Chapter 6: Parallel Regression Lines

Continuous and Categorical Predictors

Pregnancy Weight Gain Expectations - NIH FIT for delivery study

This data is from Suzanne Phelan’s NIH study that examines how various factors may influence pregnancy weight gain. Patients are recruited from 6 clinics including one “free clinic,” WIH.

> fit = read.csv('Data/fitData.csv')
> str(fit)
'data.frame': 401 obs. of 19 variables:
$ X                     : int  1 2 3 4 5 6 7 8 9 10 ...
$ IdealGain             : int  15 25 20 10 25 25 25 25 30 30 ...
$ ExpectGain            : int  10 20 18 10 30 15 20 30 35 30 ...
$ AdviceMD              : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 2 2
$ lbsAdviceMd           : int NA NA NA NA NA NA NA 25 30 NA ...
$ prepregBMI            : num 18.3 24 17.4 19.7 21.3 ...
$ age                   : int 21 22 24 20 18 19 22 25 24 21 ...
$ race                  : Factor w/ 2 levels "Other","White": 1 2 1 1 1 1
$ income                : Factor w/ 2 levels "over $25K","under $25K": NA
$ weightChange1stTrimester: num 0.455 1.545 0.545 0.545 3.364 ...
$ GestationWeeks        : int 14 16 14 15 12 12 12 15 ...
$ education             : Factor w/ 2 levels "<= hs","> hs": 1 2 1 1 1 2 2 2
$ parity                : Factor w/ 2 levels "multiparous",...: 2 2 2 2 2 2
$ dietExercise          : int 5 5 6 5 5 6 8 6 5 ...
$ CorrectIdeal          : Factor w/ 2 levels "in or under",...: 1 1 1 1 1 1
$ CorrectExpected       : Factor w/ 2 levels "in or under",...: 1 1 1 1 1 1
$ CorrectMD             : Factor w/ 2 levels "in or under",...: NA NA NA NA NA NA
$ maritalStatus         : Factor w/ 2 levels "married","not married": 2 2
$ Clinic                : Factor w/ 6 levels "Bayside","Beistle",...: 6 6 6

> stripchart(weightChange1stTrimester ~ parity, data = fit,
+     vertical = TRUE, method = 'jitter', pch = 16)
Because there is a lot of data, a boxplot might be more useful here.

```r
> boxplot(weightChange1stTrimester ~ parity, data = fit,
+     vertical = TRUE, ylab = 'First trimester weight gain (lbs.)')
```

Are there weight gain differences, on average, for the parity groups? \(F(1, 399) = 2.147, p = 0.1436\) or \(t(399) = -1.465, p = 0.144\)

```r
> m1 = lm(weightChange1stTrimester ~ parity, data = fit)
> summary(m1)
```

Coefficients:

|                  | Estimate | Std. Error | t value | Pr(>|t|) |
|------------------|----------|------------|---------|----------|
| (Intercept)      | 3.1549   | 0.3255     | 9.693   | <2e-16 *** |
| parityprimiparous| -0.5433  | 0.3708     | -1.465  | 0.144    |

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.122 on 399 degrees of freedom
Multiple R-squared: 0.005352,   Adjusted R-squared: 0.002859
F-statistic: 2.147 on 1 and 399 DF,  p-value: 0.1436

```r
> drop1(m1, test = 'F')
Single term deletions

  Model: weightChange1stTrimester ~ parity
         Df Sum of Sq  RSS    AIC  F value    Pr(F)
<none>                     3888.9 915.04
parity  1       20.925 3909.8 915.19 2.1469 0.1436
```
Is there any relationship between prepregnancy BMI and 1st trimester weight change? Let’s first plot the data. We’ll use a different color for each parity group since soon we’ll want to look at the effect of parity too.

```r
> mycolors = c('indianred', 'cornflowerblue')
> plot(weightChange1stTrimester ~ prepregBMI, data = fit,
+     xlab = 'Pre-Pregnancy BMI',
+     ylab = 'Weight Change (1st Trimester) (lbs.)',
+     col = mycolors[parity], pch = 20)
```

It’s hard to see much of a pattern in the above plot. Let’s consider a formal test to see if the slope of the regression line is different from zero (t-test) or equivalently if having prepregBMI explains a significant proportion of the variability in weight change.

```r
> m1 = lm(weightChange1stTrimester ~ prepregBMI, data = fit)
> summary(m1)
```

```
Call:
  lm(formula = weightChange1stTrimester ~ prepregBMI, data = fit)

Coefficients:  
                          Estimate Std. Error  t value   Pr(>|t|)  
(Intercept)             4.52100    0.73461   6.154  1.84e-09 ***  
prepregBMI             -0.06763    0.02721  -2.486   0.0133 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

Residual standard error: 3.106 on 399 degrees of freedom   
Multiple R-squared: 0.01525,  Adjusted R-squared: 0.01278   
F-statistic: 6.178 on 1 and 399 DF,  p-value: 0.01335
```
> drop1(m1, test = 'F')
Single term deletions

Model:
weightChange1stTrimester ~ prepregBMI
                 Df Sum of Sq    RSS    AIC F value   Pr( F)
<none>                    3850.2 911.03
prepregBMI  1     59.613 3909.8 915.19  6.1778 0.01353 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> Perhaps we could consider a separate (parallel?) line for each group.

> library(car)
> coplot(weightChange1stTrimester ~ prepregBMI | parity, data = fit,
> +       panel = panel.car)

Given: parity

```
  primiparous

  multiparous
```

> coplot(weightChange1stTrimester ~ prepregBMI | parity, data = fit,
> +       panel = panel.smooth)
Given: parity

<table>
<thead>
<tr>
<th>primiparous</th>
<th>multiparous</th>
</tr>
</thead>
</table>

Or maybe add some lines to a single plot.

```r
> plot(weightChange1stTrimester ~ prepregBMI, data = fit,
+     xlab = 'Pre-Pregnancy BMI',
+     ylab = 'Weight Change (1st Trimester) (lbs.)',
+     col = mycolors[parity], pch = 20)
> m1 = lm(weightChange1stTrimester ~ prepregBMI + parity, data = fit)
```
> summary(m1)

Call:
lm(formula = weightChange1stTrimester ~ prepregBMI + parity, 
data = fit)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)     5.3129     0.8394    6.329   6.65e-10 ***
prepregBMI     -0.0766     0.0275   -2.785   0.00561 **
parityprimiparous  -0.7195     0.3731    -1.928   0.05451 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.096 on 398 degrees of freedom 
Multiple R-squared: 0.02436, Adjusted R-squared: 0.01946
F-statistic: 4.969 on 2 and 398 DF,  p-value: 0.007384

> drop1(m1, test = 'F')

Single term deletions

Model: weightChange1stTrimester ~ prepregBMI + parity

Df  Sum of Sq   RSS  AIC  F value    Pr(F)
<none>                      3814.5 909.30
prepregBMI        1   74.331 3888.9 915.04  7.7556 0.00561  **
parityprimiparous 1   35.644 3850.2 911.03  3.7190 0.05451 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

How could abline be used to add lines to the basic plot? What are the equations of the lines?

Here is a fancy way to add the lines so that they don’t extend past the ends of the data. Plus, this
method is more general; it will work even when we have polynomials later.

```r
> XPredData = with(subset(fit, subset = parity == 'multiparous'), 
+    seq(min(prepregBMI), max(prepregBMI), length = 1000))
> Ypred = predict(m1, newdata = data.frame(prepregBMI = XPredData, 
+                  parity = 'multiparous'))
> lines(XPredData, Ypred, col = mycolors[1], lwd = 2)
>   
> XPredData = with(subset(fit, subset = parity == 'primiparous'), 
+    seq(min(prepregBMI), max(prepregBMI), length = 1000))
> Ypred = predict(m1, newdata = data.frame(prepregBMI = XPredData, 
+                  parity = 'primiparous'))
> lines(XPredData, Ypred, col = mycolors[2], lwd = 2)
>   
> legend('topleft', legend = c('multiparous', 'primiparous'), 
+    inset = 0.02, col = mycolors, pch = c(16,16), 
+    title = 'Parity')
```
Leprosy Example from book
Is there a difference in three different treatments in their ability to alter bacterial counts? That is, at the end of the study, is there a treatment that has lower average bacterial count than some others?

```r
> str(leprosy)
'data.frame': 30 obs. of 3 variables:
$ TREATMT : int 1 1 1 1 1 1 1 1 1 1 ... 
$ BACBEF  : num 7.46 8.55 10.77 8.54 16.12 ... 
$ BACAFTER: num 7.79 0 4.74 3.38 15.08 ... 

> leprosy$TREATMT = factor(leprosy$TREATMT)
> leprosy$TREATMT = C(leprosy$TREATMT, contr = treatment)
> library(RColorBrewer)
> mycolors = brewer.pal(3, 'Set1')
> plot(BACAFTER ~ BACBEF, col = mycolors[TREATMT], data = leprosy, pch = 16)
> legend('bottomright', legend = c('1', '2', '3'),
+       inset = 0.02, col = mycolors, pch = c(16,16,16),
+       title = 'Treatment')

> m1 = lm(BACAFTER ~ BACBEF + TREATMT, data = leprosy)
> summary(m1)

Call:
  lm(formula = BACAFTER ~ BACBEF + TREATMT, data = leprosy)

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)          -1.6026     1.8815  -0.852   0.4021
BACBEF               0.8831     0.1488   5.935 2.91e-06 ***
TREATMT2            -0.8641     1.7548  -0.492   0.6265
TREATMT3             3.9060     1.7326   2.254   0.0328 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.824 on 26 degrees of freedom
Multiple R-squared: 0.6383,    Adjusted R-squared: 0.5966
```
\textbf{F-statistic: 15.29 on 3 and 26 DF, p-value: 6.186e-06}

```r
> leprosy$TREATMT = C(leprosy$TREATMT, contr = sum)
> m1 = lm(BACAFTER ~ BACBEF + TREATMT, data = leprosy)
> summary(m1)
```

```
Call:
  lm(formula = BACAFTER ~ BACBEF + TREATMT, data = leprosy)

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             -0.01258    1.80608  -0.007    0.994
BACBEF                   0.88307    0.14879   5.935 2.91e-06 ***
TREATMT1                -1.59004    1.01244  -1.570    0.128
TREATMT2                -0.72589    1.00176  -0.725    0.475
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.824 on 26 degrees of freedom  
Multiple R-squared: 0.6383,    Adjusted R-squared:  0.5966
F-statistic: 15.29 on 3 and 26 DF,  p-value: 6.186e-06
```

```r
> drop1(m1, test = 'F') #Adjusted
```

```
Single term deletions

Model:  
  BACAFTER ~ BACBEF + TREATMT

Df Sum of Sq   RSS   AIC  F value Pr(F)
<none>              380.15  84.181
BACBEF             1   515.01 895.17 107.874 2.912e-06 ***
TREATMT            2   83.35  463.50  86.128  2.8502 0.076 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```