STAT302 Methods of Data Analysis 1
Module 5: Inferences in Linear Regression and Prediction

Austin Brown

University of Toronto, Department of Statistical Sciences

October 2, 2024

Week 6 / Term Test Information

- Term Test Information is posted on Quercus
- No lecture on Wednesday next week.
- I will hold additional office hours / review questions during lecture time 6pm-8pm My150. But I will not hold office hours next week at 4-5pm.

Some rough guidelines for transformations

- Finding the right transformation can change from problem to problem and there is no universal answer, but here are some guidelines:
- Issues with the residual variance/distribution: Transform the response to stabilize the variance and/or fix the distribution of the residuals.
- Issues with linearity: Transform the predictors to address linearity.

Box plots in R

If you plot residuals versus categorical predictors, you will get a boxplot. The same principles apply. However, this particular plot can be less difficult to draw conclusions from.

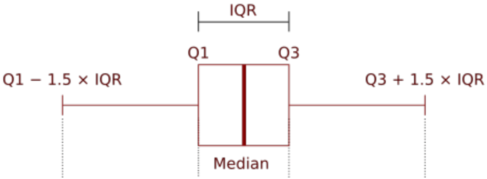


Image credit goes to Xianjun Dong on R bloggers.

Last Week Review

Review where we left off and catch up.

Lecture 1: Confidence intervals for coefficients

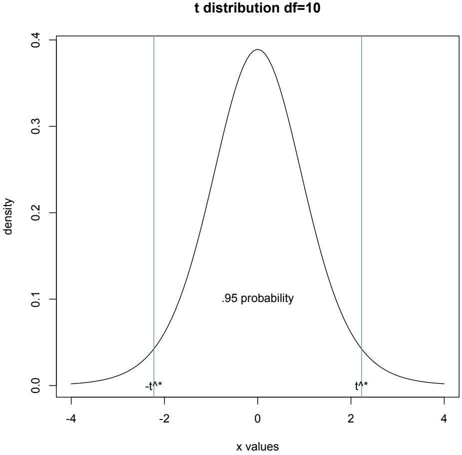
Confidence interval form

Compute a confidence interval from a data set:

[Point estimate] $\pm$ [critical value / quantile] $\times$ [estimate for the standard error]

t-distribuiton and its quantiles

t^* is the $.975$ quantile for the t-distribution with $df = 10$.




```
qt(.975, df = 10)
```

Simple linear regression: sampling distributions

Under the regression (population) model assumptions,

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{s_{XX}}} \sim t_{n-2}$$

A $(1-\alpha) * 100\%$ confidence interval for β_1 is

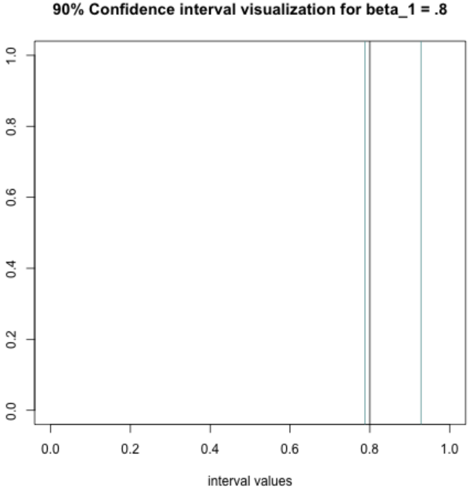
$$\hat{\beta}_1 \pm t^*_{n-2,1-\alpha/2} \cdot \hat{\sigma}\sqrt{\frac{1}{s_{XX}}}$$

95 percent confidence interval for β_1

$$\text{Prob}\left(-t^*_{n-2,.975} \leq \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{s_{XX}}} \leq t^*_{n-2,.975}\right) = .95. \text{ (Why?)}$$

Confidence interval visualization for β_1

The true $\beta_1 = .8$ and we compute 90% confidence intervals over many different observations from the population.



Interpretation

Interpret a 95% C.I.:

If we were to repeatedly randomly sample from the population corresponding to the regression model and then compute a 95% confidence interval with the fitted regression each time, then 95% of those confidence intervals would include the actual population parameter β_1.

Simple linear regression: sampling distributions

Under the regression model assumptions,

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{s_{XX}}}} \sim t_{n-2}$$

A $(1-\alpha) * 100\%$ confidence interval for β_0 is

$$\hat{\beta}_0 \pm t^*_{n-2,1-\alpha/2} \cdot \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{s_{XX}}}$$

Interpretation

Interpret a 95% C.I.:

If we were to repeatedly randomly sample from the population corresponding to the regression model and then compute a 95% confidence interval with the fitted regression each time, then 95% of those confidence intervals would include the actual population parameter β_0.

Multiple linear regression: sampling distributions

Under the regression model assumptions,

$$\frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}\sqrt{(X^T X)^{-1}_{i+1,i+1}}} \sim t_{n-(p+1)}$$

A $(1-\alpha) * 100\%$ confidence interval for β_i is

$$\hat{\beta}_i \pm t^*_{n-(p+1),1-\alpha/2} \cdot \hat{\sigma}\sqrt{(X^T X)^{-1}_{i+1,i+1}}$$

95 percent confidence interval for β_i

$$\text{Prob}\left(-t^*_{n-p-1,.975} \leq \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}\sqrt{(X^TX)^{-1}_{i+1,i+1}}} \leq t^*_{n-p-1,.975}\right) = .95. \text{ (Why?)}$$

Interpretation

Interpret a 95% C.I.:

If we were to repeatedly randomly sample from the population corresponding to the regression model and then compute a 95% confidence interval with the fitted regression each time, then 95% of the confidence intervals would include the population parameter β_i.

Confidence intervals

What if 0 is in a confidence interval for β_1?

Confidence intervals

We have evidence at level α, that the 0 is a plausible value for the population parameter β_i.

Confidence interval using R

You can compute the confidence interval using the summary output and the critical t-value:

```
summary(fit)

qt(.975, df = n - (p + 1))
```

or R can do everything:

```
confint(fit, level = .95)
```

Lecture 1: CI Example

Interpret in the context of the problem: iris data

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

t_value = qt(.975, df = 147)
se = 0.12236
beta_hat = -1.33862

c(beta_hat - t_value * se,
beta_hat + t_value * se)

# [1] -1.580432 -1.096808
```

- A 95% confidence interval is $[-1.580432, -1.096808]$.
- If we were to repeatedly randomly collect 150 iris flowers and compute a 95% confidence interval each time with this regression model, then 95% of the confidence intervals would contain the true coefficient for sepal width (cm).

Lecture 2: Prediction

Prediction

Consider a new random independent response Y^* and predictors $\boldsymbol{x}^* = (1, x_1^*, \ldots, x_p^*)^T$ from the population:

$$Y^* = \beta_0 + \beta_1 x_1^* + \cdots + \beta_p x_p^* + e^*$$

with an an independent error $e^* \sim N(0, \sigma^2)$.

Our goal is to use the existing regression model fit to predict new responses Y^* on average, that is, $E(Y^*|X = \boldsymbol{x^*})$.

Prediction

- Y^* is random. (Why?)
- $E(Y^*|X = \boldsymbol{x}^*)$ is fixed. (Why?)

Predicting the average response in multiple/multivariate linear regression

We predict the average/mean response of $Y^*, \boldsymbol{x}^*$:

$$\begin{aligned}\hat{Y}^* &= \hat{E}(Y^*|X = \boldsymbol{x}^*)\\ &= \boldsymbol{x}^{*T}\hat{\boldsymbol{\beta}}\\ &= \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \cdots + \hat{\beta}_p x_p^*\end{aligned}$$

Computing $\hat{\beta}_i = b_i$ from observed data, the prediction is:

$$b_0 + b_1 x_1^* + \cdots + b_p x_p^*.$$

Simple linear regression: sampling distribution

Under the simple linear regression model assumptions,

$$
\begin{aligned}
& E(\hat{Y^*}|\boldsymbol{X}, X = x^*) \\
& = E(\hat{\beta}_0 + \hat{\beta}_1 x^* | \boldsymbol{X}, X = x^*) && \text{(definition)} \\
& = && \text{(linearity of expected value)} \\
& = && \text{(unbiased estimators)}
\end{aligned}
$$

Simple linear regression: sampling distribution

Under the simple linear regression model assumptions,

$$
\begin{aligned}
\mathsf{Var}(\hat{Y}^*|\boldsymbol{X}, X = x^*) & \\
&= && \text{(definition)} \\
&= && \text{(See Module 4)} \\
&= && \text{(Variance property)} \\
&= \frac{\sigma^2}{n} + \frac{\sigma^2(x^* - \overline{x})^2}{s_{XX}} && \text{(See Module 4)}
\end{aligned}
$$

Simple linear regression: sampling distributions

Under the simple linear regression model assumption, the **sampling distribution** is

$$\hat{Y}^* \sim N\left[\beta_0 + \beta_1 x^*, \sigma^2 \left(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{s_{XX}}\right)\right]$$

Estimated standard error:

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

Simple linear regression: confidence interval for the mean response

Under the simple linear regression model assumptions,

$$\frac{\hat{Y}^* - [\beta_0 + \beta_1 x^*]}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{s_{XX}}}} \sim t_{n-2}$$

A $1 - \alpha$ confidence interval for $\beta_0 + \beta_1 x^*$ is

$$\hat{Y}^* \pm t^*_{n-2,1-\alpha/2} \cdot \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{s_{XX}}}$$

Interpretation

Interpret a 95% C.I.:

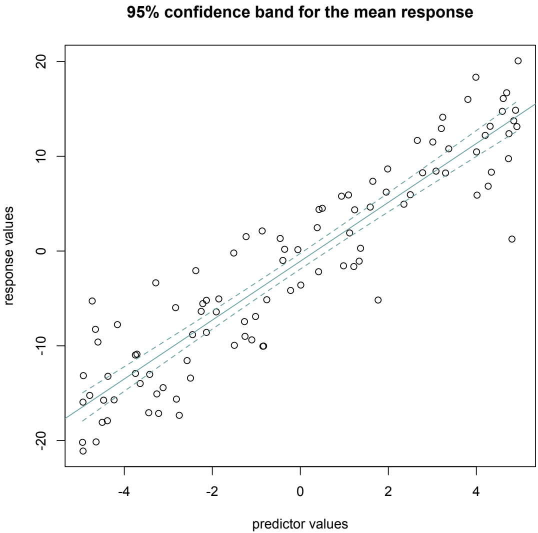
If we were to repeatedly sample from the population and compute a 95% confidence interval each time from the regression fit, then 95% of the intervals would include the population average response at x^* (the average population response is $\beta_0 + \beta_1 x^*$).

Simple linear regression: confidence interval for the mean response

```
pred_means = predict(fit, data.frame(x = new_x_values),
level = .95,
interval = "confidence")
```

Simple linear regression: confidence interval for the mean response

Confidence bands:



Multiple/multivariate linear regression: sampling distributions

Under the regression model assumptions,

$$
\begin{aligned}
& E(\hat{Y^*}|\boldsymbol{X}, X = \boldsymbol{x}^*) \\
& = \boldsymbol{x}^{*T} E(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) && \text{(linearity of expectation)} \\
& = \boldsymbol{x}^{*T} \boldsymbol{\beta} && \text{(unbiased estimator)}
\end{aligned}
$$

$$
\begin{aligned}
& \mathsf{Var}(\hat{Y^*}|\boldsymbol{X}, X = \boldsymbol{x}^*) \\
& = \boldsymbol{x}^{*T} \mathsf{Cov}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}, X = \boldsymbol{x}^*) \boldsymbol{x}^* && \text{(property of covariance)} \\
& = \boldsymbol{x}^{*T} \sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}^* && \text{(Variance of estimator)}
\end{aligned}
$$

Multiple linear regression: sampling distributions

Under the multiple/multivariate linear regression model assumption, the **sampling distribution** is

$$\hat{Y}^* \sim N\left[\boldsymbol{x^*}^T\boldsymbol{\beta}, \sigma^2\boldsymbol{x}^{*T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x^*}\right]$$

Estimated standard error:

$$\hat{se}(\hat{Y}^*) = \hat{\sigma}\sqrt{\boldsymbol{x}^{*T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x^*}}$$

Multiple/multivariate linear regression: sampling distributions

Under the regression model assumptions,

$$\frac{\hat{Y}^* - \boldsymbol{x}^{*T}\boldsymbol{\beta}}{\hat{\sigma}\sqrt{\boldsymbol{x}^{*T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}^*}} \sim t_{n-(p+1)}$$

A $(1-\alpha) * 100\%$ confidence interval for the mean response of Y^* (given $\boldsymbol{x}^*$), which is $\beta_0 + \beta_1 x_1^* + \cdots + \beta_p x_p^*$, is

$$\hat{Y}^* \pm t^*_{n-(p+1),1-\alpha/2} \cdot \hat{\sigma}\sqrt{\boldsymbol{x}^{*T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}^*}$$

95 percent confidence interval for the average response

$$\text{Prob}\left(-t^*_{n-p-1,.975} \leq \frac{\hat{Y}^* - \boldsymbol{x}^{*T}\boldsymbol{\beta}}{\hat{\sigma}\sqrt{\boldsymbol{x}^{*T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}^*}} \leq t^*_{n-p-1,.975}\right) = .95.$$

Interpretation

Interpret a 95% C.I.:

If we were to repeatedly sample the population and compute the confidence interval according to this regression mode, then 95% of the confidence intervals would include the population mean response at $\boldsymbol{x}^*$

Lecture 2: CI Example

Interpret in the context of the problem: iris data

```
data(iris)
fit = lm(Petal.Length ~ Sepal.Length + Sepal.Width, data = iris)
mean_predictions = predict(fit, data.frame(Sepal.Length = 4.5, Sepal.Width = 3),
      interval="confidence")

#         fit       lwr       upr
# 1  1.449535 1.247384 1.651686
```

- The estimated length of a petal of an iris flower in centimeters with a 4.5 centimeter sepal length and a 3 centimeter sepal width is 1.449535 (cm).
- A 95% confidence interval for the average petal length in centimeters of an iris flower with a 4.5 centimeter sepal length and 3 centimeter sepal width is $[1.247384, 1.651686]$.
- If we were to randomly gather iris flowers repeatedly according to this regression model and construct a 95% confidence interval, then 95% of the intervals would contain the population average petal length of an iris flower in centimeters with 4.5 (cm) sepal length and 3 (cm) sepal width.

Lecture 3: Prediction intervals

Prediction

Consider a new sample response Y^* given the fixed predictors value $\boldsymbol{x}^*$ from the population:

$$Y^* = \beta_0 + \beta_1 x_1^* + \cdots + \beta_p x_p^* + e^*$$

with an independent error $e^* \sim N(0, \sigma^2)$.

Our goal is to use the existing regression model fit to find an interval of possible values for Y^*.

Simple linear regression: prediction intervals

Under the simple linear regression model assumptions,

$$E(Y^* - \hat{Y}^* | \boldsymbol{X}, X = x^*)$$

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

$$= \qquad \text{(unbiased estimators)}$$

$$= 0.$$

Simple linear regression: prediction intervals

Under the simple linear regression model assumptions,

$$\mathsf{Var}(Y^* - \hat{Y}^* | \boldsymbol{X}, X = x^*)$$

$$= \qquad\qquad \text{(since } Y^* \text{ is independent)}$$

$$= \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{s_{XX}} \right). \qquad \text{(regression model)}$$

Simple linear regression: sampling distributions

Under the simple linear regression model assumption, the **sampling distribution** is

$$Y^* - \hat{Y}^* \sim N\left[0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{s_{XX}}\right)\right]$$

Estimated standard error:

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

Simple linear regression: sampling distributions

Under the simple linear regression model assumptions,

$$\frac{Y^* - \hat{Y}^*}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{s_{XX}}}} \sim t_{n-2}$$

A $(1-\alpha) * 100\%$ prediction interval for Y^* is

$$\hat{Y}^* \pm t^*_{n-2,1-\alpha/2} \cdot \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{s_{XX}}}$$

Interpretation

Interpret a 95% P.I.:

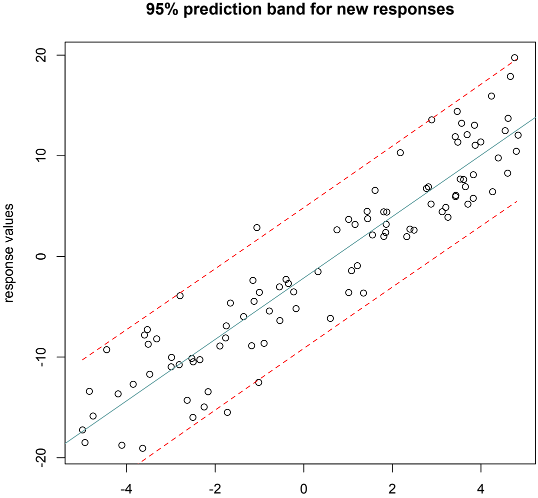
If we were to repeatedly sample from the population according to the regression model and also a new Y^*, x^* and compute a 95% prediction interval each time, then 95% of the intervals would include the population response value.

Simple linear regression: prediction interval for a new response

```
mean_predictions = predict(fit, data.frame(x = x_values),
level = .95,
interval = "prediction")
```

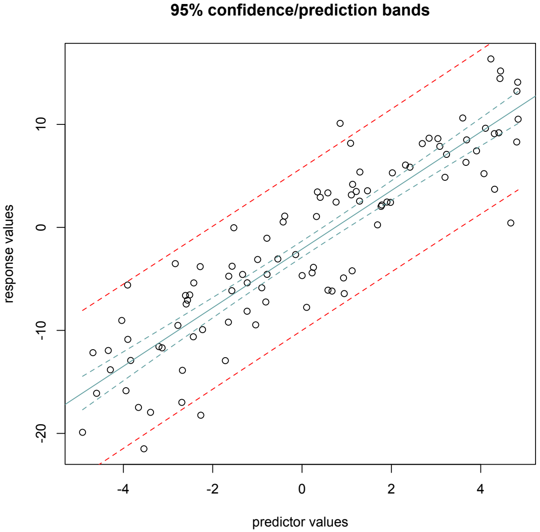
Simple linear regression: prediction interval for a new response

Generate new points and plot the prediction interval for those points:



Simple linear regression: prediction interval for a new response

Why is the prediction band wider?



Multiple linear regression: sampling distributions

Under the multiple/multivariate linear regression model assumption, the **sampling distribution** is

$$Y^* - \hat{Y}^* \sim N\left[0, \sigma^2 \left(1 + \boldsymbol{x}^{*T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}^*\right)\right]$$

Estimated standard error:

$$\hat{se}(Y^* - \hat{Y}^*) = \hat{\sigma}\sqrt{1 + \boldsymbol{x}^{*T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}^*}$$

Multiple/multivariate linear regression: sampling distributions

Under the regression model assumptions,

$$\frac{Y^* - \hat{Y}^*}{\hat{\sigma}\sqrt{1 + \boldsymbol{x}^{*T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}^*}} \sim t_{n-(p+1)}$$

A $1-\alpha$ prediction interval for Y^* is

$$\hat{Y}^* \pm t^*_{n-(p+1),1-\alpha/2} \cdot \hat{\sigma}\sqrt{1 + \boldsymbol{x}^{*T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}^*}$$

95 percent prediction interval

$$\text{Prob}\left(-t^*_{n-p-1,.975} \leq \frac{Y^* - \hat{Y}^*}{\hat{\sigma}\sqrt{1 + \boldsymbol{x}^{*T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x^*}}} \leq t^*_{n-p-1,.975}\right) = .95.$$

Prediction interval for a new responses in R

```
predictions = predict(fit, data.frame(x1 = x1_values, x2 = x2_values,),
level = .95,
interval = "prediction")
```

Lecture 3: PI Example

Interpret in the context of the problem: iris data

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

mean_predictions = predict(fit,
        data.frame(Sepal.Length = 4.5, Sepal.Width = 3),
        interval="prediction")
#         fit       lwr      upr
# 1 1.449535 0.1560447 2.743025
```

- A 95% prediction interval for the petal length in centimeters of an iris flower with 4.5 (cm) sepal length and 3 (cm) sepal width is $[0.1560447, 2.743025]$.
- If we were to randomly gather iris flowers repeatedly according to this regression model and construct a 95% prediction interval each time, then 95% of the time the petal length in centimeters of an iris flower with 4.5 (cm) sepal length and 3 (cm) sepal width would lie in this interval.

Lecture 1: Activity

Activity

Use the template in Quercus to complete the activity.

- Fit with petal length as the response and include all main effects and interaction terms.
- Plot and interpret the residuals versus fitted values and QQ plots for the fit.
- Using the **summary** and **qt**, construct a 95% confidence intervals for some of the coefficients. Check your calculation with the **confint** R function.
- Interpret the interaction between sepal length and width. Use a confidence interval to determine if there is statistical evidence for the interaction term to be included.
- Compute a confidence intervals for the average response for iris flowers with 1 (cm) sepal widths and lengths.

Module takeaways

- How did we determine the properties of the sampling distribution and where did assumptions play a role?
- Why do we use a t-distribution when working with the sampling distribution in practice?
- How do we compute confidence/prediction intervals on regression components?
- How are the inferential procedures concluded?
- What is the difference between estimating a mean response and predicting an actual response?

References I