Final Project

Zoe Aiello

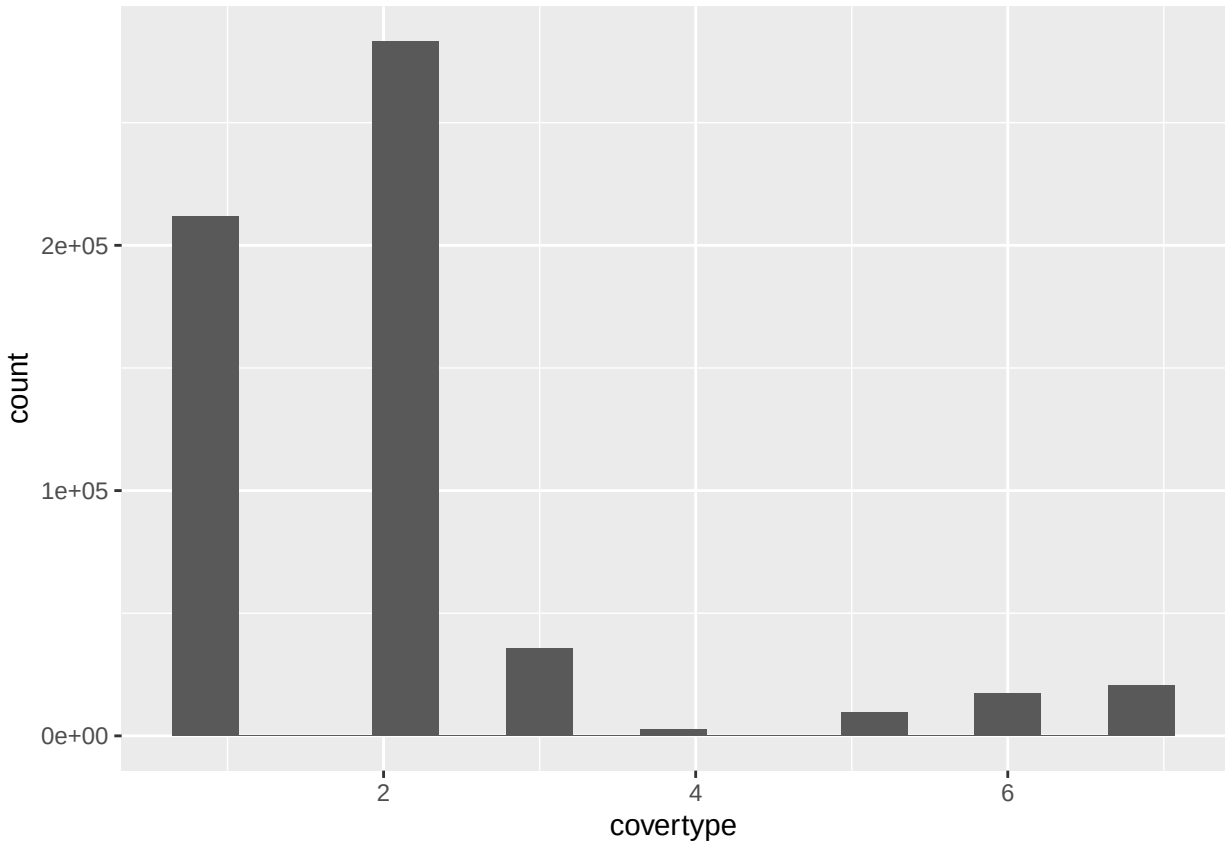
2022-05-25

Executive Summary

Summary of the study and data

The Forest Covertype data set is from Colorado State University and includes both categorical and integer variables that describe cartographic variables for 581,012 30x30 meter cells in four wilderness areas of the Roosevelt National Forest in Colorado. The forest covertype was determined by the US Forest Service Region 2 Resource Information System, a geospatial database for the Rocky Mountain Region of the US. The independent variables were pulled from data obtained by the US Geological Survey and the US Forest Service. I was not able to find the associated publication for the USGS and USFS so I'm unsure of their exact methods. I'm interested to understand how the authors determined covertype from viewing geospatial data. The USGS and USFS data record variables for each 30x30 meter cell including elevation, aspect, slope, horizontal and vertical distances to hydrology, horizontal distance to roadways, hillshade at 9am, noon, and 3pm, horizontal distance to fire points, soil type, and wilderness area. 10 of these variables are quantitative variables. Soil type is a 40-column binary variable, cover type is a qualitative variable with 7 levels, and wilderness area is a 4-column binary variable.

The distribution of the response variable, covertype is described by the following figure:



As expected, the distribution is nonnormal since it is displayed by count. Most cases are have covertypes 1 and 2 and the least cases have covertype 4. However, covertype will be described as a nominal categorical variable since its categories do not have a natural order. The categories are described as follows:

1 – Spruce/Fir

2 – Lodgepole Pine

3 – Ponderosa Pine

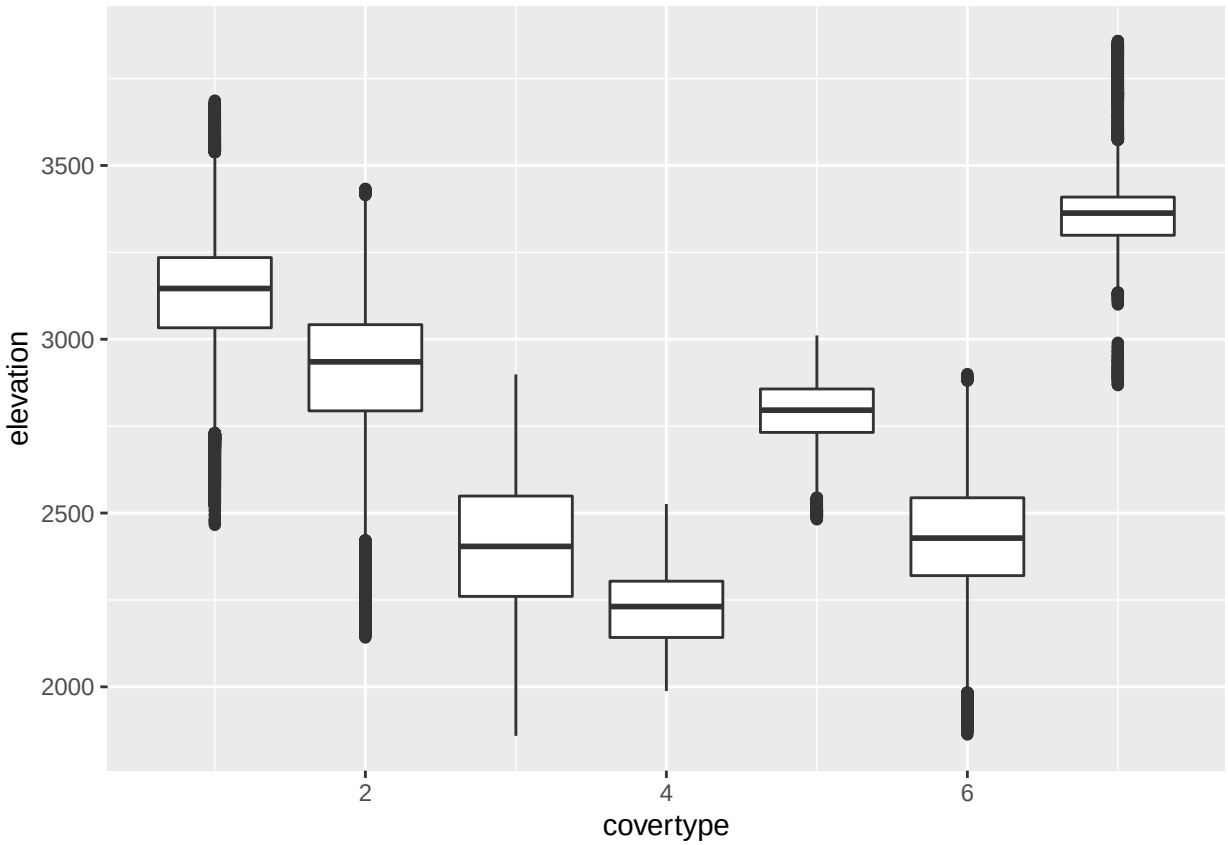
4 – Cottonwood/Willow

5 – Aspen

6 – Douglas-fir

7 – Krummholz

The relationship between covertype and elevation is described by the following figure:



Out of all the relationships between covertype and the explanatory variables, the relationship between covertype and elevation appears to be the most interesting. The relationship displayed appears to be quadratic with the only outlier being covertype 5 which is found at higher elevations than covertype 6. However, because I will be combining the binary covertype variables into one nominal categorical variable I wouldn't consider the relationship between the nonordered covertypes and elevation quadratic. Since this was the most interesting relationship, I decided to investigate the following question:

Does elevation have the strongest effect on covertype? To answer this question I fit a multinomial logistic regression model using covertype as the response variable, and tested the significance of elevation in the model using Wald's Test. I wanted to use a multinomial regression model because my response variable, cover type, is a nominal (having no natural order) categorical variable and the multinomial regression model is best for data of this type.

I performed the Wald's Test and found that elevation does have a significant effect on the success of the model in predicting covertype (p-value = 0.0) but I also found that slope is has the strongest effect on covertype according to its estimated coefficient ($\beta = -0.01545555$).

My second question about the data involved creating a predictive model for covertype:

Can we accurately predict cover type based on the given variables? To answer this question I split the data into 25% testing and 75% training sets and fit a multinomial logistic regression model using covertype as a response variable with the training data. I chose multinomial regression model because the response variable, cover type, is a nominal categorical variable. I used the training data to predict the responses and evaluated the prediction model using MSPE.

I found that the model was fairly accurate in predicting covertype with the given explanatory variables because it had an MSPE value of 0.18.

Body (Analysis and Results)

The Wald's Test concluded that elevation has a significant effect on the success of the model in predicting covertype (p-value = 0.0, df =). I took the average of the coefficients for each variable at each covertype and found that slope has the strongest effect on covertype (β = -0.01545555) and elevation has the second strongest effect (β = -0.014123). The coefficients of the model are summarized in the following table:

Coefficients	Model	Average
(Intercept):2	2.281378e+01	
(Intercept):3	6.586145e+01	
(Intercept):4	8.398127e+01	NA
(Intercept):5	2.750529e+01	
(Intercept):6	6.524793e+01	
(Intercept):7	-5.458692e+01	
elevation:2	-7.553361e-03	
elevation:3	-2.479824e-02	
elevation:4	-3.349503e-02	-0.014123
elevation:5	-1.011294e-02	
elevation:6	-2.481999e-02	
elevation:7	1.604154e-02	
aspect:2	6.601025e-04	
aspect:3	1.290574e-03	
aspect:4	-1.674109e-03	0.0001569778
aspect:5	3.442425e-05	
aspect:6	2.075637e-03	
aspect:7	-1.444762e-03	
slope:2	-1.545726e-02	
slope:3	-7.200134e-03	
slope:4	-7.874396e-02	-0.01545555
slope:5	1.372930e-02	
slope:6	-1.824206e-02	
slope:7	1.318081e-02	
vert_dist_hydro:2	6.333873e-03	
vert_dist_hydro:3	1.748994e-02	
vert_dist_hydro:4	1.512194e-02	0.008846647
vert_dist_hydro:5	7.615643e-03	
vert_dist_hydro:6	1.252636e-02	
vert_dist_hydro:7	-6.007872e-03	
horz_dist_road:2	6.897790e-05	
horz_dist_road:3	6.893344e-04	
horz_dist_road:4	1.839432e-03	0.0005071198
horz_dist_road:5	-3.953039e-04	
horz_dist_road:6	7.654432e-04	
horz_dist_road:7	7.483549e-05	
horz_dist_fire:2	9.300507e-06	
horz_dist_fire:3	-5.954351e-04	
horz_dist_fire:4	-2.372502e-04	-0.0002181427
horz_dist_fire:5	-1.853773e-04	
horz_dist_fire:6	-1.853773e-04	
horz_dist_fire:7	7.630036e-05	

For the parameters of the model I used elevation, aspect, slope, vertical distance to hydro (vert_dist_hydro), horizontal distance to a road (horz_dist_road), and horizontal distance to a fire point (horz_dist_fire). I

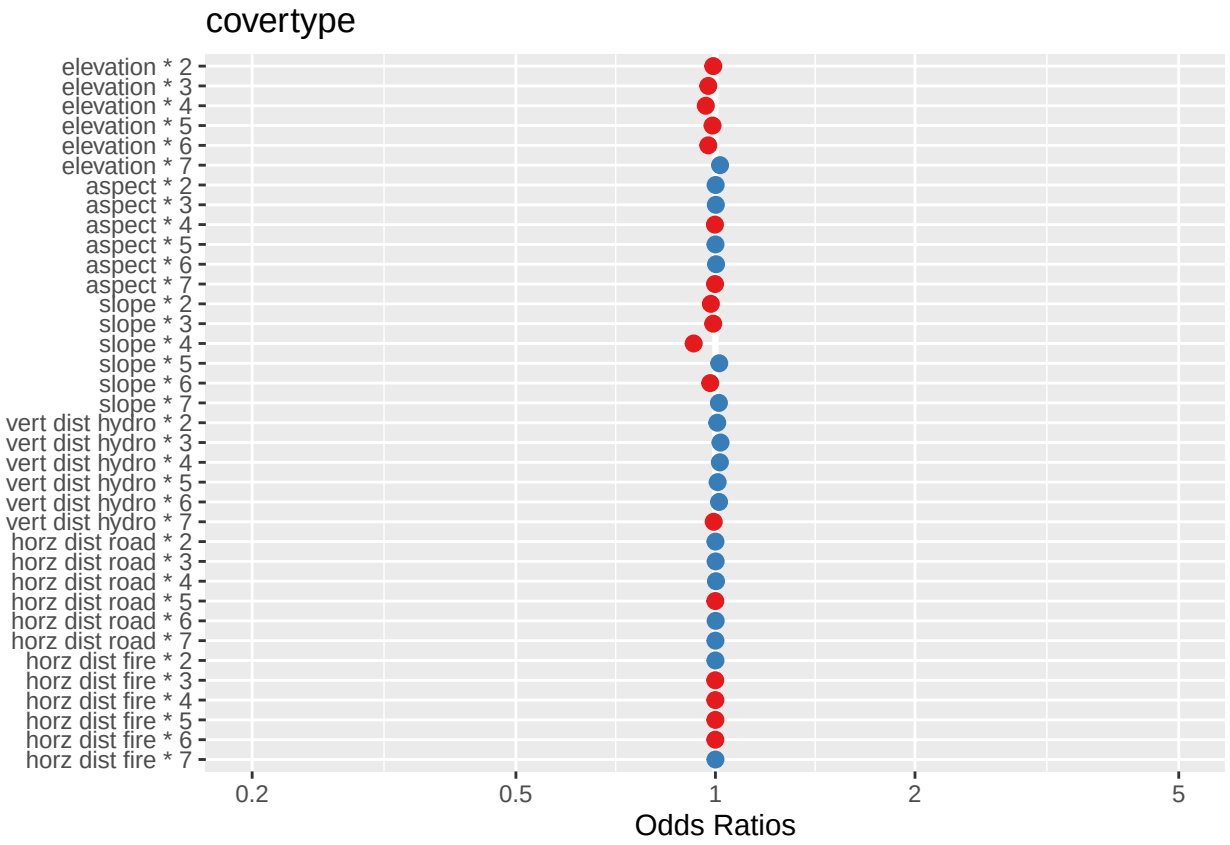
chose to exclude horizontal distance to hydro (horz_dist_hydro), shade at 3pm (shade_3pm), shade at noon (shade_noon), and shade at 9am (shade_9am) because these variables had pearson correlation coefficients greater than abs(0.5) with at least one of the other variables. More information on the exclusion of these variables is included in the appendix.

I also chose to remove the categorical variables wilderness area type (wildtype) and soiltype because the model parameters could not be estimated when they were included in the model. I believe this is because not every covertype includes every type of combination of soiltypes and wildtypes (e.g. Covertype 1 is not present in every wilderness area and included in every soiltype, etc.). The exclusion of these parameters may have affected the accuracy of the model but the McFadden's R^2 value of 0.38525 suggests that the fit is pretty good. According to McFadden (1974), McFadden R^2 values between 0.2 and 0.4 indicate an excellent model fit (https://stats.stackexchange.com/questions/82105/mcfaddens-pseudo-r2-interpretation). Another point of concern is the Independence of Irrelevant Alternatives (IIA) assumption which is an important one for multinomial regression models. The IAA states that the model's choice between two outcomes (e.g. Covertype 1 or Covertype 2) is not impacted by the alternative options (https://citeseerx.ist.psu.edu/viewdoc/download ?doi=10.1.1.951.3160&rep=rep1&type=pdf). The IIA can be tested using the Hausman-McFadden Test which compares the full model to a reduced model without certain alternatives to determine if those selected alternatives are irrelevant or not. Tests for the IIA have been rightfully criticized as being misinformative (Blue-Red Bus problem) so when I used the Hausman-McFadden Test on my model and found that IIA was rejected, I decided to proceed with my model regardless. This is an important caveat to the results from my model. More information about how I conducted the Hausman-McFadden test is included in the Appendix.

After separating my data into 75% training and 25% testing sets, I used the training data to fit the same model. Therefore, the model gives the same concerns here but the model had a Mean Squared Prediction Error that was fairly close to 0 at 0.18. However, the exponentiation coefficients for all the predictors are fairly close to 1, indicating that they have a very minute association with covertype. This is puzzling because the p-values for all the coefficients were very small and the MSPE value was also small. The p-values for the coefficients are summarized in the following table:

Coefficients	p-value
(Intercept):2	< 2.2e-16
(Intercept):3	< 2.2e-16
(Intercept):4	< 2.2e-16
(Intercept):5	< 2.2e-16
(Intercept):6	< 2.2e-16
(Intercept):7	< 2.2e-16
elevation:2	< 2.2e-16
elevation:3	< 2.2e-16
elevation:4	< 2.2e-16
elevation:5	< 2.2e-16
elevation:6	< 2.2e-16
elevation:7	< 2.2e-16
aspect:2	< 2.2e-16
aspect:3	< 2.2e-16
aspect:4	2.796e-13
aspect:5	0.8712
aspect:6	< 2.2e-16
aspect:7	< 2.2e-16
slope:2	< 2.2e-16
slope:3	8.852e-07
slope:4	< 2.2e-16
slope:5	3.331e-15
slope:6	< 2.2e-16
slope:7	< 2.2e-16

Coefficients	p-value
vert_dist_hydro:2	< 2.2e-16
vert_dist_hydro:3	< 2.2e-16
vert_dist_hydro:4	< 2.2e-16
vert_dist_hydro:5	< 2.2e-16
vert_dist_hydro:6	< 2.2e-16
vert_dist_hydro:7	< 2.2e-16
horz_dist_road:2	< 2.2e-16
horz_dist_road:3	< 2.2e-16
horz_dist_road:4	< 2.2e-16
horz_dist_road:5	< 2.2e-16
horz_dist_road:6	< 2.2e-16
horz_dist_road:7	< 2.2e-16
horz_dist_fire:2	7.787e-05
horz_dist_fire:3	< 2.2e-16
horz_dist_fire:4	1.573e-09
horz_dist_fire:5	< 2.2e-16
horz_dist_fire:6	< 2.2e-16
horz_dist_fire:7	2.220e-16



I tried fitting a model with 2-way and 3-way interactions between the variables but it resulted in an error. The large sample size of my data may have resulted in low p-values, causing the model coefficients to have statistical significance but no practical significance.

Within the covertype categories, the smallest sample size is 2,747 in covertype 4 and the largest is in covertype 1 which has 211,840 cases. Overall, there are 581,012 observations. With a large n, even very small effects can be statistically significant

Conclusions/Discussion

The data analysis found that elevation was the second most influential parameter behind slope using a Wald's Test and looking at the model's coefficients. Additionally, the model resulted in a reasonable RMSE estimate but the odds ratios for the coefficients suggested that they had little effect on the covertype. In the future, it might be useful to subset the data and create a model with that smaller subset to avoid the perils of a large sample size. The dissertation that used the Forest Cover dataset used 11,340 observations to train their model, 3,780 to develop their model, and the rest of the data for testing (565,892), this may have been a better approach to building a model with such a large sample size. Additionally, it may be useful to explore another modeling option like the Nested logit or Multinomial probit models which are alternatives to the multinomial regression model that relax the Independence of Irrelevant Alternatives assumption which may be a difficult assumption to test accurately. More data collection might include taking the temperature, precipitation, or oxygen levels in the area of the plot which may be better indicators of forest cover type.

Appendix

Multinomial Logistic Regression

https://it.unt.edu/sites/default/files/mlr_jds_aug2011.pdf

Multinomial logistic regression is an extension of the binomial logistic regression but it models nominal categorical response variables and does not assume normality, linearity or homoscedasity. It uses maximum likelihood estimation and has a few assumptions:

1. Sample size All the levels of Covertype have over 10 observations so the sample size assumption is met.

```
# get a summary of all the variables
summary(covdata3$covertype)
```

```
##      1      2      3      4      5      6      7
## 211840 283301  35754   2747   9493  17367  20510
```

```
# get a summary of the frequency of alternatives
apply(fitted(mod_1, outcome = FALSE), 2, mean)
```

2. Non-Perfect Seperation

```
##           1           2           3           4           5           6
## 0.364605206 0.487599223 0.061537455 0.004727957 0.016338733 0.029890949
##           7
## 0.035300476
```

```
# Frequencies of alternatives:choice
      # 1        2        3        4        5        6        7
# 0.364605 0.487599 0.061537 0.004728 0.016339 0.029891 0.035300
```

The frequencies of the covertype categories show that the assumption of non-perfect seperation is satisfied.

3. Multicolinearity. Use cor() to find any mulicolinear variables and remove variables that have a pearson correlation coefficient above abs(0.5).

```
##                   elevation      aspect       slope vert_dist_hydro
## elevation        1.00000000  0.01573494 -0.24269664      0.09330644
## aspect           0.01573494  1.00000000  0.07872841      0.07030512
## slope           -0.24269664  0.07872841  1.00000000      0.27497568
## vert_dist_hydro  0.09330644  0.07030512  0.27497568      1.00000000
## horz_dist_road   0.36555928  0.02512069 -0.21591416     -0.04637197
```

```
## horz_dist_fire   0.14802156 -0.10917150 -0.18566195     -0.06991257
##                 horz_dist_road horz_dist_fire
## elevation           0.36555928     0.14802156
## aspect              0.02512069    -0.10917150
## slope              -0.21591416    -0.18566195
## vert_dist_hydro    -0.04637197    -0.06991257
## horz_dist_road      1.00000000     0.33157958
## horz_dist_fire      0.33157958     1.00000000
```

These variables had a pearson correlation coefficient above abs(0.5):

- vert_dist_hydro and horz_dist_hydro
- shade 9am and aspect
- slope and shade noon
- shade_9am and shade_3pm
- shade_noon and shade_3pm
- aspect and shade_3pm

I removed horz_dist_hydro, shade_3pm, shade_noon, and shade_9am for this reason.

4. Independence The Multinomial Regression Model assumes the Independence of Irrelevant Alternatives (IIA) which states that the model's choice between two outcomes (e.g. Covertype 1 or Covertype 2) is not impacted by the alternative options (https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.951.3160&rep=rep1&type=pdf). The IIA can be tested using the Hausman-McFadden Test which compares the full model to a reduced model without certain alternatives to determine if those selected alternatives are irrelevant or not. This test can be misinformative, so we will use it with caution.

The Cache wildtype area is more unique than the other wildtype areas according to the dataset information. This wildtype area is composed of primarily covertypes 3, 6, and 4. I will try removing these covertypes for the alternative model to see if they have an effect on the IIA.

```
##
##  Hausman-McFadden test
##
## data:  covdata4
## chisq = 100.49, df = 21, p-value = 2.361e-12
## alternative hypothesis: IIA is rejected
```

The Hausman-McFadden test rejects the IIA assumption for the full model.

Since we're suspicious of the Hausman-McFadden test, let's compare the coefficients of the two models, if their coefficients aren't significantly different, the IAA assumption holds (https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.951.3160&rep=rep1&type=pdf)

```
##     (Intercept):2     (Intercept):5     (Intercept):7       elevation:2
##      2.282110e+01      2.710300e+01     -5.458437e+01     -7.556348e-03
##       elevation:5       elevation:7          aspect:2          aspect:5
##     -9.975702e-03      1.604069e-02      6.565324e-04     -8.807940e-05
##          aspect:7           slope:2           slope:5           slope:7
##     -1.445593e-03     -1.537858e-02      1.327033e-02      1.319134e-02
## vert_dist_hydro:2 vert_dist_hydro:5 vert_dist_hydro:7  horz_dist_road:2
##      6.337816e-03      7.394365e-03     -6.007248e-03      7.015838e-05
##  horz_dist_road:5  horz_dist_road:7  horz_dist_fire:2  horz_dist_fire:5
##     -3.735155e-04      7.504210e-05      8.663803e-06     -1.774370e-04
##  horz_dist_fire:7
```

```
##      7.619822e-05
## attr(,"names.sup.coef")
## character(0)
## attr(,"fixed")
##     (Intercept):2     (Intercept):5     (Intercept):7       elevation:2
##             FALSE             FALSE             FALSE             FALSE
##       elevation:5       elevation:7          aspect:2          aspect:5
##             FALSE             FALSE             FALSE             FALSE
##          aspect:7           slope:2           slope:5           slope:7
##             FALSE             FALSE             FALSE             FALSE
## vert_dist_hydro:2 vert_dist_hydro:5 vert_dist_hydro:7  horz_dist_road:2
##             FALSE             FALSE             FALSE             FALSE
##  horz_dist_road:5  horz_dist_road:7  horz_dist_fire:2  horz_dist_fire:5
##             FALSE             FALSE             FALSE             FALSE
##  horz_dist_fire:7
##             FALSE
## attr(,"sup")
## character(0)
```

```
##     (Intercept):2     (Intercept):3     (Intercept):4     (Intercept):5
##      2.281378e+01      6.586145e+01      8.398127e+01      2.750529e+01
##     (Intercept):6     (Intercept):7       elevation:2       elevation:3
##      6.524793e+01     -5.458692e+01     -7.553361e-03     -2.479824e-02
##       elevation:4       elevation:5       elevation:6       elevation:7
##     -3.349503e-02     -1.011294e-02     -2.481999e-02      1.604154e-02
##          aspect:2          aspect:3          aspect:4          aspect:5
##      6.601025e-04      1.290574e-03     -1.674109e-03      3.442425e-05
##          aspect:6          aspect:7           slope:2           slope:3
##      2.075637e-03     -1.444762e-03     -1.545726e-02     -7.200134e-03
##           slope:4           slope:5           slope:6           slope:7
##     -7.874396e-02      1.372930e-02     -1.824206e-02      1.318081e-02
## vert_dist_hydro:2 vert_dist_hydro:3 vert_dist_hydro:4 vert_dist_hydro:5
##      6.333873e-03      1.748994e-02      1.512194e-02      7.615643e-03
## vert_dist_hydro:6 vert_dist_hydro:7  horz_dist_road:2  horz_dist_road:3
##      1.252636e-02     -6.007872e-03      6.897790e-05      6.893344e-04
##  horz_dist_road:4  horz_dist_road:5  horz_dist_road:6  horz_dist_road:7
##      1.839432e-03     -3.953039e-04      7.654432e-04      7.483549e-05
##  horz_dist_fire:2  horz_dist_fire:3  horz_dist_fire:4  horz_dist_fire:5
##      9.300507e-06     -5.954351e-04     -2.372502e-04     -1.853773e-04
##  horz_dist_fire:6  horz_dist_fire:7
##     -3.763945e-04      7.630036e-05
## attr(,"names.sup.coef")
## character(0)
## attr(,"fixed")
##     (Intercept):2     (Intercept):3     (Intercept):4     (Intercept):5
##             FALSE             FALSE             FALSE             FALSE
##     (Intercept):6     (Intercept):7       elevation:2       elevation:3
##             FALSE             FALSE             FALSE             FALSE
##       elevation:4       elevation:5       elevation:6       elevation:7
##             FALSE             FALSE             FALSE             FALSE
##          aspect:2          aspect:3          aspect:4          aspect:5
##             FALSE             FALSE             FALSE             FALSE
##          aspect:6          aspect:7           slope:2           slope:3
##             FALSE             FALSE             FALSE             FALSE
```

```
##           slope:4           slope:5           slope:6           slope:7
##             FALSE             FALSE             FALSE             FALSE
## vert_dist_hydro:2 vert_dist_hydro:3 vert_dist_hydro:4 vert_dist_hydro:5
##             FALSE             FALSE             FALSE             FALSE
## vert_dist_hydro:6 vert_dist_hydro:7  horz_dist_road:2  horz_dist_road:3
##             FALSE             FALSE             FALSE             FALSE
##  horz_dist_road:4  horz_dist_road:5  horz_dist_road:6  horz_dist_road:7
##             FALSE             FALSE             FALSE             FALSE
##  horz_dist_fire:2  horz_dist_fire:3  horz_dist_fire:4  horz_dist_fire:5
##             FALSE             FALSE             FALSE             FALSE
##  horz_dist_fire:6  horz_dist_fire:7
##             FALSE             FALSE
## attr(,"sup")
## character(0)
```

Coefficients	Full Model (mod_1)	Reduced (w/o 3, 4, 6)	Difference
(Intercept):2	2.281378e+01	2.282110e+01	-0.00732
(Intercept):5	2.750529e+01	2.710300e+01	0.40229
(Intercept):7	-5.458692e+01	-5.458437e+01	-0.00255
elevation:2	-7.553361e-03	-7.556348e-03	2.987e-06
elevation:5	-1.011294e-02	-9.975702e-03	-0.000137238
elevation:7	1.604154e-02	1.604069e-02	-8.5e-07
aspect:2	6.601025e-04	6.565324e-04	3.5701e-06
aspect:5	3.442425e-05	-8.807940e-05	0.0001225036
aspect:7	-1.444762e-03	-1.445593e-03	8.31e-07
slope:2	-1.545726e-02	-1.537858e-02	-7.868e-05
slope:5	1.372930e-02	1.327033e-02	0.00045897
slope:7	1.318081e-02	1.319134e-02	-1.053e-05
vert_dist_hydro:2	6.333873e-03	6.337816e-03	-3.943e-06
vert_dist_hydro:5	7.615643e-03	7.394365e-03	0.000221278
vert_dist_hydro:7	-6.007872e-03	-6.007248e-03	-6.24e-07
horz_dist_road:2	6.897790e-05	7.015838e-05	-1.18048e-06
horz_dist_road:5	-3.953039e-04	-3.735155e-04	-2.17884e-05
horz_dist_road:7	7.483549e-05	7.504210e-05	-2.0661e-07
horz_dist_fire:2	9.300507e-06	8.663803e-06	6.36704e-07
horz_dist_fire:5	-1.853773e-04	-1.774370e-04	-7.9403e-06
horz_dist_fire:7	7.630036e-05	7.619822e-05	1.0214e-07

Most of the model coefficients seem very similar. The largest difference by far is for (Intercept):5 (0.40229). Among non-intercept coefficients, the largest difference is for slope:5 which has a difference of 0.00045897. We can look at their AIC values but obviously mod_2 will be much lower because it has less parameters.

```
AIC(mod_1)
```

```
## [1] 860991.6
```

```
AIC(mod_2)
```

```
## [1] 687812
```

Since tests for IIA are often scrutinized in the literature and the coefficients from the full and reduced models seem fairly similar. I will assume that the IIA is fulfilled in the full model and use all of the covertype categories for my model.

Computer Code for Questions of Interest

Does elevation have the strongest effect on covertype? Let's use the Wald's test to test for the significance of elevation in predicting covertype.

Null hypothesis, H_0: $\beta_{elevation} = 0$, the coefficient for elevation is 0 Alternative hypothesis, H_1: $\beta_{elevation} \neq 0$, the coefficient for elevation is not 0

```
wald.test(Sigma = vcov(mod_1), b = coef(mod_1), Terms = 1:6)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 160050.7, df = 6, P(> X2) = 0.0
```

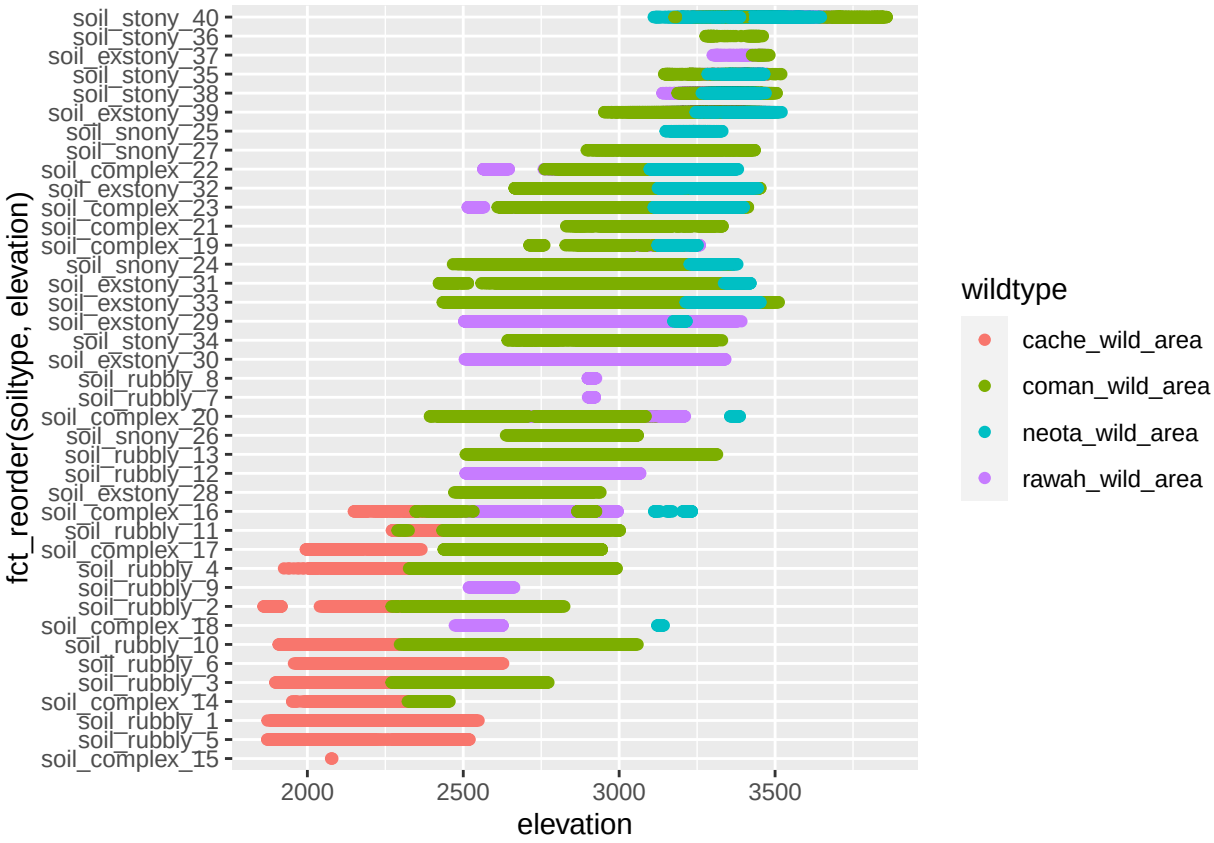
At a significance level of 0.05, we have significant evidence to reject the null hypothesis that the coefficient for elevation is zero meaning that elevation is an important predictor in our model (Wald-Test, p-value = 0.0).

```
# select the variables we used in our model plus the target (covertype)
tes_pred <- data.frame(test_data) %>% select(c(elevation,
                                               aspect,
                                               slope,
                                               vert_dist_hydro,
                                               horz_dist_road,
                                               horz_dist_fire,
                                               covertype))

# Get the Mean Squared Prediction Error
# https://stackoverflow.com/questions/64635547/mean-square-
# prediction-error-mspe-on-out-of-sample-data-in-r
mspe <- function(model, dv, data) {
  yhat <- predict(model, newdata=data)
  y <- data[[dv]]
  mean((y - yhat)^2)
}

# Get MSPE for training model for the covertype response on the testing data
# closer to 0 is better
mspe(mod_train, "covertype", data.frame(tes_pred))
```

Can we accurately predict cover type based on the given variables?

```
## [1] 0.1827033
```

Plots



It seems that elevation and soil type are correlated: rubbly soil types occur at lower elevations whereas stony and extremely stony soil types occur at higher elevations and complex soiltypes seem to occur throughout the elevations. We can also see the distribution of the different wilderness areas. Coman wild area is fairly distributed at all elevations whereas cache is at the lowest elevation and neota is at the highest. Rowah appears pretty much in the middle.