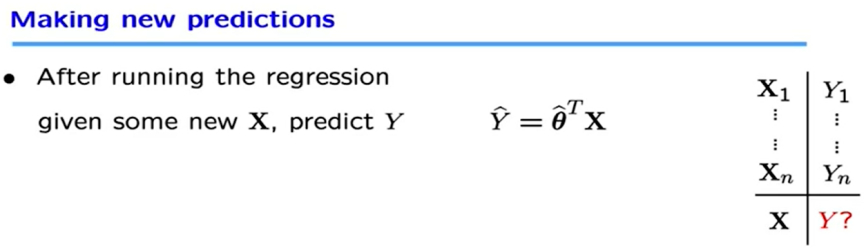
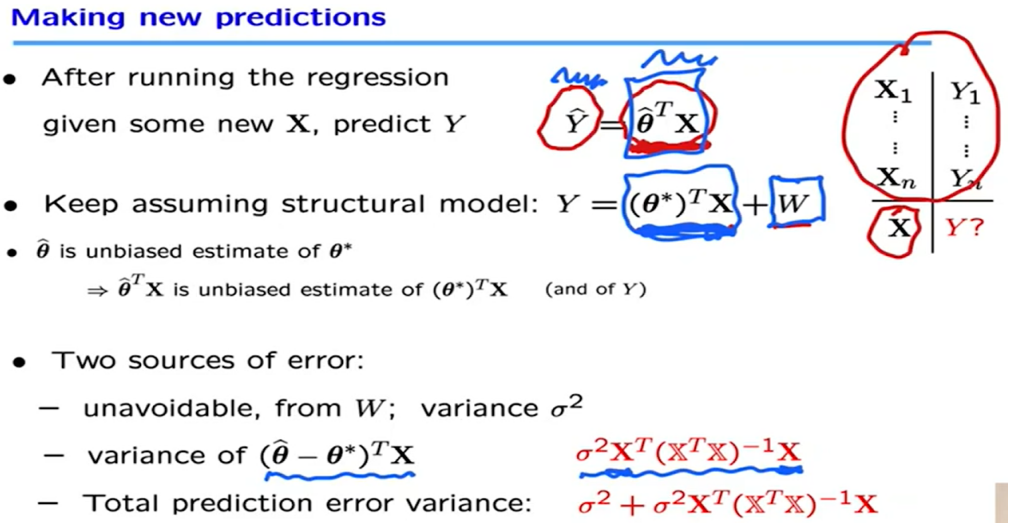
Making New Predictions

So we discussed quite extensively that theta hats are random and so they have some errors. But now let us use the estimated theta hats to actually make predictions and discuss the errors in these predictions. So we run the regression using our data set. And using that data set, we come up with some estimates of theta hat and then a new person comes in, and we want to make a prediction for their Y. And we make their prediction by using this linear relation with the coefficients that we have learned through our estimation. How good are those predictions?



To continue let us make the assumption that the world is truly linear and there is some true θ^*. The true relation is linear plus some noise. Under those circumstances, we know that theta hat is an unbiased estimate of θ^*. So on the average theta hat doesn't go higher or lower on the average is about right. Since it has no bias, that means that the terms here are also an unbiased estimate of Y hat. So on average, our predictions will be right. But that's only on the average. Sometimes they will be high, sometimes they will be low. How much of an error is there going to be to study that we need to analyze the various sources of errors.



Why do we make errors in our predictions?

One source of error is that W is unknowable. It's idiosyncratic to a particular individual. The best we could do even if we knew θ^* would be to predict Y according to this. There's nothing we can do about W, W has certain variance and so there's some error that we're going to get because of that. But there's a second source of error. Since we do not know the θ^* exactly the error between our estimate and the true value of θ^* will cause an error when predicting this term by using that term. So the difference between those terms is this expression here and this expression has some variance, some randomness that's due to the variance in theta hat while there is a formula for the variance of the theta hats, and so using that formula one actually can write down the formula for mean squared prediction error that's caused by the inaccuracy in these estimates. Mean squared prediction error actually depends on the X's. For certain X's we will expect to have bigger prediction errors and for some, we're going to have smaller prediction errors and the reason is that our estimates are produced by multiplying Thetas by X's. If we multiply with a big X, the errors in theta are going to be amplified and that gets reflected in the variance of the corresponding predictor.

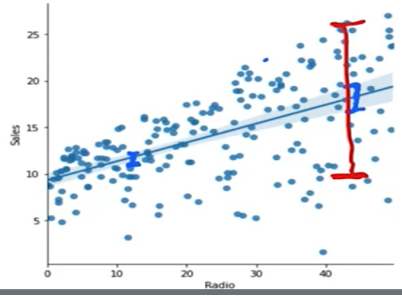
So we can add up together those two terms and we get the formula for the total error variance in our predictor.

How do we use those formulas?

We use those formulas to form so-called Confidence Bands. These are confidence intervals around the estimated prediction, they're confidence intervals about what we think about θ^*. So, it's the same recipe as before. We take the estimated value of Y, we take our prediction, and add plus or minus two standard deviations, and that gives us a confidence interval for this quantity here. This is again, a confidence interval whose width changes with X. That's for the same reasons that we've discussed before. If we're dealing with simple regression, where X is just one dimensional you can plot those confidence intervals as a function of X. You see that for large X the confidence interval tends to be wide. For smaller X's the confidence interval tends to be narrower. when you put all those confidence intervals together we get the confidence band. This confidence band basically tells you where you think or where this true line might be.

Confidence bands

- 95% confidence interval about the value of $(\theta^*)^T\mathbf{X}$:
  $(\hat{\theta})^T\mathbf{X}$ plus or minus $2 \cdot \hat{\sigma}\sqrt{\mathbf{X}^T(\mathbb{X}^T\mathbb{X})^{-1}\mathbf{X}}$
  - confidence interval width changes with $\mathbf{X}$
  - in simple regression, this gives a confidence band



The estimated regression line is certainly not exactly the same as the true line and the confidence bands give us a sense of where the true line might be. So these are confidence intervals for this part of the model. Of course, there's additional randomness involved and if we use that additional randomness, we end up getting wider confidence intervals about where we think Y is, because Y has extra noise, we would be getting a much wider confidence interval that tells us where we think Y would be falling in. To summarize what we have discussed so far, we have done most of our analysis by assuming the structural model that the world is linear. There's a true linear relation plus some noise that's idiosyncratic to each individual. Linear regression can be used for predictions, even without assuming a structural model. But when we have a structural model, we also have the hook of recovering that structural model exactly. We build an estimator, and the formula for the estimator is the same, no matter which interpretation of linear regression we're using. The formula itself doesn't really matter.

- Structural model: $Y_i = (\theta^*)^T\mathbf{X}_i + W_i$
- Estimator: $\widehat{\Theta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{Y}$
- Predictor: $\hat{Y} = \widehat{\Theta}^T\mathbf{X}$
- Standard error σ_j: standard deviation of $\widehat{\Theta}_j$
- 95% confidence interval: $\text{CI} = [\widehat{\Theta}_j - 2\sigma_j, \widehat{\Theta}_j + 2\sigma_j]$
- Wald test: reject "$\theta_j^* = 0$" if $0 \notin \text{CI}$

The important thing is that there is a formula that's easy to implement in software, and there's very fast software for doing that. Once we have an estimator, we can use it to make predictions. We have a predictor. However, this predictor is going to have some errors and one source of errors is that the theta hats may not be perfectly accurate. The accuracy or inaccuracy of the theta hats is summarized by reporting the standard error, which tells us how much these estimates vary because of the noise in our sampling and measuring procedures. These standard errors are very important quantities. They tell us something about the accuracy of the estimate. They are reported by regression software and they're used in various ways. One is to construct confidence intervals. For example, for a 95% confidence interval, we get plus or minus two standard errors from the estimated value and we also use them for hypothesis testing. Once we construct the confidence intervals, then we check if the zero value is inside the confidence interval or outside, and depending on that we accept or reject the hypothesis that the corresponding coefficient is zero or not zero. But as we have said before, the words accept or reject have to be used carefully. They have very precise meanings for what we're saying and what we are not saying and now that we have the basic pieces in our hands, we could start using them and we're going to start using them first on a conceptual level, but discussing all the various things that might go wrong.

What we've discussed here in the form of the confidence intervals, for example, the Wald test, all depend on certain mathematical assumptions and once those assumptions are violated, then we might start becoming suspicious about the results.