Covariance Matrix of Coefficients

We have discussed already that $\hat{\theta}$ is a random variable, and by using the properties of this random variable especially variance we can do useful things about confidence interval and hypothesis testing.
But in order to do that, we need to somehow get hold of the standard error of the different components of $\hat{\theta}$. Let's see how this is done.

The Covariance Matrix of $\hat{\theta}$:

Suppose we have the structural model and we believe that the world is linear. Under this assumption we will see how this can be done. Alternatively, it would have to be done in a different data driven way if we do not believe these assumptions.

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

Now, in this case what happens is, the linear regression software produces an estimate of the covariance matrix of $\hat{\theta}$.

$$\mathbb{E}\left[(\widehat{\Theta} - \theta^*)(\widehat{\Theta} - \theta^*)^T\right]$$

dimensions $(m+1) \times (m+1)$

Here, theta is a vector and then we take the transpose of that vector. So column vector times a row vector give us a matrix. So the quantity here is a matrix and we take the average value of that matrix.

This matrix gets estimated and reported by regression software. The main part of that matrix that we're interested in are the **diagonal entries**. Each one of the diagonal entries gives us an estimate of the variance of each component of the estimated parameters.

diagonal entries: $\text{Var}(\widehat{\Theta}_j)$
off-diagonal entries: $\text{Cov}(\widehat{\Theta}_i, \widehat{\Theta}_j)$

The off-diagonal entries correspond to covariances between the different estimates. We will not have any use for these. We will just concentrate and work with the diagonal entries.

So the diagonal entries give us an estimated variance and by taking the square root of that we get an estimated variance. The estimate for the standard error is given by a certain formula that takes the data i.e the X's of all the records that we have and then apply some linear algebraic matrix operation that gives us the formula at the end.

formula: $\sigma^2(\mathbb{X}^T\mathbb{X})^{-1}$

use $\hat{\sigma}^2$

But this formula involves the variance of the noises. That variance of the noises is not known. So it has to be estimated, and this is what the linear regression software does. So it uses an estimate of the noise variances instead of the true value.

To get that estimate of the variances, there's a simple formula that we use.

- Estimate σ^2 by $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{\theta}^T\mathbf{X}_i\right)^2$ slight downwards bias negligible bias if $m \ll n$

Why? For large samples, $\widehat{\Theta} \approx \theta^*$, and

$$\sigma^2 = \mathbb{E}[W_i^2] \approx \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - (\theta^*)^T\mathbf{X}_i\right)^2 \approx \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{\Theta}^T\mathbf{X}_i\right)^2$$

We take the predicted values of the different Y's and subtract them from the actual values and take the square of that. This is the prediction error and if our predictions are very good i.e. if we knew the correct θ^*, then those predictions will be the best possible and the error would be just the noise.

So if we believe that this term $\left(Y_i - \hat{\theta}^T\mathbf{X}_i\right)$ is approximately the same as noise, then we're taking the average of the square of these various noise terms and that's a good estimate of the variance of the noise terms.

In the bottom of the picture above there is an equation, and you can see sort of what the mathematics are here. σ^2 is the mean squared value of the W's i.e of the noises. We estimate the expectation by taking the average over the noises in our actual data set. And within our data set, the noises are the actual value of Y's minus $\theta^* X$, but θ^* is not known. On the other hand with a large enough data set θ^* is approximately the same as our estimate. So we plug in the estimate there and that justifies the displayed formula that we have here.

The bottom line of all of this is that we have a systematic way for estimating the noise variance σ^2, and from that we have a way i.e a formula for estimating the variances and the standard errors of the different components of θ .

And now we can use them in our example.

Error covariance matrix:

	const	TV	Radio	Newspaper
const	**9.72867479E-02**	-2.65727337E-04	-1.11548946E-03	-5.91021239E-04
TV	2.65727337E-04	**1.9457371E-06**	-4.47039463E-07	-3.26595026E-07
Radio	-1.11548946E-03	-4.47039463E-07	**7.41533504E-05**	-1.78006245E-05
Newspaper	-5.91021239E-04	-3.26595026E-07	-1.78006245E-05	**3.44687543E-05**

	coef	std err	Confidence intervals [0.025	0.975]
Intercept	2.9389	0.312	2.324	3.554
TV	0.0458	0.001	0.043	0.049
Radio	0.1885	0.009	0.172	0.206
Newspaper	-0.0010	0.006	-0.013	0.011

The software reports us the estimates of the different coefficients (**coef**).

It also takes the square root of the diagonal entries and these square roots correspond to the **standard errors (std err)** for the different components, and using the standard errors, we can create confidence intervals. So, we take the estimates that go plus or

minus two standard deviations or two standard errors away from the estimates. And confidence intervals that we get is shown in the above image. These are 95% confidence intervals and that's indicated by those entries there that we have 2.5% on each end of the confidence interval. So, there's 2.5% probability that the confidence interval would miss θ^* on the one side or the other side.

Let's look into the confidence interval.

Confidence intervals	
[0.025	0.975]
2.324	3.554
0.043	0.049
0.172	0.206
-0.013	0.011

We see that the first three confidence intervals do not contain zero. So the hypothesis that the coefficient is zero is incompatible with the data.

So when we run the wald tests, the wald test will reject the null hypothesis. It will reject the hypothesis that θ^* is zero for those three coefficients.

So we say that the constant term (intercept), the TV coefficient and the radio coefficients are significant. They do not seem to be zero. The data are incompatible with those being zeros.

On the other hand, for newspaper, we see that zero is inside the confidence interval and that means that the data are compatible with the hypothesis that the coefficients associated with newspaper budgets is zero. And so we keep that null hypothesis i.e. we do not reject it or to say maybe we accept it.

Now the meaning of those words, **reject** and **accept**, can be somewhat ambiguous, and they're the source of a lot of confusion.
So, we will briefly discuss how such things are to be interpreted.

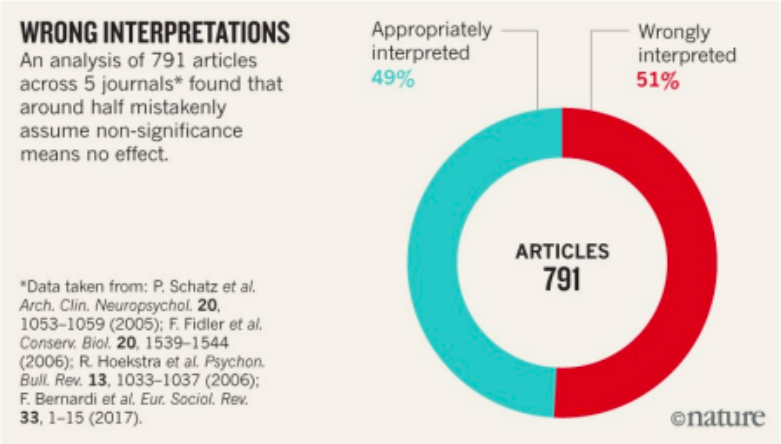
Interpretation Needs Care:

The interpretation of the results of hypothesis tests needs to be careful. It's one thing to report the results of a Wald test and it's a different thing to say what we actually mean by accepting or rejecting various hypotheses.

And statisticians have been quite unhappy with the practice in this field.

There is a claim that in a very large fraction of published papers the interpretations of the statistical results are wrong.

Scientists rise up against statistical significance, *Nature*, 20 March 2019



So what is the right interpretation? What words should we be using and what inferences can we make?

Suppose that we run the test and **we reject the null hypothesis** $\theta^*_j = 0$.

Then, it means that the data seem to be inconsistent with the null hypothesis. More exactly, it means that if θ^* was zero, we would have seen the data that we saw with a chance of less than 5%. So the data seem to be incompatible. Of course, it's possible that just by chance, or because of the noise, we do reject the null even though the null is correct, but that would happen at most 5% of the time. We need to keep in mind these outcomes as well. The fact that we reject it doesn't mean that we're perfectly sure. We can only say that 95% of those rejections will be correct, and 5% of the time, we will be incorrectly rejecting a true hypothesis.

What might be a little trickier is what does it mean **when we do not reject the null hypothesis**. So in some sense, we're sort of accepting the hypothesis of $\theta^*_j = 0$.

But we're not really accepting the hypothesis that θ^*j **= 0**. We're only saying that θ^*j equal to zero is possible.

The data do not give us any compelling evidence that $\theta^*j \neq$ **0** .

That's all that we can say. Maybe θ^*j is non-zero. But the data does not give us the evidence for that. So we cannot make that statement and we cannot reject it.

Some of the possibilities where we do not reject the null:

- **No effect**: It could be, indeed, that the null is true. Maybe θ^*j **= 0** .
- **Small effect**: It's also possible that θ^*j is nonzero. Maybe it's nonzero but it's so close to zero that the data cannot detect it. So we retain the null hypothesis. As the data doesn't give any evidence that θ^*j is big. So we do not reject the null hypothesis.
- **Too few data**: Another possibility is, maybe θ^*j is quite far from zero. But our data set is very small, and so it doesn't give us the evidence that θ^*j is non-zero.

To summarize, the cases where we sort of accept the hypothesis θ^*j **= 0** or the null hypothesis. That can happen either because the null hypothesis is true, or because even though the null hypothesis is not true, the data is not informative enough and they cannot tell us the exact case.