Multilinear Regression Extensions

Prof Wells

STA 295: Stat Learning

February 15th, 2024

Outline

In today's class, we will...

- Investigation several extensions to the linear model, including
  - Data Transformations
  - Indicator Variables
  - Non-linear models

Section 1

Diagnostic Plots

Common Problems

Most problems fall into 1 of 6 categories:

1. Non-linearity of relationship between predictors and response
2. Correlation of error terms
3. Non-constant variance in error
4. Outliers
5. High-leverage points
6. Collinearity of predictors

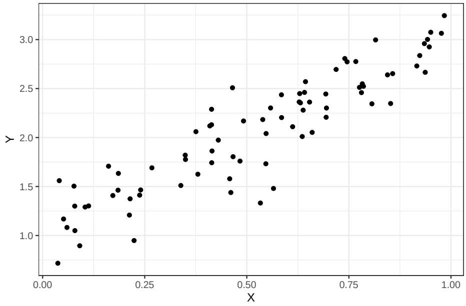
A Valid Model

Let's begin by creating a valid linear model to use as a baseline:

$$Y = 1 + 2X + \epsilon \qquad \epsilon \sim N(0, 0.25)$$

```
set.seed(700)
X <- runif(80, 0, 1)
e <- rnorm(80, 0, .25)
Y <- 1 + 2*X + e
my_data <- data.frame(X,Y)

theme_set(theme_bw())
ggplot(my_data, aes(x = X , y = Y)) + geom_point()
```



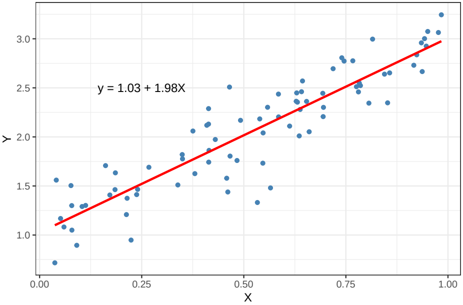
Linear Model

```r
my_mod<-lm(Y ~ X, data = my_data)
my_mod$coefficients
```

```
## (Intercept)           X
##    1.025947    1.981375
```

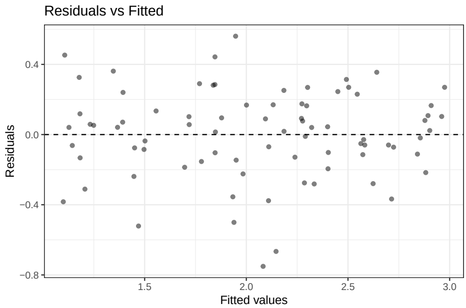
```r
summary(my_mod)$r.sq
```

```
## [1] 0.8275073
```



Residual Plot

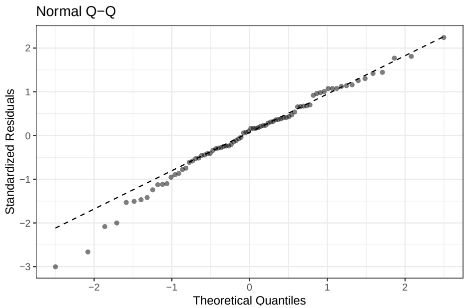
```
library(gglm)
ggplot(data = my_mod) +stat_fitted_resid()
```



- What does the plot show?
  - Fitted values are along the horizontal axis, residuals are along the vertical axis.
- What should we look for?
  - Non-linear patterns

QQ Plot

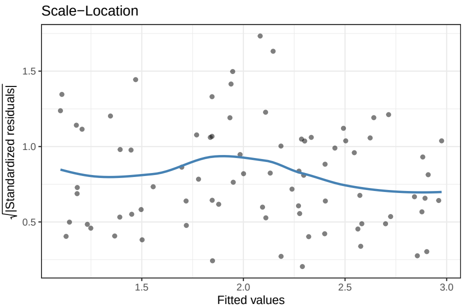
```
library(gglm)
ggplot(data = my_mod) +stat_normal_qq()
```



- What does the plot show?
    - Quantiles of reference distribution (Normal) along horizontal axis; observed quantiles of sample along vertical axis.
- What should we look for?
    - Deviations from a straight line, indicating non-Normality

Scale-Location Plot

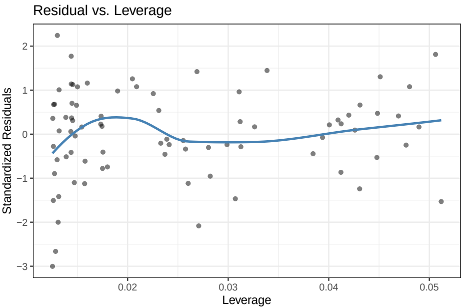
```
library(gglm)
ggplot(data = my_mod) +stat_scale_location()
```



- What does the plot show?
  - Fitted values are along the horizontal axis, **standardized** residuals are along the vertical axis.
- What should we look for?
  - Horizontal line indicates constant variability of residuals; other patterns may suggest non-constant variability

Leverage Plot

```
library(gglm)
ggplot(data = my_mod) +stat_resid_leverage()
```



- What does the plot show?
  - Leverage (distance from mean) along horizontal axis, residuals along the vertical axis.
- What should we look for?
  - Points that have high leverage and large residual are **influential**

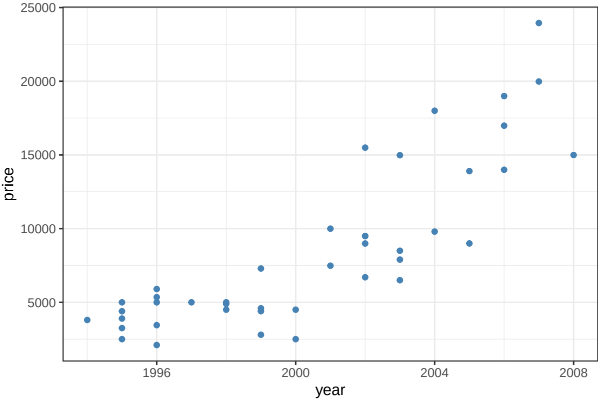
Section 2

Transformations

Example: Truck Prices

Can we use the age of a truck to predict what its price should be?

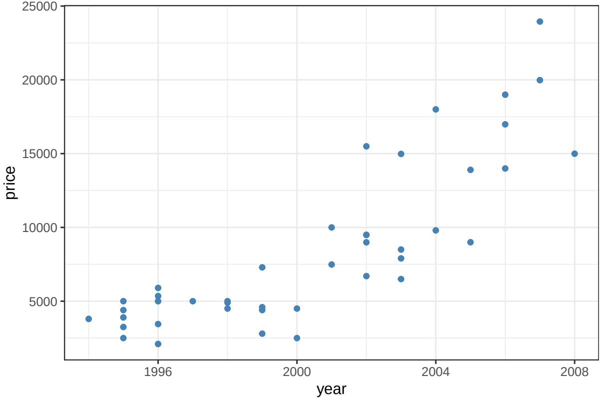
- Consider a random sample of 43 pickup trucks between 1994 and 2008.



Example: Truck Prices

Can we use the age of a truck to predict what its price should be?

- Consider a random sample of 43 pickup trucks between 1994 and 2008.

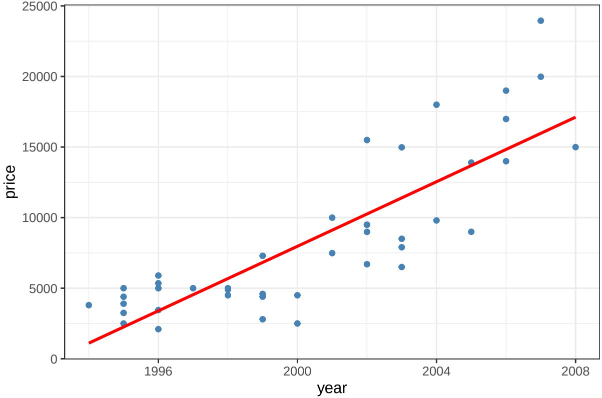


- Let's fit a linear model

Example: Truck Prices

Can we use the age of a truck to predict what its price should be?

- Consider a random sample of 43 pickup trucks between 1994 and 2008.



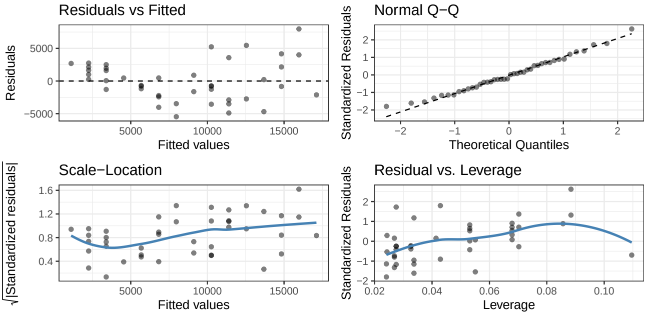
- Let's fit a linear model

Linear Model

```
truck_mod<-lm(price~year, data = pickups)
summary(truck_mod)
```

```
##
## Call:
## lm(formula = price ~ year, data = pickups)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5468.7 -2202.9  -313.6  2096.0  7977.7
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2278766.2   238325.7  -9.562 6.92e-12 ***
## year            1143.4      119.1   9.597 6.24e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3080 on 40 degrees of freedom
## Multiple R-squared:  0.6972, Adjusted R-squared:  0.6896
## F-statistic:  92.1 on 1 and 40 DF,  p-value: 6.238e-12
```

Diagnostics



- Residuals appear normally distributed.
- But data suggests a non-linear relationship
- Two observations appear influential.
- There is evidence of non-constant variance in the residuals.

Transformations

If the diagnostic plots look bad, try to transform variables by applying functions.

Transformations

If the diagnostic plots look bad, try to transform variables by applying functions.

- When the response variable spans multiple orders of magnitude, models are often non-linear and residuals often have non-constant variance.

Transformations

If the diagnostic plots look bad, try to transform variables by applying functions.

- When the response variable spans multiple orders of magnitude, models are often non-linear and residuals often have non-constant variance.
  - These variables often benefit from a log-transformation

$$Y_t = \log(Y)$$

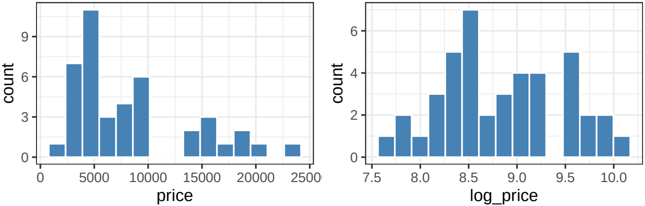
Transformations

If the diagnostic plots look bad, try to transform variables by applying functions.

- When the response variable spans multiple orders of magnitude, models are often non-linear and residuals often have non-constant variance.
  - These variables often benefit from a log-transformation

$$Y_t = \log(Y)$$

```
pickups$log_price <- log(pickups$price)
```



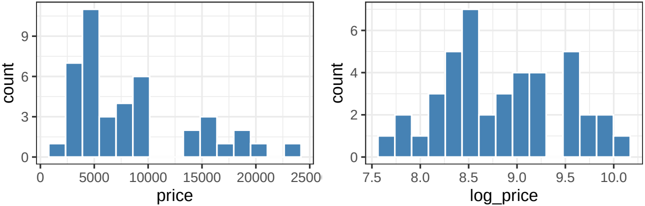
Transformations

If the diagnostic plots look bad, try to transform variables by applying functions.

- When the response variable spans multiple orders of magnitude, models are often non-linear and residuals often have non-constant variance.
  - These variables often benefit from a log-transformation

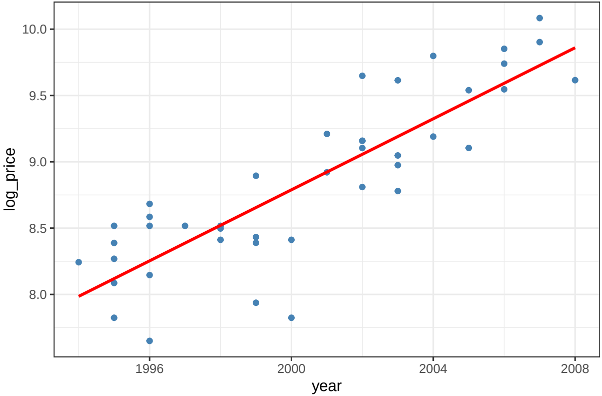
$$Y_t = \log(Y)$$

```
pickups$log_price <- log(pickups$price)
```



- *Note:* It's fine that price is not Normally distributed
  - The linear model only requires that **residuals** be Normally distributed

Log-transformed linear model




```r
truck_log_mod <- lm(log_price ~ year, data = pickups)
summary(truck_log_mod)$coef
```

```
##                 Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) -258.9980504 26.12294226 -9.914582 2.471946e-12
## year           0.1338934  0.01305865 10.253239 9.342855e-13
```

Poll: Interpretation

The slope coefficient in the log-linear model was 0.13. Which of the following interpretations are correct? Select all that apply

1. Increasing year by 1 increases price by approximately 0.13.
2. Increasing year by 1 produces a relative increase in price of approximately $e^{.13}$.
3. Increasing year by 1 increases the log-price by approximately 0.13.
4. Increasing year by $\ln(1)$ increases price by approximately 0.13.

Model Accuracy

The R^2 and RSE values for the log and original models

```
##      model      r.sq         rse
## 1      log 0.7243830    0.337582
## 2 original 0.6972079 3079.839269
```

Model Accuracy

The R^2 and RSE values for the log and original models

```
##      model     r.sq        rse
## 1      log 0.7243830    0.337582
## 2 original 0.6972079 3079.839269
```

- The log model has slight improvement in R^2. And has massive improvement in RSE...

Model Accuracy

The R^2 and RSE values for the log and original models

```
##       model      r.sq         rse
## 1       log 0.7243830    0.337582
## 2 original 0.6972079 3079.839269
```

- The log model has slight improvement in R^2. And has massive improvement in RSE...
  - Or does it? (Recall that RSE depends on the units of Y)

Model Accuracy

The R^2 and RSE values for the log and original models

```
##      model      r.sq         rse
## 1      log 0.7243830    0.337582
## 2 original 0.6972079 3079.839269
```

- The log model has slight improvement in R^2. And has massive improvement in RSE...
  - Or does it? (Recall that RSE depends on the units of Y)
  - We need to transform predicted values from log model back into original scale

Model Accuracy

The R^2 and RSE values for the log and original models

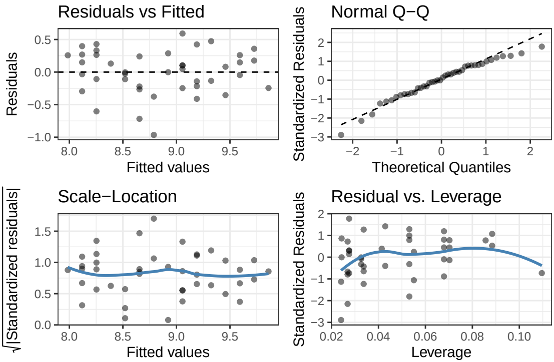
```
##      model      r.sq         rse
## 1      log 0.7243830    0.337582
## 2 original 0.6972079 3079.839269
```

- The log model has slight improvement in R^2. And has massive improvement in RSE. . .
  - Or does it? (Recall that RSE depends on the units of Y)
  - We need to transform predicted values from log model back into original scale

```
pred_price <- exp(truck_log_mod$fitted.values)
RSS <- sum((pickups$price - pred_price)^2)
RSE <- sqrt(RSS/(42-2))
RSE
```

```
## [1] 2841.049
```

Diagnostics



- The residuals from this model appear less normal
- But the quadratic trend is now less apparent.
- There are no influential points
- The variance has been stabilized

Transformations summary

- If a linear model fit to the raw data leads to questionable residual plots, consider transformations.

Transformations summary

- If a linear model fit to the raw data leads to questionable residual plots, consider transformations.
  - Count data and prices often benefit from transformations.

Transformations summary

- If a linear model fit to the raw data leads to questionable residual plots, consider transformations.
  - Count data and prices often benefit from transformations.
  - The natural log and the square root are the most common, but you can use any transformation you like.

Transformations summary

- If a linear model fit to the raw data leads to questionable residual plots, consider transformations.
  - Count data and prices often benefit from transformations.
  - The natural log and the square root are the most common, but you can use any transformation you like.
- Transformations may change model interpretations.

Transformations summary

- If a linear model fit to the raw data leads to questionable residual plots, consider transformations.
  - Count data and prices often benefit from transformations.
  - The natural log and the square root are the most common, but you can use any transformation you like.
- Transformations may change model interpretations.
- Non-constant variance is a serious problem but it can sometimes be solved by transforming the response.

Transformations summary

- If a linear model fit to the raw data leads to questionable residual plots, consider transformations.
    - Count data and prices often benefit from transformations.
    - The natural log and the square root are the most common, but you can use any transformation you like.
- Transformations may change model interpretations.
- Non-constant variance is a serious problem but it can sometimes be solved by transforming the response.
- Transformations can also fix non-linearity

Section 3

Qualitative Predictors

Qualitative Predictors

Thus far, we have assumed all predictors are quantitative, but it would be nice to include qualitative predictors also

Qualitative Predictors

Thus far, we have assumed all predictors are quantitative, but it would be nice to include qualitative predictors also

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

Qualitative Predictors

Thus far, we have assumed all predictors are quantitative, but it would be nice to include qualitative predictors also

- For binary categorical variables, we create a new *quantitative* variable I by coding the first level as 0 and the second as 1.
- Extend to multi-level categorical variables by creating binary variables I for all but one level.

Qualitative Predictors

Thus far, we have assumed all predictors are quantitative, but it would be nice to include qualitative predictors also

- For binary categorical variables, we create a new *quantitative* variable I by coding the first level as 0 and the second as 1.
- Extend to multi-level categorical variables by creating binary variables I for all but one level.
- If X_1 is quantitative and X_2 is quantitative with 3 levels (A,B,C), the resulting model will be

$$\hat{Y} = f(X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 I_B + \beta_3 I_C = \begin{cases} \beta_0 + \beta_1 X_1, & \text{if } X_2 = A, \\ (\beta_0 + \beta_2) + \beta_1 X_1, & \text{if } X_2 = B, \\ (\beta_0 + \beta_3) + \beta_1 X_1, & \text{if } X_2 = C, \end{cases}$$

Qualitative Predictors

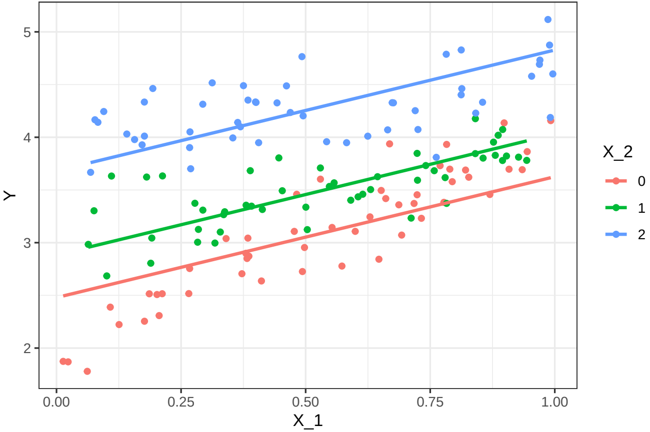
Thus far, we have assumed all predictors are quantitative, but it would be nice to include qualitative predictors also

- For binary categorical variables, we create a new *quantitative* variable I by coding the first level as 0 and the second as 1.
- Extend to multi-level categorical variables by creating binary variables I for all but one level.
- If X_1 is quantitative and X_2 is quantitative with 3 levels (A,B,C), the resulting model will be

$$\hat{Y} = f(X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 I_B + \beta_3 I_C = \begin{cases} \beta_0 + \beta_1 X_1, & \text{if } X_2 = A, \\ (\beta_0 + \beta_2) + \beta_1 X_1, & \text{if } X_2 = B, \\ (\beta_0 + \beta_3) + \beta_1 X_1, & \text{if } X_2 = C, \end{cases}$$

- Note that all 3 regression lines have the same slope, but different intercept.

Scatterplot



$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 I_1 + \hat{\beta}_3 I_2 = 2.48 + 1.14X_1 + 0.40I_1 + 1.20I_2$$

The model in R

```
cat_mod<- lm(data = my_data, Y ~ X_1 + X_2)
summary(cat_mod)
```

```
##
## Call:
## lm(formula = Y ~ X_1 + X_2, data = my_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.77071 -0.19279 -0.00376  0.18634  0.69164
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.47917    0.06238  39.742  < 2e-16 ***
## X_1          1.14670    0.08730  13.135  < 2e-16 ***
## X_21         0.40423    0.05881   6.873 1.69e-10 ***
## X_22         1.20196    0.05883  20.432  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2941 on 146 degrees of freedom
## Multiple R-squared:  0.8022, Adjusted R-squared:  0.7981
## F-statistic: 197.3 on 3 and 146 DF,  p-value: < 2.2e-16
```

Poll 3: MLR Slope Interpretation

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

a. How much we expect the response to change if we increase the value of I from 0 to 1, while holding all other variables in the model constant.
b. The difference in the average response value between observations in the baseline level vs the indicated level, while holding other variables fixed.
c. The expected value of the response variable when $I = 0$
d. The distance between the regression line for the baseline level and the indicator level on the 2d scatterplot

Section 4

Non-linearity

Interaction Effect

- In some cases, the effect of one variable on the response changes depending the values of another variable.

Interaction Effect

- In some cases, the effect of one variable on the response changes depending the values of another variable.
  - i.e. the effect of one variable is amplified in the presence of high levels of another variable

Interaction Effect

- In some cases, the effect of one variable on the response changes depending the values of another variable.
  - i.e. the effect of one variable is amplified in the presence of high levels of another variable
- Consider an investor's annual stock returns.
  - For fixed annual income, investing larger amounts of money will provide larger returns.
  - But the size of return per dollar invested **changes** depending on income. Why?

Interaction Effect

- In some cases, the effect of one variable on the response changes depending the values of another variable.
  - i.e. the effect of one variable is amplified in the presence of high levels of another variable
- Consider an investor's annual stock returns.
  - For fixed annual income, investing larger amounts of money will provide larger returns.
  - But the size of return per dollar invested **changes** depending on income. Why?
- To account for this, we include an **interaction** term in the model:

Interaction Effect

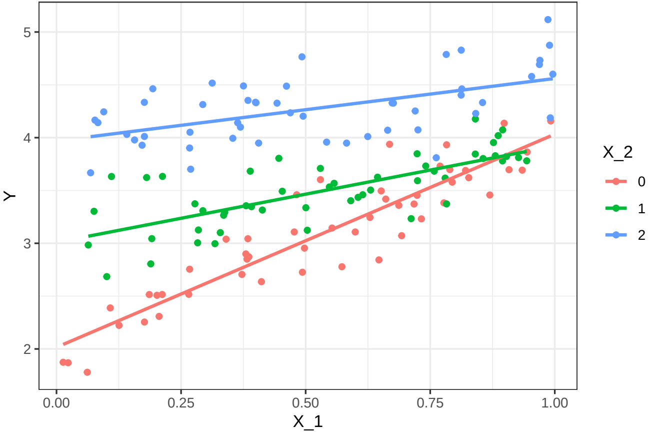
- In some cases, the effect of one variable on the response changes depending the values of another variable.
    - i.e. the effect of one variable is amplified in the presence of high levels of another variable
- Consider an investor's annual stock returns.
    - For fixed annual income, investing larger amounts of money will provide larger returns.
    - But the size of return per dollar invested **changes** depending on income. Why?
- To account for this, we include an **interaction** term in the model:

$$Y = \beta_0 + \beta_1 X_2 + \beta_2 X_2 + \epsilon \qquad \text{Old model}$$

$$Y = \beta_0 + \beta_1 X_2 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon \qquad \text{New model}$$

$$Y = \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon \qquad \tilde{\beta}_1 = \beta_1 + \beta_3 X_2$$

Interaction Terms



$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 I_1 + \hat{\beta}_3 I_2 + \beta_4 X_1 I_1 + \beta_5 X_1 I_2$$
$$= 2.02 + 2.02 X_1 + 0.99 I_1 + 1.95 I_2 - 1.10 X_1 I_1 - 1.43 X_1 I_2$$

The model in R

```
cat_mod<- lm(data = my_data, Y ~ X_1 + X_2 + X_1:X_2)
summary(cat_mod)
```

```
##
## Call:
## lm(formula = Y ~ X_1 + X_2 + X_1:X_2, data = my_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60973 -0.14215 -0.02252  0.14892  0.57340
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.01568    0.07557  26.672  < 2e-16 ***
## X_1          2.01695    0.12661  15.930  < 2e-16 ***
## X_21         0.99310    0.10784   9.209 3.58e-16 ***
## X_22         1.95331    0.10290  18.983  < 2e-16 ***
## X_1:X_21    -1.10462    0.18068  -6.114 8.67e-09 ***
## X_1:X_22    -1.42584    0.17279  -8.252 9.02e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2413 on 144 degrees of freedom
## Multiple R-squared:  0.8686, Adjusted R-squared:  0.8641
## F-statistic: 190.5 on 5 and 144 DF,  p-value: < 2.2e-16
```

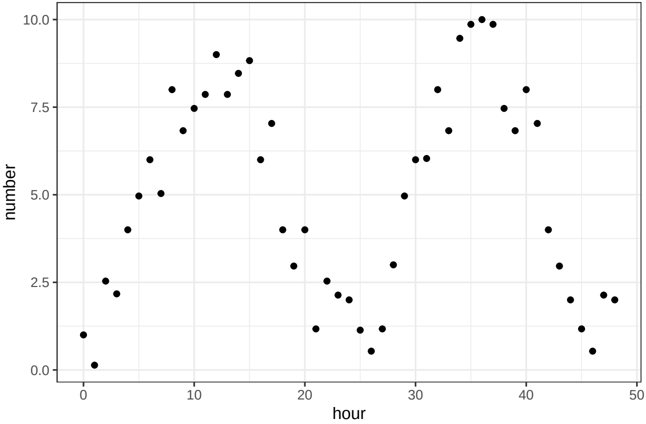
The model in R (Alternative)

```
cat_mod<- lm(data = my_data, Y ~ X_1*X_2)
summary(cat_mod)
```

```
##
## Call:
## lm(formula = Y ~ X_1 * X_2, data = my_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60973 -0.14215 -0.02252  0.14892  0.57340
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.01568    0.07557  26.672  < 2e-16 ***
## X_1          2.01695    0.12661  15.930  < 2e-16 ***
## X_21         0.99310    0.10784   9.209 3.58e-16 ***
## X_22         1.95331    0.10290  18.983  < 2e-16 ***
## X_1:X_21    -1.10462    0.18068  -6.114 8.67e-09 ***
## X_1:X_22    -1.42584    0.17279  -8.252 9.02e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2413 on 144 degrees of freedom
## Multiple R-squared:  0.8686, Adjusted R-squared:  0.8641
## F-statistic: 190.5 on 5 and 144 DF,  p-value: < 2.2e-16
```

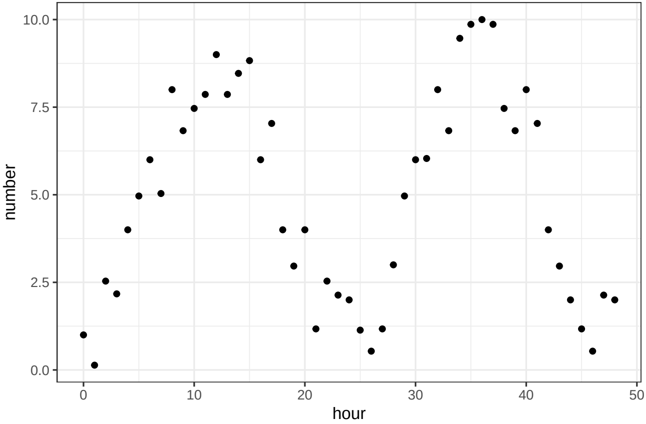
Non-linear models

The `emails` data set shows the `number` of emails recieved in a given `hour` over 2 days



Non-linear models

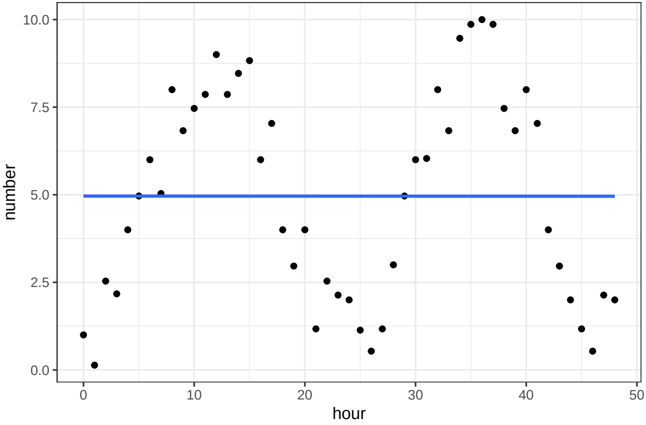
The `emails` data set shows the `number` of emails recieved in a given `hour` over 2 days



- What kind of model will accurately predict email number?

Non-linear models

The `emails` data set shows the `number` of emails recieved in a given `hour` over 2 days



- A linear model performs poorly.

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

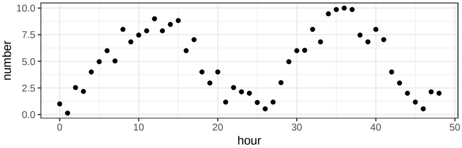
$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot X^2 + \cdots + \beta_p \cdot X^p + \epsilon$$

- 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?

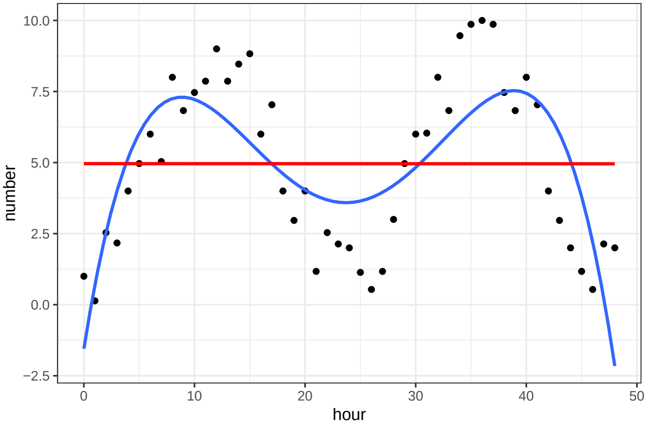
What polynomial degree seems most appropriate for the given data?

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

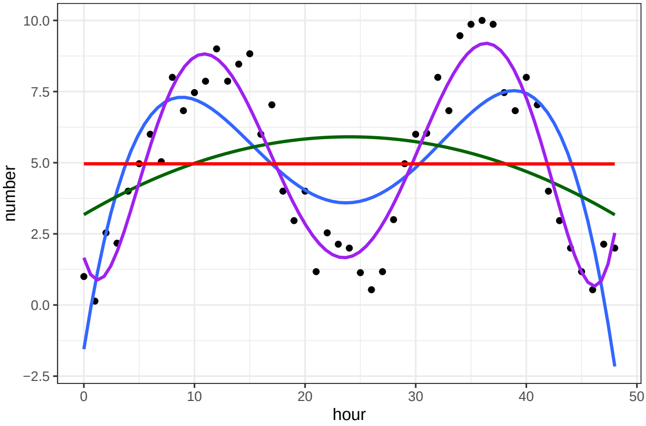


Plotting non-linear regression curves

```
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, raw= T), data = emails)
summary(emails_mod)
```

```
##
## Call:
## lm(formula = number ~ poly(hour, degree = 4, raw = T), data = emails)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2317 -1.4687 -0.0364  1.4185  4.1590
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      -1.551e+00  1.312e+00  -1.183    0.243
## poly(hour, degree = 4, raw = T)1  2.458e+00  3.870e-01   6.352 1.03e-07 ***
## poly(hour, degree = 4, raw = T)2 -2.223e-01  3.328e-02  -6.680 3.37e-08 ***
## poly(hour, degree = 4, raw = T)3  7.177e-03  1.047e-03   6.855 1.86e-08 ***
## poly(hour, degree = 4, raw = T)4 -7.536e-05  1.082e-05  -6.967 1.28e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.065 on 44 degrees of freedom
## Multiple R-squared:  0.5645, Adjusted R-squared:  0.5249
## F-statistic: 14.26 on 4 and 44 DF,  p-value: 1.536e-07
```

Section 5

Variable Selection

Improving Model Accuracy

What do we do when the full model is inaccurate on test data (i.e. high test MSE)?

Improving Model Accuracy

What do we do when the full model is inaccurate on test data (i.e. high test MSE)?

- Remove redundant variables strongly correlated with each other
- Remove spurious variables that aren't linearly related with response

Improving Model Accuracy

What do we do when the full model is inaccurate on test data (i.e. high test MSE)?

- Remove redundant variables strongly correlated with each other
- Remove spurious variables that aren't linearly related with response

Two common methods of identifying extraneous predictors:

1. **Backwards Elimination**:
   - Start with the full model, remove the variable with highest p-value, and refit. Continue to do so until all p-values are small, or accuracy ceases to improve.

Improving Model Accuracy

What do we do when the full model is inaccurate on test data (i.e. high test MSE)?

- Remove redundant variables strongly correlated with each other
- Remove spurious variables that aren't linearly related with response

Two common methods of identifying extraneous predictors:

1. **Backwards Elimination**:
   - Start with the full model, remove the variable with highest p-value, and refit. Continue to do so until all p-values are small, or accuracy ceases to improve.
2. **Forward Selection**:
   - Start with the null model, create p many SLR models (one for each predictor), and select the one with best accuracy. Repeat with this new model, creating $p - 1$ two predictor models (one for each remaining predictor). Continue until accuracy ceases to improve.

Improving Model Accuracy

What do we do when the full model is inaccurate on test data (i.e. high test MSE)?

- Remove redundant variables strongly correlated with each other
- Remove spurious variables that aren't linearly related with response

Two common methods of identifying extraneous predictors:

1. **Backwards Elimination**:
   - Start with the full model, remove the variable with highest p-value, and refit. Continue to do so until all p-values are small, or accuracy ceases to improve.
2. **Forward Selection**:
   - Start with the null model, create p many SLR models (one for each predictor), and select the one with best accuracy. Repeat with this new model, creating $p-1$ two predictor models (one for each remaining predictor). Continue until accuracy ceases to improve.

- Is it possible that none of these models will have the best possible accuracy among all subsets of predictors?

Improving Model Accuracy

What do we do when the full model is inaccurate on test data (i.e. high test MSE)?

- Remove redundant variables strongly correlated with each other
- Remove spurious variables that aren't linearly related with response

Two common methods of identifying extraneous predictors:

1. **Backwards Elimination**:
   - Start with the full model, remove the variable with highest p-value, and refit. Continue to do so until all p-values are small, or accuracy ceases to improve.
2. **Forward Selection**:
   - Start with the null model, create p many SLR models (one for each predictor), and select the one with best accuracy. Repeat with this new model, creating $p - 1$ two predictor models (one for each remaining predictor). Continue until accuracy ceases to improve.

- Is it possible that none of these models will have the best possible accuracy among all subsets of predictors?
  - Yes. But we'll cover detailed model selection in Chapter 6.