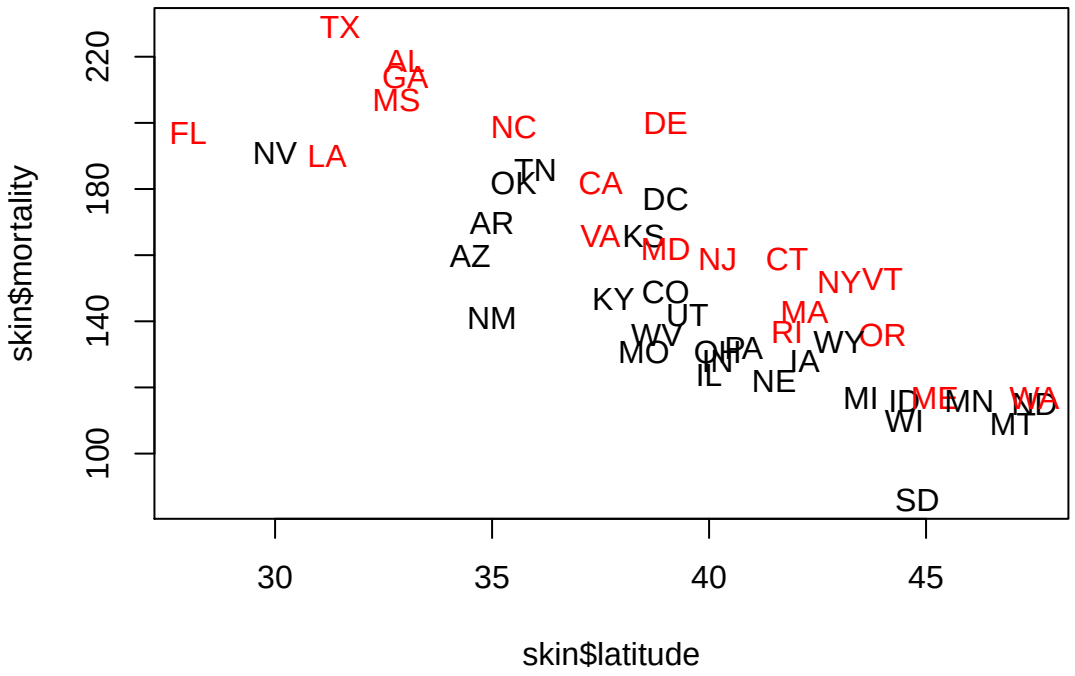
# Analysis of Covariance (ANCOVA)

*Stat 420, Dalpiaz*

*March 11, 2015*

`skin.csv` contains the average annual mortality due to malignant melanoma for white males during 1950 to 1959 per 10 million, for each state and the District of Columbia (Alaska, Hawaii, New Hampshire, and South Carolina are excluded ), the latitude at the centroid of the state, and whether the state borders an ocean. (Fisher and Van Belle (1993). Biostatistics: A methodology for the health sciences.)

```
plot(skin$latitude, skin$mortality, type = "n")
text(skin$latitude, skin$mortality, skin$state, col = skin$ocean+1)
```




Consider the model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon,$$

where $x_1$ is the latitude at the the centroid of the state and $x_2$ is a dummy variable which codes for bordering an ocean, that is,

$$x_2 = \begin{cases} 1 & \text{state borders ocean} \\ 0 & \text{state does not border ocean} \end{cases}$$

Then, for states that do not border an ocean:

$$Y = \beta_0 + \beta_1 x_1 + \epsilon$$

For states that do border an ocean:

$$Y = (\beta_0 + \beta_2) + \beta_1 x_1 + \epsilon$$

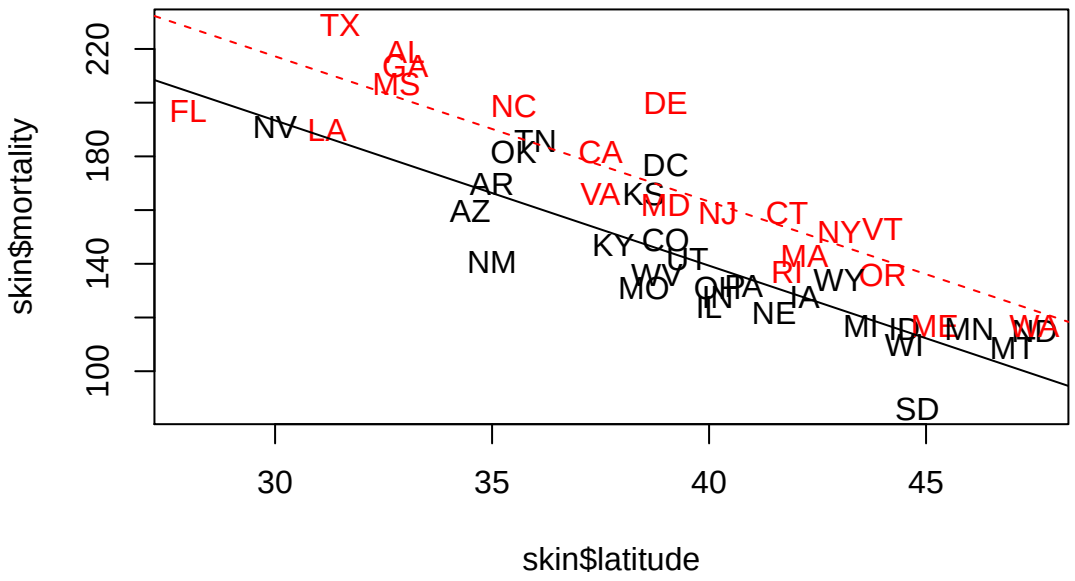
The dummy variable as used here splits the regression into two parallel lines, one for each level (0 and 1) of the qualitative variable $x_2$. The distance between the two parallel lines (measured as the distance between the two y-intercepts) is equal to the estimated coefficient of the dummy variable $x_2$.

```
fit <- lm(mortality ~ latitude + ocean, data = skin)
summary(fit)
```

```
##
## Call:
## lm(formula = mortality ~ latitude + ocean, data = skin)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.065  -9.118  -2.384  10.036  32.290
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 355.6328    18.9405  18.776  < 2e-16 ***
## latitude     -5.4083     0.4668 -11.586 5.90e-15 ***
## ocean        23.8640     4.4813   5.325 3.27e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.94 on 44 degrees of freedom
## Multiple R-squared:  0.8126, Adjusted R-squared:  0.8041
## F-statistic:  95.4 on 2 and 44 DF,  p-value: < 2.2e-16
```

We see that the coefficient for `ocean` is significant, thus using two lines is appropriate.

```
plot(skin$latitude, skin$mortality, type = "n")
text(skin$latitude, skin$mortality, skin$state, col = skin$ocean+1)
abline(fit$coeff[1],fit$coeff[2],col=1,lty=1)
abline(fit$coeff[1]+fit$coeff[3],fit$coeff[2],col=2,lty=2)
```




Now consider the model which allows for two not necessarily parallel lines

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon.$$

For states that do not border an ocean:

$$Y = \beta_0 + \beta_1 x_1 + \epsilon$$

For states that do border an ocean:

$$Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1 + \epsilon$$

We can test for non-parallel lines, or intersection, by testing

$$H_0 : \beta_3 = 0$$

which can be done using an ANOVA F-test. (Or a t-test.)

```
fit2 <- lm(mortality ~ latitude + ocean + latitude:ocean, data = skin)
#fit2 <- lm(mortality ~ latitude*ocean, data = skin)
summary(fit2)
```

```
##
## Call:
## lm(formula = mortality ~ latitude + ocean + latitude:ocean, data = skin)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -32.243  -8.829  -0.994   9.431  32.442
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    350.1125    28.2987  12.372 9.26e-16 ***
## latitude        -5.2706     0.7019  -7.509 2.38e-09 ***
## ocean           33.7399    37.5582   0.898    0.374
## latitude:ocean  -0.2511     0.9481  -0.265    0.792
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.1 on 43 degrees of freedom
## Multiple R-squared:  0.8129, Adjusted R-squared:  0.7999
## F-statistic: 62.28 on 3 and 43 DF,  p-value: 1.078e-15
```

```
anova(fit,fit2)
```

```
## Analysis of Variance Table
##
## Model 1: mortality ~ latitude + ocean
## Model 2: mortality ~ latitude + ocean + latitude:ocean
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     44 9826.5
## 2     43 9810.5  1    16.008 0.0702 0.7924
```

Here we do not reject the null hypothesis, so we prefer the model with parallel lines.

Note the two equivalent ways of specifying the interaction term in `R`. Using `*` automatically inserts the non-interaction terms. This method can be more difficult to use when there are more terms, which we will see in the next example.

A company wishes to study the effects of three different types of promotion on sales of its cookies. The three promotions were:

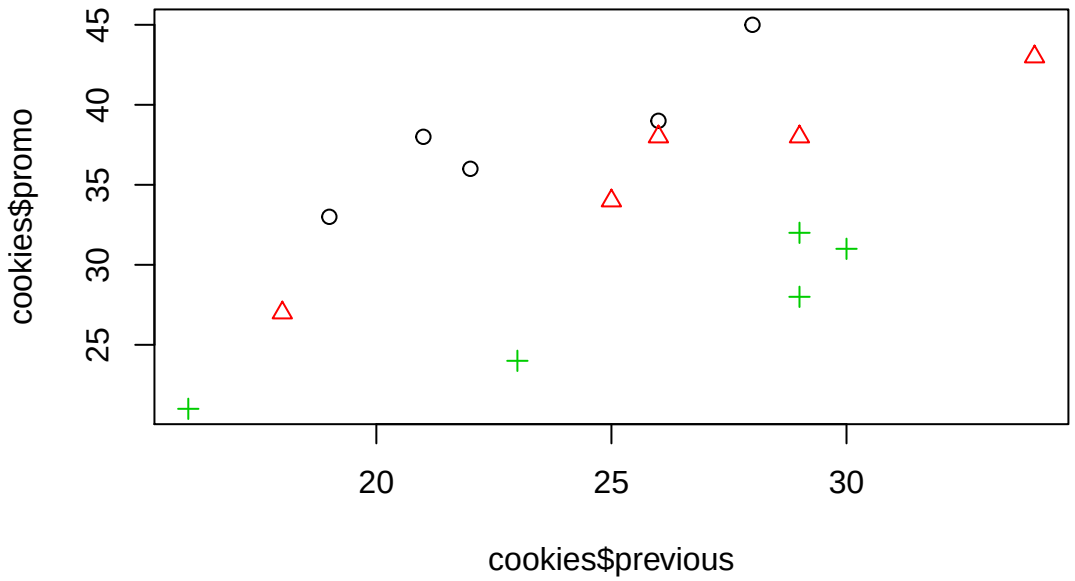
- Treatment 1: Sampling of product by customers in store and regular shelf space
- Treatment 2: Special display shelves at ends of aisle in addition to regular shelf space
- Treatment 3: Additional shelf space in regular location

Fifteen stores were selected as the experimental units. Each store was randomly assigned one of the promotion types, with five stores assigned to each type of promotion. Other relevant conditions under the control of the company, such as price and advertising, were kept the same for all stores in the experiment. Data on the number of cases of the product sold during the promotional period are recorded, denoted by $Y$, as are data on the sales of the product in the preceding period, denoted by $x$. Sales of the preceding period are to be used as the covariate variable. The data can be found in `cookies.csv`

```
cookies
```

```
##    promo previous treat1 treat2 treat3 treatment
## 1     38       21      1      0      0         a
## 2     39       26      1      0      0         a
## 3     36       22      1      0      0         a
## 4     45       28      1      0      0         a
## 5     33       19      1      0      0         a
## 6     43       34      0      1      0         b
## 7     38       26      0      1      0         b
## 8     38       29      0      1      0         b
## 9     27       18      0      1      0         b
## 10    34       25      0      1      0         b
## 11    24       23      0      0      1         c
## 12    32       29      0      0      1         c
## 13    31       30      0      0      1         c
## 14    21       16      0      0      1         c
## 15    28       29      0      0      1         c
```

```
plot(cookies$previous, cookies$promo,
     col = cookies$treat1 + 2*cookies$treat2 + 3*cookies$treat3,
     pch = cookies$treat1 + 2*cookies$treat2 + 3*cookies$treat3)
```




For now we will ignore the `treatment` variable. We have three dummy variables, one for each of the treatments.

$$v_1 = \begin{cases} 1 & \text{treatment 1} \\ 0 & \text{not treatment 1} \end{cases}$$

$$v_2 = \begin{cases} 1 & \text{treatment 2} \\ 0 & \text{not treatment 2} \end{cases}$$

$$v_3 = \begin{cases} 1 & \text{treatment 3} \\ 0 & \text{not treatment 3} \end{cases}$$

We will first test for treatment effects. (In other words, test whether or not the three promotions differ in effectiveness, do we need one regression line, or three parallel lines.)

The full model is,

$$Y = \beta_0 + \beta_1 x + \beta_2 v_2 + \beta_3 v_3 + \epsilon.$$

This gives a different line for each treatment,

- $Y = \beta_0 + \beta_1 x + \epsilon$
- $Y = (\beta_0 + \beta_2) + \beta_1 x + \epsilon$
- $Y = (\beta_0 + \beta_3) + \beta_1 x + \epsilon$

Then we will test,

$$H_0 : \beta_2 = \beta_3 = 0$$

which is only one regression line,

$$Y = \beta_0 + \beta_1 x + \epsilon.$$

```
fit <- lm(promo ~ previous + treat2 + treat3, data = cookies)
summary(fit)
```

```
##
## Call:
## lm(formula = promo ~ previous + treat2 + treat3, data = cookies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4348 -1.2739 -0.3362  1.6710  2.4869
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.3534     2.5230   6.878 2.66e-05 ***
## previous      0.8986     0.1026   8.759 2.73e-06 ***
## treat2       -5.0754     1.2290  -4.130  0.00167 **
## treat3      -12.9768     1.2056 -10.764 3.53e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.873 on 11 degrees of freedom
## Multiple R-squared:  0.9403, Adjusted R-squared:  0.9241
## F-statistic: 57.78 on 3 and 11 DF,  p-value: 5.082e-07
```

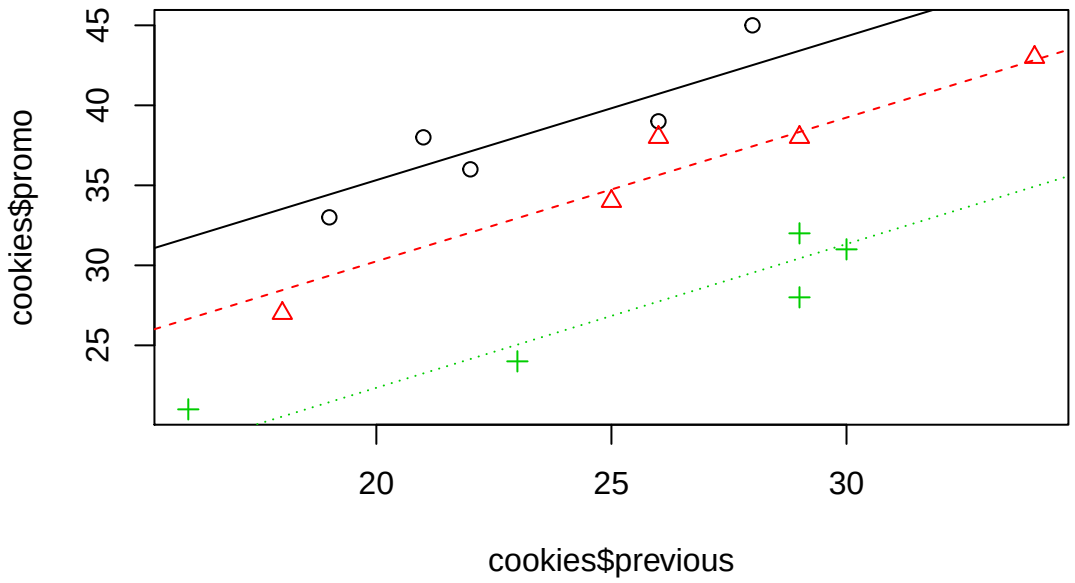
```
fit0 <- lm(promo ~ previous, data = cookies)
anova(fit0,fit)
```

```
## Analysis of Variance Table
##
## Model 1: promo ~ previous
## Model 2: promo ~ previous + treat2 + treat3
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     13 455.72
## 2     11  38.57  2    417.15 59.483 1.264e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here we clearly reject the null hypothesis at any reasonable level.

```
plot(cookies$previous, cookies$promo,
     col = cookies$treat1 + 2*cookies$treat2 + 3*cookies$treat3,
     pch = cookies$treat1 + 2*cookies$treat2 + 3*cookies$treat3)
```

```r
abline(fit$coeff[1],fit$coeff[2],col=1,lty=1)
abline(fit$coeff[1]+fit$coeff[3],fit$coeff[2],col=2,lty=2)
abline(fit$coeff[1]+fit$coeff[4],fit$coeff[2],col=3,lty=3)
```




We could have also fit and tested this model another way.

```r
fit2 <- lm(promo ~ 0 + treat1 + treat2 + treat3 + previous, data = cookies)
summary(fit2)
```

```
##
## Call:
## lm(formula = promo ~ 0 + treat1 + treat2 + treat3 + previous,
##     data = cookies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4348 -1.2739 -0.3362  1.6710  2.4869
##
## Coefficients:
##          Estimate Std. Error t value Pr(>|t|)
## treat1    17.3534     2.5230   6.878 2.66e-05 ***
## treat2    12.2780     2.8348   4.331  0.00119 **
## treat3     4.3766     2.7369   1.599  0.13811
## previous   0.8986     0.1026   8.759 2.73e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.873 on 11 degrees of freedom
```

```
## Multiple R-squared:  0.9978, Adjusted R-squared:  0.997
## F-statistic:  1265 on 4 and 11 DF,  p-value: 1.451e-14
```

```
anova(fit0,fit2)
```

```
## Analysis of Variance Table
##
## Model 1: promo ~ previous
## Model 2: promo ~ 0 + treat1 + treat2 + treat3 + previous
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     13 455.72
## 2     11  38.57  2    417.15 59.483 1.264e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here we remove the common intercept and directly get the intercepts of each regression line.

$$Y = \mu_1 v_1 + \mu_2 v_2 + \mu_3 v_3 + \beta x + \epsilon.$$

Then the individual lines are given by,

- $Y = \mu_1 + \beta x + \epsilon$
- $Y = \mu_2 + \beta x + \epsilon$
- $Y = \mu_3 + \beta x + \epsilon$

and the null hypothesis is

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

which is only one regression line,

$$Y = \mu + \beta x + \epsilon.$$

```r
fit_bad <- lm(promo ~ previous + treat1 + treat2 + treat3, data = cookies)
summary(fit_bad)
```

```
##
## Call:
## lm(formula = promo ~ previous + treat1 + treat2 + treat3, data = cookies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4348 -1.2739 -0.3362  1.6710  2.4869
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.3766     2.7369   1.599    0.138
## previous      0.8986     0.1026   8.759 2.73e-06 ***
## treat1       12.9768     1.2056  10.764 3.53e-07 ***
## treat2        7.9014     1.1887   6.647 3.63e-05 ***
## treat3            NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.873 on 11 degrees of freedom
## Multiple R-squared:  0.9403, Adjusted R-squared:  0.9241
## F-statistic: 57.78 on 3 and 11 DF,  p-value: 5.082e-07
```

Note that $\mathbf{1}$, $v_1$, $v_2$ and $v_3$ are linearly dependent since

$$\mathbf{1} = v_1 + v_2 + v_3$$

so when we attempt to fit the model above, `R` drops one of these in order avoid this issue.

Now that we have seen that the model with parallel regression lines is significant, we will look at adding interaction terms to allow for individual regression lines. We will test this model against the model with parallel lines.

$$Y = \beta_0 + \beta_1 x + \beta_2 v_2 + \beta_3 v_3 + \gamma_2 x v_2 + \gamma_3 x v_3 + \epsilon$$

This gives a different line for each treatment,

- $Y = \beta_0 + \beta_1 x + \epsilon$
- $Y = (\beta_0 + \beta_2) + (\beta_1 + \gamma_2)x + \epsilon$
- $Y = (\beta_0 + \beta_3) + (\beta_1 + \gamma_3)x + \epsilon$

Then we will test,

$$H_0 : \gamma_2 = \gamma_3 = 0$$

which is the parallel regression lines we saw before,

$$Y = \beta_0 + \beta_1 x + \beta_2 v_2 + \beta_3 v_3 + \epsilon.$$

```
fit3 <- lm(promo ~ previous + treat2 + treat3 + previous:treat2
           + previous:treat3, data = cookies)
summary(fit3)
```

```
##
## Call:
## lm(formula = promo ~ previous + treat2 + treat3 + previous:treat2 +
##     previous:treat3, data = cookies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2555 -0.7820 -0.5773  1.1295  2.3965
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     12.88321    5.92451   2.175  0.05768 .
## previous         1.09124    0.25281   4.317  0.00194 **
## treat2          -3.05231    7.32063  -0.417  0.68649
## treat3          -4.31947    7.19742  -0.600  0.56322
## previous:treat2 -0.09999    0.29906  -0.334  0.74579
## previous:treat3 -0.35753    0.29785  -1.200  0.26064
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.871 on 9 degrees of freedom
## Multiple R-squared:  0.9512, Adjusted R-squared:  0.9241
## F-statistic: 35.11 on 5 and 9 DF,  p-value: 1.216e-05
```

```
anova(fit,fit3)
```

```
## Analysis of Variance Table
##
## Model 1: promo ~ previous + treat2 + treat3
## Model 2: promo ~ previous + treat2 + treat3 + previous:treat2 + previous:treat3
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     11 38.571
## 2      9 31.521  2    7.0505 1.0065 0.4032
```

The F-test does not reject the null hypothesis, so we chose the model with parallel lines.

Equivalently,

$$Y = \mu_1 v_1 + \mu_2 v_2 + \mu_3 v_3 + \beta_1 x v_1 + \beta_2 x v_1 + \beta_3 x v_1 + \epsilon.$$

This gives a different line for each treatment,

- $Y = \mu_1 + \beta_1 x + \epsilon$
- $Y = \mu_2 + \beta_2 x + \epsilon$
- $Y = \mu_3 + \beta_3 x + \epsilon$

Then we will test,

$$H_0 : \beta_1 = \beta_2 = \beta_3$$

which is the parallel regression lines we saw before,

$$Y = \mu_1 v_1 + \mu_2 v_2 + \mu_3 v_3 + \beta x + \epsilon.$$

```
fit4 <- lm(promo ~ treat1 + treat2 + treat3 + previous:treat1
           + previous:treat2 + previous:treat3 + 0, data = cookies)
summary(fit4)
```

```
##
## Call:
## lm(formula = promo ~ treat1 + treat2 + treat3 + previous:treat1 +
##     previous:treat2 + previous:treat3 + 0, data = cookies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2555 -0.7820 -0.5773  1.1295  2.3965
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## treat1            12.8832     5.9245   2.175 0.057683 .
## treat2             9.8309     4.3002   2.286 0.048077 *
## treat3             8.5637     4.0869   2.095 0.065606 .
## treat1:previous    1.0912     0.2528   4.317 0.001943 **
## treat2:previous    0.9913     0.1598   6.204 0.000158 ***
## treat3:previous    0.7337     0.1575   4.659 0.001187 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.871 on 9 degrees of freedom
## Multiple R-squared:  0.9982, Adjusted R-squared:  0.997
## F-statistic: 844.7 on 6 and 9 DF,  p-value: 7.407e-12
```

```
anova(fit2,fit4)
```

```
## Analysis of Variance Table
##
## Model 1: promo ~ 0 + treat1 + treat2 + treat3 + previous
## Model 2: promo ~ treat1 + treat2 + treat3 + previous:treat1 + previous:treat2 +
##     previous:treat3 + 0
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     11 38.571
## 2      9 31.521  2    7.0505 1.0065 0.4032
```

Lastly, we note that `R` sometimes can handle dummy variable coding automatically when a variable is of type `factor`.

```
str(cookies)
```

```
## 'data.frame':    15 obs. of  6 variables:
##  $ promo    : int  38 39 36 45 33 43 38 38 27 34 ...
##  $ previous : int  21 26 22 28 19 34 26 29 18 25 ...
##  $ treat1   : int  1 1 1 1 1 0 0 0 0 0 ...
##  $ treat2   : int  0 0 0 0 0 1 1 1 1 1 ...
##  $ treat3   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ treatment: Factor w/ 3 levels "a","b","c": 1 1 1 1 1 2 2 2 2 2 ...
```

```
fit_factor <- lm(promo ~ previous + treatment, data = cookies)
summary(fit_factor)
```

```
##
## Call:
## lm(formula = promo ~ previous + treatment, data = cookies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4348 -1.2739 -0.3362  1.6710  2.4869
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.3534     2.5230   6.878 2.66e-05 ***
## previous      0.8986     0.1026   8.759 2.73e-06 ***
## treatmentb   -5.0754     1.2290  -4.130  0.00167 **
## treatmentc  -12.9768     1.2056 -10.764 3.53e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.873 on 11 degrees of freedom
## Multiple R-squared:  0.9403, Adjusted R-squared:  0.9241
## F-statistic: 57.78 on 3 and 11 DF,  p-value: 5.082e-07
```