STAT302 Methods of Data Analysis 1
Module 4: Sampling distributions and correcting assumptions

Austin Brown

University of Toronto, Department of Statistical Sciences

September 25, 2024

Lecture 1: Review

Review: regression analysis so far

- Exploratory analysis: Scatterplots, scatterplot matrix, boxplots, histograms, numerical summaries
- Choose a regression model: response and predictors
- Fit a regression model
- Regression model diagnostics: residual plots, QQ plots, histograms, etc.
- Draw conclusions: interpret coefficients, etc.

Lecture 1: Complex predictors

Learning goals

- Extend the models we use so far to allow interaction terms
- Understand and implement polynomial regression

Lecture 1: Interactions

Interactions

Interactions: predictor has a different effect on the response for different values of another predictor. Represented by multiplications: $x_1 x_2$, etc.

$$\hat{E}(Y|X = \boldsymbol{x}) = \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2$$

- We have 2 “main effects" and 1 “interaction" term
- Here $x_3 = x_1 x_2$ and so forth.

Interpret interactions

Fixing all the other variables, the effect of one variable differs based on the value of another fixed variable.

Given $x_2 = a$,

$$\hat{Y} = \hat{\beta}_0 + (\hat{\beta}_1 + a\hat{\beta}_3)x_1$$

- **Interpretation:** Given $x_2 = a$, the average change of the response for one unit increase in x_1 is $\hat{\beta}_1 + a\hat{\beta}_3$ with all predictors held fixed (here there are no other predictors).

Interpretation of indicator coefficients with interactions

Interactions of indicators and continuous predictor: Each continuous predictor has a different slope term for each level of the category.

$$\begin{aligned}\hat{E}(Y|X=\boldsymbol{x}) = \hat{y} = \hat{\beta}_0 &+ \hat{\beta}_1 I(\text{category B}) + \hat{\beta}_2 I(\text{category C}) \\ &+ \hat{\beta}_3 x_3 \\ &+ \hat{\beta}_4 I(\text{category B}) * x_3 + \hat{\beta}_5 I(\text{category C}) * x_3\end{aligned}$$

- Each category has a different intercept
- Each category has a different slope for the variable x_3

Interpret interactions with indicators

Fixing all the other variables, the effect of one variable differs based on the value of another fixed variable.

Given category B,

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 + (\hat{\beta}_3 + \hat{\beta}_4)x_3$$

- **Interpretation:** What is the interpretation here?

Example: iris data

Measurements (cm) of the variables sepal length and width and petal length and width for 50 flowers from each of 3 species of iris (setosa, versicolor, virginica).

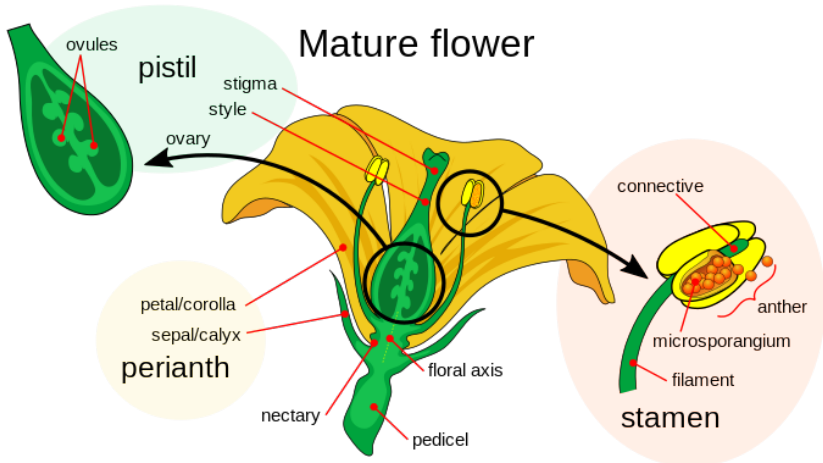


Figure: Picture by Mariana Ruiz

```
data(iris)
```

Example: iris data

$$\hat{Petal}_L = 6.4 - 4.2Sepal_W + 0.2Sepal_L + 0.5Sepal_W * Sepal_L$$

```
# Use colon for interactions. Can also use star.
fitted_model = lm(Petal.Length ~ Sepal.Width + Sepal.Length + Sepal.Width:Sepal.
    Length, data = iris)

X_matrix = model.matrix(fitted_model)
```

Question: Given a 5cm Sepal width, interpret the slope term for Sepal length.

Example: iris data

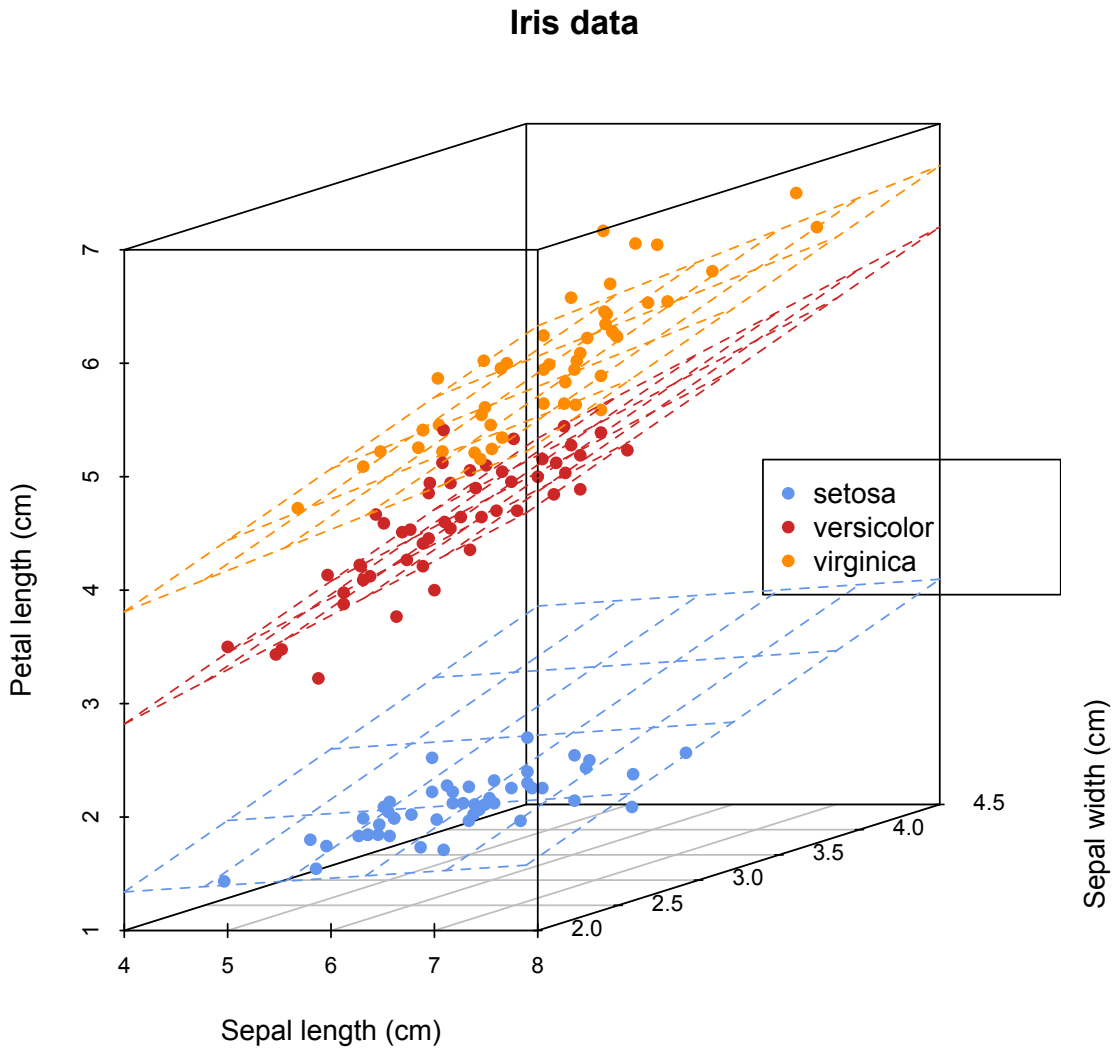
$$\hat{Petal}_L = -0.48 + 0.18I(versicolor) + 1.53I(virginica) - 0.35Sepal_W + 0.65Sepal_W I(versicolor) + 0.47Sepal_W I(virginica) + 0.63Sepal_L$$

```
fitted_model = lm(Petal.Length ~ Species + Sepal.Width + Species:Sepal.Width +
     Sepal.Length, data = iris)
X_matrix = model.matrix(fitted_model)
coef(fitted_model)
```

Question: Compute and interpret the slope term for sepal width for a virginica species.

Fit the iris data

We can have 3 completely different planes if we use different slope terms for each species of iris flowers.



Lecture 1: Polynomial regression

Professor salary example

Let's work through this example in R (See also [Sheather, 2009]).

Polynomial regression

We can add new predictors as polynomial powers of an existing predictor variable.

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + + \beta_k x^k + e$$

- This is still linear when we define $x_1 = x$ and $x_2 = x^2$, etc.
- We cannot use our current interpretation of the slope coefficients with polynomial terms!

```
fit_quadratic = lm(y ~ x + I(x^2))

fit_cubic = lm(y ~ x + I(x^2) + I(x^3))
```

Polynomial regression

The regression models can become quite complicated quickly with polynomial terms in multiple predictor variables:

$$\begin{aligned} Y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 \\ + \beta_3 x_2 + \beta_4 x_2^2 + \beta_5 x_2^3 + e \end{aligned}$$

Lecture 2: Transformations

Learning goals

- Review the importance of checking regression modeling assumptions
- Learn to implement transformations to attempt to correct deviations from the regression modeling assumptions
- Understand and implement transformations to the predictors only
- Understand and implement transformations to the response only and the Box-Cox method
- Understand and implement transformations to both response and predictors
- Understand drawbacks to transformations

Lecture 2: Checking assumptions

What if the plots look problematic?

The diagnostic plots (i.e. residual plots) are providing strong evidence against the regression model:

- Make this clear in your report/conclusions.
- Conclusions may be less reliable due to the evidence against the regression model. The estimates may be untrustworthy.

In this lecture, we try to remedy it.

Lecture 2: Transformation basics

Transformations

- Transform the predictors only
- Transform the response only
- Transform both the response and predictors

Choosing a transformation

We will study "trial and error" methods for picking transformations and the Box-Cox method for an automatic approach. See [Weisberg, 2005, Sheather, 2009] for references to more complex methodology on choosing a transformation.

Choosing a transformation basics:

- Plot the response versus predictors to look for an appropriate transformation.
- Plot residuals versus fitted and versus predictors.
- Choose a transformation to try to fix the residuals plots.

Transformations may make interpretations of the estimated coefficients difficult, but often improves prediction.

Common transformations

- Logarithmic transformation: $g(x) = \log(x)$ ($\log$ is natural log $\ln$)
- Power transformations: $g(x) = x^\lambda$ for $\lambda > 0$
- Inverse power transformations: $g(x) = 1/x^\lambda$ for $\lambda > 0$

Transformation to the predictors only

Find a transformation g for the predictors:

$$Y = \beta_0 + \beta_1 g(x) + e$$

Examples:

$$g(x) = \log(x) \qquad\qquad g(x) = x^{\lambda}$$

Transformation applied to the response

Find a transformation g for the response:

$$g(Y) = \beta_0 + \beta_1 x + e$$

Examples:

$$g(x) = \log(x) \qquad\qquad g(x) = x^{\lambda}$$

Transformation can be applied to the predictors and response

$$g(Y) = \beta_0 + \beta_1 g(x) + e$$

Examples:

$$g(y) = \log(y) \qquad\qquad g(y) = y^{\lambda}$$

- Use scatter plots and residual plots to determine a transformation

Lecture 2: Box-Cox method

Box-Cox method

The Box-Cox transformation attempts to "fix" the residuals by transforming the residuals to approximate normality:

$$\psi(\boldsymbol{Y}) = \begin{cases} GM(\boldsymbol{Y})^{1-\lambda}(\boldsymbol{Y}^{\lambda} - 1)/\lambda, \lambda \neq 0 \\ GM(\boldsymbol{Y}) \log(\boldsymbol{Y}), \lambda = 0 \end{cases}$$

where $GM(\boldsymbol{Y})$ is the geometric mean of $\boldsymbol{Y}$. Uses maximum likelihood to choose λ so the residuals are approximately normally distributed. The Box-Cox method is **not** transforming for linearity.

- Often just use the power or log transformation based on the Box-Cox suggestion in practice.

Box-Cox method

The Box-Cox method is an automatic way based on maximum likelihood to choose a transformation. Provides a confidence interval for possible λ choices.

```
# Choose lambda
boxcox(fit)
lambda = 1/2

# Transform and fit
gm = prod(y)^(1/n)
y_transformed = gm^(1-lambda) * (y^lambda - 1)/lambda
transformed_fit = lm(y_transformed ~ x)
```

Lecture 2: Activity

Predicting magazine advertisement revenue

An analyst is interested in understanding the relationship between revenue from magazine sales and that from advertising. The analyst has obtained some US data from Advertising Age's 14th annual Magazine 300 report which was released in September 2003. Data are available for 204 US magazines for the following variables: Interest centers on building a regression model to predict Ad Revenue (on average) from Ad Pages, Sub Revenue and News Revenue.

- Y = AdRevenue = Revenue from advertising (in thousands USD)
- X1= AdPages = Number of pages of paid advertising
- X2= SubRevenue = Revenue from paid subscriptions (in thousands of USD)

Activity

Work together in groups using the Rmarkdown template and find a transformation.

Lecture 2: Concluding remarks

Interpreting transformations of response and predictors

- To interpret regression models that use transformations, always remember to incorporate the variables as they are. $\log(Y)$, $\log(x)$, etc.
- If a predictor was squared, then we say "for a one-unit increase in x^2".
- When response has been transformed, predicted values only represent conditional means on the transformed scale, not the original one.
- We don't back-transform to get variables on the original measurement scale. See [Sheather, 2009, Weisberg, 2005] for some examples where the transformation has some approximate interpretation.

Summary

Choosing a transformation:

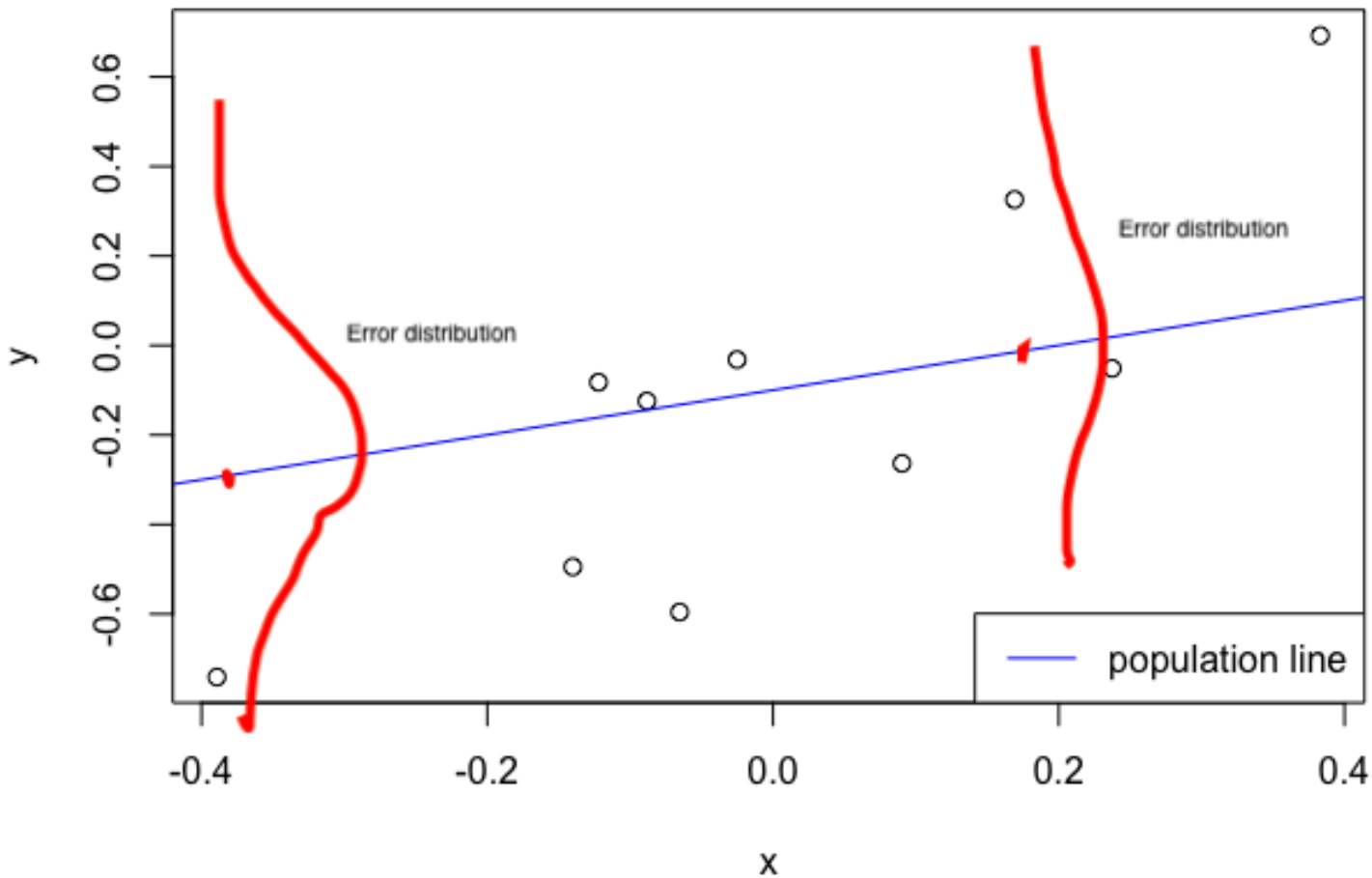
- Plot the data: scatterplot, scatterplot matrix
- Plot residuals versus fitted, predictors
- Other plots may also be helpful, Response versus fitted, QQ plot, histograms
- Determine an appropriate transformation
- See [Weisberg, 2005, Sheather, 2009] for references to more complex methodology on choosing transformations

Lecture 3: Sampling distributions of the least squares estimators

Learning goals

- Understand the properties of the least squares estimators including their sampling distribution.
- Implement, read, and interpret relevant output from the "summary" function in R.

Visualiziation regression model



Review

Consider a random vector $X = (X_1, X_2)$. The entries of the vector are random and possibly correlated random variables.
If $X \sim N_2(\mu, \Sigma)$ is multivariate normal. Then

$$E(X) = \mu = \begin{pmatrix} E(X_1) \\ E(X_2) \end{pmatrix}$$

$$\begin{aligned} \text{Cov}(X) &= E\left[(X-\mu)(X-\mu)^T\right] = \Sigma \\ &= \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) \\ Cov(X_1, X_2) & Var(X_2). \end{pmatrix} \end{aligned}$$

The entries of Σ are the covariances: $Cov(X_i, X_j)$. This is the same for dimension k vectors and dimension $k+1$ vectors and so on.

If A is a fixed matrix of dimensions 2×2,

$$E(AX) = A\mu$$

The same is true for arbitrary dimensions k, $k + 1$, etc.

If A is a fixed matrix of dimensions 2×2,

$$Cov(AX) =$$

$$=$$

$$= ACov(X)A^T.$$

The same is true for arbitrary dimensions k, $k+1$, etc.

Lecture 3: Expected values

Simple linear regression: sampling distributions

Under the simple linear regression model, recall

$$\hat{\beta}_1 = \frac{s_{XY}}{s_{XX}}$$
$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}$$

where

$$s_{XY} = \sum_{i=1}^{n} (Y_i - \overline{Y})(x_i - \overline{x})$$
$$s_{XX} = \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

Simple linear regression: unbiased proof

We will use the identity

$$\sum_i (x_i - \overline{x}) = \qquad\qquad = 0. \qquad (1)$$

Using identity (1),

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \overline{Y})(x_i - \overline{x})}{s_{XX}}$$

$$=$$

Simple linear regression: unbiased proof

$$E\left(\hat{\beta}_1|\boldsymbol{X}\right) = E\left(\frac{\sum_{i=1}^{n} Y_i(x_i - \overline{x})}{s_{XX}}|\boldsymbol{X}\right) \qquad \text{(identity (1))}$$

$$= \qquad \text{(linearity of expected value)}$$

$$= \qquad \text{(regression model)}$$

$$= \qquad \text{(identity (1))}$$

$$= \qquad \text{(identity (1))}$$

$$= \beta_1.$$

Simple linear regression: unbiased proof

Due to the regression model:

$$E\left(\overline{Y}|\boldsymbol{X}\right) =$$

Using this, we have

$$E\left(\hat{\beta}_0|\boldsymbol{X}\right) = E\left(\overline{Y} - \hat{\beta}_1\overline{x}|\boldsymbol{X}\right) \qquad \text{(definition)}$$

$$= \qquad \text{(linearity of expected value)}$$

$$= \qquad \text{(using our previous results)}$$

$$= \beta_0.$$

Lecture 3: Expected values of multiple linear regression estimators

Multiple linear regression: sampling distributions

Under the regression model, recall

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}.$$

is **random**.

Multiple linear regression: unbiased proof

$$E\left(\hat{\boldsymbol{\beta}}|\boldsymbol{X}\right) = \qquad \text{(definition)}$$

$$= \qquad \text{(linearity of expected value)}$$

$$=$$

$$= \boldsymbol{\beta}.$$

Lecture 3: Variance of simple linear regression estimators

Simple linear regression: variance proof

$$\text{Var}\left(\hat{\beta}_1|\boldsymbol{X}\right) = \text{Var}\left(\frac{\sum_{i=1}^n Y_i(x_i - \overline{x})}{s_{XX}}|\boldsymbol{X}\right) \qquad \text{(identity (1)}$$

$$= \qquad \text{(property of variance}$$

$$= \qquad \text{(variance of independent r.v.}$$

$$= \qquad \text{(property of variance}$$

$$= \qquad \text{(regression model}$$

$$= \frac{\sigma^2}{s_{XX}}.$$

Simple linear regression: variance proof

$$
\begin{aligned}
&\mathsf{Var}\left(\hat{\beta}_0|\boldsymbol{X}\right) \\
&= \mathsf{Var}\left(\overline{Y} - \hat{\beta}_1\overline{x}|\boldsymbol{X}\right) && \text{(definition)} \\
&= \mathsf{Var}\left(\overline{Y}|\boldsymbol{X}\right) + \overline{x}^2\mathsf{Var}\left(\hat{\beta}_1|\boldsymbol{X}\right) \\
&\quad - 2\mathsf{Cov}\left(\overline{Y}, \overline{x}\hat{\beta}_1|\boldsymbol{X}\right) && \text{(variance properties).}
\end{aligned}
$$

From the regression model and previous results,

$$
\begin{aligned}
\overline{x}^2\mathsf{Var}\left(\hat{\beta}_1|\boldsymbol{X}\right) &= \frac{\overline{x}^2\sigma^2}{s_{XX}} \\
\mathsf{Var}\left(\overline{Y}|\boldsymbol{X}\right) &= \sum_{i=1}^{n} \frac{\mathsf{Var}\left(Y_i|\boldsymbol{X}\right)}{n^2} = \frac{\sigma^2}{n}.
\end{aligned}
$$

Simple linear regression: variance proof

$$\mathsf{Cov}\left(\overline{Y}, \hat{\beta}_1 | \boldsymbol{X}\right) = 0$$

Simple linear regression: variance proof

Putting everything together:

$$\mathsf{Var}\left(\hat{\beta}_0 | \boldsymbol{X}\right) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{s_{XX}}\right).$$

Simple linear regression: sampling distributions

Under the simple linear regression model assumption, the **sampling distributions** are normally distributed:

$$\hat{\beta}_1 \sim N\left[\beta_1, \frac{\sigma^2}{s_{XX}}\right]$$

$$\hat{\beta}_0 \sim N\left[\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{s_{XX}}\right)\right]$$

Estimated standard errors:

$$\hat{se}(\hat{\beta}_1) = \hat{\sigma}\sqrt{\frac{1}{s_{XX}}}$$

$$\hat{se}(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{s_{XX}}}$$

R example

Simple linear regression model for iris data:

$$Petal_L = \beta_0 + \beta_1 Sepal_L + e.$$

Compute the estimated standard errors using R using the code below:

```
data(iris)
fit = lm(Petal.Length ~ Sepal.Length, data = iris)
fitted_values = fitted(fit)

y_values = iris$Petal.Length
x_values = iris$Sepal.Length
mean_x = mean(x_values)

RSS = sum( (y_values - fitted_values)^2 )

s_xx = sum( (x_values - mean_x)^2 )

# Use sqrt()
```

R example

Let's understand relevant parts of the "summary" function. Check your computation against the summary output.

Lecture 3: Covariance of multiple linear regression estimators

Multiple linear regression: variance proof

$$\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right]^T = \boldsymbol{X}\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right]^T = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}.$$

Using this, we have

$$\text{Cov}\left(\hat{\boldsymbol{\beta}}|\boldsymbol{X}\right) = \text{Cov}\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}|\boldsymbol{X}\right) \qquad \text{(definition)}$$

$$= \qquad \text{(property of covariance)}$$

$$= \qquad \text{(regression model)}$$

$$= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}.$$

Multiple linear regression: sampling distribution

Similar reasoning to simple linear regression applies to multiple linear regression. Under the multiple linear regression model assumption, the **sampling distribution** has a multivariate normal distribution in dimension $p+1$:

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} \sim N_{p+1}\left(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right).$$

Multiple linear regression: sampling distribution

Under the linear regression model assumption, each i-th entry of $\hat{\boldsymbol{\beta}}$ is normally distributed

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2 (X^T X)^{-1}_{i+1,i+1})$$

where $\sigma^2 (X^T X)^{-1}_{i+1,i+1}$ is the diagonal entry of the covariance matrix $\text{Cov}(\hat{\boldsymbol{\beta}})$.

Estimated standard errors:

$$\hat{se}(\hat{\beta}_i) = \hat{\sigma} \sqrt{(X^T X)^{-1}_{i+1,i+1}}.$$

Lecture 3: Back to simple linear regression

Example: Simple linear regression

$\hat{\beta}_0, \hat{\beta}_1$ are correlated:

$$\begin{aligned}
\mathsf{Cov}(\hat{\beta}_0, \hat{\beta}_1) &= \mathsf{Cov}(\overline{Y} - \overline{x}\hat{\beta}_1, \hat{\beta}_1) \\
&= \mathsf{Cov}(\overline{Y}, \hat{\beta}_1) - \overline{x}\mathsf{Cov}(\hat{\beta}_1, \hat{\beta}_1) \\
&= -\overline{x}\mathsf{Var}(\hat{\beta}_1) \\
&= \frac{-\overline{x}\sigma^2}{s_{XX}}.
\end{aligned}$$

Example: Simple linear regression

Let us check this matches our results for simple linear regression:

$$\mathsf{E}\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix} = \begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}$$

and

$$\begin{aligned}\mathsf{Cov}\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix} &= \begin{pmatrix}\sigma^2(1/n + \overline{x}^2/s_{XX}) & -\sigma^2\overline{x}/s_{XX}\\ -\sigma^2\overline{x}/s_{XX} & \sigma^2/s_{XX}\end{pmatrix}\\ &= \frac{\sigma^2}{ns_{XX}}\begin{pmatrix}\sum_i x_i^2 & -n\overline{x}\\ -n\overline{x} & n\end{pmatrix}\\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\end{aligned}$$

using the formula for inverse of 2×2 matrix.

Lecture 3 : R example

Activity

Linear regression model for iris data:

$$Petal_L = \beta_0 + \beta_1 Sepal_L + \beta_2 Sepal_W + e$$

Compute the standard error estimates using the code below.

```
data(iris)
fit = lm(Petal.Length ~ Sepal.Length + Sepal.Width, data = iris)

summary(fit)
```

Module takeaways

- Why is it important to attempt to correct model violations?
- How do we select and implement transformations to correct model violations?
- Are transformations always required or always a good idea?
- What does this change in how we interpret our model coefficients?
- What is a Box-Cox transformation/method trying to achieve?
- What does the "linear" mean in linear regression Polynomial regression, interaction terms, transformations, etc.

References I

Simon Sheather. *A modern approach to regression with R*. Springer Science & Business Media, 2009.

Sanford Weisberg. *Applied linear regression*, volume 528. John Wiley & Sons, 2005.