Statistics 452: Statistical Learning and Prediction

Chapter 3, Part 2: Multiple Linear Regression

Brad McNeney

2018-09-01

Multiple Linear Regression Model

- Recall our general model from Chapter 2:

$$Y = f(X) + \epsilon$$

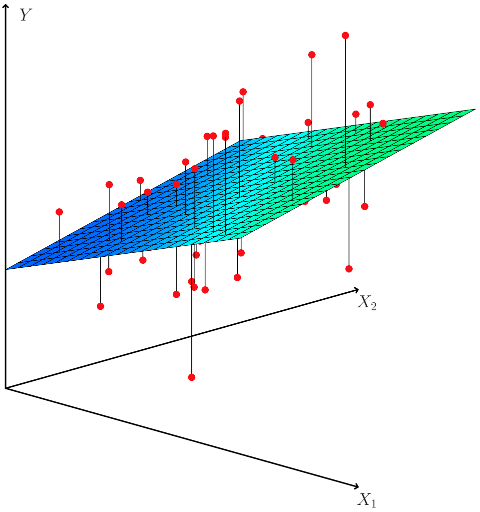
- Multiple linear regression assumes the function f is linear in $p \geq 1$ predictors $X = (X_1, \ldots, X_p)$; i.e., $f(X) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$.
  - β_0 is the intercept and
  - β_i is the slope for the ith predictor: A one unit increase in X_i *holding all other predictors fixed* is associated with a β_i increase in f.

Fitting the line

- We use the method of least squares to fit the line.
- Goal: Using observed data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ (where now x_i is a vector of length p) fit the model

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip}$$

  where $\hat{y}_i$ is the fitted value of Y for $X = x_i$.
- The residuals are still defined as vertical distances $e_i = y_i - \hat{y}_i$ (see Figure 3.4 of text, copied on next slide).
- Least squares finds the $\hat{\beta}_0, \ldots, \hat{\beta}_p$ that minimize $RSS = \sum_{i=1}^{n} e_i^2$.


Y
X_2
X_1

Advertising Example

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

```
##                 Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)  2.938889369 0.311908236  9.4222884 1.267295e-17
## TV           0.045764645 0.001394897 32.8086244 1.509960e-81
## newspaper   -0.001037493 0.005871010 -0.1767146 8.599151e-01
## radio        0.188530017 0.008611234 21.8934961 1.505339e-54
```

```
confint(afit)
```

```
##                   2.5 %     97.5 %
## (Intercept)  2.32376228 3.55401646
## TV           0.04301371 0.04851558
## newspaper   -0.01261595 0.01054097
## radio        0.17154745 0.20551259
```

- We are 95% confident that an increase of $1000 in TV advertising, holding newspaper and radio ads fixed, is associated with an increase in sales of between 43 and 49 units.
  - Compare to interval estimate (42,53) from simple linear regression.

Advertising Example: Effect of Newspaper Ads

```r
afitN <- lm(sales ~ newspaper, data=advert)
summary(afitN)$coefficients
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 12.3514071 0.62142019 19.876096 4.713507e-49
## newspaper    0.0546931 0.01657572  3.299591 1.148196e-03
```

```r
summary(afit)$coefficients
```

```
##                 Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)  2.938889369 0.311908236  9.4222884 1.267295e-17
## TV           0.045764645 0.001394897 32.8086244 1.509960e-81
## newspaper   -0.001037493 0.005871010 -0.1767146 8.599151e-01
## radio        0.188530017 0.008611234 21.8934961 1.505339e-54
```

- Newspaper ads are significantly associated with sales in the simple but not the multiple regression.

Confounding

- The effect of newspaper is different depending on whether or not we include TV and radio ads in the model.
  - TV and radio are said to **confound** the newspaper effect.
- Correlation between radio and newspaper is behind the confounding.
  - More radio ads, more sales. More radio adds more newspaper ads. $\Rightarrow$ More newspaper ads, more sales.

```r
cor(advert)
```

```
##                   TV      radio  newspaper     sales
## TV        1.00000000 0.05480866 0.05664787 0.7822244
## radio     0.05480866 1.00000000 0.35410375 0.5762226
## newspaper 0.05664787 0.35410375 1.00000000 0.2282990
## sales     0.78222442 0.57622257 0.22829903 1.0000000
```

Testing the Overall Effect of Predictors

- Hypothesis of no association between the outcome and the predictors is
$$H_0 : \beta_1 = \cdots = \beta_p = 0$$
and the alternative hypothesis is
$$H_a : \text{at least one } \beta_j \text{ is non-zero}$$
- We test H_0 *vs* H_a with an F test. The F statistic is MSM/MSE, where
  - $MSM = SSM/p$, with $SSM = TSS - RSS$, and
  - $MSE = RSS/(n - p - 1)$.
- F is compared to an F-distribution with p numerator and $n - p - 1$ denominator df.

Advertising Example

```r
summary(afit)
```

```
##
## Call:
## lm(formula = sales ~ TV + newspaper + radio, data = advert)
##
## 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 ***
## newspaper   -0.001037   0.005871  -0.177     0.86
## radio        0.188530   0.008611  21.893   <2e-16 ***
## ---
## 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
```

- There is strong evidence that TV, radio and newspaper advertising is associated with sales.

Testing the Effect of a Subset of Predictors

- E.G., suppose we are interested in testing $H_0 : \beta_2 = \beta_3 = 0$ *vs* $H_a : \beta_2$ and β_3 are not both zero.
- We do a multiple-partial F test.
- The test statistic is

$$\frac{(RSS(red) - RSS(full))/q}{RSS(full)/(n-p-1)}$$

  where

  - $RSS(red)$ is the RSS from the reduced model,
  - $RSS(full)$ is the RSS from the full model, and
  - q is the difference in the number of model parmeters in the two models.

- The test statistic is compared to an F-distribution with q numerator and $n-p-1$ denominator df.

Advertising Example

```r
afitTV <- lm(sales ~ TV, data=advert) # reduced model
anova(afitTV,afit)
```

```
## Analysis of Variance Table
##
## Model 1: sales ~ TV
## Model 2: sales ~ TV + newspaper + radio
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    198 2102.53
## 2    196  556.83  2    1545.7 272.04 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- There is strong evidence that newspaper and radio ads are associated with sales, adjusting for TV ads.

Variable Selection

- Multiple-partial F-tests can be used for selecting subsets of variables that explain the association between Y and X.
- Alternately, we can think of variable selection as selecting $f(X)$ to avoid over-fitting.
- We will study modern methods for variable selection in Chapter 6.
- Here we mention three classical approaches, forward selection, backward selection and mixed (stepwise) selection.

Selection Strategies

1. Forward selection. Start with a smallest model (e.g., null model) and try to add terms up to some largest model.
2. Backward selection. Start with a largest model and try to drop terms, down to some smallest model.
3. Stepwise selection. Try adding *and* dropping terms, staying between a smallest and largest model.

- (1-3) can be described in terms of adding (ADD1) and dropping (DROP1) steps.

ADD1 and DROP1

- ADD1:
  - Given a current model (subset of X) M_c, try to find a model term *not* in M_c that will improve the model.
  - Whether or not a particular term improves the model is a model comparison.
  - If can't find a term to add that will improve the model, do nothing.
- DROP1:
  - Given a current model M_c, find a term *in* M_c that can be dropped to improve the model.
  - Whether or not a particular term improves the model is a model comparison.
  - If can't find a term to drop that will improve the model, do nothing.
- Both ADD1 and DROP1 make model comparisons and need to know when adding/dropping terms is an improvement.

Model Comparisons

- In Chapter 6 we will consider several types of comparisons of M_{full} to M_{red}.
- Given what we know now, we could use partial F tests or t tests of the null hypothesis that the coefficient being added/dropped is zero *vs* the two-sided alternative.

Forward Selection

- Starting from the smallest model, apply ADD1 to try adding a term.
- If we can add a term, do so. Othewise stop.
- Repeat until we stop or reach the largest model.

Backward Selection

- Starting from the largest model, apply DROP1 to try dropping a term.
- If we can drop a term, do so. Othewise stop.
- Repeat until we stop or reach the smallest model.

Stepwise Selection

- Starting from either the largest or smallest model, apply ADD1 and DROP1 to try to find a better model
  - But never add to largest or drop from smallest models.
- If we can either add or drop a term, do so. Othewise stop.
- Repeat until we stop or reach the model at the opposite extreme.
  - That is, if we reach the smallest having started from the largest, or if we reach the largest having started from the smallest.

Advertising Example

- Output from `lm` summary makes it easiest to use t tests and backward selection.
  - Though not possible if $p > n$. More on this in Chapter 6.

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

```
##                 Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)  2.938889369 0.311908236  9.4222884 1.267295e-17
## TV           0.045764645 0.001394897 32.8086244 1.509960e-81
## newspaper   -0.001037493 0.005871010 -0.1767146 8.599151e-01
## radio        0.188530017 0.008611234 21.8934961 1.505339e-54
```

```
afit2<-lm(sales ~ TV + radio,data=advert)
summary(afit2)$coefficients
```

```
##               Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 2.92109991 0.294489678  9.919193 4.565557e-19
## TV          0.04575482 0.001390356 32.908708 5.436980e-82
## radio       0.18799423 0.008039973 23.382446 9.776972e-59
```

Model Fit

- Can use R^2 to describe the fit of the model

```r
summary(afit2)$r.squared
```

```
## [1] 0.8971943
```

```r
# Compare with summary(afit)$r.squared
```

- TV and radio advertising expenditures explain about 90% of the variation in sales.

Predictions

- The fitted model can be used to make predictions.
- For value x_0 of X, the prediction is

$$\hat{\beta}_0 + \hat{\beta}_1 x_{01} + \ldots + \hat{\beta}_p x_{0p}$$

```
newdat <- data.frame(TV=150,radio=20)
predict(afit2,newdata=newdat)
```

```
##        1
## 13.54421
```

Sources of Uncertainty

- Three sources:
  1. Model bias: The model $f(x_0)$ may be wrong. We will ignore this for now; i.e., assume Bias($\hat{f}(x_0)$)=0.
  2. Estimation: Our $\hat{f}$ is based on $\hat{\beta}_1, \ldots, \hat{\beta}_p$, which will not be equal to $\beta_1, \ldots, \beta_p$. Var($\hat{f}(x_0)$) is part of the reducible error.
  3. Irreducible error: $Y = f(x_0) + \epsilon$, and so even if $\hat{f} = f$, predictions will not be perfect.
- We construct confidence intervals for $f(x_0)$ to quantify estimation uncertainty, and prediction intervals to quantify estimation uncertainty plus irreducible error.

Confidence Intervals for $f(x_0)$.

- For fixed x_0, $f(x_0)$ is a function of parameters, and so is a parameter.
- We can construct a confidence interval for $f(x_0)$ based on the sampling distribution of $\hat{f}(x_0)$.
  - Details are not important. We will use R.

```
predict(afit2,newdata=newdat,interval="confidence",level=0.95)
```

```
##        fit      lwr      upr
## 1 13.54421 13.30387 13.78454
```

- We are 95% confident that the average sales for cities in which there are \$150,000 in TV ads and \$20,000 in radio ads is between 13,304 and 13,785.

Prediction Intervals

- Prediction intervals are constructed to contain a given proportion of *future* observations.
- The intervals must account for both the reducible error from estimating f and the irreducible error ϵ.
  - Details are not important. We will use R.

```
predict(afit2,newdata=newdat,interval="prediction",level=0.95)
```

```
##        fit      lwr      upr
## 1 13.54421 10.21973 16.86868
```

- We believe that 95% of future observations of cities with \$150,000 in TV ads and \$20,000 in radio ads will have sales between 10,220 and 16,869.