Now available at

`http://dicook.github.io/Statistical_Thinking/tutorials/lab06/lab6help.pdf`

We are converting a few of the variables to factors, which R interprets as dummy variables. Each different number (eg. ST04Q01 = 1 or 2) in the factor counts as its own variable in the regression. To model dummy variables, we always exclude one case and treat it as the base variable and the other dummy variable estimates are a comparision to this base. Why do we exclude one? To avoid multicollinearity, which makes us unable to invert the $X'X$ matrix in OLS estimation.

We're also using glm today instead of lm. A glm is a generalised linear model, and lets you model more types of relationships. The default family for glm is gaussian, which makes glm the same as lm. Other families let you fit different types of models, like logit or poisson regression models. We don't need that now.

We've also standardised the y values, subtracting the mean and dividing by the standard deviation. Now a zero is average, a one is a standard deviation above the mean etc.

Sketch the model - not by hand

```
ggplot(aus_nomiss, aes(x=ST28Q01, y = math_std, colour = interaction(ST04Q01, ST26Q04))) +
  geom_smooth(method = "lm", se = F)
```

Question 1
Use ggplot with the aus nomiss data and consider

```
geom_boxplot(), geom_point() and geom_smooth(method="lm", se = F)
```

Question 2 To get the coefficients, we can use this code. Remember kable in the knitr library can make it look nicer.

```
summary(your-model-here)$coefficients
```

Be careful if you choose to drop a dummy variable, not all of the categories might be insignificant. Even if they're all individually insignificant, they can still be jointly significant.

Question 3 and 4

```
library(broom)
diag_statisitcs <- augment(your-model-here)
```

Some interesting columns augment gives you are

- .fitted, the $\hat{y}$ values
- .resid, $y - \hat{y}$
- .hat, the diagonal of the hat matrix
- .cooksd, Cook's Distance

Now make scatterplots to use these statistics and the actual y values to answer both questions. Check out the lecture notes for cut-offs on what influence and leverage is significant.

For qqplots we need this line inside qnorm to get the quantiles:

```
c(1 - 0.5^(1/n), (2:(n-1) - 0.3175) / (n + 0.365), 0.5^(1/n))
```

Because we standardised y, compare this to a standard normal $\mu = 0, \sigma = 1$.

Question 5

This one is easy.

```
library(car)
vif(your-model-here)
```

Question 7

So we need to estimate another glm with everything but the books variable. Then we can calculate the ANOVA statistic with the deviances of each glm.

```
your-glm$deviance #gives you the deviance
your-glm$null.deviance #we need this too
```

$$\frac{\text{GLM-NO-BOOKS-DEVIANCE} - \text{GLM-EVERYTHING-DEVIANCE}}{\text{GLM-EVERYTHING-NULL-DEVIANCE}} \times 100\%$$

Question 8

Make a dataframe that includes **every** X variable used in the glm with values of the X variables you want to predict. You can put a vector into each X variable to predict more than one score at once.

```
pred.data = data.frame(ST04Q01 = c(number.pred1, number.pred2),
                       ST06Q01 = c(number.pred1, number.pred2),
                       include all X variables)
predict(your-glm, pred.data)
```

Remember how we standardised the y values? After you run predict you need to undo the standardisation to get the actual predicted scores.