Examples to work through. Stat 302, Week 9, Hours 1 and 2.

1. Polynomial functions, interactions, and perturbations.

2. Dummy variables in combination with numeric variables.

3. Treating a numerical value as if it's categorical.

4. Residual plots and Q-Q plots.

5. Picking the best model with stepAIC

1. Polynomial functions, interactions, and perturbations.

1. Consider the data set eden.potato in the agridat package

```
install.packages("agridat")

library(agridat)

library(car)

eden.potato = subset(eden.potato, !is.na(nitro))
```

We will construct some models of yield as a response to the amount of nitrogen (nitro) and of potash (potash).

Model 1: The first model is a model with just linear terms.

```
mod1 = lm(yield ~ nitro + potash, data=eden.potato)
```

Model 2: The second model is model with an interaction between nitrogen and potash, and squared terms for each main effect variable.

```
mod2 = lm(yield ~ nitro + potash + nitro:potash + I(nitro^2) + I(potash^2), data=eden.potato)
```

Which of these models is ‘better’? How are you measuring this? What are some advantages of each model over the other?

Now let’s make a small change to the data and try again. There are 193 observations. What happens if we remove 20 (and 50) of them at random.

```
set.seed(1234) # Makes sure we all remove the same ones

remove20 = sample(1:nrow(eden.potato), 20)

remove50 = sample(1:nrow(eden.potato), 50)


eden.less20 = eden.potato[ -remove20 , ]

eden.less50 = eden.potato[ -remove50 , ]
```

```
mod1.less20 = lm(yield ~ nitro + potash, data=eden.less20)
```

```
mod2.less20 = lm(yield ~ nitro + potash + nitro:potash + I(nitro^2) +
I(potash^2), data=eden.less20)
```

```
mod1.less50 = lm(yield ~ nitro + potash, data=eden.less50)
```

```
mod2.less50 = lm(yield ~ nitro + potash + nitro:potash + I(nitro^2) +
I(potash^2), data=eden.less50)
```

Now compare the coefficients from each model.

```
mod1
mod1.less20
mod1.less50
```

```
mod2
mod2.less20
mod2.less50
```

Are there major changes between them? Compare the size of any changes to the standard errors in mod1 and mod2? Have a look at the VIFs for each model. Comment.

```
vif(mod1)

vif(mod2)
```

Take home messages.

- Any model that uses a variable in more than one term is going to have variance inflation and be sensitive to perturbations.

- If the VIF isn’t large, you don’t have a problem.

2. Dummy variables in combination with numeric variables.

2. Construct the complex model from question 1, but include block as an independent blocking variable.

```
mod3 = lm(yield ~ nitro + potash + nitro:potash + I(potash^2) +
I(nitro^2) + block, data=eden.potato)
```

Look at the summary and the ANOVA for this model. Comment on the effect, if any from blocks. What do the Df values in the ANOVA indicate that you can also see in the linear model? What does it mean that, for example, that the value for blockB6 is -28.5 ?

Is there any evidence that yield changes by block when holding nitrogen and potash levels constant?

```
summary(mod3)

anova(mod3)
```

Try another model with the same terms as model 3, but also an interaction between potash and block.

Is there any evidence that yield changes by some interaction between block and potash?

What does the value for potash:blockB6 mean?

```
mod4 = lm(yield ~ nitro + potash + nitro:potash + I(potash^2) +
I(nitro^2) + block + block:potash, data=eden.potato)

summary(mod4)

anova(mod4)
```

Take home messages:

- Dummy variables on their own estimate differences between means. Specifically the baseline mean.

- Dummy variables are controlled for just like other variables.

- Interaction terms between dummy variables and numeric/continuous variables show the differences in slopes for different categories.

3. Treating a numerical value as if it’s categorical.

Use a table to see the values that potash takes. Does this seem like a continuous variable to, or is it set at very specific levels?

Interpret the potash levels as categories instead of a continuous scale.

```
eden.potato$potash_cat = as.factor(eden.potato$potash)
```

Model 5: Try a model that fits yield in response to nitrogen and potash (as a category). This is essentially the same as model 1, but looking at potash in categories.

The baseline category is 0 potash because ‘0’ comes first ‘alphabetically’.

Compare the coefficients for nitrogen in both models. Compare both the values as well as the interpretations.

Consider the coefficient for potash as a continuous variable. This value is positive, implying that fields with potash added have higher yields. In the potash-category variable, do the coefficients for the dummy variables confirm this? Do they ALL confirm this? Comment.

```
mod5 = lm(yield ~ nitro + potash_cat, data= eden.potato)
```

Model 6: Try model 5, but add a term for nitrogen and an interaction between nitrogen and potash (as a category). This model is basically the same as model 2, but looking at potash in categories.

Interpret and comment on the values of the potash squared and the nitrogen:potash interactions in model 2.

Compare the model 2 findings to the related ones in model 6. That is, look at values of the dummy variables for potash and for the nitrogen:potash interaction at the highest category of potash (6).

```
mod6 = lm(yield ~ nitro + I(nitro^2) + potash_cat + nitro:potash_cat,
data= eden.potato)
```

Which model better describes the effects of nitrogen and potash on yield?

Which model would be better at describing the effect of 5 units of potash, or 1.75 units of potash?

Take home messages:

- Just because a value can be interpreted as a continuous variable doesn't imply that in needs to be, especially when that variable only takes on a few values in your data.

- A few dummy variables can reveal a non-linear relationship in more detail than a polynomial. However, a model with dummy variables can't be used to predict responses for values that weren't one of the original categories used.

4. Residual plots, Q-Q plots, and the Shapiro-Wilks statistic.

Plot the residuals of model 6 over the values of potash (numeric).

Why are the residuals arranged in vertical lines?

Is there any indication of unequal variance?

```
plot( resid(mod6) ~ eden.potato$potash)

abline(h=0)
```

Confirm your findings by plotting the residuals against potash (category).

```
plot( resid(mod6) ~ eden.potato$potash_cat)
```

Plot the residuals of model 6 over the fitted (predicted) values of model 6.

Is there any indication of unequal variance? Any other patterns?

```
plot( resid(mod6) ~ predict(mod6))

abline(h=0)
```

A quantile-quantile plot, or Q-Q Plot for short, checks a set of values against some assumed distribution.

If the points on a Q-Q Plot fall into a line, then the set of values fits that distribution. If there are bends in the Q-Q plot, or far-off values at the end, then you have evidence against the points fitting the assumed distribution.

The most common distribution a Q-Q Plot is used for is the normal distribution.

A bend in the Q-Q Normal Plot would indicate that the distribution is skewed.

Extreme values at the end indicate that there are outliers.

An s-curve indicates more complicated deviations from normality, such as uneven variance.

With a Q-Q plot, check the residuals of model 6 against the assumption of normality.

```
qqnorm( resid(mod6) )
qqline( resid(mod6) )
```

A Shapiro-Wilks test is a hypothesis test against the null “this distribution is normal”.

A small p-value indicates evidence that a distribution is non-normal. However, like other hypothesis tests, small differences can be detected when there is a lot of data.

In the case of Shapiro-Wilks, this means that small deviations from normality can produce a small p-value if N is very large.

Do a Shapiro-Wilks test on the residuals of model 6. Does the p-value confirm what you saw with the Q-Q plot?

```
shapiro.test( resid(mod6) )
```

5. Picking the best model with stepAIC

Construct a ‘full’ model that includes every variable we considered in the previous models. Use the categorical version of potash. Use the stepAIC function in the car package to (try to) find the model with the best AIC.

```
mod_full = lm( yield ~ nitro + I(nitro^2) + potash_cat +
nitro:potash_cat + block + block:nitro, data=eden.potato)
mod_best = stepAIC(mod_full)
```

The Stepwise method works by comparing the current model to other models with one term removed or added. If a categorical term is removed or added, all the dummy variables are removed or added together. Whichever model has the lowest AIC (or BIC) is selected.

Then the process repeats until the best model is the current one.

The stepAIC function prints out every step to show you this process.

Which variable is removed first?

Which variable is removed second?

Are there any variables removed or added after? Why or why not?