STAT 215
Multiple Regression II

Colin Reimer Dawson

Oberlin College

October 27, 2017

Outline

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

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
```

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

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
```

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

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

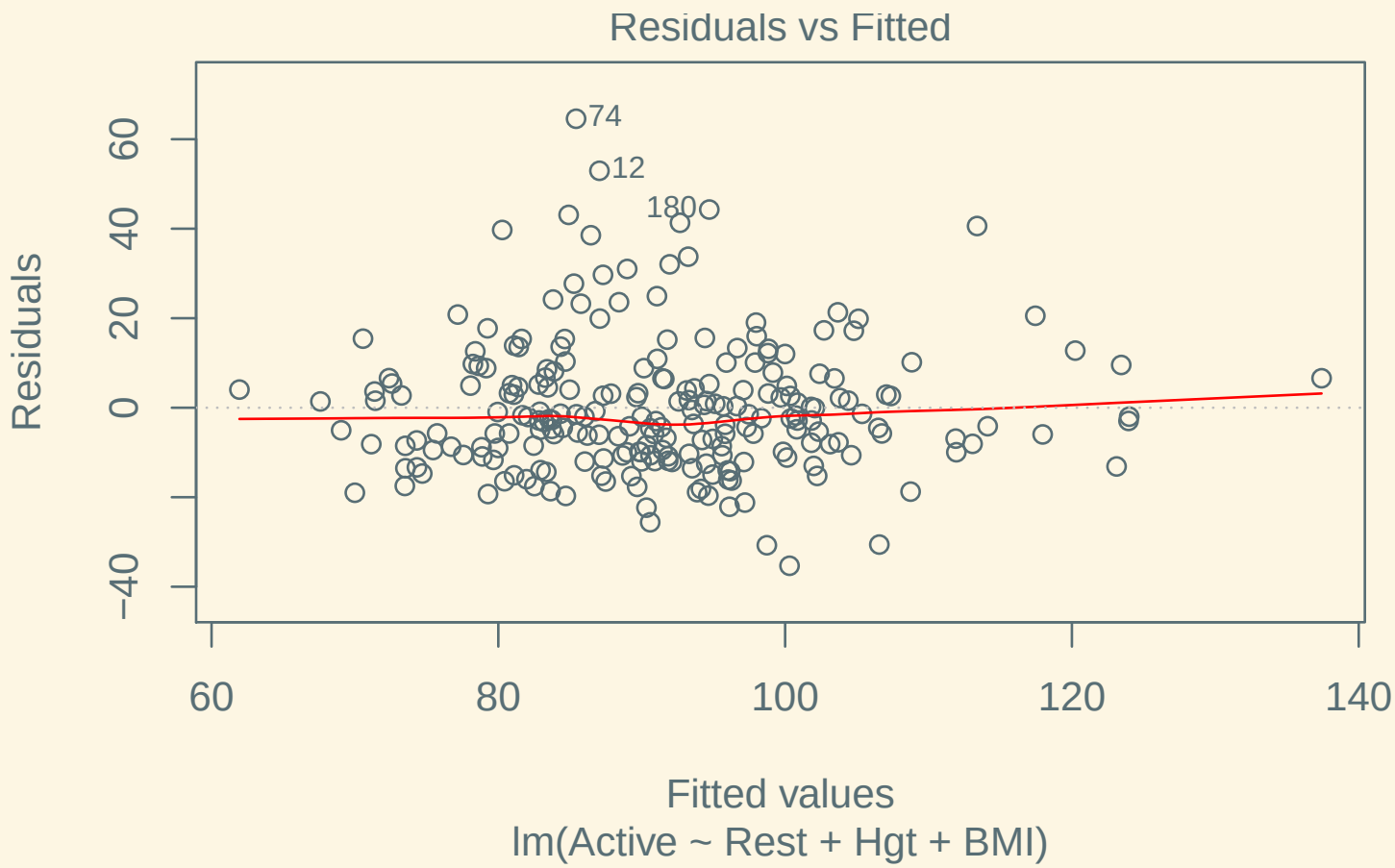
ASSESS: Check Conditions

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)

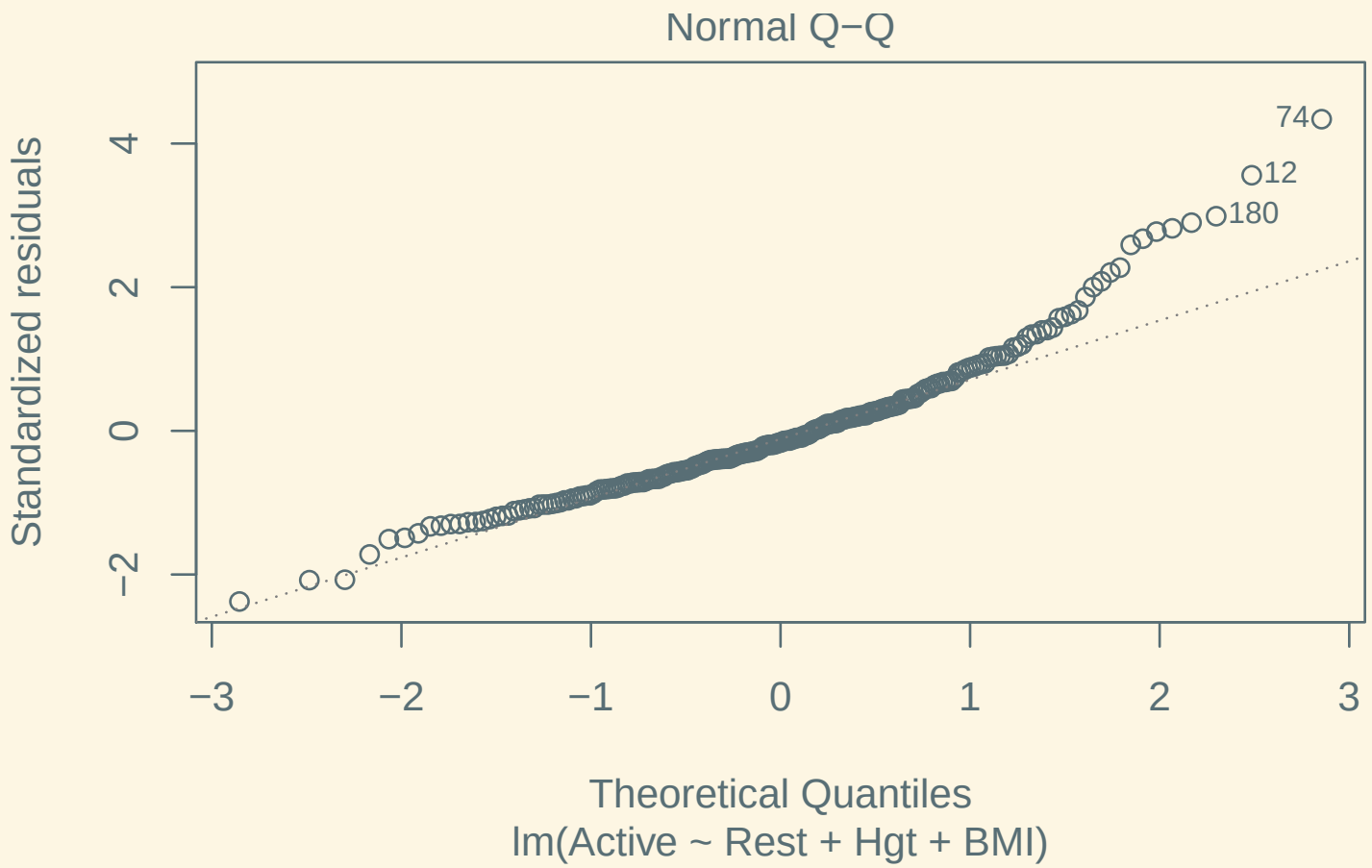
ASSESS: Check Conditions

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



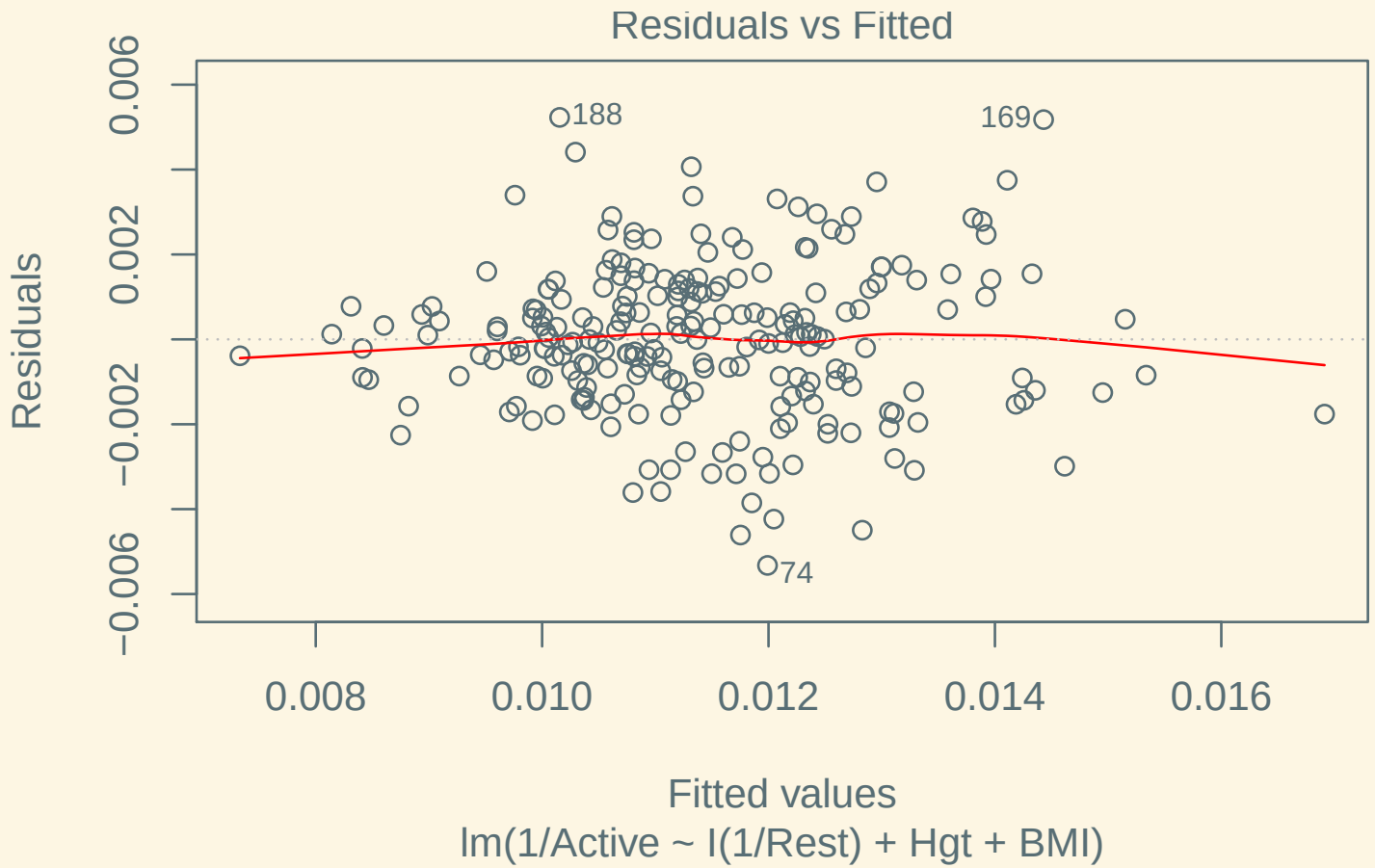
ASSESS: Check Conditions

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



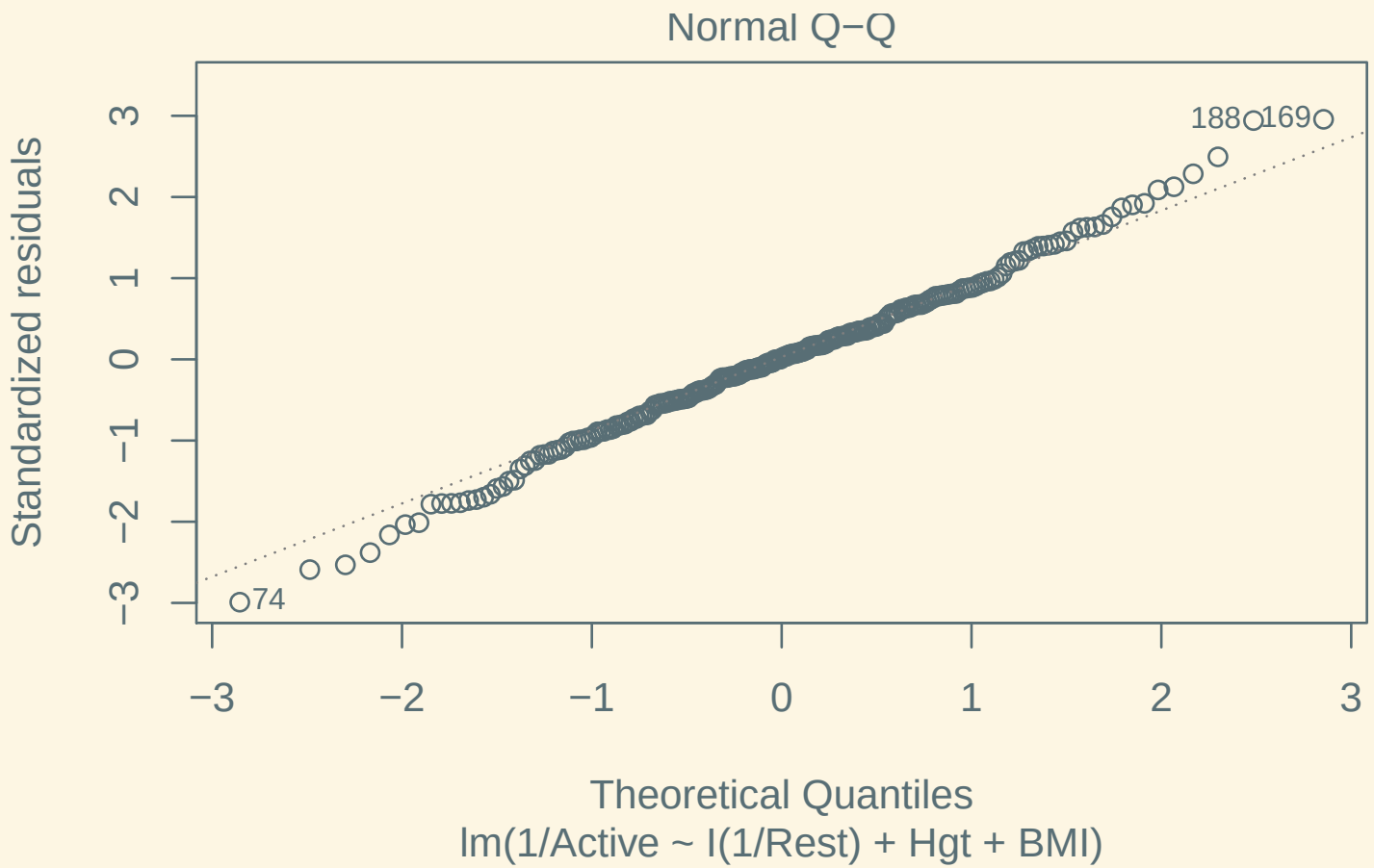
Go Back to Step 1

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

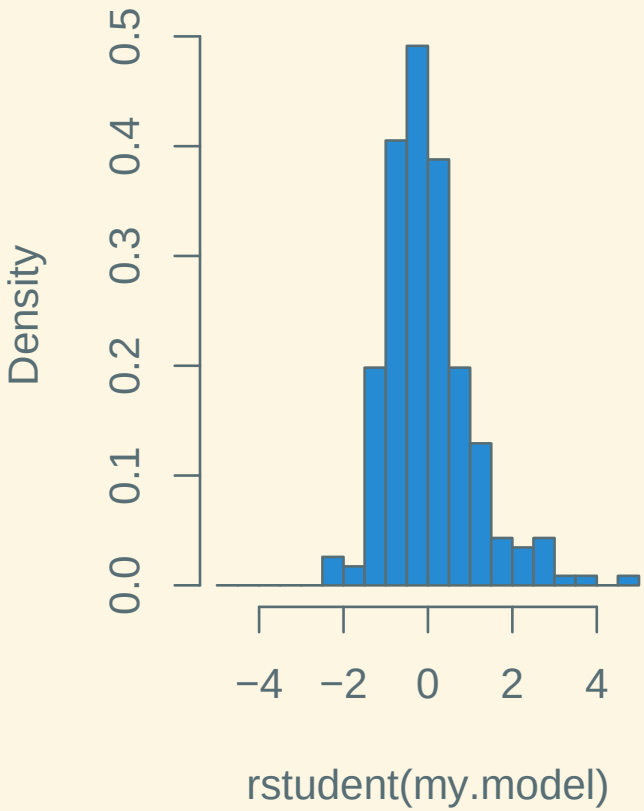


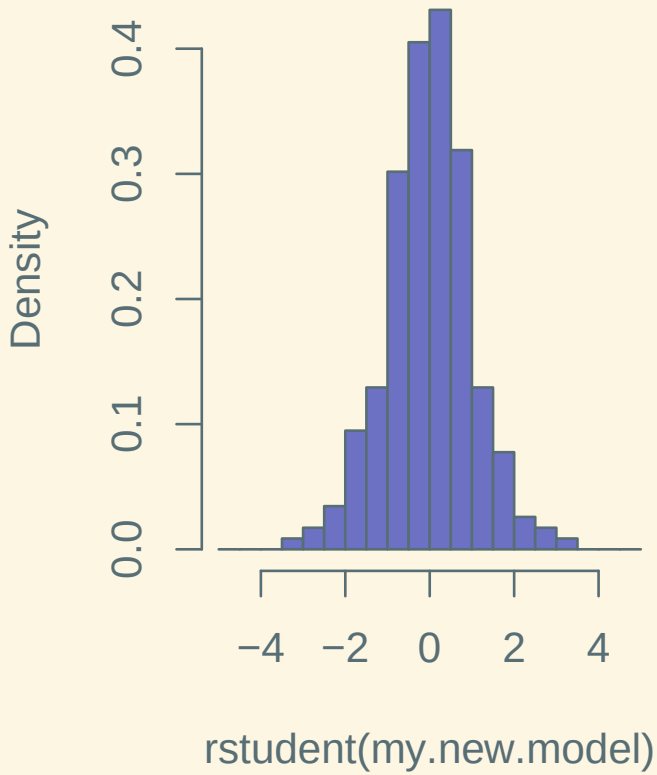
Go Back to Step 1

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



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()

                InvActive    InvRest       Hgt         BMI
InvActive  1.00000000 0.62157254 0.1775363 -0.04283408
InvRest    0.62157254 1.00000000 0.2166860  0.09276839
Hgt        0.17753629 0.21668603 1.0000000  0.31138335
BMI       -0.04283408 0.09276839 0.3113833  1.00000000
```

`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

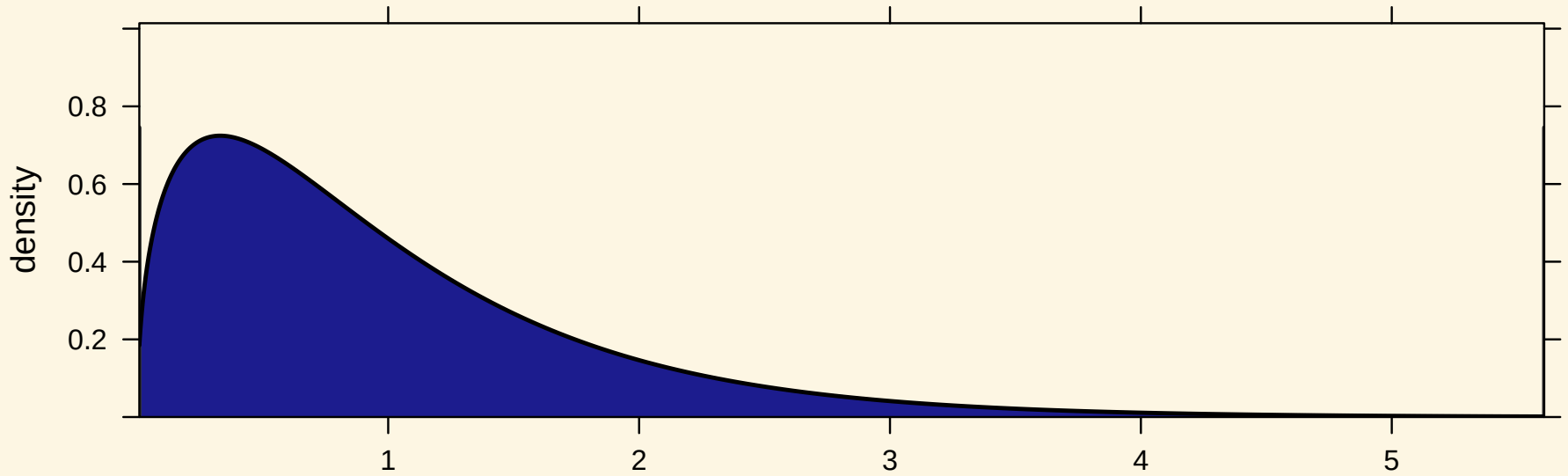
In the context of a multiple regression model, the t-test for a predictor tests for a linear association *after controlling for the other predictors*.

ASSESS: Test Overall Model

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

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

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



USE: CIs and PIs

Confidence and Prediction Intervals have same interpretation as in the single predictor case:

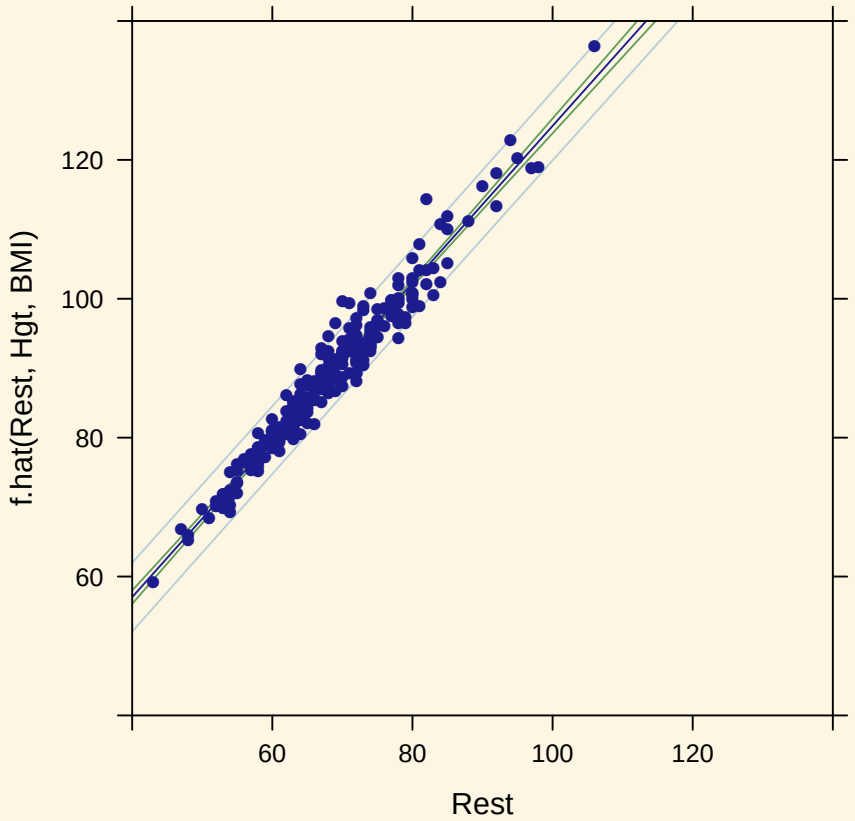
- C% CI: Procedure to produce an interval at a particular $(X_1, \ldots, X_k)$ that will contain the true $\hat{Y}$ for C% of data sets.
- C% PI: Procedure to produce an interval at a particular $(X_1, \ldots, X_k)$ that will contain the true Y for C% of "datasets plus a case".

In R

```
f.hat <- makeFun(my.new.model, transform = function(x) {1 / x})
## transform= defines the inverse of the transformation of the response
## used in the model so that we get intervals for the original variable

f.hat(Rest = 73, Hgt = 74, BMI = 25.8, interval = "confidence")

             fit      lwr      upr
     1 93.43854 97.81512 89.43684

f.hat(Rest = 73, Hgt = 74, BMI = 25.8, interval = "prediction")

             fit      lwr      upr
     1 93.43854 139.8696 70.15118
```

Bands With Some Predictors Fixed

```
xyplot(f.hat(Rest,Hgt,BMI) ~ Rest, Hgt = 74, BMI = 25.8, data = PulseWithBMI,
       panel = panel.lmbands, xlim = c(40,140), ylim = c(40,140))
```



Pulse Rates Revisited

```
library(Stat2Data); data("Pulse")
PulseWithBMI <-
    mutate(
        Pulse,
        BMI = Wgt / Hgt^2 * 703,
        InvActive = 1 / Active,
        InvRest = 1 / Rest,
        Male = 1 - Gender)
```

Active Pulse Rate by Sex

```
### Male = 1 for males, 0 for females
### factor() tells R this represents categories
apr.sex <- lm(Active ~ factor(Male), data = PulseWithBMI)
coef(apr.sex)

      (Intercept) factor(Male)1
        94.818182     -6.695231
```

What is the model here?

What does the coefficient for `Male` mean?

```
summary(apr.sex)


    Call:
    lm(formula = Active ~ factor(Male), data = PulseWithBMI)

    Residuals:
        Min      1Q  Median      3Q     Max
    -38.818 -12.894  -1.818  10.953  65.877

    Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
    (Intercept)     94.818      1.770  53.581  < 2e-16 ***
    factor(Male)1   -6.695      2.440  -2.744  0.00656 **
    ---
    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Residual standard error: 18.56 on 230 degrees of freedom
    Multiple R-squared:  0.03169,Adjusted R-squared:  0.02748
    F-statistic: 7.527 on 1 and 230 DF,  p-value: 0.006556
```

What does the t-test tell us?

Pair Discussion

(3 min.)

An environmental expert is interested in modeling the concentration of various chemicals in well water. Write down a regression model in which the amount of lead (`Lead`) depends on whether the well has been cleaned (`Iclean`, a 0/1 variable).

(5 min.)

Can you write down a single regression model that you could use to predict the amount of lead (`Lead`) in a well based on `Year` and on whether the well has been cleaned? How do you interpret each coefficient?

Another Example

A question of interest is how birth weights (`BirthWeightOz`) in North Carolina might be related to mother's race. The variable `MomRace` codes the mother's "race" as Black, Latinx, Other, or White. For the fitted model

$$\texttt{BirthWeightOz} = 117.87 + 7.96 \cdot \texttt{Latinx} + 6.58 \cdot \texttt{Other} + 7.31 \cdot \texttt{White}$$

the predictors are equal to 1 when the mother identifies with the race in question, and zero otherwise. What does each coefficient tell us about race and birth weights? (Assume that each mother picks one category to identify with.)

Combining Quantitative and Indicator Variables

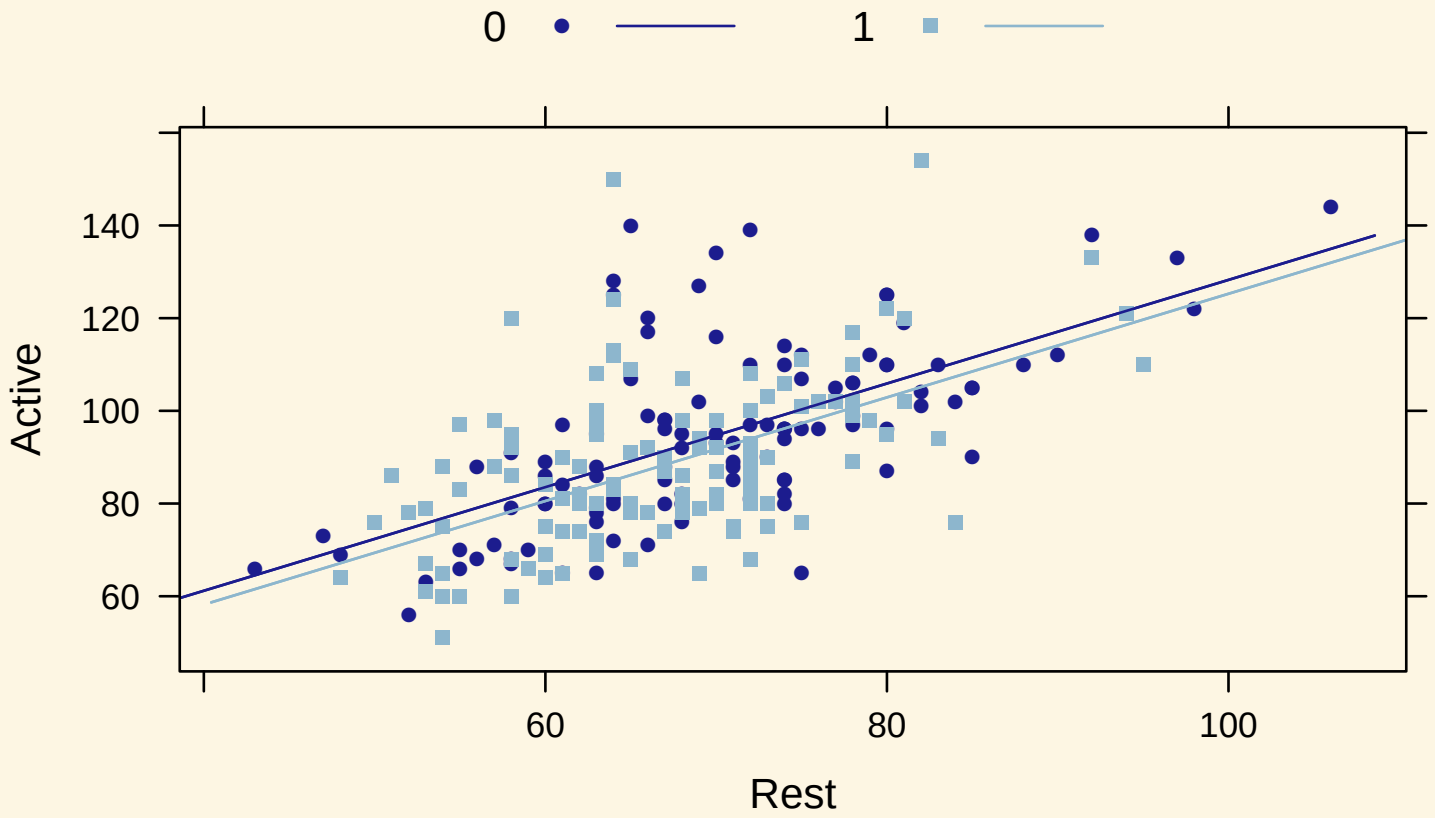
```
apr.sex.rest <- lm(Active ~ Rest + factor(Male), data = PulseWithBMI)
apr.sex.rest
```

```

Call:
lm(formula = Active ~ Rest + factor(Male), data = PulseWithBMI)

Coefficients:
  (Intercept)           Rest  factor(Male)1
       16.470          1.118         -2.993
```

$$\widehat{\texttt{Active}} = 16.47 + 1.12 \cdot \texttt{Rest} - 2.99 \cdot \texttt{Male}$$

Now what does the `Male` coefficient tell us?

```
## xyplot(Active ~ Rest, groups = Male, data = PulseWithBMI, auto.key = TRUE)
## f.hat <- makeFun(apr.sex.rest)
## lty = 1 for solid lty = 2 for dashed
## plotFun(f.hat(Rest, Male) ~ Rest, Male = 0, lty = 1, add = TRUE)
## plotFun(f.hat(Rest, Male) ~ Rest, Male = 1, lty = 2, add = TRUE)
plotModel(apr.sex.rest)
```



One Model, Two Prediction Equations

$$\widehat{\texttt{Active}} = 16.47 + 1.12 \cdot \texttt{Rest} - 2.99 \cdot \texttt{Male}$$

Females: $\widehat{\texttt{Active}} = 16.47 + 1.12 \cdot \texttt{Rest}$

Males: $\widehat{\texttt{Active}} = (16.47 - 2.99) + 1.12 \cdot \texttt{Rest}$

t-test for `Male` coefficient tests whether intercepts are different

```
summary(apr.sex.rest)


    Call:
    lm(formula = Active ~ Rest + factor(Male), data = PulseWithBMI)

    Residuals:
        Min      1Q  Median      3Q     Max
    -35.306  -9.766  -2.542   7.340  64.983

    Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
    (Intercept)    16.4703     7.1895   2.291   0.0229 *
    Rest            1.1178     0.1005  11.120   <2e-16 ***
    factor(Male)1  -2.9928     1.9987  -1.497   0.1357
    ---
    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Residual standard error: 14.99 on 229 degrees of freedom
    Multiple R-squared:  0.3712,Adjusted R-squared:  0.3657
    F-statistic: 67.59 on 2 and 229 DF,  p-value: < 2.2e-16
```

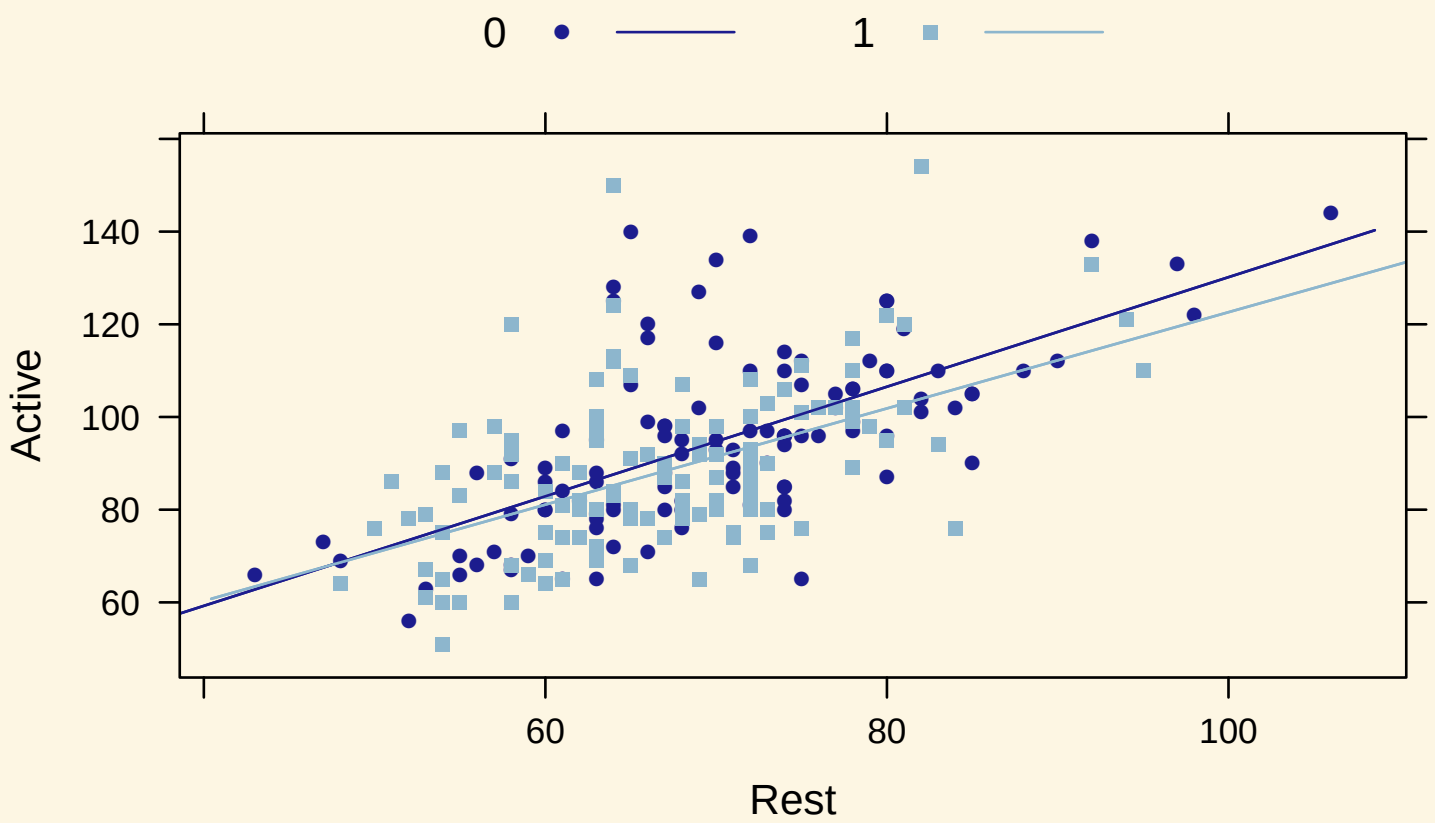
Non-Parallel Lines

```
two.lines.model <-
    lm(Active ~ Rest + factor(Male) + Rest:factor(Male),
       data = PulseWithBMI)
coef(two.lines.model)

            (Intercept)                    Rest      factor(Male)1
             11.9763226               1.1819202               6.8200842
     Rest:factor(Male)1
             -0.1437664
```

$$\texttt{Active} = 11.98 + 1.18 \cdot \texttt{Rest} + 6.82 \cdot \texttt{Male} - 0.14 \cdot \texttt{Rest} \cdot \texttt{Male}$$

Now what does the `Male` coefficient tell us? The last coefficient?

```
plotModel(two.lines.model)
```



Non-Parallel Lines

- `Male` coefficient is the difference in intercepts
- the **interaction term** is the difference in slopes

$$\widehat{\texttt{Active}} = 11.98 + 1.18 \cdot \texttt{Rest} + 6.82 \cdot \texttt{Male} - 0.14 \cdot \texttt{Rest} \cdot \texttt{Male}$$

Females: $\widehat{\texttt{Active}} = 11.98 + 1.18 \cdot \texttt{Rest}$

Males: $\widehat{\texttt{Active}} = (11.98 + 6.82) + (1.18 - 0.14) \cdot \texttt{Rest}$

t-test for $\texttt{Male} \cdot \texttt{Rest}$ coefficient tests whether slopes are different

```
summary(two.lines.model)



    Call:
    lm(formula = Active ~ Rest + factor(Male) + Rest:factor(Male),
        data = PulseWithBMI)

    Residuals:
        Min      1Q  Median      3Q     Max
    -35.620  -9.933  -2.524   6.764  64.762

    Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
    (Intercept)         11.9763     9.5839   1.250    0.213
    Rest                 1.1819     0.1352   8.742 5.08e-16 ***
    factor(Male)1        6.8201    13.9629   0.488    0.626
    Rest:factor(Male)1  -0.1438     0.2025  -0.710    0.478
    ---
    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Residual standard error: 15.01 on 228 degrees of freedom
    Multiple R-squared:  0.3726,Adjusted R-squared:  0.3643
    F-statistic: 45.13 on 3 and 228 DF,  p-value: < 2.2e-16
```

Caution

Test for different intercepts is not a test for separate lines when the fitted lines are not parallel: could be that the difference at $X = 0$ is smaller than elsewhere

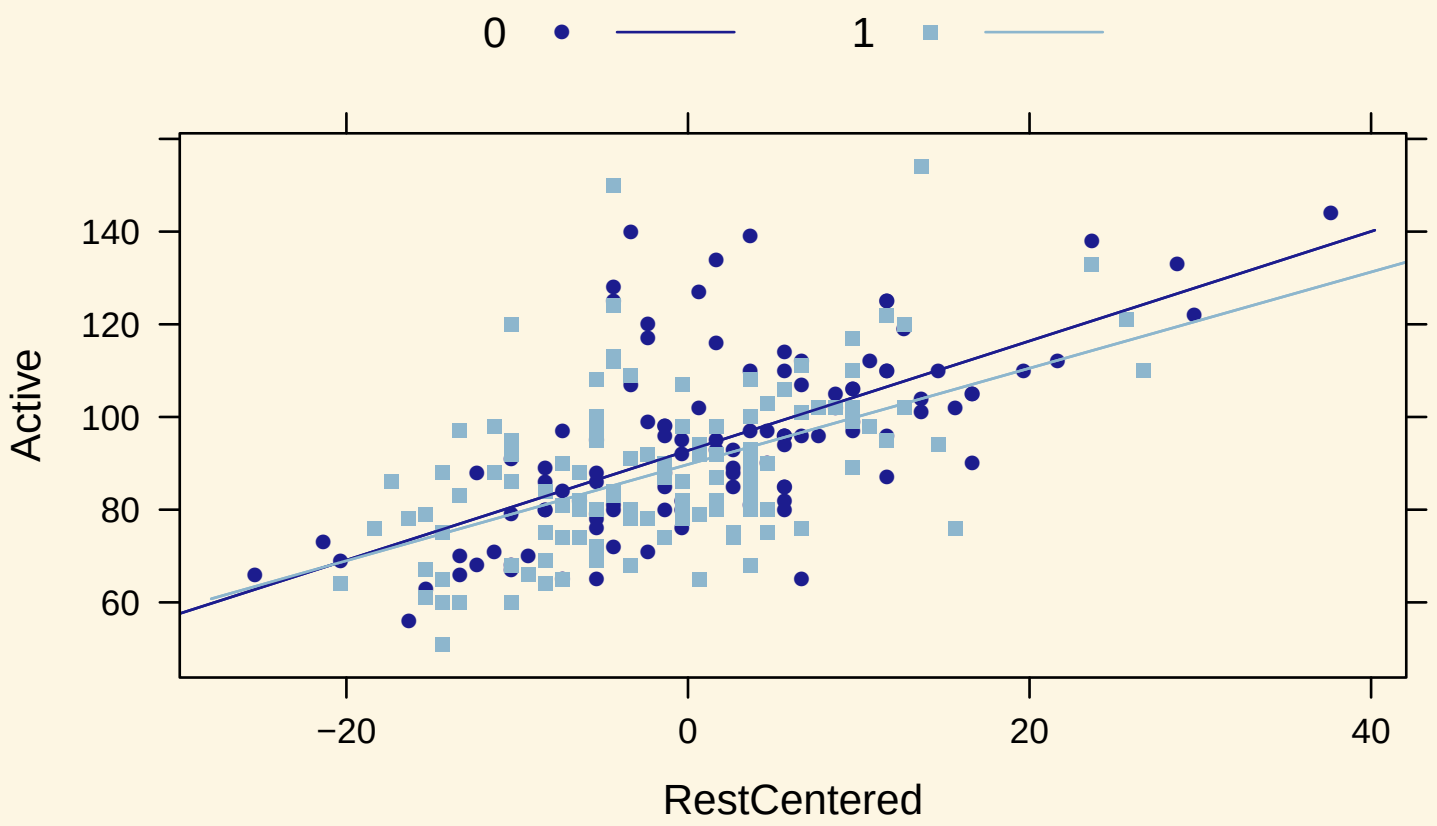
Centering a Predictor

```
PulseWithBMI <- mutate(PulseWithBMI, RestCentered = Rest - mean(Rest))
two.lines.model <-
    lm(Active ~ RestCentered + factor(Male) + RestCentered:factor(Male),
       data = PulseWithBMI)
coef(two.lines.model)

                    (Intercept)                    RestCentered
                     92.7595474                       1.1819202
                 factor(Male)1 RestCentered:factor(Male)1
                     -3.0062286                      -0.1437664
```

$$\texttt{Active} = 92.76 + 1.18 \cdot \texttt{Rest} - 3.01 \cdot \texttt{Male} - 0.14 \cdot \texttt{Rest} \cdot \texttt{Male}$$

Now what does the `Male` coefficient tell us?

```
plotModel(two.lines.model)
```



Pair Discussion Revisited

Can you write down a single regression model that you could use to predict the amount of lead (`Lead`) in a well based on `Year`, but where the trend line is different depending on whether or not the well has been cleaned (`Iclean`)? What coefficients do you need and what is their interpretation?

Testing multiple (but not all) predictors

We can test:

- one term at a time (t-test)

$$H_0 : \beta_k = 0 \qquad H_1 : \beta_k \neq 0$$

- all terms at once (F-test)

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_K = 0$$
$$H_1 : \text{Some } \beta_k \neq 0$$

- What if we want to test a *subset* of the βs together?

Nested Models

If Model B has all the terms in Model A and then some, we say that Model A is **nested** in Model B

Model A: $\texttt{Active} = \beta_0 + \beta_1 \texttt{Rest}$

Model B: $\texttt{Active} = \beta_0 + \beta_1 \texttt{Rest} + \beta_2 \texttt{Male} + \beta_3 \texttt{Male} \cdot \texttt{Rest}$

Model A is nested in Model B

Comparing Nested Models

- Is there evidence that the additional predictors in Model B are helpful?
- Some of SS_{Error} for the simpler model moves to SS_{Model} for the complex model.
- Nested F-test: is this difference more than we would expect by chance?
- $H_0 : \beta_{K_A+1} = \cdots = \beta_{K_B} = 0$

$$F_{Comparison} = \frac{MS_{Comparison}}{MSE_{Full}}$$
$$= \frac{\text{Increase in } SS_{Model}/\text{Increase in } df_{Model}}{MSE_{Full}}$$

Nested F-test

```
modelA <- lm(Active ~ Rest, data = PulseWithBMI)
modelB <- lm(Active ~ Rest + factor(Male) + factor(Male):Rest,
             data = PulseWithBMI)
anova(modelA,modelB)
```

```
Analysis of Variance Table

Model 1: Active ~ Rest
Model 2: Active ~ Rest + factor(Male) + factor(Male):Rest
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    230 51953
2    228 51335  2    617.27 1.3708  0.256
```