MSDS 596 Regression & Time Series

Lecture 07 Transformation and Model Selection

Department of Statistics
Rutgers University

Oct 31, 2022

1. Recall: Checking Error Assumption
2. Predictor multicollinearity
3. Model selection
   - Testing-based procedures

Recall: Checking Error Assumption

The estimation and inference from the regression model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ depend on several assumptions, including that the errors have distribution

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}).$$

Potential problems with the error assumption include

- Constant variance
- Normality
- Correlation in errors

Recall: Checking Error Assumption

The estimation and inference from the regression model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ depend on several assumptions, including that the errors have distribution

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}).$$

Potential problems with the error assumption include

- Constant variance – Transformations of Y; Box-Cox;
- Normality – QQ-plot; Shapiro-Wilk test;
- Correlation in errors

Correlated Errors

- This pattern is often not easily seen from the residual plot directly;
- Can plot successive pairs of residuals;
- Durbin-Watson test can be used to check the autocorrelations (`dwtest()` in R). Will see more of this in time series analysis.

Predictor Multicollinearity

- Multicollinearity arises when the predictors being considered for the regression model are highly correlated among themselves.
- An extreme example.
  - True relationship $Y = X_1 + \epsilon$. And $X_1 = 2X_2$.
  - Fit $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$.
  - Many exactly equivalent solutions

$$Y = X_1 + \epsilon \iff Y = 2X_2 + \epsilon \iff Y = 0.5X_1 + X_2 + \epsilon \iff \dots$$

  - For given data, many solutions have the same SSE, hence no unique solution to the coefficient $\boldsymbol{\beta}$;
  - In fact, $\boldsymbol{X}^T\boldsymbol{X}$ is singular and does not have inverse.
- A realistic example.
  - If $X_2 = X_1 + e$ where e is small (X_1 and X_2 are highly correlated).
  - In the data, the two predictors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ have large sample correlations.
  - Many solutions are roughly equally good, resulting in large standard errors of the estimated parameters (one is extremely unsure which one is the correct solution).

Multicollinearity

Multicollinearity might NOT affect prediction accuracy. However, there are other symptoms:

- Large changes in the estimated regression coefficients when a predictor variable is added, deleted, or altered by a small amount;
- Estimated regression coefficients have signs that are opposite of that expected from theoretical considerations or prior experience;
- The estimated coefficients of important explanatory variables are not significant;
- Wide confidence intervals for the regression coefficients of important explanatory variables;
- All coefficients are not significant, but the F-statistic is highly significant.

Multicollinearity

Diagnosing multicollinearity:

- Large correlation coefficients in the sample correlation matrix indicate strong pairwise collinearity between predictors;
- Examine the eigenvalues of $\boldsymbol{X}'\boldsymbol{X}$, $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$.
  - zero eigenvalues denote exact collinearity;
  - the presence of a few small eigenvalues indicates multicollinearity.
- The condition number of a matrix measures the relative sizes of its eigenvalues, and is defined as

$$\kappa = \sqrt{\frac{\lambda_1}{\lambda_p}}.$$

  $\kappa \geq 30$ is considered large.
- Variance Inflation Factor (VIF)

Variance Inflation Factor (VIF)

- Variance Inflation Factor (VIF) indicates how much the variance of an estimated β_j is inflated in comparison to the case that the predictors are not correlated.
- Denote R_j^2 the coefficient of determination when predictor X_j is regressed on all the other predictor X's in the model. Note that R_j^2 does not depend on Y.
- VIF for the j-th predictor is

$$\mathsf{VIF}_j = \frac{1}{1 - R_j^2}$$

which happens to be the j-th diagonal element of $[\text{corr}(\boldsymbol{X})]^{-1}$, the inverse of the correlation matrix (not covariance matrix) of $\boldsymbol{X}$.

Variance Inflation Factor (VIF)

VIF for the j-th predictor:

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}$$

- If $R_j^2 = 0$, then VIF$= 1$, i.e. X_j is linearly uncorrelated with all other predictors.
- Notice that

  $$Var\left(\hat{\beta}_j\right) = \sigma^2 \cdot \mathrm{VIF}_j \cdot \frac{1}{\sum_i \left(x_{ij} - \bar{x}_j\right)^2},$$

  hence the name "variance inflation".
- Rule-of-thumb: serious multicollinearity if
  - average VIF of the p variables $\gg 1$, or,
  - if maximum VIF > 10.

Example: seat position data

Car drivers like to adjust the seat position for their own comfort. Car designers would find it helpful to know where different drivers will position the seat depending on their size and age. Researchers at the HuMoSim laboratory at the University of Michigan collected data on 38 drivers. The dataset (`seatpos` from `faraway`) contains the following variables:

- `Age`: Age in years
- `Weight`: Weight in lbs
- `HtShoes`: Height in shoes in cm
- `Ht`: Height bare foot in cm
- `Seated`: Seated height in cm
- `Arm`: lower arm length in cm
- `Thigh`: Thigh length in cm
- `Leg`: Lower leg length in cm
- `hipcenter`: horizontal distance of the midpoint of the hips from a fixed location in the car in mm

Example: seat position data

```
> head(seatpos, 3)
  Age Weight HtShoes    Ht Seated  Arm Thigh  Leg hipcenter
1  46    180   187.2 184.9   95.2 36.1  45.3 41.3  -206.300
2  31    175   167.5 165.5   83.8 32.9  36.5 35.9  -178.210
3  23    100   153.6 152.2   82.9 26.0  36.6 31.0   -71.673

> round(cor(seatpos[,-9]), 2)
          Age Weight HtShoes    Ht Seated  Arm Thigh   Leg
Age      1.00   0.08   -0.08 -0.09  -0.17 0.36  0.09 -0.04
Weight   0.08   1.00    0.83  0.83   0.78 0.70  0.57  0.78
HtShoes -0.08   0.83    1.00  1.00   0.93 0.75  0.72  0.91
Ht      -0.09   0.83    1.00  1.00   0.93 0.75  0.73  0.91
Seated  -0.17   0.78    0.93  0.93   1.00 0.63  0.61  0.81
Arm      0.36   0.70    0.75  0.75   0.63 1.00  0.67  0.75
Thigh    0.09   0.57    0.72  0.73   0.61 0.67  1.00  0.65
Leg     -0.04   0.78    0.91  0.91   0.81 0.75  0.65  1.00
```

Example: seat position data - multicollinearity

Eigendecomposition of $\boldsymbol{X}'\boldsymbol{X}$:

```
> x <- model.matrix(lmod)[, -1]
> e <- eigen(t(x) %*% x)
> round(e$val, 3)
[1] 3653671.363   21479.480    9043.225     298.953     148.395
[6]      81.174      53.362       7.298
```

Condition number κ:

```
> sqrt(e$val[1]/e$val)
[1]   1.00000  13.04226  20.10032 110.55123 156.91171 212.15650
[7] 261.66698 707.54911
```

Example: seat position data - multicollinearity

Eigendecomposition of $\boldsymbol{X}'\boldsymbol{X}$:

```
> x <- model.matrix(lmod)[, -1]
> e <- eigen(t(x) %*% x)
> round(e$val, 3)
[1] 3653671.363   21479.480    9043.225     298.953     148.395
[6]      81.174      53.362       7.298
```

Condition number κ:

```
> sqrt(e$val[1]/e$val)
[1]   1.00000  13.04226  20.10032 110.55123 156.91171 212.15650
[7] 261.66698 707.54911
```

Example: seat position data - multicollinearity

R_1^2 and VIF for the first predictor Age:

```
> (R2.1 <- summary(lm(x[,1] ~ x[,-1]))$r.squared)
[1] 0.4994823
> (VIF.1 <- 1/(1-R2.1))
[1] 1.997931
```

VIF for all predictors:

```
> vif(x)
      Age     Weight    HtShoes         Ht     Seated        Arm
 1.997931   3.647030 307.429378 333.137832   8.951054   4.496368
    Thigh        Leg
 2.762886   6.694291
```

Interpretation: a VIF of 307.4 for HtShoes can be interpreted as follows: the standard error for the regression coefficient for "height with shoes" is $\sqrt{307.4} \approx 17.5$ times larger than it would have been without collinearity.

Example: seat position data - multicollinearity

R_1^2 and VIF for the first predictor `Age`:

```
> (R2.1 <- summary(lm(x[,1] ~ x[,-1]))$r.squared)
[1] 0.4994823
> (VIF.1 <- 1/(1-R2.1))
[1] 1.997931
```

VIF for all predictors:

```
> vif(x)
       Age      Weight     HtShoes          Ht      Seated         Arm
  1.997931    3.647030  307.429378  333.137832    8.951054    4.496368
     Thigh         Leg
  2.762886    6.694291
```

Interpretation: a VIF of 307.4 for `HtShoes` can be interpreted as follows: the standard error for the regression coefficient for "height with shoes" is $\sqrt{307.4} \approx 17.5$ times larger than it would have been without collinearity.

Example: seat position data - multicollinearity

A symptom of multicollinearity: $\hat{\boldsymbol{\beta}}$ is sensitive to small changes in $\boldsymbol{y}$.

```
Call: lm(formula = hipcenter ~ ., data = seatpos)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 436.43213  166.57162   2.620   0.0138 *
Age           0.77572    0.57033   1.360   0.1843
Weight        0.02631    0.33097   0.080   0.9372
HtShoes      -2.69241    9.75304  -0.276   0.7845
Ht            0.60134   10.12987   0.059   0.9531
Seated        0.53375    3.76189   0.142   0.8882
Arm          -1.32807    3.90020  -0.341   0.7359
Thigh        -1.14312    2.66002  -0.430   0.6706
Leg          -6.43905    4.71386  -1.366   0.1824
Residual standard error: 37.72 on 29 degrees of freedom
Multiple R-squared:  0.6866,  Adjusted R-squared:  0.6001
F-statistic:  7.94 on 8 and 29 DF,  p-value: 1.306e-05
--------------------------------------------------
Call: lm(formula = hipcenter + 10 * rnorm(38) ~ ., data = seatpos)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 426.89417  177.72299   2.402   0.0229 *
Age           0.79828    0.60851   1.312   0.1999
Weight        0.02360    0.35313   0.067   0.9472
HtShoes       0.04178   10.40597   0.004   0.9968
Ht           -1.54065   10.80803  -0.143   0.8876
Seated       -0.24235    4.01374  -0.060   0.9523
Arm          -2.47984    4.16130  -0.596   0.5558
Thigh        -0.88113    2.83810  -0.310   0.7584
Leg          -6.53601    5.02944  -1.300   0.2040
Residual standard error: 40.25 on 29 degrees of freedom
Multiple R-squared:  0.6555,  Adjusted R-squared:  0.5605
F-statistic: 6.897 on 8 and 29 DF,  p-value: 4.534e-05
```

Example: seat position data - multicollinearity

We can reduce collinearity by carefully removing some predictor variables. The six length-based variables are strongly correlated with each other:

```
> round(cor(x[,3:8]),2)
        HtShoes   Ht Seated  Arm Thigh  Leg
HtShoes    1.00 1.00   0.93 0.75  0.72 0.91
Ht         1.00 1.00   0.93 0.75  0.73 0.91
Seated     0.93 0.93   1.00 0.63  0.61 0.81
Arm        0.75 0.75   0.63 1.00  0.67 0.75
Thigh      0.72 0.73   0.61 0.67  1.00 0.65
Leg        0.91 0.91   0.81 0.75  0.65 1.00
```

Use `Ht` as proxy for the other predictors. Not much R^2 reduction:

```
lm(formula = hipcenter ~ Age + Weight + Ht, data = seatpos)
               Estimate Std. Error t value Pr(>|t|)
(Intercept) 528.297729 135.312947   3.904 0.000426 ***
Age           0.519504   0.408039   1.273 0.211593
Weight        0.004271   0.311720   0.014 0.989149
Ht           -4.211905   0.999056  -4.216 0.000174 ***
---
Residual standard error: 36.49 on 34 degrees of freedom
Multiple R-squared:  0.6562,  Adjusted R-squared:  0.6258
F-statistic: 21.63 on 3 and 34 DF,  p-value: 5.125e-08
```

Note. Removing predictor variables due to multicollinearity doesn't mean the removed variables are not associated with the response.

Example: seat position data - multicollinearity

We can reduce collinearity by carefully removing some predictor variables. The six length-based variables are strongly correlated with each other:

```
> round(cor(x[,3:8]),2)
        HtShoes   Ht Seated  Arm Thigh  Leg
HtShoes    1.00 1.00   0.93 0.75  0.72 0.91
Ht         1.00 1.00   0.93 0.75  0.73 0.91
Seated     0.93 0.93   1.00 0.63  0.61 0.81
Arm        0.75 0.75   0.63 1.00  0.67 0.75
Thigh      0.72 0.73   0.61 0.67  1.00 0.65
Leg        0.91 0.91   0.81 0.75  0.65 1.00
```

Use `Ht` as proxy for the other predictors. Not much R^2 reduction:

```
lm(formula = hipcenter ~ Age + Weight + Ht, data = seatpos)
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 528.297729 135.312947   3.904 0.000426 ***
Age           0.519504   0.408039   1.273 0.211593
Weight        0.004271   0.311720   0.014 0.989149
Ht           -4.211905   0.999056  -4.216 0.000174 ***
---
Residual standard error: 36.49 on 34 degrees of freedom
Multiple R-squared:  0.6562,  Adjusted R-squared:  0.6258
F-statistic: 21.63 on 3 and 34 DF,  p-value: 5.125e-08
```

Note. Removing predictor variables due to multicollinearity doesn't mean the removed variables are not associated with the response.

Multicollinearity

What to do when predictors appear collinear:

- Drop one or several predictors (can use variable selection techniques)
- More data might break the (near) linear pattern between some predictors.
- In polynomial regression, use scaling techniques to center and standardize the predictor variables

$$x_{ij} = \left(\frac{x_{ij} - \bar{x}_j}{s_{\boldsymbol{x}_j}} \right)^j$$

  where s_x is the standard deviation of X.

  - when $\boldsymbol{x} = (50, 51, \ldots, 70)$, $\text{corr}(\boldsymbol{x}, \boldsymbol{x}^2) = 0.99899$, $\text{corr}(\boldsymbol{x}, \boldsymbol{x}^3) = 0.996023$
  - if transformed: $\boldsymbol{x}^* = (\boldsymbol{x} - \bar{x})/s_{\boldsymbol{x}}$, $\text{corr}(\boldsymbol{x}^*, \boldsymbol{x}^{*2}) = -1.24e - 20$, $\text{corr}(\boldsymbol{x}^*, \boldsymbol{x}^{*3}) = 0.9179$
- Combine correlated predictors to obtain linear combination(s) of correlated predictors, called composite index in economics.

Towards shrinkage methods

- The mathematical reason for the difficulty with multicollinearity is that $\boldsymbol{X}^T\boldsymbol{X}$ is not invertible, or nearly non-invertible.
- Same issue arises when the number of predictors exceed that of the observations, i.e."$p > n$".
- Shrinkage methods, e.g. ridge regression and lasso, uses additional penalty terms to work around (near-)singularity.

p-values and confidence intervals

- p-values
- Confidence intervals.

Confidence region

- $100(1-\alpha)\%$ confidence interval of β_j:

$$\hat{\beta}_j \pm t_{n-(p+1)}^{(1-\alpha/2)} \widehat{\mathrm{se}}(\hat{\beta}_j).$$

- $100(1-\alpha)\%$ confidence region for $\boldsymbol{\beta}$:

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\boldsymbol{X}'\boldsymbol{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \le (p+1)\hat{\sigma}^2 F_{p+1,n-(p+1)}^{(1-\alpha)}.$$

- $100(1-\alpha)\%$ confidence region for $\boldsymbol{\gamma} := \boldsymbol{A}'\boldsymbol{\beta}$:

$$(\boldsymbol{A}'\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma})'(\boldsymbol{A}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A})^{-1}(\boldsymbol{A}'\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma}) \le r\hat{\sigma}^2 F_{r,n-(p+1)}^{(1-\alpha)}.$$

- These are concentric ellipsoids for different values of α.

Model selction

- Test all possible models.
- All subsets. Preferred, but may be computationally impossible!

k	1	10	20	30	40
# models	1	1K	1M	1B	1T
time	1/100s	0.17m	2.9h	124d	348y

(Moore's law: computer speed doubles every two years, allowing one more variable.)

Model selction

- Test all possible models.
- All subsets. Preferred, but may be computationally impossible!

k	1	10	20	30	40
# models	1	1K	1M	1B	1T
time	1/100s	0.17m	2.9h	124d	348y

(Moore's law: computer speed doubles every two years, allowing one more variable.)

Model selection

- Forward selction
- Backward selection
- Adjusted R-square.
- Mallows C_p.
- AIC
- BIC

Model Selection

We would like to select a subset of predictors.

- Why not use all of them? Occam's Razor: among several plausible explanations for a phenomenon, the simplest is the best.
- Bias and variance trade-off:
  - Will miss the true function if only few predictors are included (more bias).
  - Will introduce additional variation if too many predictors are used (more variance).
- Need some other criterion for model selection: residual sum of squares (and R^2) are always in favor of more predictors (recall homework question)
- A principle: respect the hierarchy of higher- vs lower-order terms (e.g. polynomial regression; interactions)

Step-wise selection

- Backward elimination.
  - Select a size α: 5%, 15%, etc. Usually higher for better prediction performance.
  - Start with the full model with p predictors. Perform a t-test for each predictor, and obtain p corresponding p-values. Remove the least significant predictor with the largest p-value, provided that it is larger than α.
  - Refit the model, and repeat the preceding step, until all the predictors are significantly nonzero at the level α.

Step-wise selection

- Forward selection.
  - Select a size α: 5%, 15%, etc.
  - Start with the null model with only the intercept. Add a single predictor to the model, and test whether the corresponding coefficient is nonzero. There are p tests to be carried out. If any p-value from them is smaller than α, then add the predictor with the smallest p-value to the model.
  - Now there are $p - 1$ predictors left. Add each single one of them to the model, and test its coefficients. Similarly, if the smallest p-value (there are $p - 1$ of them in this step) is less than α, add the corresponding predictor to the model.
  - Repeat this procedure until non of the remaining predictor will have a coefficient that is significantly nonzero at level α.

Example: US Census Bureau 1977

Data collected from U.S. Census Bureau on the 50 states from the 1970s.

	Population	Income	Illiteracy	Life.Exp	Murder	HS.Grad	Frost	Area
AL	3615	3624	2.1	69.05	15.1	41.3	20	50708
AK	365	6315	1.5	69.31	11.3	66.7	152	566432
AZ	2212	4530	1.8	70.55	7.8	58.1	15	113417

Example: US Census Bureau 1977

We use life expectancy as response and the rest as predictors.

```
> lmod <- lm(Life.Exp ~ ., statedata)
> summary(lmod)

               Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.094e+01  1.748e+00  40.586  < 2e-16 ***
Population    5.180e-05  2.919e-05   1.775   0.0832 .
Income       -2.180e-05  2.444e-04  -0.089   0.9293
Illiteracy    3.382e-02  3.663e-01   0.092   0.9269
Murder       -3.011e-01  4.662e-02  -6.459 8.68e-08 ***
HS.Grad       4.893e-02  2.332e-02   2.098   0.0420 *
Frost        -5.735e-03  3.143e-03  -1.825   0.0752 .
Area         -7.383e-08  1.668e-06  -0.044   0.9649
---
Residual standard error: 0.7448 on 42 degrees of freedom
Multiple R-squared:  0.7362,  Adjusted R-squared:  0.6922
F-statistic: 16.74 on 7 and 42 DF,  p-value: 2.534e-10
```

Backward elimination. At each stage, remove the predictor with the largest p-value over $\alpha = 0.05$. Area is the first to go.

Example: US Census Bureau 1977 - Backward elimination

Backward elimination. At each stage, remove the predictor with the largest p-value over $\alpha = 0.05$. Area is the first to go.

```
> lmod <- update(lmod, . ~ . - Area)
> summary(lmod)

              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.099e+01  1.387e+00  51.165  < 2e-16 ***
Population   5.188e-05  2.879e-05   1.802   0.0785 .
Income      -2.444e-05  2.343e-04  -0.104   0.9174
Illiteracy   2.846e-02  3.416e-01   0.083   0.9340
Murder      -3.018e-01  4.334e-02  -6.963 1.45e-08 ***
HS.Grad      4.847e-02  2.067e-02   2.345   0.0237 *
Frost       -5.776e-03  2.970e-03  -1.945   0.0584 .
```

Next one up is Illiteracy.

Example: US Census Bureau 1977 - Backward elimination

Backward elimination. At each stage, remove the predictor with the largest p-value over $\alpha = 0.05$. Area is the first to go.

```
> lmod <- update(lmod, . ~ . - Area)
> summary(lmod)
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.099e+01  1.387e+00  51.165  < 2e-16 ***
Population   5.188e-05  2.879e-05   1.802   0.0785 .
Income      -2.444e-05  2.343e-04  -0.104   0.9174
Illiteracy   2.846e-02  3.416e-01   0.083   0.9340
Murder      -3.018e-01  4.334e-02  -6.963 1.45e-08 ***
HS.Grad      4.847e-02  2.067e-02   2.345   0.0237 *
Frost       -5.776e-03  2.970e-03  -1.945   0.0584 .
```

Next one up is Illiteracy.

Example: US Census Bureau 1977 - Backward elimination

```
> lmod <- update(lmod, . ~ . - Illiteracy)
> summary(lmod)

               Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.107e+01  1.029e+00  69.067  < 2e-16 ***
Population   5.115e-05  2.709e-05   1.888   0.0657 .
Income      -2.477e-05  2.316e-04  -0.107   0.9153
Murder      -3.000e-01  3.704e-02  -8.099 2.91e-10 ***
HS.Grad      4.776e-02  1.859e-02   2.569   0.0137 *
Frost       -5.910e-03  2.468e-03  -2.395   0.0210 *
```

Next one up is Income.

Example: US Census Bureau 1977 - Backward elimination

```
> lmod <- update(lmod, . ~ . - Illiteracy)
> summary(lmod)

              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.107e+01  1.029e+00  69.067  < 2e-16 ***
Population   5.115e-05  2.709e-05   1.888   0.0657 .
Income      -2.477e-05  2.316e-04  -0.107   0.9153
Murder      -3.000e-01  3.704e-02  -8.099 2.91e-10 ***
HS.Grad      4.776e-02  1.859e-02   2.569   0.0137 *
Frost       -5.910e-03  2.468e-03  -2.395   0.0210 *
```

Next one up is Income.

Example: US Census Bureau 1977 - Backward elimination

```
> lmod <- update(lmod, . ~ . - Income)
> summary(lmod)

               Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.103e+01  9.529e-01  74.542  < 2e-16 ***
Population    5.014e-05  2.512e-05   1.996  0.05201 .
Murder       -3.001e-01  3.661e-02  -8.199 1.77e-10 ***
HS.Grad       4.658e-02  1.483e-02   3.142  0.00297 **
Frost        -5.943e-03  2.421e-03  -2.455  0.01802 *
```

Next one up is Population, although its p-value is close to the critical value $\alpha = 5\%$ so it's a close call.

Example: US Census Bureau 1977 - Backward elimination

```
> lmod <- update(lmod, . ~ . - Income)
> summary(lmod)

              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.103e+01  9.529e-01  74.542  < 2e-16 ***
Population   5.014e-05  2.512e-05   1.996  0.05201 .
Murder      -3.001e-01  3.661e-02  -8.199 1.77e-10 ***
HS.Grad      4.658e-02  1.483e-02   3.142  0.00297 **
Frost       -5.943e-03  2.421e-03  -2.455  0.01802 *
```

Next one up is Population, although its p-value is close to the critical value $\alpha = 5\%$ so it's a close call.

Example: US Census Bureau 1977 - Backward elimination

```
> lmod <- update(lmod, . ~ . - Population)
> summary(lmod)

             Estimate Std. Error t value Pr(>|t|)
(Intercept) 71.036379   0.983262  72.246  < 2e-16 ***
Murder      -0.283065   0.036731  -7.706 8.04e-10 ***
HS.Grad      0.049949   0.015201   3.286  0.00195 **
Frost       -0.006912   0.002447  -2.824  0.00699 **
---
Residual standard error: 0.7427 on 46 degrees of freedom
Multiple R-squared:  0.7127,  Adjusted R-squared:  0.6939
F-statistic: 38.03 on 3 and 46 DF,  p-value: 1.634e-12
```

Notice that the multiple R^2 for this model is 0.7127, whereas the full model R^2 is 0.7362.

Example: US Census Bureau 1977 - Backward elimination

Note. Again, variables removed from the model may still be related to the response. For example, even if we removed Illiteracy early on, a simpler model using it as a predictor may still be significant:

```
> summary(lm(Life.Exp ~ Illiteracy+Murder+Frost, statedata))

             Estimate Std. Error t value Pr(>|t|)
(Intercept) 74.556717   0.584251 127.611  < 2e-16 ***
Illiteracy  -0.601761   0.298927  -2.013  0.04998 *
Murder      -0.280047   0.043394  -6.454 6.03e-08 ***
Frost       -0.008691   0.002959  -2.937  0.00517 **
---
Residual standard error: 0.7911 on 46 degrees of freedom
Multiple R-squared:  0.6739,  Adjusted R-squared:  0.6527
F-statistic: 31.69 on 3 and 46 DF,  p-value: 2.915e-11
```

Caveats of testing-based procedures

- Can miss the "optimal" model because variables are added/dropped one at a time.
- *p*-values used in the procedure should not be treated too literally. Lots of multiple testing that was not accounted for properly.
- Stepwise variable selection tends to pick models that are smaller than desirable for prediction purposes.
- Variables that are dropped can still be correlated with the response. While they provide little additional explanatory effect beyond those variables already included in the model, it would be wrong to say that these variables are unrelated to the response.
- Any variable selection method must be understood in context of the underlying purpose of the investigation.