

## Lab 3

### A Quick Introduction to Multiple Linear Regression Psychology 310

*Instructions.* Work through the lab, saving the output as you go. If you work in Microsoft Word, you can easily copy any graph to Word via the clipboard. Numerical output may also be copied easily by highlighting, moving it to the clipboard, then copying into Word. However, you should format R output in TrueType Courier New font so that it is *monospaced*. Output from this lab is to be handed in by Monday, October 10. Your output file should be named LAST\_FIRST\_LAB3.DOC, where LAST is your last name, and FIRST is your first name. Any additional files should have the same naming scheme, except the file extension should be correct. You may add any description text you wish after LAB3 in the file name.

*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

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 \quad (1)$$

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

$$y_i = \hat{y}_i + \epsilon_i \quad (2)$$

where the  $\epsilon$  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 from the website, and read it into R.

```
> babies.data <- read.table("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  1.0000000
gestation    0.40754279
not.first.born -0.04390817
mom.age      0.02698291
mom.height   0.20370418
mom.weight   0.15592327
mom.smokes   -0.24679951
```

We'll want to spot any predictable nonlinear relationships as well, so let's do a scatterplot matrix. The `scatterplot.matrix` function 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.) I am not including the plot in this file because it tends to overwhelm Adobe Acrobat's graphics engine. Here are the commands.

```
> scatterplot.matrix(~birth.weight+mom.height
+ + mom.weight + mom.age + gestation +
+ not.first.born + mom.smokes)
```

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

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)

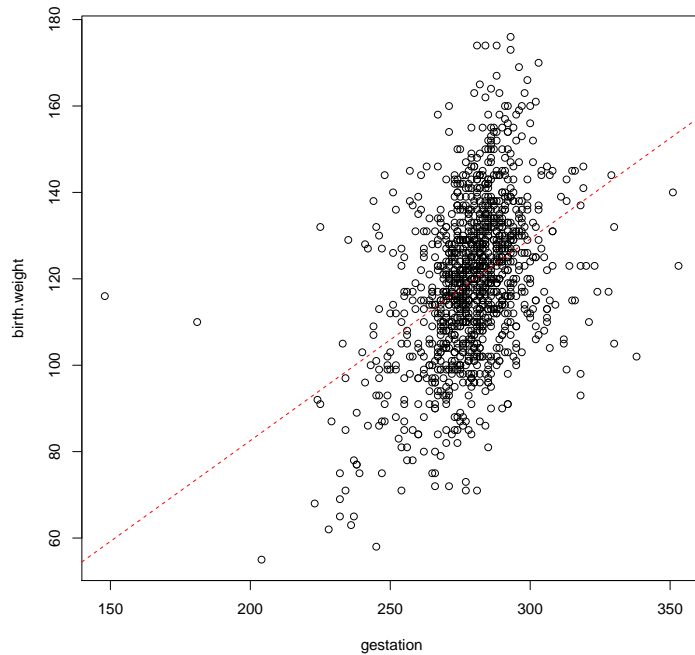
Residuals:
    Min       1Q   Median       3Q      Max
-49.348 -11.065   0.218  10.101  57.704

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.75414    8.53693   -1.26   0.208
gestation     0.46656    0.03054   15.28 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.74 on 1172 degrees of freedom
Multiple R-squared:  0.1661,    Adjusted R-squared:  0.1654
F-statistic: 233.4 on 1 and 1172 DF,  p-value: < 2.2e-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")
```



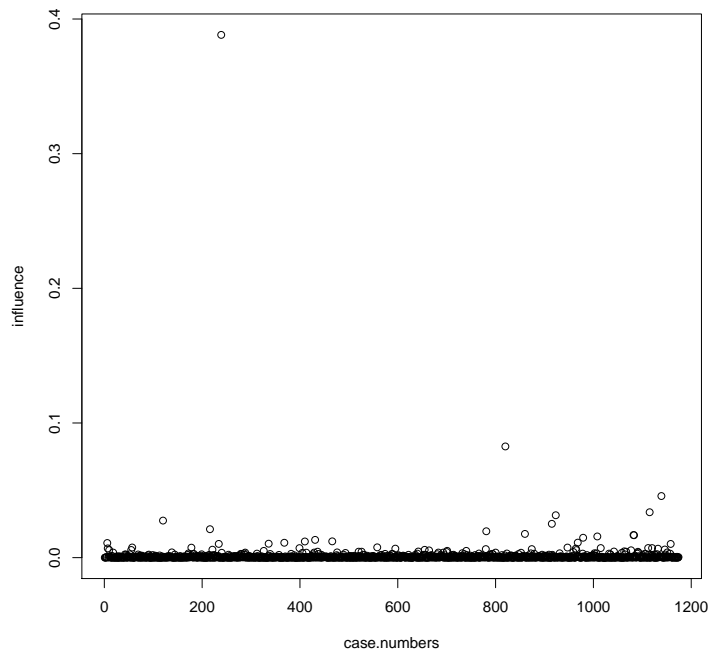
Next, use the `identify` function to identify the outliers.

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

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

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()
> 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
-49.337	-10.921	0.199	10.119	53.663

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-22.20256	8.88848	-2.498	0.0126 *
gestation	0.50726	0.03178	15.963	<2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

```
Residual standard error: 16.63 on 1170 degrees of freedom
Multiple R-squared: 0.1788,      Adjusted R-squared: 0.1781
F-statistic: 254.8 on 1 and 1170 DF,  p-value: < 2.2e-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
-50.553	-10.855	-0.178	10.013	50.495

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-13.64855	8.69303	-1.570	0.117
gestation	0.48809	0.03095	15.771	<2e-16 ***
mom.smokes	-8.17175	0.96916	-8.432	<2e-16 ***

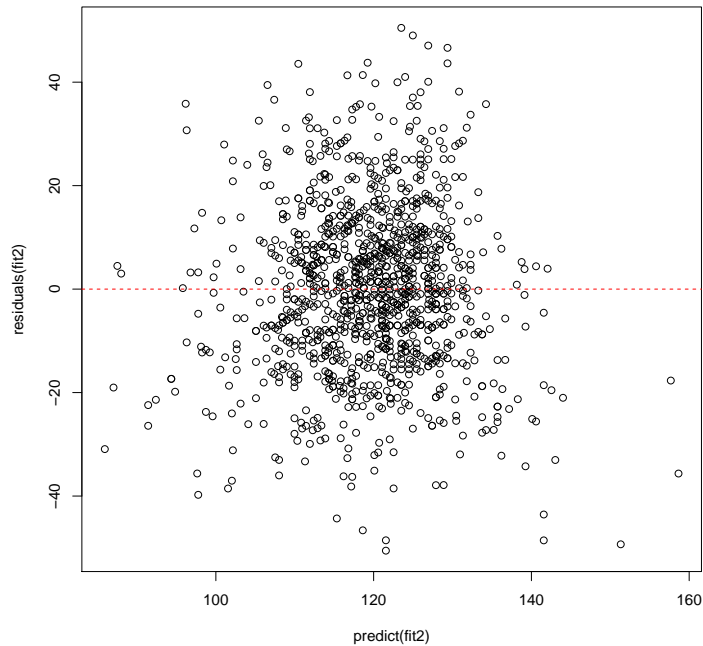
---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

```
Residual standard error: 16.15 on 1169 degrees of freedom
Multiple R-squared: 0.2259,      Adjusted R-squared: 0.2246
F-statistic: 170.6 on 2 and 1169 DF,  p-value: < 2.2e-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	270.088	< 2.2e-16 ***
3	1169	304953	1	18546	71.095	< 2.2e-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.

There is much more to say about these data, and we'll return to them in the weeks ahead.