MLR: Troubleshooting

Nate Wells

Math 243: Stat Learning

September 21st, 2020

Outline

In today's class, we will. . .

- Troubleshoot potential problems with the linear model

Section 1

Problems with Linear Model

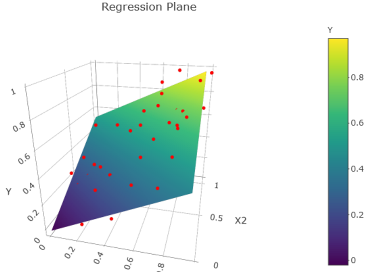
Overview

Given any data set with $n \geq p$, there is **always** a least squares regression equation

Overview

Given any data set with $n \geq p$, there is **always** a least squares regression equation

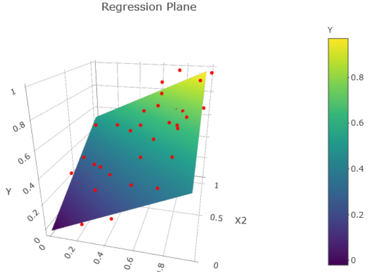
- i.e. a hyperplane in $\mathbb{R}^{p+1}$ that minimizes the squared sum of residuals.



Overview

Given any data set with $n \geq p$, there is **always** a least squares regression equation

- i.e. a hyperplane in $\mathbb{R}^{p+1}$ that minimizes the squared sum of residuals.



However, if we want to make *predictions* or perform *statistical inference* we need to make sure key assumptions of randomness are met.

Common Problems

Most problems fall into 1 of 6 categories:

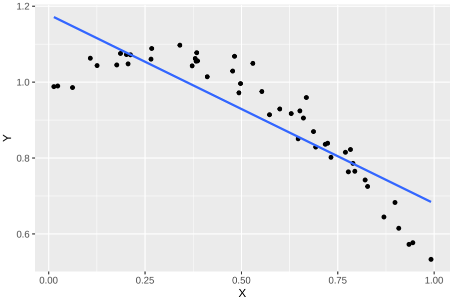
1. Non-linearity of relationship between predictors and response
2. Correlation of error terms
3. Non-constant variance in error
4. Outliers
5. High-leverage points
6. Collinearity of predictors

Non-linearity

In order to fit a linear model, we assume $Y = F(X_1, \dots, X_p) + \epsilon$, where f is linear.

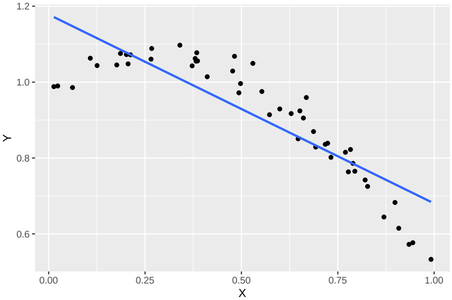
Non-linearity

In order to fit a linear model, we assume $Y = F(X_1, \ldots, X_p) + \epsilon$, where f is linear.



Non-linearity

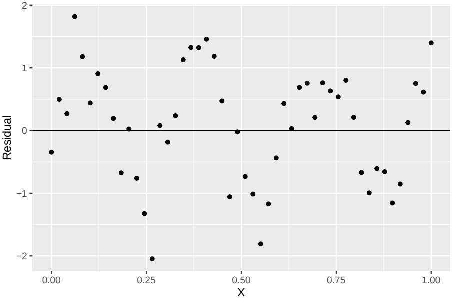
In order to fit a linear model, we assume $Y = F(X_1, \ldots, X_p) + \epsilon$, where f is linear.



But if this assumption is false, our model is likely to have high bias.

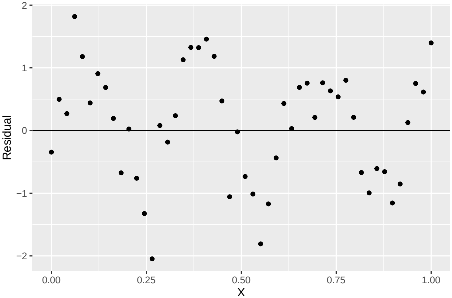
Correlation of Errors

If errors are correlated, then knowing the values of one gives extra information about values of others.



Correlation of Errors

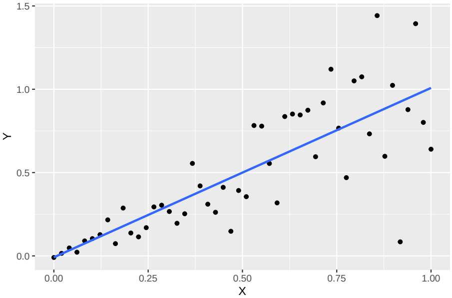
If errors are correlated, then knowing the values of one gives extra information about values of others.



Correlated errors lead to underestimates of residual standard error - Producing narrower confidence intervals and inflating test statistics

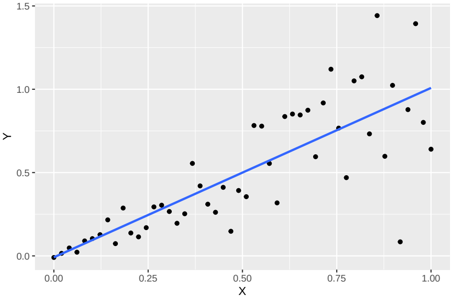
Non-constant variance

For prediction and inference with LM, we assume that all residuals have the same variance.



Non-constant variance

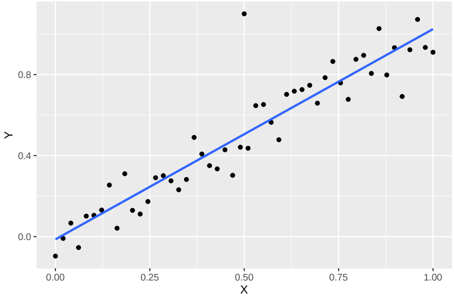
For prediction and inference with LM, we assume that all residuals have the same variance.



Least squares regression does not minimize RSS; requires more data for accurate predictions

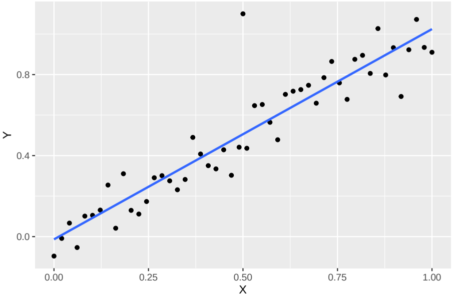
Outliers

While outliers may occur even if model assumptions are met, they do influence accuracy estimates



Outliers

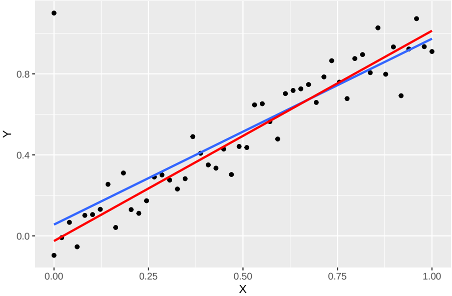
While outliers may occur even if model assumptions are met, they do influence accuracy estimates



Reduce R^2 and increase RSE estimates

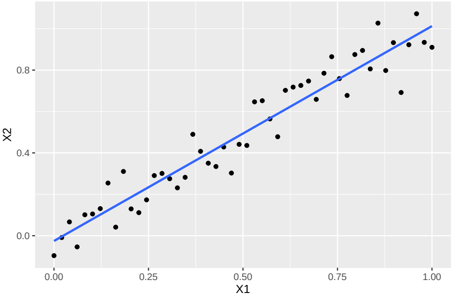
High Leverage points

Outliers which have extreme values of predictors and response are called high-leverage points



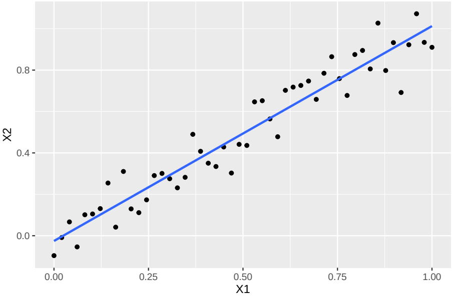
Collinearity

Collinearity occurs when predictors are highly correlated



Collinearity

Collinearity occurs when predictors are highly correlated



Collinearity produces high variance in estimates for β.

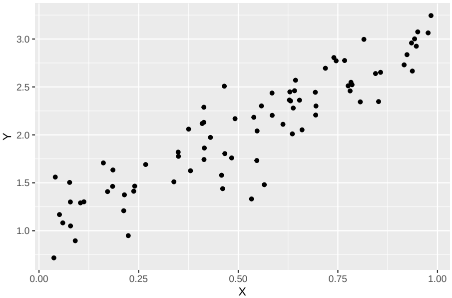
A Valid Model

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)

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

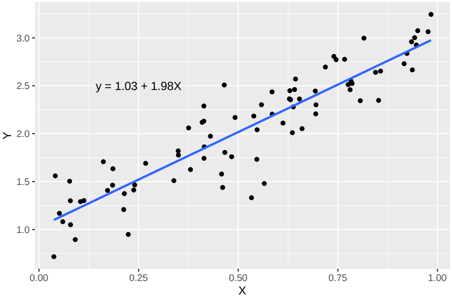


Linear Model

```
my_mod<-lm(Y ~ X, data = my_data)
beta_0 <- summary(my_mod)$coefficients[1]
beta_1 <- summary(my_mod)$coefficients[2]
c(beta_0, beta_1)
```

```
## [1] 1.025947 1.981375
```

```
ggplot(my_data, aes(x = X , y = Y)) + geom_point() + geom_smooth(method = "lm", se = F) +
  annotate(geom= "text", x = .25, y = 2.5, label = "y = 1.03 + 1.98X")
```



Model Diagnostics

Goal: Create graphics to assess how well data fits modeling assumptions.

Model Diagnostics

Goal: Create graphics to assess how well data fits modeling assumptions.

The trade-off:

- We can use the base R `plot` function to quickly create all diagnostic plots necessary

Model Diagnostics

Goal: Create graphics to assess how well data fits modeling assumptions.

The trade-off:

- We can use the base R `plot` function to quickly create all diagnostic plots necessary
  - But we then are restricted to `plot` aesthetics

Model Diagnostics

Goal: Create graphics to assess how well data fits modeling assumptions.

The trade-off:

- We can use the base R `plot` function to quickly create all diagnostic plots necessary
  - But we then are restricted to `plot` aesthetics
- On the other hand, we could create more aesthetically pleasing `ggplots`

Model Diagnostics

Goal: Create graphics to assess how well data fits modeling assumptions.

The trade-off:

- We can use the base R `plot` function to quickly create all diagnostic plots necessary
  - But we then are restricted to `plot` aesthetics
- On the other hand, we could create more aesthetically pleasing `ggplots`
  - At the cost of needing to wrangle data before plotting or use extra packages

Model Diagnostics

Goal: Create graphics to assess how well data fits modeling assumptions.

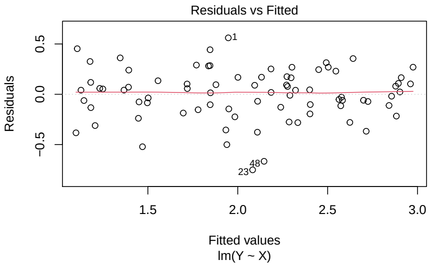
The trade-off:

- We can use the base R `plot` function to quickly create all diagnostic plots necessary
  - But we then are restricted to `plot` aesthetics
- On the other hand, we could create more aesthetically pleasing `ggplots`
  - At the cost of needing to wrangle data before plotting or use extra packages

For simplicity, we'll default to the `plot` function.

Residual Plot

```
plot(my_mod, 1)
```

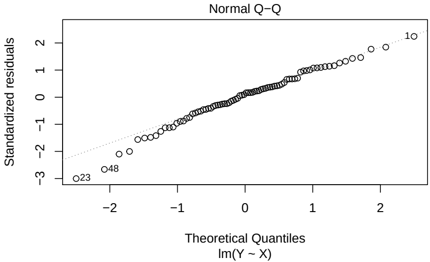


What is represented along the horizontal axis? Why?

What should we look for?

QQ Plot (Don't cry)

```
plot(my_mod, 2)
```

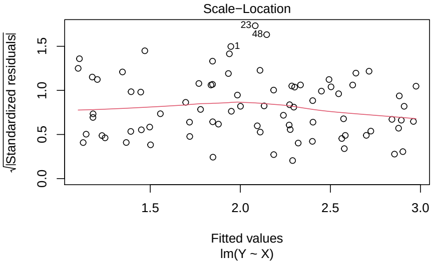


What is represented along the horizontal and vertical axes? Why?

What should we look for?

Scale-Location Plot

```r
plot(my_mod, 3)
```

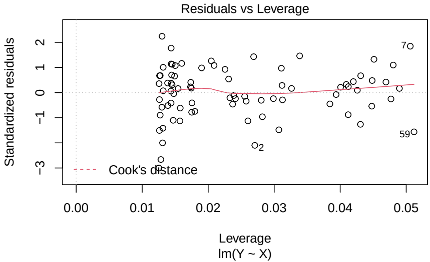


What is represented along the vertical axes? Why?

What should we look for?

Leverage Plot

```
plot(my_mod, 5)
```

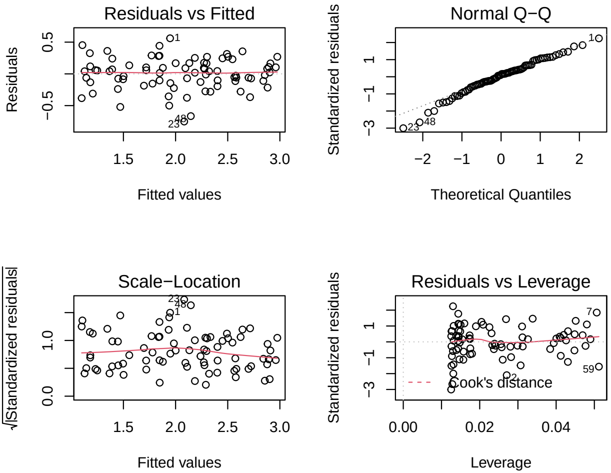


What is represented along the horizontal and vertical axes? Why?

What should we look for?

Plot Quartet

```
par(mfrow = c(2,2))
plot(my_mod)
```



Now Let's Break Things!