18.1. Models

"All models are wrong, but some are useful."
George E. P. Box (1979)

- All models are "wrong" in the sense of being a simplified representations of some reality.
- Some models are more wrong than others.
  - Let Y be the length of a string with a weight attached to the end. Let w be the weight of the weight.
    $E[Y|w] = \beta_0 + \beta_1 w$ is a pretty accurate model.
  - Contrast this with a typical model used in biostatistics or epidemiology. It's extraordinarily rare to have this kind of mechanistic justification for a model.
- "But some are useful." Linear models can be extremely useful for quantifying the association between variables, even though we don't believe the model.
- It's extremely important to interpret a model appropriately and avoid over-interpreting a model.

18.2. Example #1

Q: Does smoking affect lung function in children who smoke?

- Lung function (Y):
  - Forced expiratory volume in one second (FEV_1)
    - FEV_1 is the volume exhaled during the first second of a forced expiratory maneuver started from the level of total lung capacity.
    - Here, we use $Y = log(FEV_1)$ is a surrogate for "lung function."
    - Higher $log(FEV_1)$ denotes better lung function.
- Smoking (x):
  - $x = 1$ for children who say they smoke.
  - $x = 0$ for children who say they do not smoke.
- Age (w):
  - Age range=(8,18) years.

Suppose one models $log(FEV_1) = Y$ versus smoking status (x) for a random sample of children and obtains the following results (with estimated standard errors):

$$\hat{E}[Y \mid x] = \underbrace{1.06}_{(0.013)} + \underbrace{0.102}_{(0.033)} x.$$

Q: What is the interpretation of the estimated coefficient for smoking status (x)?

Suppose one instead models Y versus x and age (w) and obtains the following results:

$$\hat{E}[Y \mid x, w] = \underbrace{0.352}_{(0.054)} - \underbrace{0.05}_{(0.030)} x + \underbrace{0.063}_{(0.005)} w.$$

Q: What is the interpretation of the estimated coefficient for smoking status now?

Q: Can you explain such diametrically opposed results for the two models?

18.3 Example #2

Q: Is there sexual discrimination with regard to faculty salaries at the University of Washington?

- 1995 Salary ($Y = log(salary)$):
- Gender (x):
  - $x = 1$ male,
  - $x = 0$ female.
- Rank (w):
  - $x = 1$ full professor,
  - $x = 0$ assistant or associate professor.
- Confounders (skill, experience, productivity, training) ($\mathbf{c}$):

1. Fit model 1,

$$\hat{E}[Y \mid x] = \hat{\beta}_0 + \hat{\beta}_1 x$$

Q: What is the interpretation of $\hat{\beta}_1$?

2. Fit model 1 + confounders, $\mathbf{c}$,

$$\hat{E}[Y \mid x, \mathbf{c}] = \hat{\vartheta}_0 + \hat{\vartheta}_1 x + \hat{\boldsymbol{\vartheta}}_2' \mathbf{c}$$

Q: What is the interpretation of $\hat{\vartheta}_1$?

3. Fit model 2 + rank $\mathbf{w}$,

$$\hat{E}[Y \mid x, w] = \hat{\gamma}_0 + \hat{\gamma}_1 x + \hat{\gamma}_2 w + \hat{\boldsymbol{\gamma}}_3' \mathbf{c}$$

Q: Which model do you prefer for investigating sexual discrimination in salary?

18.4 Linear Models in Biostatistics

Consider the classical linear model assumptions and a typical biostatistics or epidemiological data set. It's generally a good thing to accurately model the mean: $E[Y] = \mathbf{X}\boldsymbol{\beta}$. However, we do not believe the model in the same way as we believe, for example, the model for the length of a string holding a weight. We commonly use models to measure an association of interest, while accounting for confounding variables and other important covariates (precision variables).

Thought experiment: Suppose I could fit my model to the entire population. Then I could know $\boldsymbol{\beta}$ for that population. However, I can only sample from the population, and so I get an estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$.

- $E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$ is a convenient way to measure the association I'm interested in, but I don't believe the model.

- Since I have a random sample, I am comfortable assuming my errors are uncorrelated but not necessarily homoscedastic (equal variance). $\text{cov}(\mathbf{Y}) = \text{cov}(\boldsymbol{\varepsilon}) = \mathbf{D}$, for some diagonal matrix $\mathbf{D}$

- I don't believe the normality assumption but if my sample is large I don't worry about this part.

We need methods for inference about $\boldsymbol{\beta}$ that are robust to the linear model assumptions that I don't believe.

Some possibilities:

- sandwich estimator
- bootstrapping
- permutation test

18.5 Sandwich Estimator (BIOST 570)

From the formula for $\hat{\boldsymbol{\beta}}$ we can calculate $\text{cov}(\hat{\boldsymbol{\beta}})$:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ \text{cov}(\hat{\boldsymbol{\beta}}) &= \text{cov}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{cov}(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

If we assume $\text{cov}(\mathbf{Y}) = \sigma^2\mathbf{I}$ things simplify here a lot. But without making this assumption we are left with

$$\text{cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

The idea of the sandwich estimator is to estimate $\text{cov}(\hat{\boldsymbol{\beta}})$ by estimating $\mathbf{D}$.

18.6 Bootstrapping

IDEA: We have an estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. The question is: What is the sampling distribution of $\hat{\boldsymbol{\beta}}$?

Conceptually, we could repeat our random sampling of n individuals from the population J times and fit the model each time. Then we would get $(\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_J)$ and we could use this to estimate the sampling distribution of $\hat{\boldsymbol{\beta}}$.

But we don't have the opportunity to do this.

If our sample is large enough, we pretend the sample is our population in bootstrapping. Sample *with replacement* from the subjects in our sample, taking a sample of size n each time.

This gives $(\hat{\boldsymbol{\beta}}_1^*, \ldots, \hat{\boldsymbol{\beta}}_J^*)$. Use this empirical distribution to estimate the sampling distribution of $\hat{\boldsymbol{\beta}}$.

Why is the sampling done with replacement?

1. theoretical reason: We need independent sampling.

2. practical reason: Otherwise we would just get the same sample every time.

18.7 Permutation test

Suppose we're using a model to query a possible association between predictors $\mathbf{X}$ and a response $\mathbf{Y}$.

Thought experiment: if there is no association between my response and my predictors, then the data I observe was just as likely as a dataset where the $y's$ are scrambled (permuted).

NOTE: Permutation testing is a testing method, confidence intervals are hard to come by.

ISSUE: What is the null hypothesis being tested? Is it the null hypothesis of interest?

18.8 Linear regression with heteroscedasticity

Consider simple linear regression $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. $\text{cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$ and $\text{var}(\epsilon_i) \equiv \sigma_i^2 = \alpha_i + \gamma x_i$. γ and the α_i are unknown nuisance parameters. Without loss of generality, assume $\sum_i x_i = 0$. Also assume $\boldsymbol{\alpha}'\mathbf{x} = 0$ (note this covers the case $\alpha_i = \alpha$). Let $\hat{\boldsymbol{\beta}}$ be the ordinary least squares estimate of $\boldsymbol{\beta} = (\beta_0, \beta_1)'$.

We have seen that $\hat{\boldsymbol{\beta}}$ is unbiased. For the variance of $\hat{\boldsymbol{\beta}}$ we need to calculate $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$

Notation: Let $S_{xx} = \sum_i x_i^2$ and $S_{xy} = \sum_i x_i Y_i$

Then $\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & 0 \\ 0 & S_{xx} \end{pmatrix}$, $(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{S_{xx}} \end{pmatrix}$, and $\mathbf{X}'\mathbf{Y} = \begin{pmatrix} n\bar{y} \\ S_{xy} \end{pmatrix}$. We need one more piece: $\mathbf{X}'\mathbf{V}\mathbf{X} =$

So we have $\text{var}(\hat{\boldsymbol{\beta}}) =$

$$\begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{S_{xx}} \end{pmatrix} \begin{pmatrix} \sum \alpha_i & \gamma \sum x_i^2 \\ \gamma \sum x_i^2 & \sum \alpha_i x_i^2 + \gamma \sum_i x_i^3 \end{pmatrix} \begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{S_{xx}} \end{pmatrix}$$

$$= \begin{pmatrix} \sum \alpha_i / n^2 & \gamma/n \\ \gamma/n & (\sum \alpha_i x_i^2 + \gamma \sum_i x_i^3)/S_{xx}^2 \end{pmatrix}$$

In particular, $\text{var}(\hat{\beta}_1) = (\sum \alpha_i x_i^2 + \gamma \sum_i x_i^3)/S_{xx}^2$

Now, suppose we assume the errors have constant variance, calculate the ordinary least squares $\hat{\boldsymbol{\beta}}$, and proceed to estimate

$$\hat{\sigma}^2 = \frac{1}{n-2}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

This is a consistent estimator of the average variance

$$\lim_{n\to\infty} (\sigma_i^2)/n = \lim_{n\to\infty} \frac{\sum_{i=1}^n (\alpha_i + \gamma x_i)}{n} = \bar{\alpha}$$

(assuming such a limit exists).

So if we make the false assumption of a common variance, the estimated variance of $\hat{\boldsymbol{\beta}}$ that will be used is:

$$\hat{\text{var}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1} \approx \bar{\alpha} \begin{pmatrix} 1/n^2 & 0 \\ 0 & 1/S_{xx} \end{pmatrix}$$

and in particular $\hat{\text{var}}(\hat{\beta}_1) = \frac{\bar{\alpha}}{S_{xx}}$.

In truth, the variance of $\hat{\beta}_1$ is $(\sum \alpha_i x_i^2 + \gamma \sum_i x_i^3)/S_{xx}^2$.

When will we be using the right variance in making inference about β_1?

We need the following condition to hold:

$$\frac{\bar{\alpha}}{S_{xx}} = \frac{(\sum \alpha_i x_i^2 + \gamma \sum_i x_i^3)}{S_{xx}^2},$$

i.e.,

$$\bar{\alpha} = \frac{(\sum \alpha_i x_i^2 + \gamma \sum_i x_i^3)}{S_{xx}}.$$

This will be true if the α_i's are constant and either $\gamma = 0$ or $\sum_i x_i^3 = 0$.

If the α_i's are constant and $\gamma = 0$ we have constant variance and we're not actually making any erroneous assumption.

If the α_i's are constant and $\sum_i x_i^3 = 0$ we have $x_i's$ with 0 skewness. That is, we can get away with the OLS estimate even with this kind of heteroscedasticity if the $x_i's$ are symmetric.

The requirement for the $x_i's$ to be unskewed is analogous to the balanced samples size requirement we saw in the two-sample t-test example in Lecture 17.

What are the implications for a permutation test of the significance of $\hat{\beta}_1$?

If we permute the $Y_i's$ randomly with respect to the $x_i's$, then $\text{var}(Y_i^*|x_i) = \text{var}(Y_i^*)$ is just the marginal variance of the $Y_i's$: $\bar{\alpha} + \gamma\bar{x} = \bar{\alpha}$.

By permuting the $Y_i's$, we "break" the mean variance relationship.

So $\text{var}(\hat{\beta}_1^*) = \frac{\bar{\alpha}}{S_{xx}}$

Therefore, even if the true $\beta_1 = 0$, the permutation test will "have power" if the variance of the $\hat{\beta}_1^*$ is less than the variance of $\hat{\beta}_1$.:

$$\frac{\bar{\alpha}}{S_{xx}} < \frac{\sum \alpha_i x_i^2 + \gamma \sum_i x_i^3}{S_{xx}^2},$$

equivalently

$$\bar{\alpha} < \frac{\sum \alpha_i x_i^2 + \gamma \sum_i x_i^3}{S_{xx}}.$$

If $\alpha_i = \bar{\alpha}$ this condition becomes:

$$\bar{\alpha} < \frac{(\bar{\alpha} \sum x_i^2 + \gamma \sum_i x_i^3)}{S_{xx}} = \bar{\alpha} + \frac{\gamma \sum_i x_i^3}{S_{xx}}, .$$

or $0 < \gamma \sum x_i^3$. Therefore, if γ and $\sum x_i^3$ have the same sign, then the variance of the $\hat{\beta}_1^*$ is less than the variance of $\hat{\beta}_1$