# Prestige Data Assignment

*Sergazy*

*9/23/2018*

(a) Loading the Prestige data and omitting the missing values, we have n = 98 and p = 5 for our first model which will include only the main effects.

```
library("carData")
library("car")
data = na.omit(Prestige)
prestige = data$prestige
education = data$education
income = data$income
type = data$type
fit = lm(prestige~education+income+type)
summary(fit)
```

```
##
## Call:
## lm(formula = prestige ~ education + income + type)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.9529  -4.4486   0.1678   5.0566  18.6320
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.6229292  5.2275255  -0.119    0.905
## education    3.6731661  0.6405016   5.735 1.21e-07 ***
## income       0.0010132  0.0002209   4.586 1.40e-05 ***
## typeprof     6.0389707  3.8668551   1.562    0.122
## typewc      -2.7372307  2.5139324  -1.089    0.279
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.095 on 93 degrees of freedom
## Multiple R-squared:  0.8349, Adjusted R-squared:  0.8278
## F-statistic: 117.5 on 4 and 93 DF,  p-value: < 2.2e-16
```

Now, we'll test the hypothesis that the effect on prestige of an occupation being white collar is the same as the effect on prestige of the occupation being professional. We'll use a significance level of 5%. To do this, we'll consider the model where type has only two levels, blue collar, and 1 white collar/professional. We'll combine the two levels "wc" and "prof" into a single combined level "wcProf". In doing this, we lose one degree of freedom so q = 1. We can use an F test to determine if the impact on prestige of an occupation being White Collar is the same as the effect on prestige of an occupation being Professional.

```
# Create a new factor with "wc" and "prof" combined into one level
combinedType = type
levels(combinedType)[levels(combinedType)=="wc"] = "wcProf"
levels(combinedType)[levels(combinedType)=="prof"] = "wcProf"
fitComb = lm(prestige~education+income+combinedType)
summary(fitComb)
```

```
##
## Call:
## lm(formula = prestige ~ education + income + combinedType)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.3127  -4.5080  -0.3374   5.1503  15.0086
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1.118e+01  4.215e+00  -2.653  0.00937 **
## education            4.836e+00  5.494e-01   8.803 6.42e-14 ***
## income               1.169e-03  2.255e-04   5.183 1.25e-06 ***
## combinedTypewcProf -3.283e+00  2.626e+00  -1.250  0.21433
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.428 on 94 degrees of freedom
## Multiple R-squared:  0.817,  Adjusted R-squared:  0.8112
## F-statistic: 139.9 on 3 and 94 DF,  p-value: < 2.2e-16
```

```
SSres = sum(fit$residuals^2)
SSresH0 = sum(fitComb$residuals^2)
F = ((SSresH0 - SSres) / (SSres)) * ((98-5)/(1))
1-pf(F, 1, 98-5)
```

```
## [1] 0.002082347
```

Our P-value is 0.0021, so we reject the null hypothesis that the effect on prestige is the same for both white collar and professional occupations. We have significant evidence that the effects are different, and we prefer our original model. We can alternatively ask R to run this hypothesis test, and we will get the exact same result:

```
linearHypothesis(fit, "typeprof = typewc")
```

```
## Linear hypothesis test
##
## Hypothesis:
## typeprof - typewc = 0
##
## Model 1: restricted model
## Model 2: prestige ~ education + income + type
##
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     94 5186.2
## 2     93 4681.3  1    504.93 10.031 0.002082 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The P-value matches our value from the calculation. (b) Now let's test the hypothesis that type has no effect on the prestige of an occupation. We'll make a new fit corresponding to our null hypothesis which excludes type. Then we'll perform an F test with a significance level of 5%. In this case q = 2 since we lose two degrees of freedom by reducing to the null hypothesis model.

```
fitNoType = lm(prestige~education+income)
summary(fitNoType)
```

```
##
## Call:
## lm(formula = prestige ~ education + income)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.9367  -4.8881   0.0116   4.9690  15.9280
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.6210352  3.1162309  -2.446   0.0163 *
## education    4.2921076  0.3360645  12.772  < 2e-16 ***
## income       0.0012415  0.0002185   5.682 1.45e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.45 on 95 degrees of freedom
## Multiple R-squared:  0.814,  Adjusted R-squared:  0.8101
## F-statistic: 207.9 on 2 and 95 DF,  p-value: < 2.2e-16
```

```
SSres = sum(fit$residuals^2)
SSresH0 = sum(fitNoType$residuals^2)
F = ((SSresH0 - SSres) / (SSres)) * ((98-5)/(2))
1-pf(F, 2, 98-5)
```

```
## [1] 0.003966438
```

Since our P-value is 0.0040 we reject the null hypothesis. We have statistically significant evidence that the type does have an effect on the prestige of an occupation. We can verify our result with R's calculation:

```
linearHypothesis(fit, c("typeprof = 0", "typewc = 0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## typeprof = 0
## typewc = 0
##
## Model 1: restricted model
## Model 2: prestige ~ education + income + type
##
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     95 5272.4
## 2     93 4681.3  2    591.16 5.8721 0.003966 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Again the P-value matches exactly. Next let's test the hypothesis that the regression constant for income is 0.002 and the regression constant for education is 2. In this case q = 2 when we fix both the education and income coefficients.

```
fitH0 = lm(prestige-2*education -0.002*income ~ type)
summary(fitH0)
```

```
##
## Call:
## lm(formula = prestige - 2 * education - 0.002 * income ~ type)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.7391  -4.5199   0.4978   4.9426  21.4646
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    8.060      1.175   6.860 6.98e-10 ***
## typeprof      10.501      1.828   5.745 1.10e-07 ***
## typewc         2.035      2.006   1.015    0.313
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.794 on 95 degrees of freedom
## Multiple R-squared:  0.2667, Adjusted R-squared:  0.2512
## F-statistic: 17.27 on 2 and 95 DF,  p-value: 3.995e-07
```

```
SSres = sum(fit$residuals^2)
SSresH0 = sum(fitH0$residuals^2)
F = ((SSresH0 - SSres) / (SSres)) * ((98-5)/(2))
1-pf(F, 2, 98-5)
```

```
## [1] 5.910216e-05
```

Our P-value is $< 0.0001$, so we reject the null hypothesis and conclude that the pair of regression coefficients for income and education is not 0.002, 2. Again we can check our value with R's calculation:

```
linearHypothesis(fit, c("education = 2", "income = 0.002"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## education = 2
## income = 0.002
##
## Model 1: restricted model
## Model 2: prestige ~ education + income + type
##
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     95 5771.6
## 2     93 4681.3  2    1090.3 10.831 5.91e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

And again the P-value matches our calculation. (c) Let's now consider the model where income and education are allowed to have interactions with type.

```
intFit = lm(prestige~type*(education+income))
summary(intFit)
```

```
##
## Call:
## lm(formula = prestige ~ type * (education + income))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.462  -4.225   1.346   3.826  19.631
```

```
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.276e+00  7.057e+00   0.323   0.7478
## typeprof             1.535e+01  1.372e+01   1.119   0.2660
## typewc              -3.354e+01  1.765e+01  -1.900   0.0607 .
## education            1.713e+00  9.572e-01   1.790   0.0769 .
## income               3.522e-03  5.563e-04   6.332 9.62e-09 ***
## typeprof:education   1.388e+00  1.289e+00   1.077   0.2844
## typewc:education     4.291e+00  1.757e+00   2.442   0.0166 *
## typeprof:income     -2.903e-03  5.989e-04  -4.847 5.28e-06 ***
## typewc:income       -2.072e-03  8.940e-04  -2.318   0.0228 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.318 on 89 degrees of freedom
## Multiple R-squared:  0.8747, Adjusted R-squared:  0.8634
## F-statistic: 77.64 on 8 and 89 DF,  p-value: < 2.2e-16
```

I prefer this model to the model we used in part (a) and (b), because it allows us to see the interactions, several of which are significant at the 5% level. For example, this model suggests that higher income makes a greater impact on the prestige of blue collar occupations than it does for white collar occupations and professional occupations. This perhaps makes intuitive sense to us. Blue collar occupations typically have low prestige, but higher paying blue collar occupations tend to have higher prestige (for example Tool Die Makers in this data set). Professional occupations on the other hand tend to have higher prestige overall, and even some of the lower paying professional occupations still have high prestige (for example Pychologists and Chemists in this data set).

(d) In this model with the interactions, we'll test the hypothesis that income has no effect on prestige for white collar occupations. We will do this with a T test as described in the course notes, being careful to add the vectors for "income" and "typewc:income" to get the effective coefficient for income for white collar occupations.

```
effCoef = intFit$coef[["income"]] + intFit$coef[["typewc:income"]]
cSigma = (vcov(intFit)["income", "income"]
+ vcov(intFit)["typewc:income", "typewc:income"]
+ 2*vcov(intFit)["income", "typewc:income"])
t = (effCoef) / sqrt(cSigma)
2*(1-pt(t, 89)) # n - p = 89
```

```
## [1] 0.04109131
```

Since our P-value is 0.0411, we reject the null hypothesis and conclude that income does have an effect on prestige for white collar occupations. We can compare with R's calculation:

```
linearHypothesis(intFit, 'typewc:income + income = 0')
```

```
## Linear hypothesis test
##
## Hypothesis:
## income  + typewc:income = 0
##
## Model 1: restricted model
## Model 2: prestige ~ type * (education + income)
##
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     90 3724.4
```

```
## 2     89 3552.9  1     171.5 4.2961 0.04109 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Again the P-value matches exactly with our calculations. (e) We will use our model from (c) to predict the prestige of an occupation that has an average income of 5000$, average education of 12, and is a white collar occupation.

```
newjob = c(1, 0, 1, 12, 5000, 0, 12, 0, 5000)
estPrestige = sum(intFit$coefficients * newjob)
estPrestige
```

```
## [1] 48.04124
```

Based on our linear model in (c), we estimate the prestige to be 48.04 for this occupation. (f) Next we'll test the null hypothesis that in the model in (c) the interactions between education and type are not significant, again at a 5% significance level. In the reduced model we lose 2 degrees of freedom, so q = 2.

```
fitH0 = lm(prestige~type*income + education)
summary(fitH0)
```

```
##
## Call:
## lm(formula = prestige ~ type * income + education)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.8720  -4.8321   0.8534   4.1425  19.6710
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -6.7272633  4.9515480  -1.359   0.1776
## typeprof        25.1723873  5.4669586   4.604 1.34e-05 ***
## typewc           7.1375093  5.2898177   1.349   0.1806
## income           0.0031344  0.0005215   6.010 3.79e-08 ***
## education        3.0396961  0.6003699   5.063 2.14e-06 ***
## typeprof:income -0.0025102  0.0005530  -4.539 1.72e-05 ***
## typewc:income   -0.0014856  0.0008720  -1.704   0.0919 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.455 on 91 degrees of freedom
## Multiple R-squared:  0.8663, Adjusted R-squared:  0.8574
## F-statistic: 98.23 on 6 and 91 DF,  p-value: < 2.2e-16
```

```
SSres = sum(intFit$residuals^2)
SSresH0 = sum(fitH0$residuals^2)
F = ((SSresH0 - SSres) / (SSres)) * ((98-9)/(2))
1-pf(F, 2, 98-9)
```

```
## [1] 0.0555744
```

Our P-value is 0.0556 so we fail to reject the null hypothesis. We do not have statistically significant evidence that the interactions between education and type are not zero. We can compare with R's calculation:

```
linearHypothesis(intFit, c("typeprof:education = 0", "typewc:education = 0"))
```

```
## Linear hypothesis test
```

```
##
## Hypothesis:
## typeprof:education = 0
## typewc:education = 0
##
## Model 1: restricted model
## Model 2: prestige ~ type * (education + income)
##
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     91 3791.3
## 2     89 3552.9  2     238.4 2.9859 0.05557 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We get the exact same P-value. (g) First we'll fix the model to simulate from, and define a function in R to simulate the new prestige values.

```
interc = fit$coef[["(Intercept)"]]
educ = fit$coef[["education"]]
inc = fit$coef[["income"]]
prof = fit$coef[["typeprof"]]
wc = fit$coef[["typewc"]]
stderr = 7.095
bcdata = subset(data, data$type == "bc")
wcdata = subset(data, data$type == "wc")
profdata = subset(data, data$type == "prof")
simulateData = function() {
simPrestigeBC = (interc + bcdata$education*educ + bcdata$income*inc
+ rnorm(44, mean=0, sd=stderr))
simPrestigeWC = (interc + wcdata$education*educ + wcdata$income*inc
+ wc + rnorm(23, mean=0, sd=stderr))
simPrestigeProf = (interc + profdata$education*educ + profdata$income*inc
+ prof + rnorm(31, mean=0, sd=stderr))
simData = data
simData$prestige = c(simPrestigeBC, simPrestigeWC, simPrestigeProf)
return(simData)
}
```

Next we'll define a function to run Alice's method:

```
alice = function(simData) {
aFit = lm(simData$prestige ~ simData$type * (simData$education + simData$income))
mm = model.matrix(aFit)
while(TRUE) {
aFit = lm(simData$prestige ~ mm - 1) # -1 removes the NA intercept
remainingCoef = names(aFit$coef)
highestPval = 0
highestindex = 0
index = 0
for (coef in remainingCoef) {
index = index + 1
Pval = linearHypothesis(aFit, paste(coef, " = 0"))$`Pr(>F)`[2]
if (Pval > 0.05 && Pval > highestPval) {
highestPval = Pval
highest = coef
highestindex = index
```

```
}
}
if (highestindex == 0) {
break # nothing left to remove
}
mm = mm[,-highestindex]
}
# summary(aFit)
return(aFit)
}
```

Next, the function to run Bob's method:

```
bob = function(simData) {
interactionFit = lm(simData$prestige ~ simData$type
* (simData$education + simData$income))
mm = model.matrix(interactionFit)
fullFit = lm(simData$prestige ~ mm - 1)
includedCovariates = c()
coefNames = names(fullFit$coef)
while(TRUE) {
bestPval = 1
bestIndex = 0
for (i in 1:9) {
if (!(i %in% includedCovariates)) {
covToTry = c(includedCovariates, i)
mmTry = mm[,covToTry]
bFit = lm(simData$prestige ~ mmTry - 1)
coef = names(bFit$coef)[length(includedCovariates)+1]
Pval = linearHypothesis(bFit, paste(coef, " = 0"))$`Pr(>F)`[2]
if (Pval < 0.05 && Pval < bestPval) {
bestPval = Pval
bestIndex = i
}
}
}
if (bestIndex == 0) {
break # Nothing left to add
}
includedCovariates = c(includedCovariates, bestIndex)
}
mm = mm[,includedCovariates]
bFit = lm(simData$prestige ~ mm - 1)
return (bFit)
}
```

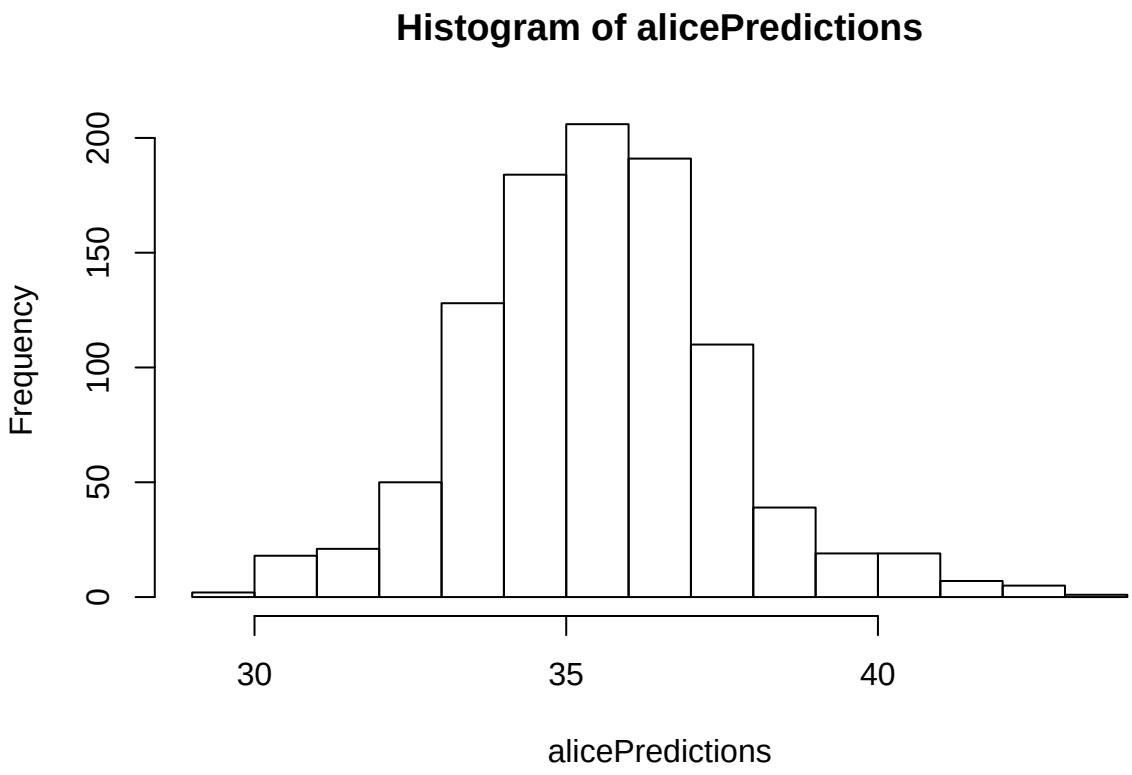
One more function which will take as an input one of the models generated by Alice's or Bob's method and return the predicted prestige for the occupation in part (e).

```
predictPrestige = function(model) {
modelInterc = 0
if ("mm(Intercept)" %in% names(model$coef)) {
modelInterc = model$coef[["mm(Intercept)"]]
}
modelWC = 0
```

```
if ("mmsimData$typewc" %in% names(model$coef)) {
modelWC = model$coef[["mmsimData$typewc"]]
}
modelEducation = 0
if ("mmsimData$education" %in% names(model$coef)) {
modelEducation = modelEducation + model$coef[["mmsimData$education"]]
}
if ("mmsimData$typewc:simData$education" %in% names(model$coef)) {
modelEducation = (modelEducation
+ model$coef[["mmsimData$typewc:simData$education"]])
}
modelIncome = 0
if ("mmsimData$income" %in% names(model$coef)) {
modelIncome = modelIncome + model$coef[["mmsimData$income"]]
}
if ("mmsimData$typewc:simData$income" %in% names(model$coef)) {
modelIncome = modelIncome + model$coef[["mmsimData$typewc:simData$income"]]
}
return (modelInterc + modelWC + modelEducation*12 + modelIncome*5000)
}
```

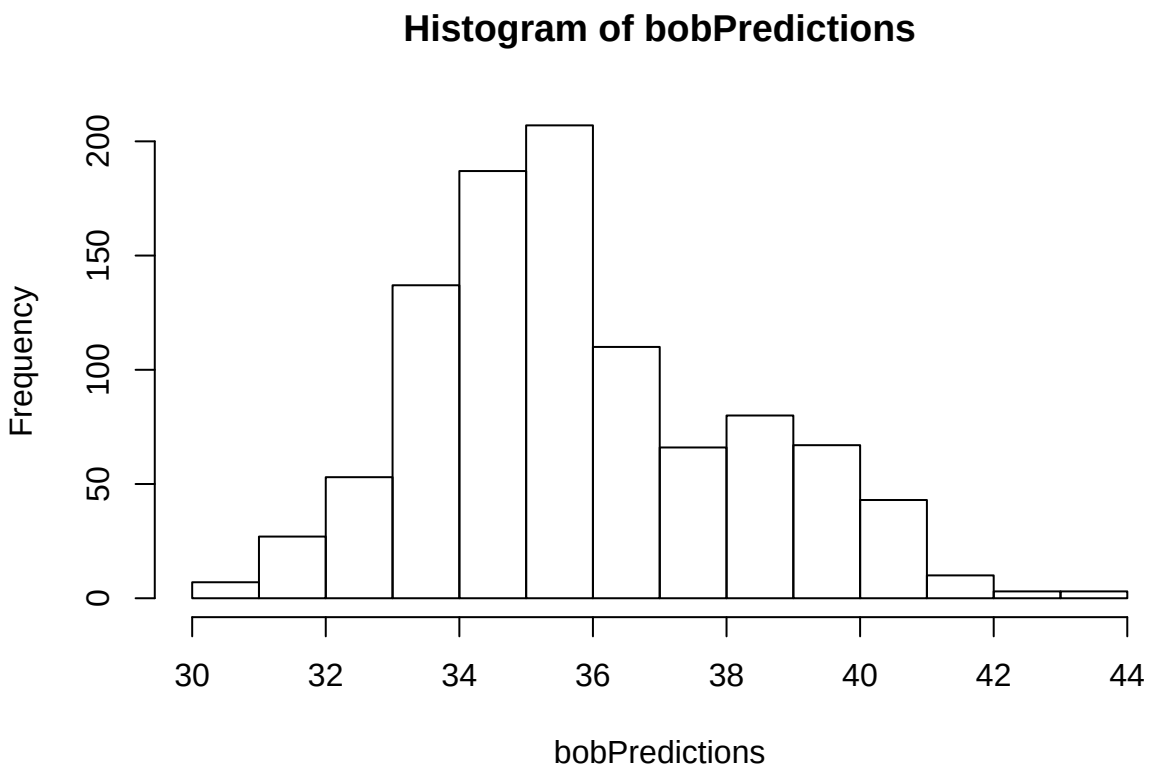
Now we're ready to simulate! Let's start with Alice's method:

```
alicePredictions = replicate(1000, predictPrestige(alice(simulateData())))
hist(alicePredictions)
```




Now we'll simulate from Bob's method:

```
bobPredictions = replicate(1000, predictPrestige(bob(simulateData())))
hist(bobPredictions)
```



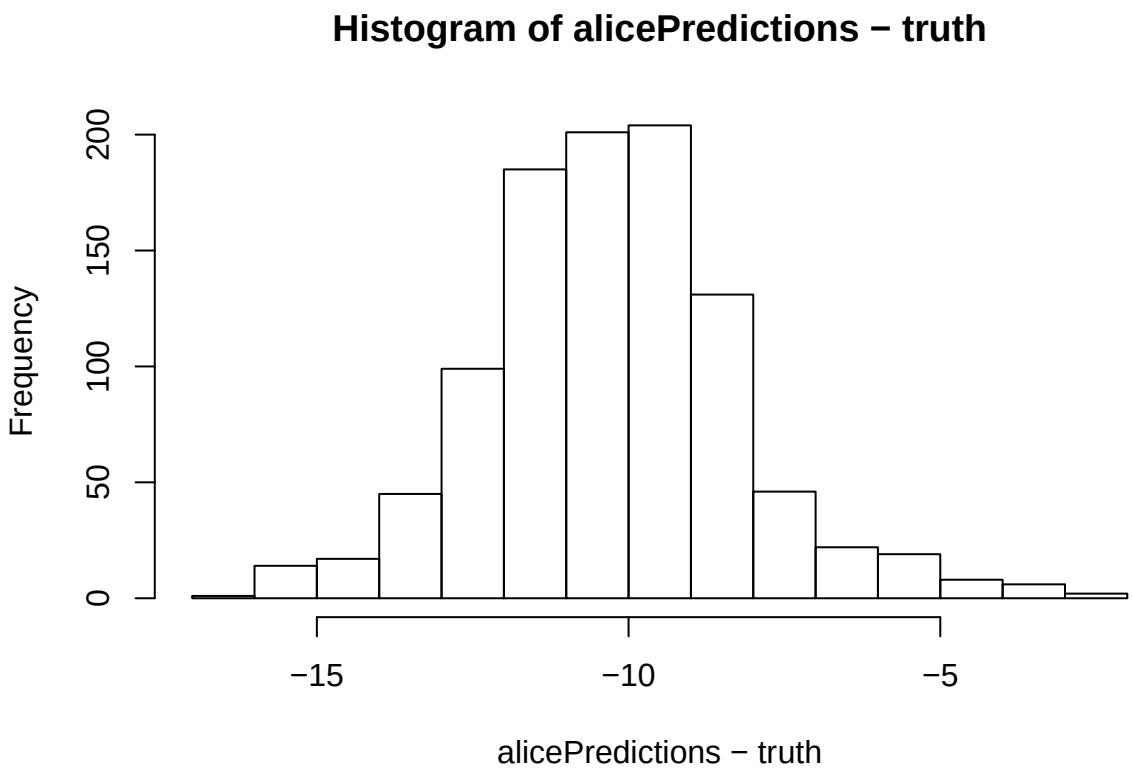

The true value from the model we simulated from is

```
truth = interc + wc + inc*5000 + educ*12
truth
```

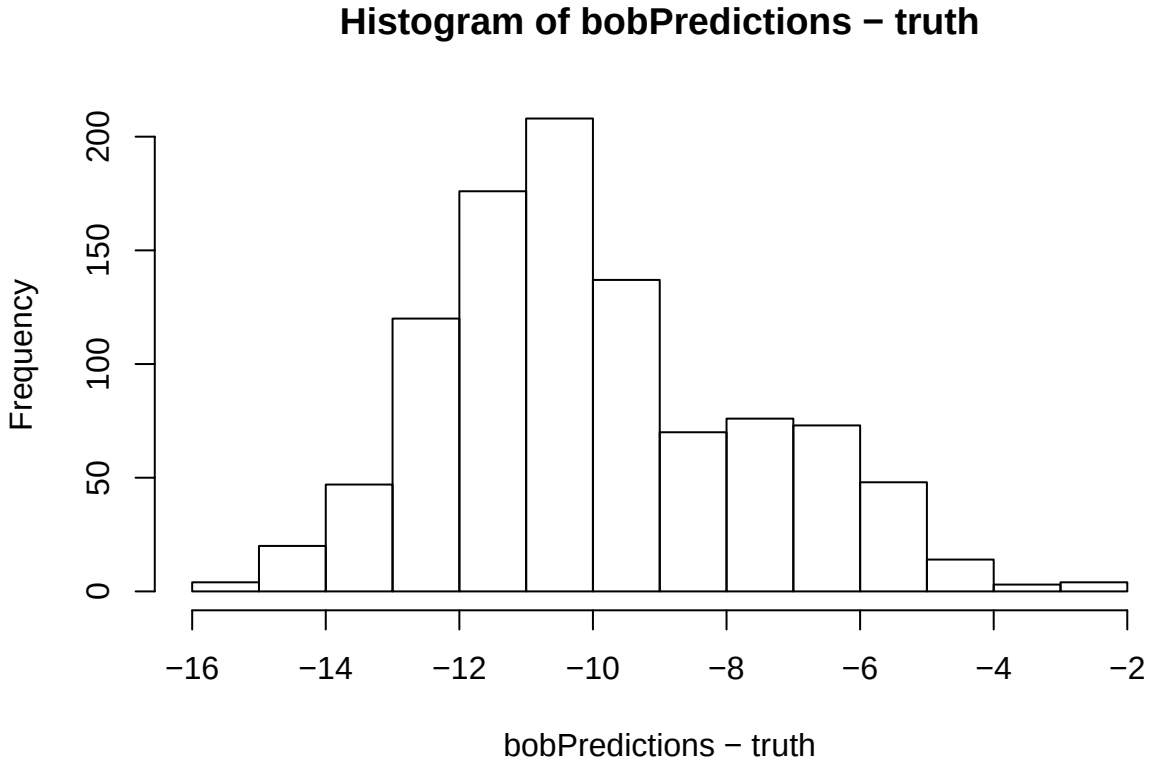
```
## [1] 45.7838
```

Let's plot the residuals for this new occupation under each method of calculation

```
hist(alicePredictions - truth)
```




```
hist(bobPredictions - truth)
```




We see there are some subtle differences between the two methods. Both methods bias about 10 below the occupations true prestige (based on the model we simulated from). Bob's method has a higher variance, while Alice's method had a lower variance and is more symmetric. Due to the higher variance Bob's method

occassionally gets better results, but overall Alice's method performs better on average for this prediction based on (e).