### 0.1 reviewer comment for reference

```
The statistical evaluation of the fitted model describes a situation to avoid,
as I have repeatedly stressed in my graduate level statistics classes.
I find that the R2 statistics is overly simplified and overly emphasized.
When using the R2 value, students will naturally equate a model with a higher
value as a better model, a simplistic, and often misleading, view.
Once this idea is introduced, we cannot easily correct the misconception.
So, please do not plant the misconception in the first place. The R2 value
is a measure of linear association. For a simple linear regression model (with one predictor),
the R2 value is the square of the correlation coefficient between x and y when the
relationship between x and y is actually linear.
```

### 0.2 Walk through the regression portion my potalk

**What is our model?**

For linear regression, the model is:

$$y \sim \theta_1 x + \theta_0 + \epsilon \tag{1}$$

Where the slope $\theta_1$ and the intercept $\theta_0$ are (in the frequentist interpretation) numbers we have estimated somehow, and $\epsilon$ is a random variable that represents variablity that isn't captured by the linear relationship. Note that this means that y is also a random variable, and we strictly need to write ($\sim$: "has the probability distribution of") rather than (=: "is equal to"). This nuance makes a big difference in everything that follows.

### 0.3 Modeling $\epsilon$

For standard ordinary least squares regression, there are some strong constraints on $\epsilon$. It is assumed to be normally distributed with mean 0 and constant variance $\sigma^2$. That is," it takes the form:

$$\epsilon(y;\sigma) \approx \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y-(\theta_1 x+\theta_0))^2}{2\sigma^2}} \tag{2}$$

and this means that the full model also has a probability distribution that is given by:

$$p\left(y|x;\theta_0,\theta_1,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y-(\theta_1 x+\theta_0))^2}{2\sigma^2}} \tag{3}$$

How do we use the model to make a prediction? If the underlying process is accurately captured by (3) then if you give me $(x,\theta_0,\theta_1,\sigma)$ I can say that, if we sampled the process repeatedly, 95% of the time we would get a measurement of y that lay in the range $\theta_0 + \theta_1 x \pm 2\sigma$ (See slide 8 )

**But I want a number not a confidence interval!**

Because we have $p(y)$, we can find $\overline{y}$, the mean value, or first moment, of y:

$$\overline{y} = \int_{-\infty}^{\infty} y\, p\left(y|x;\theta_0,\theta_1,\sigma^2\right) dy = \int_{-\infty}^{\infty} (\theta_1 x + \theta_0 + \epsilon) dy = \theta_1 x + \theta_0 \tag{4}$$

This follows since the only thing that is a function of $y$ is $\epsilon$, and $\overline{\epsilon} = 0$.

Note that we can also find the probability, for example, that if $x = 5$, $y > 10$ by doing this integral:

$$\int_{10}^{\infty} p\left(y|10;\theta_0,\theta_1,\sigma^2\right) dy \tag{5}$$

**Estimating $\theta_0$ and $\theta_1$**

How does a frequentist find estimates of the slope and the intercept? They start with this model, and assume that we have universes as plenty as blackberries each generating (x,y) values from the process described by (1.

We can imagine independently drawing a large number of measurements of (y,x) pairs from the different universes. Since we have the probability distribution of our model, we can calculate the probability that any particular sample $(x_i, y_i)$ will be observed:

$$p\left(y_i|x_i;\theta_0,\theta_1,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y_i-(\theta_1 x_i+\theta_0))^2}{2\sigma^2}} \tag{6}$$

Since we are making independent draws, the probability that will see a particular set of $X \in (x_i, y_i)$ pairs is just the product of each of their individual probabilities:

$$L_X\left(\theta_0,\theta_1,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \prod_{(x,y)\in X} e^{\frac{-(y-(\theta_1 x+\theta_0))^2}{2\sigma^2}} \tag{7}$$

This is called the “likelihood”.

**Maximum likelihood**

Solve this for the set of parameters that give the **maximum likelihood** by taking the log and finding the maximum by setting the derivative = 0 and solving for $(\theta_0, \theta_1)$. This gives you the usual relationship for the slope and intercept in terms of the sample statistics ($\overline{x}$,$\overline{y}$). Note that we don't need to know $\sigma$, because we're assuming it's constant.

$$l_X\left(\theta_0,\theta_1,\sigma^2\right) = \log\left[\frac{1}{\sqrt{2\pi\sigma^2}} \prod_{(x,y)\in X} e^{\frac{-(y-(\theta_1 x+\theta_0))^2}{2\sigma^2}}\right]$$

$$= -\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2} \sum_{(x,y)\in X} [y - (\theta_1 x + \theta_0)]^2$$

**Which yields**

$$\overline{y} = \theta_1 x + \theta_0 \tag{8}$$
$$\theta_0 = \overline{y} - \theta_1 \overline{x} \tag{9}$$
$$\theta_1 = \frac{\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_i^n (x_i - \overline{x})^2} \tag{10}$$

**To repeat,we've assumed $\sigma = constant$ so the value of particular value of $\sigma$ doesn't change the location of the maximum. That doesn't mean that it is not still part of the model though.**

### Link to ordinary least squares

Miraculously, in the specific case of this model, maximising the likelihood is the same as minimizing $\chi^2$ so we can just turn that crank, but hopefully not forget all of the above.

### Uncertainty in $\theta_0$ and $\theta_1$

We've found a single estimate of $(\theta_0, \theta_1)$ for a single sample. What if we draw other samples from different universes? Then you get something like slide 18 and slide 19

## 0.4 Estimating $\sigma^2$ (Bayesian)

How do we estimate the variance? In my talk, I show how a Bayesian would do this – see slide 25

## 0.5 Frequentist results

(for completeness, here's a typical set of textbook equations)

### Estimating $\sigma^2$ (frequentist)

Define:

$$\overline{y_i} = \theta_1 x_i + \theta_0 \tag{11}$$

and

$$\widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y_i})^2 \tag{12}$$

Then an unbiased estimater of $\sigma^2$ is:

$$\frac{n}{n-2}\widehat{\sigma^2} \tag{13}$$

**Frequentist confidence interval for $\overline{y_i}$**

The 95% confidence interval for a prediction $\overline{y_i}$ at a particular $x_i$ is

$$\overline{y_i} \pm t_{n-2} s_y \sqrt{\frac{1}{n} + \frac{\left(x_i - \bar{x}\right)^2}{(n-1)s_x^2}} \tag{14}$$

where

$$s_y = \sqrt{\frac{\sum \left(y_i - \overline{y}_i\right)^2}{n-2}} \tag{15}$$

## 0.6 Summary

In light of the above, a sentence like:

“The linear regression model can still only weakly predict the October CO2 values based on time.” isn’t quite right, because it’s ignoring the $\epsilon$ specification that is part of the model. A good model that properly captures an intrinsically large $\sigma^2$ is going to have low covariance, but that’s the right answer, not a failure of the model.