MLR: Extensions

Nate Wells

Math 243: Stat Learning

September 18th, 2020

Outline

In today's class, we will...

- Generalize MLR to include categorical variables
- Discuss non-linear "linear" regression models

Section 1

Extending the Linear Model

Qualitative Predictors

Thus far, we have assumed all predictors are quantitative (taking values on a scale).

Qualitative Predictors

Thus far, we have assumed all predictors are quantitative (taking values on a scale).

- It would nice to include qualitative predictors in our model.

Qualitative Predictors

Thus far, we have assumed all predictors are quantitative (taking values on a scale).

- It would nice to include qualitative predictors in our model.
- But if we try to include them naively, we immediately run into trouble:

Qualitative Predictors

Thus far, we have assumed all predictors are quantitative (taking values on a scale).

- It would nice to include qualitative predictors in our model.
- But if we try to include them naively, we immediately run into trouble:

$$\hat{\text{Balance}} = f(X_1, X_2, X_3) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{Income} + \hat{\beta}_2 \cdot \text{Limit} + \hat{\beta}_3 \cdot \text{Gender}$$

Qualitative Predictors

Thus far, we have assumed all predictors are quantitative (taking values on a scale).

- It would nice to include qualitative predictors in our model.
- But if we try to include them naively, we immediately run into trouble:

$$\hat{\text{Balance}} = f(X_1, X_2, X_3) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{Income} + \hat{\beta}_2 \cdot \text{Limit} + \hat{\beta}_3 \cdot \text{Gender}$$

$$\text{Suppose } \hat{\beta}^T = \begin{pmatrix} -400 & -7.5 & .25 & 2.5 \end{pmatrix}$$

$$\hat{\text{Debt}} = f(10, 4000, \text{Female}) = -400 - 7.5 \cdot 10 + .25 \cdot 4000 + 2.5 \cdot \text{Female} = ???$$

Coding and Dummy Variables

- For binary categorical variables, we create a new *quantitative* variable by coding the first level as 0 and the second as 1.

Coding and Dummy Variables

- For binary categorical variables, we create a new *quantitative* variable by coding the first level as 0 and the second as 1.
  - For 'Gender', we could code: $1 \leftarrow \text{Female} \quad 0 \leftarrow \text{Male}$

Coding and Dummy Variables

- For binary categorical variables, we create a new *quantitative* variable by coding the first level as 0 and the second as 1.
  - For 'Gender', we could code: $1 \leftarrow \text{Female} \quad 0 \leftarrow \text{Male}$

$$\hat{\text{Debt}} = f(7.5, 4000, \text{Female}) = -400 - 7.5 \cdot 10 + 0.25 \cdot 4000 + 2.5 \cdot 1 = 527.5$$

Coding and Dummy Variables

- For binary categorical variables, we create a new *quantitative* variable by coding the first level as 0 and the second as 1.
  - For 'Gender', we could code: $1 \leftarrow \text{Female} \quad 0 \leftarrow \text{Male}$

$$\hat{\text{Debt}} = f(7.5, 4000, \text{Female}) = -400 - 7.5 \cdot 10 + 0.25 \cdot 4000 + 2.5 \cdot 1 = 527.5$$

- In general, if X_1 is quantitative and X_2 is categorical, the resulting model will be

$$\hat{Y} = f(X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 = \begin{cases} (\beta_0 + \beta_2) + \beta_1 X_1, & \text{if obs. in 1st level,} \\ \beta_0 + \beta_1 X_1, & \text{if obs. in 2nd level.} \end{cases}$$

Coding and Dummy Variables

- For binary categorical variables, we create a new *quantitative* variable by coding the first level as 0 and the second as 1.
  - For 'Gender', we could code: $1 \leftarrow \text{Female} \quad 0 \leftarrow \text{Male}$

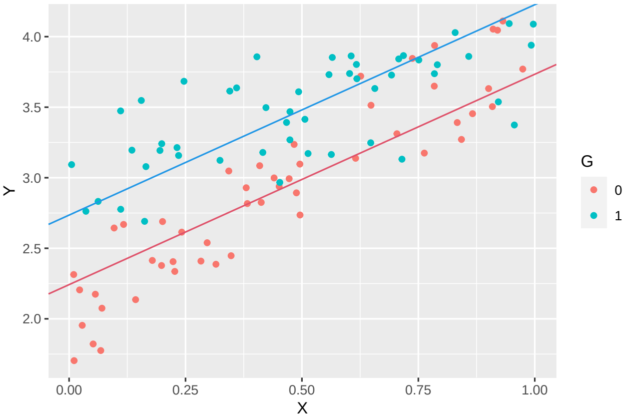
$$\hat{\text{Debt}} = f(7.5, 4000, \text{Female}) = -400 - 7.5 \cdot 10 + 0.25 \cdot 4000 + 2.5 \cdot 1 = 527.5$$

- In general, if X_1 is quantitative and X_2 is categorical, the resulting model will be

$$\hat{Y} = f(X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 = \begin{cases} (\beta_0 + \beta_2) + \beta_1 X_1, & \text{if obs. in 1st level,} \\ \beta_0 + \beta_1 X_1, & \text{if obs. in 2nd level.} \end{cases}$$

Note that both regression lines have the same slope, but different intercept.

Scatterplot



$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 G = 2.28 + 1.41X + 0.53G$$

The model in R

```r
mod_2<- lm(data = my_data, Y ~ X + G)
summary(mod_2)
```

```
##
## Call:
## lm(formula = Y ~ X + G, data = my_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.78728 -0.16815  0.00389  0.16433  0.58123
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.24222    0.06122  36.627  < 2e-16 ***
## X            1.49117    0.10168  14.665  < 2e-16 ***
## G1           0.49298    0.05873   8.394 3.87e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2925 on 97 degrees of freedom
## Multiple R-squared:  0.7618, Adjusted R-squared:  0.7569
## F-statistic: 155.1 on 2 and 97 DF,  p-value: < 2.2e-16
```

Poll 3: MLR Slope Interpretation

The slope on a (binary) categorical variable G tells us (select all that apply)

a. How much we expect the response to change if we increase the value of G from 0 to 1, while holding all else constant.

b. The difference in the average response between observations in the two categories.

c. The value of the response variable if G equals 0.

d. The distance between the two regression lines on the 2d scatterplot

Categorical Variables with more than 2 levels.

We extend to variables with more than 2 levels by creating binary variables for each level.

Categorical Variables with more than 2 levels.

We extend to variables with more than 2 levels by creating binary variables for each level.

In the `Credit` data set, the `Ethnicity` variable takes 3 levels: `African American`, `Asian`, `Caucasion`. (As with `Gender`, the levels here are incomplete)

Categorical Variables with more than 2 levels.

We extend to variables with more than 2 levels by creating binary variables for each level.

In the `Credit` data set, the `Ethnicity` variable takes 3 levels: `African American`, `Asian`, `Caucasion`. (As with `Gender`, the levels here are incomplete)

For categorical variable X_i with levels $j = 1, \ldots, k$, create a dummy variables x_{ij} by

$$x_{ij} = \begin{cases} 1, & \text{obs. in level } j, \\ 0, & \text{obs. not in level } j, \end{cases}$$

Categorical Variables with more than 2 levels.

We extend to variables with more than 2 levels by creating binary variables for each level.

In the `Credit` data set, the `Ethnicity` variable takes 3 levels: `African American`, `Asian`, `Caucasion`. (As with `Gender`, the levels here are incomplete)

For categorical variable X_i with levels $j = 1, \ldots, k$, create a dummy variables x_{ij} by

$$x_{ij} = \begin{cases} 1, & \text{obs. in level } j, \\ 0, & \text{obs. not in level } j, \end{cases}$$

For example,

$$\text{Eth}_{AA} = \begin{cases} 1, & \text{obs. is African American,} \\ 0, & \text{obs. is not African America} \end{cases}$$

$$\text{Eth}_{A} = \begin{cases} 1, & \text{obs. is Asian,} \\ 0, & \text{obs. is not Asian} \end{cases}$$

$$\text{Eth}_{C} = \begin{cases} 1, & \text{obs. is Caucasion,} \\ 0, & \text{obs. is not Caucasion} \end{cases}$$

Categorical Variables with more than 2 levels.

We extend to variables with more than 2 levels by creating binary variables for each level.

In the `Credit` data set, the `Ethnicity` variable takes 3 levels: `African American`, `Asian`, `Caucasion`. (As with `Gender`, the levels here are incomplete)

For categorical variable X_i with levels $j = 1, \ldots, k$, create a dummy variables x_{ij} by

$$x_{ij} = \begin{cases} 1, & \text{obs. in level } j, \\ 0, & \text{obs. not in level } j, \end{cases}$$

For example,

$$\text{Eth}_{AA} = \begin{cases} 1, & \text{obs. is African American}, \\ 0, & \text{obs. is not African America} \end{cases}$$

$$\text{Eth}_{A} = \begin{cases} 1, & \text{obs. is Asian}, \\ 0, & \text{obs. is not Asian} \end{cases}$$

$$\text{Eth}_{C} = \begin{cases} 1, & \text{obs. is Caucasion}, \\ 0, & \text{obs. is not Caucasion} \end{cases}$$

- Every observation evaluates to 1 in exactly 1 dummy variable.

Categorical Variables in R

```
credit_mod <- lm(Balance ~ Limit + Income + Gender + Ethnicity, data = Credit)
summary(credit_mod)$coefficients
```

```
##                         Estimate   Std. Error     t value      Pr(>|t|)
## (Intercept)         -395.7122121 25.890307793 -15.2841834  9.661647e-42
## Limit                  0.2645314  0.005894931  44.8743906 6.014584e-157
## Income                -7.6671626  0.386036409 -19.8612421  2.508448e-61
## GenderFemale           1.9069535 16.599113684   0.1148828  9.085965e-01
## EthnicityAsian        26.8788662 23.412591822   1.1480517  2.516438e-01
## EthnicityCaucasian     3.7623916 20.399222553   0.1844380  8.537648e-01
```

$$\hat{\text{Balance}} = -395.7 + 0.26 \cdot \text{L} - 7.67 \cdot \text{I} + 1.91 \cdot \text{G}_F + 26.88 \cdot \text{E}_A + 3.76 \cdot \text{E}_C$$

Categorical Variables in R

```
credit_mod <- lm(Balance ~ Limit + Income + Gender + Ethnicity, data = Credit)
summary(credit_mod)$coefficients
```

```
##                         Estimate   Std. Error     t value      Pr(>|t|)
## (Intercept)         -395.7122121 25.890307793 -15.2841834  9.661647e-42
## Limit                  0.2645314  0.005894931  44.8743906 6.014584e-157
## Income                -7.6671626  0.386036409 -19.8612421  2.508448e-61
## GenderFemale           1.9069535 16.599113684   0.1148828  9.085965e-01
## EthnicityAsian        26.8788662 23.412591822   1.1480517  2.516438e-01
## EthnicityCaucasian     3.7623916 20.399222553   0.1844380  8.537648e-01
```

$$\widehat{\text{Balance}} = -395.7 + 0.26 \cdot \text{L} - 7.67 \cdot \text{I} + 1.91 \cdot \text{G}_F + 26.88 \cdot \text{E}_A + 3.76 \cdot \text{E}_C$$

But wait, some of the levels of the categorical variables are missing!

Section 2

Non-linearity

Exam Example

The `Exam` data set gives `midterm` score, `final` exam score, and self-reported `hours` of sleep prior to the final exam.

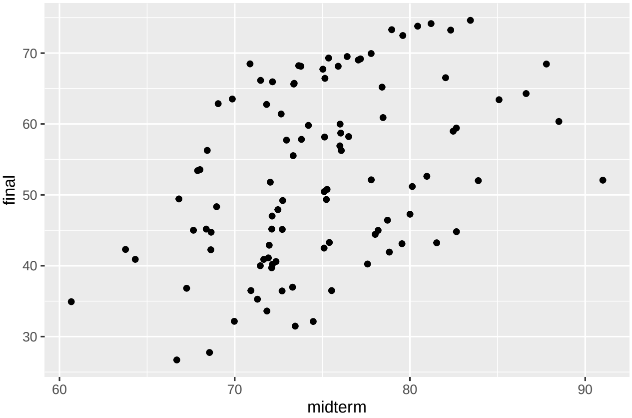
The model

```r
exam_mod<-lm(final ~ midterm + hours, data = Exam)
summary(exam_mod)
```

```
##
## Call:
## lm(formula = final ~ midterm + hours, data = Exam)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7902 -2.2642  0.0658  1.9715 10.9368
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -40.53841    4.76809  -8.502 2.28e-13 ***
## midterm       0.65929    0.06375  10.341  < 2e-16 ***
## hours         7.33650    0.23666  31.000  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.44 on 97 degrees of freedom
## Multiple R-squared:  0.9254, Adjusted R-squared:  0.9239
## F-statistic:   602 on 2 and 97 DF,  p-value: < 2.2e-16
```

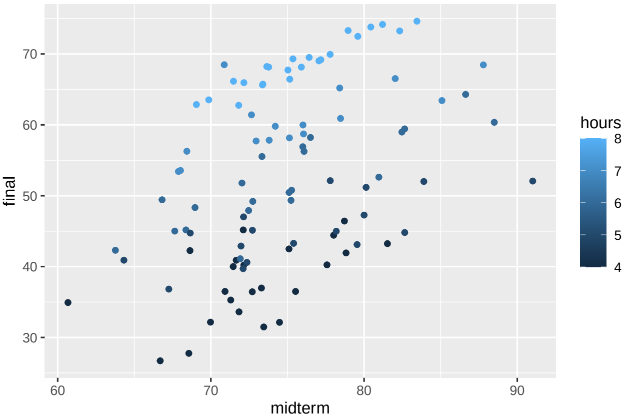
Scatterplot

```
Exam %>% ggplot(aes(x = midterm, y = final ))+geom_point()
```



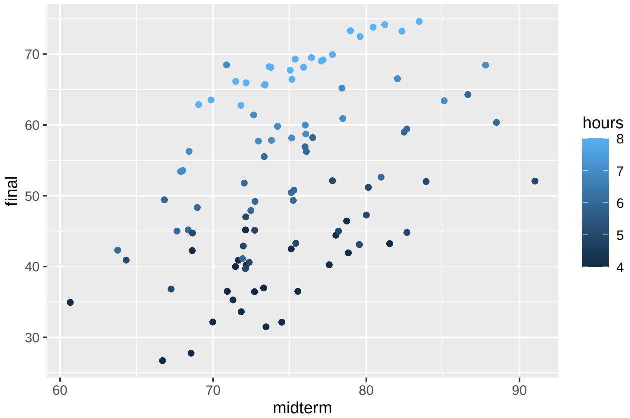
Scatterplot with hours

```
Exam %>% ggplot(aes(x = midterm, y = final, color = hours ))+geom_point()
```



Scatterplot with hours

```
Exam %>% ggplot(aes(x = midterm, y = final, color = hours ))+geom_point()
```



Does the **relationship** between midterm and final depend on hours of sleep?

Interaction Terms

To account for fact that change in final score per unit increase in midterm score depends on hours slept, we include an **interaction** term in the model:

Interaction Terms

To account for fact that change in final score per unit increase in midterm score depends on hours slept, we include an **interaction** term in the model:

$$Y = \beta_0 + \beta_1 X_2 + \beta_2 X_2 + \epsilon \qquad \text{Old model} \; Y = \quad \beta_0 + \beta_1 X_2 + \beta_2 X_2 + \beta_3 X_1 X_3 + \epsilon \qquad \text{New r}$$
$$Y = \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon \qquad \tilde{\beta}_1 = \beta_1 + \beta_3 X_3$$

Revised Model

The model

```
exam_mod_int<-lm(final ~ midterm + hours + midterm:hours, data = Exam)
summary(exam_mod_int)
```

```
##
## Call:
## lm(formula = final ~ midterm + hours + midterm:hours, data = Exam)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6515 -2.2002 -0.0273  1.7879 11.3687
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -8.91391   22.38489  -0.398    0.691
## midterm        0.23444    0.30066   0.780    0.437
## hours          1.87000    3.78889   0.494    0.623
## midterm:hours  0.07321    0.05064   1.446    0.152
##
## Residual standard error: 3.421 on 96 degrees of freedom
## Multiple R-squared:  0.927,  Adjusted R-squared:  0.9248
## F-statistic: 406.5 on 3 and 96 DF,  p-value: < 2.2e-16
```

The model

```
exam_mod_int<-lm(final ~ midterm + hours + midterm:hours, data = Exam)
summary(exam_mod_int)
```

```
##
## Call:
## lm(formula = final ~ midterm + hours + midterm:hours, data = Exam)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6515 -2.2002 -0.0273  1.7879 11.3687
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -8.91391   22.38489  -0.398    0.691
## midterm        0.23444    0.30066   0.780    0.437
## hours          1.87000    3.78889   0.494    0.623
## midterm:hours  0.07321    0.05064   1.446    0.152
##
## Residual standard error: 3.421 on 96 degrees of freedom
## Multiple R-squared:  0.927,  Adjusted R-squared:  0.9248
## F-statistic: 406.5 on 3 and 96 DF,  p-value: < 2.2e-16
```

$$Y = -2.4 + 0.1 \cdot \text{midterm} + 0.5 \cdot \text{hours} + 0.1 \cdot \text{midterm} \cdot \text{hours} + \epsilon$$

The model

```
exam_mod_int<-lm(final ~ midterm + hours + midterm:hours, data = Exam)
summary(exam_mod_int)
```

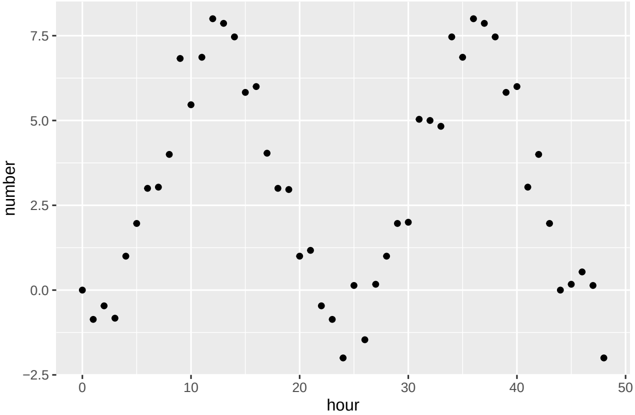
```
##
## Call:
## lm(formula = final ~ midterm + hours + midterm:hours, data = Exam)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6515 -2.2002 -0.0273  1.7879 11.3687
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -8.91391   22.38489  -0.398    0.691
## midterm        0.23444    0.30066   0.780    0.437
## hours          1.87000    3.78889   0.494    0.623
## midterm:hours  0.07321    0.05064   1.446    0.152
##
## Residual standard error: 3.421 on 96 degrees of freedom
## Multiple R-squared:  0.927,  Adjusted R-squared:  0.9248
## F-statistic: 406.5 on 3 and 96 DF,  p-value: < 2.2e-16
```

$$Y = -2.4 + 0.1 \cdot \text{midterm} + 0.5 \cdot \text{hours} + 0.1 \cdot \text{midterm} \cdot \text{hours} + \epsilon$$

- The coefficient on the interaction term measures increase in effectiveness of midterm score per unit increase in hours slept.

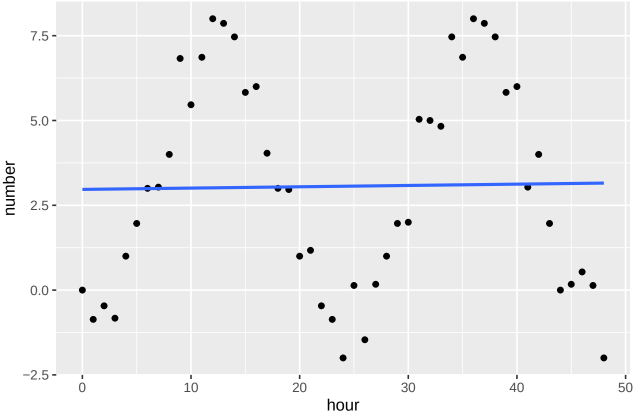
Other Non-linear models

The `emails` data set consists of the `number` of emails I receive in a given `hour` over two days



Other Non-linear models

The `emails` data set consists of the `number` of emails I receive in a given `hour` over two days



Including non-linear terms

We can theorize a polynomial model for Y

$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot X^2 + \cdots + \beta_p \cdot X^p + \epsilon$$

Including non-linear terms

We can theorize a polynomial model for Y

$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot X^2 + \cdots + \beta_p \cdot X^p + \epsilon$$

- This model is non-linear in the sense that the regression curve is not a straight line. And that there is non-constant change in Y per unit change in X.

Including non-linear terms

We can theorize a polynomial model for Y

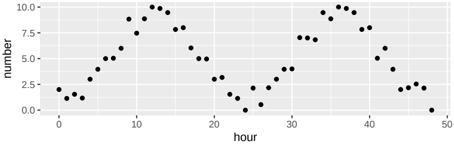
$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot X^2 + \cdots + \beta_p \cdot X^p + \epsilon$$

- This model is non-linear in the sense that the regression curve is not a straight line. And that there is non-constant change in Y per unit change in X.
- But it **is** linear in powers of the predictor.

Poll: What model?

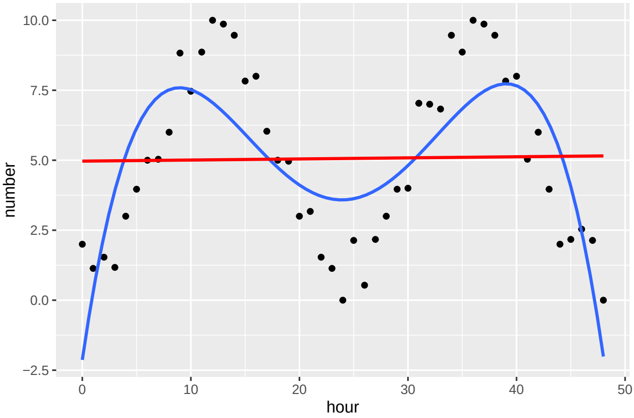
What polynomial degree seems most appropriate for the given data?

a. 1
b. 2
c. 3
d. 4
e. More than 4

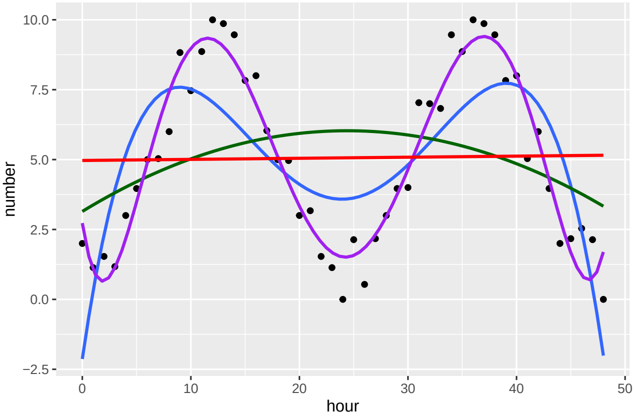


Plotting non-linear regression curves

```
ggplot(emails, aes( x = hour, y = number)) +geom_point() +
  geom_smooth(method = "lm", se = F, formula = y ~ poly(x, 4 )) +
  geom_smooth(method = "lm", se = F, color = "red")
```



Plotting non-linear regression curves II



Modeling with non-linear terms

```
emails_mod<-lm(number ~ poly(hour, degree = 4), data = emails)
summary(emails_mod)
```

```
##
## Call:
## lm(formula = number ~ poly(hour, degree = 4), data = emails)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5826 -1.8274  0.0919  1.8082  4.1322
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               5.06122    0.30649  16.514  < 2e-16 ***
## poly(hour, degree = 4)1   0.38386    2.14541   0.179  0.85882
## poly(hour, degree = 4)2  -6.06575    2.14541  -2.827  0.00704 **
## poly(hour, degree = 4)3  -0.09759    2.14541  -0.045  0.96392
## poly(hour, degree = 4)4 -15.20063    2.14541  -7.085 8.58e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.145 on 44 degrees of freedom
## Multiple R-squared:  0.5696, Adjusted R-squared:  0.5305
## F-statistic: 14.56 on 4 and 44 DF,  p-value: 1.193e-07
```