LEAST-SQUARES REGRESSION

Definition: **Regression Line**

A **regression line** is a line that describes how a response variable *y* changes as explanatory variable *x* changes. We often use a regression line to **predict** the value of *y* for a given value of *x*.

$\hat{y} = a + bx$

$\hat{y}$ (read "y hat") is the predicted value of the response variable *y* for a given value of the explanatory variable *x*.

b is the slope, the amount by which *y* is predicted to change when *x* increases by one unit.

a is the y-intercept, the predicted value of *y* when *x* = 0.

Definition: **Residual**

A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line. That is,

residual = observed - predicted = $y - \hat{y}$

Definition: **Least-squares regression line (LSRL)**

The **least-squares regression line** of y on x is the line that makes the sum of the squared residuals as small as possible.

Equation of the least-squares regression line from statistics:

slope: $b = r\frac{s_y}{s_x}$ and **y-intercept:** $a = \bar{y} - b\bar{x}$

Definition: **Residual plot**

A **residual plot** is a scatterplot of the residuals against the explanatory variable. Residual plots help us assess how well a regression line fits the data.

The residual plot should show no observable pattern and the residuals should be relatively small in size.

Definition: **Standard deviation of the residuals (s)**

If we use a least-squares regression line to predict the values of a response variable y from an explanatory variable x, the **standard deviation of the residuals (s)** is the value that approximates the size of a "typical" or "average" prediction error (residual).

Definition: **Coefficient of Determination: r^2 in regression**

The **coefficient of determination r^2** is the fraction of the variation in the values of *y* that is accounted for by the least-squares regression line of *y* on *x*.

Definition: **Extrapolation**

Extrapolation is the use of a regression line for prediction far outside the interval of values of the explanatory variable *x* used to obtain the line. *Such predictions are often not accurate*.

Definition: **Outliers and Influential observations in regression**

An observation is **influential** if removing it would markedly change the result of the calculation. Points that are outliers in the x direction of a scatterplot are often influential for the LSRL.

An **Outlier** is an observation that lies outside the overall pattern of the other observations. Points that are outliers in the *y* direction but not in the x direction of a scatterplot have large residuals.