STAT 213
Indicator Variables in MLR

Colin Reimer Dawson

Oberlin College

February 28, 2018

Notes

Outline

Notes

Outline

Notes

Pulse Rates Revisited

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


Notes

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) %>% round(digits = 2)

      (Intercept) factor(Male)1
            94.82         -6.70
```

What is the model here?
What does the coefficient for `Male` mean?

Notes


```
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?

Notes

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?

Notes

Combining Quantitative and Indicator Variables

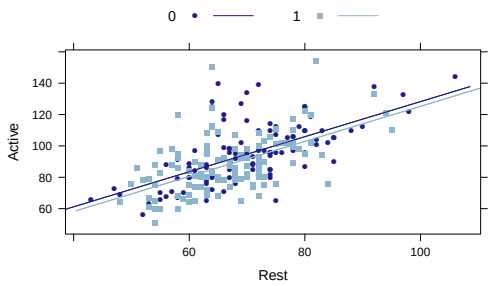
```
apr.sex.rest <- lm(Active ~ Rest + factor(Male), data = PulseWithBMI)
apr.sex.rest


     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?

Notes


```
## CAUTION: don't try to use this with multiple quantitative
## predictors; it won't make sense
plotModel(apr.sex.rest)
```



Notes

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

Notes


```
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
```


Notes

Non-Parallel Lines

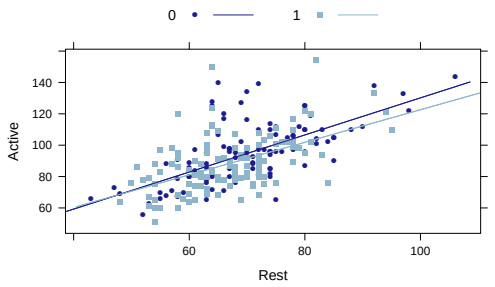
```
two.lines.model <-
    lm(Active ~ Rest + factor(Male) + Rest:factor(Male),
       data = PulseWithBMI)
coef(two.lines.model)

           (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?

Notes

```
## CAUTION: don't try to use this with multiple quantitative
## predictors; it won't make sense
plotModel(two.lines.model)
```



Notes

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

Notes


```
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
```


Notes

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

Notes

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) %>% round(digits = 2)
```

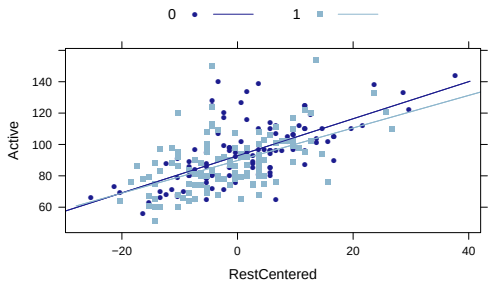
```
                  (Intercept)                  RestCentered
                        92.76                          1.18
              factor(Male)1 RestCentered:factor(Male)1
                        -3.01                         -0.14
```

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

Now what does the `Male` coefficient tell us?

Notes


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



Notes

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?

Notes

Outline

Notes

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?

Notes

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

Notes

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?
- $H_0 : \beta_{K_A+1} = \cdots = \beta_{K_B} = 0$

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

Notes

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
```

Conclusion: Little evidence that males and non-males need a different model

Notes