Lecture 7: Gelman Hill Ch 3.7 & Ch 4.4 - 4.7

Logarithmic Transformations

Additivity and linearity are necessary assumptions for linear models.

- * power basis functions for predictors
- * interactions can account for non-additive structure

Another option is to take a $\underbrace{\text{logarithmic}}_{\text{natural log}}$ transformation…

Consider a logarithmic transformation of the outcome variable.

$$\log(y_i) = B_0 + B_1 X_{i1} + \dots + B_p X_{ip} + \varepsilon_i$$

$$\Downarrow$$

$$y_i = \exp(B_0 + B_1 X_{i1} + \dots + B_p X_{ip} + \varepsilon_i)$$

.95

this is a multiplicative model

$$y_i = \exp(B_0) * \exp(B_1 X_{i1}) * \dots * \exp(\varepsilon_i)$$

$$B_0 * B_1 \exp(x_{i1})$$

The logarithmic transformation also includes the ability to force an outcome variable to be positive

Consider the Seattle housing dataset where we created the log price and scaled predictors.

```
Seattle <- read_csv('http://math.montana.edu/ahoegh/teaching/stat408/datasets/SeattleHousing.csv')
Seattle <- Seattle %>%
  mutate(log_price = log(price),
         beds_scale = scale(bedrooms),
         size_scale = scale(sqft_living))
```

```
display(lm(log_price ~ beds_scale + size_scale, data = Seattle))
```

```
## lm(formula = log_price ~ beds_scale + size_scale, data = Seattle)
##             coef.est coef.se
## (Intercept) 13.09     0.02
## beds_scale  -0.05     0.02
## size_scale   0.52     0.02
## ---
## n = 869, k = 3
## residual sd = 0.47, R-Squared = 0.52
```

- β_0: is the predicted (log) price for a house with an average number of bedrooms (3.2) and average size (2114 square feet of living space).
- β_{size} is the predicted difference (log) price for a 1sd increase in living space after controlling for the # of bedrooms,
- β_{beds}

We can use the exponential function to consider the impact on the price…

For instance exp(-.05) = 0.95, which implies that an increase in 1 unit (sd) for # of bedrooms results in a predicted decrease of about 5% … when holding everything else constant.

Additionally exp(.52) = 1.68, which implies

Interactions… see textbook.

Other Transformations and Ideas

- *Square root transformations*: The square root can also be used to transform either the outcome or predictor variables.

- *Continuous rather than discrete predictors*: often binary (or discrete) predictors can actually be seen as continuous variables. Rather than republican or democrat, we could use percent registered republican.

- *Discrete rather than continuous predictors*: in other cases, using discrete predictors might be prefered. (Especially if the response is expected to have a form that is not monotonic or quadratic)

- *Indentifiability*: If a model contains a set of regression coefficients that can be formed as a set of linear combinations is non-identifiable. This is why a baseline (intercept) term is aliased with levels of each categorical variable.

Explanatory Inference vs. Predictive Inference

When fitting models there are two common types of models

- Explanatory Inference:

assume that you work for a real estate developer and you want to understand characteristics that would lead to more expensive houses.

- Predictive Inference:

assume that you work for Trulia, the goal is to build a model to predict housing prices.

ASK RACHEL

$\boxed{\text{inputs}} \longrightarrow \boxed{f(x)} \rightarrow$ estimated price

Suppose we are interested in predicting housing prices. . .

```
Seattle <- Seattle %>% mutate(zipcode = as.factor(zipcode),
                              size_scale2 = (sqft_living - mean(sqft_living))/sd(sqft_living))
Seattle %>% group_by(zipcode) %>% summarize(mean_price = mean(price), sd_price = sd(price), num_houses =
```

```
## # A tibble: 9 x 4
##   zipcode mean_price sd_price num_houses
##   <fct>        <dbl>    <dbl>      <int>
## 1 98010      423666.  195415.        100
## 2 98014      455617.  258603.        124
## 3 98024      580638.  377595.         81
## 4 98032      251296.   64705.        125
## 5 98039     2161300  1166904.         50
## 6 98070      487480.  201698.        118
## 7 98102      901516.  786815.        105
## 8 98109      880078.  455701.        109
## 9 98148      284909.   89617.         57
```

Interpret the following model. . .

```
lm_preds <- lm(price ~  size_scale2  , data = Seattle)
display(lm_preds)
```

```
## lm(formula = price ~ size_scale2, data = Seattle)
##             coef.est  coef.se
## (Intercept) 632591.46  13424.72
## size_scale2 496558.83  13432.45
## ---
## n = 869, k = 2
## residual sd = 395744.74, R-Squared = 0.61
```

How to make predictions

```
x_new <- data.frame(size_scale2 = c(-1,0,1))
predict(lm_preds, x_new, interval ='confidence')
```

zipcode = (98039 , ,)

```
##         fit         lwr       upr
## 1  136032.6   98759.14  173306.1
## 2  632591.5  606242.71  658940.2
## 3 1129150.3 1091876.80 1166423.8
```

```
predict(lm_preds, x_new, interval ='prediction')
```

```
##         fit       lwr       upr
## 1  136032.6 -641590.9  913656.2
## 2  632591.5 -144585.1 1409768.0
## 3 1129150.3  351526.7 1906773.9
```

The confidence interval is mean price for houses with those predictor variables

The prediction interval is

price for a single house with those predictor variables

External Validation

One of the best ways to validate a model is to use an external dataset. In other words, fit the model on a dataset for housing prices in Seattle *and then* predict upcoming housing sales.

Model Building Principles

Note there can be huge issues when including too many predictors (overfitting)

It is important to think about a reasonable theoretical model in advance (including possible signs of predictors)

1. Include all input variables that, for substantive reasons, might be expected to be important in predicting the outcome.

2. It is not always necessary to include these inputs as separate predictors - for example, sometimes several inputs can be averaged or summed to create a "total score" that can be used as a single predictor in the model.

3. For inputs that have large effects, consider including their interactions as well.

4. We suggest the following strategy for decisions regarding whether to exclude a variable from a prediction model based on expected sign and statistical significance

a. If a predictor is not statistically significant and has the expected sign, it is generally fine to keep it in. It may not help predictions dramatically but is also not hurting them.

b. If a predictor is not statistically significant and does not have the expected sign, consider removing it from the model.

c. If a predictor is statistically significant and does not have the expected sign, then think hard if it makes sense. Try to gather data on potentially lurking variables and include them in the analysis.

d. If a predictor is statistically significant and has the expected sign, then by all means keep it in the model.