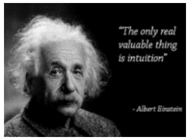
Variance, Covariance, Correlation and Analysis of Variance

Kogod School of Business
American University • Washington, DC



Explaining Variance: Intuition

- Understanding **central tendencies** (e.g., means, median) and **dispersion** around these tendencies (e.g., variance, standard deviation) is central to statistics, machine learning and predictive modeling.
- Variation of single attributes → e.g., **variance**
- Comparing variance of 2 attributes → e.g., **ANOVA**
- Analyzing how 2 or more attributes vary together → **covariance, correlation**

- **Correlation** → measure of statistical association
- **Prediction** → estimating outcomes based on observed associations – e.g., regression, decision trees, etc.
- **Causation** → harder to prove; prediction does not imply causation; need more rigorous methods to prove causation

Variance and Covariance Refresher

- A lot of what we do in analytics is understanding and explaining variance and covariance. **Knowing these concepts is key**.
- **Variance** is how much values vary relative to the mean. The value is squared so that values, both below and above the mean contribute positively to the variance statistic.

$$Var = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

- **Covariance** indicates how two variables are related – i.e., how do they "co-vary" or how they move together. If when x is above (or below) the mean y is generally above (or below) the mean, then the covariance is positive. Otherwise is negative.

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

- Covariance is a useful concept, but it has limited practical use because the covariance value is dependent on the **scale** used to measure x and y. **Correlation** takes care of this scale problem.

Correlation Refresher

- Correlation is like covariance, but the deviation from the mean of each variable are **divided by the standard deviation** of the variable → it can be shown mathematically that the correlation of two variables will not change with re-scaling.
- Mathematically, correlation statistics ranges from **-1.0** to **1.0**
- This is really a **descriptive analytics** method, but it is a necessary first step before predictive analytics
- Provides an indication for whether two variables vary in the same or opposite direction, or if they are **independent** from each other

$$\rho(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right) = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

- Analyzing the **descriptive statistics** (i.e., mean and standard deviation) and the **correlation** among all variables is a necessary **first step** (i.e., descriptive analytics) before building predictive models. This is often called **"eye-balling the data"**

Correlation Analysis – 2 Key Values

1. **Magnitude** – how large is the association

   ρ = 1.0 Perfectly positively correlated
   ρ = + Positively correlated
   ρ = Around 0 Uncorrelated (i.e., independent)
   ρ = - Negatively correlated
   ρ = - 1.0 Perfectly negatively correlated

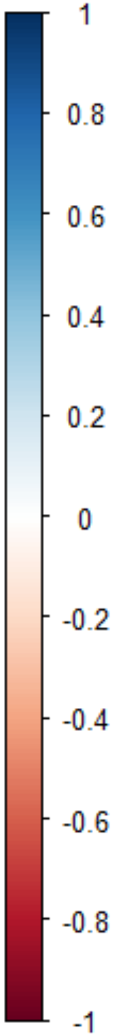
2. **Significance** – probability that the observed correlation happened by chance – i.e., **p → prob(ρ=0)**

   $p >= 0.10$ → Not significantly ≠ 0 – i.e., independent
   $p < 0.10$ → Moderately significant
   $p < 0.05$ → Significant
   $p < 0.01$ → Very significant
   $p < 0.001$ → Highly significant

It is useful to look at the correlation matrix

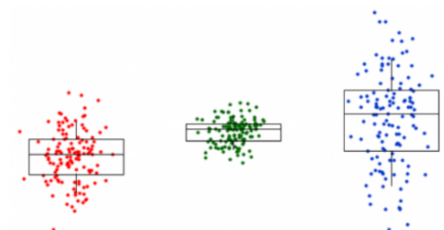
Correlation Matrix (Automobile Data)

	carb	wt	hp	cyl	disp	qsec	vs	mpg	drat	am	gear
carb		0.43	0.75	0.53	0.39	-0.66	-0.57	-0.55	-0.09	0.06	0.27
wt	0.43		0.66	0.78	0.89	-0.17	-0.55	-0.87	-0.71	-0.69	-0.58
hp	0.75	0.66		0.83	0.79	-0.71	-0.72	-0.78	-0.45	-0.24	-0.13
cyl	0.53	0.78	0.83		0.9	-0.59	-0.81	-0.85	-0.7	-0.52	-0.49
disp	0.39	0.89	0.79	0.9		-0.43	-0.71	-0.85	-0.71	-0.59	-0.56
qsec	-0.66	-0.17	-0.71	-0.59	-0.43		0.74	0.42	0.09	-0.23	-0.21
vs	-0.57	-0.55	-0.72	-0.81	-0.71	0.74		0.66	0.44	0.17	0.21
mpg	-0.55	-0.87	-0.78	-0.85	-0.85	0.42	0.66		0.68	0.6	0.48
drat	-0.09	-0.71	-0.45	-0.7	-0.71	0.09	0.44	0.68		0.71	0.7
am	0.06	-0.69	-0.24	-0.52	-0.59	-0.23	0.17	0.6	0.71		0.79
gear	0.27	-0.58	-0.13	-0.49	-0.56	-0.21	0.21	0.48	0.7	0.79	



ANOVA Refresher

- **Analysis of variance** (ANOVA) provides a statistical test of whether the mean of a given variable is equal among two or more groups.
- It is called analysis of **"variance"** and not analysis of **"means"** because it compares the variance within each group against the variance of the means between groups.
- For example, if we want to test if the mileage is different between foreign and domestic cars, or if the price of a diamond is different for various color classifications, we can do an ANOVA test.
- The **intuition** is that if the variance **between group** means is significantly larger than the variance within groups, then the means are significantly different. Otherwise they are not.



- **ANOVA** and **regression** are tightly related because ANOVA tests whether the **variance explained** by the regression line (or the various variables in the model) is significantly different than the variance of the dependent variable alone.
- As we will see later, ANOVA is very useful when **comparing** whether **one regression model** explains more variance than **another**, so it is a key test when evaluating predictive models.

R Tips

`ggplot2{ggplot}` → ggplot2 library in the ggplot package is very popular for statistical plots and graphs

`cor(), var(), and cov()` → in the {stats} library provide the correlation, variance and conariance for a matrix or data frame

`rcorr{Hmisc}` → provides more complete correlation data, like p-values

`ggpairs{GGally}, pairs{GGally}` → correlation matrix with visual scatterplots

`corrplot{corrplot}` → correlation matrix with visual scatterplots

`aov{stats}` → Traditional ANOVA to test differences in means; yields the same results as `lm(),` except with "repeated measures" (e.g., one person provides multiple observations – for example: recovery time with and without medicine) – `aov()` is preferred in such cases

`anova{stats}` → ANOVA → Used primarily to compare 2 linear models

`boxplot{graphics}` → Graphical contrast of means

KOGOD SCHOOL
of
BUSINESS