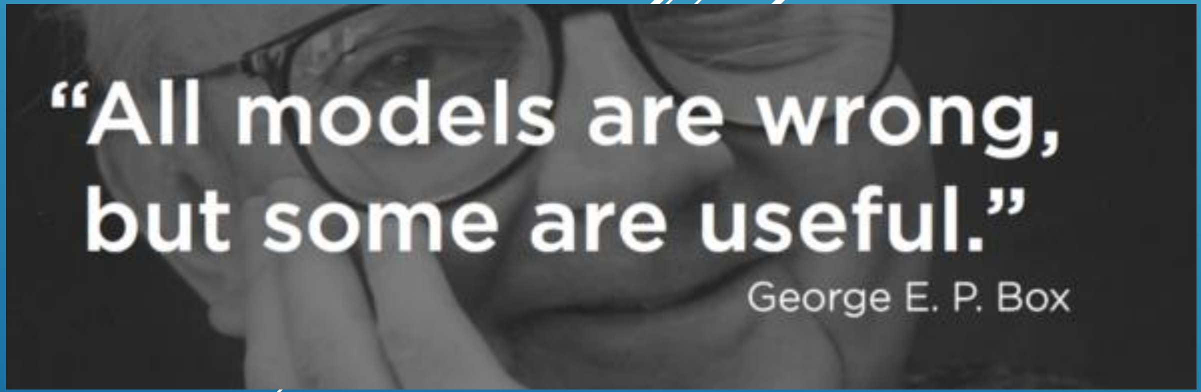
# MULTIVARIATE MODELS



# MULTIVARIATE SCENARIO

# NOTHING IN LIFE IS EVER SO SIMPLE

For 'variate', I would say this is a common way to refer to any random variable that follows a known or hypothesized distribution, e.g. we speak of gaussian variates $X_i$ as a series of observations drawn from a normal distribution (with parameters $\mu$ and $\sigma^2$). In probabilistic terms, we said that these are some random *realizations* of X, with mathematical expectation $\mu$, and about 95% of them are expected to lie on the range $[\mu - 2\sigma; \mu + 2\sigma]$ .

# MULTIVARIATE VS MULTIPLE REGRESSION

So far, we have been working with
$y = f(x) + \varepsilon$ where $f(x)$ is of the form $\beta_1 x_1 + \beta_0$
but most real world problems there would be
more than one independent variable $x_1, x_2, x_3 \ldots$
That is more than one
attribute determine the dependent variable

"Multiple regression" refers to situations in which you have more than one predictor / explanatory variable ($\boldsymbol{X}$).

"Multivariate regression" refers to situations in which you have more than one response / outcome / dependent variable ($\boldsymbol{Y}$).

Such problems are Multiple Regression problem and when there are
more than one dependent variables, it is called Multivariate.

Please note we still are considering only one dependent variable.

What to do If there are more than one dependent variables

-- at this time, the option is to run lm on each dependent variable

In what follows we continue with one dependent variable and
many independent variables.

```
> path<-"c:/users/rkannan/rk/03062015/kirpal-story-of-data/gasconsumption.csv"
> gas<-read.csv(path,sep=",",head=T)
> gas
    tax income miles driver petrol
1  9.00   3571  1976  0.525    541
2  9.00   4092  1250  0.572    524
3  9.00   3865  1586  0.580    561
4  7.50   4870  2351  0.529    414
```

My Variables
-- income -- Income
-- miles -- Number of miles of roads available
-- driver -- Number of drives living in the region
-- tax -- Tax per gallon of petrol
-- petrol -- Petrol consumed

Given more miles, more eligible drivers with more income – miles driven should be higher
Tax per gallon should discourage consumption of petrol…

# CAN I PREDICT PETROL CONSUMPTION ?

```
> head(gaspx)
   tax income miles driver petrol
1  9.0   3571  1976  0.525    541
2  9.0   4092  1250  0.572    524
3  9.0   3865  1586  0.580    561
4  7.5   4870  2351  0.529    414
5  8.0   4399   431  0.544    410
6 10.0   5342  1333  0.571    457
>
```

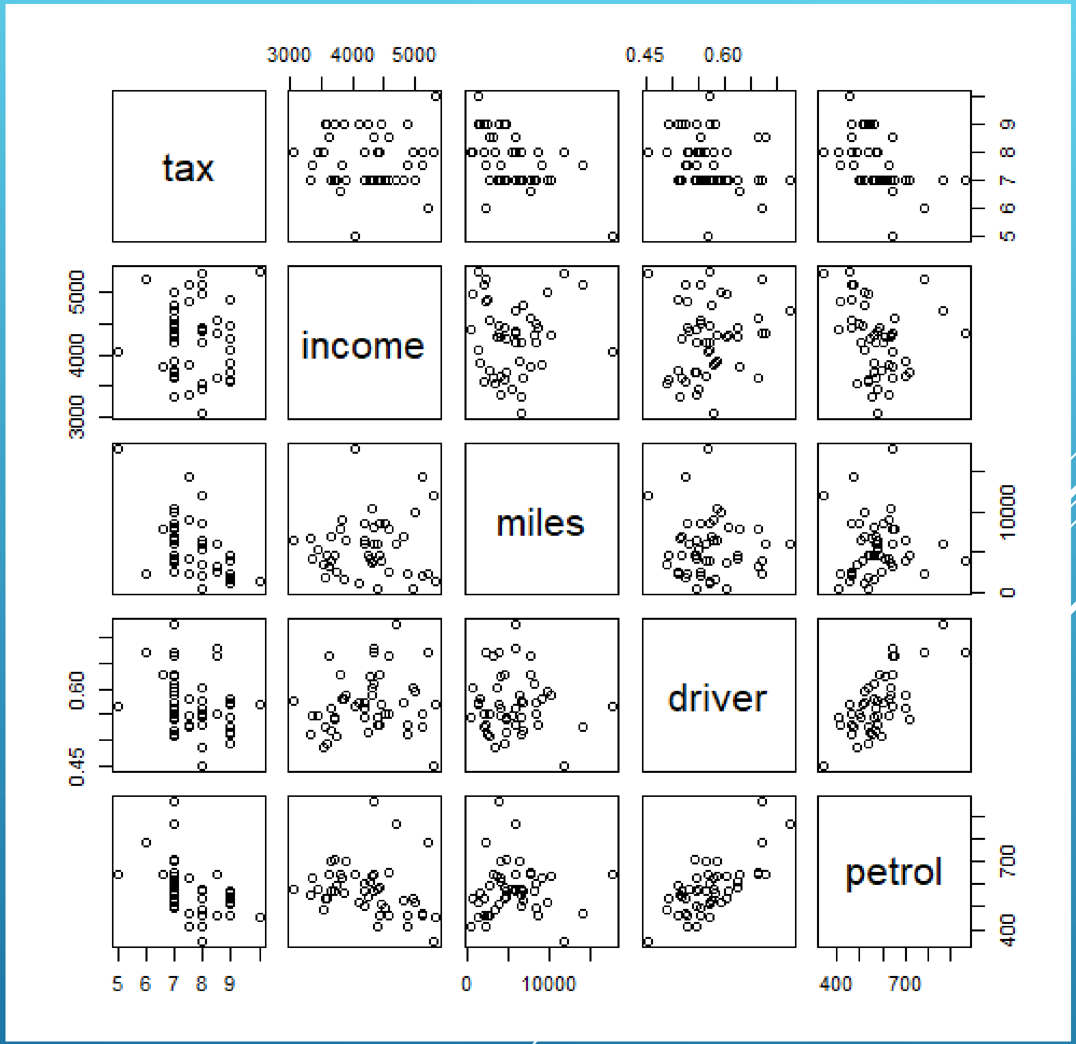
- So called wide format
- One observation per row
- Each row fully defines
- Each column is an attribute
- One dependent variable
- Other columns are independent columns

- In this dataset, petrol consumed is the independent variable
- The challenge is how much of that consumption is determined by tax, income,miles  and proportion of drivers
- Which  of these variables influence gas consumption
- Which is more dominant, sensitivity analysis

# Understanding the problem and the data

```
> summary(gas)
      tax              income          miles            driver            petrol     
 Min.   : 5.000   Min.   :3063   Min.   :  431   Min.   :0.4510   Min.   :344.0  
 1st Qu.: 7.000   1st Qu.:3739   1st Qu.: 3110   1st Qu.:0.5298   1st Qu.:509.5  
 Median : 7.500   Median :4298   Median : 4736   Median :0.5645   Median :568.5  
 Mean   : 7.668   Mean   :4242   Mean   : 5565   Mean   :0.5703   Mean   :576.8  
 3rd Qu.: 8.125   3rd Qu.:4579   3rd Qu.: 7156   3rd Qu.:0.5952   3rd Qu.:632.8  
 Max.   :10.000   Max.   :5342   Max.   :17782   Max.   :0.7240   Max.   :968.0  
```

```
> pairs(gas)
>
```




# EDA

lmds<-lm(petrol~.,data=gas)

petrol = 377.3 -34.79 tax-0.06659 income-0.002426 miles+1336 drivers

```
> lmds

Call:
lm(formula = petrol ~ ., data = gas)

Coefficients:
(Intercept)          tax       income        miles       driver
  3.773e+02   -3.479e+01   -6.659e-02   -2.426e-03    1.336e+03
```

Results are not consistent with our intuition…

Models have to be consistent with reality and to an extent our intuition…

Tax and drivers consistent

Miles and income not so…

```
> summary(lmds)

Call:
lm(formula = petrol ~ ., data = gas)

Residuals:
    Min      1Q  Median      3Q     Max
-122.03  -45.57  -10.66   31.53  234.95

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.773e+02  1.855e+02   2.033 0.048207 *
tax         -3.479e+01  1.297e+01  -2.682 0.010332 *
income      -6.659e-02  1.722e-02  -3.867 0.000368 ***
miles       -2.426e-03  3.389e-03  -0.716 0.477999
driver       1.336e+03  1.923e+02   6.950 1.52e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 66.31 on 43 degrees of freedom
Multiple R-squared:  0.6787,    Adjusted R-squared:  0.6488
F-statistic: 22.71 on 4 and 43 DF,  p-value: 3.907e-10
```

# MULTIPLE REGRESSION IN R

-- run pairwise regression and add them back
-- remove some counter-intuitive variables
-- remove collinear variables
-- scale the variables

# MULTIPLE REGRESSION IN R

```
> lmtax<-lm(gas~tax,data=gas)
Error in model.frame.default(formula = gas ~ tax, data = gas, drop.unused.levels = TRUE) :
  invalid type (list) for variable 'gas'
> lmtax<-lm(petrol~tax,data=gas)
> summary(lmtax)

Call:
lm(formula = petrol ~ tax, data = gas)

Residuals:
    Min      1Q  Median      3Q     Max
-215.16  -72.27    6.74   41.28  355.74

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   984.01     119.62   8.226 1.38e-10 ***
tax           -53.11      15.48  -3.430  0.00128 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 100.9 on 46 degrees of freedom
Multiple R-squared:  0.2037,    Adjusted R-squared:  0.1863
F-statistic: 11.76 on 1 and 46 DF,  p-value: 0.001285
```

If we raise tax, we expect people to buy less
Model is consistent with reality and our intuition.
Results are reliable –p-value less than 0.05 indicates we can reject the NULL – there is a relationship – the observed data is not due to chance and  the variable explains 20%
Of the variability in the dependent variable.

# PETROL CONSUMPTION VS GAS TAX PER GALLON

# REALITY > INTUITION > MODEL

```
> lmincome<-lm(petrol~income,data=gas)
> summary(lmincome)

Call:
lm(formula = petrol ~ income, data = gas)

Residuals:
    Min      1Q  Median      3Q     Max 
-181.32  -66.44  -20.38   43.24  396.16 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 779.36317  119.33016   6.531 4.61e-08 ***
income       -0.04776    0.02788  -1.713   0.0935 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 109.7 on 46 degrees of freedom
Multiple R-squared:  0.05996,	Adjusted R-squared:  0.03952 
F-statistic: 2.934 on 1 and 46 DF,  p-value: 0.09347
```

First, exercise caution when counter-intuitive

P-value says cannot reject NULL

NULL for lm is there is no relationship.

So the estimated coefficient for income
Is not reliable. May not have any influence.

And R-Square tells us this variable has
No explanatory potential…it is explaining
5%.

```
> lmdrivers<-lm(petrol~driver,data=gas)
> summary(lmdrivers)

Call:
lm(formula = petrol ~ driver, data = gas)

Residuals:
    Min      1Q  Median      3Q     Max 
-129.65  -60.53  -13.03   58.57  247.90 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -227.3      121.9  -1.865   0.0685 .  
driver        1409.8      212.7   6.629 3.29e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 80.88 on 46 degrees of freedom
Multiple R-squared:  0.4886,	Adjusted R-squared:  0.4774 
F-statistic: 43.94 on 1 and 46 DF,  p-value: 3.29e-08
```

# MORE DRIVERS →MORE CONSUMPTION

```
> lmmiles<-lm(petrol~miles,data=gas)
> summary(lmmiles)

Call:
lm(formula = petrol ~ miles, data = gas)

Residuals:
    Min      1Q  Median      3Q     Max 
-236.62  -66.68   -8.43   55.43  392.24 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 5.734e+02  3.094e+01  18.529   <2e-16 ***
miles       6.102e-04  4.724e-03   0.129    0.898    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 113.1 on 46 degrees of freedom
Multiple R-squared:  0.0003626, Adjusted R-squared:  -0.02137 
F-statistic: 0.01669 on 1 and 46 DF,  p-value: 0.8978
```

# MILES DO NOT HAVE BEARING ON CONSUMPTION

```
> head(gaspx)
   tax income miles driver petrol
1  9.0   3571  1976  0.525    541
2  9.0   4092  1250  0.572    524
3  9.0   3865  1586  0.580    561
4  7.5   4870  2351  0.529    414
5  8.0   4399   431  0.544    410
6 10.0   5342  1333  0.571    457
>
```

There are some nuances
Income varies at a different level
And at a different rate
Unit of change in one variable
Does not equal a unit of change in another one.
$Y=b_1x_1+b_2x_2+b_3x_3$ will not make sense

- So we have to scale them so that they are all in the same unit
- Let us take a look at the mean and standard deviation by each column

# Scaling

```
> sapply(gaspx,mean)
          tax       income        miles       driver       petrol
    7.6683333 4241.8333333 5565.4166667    0.5703333  576.7708333
> sapply(gaspx,sd)
          tax       income        miles       driver       petrol
9.507698e-01 5.736238e+02 3.491507e+03 5.547027e-02 1.118858e+02
> table(sapply(gaspx,mean)==sapply(gaspx,mean))

TRUE
   5
```

Senstivity analysis will yield wild results if we don't correct for this …in a multivariate setting

Particularly if the covariates are correlated – that is, a change in one variable results in the change of other co-variates

# Issues Particular to MultiVariate datasets

# Normalization – Scaling Multivariate datasets

What is the correlation like?

```
> cor(gaspx)
               tax      income       miles     driver      petrol
tax     1.00000000  0.01266516 -0.52213014 -0.2880372 -0.45128028
income  0.01266516  1.00000000  0.05016279  0.1570701 -0.24486207
miles  -0.52213014  0.05016279  1.00000000 -0.0641295  0.01904194
driver -0.28803717  0.15707008 -0.06412950  1.0000000  0.69896542
petrol -0.45128028 -0.24486207  0.01904194  0.6989654  1.00000000
> 
```

Do these correlations make sense?

Let us create a scaled dataframe (name, (x-mean)/std var)

```
 names(gaspx)
 scaled.gaspx<-scale(gaspx) # scale the data,it has additional attributes
scaled.gaspx.d<-scaled.gaspx[1:48,] # so we extract rows of scaled observations
names(scaled.gaspx.d)<-names(gaspx) # we copy the names of the columns
lm.scaled.gaspx<-lm(petrol~.,data=as.data.frame(scaled.gaspx[1:48,]))
```

```
> lm.gaspx<-lm(petrol~.,data=gaspx)
> summary(lm.gaspx)

Call:
lm(formula = petrol ~ ., data = gaspx)

Residuals:
    Min      1Q  Median      3Q     Max 
-122.03  -45.57  -10.66   31.53  234.95 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.773e+02  1.855e+02   2.033 0.048207 *  
tax         -3.479e+01  1.297e+01  -2.682 0.010332 *  
income      -6.659e-02  1.722e-02  -3.867 0.000368 ***
miles       -2.426e-03  3.389e-03  -0.716 0.477999    
driver       1.336e+03  1.923e+02   6.950 1.52e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 66.31 on 43 degrees of freedom
Multiple R-squared:  0.6787,	Adjusted R-squared:  0.6488 
F-statistic: 22.71 on 4 and 43 DF,  p-value: 3.907e-10
```

```
> lm.scaled.gaspx<-lm(petrol~.,data=as.data.frame(scaled.gaspx[1:48,]))
> summary(lm.scaled.gaspx)

Call:
lm(formula = petrol ~ ., data = as.data.frame(scaled.gaspx[1:48, 
    ]))

Residuals:
     Min       1Q   Median       3Q      Max 
-1.09066 -0.40732 -0.09531  0.28180  2.09988 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.879e-17  8.554e-02   0.000 1.000000    
tax         -2.956e-01  1.102e-01  -2.682 0.010332 *  
income      -3.414e-01  8.829e-02  -3.867 0.000368 ***
miles       -7.570e-02  1.058e-01  -0.716 0.477999    
driver       6.626e-01  9.534e-02   6.950 1.52e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5926 on 43 degrees of freedom
Multiple R-squared:  0.6787,	Adjusted R-squared:  0.6488 
F-statistic: 22.71 on 4 and 43 DF,  p-value: 3.907e-10
```

# LM: IS THIS A PARADOX?

```
> glm.gaspx<-glm(petrol~.,data=gaspx)
> summary(glm.gaspx)

Call:
glm(formula = petrol ~ ., data = gaspx)

Deviance Residuals:
    Min        1Q    Median       3Q      Max
-122.03    -45.57    -10.66    31.53   234.95

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.773e+02  1.855e+02   2.033 0.048207 *
tax         -3.479e+01  1.297e+01  -2.682 0.010332 *
income      -6.659e-02  1.722e-02  -3.867 0.000368 ***
miles       -2.426e-03  3.389e-03  -0.716 0.477999
driver       1.336e+03  1.923e+02   6.950 1.52e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 4396.511)

    Null deviance: 588366  on 47  degrees of freedom
Residual deviance: 189050  on 43  degrees of freedom
AIC: 545.59

Number of Fisher Scoring iterations: 2
```

```
> glm.scaled.gaspx<-glm(petrol~.,data=as.data.frame(scaled.gaspx[1:48,]))
> summary(glm.scaled.gaspx)

Call:
glm(formula = petrol ~ ., data = as.data.frame(scaled.gaspx[1:48,
    ]))

Deviance Residuals:
     Min          1Q      Median          3Q         Max
-1.09066    -0.40732    -0.09531     0.28180     2.09988

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.879e-17  8.554e-02   0.000 1.000000
tax         -2.956e-01  1.102e-01  -2.682 0.010332 *
income      -3.414e-01  8.829e-02  -3.867 0.000368 ***
miles       -7.570e-02  1.058e-01  -0.716 0.477999
driver       6.626e-01  9.534e-02   6.950 1.52e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.3512029)

    Null deviance: 47.000  on 47  degrees of freedom
Residual deviance: 15.102  on 43  degrees of freedom
AIC: 92.711

Number of Fisher Scoring iterations: 2
```

# GLM AIC AND DEVIANCE?

Both GLM the Deviance is decreasing…this is a measure of deviation and So lower deviation implies model yields closer prediction

However AIC for the scaled model is lower – indicating glm yields a better fit when presented with scaled data –in the case of MV data…

The F-Statistic is another statistic and it is the ratio of two variances (SSR/SSE),

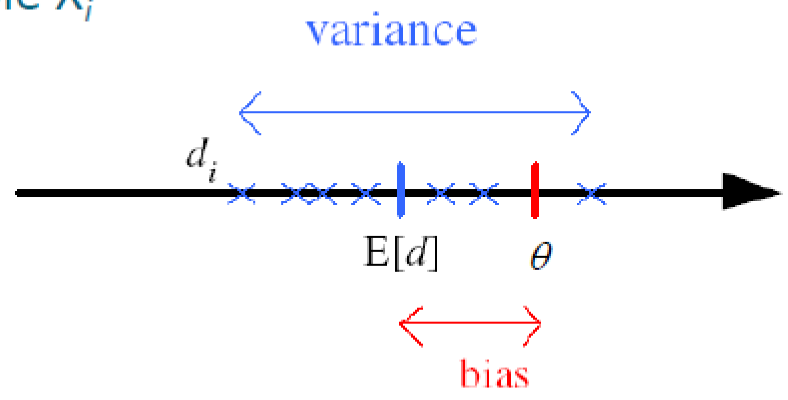
the variance explained by the parameters in the model (sum of squares of regression, SSR) and the residual or unexplained variance (sum of squares of error, SSE).

# GOODNESS OF FIT AIC/DEVIANCE

# BIAS-VARIANCE TRADE-OFF

Unknown parameter $\theta$

Estimator $d_i = d\,(X_i)$ on sample $X_i$




Bias: $b_\theta(d) = E\,[d] - \theta$

Variance: $E\,[(d-E\,[d])^2]$

Mean square error:

$r\,(d,\theta) = E\,[(d-\theta)^2]$

$= (E\,[d] - \theta)^2 + E\,[(d-E\,[d])^2]$

$=$ Bias$^2$ + Variance

From 4th chapter Ethem Alpaydin

# THE TWO SOURCES OF ERROR

**Background Information**

Consider the multiple linear regression model: $\mathbf{y} = \mathbf{X\beta} + \mathbf{e}$, where $\mathbf{y}$ is a $(n \times 1)$ vector of observations on the dependent variable, $\mathbf{X}$ is a $(n \times p)$ fixed matrix of observations on the explanatory variables, $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of unknown regression coefficients, and $\mathbf{e}$ is a $(n \times 1)$ vector of errors assumed to be normally distributed with $E(\mathbf{e}) = \mathbf{0}$ and $E(\mathbf{ee}') = \sigma^2\mathbf{I}_n$. The usual estimator for $\boldsymbol{\beta}$ is the least squares estimator given by $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

When the vector of predictor variables is multicollinear, the least squares estimates are likely to be large in absolute value and even with a wrong sign. The problem is a result of the fact that $(\mathbf{X}'\mathbf{X})$ is near singular. The Gauss–Markov property gives the assurance that the least squares estimator has minimum variance in the class of unbiased linear estimators, but there is no guarantee that this variance will be small.

One way to alleviate this problem is to drop the requirement that the estimator of $\boldsymbol{\beta}$ be unbiased. Suppose there is a biased estimator of $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}^*$, that has a smaller variance than the unbiased estimator $\hat{\boldsymbol{\beta}}$. Consider the mean squared error of the estimator $\hat{\boldsymbol{\beta}}^*$:

$$\mathrm{MSE}(\hat{\boldsymbol{\beta}}^*) = E(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta})^2 = \mathrm{Var}(\hat{\boldsymbol{\beta}}^*) + [E(\hat{\boldsymbol{\beta}}^*) - \boldsymbol{\beta}]^2$$

or

$$\mathrm{MSE}(\hat{\boldsymbol{\beta}}^*) = \mathrm{Var}(\hat{\boldsymbol{\beta}}^*) + (\text{bias in } \hat{\boldsymbol{\beta}}^*)^2$$

It should be noted that the **MSE** is just the distance from $\hat{\boldsymbol{\beta}}^*$ to $\boldsymbol{\beta}$. By allowing a small bias in $\hat{\boldsymbol{\beta}}^*$, the variance of $\hat{\boldsymbol{\beta}}^*$ can be made smaller. Consequently confidence intervals on $\boldsymbol{\beta}$ would be narrower using the biased estimator. The small variance for the biased estimator also implies that $\hat{\boldsymbol{\beta}}^*$ is a more stable estimator of $\boldsymbol{\beta}$ than is the unbiased estimator $\hat{\boldsymbol{\beta}}$. Hence a model using $\hat{\boldsymbol{\beta}}^*$ may have better predictive power.

Assumption of OLS

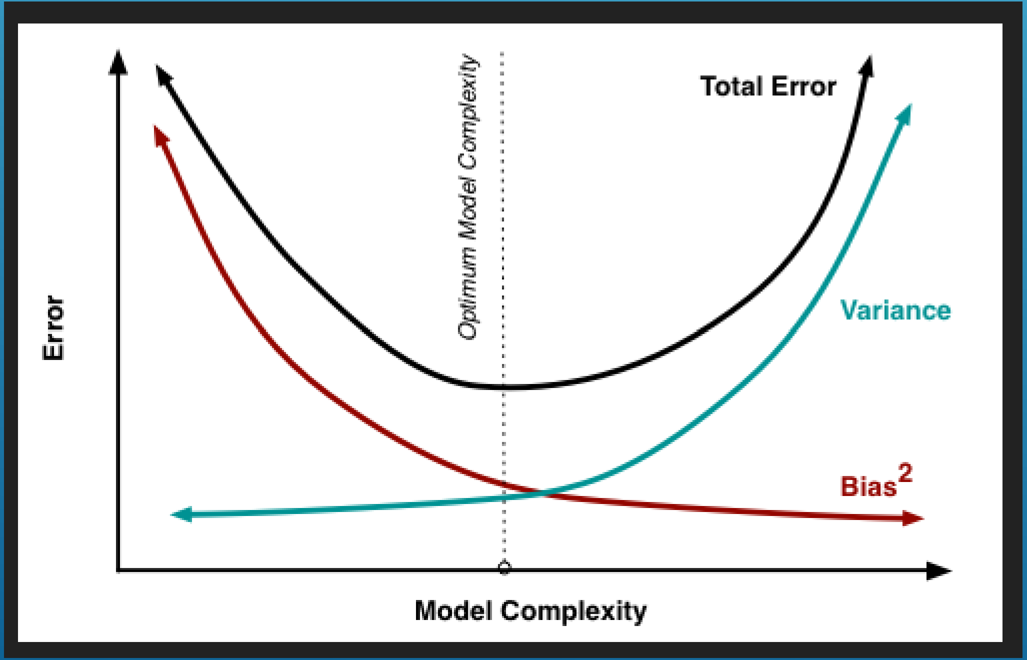
Actual-estimated = ERR where ERR is N(0,$\sigma^2$)
Normal distribution with mean is zero, variance of $\sigma^2$
E(e) = 0 is expectation (err) the mean
E(ee') is expectation ($\sigma^2$ the variance)

$$E\left[(r-g(x))^2 \mid x\right] = \underbrace{E\left[(r-E[r\mid x])^2 \mid x\right]}_{noise} + \underbrace{(E[r\mid x]-g(x))^2}_{squared\ error}$$

$$E_X\left[(E[r\mid x]-g(x))^2 \mid x\right] = \underbrace{(E[r\mid x]-E_X[g(x)])^2}_{bias} + \underbrace{E_X\left[(g(x)-E_X[g(x)])^2\right]}_{variance}$$

# BIAS/VARIANCE

- **Error due to Bias**: The error due to bias is taken as the difference between the expected (or average) prediction of our model and the correct value which we are trying to predict. Of course you only have one model so talking about expected or average prediction values might seem a little strange. However, imagine you could repeat the whole model building process more than once: each time you gather new data and run a new analysis creating a new model. Due to randomness in the underlying data sets, the resulting models will have a range of predictions. Bias measures how far off in general these models' predictions are from the correct value.
- **Error due to Variance**: The error due to variance is taken as the variability of a model prediction for a given data point. Again, imagine you can repeat the entire model building process multiple times. The variance is how much the predictions for a given point vary between different realizations of the model.



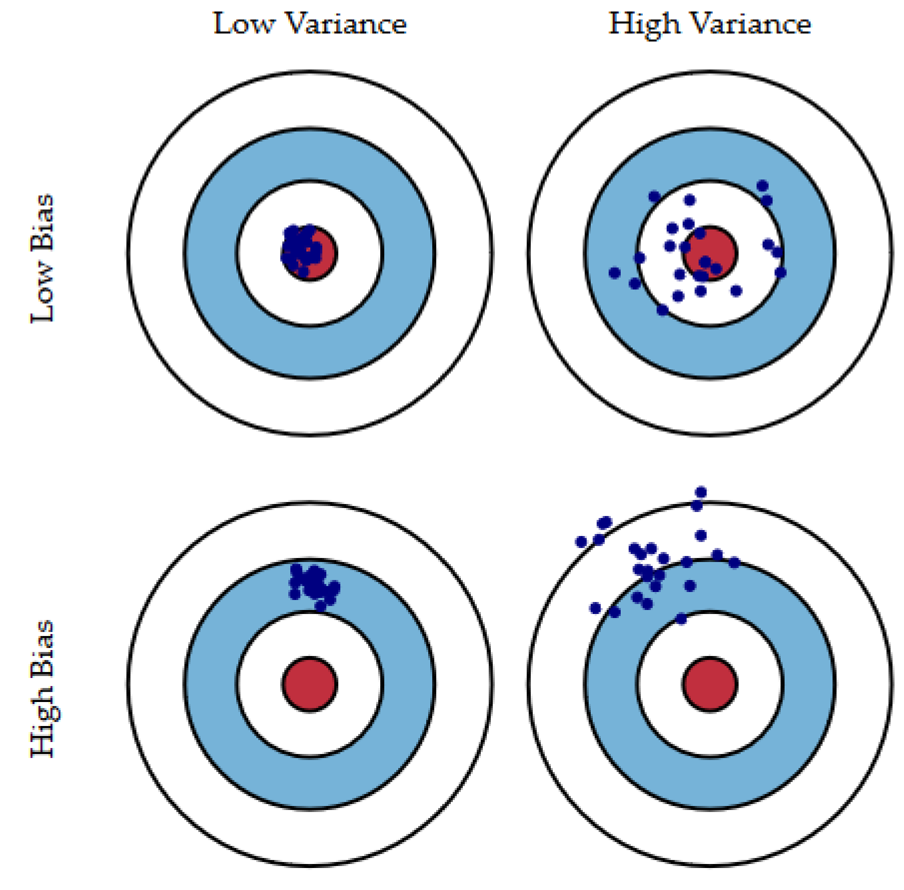




Fig. 1 Graphical illustration of bias and variance.

# UNDERFITTING/OVERFITTING

If there is bias, this indicates that our model class does not contain the solution; this is *underfitting*. If there is variance, the model class is too general and also learns the noise; this is *overfitting*. If $g(\cdot)$ is of the same hypothesis class with $f(\cdot)$, for example, a polynomial of the same order, we have an unbiased estimator, and estimated bias decreases as the number of models increase. This shows the error-reducing effect of choosing the right model (which we called *inductive bias* in chapter 2—the two uses of "bias" are different but not unrelated). As for variance, it also depends on the size of the training set; the variability due to sample decreases as the sample size increases. To sum up, to get a small value of error, we should have the proper inductive bias (to get small bias in the statistical sense) and have a large enough dataset so that the variability of the model can be constrained with the data.

Understanding bias and variance is critical for understanding the behavior of prediction models, but in general what you really care about is overall error, not the specific decomposition. The sweet spot for any model is the level of complexity at which the increase in bias is equivalent to the reduction in variance. Mathematically:

$$\frac{dBias}{dComplexity} = -\frac{dVariance}{dComplexity}$$

If our model complexity exceeds this sweet spot, we are in effect over-fitting our model; while if our complexity falls short of the sweet spot, we are under-fitting the model. In practice, there is not an analytical way to find this location. Instead we must use an accurate measure of prediction error and explore differing levels of model complexity and then choose the complexity level that minimizes the overall error. A key to this process is the selection of an *accurate* error measure as often grossly inaccurate measures are used which can be deceptive. The topic of accuracy measures is discussed here but generally resampling based measures such as cross-validation should be preferred over theoretical measures such as Aikake's Information Criteria.

https://stats.stackexchange.com/questions/5135/interpretation-of-rs-lm-output

http://people.sc.fsu.edu/%7Ejburkardt/datasets/regression/regression.html

http://scott.fortmann-roe.com/docs/BiasVariance.html

https://stats.stackexchange.com/questions/2358/explain-the-difference-between-multiple-regression-and-multivariate-regression
https://www.quora.com/What-is-the-difference-between-a-multiple-linear-regression-and-a-multivariate-regression

Gasconsumption:
http://people.sc.fsu.edu/%7Ejburkardt/datasets/regression/x16.txt

# REFERENCES

# VIF