```
Residuals:
    Min      1Q  Median      3Q     Max 
-8.8277 -0.8908  0.2418  1.1893  2.8292 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
TV           0.045765   0.001395  32.809   <2e-16 ***
radio        0.188530   0.008611  21.893   <2e-16 ***
newspaper   -0.001037   0.005871  -0.177     0.86    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,Adjusted R-squared:  0.8956 
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

The p-value associated with F-statistic is statistically significant. This suggests that at least one of the advertising media must be related to sales.

- Sometimes we want to test that a particular subset of q of the coefficients are zero. This corresponds to a null hypothesis

  $$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = ... = \beta_p = 0,$$

  versus

  $$H_a : \ H_0 \text{ is not true}$$

  where for convenience we have put the variables chosen for omission at the end of the list.

- In this case we fit a second model that uses all the variables except those last q. Suppose that the residual sum of squares for that model is RSS_0.

Then the appropriate F-statistic is

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/q}{\mathrm{RSS}/(n-p-1)}$$

For advertising data,

```
> summary(lm_fit_full)

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

Residuals:
    Min      1Q  Median      3Q     Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
TV           0.045765   0.001395  32.809   <2e-16 ***
radio        0.188530   0.008611  21.893   <2e-16 ***
newspaper   -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16

> lm_fit_r <- lm(sales ~ TV + radio , data=Adv)
> summary(lm_fit_r)

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

Residuals:
    Min      1Q  Median      3Q     Max
-8.7977 -0.8752  0.2422  1.1708  2.8328

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.92110    0.29449   9.919   <2e-16 ***
TV           0.04575    0.00139  32.909   <2e-16 ***
radio        0.18799    0.00804  23.382   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.681 on 197 degrees of freedom
Multiple R-squared:  0.8972,Adjusted R-squared:  0.8962
F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16

> anova(lm_fit_r,lm_fit_full)
Analysis of Variance Table

Model 1: sales ~ TV + radio
Model 2: sales ~ TV + radio + newspaper
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    197 556.91
2    196 556.83  1  0.088717 0.0312 0.8599
```

The anova() function performs a hypothesis test comparing the two models. The null hypothesis is that the two models fit the data equally well, and the alternative hypothesis is that the full model is superior. Here the F-statistic is 0.0312 and the associated p-value is 0.8599. This means that we don't have enough evidence to conclude that the model containing the predictors TV, radio, and newspaper is superior to the model that only contains the predictors TV and radio.

Deciding on Important Variables

- It is often the case that the response is only associated with a subset of the predictors.
- The task of determining which predictors are associated with the response, in order to fit a single model involving only those predictors, is referred to as *variable selection.*

- Ideally, we would like to perform variable selection by trying out a lot of different models, each containing a different subset of the predictors.

- Unfortunately, there are a total of 2^p models that contain subsets of p variables. This means that even for moderate p, trying out every possible subset of the predictors is infeasible.

For instance if $p=30$, then we must consider $2^{30} = 1,073,741,824$ models. This is not practical.

- Two automated approaches that search through a subset of models are:

  1. *Forward selection*
     - Begin with the null model—a model that contains an intercept but no predictors.
     - Then fit p simple linear regressions and add to the null model the variable that results in the lowest RSS.
     - Add to that model the variable that results in the lowest RSS amongst all two-variable models.
     - Continue until some stopping rule is satisfied, for example when all remaining variables have a p-value above some threshold.
  2. *Backward selection*
     - Start with all variables in the model.
     - Remove the variable with the largest p-value — that is, the variable that is the least statistically significant.
     - The new $(p-1)$-variable model is fit, and the variable with the largest p-value is removed.
     - Continue until a stopping rule is reached. For instance, we may stop when all remaining variables have a significant p-value defined by some significance threshold.

Note: Backward selection cannot be used if $p > n$, while forward selection can always be used. Forward selection is a greedy approach, and might include variables early that later become redundant.

Model Fit

- Two of the most common numerical measures of model fit are the RSE and R^2, the fraction of variance explained. These quantities are computed and interpreted in the same fashion as for simple linear regression.
- R^2 will always increase when more variables are added to the model, even if those variables are only weakly associated with the response.
- RSE may not decrease as new variables are added to the model.

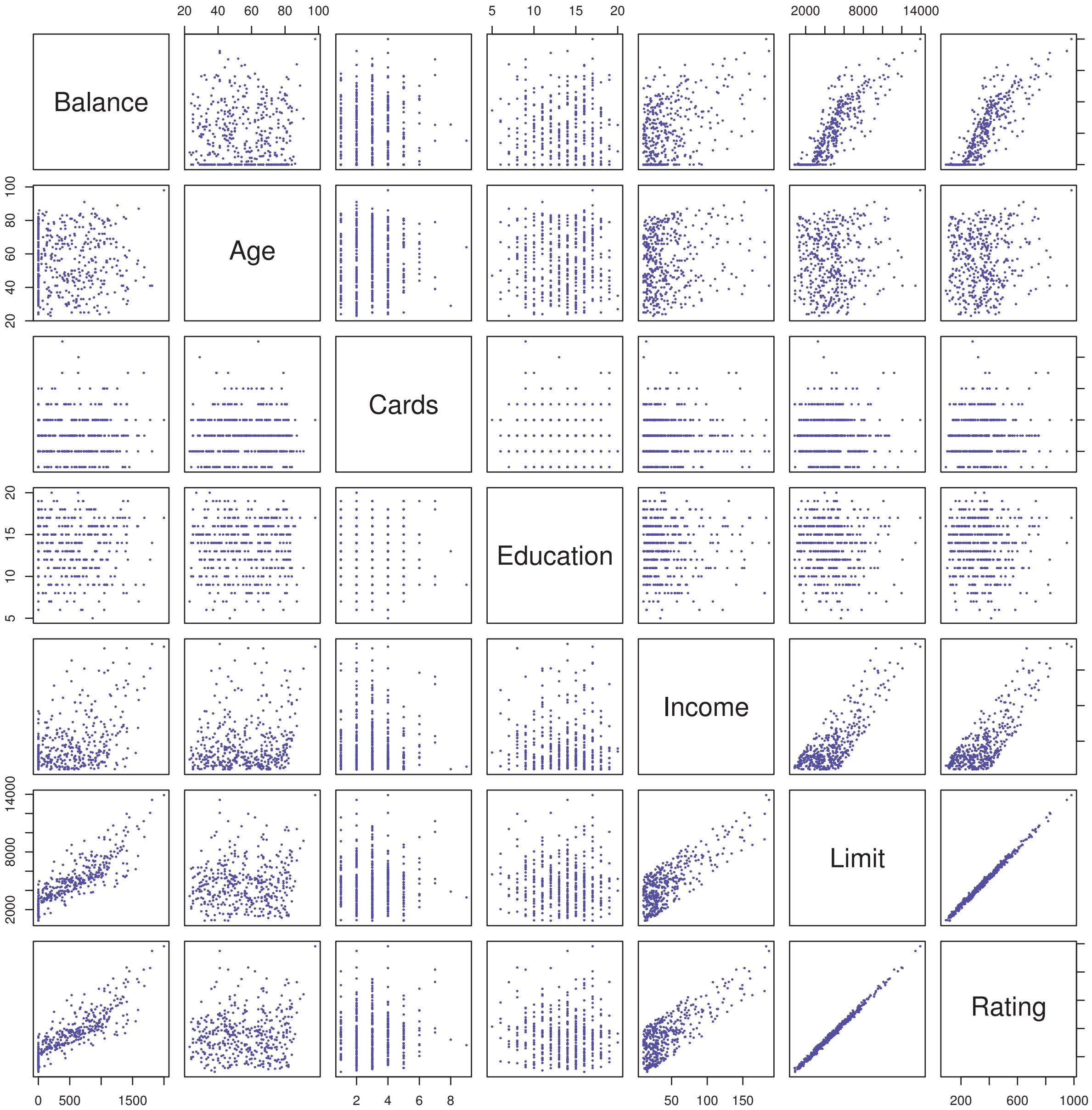
Other Considerations in the Regression Model

In our discussion so far, we have assumed that all variables in our linear regression model are quantitative. But in practice, this is not necessarily the case.

Qualitative Predictors

- Often some predictors are qualitative (They are also called categorical predictors or factor variables).
- For example, the Credit data set displayed below records variables for a number of credit card holders. The response is balance (average credit card debt for each individual).
- Quantitative predictors: age, cards (number of credit cards), education (years of education), income (in thousands of dollars), limit (credit limit), and rating (credit rating).
- Qualitative predictors: own (house ownership), student (student status), status (marital status), and region (East, West or South).

- To incorporate qualitative predictors into a linear regression model, we create dummy variables that take on possible numerical values.



Scatterplot matrix of the Credit data set containing information about balance, age, cards, education, income, limit, and rating for a number of potential customers.

Predictors with Only Two Levels

Example: Investigate differences in credit card balance between those who own a house and those who don't, ignoring the other variables.

We can create a new variable that takes the form

$$x_i = \begin{cases} 1 & \text{if } i\text{th person owns a house} \\ 0 & \text{if } i\text{th person does not own a house,} \end{cases}$$

This results in the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$= \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if ith person owns a house} \\ \beta_0 + \varepsilon_i & \text{if ith person does not} \end{cases}$$

Interpretation:

β_0: the average credit card balance among those who don't own a house

$\beta_0 + \beta_1$: the average credit card balance among those who own their house

β_1: the average difference in credit card balance between owners and non-owners.

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

```
> credit <- read.csv("Credit.csv",header=T)
> names(credit)
 [1] "Income"    "Limit"     "Rating"    "Cards"     "Age"
 [6] "Education" "Own"       "Student"   "Married"   "Region"
[11] "Balance"
```

```
> lm_fit_credit_1=lm(Balance ~ Own, data=credit)
> summary(lm_fit_credit_1)

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

Residuals:
    Min      1Q  Median      3Q     Max 
-529.54 -455.35  -60.17  334.71 1489.20 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   509.80      33.13  15.389   <2e-16 ***
OwnYes         19.73      46.05   0.429    0.669    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 460.2 on 398 degrees of freedom
Multiple R-squared:  0.0004611,Adjusted R-squared:  -0.00205 
F-statistic: 0.1836 on 1 and 398 DF,  p-value: 0.6685
```

The average credit card debt for non-owners is estimated to be $509.80, whereas owners are estimated to carry $19.73 in additional debt for a total of $509.8 + $19.73 = $529.53.

The p-value for the dummy variable is not statistically significant. This indicates that there is no statistical evidence of a difference in average credit card balance based on house ownership.

Given a qualitative variable such as Own, R generates dummy variables automatically. The contrasts() function returns the coding that R uses for the dummy variables.

```
> attach(credit)
> contrasts(Own)
    Yes
No    0
Yes   1
```

R has created an OwnYes dummy variable that takes on a value of 1 if the person owns a house, and 0 otherwise.

Qualitative Predictors with More than Two Levels

In this situation, we can create additional dummy variables.

For example, for the region variable we create two dummy variables. The first could be

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is from the South} \\ 0 & \text{if } i\text{th person is not from the South,} \end{cases}$$

and the second could be

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is from the West} \\ 0 & \text{if } i\text{th person is not from the West.} \end{cases}$$

Then both of these variables can be used in the regression equation, in order to obtain the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

$$= \begin{cases} \beta_0+\beta_1+\varepsilon_i & \text{if ith person is from the South} \\ \beta_0+\beta_2+\varepsilon_i & \text{if ith person is from the West} \\ \beta_0+\varepsilon_i & \text{if ith person is from the East} \end{cases}$$

Interpretation:

β_0: the average credit card balance for individuals from the East

β_1: the difference in the average balance between people from the South versus the East

β_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
```