Linear Models Diagnostics

Prof Wells

STA 295: Stat Learning

February 6th, 2024

Outline

In today's class, we will...

- Discuss theoretical foundation for linear regression
- Perform inference for simple linear models
- Implement simple linear regression in R

Section 1

Problems with Linear Model

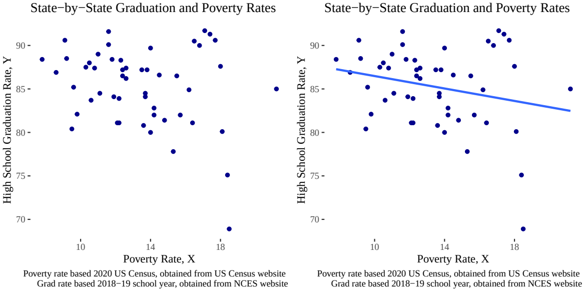
Overview

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

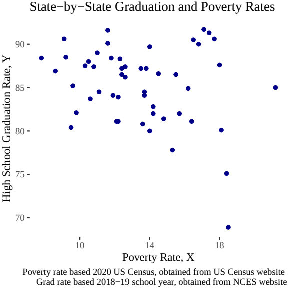
- i.e. a line that minimizes the squared sum of residuals.

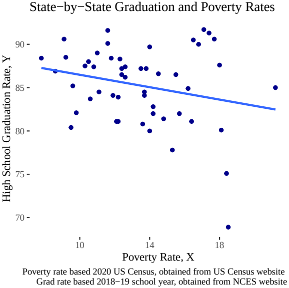


Overview

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

- i.e. a line 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

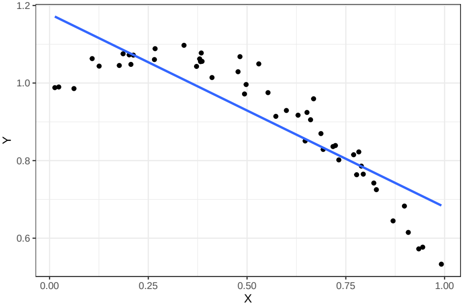
In order to fit a linear model, we assume that in $Y = F(X) + \epsilon$, we have

$$f(x) = \beta_0 + \beta_1 \cdot x$$

Non-linearity

In order to fit a linear model, we assume that in $Y = F(X) + \epsilon$, we have

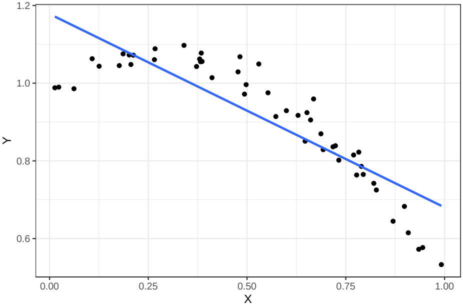
$$f(x) = \beta_0 + \beta_1 \cdot x$$



Non-linearity

In order to fit a linear model, we assume that in $Y = F(X) + \epsilon$, we have

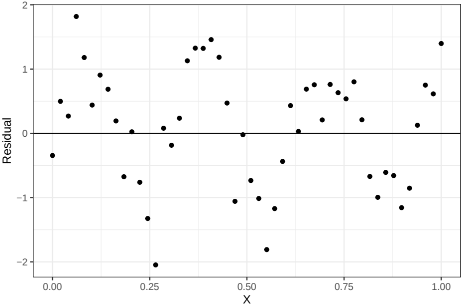
$$f(x) = \beta_0 + \beta_1 \cdot x$$



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

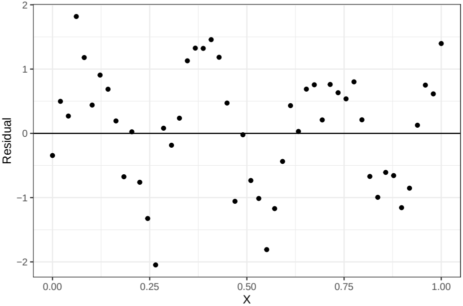
Correlation of Errors

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.

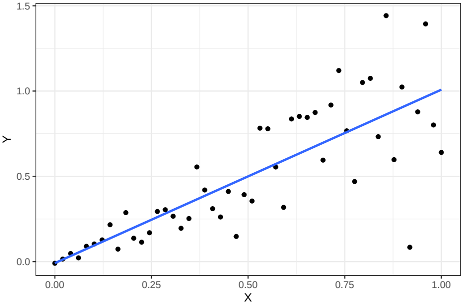


Correlated errors lead to underestimates of residual standard error

- This produces incorrectly narrow confidence intervals, as well incorrectly small p-values
- It also leads to models with higher variance

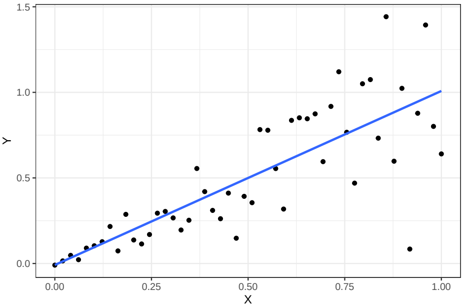
Non-constant variance

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.



Estimates for regression coefficients (β_0, β_1) are still *unbiased*; However, estimated standard errors SE are incorrect

- Confidence intervals and hypothesis tests should not be trusted
- There are other estimates for β_0 and β_1 that are still unbiased, but have lower variance (and hence, have lower test MSE)

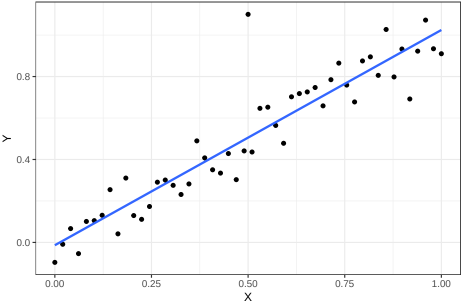
Outliers

Outliers are points which are extreme in either predictor or response values (or both)

Outliers

Outliers are points which are extreme in either predictor or response values (or both)

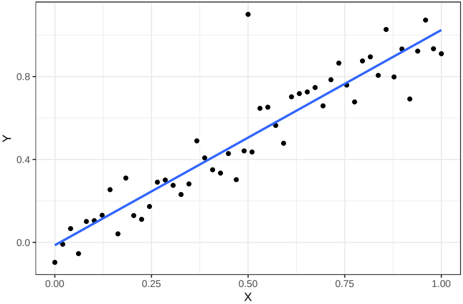
- They may occur even if model assumptions are met, but still influence accuracy estimates



Outliers

Outliers are points which are extreme in either predictor or response values (or both)

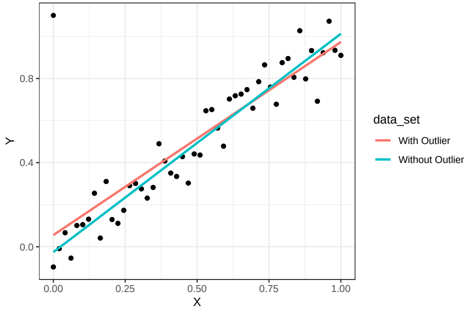
- They may occur even if model assumptions are met, but still influence accuracy estimates



- Presence of outliers decrease estimated R^2 and RSE compared to similar data set without outliers.

High Leverage points

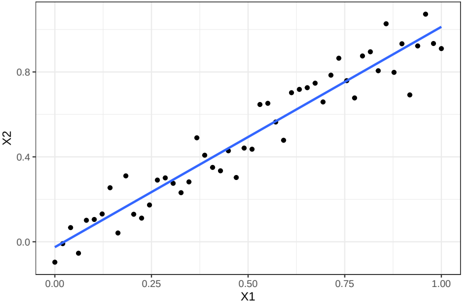
Outliers which have extreme values for **both** predictors and response are called high-leverage points



- Outliers can cause noticable changes in the parameter estimates, and can lead to less accurate models

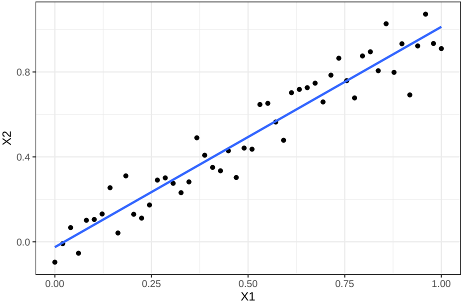
Collinearity

Collinearity occurs when predictors are highly correlated



Collinearity

Collinearity occurs when predictors are highly correlated



Collinearity produces high variance in estimates for β.

- We'll talk more about this next week.

Section 2

Diagnostic Plots

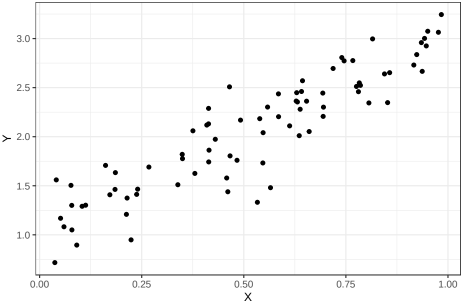
A Valid Model

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()
```



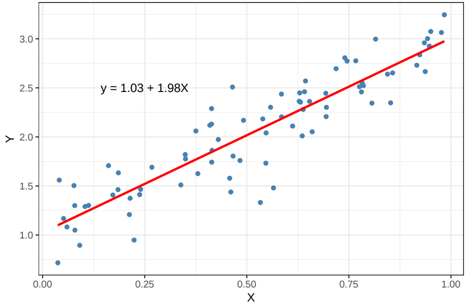
Linear Model

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

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

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

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



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:

- The base R `plot` function can be used to quickly create all diagnostic plots necessary

Model Diagnostics

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

The trade-off:

- The base R `plot` function can be used 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:

- The base R `plot` function can be used to quickly create all diagnostic plots necessary
  - But we then are restricted to `plot` aesthetics
- Alternatively, we can use the `gglm` function in the package of the same name, created and maintained by Reed alum, Grayson White.

Model Diagnostics

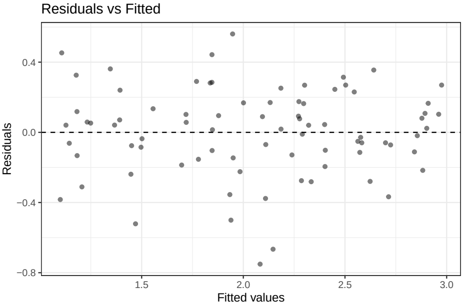
Goal: Create graphics to assess how well data fits modeling assumptions.

The trade-off:

- The base R `plot` function can be used to quickly create all diagnostic plots necessary
  - But we then are restricted to `plot` aesthetics
- Alternatively, we can use the `gglm` function in the package of the same name, created and maintained by Reed alum, Grayson White.
  - Provides the same diagnostic plots as `plot`, but with `ggplot2` appearances and customization.

Residual Plot

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

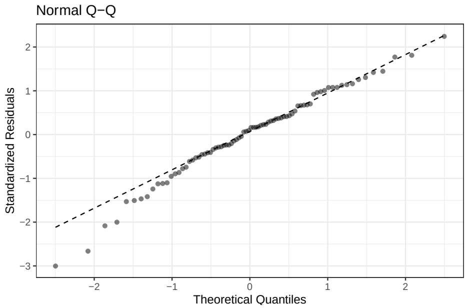


What is represented along the horizontal axis? Why?

What should we look for?

QQ Plot

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

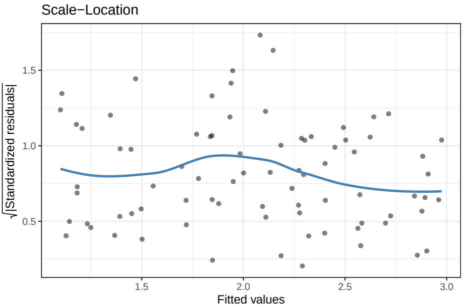


What is represented along the horizontal and vertical axes? Why?

What should we look for?

Scale-Location Plot

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

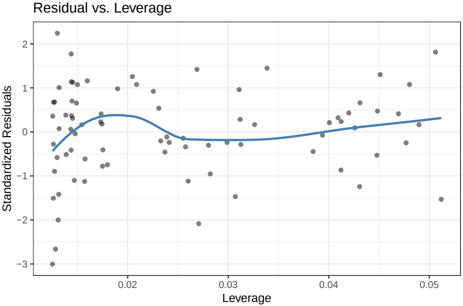


What is represented along the vertical axes? Why?

What should we look for?

Leverage Plot

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



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

What should we look for?

Plot Quartet

```
library(gglm)
gglm(my_mod)
```