LM HW 6

Joshua Ingram

3/30/2020

Problem 1

1.

Firstly, we know that $E[\epsilon_i] = 0, Var[\epsilon_i] = \sigma^2,$ and $\epsilon_i \sim N(0, \sigma^2)$

We then find that $E[y_i] = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i}$ by:

$E[y_i] = E[\alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \epsilon_i] = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + E[\epsilon_i] = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i}$

We also find that $Var[y_i] = \sigma^2$ by:

$Var[y_i] = Var[\alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \epsilon_i] = Var[\epsilon_i] = \sigma^2$

Finally, to show that $y_i \sim N(\alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i}, \sigma^2)$ we need to look at the distribution of our error term, ϵ_i:

$\epsilon_i \sim N(0, \sigma^2)$

$\epsilon_i = y_i - (\alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i})$

$y_i - (\alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i}) \sim N(0, \sigma^2)$

From earlier steps above, we showed that $E[y_i] = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i}$ and it's variance, σ^2.

Since our error has a normal distribution centered at zero, to find the distribution of y_i, we shift the normal distribution of our error term by $\alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i}$ so it will be centered at $E[y_i]$. The variance of y_i is still σ^2 and we have fixed values of x. Thus:

$y_i \sim N(\alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i}, \sigma^2)$

2.

In order to conduct a single test to determine whether either *under-18* or *urban* is significant, we would use the incremental F-test. To conduct this test, in genreal terms, we would set our hypotheses, H_0 being that our betas added in the full model are equal to zero and H_a being that at least one is not equal to zero. We would then calculate our F-statistic, compare this to it's F-distribution with q and n - (k+1) degrees of freedom, then find the p-value by calculating the area under the curve to the right of our F-statistic.

Test for significance of either under-18 or urban Null Model: $education_i = \alpha + \beta_1 income_i + \epsilon_i$

Full Model: $education_i = \alpha + \beta_1 income_i + \beta_2 under18_i + \beta_3 urban_i + \epsilon_i$

$H_0 : \beta_2 = \beta_3 = 0$

$H_a : \{\exists \beta_j \neq 0 | j = 2, 3\}$

```r
# creating the null and full models using the lm function
null_lm <- lm(education ~ income, data = anscombe)
full_lm <- lm(education ~ income + under18 + urban, data = anscombe)

# anova() to calculate the f-stat and p-value for the incremental f test
anova(null_lm, full_lm)
```

```
## Analysis of Variance Table
##
## Model 1: education ~ income
## Model 2: education ~ income + under18 + urban
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1     49 59814
## 2     47 33489  2     26325 18.472 1.204e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After creating the null and full model using the lm() function, we use the anova() function to perform the incremental f-test. Our numerator degrees of is 2 and the difference between $RegSS_1$ and $RegSS_0$ is 26,325. Our denominator df is 47 and RSS_1 is 33,489. We use these values to find our F-statistic of 18.472. We recieve a p-value of $1.204e^{-6}$ under the F distribution with 2 and 47 degreees of freedom. This is gives us significant evidence to reject the null hypothesis in favor of the alternative, that under18 and/or urban play a signifcant role in predicting education.

Problem 2

Given $y_i = \alpha + \gamma x_{sex,i} + \epsilon_i$, where i = 1 indicates male and i = 0 indicates female, we can find α and $\alpha + \gamma$

For $alpha$:

$\frac{\delta}{\delta\alpha} \sum_i (y_i - (\alpha + \gamma x_{sex,i}))^2 = \sum_i -2(y_i - (\alpha + \gamma x_{sex,i}))$

$\sum_i -2(y_i - (\alpha + \gamma x_{sex,i})) = 0$

$\sum_i y_i - \sum_i \alpha - \sum_i \gamma x_{sex,i} = 0$

$\sum_i \alpha = \sum_i y_i - \gamma \sum_i x_{sex,i}$

$n\alpha = \sum_i y_i - \gamma \sum_i x_{sex,i}$

$\hat{\alpha} = \bar{y_i} - \gamma \bar{x}_{sex,i}$

For γ:

$\frac{\delta}{\delta\gamma} \sum_i (y_i - (\alpha + \gamma x_{sex,i}))^2 = \sum_i -2x_{sex,i}(y_i - (\alpha + \gamma x_{sex,i}))$

$\sum_i -2x_{sex,i}(y_i - (\alpha + \gamma x_{sex,i})) = 0$

$\sum_i x_{sex,i}(y_i - (\alpha + \gamma x_{sex,i})) = 0$

$\sum_i x_{sex,i} y_i - \sum_i \alpha x_{sex,i} - \sum_i \gamma x_{sex,i}^2 = 0$

$\sum_i x_{sex,i} y_i - \alpha \sum_i x_{sex,i} - \gamma \sum_i x_{sex,i}^2 = 0$

$\sum_i x_{sex,i} y_i - \alpha \sum_i x_{sex,i} = \gamma \sum_i x_{sex,i}^2$

$\hat{\gamma} = \frac{\sum_i x_{sex,i} y_i - \alpha \sum_i x_{sex,i}}{\sum_i x_{sex,i}^2}$

Using the formula found for α, we show that $al\hat{p}ha$ is equal to the average height for females. Further, $\hat{\alpha} + \hat{\gamma}$ is equal to the average heightt among males.

Problem 3

1.

$$moralIntegration_i = \alpha + \beta_1 heterogeneity_i + \beta_2 mobility_i + \gamma_1 D_{MW,i} + \gamma_2 D_{S,i} + \gamma_3 D_{W,i} + \epsilon_i$$

2.

```r
moral_lm <- lm(moralIntegration ~ heterogeneity + mobility + region, data = angell)
moral_lm$coefficients
```

```
##   (Intercept) heterogeneity      mobility      regionMW       regionS 
##   17.99789250   -0.06647088   -0.06728923   -2.52496230   -4.58939053 
##       regionW 
##   -3.86451557
```

$\hat{moralIntegration}_i = 17.998 - 0.0665(heterogeneity_i) - 0.067(mobility_i) - 2.525D_{MW,i} - 4.589D_{S,i} - 3.865D_{W,i}$

3.

$H_0 : \beta_1 = \beta_2 = \gamma_1 = \gamma_2 = \gamma_3 = 0$

H_a : At least one coefficient (β_i or γ_i) $\neq 0$

```r
summary(moral_lm)
```

```
## 
## Call:
## lm(formula = moralIntegration ~ heterogeneity + mobility + region, 
##     data = angell)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.9694 -1.5521 -0.0009  1.5875  3.3807 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   17.99789    1.58947  11.323  1.4e-13 ***
## heterogeneity -0.06647    0.02839  -2.341   0.0247 *  
## mobility      -0.06729    0.05923  -1.136   0.2632    
## regionMW      -2.52496    1.07556  -2.348   0.0244 *  
## regionS       -4.58939    1.83169  -2.506   0.0168 *  
## regionW       -3.86452    1.62527  -2.378   0.0227 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.128 on 37 degrees of freedom
## Multiple R-squared:  0.6874, Adjusted R-squared:  0.6452 
## F-statistic: 16.27 on 5 and 37 DF,  p-value: 1.768e-08
```

Looking at the output for our F-test, we receive an f-statistic of 16.27 and p-value of $1.768e^{-8}$. We reject our null hypothesis in favor of the alternative. Our model is significant.

4.

Heterogeneity, regionMW, regionS, and regionW were all significant regressors according to the t-test.

Interpretations:

heterogeneity: For every one percentage point increase in heterogeneity, we expect to see, on average, a 0.066 unit (composite of crime rate and welfare expenditures) decrease in moralIntegration, holding all else constant.

regionMW: On average, we expect the MidWest to have a moralintegration value of 2.525 units lower than that of the northeast, holding all else constant.

regionS: On average, we expect the Southeat to have a moralintegration value of 4.589 units lower than that of the northeast, holding all else constant.

regionW: On average, we expect the West to have a moralintegration value of 3.865 units lower than that of the northeast, holding all else constant.

5.

I did not add the α and γ below to show where the difference comes (when dummy variable for the specified region = 1)

E region:

$moralIn\hat{t}egration_i = 17.998 - 0.0665(heterogeneity_i) - 0.067(mobility_i)$

MW region:

$moralIn\hat{t}egration_i = 17.998 - 2.525 - 0.0665(heterogeneity_i) - 0.067(mobility_i)$

S region:

$moralIn\hat{t}egration_i = 17.998 - 4.589 - 0.0665(heterogeneity_i) - 0.067(mobility_i)$

W region:

$moralIn\hat{t}egration_i = 17.998 - 3.865 - 0.0665(heterogeneity_i) - 0.067(mobility_i)$

If this were graphed, it would be 4 2-D planes that are parallel to eachother with differences of the gamma values from the Eastern region plane

6.

7.

An incremental F-test would be most appropriate here.

$H_0 : \gamma_1 = \gamma_2 = \gamma_3 = 0$

$H_a : \{\exists \gamma_j \neq 0 | j = 1, 2, 3\}$

```
# creating the null and full models using the lm function
null_lm2 <- lm(moralIntegration ~ heterogeneity + mobility, data = angell)
full_lm2 <- lm(moralIntegration ~ heterogeneity + mobility + region, data = angell)

# anova() to calculate the f-stat and p-value for the incremental f test
anova(null_lm2, full_lm2)
```

```
## Analysis of Variance Table
##
## Model 1: moralIntegration ~ heterogeneity + mobility
## Model 2: moralIntegration ~ heterogeneity + mobility + region
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     40 201.27
## 2     37 167.49  3    33.774 2.4869 0.07562 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We recieve an F-statistic of 2.4869 and p-value of 0.07562. We do not have significant evidence to reject the null and may conlcude that region is not a significant predictor of moralIntegration.

8.

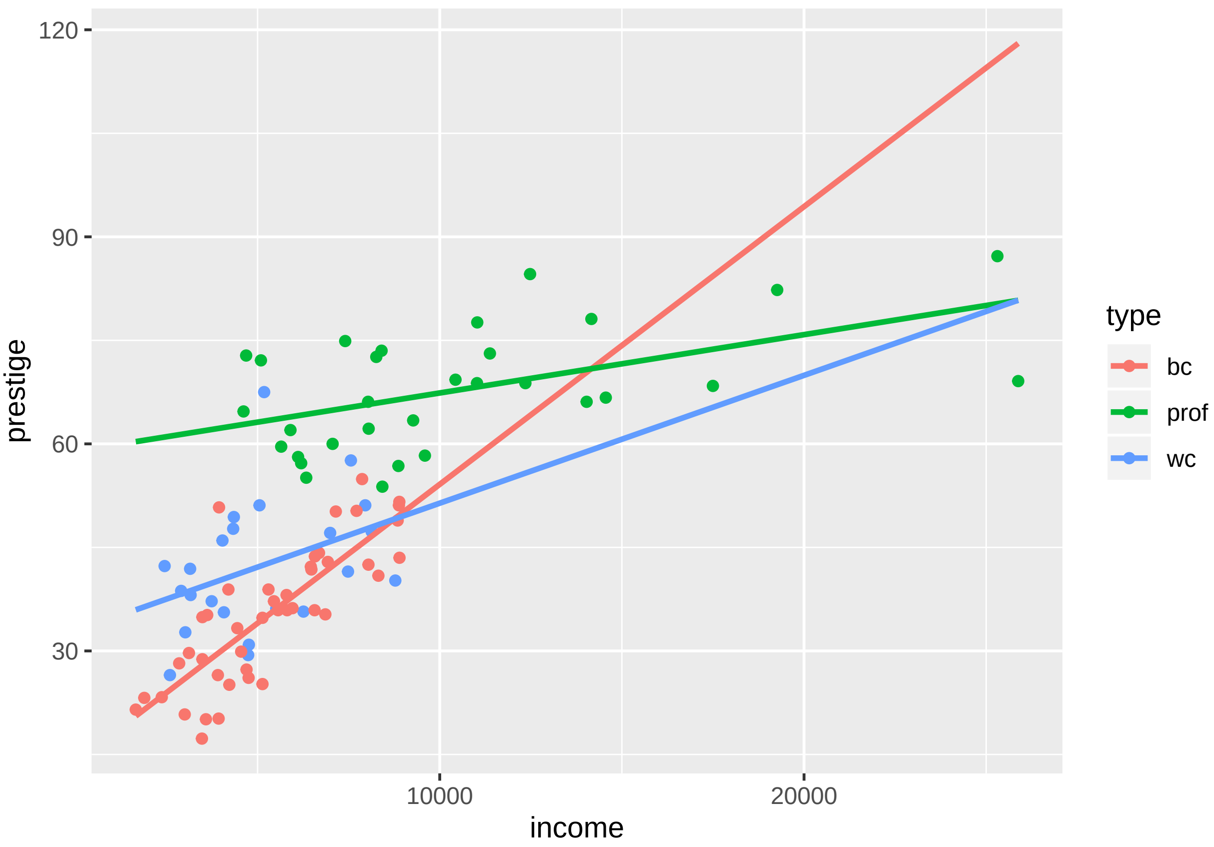
The method of using an incremental F-test to determine the significance of an individual variable in a model is more appropriate than looking at the individual t-tests. When we rely on the t-test to see if an individual predictor is significant, we run into type 1 errors, saying the predictor is significant, when it is not. By using an incremental F-test, we go around the issue of increasing chance of type 1 errors and look at the significance of the variables of interest in predicting the response by comparing a null and full model.

Problem 4

1.

a.

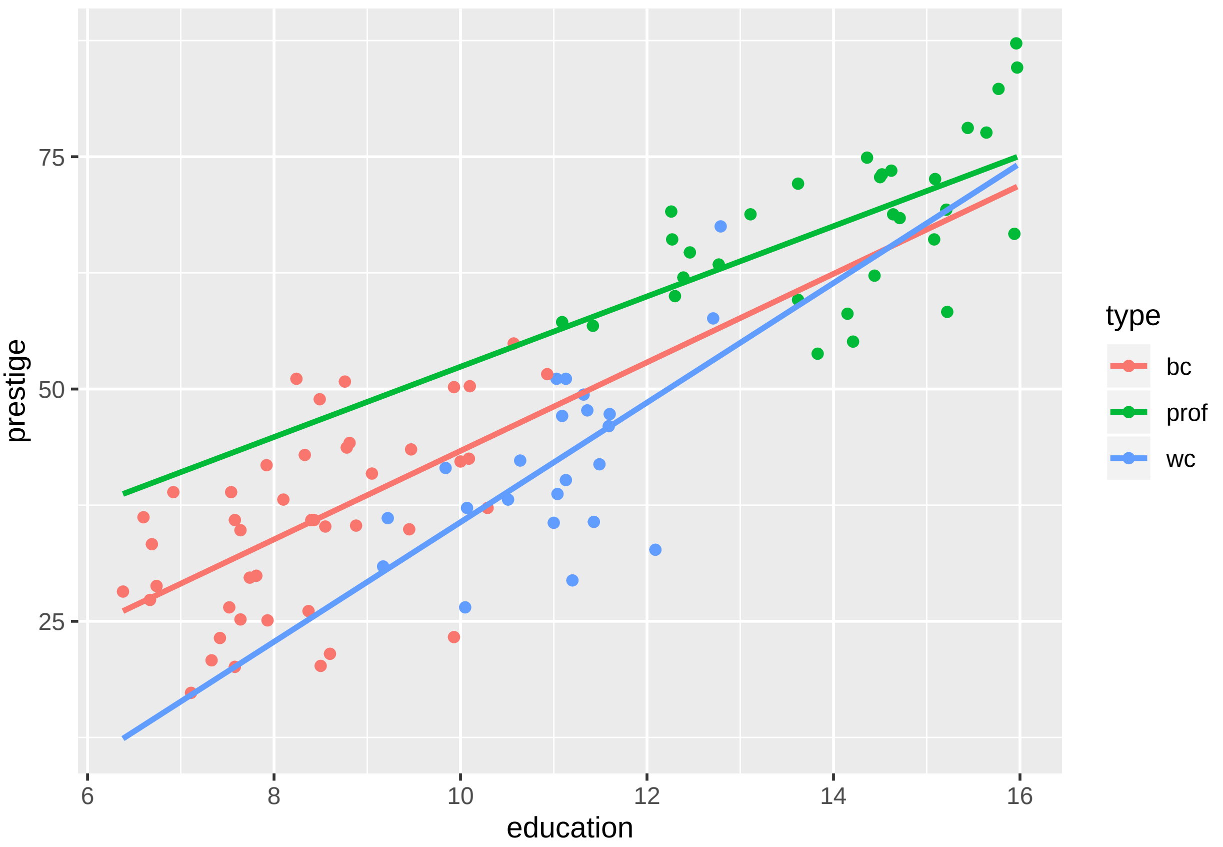
```
prestige_sub <- na.omit(subset(prestige, select = c(prestige, income, type, education)))

ggplot(data = prestige_sub, aes(x=income,  y=prestige, col=type)) + geom_point() + geom_smooth(method=l
```



There seems to be an obvious potential interaction between income and type (noticeable with or without the lines)

b.

```
ggplot(data = prestige_sub, aes(x=education,  y=prestige, col=type)) + geom_point() + geom_smooth(metho
```



There seems to be a potential and subtle interaction between education and type, but not very obvious.

2.

a.

$$prestige_i = \alpha + \beta_1(education_i) + \beta_2(income_i) + \gamma_1 D_{prof,i} + \gamma_2 D_{wc,i} + \delta_1(income_i)D_{prof,i} + \delta_2(income_i)D_{wc,i} + \epsilon_i$$

b.

```
prestige_lm <- lm(prestige ~ education + income + type + income:type, data = prestige)
prestige_lm$coefficients
```

```
##     (Intercept)       education          income       typeprof          typewc
##    -6.727263304     3.039696088     0.003134410    25.172387319     7.137509272
## income:typeprof   income:typewc
##    -0.002510174    -0.001485560
```

$\hat{prestige_i} = -6.73 + 3.037(education_i) + 0.0031(income_i) + 25.17D_{prof,i} + 7.14D_{wc,i} - 0.0025(income_i)D_{prof,i} - 0.0015(income_i)D_{wc,i}$

c.

education - for blue collar workers, for every one year increase in education we expect to see a 3.040 unit increase in prestige on average, holding all else constant.

income - for blue collar workers, for every one dollar increase in income we expect to see a 0.00313 unit increase in prestige on average, holding all else constant.

income:prof - On average, professional workers will see prestige grow at a slower rate than blue collar workers, on average, by 0.0025 for every one dollar increase in income, holding all else constant.

income:wc - On average, white collar workers will see prestige grow at a slower rate than blue collar workers, on average, by 0.0015 for every one dollar increase in income, holding all else constant.

d.

We should use the incremental F-test to determine whether or not the interactions are statistcally significant.

$H_0 : \delta_1 = \delta_2 = 0$

$H_a : \{\exists \delta_j \neq 0 | j = 1, 2\}$

```r
# creating the null and full models using the lm function
null_lm3 <- lm(prestige ~ education + income + type, data = prestige)
full_lm3 <- lm(prestige ~ education + income + type + income:type, data = prestige)

# anova() to calculate the f-stat and p-value for the incremental f test
anova(null_lm3, full_lm3)
```

```
## Analysis of Variance Table
##
## Model 1: prestige ~ education + income + type
## Model 2: prestige ~ education + income + type + income:type
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     93 4681.3
## 2     91 3791.3  2    890.02 10.681 6.809e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With an F-statistic of 10.681 and p-value of 6.809e^{-5}, we have significant evidence to reject the null hypothesis, in favor of the alternative, that the interaction between type and income is significant in predicting prestige.

e.

Blue Collar:

$\hat{prestige}_i = -6.73 + 3.037(education_i) + 0.0031(income_i)$

Professional:

$\hat{prestige}_i = -6.73 + 25.17 + 3.037(education_i) + (0.0031 - 0.0025)(income_i)$

For every one dollar increase in income, white collar workers will see a 0.0025 smaller unit increase in prestige than that of blue collar workers. (on average and holding all else constant)

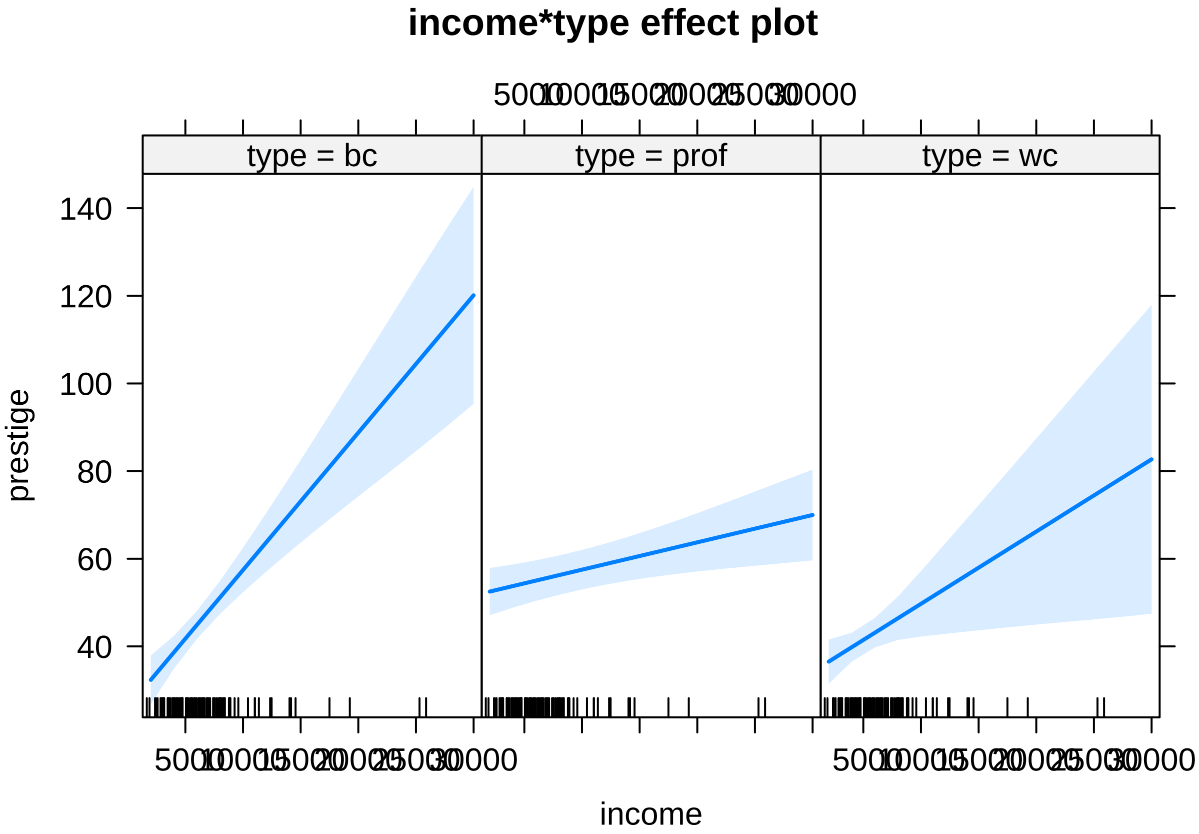
White Collar:

$\hat{prestige}_i = -6.73 + 7.14 + 3.037(education_i) + (0.0031 - 0.0015)(income_i) + 25.17D_{prof,i}$

For every one dollar increase in income, white collar workers will see a 0.0015 smaller unit increase in prestige than that of blue collar workers. (on average and holding all else constant)

f.

```r
plot(effect("income:type", prestige_lm))
```



The rate at which prestige grows is much quicker for blue collar workers than for white collar and professional workers. The shaded areas (the confidence bands), are much wider for white collar workers.