Three

Linear Regression

part(b): Considerations in the Regression Model

3.9 Qualitative Predictors

- The linear regression model can accommodate qualitative predictors.
- `Credit` data set records `balance`, the average credit card debt, along with various predictors.
  - This includes several quantitative predictors, like `age`, `income`, etc.,
  - but also qualitative predictors like `gender` and `ethnicity`.
- To make it work, we need *indicator variables*, also known as *dummy variables*.

Predictors with Only Two Levels

- Suppose we wish to investigate differences in credit card `balance` between males and females, ignoring the other variables for the moment.

- Here `gender` is a qualitative predictors with two *levels*.

- We can create a dummy variable that takes on two numerical values:

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male.} \end{cases}$$

- Now this is a numerical variable and can be inserted into our regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \varepsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

  - β_0 in this case is the average `balance` among males,

- while $\beta_0 + \beta_1$ is the average `balance` among females.
- So, β_1 here represents the average `balance` difference between females and males.
- Using this *coding*, R gives us the following output:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    509.80      33.13  15.389   <2e-16 ***
GenderFemale    19.73      46.05   0.429    0.669
```

- Males are estimated to owe \$509.80, while females are estimated to owe \$19.73 more, totaling \$529.53.
- However, note that the p-value is high here, indicating no statistical evidence for any difference in `balance` between the genders.

- The choice of coding females as 1 and males as 0 is arbitrary here.
  - The regression *fit* does not change with different coding.
  - For example, if we code males as 1 and females as 0 instead,
    - β_0 will then be the average `balance` for females,
    - and $\beta_0 + \beta_1$ for males.
    - The estimates for β_0 and β_1 will change accordingly: in this case 529.53 and -19.73.
    - This leads to the same estimates of \$529.53 for females and $\$529.53 - \$19.73 = \$509.80$ for males.

- We can also use non 0/1 coding schemes.
  - For example, say we define a dummy variable

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ -1 & \text{if } i\text{th person is male.} \end{cases}$$

  - This results in the model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is female} \\ \beta_0 - \beta_1 + \varepsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

  - Now β_0 becomes the average `balance` for the "average" `gender`, in the sense that β_1 is the amount females are above and also the amount males are below.
  - The resulting estimates from linear regression are 519.665 and 9.865, again resulting in the same *fit* for females and males.

- In general, coding does not change the regression *fit*.
  - This is true even when you include other variables in multiple linear regression.
  - The predictions will be identical regardless of the coding scheme used.
  - The difference is just in the way the coefficients (of the coded variables) are interpreted.

Predictors with More than Two Levels

- When a qualitative predictor has more than two levels, we need to use an additional dummy variable for each extra level.
- For example, in `Credit` data,
  - `ethnicity` has three levels (Asian, Caucasian, African American),
  - so we create two dummy variables:

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian.} \end{cases}$$

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

- This results in the model:

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \\
&= \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \varepsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \varepsilon_i & \text{if } i\text{th person is African American.} \end{cases}
\end{aligned}
$$

  - Here β_0 is the average `balance` for African Americans,
  - β_1 is the average `balance` difference between Asians and African Americans,
  - and β_2 is the average `balance` difference between Caucasians and African Americans.

- The resulting R output is

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           531.00      46.32  11.464   <2e-16 ***
EthnicityAsian        -18.69      65.02  -0.287    0.774
EthnicityCaucasian    -12.50      56.68  -0.221    0.826
---
...
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
```

  - It is estimated that African Americans average a `balance` of \$531.00 while Asians and Caucasians average slightly less.
  - The p-values associated with the dummy variables are very large, indicating no real difference between African Americans and each of the other two ethnicities.

- Again, the coding choice does not affect the fit or predictions.
  - If we use the resulting regression coefficient estimates, we will always end up estimating \$531.00 for African Americans, etc.
  - However, the individual coefficients and their p-values do depend on the coding.
  - To test for significance of `ethnicity`, we can use the F-test to test $H_0 : \beta_1 = \beta_2 = 0$. This does *not* depend on the coding.
  - Here we see that the F-test p-value is 0.96, indicating no relationship between `balance` and `ethnicity`.

3.10 Extensions of the Linear Model

- The standard linear regression model provides interpretable results and works quite well on many real-world problems.
- However, two important assumptions that are often violated in practice are that the relationship between the predictors and response are *additive* and *linear*.
  - The additive assumption means that the effect of changes in a particular predictor X_j on the response Y is independent of the values of the other predictors.
  - The linear assumption means that the change in the response Y due to a one-unit change in X_j is constant, regardless of the value of X_j.
- There are many ways to relax these two assumptions. Here, we look at some classical approaches that are easy to implement.

Removing the Additive Assumption

- In `Advertising`, we fitted the standard linear regression model and concluded that `TV` and `radio` are associated with `sales`.
  - The additive assumption in our model says that the average effect on `sales` of a one-unit increase in `TV` is always the same regardless of the value of `radio` (and vice versa).
  - Later, we noticed a violation of this additive assumption from our surface plot.
  - There seems to be a *synergy* effect: the observed `sales` seem to be higher than the linear model suggests when there's spending in both `TV` and `radio`.
  - How to relax the additive assumption to include this effect?

- Consider the general two variable standard linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon.$$

  - Increasing X_1 by one unit increases Y by β_1 units regardless of the value of X_2.

- We can extend this model by including a third predictor, called an *interaction term*, which is the product of X_1 and X_2. We get

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon.$$

- To see how this relaxes the additive assumption, we can rearrange and rewrite this model as

$$\begin{aligned} Y &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \varepsilon \\ &= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \varepsilon, \end{aligned}$$

  where $\tilde{\beta}_1 = \beta_1 + \beta_3 X_2$.

  - Note that $\tilde{\beta}_1$ changes with X_2.
  - This means the effect of X_1 on Y is no longer constant: the value of X_2 determines the effect of a one-unit change of X_1 on Y.

- Going back to `Advertising`, by adding an interaction term between `radio` and `TV`, we get

$$\begin{aligned}\texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{radio} \times \texttt{TV}) + \varepsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \texttt{radio}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \varepsilon.\end{aligned}$$

  - We can interpret β_3 as the increase in the effectiveness of `TV` for a one unit increase in `radio` (or vice versa).

- The resulting R output is

```
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
```

- This is strong evidence that the model with the interaction term is superior than the original one without.
  - The p-value for the interaction term is extremely low:
    - indicates strong evidence to reject $H_0 : \beta_3 = 0$.
    - i.e. reject that the true relationship is additive.
  - The R^2 goes up to 96.8% from 89.7%.
    - Nearly 70% of the remaining variance was explained by the interaction term!

- In this example, all the individual p-values are small, so it is natural to include all the terms in the model.
  - However, sometimes an interaction term has a small p-value, but the associated main effects do not.
  - In such cases, it is common practice to follow the *hierarchical principle*:
    - if we include an interaction in a model, we should also include the main effects regardless of their p-values.
  - One rationale is that if we have strong evidence to include the interaction, then we believe the predictor is relevant to our response.
  - In that case, we should require strong evidence to not include the main effect, rather than requiring strong evidence to include the main effect.

The concept of interactions applies to qualitative variables as well.

- Consider the `Credit` data set. Suppose we want to model `balance` on `income` (quantitative) and `student` (qualitative).

- The standard model takes the form

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

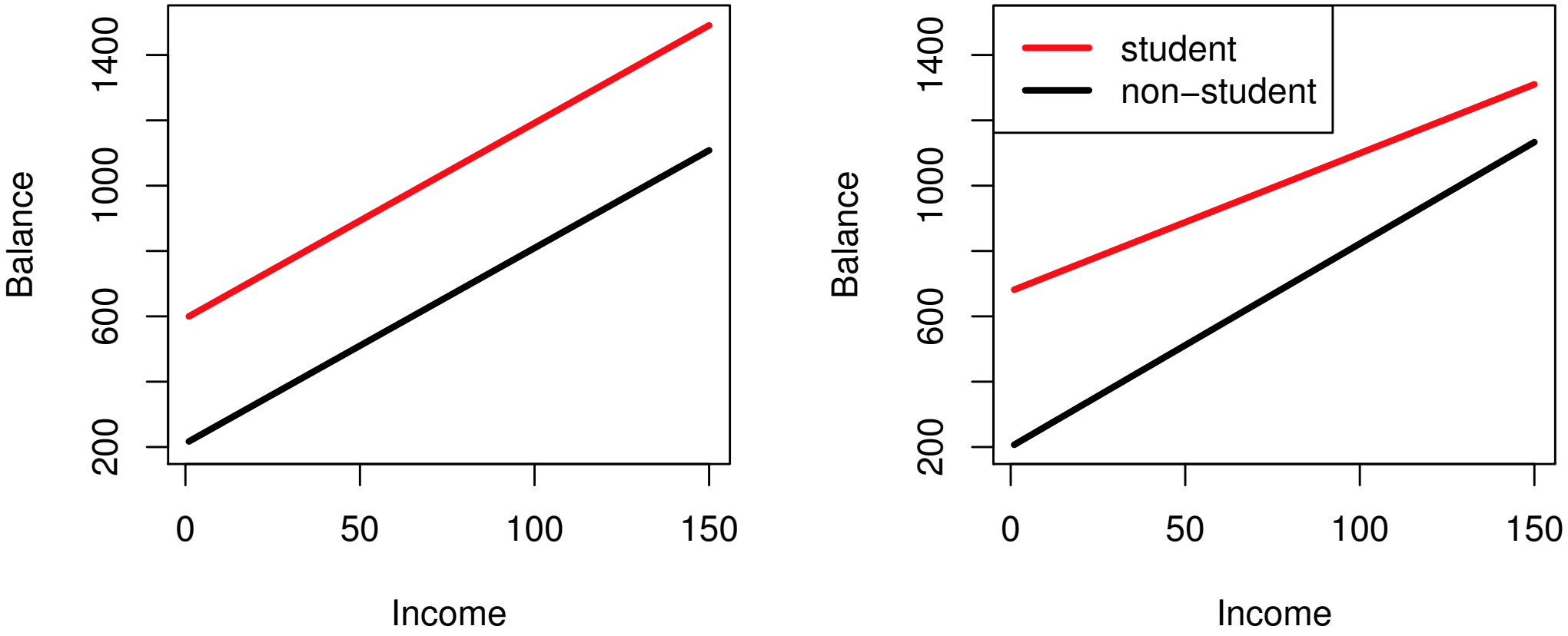
  - This is fitting two parallel lines to the data: one for students, one for non-students.
  - The two lines have the same slope β_1, but different intercepts.
  - The additive assumption holds as the effect of a one-unit increase in `income` is the same for students and non-students.

- If we add an interaction variable, by multiplying `income` with `student`, we get

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

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

  - This is again fittng two lines, but now we have different slopes as well.
  - This allows changes in `income` to affect `balance` differently depending if `student`.



– Least squares fit of `balance` on `income` without (left) and with (right) interaction with `student`.