STAT 213
Assessment of MLR Models

Colin Reimer Dawson

Oberlin College

February 23, 2018

Notes

Outline

Notes

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 (may need to return to step 1)
   - Test individual predictors (t-tests)
   - Test/measure overall fit (ANOVA, R^2)
   - Model comparison/selection
4. USE the model
   - Make predictions
   - Construct CIs and PIs

Notes

Outline

Notes

CHOOSE: Active Pulse Rate

```
library(Stat2Data); data("Pulse")
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
```

$$\texttt{Active}_i = \beta_0 + \beta_1 \cdot \texttt{Rest}_i + \beta_2 \cdot \texttt{Hgt}_i + \beta_3 \cdot \texttt{BMI}_i + \varepsilon_i$$

Notes

FIT: Estimate Coefficients

The Multiple Regression Population Model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_K X_{iK} + \varepsilon_i$$

The Multiple Regression Fitted Model

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_K X_{1K} + \hat{\varepsilon}_i$$

Choose $\hat{\beta}_k$s that minimize SSE (requires linear algebra / vector calculus)

Notes

FIT: Estimate Coefficients

```
my.model <- lm(Active ~ Rest + Hgt + BMI, data = PulseWithBMI)
coef(my.model) %>% round(digits = 1)

     (Intercept)         Rest          Hgt          BMI
            22.7          1.1         -0.4          0.7
```

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

Notes

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 $\hat{\sigma}^2_\varepsilon$.

Notes

Regression Degrees of Freedom

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

This is the number of extra "free parameters" (compared to the null model)

$df_{Error} = N - K - 1$ 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 N points exactly with N coefficients including the intercept.)

Notes

FIT: Estimate Residual Variance

$$\hat{\sigma}^2_\varepsilon = MS_{Error} = \frac{SS_{Error}}{df_{Error}} = \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{N - K - 1}$$

Notes

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
```


Notes

FIT: The Final Model

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

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

Notes

Outline

Notes

ASSESS: Check Conditions

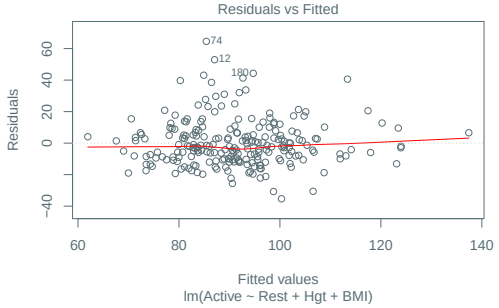
Same conditions as before apply:

1. Linearity/Zero Mean Residuals (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)

Notes

ASSESS: Check Conditions

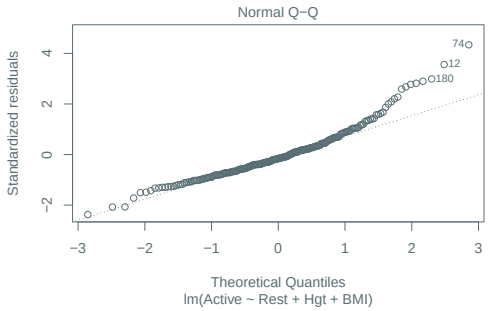
```
plot(my.model, which = 1)
```



Notes

ASSESS: Check Conditions

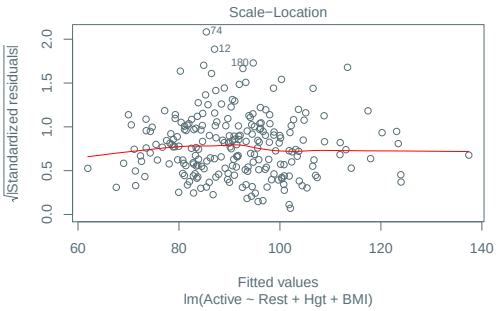
```
plot(my.model, which = 2)
```



Notes

ASSESS: Check Conditions

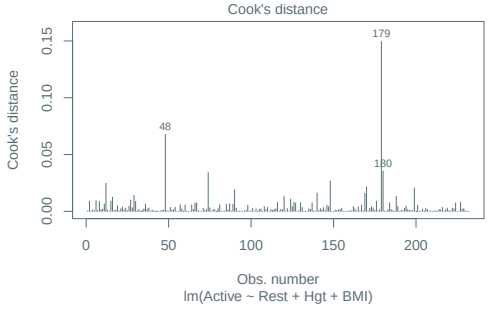
```
plot(my.model, which = 3)
```



Notes

ASSESS: Check Conditions

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



Notes

Back to CHOOSE (Transformations)

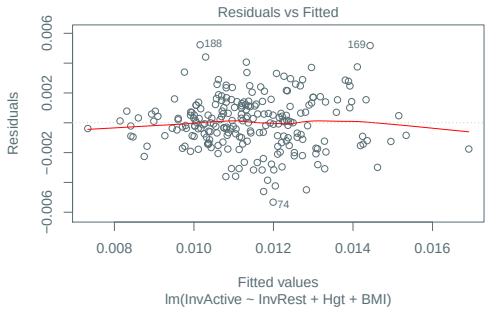
Since pulse is a ratio (beats/min), could try a reciprocal transformation (min/beat)

```
PulseWithBMI <- mutate(PulseWithBMI, InvActive = 1 / Active, InvRest = 1 / Rest)
```

$$\frac{1}{\texttt{Active}_i} = \beta_0 + \beta_1 \cdot \frac{1}{\texttt{Rest}_i} + \beta_2 \cdot \texttt{Hgt}_i + \beta_3 \cdot \texttt{BMI} + \varepsilon_i$$

Notes

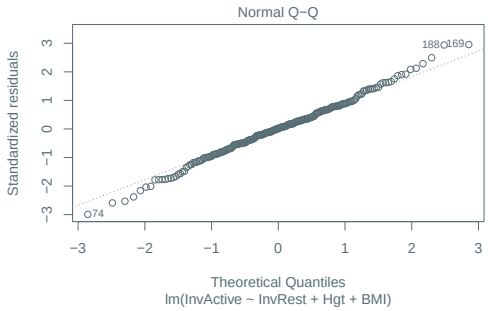

```
my.new.model <- lm(InvActive ~ InvRest + Hgt + BMI, data = PulseWithBMI)
plot(my.new.model, which = 1)
```



Notes

Re-ASSESS

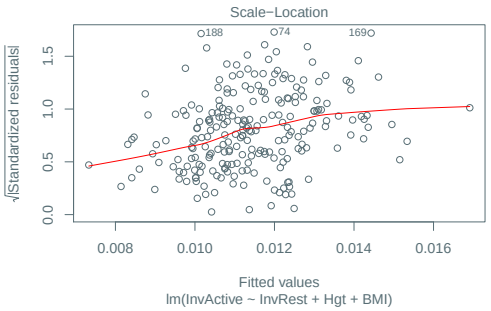
```
plot(my.new.model, which = 2)
```



Notes

Re-ASSESS

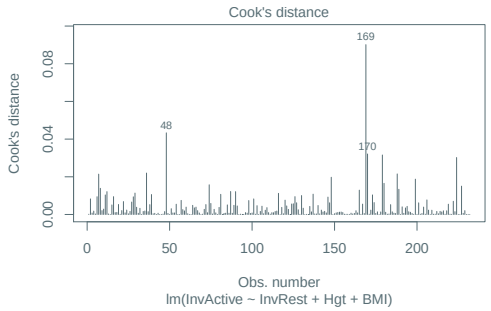
```
plot(my.new.model, which = 3)
```



Notes

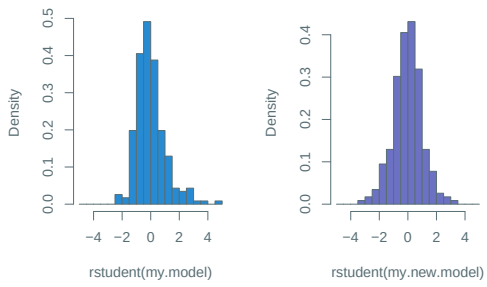
Re-ASSESS

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



Notes

Comparing Studentized Residuals



Notes

ASSESS: Test Individual Predictors (t-tests)

```
summary(my.new.model)


    Call:
    lm(formula = InvActive ~ InvRest + 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
    InvRest      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
```


Notes

Compare: t-test vs. Correlation

```
PulseWithBMI %>% 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.

Notes

Compare: t-test in MLR vs SLR models

```
summary(lm(InvActive ~ Hgt, data = PulseWithBMI))


    Call:
    lm(formula = InvActive ~ 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
```


Notes

Compare: t-test in MLR vs SLR models

```
summary(lm(InvActive ~ BMI, data = PulseWithBMI))


    Call:
    lm(formula = InvActive ~ 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
```


Notes

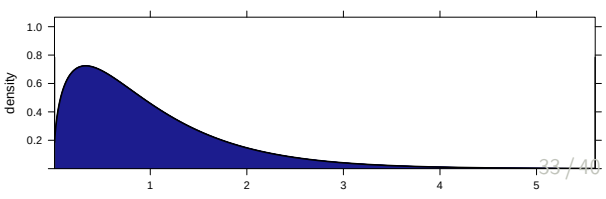
Controls

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**.

Notes

ASSESS: Test Overall Model

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_K = 0$$
$$H_1 : \text{Some } \beta_k \neq 0$$
$$F = \frac{MS_{Model}}{MS_{Error}} = \frac{\sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2/K}{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)/(N-K-1)}$$



Notes

ASSESS: Coefficient of Determination

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

Notes

Coefficient of Determination

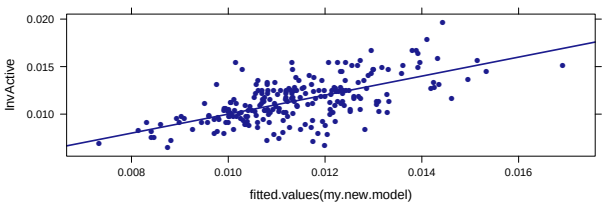
Interesting fact: R^2 is the square of $R = \text{Cor}(\hat{Y}, Y)$

```
rsquared(my.new.model)

     [1] 0.4025784

r <- cor(fitted(my.new.model) ~ InvActive, data = PulseWithBMI)
r

     [1] 0.6344907

r^2

     [1] 0.4025784
```


Notes

Coefficient of Determination

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




```
cor(InvActive ~ fitted(my.new.model), data = PulseWithBMI)^2

     [1] 0.4025784
```


Notes

Outline

Notes

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" $f(X_{*1}, \ldots, X_{*K})$ 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".

Notes

In R

```
f.hat <- makeFun(my.new.model, transformation = 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(InvRest = 1/73, Hgt = 74, BMI = 24, interval = "confidence")

           fit      lwr      upr
     1 92.03739 96.32315 88.11675

f.hat(InvRest = 1/73, Hgt = 74, BMI = 24, interval = "prediction")

           fit      lwr      upr
     1 92.03739 136.7648 69.35546
```


Notes