STAT 101

Dr. Kari Lock Morgan

Multiple Regression

- Variable selection
- Confounding variables revisited

Model Output

```
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.127e-01  8.152e-01   0.629  0.53383
HouseholdIncome   3.433e-07  2.443e-06   0.141  0.88913
IQ               -9.714e-04  1.453e-02  -0.067  0.94710
RegionNE         -3.623e-02  3.830e-02  -0.946  0.35135
RegionS           7.821e-02  3.116e-02   2.510  0.01733 *
RegionW           7.579e-02  4.061e-02   1.866  0.07119 .
Population       -2.276e-03  2.235e-03  -1.018  0.31619
X8thGradeMath     3.355e-03  3.856e-03   0.870  0.39072
HighSchool        7.909e-03  4.439e-03   1.782  0.08431 .
GSP               1.844e-06  1.803e-06   1.022  0.31434
FiveVegetables   -2.572e-03  5.371e-03  -0.479  0.63530
Smokers          -9.291e-03  4.673e-03  -1.988  0.05541 .
PhysicalActivity -1.461e-02  5.176e-03  -2.822  0.00814 **
Obese             1.542e-03  6.177e-03   0.250  0.80451
College          -6.655e-03  4.904e-03  -1.357  0.18425
NonWhite         -1.098e-03  1.691e-03  -0.649  0.52077
HeavyDrinkers    -1.757e-02  9.502e-03  -1.849  0.07378 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05462 on 32 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.7545,     Adjusted R-squared: 0.6317
F-statistic: 6.146 on 16 and 32 DF,  p-value: 6.835e-06
```

R^2 versus Adjusted R^2

If you want to evaluate the success of the model, in terms of the percentage of the variability in the response explained by the explanatory variables, you would use

a) R^2

b) Adjusted R^2

R^2 versus Adjusted R^2

If you want to compare two competing models and decide whether a certain explanatory should be included or not, you would use

a) R^2

b) Adjusted R^2

R^2 always increases or stays the same with additional explanatory variables, even if they are worthless.

Adjusted R^2 should go down if non-useful variables are added.

Variable Selection

- The p-value for an explanatory variable can be taken as a rough measure for how helpful that explanatory variable is to the model
- Insignificant variables may be pruned from the model, as long as adjusted R^2 doesn't decrease
- You can also look at relationships between explanatory variables; if two are strongly associated, perhaps both are not necessary

Variable Selection

(Some) ways of deciding whether a variable should be included in the model or not:

1. Does it improve adjusted R^2?
2. Does it have a low p-value?
3. Is it associated with the response by itself?
4. Is it strongly associated with another explanatory variables? (If yes, then including both may be redundant)
5. Does common sense say it should contribute to the model?

Stepwise Regression

- We could go through and think hard about which variables to include, or we could automate the process
- ***Stepwise regression*** drops insignificant variables one by one
- This is particularly useful if you have many potential explanatory variables

Full Model

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.940e-01  7.708e-01   0.381  0.70546
HouseholdIncome   -2.605e-07  2.310e-06  -0.113  0.91093
IQ                -7.944e-03  1.374e-02  -0.578  0.56712
RegionNE           4.545e-02  3.622e-02   1.255  0.21867
RegionS           -8.204e-02  2.947e-02  -2.784  0.00894 **
RegionW           -7.738e-02  3.840e-02  -2.015  0.05238 .
Population         1.290e-03  2.114e-03   0.610  0.54586
EighthGradeMath    4.862e-04  3.646e-03   0.133  0.89476
HighSchool        -5.364e-03  4.198e-03  -1.278  0.21050
GSP               -2.175e-06  1.705e-06  -1.275  0.21132
FiveVegetables     3.828e-03  5.079e-03   0.754  0.45659
Smokers            1.046e-02  4.419e-03   2.368  0.02409 *
PhysicalActivity   9.882e-03  4.895e-03   2.019  0.05195 .
Obese             -2.056e-03  5.841e-03  -0.352  0.72712
College            8.277e-03  4.637e-03   1.785  0.08374 .
NonWhite           2.415e-03  1.599e-03   1.510  0.14074
HeavyDrinkers      2.270e-02  8.986e-03   2.526  0.01668 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05165 on 32 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.822,      Adjusted R-squared:  0.733
F-statistic: 9.237 on 16 and 32 DF,  p-value: 6.909e-08
```

Highest p-value

Pruned Model 1

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.697e-01  7.378e-01   0.366  0.71705
HouseholdIncome  -2.883e-07  2.266e-06  -0.127  0.89953
IQ               -6.459e-03  7.916e-03  -0.816  0.42043
RegionNE          4.440e-02  3.483e-02   1.275  0.21129
RegionS          -8.205e-02  2.903e-02  -2.827  0.00793 **
RegionW          -7.682e-02  3.760e-02  -2.043  0.04910 *
Population        1.282e-03  2.081e-03   0.616  0.54219
HighSchool       -5.458e-03  4.077e-03  -1.339  0.18981
GSP              -2.138e-06  1.657e-06  -1.290  0.20597
FiveVegetables    3.864e-03  4.996e-03   0.773  0.44481
Smokers           1.043e-02  4.347e-03   2.400  0.02218 *
PhysicalActivity  9.967e-03  4.781e-03   2.085  0.04492 *
Obese            -1.778e-03  5.374e-03  -0.331  0.74284
College           8.365e-03  4.522e-03   1.850  0.07329 .
NonWhite          2.471e-03  1.521e-03   1.624  0.11379
HeavyDrinkers     2.293e-02  8.686e-03   2.639  0.01258 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05088 on 33 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.8219,     Adjusted R-squared: 0.741
F-statistic: 10.15 on 15 and 33 DF,  p-value: 2.114e-08
```

Highest p-value

Pruned Model 2

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.700e-01  7.270e-01   0.371  0.71270
IQ               -6.312e-03  7.718e-03  -0.818  0.41914
RegionNE          4.397e-02  3.416e-02   1.287  0.20672
RegionS          -8.129e-02  2.799e-02  -2.904  0.00643 **
RegionW          -7.629e-02  3.682e-02  -2.072  0.04594 *
Population        1.267e-03  2.047e-03   0.619  0.54020
HighSchool       -5.621e-03  3.813e-03  -1.474  0.14968
GSP              -2.218e-06  1.509e-06  -1.470  0.15080
FiveVegetables    3.688e-03  4.732e-03   0.779  0.44111
Smokers           1.051e-02  4.242e-03   2.478  0.01835 *
PhysicalActivity  9.893e-03  4.676e-03   2.115  0.04180 *
Obese            -1.830e-03  5.280e-03  -0.347  0.73107
College           8.269e-03  4.394e-03   1.882  0.06841 .
NonWhite          2.457e-03  1.495e-03   1.643  0.10950
HeavyDrinkers     2.318e-02  8.335e-03   2.781  0.00878 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05014 on 34 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.8218,     Adjusted R-squared: 0.7485
F-statistic:  11.2 on 14 and 34 DF,  p-value: 6.151e-09
```

Highest p-value

Pruned Model 3

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.873e-01  6.781e-01   0.276  0.78396
IQ               -5.856e-03  7.509e-03  -0.780  0.44072
RegionNE          4.629e-02  3.308e-02   1.399  0.17054
RegionS          -8.262e-02  2.737e-02  -3.019  0.00471 **
RegionW          -6.972e-02  3.117e-02  -2.237  0.03178 *
Population        1.328e-03  2.014e-03   0.659  0.51399
HighSchool       -5.750e-03  3.747e-03  -1.534  0.13390
GSP              -2.366e-06  1.430e-06  -1.654  0.10701
FiveVegetables    3.507e-03  4.643e-03   0.755  0.45512
Smokers           1.043e-02  4.183e-03   2.495  0.01748 *
PhysicalActivity  9.733e-03  4.595e-03   2.118  0.04132 *
College           8.936e-03  3.899e-03   2.292  0.02804 *
NonWhite          2.442e-03  1.476e-03   1.655  0.10683
HeavyDrinkers     2.408e-02  7.815e-03   3.082  0.00400 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0495 on 35 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.8212,     Adjusted R-squared: 0.7548
F-statistic: 12.37 on 13 and 35 DF,  p-value: 1.786e-09
```

Highest p-value

Pruned Model 4

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       9.266e-02  6.575e-01   0.141  0.88872
IQ               -3.818e-03  6.789e-03  -0.562  0.57736
RegionNE          3.708e-02  2.975e-02   1.246  0.22070
RegionS          -8.831e-02  2.578e-02  -3.426  0.00155 **
RegionW          -7.717e-02  2.882e-02  -2.677  0.01110 *
HighSchool       -6.422e-03  3.577e-03  -1.795  0.08102 .
GSP              -2.115e-06  1.368e-06  -1.546  0.13077
FiveVegetables    5.265e-03  3.772e-03   1.396  0.17126
Smokers           9.381e-03  3.835e-03   2.446  0.01947 *
PhysicalActivity  9.143e-03  4.471e-03   2.045  0.04823 *
College           7.789e-03  3.462e-03   2.250  0.03066 *
NonWhite          3.012e-03  1.186e-03   2.539  0.01559 *
HeavyDrinkers     2.428e-02  7.748e-03   3.133  0.00343 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04911 on 36 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.819,      Adjusted R-squared: 0.7586
F-statistic: 13.57 on 12 and 36 DF,  p-value: 5.727e-10
```

Highest p-value

Pruned Model 5

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -2.025e-01  3.925e-01  -0.516  0.60901
RegionNE         3.830e-02  2.940e-02   1.303  0.20063
RegionS         -8.584e-02  2.516e-02  -3.411  0.00158 **
RegionW         -7.020e-02  2.578e-02  -2.723  0.00981 **
HighSchool      -7.027e-03  3.381e-03  -2.078  0.04466 *
GSP             -2.210e-06  1.344e-06  -1.644  0.10864
FiveVegetables   5.068e-03  3.721e-03   1.362  0.18134
Smokers          9.704e-03  3.757e-03   2.583  0.01389 *
PhysicalActivity 8.650e-03  4.344e-03   1.991  0.05387 .
College          7.494e-03  3.390e-03   2.211  0.03333 *
NonWhite         3.375e-03  9.865e-04   3.421  0.00154 **
HeavyDrinkers    2.474e-02  7.633e-03   3.241  0.00252 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04866 on 37 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.8174,     Adjusted R-squared: 0.7631
F-statistic: 15.06 on 11 and 37 DF,  p-value: 1.65e-10
```

Highest p-value

Pruned Model 6

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.864e-01  3.967e-01  -0.470 0.641094
RegionNE         5.326e-02  2.758e-02   1.931 0.060930 .
RegionS         -7.881e-02  2.490e-02  -3.165 0.003054 **
RegionW         -5.980e-02  2.490e-02  -2.402 0.021322 *
HighSchool      -7.492e-03  3.401e-03  -2.203 0.033739 *
GSP             -2.666e-06  1.317e-06  -2.025 0.049906 *
Smokers          1.044e-02  3.759e-03   2.778 0.008445 **
PhysicalActivity 9.382e-03  4.359e-03   2.153 0.037771 *
College          9.499e-03  3.089e-03   3.075 0.003883 **
NonWhite         3.357e-03  9.975e-04   3.366 0.001757 **
HeavyDrinkers    2.879e-02  7.109e-03   4.050 0.000244 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0492 on 38 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.8082,     Adjusted R-squared: 0.7578
F-statistic: 16.02 on 10 and 38 DF,  p-value: 9.351e-11
```

Pruned Model 5

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -2.025e-01  3.925e-01  -0.516  0.60901
RegionNE         3.830e-02  2.940e-02   1.303  0.20063
RegionS         -8.584e-02  2.516e-02  -3.411  0.00158 **
RegionW         -7.020e-02  2.578e-02  -2.723  0.00981 **
HighSchool      -7.027e-03  3.381e-03  -2.078  0.04466 *
GSP             -2.210e-06  1.344e-06  -1.644  0.10864
FiveVegetables   5.068e-03  3.721e-03   1.362  0.18134
Smokers          9.704e-03  3.757e-03   2.583  0.01389 *
PhysicalActivity 8.650e-03  4.344e-03   1.991  0.05387 .
College          7.494e-03  3.390e-03   2.211  0.03333 *
NonWhite         3.375e-03  9.865e-04   3.421  0.00154 **
HeavyDrinkers    2.474e-02  7.633e-03   3.241  0.00252 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04866 on 37 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.8174,     Adjusted R-squared: 0.7631
F-statistic: 15.06 on 11 and 37 DF,  p-value: 1.65e-10
```

Pruned Model 7

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.1173701  0.3976552  -0.295  0.76948
RegionNE         0.0320078  0.0297918   1.074  0.28943
RegionS         -0.0832146  0.0256689  -3.242  0.00247 **
RegionW         -0.0775162  0.0259578  -2.986  0.00492 **
HighSchool      -0.0074225  0.0034468  -2.153  0.03769 *
FiveVegetables   0.0065921  0.0036831   1.790  0.08146 .
Smokers          0.0076819  0.0036288   2.117  0.04087 *
PhysicalActivity 0.0081657  0.0044297   1.843  0.07308 .
College          0.0048813  0.0030609   1.595  0.11906
NonWhite         0.0030327  0.0009857   3.077  0.00387 **
HeavyDrinkers    0.0236143  0.0077707   3.039  0.00428 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04973 on 38 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.8041,     Adjusted R-squared: 0.7525
F-statistic: 15.59 on 10 and 38 DF,  p-value: 1.382e-10
```

FINAL STEPWISE MODEL

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -2.025e-01  3.925e-01  -0.516  0.60901
RegionNE         3.830e-02  2.940e-02   1.303  0.20063
RegionS         -8.584e-02  2.516e-02  -3.411  0.00158 **
RegionW         -7.020e-02  2.578e-02  -2.723  0.00981 **
HighSchool      -7.027e-03  3.381e-03  -2.078  0.04466 *
GSP             -2.210e-06  1.344e-06  -1.644  0.10864
FiveVegetables   5.068e-03  3.721e-03   1.362  0.18134
Smokers          9.704e-03  3.757e-03   2.583  0.01389 *
PhysicalActivity 8.650e-03  4.344e-03   1.991  0.05387 .
College          7.494e-03  3.390e-03   2.211  0.03333 *
NonWhite         3.375e-03  9.865e-04   3.421  0.00154 **
HeavyDrinkers    2.474e-02  7.633e-03   3.241  0.00252 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04866 on 37 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.8174,     Adjusted R-squared: 0.7631
F-statistic: 15.06 on 11 and 37 DF,  p-value: 1.65e-10
```

Full Model

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.940e-01  7.708e-01   0.381  0.70546
HouseholdIncome  -2.605e-07  2.310e-06  -0.113  0.91093
IQ               -7.944e-03  1.374e-02  -0.578  0.56712
RegionNE          4.545e-02  3.622e-02   1.255  0.21867
RegionS          -8.204e-02  2.947e-02  -2.784  0.00894 **
RegionW          -7.738e-02  3.840e-02  -2.015  0.05238 .
Population        1.290e-03  2.114e-03   0.610  0.54586
EighthGradeMath   4.862e-04  3.646e-03   0.133  0.89476
HighSchool       -5.364e-03  4.198e-03  -1.278  0.21050
GSP              -2.175e-06  1.705e-06  -1.275  0.21132
FiveVegetables    3.828e-03  5.079e-03   0.754  0.45659
Smokers           1.046e-02  4.419e-03   2.368  0.02409 *
PhysicalActivity  9.882e-03  4.895e-03   2.019  0.05195 .
Obese            -2.056e-03  5.841e-03  -0.352  0.72712
College           8.277e-03  4.637e-03   1.785  0.08374 .
NonWhite          2.415e-03  1.599e-03   1.510  0.14074
HeavyDrinkers     2.270e-02  8.986e-03   2.526  0.01668 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05165 on 32 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.822,      Adjusted R-squared:  0.733
F-statistic: 9.237 on 16 and 32 DF,  p-value: 6.909e-08
```

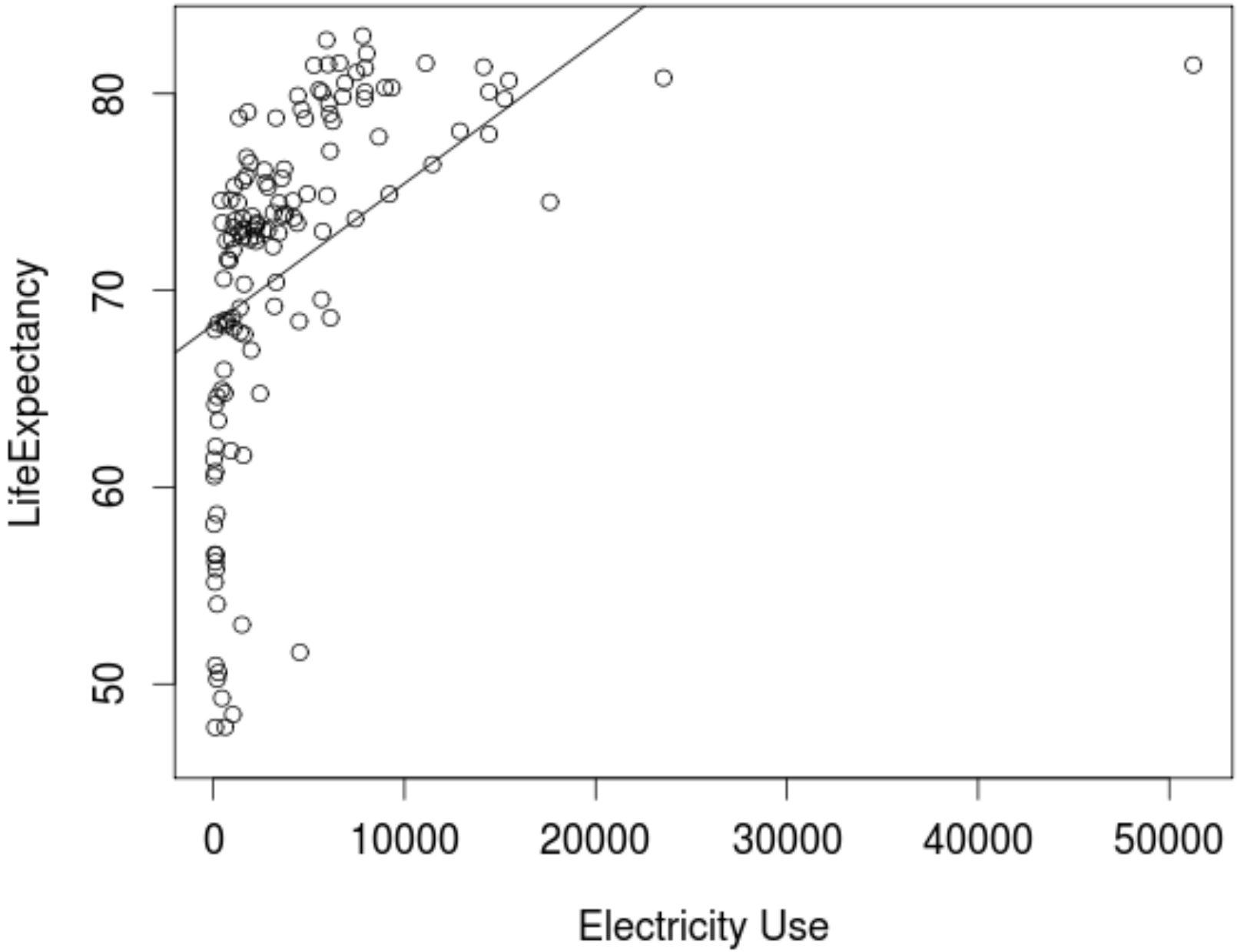
Variable Selection

- There is no one "best" model
- Choosing a model is just as much an art as a science
- Adjusted R^2 is just *one* possible criteria
- To learn much more about choosing the best model, take STAT 210

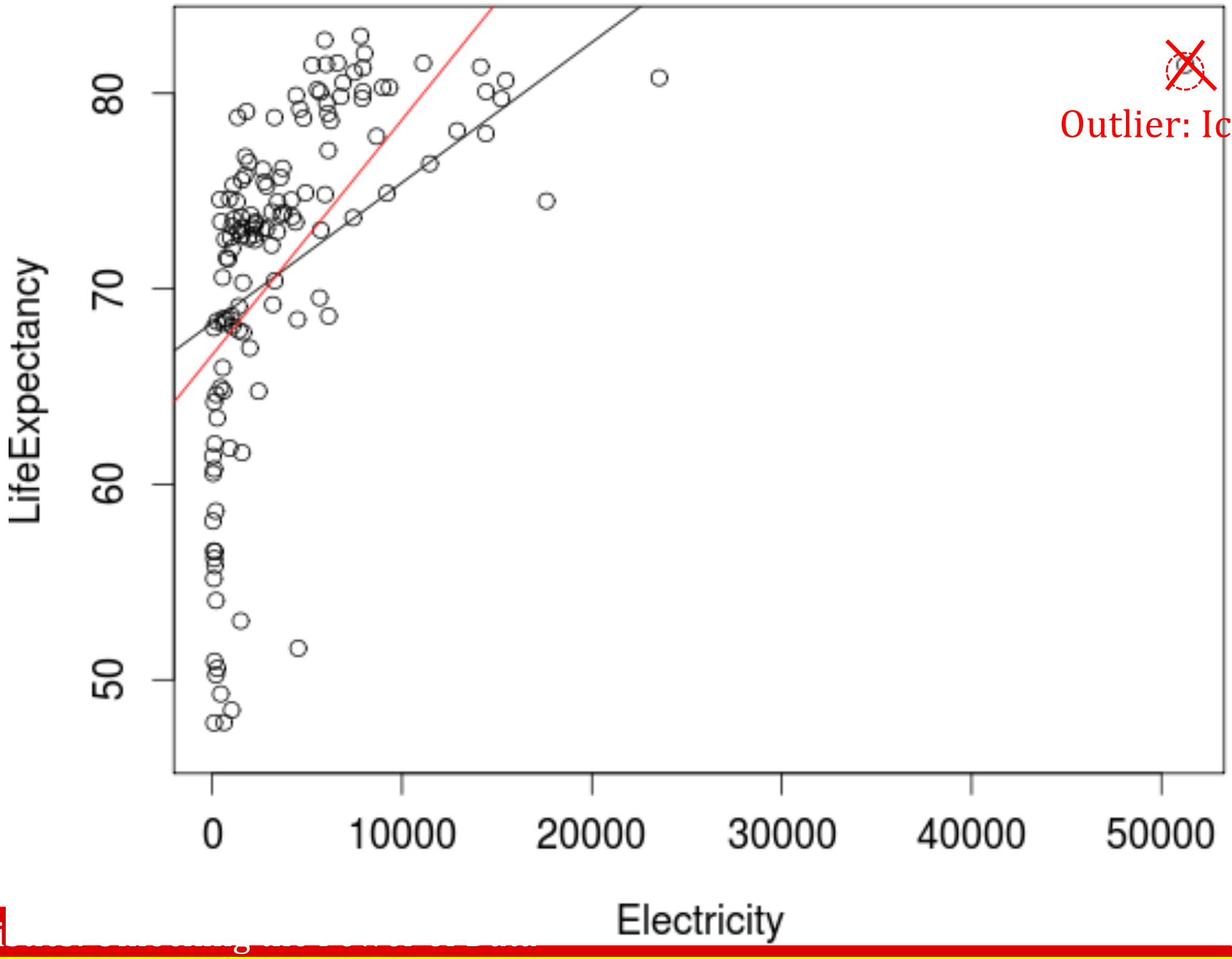
Electricity and Life Expectancy

- Cases: countries of the world
- Response variable: life expectancy
- Explanatory variable: electricity use (kWh per capita)
- Is a country's electricity use helpful in predicting life expectancy?

Electricity and Life Expectancy



Electricity and Life Expectancy



Electricity and Life Expectancy

```
> summary(lm(LifeExpectancy~Electricity))

Call:
lm(formula = LifeExpectancy ~ Electricity)

Residuals:
    Min      1Q  Median      3Q     Max
-23.576  -2.986   2.708   5.298  10.216

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.826e+01  8.172e-01  83.526  < 2e-16 ***
Electricity 7.174e-04  1.148e-04   6.251 5.23e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.738 on 132 degrees of freedom
  (82 observations deleted due to missingness)
Multiple R-squared: 0.2284,     Adjusted R-squared: 0.2226
F-statistic: 39.08 on 1 and 132 DF,  p-value: 5.231e-09
```

Electricity and Life Expectancy

Is this a good model for predicting life expectancy based on electricity use?

(a) Yes

(b) No

The association is definitely not linear.

Electricity and Life Expectancy

Is a country's electricity use helpful in predicting life expectancy?

(a) Yes

(b) No

The p-value for electricity is significant.

Electricity and Life Expectancy

```
> summary(lm(LifeExpectancy[Electricity<50000]~Electricity[Electricity<50000]))

Call:
lm(formula = LifeExpectancy[Electricity < 50000] ~ Electricity[Electricity <
    50000])

Residuals:
    Min      1Q  Median      3Q     Max
-20.459  -3.887   2.366   5.099  10.535

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     6.662e+01  8.399e-01  79.316  < 2e-16 ***
Electricity[Electricity < 50000] 1.203e-03  1.501e-04   8.019  5.2e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.205 on 131 degrees of freedom
  (82 observations deleted due to missingness)
Multiple R-squared: 0.3293,     Adjusted R-squared: 0.3242
F-statistic: 64.31 on 1 and 131 DF,  p-value: 5.197e-13
```

Electricity and Life Expectancy

If we increased electricity use in a country, would life expectancy increase?

(a) Yes
(b) No
(c) Impossible to tell

We cannot make any conclusions about causality, because this is observational data.

Electricity and Life Expectancy

If we increased electricity use in a country, would life expectancy increase?

(a) Yes
(b) No
(c) Impossible to tell

We cannot make any conclusions about causality, because this is observational data.

Confounding Variables

- Wealth is an obvious confounding variable that could explain the relationship between electricity use and life expectancy
- Multiple regression is a powerful tool that allows us to ***account for confounding variables***
- We can see whether an explanatory variable is still significant, even after including potential confounding variables in the model

Electricity and Life Expectancy

Is a country's electricity use helpful in predicting life expectancy, even after including GDP in the model?

(a) Yes (b) No

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.671e+01  8.506e-01  78.427  < 2e-16 ***
Electricity 1.899e-04  1.553e-04   1.223    0.224
GDP         2.455e-04  4.844e-05   5.068 1.52e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.323 on 117 degrees of freedom
  (96 observations deleted due to missingness)
Multiple R-squared: 0.3562,     Adjusted R-squared: 0.3452
F-statistic: 32.37 on 2 and 117 DF,  p-value: 6.461e-12
```

Once GDP is accounted for, electricity use is no longer a significant predictor of life expectancy.

Which is the "best" model?

(a)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.826e+01  8.172e-01  83.526  < 2e-16 ***
Electricity 7.174e-04  1.148e-04   6.251 5.23e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.738 on 132 degrees of freedom
  (82 observations deleted due to missingness)
Multiple R-squared: 0.2284,     Adjusted R-squared: 0.2226
F-statistic: 39.08 on 1 and 132 DF,  p-value: 5.231e-09
```

(b)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.671e+01  8.506e-01  78.427  < 2e-16 ***
Electricity 1.899e-04  1.553e-04   1.223    0.224
GDP         2.455e-04  4.844e-05   5.068 1.52e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.323 on 117 degrees of freedom
  (96 observations deleted due to missingness)
Multiple R-squared: 0.3562,     Adjusted R-squared: 0.3452
F-statistic: 32.37 on 2 and 117 DF,  p-value: 6.461e-12
```

(c)

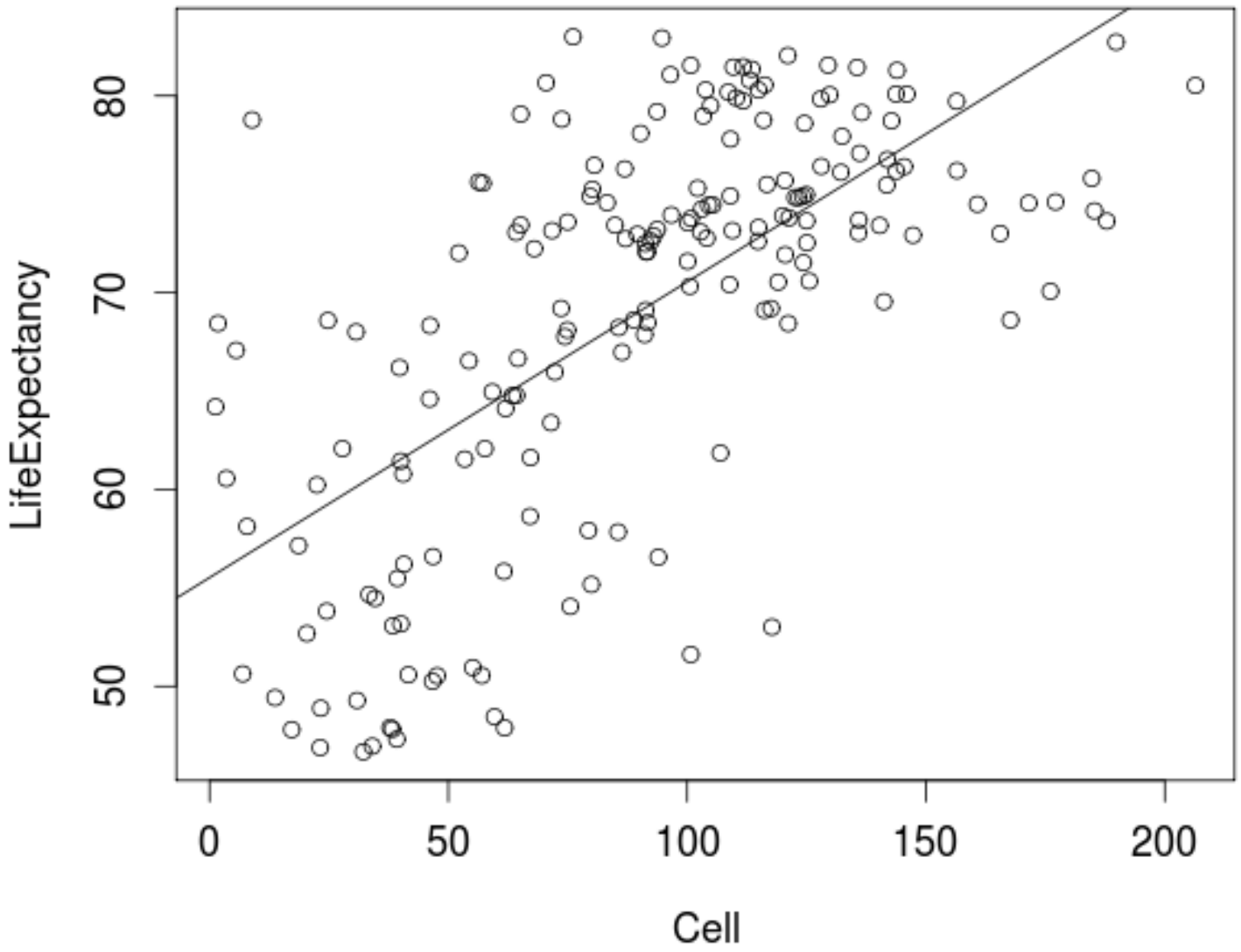
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.470e+01  7.821e-01  82.725  < 2e-16 ***
GDP         3.413e-04  3.811e-05   8.957 7.04e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.352 on 164 degrees of freedom
  (50 observations deleted due to missingness)
Multiple R-squared: 0.3285,     Adjusted R-squared: 0.3244
F-statistic: 80.23 on 1 and 164 DF,  p-value: 7.039e-16
```

You could argue for (c) as well, but I would choose (b), because it has the highest adjusted R^2

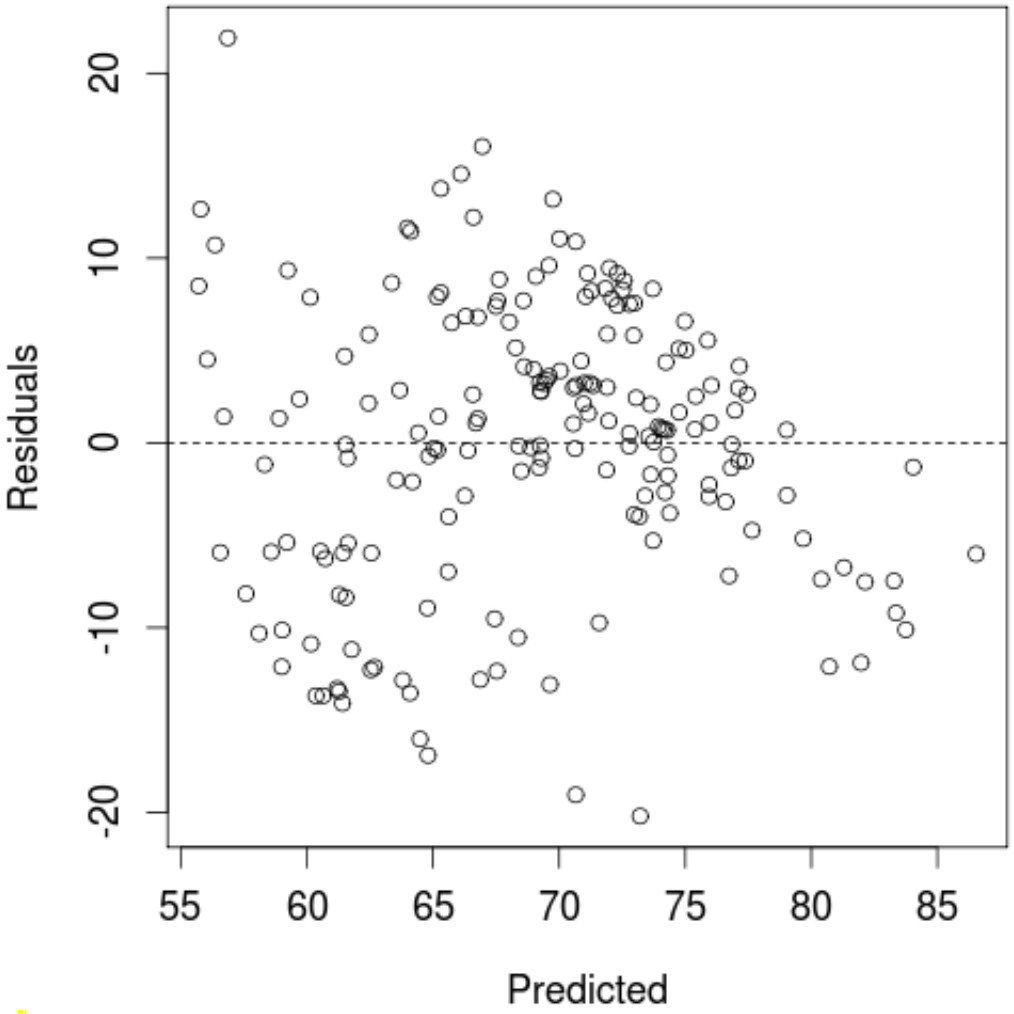
Cell Phones and Life Expectancy

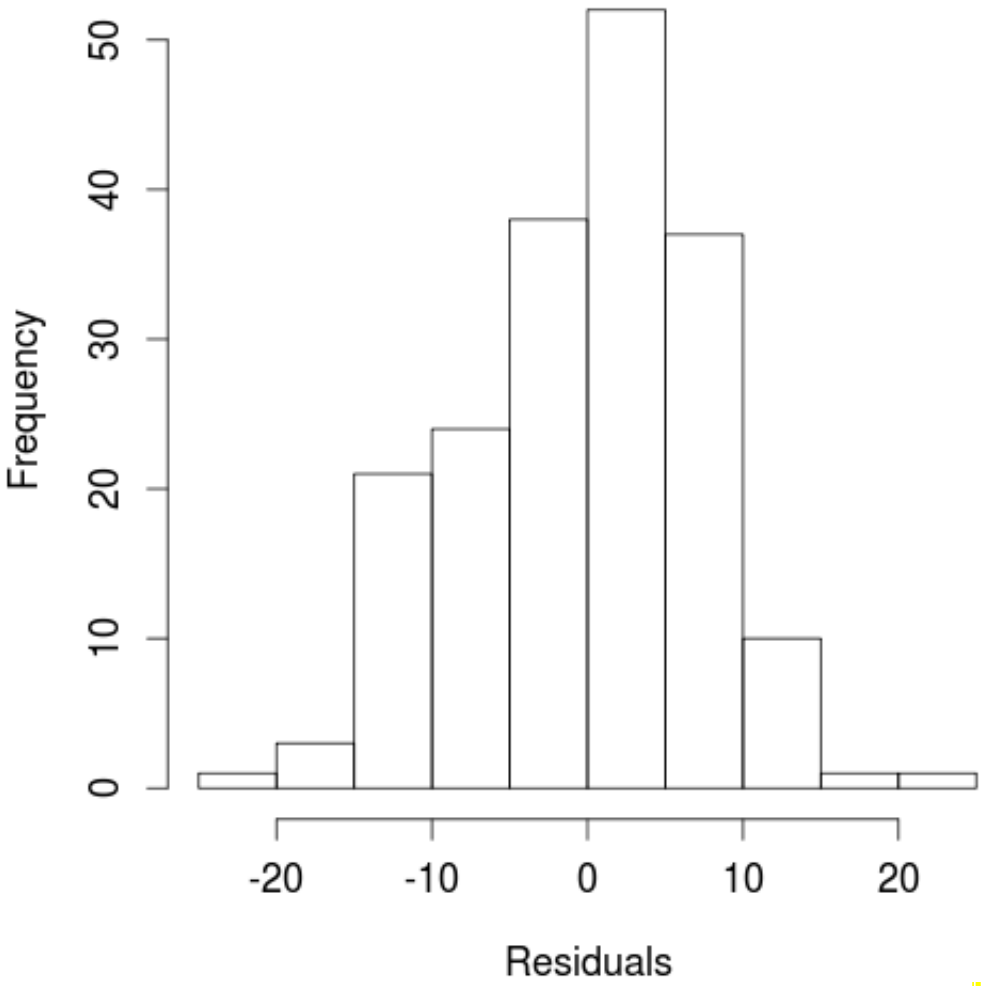
- Cases: countries of the world
- Response variable: life expectancy
- Explanatory variable: number of mobile cellular subscriptions per 100 people
- Is a country's cell phone subscription rate helpful in predicting life expectancy?

Cell Phones and Life Expectancy



Cell Phones and Life Expectancy





Cell Phones and Life Expectancy

```
> summary(lm(LifeExpectancy~Cell))

Call:
lm(formula = LifeExpectancy ~ Cell)

Residuals:
     Min       1Q   Median       3Q      Max
-20.2030  -5.3197   0.7194   5.1022  21.9136

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 55.52099    1.27221   43.64   <2e-16 ***
Cell         0.15025    0.01265   11.87   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.575 on 186 degrees of freedom
  (28 observations deleted due to missingness)
Multiple R-squared: 0.4312,     Adjusted R-squared: 0.4281
F-statistic:   141 on 1 and 186 DF,  p-value: < 2.2e-16
```

Cell Phones and Life Expectancy

Is this a good model for predicting life expectancy based on cell phone subscriptions?

(a) Yes

(b) No

The association is linear, the variability seems approximately constant, and the residuals look approximately normal.

There is a bit of concern by the slight possible downward trend towards the end of the residual plot, so if you answered no for that reason, that is okay as well.

Cell Phones and Life Expectancy

Is a country's number of cell phone subscriptions per capita helpful in predicting life expectancy?

(a) Yes

(b) No

The p-value for cell indicates strong significance.

Cell Phones and Life Expectancy

If we gave everyone in a country a cell phone and a cell phone subscription, would life expectancy in that country increase?

(a) Yes
(b) No
(c) Impossible to tell

Again, we cannot make causal conclusions.

Cell Phones and Life Expectancy

Is a country's cell phone subscription rate helpful in predicting life expectancy, even after including GDP in the model?

(a) Yes (b) No

The p-value for Cell still denotes strong significance, even with GDP in the model.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.414e+01  1.272e+00  42.573  < 2e-16 ***
Cell        1.354e-01  1.419e-02   9.539  < 2e-16 ***
GDP         1.884e-04  3.465e-05   5.439 1.95e-07 ***
```

Even after accounting for GDP, cell phone subscriptions per capita is still a significant predictor of life expectancy.

Cell Phones and Life Expectancy

- This says that wealth alone can not explain the association between cell phone subscriptions and life expectancy

- This suggests that either cell phones actually do something to increase life expectancy (causal) OR there is another confounding variable besides wealth of the country

Confounding Variables

- Multiple regression is one potential way to account for confounding variables
- This is most commonly used in practice across a wide variety of fields, but is quite sensitive to the conditions for the linear model (particularly linearity)
- You can only "rule out" confounding variables that you have data on, so it is still very hard to make true causal conclusions without a randomized experiment