Statistics 203: Introduction to Regression and Analysis of Variance

Simple Linear Regression: Inference + Diagnostics

Jonathan Taylor

Outline

- Inference for vector of coeffi cients β.
- Diagnostics: what can go wrong in our model?

Distribution of $\widehat{\beta}, e$

- The vector $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1)$ is a function of $\widehat{Y}$ so is independent of e.
- Both $\widehat{\beta}$ and $\widehat{Y}$ are linear transformations of Y so they are normally distributed.
- We will prove

$$\mathbb{E}((\widehat{\beta}_0, \widehat{\beta}_1)) = (\beta_0, \beta_1)$$

and has covariance matrix

$$\text{Var}(\widehat{\beta}) = \begin{pmatrix} \frac{\sigma^2}{n} + \sigma^2 \frac{\overline{X}^2}{S_{xx}} & -\sigma^2 \frac{\overline{X}}{S_{xx}} \\ -\sigma^2 \frac{\overline{X}}{S_{xx}} & \frac{\sigma^2}{S_{xx}} \end{pmatrix}$$

- Natural estimates of covariance matrix

$$\widehat{\text{Var}}(\widehat{\beta}) = \begin{pmatrix} \frac{\widehat{\sigma}^2}{n} + \widehat{\sigma}^2 \frac{\overline{X}^2}{S_{xx}} & -\widehat{\sigma}^2 \frac{\overline{X}}{S_{xx}} \\ -\widehat{\sigma}^2 \frac{\overline{X}}{S_{xx}} & \frac{\widehat{\sigma}^2}{S_{xx}} \end{pmatrix}$$

t-random variables

- Start with $Z \sim N(0,1)$ is standard normal and $G \sim \chi^2_\nu$, independent of Z.
- Compute

$$T = \frac{Z}{\sqrt{\frac{G}{\nu}}}.$$

- Then $T \sim t_\nu$ has a t-distribution with ν degrees of freedom.
- Where do they come up in regression?

F-random variables

- Start with $G_1 \sim \chi^2_{\nu_1}$ and another *independent* $G_2 \sim \chi^2_{\nu_2}$
- Compute

$$F = \frac{G_1/\nu_1}{G_2/\nu_2}$$

- Then $F \sim F_{\nu_1,\nu_2}$ has an F-distribution with ν_1 degrees of freedom in the numerator in ν_2 in the denominator.
- Note: if $T \sim t_\nu$ than $T^2 \sim F_{1,\nu}$.
- Where do they come up in regression?

Inference for $\widehat{\beta}$: t-statistics

- Because e is independent of $\widehat{\beta}$ it follows that $\widehat{\mathrm{Var}}(\widehat{\beta}_1)$ and $\widehat{\mathrm{Var}}(\widehat{\beta}_0)$ are independent of $\widehat{\beta}$.
- Under the hypothesis $H_0 : \beta_1 = \beta_1^0$

$$T = \frac{\widehat{\beta}_1 - \beta_1^0}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_1)}} \sim t_{n-2}.$$

(Why?)

- To test this hypothesis, compare $|T|$ to $t_{n-2,1-\alpha/2}$ the $1-\alpha/2$ quantile of the t distribution with $n-2$ degrees of freedom.
- Reject H_0 if $|T| > t_{n-2,1-\alpha/2}$.

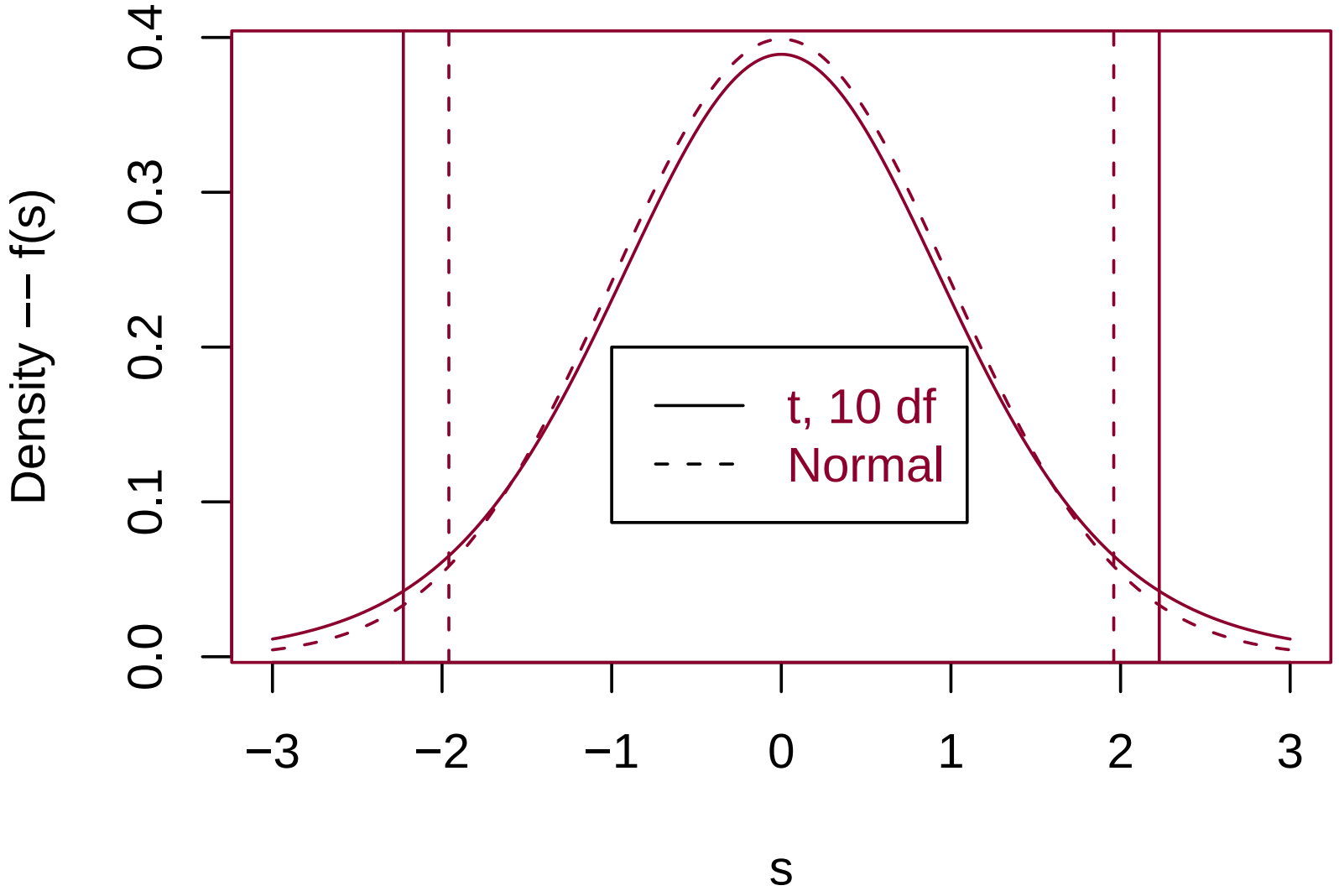
Why reject for large $|T|$?

- Observing a large $|T|$ is unlikely if $\beta_1 = \beta_1^0$: reasonable to conclude that H_0 is false.
- Common to report p-value

$$p - \text{value} = 2 \times \int_{|T|}^{\infty} f_{t_{n-2}}(s)\, ds.$$

- Above, $f_{t_{n-2}}$ is the density of a t- random variable with $n-2$ degrees of freedom.

t vs. Normal



Confidence interval for β_1

- For simplicity, write

$$SE(\widehat{\beta}_1) = \widehat{\sigma}\sqrt{\frac{1}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}.$$

- Under the model assumptions

$$1 - \alpha = \mathbb{P}\left(\left|\frac{\widehat{\beta}_1 - \beta_1}{SE(\widehat{\beta}_1)}\right| < t_{n-2,1-\alpha/2}\right)$$

$$= \mathbb{P}\left(\beta_1 \in \widehat{\beta}_1 \pm t_{n-2,1-\alpha/2} \cdot SE(\widehat{\beta}_1)\right)$$

Linear combinations of β_0, β_1

- It is not too hard to prove that $a_0\widehat{\beta}_0 + a_1\widehat{\beta}_1$ is normally distributed and its standard deviation can be estimated by

$$SE(a_0\widehat{\beta}_0 + a_1\widehat{\beta}_1) = \widehat{\sigma}\sqrt{\frac{a_0^2}{n} + \frac{(a_0\overline{X} - a_1)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}}$$

- As in last slide, confi dence interval is

$$a_0\widehat{\beta}_0 + a_1\widehat{\beta}_1 \pm t_{n-2,1-\alpha/2} \cdot SE(a_0\widehat{\beta}_0 + a_1\widehat{\beta}_1)$$

A new observation: forecasting

- New observation

$$Y_{\text{new}} = \beta_0 + \beta_1 X_{\text{new}} + \varepsilon_{\text{new}}.$$

-

$$SE(Y_{\text{new}}) = \widehat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(\overline{X} - X_{\text{new}})^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}}.$$

- Again, "prediction" interval is

$$\widehat{\beta}_0 + \widehat{\beta}_1 X_{\text{new}} \pm t_{n-2,1-\alpha/2} \cdot SE(Y_{\text{new}})$$

Goodness of fit

The variation in Y, SST, can be decomposed into two parts: one for the regression, SSR, and one for the error, SSE.

$$SST = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = SSE + SSR$$

$$SSE = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2$$

$$SSR = \sum_{i=1}^{n}(\overline{Y} - \widehat{Y}_i)^2 = \sum_{i=1}^{n}(\overline{Y} - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2$$

$$SST = SSR + SSE$$

Goodness of fit

- $$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \widehat{Cor}(X, Y)^2.$$
- R^2 tells us how much variability in the Y's is explained by the regression.

F test for "significance" of regression

- Under $H_0 : \beta_1 = 0$:

$$SSR \sim \sigma^2 \cdot \chi^2_1$$
$$SSE \sim \sigma^2 \chi^2_{n-2}$$

- Therefore

$$F = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/(n-2)} \sim F_{1,n-2}.$$

(Why?)

- Reject H_0 for large values of F.
- General form of the F: a ratio of "dispersion": numerator is the dispersion of $\hat{Y}$ around $\overline{Y}$ while denominator is disperion of e.

What can go wrong?

- Regression function can be wrong – missing predictors, nonlinear.
- Assumptions about the errors can be wrong.
- Outliers: both in predictors and observations.
- Influential points: these points have "undue" influence on the regression function.
- Examples:
  - Example #1: diagnostics for usual linear model
  - Example #2: t density
  - Example #3: misspecifi ed model

Problems in the regression function

- True regression function may have higher-order non-linear terms i.e. X_1^2, or may truly be non-linear.
- How to fi x? Sometimes things can be transformed to linearity: suppose

$$Y_i = \beta_0 e^{\beta_1 X_i} \cdot \varepsilon_i.$$

Then

$$\log Y_i = \log \beta_0 + \beta_1 X_i + \log \varepsilon_i$$

is a linear model and if ε's are independent lognormal random variables, then this transformed model has the same form as the original model!

- Later, we will see Box-Cox transformations to "choose" a transformation that "optimally" linearizes the model.

Problems with the errors

- Errors may not be normally distributed. We will look at QQplot for a graphical check. May not effect inference in large samples.
- Variance may not be constant. We will see some graphical checks of this and (later) some transformations that might help correct this.
- Errors may not be independent. This seriously affects our estimates of SE which can change t and F statistics substantially!