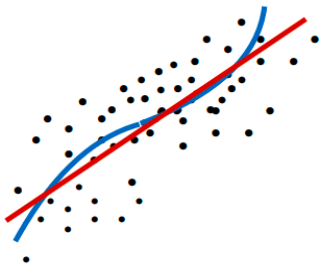
## Reliability of Coefficient Estimates

We have already discussed the prediction performance i.e how well the predictor is doing on our particular training set. That's one aspect of performance. Now we will discuss different aspects, which has to do with how accurate are the estimates of $\theta^*$.

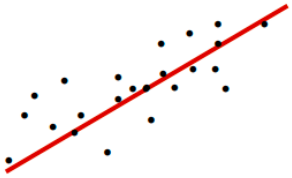
### How noisy/reliable are my estimates of weights?

This will only make sense if we believe that there's a true $\theta^*$ in there that we're trying to estimate.
In one of the settings there's a large true population and we're interested in coming up with the best linear predictor.

- Large true population

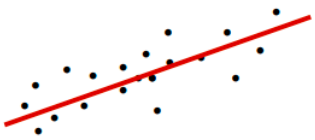


- Finite sample: find best linear fit



- true relation may be complex
- interested in best linear predictor

We have a finite sample, and we're trying to find the best linear fit on that finite sample.
If we have happened to have a different finite sample which is still representative from that population, but different individuals. Then the line that we get would be somewhat different. So there's variability in the estimates that we get i.e the estimates of the θs depending on which particular sample we ended up getting.



- Another finite sample: different results

That means that the randomness in our sampling affects the results.

So the results are somewhere random, **but how random are they or how much variation do we expect**?
Do we expect both the above red lines to be almost the same or could they vary wildly?
If they vary wildly, it means that we cannot really trust our best linear fit because we know that if we do the experiment another day, we will get something very different.
But if they do not vary too much, then we can trust the results better.

To make this way of reasoning a little more precise, we assume that there are some true $\theta^*$ back there. We assume that the world is linear, and that the model satisfies the reasonable assumptions. The Ys are linear functions of the Xs plus some idiosyncratic noise.

- Assume structural model

$$Y_i = (\theta^*)^T X_i + W_i$$
$W_i$: independent,
zero mean, variance $\sigma^2$

(If not, need to resort to simulation/bootstrap methods)

If these structural assumptions are not true then there's still things that we can do in order to assess the accuracy or the variability of our estimates. But for these we have to work more directly with the data. That's a thing for later discussion. But for now, we will try to work with formulas and analytical methods.
The analytical methods are based on the fact that the optimal estimates i.e the output of linear regression is given by a simple formula. The estimates are functions of the Xs and the Ys. Now the Ys have noise in them. So the θs are also noisy.

- Regression: $\widehat{\Theta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{Y}$

$\widehat{\Theta}$ is a random variable
(depends on random data)

So $\hat{\theta}$ is a random variable. As it is a random variable. It has some statistical properties. We hope that the estimate is close to the particular value of the parameter.

And here we are taking the index j. So we're looking at just one of the parameters, the jth parameter i.e only one component of the theta vector. There's a true value of that parameter which is $\theta^*_j$. There's the estimated value $\hat{\theta}_j$. And we look at their squared difference, which tells us how far off we are.
Now, because $\theta$ is random. So we can only talk about how far off we're going to be on the average. So we take an expected value here which is $E(\hat{\theta}j - \theta^* j)^2$

So if you're sampling individuals over a population, the expectation is that overall population from which you are sampling. So it's the expectation with respect to the randomness in our data sets. And it's also the expectation with respect to the noises W, that affects our estimates.

Now, this is the mean squared error of our estimation procedure and it tells us whenever we apply that procedure on average what kind of errors we expect to get. And using a formula from elementary probability this mean squared error can be decomposed into two terms.

$$E[(\hat{\theta}j - \theta^* j)^2] = (E[\hat{\theta}j] - \theta^* j)^2 + var(\hat{\theta}j)$$

The first term asks, if the true value and the estimate that we are going to get on the average, are they systematically different or not? If this quantity is non-zero, we say that the method is biased, and we have a non-zero bias. So the **bias** here is: $(E[\hat{\theta}j] - \theta^* j)^2$

And there's another term that has to do with the variability of our estimates. So each time that we do the experiment, we're going to get a different estimate. And how different are the estimates going to be?
That term is the **variance** of the estimates. [$var(\hat{\theta}j)$]

So that's a general situation that's happening in estimation problems. There's always some possibility for a bias which is a systematic tendency for estimates to overshoot or undershoot. And there's another term that has to do with the amount of randomness in these tests.

If the estimator is unbiased, i.e., $E[\hat{\theta}j] = \theta^{*}j$, the bias is zero and only the variance component contributes to the error. Fortunately, the estimator is unbiased in linear regression, hence we focus on the variance of $\hat{\theta}j$ [ $var$ ( $\hat{\theta}j$)].

So now, if we apply the same estimation procedure but on a new sample i.e on new data sets, how different are the $\hat{\theta}$ going to be?

Let's try to understand what's going on here.

**The distribution of $\hat{\theta}$ :**

Now, in order to talk about the variance of $\hat{\theta}$, we need to kind of understand what's the probability distribution of $\hat{\theta}$.

Remember that $\hat{\theta}$ is a random variable, because when determining $\hat{\theta}$, we take into account the Ys and the Ys have some noise terms in them.
But, what kind of random variable is it?

$$Y_i = (\boldsymbol{\theta}^*)^T \mathbf{X}_i + W_i$$
$$W_i: \text{ independent, zero mean, variance } \sigma^2$$
$$\widehat{\boldsymbol{\Theta}} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{Y}$$

Let's look at a typical component. of the vector $\hat{\theta}$, the jth components.

Then it turns out that this component is a random variable that has a normal distribution. And this is true, approximately not literally. And it is true because of the central limit theorem when we're dealing with large datasets. With a large data set there's lots of random W's that come in to formula for $\hat{\theta}$ and they add up. And when you add up many random things, in the limit you tend to get normal random variables.

$\hat{\theta}$ is going to be **exactly a normal** random variable, if we make the additional assumption that the $W_i$ are normal.

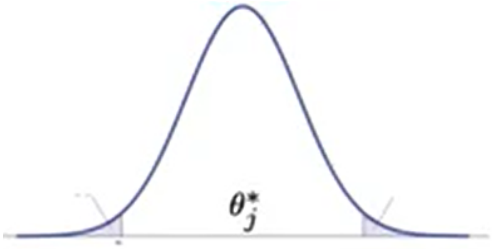
In fact, mathematically $\hat{\theta}$ is a vector that has a so-called **multivariate** normal distribution, which means that linear combinations of the different components are also normal.

So symbolically, we could write the following.

$$\widehat{\Theta}_j \sim \mathcal{N}(\theta_j^*, \sigma_j^2)$$

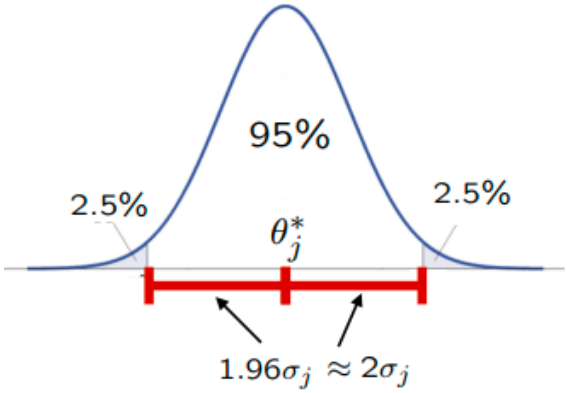
$\hat{\theta}$ is a normal random variable. On the average the mean of this normal is the true parameter( $\theta^* j$). This is the fact that the estimates are unbiased, and it has some randomness, which is captured by the variance in our estimates ($\sigma_j^2$).

Now pictorially $\hat{\theta}$ has a distribution like below, which is a plot of the normal distribution.




If we collect data, according to the above model, and construct an estimate, that estimate is going to be somewhere around the exact value of the parameter. And it's going to be a random variable. And on the average the mean of that random variable is the true value because the estimate turns out to be unbiased, and it has a certain variance.

That variance determines the width of this normal distribution. For normal distributions, we know that if we go about two standard deviations away from the mean then that covers 95% of the distribution, and leaves only 2.5% probability on either end. So that's a very useful property of normal random variables. With a probability of 95%, we fall within two standard deviations from the mean.



So this standard deviation is an important quantity. And it's called the **standard error** of the estimate.

$$\sigma_j = \sqrt{\text{var}(\widehat{\Theta}_j)} = \text{se}(\widehat{\Theta}_j)$$

standard error

The standard error basically is an indication of how large are the errors in our estimates when we're trying to estimate a particular component of the $\theta$ vector.

And what follows from here is that, the standard error is very important with respect to the accuracy. So we want to have some way of calculating that standard error. And once we have it, then we can use it for various purposes. Those purposes could be to build so-called confidence intervals, and to run hypothesis tests.