Multiple Linear Regression

Corresponds to Chapter 11 of Tamhane & Dunlop

Slides prepared by Elizabeth Newton (MIT)
with some slides by Roy Welsch (MIT).

Linear Regression

Review:

Linear Model: $y=X\beta + \varepsilon$

$y \sim N(X\beta, \sigma^2 I)$

Least squares: $\hat{\beta}$ =(X'X)X'y

$\hat{y}$ = fitted value of y = X $\hat{\beta}$ =

$X(X'X)^{-1}X'y=Hy$

e = error = residuals = y- $\hat{y}$ = y-Hy=(I-H)y

Properties of the Hat matrix

- Symmetric: H'=H
- Idempotent: HH=H
- Trace(H) = sum(diag(H)) = k+1 = number of columns in the X matrix
- 1'H=vector of 1's (hence y and $\hat{y}$ have same mean)
- 1'(I-H) = vector of 0's (hence mean of residuals is 0).
- What is H when X is only a column of 1's?

Variance-Covariance Matrices

$$\text{Cov}(\hat{\beta}) = \sigma^2 (X'X)^{-1} \text{ (as we saw last time)}$$

$$Cov(\hat{y}) = Cov(Hy) = HCov(y)H$$
$$= H\sigma^2 IH = \sigma^2 H$$

$$Cov(e) = Cov(I-H)y = (I-H)Cov(y)(I-H)$$
$$= (I-H)\sigma^2 I(I-H) = \sigma^2 (I-H)$$

Confidence and Prediction Intervals

Variance of mean response at x_0

$$Var(\hat{y}_0) = Var(x_0'\hat{\beta}) = \sigma^2 x_0'(X'X)^{-1}x_0 = \sigma^2 v_0$$

Variance of new observation at x_0, $y_0 = \hat{y}_0 + \varepsilon_0$

$$Var(\hat{y}_0 + \varepsilon_0) = Var(\hat{y}_0) + Var(\varepsilon_0) =$$

$$\sigma^2 x_0'(X'X)^{-1}x_0 + \sigma^2 = \sigma^2(x_0'(X'X)^{-1}x_0 + 1) = \sigma^2(v_0 + 1)$$

An estimate of σ^2 is s^2 = MSE = $y'(I-H)y/(n-k-1)$

Confidence and Prediction Intervals

(1-α) Confidence Interval on Mean Response at x_0:

$$\hat{y}_0 \pm cd, \text{where } \mathrm{c} = \mathrm{t}_{\mathrm{n\text{-}(k+1)}, \alpha/2} \text{ and } \mathrm{d} = \mathrm{s}\sqrt{\mathrm{v}_0}$$

(1-α) Prediction Interval on New Observation at x_0:

$$\hat{y}_0 \pm cd, \text{where } \mathrm{c} = \mathrm{t}_{\mathrm{n\text{-}(k+1)}, \alpha/2} \text{ and } \mathrm{d} = \mathrm{s}\sqrt{\mathrm{v}_0 + 1}$$

Sums of Squares

Sum of Squares Total (SST) : $\sum_{i=1}^{n}(y_i - \bar{y})^2$

Sum of Squares for Error (SSE) : $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Sum of Squares for Regression (SSR) : $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

SSR = SST - SSE

Overall Significance Test

To see if there is any linear relationship we test:

H_0: $\beta_1 = \beta_2 = . . . = \beta_k = 0$
H_1: $\beta_j \neq 0$ for some j.

Compute

$$SSE = \sum (y_i - \hat{y}_i)^2 \quad \mathrm{SST} = \sum (y_i - \bar{y}_i)^2 \quad SSR = SST - SSE$$

The F statistic is: $$\frac{SSR/k}{SSE/(n-k-1)} = \frac{MSR}{MSE}$$

with F based on k and $(n - k - 1)$ degrees of freedom.

Reject H_0 when F exceeds $F_{k,n-k-1(\alpha)}$.

Sequential Sums of Squares

SSR(x1) = SST - SSE(x1)

SSR(x2|x1) = SSR(x1,x2) - SSR(x1) =
SSE(x1) - SSE(x1,x2)

SSR(x3|x1 x2) = SSE(x1,x2) - SSE(x1,x2,x3)

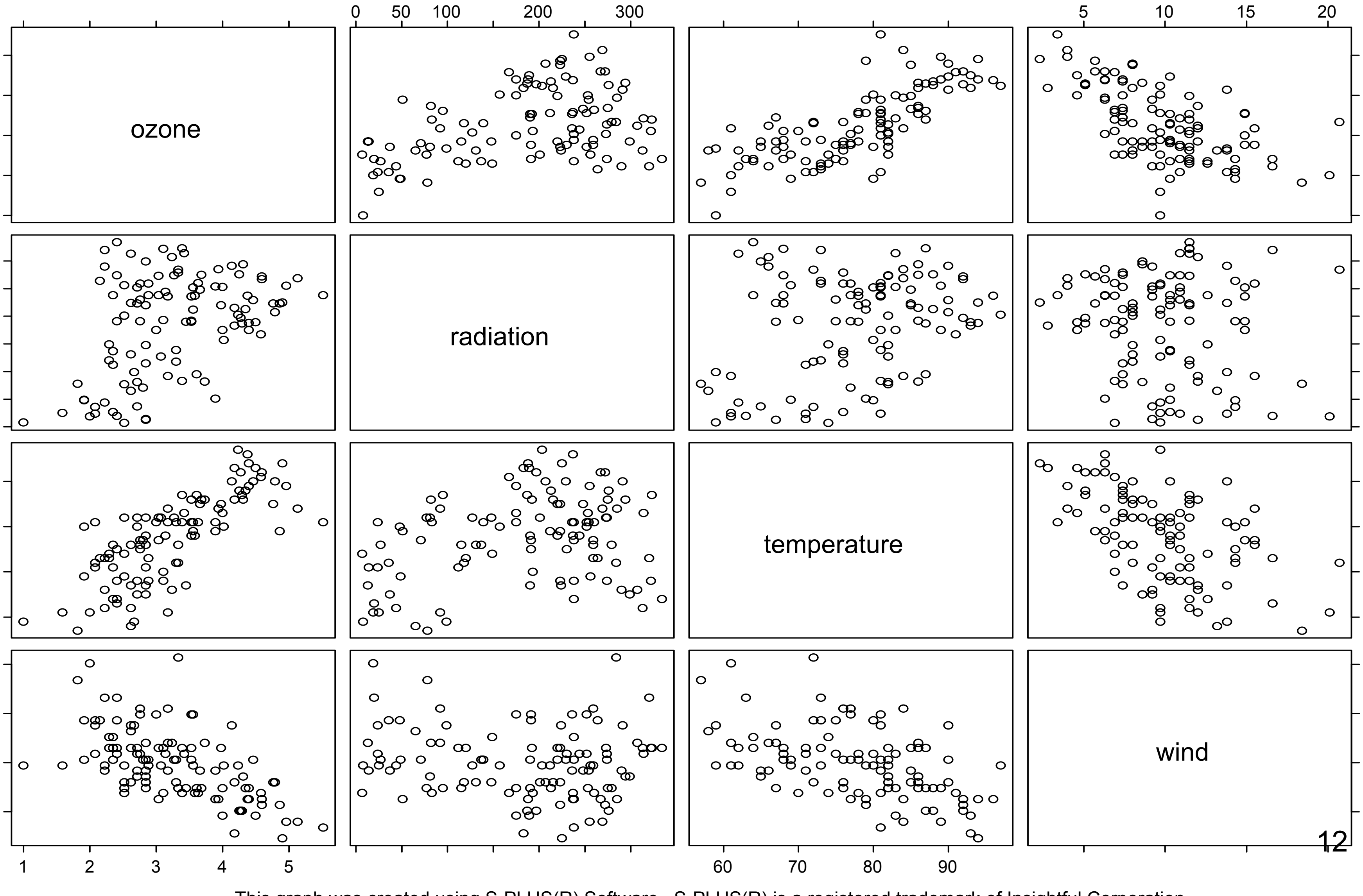
ANOVA Table
Type 1 (sequential) sums of squares

Source of Variation	SS	df
Regression	SSR(x1,x2,x3)	3
x1	SSR(x1)	1
x2|x1	SSR(x2|x1)	1
x3|x2 x1	SSR(x3|x2,x1)	1
Error	SSE(x1,x2,x3)	n-4
Total	SST	n-1

ANOVA Table
Type 3 (partial) sums of squares

Source of Variation	SS	df
Regression	SSR(x1,x2,x3)	3
x1|x2,x3	SSR(x1|x2,x3)	1
x2|x1,x3	SSR(x2|x1,x3)	1
x3|x1,x2	SSR(x3|x1,x2)	1
Error	SSE(x1,x2,x3)	n-4
Total	SST	n-1

Scatter plot Matrix of the Air Data Set in S-Plus
pairs(air)



```
air.lm<-lm(y~x1+x2+x3)

> summary(air.lm)$coef
                  Value    Std. Error    t value     Pr(>|t|)
(Intercept) -0.297329634 0.5552138923 -0.5355227 5.933998e-001
         x1  0.002205541 0.0005584658  3.9492854 1.407070e-004
         x2  0.050044325 0.0061061612  8.1957098 5.848655e-013
         x3 -0.076021950 0.0157548357 -4.8253090 4.665124e-006

> summary.aov(air.lm)
           Df Sum of Sq  Mean Sq  F Value         Pr(F)
       x1   1  15.53144 15.53144  59.6761 6.000000e-012
       x2   1  37.76939 37.76939 145.1204 0.000000e+000
       x3   1   6.05985  6.05985  23.2836 4.665124e-006
Residuals 107  27.84808  0.26026

> summary.aov(air.lm,ssType=3)
Type III Sum of Squares
           Df Sum of Sq  Mean Sq  F Value        Pr(F)
       x1   1   4.05928  4.05928 15.59685 0.0001407070
       x2   1  17.48174 17.48174 67.16966 0.0000000000
       x3   1   6.05985  6.05985 23.28361 0.0000046651
Residuals 107  27.84808  0.26026
>
```

Polynomial Models

$y=\beta_0 + \beta_1 x + \beta_2 x^2 \ldots + \beta_k x^k$

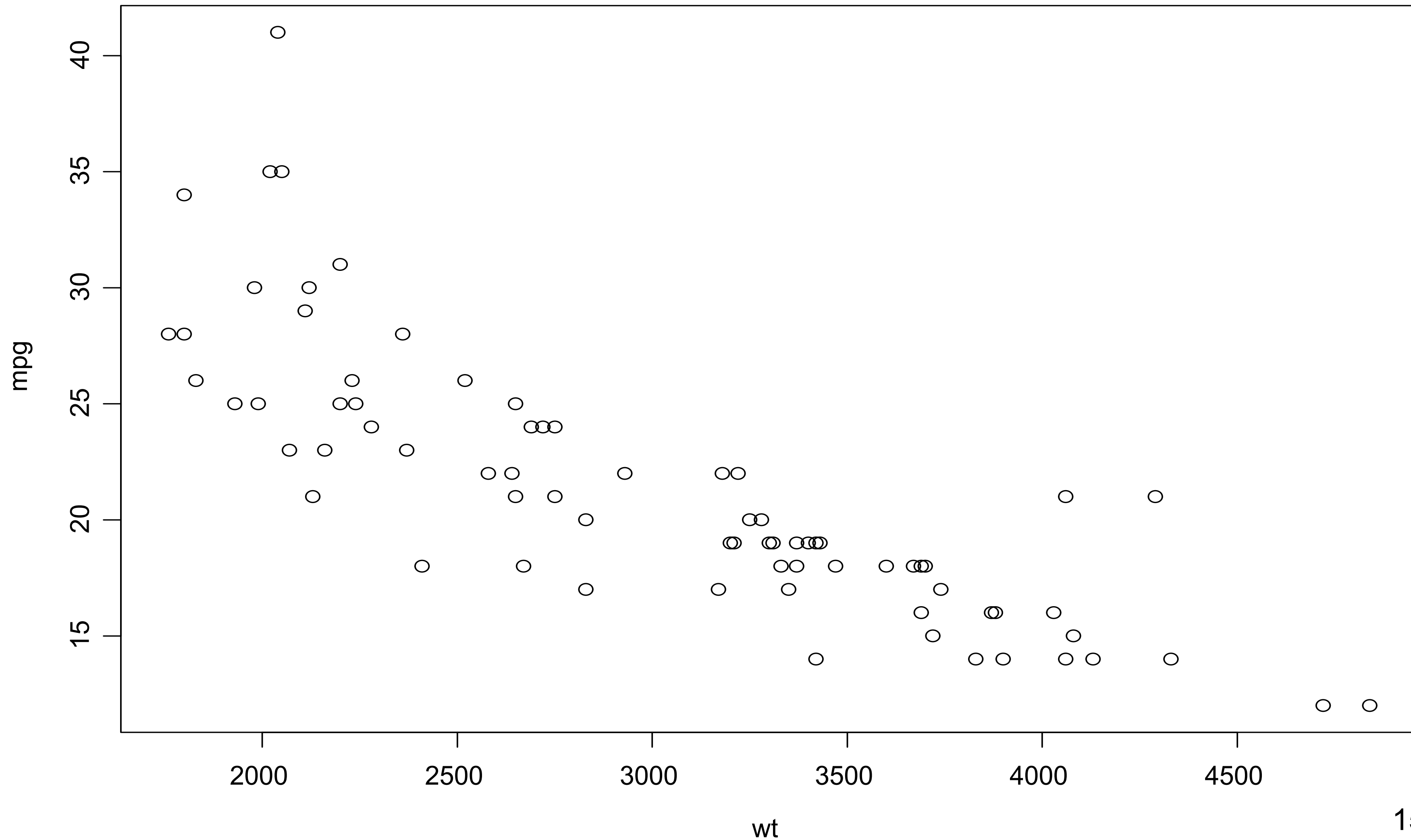
Problems:

- Powers of x tend to be large in magnitude
- Powers of x tend to be highly correlated

Solutions:

- Centering and scaling of x variables
- Orthogonal polynomials (poly(x,k) in S-Plus, see Seber for methods of generating)

Plot of mpg vs. weight for 74 autos (S-Plus dataset auto.stats)



This graph was created using S-PLUS(R) Software. S-PLUS(R) is a registered trademark of Insightful Corporation.

```
summary(lm(mpg~wt+wt^2+wt^3))

Call: lm(formula = mpg ~ wt + wt^2 + wt^3)
Residuals:
    Min     1Q  Median    3Q   Max
 -6.415 -1.556 -0.2815 1.265 13.06

Coefficients:
                Value     Std. Error   t value    Pr(>|t|)
(Intercept)  68.1797   21.4515      3.1783     0.0022
         wt   -0.0309    0.0214     -1.4430     0.1535
    I(wt^2)    0.0000    0.0000      0.9586     0.3410
    I(wt^3)    0.0000    0.0000     -0.7449     0.4588

Residual standard error: 3.209 on 70 degrees of freedom
Multiple R-Squared: 0.705
F-statistic: 55.76 on 3 and 70 degrees of freedom, the p-value is 0

Correlation of Coefficients:
       (Intercept)      wt        I(wt^2)
     wt -0.9958
I(wt^2)  0.9841     -0.9961
I(wt^3) -0.9659      0.9846   -0.9961
```

```
wts<-(wt-mean(wt))/sqrt(var(wt))

summary(lm(mpg~wts+wts^2+wts^3))

Call: lm(formula = mpg ~ wts + wts^2 + wts^3)
Residuals:
    Min      1Q  Median     3Q    Max
 -6.415 -1.556 -0.2815 1.265 13.06

Coefficients:
               Value    Std. Error  t value   Pr(>|t|)
(Intercept)  20.2331    0.5676    35.6470    0.0000
        wts  -4.4466    0.7465    -5.9567    0.0000
   I(wts^2)   1.1241    0.4682     2.4007    0.0190
   I(wts^3)  -0.2521    0.3385    -0.7449    0.4588

Residual standard error: 3.209 on 70 degrees of freedom
Multiple R-Squared: 0.705
F-statistic: 55.76 on 3 and 70 degrees of freedom, the p-value is 0

Correlation of Coefficients:
         (Intercept)     wts      I(wts^2)
      wts -0.2800
I(wts^2) -0.7490      0.4558
I(wts^3)  0.3925     -0.8596    -0.6123
```

Orthogonal Polynomials

Generation is similar to Gram-Schmidt orthogonalization (see Strang, Linear Algebra)

Resulting vectors are orthonormal $X'X=I$

Hence $(X'X)^{-1} = I$ and coefficients

$= (X'X)^{-1}X'y = X'y$

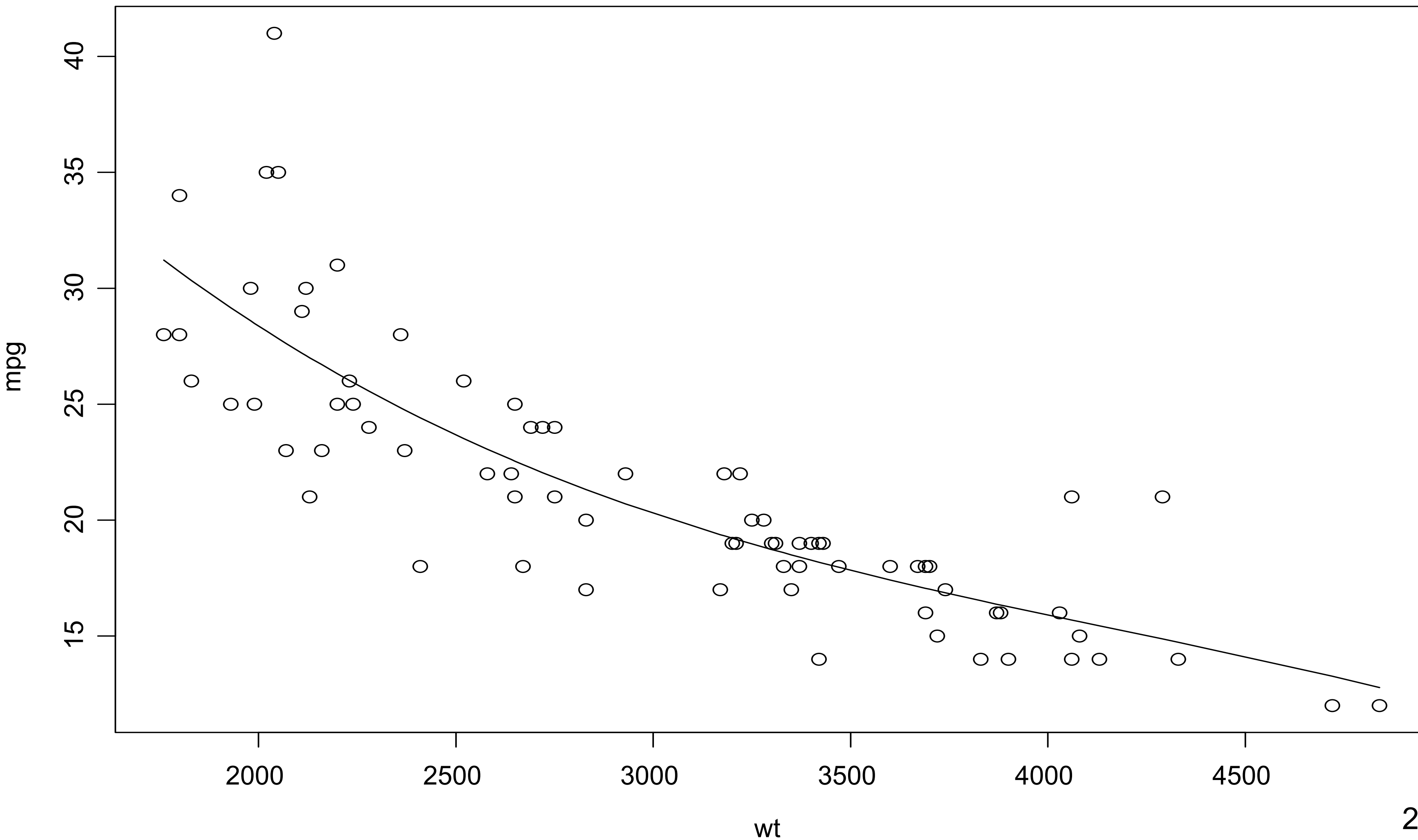
Addition of higher degree term does not affect coefficients for lower degree terms

Correlation of coefficients = I

SE of coefficients = s = sqrt(MSE)

```
summary(lm(mpg~poly(wt,3)))

Call: lm(formula = mpg ~ poly(wt, 3))
Residuals:
    Min     1Q  Median    3Q   Max
 -6.415 -1.556 -0.2815 1.265 13.06

Coefficients:
                 Value Std.   Error        t value      Pr(>|t|)
 (Intercept)   21.2973     0.3730     57.0912      0.0000
poly(wt, 3)1 -40.6769      3.2090    -12.6758     0.0000
poly(wt, 3)2    7.8926      3.2090      2.4595       0.0164
poly(wt, 3)3  -2.3904       3.2090     -0.7449      0.4588

Residual standard error: 3.209 on 70 degrees of freedom
Multiple R-Squared: 0.705
F-statistic: 55.76 on 3 and 70 degrees of freedom, the p-value is 0

Correlation of Coefficients:
             (Intercept) poly(wt, 3)1 poly(wt, 3)2
poly(wt, 3)1 0
poly(wt, 3)2 0           0
poly(wt, 3)3 0           0               0
```

Plot of mpg by weight with fitted regression line

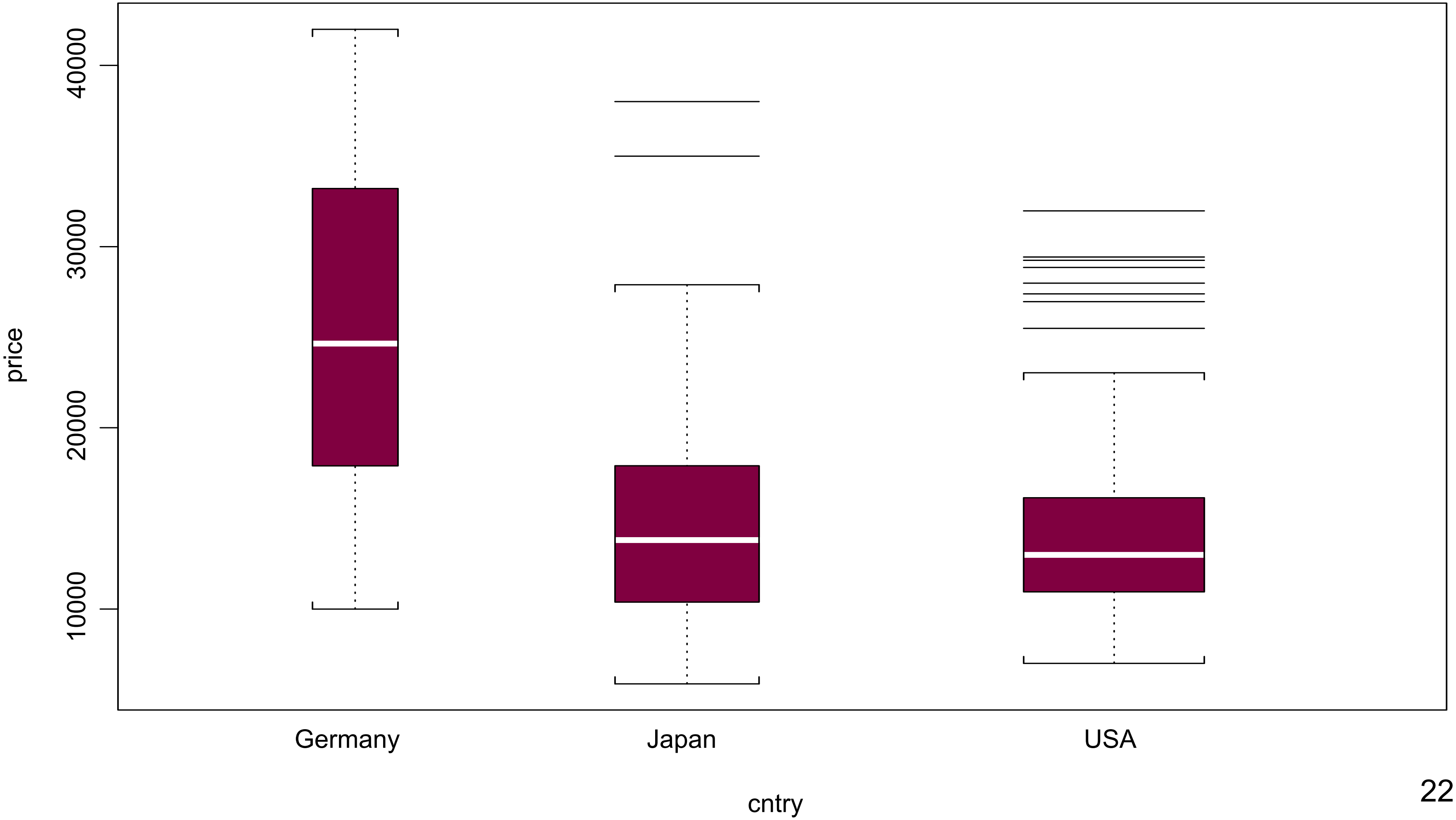


This graph was created using S-PLUS(R) Software. S-PLUS(R) is a registered trademark of Insightful Corporation.

Indicator Variables

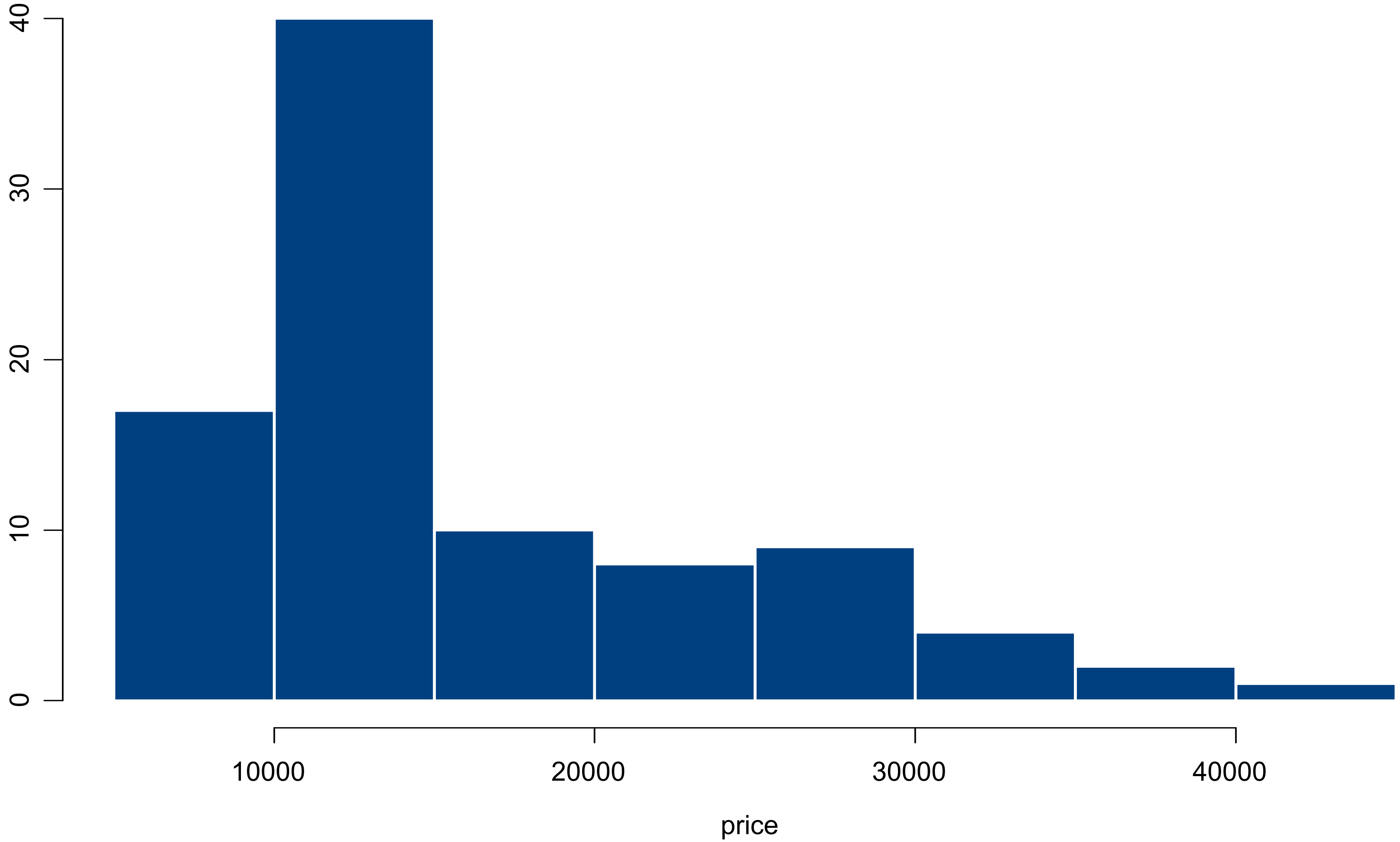
- Sometimes we might want to fit a model with a categorical variable as a predictor. For instance, automobile price as a function of where the car is made (Germany, Japan, USA).
- If there are c categories, we need c-1 indicator (0,1) variables as predictors. For instance j=1 if car is made in Japan, 0 otherwise, u=1 if car is made in USA, 0 otherwise.
- If there are just 2 categories and no other predictors, we could just do a t-test for difference in means.

Boxplots of price by country for S-Plus dataset cu.summary



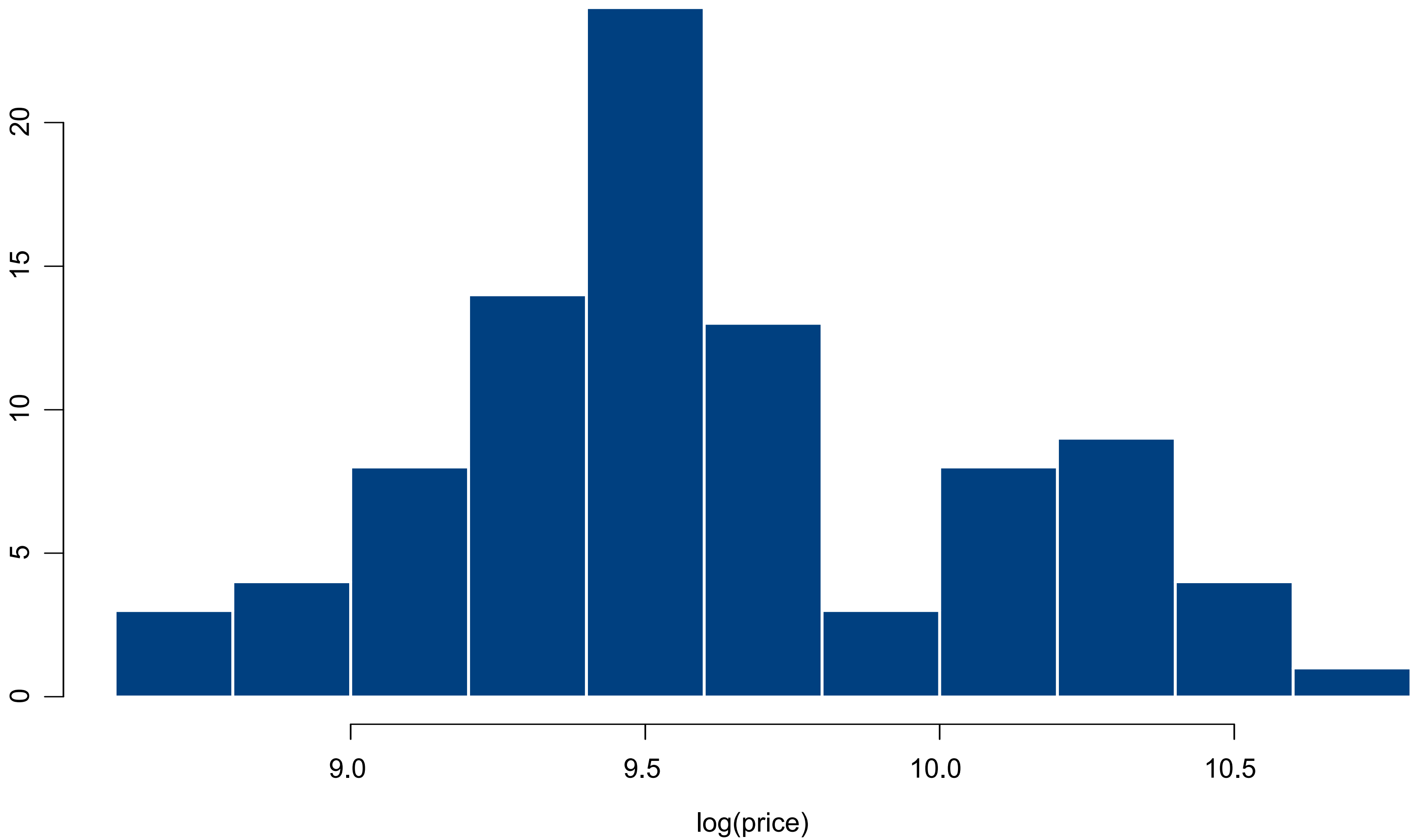
This graph was created using S-PLUS(R) Software. S-PLUS(R) is a registered trademark of Insightful Corporation.

Histogram of automobile prices for S-Plus dataset cu.summary



This graph was created using S-PLUS(R) Software. S-PLUS(R) is a registered trademark of Insightful Corporation.

Histogram of log of automobile prices for S-Plus dataset cu.summary



This graph was created using S-PLUS(R) Software. S-PLUS(R) is a registered trademark of Insightful Corporation.

```
summary(lm(price~u+j))

Call: lm(formula = price ~ u + j)
Residuals:
    Min     1Q Median   3Q   Max
 -15746 -4586  -2071 2374 22495

Coefficients:
                 Value        Std. Error   t value    Pr(>|t|)
(Intercept)  25741.3636   2282.2729    11.2788      0.0000
        u    -10520.5473   2525.4871    -4.1657      0.0001
        j    -10236.0088   2656.5095    -3.8532      0.0002

Residual standard error: 7569 on 88 degrees of freedom
Multiple R-Squared: 0.1723
F-statistic: 9.159 on 2 and 88 degrees of freedom, the p-value is
    0.0002435

Correlation of Coefficients:
  (Intercept)       u
u -0.9037
j  -0.8591      0.7764
```

```
summary(lm(price~u+g))

Call: lm(formula = price ~ u + g)
Residuals:
   Min    1Q Median   3Q   Max
 -15746 -4586  -2071 2374 22495

Coefficients:
                 Value        Std. Error    t value     Pr(>|t|)
(Intercept)  15505.3548   1359.5121    11.4051     0.0000
        u      -284.5385   1737.1208    -0.1638     0.8703
        g     10236.0088   2656.5095     3.8532     0.0002

Residual standard error: 7569 on 88 degrees of freedom
Multiple R-Squared: 0.1723
F-statistic: 9.159 on 2 and 88 degrees of freedom, the p-value is
    0.0002435

Correlation of Coefficients:
  (Intercept)      u
u -0.7826
g -0.5118      0.4005
```


This code was created using S-PLUS(R) Software. S-PLUS(R) is a registered trademark of Insightful Corporation.

Regression Diagnostics

Goal: identify remarkable observations and unremarkable predictors.

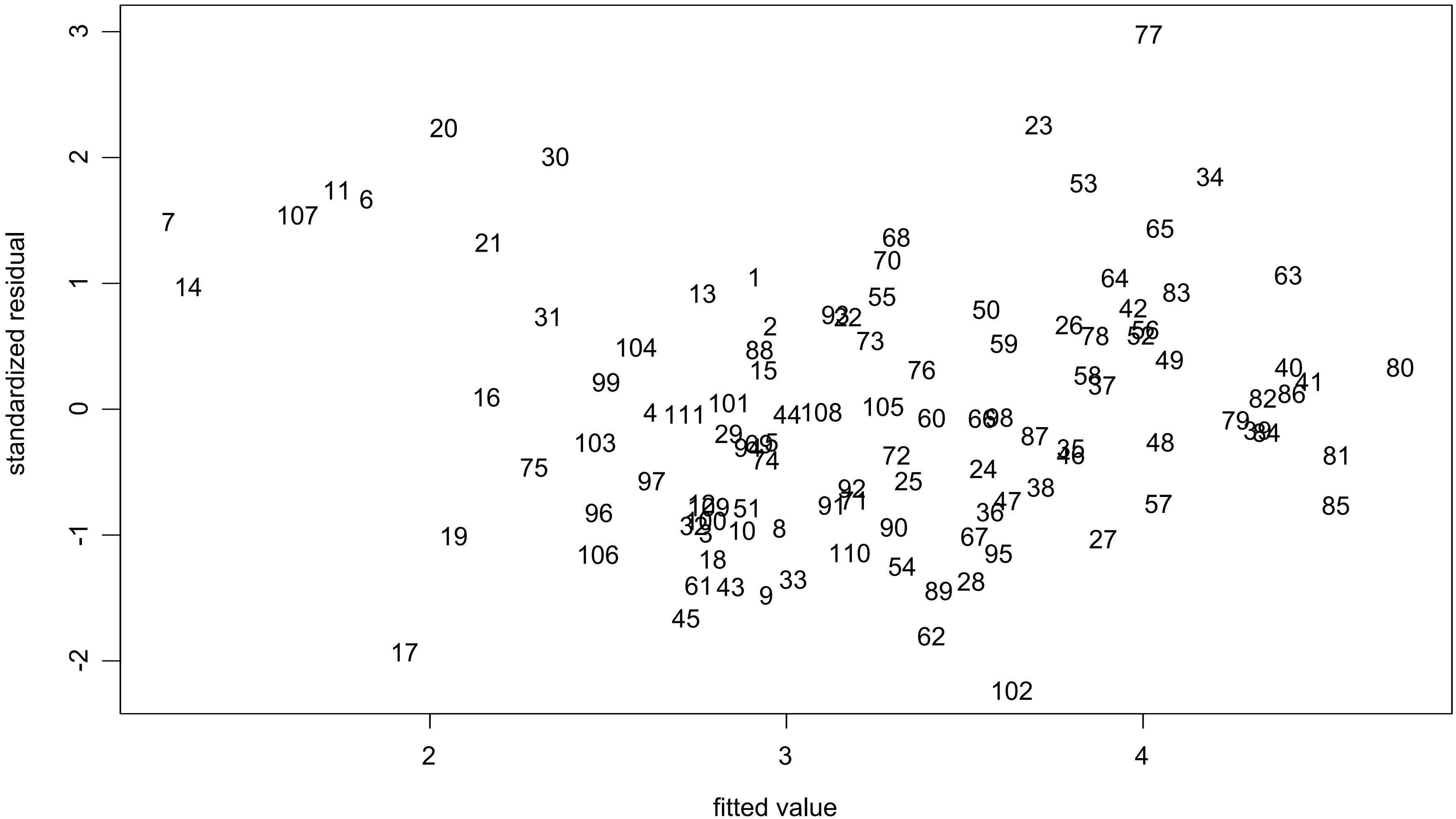
Problems with observations:

- Outliers
- Influential observations

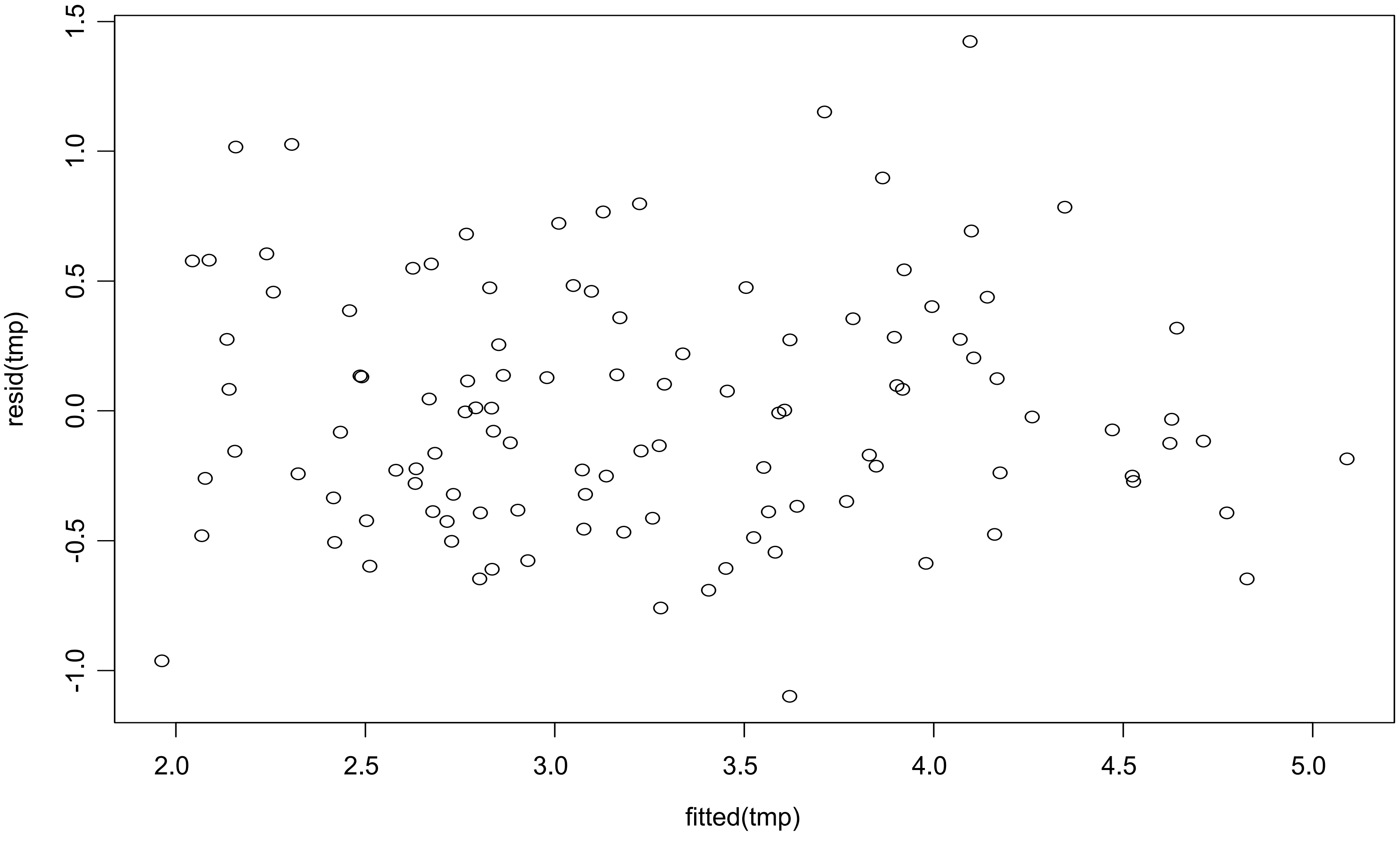
Problems with predictors:

- A predictor may not add much to model.
- A predictor may be too similar to another predictor (collinearity).
- Predictors may have been left out.

Plot of standardized residuals vs. fitted values for air dataset

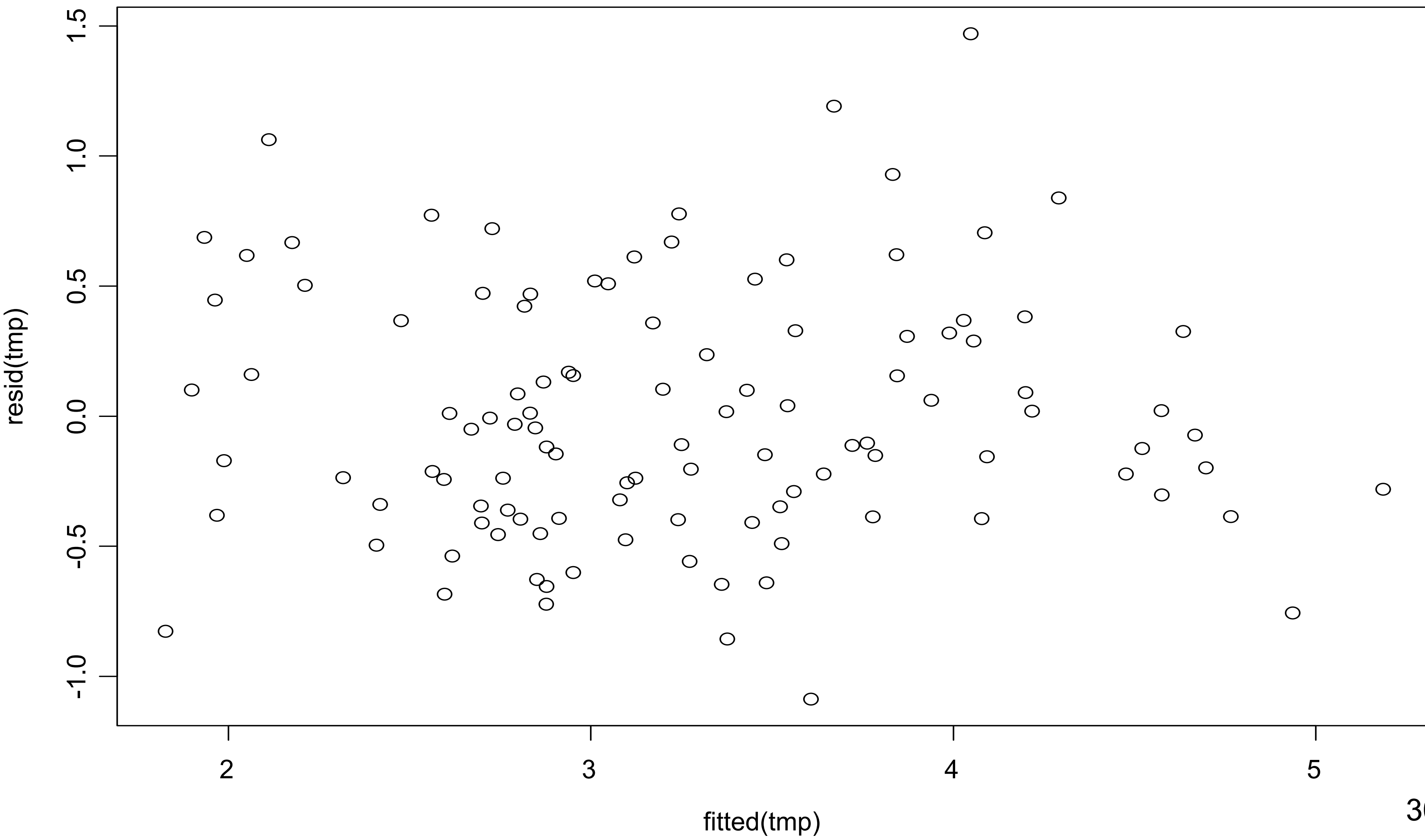


Plot of residual vs. fit for air data set with all interaction terms



This graph was created using S-PLUS(R) Software. S-PLUS(R) is a registered trademark of Insightful Corporation.

Plot of residual vs. fit for air model with x3*x4 interaction



```
Call: lm(formula = air[, 1] ~ air[, 2] + air[, 3] + air[, 4] + air[, 3] * air[, 4])
Residuals:
    Min      1Q   Median     3Q  Max
 -1.088 -0.3542 -0.07242 0.3436 1.47

Coefficients:
                  Value     Std. Error  t value  Pr(>|t|)
     (Intercept) -3.6465  1.1684    -3.1209  0.0023
        air[, 2]    0.0023   0.0005     4.3223  0.0000
        air[, 3]    0.0920   0.0143     6.4435  0.0000
        air[, 4]    0.2523   0.1031     2.4478  0.0160
air[, 3]:air[, 4] -0.0042    0.0013    -3.2201  0.0017

Residual standard error: 0.4892 on 106 degrees of freedom
Multiple R-Squared: 0.7091
F-statistic: 64.61 on 4 and 106 degrees of freedom, the p-value is 0

Correlation of Coefficients:
                 (Intercept)    air[, 2]   air[, 3]     air[, 4]
        air[, 2] -0.0361
        air[, 3] -0.9880      -0.0495
        air[, 4] -0.9268       0.0620    0.9313
air[, 3]:air[, 4]  0.8902     -0.0661  -0.9119  -0.9892
>
```

Remarkable Observations

Residuals are the key

Standardized residuals:

$$e_i^* = \frac{e_i}{SE(e_i)} = \frac{e_i}{s\sqrt{i - h_{ii}}}$$

Outlier if $|e_i^*|>2$

Hat matrix diagonals, h_{ii}

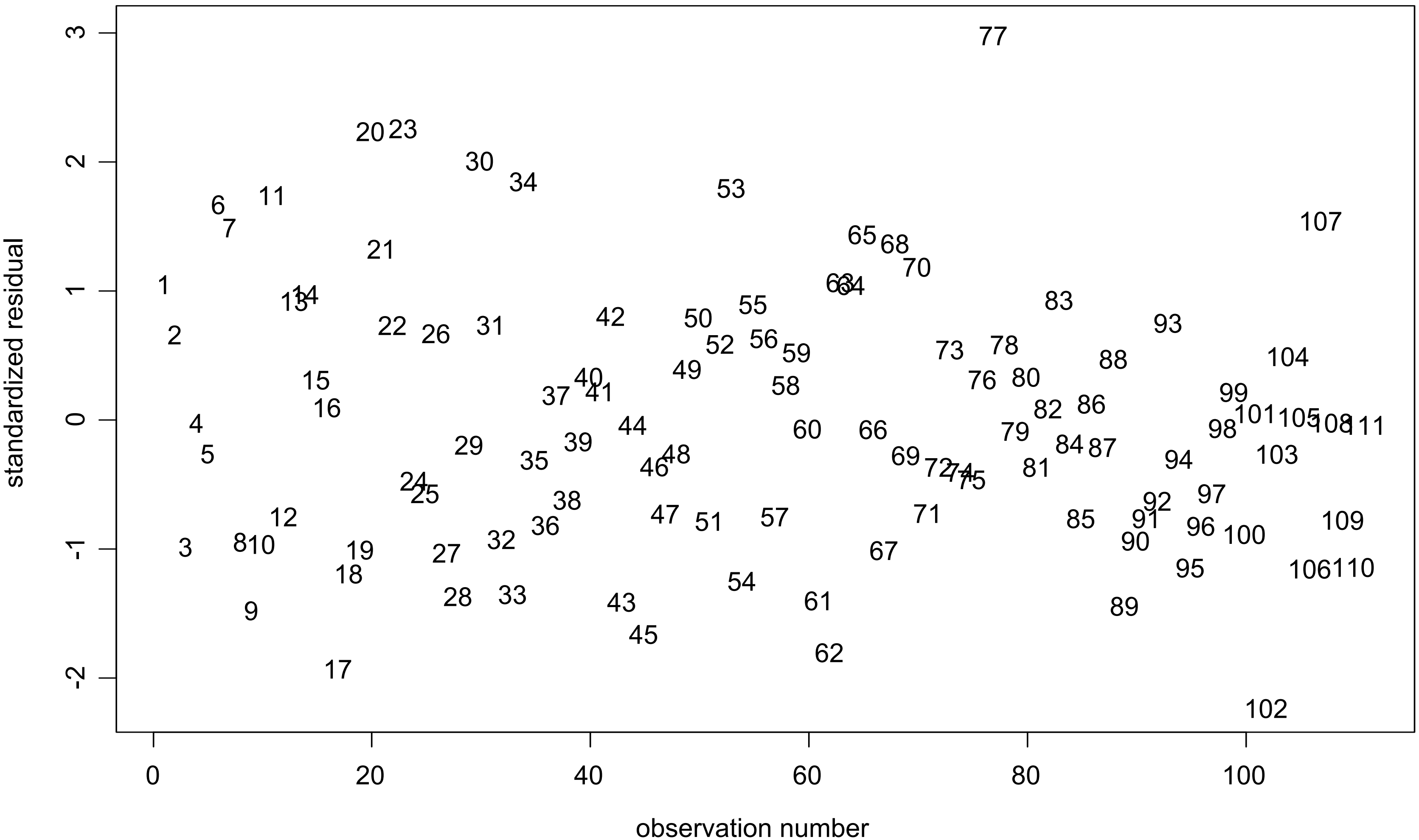
Influential if $h_{ii} > 2(k+1)/n$

Cook's Distance

$$d_i = \left(\frac{e_i^*}{\sqrt{k+1}}\right)^2 \left(\frac{h_{ii}}{1-h_{ii}}\right)$$

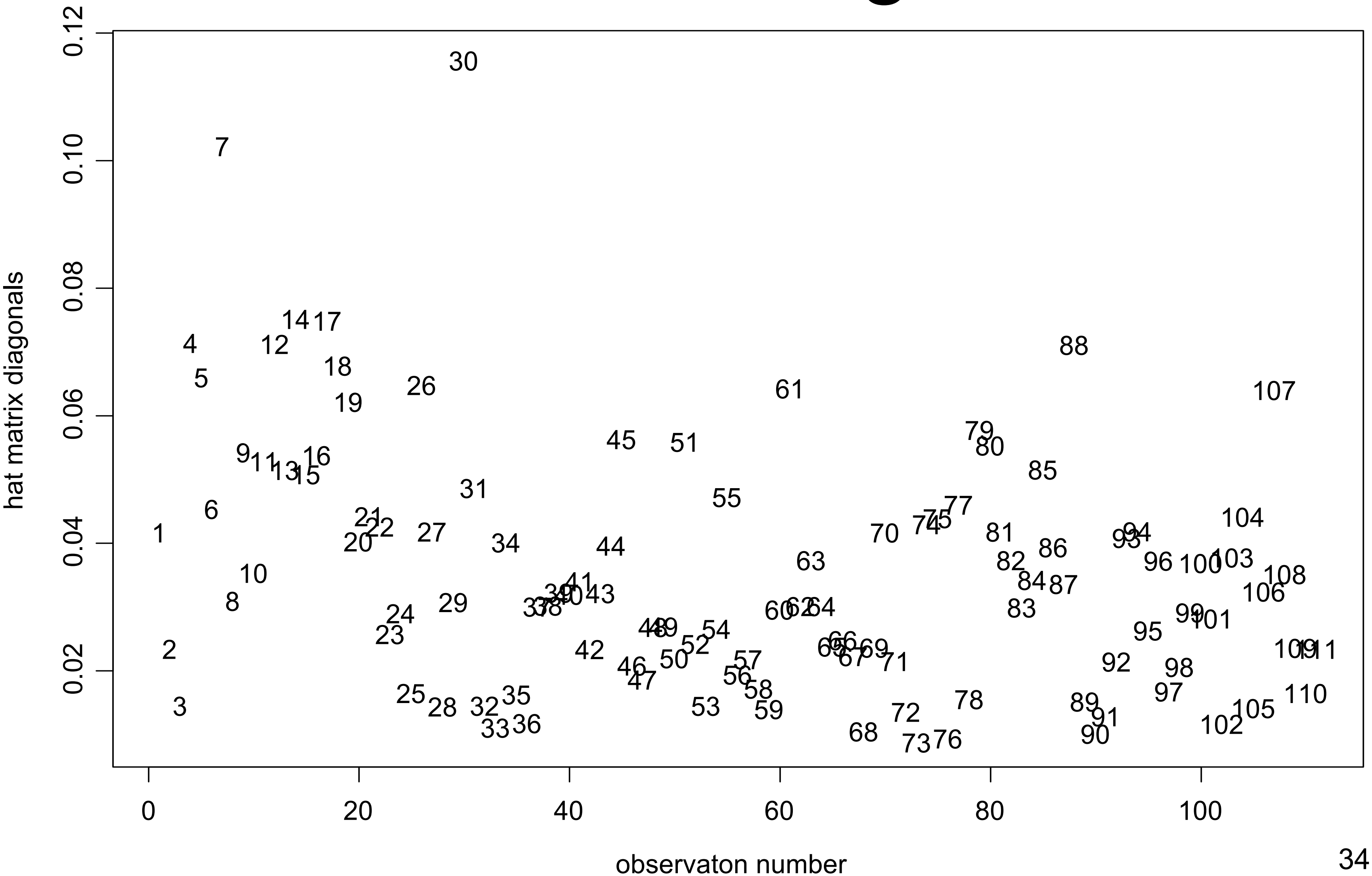
Influential if $d_i > 1$

Plot of standardized residual vs. observation number for air dataset

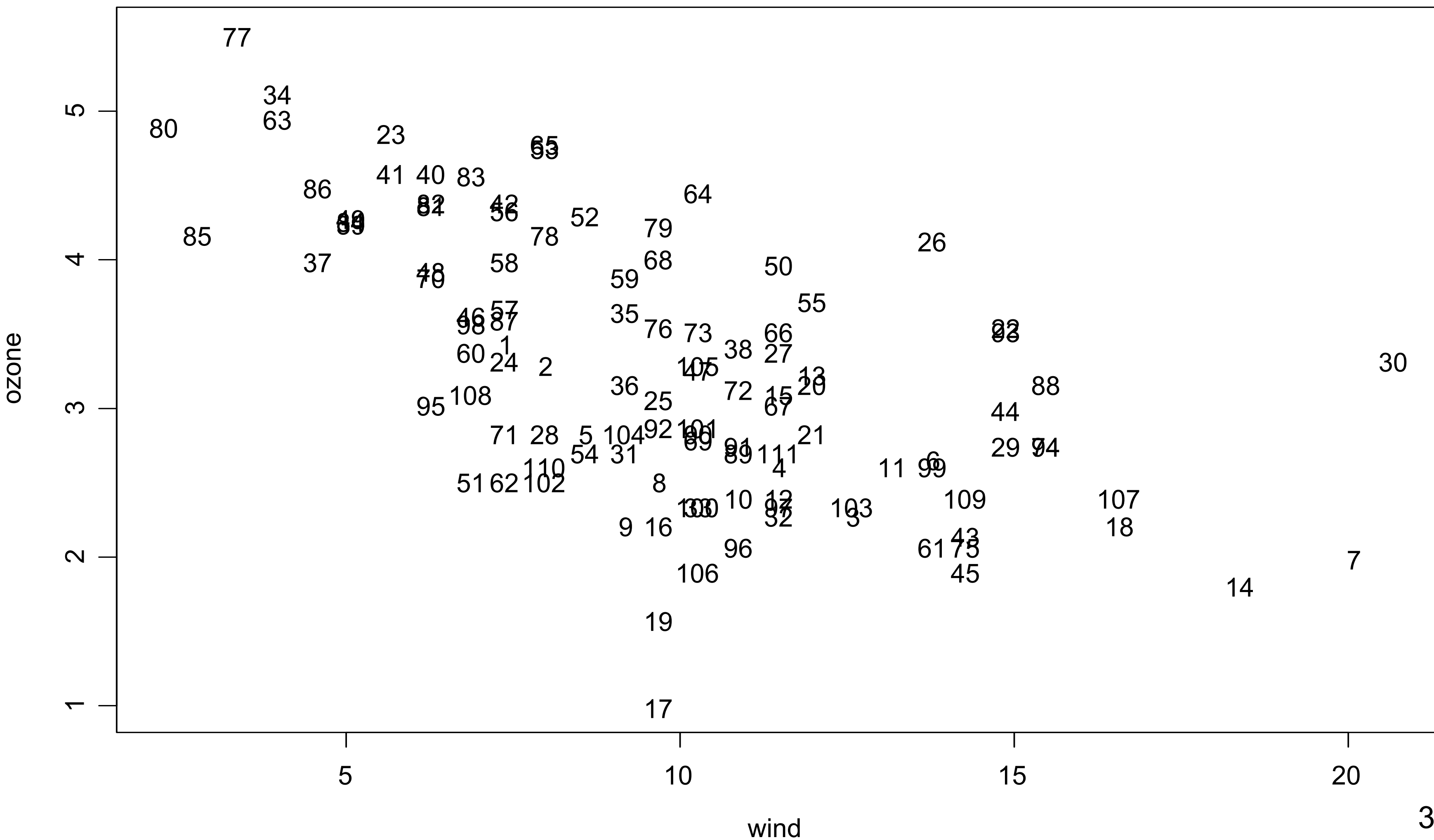


This graph was created using S-PLUS(R) Software. S-PLUS(R) is a registered trademark of Insightful Corporation.

Hat matrix diagonals

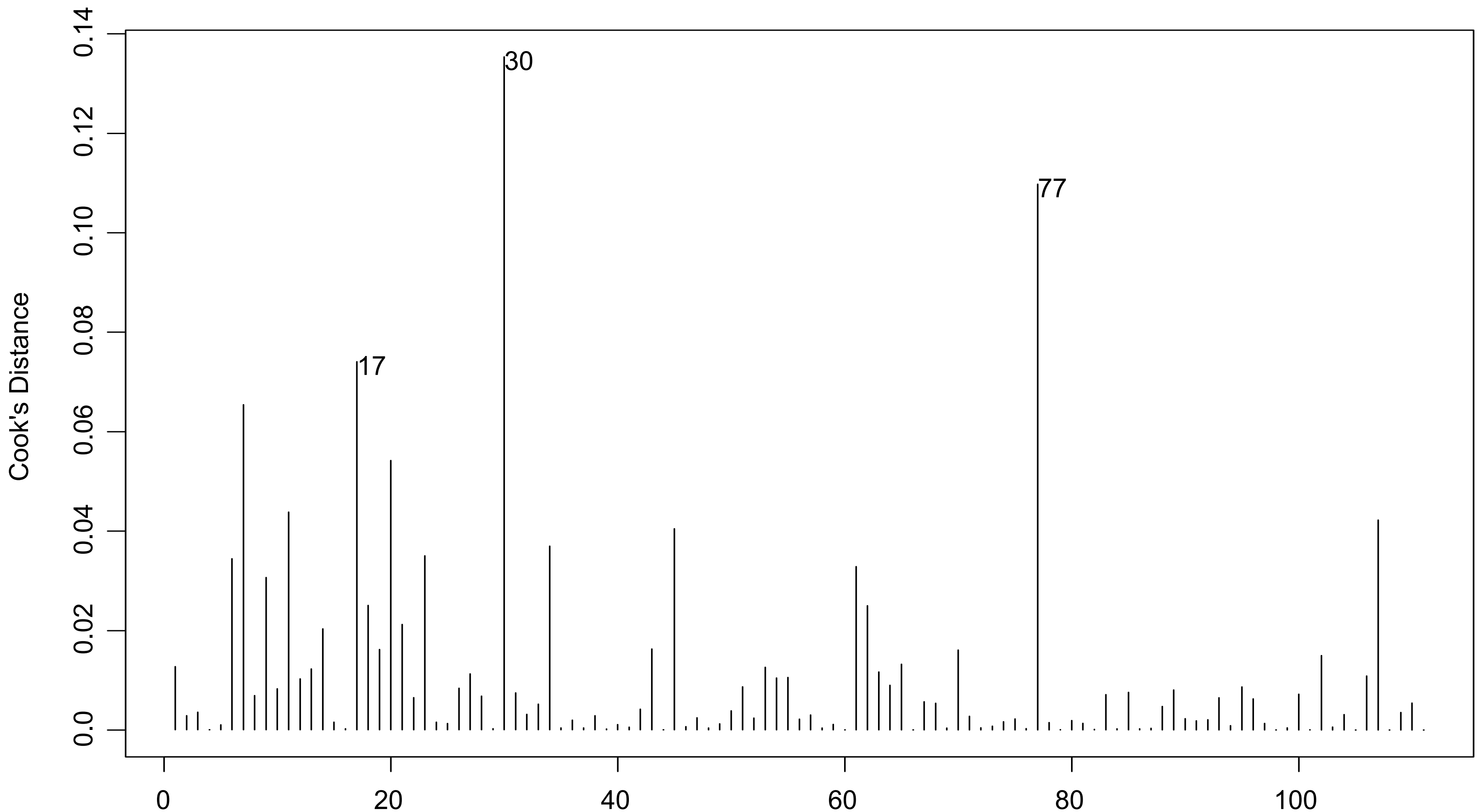


Plot of wind vs. ozone



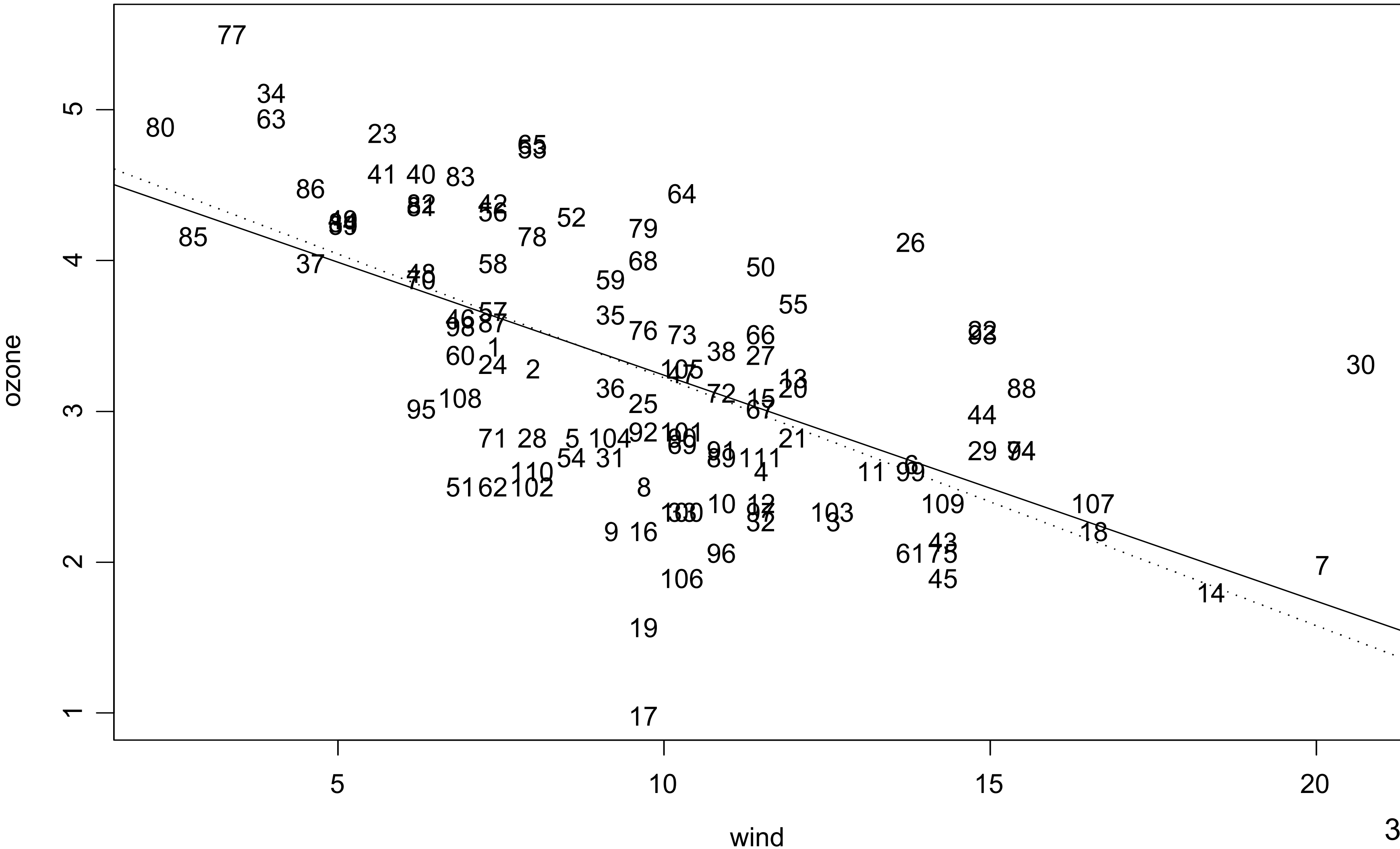
This graph was created using S-PLUS(R) Software. S-PLUS(R) is a registered trademark of Insightful Corporation.

Cook's Distance



This graph was created using S-PLUS(R) Software. S-PLUS(R) is a registered trademark of Insightful Corporation.

Plot of ozone vs. wind including fitted regression lines with and without observation 30 (simple linear regression)



This graph was created using S-PLUS(R) Software. S-PLUS(R) is a registered trademark of Insightful Corporation.

Remedies for Outliers

- Nothing?
- Data Transformation?
- Remove outliers?
- Robust Regression – weighted least squares: $b=(X'WX)^{-1}X'Wy$
- Minimize median absolute deviation

Collinearity

High correlation among the predictors can cause problems with least squares estimates (wrong signs, low t-values, unexpected results).

If predictors are centered and scaled to unit length, then X'X is the correlation matrix.

Diagonal elements of inverse of correlation matrix are called VIF's (variance inflation factors).

$$VIF_j = \frac{1}{1-R_j^2}, \; where \, \mathrm{R}_\mathrm{j}^2$$

is the coefficient of determination for the regression of the jth predictor on the remaining predictors

When $R_j^2 = .90$, VIF is about 10 and caution is advised. (Some authors say VIF = 5.) A large VIF indicates there is redundant information in the explanatory variables.

Why is this called the variance inflation factor?
We can show that

$$\operatorname{Var}\left(\hat{\beta}_j\right) = \frac{1}{1-R_j^2} \quad \frac{\sigma^2}{\sum_{i=1}^{n}\left(x_j - \overline{x}_j\right)^2}$$

$$= \mathrm{VIF}_j \ \operatorname{Var}\left[\hat{\beta}_j \text{ in simple regression}\right]$$

Thus VIF_j represents the variation inflation caused by adding all the variables other than x_j to the model.

Remedies for collinearity

1. Identify and eliminate redundant variables (large literature on this).

2. Modified regression techniques

   a. ridge regression, $b=(X'X+cI)^{-1}X'y$

3. Regress on orthogonal linear combinations of the explanatory variables

   a. principal components regression

4. Careful variable selection

Correlation and inverse of correlation matrix for air data set.

```
r<-cor(model.matrix(air.lm)[,-1])

> r
            x1          x2          x3
x1   1.0000000   0.2940876  -0.1273656
x2   0.2940876   1.0000000  -0.4971459
X3  -0.1273656  -0.4971459   1.0000000

> solve(r)
             x1          x2           x3
x1   1.09524102  -0.3357220  -0.02740677
x2  -0.33572201   1.4312012   0.66875638
x3  -0.02740677   0.6687564   1.32897882
>
```

Correlation and inverse of correlation matrix for mpg data set

```
r<-cor(model.matrix(auto1.lm)[,-1])

> r
              wt          I(wt^2)     I(wt^3)
    wt     1.0000000   0.9917756  0.9677228
I(wt^2)   0.9917756  1.0000000   0.9918939
I(wt^3)   0.9677228   0.9918939  1.0000000
```

```
➢ solve(r)
              wt          I(wt^2)      I(wt^3)
    wt      2000.377    -3951.728     1983.884
I(wt^2)   -3951.728     7868.535    -3980.575
I(wt^3)    1983.884    -3980.575     2029.459
```

Variable Selection

- We want a parsimonious model – as few variables as possible to still provide reasonable accuracy in predicting y.
- Some variables may not contribute much to the model.
- SSE never will increase if add more variables to model, however MSE=SSE/(n-k-1) may.
- Minimum MSE is one possible optimality criterion. However, must fit all possible subsets (2^k of them) and find one with minimum MSE.

Backward Elimination

1. Fit the full model (with all candidate predictors).
2. If P-values for all coefficients < α then stop.
3. Delete predictor with highest P-value
4. Refit the model
5. Go to Step 2.