Chapter 7: Problems with Predictors

Problems with Predictors

- Errors in predictors
- Change of scale
- Collinearity

Errors in Predictors

Consider simple regression as example.
The X we observe is not the X that generates the y.

$$\begin{aligned} y_i^O &= y_i^A + \epsilon_i \\ x_i^O &= x_i^A + \delta_i \end{aligned}$$

The true relationship is:

$$y_i^A = \beta_0 + \beta_1 x_i^A$$

We get:

$$y_i^O = \beta_0 + \beta_1 x_i^O + (\epsilon_i - \beta_1 \delta_i)$$

Notation

Assume $E(\epsilon_i) = E(\delta_i) = 0$
Let

$$\begin{aligned} var(\epsilon_i) &= \sigma_\epsilon^2 \\ var(\delta_i) &= \sigma_\delta^2 \\ \sigma_x^2 &= \sum \left(x_i^A - \bar{x}^A \right)^2 / n \\ \sigma_{x\delta} &= cov(x^A, \delta) \end{aligned}$$

Effect on the fit

We use least squares to estimate β_1. It turns out

$$E(\hat{\beta}_1) = \beta_1 \frac{\sigma_x^2 + \sigma_{x\delta}}{\sigma_x^2 + \sigma_\delta^2 + 2\sigma_{x\delta}}$$

Scenario 1. x^A and δ are unrelated, i.e., $\sigma_{x\delta} = 0$. Then

$$E(\hat{\beta}_1) = \beta_1 \frac{1}{1 + \sigma_\delta^2/\sigma_x^2}$$

- Shrinks toward 0
- If $\sigma_x^2 \gg \sigma_\delta^2$, the error can be ignored.

Simulation Example

```
## No error in X
> xA <- 10*runif(50)
> yA <- xA
> yO <- yA + rnorm(50)
>  summary(lm(yO ~ xA))
Coefficients:
            Estimate Std.Error t value Pr(>|t|)
(Intercept) -0.23841   0.28125  -0.848    0.401
xA           1.06733   0.05414  19.715   <2e-16
```

```
## Add errors to X
> xO <- xA + rnorm(50)
> summary(lm(yO ~ xO))
Coefficients:
            Estimate Std.Error t value Pr(>|t|)
(Intercept)  0.56790   0.33005   1.721   0.0918
xO           0.89873   0.06198  14.501   <2e-16
## Larger errors
> xO_2 <- xA + 5*rnorm(50)
> summary(lm(yO ~ xO_2))
Coefficients:
            Estimate Std.Error t value Pr(>|t|)
(Intercept)  4.34652   0.49175   8.839 1.23e-11
xO_2         0.07710   0.07035   1.096    0.279
```

Change of Scale

$$x_j \rightarrow \frac{x_j + a}{b}$$

- Predictors of similar magnitude are easier to compare.
- Numerical stability
- Can aid interpretation

Consequences

- Rescaling x_j leaves the t and F tests and $\hat{\sigma^2}$ and R^2 unchanged.

$$\hat{\beta}_j \to b\hat{\beta}_j$$

- Rescaling y leaves the t and F tests and R^2 unchanged but both $\hat{\sigma}$ and $\hat{\beta}$ rescaled by b; $\hat{\beta}_0$ is both shifted by a and rescaled by b.

Savings Example

```
> data(savings)
> result <- lm(sr ~ ., data=savings)
> summary(result)
Coefficients:
            Estimate Std.Error t value Pr(>|t|)
Intercept 28.5666100 7.3544986   3.884 0.000334
pop15     -0.4612050 0.1446425  -3.189 0.002602
pop75     -1.6915757 1.0835862  -1.561 0.125508
dpi       -0.0003368 0.0009311  -0.362 0.719296
ddpi       0.4096998 0.1961961   2.088 0.042468
Residual standard error: 3.803 on 45 degrees of freedom
Multiple R-Squared: 0.3385  Adjusted R-squared: 0.2797
F-statistic: 5.756 on 4 and 45 DF  p-value: 0.0007902
```

Savings Example

```
## Scale one predictor variable
> summary(lm(sr ~ pop15 + pop75 + I(dpi/1000)
  + ddpi, data=savings))
Coefficients:
            Estimate Std.Error t value Pr(>|t|)
(Intercept)  28.5666    7.3545   3.884 0.000334
pop15        -0.4612    0.1446  -3.189 0.002602
pop75        -1.6916    1.0836  -1.561 0.125508
I(dpi/1000)  -0.3368    0.9311  -0.362 0.719296
ddpi          0.4097    0.1962   2.088 0.042468
Residual standard error: 3.803 on 45 degrees of freedom
Multiple R-Squared: 0.3385   Adjusted R-squared: 0.2797
F-statistic: 5.756 on 4 and 45 DF    p-value: 0.0007902
```

Standardizing variables

- Convert all variables to standard units (mean 0, variance 1)
- Can compare coefficients directly
- Helps numerical stability
- Interpretation is harder

```
## Standardize all variables
> sctemp <- data.frame(scale(savings))
> summary(lm(sr ~ ., data=sctemp))
Coefficients:
           Estimate Std.Error   t value Pr(>|t|)
Intercept-2.453e-16 1.200e-01 -2.04e-15   1.0000
pop15    -9.420e-01 2.954e-01    -3.189   0.0026
pop75    -4.873e-01 3.122e-01    -1.561   0.1255
dpi      -7.448e-02 2.059e-01    -0.362   0.7193
ddpi      2.624e-01 1.257e-01     2.088   0.0425
Residual standard error:   0.8487 on 45 degrees of freedom
Multiple R-Squared: 0.3385      Adjusted R-squared: 0.2797
F-statistic: 5.756 on 4 and 45 DF       p-value: 0.0007902
```

Collinearity

- Collinearity: $X^T X$ close to singular
- Cause: some predictors are (almost) linear combinations of others.
- Detection:
  - Correlation matrix: large pairwise correlation
  - Regress x_j on other predictors – get R_j^2.
    R_j^2 close to 1 indicates a problem
  - Condition number of $X^T X$: $\kappa = \sqrt{\frac{\lambda_1}{\lambda_{p+1}}}$

Consequences of Collinearity

- Imprecise estimate of β
- t-test fails to reveal significant predictors
- Sensitivity to measurement errors
- Numerical instability

Collinearity Continued

Why? Let $S_{x_j} = \sum_i (x_{ij} - \bar{x}_j)^2$, then

$$var(\hat{\beta}_j) = \sigma^2 \left(\frac{1}{1 - R_j^2} \right) \frac{1}{S_{x_j}}$$

- Variance inflation factor: $\frac{1}{1-R_j^2}$
- Spread of x_j

Car Example

- Car drivers adjust the seat position for comfort
- Response: seat position
- Predictors: age, weight, height with and without shoes, seated height, arm length, thigh length, lower leg length

```
> data(seatpos)
> result <- lm(hipcenter ~ ., data=seatpos)
> summary(result)
```

```
Coefficients:
            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
```

```
## Correlation matrix
> round(cor(seatpos)[2:7, 2:7], 2)
        Weight HtShoes   Ht Seated  Arm Thigh
Weight    1.00    0.83 0.83   0.78 0.70  0.57
HtShoes   0.83    1.00 1.00   0.93 0.75  0.72
Ht        0.83    1.00 1.00   0.93 0.75  0.73
Seated    0.78    0.93 0.93   1.00 0.63  0.61
Arm       0.70    0.75 0.75   0.63 1.00  0.67
Thigh     0.57    0.72 0.73   0.61 0.67  1.00
```

```
## Condition number
> X <- model.matrix(result)[, -1]
> e <- eigen(t(X) %*% X)
> e$val
[1] 3.653671e+06 2.147948e+04 9.043225e+03
[4] 2.989526e+02 1.483948e+02 8.117397e+01
[7] 5.336194e+01 7.298209e+00
> round(sqrt(e$val[1]/e$val), 3)
[1]   1.000  13.042  20.100 110.551 156.912
[6] 212.156 261.667 707.549
```

```
## Variance inflation factor
> library(faraway)
> round(vif(X), 3)
    Age  Weight HtShoes      Ht  Seated
  1.998   3.647 307.429 333.138   8.951
    Arm   Thigh     Leg
  4.496   2.763   6.694
```

```
## Sensitivity to measurement errors
> junk <- lm(hipcenter + 10*rnorm(38) ~ ., data=seatpos)
> summary(junk)
Coefficients:
             Estimate Std.Error t value Pr(>|t|)
(Intercept)431.13413 176.13709   2.448   0.0207
Age           0.60041   0.60308   0.996   0.3277
Weight       -0.10886   0.34998  -0.311   0.7580
HtShoes      -3.86967  10.31311  -0.375   0.7102
Ht            1.33472  10.71159   0.125   0.9017
Seated        0.79736   3.97792   0.200   0.8425
Arm          -0.01702   4.12417  -0.004   0.9967
Thigh        -1.54993   2.81278  -0.551   0.5858
Leg          -4.73289   4.98456  -0.950   0.3502

Residual standard error:  39.89 on 29 degrees of freedom
Multiple R-Squared: 0.656     Adjusted R-squared: 0.5611
F-statistic: 6.912 on 8 and 29 DF     p-value: 4.451e-05
```

```
## Correlation of variables measuring length
> 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
```

```
## Using a subset of predictor variables
> result2 <- lm(hipcenter ~ Age + Weight + Ht,
    data=seatpos)
> summary(result2)
Coefficients:
            Estimate Std.Error t value Pr(>|t|)
Intercept 528.297729 135.31295   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
```

What to do about collinearity

- If you mostly care about prediction, drop highly correlated predictors
- Variable selection may be used (Ch 8)
- If interpretation is important and you must keep all predictors, do not use least squares. Use some other estimation method, e.g., ridge regression (Ch 9)

Chapter 8: Problems with Error

What can go wrong with the errors?

Recall we assumed $\epsilon \sim N(0, \sigma^2 I)$

- Unequal variance
- Correlated
- Heavy-tailed

Weighted Least Squares

Errors uncorrelated, but unequal variance, i.e.

$$\epsilon \sim N(0, \sigma^2 W^{-1})$$

where

$$W^{-1} = diag(1/w_1, \ldots, 1/w_n)$$

Examples:

- Error variance proportional to the response: $w_i = y_i^{-1}$
- y_i is the average of n_i observations: $w_i = n_i$

Estimates

Transformation:

$$\begin{aligned} y_i &\rightarrow \sqrt{w_i} y_i \\ x_i &\rightarrow \sqrt{w_i} x_i \end{aligned}$$

Regress $\sqrt{w_i} y_i$ on $\sqrt{w_i} x_i$. Then

$$\begin{aligned} \hat{\beta} &= (X^T W X)^{-1} X^T W y \\ var(\hat{\beta}) &= (X^T W X)^{-1} \sigma^2 \\ \hat{\sigma}^2 &= \frac{\hat{\epsilon}^T W \hat{\epsilon}}{n - (p+1)} \end{aligned}$$

French Election Example

- French presidential election in 1981
- 10 candidates in the first round, top 2 in the second round
- Who do the votes go to in the second round?

```
> data(fpe)
> fpe
       EI   A   B   C   D  E  F  G  H  J  K  A2  B2  N
Ain    260  51  64  36  23  9  5  4  4  3  3 105 114 17
Alpes   75  14  17   9   9  3  1  2  1  1  1  32  31  5
... ...
## EI: total number of registered voters
## N: difference between 1st and 2nd round totals
```

```
##Fit a linear model with no intercept
> g <- lm(A2 ~ A+B+C+D+E+F+G+H+J+K+N-1,
    data=fpe, weights=1/EI)
> round(g$coef, 3)
     A      B      C      D      E      F      G
 1.067 -0.105  0.246  0.926  0.249  0.755  1.972
     H      J      K      N
 -0.566  0.612  1.211  0.529
> lm(A2 ~ A+B+C+D+E+F+G+H+J+K+N-1,data=fpe)$coef
     A      B      C      D      E      F      G
 1.075 -0.125  0.257  0.905  0.671  0.783  2.166
     H      J      K      N
-0.854  0.144  0.518  0.558
```

```
## Remove coefficients less than 0
## Set coefficients bigger than 1 to 1
> lm(A2 ~ offset(A+G+K)+C+D+E+F+J+N-1, data=fpe,
        weights=1/EI)$coef
     C       D       E       F      J       N
  0.228   0.970   0.426   0.751  -0.177   0.615

# Now drop J
lm(A2 ~ offset(A+G+K)+C+D+E+F+N-1, data=fpe,
        weights=1/EI)$coef
    C      D      E      F      N
0.226  0.970   0.390  0.744  0.609
```

Generalized Least Squares (GLS)

In general

$$\epsilon \sim N(0, \sigma^2 \Sigma)$$

Write

$$\Sigma = SS^T$$

where S is a lower triangular matrix (the Cholesky decomposition).
Transformation:

$$\begin{aligned} y &\to S^{-1} y \\ x &\to S^{-1} x \end{aligned}$$

Generalized Least Squares Continued

Estimates:

$$\begin{aligned}
\hat{\beta} &= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y \\
var(\hat{\beta}) &= (X^T\Sigma^{-1}X)^{-1}\sigma^2 \\
\hat{\sigma}^2 &= \frac{\hat{\epsilon}^T\Sigma^{-1}\hat{\epsilon}}{n-(p+1)}
\end{aligned}$$

Employment Example

Employment data from 1947 to 1962
Response: number of people employed (yearly) Predictors: gross national product and population over 14

- Data collected over time: errors could be correlated
- One of the simplest correlation structures over time: the autoregressive model – here AR(1):

$$\epsilon_{i+1} = \rho\epsilon_i + \delta_i$$

where δ_i are i.i.d. $N(0, \tau^2)$. This gives

$$cor(\epsilon_i, \epsilon_j) = \rho^{|i-j|}.$$

Employment Example

```
> data(longley)
> g <- lm(Employed ~ GNP + Population, longley)
> summary(g)
Coefficients:
            Estimate Std.Error t value Pr(>|t|)
(Intercept) 88.93880  13.78503   6.452 2.16e-05
GNP          0.06317   0.01065   5.933 4.96e-05
Population  -0.40974   0.15214  -2.693   0.0184

Residual standard error:   0.5459 on 13 degrees of freedom
Multiple R-Squared: 0.9791      Adjusted R-squared: 0.9758
F-statistic: 303.9 on 2 and 13 DF       p-value: 1.221e-11
```

```
## Fit GLS with AR(1) structure
> library(nlme)
> g <- gls(Employed ~ GNP + Population,
   correlation=corAR1(form=~Year), data=longley)
> summary(g)
Correlation Structure: AR(1)
 Formula: ~Year
 Parameter estimate(s):
      Phi     0.6441692
Coefficients:
              Value Std.Error   t-value p-value
Intercept 101.85813 14.198932  7.173647  <.0001
GNP         0.07207  0.010606  6.795485  <.0001
Population -0.54851  0.154130 -3.558778  0.0035

Residual standard error: 0.689207
Degrees of freedom: 16 total; 13 residual
```

```
> intervals(g)
Approximate 95% confidence intervals
 Coefficients:
                  lower         est.       upper
(Intercept) 71.18320440 101.85813280 132.5330612
GNP          0.04915865   0.07207088   0.0949831
Population  -0.88149053  -0.54851350  -0.2155365
 Correlation structure:
         lower      est.     upper
Phi -0.4430373 0.6441692 0.9644866
```

Robust Regression

Main concern: heavy-tailed error distribution

1. M-estimation
2. Least trimmed squares

M-estimation

Find β to minimize

$$\sum_{i=1}^{n} L(y_i - x_i^T \beta)$$

$L(\cdot)$ is called the loss function.

M-estimation Continued

Possible loss functions:

- $L(z) = z^2$ least squares (LS)
- $L(z) = |z|$ least absolute deviations (LAD)
- Huber's method

$$L(z) = \begin{cases} z^2/2 & \text{if } |z| \le c \\ c|z| - c^2/2 & \text{otherwise} \end{cases}$$

c should be a robust estimate of σ, e.g., the median of $|\hat{\epsilon}_i|$.

Gala Example

Recall from Ch. 2: Number of species of tortoise on the various Galapagos slands

- Response: number of species of tortoise
- Predictors: number of endemic species, area of the island, highest elevation of the island, distance from the nearest island, distance from Santa Cruz Island, area of the adjacent island

```
> data(gala)
## Least squares
> g <- lm(Species ~ Area + Elevation + Nearest
    + Scruz + Adjacent, data=gala)
> summary(g)
Coefficients:
             Estimate Std.Error t value Pr(>|t|)
(Intercept) 7.068221 19.154198   0.369 0.715351
Area       -0.023938  0.022422  -1.068 0.296318
Elevation   0.319465  0.053663   5.953 3.82e-06
Nearest     0.009144  1.054136   0.009 0.993151
Scruz      -0.240524  0.215402  -1.117 0.275208
Adjacent   -0.074805  0.017700  -4.226 0.000297

Residual standard error:   60.98 on 24 degrees of freedom
Multiple R-Squared: 0.7658     Adjusted R-squared: 0.7171
F-statistic:  15.7 on 5 and 24 DF      p-value: 6.838e-07
```

```
## Huber's method
> library(MASS)
> ghuber <- rlm(Species ~ Area + Elevation + Nearest
    + Scruz + Adjacent, data=gala)
> summary(ghuber)
Coefficients:
            Value   Std.Error t value
(Intercept)  6.3611 12.3897    0.5134
Area        -0.0061  0.0145   -0.4214
Elevation    0.2476  0.0347    7.1320
Nearest      0.3592  0.6819    0.5267
Scruz       -0.1952  0.1393   -1.4013
Adjacent    -0.0546  0.0114   -4.7648
Residual standard error:  29.73 on 24 degrees of freedom
```

```
## Least absolute deviations
> library(quantreg)
> glad <- rq(Species ~ Area + Elevation + Nearest
    + Scruz + Adjacent, data=gala)
> summary(glad)
Coefficients:
            coefficients lower bd  upper bd
(Intercept)   1.31445    -19.87777  24.37411
Area         -0.00306     -0.03185   0.52800
Elevation     0.23211      0.12453   0.50196
Nearest       0.16366     -3.16339   2.98896
Scruz        -0.12314     -0.47987   0.13476
Adjacent     -0.05185     -0.10458   0.01739
```