CSI 2300: Lecture 20 – Linear Regression with Multiple Independent Variables

Outline

In this lecture we'll discuss linear regression with multiple independent variables.

- Why multiple independent variables?
- What does the mathematical model look like?
- How should we understand the model?
- Pitfalls to avoid

Multiple Independent Variables

The world is not simple, one-to-one. That is, there are usually multiple influences for something. I love Mexican food, but the probability that I'm going to eat it on any given day (my dependent variable) has **multiple** influences (independent variables), including:

- When was my last meal?
- What time of day is it?
- Where am I physically located?

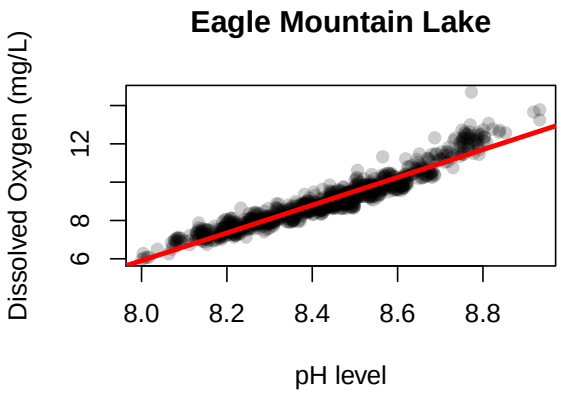
So I may want to include all of those things in predicting my likelihood to eat Mexican food soon.

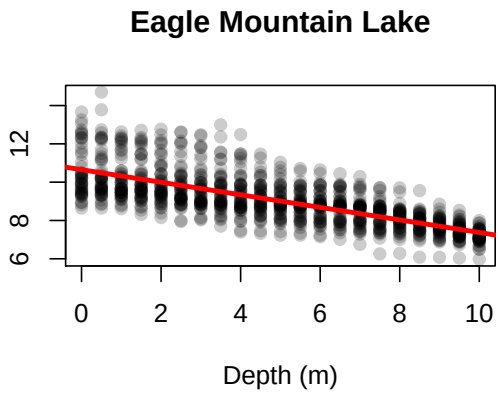
Suppose we're trying to understand the dissolved oxygen (DO) content in Eagle Mountain Lake. In the past, we've constructed the linear model which we expressed in R as `DO ~ pH`. This model has two parameters: a slope (for pH) and an (implied) intercept. In other words:

$$DO = \text{slope} \cdot pH + \text{intercept}$$

(When we run `lm(DO ~ pH, data=eml)` to construct a linear model, it is fitting a line to that model by finding the best two parameters for the observed data. Recall that the best fit is defined as minimizing the squared distance between each example and the regression line.)

But is pH the only influence? Probably not.





From these plots, we can see that **both** pH and Depth have some correlation with DO. The correlation between DO and Depth makes sense – oxygen is less prevalent in deeper water which is under more pressure.

How can we understand how pH and Depth both affect DO?

Multiple Regression Combines Independent Variables

We can construct a linear regression model that uses more than one independent variable! Let's move from one to two independent variables.

- In R: `DO ~ pH + Depth` (the intercept is still implied)
- In math: $DO = \text{slope1} \cdot pH + \text{slope2} \cdot Depth + \text{intercept}$

Note that since we have two independent variables and one intercept, we now have **three** parameters (which we called "slope1", "slope2", and "intercept") that the linear model must fit.

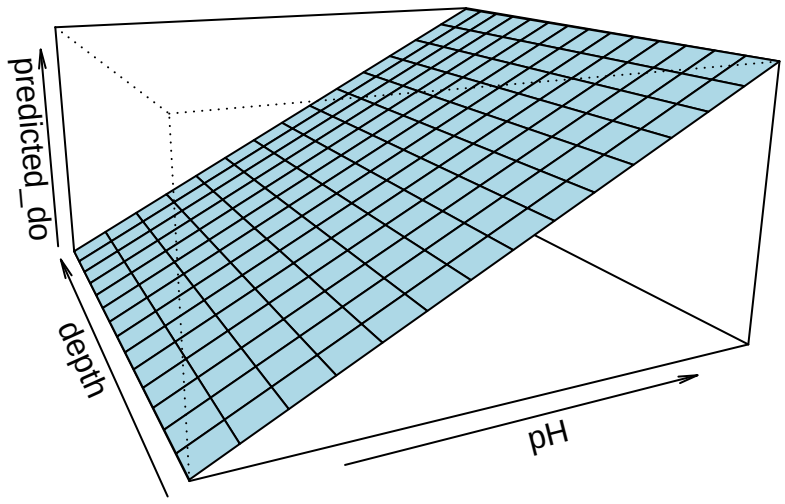
```
(m_ph_depth <- lm(DO ~ pH + Depth, data=eml_small))
```

```
##
## Call:
## lm(formula = DO ~ pH + Depth, data = eml_small)
##
## Coefficients:
## (Intercept)           pH        Depth
##  -51.767143     7.212436    -0.002831
```

What does this look like?

- Is it a line?
- Is it a bird?
- No, it's a **plane**!

For a regression model with 2 independent variables, the regression model is a 2-dimensional plane, rotated and shifted by the three coefficients. We can view this in three dimensions (two for the independent variables, a third for the dependent variable).



You'll notice that there's a lot of change in DO as pH changes, but very little as Depth changes. That's because their coefficients have very different values.

And here's an (example stolen from the web) showing the residuals. The residuals still represent exactly the same idea as when we had a straight line: they are the distance of the dependent variable from its observed value (a black point) to its predicted value (the plane).

Even More!

We can do even more interesting things with multiple regression.

- we can use more than two independent variables
- we can do more than add them together – we could multiply them, e.g.

However, when moving beyond our original (simple) model with one independent variable, it becomes increasingly hard to visualize what's going on. That makes it hard to see when things are correct (or not). So it's important to gain some intuition for what's really going on with multiple regression (or any model we use). We need to check the values going in and out of our models and make sure they make sense.

Pitfalls (and Good Practices)

One common mistake that we can face with multiple regression is called **collinearity** or **correlation** between two (or more) independent variables. That's when two independent variables are either the same, or represent the same information but in different ways. They don't have to be identical. A simple example is measuring the temperature in both Celsius and Fahrenheit, and plugging both in as independent variables.

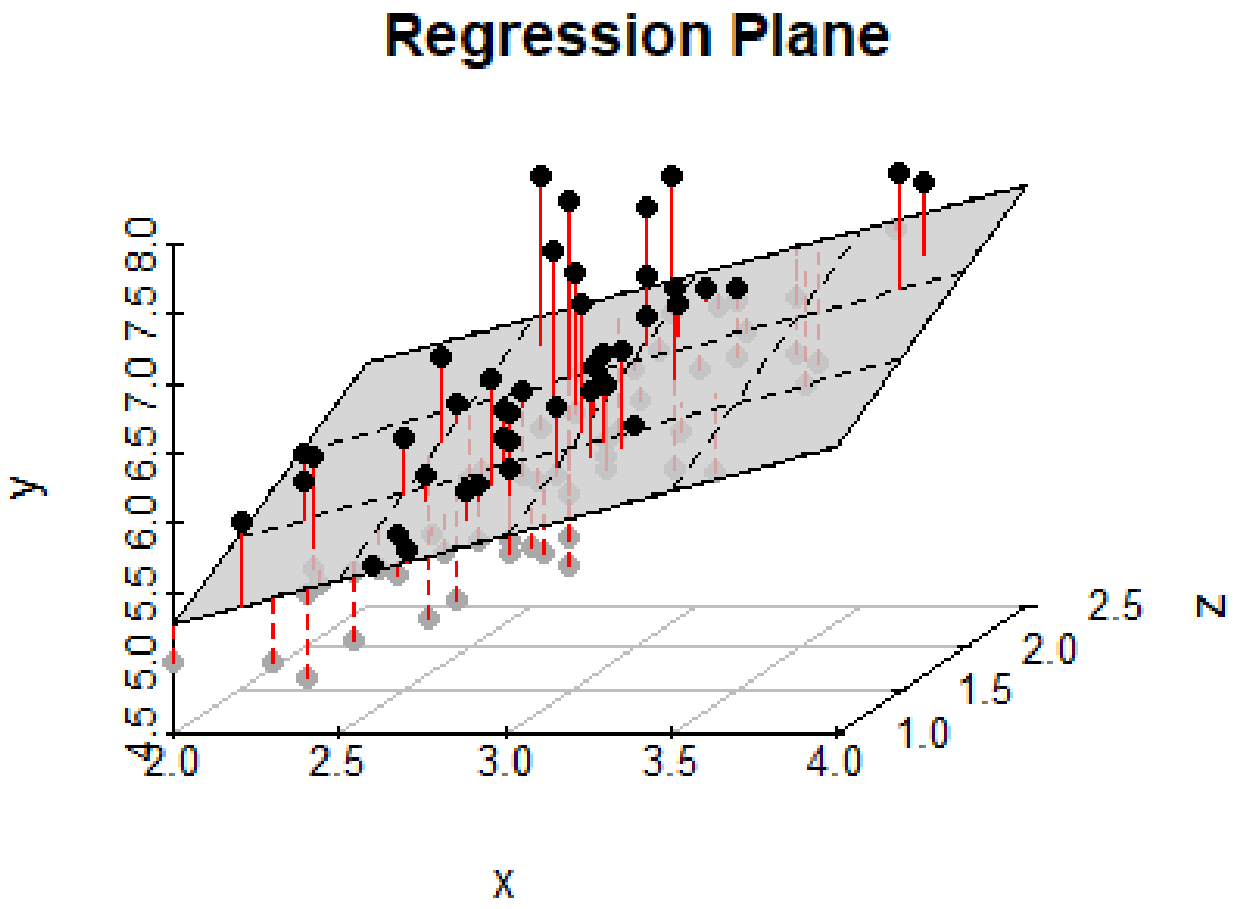


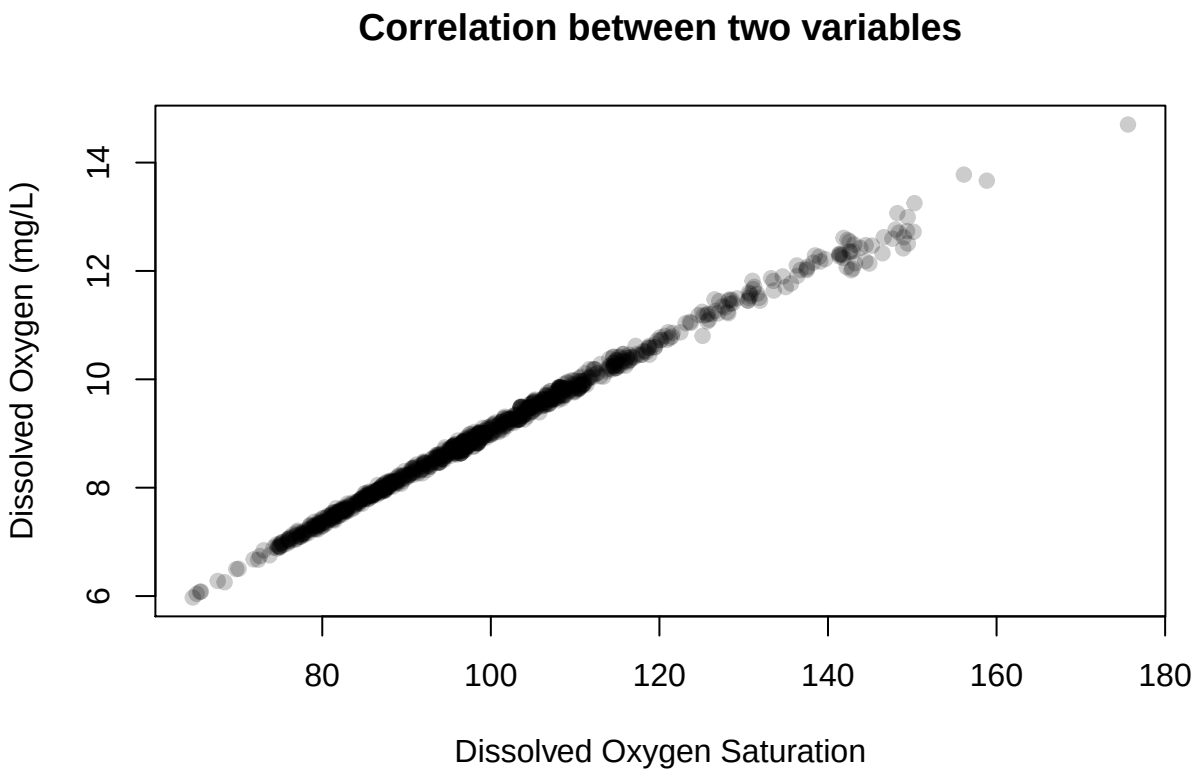
Figure 1: An example of a regression plane with residuals (from https://stackoverflow.com/a/51868640)

Rule of thumb: use independent variables that are **uncorrelated** (aka "orthogonal") with *each other*, but *are* correlated to the dependent variable. That is, independent variables should predict each other as little as possible, but should help predict the dependent variable.

An (Un-)Motivating Example – What Not To Do

The EML data has an example of correlated variables: DO and DOsat. Let's look at them again:

```
plot(DO ~ DOsat, data=eml_small, xlab="Dissolved Oxygen Saturation", ylab="Dissolved Oxygen (mg/L)", mai
```




```
# we can also measure their correlation numerically
cor(eml_small$DO, eml_small$DOsat)
```

```
## [1] 0.9976623
```

This is very close to 1. They are very highly correlated.

Let's turn the tables (for the sake of this example) and show what happens if we were to treat DO and DOsat as independent variables, with pH as the dependent variable. This is only so that we can show the effect of correlation.

```
# use only DO as the independent variable
m_do <- lm(pH ~ DO, data=eml_small)
summary(m_do)
```

```
##
## Call:
## lm(formula = pH ~ DO, data = eml_small)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37494 -0.03239  0.00230  0.03663  0.11494
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.290471   0.011059   659.2   <2e-16 ***
## DO          0.126341   0.001212   104.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05189 on 998 degrees of freedom
## Multiple R-squared:  0.9158, Adjusted R-squared:  0.9158
## F-statistic: 1.086e+04 on 1 and 998 DF,  p-value: < 2.2e-16
```

```
# use both DO and DOsat as the (correlated) independent variables
m_do_dosat <- lm(pH ~ DO + DOsat, data=eml_small)
summary(m_do_dosat)
```

```
##
## Call:
## lm(formula = pH ~ DO + DOsat, data = eml_small)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36066 -0.03366  0.00170  0.03614  0.11811
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.264946   0.019197 378.435   <2e-16 ***
## DO           0.155094   0.017726   8.750   <2e-16 ***
## DOsat       -0.002343   0.001441  -1.626    0.104
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05185 on 997 degrees of freedom
## Multiple R-squared:  0.9161, Adjusted R-squared:  0.9159
## F-statistic:  5441 on 2 and 997 DF,  p-value: < 2.2e-16
```

As you can see, the coefficient for DO went down when we added DOsat: from 0.13 to 0.16. Further, the coefficient for DOSat is slightly negative, so it's offsetting the decrease in the coefficient for DO. Finally, notice that the coefficient for DOsat is **not** statistically significant. Again, this is due to correlation.

Coefficient Variation due to Correlation and Random Samples

If we were to try this correlated-independent variable experiment multiple times with different randomly selected subsets of the data, we would likely get the coefficients to vary quite a bit (again just due to the randomness of the data selected, and the correlation). Here's that experiment with three (intentionally) small subsets of the EML data:

```r
for (i in 1:3) {
    eml_test <- eml[sample(nrow(eml), 20),]
    m_single <- lm(pH ~ DO, data=eml_test)
    m_multiple <- lm(pH ~ DO + DOsat, data=eml_test)
    print(paste("test ", i))
    print(m_single$coefficients)
    print(m_multiple$coefficients)
}
```

```
## [1] "test  1"
## (Intercept)          DO
##   7.3061376   0.1431144
## (Intercept)          DO       DOsat
## 7.306450447 0.019925551 0.009823348
## [1] "test  2"
## (Intercept)          DO
##   7.3980956   0.1281547
## (Intercept)          DO       DOsat
##  7.42763632 -0.06294640  0.01472066
## [1] "test  3"
## (Intercept)          DO
##   7.2807785   0.1391407
## (Intercept)          DO       DOsat
##  7.26088096 -0.02162361  0.01320092
```

If you look carefully, you'll notice that the coefficients for `m_multiple` (with correlation) are changing a lot more than the coefficients for `m_single` (without correlation).