B_2: the difference in the average balance between those from the west versus the East

note: There will always one fewer dummy variable than the number of levels.

the level with no dummy variable (East in this example) is known as the baseline.

Below we have the linear regression fit of Balance onto Region in the Credit data set.

```
> lm_fit_credit_2=lm(Balance ~ Region, data=credit)
> summary(lm_fit_credit_2)

Call:
lm(formula = Balance ~ Region, data = credit)

Residuals:
    Min      1Q  Median      3Q     Max
-531.00 -457.08  -63.25  339.25 1480.50

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   531.00      46.32  11.464   <2e-16 ***
RegionSouth   -12.50      56.68  -0.221    0.826
RegionWest    -18.69      65.02  -0.287    0.774
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 460.9 on 397 degrees of freedom
Multiple R-squared:  0.0002188,Adjusted R-squared:  -0.004818
F-statistic: 0.04344 on 2 and 397 DF,  p-value: 0.9575

> contrasts(Region)
      South West
East      0    0
South     1    0
West      0    1
```


The estimated balance for the baseline (East) is $531.00. It is estimated that those in the south will have $12.5 less debt than

those in the East, and that those in the West will have $18.69 less debt than those in the East.

The p-values associated with the coefficient estimates for the two dummy variables are not significant. This suggests that there is no statistical evidence of a real difference in average credit card balance between South and East or between West and East.

Note: Using this dummy variable approach, we can incorporate both quantitative and qualitative predictors.

For example, below we have the linear regression fit of Balance onto the quantitative variable Income and the qualitative variable Region in the Credit data set.

```
> lm_fit_credit_3=lm(Balance ~ Income + Region , data=credit)
> summary(lm_fit_credit_3)

Call:
lm(formula = Balance ~ Income + Region, data = credit)

Residuals:
    Min      1Q  Median      3Q     Max
-806.45 -346.85  -52.38  332.57 1097.57

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 242.4882    49.5674   4.892 1.45e-06 ***
Income        6.0507     0.5813  10.410  < 2e-16 ***
RegionSouth   6.6188    50.3215   0.132    0.895
RegionWest    2.4566    57.7230   0.043    0.966
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 408.9 on 396 degrees of freedom
Multiple R-squared:  0.215,Adjusted R-squared:  0.2091
F-statistic: 36.16 on 3 and 396 DF,  p-value: < 2.2e-16
```

Extensions of the Linear Model

The restrictive assumptions in standard linear regression model: the relationship between the predictors and response are *additive* and *linear*.

Removing the Additive Assumption: Interactions

- In our previous analysis of the Advertising data, we assumed that the effect on sales of increasing one advertising medium is independent of the amount spent on the other media.

For example, the linear model

$$sales = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \varepsilon$$

states that the average effect on sales of a one-unit increase in TV is always β_1, regardless of the amount spent on radio.

- But suppose that spending money on radio advertising actually increases the effectiveness of TV advertising, so that the slope term for TV should increase as radio increases.
- In this situation, given a fixed budget of \$100,000, spending half on radio and half on TV may increase sales more than allocating the entire amount to either TV or to radio.
- In marketing, this is known as a *synergy* effect, and in statistics it is referred to as an *interaction* effect.

- A linear model that uses radio, TV, and an interaction between the two to predict sales takes the form

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times (\text{radio} \times \text{TV}) + \epsilon$$

or

$$\text{sales} = \beta_0 + (\beta_1 + \beta_3 \times \text{radio}) \times \text{TV} + \beta_2 \times \text{radio} + \epsilon$$

(or $\text{sales} = \beta_0 + \beta_1 \times \text{TV} + (\beta_2 + \beta_3 \times \text{TV}) \times \text{radio} + \epsilon$)

we can interpret β_3 as the increase in the effectiveness of TV advertising associated with a one-unit increase in radio advertising (or vice-versa).

```
> summary(lm(sales ~  TV + radio + TV:radio , data = Adv))

Call:
lm(formula = sales ~ TV + radio + TV:radio, data = Adv)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3366 -0.4028  0.1831  0.5948  1.5246

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
TV          1.910e-02  1.504e-03  12.699   <2e-16 ***
radio       2.886e-02  8.905e-03   3.241   0.0014 **
TV:radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9435 on 196 degrees of freedom
Multiple R-squared:  0.9678,Adjusted R-squared:  0.9673
F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16
```

$\hat{\beta}_0$ $\hat{\beta}_1$ $\hat{\beta}_2$ $\hat{\beta}_3$

The p-value for the interaction term (TV×radio) is significant, indicating that there is strong evidence for $H_a: \beta_3 \neq 0$. In other words, it is clear that the true relationship is not additive. R^2 for the model with the interaction term is 96.8%

R^2 for the model without the interaction term is 89.7% (Page 49)

This means that $\frac{96.8 - 89.7}{100 - 89.7} = 69\%$ of the variability in sales that remains after fitting the additive model has been explained by the interaction term.

The coefficient estimates suggest that an increase in TV advertising of \$1000 is associated with increased sales of $(\hat{\beta}_1 + \hat{\beta}_3 \times \text{radio}) \times 1000 = 19 + 1.1 \times \text{radio}$ units. And increase in radio advertising of \$1000 will be associated with an increase in sales of $(\hat{\beta}_2 + \hat{\beta}_3 \times \text{TV}) \times 1000 = 29 + 1.1 \times \text{TV}$ units.

- The hierarchical principle states that if we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.
- The concept of interactions applies just as well to qualitative variables, or to a combination of quantitative and qualitative variables.
- For example, consider the Credit data set and suppose that we wish to predict balance using the income (quantitative) and student (qualitative) variables.
- Without an interaction term, the model takes the form

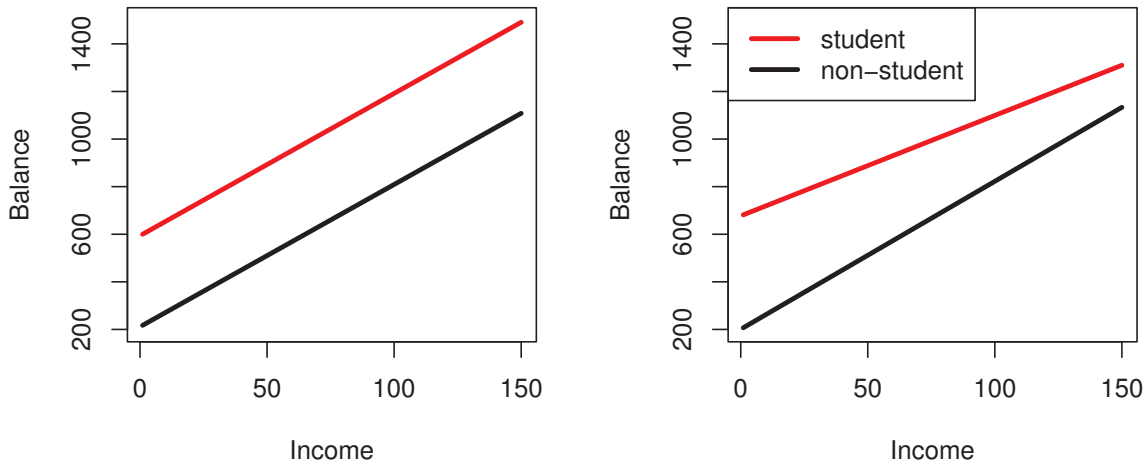
$$\text{balance}_i \approx \beta_0 + \beta_1 \times \text{income}_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases}$$

$$= \beta_1 \times \text{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{if } i\text{th person is a student} \\ \beta_0 & \text{if } i\text{th person is not a student} \end{cases}$$

- With an interaction term, the model becomes

$$\text{balance}_i \approx \beta_0 + \beta_1 \times \text{income}_i + \begin{cases} \beta_2 + \beta_3 \times \text{income}_i & \text{if student} \\ 0 & \text{if not student} \end{cases}$$

$$= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \text{income}_i & \text{if student} \\ \beta_0 + \beta_1 \times \text{income}_i & \text{if not student} \end{cases}$$



For the Credit data, the least squares lines are shown for prediction of balance from income for students and non-students. Left: The model without interaction. There is no interaction between income and student. Right: The model with interaction. There is an interaction term between income and student.