

# Get the fitted regression line
> (house.reg <- lm(y ~ x))

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
      5.432       56.083

>

# Does R do what we expect it to do?
> c(mean(x), sd(x), mean(y), sd(y), cor(x,y))
```

```
[1]   1.8829655   0.6316624 111.0344483  40.4431900
0.8759374
>
> cor(x,y)*sd(y)/sd(x)
[1] 56.08328
>
> mean(y)-(cor(x,y)*sd(y)/sd(x))*mean(x)
[1] 5.431568
>

# Add the line to the plot
plot(x, y, xlab="square footage (1000 sq feet)",
ylab="price ($1000)")
abline(house.reg)
```

Fitted regression for house price data



The estimated regression coefficients are:

$$\hat{\beta}_0 = 5.432, \ \hat{\beta}_1 = 56.083$$

Q: How do we interpret these coefficients? What is the predicted price of a house that is 3200 square feet?

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ $\quad x = 3.2$

$= (5.432) + (56)(3.2)$

$= ?$