Statistics 452: Statistical Learning and Prediction

Chapter 3, Part 3: Other Considerations in Multiple Linear Regression

Brad McNeney

2018-09-01

Topics

- Interaction and non-linear model terms
- Categorical variables as predictors
- Model diagnostics

Example Data: Credit Card Balances

- Understand which variables are associated with credit card balance.

```r
uu <- url("http://www-bcf.usc.edu/~gareth/ISL/Credit.csv")
credit <- read.csv(uu,row.names=1)
head(credit)
```

```
##    Income Limit Rating Cards Age Education Gender Student Married
## 1  14.891  3606    283     2  34        11   Male      No     Yes
## 2 106.025  6645    483     3  82        15 Female     Yes     Yes
## 3 104.593  7075    514     4  71        11   Male      No      No
## 4 148.924  9504    681     3  36        11 Female      No      No
## 5  55.882  4897    357     2  68        16   Male      No     Yes
## 6  80.180  8047    569     4  77        10   Male      No      No
##   Ethnicity Balance
## 1 Caucasian     333
## 2     Asian     903
## 3     Asian     580
## 4     Asian     964
## 5 Caucasian     331
## 6 Caucasian    1151
```

Software Note

- In R, categorical variables should be stored as **factors**.

Interaction and Non-Linear Model Terms

Interaction and Non-Linear Model Terms

- We have already seen non-linear model terms when we modelled the relationship between income and education as a polynomial.
- We now discuss interaction.

Statistical Interaction

- Start with two explanatory variables income (X_1) and student status (X_2)
  - `StudentYes=1` if the person is a student and 0 otherwise.
- X_2 is said to modify the effect of X_1 on Y if the regression slope of the regression of Y on X_1 differs in the $X_2 = 0$ and $X_2 = 1$ sub-groups.
  - If we stratify the analysis by student status and find different effects of income in the two groups, there is statistical interaction between income and student status.

Model for Stratification by Student Status

- $f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$ where
  - X_1 is income
  - X_2 is student status (1 is student, 0 is not)
  - $X_1 \times X_2$ is the statistical interaction between income and student status.
  - β_1, β_2, and β_3 are the corresponding regression coefficients.
- This model allows separate lines for the two values of student status.
  - student status = 0 model: intercept β_0 and slope β_1
  - student status = 1 model: intercept $\beta_0 + \beta_2$ and slope $\beta_1 + \beta_3$
  - Interpret β_3 as the difference between slopes.
  - If $\beta_3 = 0$, then student status does not modify effect of income on balance.
  - In practice, we test the hypothesis $H_0 : \beta_3 = 0$.

Fitted Model

```
cfit <- lm(Balance ~ Income*Student,data=credit)
summary(cfit)$coefficients
```

```
##                     Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)       200.623153   33.6983706  5.953497 5.789658e-09
## Income              6.218169    0.5920936 10.502003 6.340684e-23
## StudentYes        476.675843  104.3512235  4.567995 6.586095e-06
## Income:StudentYes  -1.999151    1.7312511 -1.154743 2.488919e-01
```

- The t-test of the hypothesis $H_0 : \beta_3 = 0$ does not reject at any of the standard levels.
- That is, we retain the hypothesis that student status does not modify the effect of income on balance.

Statistical Interaction More Generally

- Interaction terms are generally defined as products of two other model terms:
  - $f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$
- In general we allow different lines for different values of X_2.
  - $X_2 = 0$ model: intercept β_0 and slope β_1
  - $X_2 = x_2$ model: intercept $\beta_0 + \beta_2 x_2$ and slope $\beta_1 + \beta_3 x_2$
  - Interpret β_3 as the difference between slopes for a one-unit change in X_2.
  - If $\beta_3 = 0$ then X_2 does not modify effect of X_1 on Y and *vice versa*.
  - In practice, test the hypothesis $H_0 : \beta_3 = 0$.

Interaction Between TV and Radio Advertising

```
afit <- lm(sales ~ TV*radio,data=advert)
summary(afit)$coefficients
```

```
##                Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept) 6.750220203 0.2478713699 27.232755 1.541461e-68
## TV          0.019101074 0.0015041455 12.698953 2.363605e-27
## radio       0.028860340 0.0089052729  3.240815 1.400461e-03
## TV:radio    0.001086495 0.0000524204 20.726564 2.757681e-51
```

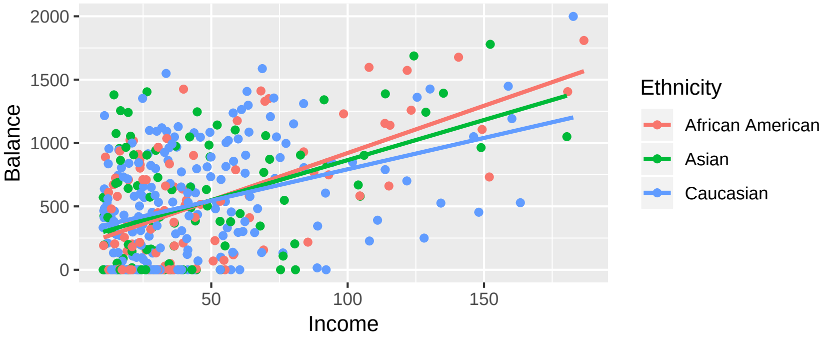
- There is statistical interaction between TV and radio ads.
  - TV modifies the effect of radio, or radio modifies the effect of TV

Categorical Variables as Predictors

Dummy Variables for Categorical Predictors

- We have seen dummy variables before.
  - A binary variable for student status, coded 0 or 1 to represent the two levels of a dichotomous variable.
- When the categorical variable has more than two values, or levels, we need more than one binary "dummy" variable.
  - Example: The `Ethnicity` variable from the credit data.

```r
ggplot(credit,aes(x=Income,y=Balance,color=Ethnicity)) +
  geom_point() + geom_smooth(method="lm",se=FALSE)
```



Regression Model for Balance

```
cfit <- lm(Balance ~ Income*Ethnicity,data=credit)
summary(cfit)$coefficients
```

```
##                               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)                 175.495201  65.192852  2.6919393 7.406626e-03
## Income                        7.455728   1.062646  7.0161899 1.000371e-11
## EthnicityAsian               57.347367  91.784711  0.6248030 5.324620e-01
## EthnicityCaucasian          122.814867  81.198364  1.5125288 1.312011e-01
## Income:EthnicityAsian        -1.131111   1.559045 -0.7255149 4.685669e-01
## Income:EthnicityCaucasian    -2.510135   1.374387 -1.8263672 6.855154e-02
```

Model Details

- Model requires too much notation to write out in detail; will explain in the context of this example.

```r
coefficients(cfit)
```

```
##              (Intercept)                   Income 
##               175.495201                 7.455728 
##           EthnicityAsian       EthnicityCaucasian 
##                57.347367               122.814867 
##    Income:EthnicityAsian Income:EthnicityCaucasian 
##                -1.131111                -2.510135
```

- The model uses African American as a "baseline".
  - The intercept=175 and slope `Income`=7.46 terms are the fitted model for mean `Balance` in African Americans.
- The model for mean `Balance` in another ethnic group is the baseline plus ethnic-group-specific intercept and slope
  - E.G., for the Asians, add 57.3 to the African American intercept and -1.13 to the African American slope

Dummy Variables for Ethnic Group

- Create a binary variable for each non-baseline ethnic group that takes value 1 if the person is from that ethnic group and 0 otherwise.

```
##   Ethnicity  Income EthnicityAsian EthnicityCaucasian
## 1 Caucasian  14.891              0                  1
## 2     Asian 106.025              1                  0
## 3     Asian 104.593              1                  0
## 4     Asian 148.924              1                  0
## 5 Caucasian  55.882              0                  1
## 6 Caucasian  80.180              0                  1
```

Model with separate lines for each continent

```
##   Ethnicity  Income EthnicityAsian EthnicityCaucasian
## 1 Caucasian  14.891              0                  1
## 2     Asian 106.025              1                  0
## 3     Asian 104.593              1                  0
## 4     Asian 148.924              1                  0
## 5 Caucasian  55.882              0                  1
## 6 Caucasian  80.180              0                  1
##   Income.EthnicityAsian Income.EthnicityCaucasian
## 1                 0.000                    14.891
## 2               106.025                     0.000
## 3               104.593                     0.000
## 4               148.924                     0.000
## 5                 0.000                    55.882
## 6                 0.000                    80.180
```

Multiple-Partial F-test for Interaction

```
cfitReduced <- lm(Balance ~ Income + Ethnicity,data=credit)
anova(cfitReduced,cfit)
```

```
## Analysis of Variance Table
##
## Model 1: Balance ~ Income + Ethnicity
## Model 2: Balance ~ Income * Ethnicity
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1    396 66205558
## 2    394 65636921  2    568637 1.7067 0.1828
```

- There is little evidence that ethnicity modifies the effect of income on balance.

Model Diagnostics based on Residuals

Model Diagnostics based on Residuals

- Residuals are the primary tool for
  - checking model assumptions (correct linear model, constant error SD, and normal errors) and
  - identifying unusual observations.
- Residuals may also be useful for detecting correlation in the errors, but this is a more specialized topic not discussed.

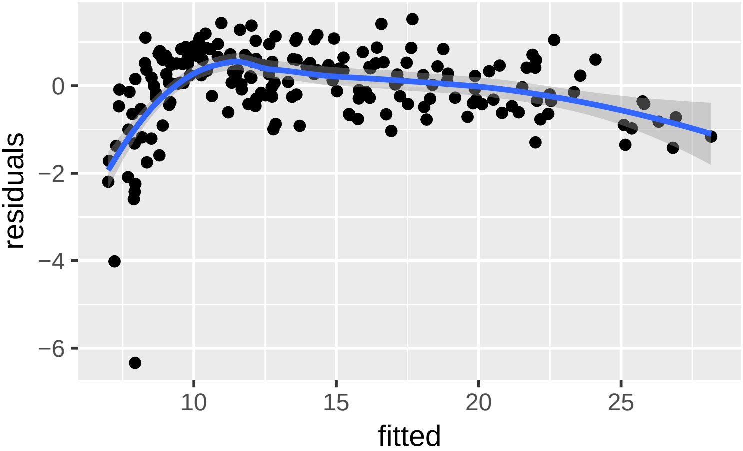
Residuals versus fitted values

- Plot residuals *vs* fitted values to assess adequacy of the linear model and constant error SD.
  - A plot of ϵ_i *vs* $f(x_i)$ would show no pattern, because the ϵ_i's are random noise.
  - If the linear model is adequate, we should not see any trends or patterns in the residuals *vs* fitted values $\hat{y}_i = \hat{f}(x_i)$.
  - Also, if the error SD is constant, the variation in residuals *vs* $\hat{y}_i$ should look roughly equal.
- We may also see outliers in the regression sense.

Residuals versus Fitted Values - Advertising

- Use the `residual()` and `fitted()` extractor functions.

```
adAug <- data.frame(advert,fitted=fitted(afit),residuals=residuals(afit))
ggplot(adAug,aes(x=fitted,y=residuals)) +
  geom_point() + geom_smooth()
```



Residual vs fitted – comments

- Horizontal line at zero is outside the error bands around smoother line.
  - Suggests we have missed a non-linear trend.
- Spread of residuals fairly constant over range of fitted values, so constand SD assumptions appears reasonable.

Q-Q Plots

- A quantile-quantile (Q-Q) plot is a plot of the quantiles of one distribution to another.
  - If the two distributions have the similar shape, the points should fall roughly on a straight line.
- Our interest is in comparing the quantiles of the distribution of residuals to the quantiles of the distribution they should have under normal errors.
  - One can argue that the residuals don't have the same distribution (those closer to the centre of the plot are slightly more variable).
- However, Studentized residuals do – they have a t distribution with $n - k - 2$ degrees of freedom.
  - The Studentized residual for the ith case is

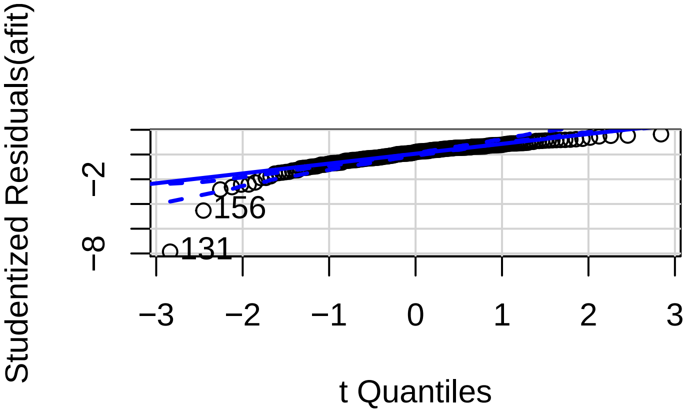
$$\frac{y_i - \hat{y}_i}{\hat{\sigma}_{-i}\sqrt{1 - h_{ii}}}.$$

where h_{ii} is the leverage or hat value and $\hat{\sigma}_{-i}$ is the estimate of the error SD **without** case i

Q-Q plot of Studentized residuals

- The `qqPlot()` function from the `car` package plots Studentized residuals against quantiles of the appropriate t distribution.
  - Also adds error bands: If all points fit within bands, it is plausible that the sample is from the t.

```r
library(car) # Use install.packages("car") to install
qqPlot(afit)
```



Identifying unusual observations

- Studentized residuals can identify outliers
  - Rule-of-thumb: Residuals beyond ± 2 are moderate outliers and beyond ± 3 are serious outliers.
- Leverage (h_i) is a measure of how atypical an observation's X values are.
  - Rule-of-thumb: $h_i > 2(p+1)/n$ is somewhat high leverage and $h_i > 3(p+1)/n$ is very high leverage.
- Cook's distance measures the influence of an observation; i.e., how much the estimated regression coefficients change when the observation is removed.
  - Rule-of-thumb: Cook's distance > 0.5 is moderately influential, and > 1 is highly influential

Identify Unusual Observations in Advertising Data

- Augment the dataset and `View()` to identify cases that are unusual according to our rules-of-thumb.
  - Several "moderate" residuals, and two severe residuals of -4.5 and -7.9.
  - For leverage, $p = 2$, $n = 200$, $2(p+1)/n = 0.03$, $3(p+1)/n = 0.045$: 27 cases with moderate leverage, 12 with very high leverage!
  - One moderately influential case.

```r
adAug <- data.frame(advert,studRes = rstudent(afit),
                    hats = hatvalues(afit),
                    cooks = cooks.distance(afit))
# Now View(adAug)
```

Correlated Predicters, or Collinearity

- The distribution of X's can affect stability of the least squares estimates.
  - For simple linear regression one can show that:

    $$SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{S_X\sqrt{n-1}}$$

    where S_X is the SD of the X's in the dataset.
  - Implies that the larger the S_X the smaller the SE (i.e., the more stable the fit).
- In general can think of the positioning of X's as a "foundation" that supports the least squares surface – the broader the base, the more stable the estimates.
- Collinearity, or correlation between predictors, yields an unstable foundation and hence unstable estimates.

More on SEs

- With two explanatory variables, X_1 and X_2, can show that

$$SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{S_{X_1}\sqrt{n-1}\sqrt{1-r_{12}^2}}$$

and

$$SE(\hat{\beta}_2) = \frac{\hat{\sigma}}{S_{X_2}\sqrt{n-1}\sqrt{1-r_{12}^2}}$$

where r_{12} is the correlation between X_1 and X_2.
- In addition to S_{X_1} and S_{X_2} we must consider the correlation between X_1 and X_2.
- The larger the squared correlation, the larger the SEs

Variance Inflation Factors (VIFs)

- In a multiple regression with $X_1, \ldots, X_p$,

$$SE(\hat{\beta}_j) = \frac{\hat{\sigma}}{S_{X_j}\sqrt{n-1}\sqrt{1-R_j^2}}$$

  where R_j^2 is the R^2 from the regression of X_j on $X_{(-j)}$.
- The term $1/\sqrt{1-R_j^2}$, is the factor by which the SE of $\hat{\beta}_j$ is inflated over the SE from a simple linear regression by correlation between X_j and the other X's.
- The variance inflation factor for X_j, VIF_j, is defined to be $1/(1-R_j^2)$; i.e., the SE of $\hat{\beta}_j$ is inflated by $\sqrt{VIF_j}$.
- High VIFs indicate instability.
  - One rule of thumb is that a $VIF_j > 10$ is cause for concern.

VIFs and Other Diagnostics in the car Package

- If you haven't already done so, install the R package car.

```
library(car)
vif(afit)
```

```
##       TV    radio TV:radio
## 3.727848 3.907651 6.937860
```

- The VIFs suggest the interaction is quite well predicted by TV and radio, but the VIF is less than our threshold so we are not concerned.

Collinearity with Polynomial Terms

```r
uu <- url("http://www-bcf.usc.edu/~gareth/ISL/Income1.csv")
income <- read.csv(uu,row.names=1)
ifit<- lm(Income ~ Education + I(Education^2) + I(Education^3), data=income)
vif(ifit)
```

```
##       Education I(Education^2) I(Education^3)
##        5612.306      23764.744       6449.139
```

Remedies for collinearity

- When the collinearity arrises from explanatory variables that are products of other variables, centering can help.

```
centre <- function(x) { x - mean(x)}
income <- data.frame(income, cEducation = centre(income$Education))
ifit<- lm(Income ~ cEducation + I(cEducation^2) + I(cEducation^3), data=income)
vif(ifit)
```

```
##      cEducation I(cEducation^2) I(cEducation^3)
##        6.275936        1.000000        6.275936
```

Collinearity: If centering doesn't help

- May need to exclude a variable.
  - Sounds drastic, but high R_j^2 indicates X_j is very well predicted by $X_{(-j)}$, so nothing really lost.
- Which variable to exclude?
- First use common sense:
  - If one variable is a surrogate for another, drop the surrogate.
  - For example, if we are modeling house prices with (i) size of the house in square feet and (ii) the number of bedrooms, we may think bedrooms is just a surrogate for size.
- If no obvious candidate to drop, use model selection.

Fit of the Advertising Data Revisited

```r
library(dplyr)
advert <- mutate(advert,cTV = TV - mean(TV),
         cRadio = radio - mean(radio))
afitC <- lm(sales ~ cTV*cRadio+I(cTV^2) +I(cTV^3)+I(cTV^4),
            data=advert)
summary(afitC)$coefficients
```

```
##                  Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)  1.449416e+01 6.224612e-02 232.852415 2.089992e-238
## cTV          3.355736e-02 8.898603e-04  37.710818  5.634012e-91
## cRadio       1.960573e-01 2.130075e-03  92.042410 2.118337e-161
## I(cTV^2)     1.829559e-05 1.766454e-05   1.035724  3.016270e-01
## I(cTV^3)     7.812590e-07 6.237580e-08  12.525033  9.967540e-27
## I(cTV^4)    -6.750627e-09 8.823011e-10  -7.651160  9.242163e-13
## cTV:cRadio   1.041128e-03 2.468919e-05  42.169394  2.478978e-99
```

```r
vif(afitC)
```

```
##       cTV    cRadio   I(cTV^2)   I(cTV^3)   I(cTV^4) cTV:cRadio
##  5.961561  1.021525  13.337249   5.992502  13.370396   1.020922
```

```r
mutate(advert,fitted=fitted(afitC),residuals=residuals(afitC)) %>%
  ggplot(aes(x=fitted,y=residuals)) + geom_point() + geom_smooth()
```