Modeling

Class 1:
Multiple Linear Regression

Vedaste Ndahindwa

31 Mar 2016

Linear Regression Line

- Population regression model:

$$\mu_{Y|\mathbf{X}}=\alpha+\beta_1X_1+\beta_2X_2+\ldots\beta_qX_q$$

- Fitted regression model:

$$Y=\alpha+\beta_1X_1+\beta_2X_2+\ldots\beta_qX_q+\varepsilon$$

- In these models, Y is a continuous response variable.

Multiple regression

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_k X_k + \varepsilon$$

(individual regression model)

- The value of Y is now a function of multiple covariates.

Model Framework

- $\mu_{Y|X1,X2,\ldots,Xk}$ is <u>linear</u> in $X_1, X_2, \ldots, X_k$
- The residuals are <u>homoscadastic</u> ($\sigma_{Y|X1,X2,\ldots,Xk}$ are constant)

- For fixed covariates, $X_1, X_2, \ldots, X_{k,}$ Y is <u>normally</u> distributed with
  - mean, $\mu_{Y|X1,X2,\ldots,Xk}$
  - standard deviation, $\sigma_{Y|X1,X2,\ldots,Xk}$

- Observations are <u>independent</u>

```
. regress sysbp age heartrate

      Source |       SS           df       MS      Number of obs   =        25
-------------+----------------------------------   F(2, 22)        =     68.88
       Model |  27386.2231         2  13693.1115   Prob > F        =    0.0000
    Residual |  4373.77576        22  198.807989   R-squared       =    0.8623
-------------+----------------------------------   Adj R-squared   =    0.8498
       Total |  31759.9988        24  1323.33329   Root MSE        =      14.1

------------------------------------------------------------------------------
       sysbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    1.95156   .3208844     6.08   0.000     1.286087    2.617034
   heartrate |  -.8748746   .1368827    -6.39   0.000    -1.158752   -.5909973
       _cons |   200.6998   21.68584     9.25   0.000     155.7261    245.6735
------------------------------------------------------------------------------
```

$$\hat{Y} = SBP = 200.7 + 1.95age + -0.875heartrate$$

Special case of “dummy” variables

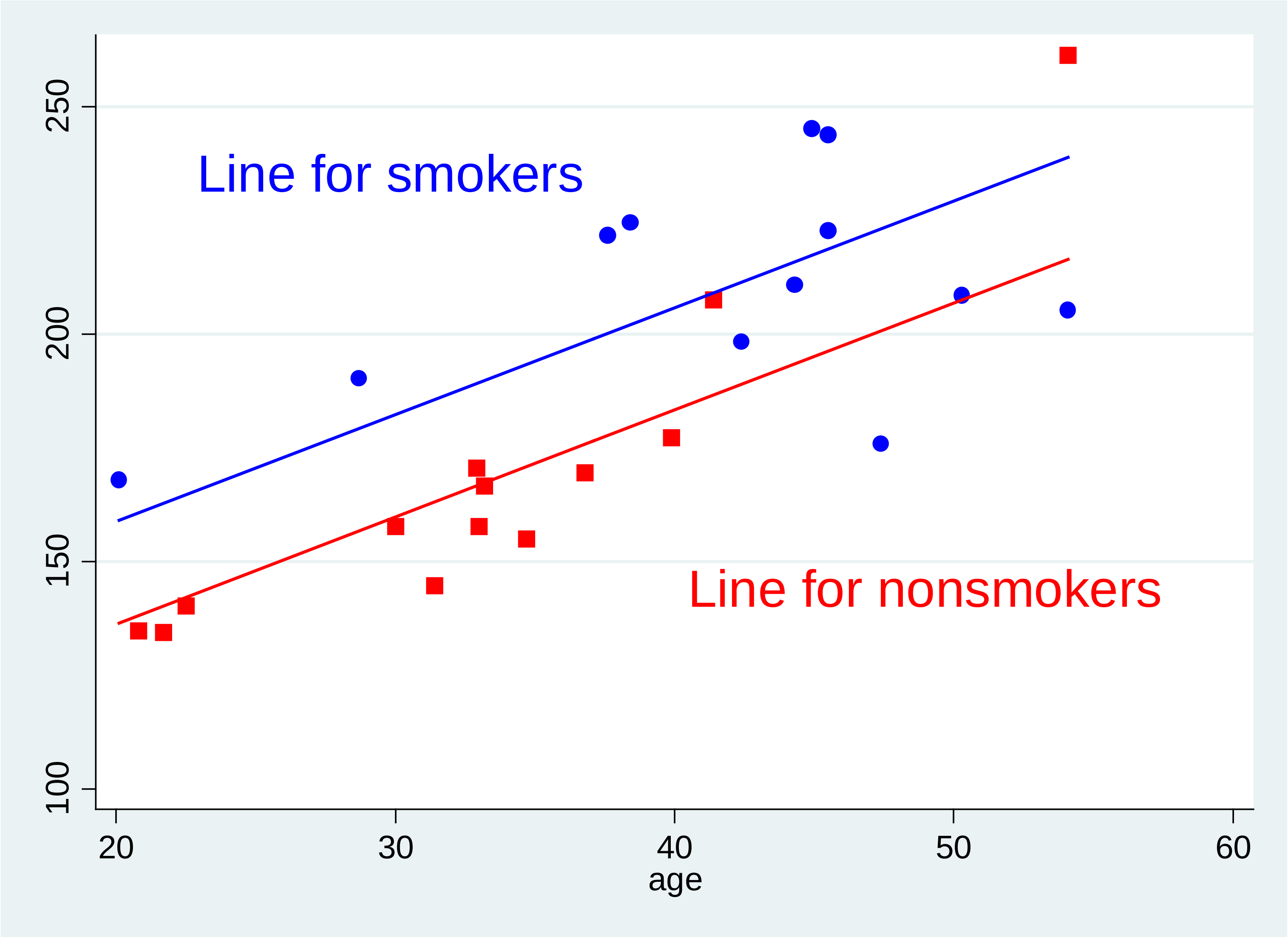
```
. regress sysbp age smoke

      Source |       SS           df       MS      Number of obs   =        25
-------------+----------------------------------   F(2, 22)        =     24.26
       Model |  21850.9395         2  10925.4698   Prob > F        =    0.0000
    Residual |  9909.05934        22  450.411788   R-squared       =    0.6880
-------------+----------------------------------   Adj R-squared   =    0.6596
       Total |  31759.9988        24  1323.33329   Root MSE        =    21.223

------------------------------------------------------------------------------
       sysbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    2.35224   .4808594     4.89   0.000     1.354998    3.349481
       smoke |   22.51094   9.394599     2.40   0.026     3.027734    41.99415
       _cons |   89.26089   17.04286     5.24   0.000     53.91616    124.6056
------------------------------------------------------------------------------
```

$$\hat{Y} = SBP = 200.7 + 2.35age + 22.5smoke$$

Dummy variables

- Take on two values
- Result in two linear models, one for when dummy=0 and one for when dummy=1

- For smoke =0
  - $SBP = 200.7 + 2.35age + 22.5 * 0$
  - $SBP = 200.7 + 2.35age$
- For smoke =1
  - $SBP = 200.7 + 2.35age + 22.5 * 1$
  - $SBP = 223.2 + 2.35age$



250
200
150
100
Line for smokers
Line for nonsmokers
20
30
40
50
60
age

What about interactions?

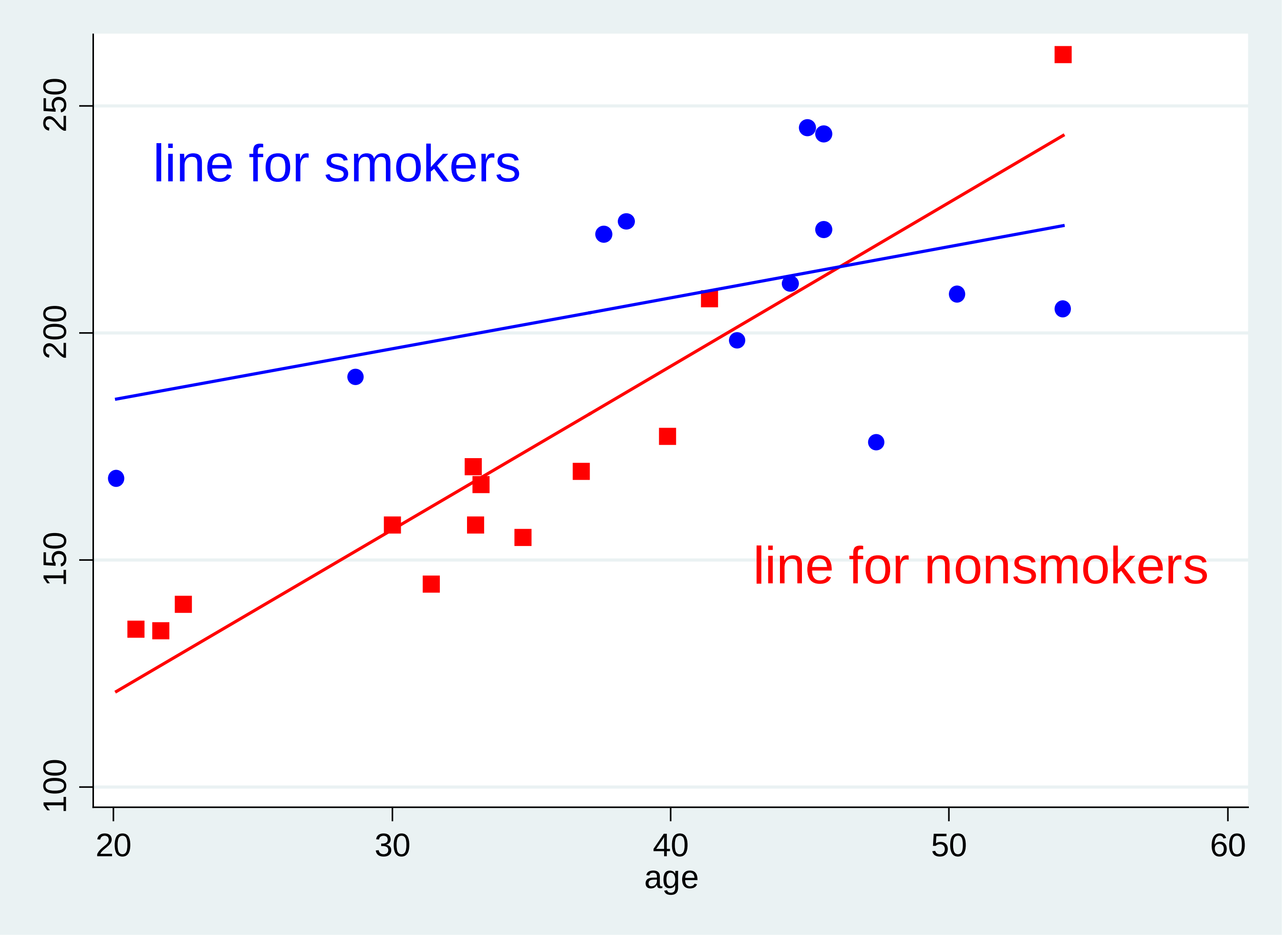
```
. generate age_smoke=age*smoke

. regress sysbp age smoke age_smoke

      Source |       SS           df       MS      Number of obs   =        25
-------------+----------------------------------   F(3, 21)        =     24.84
       Model |  24777.0793         3  8259.02643   Prob > F        =    0.0000
    Residual |  6982.91956        21  332.519979   R-squared       =    0.7801
-------------+----------------------------------   Adj R-squared   =    0.7487
       Total |  31759.9988        24  1323.33329   Root MSE        =    18.235

------------------------------------------------------------------------------
       sysbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   3.572543   .5830324     6.13   0.000     2.360061    4.785025
       smoke |   114.3092   31.98084     3.57   0.002     47.80141     180.817
   age_smoke |  -2.451291   .8263352    -2.97   0.007    -4.169749   -.7328331
       _cons |   48.67173    20.0412     2.43   0.024     6.993777    90.34968
------------------------------------------------------------------------------
```

$$\hat{Y} = SBP = 48.7 + 3.57age + 114.3smoke + -2.45age\_smoke$$

Dummy variables with interactions

- Take on two values
- Result in two linear models, one for when dummy=0 and one for when dummy=1

- For smoke =0
  - $SBP = 48.7 + 3.57age + 114.3 * 0 + -2.45 * age * 0$
  - $SBP = 48.7 + 3.57age$
- For smoke =1
  - $SBP = 48.7 + 3.57age + 114.3 * 1 + -2.45 * age * 1$
  - $SBP = 163.0 + 1.12age$



line for smokers
line for nonsmokers
250
200
150
100
20
30
40
50
60
age

Building models

- Look at predictor variables one at a time with outcomes to see what is significant.
- Look at interactions to see what is significant.

- Forward selection –
  1. Start with most significant predictor.
  2. Add next most significant predictor.
  3. Keep adding until there are no more significant predictors.
- Backward selection –
  1. Start with everything in the model.
  2. Drop least significant predictor.
  3. Keep dropping until only significant predictors remain.

Final model

What gets dropped?

- Collinearity – Two or more variables are highly correlated and essentially convey the same information.

Can non-significant things remain?

- If a variable is a not significant predictor of the outcome but changes (confounds) the relationship between another predictor and outcome, then it often will be kept in the model.