MPP-E1180 Lecture 8: Statistical Modeling with R

Christopher Gandrud

18 November 2016

Objectives for the class

- Assignment 3
- Review
- Intro to the general syntax for statistical modelling in R.
- Specific examples using:
  - Normal linear regression
  - Logistic regression
  - Panel data

Collaborative Research Project (1)

Purposes: Pose an interesting research question and try to answer it using data analysis and standard academic practices. Effectively communicate your results to a **variety of audiences** in a **variety of formats**.

Deadline:

- Presentation: In-class 2 December
- Website/Paper: 16 December 2016

Collaborative Research Project (2)

The project can be thought of as a ‘dry run’ for your thesis with multiple presentation outputs.

Presentation: 10 minutes **maximum**. **Engagingly** present your research question and key findings to a general academic audience (fellow students).

Paper: 5,000 words maximum. **Standard academic paper**, properly cited laying out your research question, literature review, data, methods, and findings.

Website: An engaging website designed to convey your research to **a general audience**.

Collaborative Research Project (3)

As always, you should **submit one GitHub repository** with all of the materials needed to **completely reproduce** your data gathering, analysis, and presentation documents.

Note: Because you've had two assignments already to work on parts of the project, I expect **high quality work**.

Collaborative Research Project (4)

Find one other group to be a **discussant** for your presentation.

The discussants will provide a quick (max 2 minute) critique of your presentation–ideas for things you can improve on your paper.

Office hours

I will have normal office hours every week for the rest of the term **except 9 December**.

Please take advantages of this opportunity to **improve your final project**.

Review

- What is **web scraping**? What are some of tools R has for web scraping?
- What are **regular expressions** (give at least two examples)? Why are they useful?
- What dplyr function can you use to create a **new variable** in a data frame by running a command on values from groups in that data frame?

Statistical Modelling in R

Caveat: We are **definitely not** going to cover anywhere near R's full capabilities for statistical modeling.

We are also **not going to cover** all of the **modeling concerns/diagnostics** you need to consider when using a given model.

You will need to **rely on your other stats courses** and **texts**.

What are we going to do?

- Discuss the basic syntax and capabilities in R for estimating normal linear and logistic regressions.
- Basic model checking in R.
- Discuss basic ways of interpreting results (we'll do more on this next week).

The basic model

Most statistical models you will estimate are from a general class (**Generalised Linear Model**) that has **two parts**:

Stocastic Component (e.g. randomly determined) assumes the dependent variable Y_i is generated from as a random draw from the probability density function:

$$Y_i \sim f(\theta_i, \alpha)$$

- θ_i: parameter vector of the part of the function that **varies between observations**.
- α: matrix of **non-varying parameters**.

Sometimes referred to as the '**error structure**'.

The basic model

The **Systematic Component** indicating how θ_i varies across observations depending on values of the explanatory variables and (often) some constant:

$$\theta_i = g(X_i,\ \beta)$$

- X_i: a $1 \text{ x } k$ vector of **explanatory variables**.
- β: a $1 \text{ x } k$ vector of **parameters** (i.e. coefficients).
- $g(.,.)$: the **link function**, specifying how the explanatory variables and parameters are translated into θ_i.

Today

Today we will cover two variations of this general model:

- linear-normal regression (i.e. ordinary least squares)
- logit model

Linear-normal regression

For continuous dependent variables assume that Y_i is from the **normal distribution** $(N(.,.))$.

Set the main parameter vector θ_i to the **scalar mean** of: $\theta_i = E(y_i) = \mu_i$.

- **Scalar**: a real number (in R-language: a vector of length 1)

Assume the ancillary parameter matrix is the scalar homoskedastic variance: $\alpha = V(Y_i) = \sigma^2$.

- **Homoskedastic variance**: variance does not depend on the value of x. The standard deviation of the error terms is constant across values of x.

Set the systematic component to the linear form: $g(X_i, \beta) = X_i\beta = \beta_0 + X_{i1}\beta_1 + \dots$

Linear-normal regression

So:

$$Y_i \sim N(\mu_i,\, \sigma^2), \quad \mu_i = X_i \beta$$

Logit regression

For binary data (e.g. 0, 1) we can assume that the stochastic component has a Bernoulli distribution.

The main parameter is $\pi_i = \Pr(Y_i = 1)$.

The systematic component is set to a logistic form: $\pi_i = \frac{1}{1+e^{-X_i\beta}}$.

So:

$$Y_i \sim \text{Bernoulli}(\pi_i), \quad \pi_i = \frac{1}{1 + e^{-X_i\beta}}$$

Example error structure families and link functions

Error Family	Canonical link
Normal	identity
binomial	logit
poisson	log

R syntax

The general syntax for estimating statistical models in R is:

```
response variable ~ explanatory variable(s)
```

Where '~' reads 'is modelled as a function of'.

In the Generalised Linear Model context, either explicitly or implicitly:

```
response variable ~ explanatory variable(s), family = error
```

Model functions

We use model functions to specify the model structure.

Basic model functions include:

- `lm`: fits a linear model where Y is assumed to be normally distributed and with homoskedastic variance.
- `glm`: allows the fitting of many Generalised Linear Models. Lets you specify the error `family`.
- `plm` (package and function): panel data Linear Models
- `pglm` (package and function): panel data Generalised Linear Models

Example of `lm`

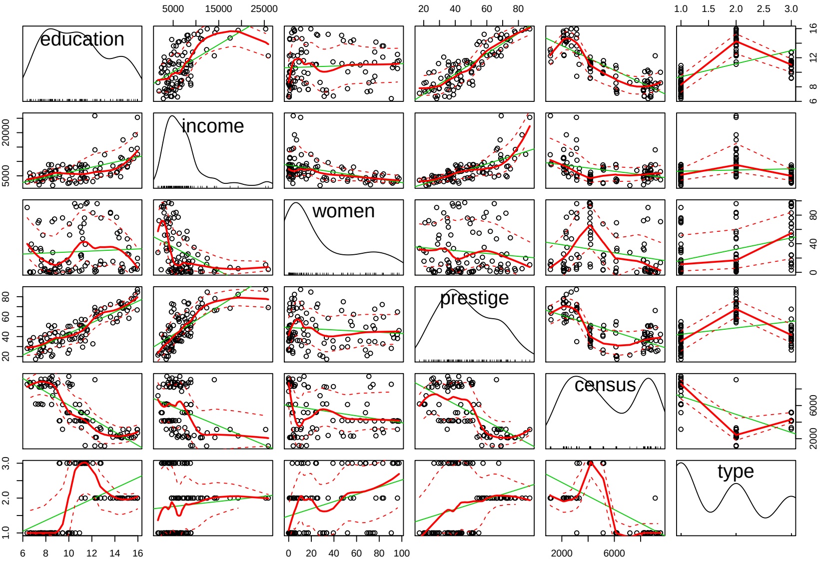
Example data *Prestige* (example based on `http://www.princeton.edu/~otorres/Regression101R.pdf`).

The observations are **occupations** and the dependent variable is a score of each occupation's **prestige**.

```r
library(car)
data(Prestige)
```

Examine correlation matrix

```
car::scatterplotMatrix(Prestige)
```



Example of `lm`

Estimate simple model (education is in years):

```
M1 <- lm(prestige ~ education, data = Prestige)
```

```r
summary(M1)
```

```
##
## Call:
## lm(formula = prestige ~ education, data = Prestige)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.0397  -6.5228   0.6611   6.7430  18.1636
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -10.732      3.677  -2.919  0.00434 **
## education      5.361      0.332  16.148  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.
##
## Residual standard error: 9.103 on 100 degrees of freedo
## Multiple R-squared:  0.7228, Adjusted R-squared:   0.72
```

Confidence intervals of parameter point estimates

Note: **Always prefer estimation intervals** over point estimates.

Deal with your **uncertainty**!

About **95%** of the time the population parameter will be within **about 2 standard errors** of the point estimate.

Using **Central Limit Theorem** (at least about 50 observations and the data is not extremely skewed):

$$CI\_95 = \text{point estimate} \pm 1.96 * SE$$

Confidence intervals of parameter point estimates

```r
confint(M1)
```

```
##                  2.5 %    97.5 %
## (Intercept) -18.027220 -3.436744
## education     4.702223  6.019533
```

Example of `lm`

Estimate model with categorical (factor) variable:

```
M2 <- lm(prestige ~ education + type,
         data = Prestige)
```

```r
summary(M2)
```

```
##
## Call:
## lm(formula = prestige ~ education + type, data = Prestig
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.410  -5.508   1.360   5.694  17.171
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.6982     5.7361  -0.470   0.6392
## education     4.5728     0.6716   6.809 9.16e-10 ***
## typeprof      6.1424     4.2590   1.442   0.1526
## typewc       -5.4585     2.6907  -2.029   0.0453 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.
##
```

Create categorical variable from continuous variable

Use the cut function to create a categorical (factor) variable from a continuous variable.

```
Prestige$income_cat <- cut(Prestige$income,
                breaks = c(0, 4999, 9999, 14999, 30000),
                labels = c('< 5,000', '< 10,000', '< 15,00
                            '>= 15,000'))
summary(Prestige$income_cat)
```

```
##   < 5,000  < 10,000  < 15,000 >= 15,000
##        38        51         9         4
```

Note: cut excludes the left value and includes the right value, e.g. (0, 4999].

Example of lm

```
M3 <- lm(prestige ~ education + income_cat,
         data = Prestige)
confint(M3)
```

```
##                              2.5 %    97.5 %
## (Intercept)             -10.914031  3.182500
## education                 3.350201  4.778444
## income_cat< 10,000        6.085375 13.030546
## income_cat< 15,000        9.584532 23.391854
## income_cat>= 15,000      12.040936 29.902733
```

Example of `lm`

Estimate models with polynomial transformations:

```r
# Cubic polynomial transformation
M4  <- lm(prestige ~ education + poly(income, 2),
          data = Prestige)
confint(M4)
```

```
##                        2.5 %    97.5 %
## (Intercept)        -1.470988  13.33552
## education           3.132827   4.48515
## poly(income, 2)1  45.174019  81.39445
## poly(income, 2)2 -43.150740 -12.87994
```

Example of `lm`

Estimate models with (natural) logarithmic transformations:

```r
# Cubic polynomial transformation
M5  <- lm(prestige ~ education + log(income),
          data = Prestige)
```

```r
summary(M5)
```

```
##
## Call:
## lm(formula = prestige ~ education + log(income), data =
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.0346  -4.5657  -0.1857   4.0577  18.1270
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -95.1940    10.9979  -8.656 9.27e-14 ***
## education     4.0020     0.3115  12.846  < 2e-16 ***
## log(income)  11.4375     1.4371   7.959 2.94e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.
##
## Residual standard error: 7.145 on 99 degrees of freedom
```

Example of `lm`

Estimate model with interactions:

```
M6 <- lm(prestige ~ education * type,
         data = Prestige)
```

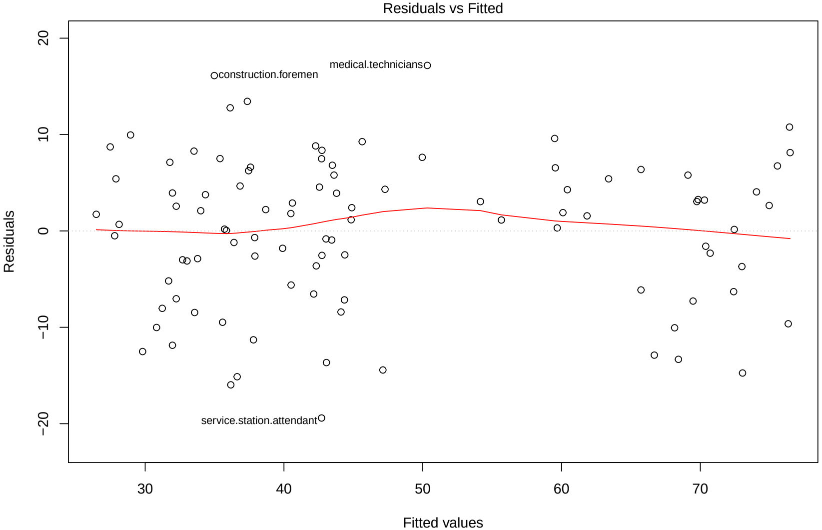
```r
summary(M6)
```

```
##
## Call:
## lm(formula = prestige ~ education * type, data = Prestig
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.7095  -5.3938   0.8125   5.3968  16.1411
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -4.2936     8.6470  -0.497    0.621
## education             4.7637     1.0247   4.649 1.11e-05
## typeprof             18.8637    16.8881   1.117    0.267
## typewc              -24.3833    21.7777  -1.120    0.266
## education:typeprof   -0.9808     1.4495  -0.677    0.500
## education:typewc      1.6709     2.0777   0.804    0.423
## ---
```

Diagnose heteroscedasticity

Use `plot` on a model object to run visual diagnostics.

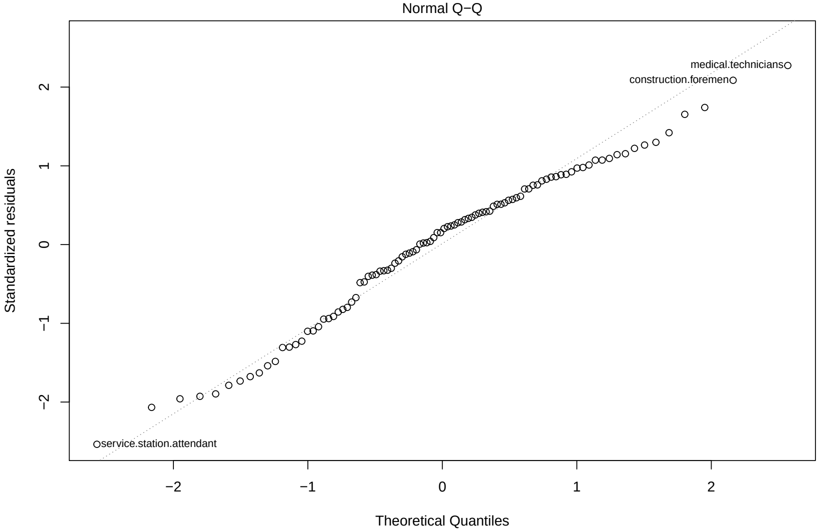
```
plot(M2, which = 1)
```



Diagnose non-normality of errors

`plot` to see if a model's errors are normally distributed.

```
plot(M2, which = 2)
```



Example of logistic regression with glm

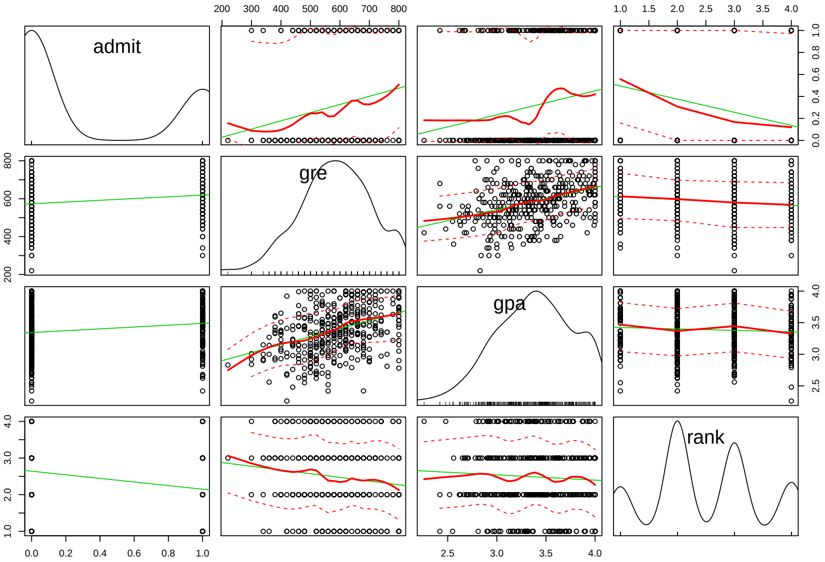
Example from UCLA IDRE.

Simulated data of admission to grad school.

```r
# Load data
URL <- 'http://www.ats.ucla.edu/stat/data/binary.csv'
Admission <- read.csv(URL)
```

Example of logistic regression with `glm`

```
car::scatterplotMatrix(Admission)
```



Contingency table for school rank and admission

```r
admit_table <- xtabs(~admit + rank, data = Admission)
admit_table
```

```
##       rank
## admit  1  2  3  4
##     0 28 97 93 55
##     1 33 54 28 12
```

Row and column proportions

```
# Row proportions
prop.table(admit_table, margin = 1)
```

```
##       rank
## admit          1          2          3          4
##     0 0.10256410 0.35531136 0.34065934 0.20146520
##     1 0.25984252 0.42519685 0.22047244 0.09448819
```

```
# Column proportions
prop.table(admit_table, margin = 2)
```

```
##       rank
## admit         1         2         3         4
##     0 0.4590164 0.6423841 0.7685950 0.8208955
##     1 0.5409836 0.3576159 0.2314050 0.1791045
```

Summary of contingency table for school rank and admission

```r
summary(admit_table)
```

```
## Call: xtabs(formula = ~admit + rank, data = Admission)
## Number of cases in table: 400
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 25.242, df = 3, p-value = 1.374e-05
```

Example of logistic regression with `glm`

```r
Logit1 <- glm(admit ~ gre + gpa + as.factor(rank),
              data = Admission, family = 'binomial')
```

Note: Link function is assumed to be logit if `family = 'binomial'`.

Example of logistic regression with `glm`

```r
confint(Logit1)
```

```
##                           2.5 %       97.5 %
## (Intercept)       -6.2716202334 -1.792547080
## gre                0.0001375921  0.004435874
## gpa                0.1602959439  1.464142727
## as.factor(rank)2 -1.3008888002 -0.056745722
## as.factor(rank)3 -2.0276713127 -0.670372346
## as.factor(rank)4 -2.4000265384 -0.753542605
```

Interpreting logistic regression results

β's in logistic regression are interpretable as **log odds**. These are weird.

If we exponentiate log odds we get **odds ratios**.

```r
exp(cbind(OddsRatio = coef(Logit1), confint(Logit1)))
```

```
##                  OddsRatio       2.5 %    97.5 %
## (Intercept)      0.0185001 0.001889165 0.1665354
## gre              1.0022670 1.000137602 1.0044457
## gpa              2.2345448 1.173858216 4.3238349
## as.factor(rank)2 0.5089310 0.272289674 0.9448343
## as.factor(rank)3 0.2617923 0.131641717 0.5115181
## as.factor(rank)4 0.2119375 0.090715546 0.4706961
```

These are **also weird**.

Interpreting logistic regression results

What we really want are **predicted probabilities**

First create a data frame of fitted values:

```
fitted <- with(Admission,
               data.frame(gre = mean(gre),
                          gpa = mean(gpa),
                          rank = factor(1:4)))
fitted
```

```
##     gre    gpa rank
## 1 587.7 3.3899    1
## 2 587.7 3.3899    2
## 3 587.7 3.3899    3
## 4 587.7 3.3899    4
```

Interpreting logistic regression results

Second predict probability point estimates for each fitted value.

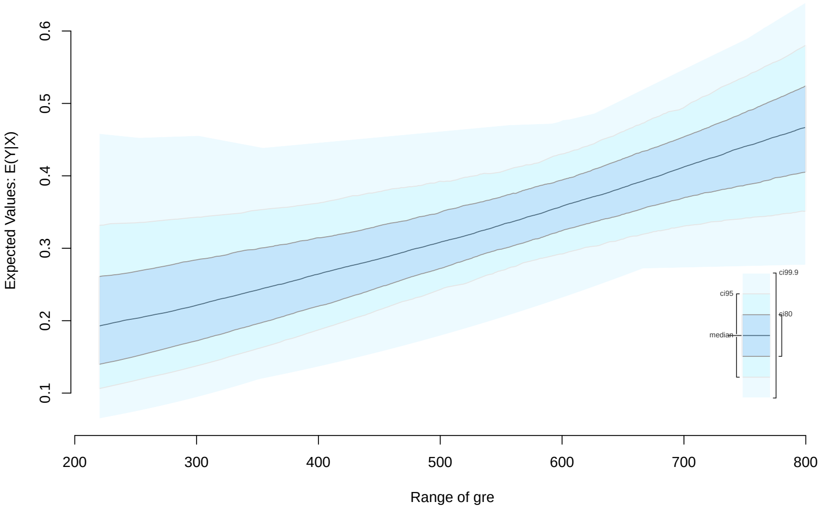
```
fitted$predicted <- predict(Logit1, newdata = fitted,
                            type = 'response')

fitted
```

```
##     gre    gpa rank predicted
## 1 587.7 3.3899    1 0.5166016
## 2 587.7 3.3899    2 0.3522846
## 3 587.7 3.3899    3 0.2186120
## 4 587.7 3.3899    4 0.1846684
```

More interpretation

Next class we will explore other methods of interpreting results from regression models.



Showing results from regression models

@King2001 argue that **post-estimation simulations** can be used to effectively communicate **results from regression models**.

Steps

1. Estimate our parameters' point estimates for $\hat{\beta}_{1 \ldots k}$.
2. Draw *n* values of the point estimates from multivariate normal distributions with means $\bar{\beta}_{1 \ldots k}$ and variances specified by the parameters' estimated co-variance.
3. Use the simulated values to calculate quantities of interest (e.g. predicted probabilities).
4. Plot the simulated distribution using **visual weighting**.

Notes

Post-estimation simulations allow us to effectively communicate our estimates and the **uncertainty around them**.

This method is broadly similar to a fully Bayesian approach with Markov-Chain Monte Carlo or bootstrapping. Just differ on **how the parameters are drawn**.

Implementation

1. Find the coefficient estimates from an estimated model with `coef`.
2. Find the co-variance matrix with `vcov`.
3. Draw point estimates from the multivariate normal distribution with `mvrnorm`.
4. Calculate the quantity of interest with the draws + fitted values using and plot the results.

Simulations: estimate model

First estimate your model as normal and create fitted values:

```r
library(car) # Contains data
library(dplyr) # Piping function
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:car':
##
##     recode

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Simulations: extract estimates

Extract point estimates (coefficients) and co-variance matrix:

```
mp_coef <- matrix(coef(M_prest))
mp_vcov <- vcov(M_prest)
```

Now draw 1,000 simulations of your point estimates:

```
library(MASS) # contains the mvrnorm function
drawn <- mvrnorm(n = 1000, mu = mp_coef, Sigma = mp_vcov) %
            data.frame
head(drawn)[1:3, ]
```

```
##   X.Intercept. education   typeprof     typewc
## 1    -7.638579  5.443553 -0.2803541 -8.0892620
## 2    -8.137030  5.303804  0.6064953 -8.2337444
## 3     2.922903  3.745013 12.1953685 -0.5905028
```

Simulations: merge in fitted values

Now we can add in our fitted values to the simulation data frame:

```
drawn_sim <- merge(drawn, edu_range)

# Rename the fitted value variable
drawn_sim <- dplyr::rename(drawn_sim, fitted_edu = y)

nrow(drawn)
```

```
## [1] 1000
```

```
nrow(drawn_sim)
```

```
## [1] 11000
```

Simulations: calculate quantity of interest

Using the normal linear regression formula ($\hat{y}_i = \hat{\alpha} + X_{i1}\hat{\beta}_1 + \ldots$) we can find the quantity of interest for white collar workers:

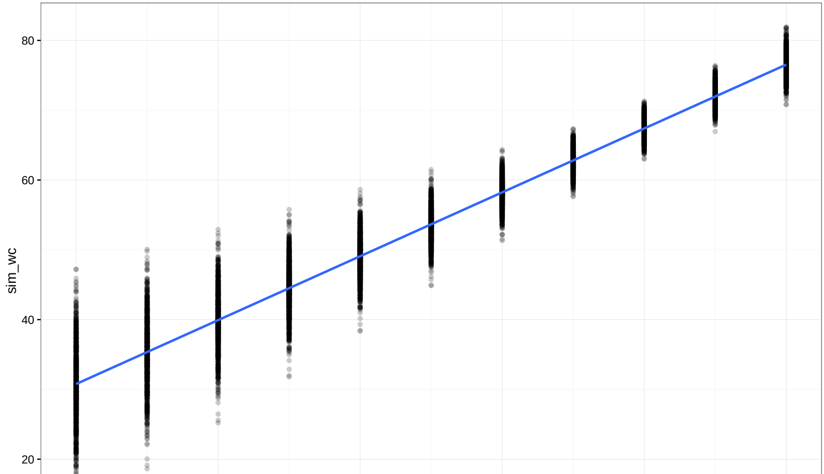
```
names(drawn_sim)
```

```
## [1] "X.Intercept." "education"    "typeprof"     "typewc
## [5] "fitted_edu"
```

```
drawn_sim$sim_wc <- drawn_sim[, 1] + drawn_sim[, 2] * drawn
                         drawn_sim[, 3]
```

Simulations: plot points

```
library(ggplot2)
ggplot(drawn_sim, aes(fitted_edu, sim_wc)) +
        geom_point(alpha = 0.2) + stat_smooth(se = FALSE) +
        theme_bw()
```

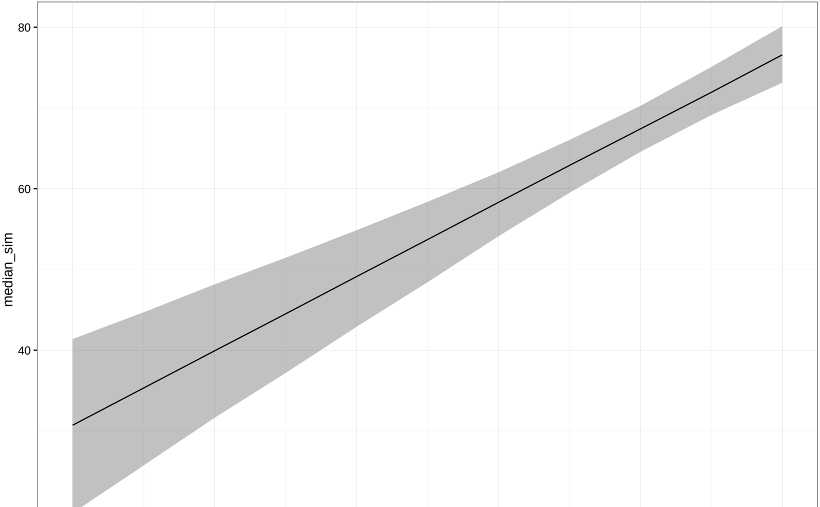


Simulations: find 95% central interval ribbons

```
# Find 95% central interval and median at each fitted value
central <- drawn_sim %>% group_by(fitted_edu) %>%
           summarise(median_sim = median(sim_wc),
                     lower_95 = quantile(sim_wc, probs = 0
                     upper_95 = quantile(sim_wc, probs = 0
           )
```

Simulations: plot 95% central interval

```
ggplot(central, aes(fitted_edu, median_sim)) +
    geom_ribbon(aes(ymin = lower_95, ymax = upper_95), alph
    geom_line() + theme_bw()
```



Predictions from logistic regression

Use the same steps for simulating predicted outcomes from logistic regression models. The only difference is that the equation for the quantity of interest is:

$$P(y_i = 1) = \frac{\exp(\hat{\alpha} + X_{i1}\hat{\beta}_1 + \ldots)}{1 - \exp(\hat{\alpha} + X_{i1}\hat{\beta}_1 + \ldots)}$$

Easier Implementation

The Zelig package streamlines the simulation process.

Zelig (1)

First estimate your regression model using `zelig`.

```
library(Zelig)

# Have to explicitly declare rank as factor
Admission$rank <- as.factor(Admission$rank)

Z1 <- zelig(admit ~ gre + gpa + rank, cite = FALSE,
              data = Admission, model = 'logit')
```

Zelig (2)

Then set the fitted values with `setx`.

```
setZ1 <- setx(Z1, gre = 220:800)
```

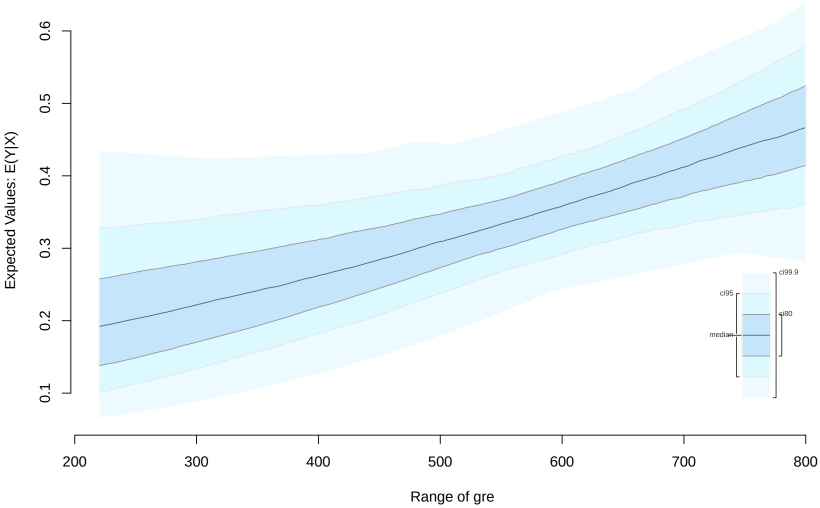
And run the simulations (1,000 by default) with `sim`.

```
simZ1 <- sim(Z1, x = setZ1)
```

Zelig (3)

Plot:

```r
ci.plot(simZ1)
```



Seminar: modeling

Begin working on the statistical models for **your project**.

and/or

Out of Lecture Challenge: Estimate a normal regression model and **plot predicted values** across a range of fitted values. Bonus: do so with a measure of uncertainty.