Fitting Simple Linear Regresion Models

Math 430, Winter 2017

Predicting fuel economy

- **Task:** predict the fuel economy of a vehicle based on its weight
    - i.e. find $\widehat{\beta}_0$ and $\widehat{\beta}_1$

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

- **Approach:** minimize the residual sums of squares

$$RSS = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

- This is called least squares (LS) estimation

Linear models in R

`lm` is our workhorse function

```
mod <- lm(mpg ~ weight, data = mpg)
```

- The formula is of the form `response ~ predictor`
- The result is an object of class `lm`

```
names(mod)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

Linear models in R

You have a few options to the results

1. **Print**: print `mod` to see the estimated regression coefficients

2. **Summary**: `summary(mod)` displays the most useful information about the model

3. **Attributes**: extract the attribute of interest using the `$` operator

```r
summary(mod)
```

```
##
## Call:
## lm(formula = mpg ~ weight, data = mpg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7011  -3.3404  -0.5987   2.3588  16.0605
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 51.5871689  1.4835394   34.77   <2e-16 ***
## weight      -0.0098334  0.0005749  -17.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.723 on 287 degrees of freedom
## Multiple R-squared:  0.5048,	Adjusted R-squared:  0.5031
## F-statistic: 292.6 on 1 and 287 DF,  p-value: < 2.2e-16
```

Interpreting the slope

Interpreting the intercept

Making predictions

Once we have our estimated regression coefficients, $\hat{\beta}_0$ and $\hat{\beta}_1$, obtaining a prediction is easy.

Example predict the MPG for a car weighing 2,500 lbs

Making predictions

Once we have our estimated regression coefficients, $\hat{\beta}_0$ and $\hat{\beta}_1$, obtaining a prediction is easy.

Example predict the MPG for a car weighing 2,500 lbs

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1(2500)$$

Making predictions

Once we have our estimated regression coefficients, $\hat{\beta}_0$ and $\hat{\beta}_1$, obtaining a prediction is easy.

Example predict the MPG for a car weighing 2,500 lbs

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1(2500)$$

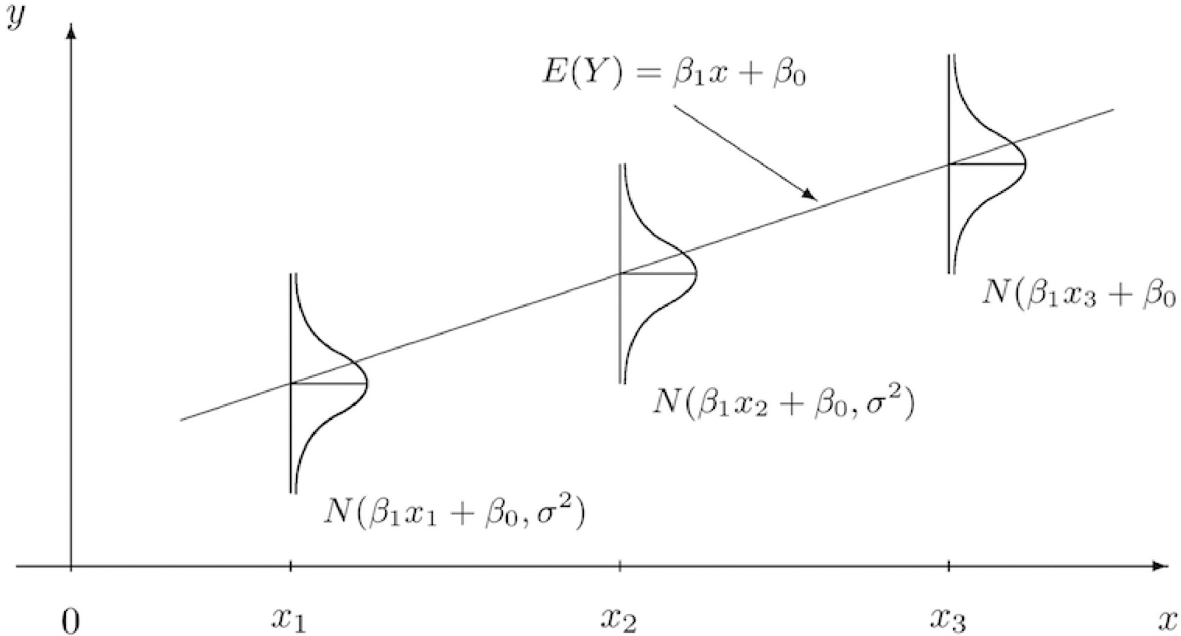
In R, we use the `predict` function

```
predict(mod, newdata = data.frame(weight = 2500))
```

```
##        1
## 27.00371
```

The full SLR model

- LS only assumes that there is a linear relationship between x and y
- Additional assumptions are needed to understand the uncertainty of our predictions
- The SLR model can be written in a few forms
    - $Y_i = \beta_0 + \beta_1 x_i + e_i$ where $e \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$
    - $Y_i \overset{iid}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$


y
$E(Y) = \beta_1 x + \beta_0$
$N(\beta_1 x_3 + \beta_0, \sigma^2)$
$N(\beta_1 x_2 + \beta_0, \sigma^2)$
$N(\beta_1 x_1 + \beta_0, \sigma^2)$
0
x_1
x_2
x_3
x

Regression assumptions

1. **Linearity**: $E(Y|X = x_i) = \beta_0 + \beta_1 x$
2. **Independence**: $e_1, \ldots, e_n$ are independent
3. **Constant error variance**: $Var(e_1) = \cdots = Var(e_n) = \sigma^2$
4. **Normal error terms**: $e \sim \mathcal{N}(0, \sigma^2)$

ML estimation

We cannot obtain an estimate of σ^2 through LS, so instead we can use maximum likelihood (ML)

To do this, we simply maximize the likelihood function

$$L(\beta_0, \beta_1, \sigma) = \prod_{i=1}^{n} f(y_i | x_i, \beta_0, \beta_1, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(y_i - \beta_0 - \beta_1 x_i)/2\sigma^2}$$

Idea: finding the values of β_0, β_1, and σ that make our data most likely

ML estimation

It's often easier to work with the log likelihood

$$\ell(\beta_0, \beta_1, \sigma) = \log L(\beta_0, \beta_1, \sigma) = \sum_{i=1}^{n} \log(\sigma) - \frac{1}{2}\log(2\pi) - (y_i - \beta_0 - \beta_1 x_i)^2/2\sigma^2$$

Taking partial derivatives we find

$$\frac{\partial \ell}{\partial \beta_0} = \frac{1}{\sigma^2} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\partial \ell}{\partial \beta_1} = \frac{1}{\sigma^2} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i$$

$$\frac{\partial \ell}{\partial \sigma} = \frac{-n}{\sigma} - \frac{1}{\sigma^3} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 = \frac{1}{\sigma^2}\left(n\sigma^2 - \sum_{i=1}^{n} e_i^2\right)$$

ML estimation

Setting the derivatives to 0 and solving yields

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} \qquad \widehat{\beta}_1 = \frac{SXY}{SXX} \qquad \widehat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n}$$

- $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the LS estimates
- The above estimate of σ^2 is biased, so we must make an adjustment to obtain an unbiased estimator

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

Properties of our estimators

- $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are **unbiased estimates** of β_0 and β_1
- $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the *best linear unbiased estimates* (BLUE); that is, they have the smallest variance of all linear unbiased estimates
- $\widehat{\sigma}_\varepsilon$ is an unbiased estimate of σ_ε