Lecture Notes 3

Zach Fechko

03 October, 2022

Simple Linear Regression with a qualitative regressor

Summary

Simple linear regression with a qualitative regressor is used to model the **mean** of a quantitative random variable as a linear function of another *qualitative* random variable

True model

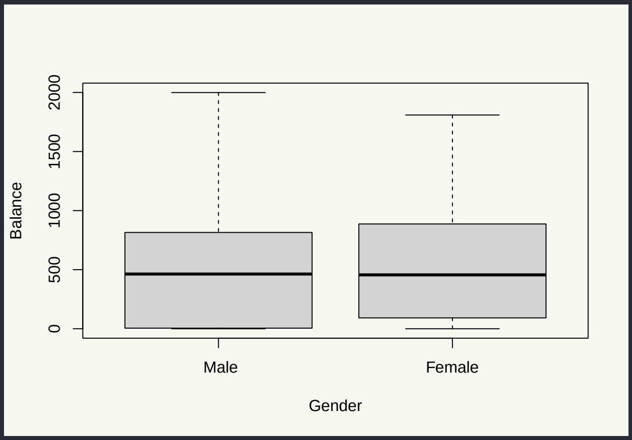
- For two quantitative random variables X and Y, a simple linear model is
  $E(Y) = \beta_0 + \beta_1 X$
  where β_0 and β_1 are unknown, true model parameters (or coefficients), and β_1 is called the regression coefficient.
- The above model is equivalent to
  $Y = \beta_0 + \beta_1 X + \epsilon$ with $E(\epsilon) = 0$
  which is called the population regression line

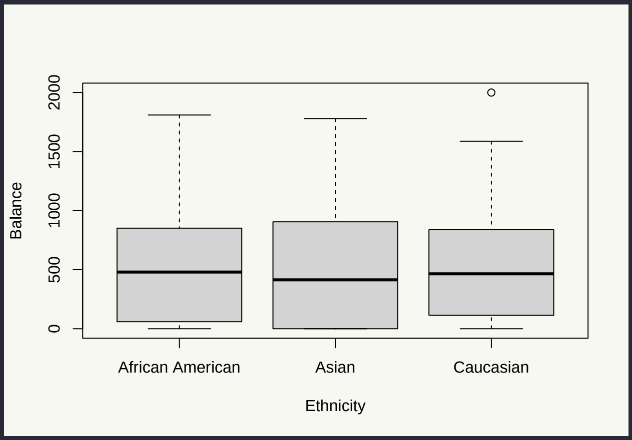
Linear regression with a qualitative regressor

Motivation

- How is the `Balance` of a credit card related to its holder's `Gender`?
- How is the `Balance` of a credit card related to its holder's `Ethnicity`?

```r
par(bg = '#f8f8f2')
boxplot(Balance ~ Gender, data = Credit)
boxplot(Balance ~ Ethnicity, data = Credit)
```





Model 1: 2 levels

- Coding: `Gender` has 2 levels: `Male`, and `Female`
- *dummy variable*: $x_i = 0$ if ith person is `Female`, and $x_i = 1$ if ith person is `Male`
- Model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, which induces 2 submodels:
  - $y_i = \beta_0 + \beta_1 + \epsilon_i$ if ith person is `Male`
  - $y_i = \beta_0 + \epsilon_i$ if ith person is `Female`

Fitting the Model 1

```
lm(formula = Balance ~ Gender, data = Credit)
```

```
##
## Call:
## lm(formula = Balance ~ Gender, data = Credit)
##
## Coefficients:
##  (Intercept)  GenderFemale
##       509.80         19.73
```

- `Male` card holders have a `Balance` of \$509.80
- `Female` card holders have a `Balance` of $(\$509.80 + \$19.73) = \$529.53$

Testing the Model 1

```
lm(formula = Balance ~ Gender, data = Credit)
```

```
##
## Call:
## lm(formula = Balance ~ Gender, data = Credit)
##
## Coefficients:
##  (Intercept)  GenderFemale
##       509.80         19.73
```

```r
summary(lm(formula = Balance ~ Gender, data = Credit))
```

```
##
## Call:
## lm(formula = Balance ~ Gender, 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 ***
## GenderFemale    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
```

- If model assumptions are met, `Gender` is not significant on affecting average `Balance` at 5% significance level based on the F-Statistic

Model 2: 3 levels

- `Ethnicity` has 3 levels, `African American`, `Asian`, and `Caucasian`. *2 dummy variables are needed*
  - $x_{i,1} = 0$ if the ith person is *not* `Asian` and $x_{i,1} = 1$ if ith person is `Asian`
  - $x_{i,2} = 0$ if the ith person is *not* `Caucasian` and $x_{i,2} = 1$ if ith person is `Caucasian`
- Model:
  $y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i$
- Codings on previous slide
- β_0: average `Balance` of `African American` card holders
- β_1: average difference in `Balance` between `Asian` and `African American` card holders
- β_2: average difference in `Balance` between `Caucasian` and `African American` card holders

Fitting the Model 2

```r
lm(formula = Balance ~ Ethnicity, data = Credit)
```

```
##
## Call:
## lm(formula = Balance ~ Ethnicity, data = Credit)
##
## Coefficients:
##        (Intercept)      EthnicityAsian  EthnicityCaucasian
##             531.00              -18.69              -12.50
```

- `African American` card holders have an average balance of \$531
- `Asian` card holders have an average balance of $\$(531 - 18.69) = \$512.31$
- `Caucasian` card holders have an average balance of $\$(531 - 12.50) = \$518.50$

Testing the Model 2

```r
lm(formula = Balance ~ Ethnicity, data = Credit)
```

```
##
## Call:
## lm(formula = Balance ~ Ethnicity, data = Credit)
##
## Coefficients:
##        (Intercept)      EthnicityAsian  EthnicityCaucasian
##             531.00              -18.69              -12.50
```

```r
summary(lm(formula = Balance ~ Ethnicity, data = Credit))
```

```
##
## Call:
## lm(formula = Balance ~ Ethnicity, 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 ***
## EthnicityAsian       -18.69      65.02  -0.287    0.774
## EthnicityCaucasian   -12.50      56.68  -0.221    0.826
## ---
## 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
```

- If model assumptions are met, at 5% significance level, `Ethnicity` does not significantly affect average `Balance` based on the F-Statistic

```r
tidy(lm(formula = Balance ~ Ethnicity, data = Credit))
```

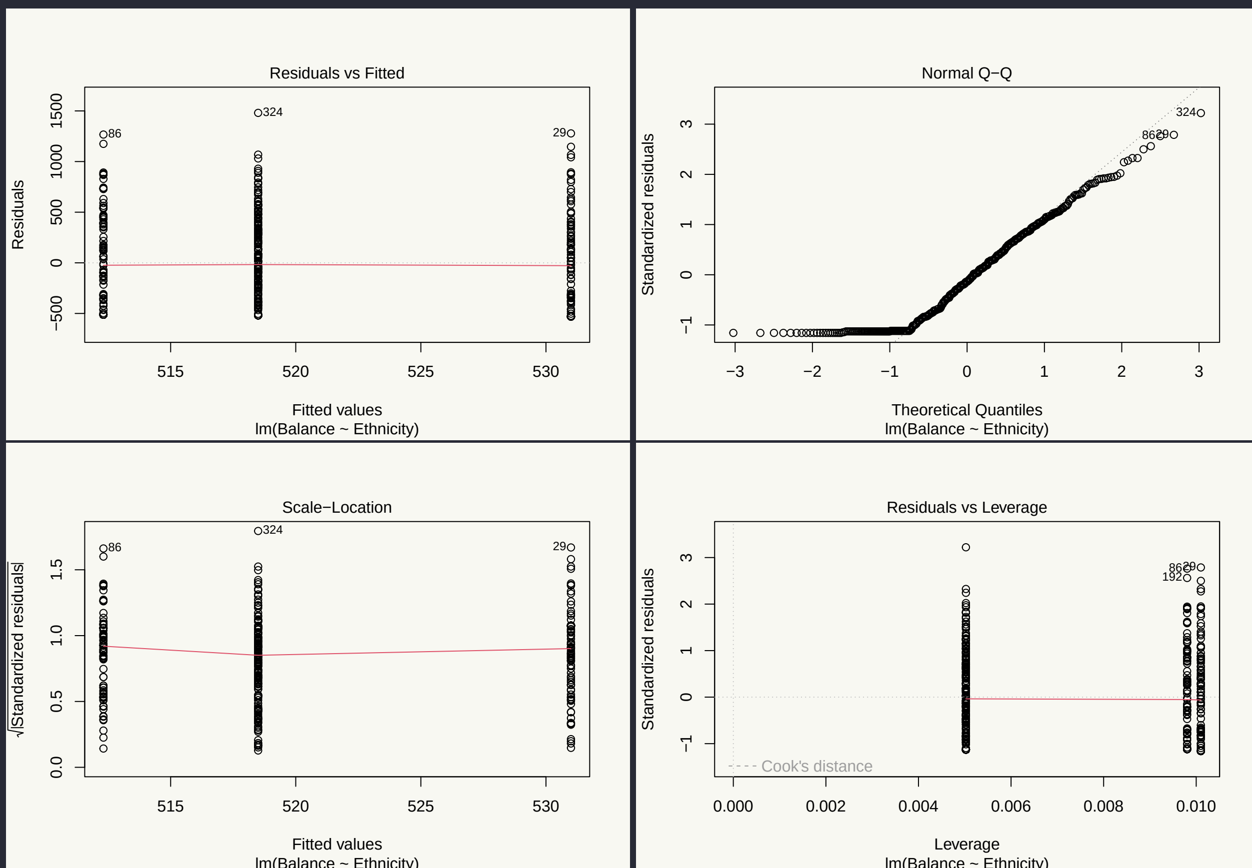
```
## # A tibble: 3 x 5
##   term               estimate std.error statistic  p.value
##   <chr>                 <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)           531        46.3    11.5   1.77e-26
## 2 EthnicityAsian        -18.7      65.0    -0.287 7.74e- 1
## 3 EthnicityCaucasian    -12.5      56.7    -0.221 8.26e- 1
```

- If model assumptions are met and `Ethnicity` does not significantly affect average `Balance`, there is no need to check:
  - whether there is a significant difference in average `Balance` between `Asian` and `African American` card holders or between `Caucasian` and `African American` card holders

Diagnostics

- Diagnostics are the same as those for simple linear regression with a quantitative predictor

```
par(bg = '#f8f8f2')
plot(lm(Balance ~ Ethnicity, data = Credit)) # I only had to use this to generate all 4 plots
```



Multiple Linear Regression

Motivation

- How is `sales` (in thousands of units) for a particular product related to advertising budgets (in thousands of dollars) for `TV`, `Radio`, and `Newspaper`?
- Model:

  $\text{sales} = \beta_0 + \beta_1 \text{TV} + \beta_2 \text{Radio} + \beta_3 \text{Newspaper} + \epsilon$
- We want to examine the relationship between `sales` and budgets for `TV`, `Radio`, and `Newspaper` *jointly* instead of *marginally*

Model

- Response Y and p predictors $X_1, X_2, \ldots, X_p$ are bound by the model:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$

- β_j change in units in $E(Y)$ for a unit change in X_j with all other predictors held constant
- ϵ: error term with $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$
- Estimate coefficient vector $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ by the *least squares method*; estimate $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ as *LSE* (least squares estimator)

Fitting the Model

- Joint model vs marginal model