Stat 435 Intro to Statistical Machine Learning

Week 2: Linear Regression

Richard Li

April 4, 2017

Plan for today

- A review of multiple linear regression (Sec 3.2)
  - Parameter estimation
  - Prediction
- Complications/Problems in regression (Sec 3.3)

Linear Regression: an overview

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x}_1 + ... + \beta_p \boldsymbol{x}_p + \boldsymbol{\epsilon}$$

- Or in matrix form:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- Notice about the intercept...
- What are the dimensions of the bold letters?

Linear Regression: an overview

1. Is there a relationship between response Y and predictors $X_1, .., X_p$.
2. How accurately can we estimate the effects of X on Y?
3. Which predictors predict/explain Y?
4. How accurate can we predict (future) Y?
5. Other considerations?

Estimating the Regression Coefficients: Reviews

- Estimate $\hat{\beta}_0, ..., \hat{\beta}_p$
    - If we write the intercept as the first column of $\boldsymbol{X}$, $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$
    - Derived from minimizing sum of squared residuals.
- What is sum of squared residuals?
    - $RSS = \sum_i^n (y_i - \hat{y}_i)^2)$
- What is total sum of squares?
    - $TSS = \sum_i^n (y_i - \bar{y})^2$

Hypothesis testing of model utility

1. Is there a relationship between response Y and any predictors $X_1, .., X_p$.
2. How accurately can we estimate the effects of X on Y?
3. Which predictors predict/explain Y?
4. How accurate can we predict (future) Y?
5. Other considerations?

Hypothesis testing of model utility

- Null hypothesis

$$H_0 : \text{There is no relationship between any } X \text{ and } Y\text{, or}$$

$$H_0 : \beta_0 = \beta_1 = ... = \beta_p = 0.$$

- Alternative hypothesis

$$H_1 : \text{There is some relationship between some } X \text{ and } Y\text{, or}$$

$$H_1 : \text{at least one } \beta_j \neq 0.$$

- **F-statistics**

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

Hypothesis testing of model utility

- If F-statistic is large, we have more evidence to reject H_0.
- F-statistics follow a F distribution (df: $(p, n-p-1)$) if
  - ϵ_i follow a normal distribution (our assumption),
  - or if ϵ_i are not normally distributed, F-statistic still follow F distribution *approximated* if n is large.
- How large is large? Look for p-values!
- Caveat: notice we need degree of freedom $n - p - 1 > 0 \rightarrow$ F-test not useful when $n \leq p$!

Hypothesis testing of single regression coefficient

1. Is there a relationship between response Y and any predictors $X_1, .., X_p$.
2. How accurately can we estimate the effects of X on Y?
3. Which predictors predict/explain Y?
4. How accurate can we predict (future) Y?
5. Other considerations?

Hypothesis testing of single regression coefficient

- $\hat{\beta}_j$: additional contribution of $\boldsymbol{x}_j$ on $\boldsymbol{y}$ conditional or $\boldsymbol{x}_0, ..., \boldsymbol{x}_{j-1}, \boldsymbol{x}_{j+1}, ..., \boldsymbol{x}_p$.
- $\hat{\beta}_j$ follows a t-distribution (usually approximated by Normal distribution).
- Testing $H_0 : \beta_j = 0$ v.s. $H_1 : \beta_j \neq 0$ adjusting other predictors
    - F-test, or Analysis of Variance (ANOVA)
    - t-test

Hypothesis testing of single regression coefficient

- F-test
  - Comparing the full model ($\boldsymbol{y} = f(\boldsymbol{x}_1, ..., \boldsymbol{x}_p)$) with the reduced model ($\boldsymbol{y} = f(\boldsymbol{x}_1, ..., \boldsymbol{x}_{j-1}, \boldsymbol{x}_{j+1}, ..., \boldsymbol{x}_p)$).
  - *Formula for test statistic F^* omitted here.*
- t-test

$$t^* = \frac{\hat{\beta}}{se(\hat{\beta})}$$

- Interpretation of p-value: conditional on all other estimated coefficients, when the true $\beta_j = 0$, obtaining a $\hat{\beta}_j \neq 0$ has probability ...

Example: Carseats dataset

- *Carseats*: dataset in the *ISLR* library
- *Sales*: unit sales (in thousands) of child car seats at each location.
- *Price*: Price company charges for car seats at each site
- *CompPrice*: Price charged by competitor at each location
- *Income*: Community income level (in thousands of dollars)
- *Advertising*: Local advertising budget for company at each location (in thousands of dollars)
- *Population* Population size in region (in thousands)
- ...

```
library(ISLR)
data(Carseats)
?Carseats
```

Example: Regression in R

```
fit <- lm(Sales ~ Age + Price + CompPrice + Population + Income:Advertising,
          data = Carseats)
summary(fit)
```

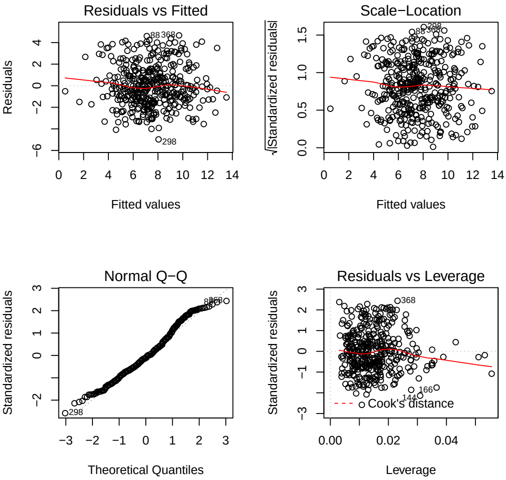
```
##
## Call:
## lm(formula = Sales ~ Age + Price + CompPrice + Population + Income:Advertising,
##     data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9767 -1.3119 -0.2094  1.2252  4.6683
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         7.916e+00  9.368e-01   8.450 5.68e-16 ***
## Age                -4.355e-02  6.040e-03  -7.211 2.87e-12 ***
## Price              -9.204e-02  5.072e-03 -18.144  < 2e-16 ***
## CompPrice           9.409e-02  7.876e-03  11.948  < 2e-16 ***
## Population         -4.293e-05  6.826e-04  -0.063     0.95
## Income:Advertising  1.740e-03  1.856e-04   9.376  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.939 on 394 degrees of freedom
## Multiple R-squared:  0.5345,Adjusted R-squared:  0.5286
## F-statistic: 90.49 on 5 and 394 DF,  p-value: < 2.2e-16
```

Example: F-test of single coefficient

```r
fit_no_pop <- lm(Sales ~ Age + Price + CompPrice + Income:Advertising,
           data = Carseats)
anova(fit, fit_no_pop)

## Analysis of Variance Table
##
## Model 1: Sales ~ Age + Price + CompPrice + Population + Income:Advertising
## Model 2: Sales ~ Age + Price + CompPrice + Income:Advertising
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1    394 1481.2
## 2    395 1481.2 -1  -0.01487 0.004 0.9499
```

Example: Diagnostic plots

```
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(fit)
```



Select subset of useful predictors

1. Is there a relationship between response Y and any predictors $X_1, .., X_p$.
2. How accurately can we estimate the effects of X on Y?
3. Which predictors predict/explain Y?
4. How accurate can we predict (future) Y?
5. Other considerations?

Variable selection

- If p is large, looking at individual p-values is misleading
  - p-value < 0.05: When $\beta_j = 0$, obtaining a $\hat{\beta}_j \neq 0$ has probability smaller than 0.05.
  - i.e., if $p = 100$, we expect 5 variables that are wrongly discovered as useful predictors *by chance*.
- F-test does not suffer from this problem, but require $n > p$.
- Also a problem for interpretation.
- How to reduce p?

Variable selection

- Best-subset selection
  - Try every possible subset of size $1, 2, 3, ..., p$.
  - Only feasible when p is small.
- Forward selection
  - Begin with a model with only intercept.
  - Try each one of the p variables, and **add** the variable that result in lowest RSS.
  - Repeat to add more.
- Backward selection (if $p \leq n$)
  - Begin with a model with all predictors.
  - Try each one of the p variables, and **remove** the variable that result in lowest RSS.
  - Repeat to remove more.
- More elegant approaches later in the course (LASSO, ridge regression, etc.)

Example: Backward selection (FYI)

```
library(MASS)
step <- stepAIC(fit, direction="backward", trace=FALSE)
step$anova

## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## Sales ~ Age + Price + CompPrice + Population + Income:Advertising
##
## Final Model:
## Sales ~ Age + Price + CompPrice + Income:Advertising
##
##
##           Step Df  Deviance Resid. Df Resid. Dev      AIC
## 1                                 394   1481.229 535.6652
## 2 - Population  1 0.0148696       395   1481.244 533.6692
```

Prediction interval

1. Is there a relationship between response Y and any predictors $X_1, .., X_p$.
2. How accurately can we estimate the effects of X on Y?
3. Which predictors predict/explain Y?
4. How accurate can we predict (future) Y?
5. Other considerations?

Prediction interval

- $\hat{y}|\boldsymbol{x} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_p x_p$
- Confidence interval of $\hat{y}|\boldsymbol{x}$
    - Given a fixed $\boldsymbol{x}$, $\hat{y}$ can mean the average response for all observations with $\boldsymbol{x}$.
    - e.g., on average, how many species live on an island with characteristics a, b, and c.
- Prediction interval of $\hat{y}|\boldsymbol{x}$
    - Given a fixed $\boldsymbol{x}$, $\hat{y}$ can also mean the response for one particular observations with $\boldsymbol{x}$.
    - e.g., predict how many species live on Long Island, given we know that it has characteristics a, b, and c.
- Different variance calculation of $\hat{y}$.
- The later is always wider.

Example: Prediction

```
newdata = data.frame(Age = 50, Price = 100, CompPrice = 110, Income = 100,
          Advertising = 10, Population = 300)
predict(fit, newdata, interval = "confidence")

##        fit      lwr      upr
## 1 8.612367 8.285913 8.938821

predict(fit, newdata, interval = "predict")

##        fit      lwr      upr
## 1 8.612367 4.786463 12.43827
```

Other complications

1. Is there a relationship between response Y and any predictors $X_1, .., X_p$.
2. How accurately can we estimate the effects of X on Y?
3. Which predictors predict/explain Y?
4. How accurate can we predict (future) Y?
5. Other considerations?
   - A lot of them!

What assumptions of linear regression are we making?

...and what if they are violated.

- Linear relationship (*nonlinearity*)
- Independence of errors (*correlation of errors*)
- Constant variance of errors, or homoscedasticity (*non-constant variance of errors*)
- No or little multicollinearity (*multicollinearity*)

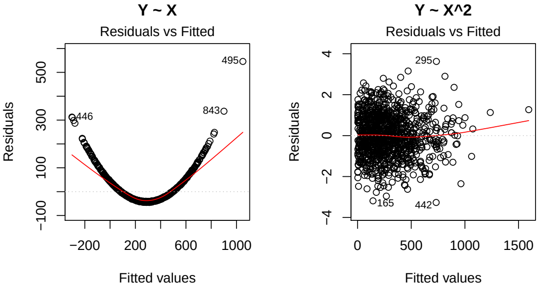
Caveat: Normality of error terms is also an assumption in the standard setting, but inference is still valid if normality is violated but sample size is large. However, prediction interval is problematic in this case.

Non-linearity

- Problem: if the 'straight-line' relationship does not hold, *all conclusions we draw may be wrong!*
- Identify non-linearity
  - Residual plot
- Possible treatment:
  - Transforming your data, e.g., $\sqrt{X}, log(X)$, ...

Example: Residual plot

```r
set.seed(1)
x <- rnorm(1000, mean = 5, sd = 2)
y <- 2 + x^2 * 10 + rnorm(1000)
fit1 <- lm(y ~ x)
fit2 <- lm(y ~ I(x ^ 2))
par(mfrow = c(1, 2))
plot(fit1, 1, main = "Y ~ X")
plot(fit2, 1, main = "Y ~ X^2")
```



Correlation of error terms

- Problem: if error terms are correlated, estimated standard errors will tend to be smaller, i.e., smaller CI and PI.
- Examples:
  - Time-series: Y_i: temperature of Seattle on day i.
  - Clustered: Y_i: scores of every homework of everyone in class.
  - Extreme errors: replicated data, ...
- Possible treatment: many but outside the scope of this class.

Example: Effect of replicating the same data

```
confint(fit2)

##                 2.5 %    97.5 %
## (Intercept) 1.855079  2.072886
## I(x^2)      9.997661 10.003699

x2 <- rep(x, 10)
y2 <- rep(y, 10)
fit3 <- lm(y2 ~ I(x2 ^ 2))
confint(fit3)

##                 2.5 %    97.5 %
## (Intercept) 1.929613  1.998352
## I(x2^2)     9.999727 10.001633
```

Example: Effect of replicating the same data

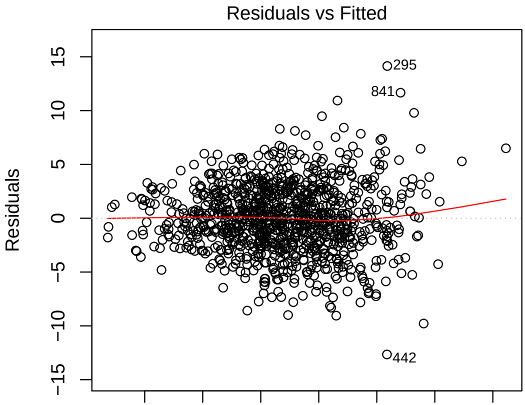
```
set.seed(1)
n0 <- 0
n1 <- 0
for(sim in 1:1000){
        x <- rnorm(1000, mean = 5, sd = 2)
        y <- 2 + x^2 * 10 + rnorm(1000)
        m0 <- lm(y ~ I(x ^ 2))
        CI0 <- confint(m0)[1, ]

        x2 <- rep(x, 10)
        y2 <- rep(y, 10)
        m1 <- lm(y2 ~ I(x2 ^ 2))
        CI1 <- confint(m1)[1, ]

        if(CI0[1] < 2 && CI0[2] > 2) n0 <- n0 + 1
        if(CI1[1] < 2 && CI1[2] > 2) n1 <- n1 + 1
}
c(n0, n1) / 1000

## [1] 0.954 0.466
```

Non-constant variance of error terms

- Problem: if error terms do not have constant variance (*heteroscedasticity*), again everything could go wrong.
- Possible treatment:
  - Transform the response variable, e.g., $\sqrt{Y}$, $log(Y)$, ...
  - Weighted least squares (if you know a reasonable set of weights)

Example: More residual plot

```
set.seed(1)
x <- rnorm(1000, mean = 3, sd = 0.5)
y <-  2 + x * 10 + rnorm(1000, sd = x)
fit1 <- lm(y ~ x)
plot(fit1, 1)
```



Example: More residual plot

```
summary(fit1)$coefficients

##              Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)  1.766406  0.5898077  2.994884  2.813248e-03
## x           10.062904  0.1941104 51.841139 2.264075e-285

# use inverse variance as the weight
fit2 <- lm(y ~ x, weights = 1/(x^2))
summary(fit2)$coefficients

##              Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)  1.942002  0.4965033  3.911357  9.798986e-05
## x           10.003749  0.1744350 57.349439 3.831311e-318
```

The weighted least square correction is only FYI.

Collinearity

- Problem: Some predictors are correlated, which can reduce the accuracy of the $\hat{\beta}_j$.
- Intuitively, it is difficult to separate effects of individual predictors if they are correlated.
- Possible treatment: read the textbook about *variance inflation factor*.

Outliers

- Problem: some points have very different y_i than $\hat{y}_i$, which can change RSE and R^2 significantly.
- Example: regression of height on weight using a dataset containing Manute Bol ($y = 7'7''$, $x = 210$ lb).
- Possible treatment: double check your data is correct.

High-leverage points

- Problem: some points have very unusual x_i, which can change the regression line significantly.
- Example: regression of height on weight using a dataset containing the Hulk ($y = 7'6''$, $x = 1150$ lb).
- High-leverage points are hard to eyeball in multivariate regression.
- Read more in textbook about *leverage statistic*.
- Possible treatment: double check your data is correct.