# Linear regression

Linear regression assumes that the relationship between two variables, $x$ and $y$, can be modelled by a straight line:

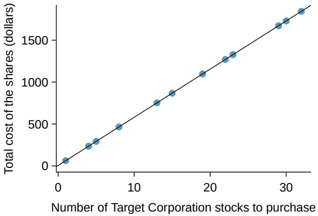
$$y = \beta_0 + \beta_1 x$$

where $\beta_0$ is the intercept and $\beta_1$ the slope.

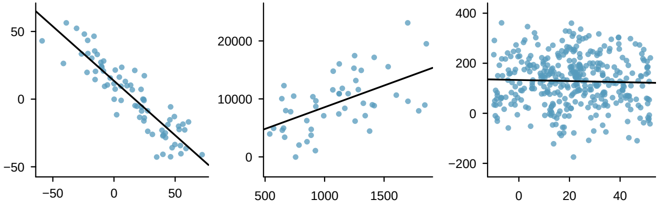
- $x$ is called the **explanatory** or the **predictor** variable,
- $y$ is called the **response** variable.

# Linear regression

## Theory

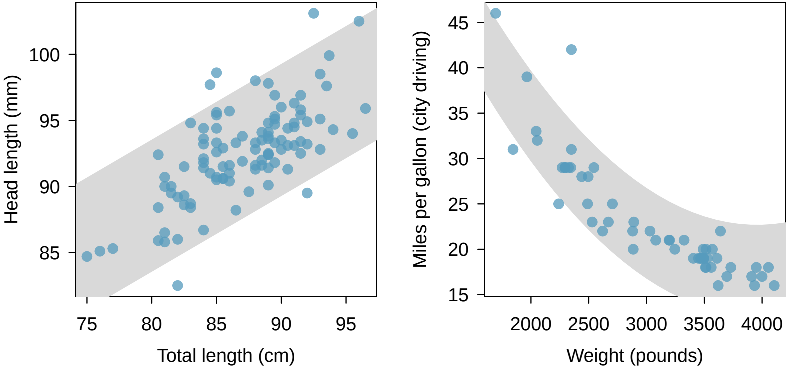



## Reality



# Look for linear trend

Examine the **scatter plot**



# Fitting a line

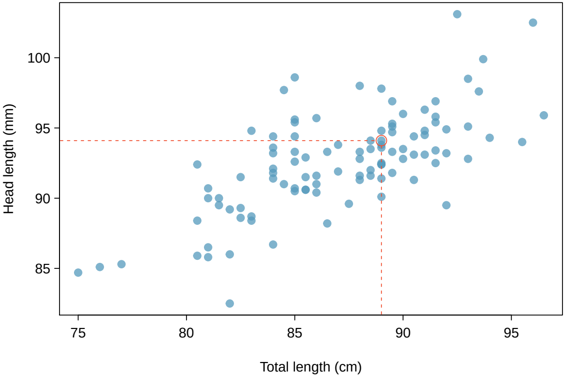



Figure: A scatterplot showing head length against total length for 104 brushtail possums. A point representing a possum with head length 94.1mm and total length 89cm is highlighted.

# Fitting a line

**Probabilistic interepretation**

$$y = \beta_0 + \beta_1 x + \epsilon,$$

where $\epsilon$ is a random variable with mean 0.
The fitted line

$$\hat{y} = b_0 + b_1 x,$$

is the expected value of y for fixed x.

# Residuals

**Residuals** are the leftover variation in the data after accounting for the model fit:

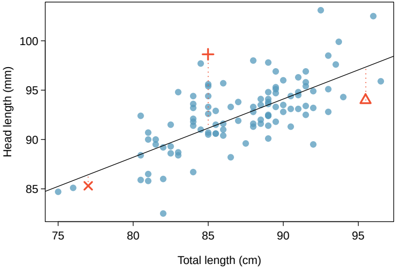
$$\text{Data} = \text{Fit} + \text{Residual}$$

The residual of the $i^{th}$ observation $(x_i, y_i)$ is the difference of the observed response $(y_i)$ and the response we would predict based on the model fit $(\hat{y}_i)$:

$$e_i = y_i - \hat{y}_i$$

We typically identify $\hat{y}_i$ by plugging $x_i$ into the model.

# Residuals



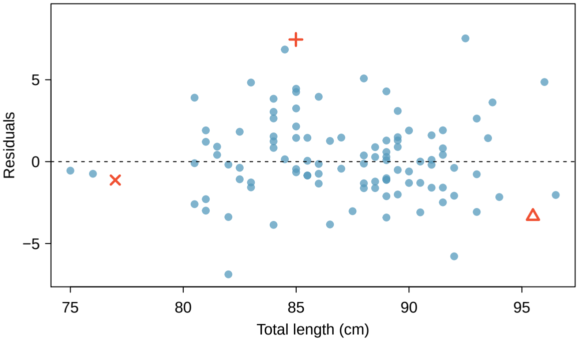

Fitted line:

$$\hat{y} = 41 + 0.59x$$

Observations:

- $\times$ is $(77.0, 85.3) \Rightarrow e_\times = y_\times - (41 + 0.59x_\times) = -1.1$
- $+$ is $(85.0, 98.6) \Rightarrow e_+ = y_+ - \hat{y}_+ = 7.45$
- $\triangle$ is $(95.5, 94.0) \Rightarrow e_\triangle = y_\triangle - \hat{y}_\triangle = -3.3$

# Residual plot



# Residual plots-Linear models

**Q:** How well does a linear model fit the data?

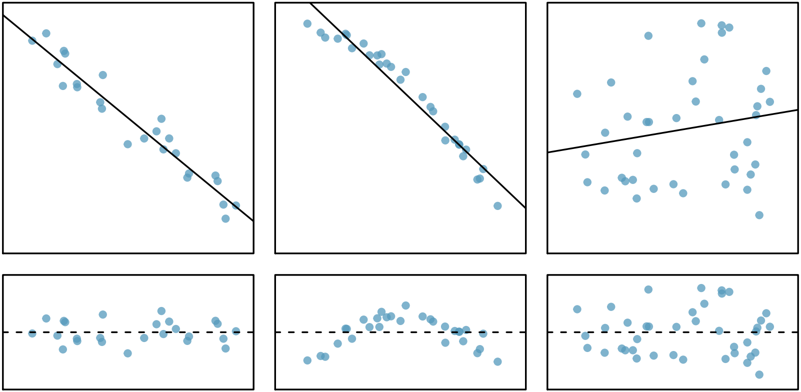


Figure: Sample data with their best fitting lines (top row) and their corresponding residual plots (bottom row).

# Correlation $R$

**Correlation**, which always takes values between -1 and 1, describes the *strength of the linear relationship* between two variables. We denote the correlation by $R$.

**Formally:**
The **correlation**, $R$, for observations $(x_1, y_1)$, $(x_2, y_2)$, ..., $(x_n, y_n)$ is:

$$R = \frac{1}{n-1} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} \frac{y_i - \bar{y}}{s_y}$$

where $\bar{x}$, $\bar{y}$, $s_x$, and $s_y$ are the sample means and standard deviations for each variable.

# Correlation

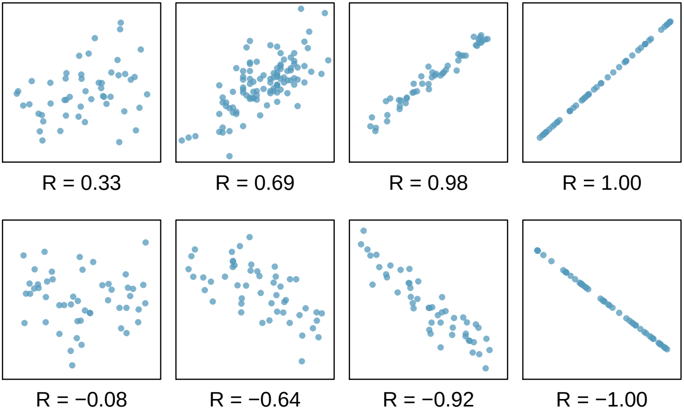



Figure: Sample scatterplots and their correlations. The first row shows variables with a positive relationship, represented by the trend up and to the right. The second row shows variables with a negative trend, where a large value in one variable is associated with a low value in the other.

# Correlation

The correlation quantifies the strength of a **linear** trend!!!!

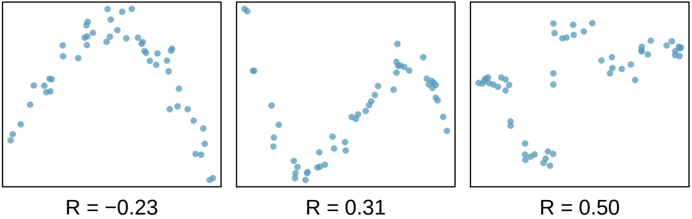



Figure: Sample scatterplots and their correlations. In each case, there is a strong relationship between the variables. However, the correlation is not very strong, and the relationship is not linear.

# Fitting a line

Which line is optimal?

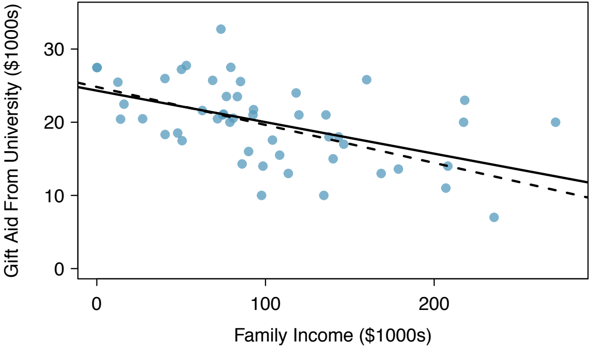
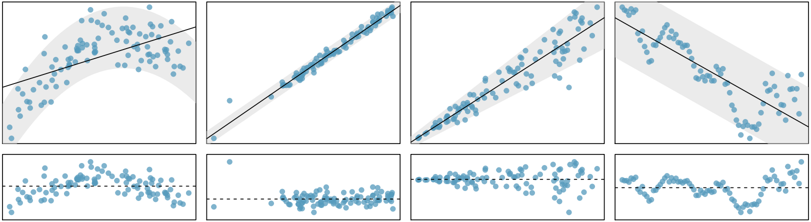



Figure: Gift aid and family income for a random sample of 50 freshman students from Elmhurst College. Two lines are fit to the data, the solid line being the *least squares line*.

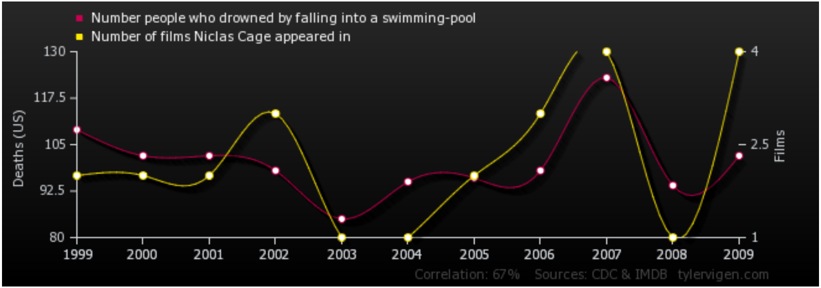
# Conditions

- **Linearity** The data should show a linear trend.
- **Nearly normal residuals** The residuals must be nearly normal. When this condition is found to be unreasonable, it is usually because of outliers or concerns about influential points.
- **Constant variability** The variability of points around the least squares line remains roughly constant.

# Counterexamples



# Correlation




$R = 0.66$[1]

[1] http://tylervigen.com/

# Least squares

**Least Squares regression**
Choose the line that **minimises** the sum of the square residuals:

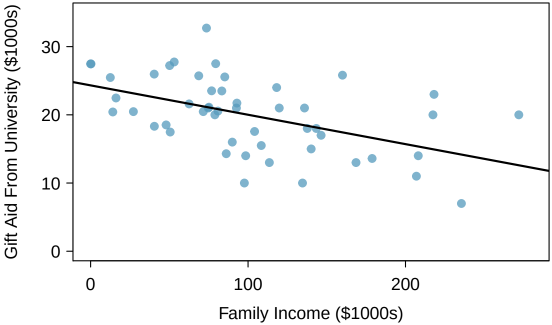
$$e_1^1 + e_2^2 + \cdots + e_n^2$$

This is the **least square line**, with parameters

$$b_0 = \bar{y} - b_1\bar{x} \text{ and } b_1 = \frac{s_y}{s_x}R,$$

where $\bar{x}$, $\bar{y}$, $s_x$, and $s_y$ are the sample means and standard deviations for each variable, and $b_0, b_1$ are the point estimates of $\beta_0$ and $\beta_1$ respectively.

# How to read the tables




| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 24.3193 | 1.2915 | 18.83 | 0.0000 |
| family_income | -0.0431 | 0.0108 | -3.98 | 0.0002 |

# Interpreting regression

From the table $b_0 = 24.32$ and $b_1 = -0.0431$, hence the least square line is:

$$\hat{y} = 24.32 - 0.0431x$$

Or

$$\widehat{aid} = 24.3 - 0.0431 \times family\_income$$

**Interpretation of slope** For each additional \$1,000 of family income, we would expect a student to receive a net difference of $\$1{,}000 \times (-0.0431) = -\$43.10$ in aid on average, i.e. \$43.10 *less*.

**Interpretation of intercept** The estimated intercept $b_0 = 24.3$ (in \$1000s) describes the average aid if a student's family had no income.

# Strength of fit

Sum of square errors:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

**SSE** provides a measure of variation in the $y$ values that remains unexplained after using the linear regression model.

**R-squared**

$$R^2 = 1 - \frac{SSE}{s_y^2}$$

**R-squared** can be interpreted as the proportion of the total variation in the $y$ values that is explained by the variable $x$ in least square line.

**Example:** For the Gift aid and family income data the correlation is $R = -0.499$ and the strength of fit is $R^2 = 0.25$.

# Extraplation

**Extrapolation** is applying a model estimate to values outside the realm of the original data.

**Beware!!!**

**Example**

The fitted line for the gift aid at Elmhurst college, with respect to family income, is

$$\hat{y} = 24.32 - 0.0431x$$

Can we predict the gift aid of a student with family income of \$ 1million?

- Apply the model, the financial aid for the student is

$$\hat{y} = -18.8$$

- The student must pay extra -\$ 18.800
- This is unrealistic, since Elmhurst college only charges a tuition fee.

# Categorical predictors

Categorical variables can be incorporated in linear models.
**Data** Ebay auctions for a video game, *Mario Kart* for the Nintendo Wii, where both the total price of the auction and the condition of the game were recorded.
**Goal** Fit a linear model of the form:

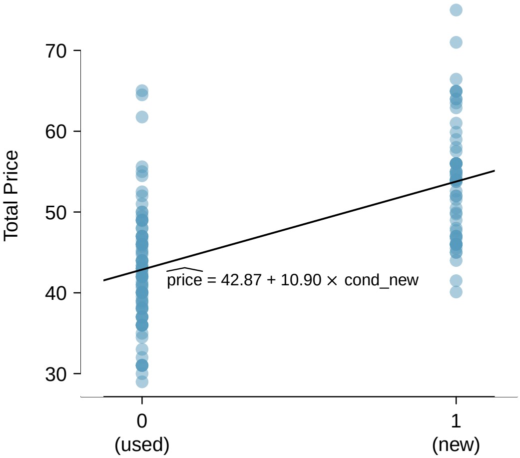
$$\widehat{price} = \beta_0 + \beta_1 \times \text{cond\_new},$$

Formally,

$$\hat{y} = \beta_0 + \beta_1 x,$$

where $y$ is the price of the game in dollars and $x$ is a categorical variable, with $x = 0$ if the game is used and $x = 1$ if the game is new.

# Categorical predictors



# Least square regression

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 42.87 | 0.81 | 52.67 | 0.0000 |
| cond_new | 10.90 | 1.26 | 8.66 | 0.0000 |

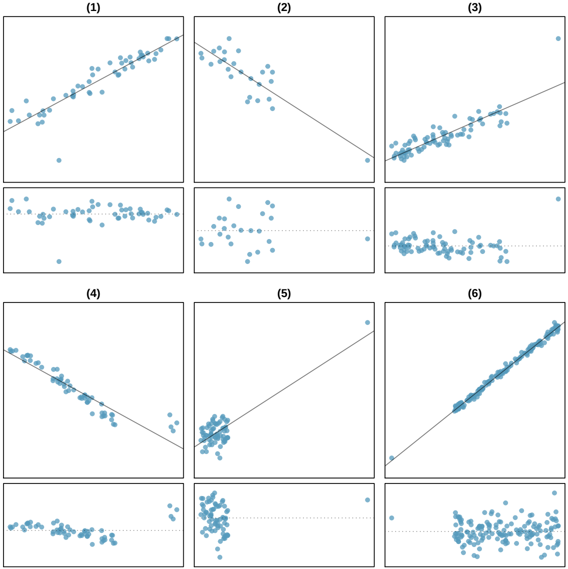
### Least square line

$$\hat{y} = 42.87 + 10.9x$$

### Interpretation

- The intercept indicates that, the average selling price of a used version of the game is \$42.87.
- The slope indicates that, on average, new games sell for about \$10.90 more than used games.

# Residuals



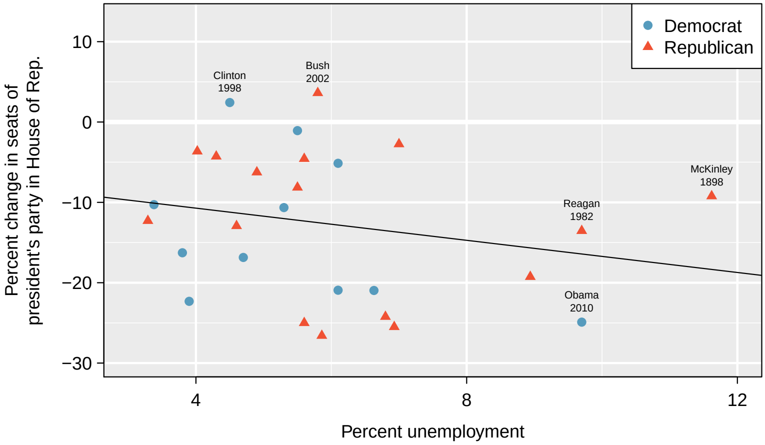
# Residuals

- Points that fall horizontally away from the centre of the cloud tend to pull harder on the line, so we call them points with **high leverage**.
- If one of these high leverage points does appear to actually invoke its influence on the slope of the line then we call it an **influential point**[2].
- If there are outliers in the data, they should not be removed or ignored without a good reason. Whatever final model is fit to the data would not be very helpful if it ignores the most exceptional cases.
- Be cautious about using a categorical predictor when one of the levels has very few observations. When this happens, those few observations become influential points.

[2] cases (3), (4), and (5)

# Inference for linear regression

**Hypothesis:** In America's two-party system, one political theory suggests the higher the unemployment rate, the worse the President's party will do in the midterm elections.



# Read the tables

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -6.7142 | 5.4567 | -1.23 | 0.2300 |
| unemp | -1.0010 | 0.8717 | -1.15 | 0.2617 |
| | | | | $df = 25$ |

**Least square line**

$$\hat{y} = -6.71 - 1.00 \times x,$$

where $y$ denotes the % change in House seats for President's party and $x$ the unemployment rate.

# Hypothesis testing

$H_0$: $\beta_1 = 0$.

$H_A$: $\beta_1 < 0$.

**Or**

$H_0$: The true linear model has slope zero.

$H_A$: The true linear model has a slope less than zero. The higher the unemployment, the greater the loss for the President's party in the House of Representatives.

# Hypothesis Testing

- From the table $P(> |t|) = 0.2617$.
- For a one-sided test
  $p - value = P(> |t|)/2 = 0.13 > 0.05$
- There is no strong evidence to suggests that the higher the unemployment rate, the worse the President's party will do in the midterm elections.