1 regression and analysis of variance

Fitting a multiple regression model is done using `lm()`. For instance using the data set `nlschools` with variables `lang` for a language score on a standardized test and predictors `IQ`, `GS` for class size, and `SES` for a socio-economic factor we have the model outputL

```
> library(MASS)
> data(nlschools)
> res = lm(lang ~ IQ + SES + GS, data = nlschools)
> summary(res)


Call:
lm(formula = lang ~ IQ + SES + GS, data = nlschools)

Residuals:
     Min       1Q   Median       3Q      Max
-28.1066  -4.4640   0.4572   4.9278  25.5800

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.96611    1.06739   8.400   <2e-16 ***
IQ           2.40544    0.07430  32.376   <2e-16 ***
SES          0.15015    0.01416  10.604   <2e-16 ***
GS          -0.02539    0.02560  -0.992    0.321
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.971 on 2283 degrees of freedom
Multiple R-Squared: 0.4014,       Adjusted R-squared: 0.4006
F-statistic: 510.2 on 3 and 2283 DF,  p-value: < 2.2e-16
```

Some questions: How do you read this output? Is the language score dependent on all three variables? How would you make predictions?

Let's warm up first by looking a model of `lang` modeled by `IQ`.

```
> res.min = lm(lang ~ IQ, data = nlschools)
> summary(res.min)
```

```
Call:
lm(formula = lang ~ IQ, data = nlschools)

Residuals:
     Min       1Q   Median       3Q      Max
-28.7022  -4.3944   0.6056   5.2595  26.2212

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.52848    0.86682   10.99   <2e-16 ***
IQ           2.65390    0.07215   36.78   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.137 on 2285 degrees of freedom
Multiple R-Squared: 0.3719,       Adjusted R-squared: 0.3716
F-statistic:  1353 on 1 and 2285 DF,  p-value: < 2.2e-16
```

1. Write the equation of the regression line.
2. What is the predicted score of a person with IQ of 10?
3. Make a scatterplot and add the regression line
4. What is r^2?
5. Perform the statistical test

   $$H_0 : \beta_1 = 0, \quad H_A : \beta_1 \neq 0$$

   What is the p-value?

To read the output of a multiple regression model is similar. In the output for `res` find the following:

1. What is the estimated intercept?
2. What is the coefficient in front of `SES`? What is its SE?

3. What is the coefficient in front of `GS`? What is its SE?

4. What does the value of .321 for the last entry for `GS` mean?

We can compare models using a significance test called the *F*-test. It is implemented in `anova()`. We simply use two nested models.

```
> anova(res, res.min)

Analysis of Variance Table

Model 1: lang ~ IQ + SES + GS
Model 2: lang ~ IQ
  Res.Df    RSS   Df Sum of Sq      F    Pr(>F)
1   2283 110938
2   2285 116402   -2     -5464 56.219 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null hypothesis is that the coefficients in the larger model are 0. (That is the extra parameters are not needed). In this case, the small *p*-value indicates that the extra two variables are good for the model.

1. Make the model

   ```
   > res.int = lm(lang ~ IQ + SES, nlschools)
   ```

   Compare this to the full model in `res`. Is the extra variable adding anything? Write the significance test, and your answer.

2 Analysis of variance

The variable `SES` is a actually a factor – categorical, not numeric. How does this change things?

1. Compare the plots produced by

   ```
   > plot(lang ~ SES, nlschools)
   > plot(lang ~ factor(SES), nlschools)
   ```

What is different?

When we have a grouping variable which is a factor, we are really comparing populations for a discrete set of levels. If there were only two we could use a t-test to compare the centers. In the case when there are more, we use an analysis of variance (one way) instead. The `oneway.test()` does all the work of testing

$$H_0 : \mu_1 = \cdots = \mu_k, \quad H_A : \text{Atleast one is not equal}$$

To see it in action we have

```
> oneway.test(lang ~ factor(SES), nlschools)


        One-way analysis of means (not assuming equal variances)

data:  lang and factor(SES)
F = 18.6185, num df = 20.000, denom df = 427.462, p-value < 2.2e-16
```

The small p-value is consistent with the boxplots — the centers appear to depend on the level of `SES`.

1. The variable `COMB` records a 1 if the student was in a combined class. Make a graph based on the levels of `COMB` and then perform a oneway analysis of variance.

3 Misc. problems

1. Load the data set `forbes` and model boiling point (`bp`) by atmospheric pressure (`pres`).

   ```
   > data(forbes)
   > res = lm(pres ~ bp, data = forbes)
   ```

   What is R^2? Find a 95% CI for the slope.

2. Now try to fit the quadratic model for the same data set:

   ```
   > res.q = lm(pres ~ bp + I(bp^2), data = forbes)
   ```

Compare the two models using `anova()`. What is the p-value? What does it say about the extra term?

3. The data set `survey` contains responses of 237 Statistics I students at the University of Adelaide to a number of questions, including the span of the writing hand (Wr.Hnd) and non-writing hand (NW.Hnd), `Pulse`, `Smoke`, `Height` and `Age`.

   Make two regression models:

   ```
   > res.full = lm(Pulse ~ Wr.Hnd + NW.Hnd + Height + Age, data = survey)
   > res.min = lm(Pulse ~ Age, data = survey)
   ```

   (a) From the output of `summary()` on `res.full`, which variables are flagged in the two-sided test of $H_o : \beta_i = 0$?

   (b) Compare the two models using `anova`. What does this say about the presence of extra variables?

4. Again for the `survey` data set, perform a one-way anova significance test to see if the `Smoke` variable has an effect on the students `Pulse` rate. A boxplot of the data can be made as

   ```
   > plot(Pulse ~ Smoke, data = survey)
   ```

5. Load in the data set `anorexia` and plot the difference in pre and post weights by the treatment:

   ```
   > plot(Postwt - Prewt ~ Treat, data = anorexia)
   ```

   Does it appear that the centers of the three distributions are the same?

   Answer this using a one-way analysis of variance test at the $\alpha = 0.05$ level.

6. The data set `michelson` contains measurements on the speed of light performed by michelson. The code speed is contained in `Speed`. Different experimental days are recorded in `Expt`. A plot of the different speeds measured during the separate experiments can be made with

   ```
   > plot(Speed ~ Expt, data = michelson)
   ```

   (a) Based on the boxplots, does it appear that the center (implied population center) of each data set is the same?

(b) Perform a one-way analysis of variance significance test. What is the p-value?