Lab 3

A Quick Introduction to Multiple Linear Regression
Psychology 310

*Instructions.*Work through the lab, saving the output as you go. You will be submitting your assignment as an R Markdown document.

Preamble. Today's assignment involves looking at multiple linear regression as an analysis technique.

1 The Multiple Linear Regression Model

Multiple linear regression is a generalization of simple bivariate linear regression. Instead of having just one predictor, we have two or more. For example, suppose we have 3 predictors, X_1, X_2, and X_3. The prediction equation becomes

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \tag{1}$$

The standard multiple regression model tested by commercial statistical programs does not actually assume anything about the distribution of the X variables. Actually, they are treated as fixed observations, or as "already given." Strictly speaking, this model is not correct for many studies in which the X variables and Y scores are gathered simultaneously, and all are subject to random variation. Fortunately, it often doesn't matter too much in practice.

The fixed regressor model does assume that observations on Y and X follow the rule

$$y_i = \hat{y}_i + \epsilon_i \tag{2}$$

where the ϵ values have a normal distribution that has a mean of zero and a variance that is constant across values of the predictor variables. Under these assumptions, we estimate the regression coefficients using a multivariate equivalent of the least squares regression equations we learned in our module on bivariate linear regression.

Fitting a multiple linear regression model in R is really simple.

2 An Example – Predicting Birth Weight

As an illustration, we shall examine some data on birth weight of babies. Download the *babies.data.txt* file, read it into R, and attach it with the following commands.

```
> babies.data <- read.table("http://www.statpower.net/data/babies.data.txt",
+     header = TRUE)
> attach(babies.data)
```

It is reasonable to expect many if not all of these variables to have an impact on birth weight. Length of the gestation period is obviously a key factor, but physical size is inherited, so we would expect the mom's height and weight to also be predictors. Smoking has been identified as a potential cause of reduced birth weight as well.

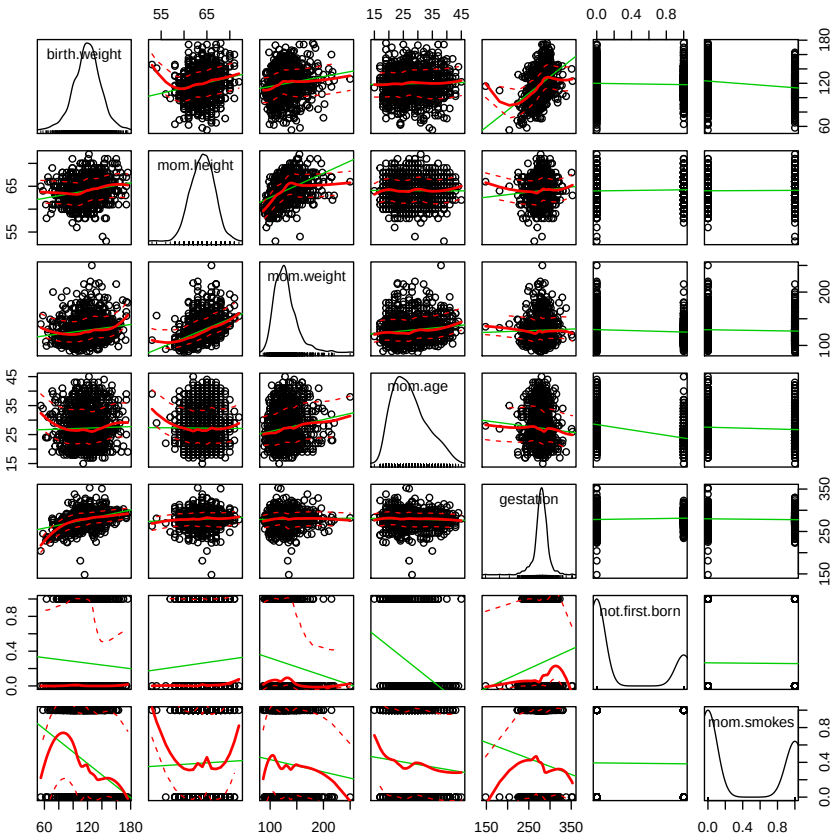
After loading the variables, let's take a quick look at all the correlations. We can look at *all* the intercorrelations by giving the command `cor(babies.data)`, but a more efficient command will display only the correlations between the criterion, `birth.weight`, and the predictors.

```
> cor(babies.data, birth.weight)

                   [,1]
birth.weight    1.00000
gestation       0.40754
not.first.born -0.04391
mom.age         0.02698
mom.height      0.20370
mom.weight      0.15592
mom.smokes     -0.24680
```

As you can see from the correlations, the `gestation` variable is the strongest predictor.

We'll want to spot any predictable nonlinear relationships as well, so let's do a scatterplot matrix. The `scatterplotMatrix` function in the `car` library is especially good, because it includes a linear fit *and* a nonparametric regression line called a *loess* fit in each little scatterplot. (You will need to have the `car` library installed and running to execute this command.) Here are the commands.

```
> library(car)
> scatterplotMatrix(~birth.weight + mom.height + mom.weight + mom.age + gestation +
+     not.first.born + mom.smokes)
```



We begin by fitting two simple models, one with only the intercept term, and one with only one predictor, `gestation`.

```
> fit0 <- lm(birth.weight ~ 1)
> fit1 <- lm(birth.weight ~ gestation)
```

Note that as long as you have one predictor, you do not need the intercept term, which is represented by a 1. The model with no predictors and only an intercept is frequently referred to as the "null" model in discussions of regression.

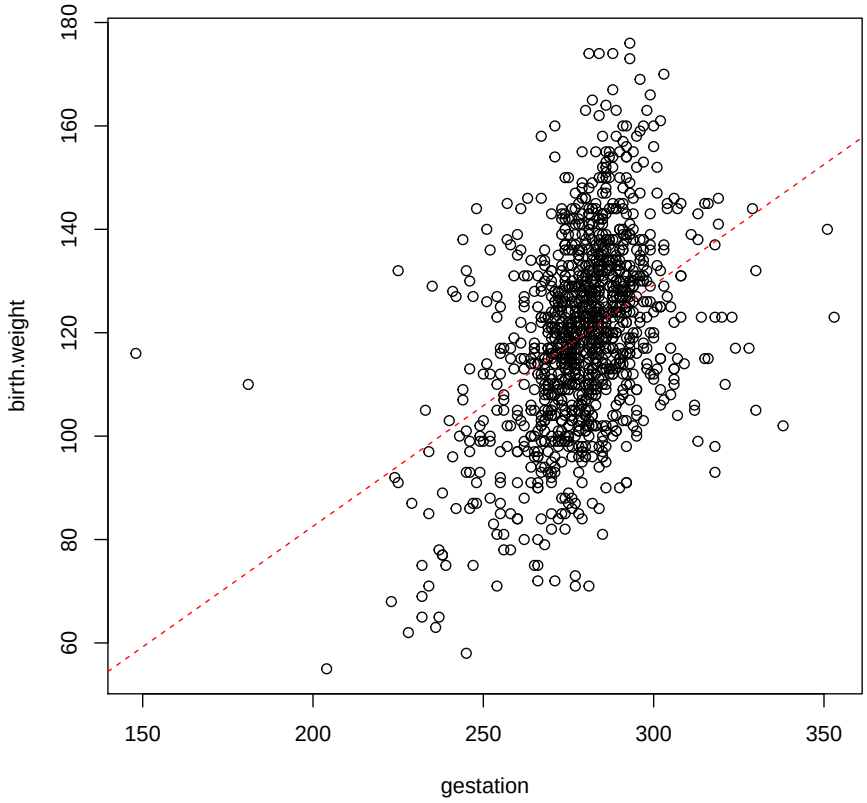
Let's examine the fit of our first non-null model.

```
> summary(fit1)


Call:
lm(formula = birth.weight ~ gestation)

Residuals:
   Min     1Q Median     3Q    Max
```

```
-49.35 -11.07   0.22  10.10  57.70

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.7541     8.5369   -1.26     0.21
gestation     0.4666     0.0305   15.28   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.7 on 1172 degrees of freedom
Multiple R-squared:  0.166,Adjusted R-squared:  0.165
F-statistic:  233 on 1 and 1172 DF,  p-value: <2e-16
```

The squared multiple correlation of 0.166 is highly significant. However, the scatterplot of `birth.weight` versus `gestation` gives us pause, because two observations appear to be outliers. You can see this clearly by loading the `alr3` library and using their `residual.plots` function, or simply by plotting the two variables

```
> plot(gestation, birth.weight)
> abline(fit1, lty = 2, col = "red")
```



Next, use the `identify` function to identify the outliers.

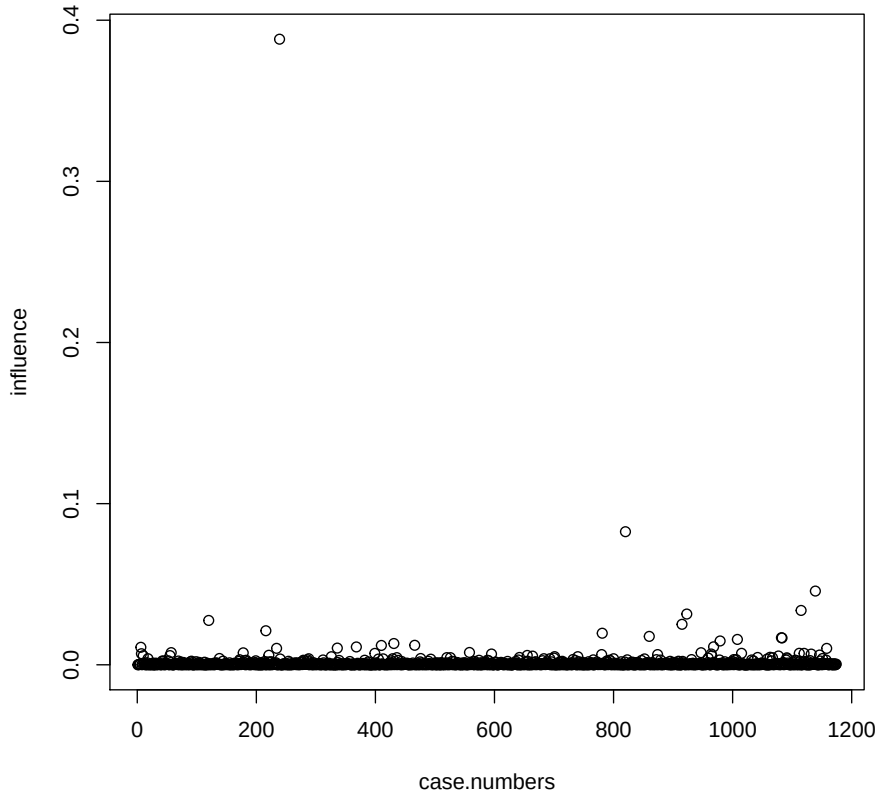
```
> identify(gestation, birth.weight)
```

Click on the points and you'll quickly identify them as points 239 and 820.

Frequently outliers are also *high influence* observations. An observation is high influence if it exerts a strong effect on the values of the estimated β coefficients. This is measured by taking a standardized summary measure comparing $\hat{\boldsymbol{\beta}}$, the vector of estimated beta weights using all the data, with $\hat{\boldsymbol{\beta}}_i$, the vector of estimates based on all the data but the ith observation.

You can verify that point 239 is definitely an *outlier* and a *high influence* observation by loading the `alr3` library, then displaying the Cook's Distance measure against the case number, as follows:

```
> case.numbers <- 1:length(gestation)
> influence <- cooks.distance(fit1)
> plot(case.numbers, influence)
```



Let's remove these two cases and continue. We use a command that tells R to exclude the 239th and 820th rows of the data. Then we `detach` the current file and `attach` the trimmed data.

```
> trimmed.data <- babies.data[c(-239, -820), ]
> detach(babies.data)
> attach(trimmed.data)
> fit0 <- lm(birth.weight ~ 1)
> fit1 <- lm(birth.weight ~ gestation)
> summary(fit1)


Call:
lm(formula = birth.weight ~ gestation)

Residuals:
   Min     1Q Median     3Q    Max
 -49.3  -10.9    0.2   10.1   53.7

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -22.2026     8.8885    -2.5    0.013 *
gestation     0.5073     0.0318    16.0   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.6 on 1170 degrees of freedom
Multiple R-squared:  0.179,Adjusted R-squared:  0.178
F-statistic:  255 on 1 and 1170 DF,  p-value: <2e-16
```

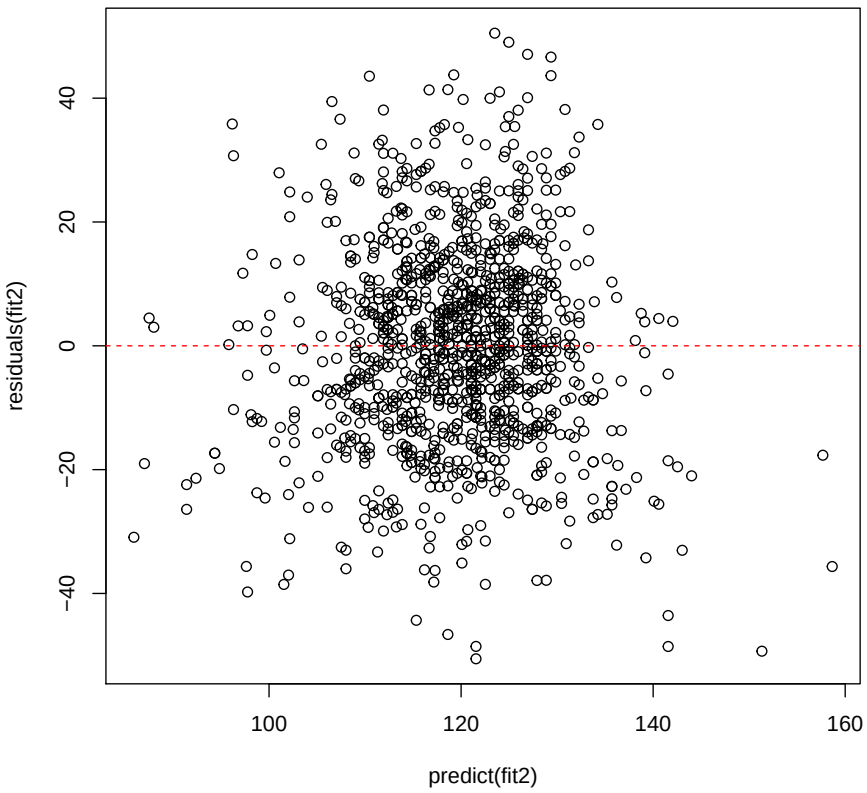
Notice how the squared multiple R has improved a bit, and the slope of the regression line has increased.

Our next step is to add a predictor to the regression equation. Let's add mom's smoking as a predictor, and store the fit.

```
> fit2 <- lm(birth.weight ~ gestation + mom.smokes)
> summary(fit2)


Call:
lm(formula = birth.weight ~ gestation + mom.smokes)

Residuals:
   Min     1Q Median     3Q    Max
-50.55 -10.86  -0.18  10.01  50.49

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -13.6485     8.6930   -1.57     0.12
gestation     0.4881     0.0309   15.77   <2e-16 ***
mom.smokes   -8.1717     0.9692   -8.43   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.2 on 1169 degrees of freedom
Multiple R-squared:  0.226,Adjusted R-squared:  0.225
F-statistic:  171 on 2 and 1169 DF,  p-value: <2e-16
```

We can look for outliers in this multiple regression situation by plotting predicted scores versus residual scores. Adding a horizontal line at zero helps interpretation of the plot.

```
> plot(predict(fit2), residuals(fit2))
> abline(0, 0, lty = 2, col = "red")
```



This regression suggests that kids grow about .45 ounces a day within the range of the data, and that smoking reduces birth weight by about a half a pound. Note that the R^2 value increased, indicating that smoking adds about 4% more variance to what is predicted by gestation period.

We can compare these linear fits by significance with the `anova` command.

```
> anova(fit0, fit1, fit2)

Analysis of Variance Table

Model 1: birth.weight ~ 1
Model 2: birth.weight ~ gestation
Model 3: birth.weight ~ gestation + mom.smokes
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1   1171 393956
2   1170 323499  1     70457 270.1 <2e-16 ***
3   1169 304953  1     18546  71.1 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The two F tests for these models are both highly significant, indicating that each model is significantly better than its predecessor.