STAT 302: Linear Regression

Sheridan Grant

University of Washington Statistics Department

slgstats@uw.edu

April 15, 2020

Transformations of Variables

Linear models are much more comprehensive than you might guess. Suppose you thought the X, Y relationship was quadratic, i.e.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$$

Then just define a new covariate, X^2, by considering the squares of the X_i!

Transformations of Variables

Linear models are much more comprehensive than you might guess. Suppose you thought the X, Y relationship was quadratic, i.e.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$$

Then just define a new covariate, X^2, by considering the squares of the X_i! Interpretation: an increase of SAT from x_0 to x_1 points is associated with an expected increase in GPA of

$$[\beta_0 + \beta_1 x_1 + \beta_2 x_1^2] - [\beta_0 + \beta_1 x_0 + \beta_2 x_0^2] = \beta_1 [x_1 - x_0] + \beta_2 (x_1^2 - x_0^2)$$

Transformations of Variables

You can transform the response variable too:

$$\log(Y) = \beta_0 + \beta_1 X + \epsilon$$

Transformations of Variables

You can transform the response variable too:

$$\log(Y) = \beta_0 + \beta_1 X + \epsilon$$

This cannot be interpreted on the Y scale, however, because

$$\begin{aligned} Y &= \exp(\beta_0 + \beta_1 X + \epsilon) \\ &= \exp(\beta_0) \exp(\beta_1) X \exp(\epsilon) \end{aligned}$$

is not a linear model: it is not additive, but multiplicative!
Such models are common in, e.g., finance, because investments can be expected to grow exponentially in the long run, but noisily. When else?

Transformations of Variables

If we suspect that the linear relationship between X and Y differs based on a second covariate Z, we can fit an *interaction* term:

$$Y = \beta_0 + \beta_X X + \beta_Z Z + \beta_{XZ} XZ + \epsilon$$

Transformations of Variables

If we suspect that the linear relationship between X and Y differs based on a second covariate Z, we can fit an *interaction* term:

$$Y = \beta_0 + \beta_X X + \beta_Z Z + \beta_{XZ} XZ + \epsilon$$

If Y is College GPA, X is HS GPA, and Z is sex ($0 ==$ male, $1 ==$ female), then the linear relationship between HS and College GPA can differ by sex:

$$\begin{aligned} \text{(men): } Y &= \beta_0 + \beta_X X + \beta_Z Z \\ \text{(women): } Y &= \beta_0 + (\beta_X + \beta_{XZ}) X + \beta_Z Z \end{aligned}$$

Inference

How certain are we about $\hat{Y}_i$ for any given i? How certain are we that a change of a unit in X leads to a change of $\hat{\beta}$ in Y on average? There are 2 facts we need in the univariate case:

1. Regression coefficients obey a CLT, just like the sample mean (regression with just β_0 and no covariates is computing the sample mean). There are assumptions...
2. $\text{sd}(\hat{\beta}x) = xsd(\hat{\beta})$

So if the S.E. estimate for $\hat{\beta}$ is $\hat{\sigma}$, then

1. a 95% confidence interval for $\hat{\beta}$ is $[\hat{\beta} - 1.96\hat{\sigma}, \hat{\beta} + 1.96\hat{\sigma}]$
2. a 95% confidence interval for the expected change in Y associated with a change in X of a units is $[a\hat{\beta} - 1.96a\hat{\sigma}, a\hat{\beta} + 1.96a\hat{\sigma}]$

Multivariate inference

To get inference for $\hat{Y}_i$ or complex quantities in multivariate models, we need one more concept:

Definition

Variance-covariance matrix Let $\hat{\beta}$ be a random vector (a vector of random variables). Then

$$\mathrm{Var}(\hat{\beta}) = \hat{\Sigma} = \begin{bmatrix} \hat{\mathrm{Var}}(\hat{\beta}_0) & \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \cdots & \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_d) \\ \mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_0) & \mathrm{Var}(\hat{\beta}_1) & \cdots & \vdots \\ \vdots & & \ddots & \\ \mathrm{Cov}(\hat{\beta}_d, \hat{\beta}_0) & \cdots & & \mathrm{Var}(\hat{\beta}_d) \end{bmatrix}$$

All you need to know is: if $a \in \mathbb{R}^{d+1}$, then $\hat{\mathrm{Var}}(a^T \hat{\beta}) = a^T \hat{\Sigma} a$. And if `lmod` is a linear model object, then `vcov(lmod)` gives you $\hat{\Sigma}$.