Chapter 11: Linear Regression and Model Building

Answers to Exercises

Brian Fogarty

Contents

Exercise 1

Exercise 1.a

We multiple these three variables by 100 so they are more intuitive as percentages.

```
library(tidyverse)
simd <- read_csv("simd2020.csv", na = "*")

simd <- simd %>%
  mutate(pct_income_deprived = Income_rate*100,
         pct_employment_deprived = Employment_rate*100,
         pct_not_participating = not_participating*100)
```

Exercise 1.b

```
summary(model.1 <- lm(crime_rate ~ pct_income_deprived + pct_employment_deprived +
                        pct_not_participating, data = simd))
```

```
Call:
lm(formula = crime_rate ~ pct_income_deprived + pct_employment_deprived +
    pct_not_participating, data = simd)

Residuals:
    Min      1Q  Median      3Q     Max
 -886.4  -134.9   -59.5    45.2 12322.7

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)               84.748      8.128  10.426  < 2e-16 ***
pct_income_deprived       -5.179      1.758  -2.945  0.00324 **
pct_employment_deprived   21.746      2.382   9.128  < 2e-16 ***
pct_not_participating     19.348      1.412  13.706  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 377.6 on 6470 degrees of freedom
  (502 observations deleted due to missingness)
Multiple R-squared:  0.1592,    Adjusted R-squared:  0.1588
F-statistic: 408.4 on 3 and 6470 DF,  p-value: < 2.2e-16
```

We see $R^2 = 0.1592$, which we interpret as *our model explains 15.92% of the variance in crime rates in Scottish datazones.* We see that the p-value for the F-test is below 0.05 and thus our overall model is statistically significant. Again, this means that our model is better than a model where all the predictors equal 0.

All three predictors have a statistically significant effect on crime rates. The coefficient interpretation for `pct_income_deprived` is *for a one-unit increase in the percentage of income deprivation in a datazone, the crime rate is expected to decrease by 5.18 crimes (per 10,000 residents).* This is an unexpected result. Although income deprivation includes government support for the working poor and certain allowances for job seekers, we would not expect to find datazones with higher income deprivation have lower crime. We might speculate that this result is due to the working poor being too busy to be drawn into criminal activity, but we observed a positive correlation between income deprivation and crime rates in the Chapter 10 exercises. Therefore, we should follow-up on this finding to check its robustness.

The coefficient interpretation for `pct_employment_deprived` is *for a one-unit increase in the percentage of employment deprivation in a datazone, the crime rate is expected to increase by 21.75 crimes (per 10,000 residents).* Unlike the effect for income deprivation, this finding is inline with expectations. This result suggests that datazones with high employment deprivation are expected to have higher crime rates. Employment deprivation includes government support for people out of work, for a variety of reasons, and is less about government support for the working poor (that is, income deprivation). However, we expect that income and employment deprivation would have similar effects on crime rates; we'll look at this further in the Chapter 12 exercises.

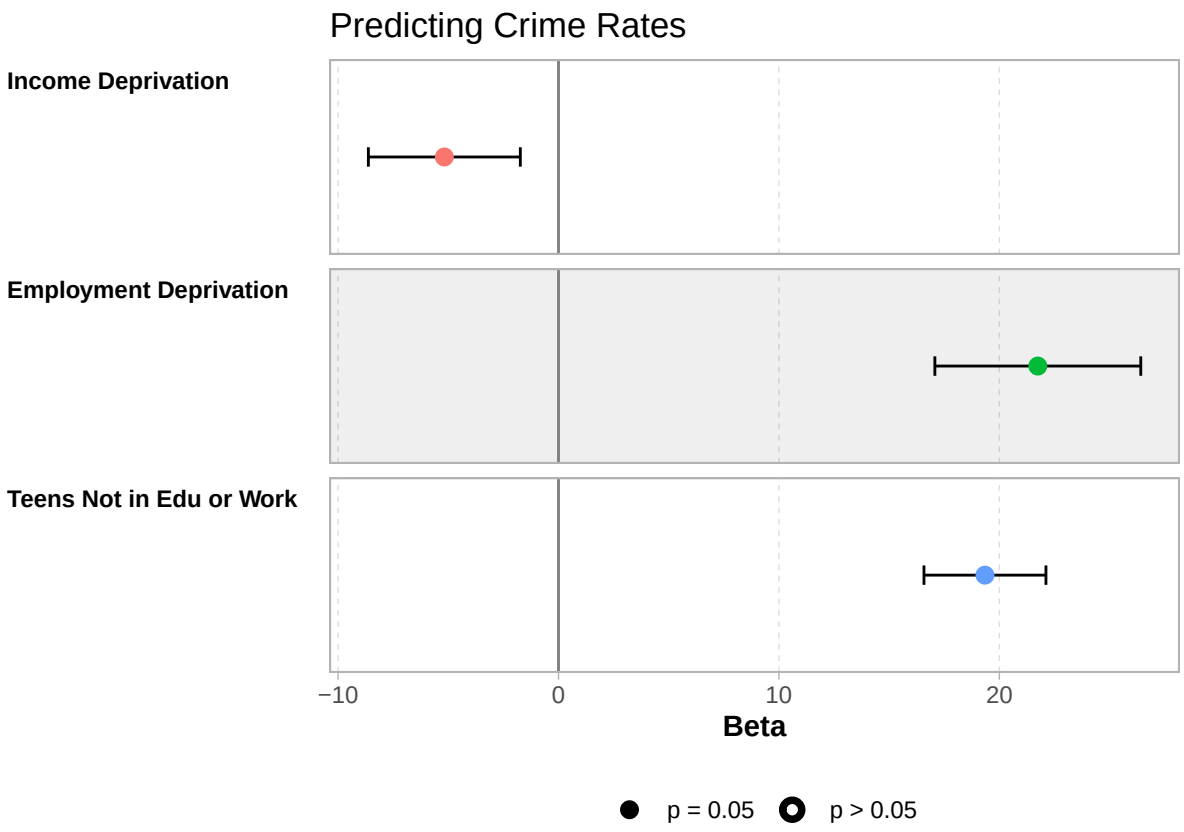
Lastly, the coefficient interpretation for `pct_not_participating` is *for a one-unit increase in the percentage of teens not in education, work, or training in a datazone, the crime rate is expected to increase by 19.35 crimes (per 10,000 residents).* This result fits the common stereotype of non-engaged teens - teens not in school, training, or working fill their time by committing crimes.[1]

Exercise 1.c

```
library(GGally)

ggcoef_model(model.1,
             variable_labels = c(
               pct_income_deprived = "Income Deprivation",
               pct_employment_deprived = "Employment Deprivation",
               pct_not_participating = "Teens Not in Edu or Work"),
             show_p_values = FALSE,
             signif_stars = FALSE) +
  labs(title = "Predicting Crime Rates")
```

[1] Every country seems to have their own usually derogatory name for such teens (e.g., neds, chavs, etc.).



Exercise 2

Exercise 2.a

We should label the values in `urban` so the regression results are clearer.

```
simd <- simd %>%
  mutate(urban_fct = recode(urban, `1` = "Urban", `0` = "Rural"))

summary(model.2 <- lm(crime_rate ~ pct_income_deprived + pct_employment_deprived +
                        pct_not_participating + urban_fct, data = simd))
```

```
Call:
lm(formula = crime_rate ~ pct_income_deprived + pct_employment_deprived +
    pct_not_participating + urban_fct, data = simd)

Residuals:
    Min      1Q  Median      3Q     Max
 -881.1  -134.8   -54.8    48.0 12286.9

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)              35.029     10.091   3.471 0.000521 ***
pct_income_deprived      -6.324      1.755  -3.604 0.000316 ***
```

```
pct_employment_deprived   22.384      2.371   9.439  < 2e-16 ***
pct_not_participating     18.575      1.408  13.196  < 2e-16 ***
urban_fctUrban            86.582     10.512   8.237  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 375.7 on 6469 degrees of freedom
  (502 observations deleted due to missingness)
Multiple R-squared:  0.1679,    Adjusted R-squared:  0.1674
F-statistic: 326.4 on 4 and 6469 DF,  p-value: < 2.2e-16
```

We see $R^2 = 0.1679$, which we interpret as *our model explains 16.79% of the variance in crime rates in Scottish datazones*. We see that the p-value for the F-test is below 0.05 and thus our overall model is statistically significant. Again, this means that our model is better than a model where all the predictors equal 0.

We do not see any changes in the statistical significance of the predictors in model.1. `urban_fct` has a statistical significant effect and we interpret its coefficient as *urban datazones are expected to have 86.58 more crimes (per 10,000 residents) than rural datazones*. Hence, urban areas are expected to have more crime than rural areas. This result matches our expectations.

Exercise 2.b

```
ggcoef_model(model.2,
             variable_labels = c(
               pct_income_deprived = "Income Deprivation",
               pct_employment_deprived = "Employment Deprivation",
               pct_not_participating = "Teens Not in Edu or Work",
               urban_fct = "Urban/Rural"),
             no_reference_row = "urban_fct",
             show_p_values = FALSE,
             signif_stars = FALSE) +
  labs(title = "Predicting Crime Rates")
```


Predicting Crime Rates
Income Deprivation
Employment Deprivation
Teens Not in Edu or Work
Urban/Rural
Urban
0
40
80
Beta
p = 0.05
p > 0.05