Introduction to Multiple Regression

Keegan Korthauer

Department of Statistics

UW Madison

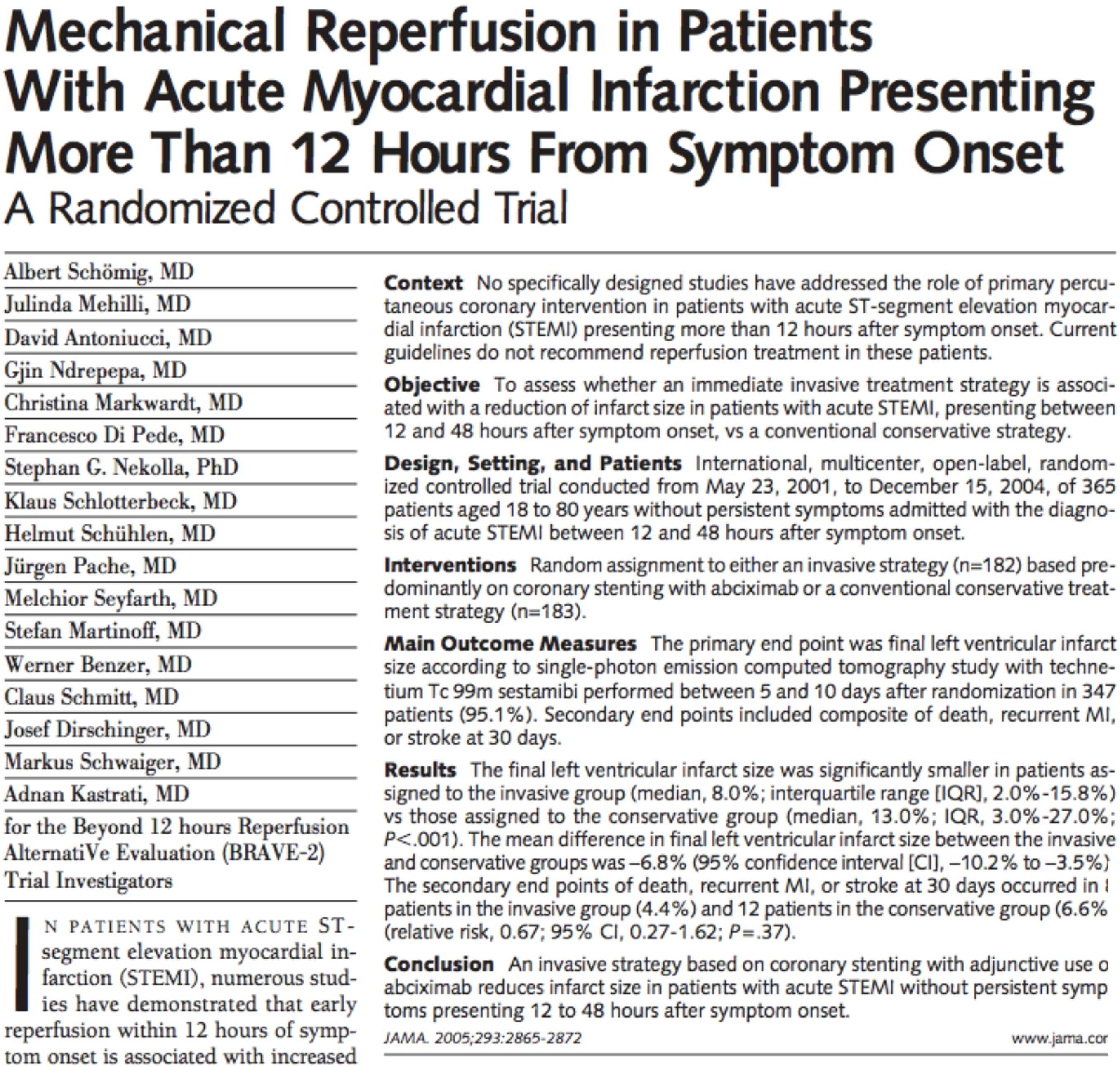
Basic MLR Model

- Dependent continuous variable **y**
- *p* independent continuous variables $\mathbf{x_1, x_2, \ldots, x_p}$
- *n* observations: ordered pairs $\mathbf{(y_i, x_{1i}, x_{2i}, \ldots, x_{pi})}$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \varepsilon_i$$

- Predicted y_i for a set of $x_{1i}, x_{2i}, \ldots, x_{pi}$:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_p x_{pi}$$



Mechanical Reperfusion in Patients With Acute Myocardial Infarction Presenting More Than 12 Hours From Symptom Onset

A Randomized Controlled Trial

Albert Schömig, MD
Julinda Mehilli, MD
David Antoniucci, MD
Gjin Ndrepepa, MD
Christina Markwardt, MD
Francesco Di Pede, MD
Stephan G. Nekolla, PhD
Klaus Schlotterbeck, MD
Helmut Schühlen, MD
Jürgen Pache, MD
Melchior Seyfarth, MD
Stefan Martinoff, MD
Werner Benzer, MD
Claus Schmitt, MD
Josef Dirschinger, MD
Markus Schwaiger, MD
Adnan Kastrati, MD
for the Beyond 12 hours Reperfusion AlternatiVe Evaluation (BRAVE-2) Trial Investigators

Context No specifically designed studies have addressed the role of primary percutaneous coronary intervention in patients with acute ST-segment elevation myocardial infarction (STEMI) presenting more than 12 hours after symptom onset. Current guidelines do not recommend reperfusion treatment in these patients.

Objective To assess whether an immediate invasive treatment strategy is associated with a reduction of infarct size in patients with acute STEMI, presenting between 12 and 48 hours after symptom onset, vs a conventional conservative strategy.

Design, Setting, and Patients International, multicenter, open-label, randomized controlled trial conducted from May 23, 2001, to December 15, 2004, of 365 patients aged 18 to 80 years without persistent symptoms admitted with the diagnosis of acute STEMI between 12 and 48 hours after symptom onset.

Interventions Random assignment to either an invasive strategy (n=182) based predominantly on coronary stenting with abciximab or a conventional conservative treatment strategy (n=183).

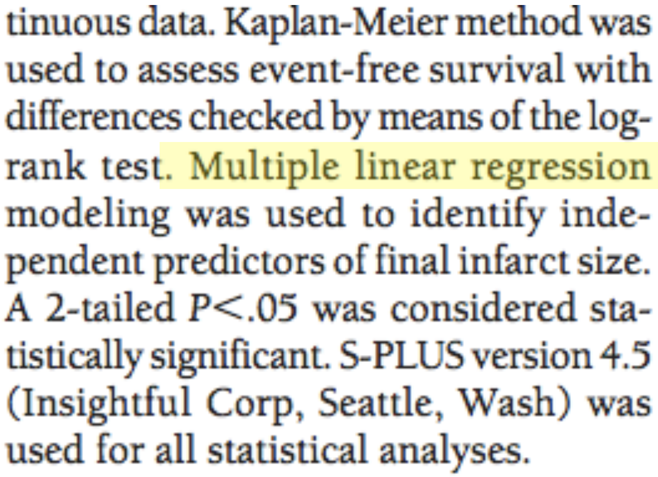
Main Outcome Measures The primary end point was final left ventricular infarct size according to single-photon emission computed tomography study with technetium Tc 99m sestamibi performed between 5 and 10 days after randomization in 347 patients (95.1%). Secondary end points included composite of death, recurrent MI, or stroke at 30 days.

Results The final left ventricular infarct size was significantly smaller in patients assigned to the invasive group (median, 8.0%; interquartile range [IQR], 2.0%-15.8%) vs those assigned to the conservative group (median, 13.0%; IQR, 3.0%-27.0%; $P<.001$). The mean difference in final left ventricular infarct size between the invasive and conservative groups was −6.8% (95% confidence interval [CI], −10.2% to −3.5%) The secondary end points of death, recurrent MI, or stroke at 30 days occurred in 8 patients in the invasive group (4.4%) and 12 patients in the conservative group (6.6% (relative risk, 0.67; 95% CI, 0.27-1.62; $P=.37$).

Conclusion An invasive strategy based on coronary stenting with adjunctive use o abciximab reduces infarct size in patients with acute STEMI without persistent symptoms presenting 12 to 48 hours after symptom onset.

JAMA. 2005;293:2865-2872 www.jama.cor

IN PATIENTS WITH ACUTE ST-segment elevation myocardial infarction (STEMI), numerous studies have demonstrated that early reperfusion within 12 hours of symptom onset is associated with increased



tinuous data. Kaplan-Meier method was used to assess event-free survival with differences checked by means of the log-rank test. Multiple linear regression modeling was used to identify independent predictors of final infarct size. A 2-tailed $P<.05$ was considered statistically significant. S-PLUS version 4.5 (Insightful Corp, Seattle, Wash) was used for all statistical analyses.

Least Squares Coefficients

- Minimize the **sum of squared residuals (SSE)** to obtain coefficient point estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$
  - Analogous to SLR, involves taking p+1 partial derivatives, setting them equal to zero, and solving a system of p+1 equations…
  - OR a much more elegant expression using linear algebra…
  - But we'll rely on R to calculate the values for us

- SSE is still the sum of squared differences between observed y and predicted ŷ – no longer can visualize in 2D

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \ldots - \hat{\beta}_p x_{pi}$$

$$SSE = \sum_{i=1}^{n} e_i^2$$

Interpreting Coefficients

- $\hat{\beta}_0$ is the predicted value of y when all x_i are equal to zero
  - Often this is nonsensical or involves extrapolation

- $\hat{\beta}_i$ is the predicted change in y for every one unit increase in x_i *while holding all other predictors constant*
  - Or *after adjusting for all other predictors in the model*
  - Often of the most interest to us

- We can use these estimates to predict the value of y for a new observation with $x_1,\ldots,x_n$ (*as long as each* $x_1,\ldots,x_n$ value *is within the range of the observed values*)

Interpreting Interactions

We can rearrange the following model:

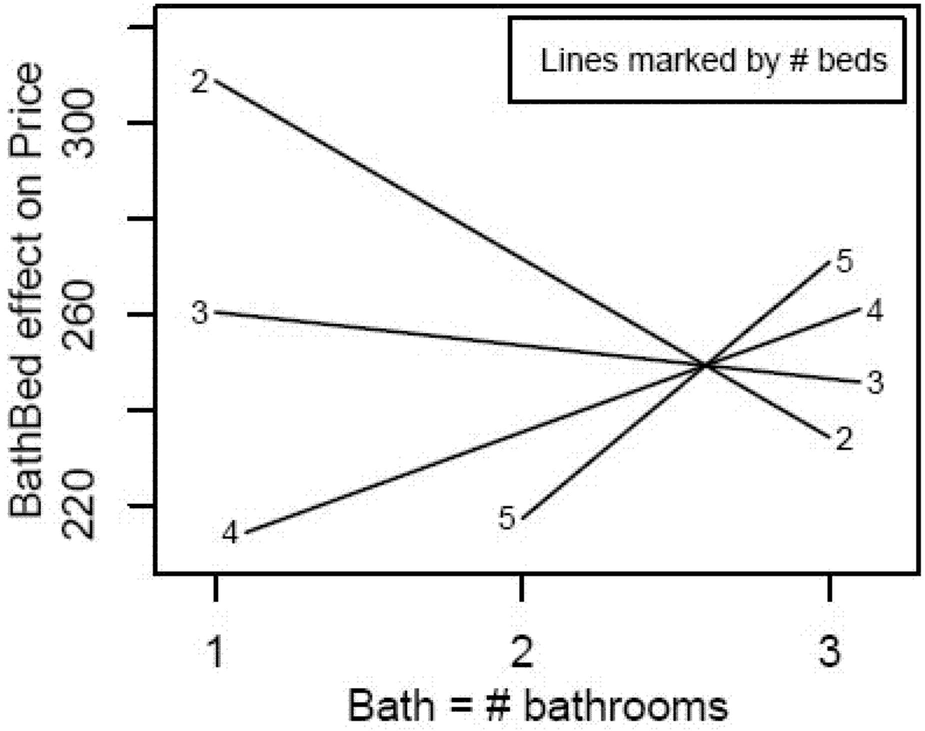
$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \varepsilon_i$$

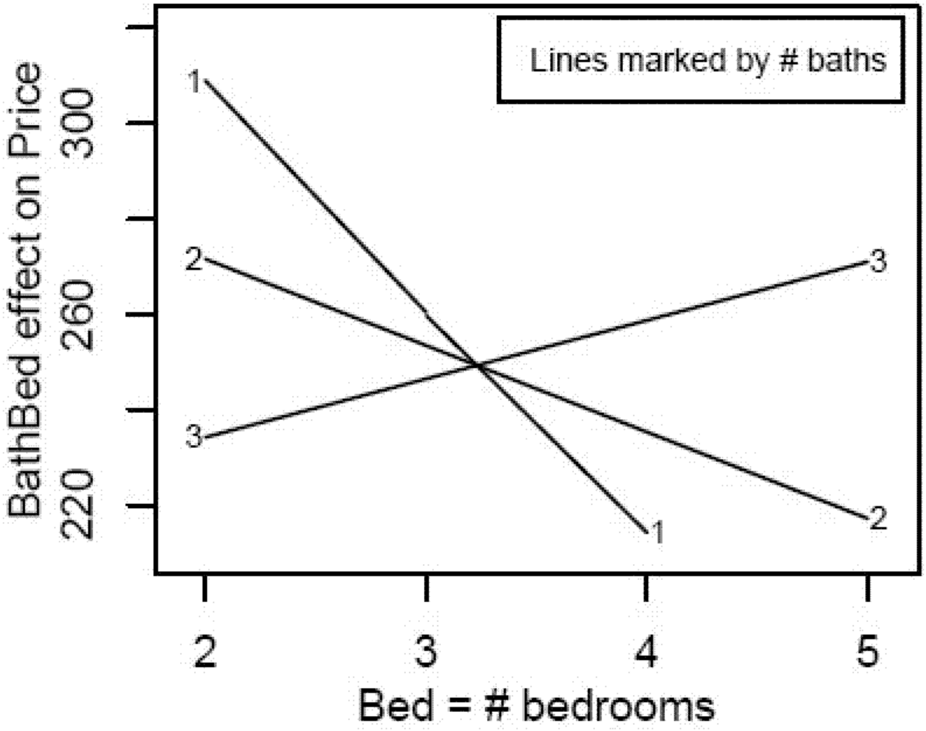
$$= \beta_0 + (\beta_1 + \beta_3 x_{2i}) x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

$$= \beta_0 + \beta_1 x_{1i} + (\beta_2 + \beta_3 x_{1i}) x_{2i} + \varepsilon_i$$

- The presence of β_3 here indicates that the effect of x_1 depends on the value of x_2
- In other words, the model predicts that y will change by $\hat{\beta}_1 + \hat{\beta}_3 x_2$ units for every one unit increase in x_1

What Does Interaction Look Like?





Price = 504.2 – 98.16*Bath – 78.91*Bed + 30.39*Bath*Bed

Example – Patient Satisfaction

A hospital administrator wished to study the relation between

- Y: patient satisfaction (percent)
- X_1: patient's age
- X_2: severity of illness (an index)
- X_3: patient's anxiety level (an index)

They randomly selected 46 patients and collected each measurement above – the first 5 observations are:

```
Satis Age Sev Anx
48    50   51 2.3
57    36   46 2.3
66    40   48 2.2
70    41   44 1.8
89    28   43 1.8
          ...
```

Estimating Coefficients in R

```
> ptsat <- read.table("patient_satisfaction.txt", header=T)
> attach(ptsat)
> fit1 <- lm(Satis ~ Age + Sev + Anx)
> summary(fit1)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 158.4913    18.1259   8.744 5.26e-11 ***
Age          -1.1416     0.2148  -5.315 3.81e-06 ***
Sev          -0.4420     0.4920  -0.898   0.3741    
Anx         -13.4702     7.0997  -1.897   0.0647 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.06 on 42 degrees of freedom
Multiple R-squared: 0.6822,Adjusted R-squared: 0.6595 
F-statistic: 30.05 on 3 and 42 DF,  p-value: 1.542e-10
```

$\hat{\beta}_i$

What about those other Sums of Squares?

We’re in luck!

They’re exactly the same as in SLR

Sums of Squares

- Error Sum of Squares $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} \hat{y}_i^2$
- Total Sum of Squares $SST = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2$
- Regression Sum of Squares $SSR = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$

Analysis of Variance property: $SST = SSR + SSE$

Coefficient of Determination (Goodness-of-fit measure):

$$R^2 = \frac{\text{Regression sum of squares}}{\text{Total sum of squares}} = \frac{SSR}{SST}$$

Assumptions for Inference

- Again we have to make some assumptions in order to perform any inference (CIs/PIs/HTs)

- Simplest case: same four assumptions on the errors:

1. Errors $\varepsilon_1,\ldots, \varepsilon_n$ are **random** and **independent**. In particular, the magnitude of any error ε_i does not influence the value of the next error ε_{i+1}
2. Errors $\varepsilon_1,\ldots, \varepsilon_n$ all have **mean 0**
3. Errors $\varepsilon_1,\ldots, \varepsilon_n$ all have the **same variance** denoted by σ^2
4. Errors $\varepsilon_1,\ldots, \varepsilon_n$ are **normally distributed**

Consequences of the Assumptions

- The errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent normal random variables with mean zero and variance σ^2:

$$e_i \sim N(0,\ \sigma^2)$$

- Since $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \varepsilon_i$ the y_i are a linear combination of ε_i so they are also normally distributed:

$$y_i \sim N(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi},\ \sigma^2)$$

Estimating the Error Variance σ²

- In SLR: $\hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{SSE}{n-2}$

- In MLR: $\hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p-1} = \frac{SSE}{n-p-1}$

- Estimates of $s_{\hat{\beta}_0}$ and $s_{\hat{\beta}_1}$ are the same as before, but using the appropriate value of s
- We'll obtain them from R

Other MLR Quantities in R

```
> ptsat <- read.table("patient_satisfaction.txt", header=T)
> attach(ptsat)
> fit1 <- lm(Satis ~ Age + Sev + Anx)
> summary(fit1)
```

```
Coefficients:
            Estimate Std. Error     t value Pr(>|t|)
(Intercept) 158.4913    18.1259      8.744 5.26e-11 ***
Age          -1.1416     0.2148     -5.315 3.81e-06 ***
Sev          -0.4420     0.4920     -0.898   0.3741
Anx         -13.4702     7.0997     -1.897   0.0647 .
```

$\hat{\beta}_i$ $s_{\hat{\beta}_i}$

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

s

```
Residual standard error: 10.06 on 42 degrees of freedom
```

R^2

```
Multiple R-squared: 0.6822,Adjusted R-squared: 0.6595
F-statistic: 30.05 on 3 and 42 DF,  p-value: 1.542e-10
```

HTs for Coefficients (One at a Time)

- Under assumptions 1-4,

$$\frac{(\hat{\beta}_i - \beta_i)}{s_{\hat{\beta}_i}} \sim t_{n-p-1}$$

- We can test a hypothesis for any of the β_i (one at a time) using a t-test where the quantity above is the test statistic

Other MLR Quantities in R

```
> ptsat <- read.table("patient_satisfaction.txt", header=T)
> attach(ptsat)
> fit1 <- lm(Satis ~ Age + Sev + Anx)
> summary(fit1)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 158.4913    18.1259   8.744 5.26e-11 ***
Age          -1.1416     0.2148  -5.315 3.81e-06 ***
Sev          -0.4420     0.4920  -0.898   0.3741    
Anx         -13.4702     7.0997  -1.897   0.0647 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.06 on 42 degrees of freedom
Multiple R-squared: 0.6822,Adjusted R-squared: 0.6595
F-statistic: 30.05 on 3 and 42 DF,  p-value: 1.542e-10
```

$\hat{\beta}_i$

$s_{\hat{\beta}_i}$

Test statistics and p-values for the null that the coefficient=0

s

R^2

HTs for Coefficients (Globally)

- Say we want to test whether all the predictor coefficients are equal to zero, i.e.

$$H_0 : \beta_1 = \ldots = \beta_p = 0, \text{ versus}$$

$$H_1 : \text{at least one of the } \beta_i \text{ is not zero}$$

Looks like a ratio of 'variances'!

- The test statistic is: $$F = \frac{SSR / p}{SSE / (n - p - 1)} \sim F_{p,\, n-p-1}$$

- This uses the F distribution, that we used previously to test whether the ratio of two (normal) variances was different than 1 (section 6.11)

If this test is not rejected, the model may not be useful

Other MLR Quantities in R

```
> ptsat <- read.table("patient_satisfaction.txt", header=T)
> attach(ptsat)
> fit1 <- lm(Satis ~ Age + Sev + Anx)
> summary(fit1)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 158.4913    18.1259   8.744 5.26e-11 ***
Age          -1.1416     0.2148  -5.315 3.81e-06 ***
Sev          -0.4420     0.4920  -0.898   0.3741
Anx         -13.4702     7.0997  -1.897   0.0647 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$\hat{\beta}_i$

$s_{\hat{\beta}_i}$

Test statistics and p-values for the null that the coefficient=0

s

```
Residual standard error: 10.06 on 42 degrees of freedom
Multiple R-squared: 0.6822,Adjusted R-squared: 0.6595
F-statistic: 30.05 on 3 and 42 DF,  p-value: 1.542e-10
```

R^2

F

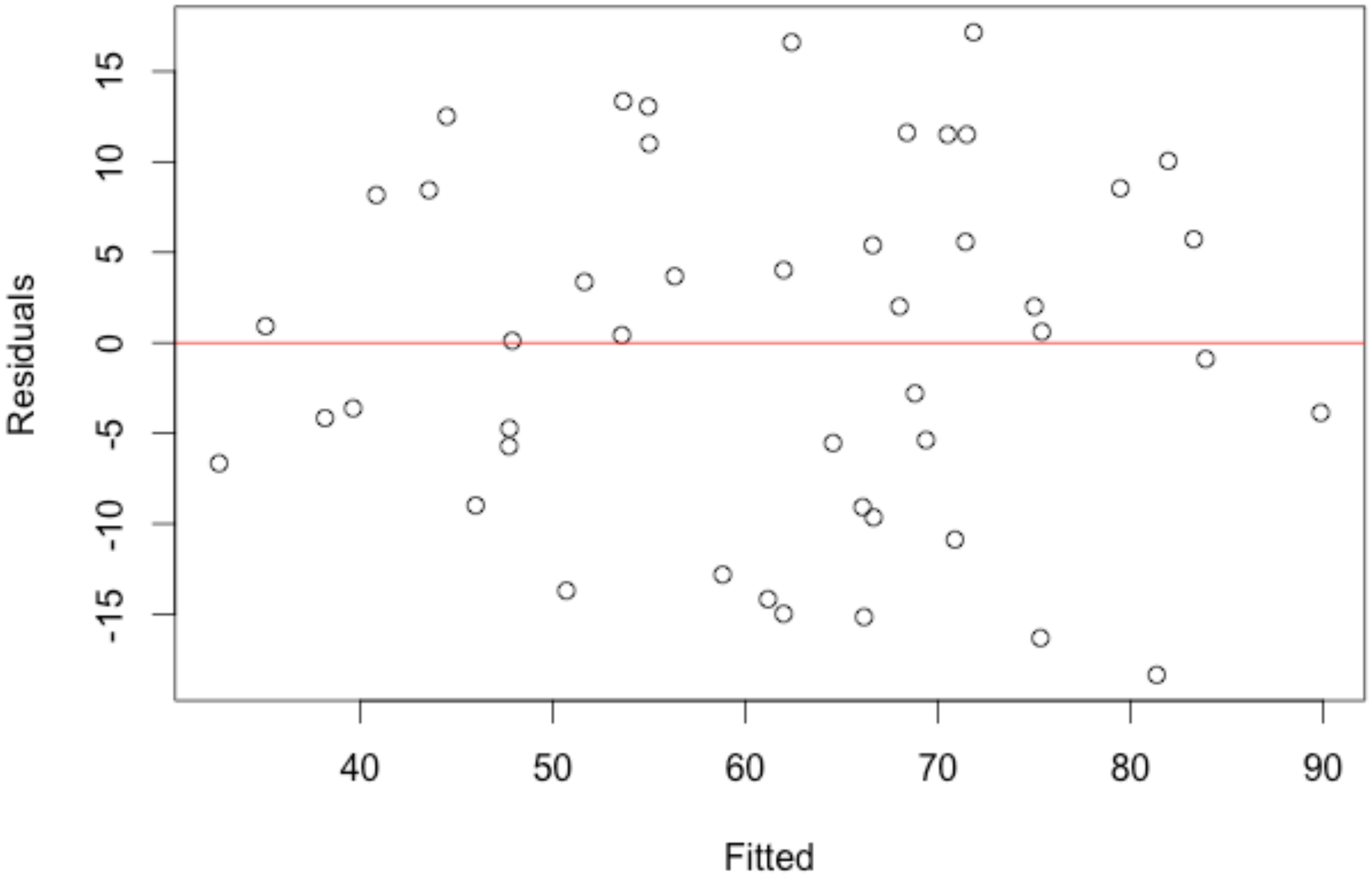
Checking Assumptions

- All of what we discussed for SLR diagnostics applies

- In addition, look at a plot of each of the independent variables against the residuals
  - look for trends or heteroscedasticity

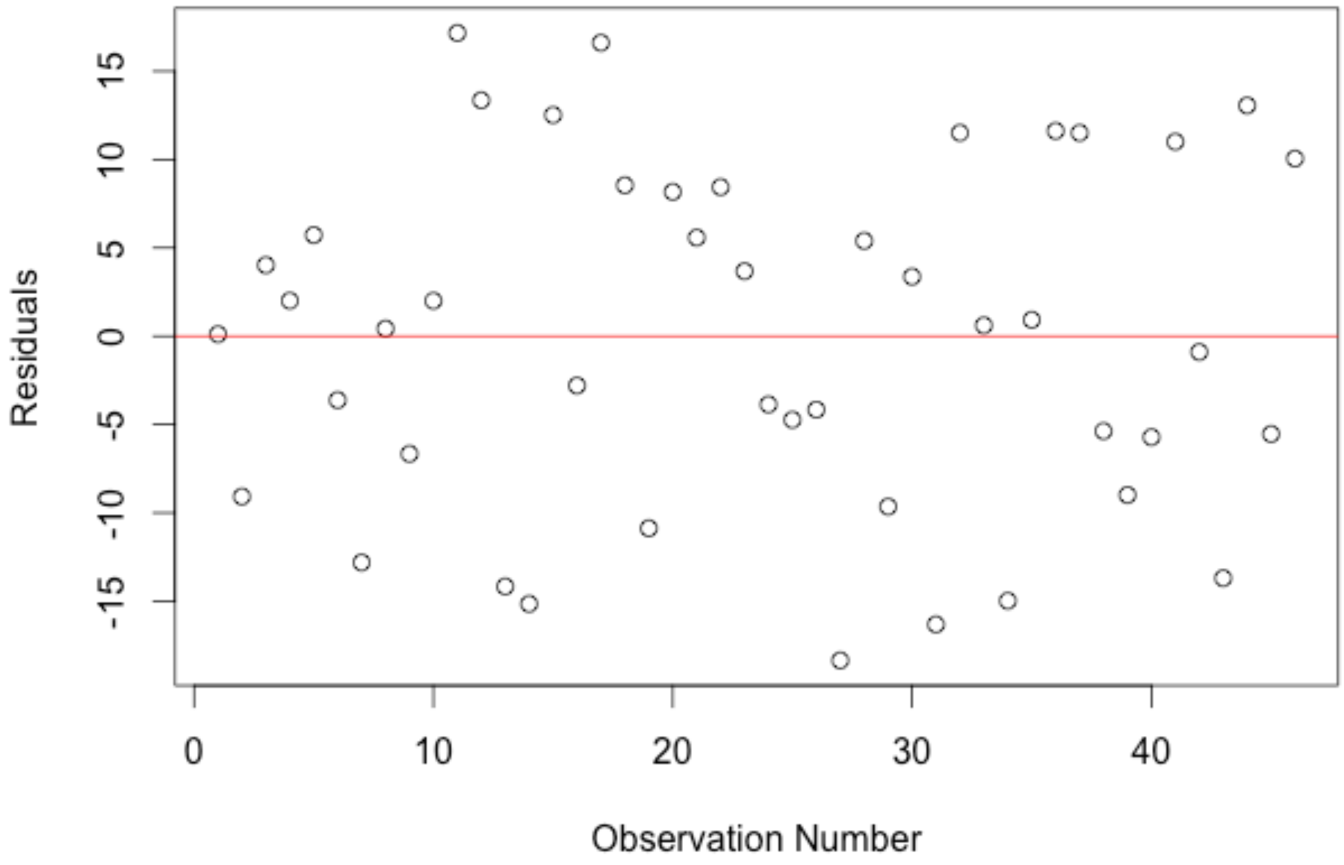
Residual Plot

```
# residual plot
plot(fit1$fitted, fit1$residuals, xlab="Fitted", ylab="Residuals")
abline(h=0, col="red")
```



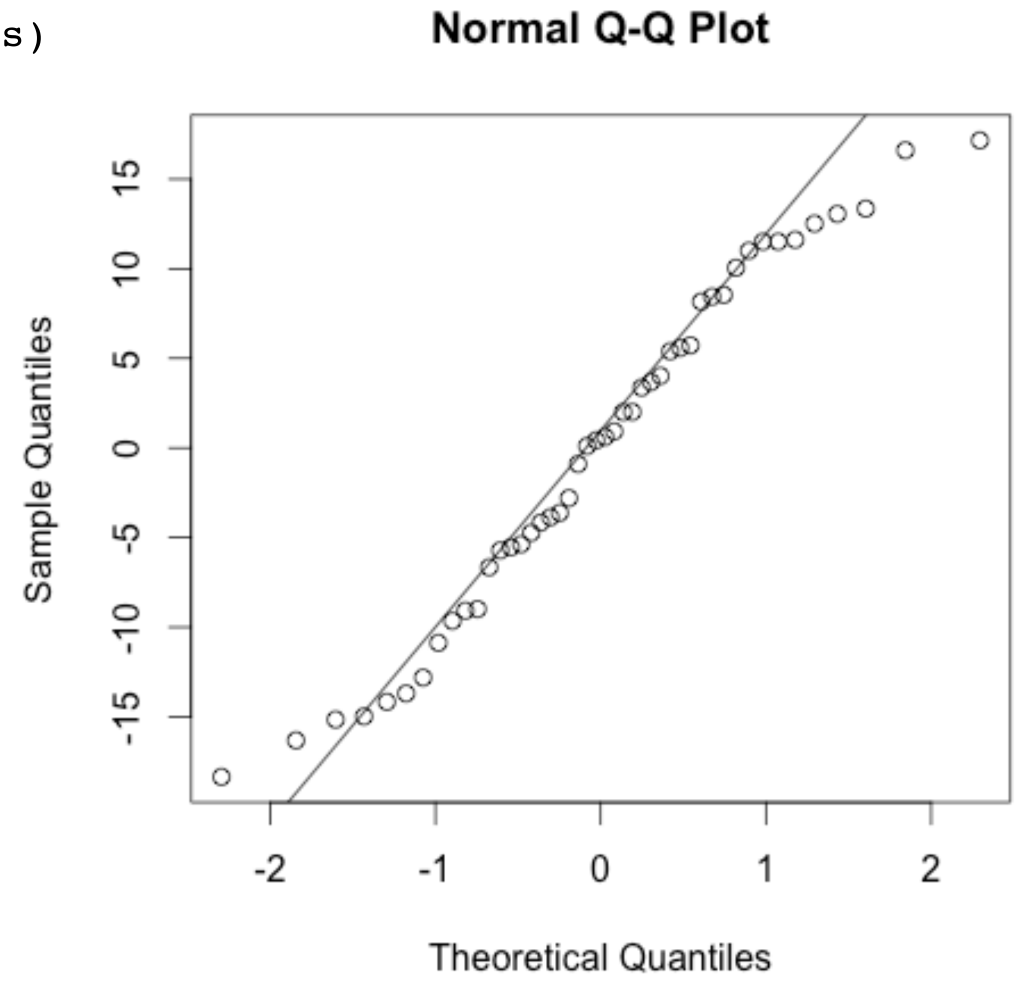
Time Order (Independence) Plot

```
# time order plot
plot(1:46, fit1$residuals, xlab=”Observation Number", ylab="Residuals")
abline(h=0, col="red")
```



QQ (Normality) Plot

```
# QQ plot
qqnorm(fit1$residuals)
qqline(fit1$residuals)
```

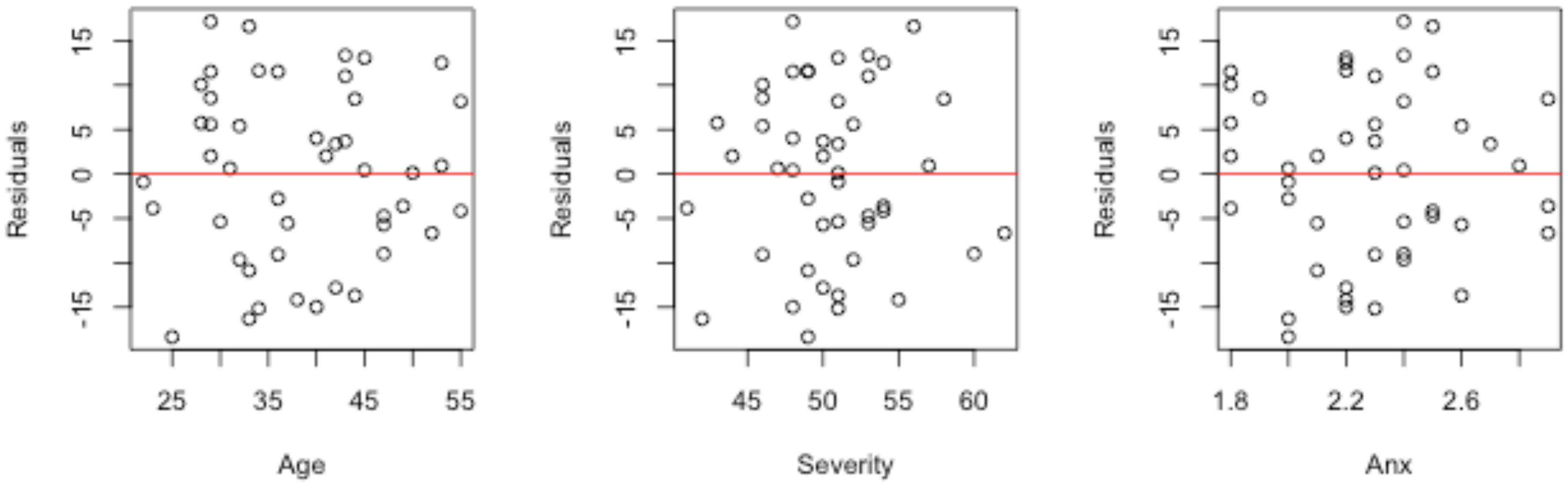


Residuals Against All Predictors

```
#residuals against all predictors
plot(Age, fit1$residuals, xlab="Age", ylab="Residuals")
abline(h=0, col="red")

plot(Sev, fit1$residuals, xlab="Severity", ylab="Residuals")
abline(h=0, col="red")

plot(Anx, fit1$residuals, xlab="Anx", ylab="Residuals")
abline(h=0, col="red")
```



Next

- More on Multiple Linear Regression
  - Collinearity and Confounding
  - Model Selection