R notes for Lecture 16

The data we will work with consists of measurements of the thickness of mussel beds at various sites along the California coast, as well as measurements of variables believed to contribute or affect the health of mussels.

```
> mussel <- read.table("mussels.short", header=T)
> names(mussel)
[1] "site"      "thickness" "food"      "temp"      "waves"
"human.use"
```

1. Prediction and Confidence Intervals
These work the same as for simple linear regresion. First we fit a model:

```
> full <- lm(thickness~food+temp+waves+human.use)
```

Then we can predict the thickness of a mussel bed at a site for which teh food supply is 2.0, temperature is 13.5, wave activity is 3, and human use is 4. The first step is to create a new data frame that has the predictor values we want to use. Then we ask for the prediction.

```
> new <- data.frame(food=2, temp=13.5, waves=3,
human.use=4)
> predict(full, new, interval="prediction")
          fit       lwr      upr
[1,] 42.24812 -13.33906 97.83531
```

We predict the mussel bed will be 42.4cm thick. But a 95% prediction interval is, well, problematic.

We can use the same command to get a confidence interval for the mean mussel bed thickness of all beds that have those values for the predictors:

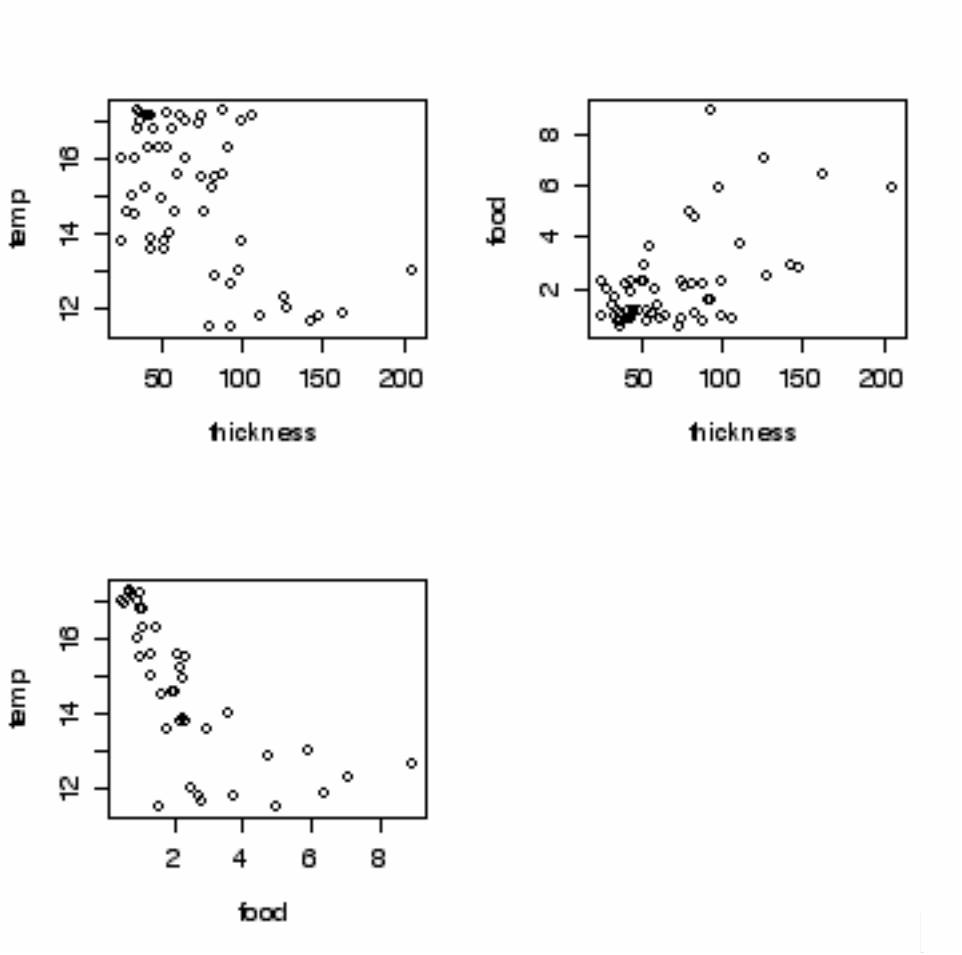
```
> predict(full, new, interval="confidence")
          fit      lwr      upr
[1,] 42.24812 21.87269 62.62355
```

We are 95% confident that the true mean thickness is between 21.8 and 62.6.

2. Added Variable Plots
Added-Variable plots are a graphical tool that helps you assess whether a "new" variable is worth adding to the model. We begin with the idea that we have decided to use temperature to predict thickness. We think this is a good idea because a scatterplot suggests they are in fact related. But then we want to add food to the model. This seems

like a good idea because the scatterplot (upper-right) shows that food and thickness are related. But is it necessary, given that we've already included temp in the model? The final scatterplot shows that temp and food are related, which means perhaps there is no new information to be gained from including food once temp is in the model.

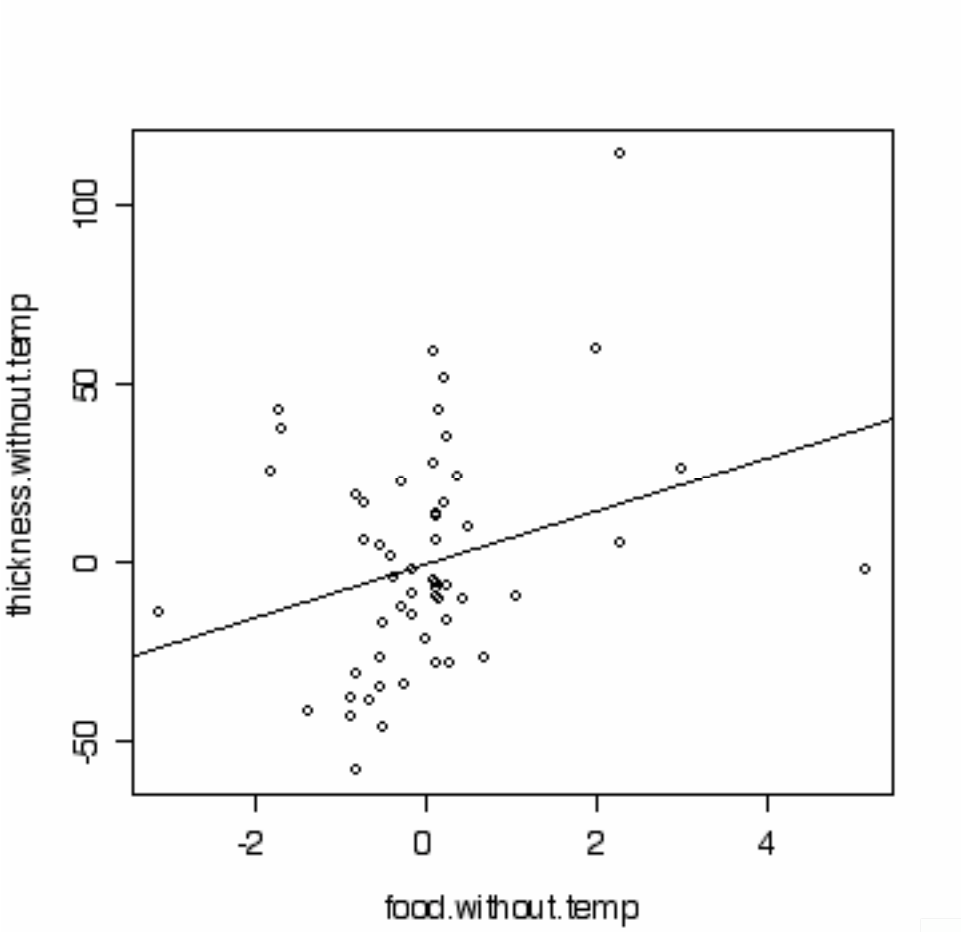
```
> par(mfrow=c(2,2))
> plot(thickness, temp)
> plot(thickness,food)
> plot(food, temp)
```




```
> simple.temp <- lm(thickness~temp)
> thickness.without.temp <- residuals(simple.temp)
> simple.food <- lm(food~temp)
> food.without.temp <- residuals(simple.food)
> par(mfrow=c(1,1))

> plot(food.without.temp, thickness.without.temp)
> simple <- lm(thickness.without.temp ~ food.without.temp)
```

```
> abline(simple)
```



The slope of this line is the same slope we would get for food if we were to put it into the model with temp. The fact that the slope seems vnon-zero suggests that food is indeed useful information even given that we've already used temperature.

```
> summary(simple)

Call:
lm(formula = thickness.without.temp ~ food.without.temp)

Residuals:
    Min      1Q  Median      3Q     Max
-51.817 -22.456  -6.607  14.877  97.112

Coefficients:
                    Estimate Std. Error   t value Pr(>|t|)
(Intercept)       -2.824e-15  4.029e+00 -7.01e-16   1.0000
```

```
food.without.temp  7.469e+00  3.412e+00     2.189   0.0329
*
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `
' 1

Residual standard error: 30.42 on 55 degrees of freedom
Multiple R-Squared: 0.08013,  Adjusted R-squared: 0.0634
F-statistic: 4.791 on 1 and 55 DF,  p-value: 0.03287
```

Note that the slope is 7.469 and the intercept is essentially 0.

Now look at the "full" model.

```
> full <- lm(thickness~temp+food)
> summary(full)

Call:
lm(formula = thickness ~ temp + food)

Residuals:
    Min      1Q  Median      3Q     Max
-51.817 -22.456  -6.607  14.877  97.112

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  135.723     55.163   2.460   0.0171 *
temp          -5.430      3.276  -1.657   0.1032
food           7.469      3.444   2.169   0.0345 *
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `
' 1

Residual standard error: 30.7 on 54 degrees of freedom
Multiple R-Squared: 0.3519,   Adjusted R-squared: 0.3279
F-statistic: 14.66 on 2 and 54 DF,  p-value: 8.207e-06
```

The slope is indeed 7.469