STA302 METHODS OF DATA ANALYSIS 1

MODULE 4: MITIGATING VIOLATED ASSUMPTIONS

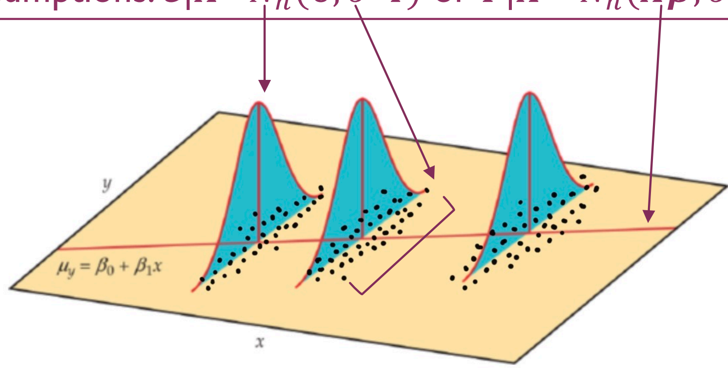
PROF. KATHERINE DAIGNAULT

MODULE 4 OUTLINE

1. Variance Stabilizing Transformations and Other Common Functions
2. Box-Cox Transformations for Normality
3. Application and Interpretation of Transformed Models
4. Impact of Violations on Sampling Distributions

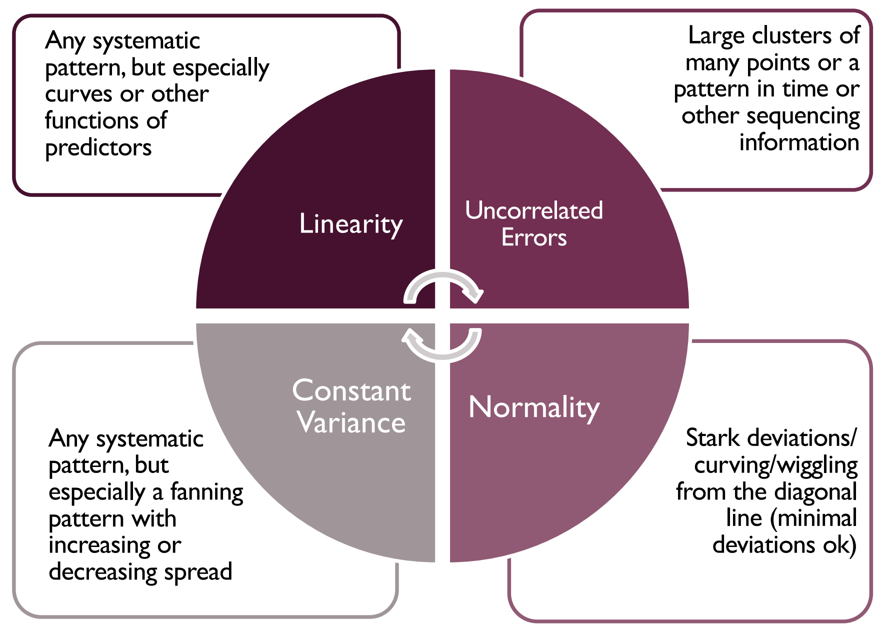
REVIEW OF LINEAR REGRESSION ASSUMPTIONS

Assumptions: $\boldsymbol{\varepsilon}|\boldsymbol{X} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ or $\boldsymbol{Y}|\boldsymbol{X} \sim N_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$



- States the true relationship and structure of information in population
- Implicitly made every time a model is fit to data
- Use residual plots to identify violations

Random bands of residuals indicates no violations



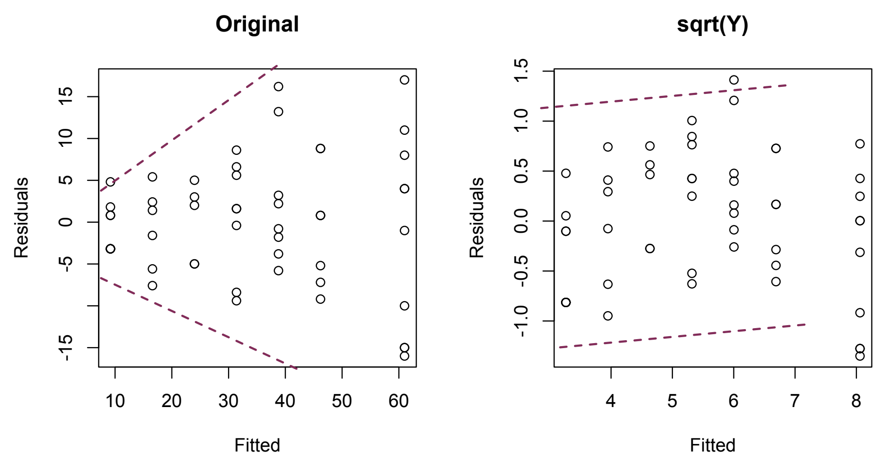
COMMON TRANSFORMATIONS BASED ON DATA EXPLORATION

- Assumptions are formally checked using residual plots but knowing the data can also help
- Always conduct an exploratory data analysis before fitting a model
  - Skew in response variable → probably will have an issue with Normality or Linearity
  - Skews in predictor variables → may see an issue with Linearity
  - Think about underlying characteristic → may help deciding how to include predictor
- Existing literature informs your knowledge about true relationship

- Way to correct or improve violated assumptions is via a transformation on relevant variables
  - E.g., instead of using X we use $X^* = f(X)$
- Transformations can be made to any numerical variable
- Common transformations include:
  - Natural logarithm (ln) or square root will improve most right skews
  - Squares, cube roots, or other logarithms can improve left skews
- These will feature in our analytical methods of selecting transformations.

VARIANCE STABILIZING TRANSFORMATION: WHAT IS IT?

- It's best practice to select transformations that target the violation you observe
- Variance Stabilizing Transformations specifically target the violation of constant variance.
- Transformation would be applied only to the response
- Choice depends on exact situation and data
  - Identifiable data type (e.g. count data) or specific variable context
  - Experience from similar contexts or data types
  - No one transformation will work for every model



VARIANCE STABILIZING TRANSFORMATION: HOW IT WORKS

- Works by exploiting the connection between the mean and the variance of the response
  - Non-constant variance implies both $E(Y|X)$ and $Var(Y|X)$ change with the same X.
- Consider the first-order Taylor series expansion of $f(Y)$ around the mean

  $$f(Y) = f\big(E(Y)\big) + f'\big(E(Y)\big)\big(Y - E(Y)\big) + \cdots$$

- See what happens if we look at the variance of $f(Y)$, i.e. what happens if variance stabilizing transformation applied

- Take variance of both sides:

  $$Var\big(f(Y)\big) = \cancel{Var\Big(f\big(E(Y)\big)\Big)} + Var\big[\underbrace{f'\big(E(Y)\big)}\big(Y - \cancel{E(Y)}\big)\big]$$

  - $f(E(Y))$ is fixed so variance is 0
  - $f'(E(Y))$ also constant so gets pulled out of variance and squared
  - $Var\big(Y - E(Y)\big) = Var(Y)$ as constants added inside variance disappear
- So we get $Var\big(f(Y)\big) = \boxed{\big[f'\big(E(Y)\big)\big]^2} Var(Y)$
- The highlighted term is what undoes the non-constant variance of Y

VARIANCE STABILIZING TRANSFORMATION: EXAMPLE

- Let's show how a specific transformation, applied to Y, could remove the violation of constant variance.
- Suppose $Y \sim Poi(\lambda)$ where $E(Y|X) = Var(Y|X) = \lambda$
  - Yes this does violate Normality, but we'll ignore this for now
- What if we chose to apply a square root transformation (i.e. $f(Y) = \sqrt{Y}$)?
- Using the expression

$$Var\big(f(Y)\big) = \left[f'\big(E(Y)\big)\right]^2 Var(Y)$$

we can check that this function gives us a constant variance for $f(Y)$

- We need the derivative of the transformation:

$$f(Y) = y^{1/2} \Rightarrow \frac{d}{dy} y^{1/2} = \frac{1}{2} y^{-1/2}$$

- Then use $Var(Y|X) = \lambda$ and plug in to get

$$Var\left(y^{1/2}\right) = \left[\frac{1}{2}\lambda^{-1/2}\right]^2 \lambda = \frac{1}{4}\lambda^{-1}\lambda = \frac{1}{4}$$

- We get that the variance of $f(Y)$ is constant so using this as a response should satisfy this assumption.
- Of course, we don't always know what this function is
  - But often common transformations do a decent job.

MODULE 4 OUTLINE

1. Variance Stabilizing Transformations and Other Common Functions
2. Box-Cox Transformations for Normality
3. Application and Interpretation of Transformed Models
4. Impact of Violations on Sampling Distributions

BOX-COX TRANSFORMATION FOR NORMALITY/LINEARITY

- Variance stabilizing transformations are used to specifically improve non-constant variance violations
- To try to improve Normality (and often) also linearity violations, a Box-Cox transformation can be used
  - Although we will not use it specifically as intended due to its complexity
- These transformations are power transformations, taking variables to suggested powers
- Can be used on the response, on the predictor(s), or on response and predictors simultaneously.

Maximum Likelihood Estimation

- Box-Cox method uses Maximum Likelihood to estimate the power transformation
  - This estimated power gives the best approximation to Normality of all possible powers from -5 to 5
- Log-likelihood in our simple linear regression

$$\log(L(\beta_0, \beta_1, \sigma^2 | Y))$$
$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2)$$
$$-\frac{1}{2\sigma^2}\log(\underbrace{\boxed{\sum (y_i - \beta_0 - \beta_1 x_i)^2}}_{\text{RSS}})$$

BOX-COX TRANSFORMATIONS: HOW THEY WORK

- The RSS is contained in the log-likelihood
  - Could use Maximum Likelihood Process to find estimates of coefficients (end up being the same as LS!)
- To find a transformation on e.g. Y, we would modify our RSS (i.e. fit a new model with a different response)

$$RSS = \sum (\Psi_M(Y,\lambda) - \beta_0 - \beta_1 x_i)^2$$

- $\Psi_M(Y,\lambda)$ is a modified power transformation being applied to Y
  - We are using ML to estimate the best value of λ
- This $\Psi_M(Y,\lambda)$ is not pretty or easy to use

$$\Psi_M(Y,\lambda) = \begin{cases} {e^{\frac{1}{n}\Sigma \log(Y)}}^{1-\lambda} \frac{(Y^{\lambda}-1)}{\lambda}, & \lambda \neq 0 \\ {e^{\frac{1}{n}\Sigma \log(Y)}}^{1-\lambda} \log(Y), & \lambda = 0 \end{cases}$$

🤮😰

- So instead, once we get a maximum likelihood estimate for λ we just take Y^{λ} as our transformation
  - Still gives us a line of best that minimizes RSS and gets close to Normality
  - Easier to interpret than above
  - When $\lambda = 0$ we use $\ln(Y)$ as our transformation

BOX-COX ON X'S AND BOTH Y & X SIMULTANEOUSLY

- To transform only predictors, you'd follow the same
- Define RSS with a transformed predictor, e.g. in SLR:

$$RSS = \sum (y_i - \beta_0 - \beta_1 \Psi_S(X, \lambda))^2$$

- Continue to estimate the value of λ that gives minimal RSS, maximal likelihood, and gets as close to Normality as possible

$$\Psi_S(X, \lambda) = \begin{cases} \frac{(X^\lambda - 1)}{\lambda}, & \lambda \neq 0 \\ \ln(X), & \lambda = 0 \end{cases}$$

- Just use X^λ or $\ln(X)$ as before for simplicity

- Can also try transformations to both Y and X simultaneously
- If working in the SLR setting, we'd define RSS as

$$RSS = \sum (\Psi_M(Y, \lambda_y) - \beta_0 - \beta_1 \Psi_M(X, \lambda_x))^2$$

- The ML process would yield an estimate for (λ_y, λ_x) that would use the same yucky formulae as before
  - Instead we use simpler option of Y^{λ_y} and X^{λ_x}
- For each version of Box-Cox, process would be the same for multiple linear models, just more predictors being transformed and more λs to estimate.

BOX-COX IN PRACTICE

- The ML process to estimate λs does not yield a closed-form expression so requires a computer
- As you noticed, the actual Box-Cox transformations are not very nice and are very difficult to work with
- Using simpler powers as transformations is much easier to interpret and apply
  - You still get close to Normality like the original Box-Cox power transformation
  - Avoid the headache of figuring out what your variable now means
- There are a few guidelines we can use to select a simple power

How to Select a Simple Power

- Pick something even simpler than the estimated λ:
  - E.g. estimates $\lambda = 0.103$ use $\ln(Y)$ since close to 0
  - Aim for values near
    - $\lambda = 0.5$ a square root transformation
    - $\lambda = 0.33$ a cube root transformation
    - $\lambda = 0.25$ a fourth root transformation
    - $\lambda = -0.5$ a reciprocal square root
    - $\lambda = -1$ a reciprocal (inverse) transformation
  - Function will report confidence intervals so can use those as a range of possibilities

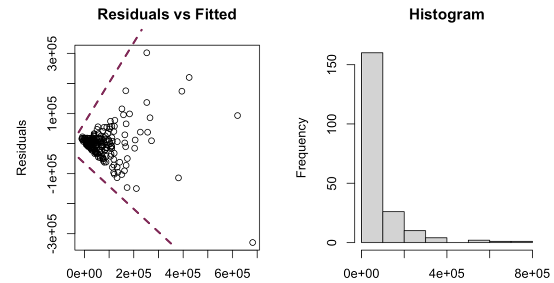
MODULE 4 OUTLINE

1. Variance Stabilizing Transformations and Other Common Functions
2. Box-Cox Transformations for Normality
3. Application and Interpretation of Transformed Models
4. Impact of Violations on Sampling Distributions

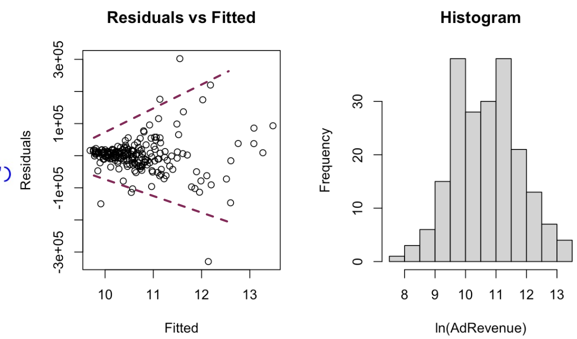
EXAMPLE: VARIANCE STABILIZING TRANSFORMATION

- Suppose we wish to predict the revenue from advertising based on number of ad pages, subscription revenue and newsstand revenue.
- We fit a model and check one of our residual plots
- Based on skew, consider ln() transformation to response
- We see some improvement, but other violations may be present

```
> model1 <- lm(AdRevenue ~ AdPages + SubRevenue + NewsRevenue, data=mag)
> e_hat <- resid(model1)
> y_hat <- fitted(model1)
>
> par(mfrow=c(1,2))
> plot(e~y_hat, main="Residuals vs Fitted", ylab="Residuals", xlab="Fitted")
> hist(mag$AdRevenue, main="Histogram", xlab="AdRevenue")
```




```
> mag$lnAdRevenue <- log(mag$AdRevenue)
> model2 <- lm(lnAdRevenue ~ AdPages + SubRevenue + NewsRevenue, data=mag)
> e_hat <- resid(model2)
> y_hat <- fitted(model2)
>
> par(mfrow=c(1,2))
> plot(e~y_hat, main="Residuals vs Fitted", ylab="Residuals", xlab="Fitted")
> hist(mag$lnAdRevenue, main="Histogram", xlab="ln(AdRevenue)")
```

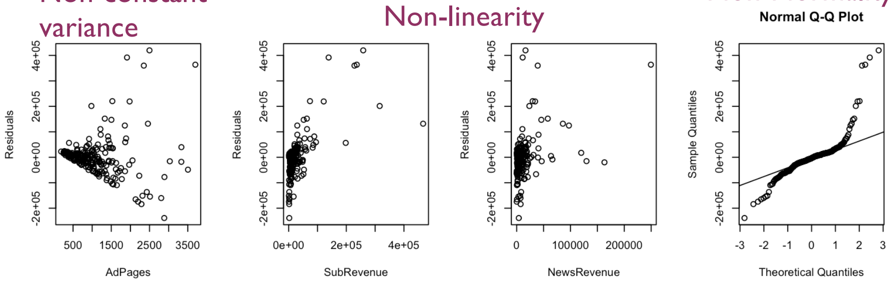


EXAMPLE: BOX-COX ON Y AND ON X

Non-constant variance

Non-linearity

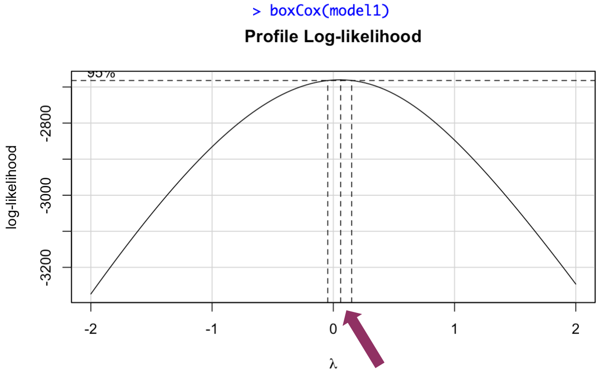
Non-Normality



Try Box-Cox Transformation on Y:

```
> library(car)
Loading required package: carData
> boxCox(model1)
```

95% CI on MLE overlaps 0, so ln transformation reasonable

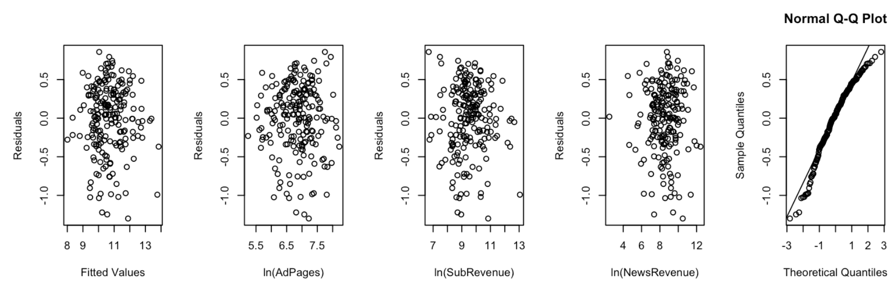


Try Box-Cox Transformation on the predictors:

```
> p <- powerTransform(cbind(mag[,3:5]))
> summary(p)
bcPower Transformations to Multinormality
            Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
AdPages        0.1119        0.00      -0.0869       0.3107
SubRevenue    -0.0084        0.00      -0.0973       0.0804
NewsRevenue    0.0759        0.08       0.0106       0.1412
```

Re-check residual plots:

```
> model3 <- lm(lnAdRevenue ~ log(AdPages) + log(SubRevenue) + log(NewsRevenue), data=mag)
> e_hat <- resid(model3)
> y_hat <- fitted(model3)
>
> par(mfrow=c(1,5))
> plot(e_hat ~ y_hat, xlab="Fitted Values", ylab="Residuals")
> plot(e_hat ~ log(mag[,3]), xlab="ln(AdPages)", ylab="Residuals")
> plot(e_hat ~ log(mag[,4]), xlab="ln(SubRevenue)", ylab="Residuals")
> plot(e_hat ~ log(mag[,5]), xlab="ln(NewsRevenue)", ylab="Residuals")
> qqnorm(e_hat); qqline(e_hat)
```



EXAMPLE: BOX-COX ON Y AND X SIMULTANEOUSLY

- There are reasons to only attempt to estimate a Box-Cox transformation on Y or only on the predictors (or even some of the predictors)
- But we can also do them simultaneously.
- Even here, we seem to select the same transformations
  - 0.11 is close enough to 0 to select a ln() transformation
  - An estimate of 0 tells us to use a ln() transformation
  - 0.08 is also close enough to 0 to use a ln()
- However, it is not always true that you'll estimate the same powers

```
> p <- powerTransform(cbind(mag[,2:5]))
> summary(p)
bcPower Transformations to Multinormality
            Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
AdRevenue      0.1071        0.11       0.0299       0.1843
AdPages        0.0883        0.00      -0.0755       0.2521
SubRevenue    -0.0153        0.00      -0.0862       0.0557
NewsRevenue    0.0763        0.08       0.0115       0.1410
```

INTERPRETATIONS WITH TRANSFORMATIONS

- To interpret models that use transformations, always remember to incorporate the variables as they now are
- E.g. if a predictor was squared, then we say "for a one-unit increase in x^2", not in x.
- We don't back-transform to get variables on the original measurement scale
  - When response has been transformed, predicted values only represent conditional means on the transformed scale, not the original one

From Previous Example

- Intercept: *The expected log of Revenue due to Advertising when log Ad Pages, log Subscription revenue and log Newsstand Revenue are 0.*

- Slope: *For a one unit increase in the log of Advertising Pages, the slope is the expected change in log of Advertising Revenue for a fixed log Subscription Revenue and log Newsstand Revenue.*

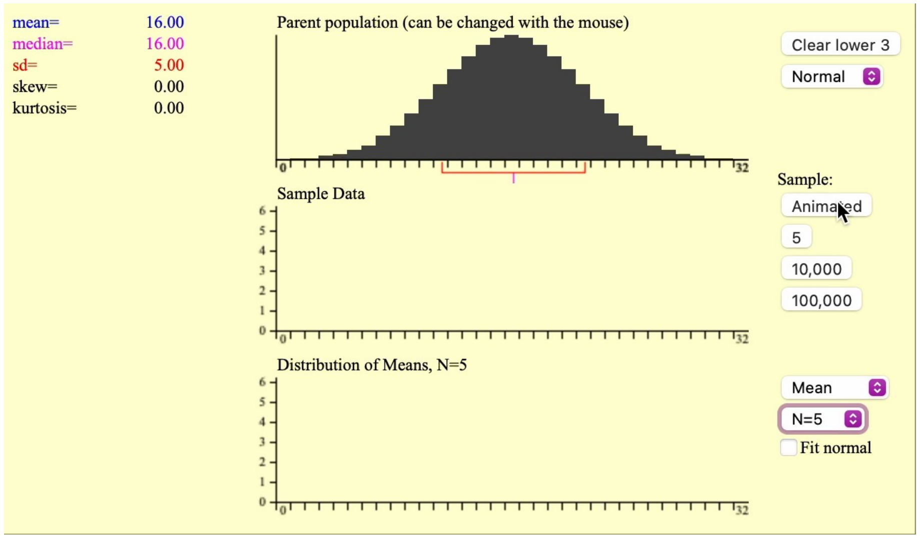
MODULE 4 OUTLINE

1. Variance Stabilizing Transformations and Other Common Functions
2. Box-Cox Transformations for Normality
3. Application and Interpretation of Transformed Models
4. Impact of Violations on Sampling Distributions

SAMPLING DISTRIBUTION PURPOSE

- With any estimate, we need a measure of variation or error as it is based on data
- Specifically, need to describe how the value varies from one sample to another
- Any time we define a sampling distribution, we utilize assumptions about the population
  - Even if we didn't know it when we use a confidence interval
  - Assumptions give us properties of this distribution, like mean or variance or shape
- When assumptions don't hold, these properties are no longer true.

https://onlinestatbook.com/stat_sim/sampling_dist/



SAMPLING DISTRIBUTION OF ESTIMATED COEFFICIENTS

- Our assumptions say $\boldsymbol{Y}|\boldsymbol{X} \sim N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$
- Recall our estimates are found as $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$
  - So the $\widehat{\boldsymbol{\beta}}$ are a function of $\boldsymbol{Y}$
- Assuming the assumptions hold in our population and for our model, we can find our sampling distribution of $\widehat{\boldsymbol{\beta}}$
  - Use the property of linearity of Normal random variables

$$\boldsymbol{AY} \sim N(\boldsymbol{A\mu_Y}, \boldsymbol{A\Sigma A}) \quad \text{or} \quad \sum a_i Y_i \sim N\left(\sum a_i \mu_i, \sum a_i^2 \sigma_i^2\right)$$

where $\boldsymbol{A}$ is a matrix of constants.

<u>Sampling Distribution of $\widehat{\boldsymbol{\beta}}$</u>

- We will see this derived in more detail later.
- Using linearity of Normals, we get

$$\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$$

- This tells us that, when assumptions hold, our estimators are:
  - Unbiased (i.e. estimate on average the value they should)
  - Are generally correlated (i.e. off-diagonals of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ not guaranteed to be 0)
  - Use the same constant variance as the errors.

VIOLATIONS BREAK THE SAMPLING DISTRIBUTION

$$\boxed{\widehat{\boldsymbol{\beta}} \sim \cancel{N}(\cancel{\boldsymbol{\beta}}, \cancel{\sigma}^2(\cancel{\boldsymbol{X}^T}\boldsymbol{X})^{-1})}$$

?

- What happens if we have model violations?
  - Linearity: mean is no longer $\boldsymbol{\beta}$ and so our estimates are no longer unbiased
  - Constant variance: we no longer have a single σ^2 as part of the variance
    - Often means we either under- or over-estimate the variation in our estimator
  - Uncorrelated errors: the variance in our estimators will be under- or over-estimated because we are borrowing information across individuals/measurements
  - Normality: our estimators would not have a Normal sampling distribution, although with large samples it might be approximately Normal.
- Essentially, when assumptions are violated, we don't know what the correct way is to describe variation due to sampling and so inference will be incorrect or misleading.

MODULE TAKE-AWAYS

1. Why does a Variance Stabilizing Transformation correct non-constant variance?
2. What is a Box-Cox transformation trying to achieve?
3. How do we select and implement transformations to correct model violations?
4. What does this change in how we interpret our model coefficients?
5. Why is it important to attempt to correct model violations?