# STAT 420: Methods of Applied Statistics

## Inference of Linear Regressions

Ruoqing Zhu, Ph.D. <rqzhu@illinois.edu>

Course website: https://sites.google.com/site/teazrq/teaching/STAT420

University of Illinois at Urbana-Champaign
February 16, 2017

# Linear Regressions

- We have already learned how to estimate the parameters $\boldsymbol{\beta}$ of a linear regression, and to assess the goodness-of-fit $R^2$.
- Now we are going to learn how make inference on the parameters.
- Recall that we used simulation study to estimate the $p$-values, now we will derive a rigorous formula for testing:
    - Distribution of the parameter estimates
    - Testing for individual coefficient
    - Confidence interval for individual coefficient
    - Confidence interval for predicting new subject
    - Testing for multiple coefficients and the entire model
- We will derive the results for both SLR and MLR.
- Simulation study can still help us understand these results.

# Distribution of $\widehat{\beta}$

- For now, we will make several assumption:
- The error terms $\epsilon$'s follow i.i.d. Normal distribution, with mean 0 and variance $\sigma^2$:

$$\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2).$$

- We also assume that the covariates $X_i$'s are fixed (not a random variable), with no measurement errors.
- The random errors $\boldsymbol{\epsilon}$ are independent of the covariates $\mathbf{X}$.
- The design matrix $\mathbf{X}$ has full rank.

# Distribution of $\widehat{\beta}$

- Some consequences:
  - The outcome variable $Y_i = X_i^\intercal \boldsymbol{\beta} + \epsilon_i$ also follows a normal distribution:

    $$Y_i \sim \mathcal{N}(X_i^\intercal \boldsymbol{\beta}, \sigma^2),$$

  - The error vector $\boldsymbol{\epsilon}$ follows a multivariate normal distribution:

    $$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}_{n\times 1}, \sigma^2 \mathbf{I}_{n\times n})$$

    where $\mathbf{I}_{n\times n}$ is a diagonal matrix with size $n$, and all diagonal elements are 1.
  - Question: What is the distribution of the outcome vector $\mathbf{Y}$?

# Distribution of $\widehat{\beta}$

- Consider the vector $\mathbf{Y}$, we want to find its distribution.
- Since $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, we have

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}),$$

  independent normal variables, with means $X_i^{\mathsf{T}}\boldsymbol{\beta}$, $i = 1, \ldots, n$ and variance all equal to $\sigma^2$.
- Then the next step is to derive the distribution of $\widehat{\boldsymbol{\beta}}$, which is a linear transformation of $\mathbf{Y}$.
- Recall the multivariate Normal distribution results, since

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{Y}$$

  is a linear transformation of the vector $\mathbf{Y}$, where we let the transformation matrix $\mathbf{A} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}$. Then, what is the distribution of $\mathbf{AY}$?

# Distribution of $\widehat{\beta}$

- When the mean of $\mathbf{Y}$ is $\boldsymbol{\mu}$, and variance-covariance matrix is $\boldsymbol{\Sigma}$, the distribution of $\mathbf{AY}$ is

$$\mathcal{N}(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\mathsf{T})$$

- In our case, $\boldsymbol{\mu} = \mathbf{X}\beta$, $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, and $\mathbf{A} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}$, so

$$\begin{aligned}
&\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{Y} \\
&\quad \sim \mathcal{N}\left((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{\beta}, (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\sigma^2\mathbf{I}((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T})^\mathsf{T}\right) \\
\implies \quad &\widehat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right)
\end{aligned}$$

# Distribution of $\widehat{\beta}$

- Properties of the distribution $\mathcal{N}\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right)$ ?
  1. $\mathsf{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, so $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator of the parameter $\boldsymbol{\beta}$.
  2. $\mathbf{X}^\mathsf{T}\mathbf{X}$ grows in the same rate as the sample size $n$ (essentially each element of $\mathbf{X}^\mathsf{T}\mathbf{X}$ is a sum of $n$ terms. So $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$ is in the order of $\mathcal{O}(1/n)$. Hence the variance of goes to 0, i.e., $\widehat{\boldsymbol{\beta}}$ becomes more accurate as we collect more samples.
- However, we still don't know what $\sigma^2$ is. What to do?
  - We have $\widehat{\sigma}^2$
  - After replacing $\sigma^2$ with $\widehat{\sigma}^2$, is it still normal? $z$-test vs. $t$-test?

# Distribution of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ in SLR

- In a SLR, we can explicitly write out the inverse matrix:

$$\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} = \sigma^2 \begin{pmatrix} \frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2} & -\frac{\bar{x}^2}{(n-1)s_x^2} \\ -\frac{\bar{x}^2}{(n-1)s_x^2} & \frac{1}{(n-1)s_x^2} \end{pmatrix}$$

- What is the variance of $\widehat{\beta}_1$?

$$\text{Var}(\widehat{\beta}_1) = \frac{\sigma^2}{(n-1)s_x^2}$$

$$\Rightarrow \quad \frac{\widehat{\beta}_1 - \beta_1}{\sigma/(\sqrt{(n-1)}s_x)} \sim \mathcal{N}(0,1)$$

- Try our simulation study to confirm this.

# Distribution of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ in SLR

- Replace $\sigma^2$ with $\widehat{\sigma}^2 = \frac{\text{SSE}}{n-2}$. Why $n-2$? What's its distribution?

$$\frac{(n-2)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$$

- Then we have

$$\frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_r/r}} \sim t(r)$$

$$\frac{\widehat{\beta}_1 - \beta_1}{\sigma/(\sqrt{(n-1)}s_x)} \Bigg/ \sqrt{\frac{\widehat{\sigma}^2}{\sigma^2}} \sim t(n-2)$$

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}/(\sqrt{(n-1)}s_x)} \sim t(n-2)$$

- Similarly, we have

$$\frac{\widehat{\beta}_0 - \beta_0}{\widehat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}}} \sim t(n-2)$$

# CI for $\widehat{\beta}_0$ and $\widehat{\beta}_1$ in SLR

- Now we know the distributions of both $\widehat{\beta}_0$ and $\widehat{\beta}_1$, we can construct the distributions of them accordingly.
- The $(1-\alpha)100\%$ two-sided confidence intervals for $\widehat{\beta}_0$ and $\widehat{\beta}_1$:

$$\widehat{\beta}_0 \pm t_{\alpha/2,n-2}\,\widehat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{(n-1)s_x^2}}$$

$$\widehat{\beta}_1 \pm t_{\alpha/2,n-2}\,\frac{\widehat{\sigma}}{\sqrt{(n-1)}s_x}$$

- Similarly, $p$-values can be computed.

# Example

- Use the cheddar data to fit SLR with Lactic, and calculate:
  - The variance of both parameter estimates
  - The 95% confidence intervals
  - $p$-value for testing $\mathrm{H}_0 : \beta_1 = 0$ vs. $\mathrm{H}_1 : \beta_1 \neq 0$
- See R code.
- Fit the mode using Acetic and construct a 90% CI for the parameter estimate. Does the interval include 8?

A: Yes; B: No;

# Example

- The R function for calculating the CIs:

```
> fit = lm(taste ~ Lactic, data= cheddar)
> confint(fit, level = 0.95)
                 2.5 %     97.5 %
(Intercept) -51.53573 -8.181935
Lactic       22.99928 52.440613
```

- The $p$-values can be found in

```
> summary(fit)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -29.859     10.582  -2.822  0.00869 **
Lactic        37.720      7.186   5.249 1.41e-05 ***
---
Signif. codes:  0 ?**?0.001 ?*?0.01 ??0.05 ??0.1 ??1
```

- You still need to know how to calculate them by hand if given sufficient information

# Example

- Calculate the $t$-statistic and $p$-value based on the following $\text{lm()}$ fitting results:

```
Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   4.84691     0.37422   12.952    <2e-16 ***
light        -0.10712     0.07419
---
Signif. codes:
0  ***  0.001  **  0.01  *  0.05  .  0.1     1
Residual standard error: 0.2875 on 45 degrees of freedom
Multiple R-squared:  0.04427, Adjusted R-squared:  0.02304
F-statistic: 2.085 on 1 and 45 DF,  p-value: 0.1557
```

- Calculate the sum of squares total, SST.

# Inference about predictions: $\mu_{\text{new}}$

- Suppose we have a new subject with Lactic $= x_{\text{new}}$, can we predict its mean taste score? How accurate is that prediction?
- Let $\mu_{\text{new}} = \beta_0 + \beta_1 x_{\text{new}}$ be the mean taste score of cheddar if the Lactic level is $x$
- The prediction is $\widehat{\mu}_{\text{new}} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{\text{new}}$, i.e.

$$\widehat{\mu}_{\text{new}} = (1, x_{\text{new}}) \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix}$$

  and we already know the distribution of $\widehat{\boldsymbol{\beta}}$. This is yet another linear transformation of MVN variables.

$$\widehat{\mu}_{\text{new}} \sim \mathcal{N}\left(\beta_0 + \beta_1 x_{\text{new}}, \sigma^2 \left(\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{(n-1)s_x^2}\right)\right)$$

- The CI for $\mu_{\text{new}}$ is

$$\widehat{\mu}_{\text{new}} \pm t_{\alpha/2, n-2}\, \widehat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{(n-1)s_x^2}}$$

# Example

- Construct a 90% CI for mean taste with Lactic = 1

```
> xnew <- data.frame(Lactic = 1)
> predict.lm(fit, xnew, interval = c("confidence"), level =
    0.9)
       fit      lwr      upr
1 7.861116 1.341625 14.38061
```

- Construct a 95% CI for mean taste with Lactic = 2

```
> xnew <- data.frame(Lactic = 2)
> predict.lm(fit, xnew, interval = c("confidence"), level =
    0.95)
       fit      lwr      upr
1 45.58106 36.26624 54.89588
```

# Hypothesis test for $\mu_{\text{new}}$

- The data star in the faraway package is trying to model the surface temperature of a star with its light intensity. The stars are in the star cluster CYG OB1, which is in the direction of Cygnus.
- Fit linear regression and test for $x_{\text{new}} = 6$ at 95% confidence level:

$$\text{H}_0 : \mu_{\text{new}} = 4.06 \quad \text{vs.} \quad \text{H}_1 : \mu_{\text{new}} > 4.06$$

- Do we reject the hypothesis?

A: Yes; B: No;

# Inference about predictions: $Y_{\text{new}}$

- What if we want to predict the observed value $Y_{\text{new}}$ for a new subject with $x_{\text{new}}$, there will be additional variations.
- Since $\widehat{Y}_{\text{new}} = \widehat{\mu}_{\text{new}} + \epsilon_{\text{new}}$, this is only adding additional $\sigma^2$ to the variance component. Also $\epsilon_{\text{new}}$ is independent of everything else.
- Hence the CI for a future observed value $Y_{\text{new}}$

$$\widehat{\mu}_{\text{new}} \pm t_{\alpha/2, n-2}\, \widehat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{(n-1)s_x^2}}$$

# Example

- Construct a 90% CI for future taste score with Lactic = 2

```
> xnew <- data.frame(Lactic = 2)
> predict.lm(fit, xnew, interval = c("prediction"), level =
    0.95)
       fit      lwr      upr
1 45.58106 19.78216 71.37997
```

# Example

- Suppose the following data are observed.

```
> x = c(-1, -1, -1, -1, 1, 1, 1, 1)
> y = c(1.3, 0.8, 1.2, 0.6, 2.3, 2.5, 1.8, 1.6)
```

- Construct a 90% confidence interval for the $\widehat{\beta}_1$ estimation.
- Predict a future outcome value at $x = 0$.

# The Gauss-Markov Theorem

- We know that $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{Y}$ and derived its properties. However, is this the best estimator? Is it possible that we can have some other estimators that are also unbiased and more accurate (smaller variance)?
- Turns out that this is the best we can do if $\widehat{\boldsymbol{\beta}}$ is unbiased and also a “linear estimator”, i.e., $\widehat{\boldsymbol{\beta}} = \mathbf{C}\mathbf{y}$ for some matrix $\mathbf{C}$.
- This is guaranteed by the Gauss-Markov Theorem.
- $\widehat{\boldsymbol{\beta}}$ is the BLUE (best linear unbiased estimate).