36-617: Applied Linear Models

Regression Basics

Brian Junker
132E Baker Hall
brian@stat.cmu.edu

Monday is a holiday (Labor Day)

- No class, no office hours
- HW01 will be due Tuesday evening
- Since there are no office hours Monday, I will hold an extra office hour:
  - Tuesday, 11:30am-12:30pm
  - 132E Baker Hall

Reading, HW & TA Office Hours

- Reading:
  - Sheather Ch's 1-2 – for this week!
  - Handouts in week02 folder in Canvas files area – for next week!
    - I will not cover everything in the chapters
    - You will need to read & try some things on your own!
- HW01 out later today; due next Tue at midnight
  - Several "technical" exercises (math, R, data analysis, thinking)
  - Normally HW due on Mondays, but this Monday is a holiday…
- TA (Lorenzo – lthomasel@andrew.cmu.edu) Office hours
  - 8:45-9:45 Mondays
  - 12-1 Thursdays
  - Location TBA
  - *Let's see if the Monday office hour works…*

Outline

- The Linear Regression Model
  - Lazy formulation
  - Long formulation
  - Matrix formulation
- Least Squares and Maximum Likelihood
- Distribution of Estimated Coefficients
- SST = SSreg + RSS, R^2, Anova Table
- Confidence & Prediction Intervals for y's
- Example….

The Linear Regression Model

A lazy formulation

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

- *y* is the response variable, a.k.a. dependent variable
- β_0 = intercept; $\beta_1, \ldots, \beta_p$ = coefficients/slopes/effects
- $X_1, X_2, \ldots, X_p$ also have many equivalent names:
  - predictors / Independent variables / Regressors / Covariates
- ϵ is "error" (better: random deviation from mean)
- Quick, easy hand-waving, emphasizes functional form
- Easy to replace abstract letters with meaningful words…

```
(kid.score) = (intcpt) + (coef1)(mom.hs) +
                         (coef2)(mom.iq) + (error)
```

The Linear Regression Model

A longer, more accurate formulation

$$
\begin{aligned}
y_1 &= \beta_0 X_{10} + \beta_1 X_{11} + \cdots + \beta_p X_{1p} + \epsilon_1 \\
y_2 &= \beta_0 X_{20} + \beta_1 X_{21} + \cdots + \beta_p X_{2p} + \epsilon_2 \\
&\vdots \\
y_n &= \beta_0 X_{n0} + \beta_1 X_{n1} + \cdots + \beta_p X_{np} + \epsilon_n
\end{aligned}
$$

- $X_{i0} \equiv 1$, a column of 1's (for the intercept!)
- *n* = number of "units" or "cases";
- *k* = *p+1* = number of "predictors" or "covariates"
- $y_i = \beta_0 X_{i0} + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i, \ i = 1, \ldots n$

The Linear Regression Model

A matrix formulation…

- Let $X_i = (X_{i0}, \ldots, X_{ip})$ and $\beta = (\beta_0, \ldots, \beta_p)^T$; then

$$\begin{aligned} y_i &= \beta_0 X_{i0} + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i \\ &= X_i \beta + \epsilon_i \end{aligned}$$

- If we also stack $Y = (Y_1, \ldots, Y_n)^T$, $X = (X_1^T, \ldots, X_n^T)^T$, and $\epsilon = (\epsilon_1, \ldots \epsilon_n)^T$, we can write

$$Y = X\beta + \epsilon$$

- Has all the information of the Long Formulation
- Lends itself to compact formulas & computation

Linear Regression Model: Mean + Error Distribution

- In the model

$$y_i = X_i\beta + \epsilon_i, i = 1, \ldots, n$$

It is usual to assume $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

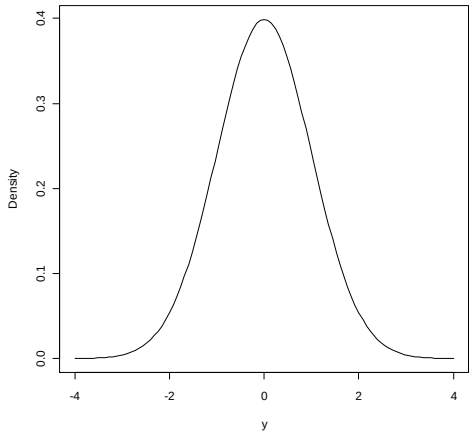
- We can write this as

$$\begin{aligned} y_i &\overset{indep}{\sim} N(\theta_i, \sigma^2), \ i = 1, \ldots, n \\ \theta_i &= X_i\beta = \beta_0 X_{i0} + \cdots \beta_p X_{ip} \end{aligned}$$

  - Each $y_i \in (-\infty, \infty)$ has some mean $\theta_i = E[y_i|X_i]$
  - Each θ_i has some linear structure
  - There is a statistical distribution N(*, σ^2) that describes unmodeled variation around $\theta_i = E[y_i|X_i]$

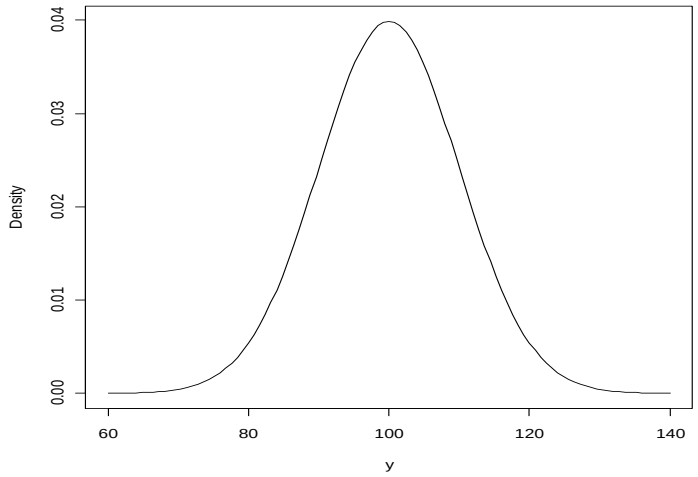
Aside: The Normal Distribution

- Y ~ N(0,1) iff $f(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$



- Y ~ N(μ,σ^2) iff $\frac{y-\mu}{\sigma} \sim N(0,1)$

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$



Least Squares for Simple Linear Regression, From the Model

- Starting with the simple linear regression model

$$\begin{aligned} y_i &\overset{indep}{\sim} N(\theta_i, \sigma^2), \; i = 1, \ldots, n \\ \theta_i &= E[y_i | X_i] = \beta_0 + \beta_1 X_i \end{aligned}$$

we see that the density for y_i is

$$\begin{aligned} f(y_i | X_i) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - \theta_i)^2}{2\sigma^2}} \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2}} \end{aligned}$$

Least Squares & Maximum Likelihood

- Independence allows us to write the likelihood as

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} f(y_i|X_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 X_i)^2\right\}$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 X_i)^2\right)$$

- The log-likelihood is then

$$\log L(\beta_0, \beta_1, \sigma^2)$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i-1}^{n}(y_i - \beta_0 - \beta_1 X_i)^2$$

Least Squares & Maximum Likelihood

- The MLE's $\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\sigma}^2$ are found by maximizing

$$\log L(\beta_0, \beta_1, \sigma^2)$$

$$= \quad -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i-1}^{n}(y_i - \beta_0 - \beta_1 X_i)^2$$

  - MLE's $\hat{\beta}_0,\ \hat{\beta}_1$ minimize the residual sum of squares

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\hat{e}_i^2$$

    for any positive σ^2.
  - The MLE $\hat{\sigma}^2$ will then just depend on the minimized RSS.

Least Squares & Maximum Likelihood

- After a little calculus,

$$\begin{aligned}
\hat{\beta}_0 &= \overline{y} - \hat{\beta}_1 \overline{X} \\
\hat{\beta}_1 &= \frac{\sum_i X_i y_i - n\overline{X}\overline{y}}{\sum_i X_i^2 - n\overline{X}^2} = \frac{\sum_i (X_i - \overline{X})(y_i - \overline{y})}{\sum_i (X_i - \overline{X})^2} = \frac{\mathsf{SXY}}{\mathsf{SXX}} \\
\hat{\sigma}^2 &= \frac{1}{n}\mathsf{RSS}
\end{aligned}$$

- It can be shown that $\hat{\sigma}^2$ is ***biased***: $E[\hat{\sigma}^2] \neq \sigma^2$; instead we usually use the ***unbiased*** estimator[1]

$$S^2 = \frac{1}{n-2}\mathsf{RSS}$$

(*2 = k = p+1 =* number of estimated coefs = df)

[1]See e.g. Weissberg (2013) *Applied Linear Regression*.

Distribution of Estimated Coefficients

- $\hat{\beta}_1 = \frac{SXY}{SXX} = \sum_i \frac{(X_i - \overline{X})}{SXX} y_i$, and hence
  - $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{SXX}\right)$
  - $\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{SXX}\right)\right)$
- And therefore[1]
  - $T_1 = \frac{\hat{\beta}_1 - \beta_1}{SE(\beta_1)} \sim t_{n-2}$, $SE(\beta_1) = S\sqrt{1/SXX}$
  - $T_0 = \frac{\hat{\beta}_0 - \beta_0}{SE(\beta_0)} \sim t_{n-2}$, $SE(\beta_0) = S\sqrt{1/n + \overline{X}^2/SXX}$

This gives rise to the usual confidence intervals and hypothesis tests (more to come. . .)

[1]See e.g. Weissberg (2013) *Applied Linear Regression*.

Partitioning Sums of Squares…

- It's easy to see that $y_i - \overline{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \overline{y})$
- If we square both sides and do a little algebra we get SST = SSreg + RSS, where
  - $SST(= SYY) = \sum_{i=1}^{n}(y_i - \overline{y})^2$
  - $SSreg = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$
  - $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

  (why isn't this obvious? – you will show it in HW01!)
- Moreover[1] SSreg is *independent* of RSS, and
  - $RSS/\sigma^2 \sim \chi^2_{n-2}$
  - When $\beta_1 = 0$, $SSreg/\sigma^2 \sim \chi^2_1$

[1]See e.g. Weissberg (2013) *Applied Linear Regression*.

From SST = SSreg + RSS, Anova Table

- $R^2 = \frac{SSreg}{SST} = 1 - \frac{RSS}{SST}$ = portion of variation in y due to (or explained by) $\hat{y}$
  - Also $R^2 = [Corr(X,y)]^2$ in the data (algebra!)
- When H_0: $\beta_1 = 0$ is true, $F = \frac{SSreg/1}{RSS/(n-2)} \sim F_{1,n-2}$
- The traditional Analysis of Variance table

Source of variation	Degrees of freedom (df)	Sums of squares (SS)	Mean square (MS)	F
Regression	1	$SSreg$	$SSreg/1$	$F = \frac{SSreg/1}{RSS/(n-2)}$
Residual	$n-2$	RSS	$RSS/(n-2)$	
Total	$n-1$	SST		

will later be generalized to $p>1$ predictors $X_1, \ldots, X_p$

Distribution of $\hat{y}_i$

- Combining

$$\begin{aligned}\hat{\beta}_0 &= \overline{y} - \hat{\beta}_1 \overline{X} \\ \hat{\beta}_1 &= \sum_{j=1}^{n} \frac{X_j - \overline{X}}{SXX} y_j\end{aligned}$$

with a little algebra, we can show

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = \sum_{j=1}^{n} \left[\frac{1}{n} + \frac{(X_i - \overline{X})(X_j - \overline{X})}{SXX} \right] y_j = \sum_{j=1}^{n} h_{ij} y_j$$

- Therefore

$$\begin{aligned}E[\hat{y}_i] &= \beta_0 + \beta_1 X_i = E[y_i | X_i] \\ \text{Var}[\hat{y}_i] &= \sigma^2 \left[\frac{1}{n} + \frac{(X_i - \overline{X})^2}{SXX} \right] = h_{ii}\sigma^2 \\ \hat{y}_i &\sim N\left(\beta_0 + \beta_1 X_i, \sigma^2 \left[\frac{1}{n} + \frac{(X_i - \overline{X})^2}{SXX} \right] \right) = N\left(\beta_0 + \beta_1 X_i, h_{ii}\sigma^2 \right)\end{aligned}$$

Confidence Interval for a point y* on the regression line

- Let X* be a new X value, and let y* be the new y value: $y^* = \beta_0 + \beta_1 X^* + \epsilon^*$
- A CI for the point $E[y^*|X^*] = \beta_0 + \beta_1 X^*$ *on the regression line* is

$$(\hat{y}^* - t \cdot SE(\hat{y}^*), \hat{y}^* + t \cdot SE(\hat{y}^*))$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X^*$ and $SE(\hat{y}^*) = \sqrt{h_{ii}} S$, and *t* is an appropriate cutoff (around 2 for a 95% interval, e.g.)

Prediction Interval for a new obs. y*

- Again let X* be a new value, and now consider predicting y* itself at that X*:

$$\hat{y}^*_{pred} = \hat{\beta}_0 + \hat{\beta}_1 X^* + \epsilon^*_{pred}$$

- Using the same sorts of calculations as before,

$$E[\hat{y}^*_{pred}] = E[\hat{y}^*] = \beta_0 + \beta_1 X^*$$

$$\text{Var}\,(\hat{y}^*_{pred}) = \text{Var}\,(\hat{\beta}_0 + \hat{\beta}_1 X^* + \epsilon_{pred}) = (h_{ii} + 1)\sigma^2$$

- So a prediction interval for y^*_{pred} would be

$$(\hat{y}^* - t \cdot SE(\hat{y}^*_{pred}), \hat{y}^* + t \cdot SE(\hat{y}^*_{pred})),$$

where $SE(\hat{y}^*_{pred}) = \sqrt{h_{ii} + 1}S$

Example…

Summary

- The Linear Regression Model
  - The Lazy, Long and Matrix Formulations
  - Least Squares and Maximum Likelihood
  - Distribution of Estimated Coefficients
  - SST = SSreg + RSS, R^2, Anova Table
  - Confidence & Prediction Intervals for y's
- HW01 out later today – due next Tues
- Read Ch 3 for next week
- TA Office hours?