5. Measuring the Strength of a Relationship

   a. Recall that the $F$ test tested whether any of the regression parameters are non-zero

   b. Recall that for one parameter, $R^2$ measured the strength rather than tested for the strength.

   c. Do this for more explanatory variables: $R^2 = SS_R / SS_t = 1 - SS_{Res} / SS_t$ .

      i. $SS_{Res} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

      ii. $SS_R = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$

      iii. $SS_t = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$

   d. $R^2 \in [0,1]$ , represents the proportion of variablity explained by explanatory variables.

   e. Any additional variable will result in $R^2$ no smaller.

      i. Almost all will result in $R^2$ somewhat larger, even if variable is probabilistically unrelated.

      ii. Adjust $R^2$ to penalize it for larger numbers of parameters:

         - $R^2_{Adj} = 1 - (SS_{Res}/(n-p))/(SS_t/(n-1)) = 1 - (SS_{Res}/SS_t)((n-1)/(n-p))$ .
         - $((n-1)/(n-p)) \approx 1$ if $n$ large and $p$ small.

- $R^2_{Adj}$ drops if $SS_{Res}$ fixed and $p$ increases.

6. Confidence Regions for Parameters

a. Recall that confidence intervals were found by determining set of univariate null hypotheses that were not rejected

i. Constructed so that probability of incorrect rejection is $\alpha$ for variables one at a time.

ii. Neither tests nor intervals had type I error/coverage controled for sets of variables.

iii. That is, with two explanatory variables, $\mathrm{P}[\beta_1 \in \mathcal{I}_1] = \mathrm{P}[\beta_2 \in \mathcal{I}_2] = 1 - \alpha$, but $\mathrm{P}[\beta_1 \in \mathcal{I}_1$ and $\beta_2 \in \mathcal{I}_2] < 1 - \alpha$.

iv. Also, $\mathrm{P}_{\beta_1^\circ}\left[H_0 : \beta_1 = \beta_1^\circ \text{ not rejected}\right]$

$$= \mathrm{P}_{\beta_2^\circ}\left[H_0 : \beta_2 = \beta_2^\circ \text{ not rejected}\right] = 1 - \alpha,$$

but $\mathrm{P}_{\beta_1^\circ,\beta_2^\circ}[\text{Neither null hypothesis rejected}] < 1 - \alpha$

b. $F$ procedure tests for multiple included parameters simultaneously.

i. Confidence region is the set of null hypotheses not rejected.

- $\boldsymbol{\gamma}$ is set of $m$ parameters you want to bound
- $\boldsymbol{W}$ is the corresonding submatrix of $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$
- $\mathcal{R} = \{\boldsymbol{\gamma} | (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^\top \boldsymbol{W}^{-1} (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \le \hat{\sigma}^2 F_{m,n-k,\alpha/2}\}$, an

elipse.

MPV: 3.5

7. Confidence Intervals for Fitted Values

a. Predict and bound the fitted value for explanatory variables $(1, x_1, \ldots, x_{k-1})$.

i. Represent this as a row vector $\boldsymbol{x}$.

b. Prediction is $\hat{Y}(\boldsymbol{x}) = \boldsymbol{x}\hat{\boldsymbol{\beta}} = \sum_{j=0}^{k-1} \hat{\beta}_j x_j$

i. $\operatorname{Var}\left[\hat{Y}(\boldsymbol{x})\right] = \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} x_i x_j m_{ij} \sigma^2$

- for $m_{ij}$ the entry in row $i$ and column $j$ of $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$

ii. Can be written as $\boldsymbol{x}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x} \sigma^2$

- If $\boldsymbol{x}$ is a covariate pattern in the data set, this is the diagonal element of the hat matrix, times $\sigma^2$.

c. Prediction is unbiased estimator of $\boldsymbol{x}\boldsymbol{\beta}$

d. When observations are normally distributed, prediction is also normally distributed.

e. So level $1-\alpha$ CI for $\boldsymbol{x}\boldsymbol{\beta}$ is $\boldsymbol{x}\hat{\boldsymbol{\beta}} \pm z_{\alpha/2}\sigma\sqrt{\boldsymbol{x}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}} = \boldsymbol{x}\hat{\boldsymbol{\beta}} \pm \sigma z_{\alpha/2}\sqrt{1/n + (\boldsymbol{x}^*)^\top (\boldsymbol{X}_c^\top \boldsymbol{X}_c)^{-1} \boldsymbol{x}^*}$.

i. $\boldsymbol{x}^*$ is vector of covariates with intercept part removed and average subtracted off.

ii. In the more-common case of $\sigma$ unknown, level $1-\alpha$ confidence interval for $\boldsymbol{x\beta}$ is $\boldsymbol{x}\hat{\boldsymbol{\beta}} \pm t_{n-k,\alpha/2}\sqrt{\boldsymbol{x}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}}\hat{\sigma}$ .

f. Prediction interval is available as in the one-explanatory-variable case.

i. Level $1-\alpha$ confidence interval for $\boldsymbol{x\beta}$ is

$$\boldsymbol{x}\hat{\boldsymbol{\beta}} \pm t_{n-k,\alpha/2}\sqrt{1+\boldsymbol{x}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}}\hat{\sigma}.$$

MPV: 3.8

g. Hidden extrapolation:

i. Predict response at a set of explanatory variables separately typical

ii. an outlier in a bivariate sense.

iii. Distance from center of data set may be indicated by diagonal element of hat matrix.

MPV: 3.11

8. Adding extra explanatory variables can change the sign of other variables?

a. Adding a covariate may remove or change the direction of an effect.

i. Categorical response and two categorical explanatory variables: Simpson's paradox

b. Adding a baseline value is not the same as looking at a change score

i. Continuous response and one categorical explanatory variables and one continuous variable: Lord's paradox

ii. Effect of the categorical variable is generally attenuated, because of regression to the mean.

E. Linear Algebra Concepts

1. An ordered list of observations called a vector.

a. With $n$ entries, call it an $n$ vector.

b. Describe it by giving value for entry in place $j$ , for each $j$

2. A grid of observations, with entries in each row and column, is called a matrix.

a. Describe by generic entry in row $i$ and column $j$

b. With $n$ rows and $k$ columns called a $n \times k$ matrix.

i. This is not really multiplication: here if $n = 10$ and $k = 3$ , read this as "10 by 1" and not "30".

3. Juxtaposition of a matrix $\boldsymbol{X}$ and vector $\boldsymbol{\beta}$ represents matrix multiplication
   a. Defined only if $\boldsymbol{X}$ has as many columns (that is, second dimension) as $\boldsymbol{\beta}$ has entries.
   b. That is, if $= \boldsymbol{X}\boldsymbol{\beta}$, then is the vector
      i. with as many components as $\boldsymbol{X}$ has rows
      ii. Value in slot $i$ is $\sum_j x_{ij}\beta_j$
         - sum of elements in row $i$ of the matrix times corresponding elements of vector.
4. $+$ is vector addition:
   a. need both sides to have same number of components
   b. Result is component-wise sum.
5. A matrix like $\boldsymbol{X}$ with rows and columns interchanged is called the transpose of $\boldsymbol{X}$
   a. $(\boldsymbol{AB})^\top = \boldsymbol{B}^\top \boldsymbol{A}^\top$.
   b. Denote by $\boldsymbol{X}^\top$
6. Just as with scalars, matrix multiplication is distributive: $\boldsymbol{A}(\boldsymbol{b} - \boldsymbol{c}) = \boldsymbol{Ab} + \boldsymbol{Ac}$ if $\boldsymbol{b}$ and $\boldsymbol{c}$ both have as many components as $\boldsymbol{A}$ has columns.

7. Sequential multiplications of the vector can be expressed as a matrix.
   a. Formulate more generally: Find $\boldsymbol{C}$ so that $\boldsymbol{A}(\boldsymbol{B}\boldsymbol{\beta}) = \boldsymbol{C}\boldsymbol{\beta}$
      i. Let $\boldsymbol{A}$ have entries in row $i$ and column $j$ $a_{ij}$
      ii. Let $\boldsymbol{B}$ have entries in row $i$ and column $j$ $b_{ij}$
      iii. Recall that entry $i$ in $\boldsymbol{B}\boldsymbol{\beta}$ is $\sum_j b_{ij}\beta_j$
      iv. Then entry $l$ in $\boldsymbol{A}(\boldsymbol{B}\boldsymbol{\beta})$ is $\sum_i a_{li}(\sum_j b_{ij}\beta_j)$
      v. Rearrange terms in sum to do summation over $i$ for $j = 1$ first, then summation over $i$ for $j = 2$, then ... :
      vi. $\sum_j(\sum_i a_{li}b_{ij}\beta_j)$ : Commutative property of addition
      vii. Factor out $\beta_j$ from multiple terms that contain it: $\sum_j(\sum_i a_{li}b_{ij})\beta_j$ : Distributive Property
   b. Result is $\sum_j c_{lj}\beta_j$ for $c_{lj} = \sum_i a_{li}b_{ij}$
   c. So define the matrix product $\boldsymbol{AB}$ to be the matrix with entry $c_{lj} = \sum_i a_{li}b_{ij}$ in row $l$ column $j$.
   d. Defined only if number of rows of $\boldsymbol{B}$ is the number of columns of $\boldsymbol{A}$.
   e. Result has number of rows of first matrix and number of columns of second matrix.

f. The definition of multiplication of a matrix and a vector is a special case, if the vector is viewed as having one column.

g. Same argument shows matrix multiplication is associative: $\boldsymbol{A}(\boldsymbol{BC}) = (\boldsymbol{AB})\boldsymbol{C}$ .

8. Matrix multiplication is NOT commutative.

a. The product of a $2 \times 3$ and a $3 \times 4$ matrix is a $2 \times 4$ ,

b. but the product with the orders reversed is not defined, because $3 \times 4$ and $2 \times 3$ matrices do not have the number of columns of first matching number of rows of second.

9. Some square matrices have an inverse.

a. Take $\boldsymbol{A}$ a matrix with the same number of rows and columns.

b. Can we find $\boldsymbol{C}$ such that $\boldsymbol{CA}$ is of form $\begin{pmatrix} 1 & 0 & \cdots \\ 0 & 1 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$ .

i. All zeros except 1 if row number matches column number

ii. Such positions are called the diagonal.

iii. A matrix of all zeros except all $1$ on the diagonal is called an identity matrix, because

- if $\boldsymbol{B}$ is a $n \times m$ matrix, and if $\boldsymbol{I}$ is a $n \times n$ identity matrix, then $\boldsymbol{IB} = \boldsymbol{B}$ .
- if $\boldsymbol{B}$ is a $n \times m$ matrix, and if $\boldsymbol{I}$ is a $m \times m$ identity matrix,

then $\boldsymbol{BI} = \boldsymbol{B}$.

iv. So I want to find $\boldsymbol{CA} = \boldsymbol{I}$ (if it exists).

- Such a matrix $\boldsymbol{C}$ is called the matrix inverse: $\boldsymbol{C} = \boldsymbol{A}^{-1}$.
- A matrix without such an inverse is called singular

v. Whether this matrix exists, and its value in this case, is usually straight-forward to compute.

- We will leave these details to the numerical linear algebrists.

10. Other definitions

a. A matrix that is its own square is called idempotent.

MPV: 4.1

F. Model Checking

1. Recall Regression Assumptions

a. Response expectation is approximately linearly in explanatory variables.

i. To make this sensible, the center of the error distribution needs to be zero.

ii. For prediction purposes, this assumption is quite important.

iii. For testing purposes, this is less important.

b. Errors are uncorrelated

   i. A moderate correlation can make standard errors misleading.
   ii. A formal test for one type of deviation will come later.
 c. Errors have a constant variance
   i. This is not so important for a large sample.
 d. Errors are normal.
   i. This is not so important for a moderate sample.

MPV: 4.2

2. Check via residual plot.
 a. Plot vs. fitted value
 b. Or vs. separate covariates.
 c. Expect residuals
   i. exhibit no pattern,
   ii. Be approximately evenly spread out