Jonathan Go

STA 371H Notes 1/28/15

Reminders from Last Class:
-Statistical Models are used to partition variation
-Takes observation Y_i and decomposes it into two parts:
$\hat{Y}_i + e_i$ = systematic + unpredictable = fitted/model + residual
-Can use Standard Deviation: compare SD of $\hat{Y}_i$ to SD of e_i

Dummy Variable:
-aka "baseline/offset form"
-choose a baseline (B_0) and then compare everything to that baseline
-Indicator Variable:
e.g. two groups
X_i=1 if case i is in group 1
X_i=0 if case i is in group 2
$Y_i = B_0 + B_1 X_i + e_i$
If $X_i = 0$, $Y_i = B_0 + e_i$
(group mean is B_0=regression coefficient)
If $X_i = 1$, $Y_i = B_0$ + beta1 + e_i
X_i = dummy variable, B_0 = baseline/intercept, B_1 =slope/coefficient on the dummy variable
e.g. three groups
X_i1=1 if case i is in group 1, otherwise 0
X_i2=1 if case i is in group 2, otherwise 0
$Yi = B_0 + B1(Xi1) + B_2(X_i2) + e_i$
Group 0
$Yi = B_0 + e_i$
Group 1
$Yi = B_0 + B_1 + e_i$
Group 2
$Yi = B_0 + B_2 + e_i$

Jonathan Go

R Software:

-Change "mean" to "lm" to get baseline offset form of data

-Change "plot" to "lm" to fit line onto scatterplot

Fitted line + residual = original data

-Story 1: Plug In Prediction

Newx=c(#, #, #)

c="combined"

Yhat=_________

-Story 2: Summarizing the Trend (quantify rate of change=derive)

-Derivative of $\hat{Y}=B_1$

-Don't need the intercept because the baseline is only a baseline:

-Can mean nothing or anything or both

-Story 3: Taking the "X-ness" out of Y

-"Statistical Adjustment": adjust for something in the y-variable

-Residual (e_i)=y adjusted for x

$Y_i = B_0 + B_1X_1 + e_i$

$Y_i - B_0 - B_1X_1 = e_i$

e.g. used car salesman:

-plot residual vs miles:

-find the lowest residual for a given level of miles=best deal

(not necessarily the cheapest)

-find what it ought to be cost and then find the one that has the lowest residual

(how much the price is below that)

Use min(resid(model)) to find value

Use which.min(resid(model)) to find which one has that value

-Story 4: Quantify the Reduction in Uncertainty

Using the mean

Use Standard deviation:

-we're gonna guess the mean, but we're going to expect on average that we will miss by _______ amount

sd(pickup$price)

Using the model

Use residuals

sd(fitted(model1))

sd(resid(model1))

This quantifies the information content of the model: that is, how much our uncertainty in a truck's price is reduced by knowing its mileage, and how much remains in the residuals.

Don't forget to assign variables to store

Model1 = lm(…………….)

Jonathan Go

OLS=ordinary least squares

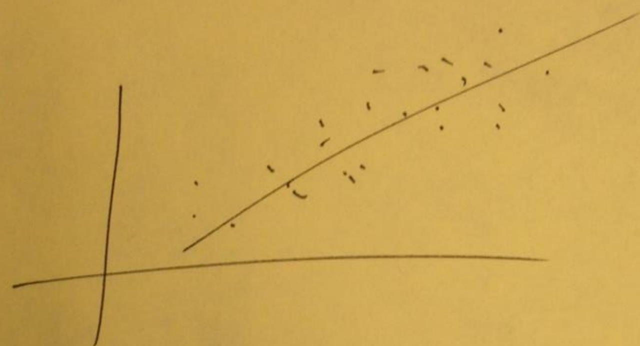
- Y_i=response
- X_i=predictor/feature/covariant/independent variable
    - ingest into model to make a prediction
- $Y_i=B_0 + B_1(X_i) + e_i$
- Created by Legendre-french mathematician)
    - Minimize the summation of e_i^2
    - Why the square? Instead of something else, e.g. the absolute value
        1) Punish the outliers, "how much does it hurt to miss"
        2) Normalize both negative and positive differences
        3) Easier to do calculus (take derivative and set=0)
        4) Special relationship between sum of square errors and
        5) Variance decomposition=Pythagorean theorem

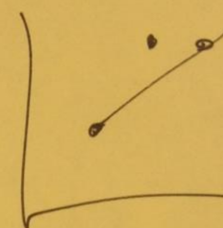
Jonathan Go

OLS

y_i : response

x_i : predictor





$$y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\hat{y}_i} + e_i$$

Legendre

$$\text{Minimize} \quad \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} (y_i - [\beta_0 + \beta_1 x_i])^2$$

① ~~Plugi~~ Plug-in prediction

② Summarizing the trend

$$\hat{y} = \beta_0 + \beta_1 x$$

$$\frac{d\hat{y}}{dx} = \beta_1$$

Jonathan Go

③ Taking the "x"-ness of y.

Statistical adjustment

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

$$y_i - \beta_0 - \beta_1 x_i = e_i$$

④ Quantifying the Reduction in Uncertainty