Multiple Linear Regression

Nate Wells

Math 243: Stat Learning

September 20th, 2021

Outline

In today's class, we will...

- Quantify model accuracy for linear regression models (both simple and multiple)
- Troubleshoot potential problems with the linear model

Section 1

Assessing Model Accuracy

How Strong is a Linear Model?

- In an linear model model,
$$Y = f(X) + \epsilon$$
So even if we could perfectly predict f using $\hat{f}$, our model would still have non-zero MSE.

How Strong is a Linear Model?

- In an linear model model,
$$Y = f(X) + \epsilon$$
So even if we could perfectly predict f using $\hat{f}$, our model would still have non-zero MSE.

- The **Residual Standard Error** (RSE) measures the average size of deviations of the response from the linear regression line. It is given by

$$\text{RSE} = \sqrt{\frac{1}{n-1-p}\text{RSS}} = \sqrt{\frac{1}{n-1-p}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

How Strong is a Linear Model?

- In an linear model model,

$$Y = f(X) + \epsilon$$

So even if we could perfectly predict f using $\hat{f}$, our model would still have non-zero MSE.

- The **Residual Standard Error** (RSE) measures the average size of deviations of the response from the linear regression line. It is given by

$$\mathrm{RSE} = \sqrt{\frac{1}{n-1-p}\mathrm{RSS}} = \sqrt{\frac{1}{n-1-p}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- It has the property that

$$E(\mathrm{RSE}^2) = \mathrm{Var}(\epsilon)$$

- Which means that $E(\mathrm{RSE}) \approx \mathrm{sd}(\epsilon)$

Five Flavors of Error

Which of the following are most likely to decrease as more and more predictors are added to a linear model (select all that apply)?

a. test MSE
b. training MSE
c. RSS
d. RSE
e. $\mathrm{Var}(\epsilon)$

The R^2 statistic

Large RSE indicates poor model fit, while small RSE indicates good fit. But how do we determine how small is **small**?

The R^2 statistic

Large RSE indicates poor model fit, while small RSE indicates good fit. But how do we determine how small is **small**?

- The answer depends on the units of Y

The R^2 statistic

Large RSE indicates poor model fit, while small RSE indicates good fit. But how do we determine how small is **small**?

- The answer depends on the units of Y

An alternative, standardized measure of goodness of fit is the R^2 statistic:

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \qquad \text{where TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

The R^2 statistic

Large RSE indicates poor model fit, while small RSE indicates good fit. But how do we determine how small is **small**?

- The answer depends on the units of Y

An alternative, standardized measure of goodness of fit is the R^2 statistic:

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \qquad \text{where } \mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

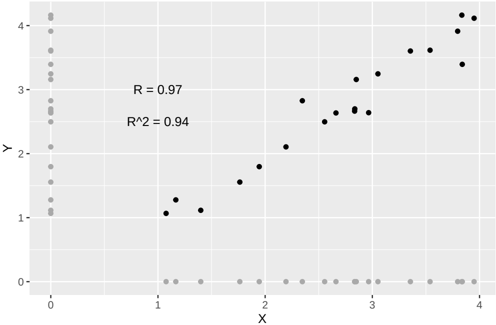
- The value of R^2 is always between 0 and 1, and represents the percentage of variability in values of the response just due to variability in the predictors.

Values of R^2

If $R^2 \approx 1$: nearly all the variability in response is due to variability in the predictor variable.

Values of R^2

If $R^2 \approx 1$: nearly all the variability in response is due to variability in the predictor variable.

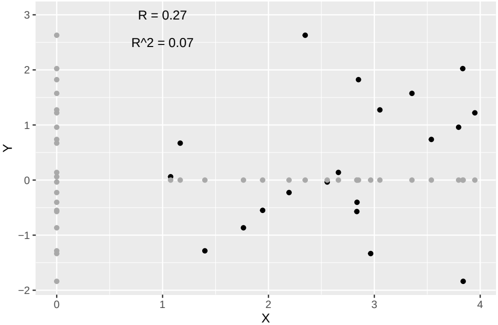


Values of R^2

If $R^2 \approx 0$: almost none of the variability in response is due to variability in the predictor variable.

Values of R^2

If $R^2 \approx 0$: almost none of the variability in response is due to variability in the predictor variable.



Formulas for R^2 in terms of correlation

For SLR,

$$R^2 = [\mathrm{Cor}(X, Y)]^2 = \left[\frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}\right]^2 = \left[\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}\right]^2$$

Formulas for R^2 in terms of correlation

For SLR,

$$R^2 = [\mathrm{Cor}(X, Y)]^2 = \left[\frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}\right]^2 = \left[\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}\right]^2$$

For MLR,

$$R^2 = \left[\mathrm{Cor}(Y, \hat{Y})\right]^2$$

Formulas for R^2 in terms of correlation

For SLR,

$$R^2 = [\mathrm{Cor}(X, Y)]^2 = \left[\frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}\right]^2 = \left[\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}\right]^2$$

For MLR,

$$R^2 = \left[\mathrm{Cor}(Y, \hat{Y})\right]^2$$

We will usually use software to compute R^2.

Model Accuracy in R

```
mod_credit<-lm(Balance ~ Income + Limit , data = Credit)

summary(mod_credit)
```

```
##
## Call:
## lm(formula = Balance ~ Income + Limit, data = Credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -232.79 -115.45  -48.20   53.36  549.77
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept) -385.17926   19.46480  -19.79   <2e-16 ***
## Income        -7.66332    0.38507  -19.90   <2e-16 ***
## Limit          0.26432    0.00588   44.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 165.5 on 397 degrees of freedom
## Multiple R-squared:  0.8711, Adjusted R-squared:  0.8705
## F-statistic:  1342 on 2 and 397 DF,  p-value: < 2.2e-16
```

Model Accuracy in R

```
mod_credit<-lm(Balance ~ Income + Limit , data = Credit)

summary(mod_credit)
```

```
##
## Call:
## lm(formula = Balance ~ Income + Limit, data = Credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -232.79 -115.45  -48.20   53.36  549.77
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -385.17926   19.46480  -19.79   <2e-16 ***
## Income        -7.66332    0.38507  -19.90   <2e-16 ***
## Limit          0.26432    0.00588   44.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 165.5 on 397 degrees of freedom
## Multiple R-squared:  0.8711, Adjusted R-squared:  0.8705
## F-statistic:  1342 on 2 and 397 DF,  p-value: < 2.2e-16
```

We can use `summary(mod)$r.sq` or `summary(mod)$sigma` to access R^2 and RSE directly.

Adjusted R^2

- It turns out that the samples's R^2 gives a **biased** estimate of the variability in the *population* explained by the model.

Adjusted R^2

- It turns out that the samples's R^2 gives a **biased** estimate of the variability in the *population* explained by the model.
- Instead, we use the adjusted R:

$$R^2_{\text{adjusted}} = 1 - \frac{\text{RSS}}{\text{TSS}} \frac{n-1}{n-p-1}$$

Adjusted R^2

- It turns out that the samples's R^2 gives a **biased** estimate of the variability in the *population* explained by the model.
- Instead, we use the adjusted R:

$$R^2_{\text{adjusted}} = 1 - \frac{\text{RSS}}{\text{TSS}} \frac{n-1}{n-p-1}$$

- This adjusted R^2 is usually a bit smaller than R^2, and the difference decreases as n gets large.

Testing Significance

Suppose we wish to test whether at least one predictor has a significant linear relationship with the response.

Testing Significance

Suppose we wish to test whether at least one predictor has a significant linear relationship with the response.

Why would it be incorrect to conduct p many significant tests comparing each predictor to the response?

The Hypothesis Test

Goal: test whether any predictors are significant.

The Hypothesis Test

Goal: test whether any predictors are significant.

Hypotheses:

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \qquad H_a : \text{ at least one of } \beta_i \neq 0$$

The Hypothesis Test

Goal: test whether any predictors are significant.

Hypotheses:

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \qquad H_a : \text{ at least one of } \beta_i \neq 0$$

Test statistic:

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n - p - 1)}$$

The Hypothesis Test

Goal: test whether any predictors are significant.

Hypotheses:

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \qquad H_a : \text{ at least one of } \beta_i \neq 0$$

Test statistic:

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)}$$

Under the null hypothesis, F is approximately F-distributed with $p, n-p-1$ parameters.

The Hypothesis Test

Goal: test whether any predictors are significant.

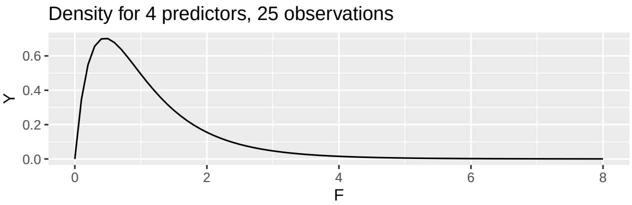
Hypotheses:

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \qquad H_a : \text{ at least one of } \beta_i \neq 0$$

Test statistic:

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)}$$

Under the null hypothesis, F is approximately F-distributed with $p, n-p-1$ parameters.



Typical Values of the F statistic

Provided conditions for linear regression are met,

$$E\left[\frac{\mathrm{RSS}}{n-p-1}\right] = \sigma^2 = \mathrm{Var}(\epsilon)$$

Typical Values of the F statistic

Provided conditions for linear regression are met,

$$E\left[\frac{\mathrm{RSS}}{n-p-1}\right] = \sigma^2 = \mathrm{Var}(\epsilon)$$

And if H_0 is also true, then

$$E\left[\frac{\mathrm{TSS}-\mathrm{RSS}}{p}\right] = \sigma^2 = \mathrm{Var}(\epsilon)$$

Typical Values of the F statistic

Provided conditions for linear regression are met,

$$E\left[\frac{\mathrm{RSS}}{n-p-1}\right] = \sigma^2 = \mathrm{Var}(\epsilon)$$

And if H_0 is also true, then

$$E\left[\frac{\mathrm{TSS}-\mathrm{RSS}}{p}\right] = \sigma^2 = \mathrm{Var}(\epsilon)$$

Hence, if there is truly no relationship between any of the predictors and the response, then on average,

$$F = \frac{(\mathrm{TSS}-\mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)} = 1$$

Typical Values of the F statistic

Provided conditions for linear regression are met,

$$E\left[\frac{\mathrm{RSS}}{n-p-1}\right] = \sigma^2 = \mathrm{Var}(\epsilon)$$

And if H_0 is also true, then

$$E\left[\frac{\mathrm{TSS}-\mathrm{RSS}}{p}\right] = \sigma^2 = \mathrm{Var}(\epsilon)$$

Hence, if there is truly no relationship between any of the predictors and the response, then on average,

$$F = \frac{(\mathrm{TSS}-\mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)} = 1$$

Moreover, it is unlikely that F is drastically larger than 1.

Poll 2: TSS and RSS

Suppose we have a linear model with 25 observations and 4 predictors. Which of the following provides the best evidence of a relationship between the response and at least 1 of the predictors?

a. $\text{TSS} = 64$, $\text{RSS} = 4$
b. $\text{TSS} = 4$, $\text{RSS} = 16$
c. $\text{TSS} = 48$, $\text{RSS} = 8$
d. $\text{TSS} = 4$, $\text{RSS} = 4$

Improving Model Accuracy

What do we do when model accuracy is low (either high RSE or low R^2)?

Improving Model Accuracy

What do we do when model accuracy is low (either high RSE or low R^2)?

- If some variables are strongly correlated, remove some redundant ones.

Improving Model Accuracy

What do we do when model accuracy is low (either high RSE or low R^2)?

- If some variables are strongly correlated, remove some redundant ones.
  - This process is known as *backwards elimination*.
  - Start with the full model, remove the variable with highest p-value, and refit. Continue to do so until accuracy ceases to improve.

Improving Model Accuracy

What do we do when model accuracy is low (either high RSE or low R^2)?

- If some variables are strongly correlated, remove some redundant ones.
    - This process is known as *backwards elimination*.
    - Start with the full model, remove the variable with highest p-value, and refit. Continue to do so until accuracy ceases to improve.
- If ϵ is too large, add further variables.

Improving Model Accuracy

What do we do when model accuracy is low (either high RSE or low R^2)?

- If some variables are strongly correlated, remove some redundant ones.
  - This process is known as *backwards elimination*.
  - Start with the full model, remove the variable with highest p-value, and refit. Continue to do so until accuracy ceases to improve.
- If ϵ is too large, add further variables.
  - This process is known as *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 model accuracy is low (either high RSE or low R^2)?

- If some variables are strongly correlated, remove some redundant ones.
  - This process is known as *backwards elimination*.
  - Start with the full model, remove the variable with highest p-value, and refit. Continue to do so until accuracy ceases to improve.
- If ϵ is too large, add further variables.
  - This process is known as *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 model accuracy is low (either high RSE or low R^2)?

- If some variables are strongly correlated, remove some redundant ones.
  - This process is known as *backwards elimination*.
  - Start with the full model, remove the variable with highest p-value, and refit. Continue to do so until accuracy ceases to improve.
- If ϵ is too large, add further variables.
  - This process is known as *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.

Section 2

Problems with Linear Model

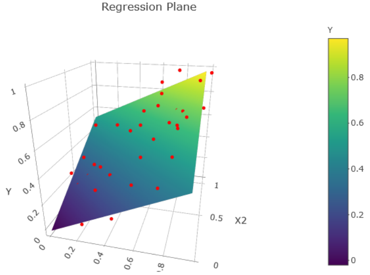
Overview

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

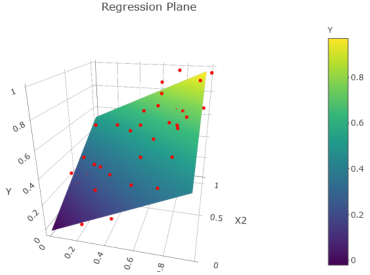
- i.e. a hyperplane in $\mathbb{R}^{p+1}$ that minimizes the squared sum of residuals.



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:

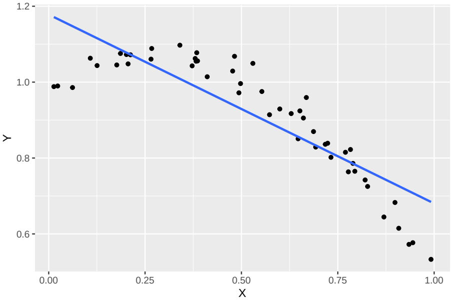
1. Non-linearity of relationship between predictors and response
2. Correlation of error terms
3. Non-constant variance in error
4. Outliers
5. High-leverage points
6. Collinearity of predictors

Non-linearity

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

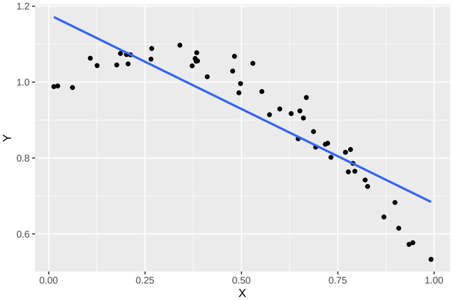
Non-linearity

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



Non-linearity

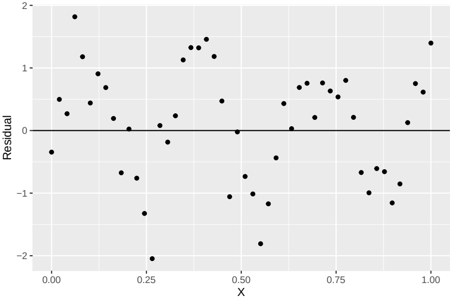
In order to fit a linear model, we assume $Y = F(X_1, \ldots, X_p) + \epsilon$, where f is linear.



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

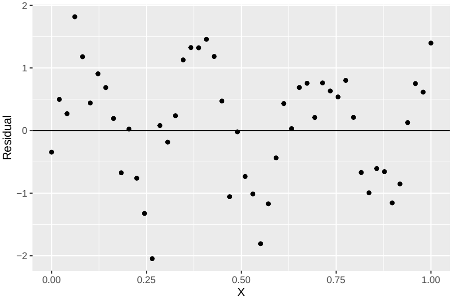
Correlation of Errors

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



Correlation of Errors

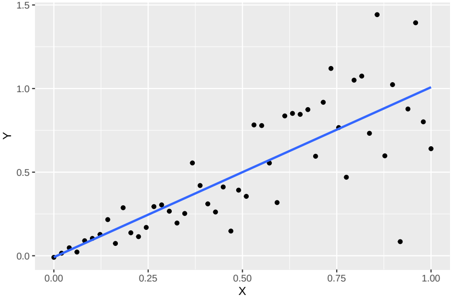
If errors are correlated, then knowing the values of one gives extra information about values of others.



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

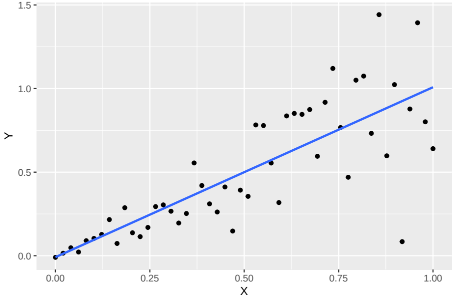
Non-constant variance

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



Non-constant variance

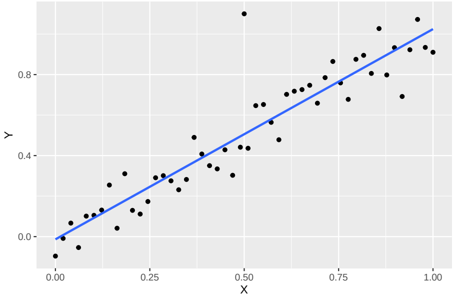
For prediction and inference with LM, we assume that all residuals have the same variance.



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

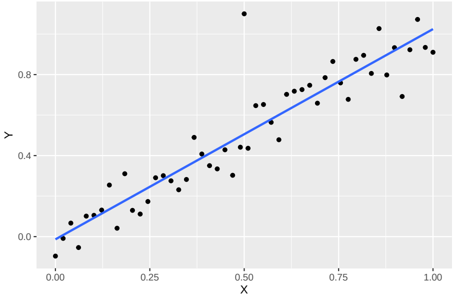
Outliers

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



Outliers

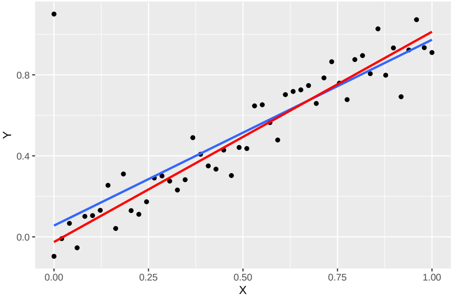
While outliers may occur even if model assumptions are met, they do influence accuracy estimates



Reduce R^2 and increase RSE estimates

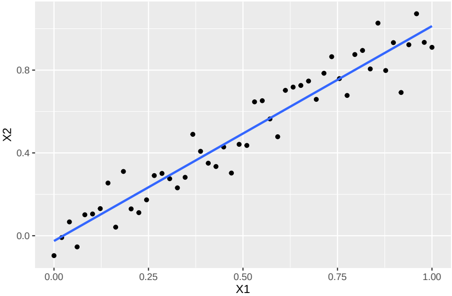
High Leverage points

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



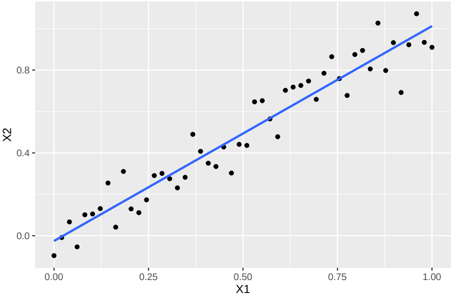
Collinearity

Collinearity occurs when predictors are highly correlated



Collinearity

Collinearity occurs when predictors are highly correlated



Collinearity produces high variance in estimates for β.

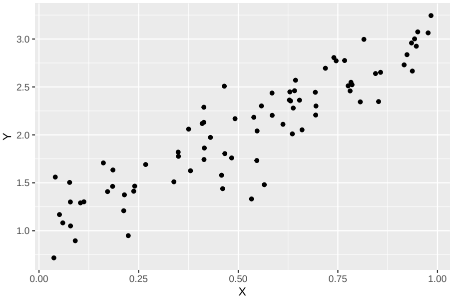
A Valid Model

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()
```

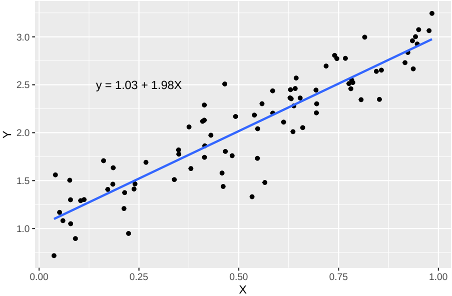


Linear Model

```
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:

- 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 created and maintained by Reed Alum, Grayson White.

Model Diagnostics

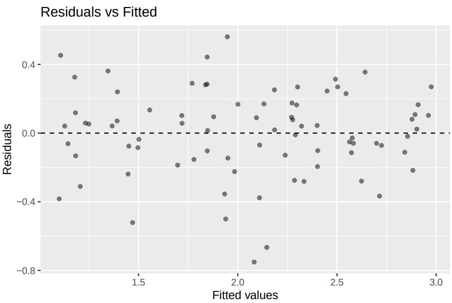
Goal: Create graphics to assess how well data fits modeling assumptions.

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 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()
```

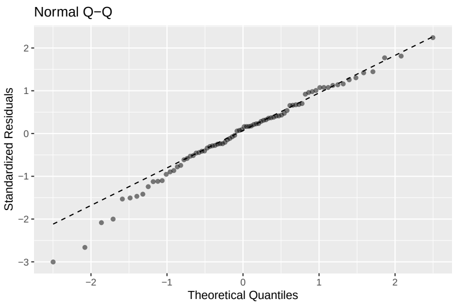


What is represented along the horizontal axis? Why?

What should we look for?

QQ Plot

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

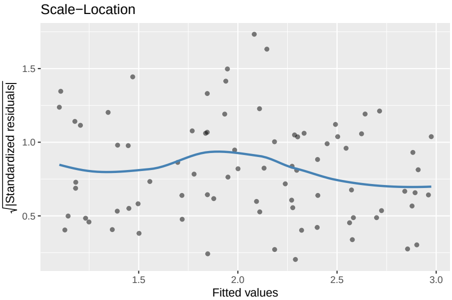


What is represented along the horizontal and vertical axes? Why?

What should we look for?

Scale-Location Plot

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

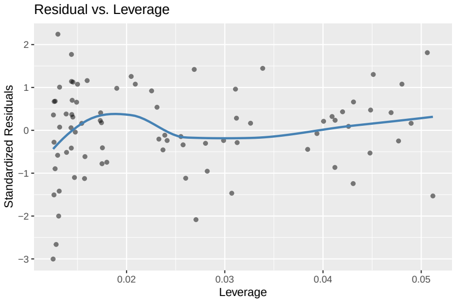


What is represented along the vertical axes? Why?

What should we look for?

Leverage Plot

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



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

What should we look for?

Plot Quartet

```
gglm(my_mod)
```