Linear Regression
(Multiple Regression Case)

Additional Considerations

- Have multiple possible explanatory variables
- Assume that explanatory variables are (roughly) independent
- Will need to select best subset of explanatory variables to use

- Interpretation of β: same as simple linear regression but when the condition that *all other predictors are fixed*

Exercises: US Crime Data

- Built-in data in package "MASS"
  - ?UScrime for detailed information
  - Aggregate data on 47 states of the USA for 1960
- Response: rate of crimes **y** (scaled)
- Fifteen possible explanatory variables
- Will want to choose best subset of these 15 variables for modeling crime rate

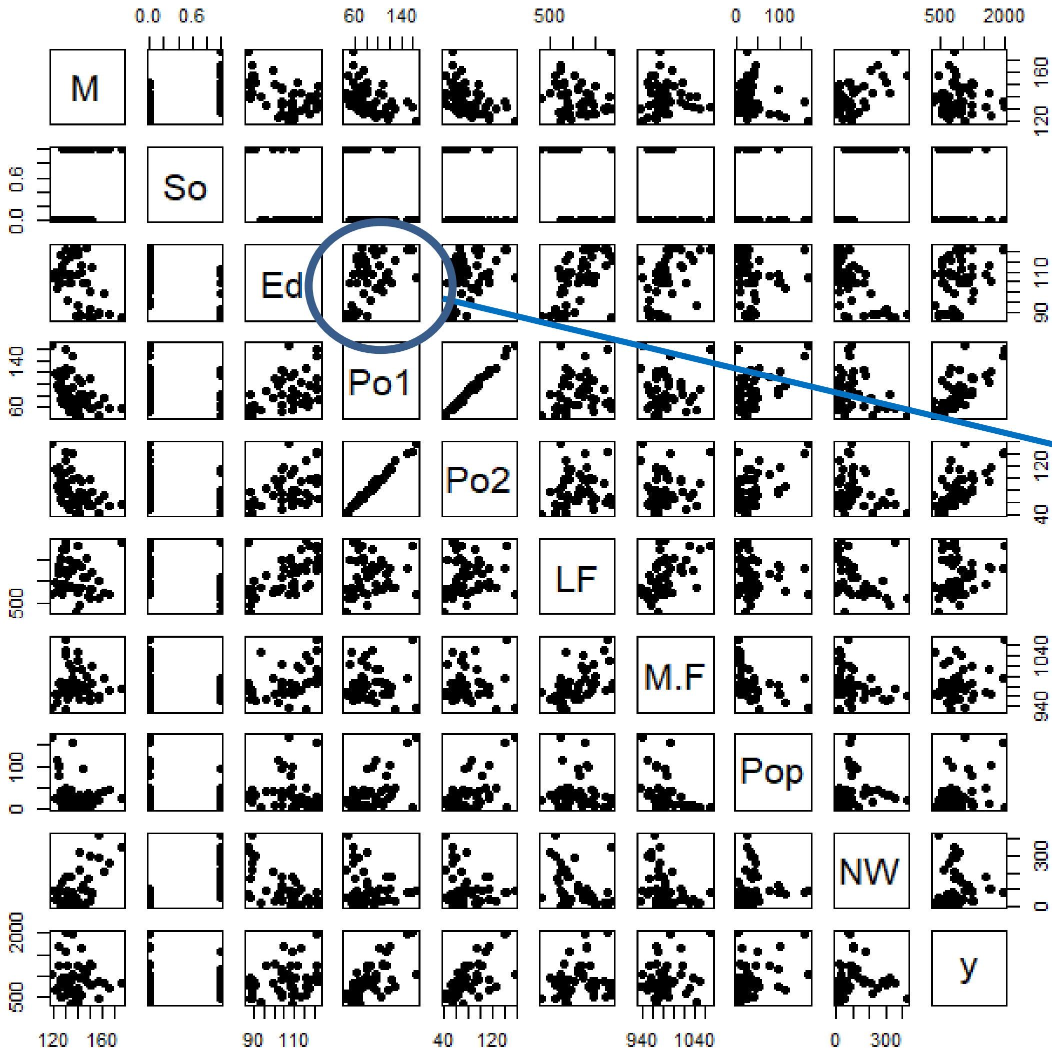
Exercises: US Crime Data

Aggregate data on 47 states of the USA for 1960

M: percentage of males aged 14–24.
So: indicator variable for a Southern state.
Ed: mean years of schooling.
Po1: police expenditure in 1960.
Po2: police expenditure in 1959.
LF: labour force participation rate.
M.F: number of males per 1000 females.

Pop: state population.
NW: number of non-whites per 1000 people.
U1: unemployment rate of urban males 14–24.
U2: unemployment rate of urban males 35–39.
GDP: gross domestic product per head.
Ineq: income inequality.
Prob: probability of imprisonment.
Time: average time served in state prisons.
y: rate of crimes.

Exercise: Visual Inspection



Pairwise scatter plot for first 10 variables.

e.g.,
X-axis: Po1
Y-axis: Ed

Exercise: Visual Inspection

- What does this plot tell us about relationships among the various predictors?
- What does the plot tell us about possible predictors for crime rate?

Multicollinearity

Problems with highly correlated predictors:

- Model is more complicated to interpret
- **Predictors confound each other**
- Variance estimates will be larger
- Predictions will be less reliable
- Want predictors to not be highly dependent on each other
- Can **not** be detected from diagnostics plots

Simulation studies: Why multicollinearity matters?

- Use drinking data from chapter6
- Download data SimData_multicoll.csv

- What happens if two variables are <u>perfectly correlated</u>? (i.e., corr=1)
  1. Generate a new variable **alcohol2**=3*alcohol
  2. Run a regression on **cirrhosis** with **alcohol** and **alcohol2**
  3. Check the result

- What happens if two variables are <u>highly correlated</u>?
  1. Generate a new variable **alcohol3**=3*alcohol + e, e ~N(0,0.1)
  2. Check their correlation
  3. Run a regression on **cirrhosis** with **alcohol** and **alcohol3**
  4. Run a simple regression, separately
  5. Compare results

Simulation studies:
Why multicollinearity matters?

(i) Regression with perfectly correlated predictors

```
lm.multicol1 = lm(cirrhosis~alcohol+alcohol2, data=drinking
.new)

summary(lm.multicol1)


##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5.9958     2.0977  -2.858   0.0134 *
## alcohol       1.9779     0.2012   9.829  2.2e-07 ***
## alcohol2          NA         NA      NA       NA
## ---
```

Simulation studies: Why multicollinearity matters?

(i) Regression with highly correlated predictors

```
lm.multicol2 = lm(cirrhosis~alcohol+alcohol3, data=dr
inking.new)

summary(lm.multicol2)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.99254    2.21610  -2.704   0.0192 *
## alcohol      1.92351    6.42640   0.299   0.7698
## alcohol3     0.01814    2.14085   0.008   0.9934
## ---
## F-statistic: 44.59 on 2 and 12 DF,  p-value: 2.783e-06
```

Checking for High Correlation

- Pairwise scatter plot for correlation between pairs of variables
- Use variance inflation factors (VIFs)
  - vif() in package “car” for calculating VIFs
- $VIF_j = \frac{1}{1-{R_j}^2}$
- **VIF_j>10 means at least 90% of x_j explained by other predictors**

Exercise: US Crime Data

- Fit **y** as a function of all the other variables and obtain the VIF values
- Predictors exceeding the VIF cutoff of 10?
- Omit the predictor **with largest VIF and refit**

  Do not remove variables all at once!

  Remove sequentially
- Any terms with VIF above the cutoff now?
- Which terms seem to be significant in this model?
- Any noticeable issues in the diagnostics?

R output

```
lm.crime3=lm(y~.-Po2-GDP, data=UScrime)

summary(lm.crime3)
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6041.0176  1515.7345  -3.986 0.000351 ***
## M               8.4035     4.0896   2.055 0.047879 *
## So             35.2894   143.7092   0.246 0.807543
## Ed             18.5920     5.9820   3.108 0.003861 **
## Po1            10.5094     2.1766   4.828 3.06e-05 ***
## LF             -0.1280     1.3924  -0.092 0.927317
## M.F             2.0125     2.0107   1.001 0.324141
## Pop            -0.6822     1.2761  -0.535 0.596494
## NW              0.1391     0.6048   0.230 0.819502
## U1             -5.7484     4.1469  -1.386 0.174980
## U2             18.0736     8.0840   2.236 0.032251 *
## Ineq            6.0732     1.7917   3.390 0.001829 **
## Prob        -4517.0792  2160.3360  -2.091 0.044315 *
## Time           -0.5337     6.6346  -0.080 0.936366
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 207.9 on 33 degrees of freedom
## Multiple R-squared:  0.7927, Adjusted R-squared:  0.711
## F-statistic: 9.707 on 13 and 33 DF,  p-value: 7.32e-08
```

Model Hypothesis Test (F-test in multiple linear regression)

- **Null Hypothesis:**
  - ✓ **Conceptually:** The multiple linear regression model does ***not*** fit the data better than the baseline model.
  - ✓ **Conceptually:** There is no linear relationship between x's and y
  - ✓ **Statistically:** $\beta_1 = \beta_2 = \ldots = \beta_p = 0$

- **Alternative Hypothesis:**
  - ✓ **Conceptually:** The multiple linear regression model does fit the data better than the baseline model.
  - ✓ **Conceptually:** There is linear relationship between some of x's and y
  - ✓ **Statistically:** at least one β is non zero

Individual Term Hypothesis Test (t-test for each predictor)

For example, for j-th predictor x_j

- **Null Hypothesis:**
  - ✓ **Conceptually:** There is no linear relationship between x_j and y
  - ✓ **Statistically:** $\beta_j = 0$
- **Alternative Hypothesis:**
  - ✓ **Conceptually:** There is linear relationship between x_j and y
  - ✓ **Statistically:** $\beta_j \neq 0$

Model Selection

- This is an "unsolved" problem in statistics: there are <u>no magic procedures to get you the "best model."</u>
- **Two approaches** in our class
  - Automatic selection
    - based on significance of the predictors
  - Best subset approach
    - based on penalized goodness of fit measure (simple but good prediction model)
- Depending on the goal of the study

Model Selection: Automatic Variable Selection

- Through algorithms that **pick the variables to include/ remove in your regression model**

  - **Forward Selection** -- start with no terms, sequentially add significant terms
  - **Backward Selection** -- start with all terms, sequentially remove insignificant terms
  - **Stepwise Selection** – start with no terms, alternate between forward and backward steps

- Final model will depend on significance level (threshold for enter/removal) that we set for algorithms
- “Greedy” search - always take the biggest jump (up or down)

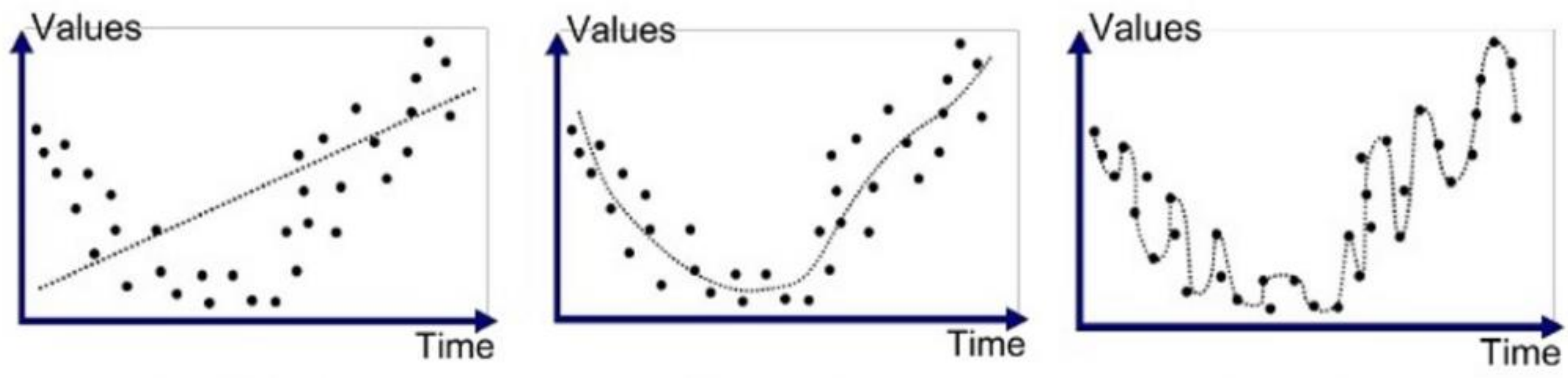
Model Selection:
Automatic Variable Selection

```
##
##                               Stepwise Selection Summary
## -------------------------------------------------------------------------------------
##                       Added/                    Adj.
## Step    Variable     Removed     R-Square     R-Square     C(p)        AIC         RMSE
## -------------------------------------------------------------------------------------
##    1       Po1       addition       0.473        0.461    40.9230    668.3155    283.9259
##    2       Ineq      addition       0.580        0.561    25.8080    659.5957    256.1874
##    3        Ed       addition       0.666        0.642    14.2270    650.9145    231.3136
##    4        M        addition       0.700        0.672    10.6880    647.7503    221.5397
##    5       Prob      addition       0.738        0.706     6.7180    643.4641    209.7205
##    6        U2       addition       0.766        0.731     4.2710    640.1661    200.6899
## --------------------------------------------------------------------------------------
```

- Final model from stepwise selection
  - Y ~ Po1 + Ineq + Ed + M + Prob + U2
- Additional information – R-square, Adj. R-square, C(p) etc.

Model Selection: Best Subset Approach

- **Compare all possible models** using a set of predictors, and displays **the best-fitting models** that contain one predictor, two predictors, and so on.
- Search all possible models (not “Greedy”) thus computation is heavy (takes longer time in R)



Underfitted Good Fit/Robust Overfitted

Model Selection: Best Subset Approach

- Search all possible models (not "Greedy") thus computation is heavy (takes longer time in R)

```
##                  Best Subsets Regression
## ------------------------------------------------------------
## Model Index    Predictors
## ------------------------------------------------------------
##     1         Po1
##     2         Po1 Ineq
##     3         Ed Po1 Ineq
##     4         M Ed Po1 Ineq
##     5         M Ed Po1 Ineq Prob
##     6         M Ed Po1 U2 Ineq Prob
##     7         M Ed Po1 U1 U2 Ineq Prob
##     8         M Ed Po1 M.F U1 U2 Ineq Prob
##     9         M Ed Po1 M.F Pop U1 U2 Ineq Prob
##    10         M So Ed Po1 M.F Pop U1 U2 Ineq Prob
##    11         M So Ed Po1 M.F Pop NW U1 U2 Ineq Prob
##    12         M So Ed Po1 LF M.F Pop NW U1 U2 Ineq Prob
##    13         M So Ed Po1 LF M.F Pop NW U1 U2 Ineq Prob Time
## ------------------------------------------------------------
```

Model Selection: Best Subset Approach

- For each candidate, we compare
  - **Adjusted R-square**
  - **AIC, BIC, SBC..**
  - **Mallow's Cp**

- **Goodness of fit + model simplicity**
  - Model with small errors but as simple as possible

- Compare different criteria and choose one. Sometimes the results do not point to one best model and your judgment is required.
- Predictors in the final model are always significant? (**FALSE**)

Adjusted R^2

- When **more variables** are added, R^2 values always **increase**
  - R^2 as a criterion, "optimum" is to take the biggest model
  - When p=n, R^2 =1 (perfect fit but not meaningful)

- Impose penalty for larger model (large p) and define adj- R^2

$$\overline{R}^2 = 1 - \frac{(n-i)(1-R^2)}{n-p}$$

n = the number of observations

***p* = the number of parameters in the model**

- This value will **not necessarily increase as additional terms are introduced into the model**, thus we want a model with the maximum adjusted R^2
- **Bigger the better**

AIC, BIC, SBC ..

- Penalized-likelihood criteria

$$AIC(M) = -2 \log L(M) + 2 \cdot p$$

$$BIC(M) = -2 \log L(M) + \log n \cdot p$$

- **$-2 \log L(M)$:** approximates regression error (model performace)
- **$2 \cdot p$** or **$\log n \cdot p$:** penalty term - proportional to p (model complexty)

- **Smaller the better**

Mallows' C_p

- Mallows' C_p is a simple indicator of effective variable selection

$$C_p = \frac{SSE_p}{S^2} - N + 2P,$$

- Again, penalized goodness-of-fit measure
- Look for models with **$C_p \le p$**, where *p* equals the number of predictors in the model, including the intercept.
  - Mallows **recommends choosing the first (fewest variables or simpler) model where C_p approaches p**.

Part of R output – Best Subset Approach

Bigger the better

Cp<=P

Smaller the better

```
## -----------------------------------------------------------------------------------------------------------
-------------------------------
##                              Adj.        Pred
## Model     R-Square     R-Square     R-Square       C(p)          AIC       SBIC       SBC             MSEP
## -----------------------------------------------------------------------------------------------------------
-------------------------------
##   1        0.4728       0.4611       0.3926     40.9229     668.3155        NA   673.8659     3789009.5774
##   2        0.5803       0.5612       0.4856     25.8077     659.5957        NA   666.9963     3086424.3129
##   3        0.6656       0.6423       0.5748     14.2266     650.9145        NA   660.1652     2517546.1300
##   4        0.7004       0.6719       0.6089     10.6882     647.7503        NA   658.8512     2310597.8418
##   5        0.7379       0.7060       0.6412      6.7180     643.4641        NA   656.4151     2071865.4454
##   6        0.7659       0.7307       0.6662      4.2708     640.1661        NA   654.9673     1898463.0711
##   7        0.7738       0.7332       0.6622      5.0024     640.5387        NA   657.1900     1882110.8120
##   8        0.7888       0.7444       0.6676      4.6158     639.3151        NA   657.8166     1804845.5315
##   9        0.7913       0.7405       0.6506      6.2296     640.7719        NA   661.1235     1833665.3821
##  10        0.7923       0.7347       0.6362      8.0554     642.5250        NA   664.7267     1876172.0917
##  11        0.7926       0.7274       0.6118     10.0155     644.4681        NA   668.5201     1929019.7157
##  12        0.7927       0.7195       0.5715     12.0065     646.4553        NA   672.3574     1986931.4697
##  13        0.7927       0.7110       0.5175     14.0000     648.4461        NA   676.1983     2048621.2922
## -----------------------------------------------------------------------------------------------------------
-------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##
```

Example: Stepwise Selection

- Start with all the predictors and significance levels of .1 for adding and for retaining terms
- What is the final model?
- Amount of variation in crime rate described by model?
- Problems in the diagnostics for this model?
- **Stepwise can handle highly correlated variables** (usually one of them has large p-value)

Exercise: Forward Selection

- Use forward selection and entry significance level of .1
- Compare steps of the selection process
- What is the final model?
- Amount of variation in crime rate described by model?
- Problems in the diagnostics for this model?

Exercise: Backward Selection

- Use backward selection and significance level of .1 for keeping terms
- Compare steps of the selection process
- What is the final model?
- Amount of variation in crime rate described by model?
- Problems in the diagnostics for this model?

Exercise: Selection through Best Subset Approach

- Best model with adj. R^2 criterion
  - Bigger the better
- Best model with AIC criterion
  - Smaller the better
- Best model with Mallow's Cp
  - Model with $\mathbf{C_p} \leq \boldsymbol{p}$ with the fewest variables (i.e., smallest p satisfying $C_p \leq p$)
- Compare selected models from each criterion

Regression with categorical variables

- Coding of categorical variables
  - 0/1 (this is **reference cell coding**): **default**
  - -1/1 (**deviations from means coding**)

- ANOVA is a special case of linear regression model with only categorical variables
- Interpretation of coefficient should be different
  - One unit increase in x … does not make sense for categorical variable
- Estimated coefficient represents difference between **Reference (a group coded as 0)** vs. **Comparison group (the other group coded as 1)**

Example with categorical variable: professor.csv

- **SALARY: continuous response**
- **Gender: categorical predictor**
- **TIME/ CITS/ PUBS: continuous predictors**

(1) SALARY ~ Gender + PUBS **(w/out interaction)**

(2) final ~ Gender + PUBS + Gender*PUBS **(w/ interaction)**

- R automatically handles factor variable (with reference cell coding) and we can check which group is set as a reference group (coded as 0) from the output
- If interaction term is significant, it implies: **the effect of PUBS differs depending on Gender**

Example with categorical variable: w/out interaction

```
summary(lm(SALARY~Gender+PUBS, data=professor)) ##

##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 41911.98    2353.34  17.810  < 2e-16 ***
## GenderMale   7143.82    2533.25   2.820  0.00653 **
## PUBS          470.28      90.42   5.201 2.61e-06 ***
## ---
```

- **Gender=Female (reference group - coded as 0)**

$$\hat{Y} = 41911.9 + 470.28*\text{PUBS}$$

- **Gender=Male (comparison group - coded as 1)**

$$\hat{Y} = (41911.9+7143) + 470.28*\text{PUBS}$$

- Interpretation: On average, males earn 7143 more than females do when the number of publications is same

Example with categorical variable: w/ interaction

```
summary(lm(SALARY~Gender*PUBS, data=professor)) ##

## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         47680.3     2969.8  16.055   <2e-16 ***
## GenderMale          -1998.6     3937.9  -0.508    0.614
## PUBS                  102.1      152.2   0.671    0.505
## GenderMale:PUBS       535.9      183.6   2.918    0.005 **
```

- Decide to include interaction -> include main effects
- **Gender=Female (reference group - coded as 0)**

$$\hat{Y} = 47680.3 + 102.1 * \text{PUBS}$$

- **Gender=Male (comparison group – coded as 1)**

$$\hat{Y} = (47680.3 - 1998.6) + (102.1 + 535.9) * \text{PUBS}$$

Regression with categorical variables: More than 2 levels

- What if a categorical variable has more than two levels in it? E.g., A1, A2, A3
- Same manner – one reference group (e.g., A1) and two estimates are **expected differences between (A2 vs. A1) and (A3 vs. A1)**
- Example with ToothGrowth data (used in ANOVA)

Regression with categorical variables: ToothGrowth.csv

```
## 'data.frame':    60 obs. of  3 variables:
##  $ Toothlength: num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ Supplement : Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2
##  $ Dose       : Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1
```

```
summary(lm(Toothlength ~ Dose, data=tooth))
```

```
##
## Call:
## lm(formula = Toothlength ~ Dose, data = tooth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6000 -3.2350 -0.6025  3.3250 10.8950
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.6050     0.9486  11.180 5.39e-16 ***
## Dose1         9.1300     1.3415   6.806 6.70e-09 ***
## Dose2        15.4950     1.3415  11.551  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With log transformation

- Additive change
  - ✓ Same amount of change in Y regardless of x values
- Multiplicative change
  - ✓ Change of Y depends on x values

- We expect to see β increase in log(Y) with one unit increase in X (-> hard to get what it means)
- Important to interpret with **original scale of Y**
  - Meaning of positive or negative β
  - Multiplicative change instead of additive change