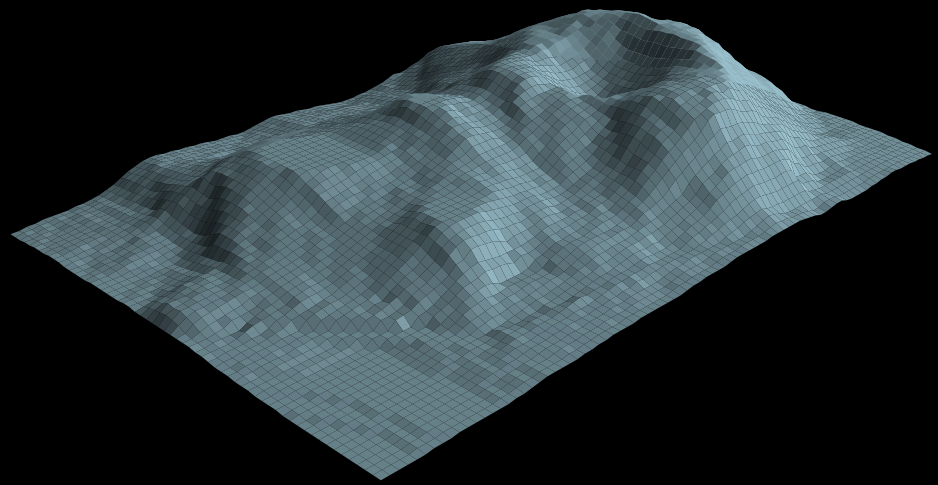
Introducing 

JEFF GILL
Department of Political Science
Division of Biostatistics
Department of Surgery (Public Health Sciences)
Washington University, St. Louis



Reminders

- Thank R Its Friday (TRIF) is supported by the Washington University TREC Center (`www.obesity-cancer.wustl.edu`), School of Medicine.
- Meeting times and topics (3-4PM):
  - January 31, 2014: Analysis of Variance
  - February 28, 2014: Logistic Regression
  - March 28, 2014: Principle Components Analysis
  - April 25, 2014: Survival Analysis
  - May 30, 2014: Nonparametric Data Analysis
- The slides for today are available at `http://jgill.wustl.edu/slides/trif3.pdf`.

General Modeling Language

- Basic Structure:

```
OV ~ EV1 + EV2
```

  where the tilde is a function that saves the formula as an unevaluated expression: formula object.

- Note that no actual "adding" is being done here in the arithmetic sense.

- A constant is automatically implied, same as:

```
OV ~ 1 + EV1 + EV2.
```

  but we can explicitly exclude the constant by:

```
OV ~ -1 + EV1 + EV2.
OV ~  0 + EV1 + EV2.
```

- Special characters:

```
+ - : * / . ^
```

Inline Math Formulas

```
log(OV) ~ EV1 + EV2

OV ~ exp(EV1) + cos(EV2)

OV ~ I(EV1/2) + sqrt(EV2 - mean(EV2))

osha.ols <- lm(INSPT ~ AP + (DI>40) ,data=osha.df)

osha.ols <- lm(INSPT ~ AP + cut(DI,3) ,data=osha.df)
```

Specifying Interactions

```
OV ~ EV1 + EV2 + EV1:EV2

OV ~ EV1 * EV2

osha.ols <- lm(INSPT ~ D1 + D1:D2, data=osha.df)

osha.ols <- lm(INSPT ~ AP * DI, data=osha.df)

osha.ols <- lm(INSPT ~ AP * DI * SIC1, data=osha.df)
```

Nesting

```
OV ~ EV1 + county %in% state

OV ~ EV1 + state/county

osha.ols <- lm(INSPT ~ D1 + D2 %in% D1, data=osha.df)

osha.ols <- lm(INSPT ~ D1/D2, data=osha.df)
```

Gauss-Markov Assumptions for Classical Linear Regression

- Functional Form: $\underset{(n\times 1)}{\mathbf{Y}} = \underset{(n\times k)(k\times 1)}{\mathbf{X}\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\epsilon}}$ ($\mathbf{X}$ has a leading column of 1's)
- Mean Zero Errors: $\mathbf{E}[\boldsymbol{\epsilon}] = \mathbf{0}$
- Homoscedasticity: $\mathtt{Var}[\boldsymbol{\epsilon}] = \sigma^2\mathbf{I}$
- Non-Correlated Errors: $\mathtt{Cov}[\epsilon_i, \epsilon_j] = 0, \qquad \forall i \neq j$
- Exogeneity of Explanatory Variables: $\mathtt{Cov}[\epsilon_i, \mathbf{X}] = 0, \qquad \forall i$

  - Note that every one of these lines has $\boldsymbol{\epsilon}$ in it, meaning that these are assumptions about the underlying population values.

Other Considerations

- Requirements:
  - conformability of matrix/vector objects
  - $\mathbf{X}$ is full rank: k, so $\mathbf{X}'\mathbf{X}$ is invertible (nonsingular).
  - identification condition: not all points lie on a vertical line.
- Freebee: eventual normality. . . $\boldsymbol{\epsilon}|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$
- Toughness: the linear model is both *robust* to minor violations of the Gauss-Markov assumptions and *resistant* to outlying values.

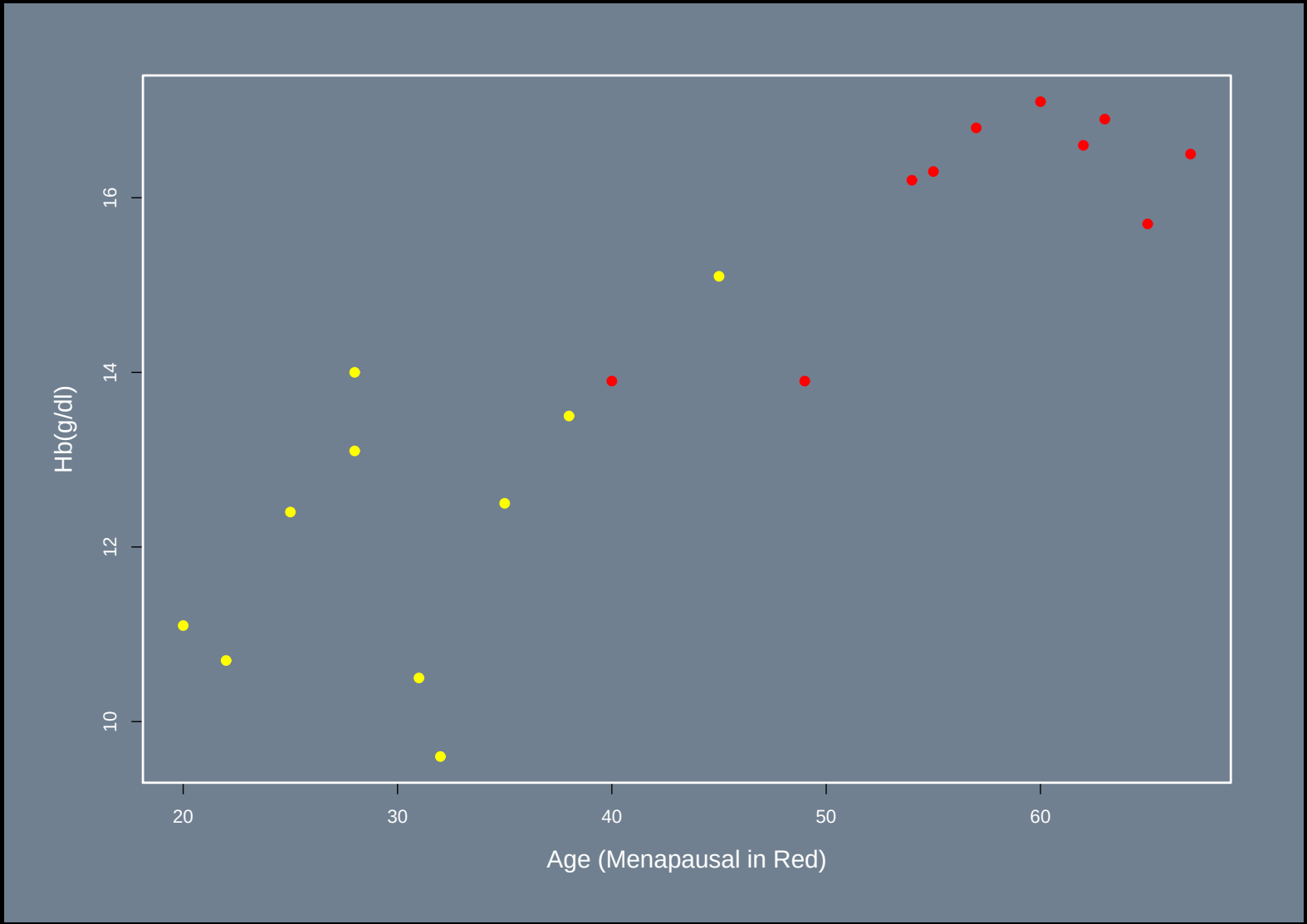
Anaemia Data

- Consider a study of anaemia in women in a given clinic where 20 cases are chosen at random from the full study to get the data here.
- From a blood sample we get:
  - haemoglobin level (Hb) in grams per deciliter (12–15 g/dl is normal in adult females)
  - packed cell volume (PCV) in percent of blood volume that is occupied by red blood cells (also called hematocrit, Ht or HCT, or erythrocyte volume fraction, EVF). 38% to 46% is normal in adult females.
- We also have:
  - age in years
  - menopausal (0=no, 1=yes)
- There is an obvious endogeneity problem in modeling Hb(g/dl) versus PCV(%).

Anaemia Data

Subject	Hb(g/dl)	PCV(%)	Age	Menopausal
1	11.1	35	20	0
2	10.7	45	22	0
3	12.4	47	25	0
4	14.0	50	28	0
5	13.1	31	28	0
6	10.5	30	31	0
7	9.6	25	32	0
8	12.5	33	35	0
9	13.5	35	38	0
10	13.9	40	40	1
11	15.1	45	45	0
12	13.9	47	49	1
13	16.2	49	54	1
14	16.3	42	55	1
15	16.8	40	57	1
16	17.1	50	60	1
17	16.6	46	62	1
18	16.9	55	63	1
19	15.7	42	65	1
20	16.5	46	67	1

Scatterplot of the Anaemia Data

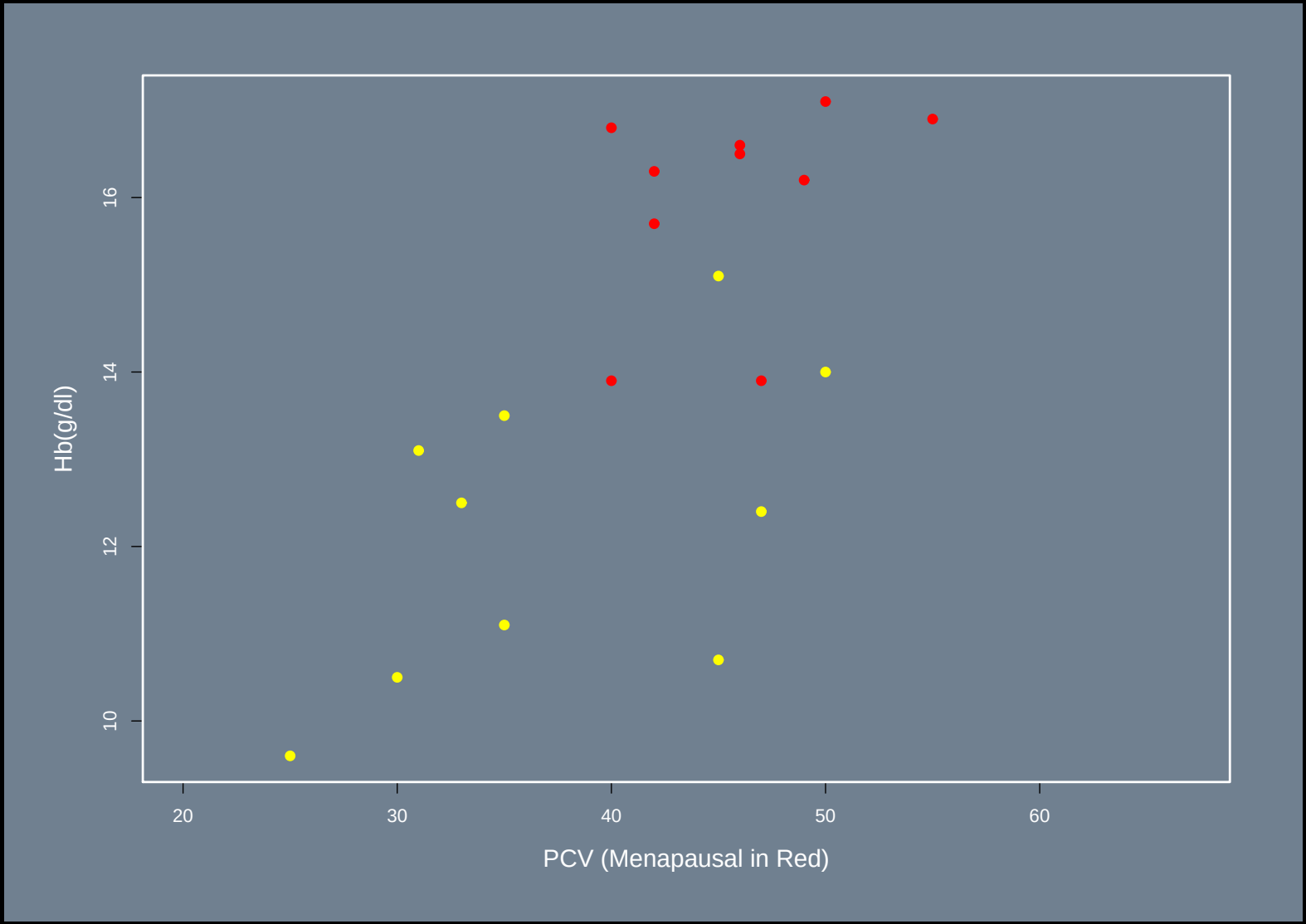


Scatterplot of the Anaemia Data

```
anaemia <- read.table("http://jgill.wustl.edu/data/anaemia.dat",
                      header=TRUE,row.names=1)

postscript("Class.PreMed.Stats/Images/anaemia1.fig.ps")
par(mfrow=c(1,1),mar=c(5,5,2,2),lwd=2,col.axis="white",col.lab="white",
    col.sub="white",col="white",bg="slategray", cex.lab=1.3)
plot(anaemia$Age[anaemia$Menapause==0],anaemia$Hb[anaemia$Menapause==0],pch=19,
     col="yellow", xlim=range(anaemia$Age),ylim=range(anaemia$Hb),
     xlab="Age (Menapausal in Red)",ylab="Hb(g/dl)")
points(anaemia$Age[anaemia$Menapause==1],anaemia$Hb[anaemia$Menapause==1],pch=19,
     col="red")
dev.off()
```

Scatterplot of the Anaemia Data



Scatterplot of the Anaemia Data

```
postscript("Class.PreMed.Stats/Images/anaemia2.fig.ps")
par(mfrow=c(1,1),mar=c(5,5,2,2),lwd=2,col.axis="white",col.lab="white",
    col.sub="white",col="white",bg="slategray", cex.lab=1.3)
plot(anaemia$PCV[anaemia$Menapause==0],anaemia$Hb[anaemia$Menapause==0],pch=19,
     col="yellow", xlim=range(anaemia$Age),ylim=range(anaemia$Hb),
     xlab="PCV (Menapausal in Red)",ylab="Hb(g/dl)")
points(anaemia$PCV[anaemia$Menapause==1],anaemia$Hb[anaemia$Menapause==1],pch=19,
       col="red")
dev.off()
```

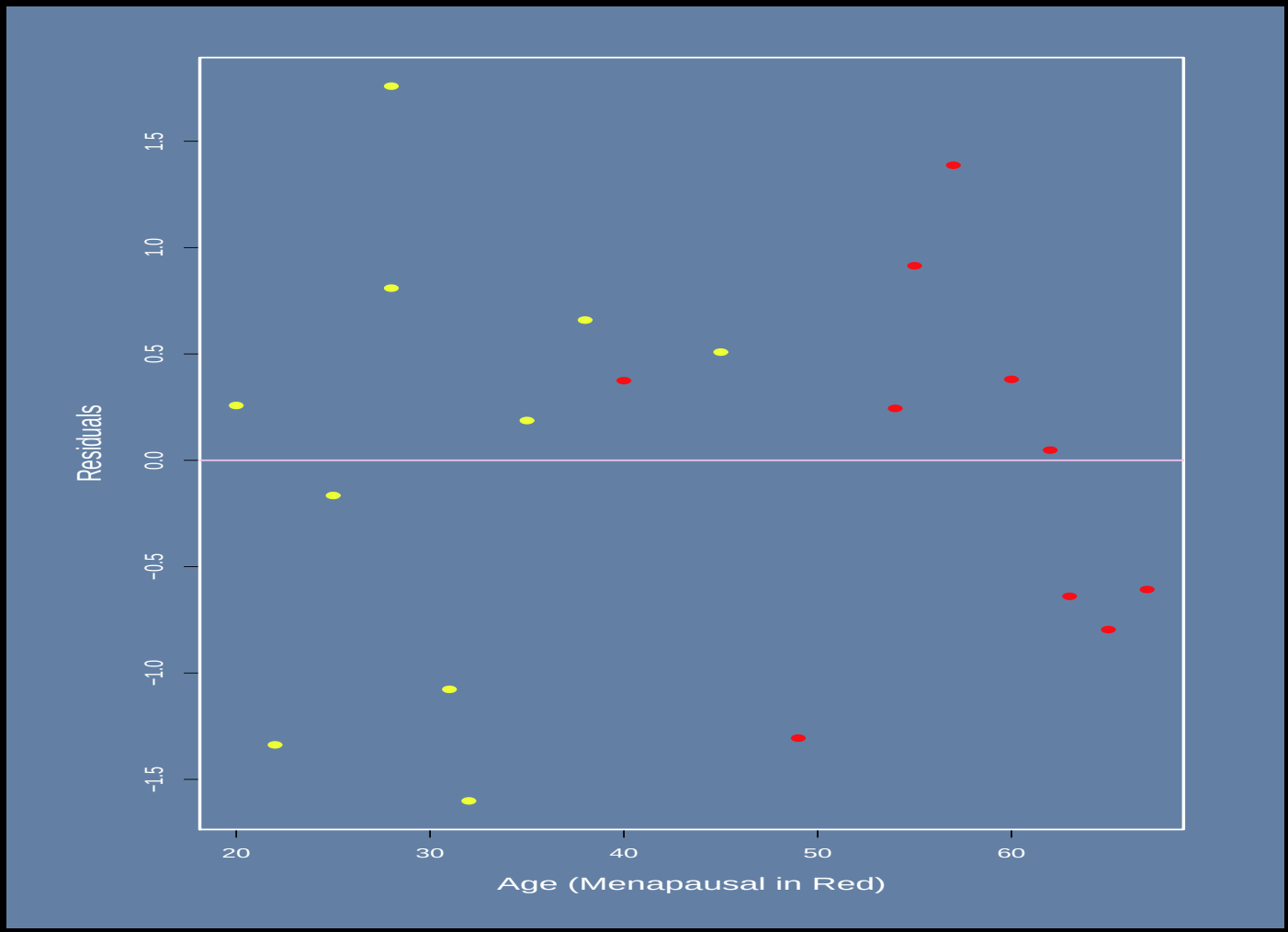
Linear Model of Anaemia

```
anaemia.lm <- lm(Hb ~ PCV + Age + Menapause, data=anaemia)
summary(anaemia.lm)
Residuals:
   Min     1Q Median     3Q    Max
-1.601 -0.678  0.216  0.546  1.759

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.2146     1.5718    3.32   0.0044
PCV           0.0973     0.0346    2.81   0.0125
Age           0.1110     0.0303    3.66   0.0021
Menapause    -0.0241     0.9540   -0.03   0.9802

Residual standard error: 1.01 on 16 degrees of freedom
Multiple R-squared: 0.851,   Adjusted R-squared: 0.823
F-statistic: 30.5 on 3 and 16 DF,  p-value: 7.46e-07
```

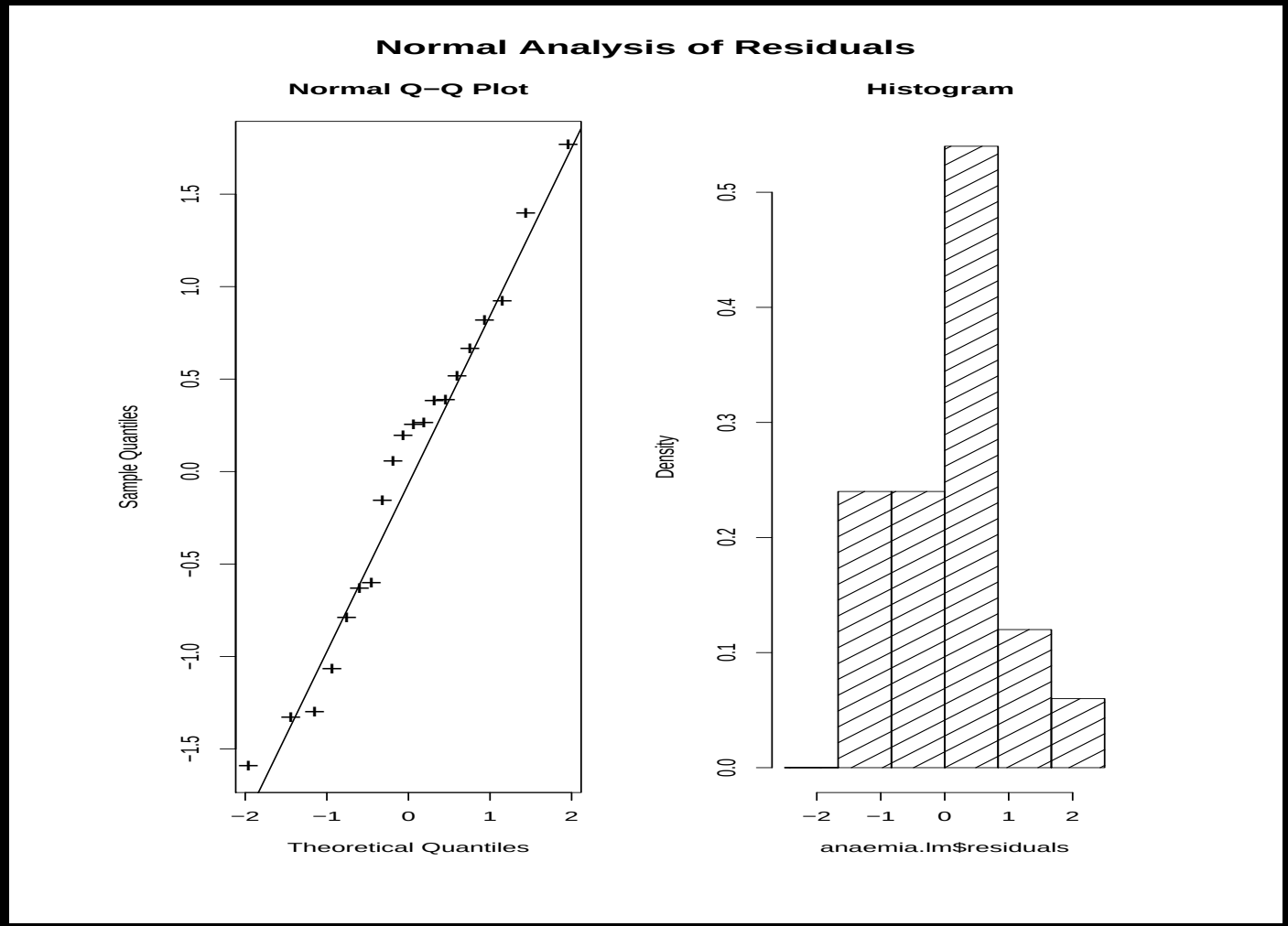
Trends In The Residuals



Trends In The Residuals

```
postscript("Class.PreMed.Stats/Images/anaemia.resids.fig.ps")
par(mfrow=c(1,1),mar=c(5,5,2,2),lwd=2,col.axis="white",col.lab="white",
    col.sub="white",col="white",bg="slategray", cex.lab=1.3)
plot(anaemia$Age[anaemia$Menapause==0],anaemia.lm$residuals[anaemia$Menapause==0],
     pch=19,col="yellow", xlim=range(anaemia$Age),ylim=range(anaemia.lm$residuals),
     xlab="Age (Menapausal in Red)",ylab="Residuals")
points(anaemia$Age[anaemia$Menapause==1],anaemia.lm$residuals[anaemia$Menapause==1],
       pch=19,col="red")
abline(h=0,lwd=2,col="pink")
dev.off()
```

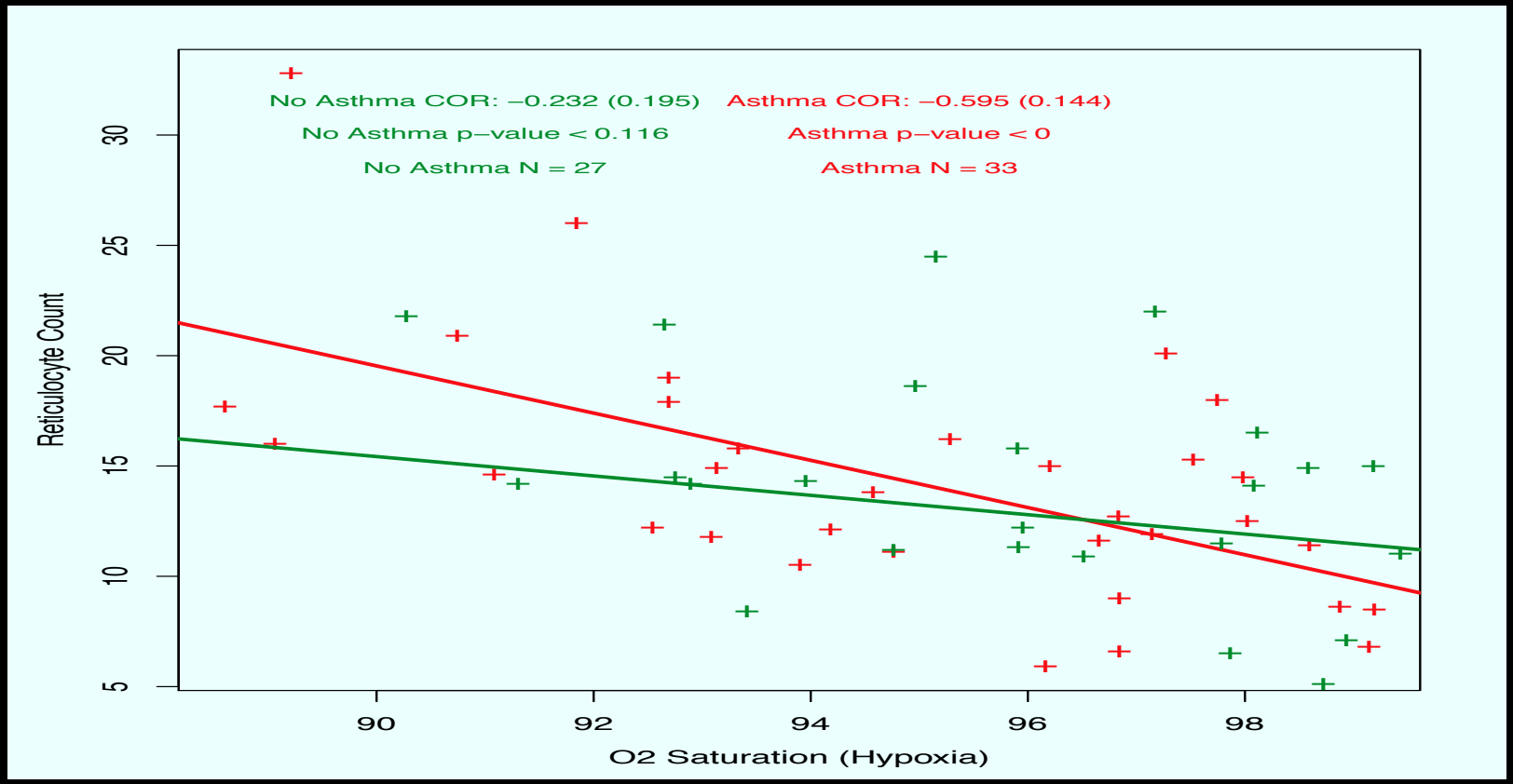
Normality Of The Residuals



Normality Of The Residuals

```
postscript("Class.Stat.Comp/Images/anaemia.qqnorm.ps")
par(mfrow=c(1,2),oma=c(2,2,2,2),bg="white")
qqnorm(anaemia.lm$residuals,pch="+",cex=1.5)
qqline(anaemia.lm$residuals)
hist(anaemia.lm$residuals,breaks=seq(-2.5,2.5,length=7),freq=FALSE,density=7,
     main="Histogram ")
mtext(side=3,font=2,cex=1.5,"Normal Analysis of Residuals",outer=TRUE)
dev.off()
```

Another Regression Plot



Code For This Graph With Regression

```
sickle0 <- sickle[sickle$asthma == "noasthma" | sickle$asthma == "noclass",]
sickle1 <- sickle[sickle$asthma == "asthma",]
pdf("Article.Sickle.Cell/linear.plot.asthma.pdf")
par(mfrow=c(1,1),mar=c(4,4,2,2),col.axis="black",col.lab="black",col.sub="black",
    col="black",bg="white")
plot(sickle1$sac2.satmeantst,sickle1$reticulocyte,pch="+",xlab="",ylab="",
    cex=1.25,col="red")
mtext(side=1,line=2.5,"O2 Saturation (Hypoxia)"); mtext(side=2,line=2.5,
    "Reticulocyte Count")
abline(lm(sickle1$reticulocyte ~ sickle1$sac2.satmeantst),col="red",lwd=3)
sickle.cor <- cor(sickle1$reticulocyte, sickle1$sac2.satmeantst)
se.cor <- sqrt((1-sickle.cor^2)/(nrow(sickle1)-2))
p.cor <- pnorm(sickle.cor/se.cor)
text(95,31.5,paste("Asthma COR: ",round(sickle.cor,3)," (",round(se.cor,3),")",sep=""),
    cex=0.8,col="red")
text(95,30,paste("Asthma p-value < ",round(p.cor,3),sep=""),cex=0.8,col="red")
text(95,28.5,paste("Asthma N = ",nrow(sickle1),sep=""),cex=0.8,col="red")
```

Code For This Graph With Regression, Continued

```
points(sickle0$sac2.satmeantst,sickle0$reticulocyte,pch="+",xlab="",ylab="",cex=1.25,
    col="forestgreen")
abline(lm(sickle0$reticulocyte ~ sickle0$sac2.satmeantst),col="forestgreen",lwd=3)
sickle.cor <- cor(sickle0$reticulocyte, sickle0$sac2.satmeantst)
se.cor <- sqrt((1-sickle.cor^2)/(nrow(sickle0)-2))
p.cor <- pnorm(sickle.cor/se.cor)
text(91,31.5,paste("No Asthma COR: ",round(sickle.cor,3)," (",round(se.cor,3),")",
    sep=""), cex=0.8,col="forestgreen")
text(91,30,paste("No Asthma p-value < ",round(p.cor,3),sep=""),cex=0.8,
    col="forestgreen")
text(91,28.5,paste("No Asthma N = ",nrow(sickle0),sep=""),cex=0.8,col="forestgreen")
dev.off()
```

What Happens When it Doesn't Work?

- The New York Times Magazine, August 7, 2011, page 13.
- 24 countries: average survey review of restaurant service quality and a tipping index from three travel etiquette web sites.
- The data

Country	Quality	Tip	Country	Quality	Tip
Japan	4.4	0.00	Thailand	3.9	0.03
Canada	3.7	0.16	New_Zealand	3.7	0.07
UAE	3.6	0.10	Germany	3.6	0.08
USA	3.6	0.18	South_Africa	3.5	0.11
Australia	3.4	0.08	Argentina	3.4	0.10
Morocco	3.4	0.07	Turkey	3.4	0.08
India	3.3	0.10	Brazil	3.3	0.07
Vietnam	3.2	0.05	England	3.2	0.10
Greece	3.2	0.08	Spain	3.1	0.08
France	3.1	0.08	Italy	3.0	0.07
Egypt	3.0	0.08	Mexico	3.0	0.13
China	2.9	0.03	Russia	1.7	0.10

What Happens When it Doesn't Work?

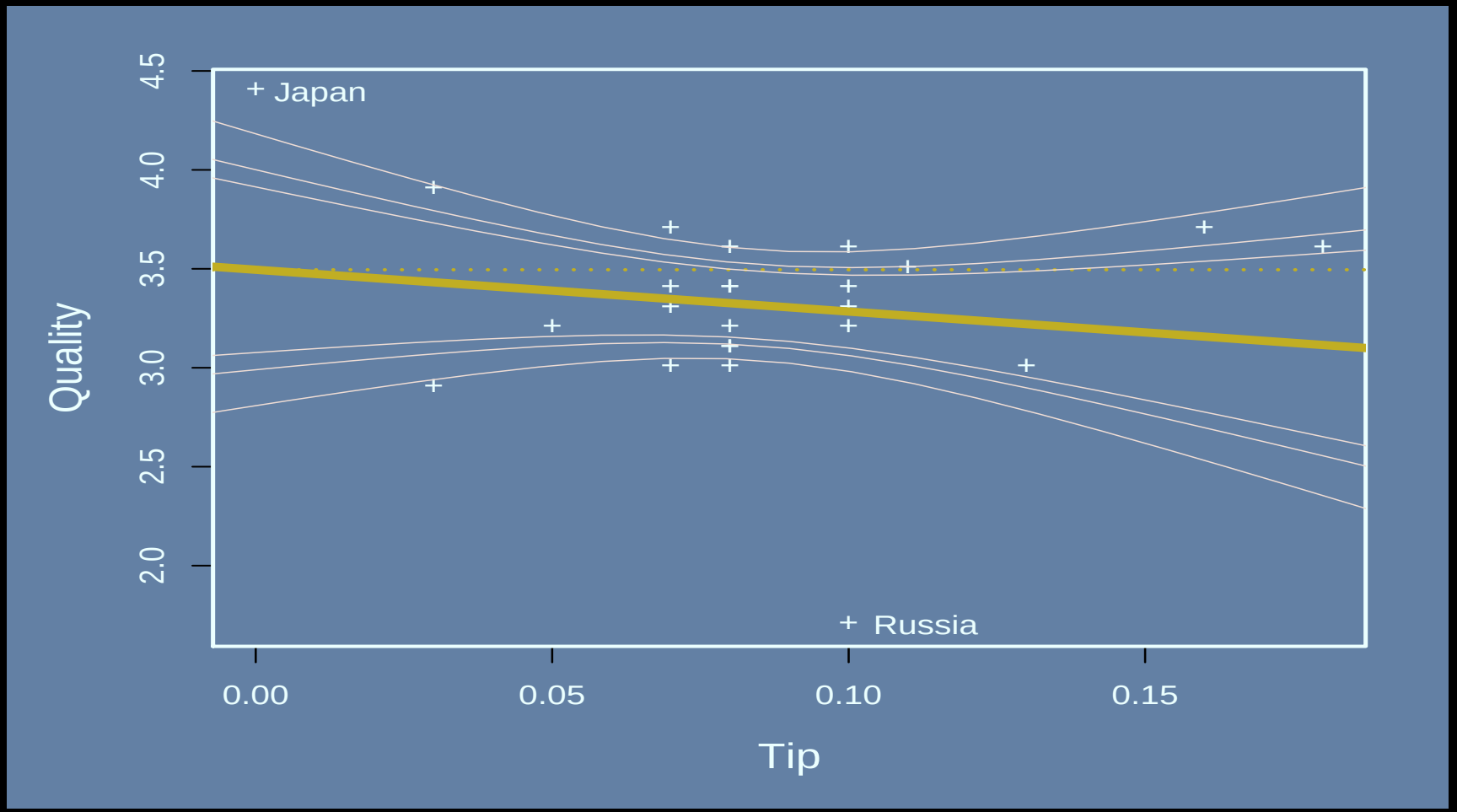
```
service <- read.table("http://jgill.wustl.edu/data/service.dat",
                      header=TRUE,row.names=1)
service.lm <- lm(Quality~Tip,data=service)
source("./Class.MLE/graph.summary.R")
graph.summary(service.lm)
```

```
Family: gaussian
Link function: identity

             Coef Std.Err. 0.95 Lower 0.95 Upper   CIs:  ZE+RO
(Intercept)  3.495    0.244      3.018      3.973              |o|
Tip         -2.113    2.632     -7.272      3.046  |-----o-----|

N: 24    Estimate of Sigma: 0.485
```

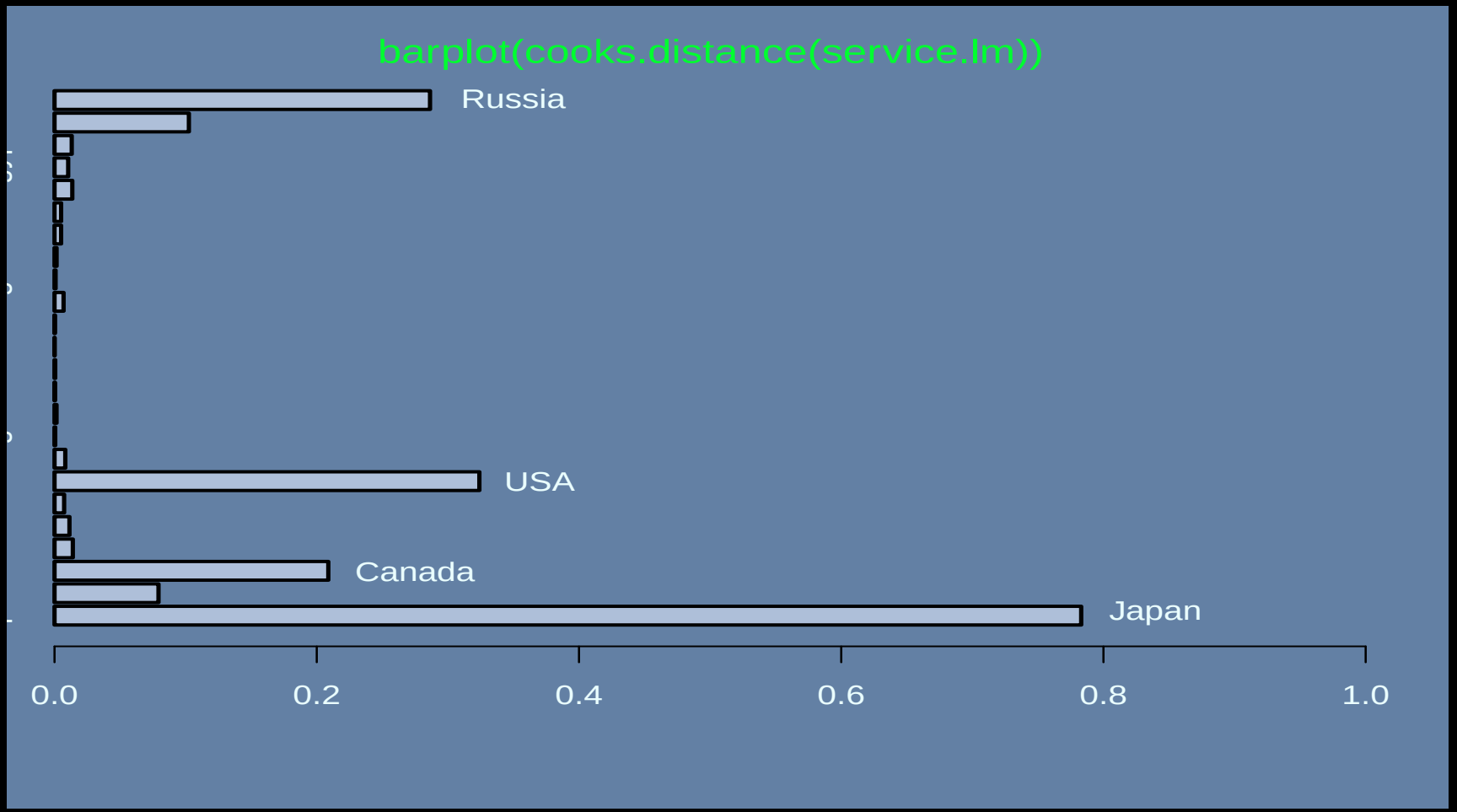
What Happens When it Doesn't Work?



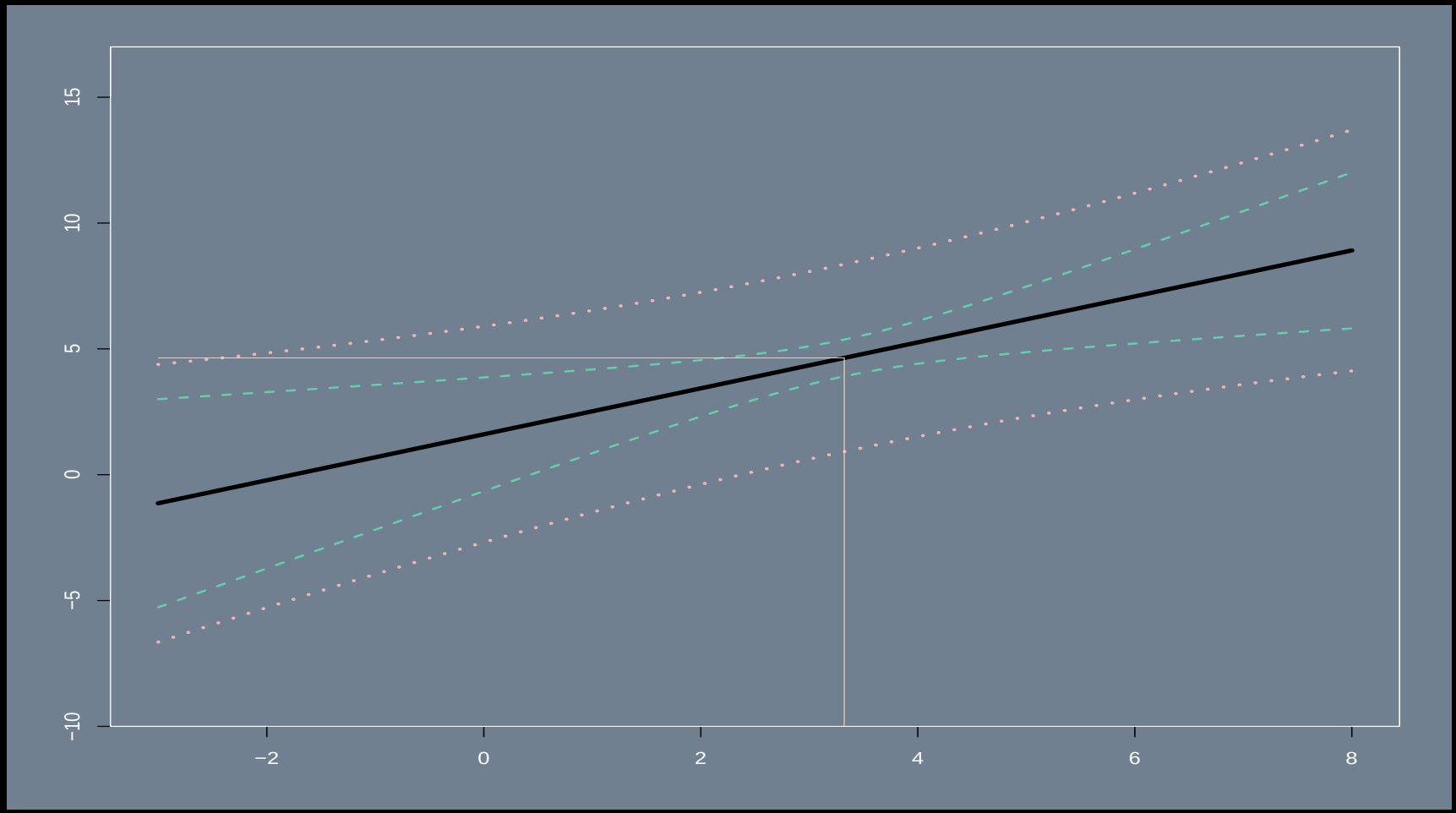
What Happens When it Doesn't Work?

```
postscript("Class.Multilevel/Images/tipping.ps",height=5,width=7)
par(mfrow=c(1,1),mar=c(5,5,2,2),lwd=2,col.axis="white",col.lab="white",
    col.sub="white",col="white",bg="slategray", cex.lab=1.3)
# PLOT POINTS AND REGRESSION LINES
plot(service$Tip,service$Quality,pch="+",xlab="Tip",ylab="Quality")
abline(service.lm,col="gold3",lwd=5)
abline(h=service.lm$coef[1],col="gold3",lty=3,lwd=2)
# ADD CONFIDENCE BOUNDS AT THREE LEVELS
ruler.df <- data.frame(Tip = seq(-0.1, 2,length=200))
for (k in c(0.99,0.95,0.90))  {
    confidence.interval <- predict(service.lm, ruler.df, interval="confidence",
        level=k)
    lines(ruler.df[,1],confidence.interval[,2],col="peachpuff",lwd=0.75)
    lines(ruler.df[,1],confidence.interval[,3],col="peachpuff",lwd=0.75)
}
# IDENTIFY POTENTIAL OUTLIERS
text(0.113,1.7,"Russia")
text(0.011,4.38,"Japan")
dev.off()
```

How Influential is Japan?



Linear Model Predictions/Forecasts



Linear Model Predictions/Forecasts

- The R code for these intervals can be produced by:

```
postscript("Class.Multilevel/linear.prediction.ps")
X <- rnorm(25,3,1); Y <- X + rnorm(25,2,2)
ruler <- data.frame(X = seq(-3, 8,length=200))
predict.interval <- predict(lm(Y ~ X), ruler, interval="prediction")
confidence.interval <- predict(lm(Y ~ X), ruler, interval="confidence")
par(mar=c(1,1,1,1),oma=c(3,3,1,1),mfrow=c(1,1),col.axis="white",col.lab="white",
   col.sub="white",col="white",bg="slategray")
plot(ruler[,1], confidence.interval[,1], type="l",lwd=4,ylim=c(-9,16),col="black")
lines(ruler[,1],confidence.interval[,2], lwd=2, lty=2, col="aquamarine3")
lines(ruler[,1],confidence.interval[,3], lwd=2, lty=2, col="aquamarine3")
lines(ruler[,1],predict.interval[,2], lwd=3, lty=3, col="rosybrown2")
lines(ruler[,1],predict.interval[,3], lwd=3, lty=3, col="rosybrown2")
segments(mean(X),-10,mean(X),mean(Y), lwd=0.5, col="peachpuff")
segments(-3,mean(Y),mean(X),mean(Y),  lwd=0.5, col="peachpuff")
dev.off()
```

Contrasts

```
options()$contrasts
        unordered           ordered
"contr.treatment"      "contr.poly"
options(contrasts=c("contr.treatment","contr.treatment"))

# POLYNOMIAL: linear, quadratic, cubic,... terms in hypothetical underlying numeric
# variable that takes on equally spaced values for lelves of levels of the factor.
# HELMERT: the difference between negative higher levels where abs(row)=#cats

N <- factor(Nlevs <- c("men","women"))
contr.sum(N)
      [,1]
men      1
women   -1

contr.treatment(N)
women
men       0
women     1
```

Contrasts

```
contr.helmert(N)
      [,1]
men     -1
women    1

contr.poly(N)
           .L
[1,] -0.70711
[2,]  0.70711

N <- factor(Nlevs <- c(1,4,8))
contr.sum(N)
   [,1] [,2]
1     1    0
4     0    1
8    -1   -1
```

Contrasts

```
contr.treatment(N)
  4 8
1 0 0
4 1 0
8 0 1

contr.poly(N)
               .L        .Q
[1,] -7.0711e-01  0.40825
[2,] -7.8505e-17 -0.81650
[3,]  7.0711e-01  0.40825

contr.helmert(N)
  [,1] [,2]
1   -1   -1
4    1   -1
8    0    2
```

Contrasts

```
contr.helmert(4)
  [,1] [,2] [,3]
1   -1   -1   -1
2    1   -1   -1
3    0    2   -1
4    0    0    3

contr.helmert(5)
  [,1] [,2] [,3] [,4]
1   -1   -1   -1   -1
2    1   -1   -1   -1
3    0    2   -1   -1
4    0    0    3   -1
5    0    0    0    4
```

Consequences For a Linear Model

```
Y <- rnorm(100); X1 <- rgamma(100,3,2); X2 <- factor(rbinom(100,2,.6))
contrasts(X2) <- contr.sum(3)
summary(lm(Y~X1+X2))
Call:
lm(formula = Y ~ X1 + X2)
Residuals:
    Min      1Q  Median      3Q     Max
-3.0125 -0.5853 -0.0534  0.8055  2.3665

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.00900    0.21297   -0.04     0.97
X1          -0.00147    0.12394   -0.01     0.99
X21          0.21212    0.18800    1.13     0.26
X22         -0.11254    0.14622   -0.77     0.44

Residual standard error: 1.05 on 96 degrees of freedom
Multiple R-squared: 0.0135,        Adjusted R-squared: -0.0173
F-statistic: 0.439 on 3 and 96 DF,  p-value: 0.726
```

Consequences For a Linear Model

```
contrasts(X2) <- contr.treatment(3)
summary(lm(Y~X1+X2))

Call:
lm(formula = Y ~ X1 + X2)
Residuals:
    Min      1Q  Median      3Q     Max
-3.0125 -0.5853 -0.0534  0.8055  2.3665

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.20311    0.29927    0.68     0.50
X1          -0.00147    0.12394   -0.01     0.99
X22         -0.32465    0.29829   -1.09     0.28
X23         -0.31170    0.31345   -0.99     0.32

Residual standard error: 1.05 on 96 degrees of freedom
Multiple R-squared: 0.0135,        Adjusted R-squared: -0.0173
F-statistic: 0.439 on 3 and 96 DF,  p-value: 0.726
```

Consequences For a Linear Model

```
contrasts(X2) <- contr.poly(3)
summary(lm(Y~X1+X2))

Call:
lm(formula = Y ~ X1 + X2)
Residuals:
    Min      1Q  Median      3Q     Max
-3.0125 -0.5853 -0.0534  0.8055  2.3665

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.00900    0.21297   -0.04     0.97
X1          -0.00147    0.12394   -0.01     0.99
X2.L        -0.22040    0.22164   -0.99     0.32
X2.Q         0.13783    0.17908    0.77     0.44

Residual standard error: 1.05 on 96 degrees of freedom
Multiple R-squared: 0.0135,        Adjusted R-squared: -0.0173
F-statistic: 0.439 on 3 and 96 DF,  p-value: 0.726
```

Consequences For a Linear Model

```
contrasts(X2) <- contr.helmert(3)
summary(lm(Y~X1+X2))

Call:
lm(formula = Y ~ X1 + X2)
Residuals:
    Min      1Q  Median      3Q     Max
-3.0125 -0.5853 -0.0534  0.8055  2.3665

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.00900    0.21297   -0.04     0.97
X1          -0.00147    0.12394   -0.01     0.99
X21         -0.16233    0.14915   -1.09     0.28
X22         -0.04979    0.07822   -0.64     0.53

Residual standard error: 1.05 on 96 degrees of freedom
Multiple R-squared: 0.0135,        Adjusted R-squared: -0.0173
F-statistic: 0.439 on 3 and 96 DF,  p-value: 0.726
```

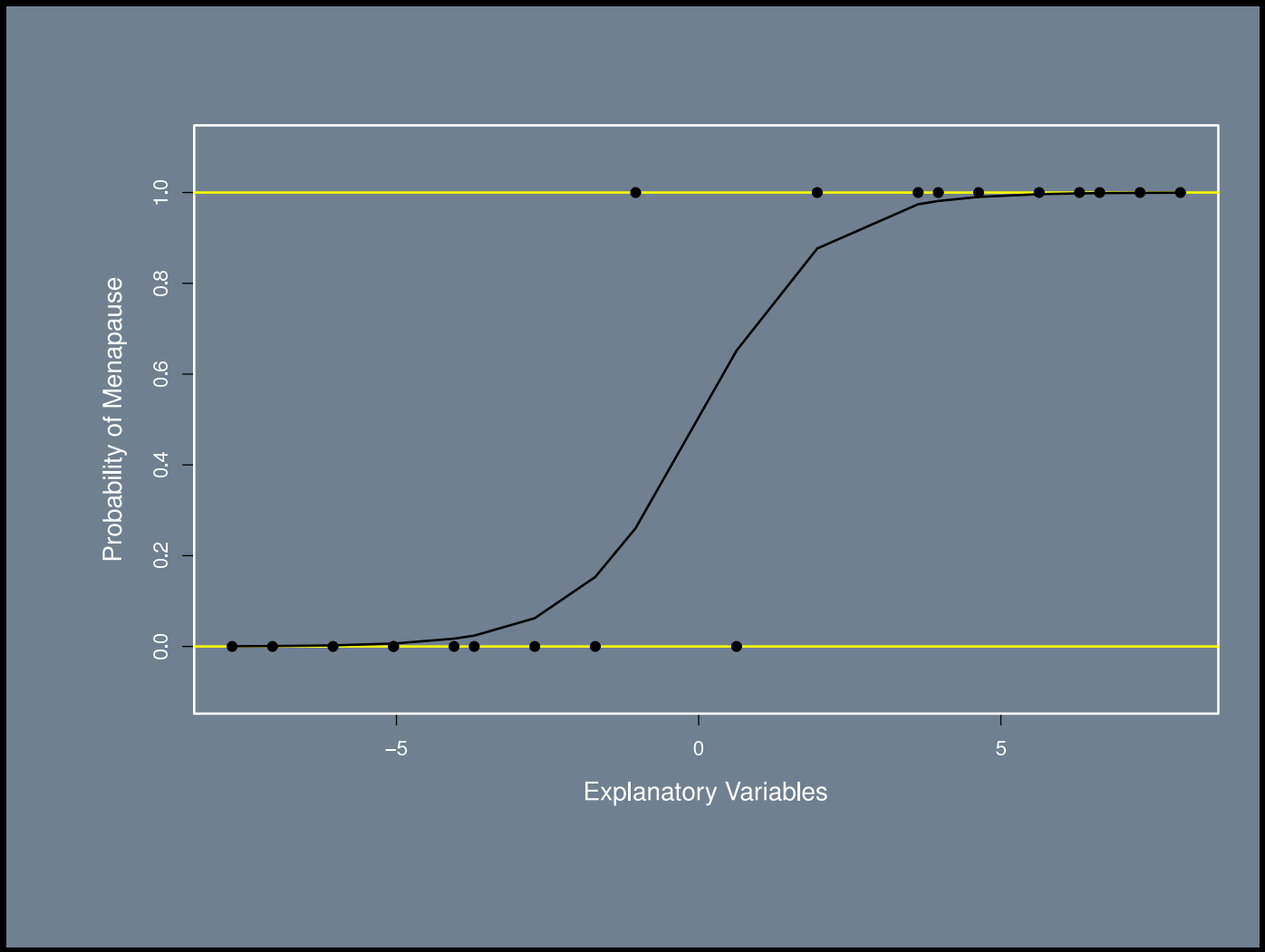
Logistic Regression: Anaemia Example

```
summary( glm(Menapause~Age, data=anaemia, family=binomial(link=logit)) )
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.45227  -0.13139  -0.00176   0.09818   1.63990

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -14.395      7.462   -1.93    0.054
Age            0.334      0.174    1.92    0.055

    Null deviance: 27.7259  on 19  degrees of freedom
Residual deviance:  5.7632  on 18  degrees of freedom
```

Logistic Illustration



Logistic Illustration

```
inv.logit <- function(mu)  log(mu/(1-mu))
logit <- function(Xb)  1/(1+exp(-Xb))
ana.logit <- glm(Menapause ~ Age, data=anaemia, family=binomial(link=logit))
postscript("Class.PreMed.Stats/Images/logit.anaemia1.fig.ps")
par(mfrow=c(1,1),mar=c(5,5,2,2),lwd=2,col.axis="white",col.lab="white",
        col.sub="white",col="white",bg="slategray",
        cex.lab=1.3,oma=c(4,2,2,2))
xbeta <- as.matrix(cbind(rep(1,length=nrow(anaemia)),anaemia$Age))
        %*% coef(ana.logit)
plot(range(xbeta),c(-0.1,1.1),type="n",xlab="Explanatory Variables",
        ylab="Probability of Menapause")
abline(h=c(0,1),col="yellow")
x <- seq(from=min(xbeta),to=max(xbeta),length=100)
points(xbeta,anaemia$Menapause,col="black",pch=19)
lines(xbeta,logit(xbeta),col="black")
dev.off()
```

Mouse Data

- This research project looks at the impact of high-fat diet of the mother on prostate cancer outcomes in male pups.
- Treatment: the "dam" is fed a high fat diet between 54 and 209 weeks.
- Control: regular mouse chow.
- The pups are sacrificed between 116 and 446 weeks after birth (includes 15 weeks of weaning).

```
M1 <- glm(Hyper.proliferation ~ Age.Group.Weeks + Diet.Treatment + Days.Old.When.Used
          + Body.Weight + Male.Pups.In.Cage + log(Days.Parents.On.Diet.Before.Birth),
          family="poisson", data=current.mouse.dat)
```

Poisson GLM

```
summary(M1)
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                              -0.4388     1.2789 -0.3431   0.7315
Age.Group.Weeks1                          1.3191     0.3370  3.9146   0.0001
Diet.Treatment                            0.8696     0.2311  3.7624   0.0002
Days.Old.When.Used                       -0.0038     0.0015 -2.5532   0.0111
Body.Weight                              -0.0248     0.0333 -0.7454   0.4583
Male.Pups.In.Cage                         0.2919     0.1393  2.0962   0.0361
log(Days.Parents.On.Diet.Before.Birth)    0.1262     0.2647  0.4767   0.6336

# DEVIANCE COMPARISON
   Null deviance: 106.662  on 52  degrees of freedom
   Residual deviance:  76.369  on 46  degrees of freedom
   AIC: 200.4
pchisq(106.662-76.369,df=6,lower.tail=FALSE)
[1] 3.4574e-05
# CHECK FOR OVERDISPERSION
sum(residuals(M1,type="pearson")^2)
[1] 64.559
```

Gamma GLM of Electoral Politics in Scotland

- On September 11, 1997 Scottish voters overwhelming (74.3%) approved the establishment of the first Scottish national parliament in nearly three hundred years.
- On the same ballot, the voters gave strong support (63.5%) to granting this parliament taxation powers.
- Data: 32 *Unitary Authorities* (also called council districts), U.K. government sources, includes 40 potential explanatory variables

The model for these data using the gamma link function is produced by:

$$\begin{aligned}\underbrace{g^{-1}(\boldsymbol{\theta})}_{32\times 1} &= g^{-1}(\boldsymbol{X\beta}) \\ &= -\frac{1}{\boldsymbol{X\beta}} \\ &= -\left[\mathbf{1}\beta_0 + \mathbf{COU}\beta_1 + \mathbf{UNM}\beta_2 + \mathbf{MOR}\beta_3 + \mathbf{ACT}\beta_4 + \mathbf{AGE}\beta_5\right]^{-1} \\ &= E[\mathbf{Y}] = E[\mathbf{YES}].\end{aligned}$$

The systematic component here is $\boldsymbol{X\beta}$, the stochastic component is $\mathbf{Y} = \mathbf{YES}$, and the link function is $\boldsymbol{\theta} = -\frac{1}{\boldsymbol{\mu}}$.

Gamma GLM

```
scotland.df <- read.table("http://jgill.wustl.edu/data/scotvote.dat",header=TRUE)
scottish.vote.glm <- glm((PerYesTax/100) ~ CouncilTax * PerClaimantFemale
                         + StdMortalityRatio + Active + GDP + Percentage5to15,
                         family=Gamma, data=scotland.df)
graph.summary(scottish.vote.glm)
Family: Gamma      Link function: inverse
                              Coef Std.Err. 0.95 Lower 0.95 Upper CIs:ZE+RO
(Intercept)                 -1.777    1.148     -4.026      0.473  |--o--|
CouncilTax                   0.005    0.002      0.002      0.008      |o|
PerClaimantFemale            0.203    0.053      0.099      0.308      |o|
StdMortalityRatio           -0.007    0.003     -0.012     -0.002      |o|
Active                       0.011    0.004      0.003      0.019      |o|
GDP                          0.000    0.000      0.000      0.000      |o|
Percentage5to15             -0.052    0.024     -0.099     -0.005      |o|
CouncilTax:PerClaimantFemale 0.000    0.000      0.000      0.000      |o|
N: 32    log-likelihood: 59.892    AIC: -111.784    Dispersion Parameter: 0.0035842
    Null deviance: 0.536 on 31 degrees of freedom
Residual deviance:  0.087 on 24 degrees of freedom
```

New Prostate Dataset

- Byar DP, Green SB (1980): Bulletin Cancer, Paris, 67:477-488
- `bm`, Bone Metastases: no=0 (420), yes=1 (82), the outcome variable.
- `stage`, M0: The cancer has not spread past nearby lymph nodes (289), M1: The cancer has spread beyond the nearby lymph nodes (213).
- `pf`, normal activity 0 (450), some required bed-rest 1 (52).
- `sz`, Size of Primary Tumor (cm^2), median=11.
- `ap`, Serum Prostatic Acid Phosphatase, median=0.7.
- `hg`, Serum Hemoglobin (g/100ml), median=13.7.

Generalized Additive Models

```
library(mgcv)
prostate.df <- read.table("http://jgill.wustl.edu/data/prostate.full.dat",header=TRUE)
prostate.gam1 <- gam(bm ~ stage + pf + log1p(sz) + s(ap) + s(hg),
    family=binomial(link=logit), data=prostate.df)
summary(prostate.gam1)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -20.017      4.087   -4.90  9.7e-07
log1p(sz)      0.375      0.197    1.91    0.057
stage          4.495      1.024    4.39  1.1e-05
pf             1.036      0.464    2.23    0.026

Approximate significance of smooth terms:
       edf Ref.df Chi.sq p-value
s(ap) 3.32   4.03   11.0  0.0270
s(hg) 1.72   2.20   12.2  0.0029
```