# Topic VIII:
# Linear Regression

Wei YOU

香港科技大學
THE HONG KONG UNIVERSITY OF
SCIENCE AND TECHNOLOGY

Spring, 2025

# Beyond Categorical Comparisons

**Limitations of Dichotomous/discrete Grouping**

- In earlier topics we learned how to test for differences between two or more populations
  **Example:** smoking vs. non-smoking and heart-disease risk.
- Such methods ($\chi^2$, two-sample $t$, ANOVA) tell us *whether* groups differ, but they reduce rich data to simple "yes/no" or group-labels.
- When we dichotomize a continuous variable (e.g. smoker vs. non-smoker), we lose information about the intensity or degree of that variable.
- As a result, we cannot make precise predictions for individuals based on the full range of their measurements.

**Looking Ahead:** We now introduce *regression* methods that use continuous predictors directly, preserving information and enabling individualized prediction.

# Moving Beyond Pairwise Comparisons

- Most methods we've seen so far fall into two categories:
  - **Summarize the data:** descriptive statistics, point & interval estimation, one-sample hypothesis tests.
  - **Identify connections:** two-sample tests, ANOVA, contingency-table ($\chi^2$) tests.
- To predict outcomes, we must model how multiple variables relate to each other.
- **Example:** Prediction Given a person's years of smoking, estimate their probability of developing heart disease. $\Rightarrow$ Connection between two continuous variables.

# Overview

| $Y$ | – | $X$ |
|---|---|---|
| Response | – | Explanatory |
| Dependent | – | Independent |
| Outcome | – | Predictor |

| $Y$ | $X$ | Test | Lecture |
|---|---|---|---|
| Categorical | Categorical | Pearson's $\chi^2$-test | Previous lectures |
| Continuous | Binary | Two-sample $t$-test | |
| | Categorical | (One-way) ANOVA | |
| Continuous | Continuous | Correlation analysis | This lectures |
| | | Simple linear regression | |
| Binary | Continuous | Logistic regression | (Probably) not covered |

# From Correlation to Regression

- In prediction tasks, the accuracy of our predictions depends on the strength of the relationship between variables.
- To quantify how strongly two variables move together, we use **correlation**.
- If a strong correlation is found, we then model the precise form of that relationship—this is called **regression**.

# Correlation

In studying correlation, we look at samples where each subject has provided values on two (or more) different variables.

- **Example:** Test intelligence and manual dexterity for 30 students, yielding 30 pairs of values.
- **Example:** Compare crime-rate and unemployment-rate for 20 large cities.

In each case, we examine whether larger values on one variable are associated with larger (or smaller) values on the other.

# Example: Radius vs. Circumference

For a concrete example, consider several circles of different radii.

- For each circle we measure its radius $r$ and its circumference $C$.

| Radius (cm) | $r = 1$ | $r = 3$ | $r = 5$ | $r = 8$ | $r = 10$ |
|---|---|---|---|---|---|
| Circumference (cm) | 6.28 | 18.85 | 31.41 | 50.26 | 62.83 |

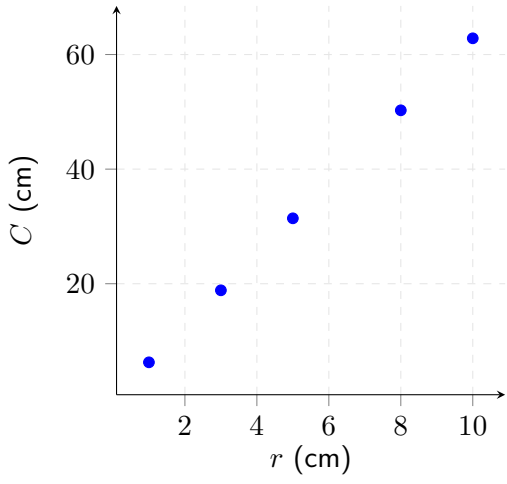
- The theoretical relationship is

$$C = 2\pi r.$$

- In practice, measurement error means the points won't lie exactly on this line, but we still expect a strong positive association: as $r$ increases, $C$ increases.

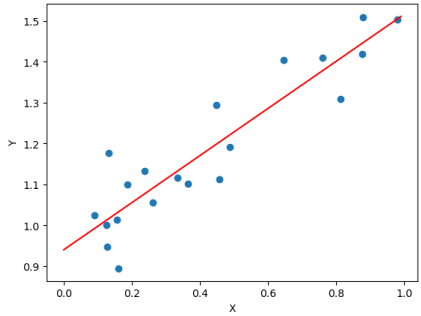
# **Example:** Illustrating the Relationship

- We can illustrate the relationship between two variables using a *scatter plot*.
- As expected, the points align very well along an increasing trend, showing a very strong relationship.
- Once a strong association is identified, we can characterize its precise form: $C = 2\pi r$.
- This formula then allows us to make predictions.



# Example: Trend Indication with a Line

- Often, the relationship is not so clear-cut.
- A scatter plot gives us a first glance at whether a trend exists.
- We can overlay a line to indicate the trend.
- (More on how to draw this line later.)

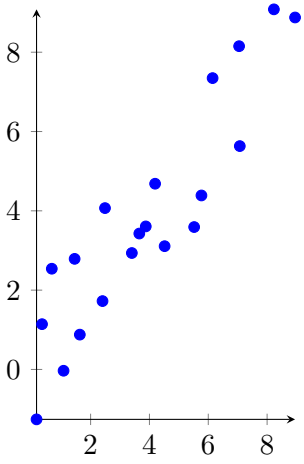


# **Example:** Types of Correlation

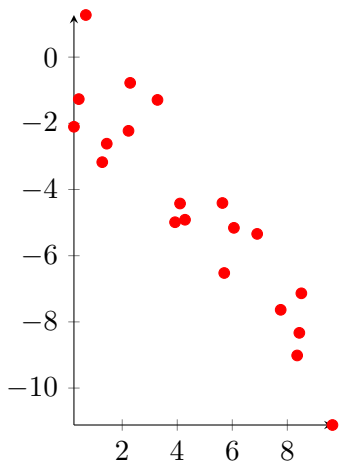
- *Positive correlation*: larger values of one variable accompany larger values of the other (**Example:** radius vs circumference).
- *Negative correlation*: larger values of one variable accompany smaller values of the other (**Example:** age vs running speed).
- *Zero correlation*: no clear tendency for the two variables to move together (**Example:** shoe size vs intelligence).

# Example: Types of Correlation

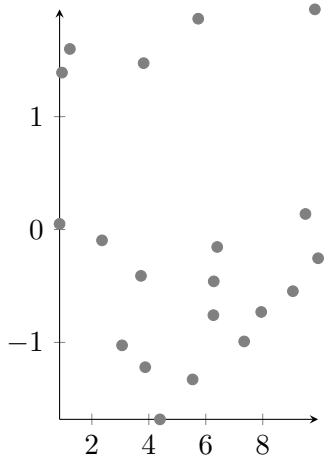
Positive correlation




Negative correlation

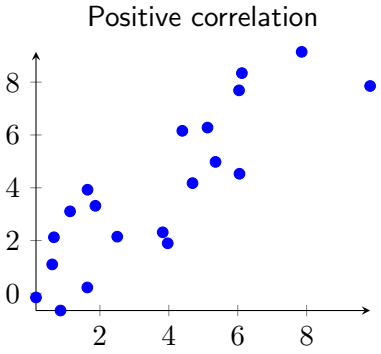


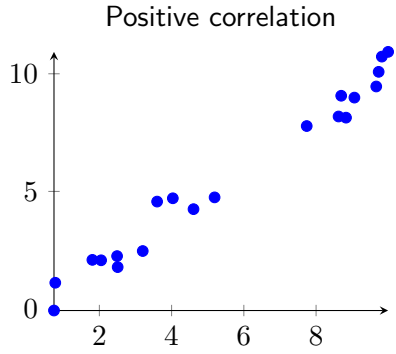

No correlation



# Correlation Coefficient

- The scatter plot not only presents you the direction of the correlation, but also a sense of the strength.
- How do we quantitatively measure the strength?



# Correlation Coefficient

Correlation coefficient is widely used as a measure of the strength and direction of the linear dependence between two variables $X$ and $Y$.

## Pearson's product-moment Correlation Coefficient

The population Pearson's correlation coefficient, denoted as $\rho$, of two variables $X$ and $Y$ is

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\mathbb{E}[(X-\mathbb{E}X)(Y-\mathbb{E}Y)]}{\sigma_X \sigma_Y}.$$

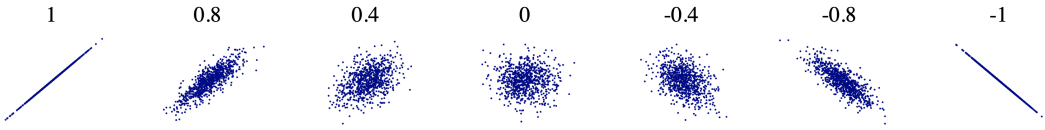
The estimated (sample) Pearson's correlation coefficient, denoted as $r$, with a sample $((X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n))$ is

$$r = \frac{\sum_{i=1}^n (X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^n (X_i-\overline{X})^2}\sqrt{\sum_{i=1}^n (Y_i-\overline{Y})^2}}$$

# Understanding Sample Correlation

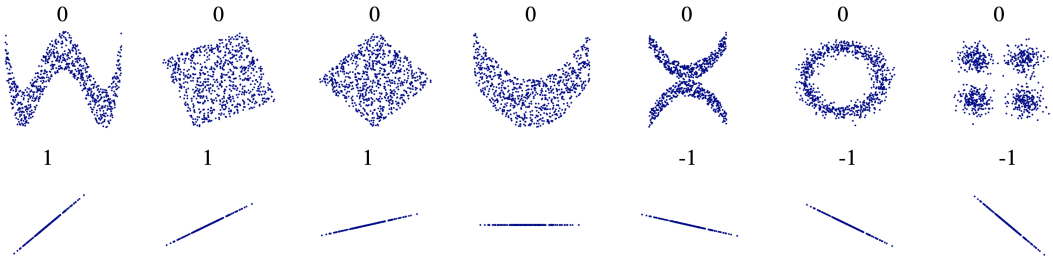
$$R = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{1}{n-1}\sum_{i=1}^{n}\frac{X_i - \bar{X}}{S_X}\frac{Y_i - \bar{Y}}{S_Y}.$$

- If $Y$ tends to increase when $X$ increases, then $(X_i - \bar{X})$ and $(Y_i - \bar{Y})$ are typically of the same sign, so $\sum(X_i - \bar{X})(Y_i - \bar{Y})$ is *large and positive*, hence $R > 0$.
- If $Y$ tends to increase when $X$ decreases, then $(X_i - \bar{X})$ and $(Y_i - \bar{Y})$ are typically of opposite sign, so $\sum(X_i - \bar{X})(Y_i - \bar{Y})$ is *large and negative*, hence $R < 0$.
- The denominator $\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}$ *scales* the numerator to ensure $R$ lies in $[-1, 1]$.




- Linear relationship between two numeric variables.
- The sign indicate the direction of the linear relationship
  - Negative $\Rightarrow$ $X$ increases, $Y$ decreases.
  - Positive $\Rightarrow$ $X$ increases, $Y$ increases.
- The absolute value indicate the strength of the linear relationship
  - Larger the absolute value $\Rightarrow$ stronger linear relationship.
  - $-1 \Rightarrow$ perfectly negative correlation.
  - $+1 \Rightarrow$ perfectly positive correlation.
  - $0 \Rightarrow$ the variables are not linearly correlated.

# Potential Problems of the Correlation Coefficient




- The correlation coefficient can only detect if a linear relationship exists.
- The value of these coefficients really does not tell us about the exact relationship, other than an abstract summary such as "$Y$ is some liner function of $X$."

$\Rightarrow$ **Correlation analysis cannot be used to predict $Y$ with $X$!**

# From Correlation to Regression

The sample correlation coefficient $R$ measures the *strength* and *direction* of a linear association between the explanatory variable $X$ and the response variable $Y$. However, it does *not* tell us the *specific* form of that relationship or how to make predictions.

- Pearson's correlation coefficient$R$ is a summary statistic: it quantifies *how closely* the data lie along some line, i.e., if a linear relationship between the explanatory variable $X$ and and the response variable $Y$.
- However, it does not tell us the *specific* form of that relationship or how to make predictions.
- To identify the precise linear relation, we use **simple linear regression**.

# Introduction – Simple Linear Regression

## Linear relationship

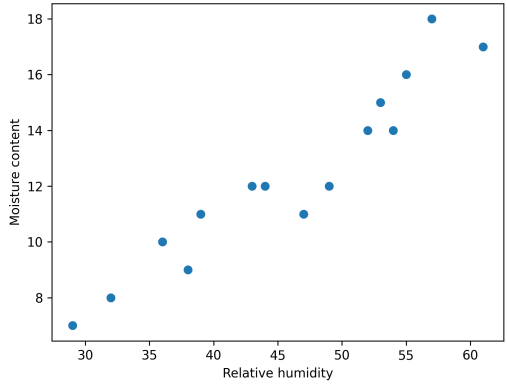
Linear relationship can be summarized by two numbers, the intercept $\alpha$ and the slope $\beta$:

$$Y = \alpha + \beta X$$

- The **intercept** $\alpha$ is the value of $Y$ when the line crosses the $y$-axis, i.e. $Y$ value when $X = 0$.
- The **slope** $\beta$ is a measure of the *steepness* of a line, i.e. the change in $Y$ when $X$ changes by one unit.

**Example:** The raw material used in the production of a certain synthetic fiber is stored in a location without a humidity control. Measurements of the relative humidity ($X$) in the storage location and the moisture content ($Y$) of a sample of the raw material were taken over 15 days with the following data.

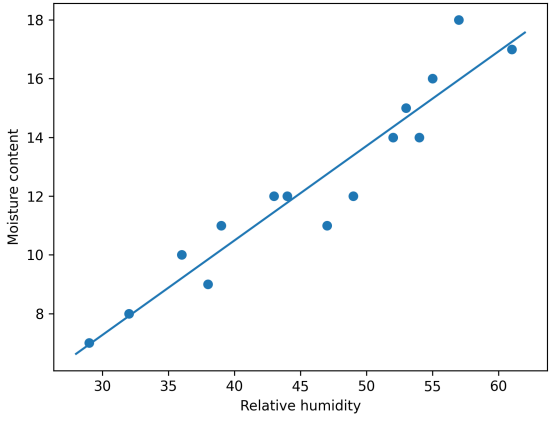
- The Pearson's correlation coefficient $\rho = 0.95$.
- Perform $T$-test for the Pearson's correlation coefficient, we have $p \approx 0$.
- There are strong positive linear relationship!



If we "fit" a straight line to the scatter plot, we may have

$$Y = -2.38 + 0.32X.$$

- **How do we find the "best" line that fits our data?**
- Notice that no matter which line we choose, there is always error!
- **How good does the line "explain/predict" $Y$?**



# Simple Linear Regression

To address the error, one choice is to include the error in the model!

## Simple linear regression model

We assume that the $i$-th observation $(Y_i, X_i)$ follows

$$\begin{array}{ccccc} Y_i & = & \alpha + \beta X_i & + & \varepsilon_i \\ \text{(Response)} & = & \text{(Linear Model)} & + & \text{(Error)} \end{array}$$

Note that $\alpha$ and $\beta$ are shared across different observations.

Let us answer first the question below:

**How do we find the "best" line that fits our data?**

# Intuition

For a line $\alpha + \beta X$ to fit the data well, we wish that the error

$$\varepsilon_i = Y_i - (\alpha + \beta X_i)$$

are "minimized" across all the observed data.

We shall measure the overall error by the sum of squared error ($SS_E$), or residual sum of squares (RSS):

$$SS_E = RSS = \sum_{i=1}^{n} \varepsilon^2 = \sum_{i=1}^{n} (Y_i - (\alpha + \beta X_i))^2$$

- We want to find the $\alpha$ and $\beta$ such that the $SS_E$ is minimized.

# Ordinary Least Square Estimators

The estimators that can minimize the $SS_E$ are called the (ordinary) least square[1] (OLS) estimators. The name is self-explanatory.

For a dataset $\{(X_i, Y_i) : i = 1, 2, \ldots, n.\}$, let's calculate the OLS estimators.

[1] There is a generalization of this method called the generalized least square (GLS) estimation.

# Example: Ordinary Least Squares (OLS) Estimators

The OLS estimators $\hat{\alpha}$, $\hat{\beta}$ minimize the sum of squared errors

$$SS_E(\alpha, \beta) = \sum_{i=1}^{n} \big(Y_i - (\alpha + \beta X_i)\big)^2.$$

Compute the first-order conditions:

$$\frac{\partial SS_E}{\partial \alpha} = -2 \sum_{i=1}^{n} \big(Y_i - (\alpha + \beta X_i)\big) = -2n\bar{Y} + 2n\alpha + 2\beta n\bar{X} = 0,$$

$$\frac{\partial SS_E}{\partial \beta} = -2 \sum_{i=1}^{n} X_i \big(Y_i - (\alpha + \beta X_i)\big) = -2 \sum_{i=1}^{n} X_i Y_i + 2\alpha n\bar{X} + 2\beta \sum_{i=1}^{n} X_i^2 = 0.$$

Solving these two **normal equations** simultaneously to derive the OLS estimates.

# Example: Derivation of the Normal Equations

From the normal equations:

$$-2n\bar{Y} + 2n\alpha + 2\beta\, n\bar{X} = 0 \quad \Longrightarrow \quad \alpha + \beta\,\bar{X} = \bar{Y} \quad \Longrightarrow \quad \alpha = \bar{Y} - \beta\,\bar{X}.$$

$$-2\sum_{i=1}^{n} X_i Y_i + 2\alpha\, n\bar{X} + 2\beta \sum_{i=1}^{n} X_i^2 = 0 \quad \Longrightarrow \quad \sum_{i=1}^{n} X_i Y_i = \alpha\, n\bar{X} + \beta \sum_{i=1}^{n} X_i^2.$$

Substitute $\alpha = \bar{Y} - \beta\bar{X}$:

$$\sum_{i=1}^{n} X_i Y_i = (\bar{Y} - \beta\bar{X})\, n\bar{X} + \beta \sum_{i=1}^{n} X_i^2 = n\bar{X}\,\bar{Y} - \beta\, n\bar{X}^2 + \beta \sum_{i=1}^{n} X_i^2$$

Rearranging gives

$$\beta = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\,\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} \quad \text{and} \quad \alpha = \bar{Y} - \beta\,\bar{X}.$$

Let $\bar{X}$ and $\bar{Y}$ be the sample means and (see next slide for the derivation)

$$S_{xx} = \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2, \qquad S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2$$

$$S_{xY} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}$$

## Least Square Estimators

$$\widehat{\beta} = \frac{S_{xY}}{S_{xx}}, \quad \widehat{\alpha} = \bar{Y} - \widehat{\beta}\bar{X}, \quad SS_E = \sum_{i=1}^{n}(Y_i - \widehat{\alpha} - \widehat{\beta}X_i)^2 = \frac{S_{xx}S_{YY} - S_{xy}^2}{S_{xx}}$$

Here we abused the notation $SS_E$ to denote the $SS_E$ measured for the best line as specified by the OLS estimator.

# Example: Sums of Squares Identities

We show the standard shortcuts for centering sums:

$$S_{xx} = \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}\left(X_i^2 - 2\bar{X}\,X_i + \bar{X}^2\right) = \sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum_{i=1}^{n} X_i + n\bar{X}^2$$

Since $\sum_{i=1}^{n} X_i = n\bar{X}$, this becomes

$$S_{xx} = \sum_{i=1}^{n} X_i^2 - 2\,\bar{X}\,(n\bar{X}) + n\bar{X}^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2.$$

Similarly,

$$S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2,$$

and for the cross-term,

$$S_{xY} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n}\left(X_iY_i - \bar{X}\,Y_i - \bar{Y}\,X_i + \bar{X}\,\bar{Y}\right) = \sum_{i=1}^{n} X_iY_i - n\bar{X}\,\bar{Y}.$$

# *An Alternative Way to Derive OLS Estimator

We know that the mean minimize the squared error, so for any fixed $\beta$, letting $\alpha = \widehat{\alpha}(\beta) = \bar{Y} - \beta\bar{X}$ minimizes $SS_E$ as a function of $\alpha$

$$\begin{aligned}
\min_{\alpha} SS_E(\alpha) &= \sum_{i=1}^{n}((Y_i - \beta X_i) - \widehat{\alpha}(\beta))^2 \\
&= \sum_{i=1}^{n}((Y_i - \beta X_i) - (\bar{Y} - \beta\bar{X}))^2 \\
&= \sum_{i=1}^{n}((Y_i - \bar{Y}) - \beta(X_i - \bar{X}))^2 \\
&= S_{YY} - 2\beta S_{xY} + \beta^2 S_{xx}
\end{aligned}$$

So choosing $\beta = \widehat{\beta} = \frac{S_{xY}}{S_{xx}}$ minimize the above.

# Coefficient of Determination

Having found the best line to fit the data, we seek to answer the second question:

**How good does the line "explain" $Y$?**

## Residual

$$\text{Residual} = Y_i - \widehat{Y}_i = Y_i - \widehat{\alpha} - \widehat{\beta} X_i$$

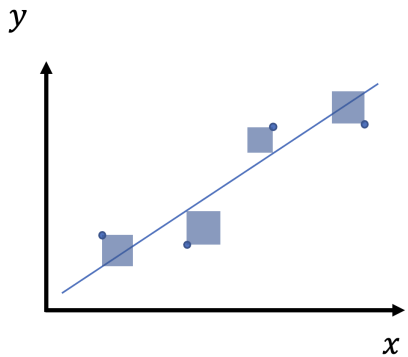
here $\widehat{Y}_i$ is called the fitted value of the $i$-th observation.

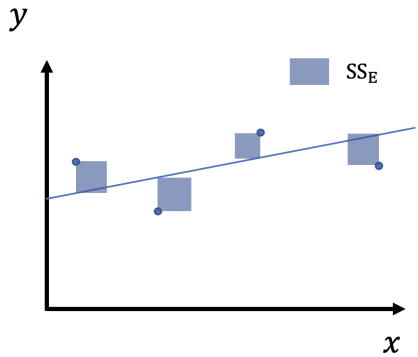
Taking the sum of squares, we have a characterization of the overall error that remains after our regression.

## Residual sum of squares (also called Error sum of squares)

$$SS_E = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2$$

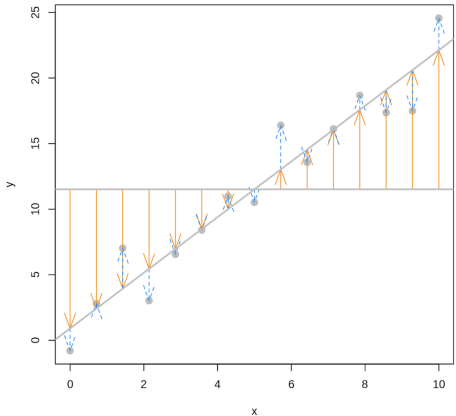
# Residual (Error) Sum of Square



We can consider the following decomposition

$$\begin{aligned}(\text{Deviation}) &= Y_i - \overline{Y} \\ &= (Y_i - \widehat{Y}_i) + (\widehat{Y}_i - \overline{Y}) \\ &= (\text{Residual/Error}) + (\text{Regression})\end{aligned}$$



As usual, we shall measure these in terms of the sum of squares.

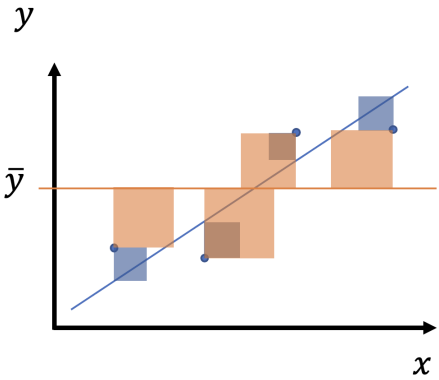
## Decomposition of the Total Sum of Squares ($SS_T$)

$$\text{Total SS} = \text{Regression SS} + \text{Residual/Error SS}$$

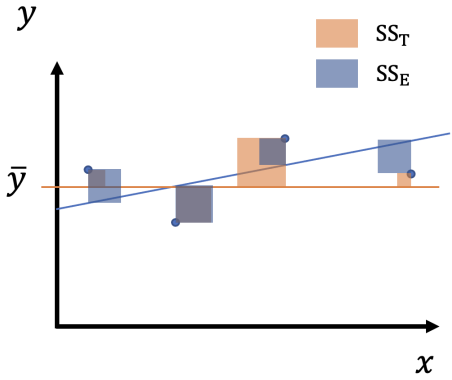
$$SS_T = SS_R + SS_E$$

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

- The cross term $2\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})(Y_i - \widehat{Y}_i)$ vanishes. *A brute-force proof is not entirely straightforward. It is usually proved using linear algebra arguements.




Well explained




Poorly explained

## Coefficient of Determination

$$R^2 = \frac{SS_T - SS_E}{SS_T} = \frac{SS_R}{SS_T}$$

Interpretation of $R^2$

- Indicate how well the variation in $Y$ is explained by $X$.
- Interpreted as the proportion of total variation in the response variable $Y$ that is "explained" by the regressors $X$ in the model.
- $0 \leq R^2 \leq 1$
- $R^2 = 1$: The data fall exactly on a straight line.
- $R^2 = 0$: No "linear" relationship.

# *Connection with Sample Correlation

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} = \frac{S_{xY}}{\sqrt{S_{xx} S_{YY}}}$$

Recall that $SS_E = \sum_{i=1}^n (Y_i - \widehat{\alpha} - \widehat{\beta} X_i)^2 = \frac{S_{xx} S_{YY} - S_{xy}^2}{S_{xx}}$.

$$r^2 = \frac{S_{xy}^2}{S_{xx} S_{YY}} = \frac{S_{xx} S_{YY} - S_{xx} SS_E}{S_{xx} S_{YY}} = \frac{SS_T - SS_E}{SS_T} = R^2$$

**Coefficient of Determination is the square of (Pearson's) coefficeint of correlation!**

Up until this point, we have answered two main questions:

- **How do we find the "best" line that fits our data?**
- **How good does the line "explain" $Y$?**

We will need more probability assumptions in order to answer other statistical questions such as

- construct confidence interval for the parameters;
- perform statistical hypothesis tests.

# Probabilistic Modeling of the Simple Linear Regression

**Assumptions**

- **Linearity**: The data actually exhibit a linear relationship, i.e., $Y_i = \alpha + \beta x_i + \varepsilon$.
- **Independency**: The error $\{\varepsilon_i : i = 1, 2, \ldots, n\}$ are independent.
- **Homoscedasticity** (Equal variance): the variance of the error $\varepsilon_i$ should be the same.
- **Normality**: each error $\varepsilon_i$ is normally distributed.

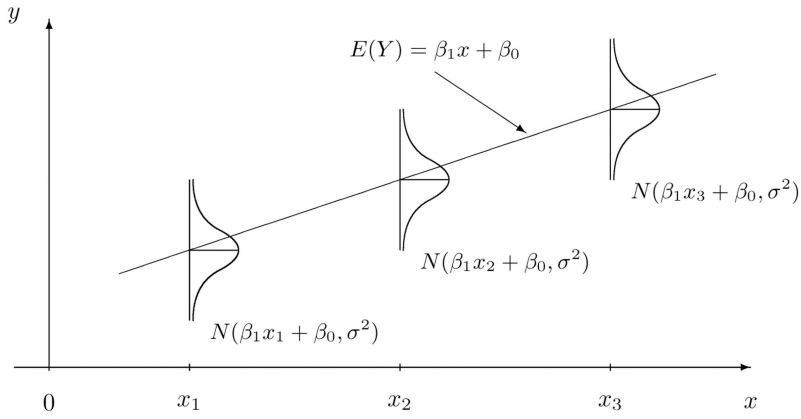
The last 3 assumptions can summarized as

$$\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2), \quad i = 1, 2, \ldots, n.$$

**We usually assume the independent variable $x_i$ to be deterministic numbers rather than random variables. That's why we shall use lower case.**

The above assumptions implies that

$$Y_i = \alpha + \beta x_i + \varepsilon_i \sim N(\alpha + \beta x_i, \sigma^2), \quad i = 1, 2, \ldots, n.$$



# Maximum Likelihood Estimator for $\alpha$ and $\beta$

Because of the i.i.d. assumption, the likelihood of the observation is

$$L(\alpha, \beta, \sigma; x, Y) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \alpha - \beta x_i)^2\right).$$

Take logarithm

$$l(\alpha, \beta, \sigma; x, Y) = \log L(\alpha, \beta; x, y) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \alpha - \beta x_i)^2.$$

Maximizing the log-likelihood with respect to $\alpha$ and $\beta$ is equivalent to minimizing the term

$$\sum_{i=1}^{n}(Y_i - \alpha - \beta x_i)^2$$

**This is exactly what we did in OLS estimation!**

$$\frac{\partial \log L(\alpha,\beta,\sigma^2)}{\partial \alpha} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \alpha - \beta x_i)$$

$$\frac{\partial \log L(\alpha,\beta,\sigma^2)}{\partial \beta} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i)(Y_i - \alpha - \beta x_i)$$

$$\frac{\partial \log L(\alpha,\beta,\sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(Y_i - \alpha - \beta x_i)^2$$

## MLE

The maximum likelihood estimators for $\alpha$, $\beta$ and $\sigma^2$ are

$$\widehat{\beta}_{\mathsf{MLE}} = \frac{S_{xY}}{S_{xx}}, \quad \widehat{\alpha}_{\mathsf{MLE}} = \bar{Y} - \widehat{\beta}_{\mathsf{MLE}}\bar{x},$$

$$\widehat{\sigma}^2_{\mathsf{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{\alpha}_{\mathsf{MLE}} - \widehat{\beta}_{\mathsf{MLE}}x_i)^2 = \frac{SS_E}{n} = \frac{S_{xx}S_{YY} - S_{xY}^2}{nS_{xx}}$$

# Estimation of the Variance

We have derived the MLE $\widehat{\sigma}^2_{\text{MLE}}$ of the variance. However, this estimator is actually biased. In the setting of simple linear regression, people usually use an unbiased version of it instead:

$$\widehat{\sigma}^2 = s_y^2 = \frac{\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2}{n-2} = \frac{SS_E}{n-2}.$$

**Notation simplification**

- Since the MLE for $\alpha$ and $\beta$ coincides with that derived from the OLS method, we shall omit the subscript and write

$$\widehat{\alpha} = \widehat{\alpha}_{\text{MLE}}, \quad \widehat{\beta} = \widehat{\beta}_{\text{MLE}}$$

**Theorem (Distributions of the Estimators)**

*$\widehat{\alpha}$, $\widehat{\beta}$ and $\widehat{\sigma}^2$ have the following distributions*

$$\widehat{\alpha} \sim N\left(\alpha, \frac{\sigma^2 \sum_{i=1}^n x_i^2}{nS_{xx}}\right), \quad \widehat{\beta} \sim N(\beta, \frac{\sigma^2}{S_{xx}}), \quad \frac{(n-2)\widehat{\sigma}^2}{\sigma^2} = \frac{SS_E}{\sigma^2} \sim \chi^2_{n-2}$$

*Moreover, $Cov(\widehat{\alpha}, \widehat{\beta}) = \frac{-\sigma^2 \bar{x}}{S_{xx}}$ and $\widehat{\sigma}^2$ is independent of $(\widehat{\alpha}, \widehat{\beta})$.*

- $\widehat{\alpha}$ and $\widehat{\beta}$ are just linear combination of $Y_i$'s, it is straightforward (but tiresome) to find the marginal and joint distribution.
- The proof for $\frac{SS_E}{\sigma^2}$ is a bit complicated.

# Hypothesis Test for the Slope $\beta$

In the simple linear regression model

$$Y = \alpha + \beta x + \varepsilon.$$

We are interested in testing the hypothesis $\beta = 0$. (What does it mean?)

Since $\widehat{\beta} \sim N(\beta, \frac{\sigma^2}{S_{xx}})$, we have

$$\frac{\widehat{\beta} - \beta}{\sqrt{\sigma^2/S_{xx}}} \sim N(0,1)$$

Since $\frac{SS_E}{\sigma^2} = \frac{(n-2)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$, we conclude (by the definition of $t$-distribution) that

$$\frac{\widehat{\beta} - \beta}{\sqrt{\widehat{\sigma}^2/S_{xx}}} = \frac{(\widehat{\beta} - \beta)/\sqrt{\sigma^2/S_{xx}}}{\sqrt{\widehat{\sigma}^2/\sigma^2}} = \frac{N(0,1)}{\sqrt{\chi^2_{n-2}/(n-2)}} \sim T_{n-2}$$

**Null hypothesis**:

$$H_0 : \beta = 0$$

**Test statistic**

$$T = \sqrt{\frac{S_{xx}}{\widehat{\sigma}^2}}\widehat{\beta} = \sqrt{\frac{(n-2)S_{xx}}{SS_E}}\widehat{\beta} \sim T_{n-2} \text{ under } H_0$$

| $H_1$ | Rejection region | $p$-value |
|---|---|---|
| $\beta \neq 0$ | $\lvert t\rvert > t_{n-2,\gamma/2}$ | $P(\lvert T\rvert > \lvert t\rvert \mid H_0)$ |
| $\beta > 0$ | $t > t_{n-2,\gamma}$ | $P(T > t \mid H_0)$ |
| $\beta < 0$ | $t < t_{n-2,\gamma}$ | $P(T < t \mid H_0)$ |

*$T \sim t_{n-2}$, $\mathbb{P}(T > t_{n-2,\gamma}) = \gamma$.

**Example:** An individual claims that the fuel consumption of his automobile does not depend on how fast the car is driven. To test the plausibility of this hypothesis, the car was tested at various speeds between 45 and 70 miles per hour. The miles per gallon (MPG) attained at each of these speeds was as the follows:

| Speed | Miles per Gallon |
|---|---|
| 45 | 24.2 |
| 50 | 25.0 |
| 55 | 23.3 |
| 60 | 22.0 |
| 65 | 21.5 |
| 70 | 20.6 |
| 75 | 19.8 |

Let $Y$ denote MPG and $x$ denote the speed. Suppose a simple linear model

$$Y = \alpha + \beta x + \varepsilon.$$

We want to test $H_0 : \beta = 0 \quad \text{v.s.} \quad H_1 : \beta \neq 0.$

Calculate

$$s_{xx} = 700, \quad S_{YY} = 21.757, \quad S_{xY} = -119$$

According to the formulas

$$ss_E = [s_{xx}S_{YY} - s_{xY}^2]/s_{xx} = [700(21.757) - (-119)^2]/700 = 1.527$$
$$\widehat{\beta} = S_{xY}/s_{xx} = -119/700 = -0.17$$

So

$$t = \sqrt{\frac{(7-2)s_{xx}}{ss_E}}\widehat{\beta} = -8.139$$

The $p$-value$= 2P\{T_{n-2} > |t|\} = 0.00045$. Reject $H_0$ for all $\alpha > 0.00045$.

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | mpg | R-squared: | 0.930 |
| Model: | OLS | Adj. R-squared: | 0.916 |
| Method: | Least Squares | F-statistic: | 66.23 |
| Date: | | Prob (F-statistic): | 0.000455 |
| Time: | | Log-Likelihood: | -4.6038 |
| No. Observations: | 7 | AIC: | 13.21 |
| Df Residuals: | 5 | BIC: | 13.10 |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 32.5429 | 1.271 | 25.612 | 0.000 | 29.277 | 35.809 |
| speed | -0.1700 | 0.021 | -8.138 | 0.000 | -0.224 | -0.116 |

| | | | |
|---|---|---|---|
| Omnibus: | nan | Durbin-Watson: | 2.472 |
| Prob(Omnibus): | nan | Jarque-Bera (JB): | 0.604 |
| Skew: | 0.708 | Prob(JB): | 0.739 |
| Kurtosis: | 3.253 | Cond. No. | 370. |

# ANOVA Table for Simple Linear Regression

**Null and Alternative Hypotheses**:

$$H_0 : \beta = 0, \ H_1 : \beta \neq 0$$

**ANOVA table**

| Source | df | SS | MS |
|---|---|---|---|
| Regression | 1 | $\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$ | $SS_R/1$ |
| Error | $n-2$ | $\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$ | $SS_E/(n-2)$ |
| Total | $n-1$ | $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ | |

**Test Statistics**:

$$F = \frac{MS_R}{MS_E} \sim F_{1,n-2} \text{ under } H_0$$

Reject $H_0$ if the observed statistic $f > F_{1,n-2,\alpha}$ where $\mathbb{P}(F > F_{1,n-2,\alpha}) = \alpha$

# Connection Between ANOVA Table and $T$-Test

Recall the $T$-statistic for testing $\beta$:

$$T^2 = \left(\sqrt{\frac{S_{xx}}{\widehat{\sigma}^2}}\widehat{\beta}\right)^2 = \frac{S_{xx}}{SS_E/(n-2)}\frac{S_{xY}^2}{S_{xx}^2} = \frac{S_{xY}^2/S_{xx}}{SS_E/(n-2)} = \frac{SS_R}{SS_E/(n-2)} = F$$

Above follows from $SS_E = \frac{S_{xx}S_{YY}-S_{xY}^2}{S_{xx}} = S_{YY} - S_{xY}^2/S_{xx} = SS_T - S_{xY}^2/S_{xx}$, so $MS_R = SS_R/1 = S_{xY}^2/S_{xx}$.

In the ANOVA test and (two-sided) $T$-test for the significance of linear regression

- The $F$-statistic and $T$-statistic is connected by $T^2 = F$.
- The two tests are equivalent.

# Derivation of the Error Sum of Squares

Recall the definitions

$$S_{xx} = \sum_{i=1}^{n}(X_i - \bar{X})^2, \quad S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2, \quad S_{xY} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}),$$

and the OLS slope

$$\hat{\beta} \;=\; \frac{S_{xY}}{S_{xx}}.$$

Substitute $\hat{\alpha}$:

$$Y_i - \hat{\alpha} - \hat{\beta}X_i = Y_i - (\bar{Y} - \hat{\beta}\,\bar{X}) - \hat{\beta}X_i = (Y_i - \bar{Y}) - \hat{\beta}\,(X_i - \bar{X}).$$

Hence

$$\begin{aligned}
\mathrm{SS}_E &= \sum_{i=1}^{n}\left[(Y_i - \bar{Y}) - \hat{\beta}\,(X_i - \bar{X})\right]^2 \\
&= \sum (Y_i - \bar{Y})^2 - 2\,\hat{\beta}\sum (Y_i - \bar{Y})(X_i - \bar{X}) + \hat{\beta}^2 \sum (X_i - \bar{X})^2 \\
&= S_{yy} - 2\,\hat{\beta}\,S_{xy} + \hat{\beta}^2\,S_{xx}.
\end{aligned}$$

# Derivation of the Error Sum of Squares

Substituting $\hat{\beta} = S_{xY}/S_{xx}$ yields

$$\mathrm{SS}_E = S_{YY} - \frac{S_{xY}}{S_{xx}} S_{xY} = S_{YY} - \frac{S_{xY}^2}{S_{xx}} = \frac{S_{xx}\, S_{YY} - S_{xY}^2}{S_{xx}}.$$

# Confidence Interval for $\beta$

The $(1-\gamma)\times 100\%$ confidence interval for $\beta$ can be constructed as before using the sample distribution of $\widehat{\beta}$ as before. We can write

$$\mathbb{P}\left(-t_{n-2,\gamma/2} \leq \frac{\widehat{\beta}-\beta}{\sqrt{\widehat{\sigma}^2/S_{xx}}} \leq t_{n-2,\gamma/2}\right) = \mathbb{P}(-t_{n-2,\gamma/2} \leq T \leq t_{n-2,\gamma/2}) = \gamma.$$

Rearrange terms and we obtain the two-sided CI

$$\beta \in \left[\widehat{\beta} - t_{n-2,\gamma/2}\widehat{\sigma}\sqrt{\frac{1}{S_{xx}}}, \widehat{\beta} + t_{n-2,\gamma/2}\widehat{\sigma}\sqrt{\frac{1}{S_{xx}}}\right].$$

Similarly, we can develop one-sided CIs

$$\beta \in \left(-\infty, \widehat{\beta} + t_{n-2,\gamma}\widehat{\sigma}\sqrt{\frac{1}{S_{xx}}}\right], \quad \beta \in \left[\widehat{\beta} - t_{n-2,\gamma}\widehat{\sigma}\sqrt{\frac{1}{S_{xx}}}, +\infty\right).$$

# Hypothesis Tests for $\alpha$

Reacall that

$$\widehat{\alpha} \sim N\left(\alpha, \frac{\sigma^2 \sum_{i=1}^n x_i^2}{nS_{xx}}\right)$$

We have

$$\frac{\widehat{\alpha} - \alpha}{\sigma\sqrt{\frac{\sum_{i=1}^n x_i^2}{nS_{xx}}}} \sim N(0,1), \quad \frac{\widehat{\alpha} - \alpha}{\widehat{\sigma}\sqrt{\frac{\sum_{i=1}^n x_i^2}{nS_{xx}}}} \sim T_{n-2}.$$

**Null hypothesis**:

$$H_0 : \alpha = 0.$$

**Test statistic**

$$T = \frac{\widehat{\alpha}}{\widehat{\sigma}\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}}}} \sim T_{n-2} \text{ under } H_0$$

| $H_1$ | Rejection region | $p$-value |
|---|---|---|
| $\alpha \neq 0$ | $\lvert t\rvert > t_{n-2,\gamma/2}$ | $P(\lvert T\rvert > \lvert t\rvert \mid H_0)$ |
| $\alpha > 0$ | $t > t_{n-2,\gamma}$ | $P(T > t \mid H_0)$ |
| $\alpha < 0$ | $t < t_{n-2,\gamma}$ | $P(T < t \mid H_0)$ |

*$T \sim t_{n-2}$, $\mathbb{P}(T > t_{n-2,\gamma}) = \gamma$.

# Confidence Interval for $\alpha$

Two-sided

$$\alpha \in \left[ \widehat{\alpha} - t_{n-2,\gamma/2}\widehat{\sigma}\sqrt{\frac{\sum_{i=1}^n x_i^2}{nS_{xx}}}, \widehat{\alpha} + t_{n-2,\gamma/2}\widehat{\sigma}\sqrt{\frac{\sum_{i=1}^n x_i^2}{nS_{xx}}} \right].$$

One-sided

$$\alpha \in \left( -\infty, \widehat{\alpha} + t_{n-2,\gamma}\widehat{\sigma}\sqrt{\frac{\sum_{i=1}^n x_i^2}{nS_{xx}}} \right], \quad \alpha \in \left[ \widehat{\alpha} - t_{n-2,\gamma}\widehat{\sigma}\sqrt{\frac{\sum_{i=1}^n x_i^2}{nS_{xx}}}, +\infty \right).$$

# Motivation

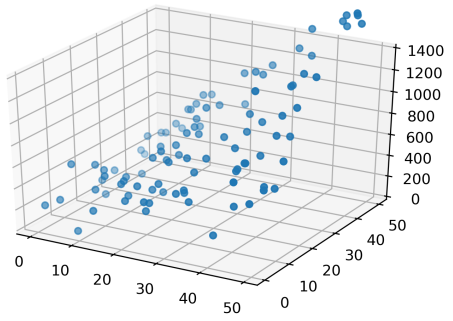
We have talked about how to find linear relationship between the dependent variable $Y$ and a single predictor $X$:

$$Y = \alpha + \beta X + \varepsilon.$$

- If we have multiple predictors that may help to predict $Y$, how can we generalize the model above to incorporate all of the variables in a linear model?

**Example:** We want to model/predict the sales $Y$ of a product, and we have data of the advertisement budget spent in multiple ways: from TV $X_1$, and on the newspapers $X_2$ and from the radio $X_3$.

- We can plot a scatter plot for the data of sales versus the two ads.
- We can clearly see an increasing trend for sales with respect to the increase of ad budget in both media formats. Can we fit a linear function to it?




We may want to use the following

$$\texttt{Sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{Newspaper} + \beta_3 \times \texttt{Radio} + \epsilon.$$

The interpretation of $\beta_i$ is, if the budget spent on media $i$ is increased by one unit, while everything else is fixed, then on average sales is increased by $\beta_i$ units.

# Multiple Linear Regression

Given

- A single dependent/target variable $Y$.
  - Let $Y_i$ denote the dependent variable for the $i$-th observation.
- Several independent/explanatory variables $x_1, x_2, \ldots, x_p$.
  - Let $x_{ij}$ denote the $j$-the dependent variable for the $i$-th observation.

Assuming a general linear regression model: for $i = 1, \ldots, n$

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

- $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ is the random error of the $i$-th observation.

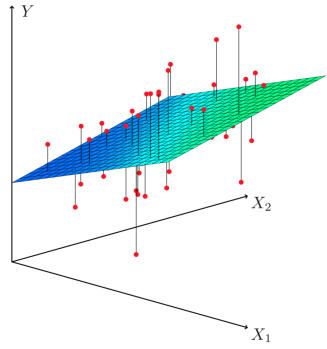
Want

- Estimation: the parameters $\beta_0, \ldots, \beta_p$ that best describe this linear dependence, and the variance $\sigma^2$ of the error.
- Inference: confidence interval, hypothesis testing for the parameters.

# Ordinary Least Square Estimator

To esitimate $\beta_i$, minimize the RSS

$$\begin{aligned}\text{RSS} = SS_E &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \\ &= \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip})^2\end{aligned}$$




This can be done through Python:

- `sklearn.linear_model.LinearRegression`
- `statsmodels.regression.linear_model.OLS`

# Example:

```
import statsmodels.formula.api as smf
import numpy as np
import pandas as pd

# load data
datas_url = 'https://www.statlearning.com/s/Advertising.csv'
df = pd.read_csv(datas_url).drop('Unnamed: 0',axis=1)

df.head()

# fit linear regression model
res = smf.ols('sales ~ TV + radio + newspaper', data=df).fit()
res.summary()
```

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | sales | R-squared: | 0.897 |
| Model: | OLS | Adj. R-squared: | 0.896 |
| Method: | Least Squares | F-statistic: | 570.3 |
| Date: | Wed, 07 May 2025 | Prob (F-statistic): | 1.58e-96 |
| Time: | 12:50:53 | Log-Likelihood: | -386.18 |
| No. Observations: | 200 | AIC: | 780.4 |
| Df Residuals: | 196 | BIC: | 793.6 |
| Df Model: | 3 | | |
| Covariance Type: | nonrobust | | |

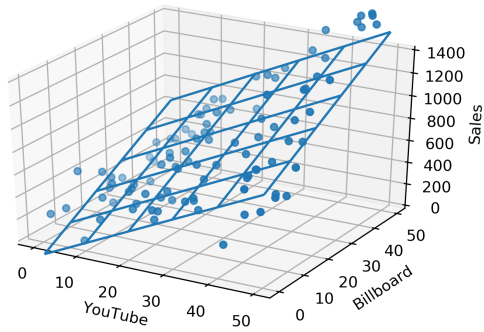
| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 2.9389 | 0.312 | 9.422 | 0.000 | 2.324 | 3.554 |
| TV | 0.0458 | 0.001 | 32.809 | 0.000 | 0.043 | 0.049 |
| radio | 0.1885 | 0.009 | 21.893 | 0.000 | 0.172 | 0.206 |
| newspaper | -0.0010 | 0.006 | -0.177 | 0.860 | -0.013 | 0.011 |

| | | | |
|---|---|---|---|
| Omnibus: | 60.414 | Durbin-Watson: | 2.084 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 151.241 |
| Skew: | -1.327 | Prob(JB): | 1.44e-33 |
| Kurtosis: | 6.332 | Cond. No. | 454. |

- The Coefficient of Determination

$$R^2 = \frac{SS_T - SS_E}{SS_T} = 0.842.$$

  - The larger $R$ is, the better the model fits the data.
- The $F$-statistic $= 259.1$.
  - The larger $F$ is, the more significant the model is in compare with simple model with only a intercept.



## Null and Alternative Hypotheses

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \; vs. \; H_1 : \text{at least one } \beta_i \text{ is not 0}$$

## ANOVA Table

| Source | df | SS | MS |
|---|---|---|---|
| Regression | $p$ | $\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$ | $SS_R/p$ |
| Residual | $n-(p+1)$ | $\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$ | $SS_E/(n-(p+1))$ |
| Total | $n-1$ | $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ | |

## Test Statistics

$$F = \frac{MS_R}{MS_E} \sim F_{p,n-(p+1)} \text{ under } H_0$$

Reject $H_0$ if $f > F_{p,n-(p+1),\alpha}$ where $\mathbb{P}(F > F_{p,n-(p+1),\alpha}) = \alpha$

# Regression Diagnostics

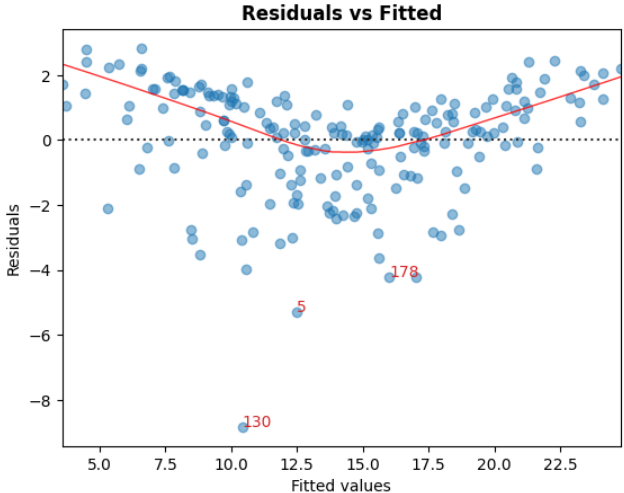
Recall that linear regression model have assumptions:

- **Linearity**: The data actually exhibit a linear relationship.
- **Independency**: Observations should be independent.
- **Homoscedasticity**: The variance of the errors $\varepsilon$ should be constant across all levels of the independent variables.

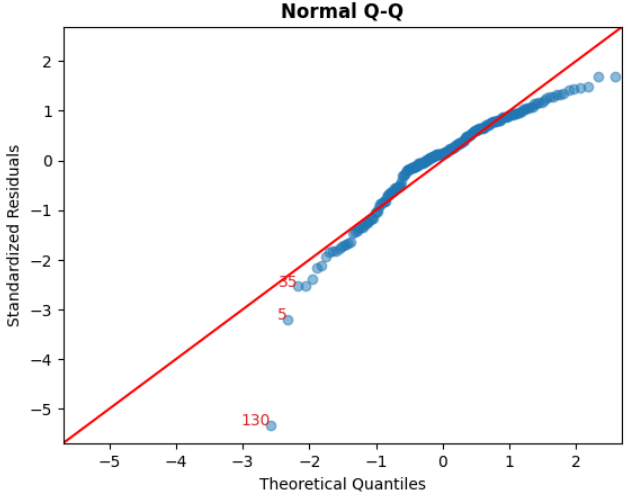
For all our statistical analysis to be valid, these assumptions must be met by the data. We need **diagnostic plots** to validate these assumptions.

# Diagnose Plots for Linear Regression

**Plot residuals against fitted value**: Check the assumption of linear model.

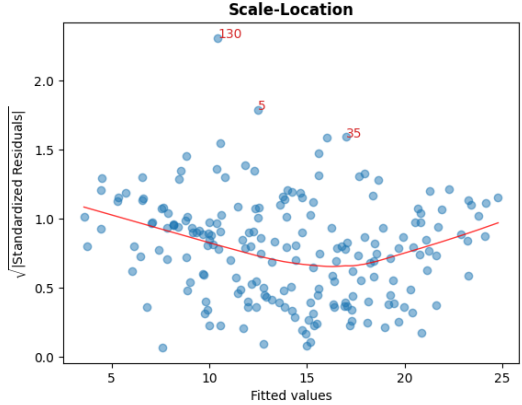


**Q-Q plot for the residuals**: Check the assumption of normality.

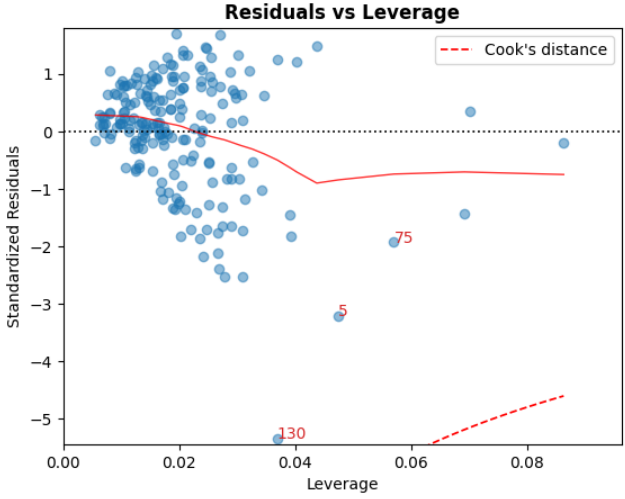


**Scale-location plot**: plot the scale (standard deviation) of the residual against the location (fitted value of $Y$). Check the assumption of equal variance.

$$\text{studendized residual} = \frac{\text{residual}}{\text{sample std. dev. of residual}}$$

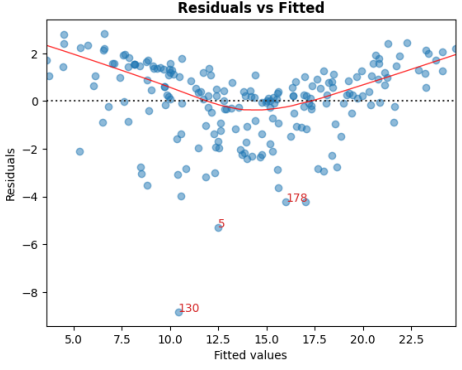


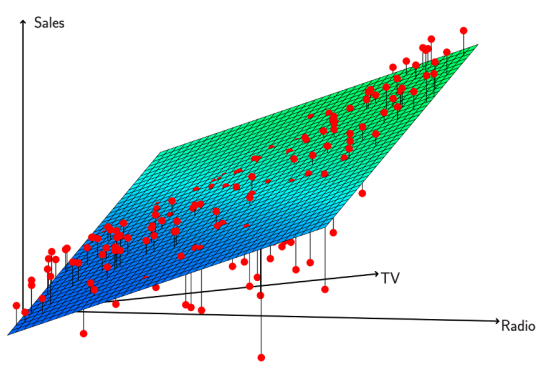
**Leverage plot**: Check the outliers.



# Non-Linear Relationship

If the relationship between $Y$ and $X_1, X_2, \ldots, X_p$ is nonlinear, can we find a way to still use linear models to model the relationship?



# Modeling Synergy via Interaction Terms

In our advertising example the simple additive model

$$\hat{y} = \beta_0 + \beta_1 \,\mathsf{TV} + \beta_2 \,\mathsf{Radio} + \beta_3 \,\mathsf{Newspaper}$$

systematically *overestimates* when one medium dominates and *underestimates* when the budget is split—indicating **synergy** between channels.

- To capture this, introduce interaction terms:

$$\hat{y} = \beta_0 + \beta_1 \,\mathsf{TV} + \beta_2 \,\mathsf{Radio} + \beta_{12} \,(\mathsf{TV} \times \mathsf{Radio}) + \beta_{13} \,(\mathsf{TV} \times \mathsf{Newspaper}) + \cdots$$

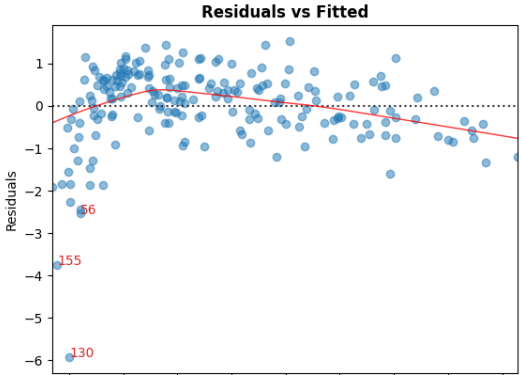
- Here $\beta_{12}$ measures the extra lift when TV and Radio spend are both high.
- You can also include higher-order terms (e.g.$\mathtt{TV}^2$) or other interactions to model more complex nonlinearity.
- Estimation proceeds as before (ordinary least squares).

## Example:

$$Y_i = \beta_0 + \beta_1 \,\mathsf{TV}_i + \beta_2 \,\mathsf{Newspaper}_i + \beta_3 \,\mathsf{Radio}_i + \beta_4 \,(\mathsf{TV}_i \times \mathsf{Radio}_i) + \beta_5 \,(\mathsf{TV}_i \times \mathsf{Newspaper}_i) + \beta_6 \,(\mathsf{Newspaper}_i \times \mathsf{Radio}_i) + \varepsilon_i$$

| | coef | std err | $t$ | $\mathbb{P}(\lvert T\rvert > \lvert t\rvert)$ | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 6.4602 | 0.318 | 20.342 | 0.000 | 5.834 | 7.087 |
| TV | 0.0203 | 0.002 | 12.633 | 0.000 | 0.017 | 0.024 |
| radio | 0.0229 | 0.011 | 2.009 | 0.046 | 0.000 | 0.045 |
| newspaper | 0.0170 | 0.010 | 1.691 | 0.092 | -0.003 | 0.037 |
| TR | 0.0011 | 5.72e-05 | 19.930 | 0.000 | 0.001 | 0.001 |
| TN | -7.971e-05 | 3.58e-05 | -2.227 | 0.027 | -0.000 | -9.12e-06 |
| RN | -0.0001 | 0.000 | -0.464 | 0.643 | -0.001 | 0.000 |

- The Coefficient of Determination $R^2 = \frac{SS_T - SS_E}{SS_T} = 0.969$.
  - Compare with the model without cross-term, where $R^2 = 0.897 < 0.969$.
  - The model fits the data better.
- The $F$-statistic $= 993.3$.
- We also see that the residual-versus-fitted plot is more linear than the one without cross-term. Though still not perfect.



**Example:** The effect of (newspsper) and (radio)×(newspaper) is not significant at 5% significance level. We should *consider removing them for interpretability and model simplicity*.

$$Y_i = \beta_0 + \beta_1 \,\mathsf{TV}_i + \beta_2 \,\mathsf{Radio}_i + \beta_3 \,(\mathsf{TV}_i \times \mathsf{Radio}_i) + \beta_4 \,(\mathsf{TV}_i \times \mathsf{Newspaper}_i) + \varepsilon_i$$

| | coef | std err | $t$ | $\mathbb{P}(\lvert T\rvert > \lvert t\rvert)$ | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 6.7491 | 0.248 | 27.195 | 0.000 | 6.260 | 7.239 |
| TV | 0.0194 | 0.002 | 12.494 | 0.000 | 0.016 | 0.022 |
| radio | 0.0288 | 0.009 | 3.225 | 0.001 | 0.011 | 0.046 |
| TR | 0.0011 | 5.34e-05 | 20.481 | 0.000 | 0.001 | 0.001 |
| TN | -1.335e-05 | 1.82e-05 | -0.733 | 0.465 | -4.93e-05 | 2.26e-05 |

- $R^2 = 0.968$.
- $F = 1469$.

**Example:** The effect of (TV)×(newspaper) is not significant at 5% significance level. We should *consider removing them for interpretability and model simplicity*.

$$Y_i = \beta_0 + \beta_1 \,\mathsf{TV}_i + \beta_2 \,\mathsf{Radio}_i + \beta_3 \,(\mathsf{TV}_i \times \mathsf{Radio}_i) + \varepsilon_i$$

| | coef | std err | $t$ | $\mathbb{P}(\lvert T\rvert > \lvert t\rvert)$ | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 6.7502 | 0.248 | 27.233 | 0.000 | 6.261 | 7.239 |
| TV | 0.0191 | 0.002 | 12.699 | 0.000 | 0.016 | 0.022 |
| radio | 0.0289 | 0.009 | 3.241 | 0.001 | 0.011 | 0.046 |
| TR | 0.0011 | 0.000 | 20.727 | 0.000 | 0.001 | 0.001 |

- $R^2 = 0.968$.
- $F = 1963$.
- Now all variables are significant at 5% significance level.

# Variable Selection

- Select the "best" subset of predictors to explain the response with maximal predictive power.
- Exclude unnecessary predictors that add noise and reduce estimation accuracy.
- Prevent collinearity by avoiding redundant variables that capture the same signal.
- Simplify the model to save on data-collection costs and improve interpretability.
- Aim for the simplest model that retains strong predictive power.

# Backward Elimination

What we just illustrated using the advertising data is a **backward elimination** approach to variable selection.

1. Start with the full model containing all candidate predictors.
2. Identify the predictor with the largest $p$-value.
   - If that $p$-value $> \alpha_{\text{remove}}$, remove this predictor.
3. Refit the reduced model.
4. Repeat steps 2-3 until every remaining predictor has $p$-value $\leq \alpha_{\text{remove}}$.

# Comparing models

Coefficient of determination

Adjusted Coefficient of determination: The closer to 1 the better.

F-Statistic: The larger the better. Tests if all $\beta_i = 0$.

Information Criterions: The smaller the better

≈ Model accuracy+penalty for model complexity.

**In selecting the models, we choose the model with the smallest AIC/BIC**

OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | Sales | **R-squared:** | 0.897 |
| **Model:** | OLS | **Adj. R-squared:** | 0.896 |
| **Method:** | Least Squares | **F-statistic:** | 570.3 |
| **Date:** | Fri, 04 Nov 2022 | **Prob (F-statistic):** | 1.58e-96 |
| **Time:** | 11:14:41 | **Log-Likelihood:** | -386.18 |
| **No. Observations:** | 200 | **AIC:** | 780.4 |
| **Df Residuals:** | 196 | **BIC:** | 793.6 |
| **Df Model:** | 3 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Intercept** | 2.9389 | 0.312 | 9.422 | 0.000 | 2.324 | 3.554 |
| **TV** | 0.0458 | 0.001 | 32.809 | 0.000 | 0.043 | 0.049 |
| **Newspaper** | -0.0010 | 0.006 | -0.177 | 0.860 | -0.013 | 0.011 |
| **Radio** | 0.1885 | 0.009 | 21.893 | 0.000 | 0.172 | 0.206 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 60.414 | **Durbin-Watson:** | 2.084 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 151.241 |
| **Skew:** | -1.327 | **Prob(JB):** | 1.44e-33 |
| **Kurtosis:** | 6.332 | **Cond. No.** | 454. |

```
# Fitting linear model
res = smf.ols(formula= "Sales ~ TV + Newspaper + Radio", data=df).fit()
res.summary()
```

OLS Regression Results

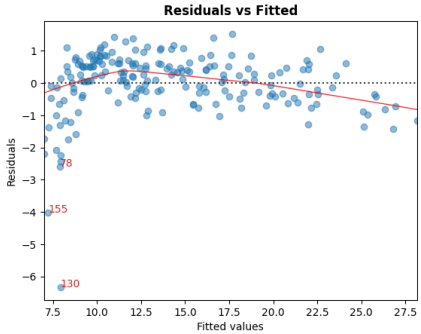
| | | | |
|---|---|---|---|
| **Dep. Variable:** | Sales | **R-squared:** | 0.968 |
| **Model:** | OLS | **Adj. R-squared:** | 0.967 |
| **Method:** | Least Squares | **F-statistic:** | 1963. |
| **Date:** | Fri, 04 Nov 2022 | **Prob (F-statistic):** | 6.68e-146 |
| **Time:** | 12:48:24 | **Log-Likelihood:** | -270.14 |
| **No. Observations:** | 200 | **AIC:** | 548.3 |
| **Df Residuals:** | 196 | **BIC:** | 561.5 |
| **Df Model:** | 3 | | |
| **Covariance Type:** | nonrobust | | |

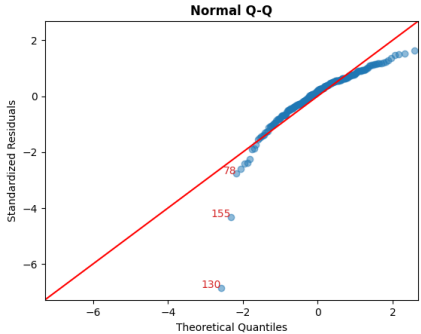
| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Intercept** | 6.7502 | 0.248 | 27.233 | 0.000 | 6.261 | 7.239 |
| **TV** | 0.0191 | 0.002 | 12.699 | 0.000 | 0.016 | 0.022 |
| **Radio** | 0.0289 | 0.009 | 3.241 | 0.001 | 0.011 | 0.046 |
| **TR** | 0.0011 | 0.000 | 7.545 | 0.000 | 0.001 | 0.001 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 128.132 | **Durbin-Watson:** | 2.224 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 1183.719 |
| **Skew:** | -2.323 | **Prob(JB):** | 9.09e-258 |
| **Kurtosis:** | 13.975 | **Cond. No.** | 1.80e+04 |

```
# Fitting linear model
res = smf.ols(formula= "Sales ~ TV + Radio + TR", data=df).fit()
res.summary()
```

Let's check the diagnostic plots again for the refined model:



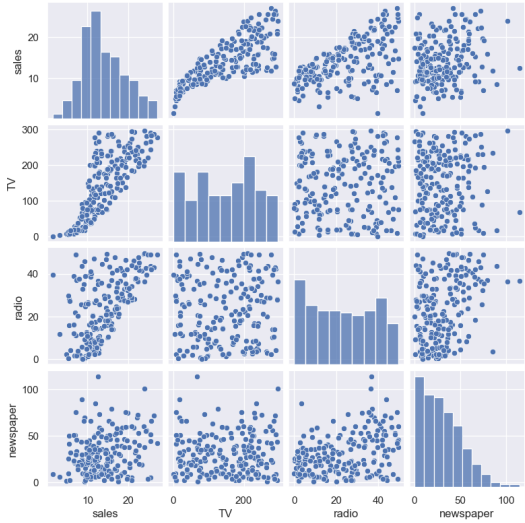

**Non-linearity exists!**

# Visualize the non-linearity

- The raw scatter of *Sales* vs. *TV* ad spend shows a nonlinear curve.
- If we plot *Sales* against $\log(\text{TV})$, the relationship appears approximately linear:

$$\text{Sales} \approx \beta_0 + \beta_1 \log(\text{TV}).$$

- This suggests using $\log(\text{TV})$ as the predictor in our regression model.



# Linearity Check

## Interpretation of the log transform

$$\log\left(x^2\right) = 2\,\log(x).$$

Thus, multiplying the TV budget by a factor of 2 increases $\log(\text{TV})$ by

$$\log(2x) - \log(x) = \log(2),$$

which corresponds to a fixed additive increment in predicted sales. In the original scale, this means *diminishing marginal returns*: each additional dollar spent on TV yields a smaller increment in sales the larger the budget already is.

# Regression with log TV, Radio, and TR

$$Y_i = \beta_0 + \beta_1 \log(\text{TV}_i) + \beta_2 \,\text{Radio}_i + \beta_3 \,(\text{TV}_i \times \text{Radio}_i) + \varepsilon_i$$

| | coef | std err | $t$ | $\mathbb{P}(\lvert T\rvert > \lvert t\rvert)$ | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 0.1886 | 0.168 | 1.125 | 0.262 | -0.142 | 0.519 |
| log_TV | 1.9670 | 0.034 | 57.041 | 0.000 | 1.899 | 2.035 |
| radio | 0.0458 | 0.003 | 17.410 | 0.000 | 0.041 | 0.051 |
| TR | 0.0010 | 1.41e-05 | 72.756 | 0.000 | 0.001 | 0.001 |

- $R^2 = 0.997, \quad F = 1952.$
- The intercept is not significant at the 5% level.

# Regression with log TV, Radio, and TR and no intercept

Removing the non-significant intercept, we get the following model:

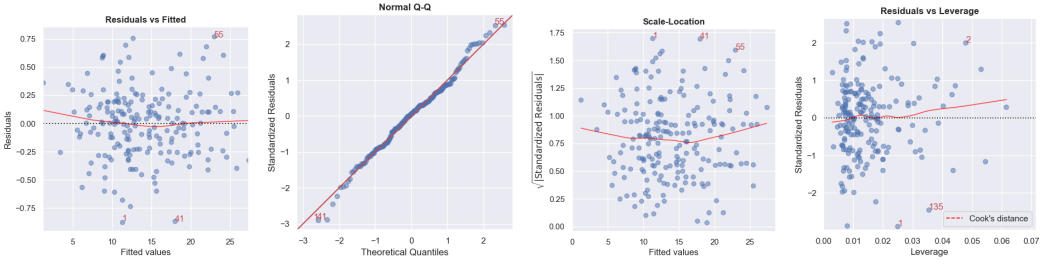
$$Y_i = \beta_1 \log(\text{TV}_i) + \beta_2 \,\text{Radio}_i + \beta_3 \,(\text{TV}_i \times \text{Radio}_i) + \varepsilon_i$$

| | coef | std err | $t$ | $\mathbb{P}(\lvert T\rvert > \lvert t\rvert)$ | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| log_TV | 2.0047 | 0.008 | 243.507 | 0.000 | 1.988 | 2.021 |
| radio | 0.0480 | 0.002 | 27.691 | 0.000 | 0.045 | 0.051 |
| TR | 0.0010 | 9.09e-06 | 111.447 | 0.000 | 0.001 | 0.001 |

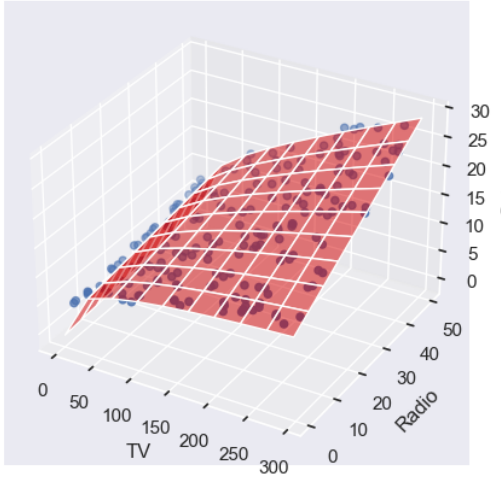
- $R^2 = 1.000, \quad F = 16150.$
- All predictors are significant at the 5% level.
- The model is now interpretable and parsimonious.

# The diagnostic plots of the final model

All four plots look almost perfect.



# Visualize the final model



# The More the Better?

If we have a lot of variables, it is tempting to include them all and conduct regression analysis for a complex model.

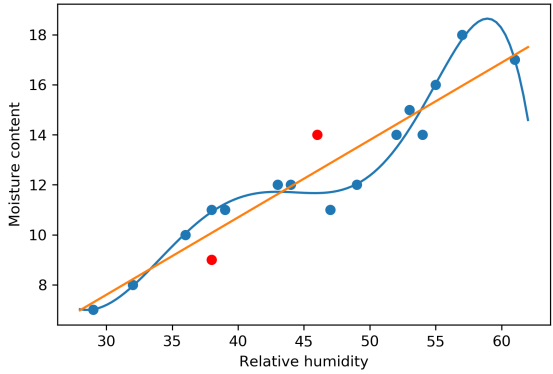
$$Y_i = \beta_0 + \beta_1 \log(\text{TV}_i) + \beta_2 \text{TV}_i + \beta_3 \text{newspaper}_i + \beta_4 \text{Radio}_i + \beta_5 (\text{TV} \times \text{Radio}) + \beta_6 (\text{TV} \times \text{newspaper}) + \beta_7 (\text{Radio} \times \text{newspaper}) + \varepsilon_i$$

| | coef | std err | $t$ | $\mathbb{P}(\|T\| > \|t\|)$ | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 0.2427 | 0.186 | 1.307 | 0.193 | -0.123 | 0.609 |
| log(TV) | 1.9377 | 0.048 | 40.344 | 0.000 | 1.843 | 2.032 |
| TV | 0.0007 | 0.001 | 1.001 | 0.318 | -0.001 | 0.002 |
| newspaper | 0.0009 | 0.003 | 0.258 | 0.797 | -0.006 | 0.007 |
| radio | 0.0448 | 0.004 | 11.916 | 0.000 | 0.037 | 0.052 |
| TR | 0.0010 | 1.88e-05 | 54.400 | 0.000 | 0.001 | 0.001 |
| TN | -9.779e-06 | 1.18e-05 | -0.830 | 0.408 | -3.30e-05 | 1.35e-05 |
| RN | 2.825e-05 | 7.70e-05 | 0.367 | 0.714 | -0.000 | 0.000 |

- $R^2 = 0.997 < 1.000$. $F = 8260 < 16150$.
- $AIC = 101.2 > 94.05, BIC = 127.6 > 103.9$. These two metrics are the smaller the better.
- Not all variables are significant at 5% significance level.

# The More the Better?

If we have a lot of variables, it is tempting to include them all and conduct regression analysis for a complex model.



**Complex model:** $R^2 = 0.98$

| | coef | std err | $t$ | $p$-value | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | -561.7124 | 2941.883 | -0.191 | 0.853 | -7345.707 | 6222.282 |
| $X$ | 117.4390 | 418.101 | 0.281 | 0.786 | -846.704 | 1081.582 |
| $X^2$ | -9.2739 | 24.447 | -0.379 | 0.714 | -65.649 | 47.101 |
| $X^3$ | 0.3662 | 0.753 | 0.486 | 0.640 | -1.370 | 2.103 |
| $X^4$ | -0.0077 | 0.013 | -0.599 | 0.566 | -0.037 | 0.022 |
| $X^5$ | 8.315e-05 | 0.000 | 0.715 | 0.495 | -0.000 | 0.000 |
| $X^6$ | -3.596e-07 | 4.32e-07 | -0.832 | 0.430 | -1.36e-06 | 6.37e-07 |

**Simple model:** $R^2 = 0.90$

| | coef | std err | $t$ | $p$-value | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | -1.6877 | 1.322 | -1.277 | 0.224 | -4.544 | 1.168 |
| X | 0.3096 | 0.028 | 10.975 | 0.000 | 0.249 | 0.371 |

# Overfitting

**High $R^2$ is not always the only objective**

- Too many "useless" variables $\Rightarrow$ overfitting $\Rightarrow$ less accurate prediction.
- Unnecessary predictors can add noise to the estimation of other important quantities, and make them unstable (having extremely large parameters with extremely large std. err.)
- Waste time and/or money to measure redundant predictors.

**General rule of thumb:**

**Find the simplest model with satisfactory prediction accuracy.**

**Never use more than you need.**

# Estimating the Prediction Accuracy

What we really care is how we can predict the dependent variable $Y_i$ for a future observation, i.e. the prediction error on a sample completely independent of the data we used to train/fit the model.

Is the $MS_E = SS_E/(n-p-1)$ a good estimation?

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

- No! In estimation of the parameters, we minimized $SS_E$, which makes it a optimistically biased assessment of the prediction error.
- This biased estimate is called the in-sample estimate of the fit.

# Estimating the Prediction Accuracy – Cross-Validation

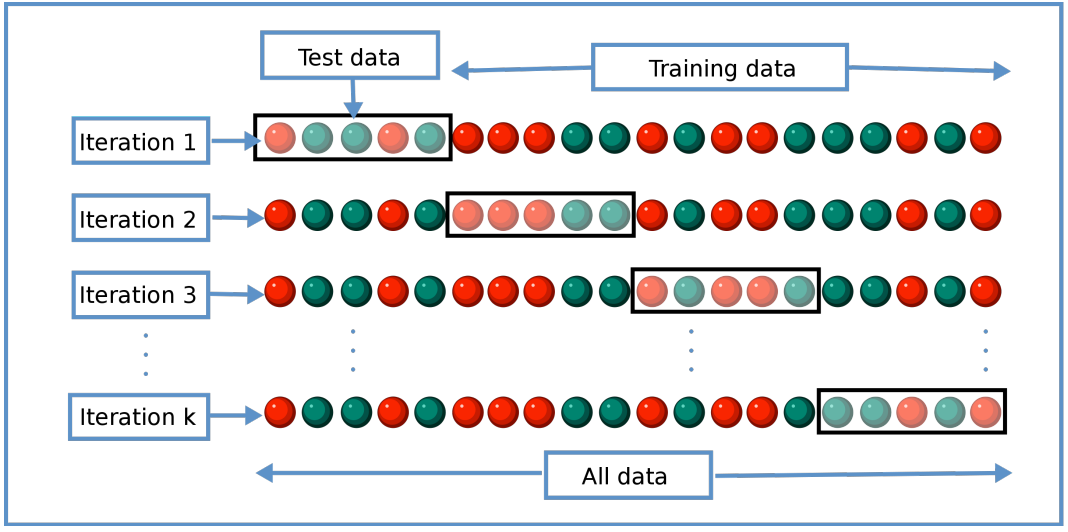
We can use **cross-validation** to obtain a out-of-sample estimate of the prediction error.

**Key idea**

- Put aside a subset of the observations (called "testing set").
- Fit the model using the rest of the observations (called "training set").
  - So the testing set is independent of the training set.
- Calculate the mean squared errors on the testing set.

In practice, people usually repeat the above for different partitioning of training and testing sets. Then use the average mean squared errors as a out-of-sample estimate of the prediction error.

# $k$-Fold Cross-Validation



# Model Selection

**Backward elimination**

Step 1 Start with all the predictors in the model. (full model)

Step 2 Remove a predictor with highest p-value greater than $\alpha_{remove}$.

Step 3 Refit the model and go to Step 2.

Step 4 Stop when all p-values are less than $\alpha_{remove}$.

**Forward selection**

Step 1 Start with no variables in the model.

Step 2 For all predictors not in the model, check their p-values if they are added to the model. Choose the one with lowest p-value less than $\alpha_{enter}$.

Step 3 Repeat Step 2 until no new predictors can be added.

**Stepwise selection**

- Combination of backward elimination and forward selection.
- At each stage, a variable is added or removed.
- Stepwise procedures are relatively cheap computationally.
- In practice, we can choose $\alpha_{remove}$ and $\alpha_{enter}$ at around $15 - 20\%$.

**Methods for model selection**

- Cross-validation
- Others: AIC, BIC, Mallow's $C_p$...
- Shrinkage methods: LASSO...