Inference About a Future Value of $Y$

A regression model may be fitted to learn about the association of $Y$ and $x$, represented by $\beta_0$ and especially $\beta_1$.

However, sometimes the intent is to make inferences about the likely values of $Y$ under new conditions.

We might want to learn about the distribution of $Y$ when pH = 7.5, which is not one of the values in the data set.
In the regression model, when $x$ has some new value $x^*$,

$$E(Y) = \beta_0 + \beta_1 x^*,$$

so the natural estimator of $E(Y)$ is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*.$$

We can show that $E(\hat{Y}) = \beta_0 + \beta_1 x^* = E(Y)$, so $\hat{Y}$ is an unbiased estimator of $E(Y)$.

To construct confidence intervals for $E(Y)$, we need the standard error of $\hat{Y}$; the formula is known, but using software is simpler.
In R

```r
arsenicLm <- lm(Percent ~ pH, arsenic)
predict(arsenicLm, data.frame(pH = 7.5), se.fit = TRUE,
        interval = "confidence")
```

Output

```r
$fit
   fit  lwr  upr
1 55.01145 50.67454 59.34837

$se.fit
[1] 2.045806

$df
[1] 16

$residual.scale
[1] 6.125584
```
In the R output, `fit` is $\hat{Y}$, and `se.fit` is its estimated standard error.

`lwr` and `upr` are the endpoints of the confidence interval for $E(Y)$, by default the 95% confidence interval.
Predicting the Future Value of $Y$

Note: $E(Y)$ is the expected value of $Y$ when $x = x^*$; in the example, it is the capability of the process to remove arsenic from water with a pH of $x^* = 7.5$.

Sometimes we need to predict the observed value of $Y$ in a future experiment with $x = x^*$.

Since $Y = E(Y) + \epsilon$ and $E(\epsilon) = 0$, the best predictor of $Y$ is still $\hat{Y}$. 
But

\[ V(Y - \hat{Y}) = V\{[Y - E(Y)] + [E(Y) - \hat{Y}]\} \]
\[ = V[Y - E(Y)] + V[E(Y) - \hat{Y}] \]
\[ = \sigma^2 + V[\hat{Y}]. \]

The *prediction interval* for \( Y \) is also centered at \( \hat{Y} \), but is wider than the confidence interval.
In R

The same `predict()` method is used, but with an option to make the interval appropriately wider:

```r
predict(arsenicLm, data.frame(pH = 7.5),
        interval = "prediction")
```

Output

```
  fit  lwr  upr
1 55.01 41.32 68.70
```

Note that the prediction interval has a width of 27.4, whereas the confidence interval has a width of 8.7.