STAT 215
Multiple Regression I

Colin Reimer Dawson

Oberlin College

October 26, 2017

Outline

Quantitative Vs. Categorical Predictor and Response

		Response	
		Quantitative	**Categorical**
Predictor(s)	**Quantitative**	SLR	Logistic Reg.
	Categorical	ANOVA	
	Multiple	Multiple Reg.	

The Multiple Regression Model

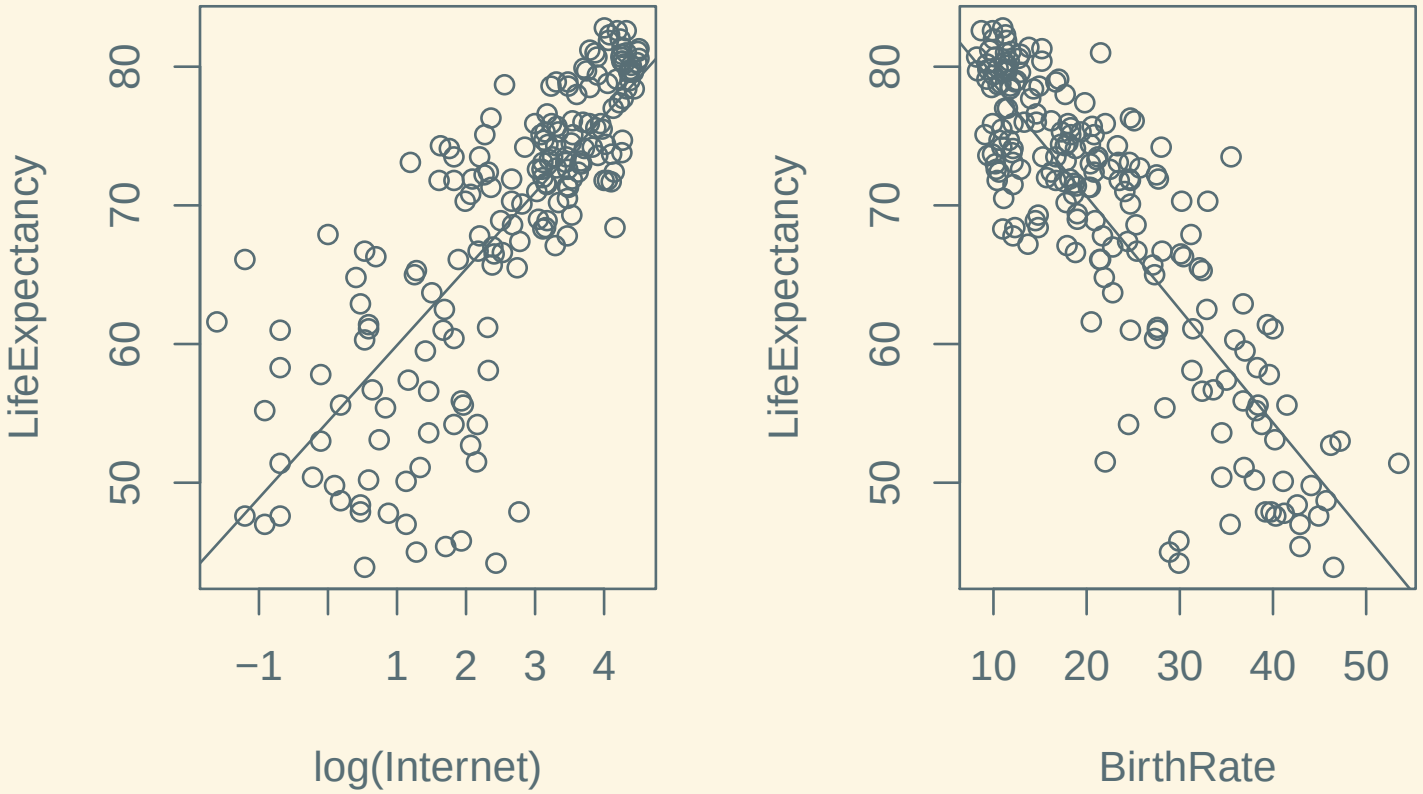
DATA $=$ PATTERN $+$ IDIOSYNCRACIES

The Multiple Regression Population Model

$$Y = f(X_1, \ldots, X_K) + \varepsilon$$
$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$$

One β_j for each predictor X_j

Example: Life Expectancy



We could fit separate regression models for each predictor...

SLR Model Using log(Internet)

```
model1 <- lm(LifeExpectancy ~ log(Internet), data = AllCountries)
summary(model1)
```

```
Call:
lm(formula = LifeExpectancy ~ log(Internet), data = AllCountries)

Residuals:
     Min       1Q   Median       3Q      Max
-23.6029  -2.6718   0.6961   3.7591  18.3274

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    54.4022     0.9474   57.42   <2e-16 ***
log(Internet)   5.5065     0.3152   17.47   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.395 on 187 degrees of freedom
  (24 observations deleted due to missingness)
Multiple R-squared:   0.62,Adjusted R-squared:  0.618
F-statistic: 305.1 on 1 and 187 DF,  p-value: < 2.2e-16
```

SLR Model Using BirthRate

```
model2 <- lm(LifeExpectancy ~ BirthRate, data = AllCountries)
summary(model2)
```

```
Call:
lm(formula = LifeExpectancy ~ BirthRate, data = AllCountries)

Residuals:
     Min       1Q   Median       3Q      Max
-18.3822  -3.3734   0.5593   3.6471  15.5005

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 86.95147    0.88338   98.43   <2e-16 ***
BirthRate   -0.81555    0.03602  -22.64   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.381 on 194 degrees of freedom
  (17 observations deleted due to missingness)
Multiple R-squared:  0.7255,Adjusted R-squared:  0.724
F-statistic: 512.6 on 1 and 194 DF,  p-value: < 2.2e-16
```

Or Fit Both at Once

```
model3 <- lm(LifeExpectancy ~ log(Internet) + BirthRate, data = AllCountries)
summary(model3)
```

```
Call:
lm(formula = LifeExpectancy ~ log(Internet) + BirthRate, data = AllCountri

Residuals:
     Min       1Q   Median       3Q      Max
-19.5702  -2.8414   0.6454   3.1477  13.6028

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   77.23444    2.33017  33.145  < 2e-16 ***
log(Internet)  1.95223    0.42551   4.588 8.21e-06 ***
BirthRate     -0.60920    0.05881 -10.360  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.106 on 186 degrees of freedom
  (24 observations deleted due to missingness)
Multiple R-squared:  0.7591,Adjusted R-squared:  0.7565
F-statistic:   293 on 2 and 186 DF,  p-value: < 2.2e-16
```

Single Multiple Regression vs. Multiple Single Regressions

- What is the added value in using both predictors together, vs. fitting each separately?
- There are disadvantages:
  - Harder to interpret
  - Can't easily plot
- Advantages...
  - If we actually know both predictors, we get a single prediction, instead of two conflicting ones
  - Can "control for" one predictor and test the other

What is a Case?

- Q: What does a single case consist of for a multiple regression model?
- A: A complete case has a value for Y, and for *each* X.

The Four-Step Process: Multiple Regression

1. CHOOSE a form of the model
   - Select predictors
   - Choose any transformations of predictors
2. FIT: Estimate
   - coefficients: $\hat{\beta}_1, \hat{\beta}_1, \ldots, \hat{\beta}_k$
   - residual variance $\hat{\sigma}^2_\varepsilon$
3. ASSESS the fit
   - Examine residuals
   - Test individual predictors (t-tests)
   - Test overall fit (ANOVA, R^2)
4. USE the model
   - Make predictions
   - Construct CIs and PIs

CHOOSE: Active Pulse Rate

```
library(Stat2Data); data("Pulse")
head(Pulse, n = 3)
```

```
  Active Rest Smoke Gender Exercise Hgt Wgt
1     97   78     0      1        1  63 119
2     82   68     1      0        3  70 225
3     88   62     0      0        3  72 175
```

$$\texttt{Active} = \beta_0 + \beta_1 \cdot \texttt{Rest} + \beta_2 \cdot \texttt{Hgt} + \beta_3 \cdot \texttt{Wgt} \quad (1)$$

$$\texttt{Active} = \beta_0 + \beta_1 \cdot \texttt{Rest} + \beta_2 \cdot \texttt{Hgt} + \beta_3 \cdot \texttt{BMI} \quad (2)$$

CHOOSE: Apply Transformations

```
PulseWithBMI <- mutate(Pulse, BMI = Wgt / Hgt^2 * 703)
head(PulseWithBMI, n = 3)
```

```
  Active Rest Smoke Gender Exercise Hgt Wgt      BMI
1     97   78     0      1        1  63 119 21.07760
2     82   68     1      0        3  70 225 32.28061
3     88   62     0      0        3  72 175 23.73167
```

FIT: Estimate Coefficients

The Multiple Regression Population Model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$$

The Multiple Regression Fitted Model

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k + \hat{\varepsilon}$$

How to choose $\hat{\beta}_j$s? Minimize SSE! (Requires linear algebra / vector calculus)

FIT: Estimate Coefficients

```
(my.model <- lm(Active ~ Rest + Hgt + BMI, data = PulseWithBMI))



    Call:
    lm(formula = Active ~ Rest + Hgt + BMI, data = PulseWithBMI)

    Coefficients:
    (Intercept)         Rest          Hgt          BMI
        22.7213       1.1291      -0.3634       0.6850
```

$$\widehat{\texttt{Active}} = 22.7 + 1.1 \cdot \texttt{Rest} + -0.4 \cdot \texttt{Hgt} + 0.7 \cdot \texttt{BMI}$$

FIT: Estimate Residual Variance

Recall Variance Decomposition for Regression:

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (\hat{Y}_i - \bar{Y})^2 + \sum_i (Y_i - \hat{Y}_i)^2$$

$$SS_{Total} = SS_{Model} + SS_{Error}$$

Recall ANOVA Table:

$$MS_{Model} = SS_{Model}/df_{Model}$$
$$MS_{Error} = SS_{Error}/df_{Error}$$

where MS_{Error} represents $\hat{\sigma}^2_\varepsilon$. So... what are df_{Model} and df_{Error}?

Regression Degrees of Freedom

$df_{Model} = k$ where k is the number of predictors

This is the number of things "free to vary" (constraint is that $f(\bar{X}_1, \ldots, \bar{X}_k) = \bar{Y}$)

$df_{Error} = n - k - 1$ where n is the sample size

This is the number of "pieces of information" we have about the sizes of the residuals. (Can fit any p points exactly with p coefficients.)

FIT: Estimate Residual Variance

$$\hat{\sigma}^2_\varepsilon = MS_{Error} = \frac{SS_{Error}}{df_{Error}} = \frac{\sum_i (Y_i - \hat{Y})^2}{n - k - 1}$$

FIT: Estimate Residual Variance

```
summary(my.model)


    Call:
    lm(formula = Active ~ Rest + Hgt + BMI, data = PulseWithBMI)

    Residuals:
        Min      1Q  Median      3Q     Max
    -35.308  -9.917  -2.370   6.569  64.578

    Coefficients:
                Estimate Std. Error t value Pr(>|t|)
    (Intercept)  22.7213    21.3864   1.062   0.2892
    Rest          1.1291     0.1018  11.090   <2e-16 ***
    Hgt          -0.3634     0.2840  -1.279   0.2021
    BMI           0.6850     0.3238   2.115   0.0355 *
    ---
    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Residual standard error: 14.93 on 228 degrees of freedom
    Multiple R-squared:  0.3785,Adjusted R-squared:  0.3703
    F-statistic: 46.28 on 3 and 228 DF,  p-value: < 2.2e-16
```

FIT: The Final Model

$$\texttt{Active} = 22.7 + 1.1 \cdot \texttt{Rest} + -0.4 \cdot \texttt{Hgt} + 0.7 \cdot \texttt{BMI} + \varepsilon$$

where $\varepsilon \sim \mathcal{N}(0, 14.9)$

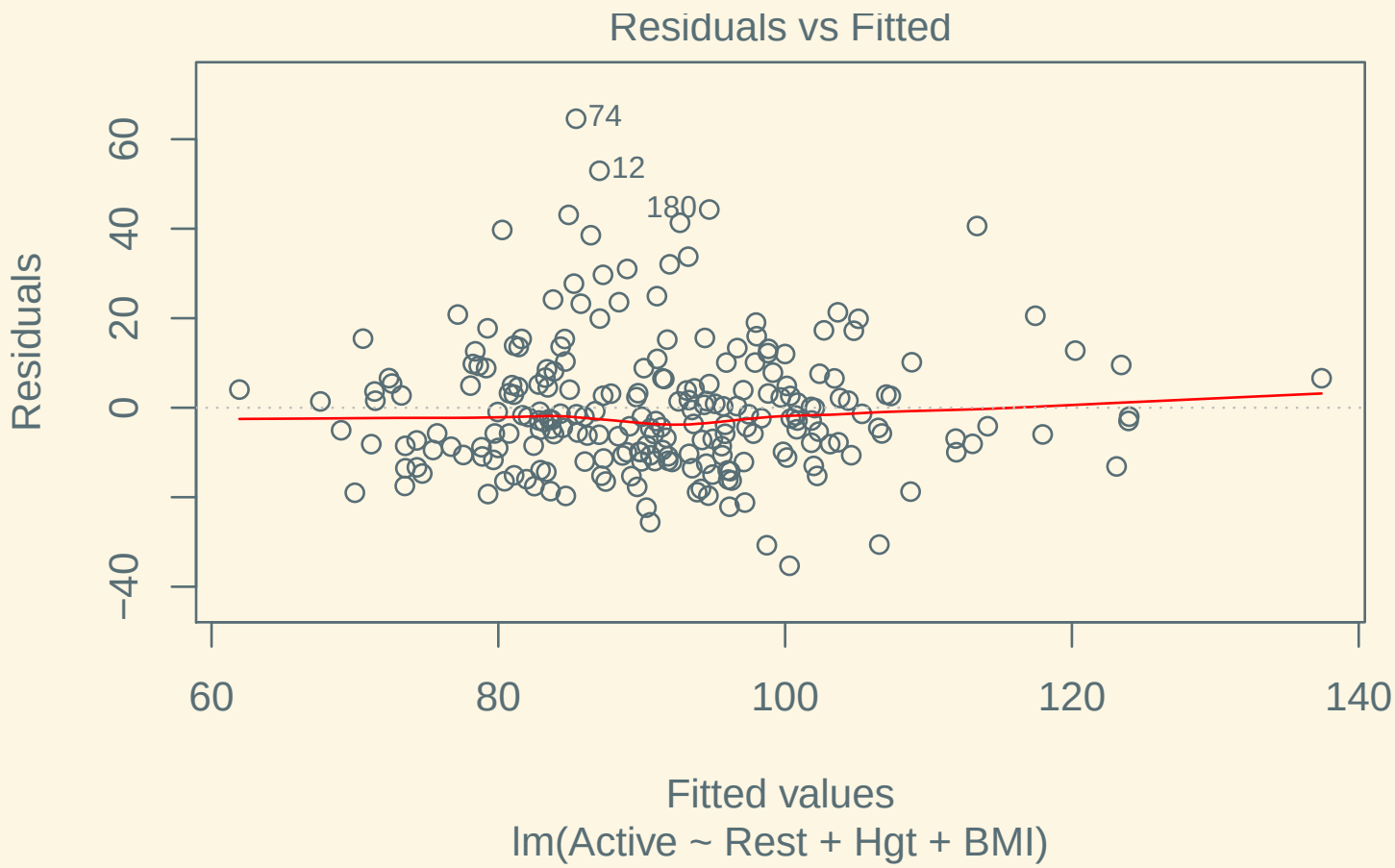
ASSESS: Check Conditions

Same conditions as always apply:

1. Linearity (mean of Y is given by some linear model)
2. Independence (residuals are not correlated)
3. Homoskedasticity (same variance at all combinations of X)
4. Normality (residuals normally distributed)

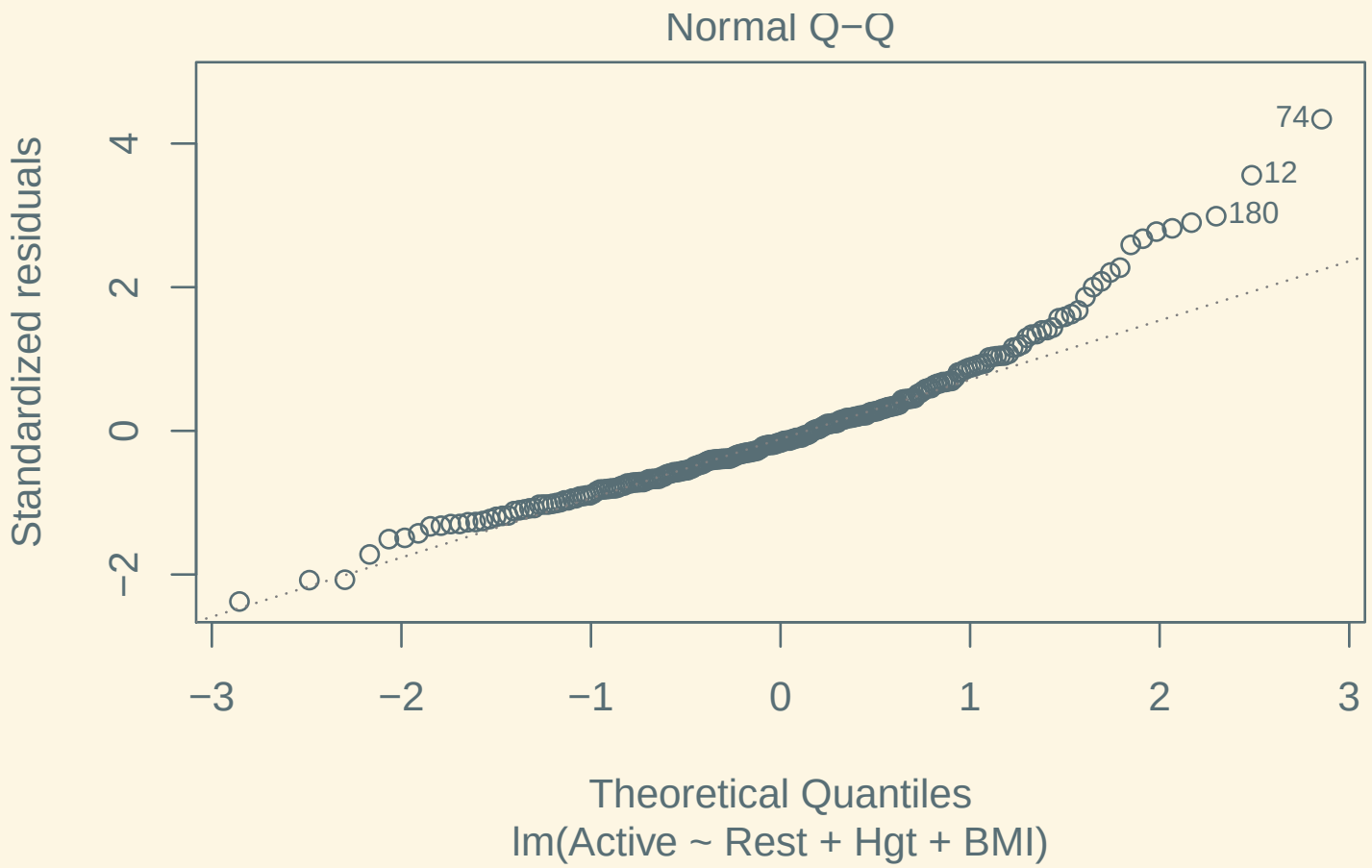
ASSESS: Check Conditions

```
plot(my.model, which = 1)
```



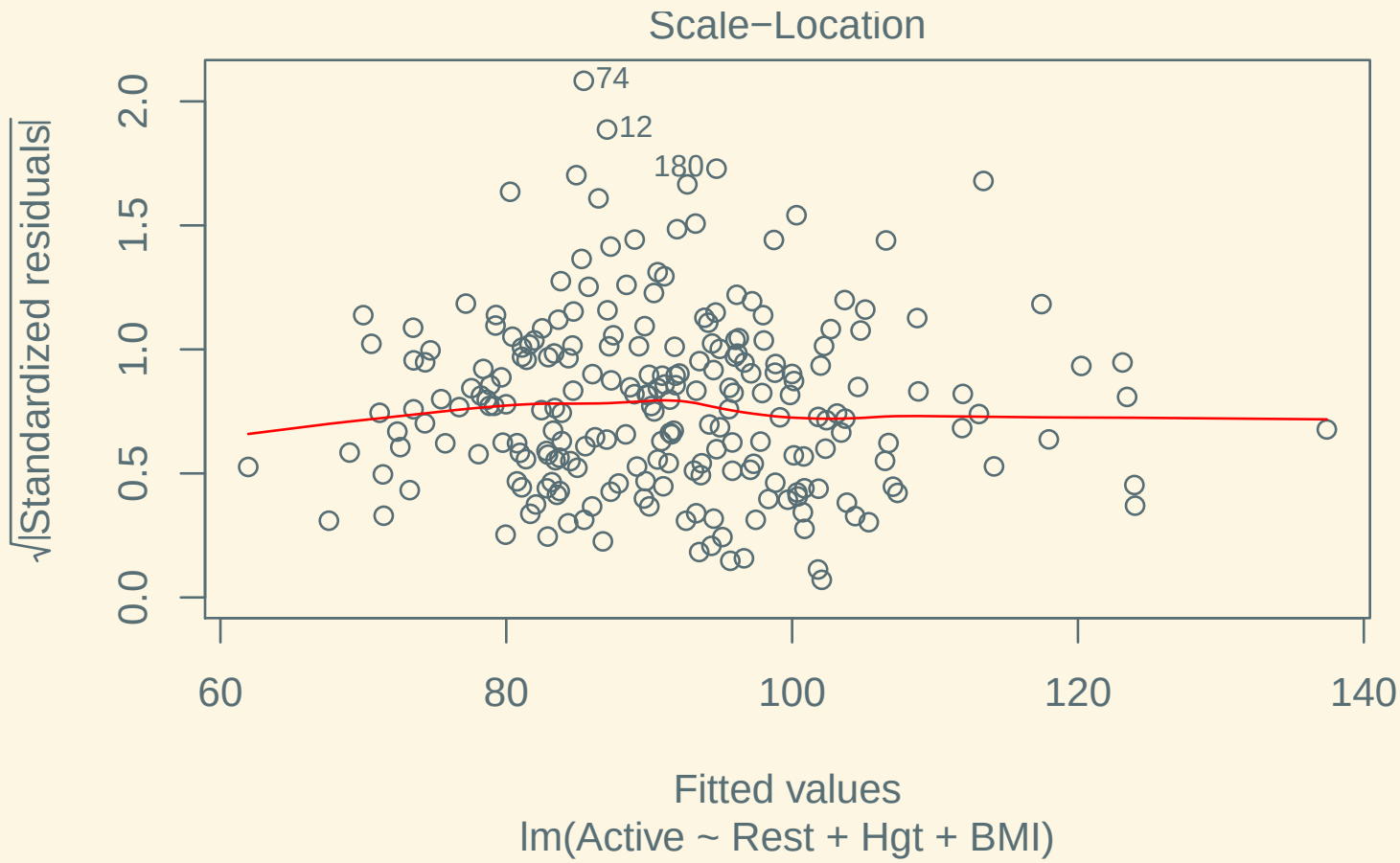
ASSESS: Check Conditions

```
plot(my.model, which = 2)
```



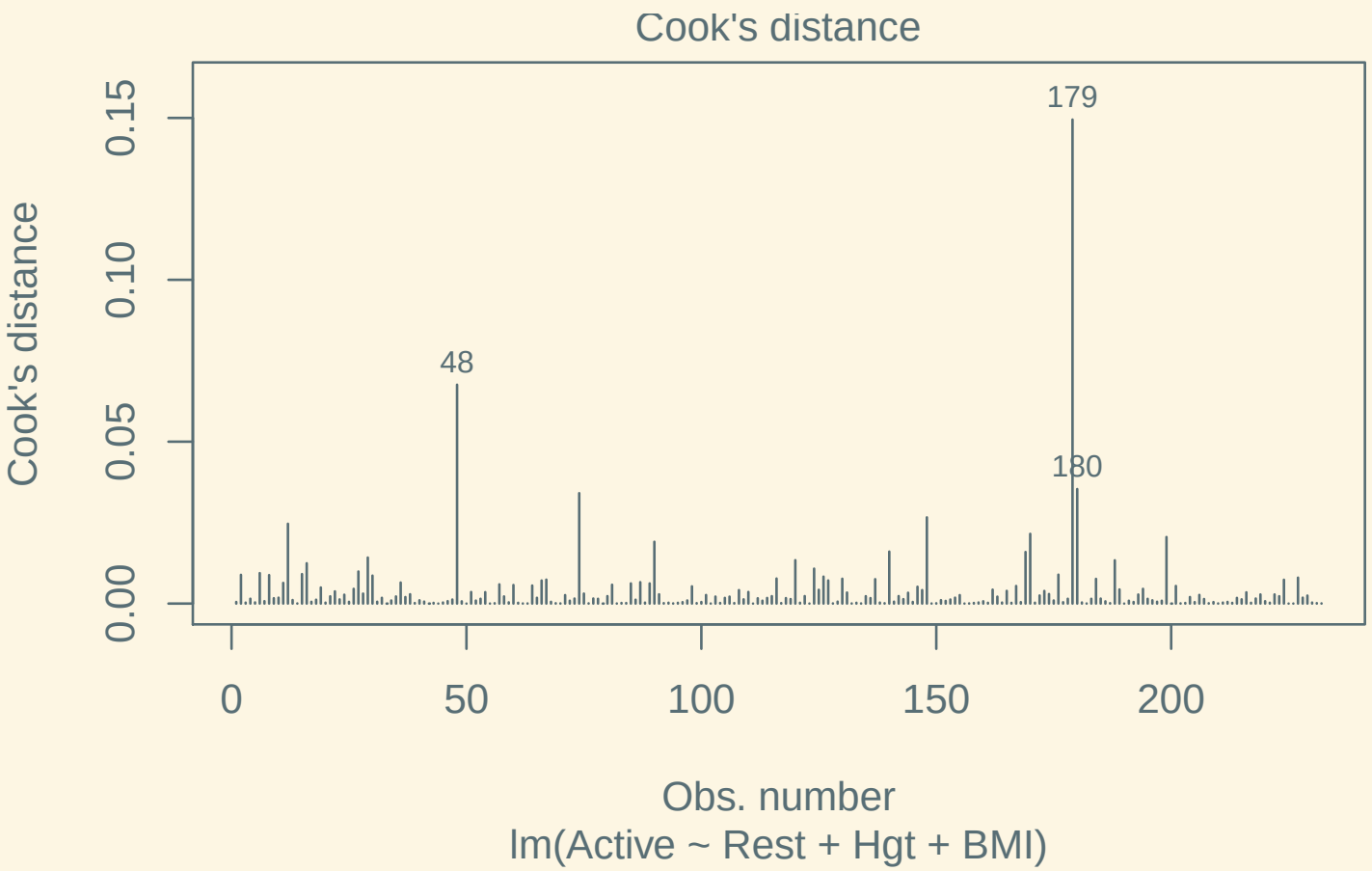
ASSESS: Check Conditions

```
plot(my.model, which = 3)
```



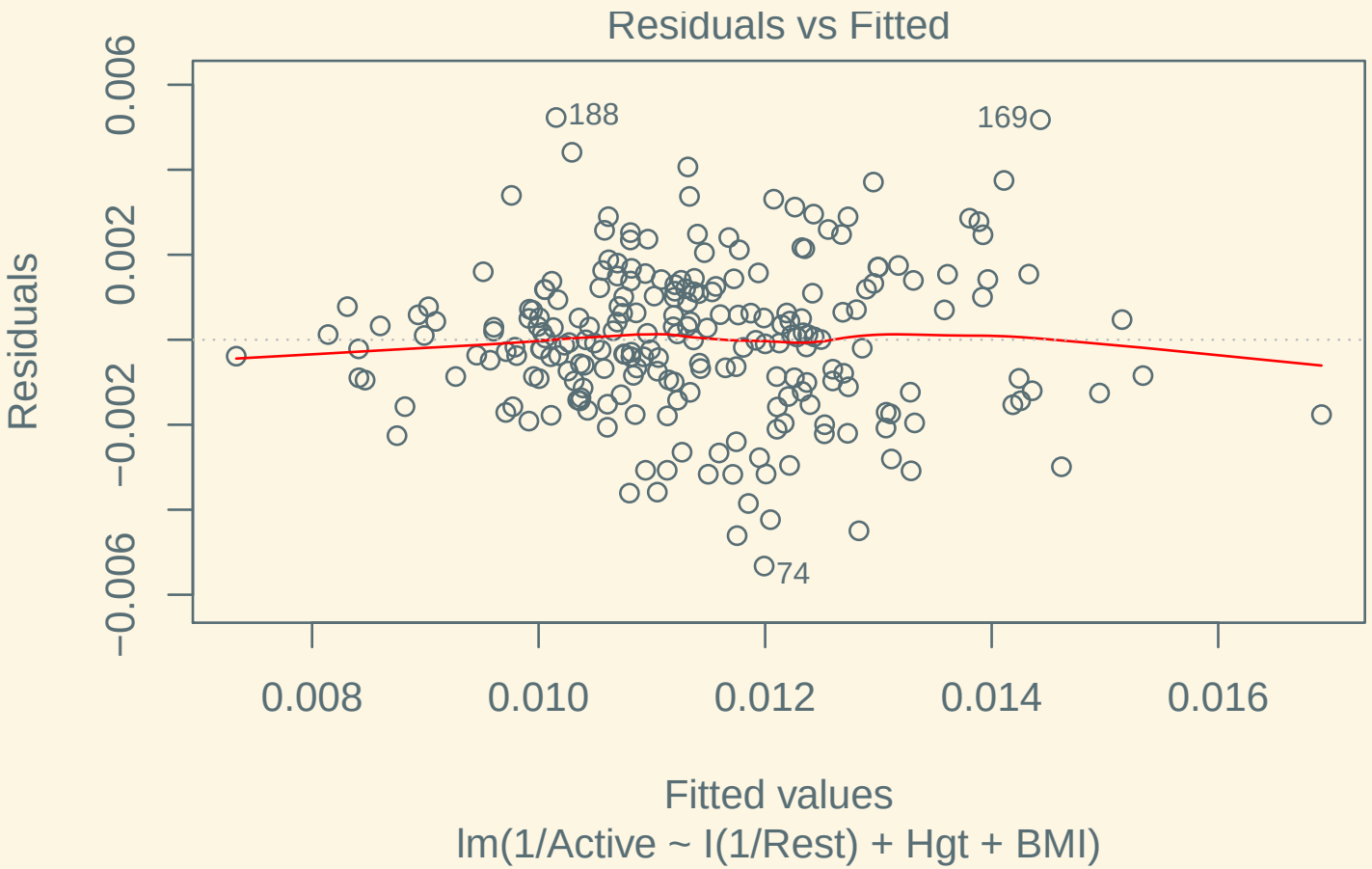
ASSESS: Check Conditions

```
plot(my.model, which = 4)
```



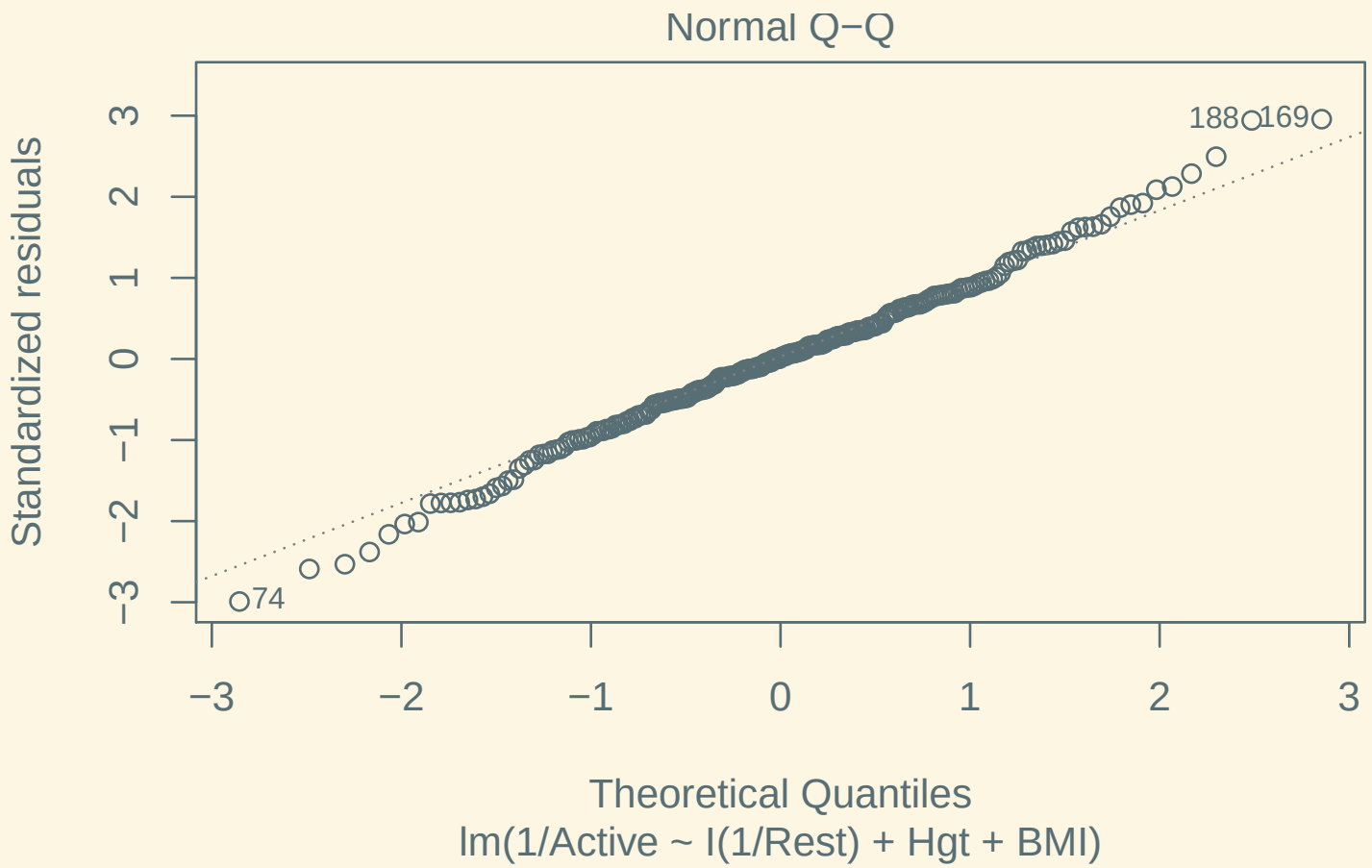
Go Back to Step 1

```
my.new.model <- lm(1 / Active ~ I(1 / Rest) + Hgt + BMI, data = PulseWithBMI)
plot(my.new.model, which = 1)
```



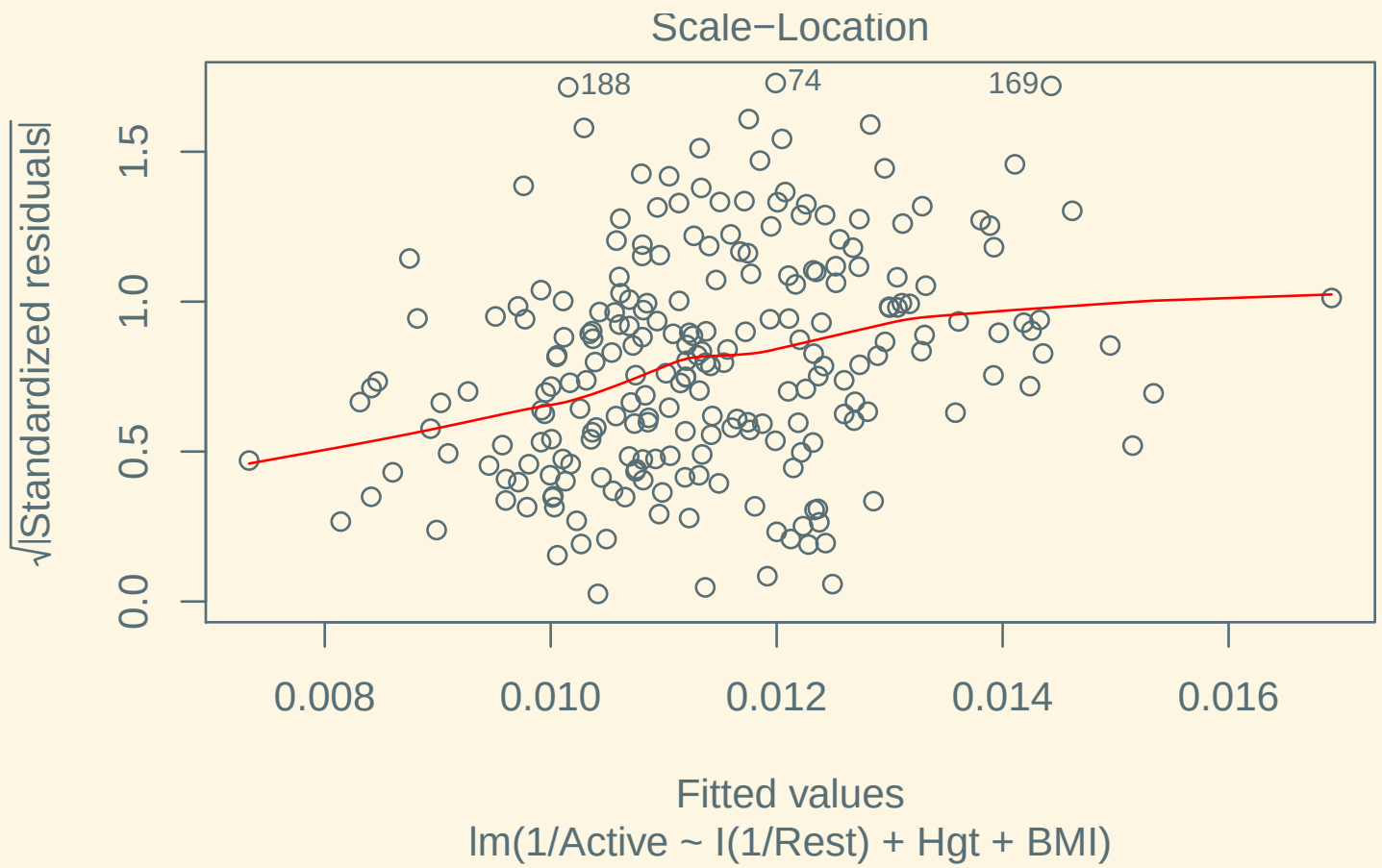
Re-ASSESS

```
plot(my.new.model, which = 2)
```



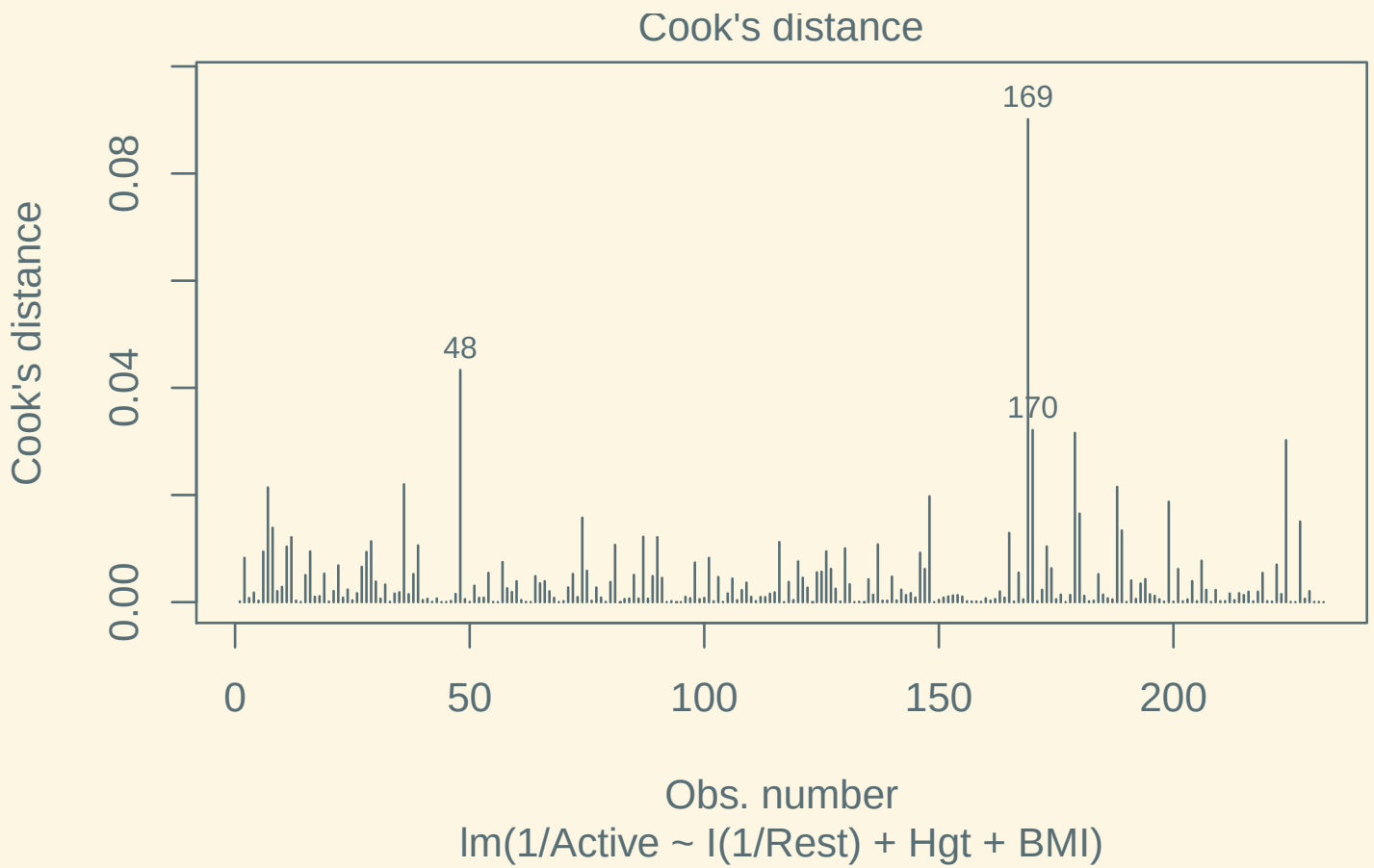
Go Back to Step 1

```
plot(my.new.model, which = 3)
```

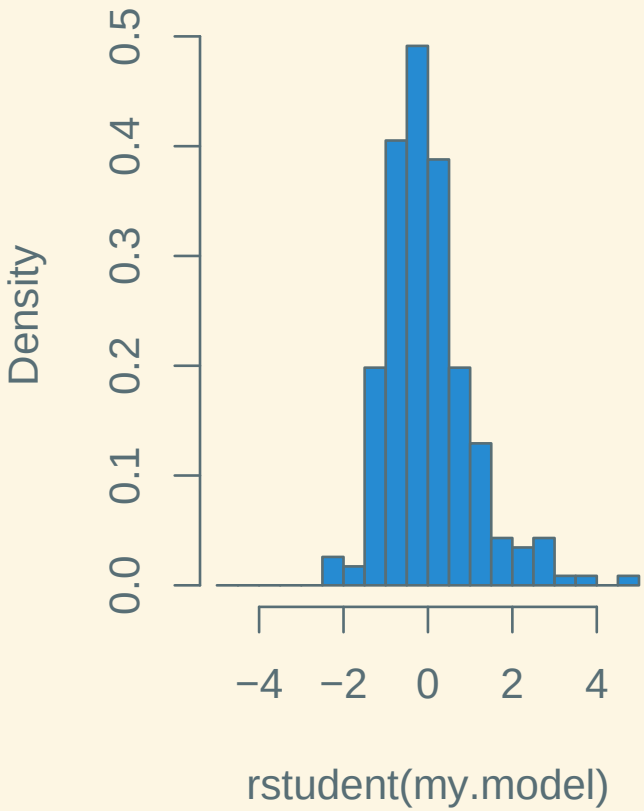


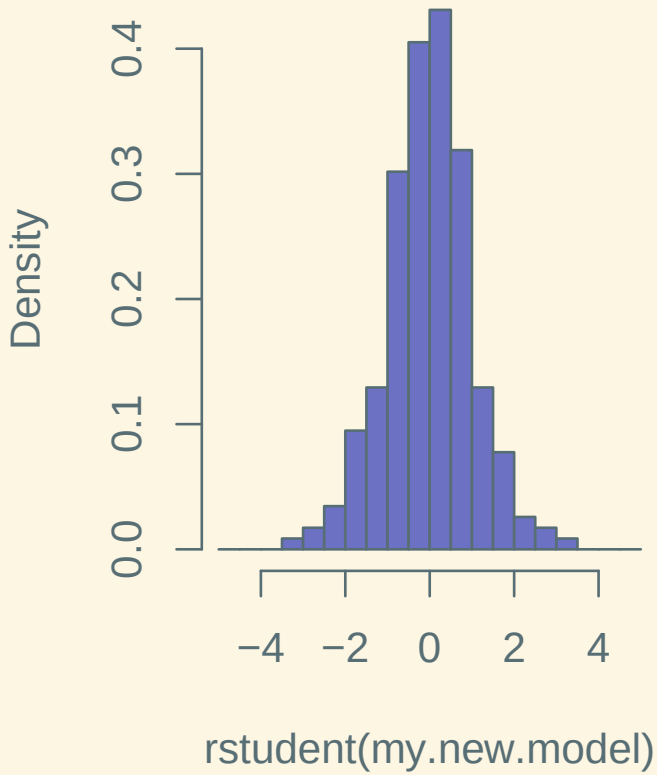
Go Back to Step 1

```
plot(my.new.model, which = 4)
```



Comparing Studentized Residuals





ASSESS: Test Individual Predictors (t-tests)

```
summary(my.new.model)


    Call:
    lm(formula = 1/Active ~ I(1/Rest) + Hgt + BMI, data = PulseWithBMI)

    Residuals:
           Min          1Q      Median          3Q         Max
    -0.0053245 -0.0010301  0.0000241  0.0011322  0.0052298

    Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
    (Intercept)  3.333e-04  2.187e-03   0.152   0.8790
    I(1/Rest)    6.506e-01  5.547e-02  11.728   <2e-16 ***
    Hgt          5.125e-05  3.376e-05   1.518   0.1304
    BMI         -9.052e-05  3.875e-05  -2.336   0.0204 *
    ---
    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Residual standard error: 0.001787 on 228 degrees of freedom
    Multiple R-squared:  0.4026,Adjusted R-squared:  0.3947
    F-statistic: 51.21 on 3 and 228 DF,  p-value: < 2.2e-16
```

Compare: t-test vs. Correlation

```
PulseWithBMI %>% mutate(InvActive = 1 / Active, InvRest = 1 / Rest) %>%
    select(InvActive, InvRest, Hgt, BMI) %>% cor() %>% round(digits = 2)
```

	InvActive	InvRest	Hgt	BMI
InvActive	1.00	0.62	0.18	-0.04
InvRest	0.62	1.00	0.22	0.09
Hgt	0.18	0.22	1.00	0.31
BMI	-0.04	0.09	0.31	1.00

BMI is more weakly correlated with InvActive than is Hgt, but yields a significant t-test, where Hgt does not.

Compare: t-test in MLR vs SLR models

```
summary(lm(1 / Active ~ Hgt, data = PulseWithBMI))



    Call:
    lm(formula = 1/Active ~ Hgt, data = PulseWithBMI)

    Residuals:
           Min          1Q      Median          3Q         Max
    -0.0052145 -0.0016130 -0.0001725  0.0012571  0.0075727

    Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
    (Intercept) 3.966e-03  2.724e-03   1.456  0.14678
    Hgt         1.090e-04  3.986e-05   2.736  0.00671 **
    ---
    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Residual standard error: 0.002265 on 230 degrees of freedom
    Multiple R-squared:  0.03152,Adjusted R-squared:  0.02731
    F-statistic: 7.485 on 1 and 230 DF,  p-value: 0.006706
```

Compare: t-test in MLR vs SLR models

```
summary(lm(1 / Active ~ BMI, data = PulseWithBMI))



    Call:
    lm(formula = 1/Active ~ BMI, data = PulseWithBMI)

    Residuals:
           Min          1Q      Median          3Q         Max
    -0.0047395 -0.0015873 -0.0001163  0.0012416  0.0080646

    Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
    (Intercept)  1.214e-02  1.131e-03   10.73   <2e-16 ***
    BMI         -3.080e-05  4.737e-05   -0.65    0.516
    ---
    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Residual standard error: 0.002299 on 230 degrees of freedom
    Multiple R-squared:  0.001835,Adjusted R-squared:  -0.002505
    F-statistic: 0.4228 on 1 and 230 DF,  p-value: 0.5162
```

Controls

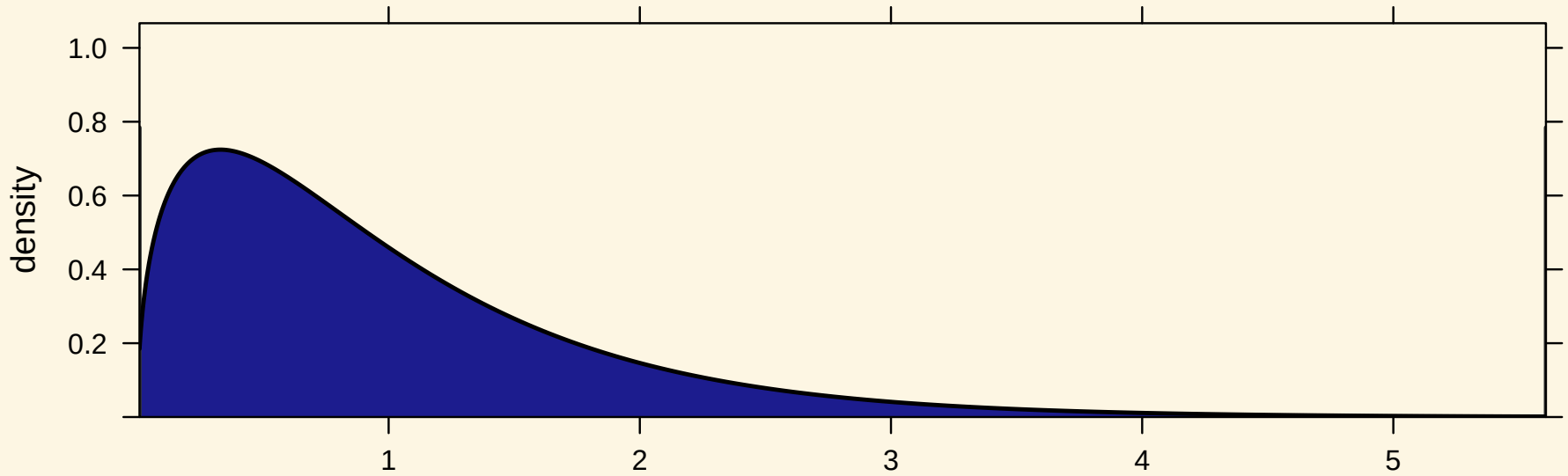
In the context of a multiple regression model, the t-test for a predictor tests for a linear association *after controlling for the other predictors*.

ASSESS: Test Overall Model

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$$H_1 : \text{Some } \beta_j \neq 0$$

$$F = \frac{MS_{Model}}{MS_{Error}} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2/k}{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)/(n-k-1)}$$



ASSESS: Coefficient of Determination

As before, $R^2 = \frac{SS_{Model}}{SS_{Total}} = 1 - \frac{SS_{Error}}{SS_{Total}}$

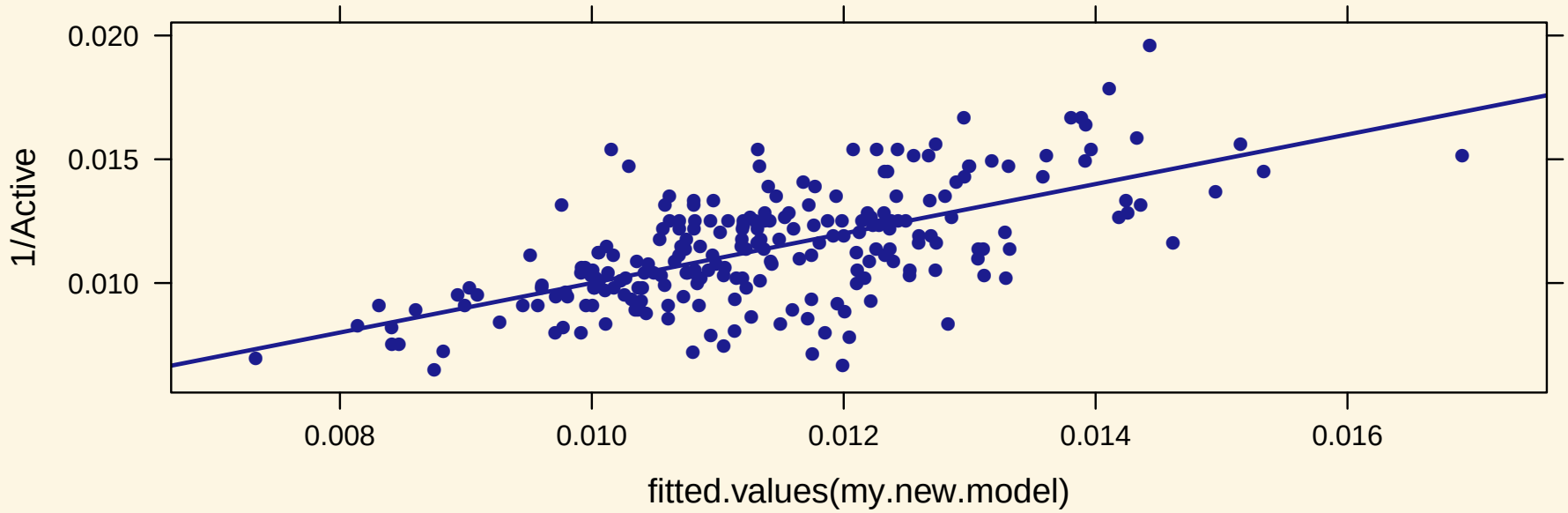
Coefficient of Determination

Interesting fact: R^2 is the square of $R = \text{Cor}(\hat{Y}, Y)$

```
summary(my.new.model)


    Call:
    lm(formula = 1/Active ~ I(1/Rest) + Hgt + BMI, data = PulseWithBMI)

    Residuals:
           Min          1Q      Median          3Q         Max
    -0.0053245 -0.0010301  0.0000241  0.0011322  0.0052298

    Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
    (Intercept)  3.333e-04  2.187e-03   0.152   0.8790
    I(1/Rest)    6.506e-01  5.547e-02  11.728   <2e-16 ***
    Hgt          5.125e-05  3.376e-05   1.518   0.1304
    BMI         -9.052e-05  3.875e-05  -2.336   0.0204 *
    ---
    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Residual standard error: 0.001787 on 228 degrees of freedom
    Multiple R-squared:  0.4026,Adjusted R-squared:  0.3947
    F-statistic: 51.21 on 3 and 228 DF,  p-value: < 2.2e-16
```

Coefficient of Determination

```
xyplot(1 / Active ~ fitted.values(my.new.model),
       data = PulseWithBMI, type = c("p","r"))
```




```
cor(1 / Active ~ fitted.values(my.new.model), data = PulseWithBMI)^2

     [1] 0.4025784
```