
Lecture 6: Some Distributional Results for the GLM

Reading Assignment:

- Muller and Fetterman Chapter 3: “Some Distributions for the GLM”

In analysis with the GLM, we use three kinds of distributions: multivariate Gaussian, χ^2 , and F .

For now, assume all assumptions hold:

- HILE for estimation and
- Gaussian errors for testing.

A Full Rank Basis for Less Than Full Rank Models

If \mathbf{X} is $n \times p$, with $n \geq p$ and $\text{rank}(\mathbf{X}) = r \leq p$, then $\text{rank}(\mathbf{X}'\mathbf{X}) = r$.

If \mathbf{X} is less than full rank ($r < p$), then *collinearity* exists among columns of \mathbf{X} . If \mathbf{X} is less than full rank, then we say the model is also less than full rank. Also, $r = \text{rank}(\mathbf{X}) = \text{rank}(\mathbf{X}'\mathbf{X})$ is the # of estimable parameters.

For every less than full rank model, there exists a corresponding full rank model with r estimable parameters. That is, for less than full rank \mathbf{X} , there exists a $p \times r$ matrix \mathbf{V}_+ such that

$$\mathbf{X}_{n \times p} = \mathbf{X}_{*,(n \times r)} \mathbf{V}'_{+, (r \times p)} \quad \text{🗨️}$$

with $\text{rank}(\mathbf{X}_*) = \text{rank}(\mathbf{V}_+) = r < p$. \mathbf{X}_* provides a *full rank basis* for \mathbf{X} .


Suppose that we have the model

$$\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times p} \boldsymbol{\beta}_{p \times 1} + \boldsymbol{\varepsilon}_{n \times 1},$$

where $\text{rank}(\mathbf{X}) = r < p$.

Then, defining $\mathbf{X}_{*,(n \times r)} = \mathbf{X}_{n \times p} \mathbf{V}_{+, (p \times r)}$ with corresponding parameter vector $\boldsymbol{\beta}_{*,(r \times 1)}$, an equivalent full-rank model is given by

$$\mathbf{y}_{n \times 1} = \mathbf{X}_{*,(n \times r)} \boldsymbol{\beta}_{*,(r \times 1)} + \boldsymbol{\varepsilon}_{n \times 1},$$

with $\widehat{\boldsymbol{\beta}}_* = (\mathbf{X}'_* \mathbf{X}_*)^{-1} \mathbf{X}'_* \mathbf{y}$. 

Many possible choices of the matrix \mathbf{V}_+ exist, such as the set of eigenvectors of $\mathbf{X}'\mathbf{X}$ corresponding to non-zero eigenvalues.

Every parameter estimable in the original (less than full rank) model is also estimable in the full rank model, and any estimable parameter is expressible as a linear combination of the $\boldsymbol{\beta}_*$'s.

More About the Multivariate Gaussian Distribution

If Σ is not full rank (e.g., covariance matrix of residuals), then the multivariate Gaussian distribution is said to be *singular normal*. In this case, the density does not exist. (The multivariate Gaussian density exists only when Σ_z is non-singular. For example, if $y \sim N_1(\mu, 0)$, we have a discrete distribution with $P(y = \mu) = 1$, a point mass at μ .)

We define a singular multivariate Gaussian distribution for a vector \mathbf{z} in terms of a particular linear transformation $\mathbf{U}\mathbf{z}$ that leads to a full rank covariance matrix $\mathbf{U}\Sigma\mathbf{U}'$ so that we can define a density for the transformed random vector $\mathbf{U}\mathbf{z}$ (with redundancies eliminated).

Definition of Singular and Nonsingular Multivariate Gaussian

1. Full Rank Case

If $\text{rank}(\Sigma_z) = n$ and the density of z is

$$p(z) = (2\pi)^{-n/2} |\Sigma_z|^{-1/2} \exp[-(z - \mu_z)' \Sigma_z^{-1} (z - \mu_z)/2],$$

then z is distributed multivariate Gaussian, indicated

$$\mathbf{z}_{n \times 1} \sim \mathcal{N}_n(\mu_{z, (n \times 1)}, \Sigma_{z, (n \times n)}).$$

For example, with $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and HILE Gauss, $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ and $\mathbf{y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$.

2. Less than Full Rank Case

If $1 \leq \text{rank}(\Sigma_z) = n_* < n$, then a fixed matrix $\mathbf{U}_{n_* \times n}$ exists such that $\text{rank}(\mathbf{U}\Sigma_z\mathbf{U}') = n_*$ (full rank).

If the density of $U\mathbf{z}$ is

$$p(\mathbf{U}\mathbf{z}) = (2\pi)^{-\frac{n_*}{2}} |\mathbf{U}\Sigma_z\mathbf{U}'|^{-\frac{1}{2}} \exp \left[-\frac{1}{2}(\mathbf{U}\mathbf{z} - \mathbf{U}\boldsymbol{\mu}_z)'(\mathbf{U}\Sigma_z\mathbf{U}')^{-1}(\mathbf{U}\mathbf{z} - \mathbf{U}\boldsymbol{\mu}_z) \right],$$

then \mathbf{z} is distributed as a singular multivariate Gaussian, indicated by

$$\mathbf{z}_{n \times 1} \sim \mathcal{SN}_n(\boldsymbol{\mu}_{z, (n \times 1)}, \Sigma_{z, (n \times n)}).$$

We must add side conditions to specify $U_{n_* \times n}$ uniquely.

Example: Pick U to contain the eigenvectors for non-zero eigenvalues of Σ_z .

Sampling Distributions of Estimators

β Estimators $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

From properties of the multivariate Gaussian distribution,

$$\begin{aligned} E[\hat{\beta}] &= [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] E(\mathbf{y}) \\ &= [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] \mathbf{X}\beta = \beta, \end{aligned}$$

and

$$\begin{aligned} \text{Cov}(\hat{\beta}) &= [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] \text{Cov}(\mathbf{y}) [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' \\ &= [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] (\sigma^2\mathbf{I}_n) [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}, \end{aligned}$$

with

$$\hat{\beta} \sim \mathcal{N}_p(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

because $\hat{\beta}$ is a full rank linear transformation of the normal random vector \mathbf{y} .

🗨 For \mathbf{X} less than full rank and appropriate choice of generalized inverse, we have

$$\hat{\boldsymbol{\beta}} \sim \mathcal{SN}_p [(\mathbf{X}'\mathbf{X})^- (\mathbf{X}'\mathbf{X})\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^-] ,$$

which is a singular multivariate Gaussian because $(\mathbf{X}'\mathbf{X})$ is less than full rank.

For a Gaussian distribution, we know that most ($> 95\%$) of its mass lies within two standard deviations of the mean. So if our null hypothesis is for no effect of the covariate corresponding to β_j and

$$|\hat{\beta}_j| > 0 + 1.96 * \sqrt{\sigma^2(\mathbf{X}'\mathbf{X})_{j,j}^{-1}},$$

where $(\mathbf{X}'\mathbf{X})_{j,j}^{-1}$ refers to the $(j, j)^{th}$ element of $(\mathbf{X}'\mathbf{X})^{-1}$, then we have evidence that the covariate of interest has an effect on the response.

Example: Covariance Matrix of $\hat{\beta}$

The following R and PROC REG code below may be used to obtain the estimated covariance matrix of $\hat{\beta}$ for the ozone data with $O_{PERSONAL}$ as the outcome and $O_{OUTDOOR}$, O_{HOME} , and $TIME_{OUTDOORS}$ as predictors.

```

> ozone = read.table("ozone.txt", header = T, sep = " ") # space delimited
> n = nrow(ozone) # number of rows of ozone (obs)
> X = model.matrix(~ outdoor + home + time_out, data = ozone) # predictor matrix
> y = ozone$personal # response
> head(X)

  (Intercept)  outdoor   home time_out
1           1 35.87771 22.29      0.57
2           1 43.79189 13.97      0.90
3           1 49.81255 18.96      0.55
4           1 34.37366 22.27      0.17
5           1 45.95496 23.40      0.00
6           1 64.76558 39.62      0.30

> bhat = solve(t(X) %*% X) %*% t(X) %*% y # from notes
> yhat=X%*%bhat; # predicted values
> ehat=y-yhat; # residuals
> p=ncol(X); # num columns of X
> df=n-p; # df
> sse = t(y) %*% y - t(bhat) %*% t(X) %*% y # SSE
> mse=sse/df; # MSE
> covbhat=solve(t(X) %*% X)*as.numeric(mse) # scalar multiplication
> print(covbhat)

```



| | (Intercept) | outdoor | home | time_out |
|-------------|-------------|--------------|-------------|--------------|
| (Intercept) | 18.8534457 | -0.183111754 | -0.18665969 | -15.18741728 |
| outdoor | -0.1831118 | 0.008175203 | -0.00831528 | -0.06888717 |
| home | -0.1866597 | -0.008315280 | 0.02715354 | 0.07755804 |
| time_out | -15.1874173 | -0.068887173 | 0.07755804 | 59.43995287 |



We may also obtain the estimated covariance matrix of the β 's using the COVB option in SAS PROC REG.

```
proc reg;  
model personal=outdoor home time_out/covb;  
run;
```

The REG Procedure
 Model: MODEL1
 Dependent Variable: personal

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|-----------------|----|----------------|-------------|---------|--------|
| Model | 3 | 5034.90667 | 1678.30222 | 9.92 | <.0001 |
| Error | 60 | 10148 | 169.13652 | | |
| Corrected Total | 63 | 15183 | | | |
| Root MSE | | 13.00525 | R-Square | 0.3316 | |
| Dependent Mean | | 23.54578 | Adj R-Sq | 0.2982 | |
| Coeff Var | | 55.23389 | | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t |
|-----------|----|--------------------|----------------|---------|---------|
| Intercept | 1 | 3.78349 | 4.34206 | 0.87 | 0.3870 |
| outdoor | 1 | 0.09142 | 0.09042 | 1.01 | 0.3160 |
| home | 1 | 0.59544 | 0.16478 | 3.61 | 0.0006 |

| | | | | | |
|----------|---|----------|---------|------|--------|
| time_out | 1 | 13.64454 | 7.70973 | 1.77 | 0.0818 |
|----------|---|----------|---------|------|--------|

Model: MODEL1

Dependent Variable: personal Personal Ozone Exposure (ppb)

Covariance of Estimates

| Variable | Label | Intercept | outdoor |
|-----------|---------------------------------------|--------------|--------------|
| Intercept | Intercept | 18.853445739 | -0.183111754 |
| outdoor | Outdoor Ozone Concentration (ppb) | -0.183111754 | 0.0081752029 |
| home | Home Indoor Ozone Concentration (ppb) | -0.186659692 | -0.00831528 |
| time_out | Proportion of Time Spent Outdoors | -15.18741728 | -0.068887173 |

Covariance of Estimates

| Variable | Label | home | time_out |
|-----------|---------------------------------------|--------------|--------------|
| Intercept | Intercept | -0.186659692 | -15.18741728 |
| outdoor | Outdoor Ozone Concentration (ppb) | -0.00831528 | -0.068887173 |
| home | Home Indoor Ozone Concentration (ppb) | 0.0271535425 | 0.0775580389 |
| time_out | Proportion of Time Spent Outdoors | 0.0775580389 | 59.439952865 |

θ Estimator

Again using properties of the multivariate Gaussian distribution, for $\mathbf{C}_{a \times p}$ a matrix of constants and $\boldsymbol{\theta} = \mathbf{C}\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\theta}} \sim N_a(\boldsymbol{\theta}, \sigma^2 \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}').$$

We should be careful to verify that $\boldsymbol{\theta}$ is both estimable and testable.

Example: Estimating θ

Suppose we wish to test the hypothesis that the $O_{OUTDOOR}$, O_{HOME} , and $TIME_{OUTDOORS}$ coefficients are all equal in the ozone data. So

$$\mathbf{C} = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix},$$

$$\boldsymbol{\theta} = \begin{bmatrix} \beta_1 - \beta_2 \\ \beta_2 - \beta_3 \end{bmatrix},$$

and

$$\hat{\boldsymbol{\theta}} \sim \mathcal{N}_2(\boldsymbol{\theta}, \sigma^2 \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{C}').$$

We may obtain the estimated covariance matrix of $\hat{\boldsymbol{\theta}}$ with the following additional R code.

```
> C=matrix(c(0, 1, -1, 0, 0, 0, 1, -1), nrow = 2, byrow = T);
> print(C)

      [,1] [,2] [,3] [,4]
[1,]    0    1   -1    0
[2,]    0    0    1   -1

> covhat=as.numeric(mse)*C*%solve(t(X) %*% X)%*%t(C) # sigma^2 * M
> print(covhat)

      [,1]      [,2]
[1,] 0.05195931 0.1109764
[2,] 0.11097639 59.3119903
```

Predicted Values: Conditional Means and Future Observations

For the GLM

$$\begin{aligned} E(\mathbf{y}_{n \times 1}) &= E(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\mu}(\mathbf{y} \mid \mathbf{X}). \end{aligned}$$



We write the estimator of the expected values as

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} = [\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] \mathbf{y} \\ &= \mathbf{H}\mathbf{y}. \end{aligned}$$



Prediction is often an important goal of modeling. For example, an actuary might want to predict medical costs for health insurance patients covered by a given insurance plan given their characteristics (including BMI, smoking status, and age).



Caution: prediction outside the range of observed data can be extremely dangerous!

To find the distribution of $\hat{\mathbf{y}}$,



$$\begin{aligned} E(\hat{\mathbf{y}}) &= \mathbf{H}E(\mathbf{y}) \\ &= [\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] \mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}. \end{aligned}$$

In addition, for $\hat{\mathbf{y}}$ an estimated conditional mean response,

$$\begin{aligned} \text{cov}(\hat{\mathbf{y}}) &= \mathbf{H}(\sigma^2\mathbf{I})\mathbf{H}' \\ &= \sigma^2\mathbf{H}\mathbf{H}' \text{  } \sigma^2\mathbf{H} \\ &= \sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \text{  } \end{aligned}$$

as \mathbf{H} is symmetric and idempotent.

As a transformation of a **singular** multivariate Gaussian ($\mathbf{H}_{n \times n}$ has rank $p < n$ when \mathbf{X} is full rank), the distribution of $\hat{\mathbf{y}}$ is given by

$$\hat{\mathbf{y}} \sim SN_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{H}).$$

In contrast, consider predicting future observations at time f with covariate values \mathbf{X}_f . In this case, we predict an *individual outcome* rather than a mean outcome. The predicted future response $\hat{\mathbf{y}}_f$ involves variance due to

1. estimating β (the fitted line is not the exact true line)
2. observing ε_f in the future sample (even if the fitted line is the exact true line, there is still variability about it), and
3. changing the design (predictor values).

Now

$$\hat{\mathbf{y}}_f = \mathbf{X}_f \hat{\beta},$$

where \mathbf{X}_f is $n_f \times p$, while \mathbf{X} is $n \times p$. ($\mathbf{X}_f \neq \mathbf{X}$). These predictions will have additional error (as opposed to estimated conditional means) ε_f : the errors to be observed at future time f ($\varepsilon_f \neq \varepsilon$), and ε_f are independent of the errors at the current time.

Thus we have

$$\hat{\mathbf{y}}_f \sim \mathcal{SN}_{n_f}\{\mathbf{X}_f\boldsymbol{\beta}, \sigma^2[\mathbf{X}_f(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_f' + \mathbf{I}]\}$$

$\mathbf{X}_f = \mathbf{X}$ implies

$$\hat{\mathbf{y}}_f \sim \mathcal{SN}_n\{\mathbf{X}\boldsymbol{\beta}, \sigma^2[\mathbf{H} + \mathbf{I}]\},$$

so we see the additional uncertainty in predicting a particular future observation rather than estimating the mean.

It would be unambiguous to discuss estimated conditional means, $\hat{\boldsymbol{\mu}}|\mathbf{X}$, or predicted future observations, $\hat{\mathbf{y}}_f$; however, describing $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ as predicted values has become standard, so that you must rely on the context to decide whether estimated conditional means or predicted future observations are being discussed.

Definitions and Properties of Residuals

The residuals $\hat{\varepsilon}$ are defined as

$$\begin{aligned}\hat{\varepsilon} &= (\mathbf{y} - \hat{\mathbf{y}}) \\ &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= (\mathbf{I} - \mathbf{H})\mathbf{y},\end{aligned}$$

with

$$\hat{\varepsilon} \sim \mathcal{SN}_n(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{H})).$$

$\hat{\varepsilon}_i$ is not independent of $\hat{\varepsilon}_j$ ($i \neq j$) unless $p = 0$ or $n \rightarrow \infty$.

$\hat{\varepsilon}_i \sim \mathcal{N}[0, \sigma^2(1 - h_i)]$, where $h_i = \mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i'$, the (i, i) element of $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

If the model spans an intercept, $\sum_{i=1}^n \hat{\varepsilon}_i = 0$.

Residual Variance


Let $r = \text{rank}(\mathbf{X})$. Recall that

$$\begin{aligned}\hat{\sigma}^2 &= \frac{SSE}{n - r} = \frac{\hat{\boldsymbol{\varepsilon}}' \hat{\boldsymbol{\varepsilon}}}{n - r} \\ &= \frac{\mathbf{y}'(\mathbf{I} - \mathbf{H})'(\mathbf{I} - \mathbf{H})\mathbf{y}}{n - r} = \frac{\mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}}{n - r}\end{aligned}$$

as $(\mathbf{I} - \mathbf{H})$ is symmetric and idempotent.

Now

$$\hat{\sigma}^2 \left(\frac{n - r}{\sigma^2} \right) \sim \chi^2(n - r).$$

Because $E(\chi^2(\nu)) = \nu$, $\hat{\sigma}^2$ is unbiased. In addition, $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are statistically independent. 

Standardized Residuals

Replacing σ^2 with $\hat{\sigma}^2$ leads some distributions to change from Gaussian to Student's T .

Because $\hat{\varepsilon}_i \sim \mathcal{N}[0, \sigma^2(1 - h_i)]$, it follows that

$$\frac{\hat{\varepsilon}_i}{\sqrt{\sigma^2(1 - h_i)}} \sim \mathcal{N}(0, 1).$$

Define the standardized residual as

$$r_i = \frac{\hat{\varepsilon}_i}{\sqrt{\hat{\sigma}^2(1 - h_i)}} \quad \text{🗨️}$$

The standardized residuals do not follow a T distribution; in fact,

$$\frac{r_i^2}{n - r} \sim \text{Beta} \left(\frac{1}{2}, \frac{n - r - 1}{2} \right),$$

and the r_i are bounded between $-\sqrt{n - r}$ and $\sqrt{n - r}$, where $r = \text{rank}(\mathbf{X})$.

Example: Calculating Standardized Residuals

To calculate the standardized residuals for the ozone example, we add the following R code to the previous code. In addition, we may obtain the standardized residuals from PROC REG by using the “R” option with the model statement.

```
> H=X%*%solve(t(X) %*% X) %*% t(X) # calculate Hat Matrix
> h_i=diag(H) # get the diagonal
> r_i=ehat/(sqrt(as.numeric(mse)*(1-h_i)));
> head(r_i,5)

      [,1]
1 -0.1441837
2 -1.1486537
3 -0.4461650
4 -1.4976573
5  0.5791377
```

The first five residuals are printed above. You may also get these in R from your fitted linear model using the MASS package.

```
> out = lm(personal ~ outdoor + home + time_out, data = ozone)
> # equivalent to out = lm(y ~ X - 1)
```

```
> summary(out)
```

```
Call:
```

```
lm(formula = personal ~ outdoor + home + time_out, data = ozone)
```

```
Residuals:
```

| Min | 1Q | Median | 3Q | Max |
|---------|--------|--------|-------|--------|
| -25.930 | -7.855 | -4.257 | 4.880 | 36.295 |

```
Coefficients:
```

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|----------|------------|---------|--------------|
| (Intercept) | 3.78349 | 4.34206 | 0.871 | 0.387032 |
| outdoor | 0.09142 | 0.09042 | 1.011 | 0.316031 |
| home | 0.59544 | 0.16478 | 3.613 | 0.000619 *** |
| time_out | 13.64454 | 7.70973 | 1.770 | 0.081845 . |

```
---
```

```
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
```

```
Residual standard error: 13.01 on 60 degrees of freedom
```

```
Multiple R-squared:  0.3316,      Adjusted R-squared:  0.2982
```

```
F-statistic: 9.923 on 3 and 60 DF,  p-value: 2.094e-05
```

```
> library(MASS) #load MASS package, or install using install.packages("MASS")
```

```
> r_i_MASS = stdres(out) # stdres function from MASS
> head(r_i_MASS,5)
```

| 1 | 2 | 3 | 4 | 5 |
|------------|------------|------------|------------|-----------|
| -0.1441837 | -1.1486537 | -0.4461650 | -1.4976573 | 0.5791377 |

The following PROC REG code obtain the residuals.

```
proc reg;  
model personal=outdoor home time_out/r;  
run;
```

The first five residuals are below. Note that the standardized residuals are labeled 'student residual.'

The REG Procedure

Model: MODEL1

Dependent Variable: personal Personal Ozone Exposure (ppb)

Output Statistics


| Obs | Dep Var personal | Predicted Value | Std Error Mean Predict | Residual | Std Error Residual | Student Residual |
|-----|---------------------|--------------------|---------------------------|----------|-----------------------|---------------------|
| 1 | 26.2900 | 28.1131 | 3.0429 | -1.8231 | 12.644 | -0.144 |
| 4 | 3.3000 | 22.5059 | 2.1640 | -19.2059 | 12.824 | -1.498 |
| 5 | 29.2800 | 21.9179 | 2.7454 | 7.3621 | 12.712 | 0.579 |
| 9 | 28.5500 | 20.0446 | 5.9850 | 8.5054 | 11.546 | 0.737 |
| 13 | 38.2800 | 37.0471 | 5.1369 | 1.2329 | 11.948 | 0.103 |

Jackknifing

Jackknifing usually involves computing a statistic with one observation deleted from a sample, once for each observation. Let the subscript $(-i)$ indicate having deleted the i th observation. For example, $\mathbf{y}_{(-i)}$ and $\mathbf{X}_{(-i)}$ have $n - 1$ rows.

Compute

- $\hat{\boldsymbol{\beta}}_{(-i)} = (\mathbf{X}'_{(-i)}\mathbf{X}_{(-i)})^{-1}\mathbf{X}'_{(-i)}\mathbf{y}_{(-i)}$
- $\hat{\mathbf{y}}_{(-i)} = \mathbf{X}\hat{\boldsymbol{\beta}}_{(-i)}$
- $\hat{\boldsymbol{\varepsilon}}_{(-i)} = \mathbf{y}_{(-i)} - \hat{\mathbf{y}}_{(-i)}$ and
- $\hat{\sigma}^2_{(-i)} = \hat{\boldsymbol{\varepsilon}}'_{(-i)}\hat{\boldsymbol{\varepsilon}}_{(-i)} / (n - r - 1),$

where $r = \text{rank}(\mathbf{X})$. 

If $\hat{\boldsymbol{\beta}}_{(-i)}$ differs greatly from $\hat{\boldsymbol{\beta}}$, then we see that observation i has a good deal of influence on the analysis.

Studentized Residuals

Standardizing the jackknifed residuals (by dividing by an estimate of the standard deviation) yields a set of residuals (the *Studentized residuals*) which follow a Student's T distribution:

$$r_{(-i)} = \frac{\hat{\varepsilon}_i}{\sqrt{\hat{\sigma}_{(-i)}^2 (1 - h_i)}} \sim T(n - r - 1)$$

Chapter 7 in MF has extensive discussion of this most important tool for assumption evaluation.

Example: Calculating Studentized Residuals

The following code may be added to the previous code to obtain the studentized residuals in PROC IML. The code uses the following formula for the studentized residuals given in Muller and Fetterman:

$$r_{(-i)} = r_i \left(\frac{(n - r) - 1}{(n - r) - r_i^2} \right)^{\frac{1}{2}},$$

where $r = \text{rank}(\mathbf{X})$.

```
> library(Matrix) #install.packages("Matrix")
> r_i2=r_i^2; # square each element of r_i
> r = rankMatrix(X) # full rank
> r_mi=r_i*sqrt((n-r-1)/(n-r-r_i2))
> head(r_mi,5)
```

```
      [,1]
1 -0.1430019
2 -1.1517756
3 -0.4431671
4 -1.5136869
5  0.5759032
```



The following SAS PROC REG code may be used to obtain the studentized residuals.



```
proc reg;  
model personal=outdoor home time_out;  
output out=resid student=standresid rstudent=studresid;  
run;
```



```
proc print data=resid;  
var personal standresid studresid;  
run;
```

The first five observations are given below.

| Obs | personal | standresid | studresid |
|-----|----------|------------|-----------|
| 1 | 26.29 | -0.14418 | -0.14300 |
| 4 | 3.30 | -1.49766 | -1.51369 |
| 5 | 29.28 | 0.57914 | 0.57590 |
| 9 | 28.55 | 0.73663 | 0.73379 |
| 13 | 38.28 | 0.10319 | 0.10233 |

Next: Multiple Regression

Reading Assignment:

- Muller and Fetterman, Chapter 4: “Multiple Regression”