1. **Read this dataset into R or Python. (See Python code).**
2. **Generate summary statistics:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | RNS | MRT | SMSA | MED | IQ | KWW | AGE | S | EXPR | LW |
| **count** | 758 | 758 | 758 | 758 | 758 | 758 | 758 | 758 | 758 | 758 |
| **mean** | 0.27 | 0.51 | 0.70 | 10.91 | 103.86 | 36.57 | 21.84 | 13.41 | 1.74 | 5.69 |
| **std** | 0.44 | 0.50 | 0.46 | 2.74 | 13.62 | 7.30 | 2.98 | 2.23 | 2.11 | 0.43 |
| **min** | 0 | 0 | 0 | 0 | 54 | 12 | 16 | 9 | 0 | 4.61 |
| **25%** | 0 | 0 | 0 | 9 | 95.25 | 32 | 20 | 12 | 0.28 | 5.38 |
| **50%** | 0 | 1 | 1 | 12 | 104 | 37 | 22 | 12 | 0.96 | 5.68 |
| **75%** | 1 | 1 | 1 | 12 | 113.75 | 41 | 24 | 16 | 2.44 | 5.99 |
| **max** | 1 | 1 | 1 | 18 | 145 | 56 | 30 | 18 | 11.44 | 7.05 |

1. **Generate scatter plots.**
2. **Fit linear models.**  
    The linear trends look consistent with scatterplots and don’t seem unreasonable.   
   
3. **Estimate a bivariate LS model relating log wages to schooling. Calculate a 95%CI.**

95% CI for : (0.084762, 0.108487)

95% CI for : (4.230288, 4.552684) (using R code)

1. Estimate a multivariate least squares model relating log wages to the variables in (b).

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.414876 0.122668 27.838 < 2e-16 \*\*\*

rns -0.087742 0.027389 -3.204 0.001415 \*\*

mrt 0.100669 0.027090 3.716 0.000217 \*\*\*

smsa 0.136766 0.026586 5.144 3.43e-07 \*\*\*

med 0.005885 0.004676 1.258 0.208630

iq 0.004189 0.001048 3.998 7.01e-05 \*\*\*

kww -0.002269 0.001932 -1.174 0.240684

age 0.049705 0.005959 8.342 3.52e-16 \*\*\*

s 0.047888 0.007776 6.159 1.20e-09 \*\*\*

expr 0.002206 0.007127 0.310 0.756979

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.325 on 748 degrees of freedom

Multiple R-squared: 0.4326, Adjusted R-squared: 0.4258

F-statistic: 63.38 on 9 and 748 DF, p-value: < 2.2e-16

95%CI of returns to schooling: (0.0326, 0.0632)

1. **Generate a variable that is age raised to the power of two. Re-estimate (f) including age-squared.**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.862755 0.557147 8.728 < 2e-16 \*\*\*

rns -0.084708 0.027302 -3.103 0.001990 \*\*

mrt 0.111797 0.027302 4.095 4.69e-05 \*\*\*

smsa 0.139964 0.026505 5.281 1.69e-07 \*\*\*

med 0.005623 0.004658 1.207 0.227753

iq 0.004053 0.001045 3.880 0.000114 \*\*\*

kww -0.001998 0.001927 -1.037 0.300064

age -0.083783 0.050467 -1.660 0.097298 .

agesq 0.002923 0.001097 2.664 0.007898 \*\*

s 0.051085 0.007837 6.519 1.31e-10 \*\*\*

expr 0.003669 0.007119 0.515 0.606445

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3237 on 747 degrees of freedom

Multiple R-squared: 0.438, Adjusted R-squared: 0.4305

F-statistic: 58.21 on 10 and 747 DF, p-value: < 2.2e-16

1. **Why do the estimates of return to schooling differ in (e) and (h)?**

Variables age, agesq, and s, are all correlated. The extent of correlation is described in the correlation matrix and scatterplots below. Since agesq and s are far from orthogonal, adding agesq to the model can affect the coefficient of s. The change in coefficient and standard error of s do not seem large enough to worry about multicollinearity.

s age agesq

s 1.000 0.448 0.438

age 0.448 1.000 0.997

agesq 0.438 0.997 1.000



## Problem 2

Assume the following DGP: , for , and suppose for .

1. **Suppose and independently. Suppose you estimate the following using least squares: . What number do you think your least squares estimate of should be? Simulate this DGP assuing 10,000 observations and compute LSE of .**

Since and are independent, the estimate of should be unbiased, close to The model fit below lends supporting evidence.

Call:

lm(formula = y ~ x1)

Residuals:

Min 1Q Median 3Q Max

-5.2653 -0.9531 0.0119 0.9636 5.1901

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.01076 0.01414 71.47 <2e-16 \*\*\*

x1 1.01545 0.01422 71.43 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.414 on 9998 degrees of freedom

Multiple R-squared: 0.3379, Adjusted R-squared: 0.3378

F-statistic: 5103 on 1 and 9998 DF, p-value: < 2.2e-16

1. **Suppose instead , and where, are independent standard normal variables. Using the model . What can you say about the LSE ? Simulate DGP assuming 10,000 observations and estimate .**

Except for the error terms and , any change in due to will be exactly offset by . For the model this means that the error term is not independent of . My intuition is that the value of would be less than since some of the contribution of to would be negated by and that would be volatile. Simulations did not show evidence of volatility, but it did show that was biased downward.

Call:

lm(formula = y ~ x1)

Residuals:

Min 1Q Median 3Q Max

-6.8562 -1.0695 -0.0162 1.0653 6.4915

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.97825 0.01577 62.03 <2e-16 \*\*\*

x1 0.52926 0.01118 47.33 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.577 on 9998 degrees of freedom

Multiple R-squared: 0.183, Adjusted R-squared: 0.1829

F-statistic: 2240 on 1 and 9998 DF, p-value: < 2.2e-16.

I ran several simulations changing the standard deviation of and . The estimate increased as these standard deviations increased (with large deviations, the contribution of and becomes comparatively small, with small deviations, and cancel each other out more effectively). Except with very small standard deviations of and , the coefficients and in the multiple regression were both close to 1.

1. **Suppose I say that any statistical estimates you put in front of me can be dismissed by claiming you haven’t included everything in the world that is relevant. How do you respond?**

Response 1. Ockham’s razor. One of the main functions of models is to simplify the world by leaving out some details in order to emphasize the most important relationships. Since most modeled problems have uncertainty, simpler models offer the advantage of having fewer sources of uncertainty, so we can better understand the structure of uncertainty of the model.

Response 2. Predictive power. If I can demonstrate that the model makes accurate and useful predictions, then I can argue the model is useful.

Response 3. If we included everything that is relevant, we might run into two problems. We may have more variables than observations, which may lead to overfitting unless we somehow select a subset of variables, which will again raise your objection that we are not including everything relevant. The other problem is that our ability to collect data is limited by the observation effect.

Problem 3.

1. **Read data. (See code).**
2. **Fit linear and logistic models to the training set.**

**Linear model:**

Call:

lm(formula = union ~ ., data = union.train)

Residuals:

Min 1Q Median 3Q Max

-0.43587 -0.23453 -0.19253 -0.06148 1.00101

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.433e-01 1.021e-01 2.382 0.01723 \*

idcode 5.084e-06 2.875e-06 1.769 0.07700 .

year -3.409e-03 1.599e-03 -2.132 0.03301 \*

age 3.913e-03 1.237e-03 3.162 0.00157 \*\*

grade 9.933e-03 1.671e-03 5.945 2.84e-09 \*\*\*

south -1.584e-01 8.732e-03 -18.139 < 2e-16 \*\*\*

black 1.207e-01 8.596e-03 14.046 < 2e-16 \*\*\*

smsa 1.319e-02 8.443e-03 1.563 0.11812

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3951 on 12121 degrees of freedom

Multiple R-squared: 0.04338, Adjusted R-squared: 0.04283

F-statistic: 78.52 on 7 and 12121 DF, p-value: < 2.2e-16

**Logistic model:**

Call:

glm(formula = union ~ ., family = binomial, data = union.train)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.1974 -0.7148 -0.6217 -0.4175 2.3946

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.116e+00 6.554e-01 -1.703 0.08866 .

idcode 2.448e-05 1.774e-05 1.380 0.16759

year -2.248e-02 1.029e-02 -2.185 0.02888 \*

age 2.511e-02 7.990e-03 3.143 0.00167 \*\*

grade 6.511e-02 1.098e-02 5.930 3.02e-09 \*\*\*

south -1.064e+00 5.990e-02 -17.755 < 2e-16 \*\*\*

black 7.684e-01 5.434e-02 14.141 < 2e-16 \*\*\*

smsa 6.668e-02 5.654e-02 1.179 0.23834

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 12309 on 12128 degrees of freedom

Residual deviance: 11760 on 12121 degrees of freedom

AIC: 11776

Number of Fisher Scoring iterations: 4

1. **Predict for test set. (See code).**
2. **Compare prediction accuracy with threshold of 0.2.**

|  |  |  |
| --- | --- | --- |
| SVM | Number of Union Members (Predicted) | Number of Union Members (Actual) |
| Linear | 9136 | 3323 |
| Logit | 8164 | 3323 |

With threshold of 0.25

|  |  |  |
| --- | --- | --- |
| SVM | Number of Union Members (Predicted) | Number of Union Members (Actual) |
| Linear | 4311 | 3323 |
| Logit | 3694 | 3323 |

Confusion matrices for threshold of 0.25.

|  |  |  |  |
| --- | --- | --- | --- |
| Linear prediction | | Observed | |
| True | False |
| Predicted | True | 1481 | 2830 |
| False | 1842 | 7918 |

|  |  |  |  |
| --- | --- | --- | --- |
| Logit prediction | | Observed | |
| True | False |
| Predicted | True | 1339 | 2355 |
| False | 1842 | 8393 |