Stat 206: Linear Models

Lecture 4

October 7, 2019

ReCap: Sampling Distributions of LS Estimators

Under the Normal error model:

- $\hat{\beta}_0, \hat{\beta}_1$ are normally distributed:

$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2\{\hat{\beta}_0\}), \quad \hat{\beta}_1 \sim N(\beta_1, \sigma^2\{\hat{\beta}_1\}).$$

- SSE/σ^2 follows a χ^2 distribution with $n-2$ degrees of freedom, denoted by $\chi^2_{(n-2)}$.
- Moreover, SSE is independent with both $\hat{\beta}_0$ and $\hat{\beta}_1$ (because residuals e_i's are independent with $\hat{\beta}_0$ and $\hat{\beta}_1$).

Recap: Confidence Interval

$(1-\alpha)$-Confidence interval of β_1:

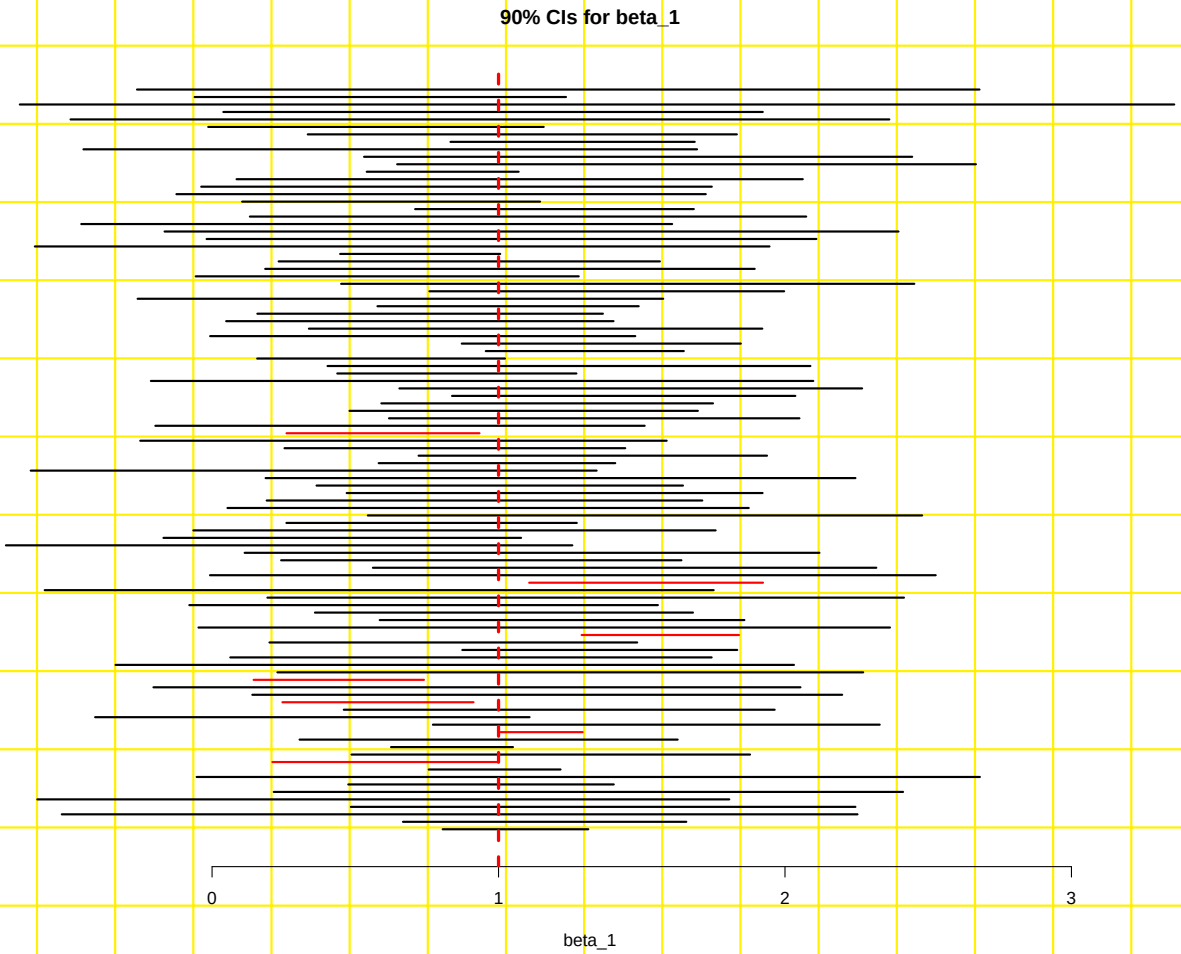
$$\hat{\beta}_1 \pm t(1-\alpha/2; n-2)s\{\hat{\beta}_1\},$$

where $t(1-\alpha/2; n-2)$ is the $(1-\alpha/2)$th percentile of $t_{(n-2)}$.

How to construct confidence intervals for β_0?

Interpretation of Confidence Intervals

Figure: A Simulation Study



Heights

- Recall $n = 928$, $\overline{X} = 68.316$, $\sum_i X_i^2 = 4334058$, $\sum_{i=1}^n (X_i - \overline{X})^2 = \sum_{i=1}^n X_i^2 - n(\overline{X})^2 = 3038.761$. Also

$$\hat{\beta}_0 = 24.54, \quad \hat{\beta}_1 = 0.637, \quad MSE = 5.031.$$

So

$$s\{\hat{\beta}_1\} = \sqrt{\frac{5.031}{3038.761}} = 0.0407.$$

- 95%-confidence interval of β_1:

$$\begin{aligned} 0.637 \pm t(0.975; 926) \times 0.0407 &= 0.637 \pm 1.963 \times 0.0407 \\ &= [0.557, 0.717]. \end{aligned}$$

We are 95% confident that the regression slope is in between 0.557 and 0.717.

T-tests

- Null hypothesis: $H_0 : \beta_1 = \beta_1^{(0)}$, where $\beta_1^{(0)}$ is a given constant.
- T-statistic:

$$T^* = \frac{\hat{\beta}_1 - \beta_1^{(0)}}{s\{\hat{\beta}_1\}}.$$

- **Null distribution** of the T-statistic:

$$\text{Under } H_0 : \beta_1 = \beta_1^{(0)}, \quad T^* \sim t_{(n-2)}.$$

Can you derive the null distribution?

Decision rule at significance level α.

- **Two-sided alternative** $H_a : \beta_1 \neq \beta_1^{(0)}$: Reject H_0 if and only if $|T^*| > t(1-\alpha/\mathbf{2}; n-2)$, or equivalently, reject H_0 if and only if pvalue:$= P(|t_{(n-2)}| > |T^*|) < \alpha$.
- **Left-sided alternative** $H_a : \beta_1 < \beta_1^{(0)}$: Reject H_0 if and only if $T^* < t(\alpha; n-2)$, or equivalently, reject H_0 if and only if pvalue:$= P(t_{(n-2)} < T^*) < \alpha$.
- **Right-sided alternative** $H_a : \beta_1 > \beta_1^{(0)}$: Reject H_0 if and only if $T^* > t(1-\alpha; n-2)$, or equivalently, reject H_0 if and only if pvalue:$= P(t_{(n-2)} > T^*) < \alpha$.

The decision rule depends on the form of the alternative hypothesis

Why are the critical value approach and the pvalue approach equivalent? How to conduct hypothesis testing with regard to β_0?

Heights

Test whether there is a linear association between parent's height and child's height. Use significance level $\alpha = 0.01$.

- The hypotheses: $H_0 : \beta_1 = 0$ *vs.* $H_a : \beta_1 \neq 0$.
- T statistic: $T^* = \frac{\hat{\beta}_1 - 0}{s\{\hat{\beta}_1\}} = \frac{0.637}{0.0407} = 15.7$.
- Critical value: $t(1 - 0.01/2; 928 - 2) = 2.58$. Since the observed $T^* = |15.7| > 2.58$, reject the null hypothesis at level 0.01.
- Or the pvalue $= P(|t_{(926)}| > |15.7|) \approx 0$. Since $pvalue < \alpha = 0.01$, reject the null hypothesis at level 0.01.
- Conclude that there is a significant association between parent's height and child's height.

Estimation of Mean Response

Given $X = X_h$, the mean response is $E(Y_h) = \beta_0 + \beta_1 X_h$.

- An unbiased point estimator for $E(Y_h)$ is :

$$\widehat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h = \overline{Y} + \hat{\beta}_1 (X_h - \overline{X}).$$

$$E(\widehat{Y}_h) = \beta_0 + \beta_1 X_h = E(Y_h).$$

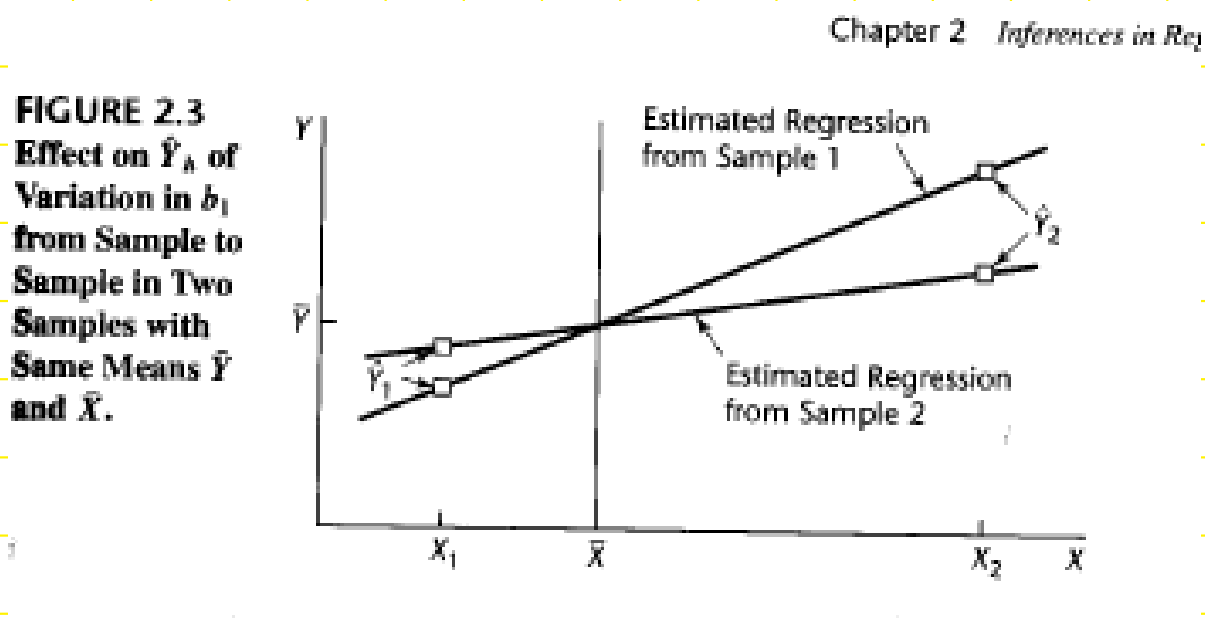
- Variance of $\widehat{Y}_h$ is:

$$\sigma^2\{\widehat{Y}_h\} = \sigma^2 \left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \right].$$

Notes: Use the fact that $\overline{Y}$ and $\hat{\beta}_1$ are uncorrelated.

- The larger the sample size and/or the larger the dispersion of X_is, the smaller
the variance of $\widehat{Y}_h$.

Figure: Effects of the distance of X_h from $\overline{X}$ on variability of $\widehat{Y}_h$.



From Applied Linear Statistical Models by Kutner, Nachtsheim, Neter and Li

The further is X_h from $\overline{X}$, the larger
is the variance of $\widehat{Y}_h$: The variability in the estimated slope $\hat{\beta}_1$ has a larger
effect on $\widehat{Y}_h$ when X_h is further away from the sample mean $\overline{X}$.

- Standard error of $\widehat{Y}_h$:

$$s\{\widehat{Y}_h\} = \sqrt{MSE\left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]}.$$

- Under the Normal error model, $\widehat{Y}_h$ is normally distributed.
  - Studentized quantity:

$$\frac{\widehat{Y}_h - E(Y_h)}{s(\widehat{Y}_h)} \sim t_{(n-2)}.$$

  - $(1-\alpha)$- C.I.

$$\widehat{Y}_h \pm t(1-\alpha/2; n-2)s(\widehat{Y}_h).$$

  - The half-width of $(1-\alpha)$- C.I., $t(1-\alpha/2; n-2)s(\widehat{Y}_h)$, increases with the confidence coefficient $(1-\alpha)$ and the standard error $s(\widehat{Y}_h)$.

Heights

Estimate the average height of children of $70in$ tall parents.

- Recall: $n = 928$, $\overline{X} = 68.316$, $\sum_{i=1}^{n}(X_i - \overline{X})^2 = 3038.761$, $\widehat{E(Y)} = 24.54 + 0.637X$ and $MSE = 5.031$.
- $\widehat{Y}_h = 24.54 + 0.637 \times 70 = 69.2$.
- Standard error:

$$s\{\widehat{Y}_h\} = \sqrt{5.031 \times \left\{\frac{1}{928} + \frac{(70 - 68.316)^2}{3038.761}\right\}} = 0.1.$$

- 95%-confidence interval of $E(Y_h)$:

$$69.2 \pm 1.8831 \times 0.1 = [68.96, 69.35], \quad t(0.975; 926) = 1.8831.$$

- We are 95% confident that the average height of children of $70in$ parents is between $[68.96in, 69.35in]$.

Prediction of New Observation

Predict a **future observation** $Y_{h(new)}$ of the response variable corresponding to a given level of the predictor variable $X = X_h$.

- $Y_{h(new)} = \beta_0 + \beta_1 X_h + \epsilon_h$.
  - This is a new observation, so ϵ_h is assumed to be uncorrelated with ϵ_is.
  - Consequently, $Y_{h(new)}$ is uncorrelated with the observed Y_is.
- The **predicted value** for $Y_{h(new)}$ is simply the estimated mean response when $X = X_h$:

$$\widehat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h = \overline{Y} + \hat{\beta}_1 (X_h - \overline{X}).$$

Distinction between prediction and mean estimation.

- $Y_{h(new)}$ is a "moving target" as it is a random variable. On the contrary, $E(Y_h)$ is a fixed non-random quantity.
- There are two sources of variations in the prediction process: Variability from the predicted value $\widehat{Y}_h$ and variability from the target $Y_{h(new)}$.

$$\sigma^2(pred_h) := Var(\widehat{Y}_h - Y_{h(new)}) = \sigma^2(\widehat{Y}_h) + \sigma^2(Y_{h(new)}) = \sigma^2(\widehat{Y}_h) + \sigma^2.$$

$$s^2\{pred_h\} = s^2(\widehat{Y}_h) + MSE = MSE\left[1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]. \quad (1)$$

Note the difference between $s^2\{\widehat{Y}_h\}$ and $s^2\{pred_h\}$.

Prediction Intervals

- Studentized quantity:

$$\frac{\widehat{Y}_h - Y_{h(new)}}{s(pred_h)}.$$

- Under the Normal error model, it follows a $t_{(n-2)}$ distribution.
- $(1-\alpha)$– prediction interval of $Y_{h(new)}$:

$$\widehat{Y}_h \pm t(1-\alpha/2; n-2)s(pred_h).$$

- Prediction interval is wider
  than the corresponding confidence interval of the mean response.
- With sample size becomes very large, the width of the confidence interval tend to vanish,
  but this would not happen for the prediction interval.

Heights

What would be the predicted height of the child of a 70*in* tall couple?

- Predicted height: $\widehat{Y}_h = 24.54 + 0.637 \times 70 = 69.2$. Standard error:

$$s\{pred_h\} = \sqrt{5.031 \times \left\{\mathbf{1} + \frac{1}{928} + \frac{(70 - 68.316)^2}{3038.761}\right\}} = 2.25.$$

- 95% prediction interval:

$$69.2 \pm 1.8831 \times 2.25 = [64.75, 73.56], \quad t(0.975; 926) = 1.8831.$$

- We are 95% confident that the child's height will be in between $[64.75in, 73.56in]$.

Extrapolation

Extrapolation occurs when predicting the response variable for values of the predictor variable lying outside of the range of the observed data.

- Every model has a range of validity. Particularly, a model may be inappropriate when it is extended outside of the range of the observations upon which it was built.
- Extrapolations are often much less reliable than interpolation and need to be handled with caution, even though they can be of more interests to us (e.g. fortune telling).
- In the Heights example: Extrapolation would happen if we use the fitted regression line to predict heights of children of very tall or very short parents.