

# Lab2

*Megan Jasek, Rohan Thakur, Charles Kekeh*

*Friday, March 04, 2016*

## Question 1

### Part 1

$$E(Y|X) = \int_0^x y * \frac{1}{x} dy = \frac{y^2}{2x} \Big|_0^x = \frac{x}{2} - 0$$
$$\mathbf{E}(\mathbf{Y}|\mathbf{X}) = \frac{\mathbf{x}}{2}$$

### Part 2

$$E(Y) = E(E(Y|X)) = E\left(\frac{x}{2}\right) = \int_0^1 \frac{x}{2} dx = \frac{x^2}{4} \Big|_0^1 = \frac{1}{4} - 0$$
$$\mathbf{E}(\mathbf{Y}) = \frac{1}{4}$$

### Part 3

$$f_{X,Y}(x,y) = f_{Y|X}(y|x) * f_X(x)$$

We know that

$$f_{Y|X}(y|x) = \frac{1}{x} \text{ and } f_X(x) = 1$$

Substituting these values in to the equation, we get

$$\mathbf{f}_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y}) = \frac{1}{\mathbf{x}}$$

### Part 4

$$f_Y(y) = \int_y^1 f_{Y|X}(y|x) * f_X(x) dx = \int_y^1 \frac{1}{x} * 1 dx$$
$$= \log(x) \Big|_y^1 = \log(1) - \log(y) = 0 - \log(y) = \log\left(\frac{1}{y}\right)$$
$$f_Y(y) = \log\left(\frac{1}{y}\right)$$

We know that

$$f_{X,Y}(x,y) = f_{X|Y}(x|y) * f_Y(y)$$

Solving for  $f_{X|Y}(x|y)$ , we get

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

Substituting, we get

$$f_{X|Y}(x|y) = \frac{\frac{1}{x}}{\log\left(\frac{1}{y}\right)}$$
$$\mathbf{f}_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) = \frac{1}{\mathbf{x} \log\left(\frac{1}{\mathbf{y}}\right)}$$

## Part 5

$$\begin{aligned} E(X|Y = \frac{1}{2}) &= \int_{\frac{1}{2}}^1 \frac{1}{x \log(2)} dx = \frac{1}{\log(2)} \int_{\frac{1}{2}}^1 \frac{1}{x} dx \\ &= \frac{1}{\log(2)} * (\log(x)|_{\frac{1}{2}}^1) = \frac{1}{\log(2)} * (\log(1) - \log(\frac{1}{2})) \\ &= \frac{1}{\log(2)} * (0 + \log(2)) = \frac{1}{\log(2)} * \log(2) \\ \mathbf{E}(\mathbf{X}|\mathbf{Y} = \frac{1}{2}) &= 1 \end{aligned}$$

## Question 2

## Question 3

$y_i, i = 1, \dots, n$  random uniform variables.

## Part 1 - Likelihood Function

$L(\theta)$  being the likelihood function, we know we have:

$$L(\theta) = f(y_1, \dots, y_n | \theta) = f(y_1 | \theta) f(y_2 | \theta) \cdots f(y_n | \theta)$$

Where  $f$  is the uniform probability density function with parameter  $\theta$ .

$$f(y_i, \theta) = \begin{cases} \frac{1}{\theta} & \text{for } 0 \leq y_i \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

Making

$$L(\theta) = \begin{cases} \frac{1}{\theta^n} & \text{for } 0 \leq y_i \leq \theta, i \in 1, \dots, n \\ 0 & \text{otherwise} \end{cases}$$

## Part 2 - MLE

Based on  $L(\theta)$  The MLE of  $\theta$  is a value of  $\theta$  for which  $\theta \geq y_i$  for  $i \in 1, \dots, n$  and which maximizes  $1/\theta^n$ .  $MLE(\theta)$  is the smallest of such values of  $\theta$  such that  $\theta \geq y_i$  for  $i \in 1, \dots, n$ . Therefore:

$$MLE(\theta) = \hat{\theta} = \max(y_1, \dots, y_n)$$

## Part 3 - Expectation n=1

Taking  $\hat{\theta} = \max(y_1, \dots, y_n)$  and  $n=1$

We have

$$\hat{\theta} = y_1$$

And

$$E[\hat{\theta}] = E[y_1] = \frac{\theta}{2}$$

Knowing that  $y_i$  is from a random uniform distribution over  $[0, \theta]$

## Part 4 - Bias

Yes, from the above,  $\hat{\theta}$  is biased. For any  $y_1, \dots, y_n$ , we expect  $\max y_1, \dots, y_n < \theta$  with probability 1. Hence  $\hat{\theta}$  underestimates  $\theta$  and we have just proven that for  $n=1$ ,  $E(\hat{\theta}) \neq \theta$

## Part 5 - Expectation general case

Taking  $\hat{\theta} = \max(y_1, \dots, y_n)$  and assuming  $n \geq 1$ .

$$E(\hat{\theta}) = E[\max(y_1, \dots, y_n)]$$

## Question 4

### 4.1 Univariate Analysis

- **wage** - The wage variable has a range from \$127 to \$2,404 with a mean of \$579 and median of \$543 with most values occurring between \$250 and \$750. The histogram shows a data distribution that's positively skewed.
- **logWage** - The logWage variable has a range from \$4.844 to \$7.785 with a mean of \$6.263 and median of \$6.297. The histogram shows a data distribution that's approximately normal.
- **education** - The education variable is an integer and has a range from 2 to 18 with a mean of 12 and median of 12. The histogram shows a data distribution that is slightly negatively skewed. There is a spike at 12 and a smaller spike at 16.
- **experience** - The experience variable is an integer and has a range from 0 to 23 with a mean of 8.788 and median of 8. The histogram shows a data distribution that is slightly positively skewed.
- **experienceSquare** - The experience variable is an integer and has a range from 0 to 529 with a mean of 95.03 and median of 64. The histogram shows a data distribution that is positively skewed. There is a spike at about 50.
- **IQscore** - The IQscore variable is an integer and has a range from 50 to 144 with a mean of 102.3 and median of 103. The histogram shows a data distribution that is approximately normal. There are 316 missing values.
- **dad\_education** - The dad\_education variable is an integer and has a range from 0 to 18 with a mean of 10.18 and median of 11. The histogram shows a data distribution that has many frequencies at about count 30 and spikes at 8 and 12. These spikes make intuitive sense because these are natural education breakpoints for people. Eight years signifying the end of middle school and 12 years indicating the end of high school. There are 239 missing values.
- **mom\_education** - The mom\_education variable is an integer and has a range from 0 to 18 with a mean of 10.45 and median of 12. The histogram shows a data distribution that has many frequencies at about count 50 and spikes at 12. This spike makes intuitive sense because 12 years indicates the end of high school which is a natural education break point for people. There are 128 missing values.
- **age** - The age variable is an integer and has a range from 24 to 34 with a mean of 28.01 and median of 27. For the ages between 24 and 28, the frequency is around 105. For the ages between 29 and 34, the frequency is around 65.
- **raceColor** - The raceColor variable is a binary variable with values 0 or 1 and mean 0.238. This means that there are about 24% 1's and 76% 0's.
- **rural** - The rural variable is a binary variable with values 0 or 1 and mean 0.391. This means that there are about 39% 1's and 61% 0's. 39% of the participants live in a rural area and 61% do not.
- **city** - The city variable is a binary variable with values 0 or 1 and mean 0.712. This means that there are about 71% 1's and 29% 0's. 71% of the participants live in a city and 29% do not.
- **z1** - The z1 variable is a binary variable with values 0 or 1 and mean 0.44. This means that there are about 44% 1's and 56% 0's.

- **z2** - The z2 variable is a binary variable with values 0 or 1 and mean 0.686. This means that there are about 69% 1's and 31% 0's.

```
# Load the data in to the df dataframe
data = read.csv("WageData2.csv", header = TRUE)
# There was already a logWage variable in the dataset, so set that one
# to logWageOLD
data$logWageOLD = data$logWage
# Create a logWage variable to use for the rest of the problem
data$logWage = log(data$wage)
# Create the experienceSquare variable
data$experienceSquare = data$experience * data$experience
```

```
# wage variable
summary(data$wage)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  127.0   400.0   543.0   578.8   702.5  2404.0
```

```
print(quantile(data$wage, probs = c(0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9,
0.95, 0.99, 1)))
```

```
##      1%      5%     10%     25%     50%     75%     90%     95%     99%
##  187.92  244.90  289.00  400.00  543.00  702.50  914.00 1068.70 1402.23
##    100%
## 2404.00
```

```
# Plot the histogram of apps at 30 bins
wage.hist <- ggplot(data, aes(wage)) + theme(legend.position = "none") +
  geom_histogram(fill = "Blue", colour = "Black", binwidth = (range(data$wage)[2] -
    range(data$wage)[1])/30) + labs(title = "Distribution of wage",
    x = "wage ($)", y = "Frequency")

plot(wage.hist)
```



```
# logWage variable
summary(data$logWage)
```

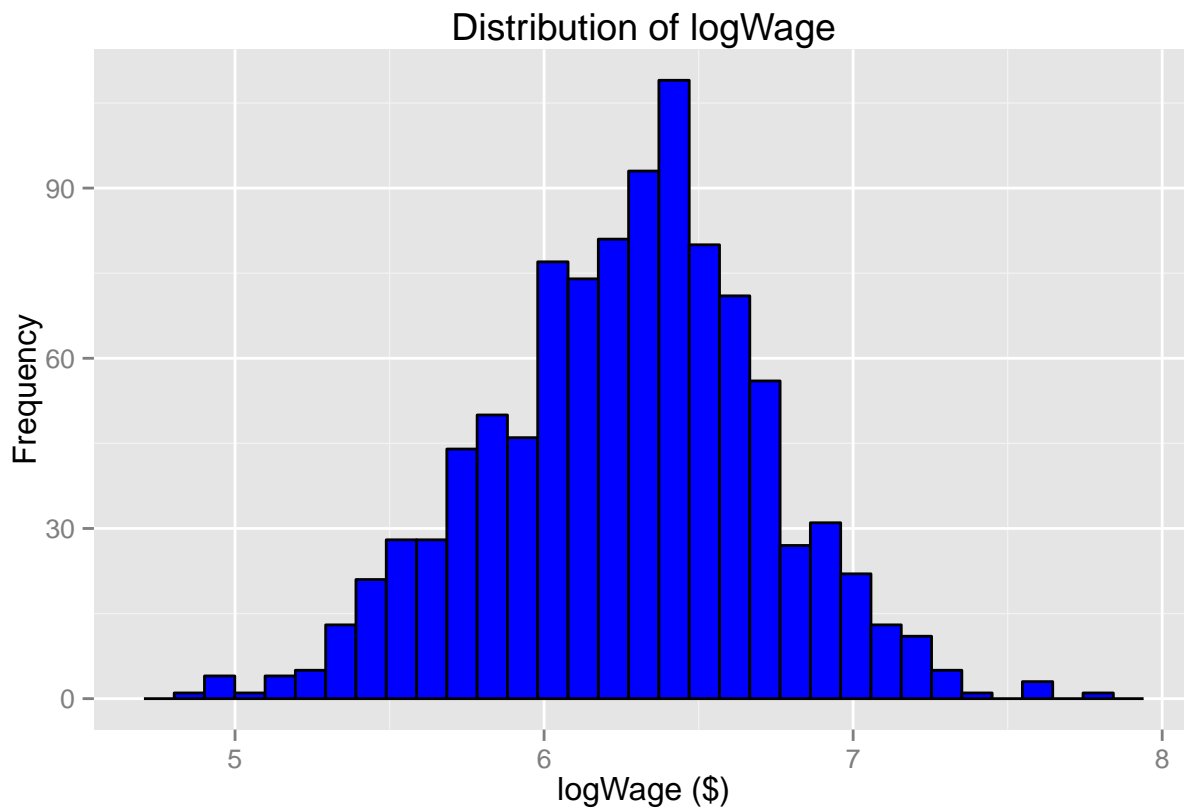
```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  4.844   5.991   6.297   6.263   6.555   7.785
```

```
print(quantile(data$logWage, probs = c(0.01, 0.05, 0.1, 0.25, 0.5, 0.75,
    0.9, 0.95, 0.99, 1)))
```

```
##      1%      5%     10%     25%     50%     75%     90%     95%
## 5.236007 5.500848 5.666427 5.991465 6.297109 6.554645 6.817825 6.974194
##      99%     100%
## 7.245818 7.784889
```

```
# Plot the histogram of apps at 30 bins
logWage.hist <- ggplot(data, aes(logWage)) + theme(legend.position = "none") +
  geom_histogram(fill = "Blue", colour = "Black", binwidth = (range(data$logWage)[2] -
    range(data$logWage)[1])/30) + labs(title = "Distribution of logWage",
    x = "logWage ($)", y = "Frequency")

plot(logWage.hist)
```



```
# education variable
summary(data$education)
```

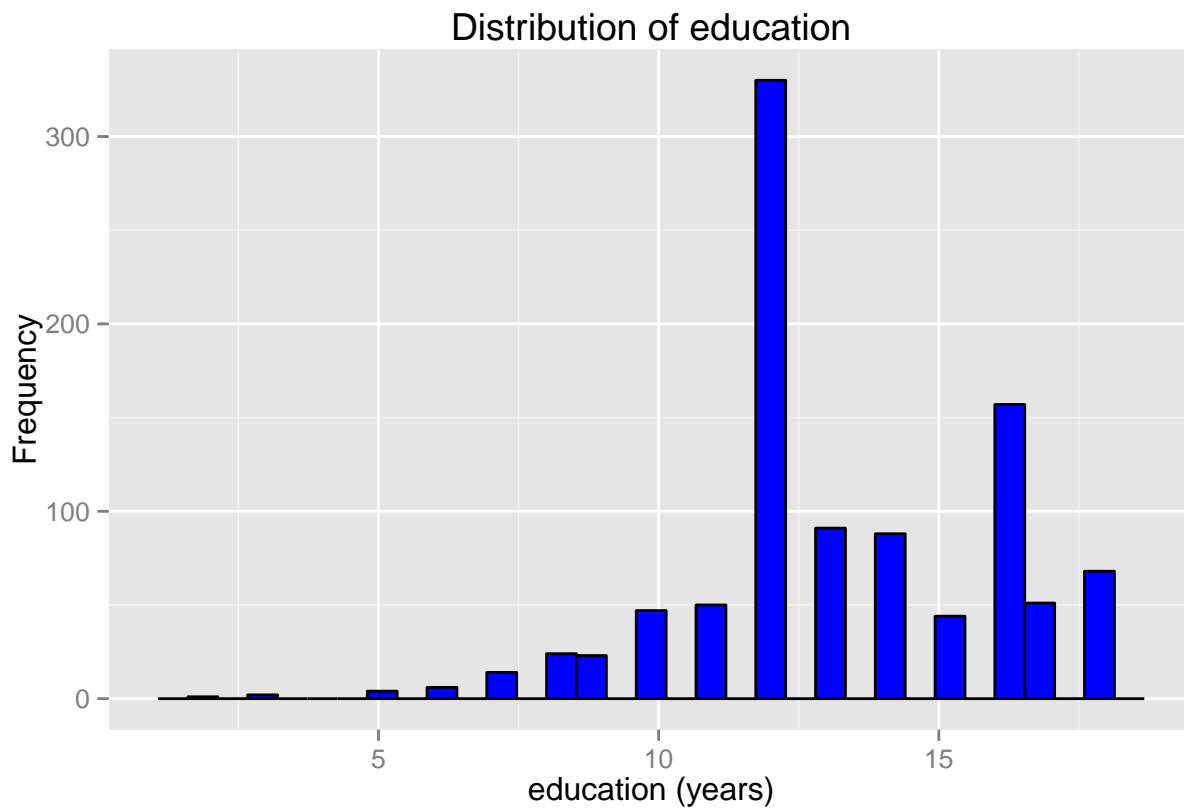
```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      2.00  12.00   12.00   13.22  16.00   18.00
```

```
print(quantile(data$education, probs = c(0.01, 0.05, 0.1, 0.25, 0.5, 0.75,
0.9, 0.95, 0.99, 1)))
```

```
##      1%    5%   10%   25%   50%   75%   90%   95%   99%  100%
##      6     8    10    12    12    16    17    18    18    18
```

```
# Plot the histogram of apps at 30 bins
education.hist <- ggplot(data, aes(education)) + theme(legend.position = "none") +
  geom_histogram(fill = "Blue", colour = "Black", binwidth = (range(data$education)[2] -
    range(data$education)[1])/30) + labs(title = "Distribution of education",
    x = "education (years)", y = "Frequency")

plot(education.hist)
```



```
# experience variable
summary(data$experience)
```

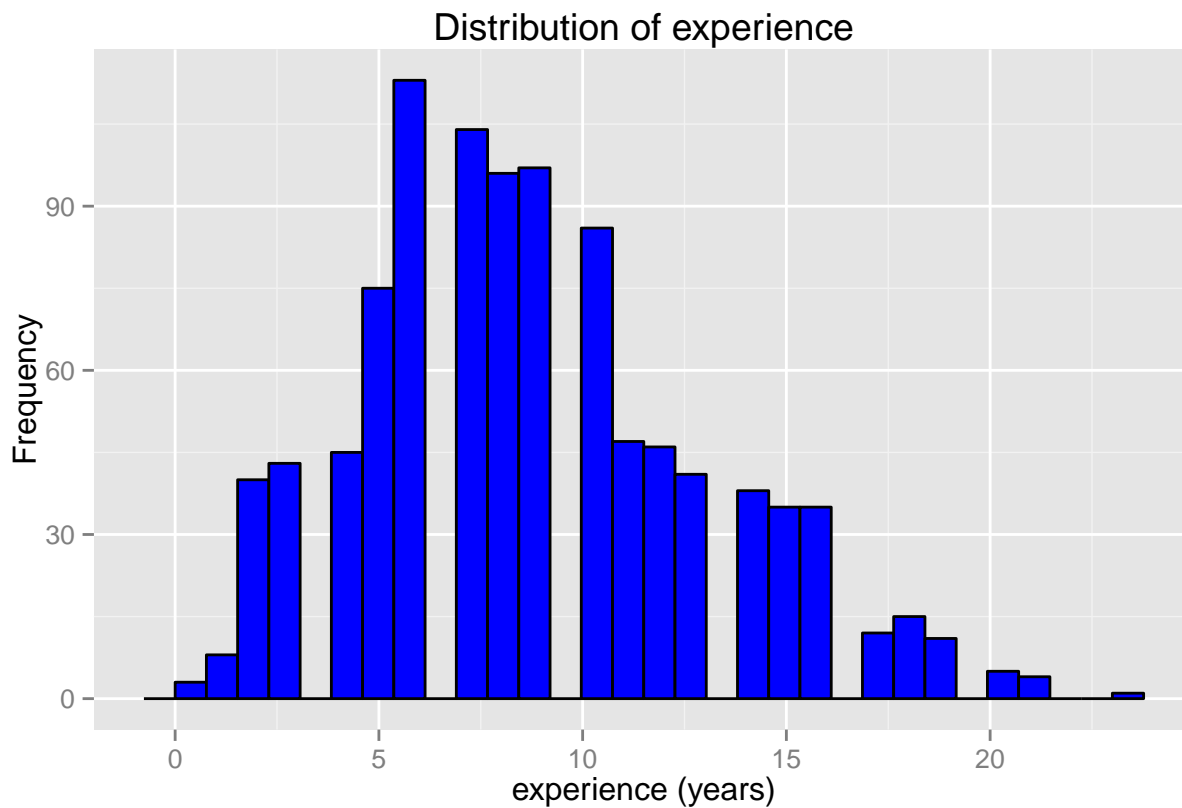
```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      0.000   6.000   8.000   8.788  11.000  23.000
```

```
print(quantile(data$experience, probs = c(0.01, 0.05, 0.1, 0.25, 0.5, 0.75,
      0.9, 0.95, 0.99, 1)))
```

```
##      1%      5%     10%     25%     50%     75%     90%     95%     99%    100%
##      1.00     2.00     4.00     6.00     8.00    11.00    15.00    16.00    19.01    23.00
```

```
# Plot the histogram of apps at 30 bins
experience.hist <- ggplot(data, aes(experience)) + theme(legend.position = "none") +
  geom_histogram(fill = "Blue", colour = "Black", binwidth = (range(data$experience)[2] -
    range(data$experience)[1])/30) + labs(title = "Distribution of experience",
    x = "experience (years)", y = "Frequency")

plot(experience.hist)
```



```
# experienceSquare variable
summary(data$experienceSquare)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      0.00  36.00   64.00   95.03 121.00  529.00
```

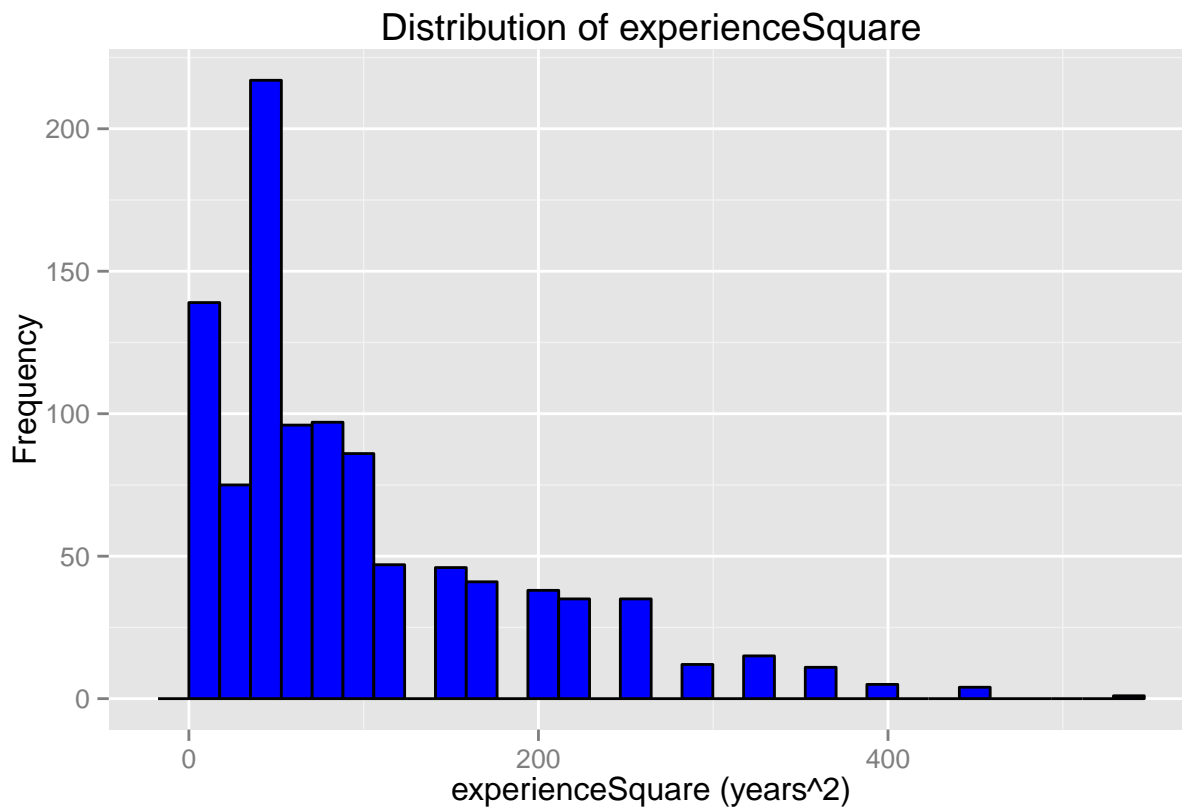
```
print(quantile(data$experienceSquare, probs = c(0.01, 0.05, 0.1, 0.25,
  0.5, 0.75, 0.9, 0.95, 0.99, 1)))
```

```
##      1%      5%     10%     25%     50%     75%     90%     95%     99%    100%
##      1.00     4.00    16.00    36.00    64.00   121.00  225.00  256.00  361.39  529.00
```

```
# Plot the histogram of apps at 30 bins
experienceSquare.hist <- ggplot(data, aes(experienceSquare)) + theme(legend.position = "none") +
  geom_histogram(fill = "Blue", colour = "Black", binwidth = (range(data$experienceSquare)[2] -
    range(data$experienceSquare)[1])/30) + labs(title = "Distribution of experienceSquare",
    x = "experienceSquare (years^2)", y = "Frequency")

plot(experienceSquare.hist)
```





```
# IQscore variable
summary(data$IQscore)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     NA's
##      50.0   93.0   103.0   102.3  113.0   144.0     316
```

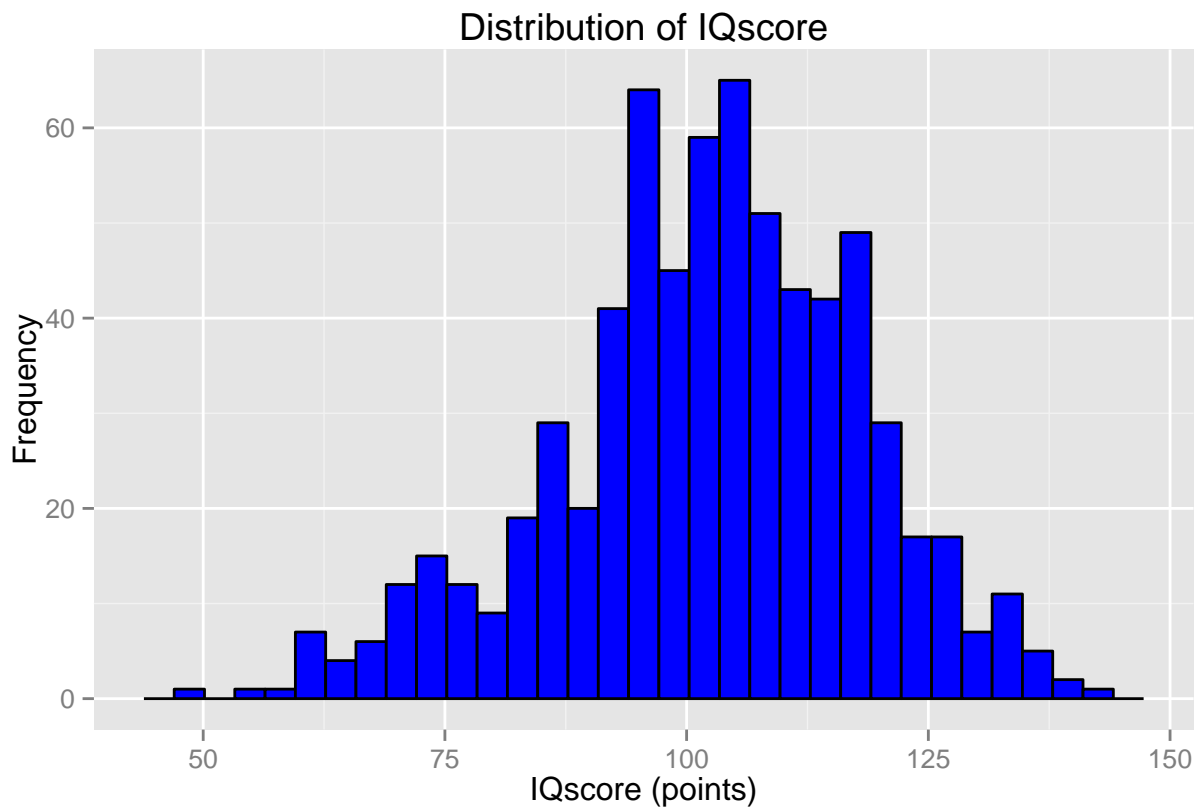
```
print(quantile(data$IQscore, probs = c(0.01, 0.05, 0.1, 0.25, 0.5, 0.75,
    0.9, 0.95, 0.99, 1), na.rm = TRUE))
```

```
##      1%      5%     10%     25%     50%     75%     90%     95%     99%    100%
##  61.83  73.15  82.00  93.00 103.00 113.00 122.00 126.85 135.00 144.00
```

```
# Plot the histogram of apps at 30 bins
IQscore.hist <- ggplot(data, aes(IQscore)) + theme(legend.position = "none") +
  geom_histogram(fill = "Blue", colour = "Black") + labs(title = "Distribution of IQscore",
    x = "IQscore (points)", y = "Frequency")

plot(IQscore.hist)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



```
# dad_education variable
summary(data$dad_education)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
##      0.00   8.00   11.00   10.18   12.00   18.00    239
```

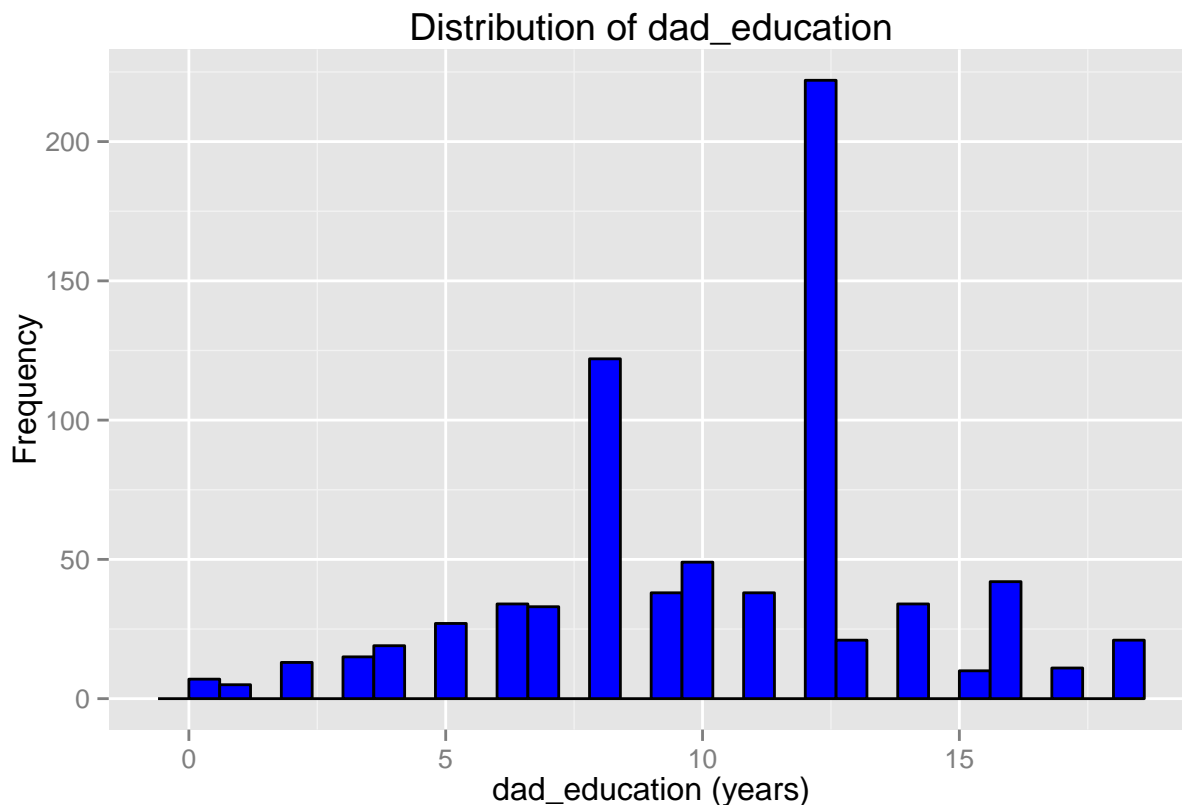
```
print(quantile(data$dad_education, probs = c(0.01, 0.05, 0.1, 0.25, 0.5,
      0.75, 0.9, 0.95, 0.99, 1), na.rm = TRUE))
```

```
##      1%    5%   10%   25%   50%   75%   90%   95%   99%  100%
##      1     3     5     8    11    12    15    16    18    18
```

```
# Plot the histogram of apps at 30 bins
dad_education.hist <- ggplot(data, aes(dad_education)) + theme(legend.position = "none") +
  geom_histogram(fill = "Blue", colour = "Black") + labs(title = "Distribution of dad_education",
    x = "dad_education (years)", y = "Frequency")

plot(dad_education.hist)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



```
# mom_education variable
summary(data$mom_education)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
##      0.00   8.00   12.00  10.45  12.00   18.00    128
```

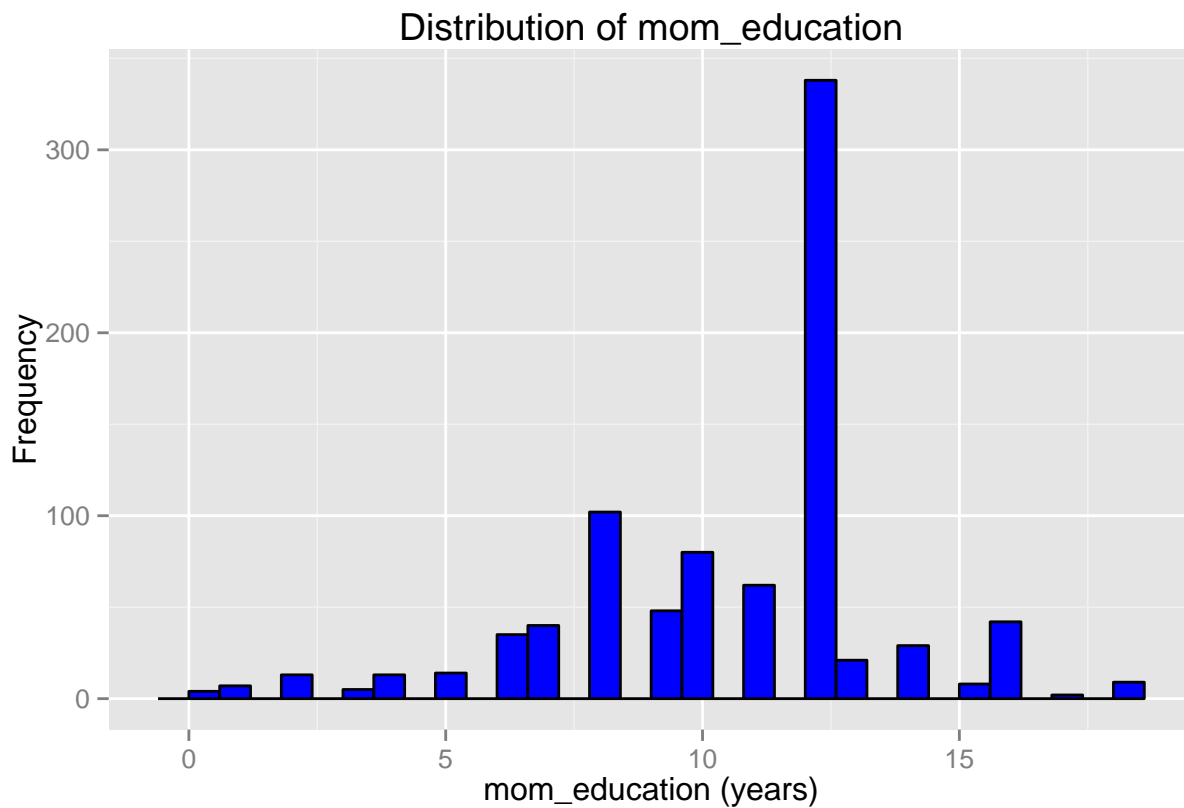
```
print(quantile(data$mom_education, probs = c(0.01, 0.05, 0.1, 0.25, 0.5,
      0.75, 0.9, 0.95, 0.99, 1), na.rm = TRUE))
```

```
##      1%      5%     10%    25%    50%    75%    90%    95%    99%   100%
##      1.00    5.00    6.00    8.00   12.00   12.00   14.00   16.00   17.29   18.00
```

```
# Plot the histogram of apps at 30 bins
mom_education.hist <- ggplot(data, aes(mom_education)) + theme(legend.position = "none") +
  geom_histogram(fill = "Blue", colour = "Black") + labs(title = "Distribution of mom_education",
    x = "mom_education (years)", y = "Frequency")

plot(mom_education.hist)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



```
# age variable
summary(data$age)
```

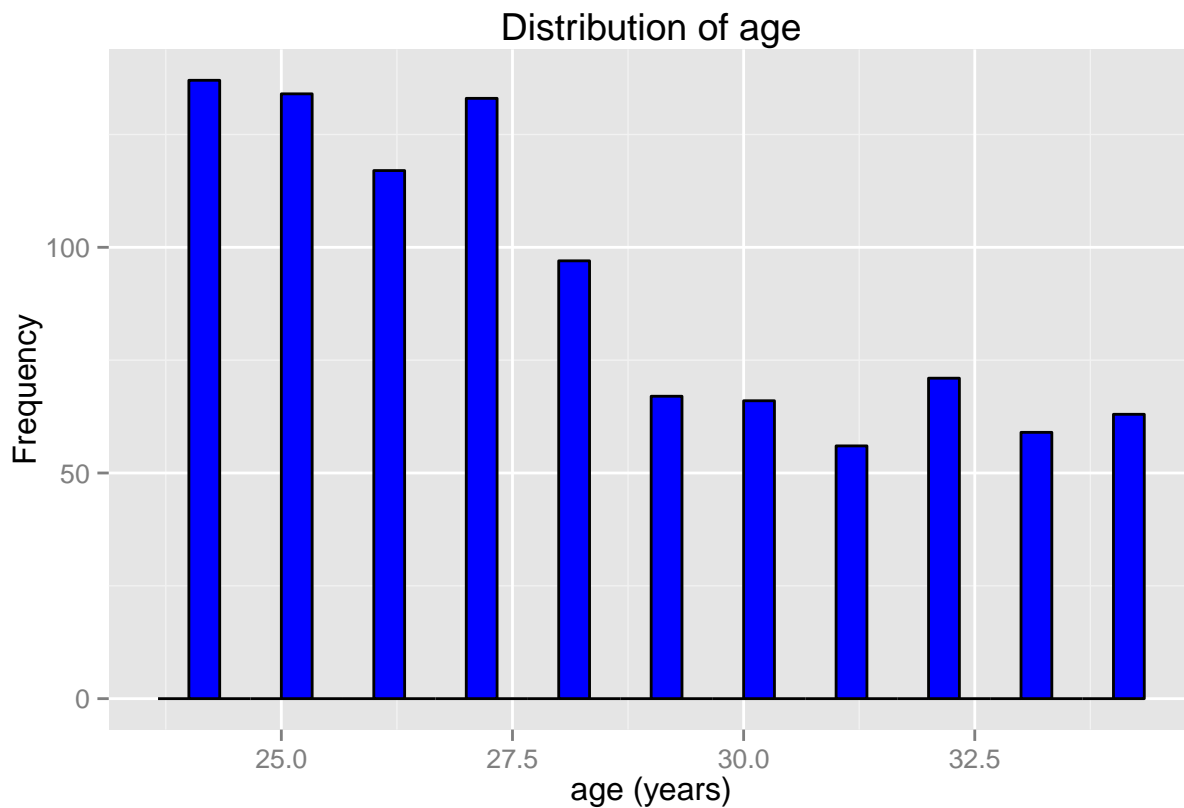
```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  24.00  25.00   27.00   28.01  30.00   34.00
```

```
print(quantile(data$age, probs = c(0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9,
  0.95, 0.99, 1), na.rm = TRUE))
```

```
##      1%    5%   10%   25%   50%   75%   90%   95%   99%  100%
##      24    24    24    25    27    30    33    34    34    34
```

```
# Plot the histogram of apps at 30 bins
age.hist <- ggplot(data, aes(age)) + theme(legend.position = "none") +
  geom_histogram(fill = "Blue", colour = "Black", binwidth = (range(data$age)[2] -
    range(data$age)[1])/30) + labs(title = "Distribution of age", x = "age (years)",
    y = "Frequency")

plot(age.hist)
```



```
# raceColor variable
summary(data$raceColor)
```

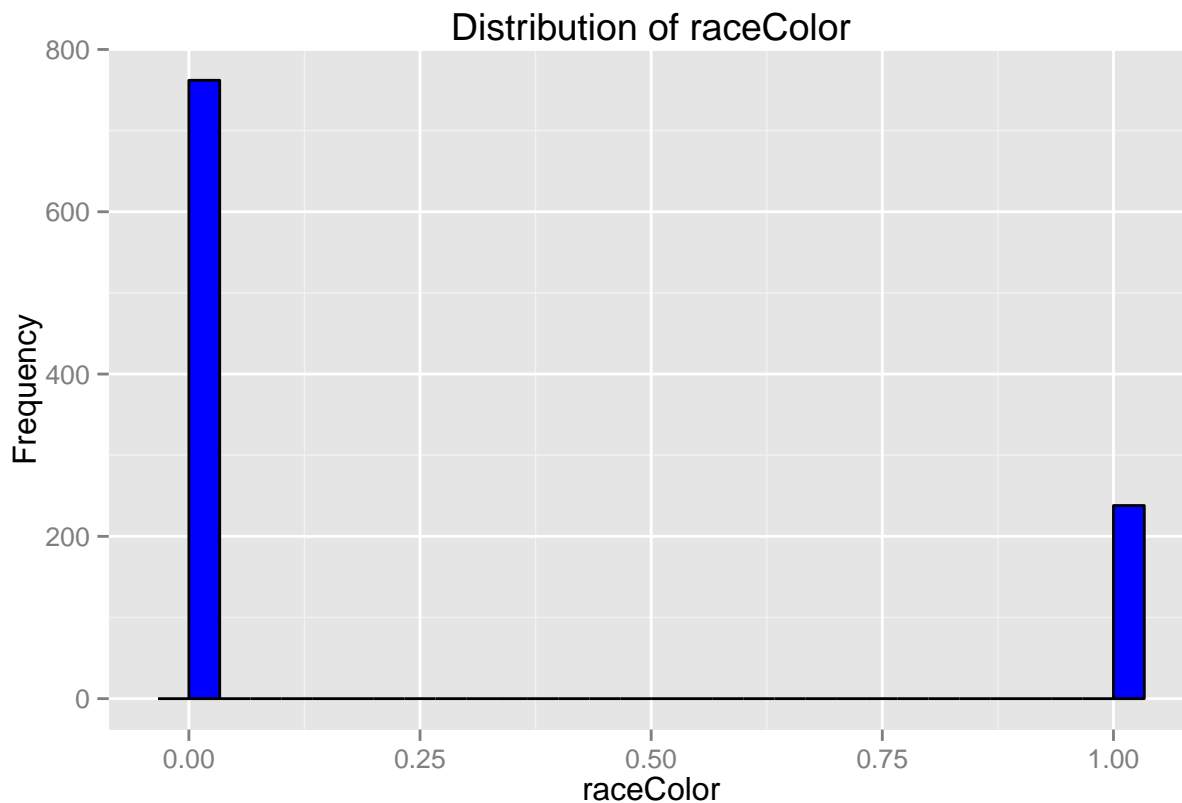
```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      0.000  0.000   0.000  0.238  0.000   1.000
```

```
print(quantile(data$raceColor, probs = c(0.01, 0.05, 0.1, 0.25, 0.5, 0.75,
      0.9, 0.95, 0.99, 1), na.rm = TRUE))
```

```
##      1%   5%  10%  25%  50%  75%  90%  95%  99% 100%
##      0    0    0    0    0    0    1    1    1    1
```

```
# Plot the histogram of apps at 30 bins
raceColor.hist <- ggplot(data, aes(raceColor)) + theme(legend.position = "none") +
  geom_histogram(fill = "Blue", colour = "Black", binwidth = (range(data$raceColor)[2] -
    range(data$raceColor)[1])/30) + labs(title = "Distribution of raceColor",
    x = "raceColor", y = "Frequency")

plot(raceColor.hist)
```



```
# rural variable
summary(data$rural)
```

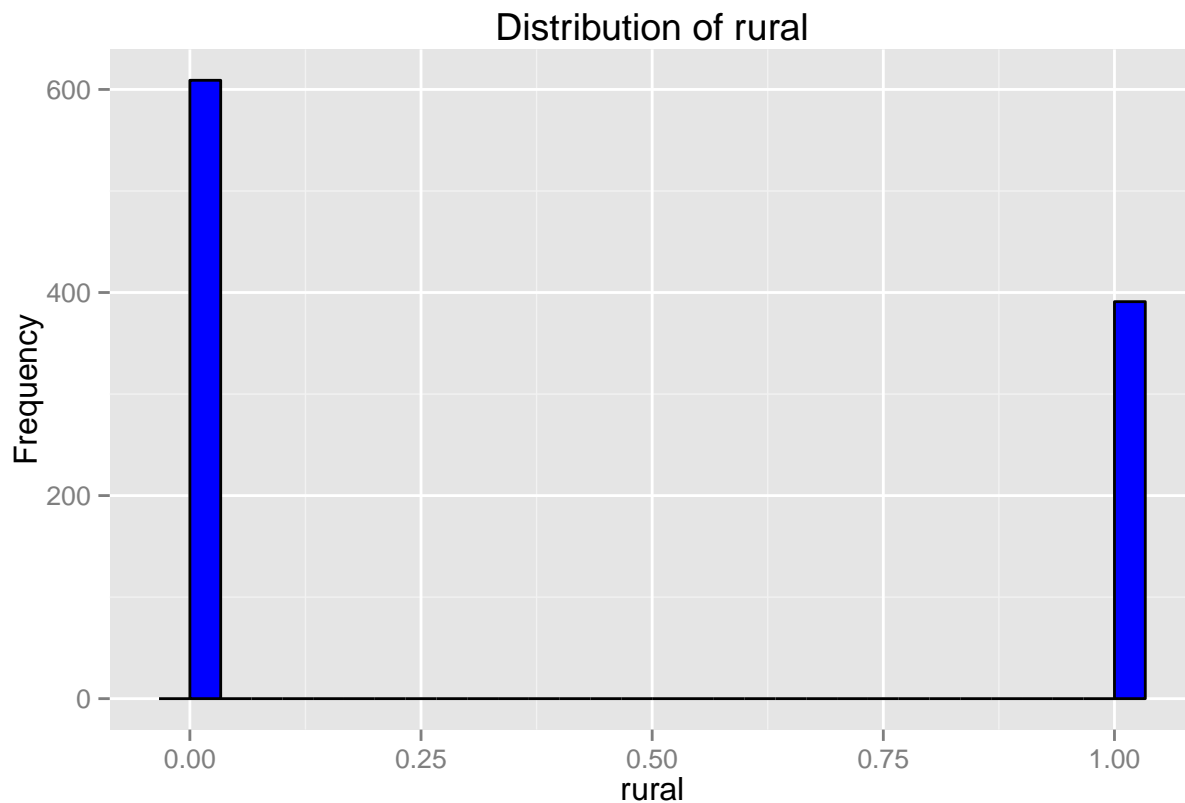
```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.000  0.000   0.000   0.391  1.000   1.000
```

```
print(quantile(data$rural, probs = c(0.01, 0.05, 0.1, 0.25, 0.5, 0.75,
  0.9, 0.95, 0.99, 1), na.rm = TRUE))
```

```
##      1%   5%  10%  25%  50%  75%  90%  95%  99% 100%
##      0    0    0    0    0    1    1    1    1    1
```

```
# Plot the histogram of apps at 30 bins
rural.hist <- ggplot(data, aes(rural)) + theme(legend.position = "none") +
  geom_histogram(fill = "Blue", colour = "Black", binwidth = (range(data$rural)[2] -
    range(data$rural)[1])/30) + labs(title = "Distribution of rural",
    x = "rural", y = "Frequency")

plot(rural.hist)
```



```
# city variable
summary(data$city)
```

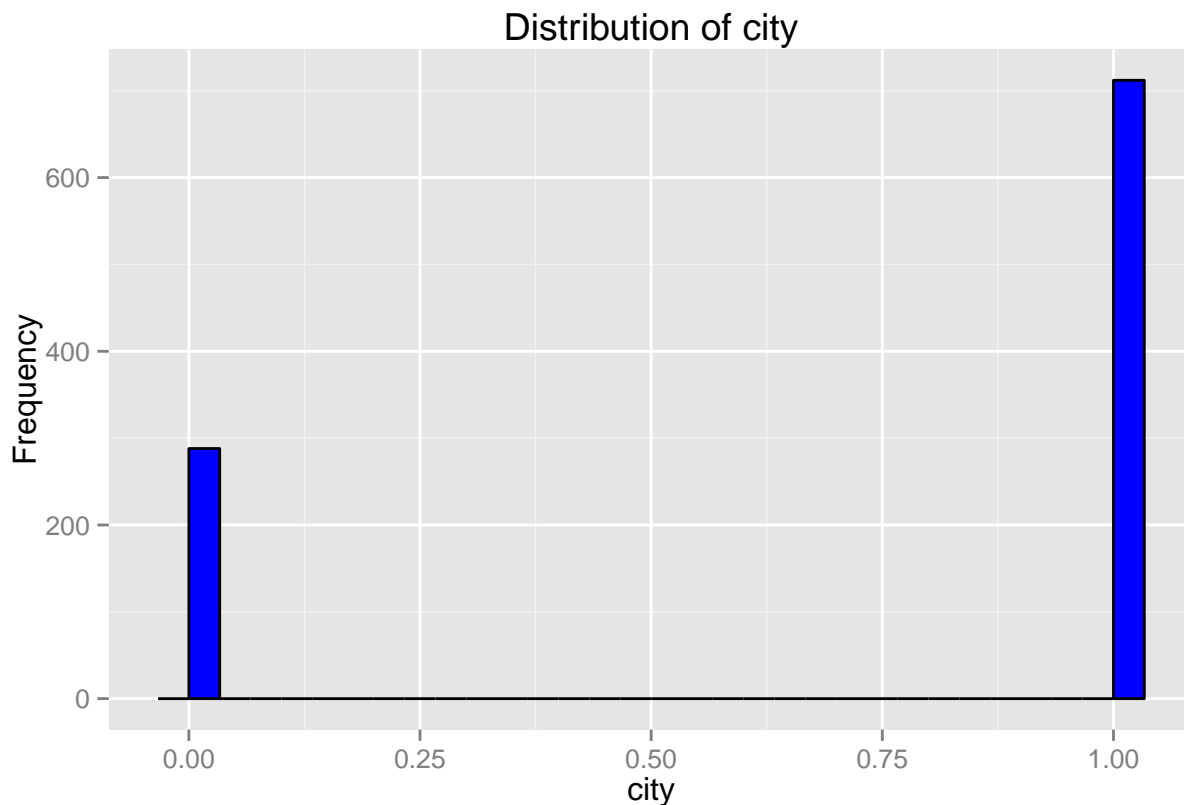
```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      0.000  0.000   1.000   0.712  1.000   1.000
```

```
print(quantile(data$city, probs = c(0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9,
  0.95, 0.99, 1), na.rm = TRUE))
```

```
##      1%   5%  10%  25%  50%  75%  90%  95%  99% 100%
##      0    0    0    0    1    1    1    1    1    1
```

```
# Plot the histogram of apps at 30 bins
city.hist <- ggplot(data, aes(city)) + theme(legend.position = "none") +
  geom_histogram(fill = "Blue", colour = "Black", binwidth = (range(data$city)[2] -
    range(data$city)[1])/30) + labs(title = "Distribution of city",
    x = "city", y = "Frequency")

plot(city.hist)
```



```
# z1 variable
summary(data$z1)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      0.00   0.00   0.00   0.44   1.00   1.00
```

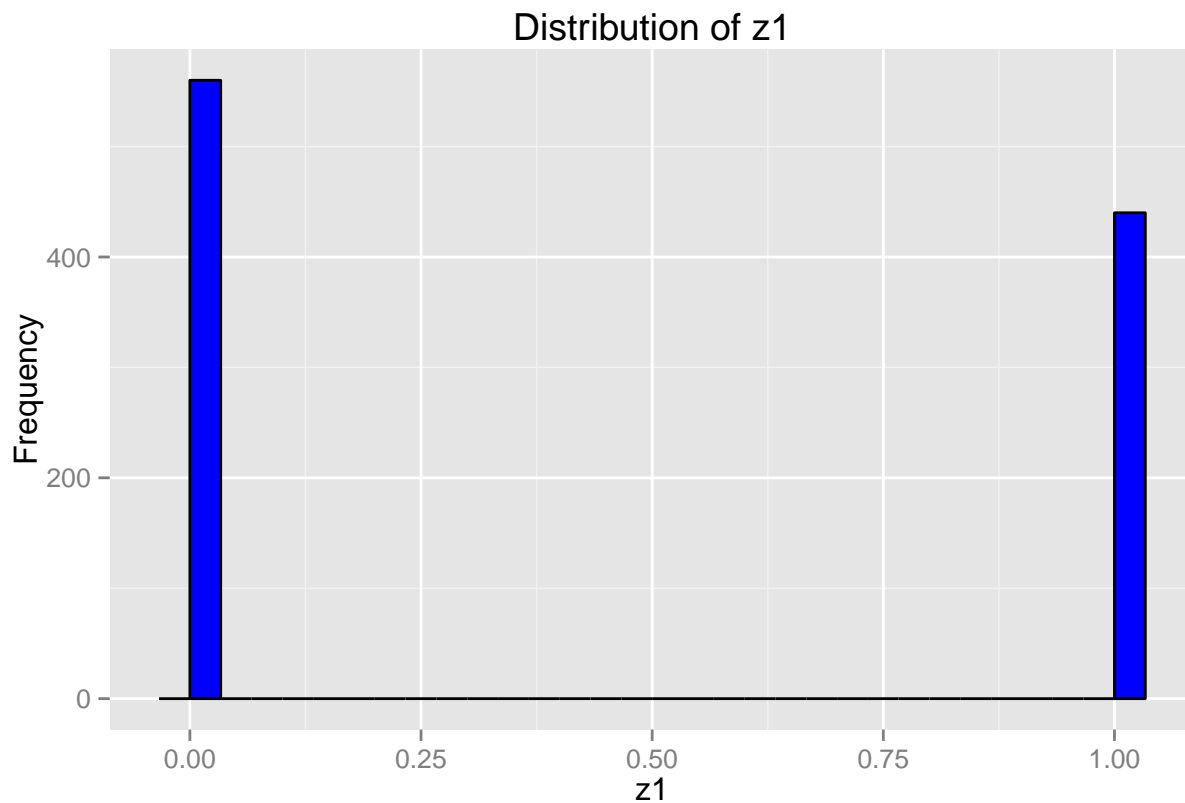
```
print(quantile(data$z1, probs = c(0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9,
0.95, 0.99, 1), na.rm = TRUE))
```

```
##      1%   5%  10%  25%  50%  75%  90%  95%  99% 100%
##      0    0    0    0    0    1    1    1    1    1
```

```
# Plot the histogram of apps at 30 bins
z1.hist <- ggplot(data, aes(z1)) + theme(legend.position = "none") + geom_histogram(fill = "Blue",
  colour = "Black", binwidth = (range(data$z1)[2] - range(data$z1)[1])/30) +
  labs(title = "Distribution of z1", x = "z1", y = "Frequency")

plot(z1.hist)
```





```
# z2 variable
summary(data$z2)
```

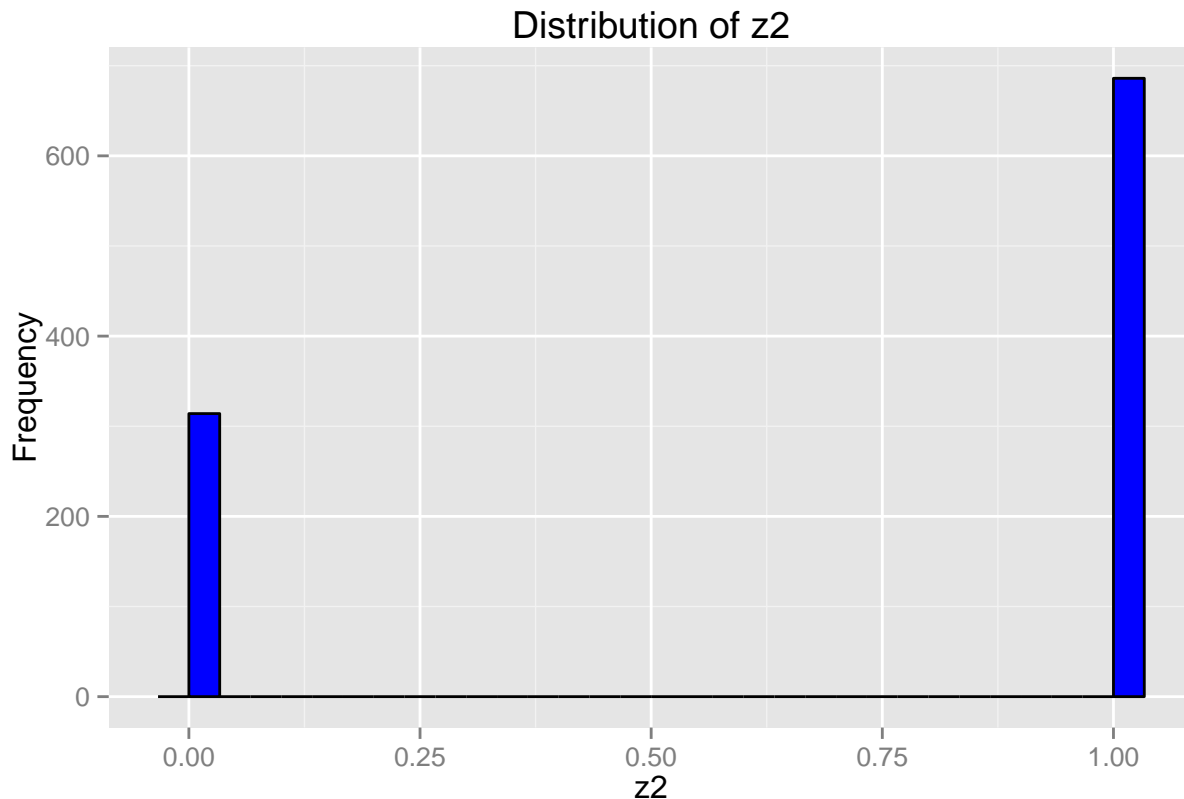
```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      0.000  0.000   1.000   0.686   1.000   1.000
```

```
print(quantile(data$z2, probs = c(0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9,
                                0.95, 0.99, 1), na.rm = TRUE))
```

```
##      1%   5%  10%  25%  50%  75%  90%  95%  99% 100%
##      0    0   0    0    1    1    1    1    1    1
```

```
# Plot the histogram of apps at 30 bins
z2.hist <- ggplot(data, aes(z2)) + theme(legend.position = "none") + geom_histogram(fill = "Blue",
  colour = "Black", binwidth = (range(data$z2)[2] - range(data$z2)[1])/30) +
  labs(title = "Distribution of z2", x = "z2", y = "Frequency")

plot(z2.hist)
```

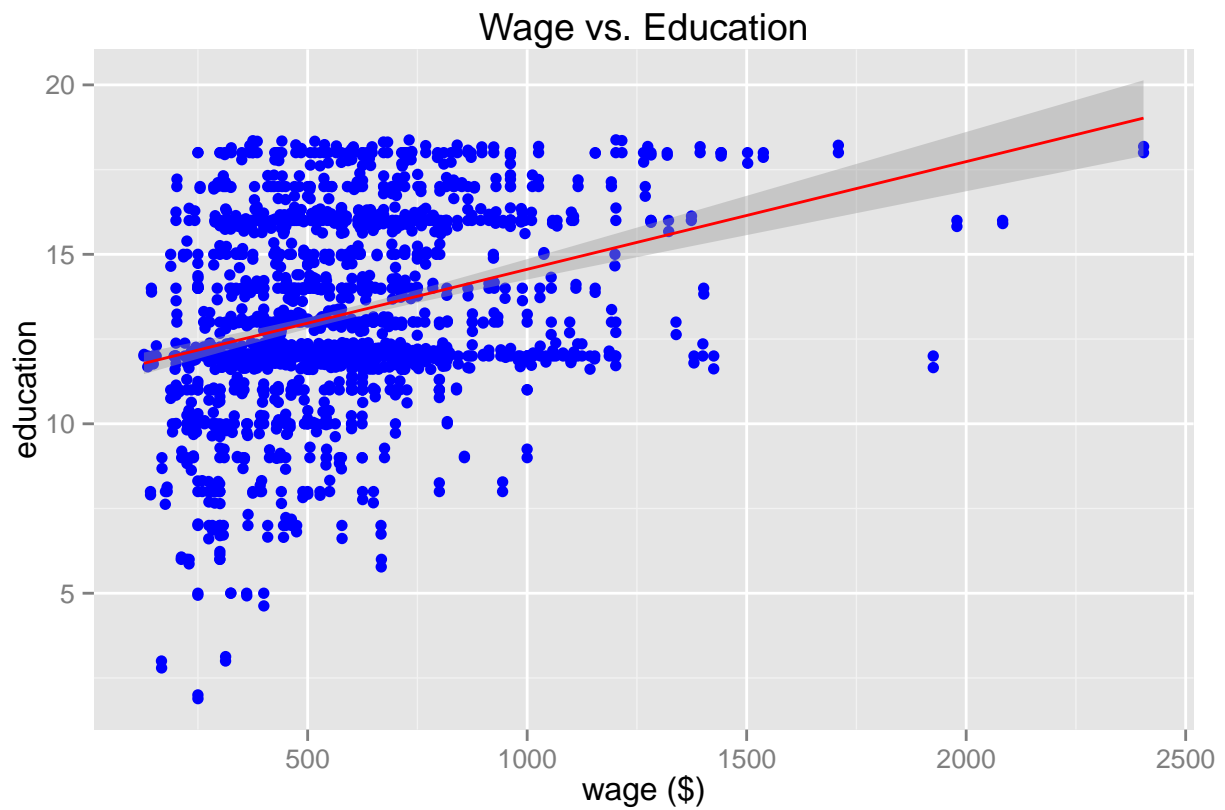


## 4.2 Bivariate Analysis

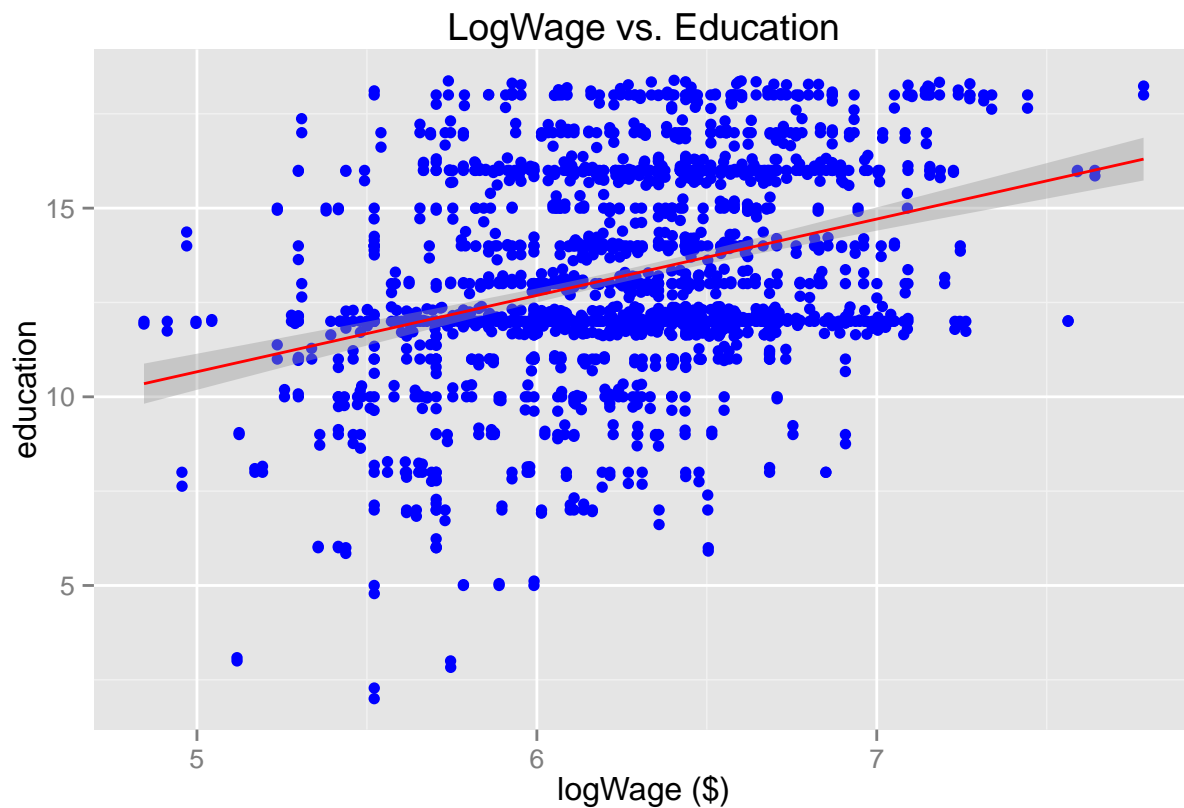
- **wage, logWage vs. education** - Both wage and logWage are weakly correlated with education with a correlation value of about 0.3. The wage vs. education scatterplot shows a possible linear trend.
- **wage, logWage vs. experience** - Both wage and logWage appear uncorrelated with experience with very low correlation values of -0.0060 and -0.0290, respectively. The wage vs. experience scatterplot shows that experience is not affected by wage for the most part. The logWage vs. experience scatterplot shows that experience is not affected by logWage as well.
- **wage, logWage vs. experienceSquare** - Both wage and logWage appear uncorrelated with experienceSquare with very low correlation values of -0.043 and -0.065, respectively. The wage vs. experienceSquare scatterplot shows that experienceSquare is not affected by wage for the most part. The logWage vs. experienceSquare scatterplot shows that experience is not affected by logWage as well.
- **wage, logWage vs. IQscore** - Both wage and logWage are weakly correlated with IQscore with low correlation values of 0.186 and 0.201, respectively. The wage and logWage vs. IQscore scatterplots show that IQscore affects wage and logWage slightly. As wage or logWage go up, IQscore increases by a small amount.
- **wage, logWage vs. dad\_education** - Both wage and logWage are weakly correlated with dad\_education with low correlation values of 0.19 and 0.19, respectively. The wage and logWage vs. dad\_education scatterplots show that dad\_education affects wage and logWage slightly. As wage or logWage go up, dad\_education increases by a small amount.
- **wage, logWage vs. mom\_education** - Both wage and logWage are weakly correlated with mom\_education with low correlation values of 0.20 and 0.21, respectively. The wage and logWage vs. mom\_education scatterplots show that mom\_education affects wage and logWage slightly. As wage or logWage go up, mom\_education increases by a small amount.

- **wage, logWage vs. age** - Both wage and logWage are weakly correlated with age with low correlation values of 0.26 and 0.25, respectively. The wage and logWage vs. age scatterplots show that age affects wage and logWage slightly. As wage or logWage go up, age increases by a small amount.
- **wage, logWage vs. raceColor** - Both wage and logWage are weakly correlated with raceColor with low correlation values of -0.30 and -0.34, respectively. The wage and logWage vs. raceColor scatterplots show that raceColor affects wage and logWage slightly. As wage or logWage go up, there are fewer people that have the raceColor variable set to 1.
- **wage, logWage vs. rural** - Both wage and logWage are weakly correlated with rural with low correlation values of -0.22 and -0.25, respectively. The wage and logWage vs. rural scatterplots show that rural affects wage and logWage slightly. As wage or logWage go up, there are fewer people that have the rural variable set to 1.
- **wage, logWage vs. city** - Both wage and logWage are weakly correlated with city with low correlation values of 0.22 and 0.24, respectively. The wage and logWage vs. rural scatterplots show that city affects wage and logWage slightly. As wage or logWage go up, there are more people that have the city variable set to 1.
- **wage, logWage vs. z1** - Both wage and logWage are weakly correlated with z1 with low correlation values of 0.101 and 0.087, respectively. The wage and logWage vs. z1 scatterplots show that z1 affects wage and logWage slightly. As wage or logWage go up, there are more people that have the z1 variable set to 1.
- **wage, logWage vs. z2** - Both wage and logWage are weakly correlated with z2 with low correlation values of 0.17 and 0.18, respectively. The wage and logWage vs. z2 scatterplots show that z2 affects wage and logWage slightly. As wage or logWage go up, there are more people that have the z2 variable set to 1. z2 shows a slightly stronger correlation with wage and logWage than z1.

```
# Scatter plot with wage variable
wage.education.plot = ggplot(data, aes(x = wage, y = education)) + theme(legend.position = "none") +
  geom_point(colour = "Blue") + geom_jitter(colour = "Blue") + geom_smooth(colour = "red",
    method = "lm") + labs(title = "Wage vs. Education", x = "wage ($)",
    y = "education")
plot(wage.education.plot)
```



```
# Scatter plot with logWage variable
lwage.education.plot = ggplot(data, aes(x = logWage, y = education)) +
  theme(legend.position = "none") + geom_point(colour = "Blue") + geom_jitter(colour = "Blue") +
  geom_smooth(colour = "red", method = "lm") + labs(title = "LogWage vs. Education",
  x = "logWage ($)", y = "education")
plot(lwage.education.plot)
```



```
# Run correlations with wage and logWage variables
cor(data$wage, data$education)
```

```
## [1] 0.3103986
```

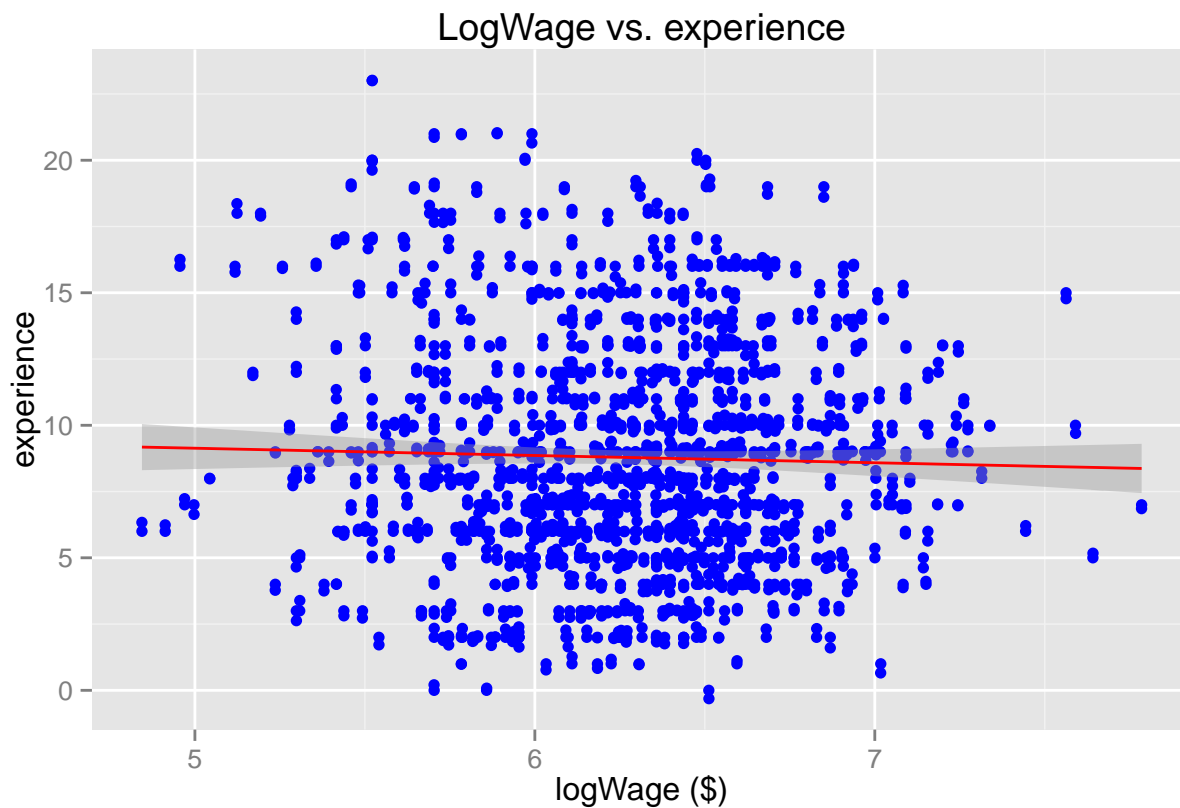
```
cor(data$logWage, data$education)
```

```
## [1] 0.3318494
```

```
# Scatter plot with wage variable
wage.experience.plot = ggplot(data, aes(x = wage, y = experience)) + theme(legend.position = "none") +
  geom_point(colour = "Blue") + geom_jitter(colour = "Blue") + geom_smooth(colour = "red",
    method = "lm") + labs(title = "Wage vs. experience", x = "wage ($)",
    y = "experience")
plot(wage.experience.plot)
```



```
# Scatter plot with logWage variable
lwage.experience.plot = ggplot(data, aes(x = logWage, y = experience)) +
  theme(legend.position = "none") + geom_point(colour = "Blue") + geom_jitter(colour = "Blue") +
  geom_smooth(colour = "red", method = "lm") + labs(title = "LogWage vs. experience",
  x = "logWage ($)", y = "experience")
plot(lwage.experience.plot)
```



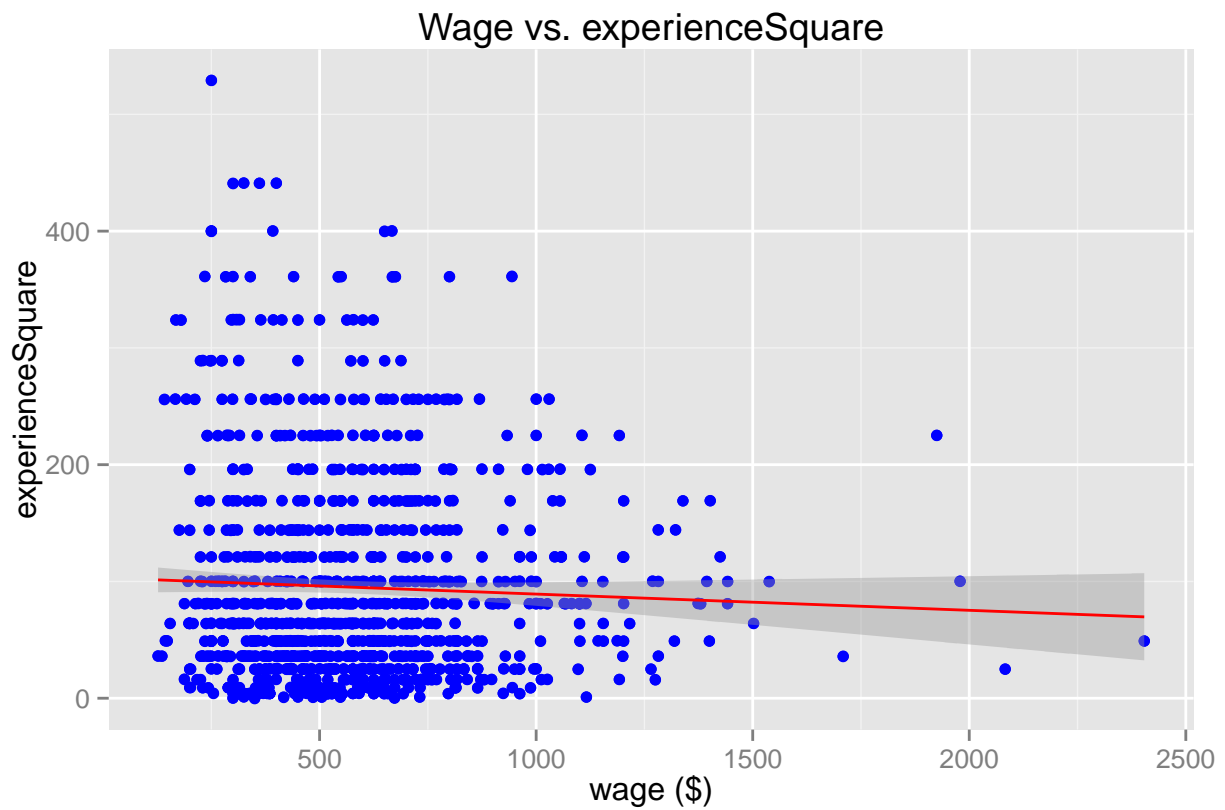
```
# Run correlations with wage and logWage variables
cor(data$wage, data$experience)
```

```
## [1] -0.005985988
```

```
cor(data$logWage, data$experience)
```

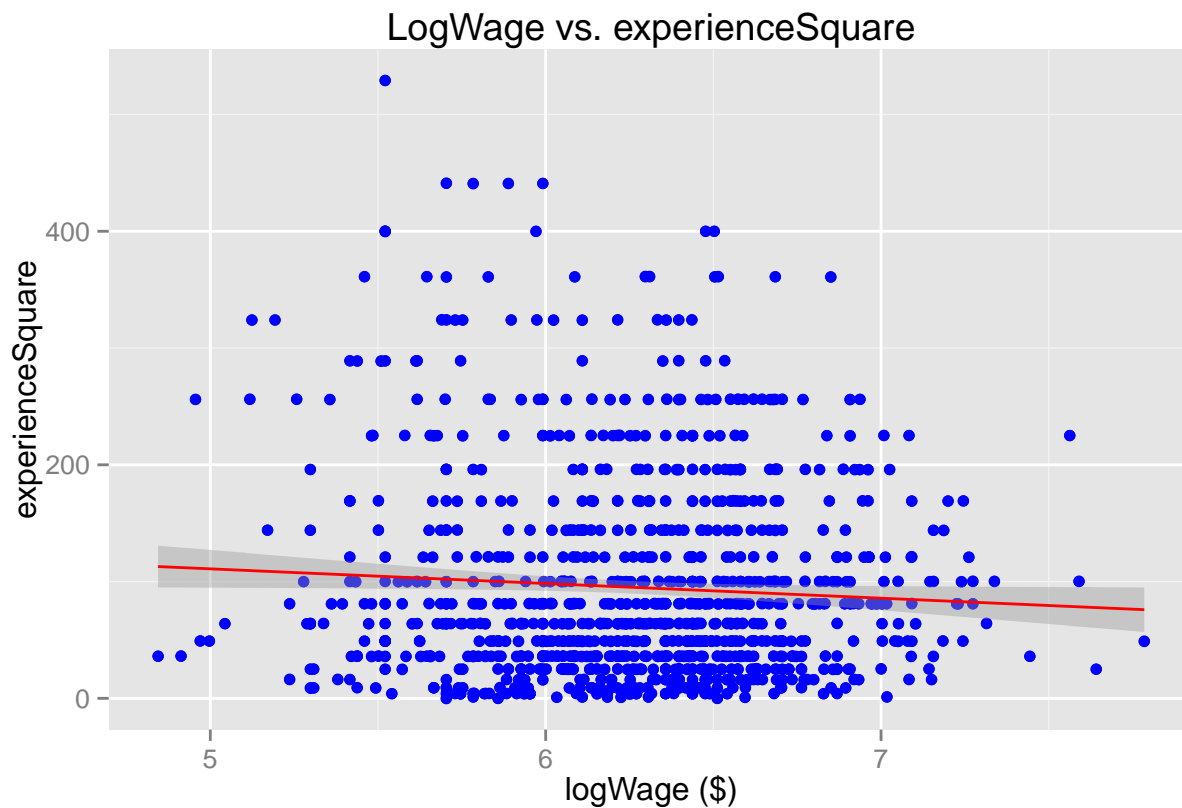
```
## [1] -0.02905727
```

```
# Scatter plot with wage variable
wage.experienceSquare.plot = ggplot(data, aes(x = wage, y = experienceSquare)) +
  theme(legend.position = "none") + geom_point(colour = "Blue") + geom_jitter(colour = "Blue") +
  geom_smooth(colour = "red", method = "lm") + labs(title = "Wage vs. experienceSquare",
  x = "wage ($)", y = "experienceSquare")
plot(wage.experienceSquare.plot)
```



```
# Scatter plot with logWage variable
lwage.experienceSquare.plot = ggplot(data, aes(x = logWage, y = experienceSquare)) +
  theme(legend.position = "none") + geom_point(colour = "Blue") + geom_jitter(colour = "Blue") +
  geom_smooth(colour = "red", method = "lm") + labs(title = "LogWage vs. experienceSquare",
  x = "logWage ($)", y = "experienceSquare")
plot(lwage.experienceSquare.plot)
```





```
# Run correlations with wage and logWage variables
cor(data$wage, data$experienceSquare)
```

```
## [1] -0.04270455
```

```
cor(data$logWage, data$experienceSquare)
```

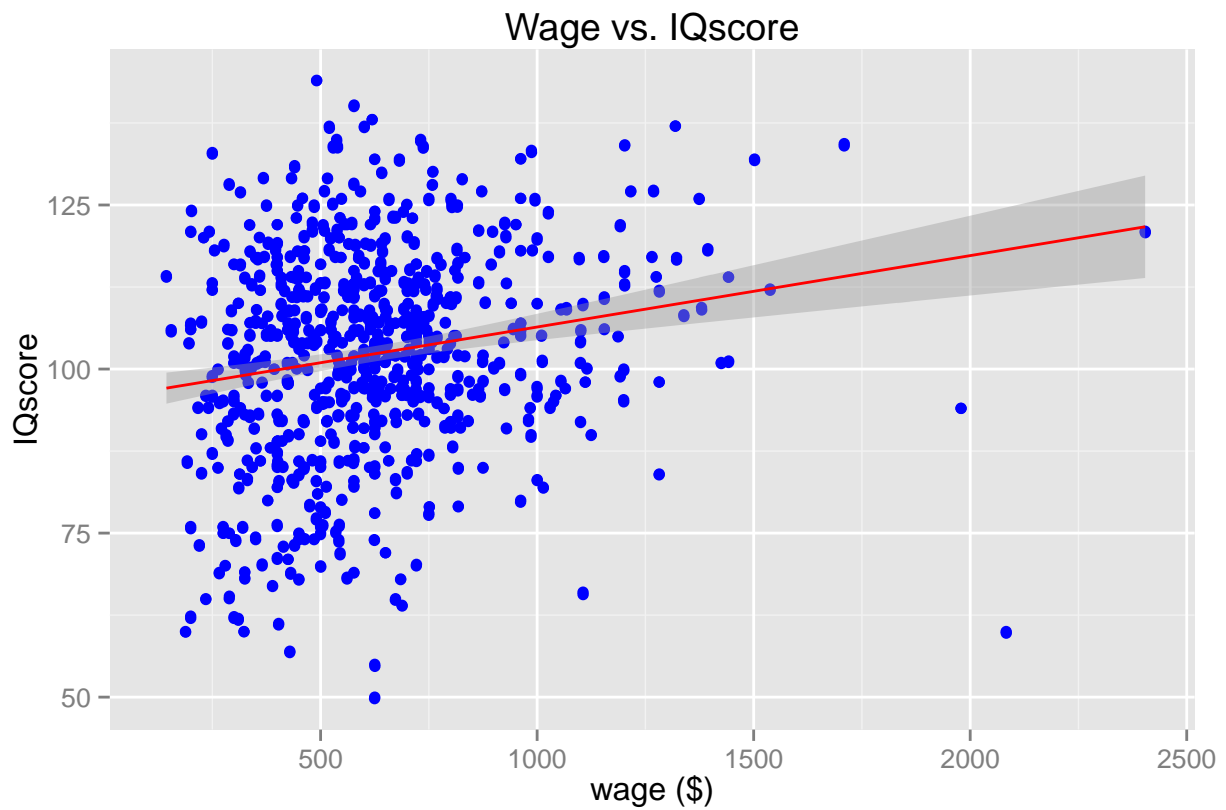
```
## [1] -0.0647476
```

```
# Scatter plot with wage variable
wage.IQscore.plot = ggplot(data, aes(x = wage, y = IQscore)) + theme(legend.position = "none") +
  geom_point(colour = "Blue") + geom_jitter(colour = "Blue") + geom_smooth(colour = "red",
    method = "lm") + labs(title = "Wage vs. IQscore", x = "wage ($)", y = "IQscore")
plot(wage.IQscore.plot)
```

```
## Warning: Removed 316 rows containing missing values (stat_smooth).
```

```
## Warning: Removed 316 rows containing missing values (geom_point).
```

```
## Warning: Removed 316 rows containing missing values (geom_point).
```

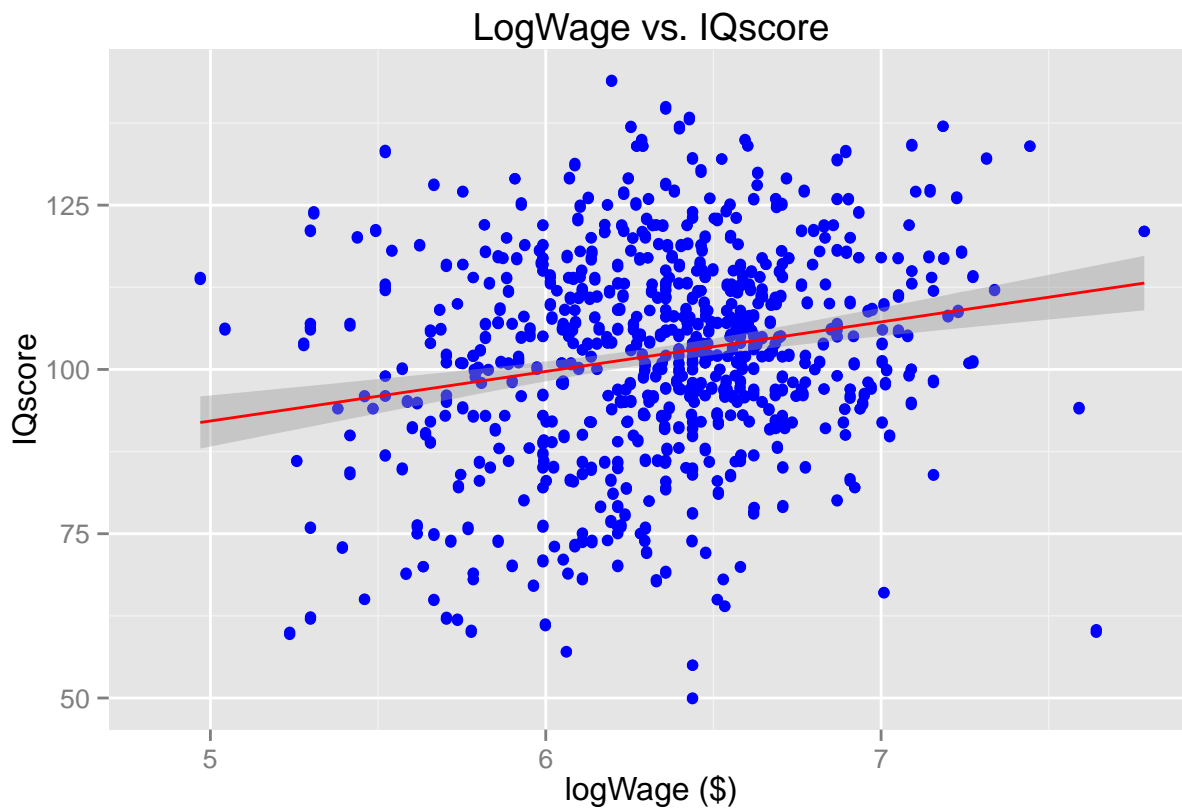


```
# Scatter plot with logWage variable
lwage.IQscore.plot = ggplot(data, aes(x = logWage, y = IQscore)) + theme(legend.position = "none") +
  geom_point(colour = "Blue") + geom_jitter(colour = "Blue") + geom_smooth(colour = "red",
  method = "lm") + labs(title = "LogWage vs. IQscore", x = "logWage ($)",
  y = "IQscore")
plot(lwage.IQscore.plot)
```

```
## Warning: Removed 316 rows containing missing values (stat_smooth).
```

```
## Warning: Removed 316 rows containing missing values (geom_point).
```

```
## Warning: Removed 316 rows containing missing values (geom_point).
```



```
# Run correlations with wage and logWage variables
cor(data$wage, data$IQscore, use = "complete.obs")
```

```
## [1] 0.1858557
```

```
cor(data$logWage, data$IQscore, use = "complete.obs")
```

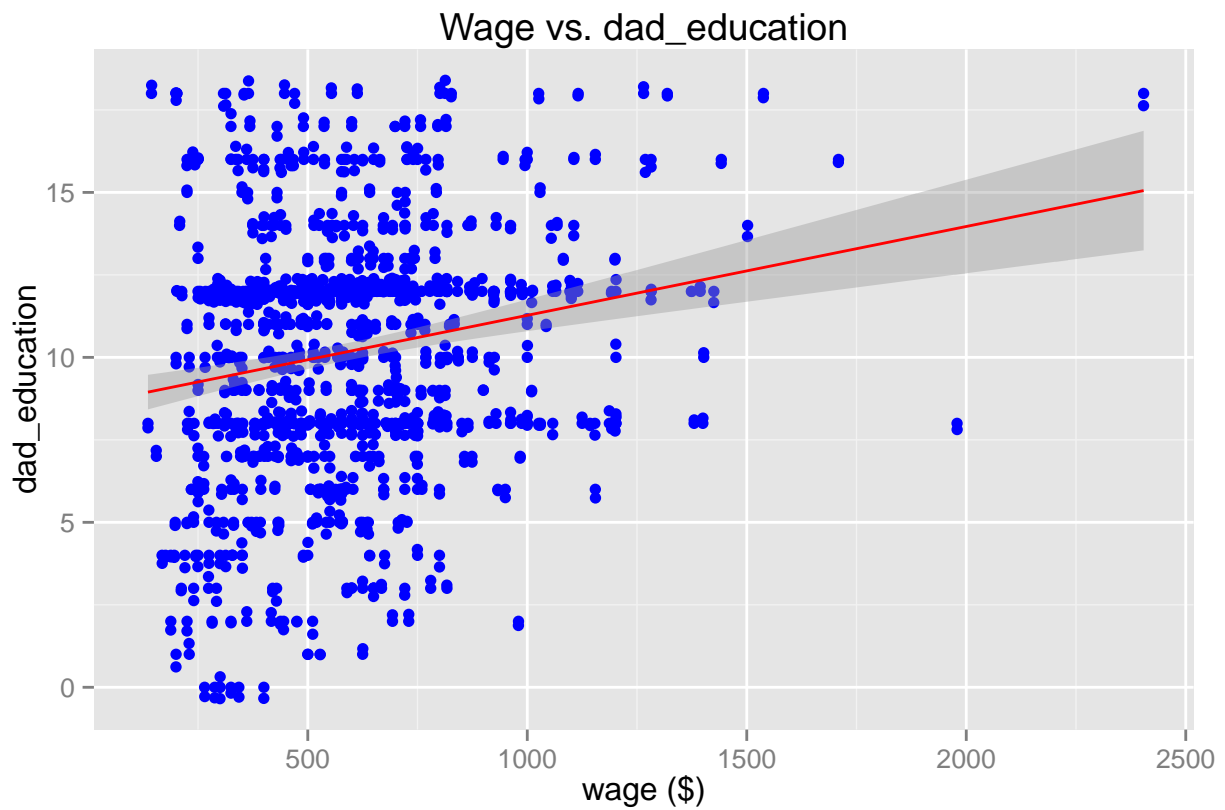
```
## [1] 0.2009578
```

```
# Scatter plot with wage variable
wage.dad_education.plot = ggplot(data, aes(x = wage, y = dad_education)) +
  theme(legend.position = "none") + geom_point(colour = "Blue") + geom_jitter(colour = "Blue") +
  geom_smooth(colour = "red", method = "lm") + labs(title = "Wage vs. dad_education",
  x = "wage ($)", y = "dad_education")
plot(wage.dad_education.plot)
```

```
## Warning: Removed 239 rows containing missing values (stat_smooth).
```

```
## Warning: Removed 239 rows containing missing values (geom_point).
```

```
## Warning: Removed 239 rows containing missing values (geom_point).
```

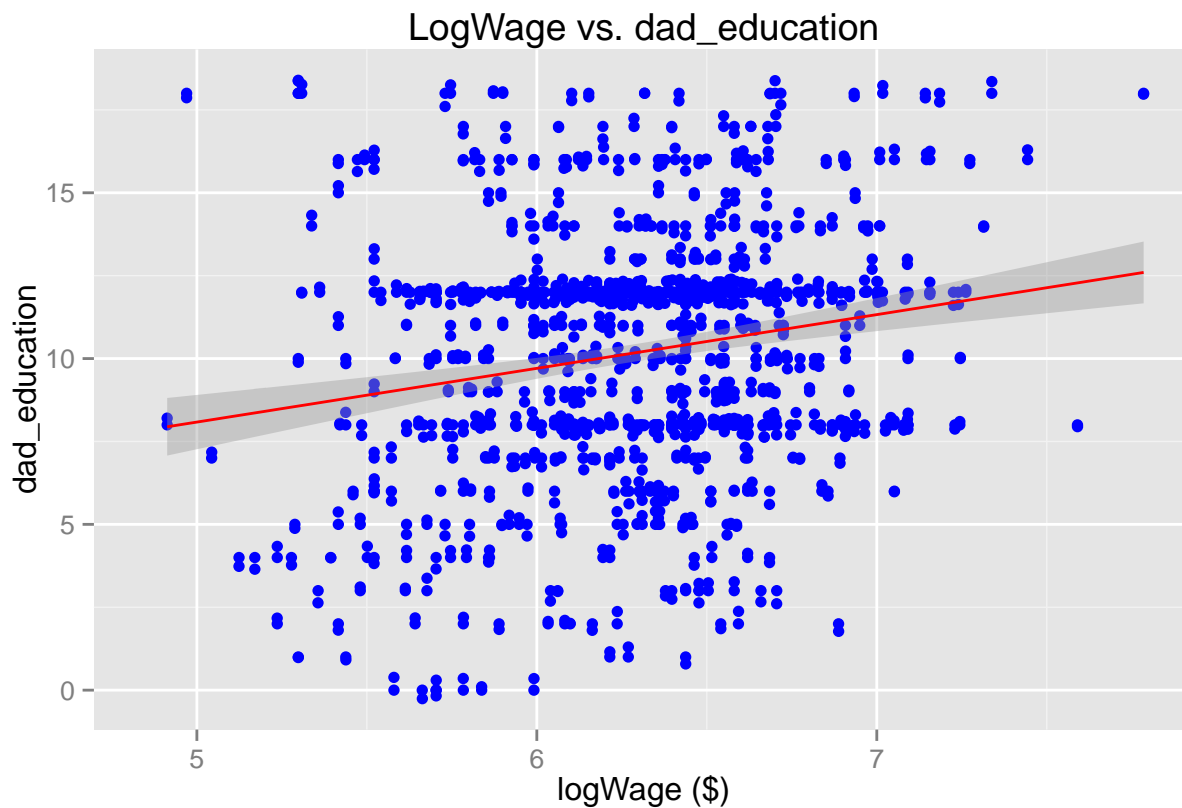


```
# Scatter plot with logWage variable
lwage.dad_education.plot = ggplot(data, aes(x = logWage, y = dad_education)) +
  theme(legend.position = "none") + geom_point(colour = "Blue") + geom_jitter(colour = "Blue") +
  geom_smooth(colour = "red", method = "lm") + labs(title = "LogWage vs. dad_education",
  x = "logWage ($)", y = "dad_education")
plot(lwage.dad_education.plot)
```

```
## Warning: Removed 239 rows containing missing values (stat_smooth).
```

```
## Warning: Removed 239 rows containing missing values (geom_point).
```

```
## Warning: Removed 239 rows containing missing values (geom_point).
```



```
# Run correlations with wage and logWage variables
cor(data$wage, data$dad_education, use = "complete.obs")
```

```
## [1] 0.1901681
```

```
cor(data$logWage, data$dad_education, use = "complete.obs")
```

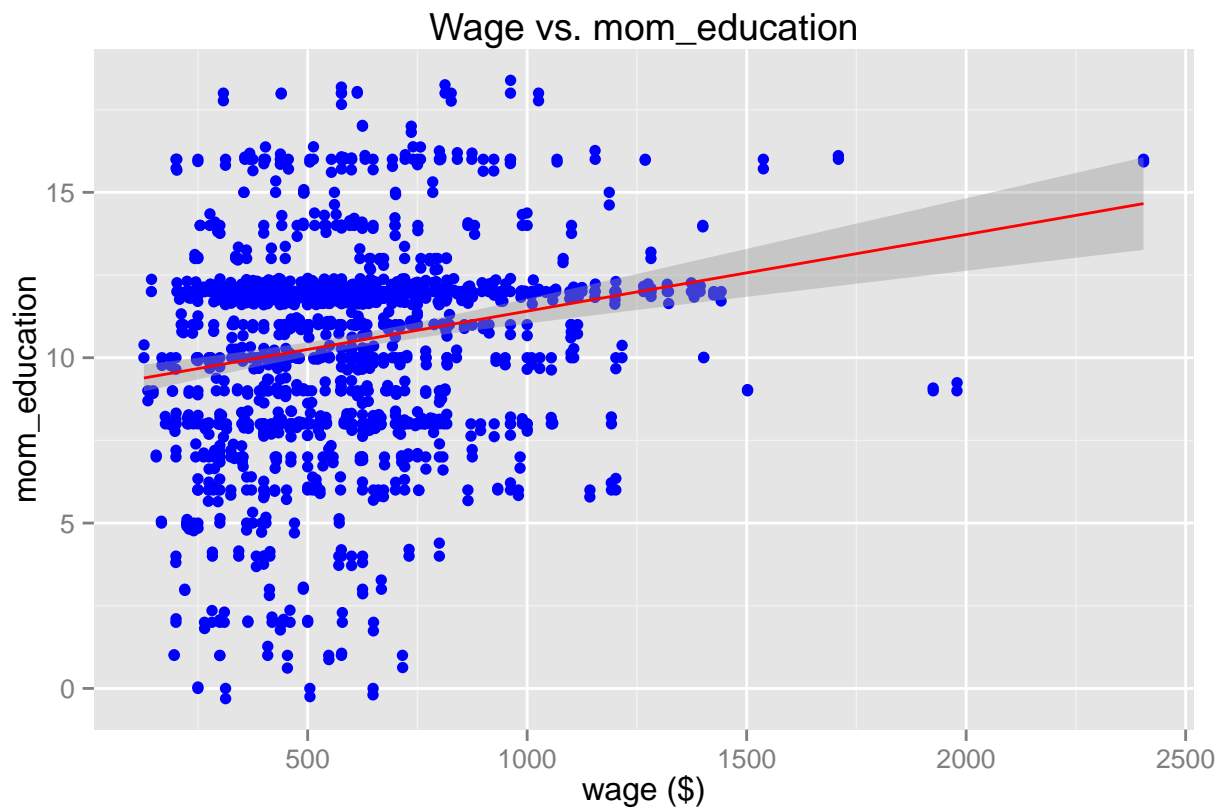
```
## [1] 0.18908
```

```
# Scatter plot with wage variable
wage.mom_education.plot = ggplot(data, aes(x = wage, y = mom_education)) +
  theme(legend.position = "none") + geom_point(colour = "Blue") + geom_jitter(colour = "Blue") +
  geom_smooth(colour = "red", method = "lm") + labs(title = "Wage vs. mom_education",
  x = "wage ($)", y = "mom_education")
plot(wage.mom_education.plot)
```

```
## Warning: Removed 128 rows containing missing values (stat_smooth).
```

```
## Warning: Removed 128 rows containing missing values (geom_point).
```

```
## Warning: Removed 128 rows containing missing values (geom_point).
```

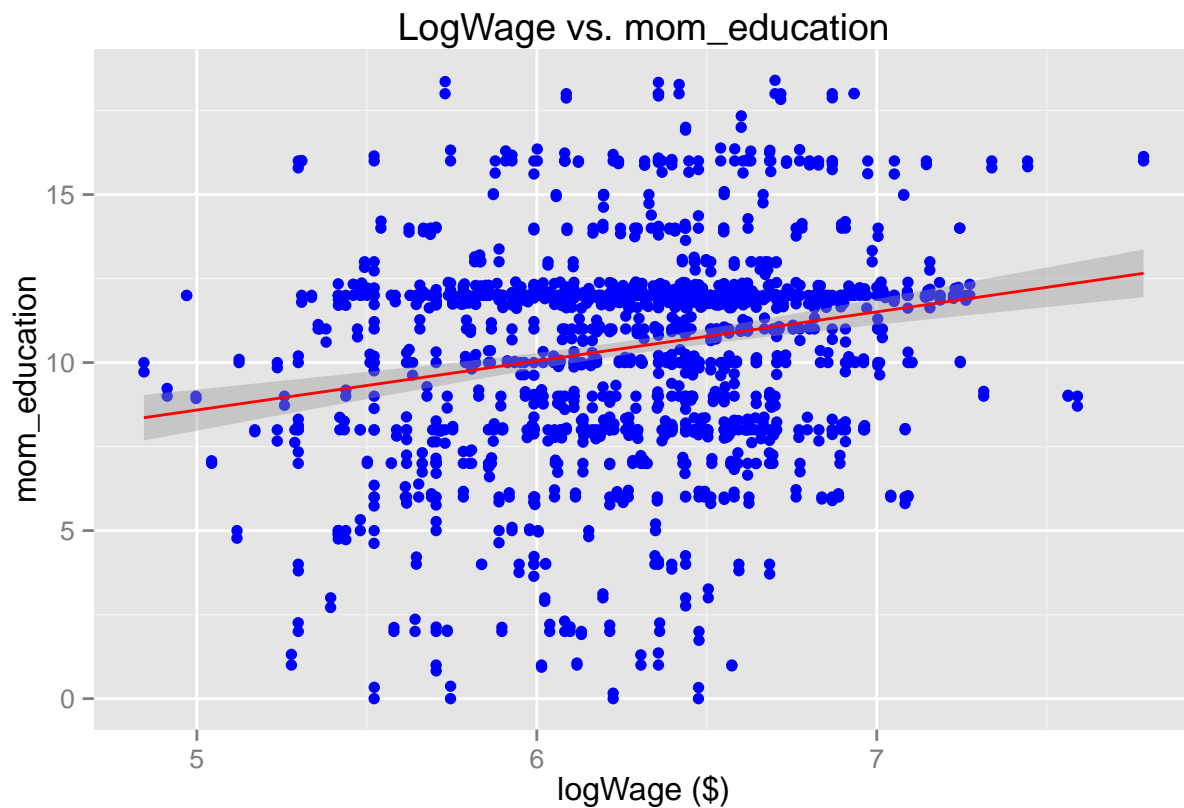


```
# Scatter plot with logWage variable
lwage.mom_education.plot = ggplot(data, aes(x = logWage, y = mom_education)) +
  theme(legend.position = "none") + geom_point(colour = "Blue") + geom_jitter(colour = "Blue") +
  geom_smooth(colour = "red", method = "lm") + labs(title = "LogWage vs. mom_education",
  x = "logWage ($)", y = "mom_education")
plot(lwage.mom_education.plot)
```

```
## Warning: Removed 128 rows containing missing values (stat_smooth).
```

```
## Warning: Removed 128 rows containing missing values (geom_point).
```

```
## Warning: Removed 128 rows containing missing values (geom_point).
```



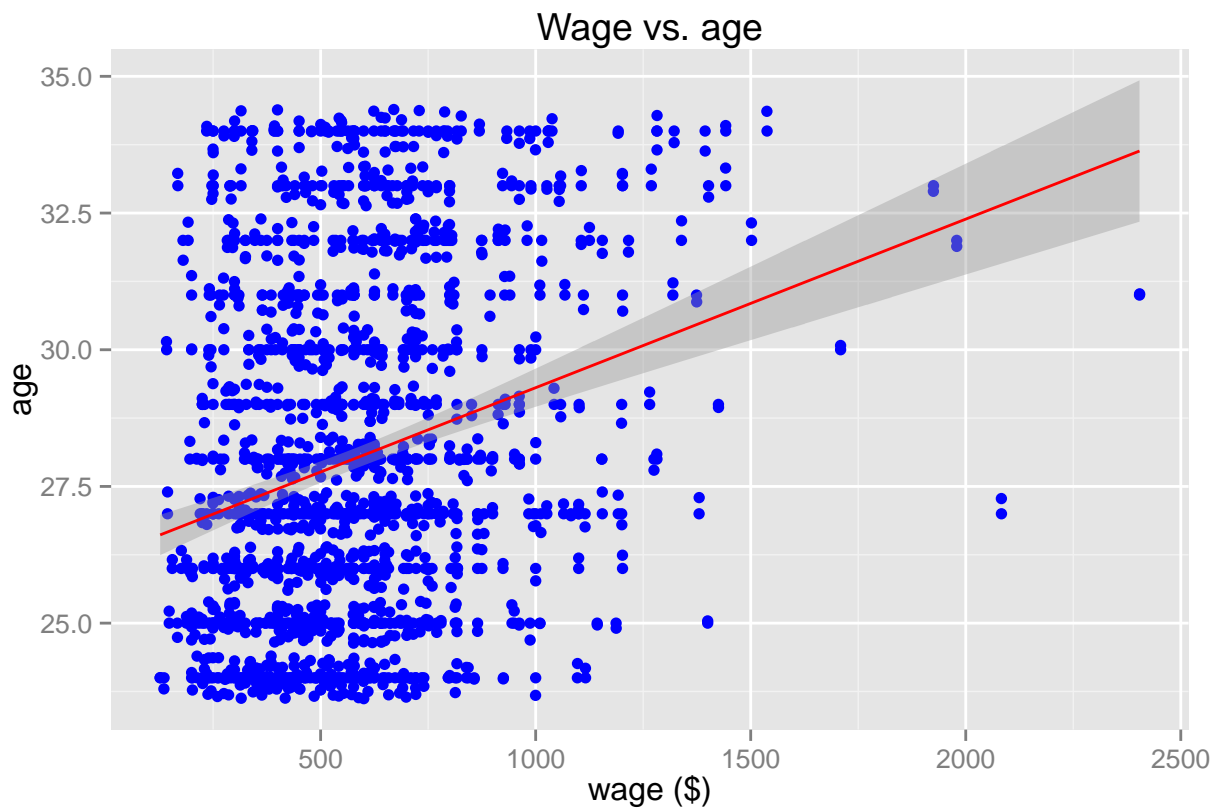
```
# Run correlations with wage and logWage variables
cor(data$wage, data$mom_education, use = "complete.obs")
```

```
## [1] 0.1983845
```

```
cor(data$logWage, data$mom_education, use = "complete.obs")
```

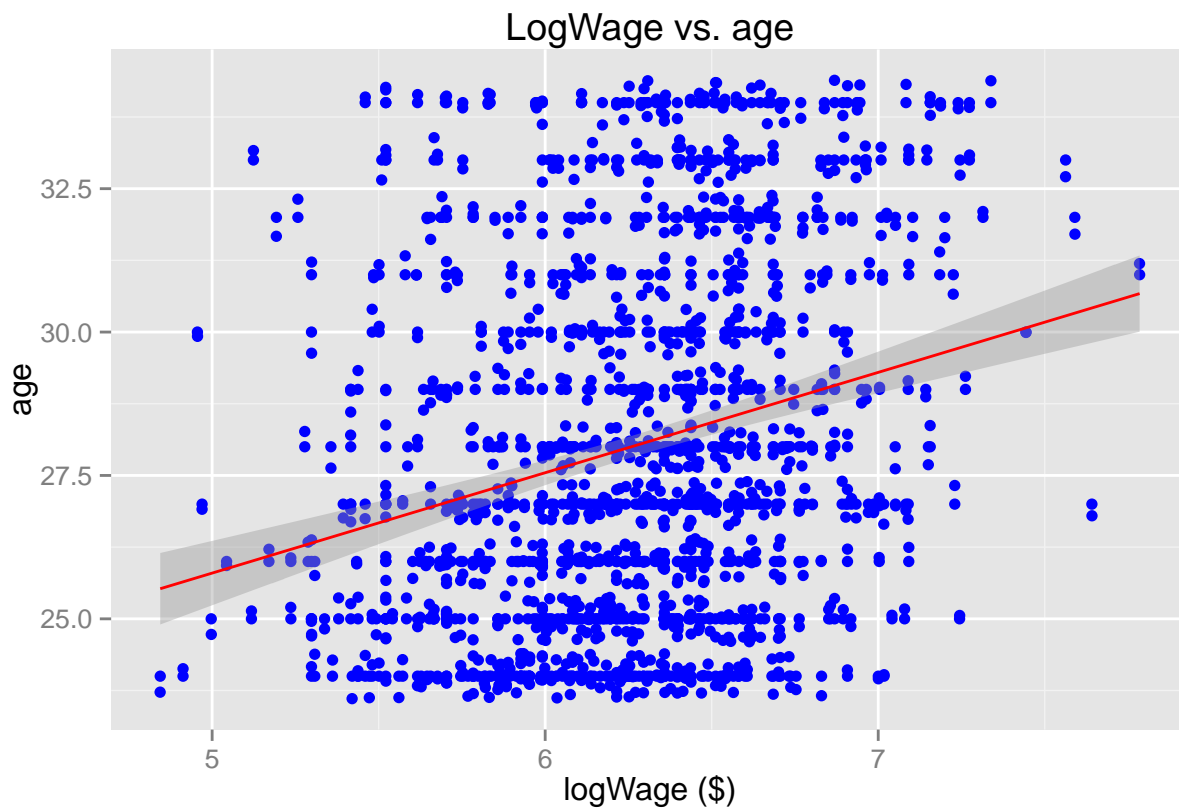
```
## [1] 0.2104614
```

```
# Scatter plot with wage variable
wage.age.plot = ggplot(data, aes(x = wage, y = age)) + theme(legend.position = "none") +
  geom_point(colour = "Blue") + geom_jitter(colour = "Blue") + geom_smooth(colour = "red",
    method = "lm") + labs(title = "Wage vs. age", x = "wage ($)", y = "age")
plot(wage.age.plot)
```



```
# Scatter plot with logWage variable
lwage.age.plot = ggplot(data, aes(x = logWage, y = age)) + theme(legend.position = "none") +
  geom_point(colour = "Blue") + geom_jitter(colour = "Blue") + geom_smooth(colour = "red",
  method = "lm") + labs(title = "LogWage vs. age", x = "logWage ($)",
  y = "age")
plot(lwage.age.plot)
```





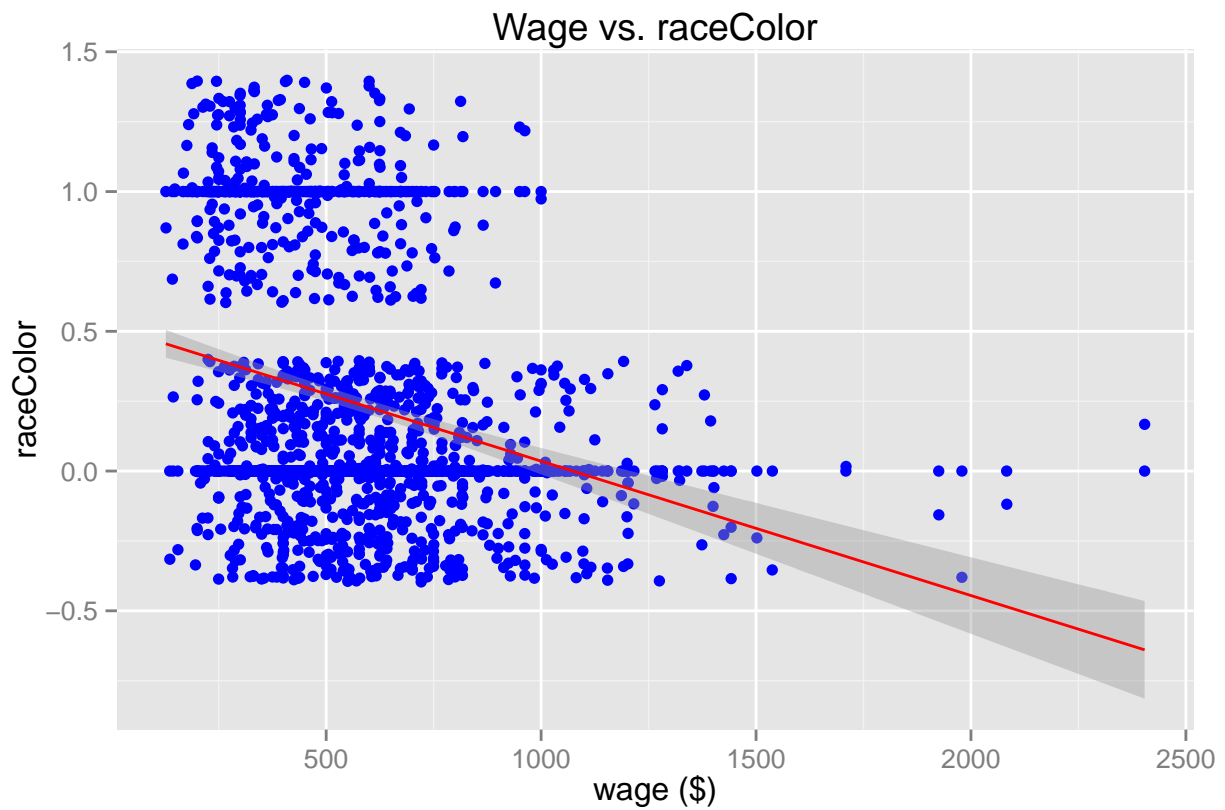
```
# Run correlations with wage and logWage variables
cor(data$wage, data$age)
```

```
## [1] 0.2635783
```

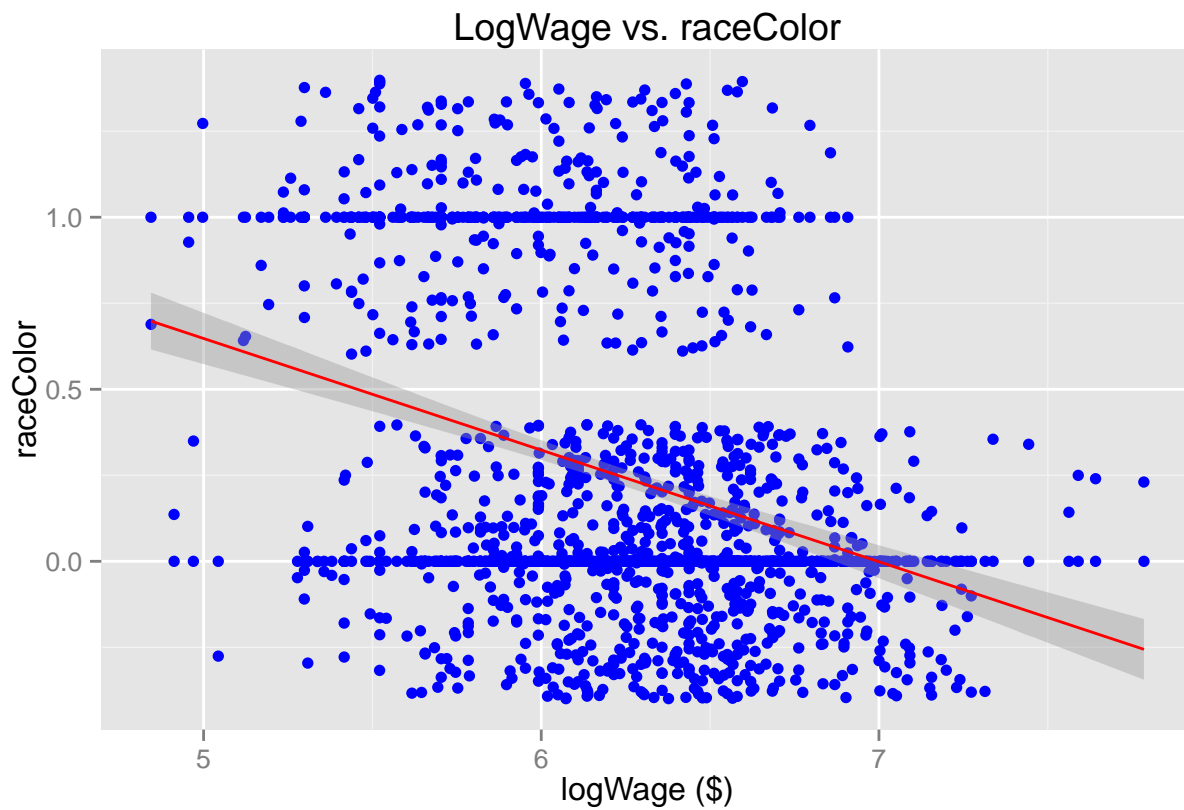
```
cor(data$logWage, data$age)
```

```
## [1] 0.2511202
```

```
# Scatter plot with wage variable
wage.raceColor.plot = ggplot(data, aes(x = wage, y = raceColor)) + theme(legend.position = "none") +
  geom_point(colour = "Blue") + geom_jitter(colour = "Blue") + geom_smooth(colour = "red",
  method = "lm") + labs(title = "Wage vs. raceColor", x = "wage ($)",
  y = "raceColor")
plot(wage.raceColor.plot)
```



```
# Scatter plot with logWage variable
lwage.raceColor.plot = ggplot(data, aes(x = logWage, y = raceColor)) +
  theme(legend.position = "none") + geom_point(colour = "Blue") + geom_jitter(colour = "Blue") +
  geom_smooth(colour = "red", method = "lm") + labs(title = "LogWage vs. raceColor",
  x = "logWage ($)", y = "raceColor")
plot(lwage.raceColor.plot)
```



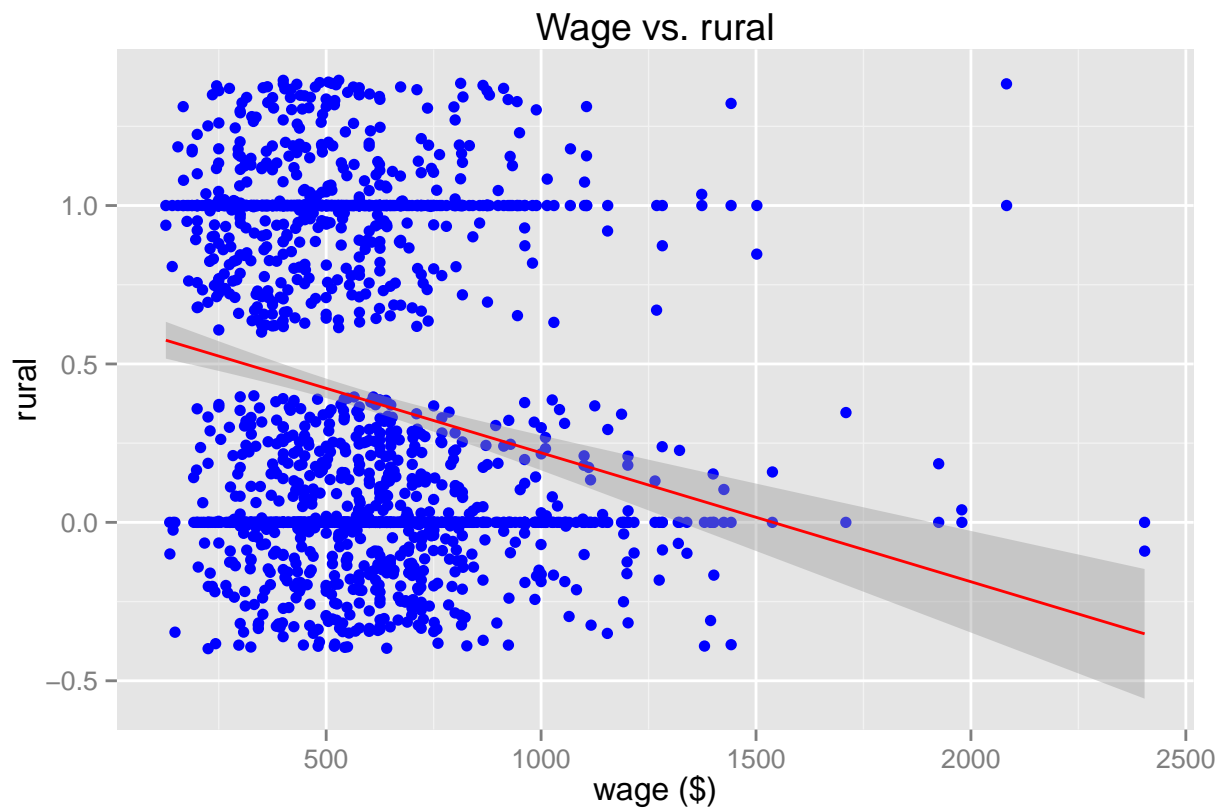
```
# Run correlations with wage and logWage variables
cor(data$wage, data$raceColor)
```

```
## [1] -0.3008475
```

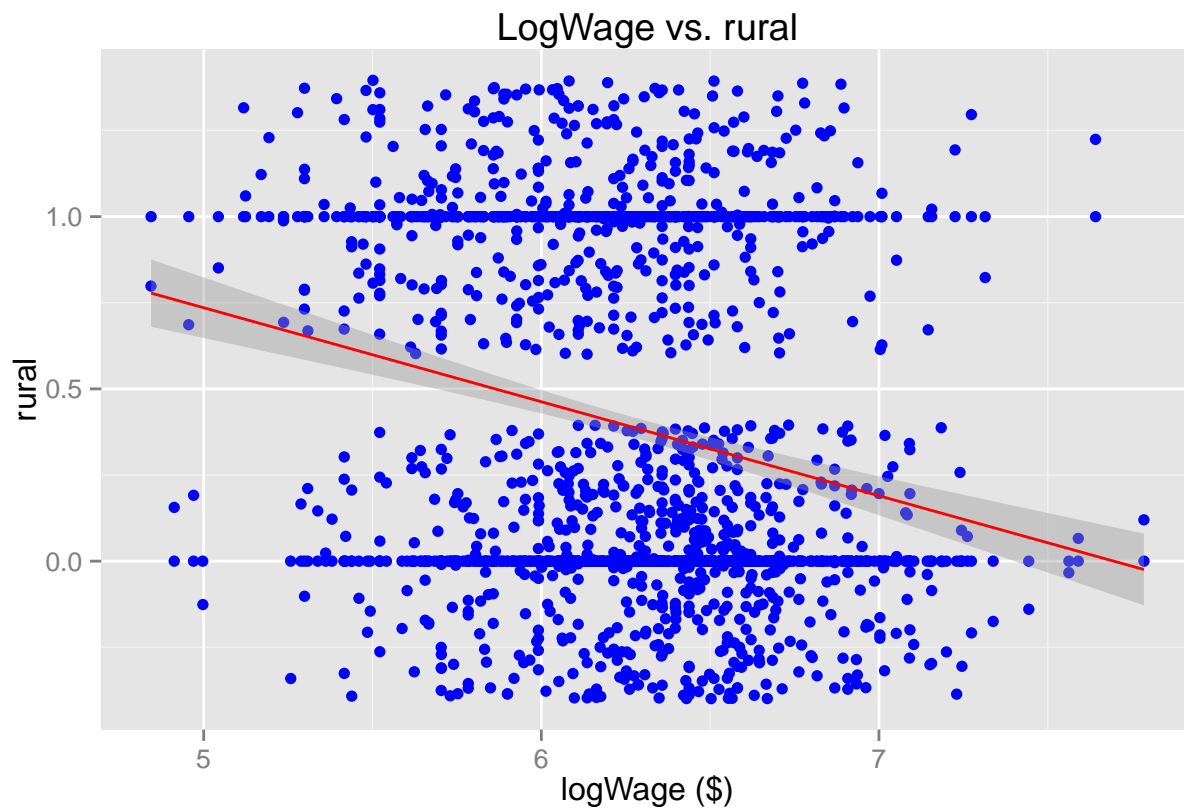
```
cor(data$logWage, data$raceColor)
```

```
## [1] -0.3407361
```

```
# Scatter plot with wage variable
wage.rural.plot = ggplot(data, aes(x = wage, y = rural)) + theme(legend.position = "none") +
  geom_point(colour = "Blue") + geom_jitter(colour = "Blue") + geom_smooth(colour = "red",
    method = "lm") + labs(title = "Wage vs. rural", x = "wage ($)", y = "rural")
plot(wage.rural.plot)
```



```
# Scatter plot with logWage variable
lwage.rural.plot = ggplot(data, aes(x = logWage, y = rural)) + theme(legend.position = "none") +
  geom_point(colour = "Blue") + geom_jitter(colour = "Blue") + geom_smooth(colour = "red",
    method = "lm") + labs(title = "LogWage vs. rural", x = "logWage ($)",
    y = "rural")
plot(lwage.rural.plot)
```



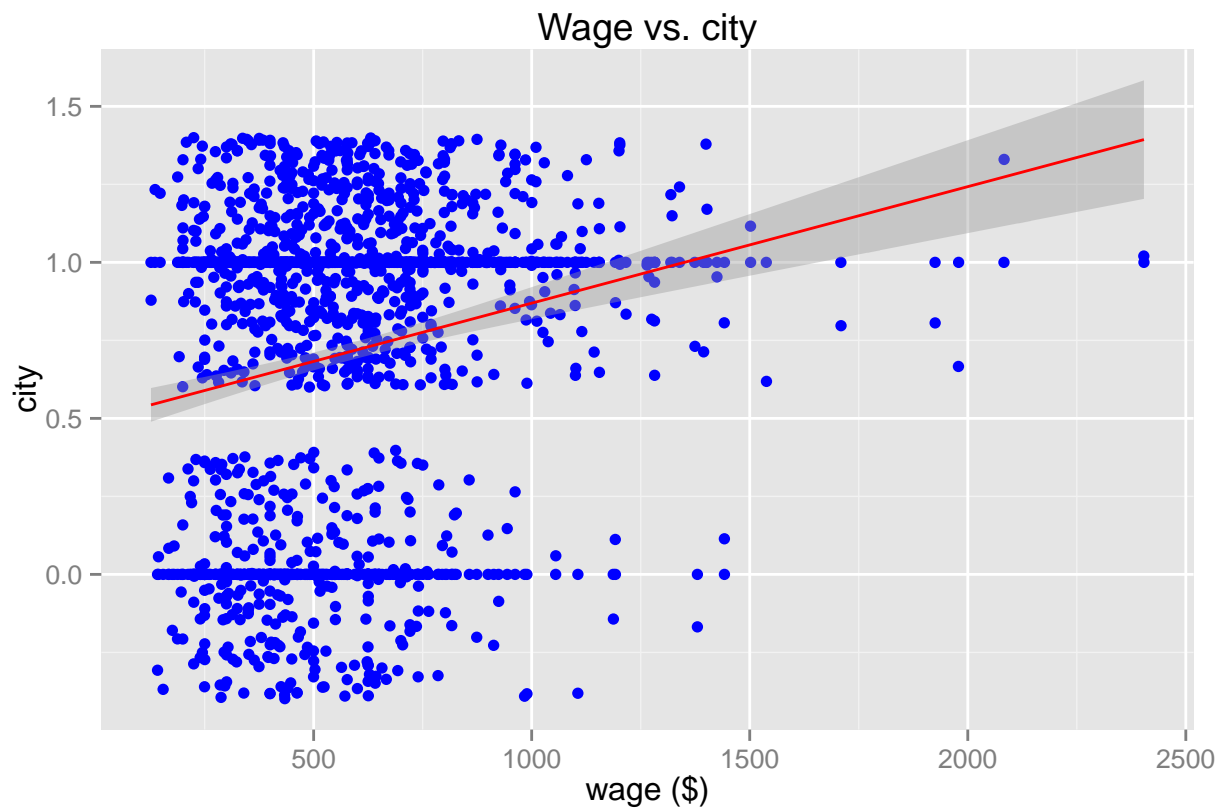
```
# Run correlations with wage and logWage variables
cor(data$wage, data$rural)
```

```
## [1] -0.2222085
```

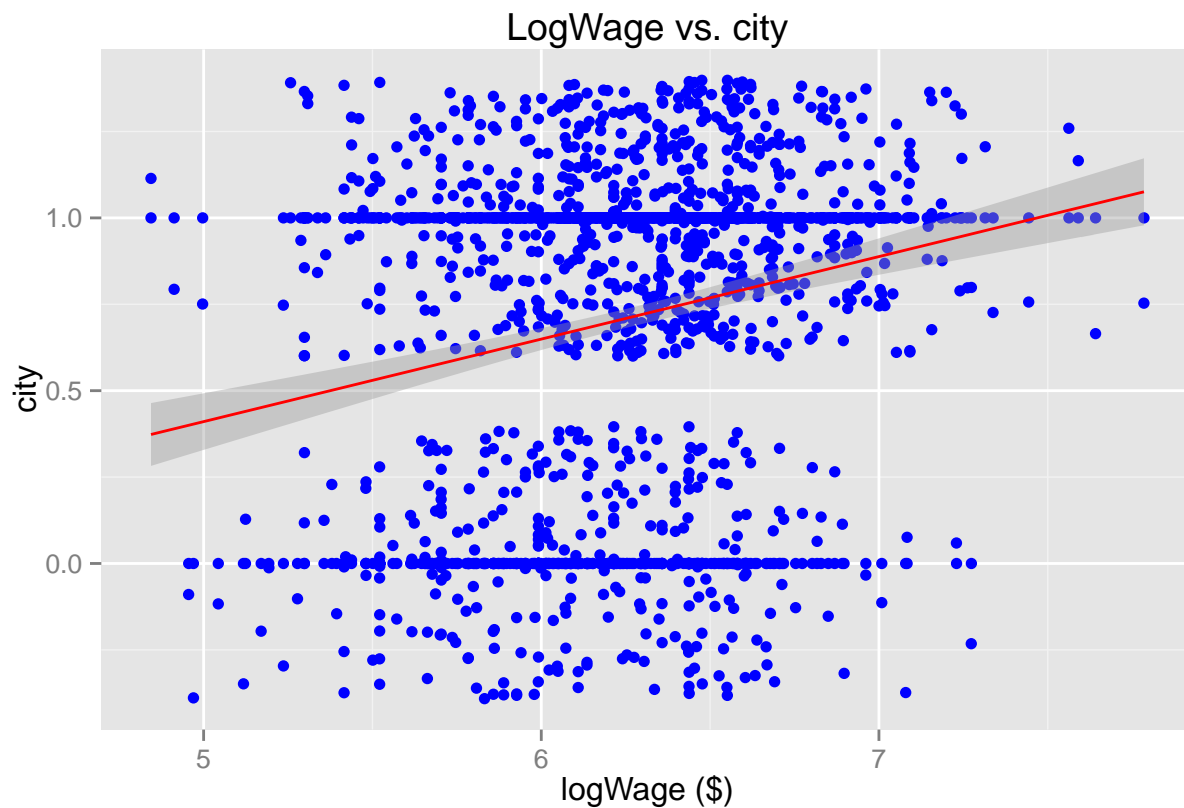
```
cor(data$logWage, data$rural)
```

```
## [1] -0.2501131
```

```
# Scatter plot with wage variable
wage.city.plot = ggplot(data, aes(x = wage, y = city)) + theme(legend.position = "none") +
  geom_point(colour = "Blue") + geom_jitter(colour = "Blue") + geom_smooth(colour = "red",
  method = "lm") + labs(title = "Wage vs. city", x = "wage ($)", y = "city")
plot(wage.city.plot)
```



```
# Scatter plot with logWage variable
lwage.city.plot = ggplot(data, aes(x = logWage, y = city)) + theme(legend.position = "none") +
  geom_point(colour = "Blue") + geom_jitter(colour = "Blue") + geom_smooth(colour = "red",
    method = "lm") + labs(title = "LogWage vs. city", x = "logWage ($)",
    y = "city")
plot(lwage.city.plot)
```



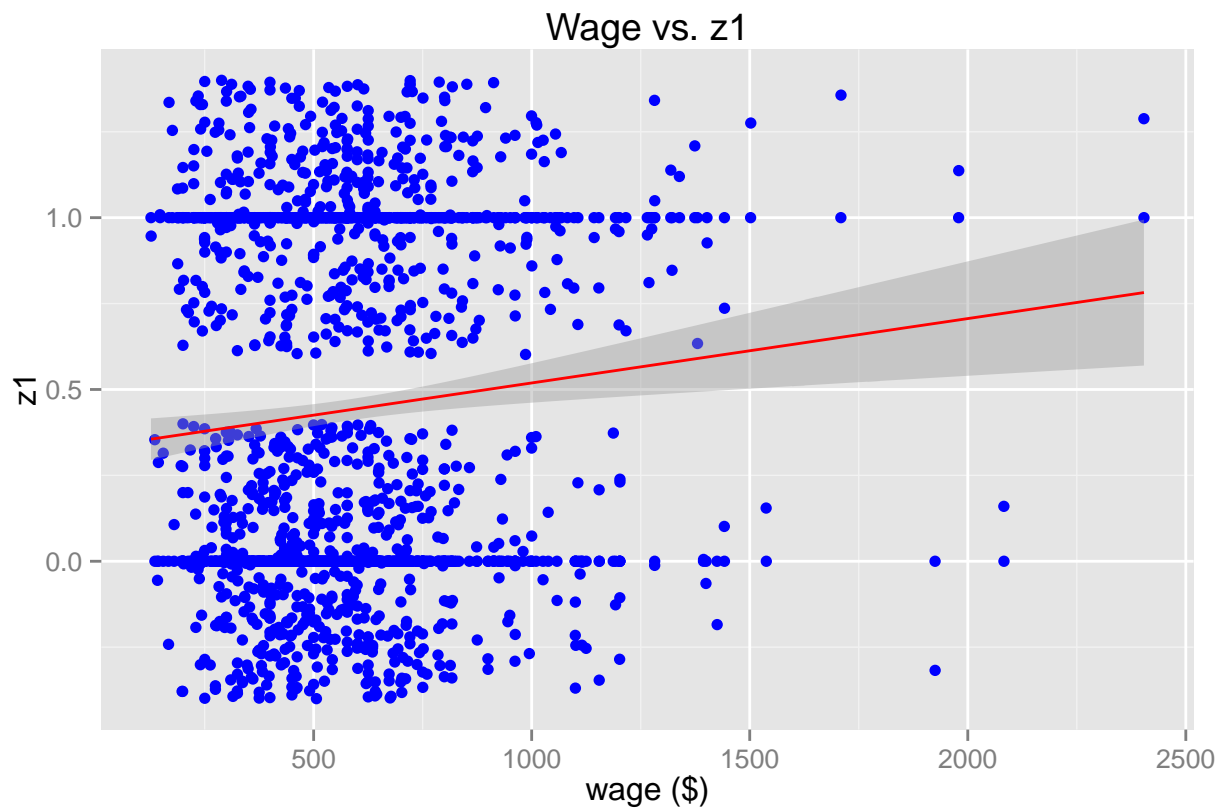
```
# Run correlations with wage and logWage variables
cor(data$wage, data$city)
```

```
## [1] 0.2196804
```

```
cor(data$logWage, data$city)
```

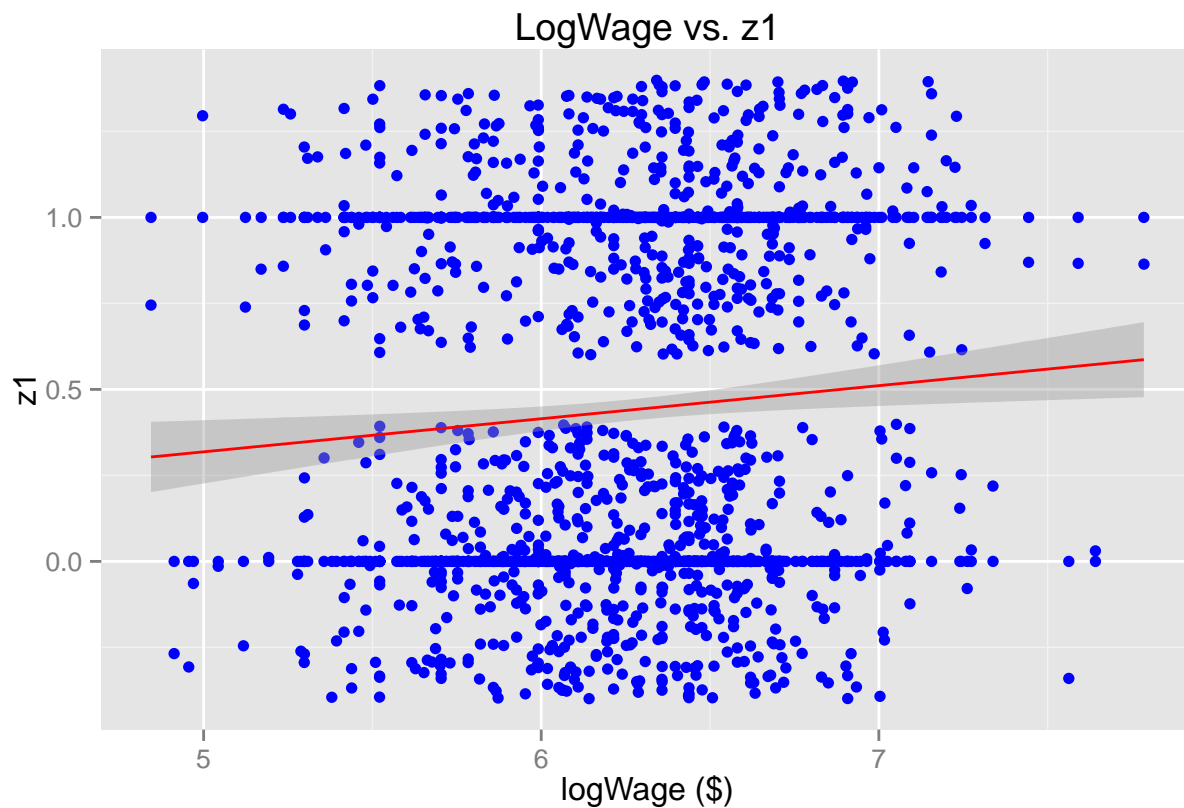
```
## [1] 0.2358269
```

```
# Scatter plot with wage variable
wage.z1.plot = ggplot(data, aes(x = wage, y = z1)) + theme(legend.position = "none") +
  geom_point(colour = "Blue") + geom_jitter(colour = "Blue") + geom_smooth(colour = "red",
    method = "lm") + labs(title = "Wage vs. z1", x = "wage ($)", y = "z1")
plot(wage.z1.plot)
```



```
# Scatter plot with logWage variable
lwage.z1.plot = ggplot(data, aes(x = logWage, y = z1)) + theme(legend.position = "none") +
  geom_point(colour = "Blue") + geom_jitter(colour = "Blue") + geom_smooth(colour = "red",
  method = "lm") + labs(title = "LogWage vs. z1", x = "logWage ($)",
  y = "z1")
plot(lwage.z1.plot)
```





```
# Run correlations with wage and logWage variables
cor(data$wage, data$z1)
```

```
## [1] 0.1005669
```

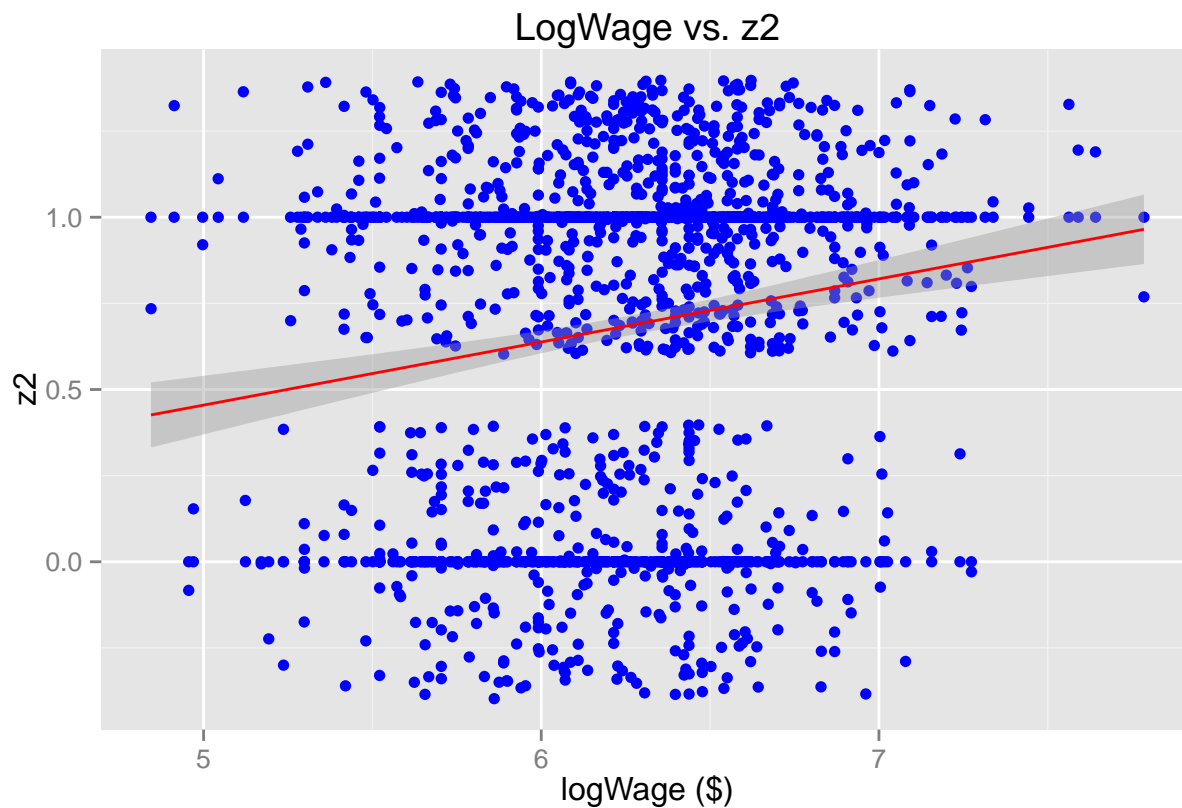
```
cor(data$logWage, data$z1)
```

```
## [1] 0.08668558
```

```
# Scatter plot with wage variable
wage.z2.plot = ggplot(data, aes(x = wage, y = z2)) + theme(legend.position = "none") +
  geom_point(colour = "Blue") + geom_jitter(colour = "Blue") + geom_smooth(colour = "red",
    method = "lm") + labs(title = "Wage vs. z2", x = "wage ($)", y = "z2")
plot(wage.z2.plot)
```



```
# Scatter plot with logWage variable
lwage.z2.plot = ggplot(data, aes(x = logWage, y = z2)) + theme(legend.position = "none") +
  geom_point(colour = "Blue") + geom_jitter(colour = "Blue") + geom_smooth(colour = "red",
    method = "lm") + labs(title = "LogWage vs. z2", x = "logWage ($)",
    y = "z2")
plot(lwage.z2.plot)
```



```
# Run correlations with wage and logWage variables
cor(data$wage, data$z2)
```

```
## [1] 0.1711982
```

```
cor(data$logWage, data$z2)
```

```
## [1] 0.1765267
```

### 4.3 Regress log(wage) on education, experience, age, and raceColor

#### Part 1

Report all the estimated coefficients, their standard errors, t-statistics, F-statistic of the regression, R<sup>2</sup>, adjustedR<sup>2</sup>, and degrees of freedom

The requested information is shown in the summary information below.

```
OLS.logWage.educ.exper.age.race = lm(logWage ~ education + experience +
  age + raceColor, data = data)
summary(OLS.logWage.educ.exper.age.race)
```

```
##
```

```
## Call:
```

```
## lm(formula = logWage ~ education + experience + age + raceColor,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.35396 -0.25550  0.01074  0.24867  1.22932
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.961661   0.113346  43.774  <2e-16 ***
## education    0.079608   0.006376  12.486  <2e-16 ***
## experience    0.035372   0.003988   8.869  <2e-16 ***
## age          NA         NA        NA      NA
## raceColor   -0.260813   0.030453  -8.564  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3917 on 996 degrees of freedom
## Multiple R-squared:  0.236, Adjusted R-squared:  0.2337
## F-statistic: 102.6 on 3 and 996 DF, p-value: < 2.2e-16
```

## Part 2

Degress of freedom = 996. This value is calculated from the following formula  $df = n - k - 1$  where n is the number of observations (n=1000). k is the number of independent variables (k=4). Plugging in these values we get,  $996 = 1000 - 4 - 1$ .

## Part 3

The unexpected results from the regression are that the age variable has coefficient estimates that are NA. This is because age is a linear combination of the education and experience variables as expressed by the formula  $age = education + experience + 6$ . To resolve this issue one of these 3 variables needs to be removed from the regression. Since the intent is to estimate return to education on race and experience, then the age variable can be removed.

```
# Create a new variable that represents the linear combination of age
# with education and experience.
data$age.formula = data$education + data$experience + 6
# Show that this new variable is dataeed the same as the age variable to
# subtracting the two variables.
data$age.difference = data$age - data$age.formula
# Now in the summary of the difference variable, all of the values are
# 0 indicating that the age.formula variable is the same as the age
# variable.
summary(data$age.difference)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##         0         0         0         0         0         0
```

## Part 4 - Interpret the coefficient estimate associated with education

The estimate for the education coefficient is 0.079608. This means that for every unit change in education, there is an 8.00% change in logWage. This value is significant at the 0.1% significance level. This is a small practical effect.

## Part 5 - Interpret the coefficient estimate associated with experience

The estimate for the experience coefficient is 0.035372. This means that for every unit change in experience, there is a 3.53% change in logWage. This value is significant at the 0.1% significance level. This is a small practical effect.

### Question 4.4

#### Part 1

See graph below of the estimated effect of experience on wage.

$$\frac{\delta \log Wage}{\delta experience} = 0.0924 - 2 * (0.00288) * experience$$

#### Part 2

$$d\log Wage_{10} = 0.0924 - 2 * (0.00288) * 10 = 0.0348$$

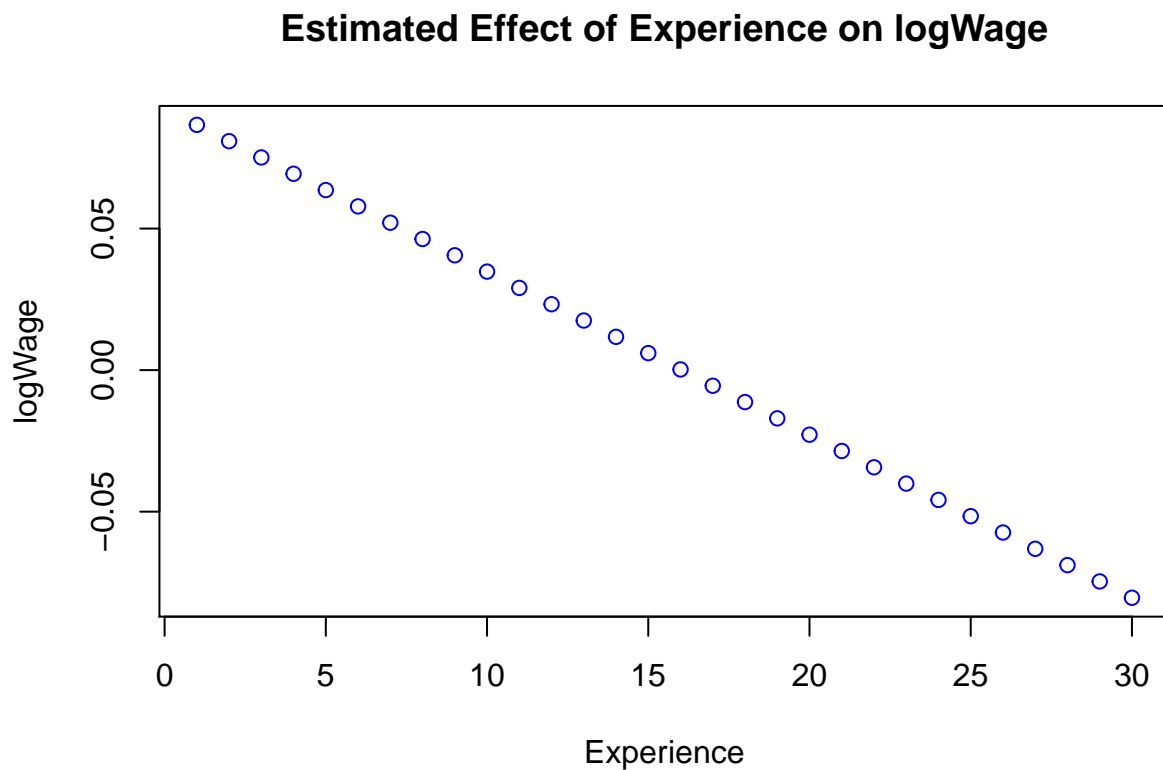
The estimated effect of experience on wage when experience is 10 years is 3.48%.

```
# Create the model
OLS.logWage.educ.exper.exper2.race = lm(logWage ~ education + experience +
  experienceSquare + raceColor, data = data)
# Print the summary of the model
summary(OLS.logWage.educ.exper.exper2.race)

##
## Call:
## lm(formula = logWage ~ education + experience + experienceSquare +
##     raceColor, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.38464 -0.25558  0.01909  0.25782  1.24410
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.7355175   0.1197719   39.538  < 2e-16 ***
## education     0.0794641   0.0062917   12.630  < 2e-16 ***
## experience    0.0924930   0.0115147    8.033 2.68e-15 ***
## experienceSquare -0.0028779  0.0005452   -5.279 1.60e-07 ***
## raceColor    -0.2627226   0.0300528   -8.742  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.3865 on 995 degrees of freedom
## Multiple R-squared:  0.2569, Adjusted R-squared:  0.2539
## F-statistic: 85.98 on 4 and 995 DF,  p-value: < 2.2e-16
```

```
# Create a variable dlogWage the represents the line created by the
# change in logWage with respect to a change in experience
dlogWage = 0
for (experience in 1:30) {
  dlogWage[experience] = 0.0924 - 2 * (0.00288) * experience
}
# Graph the line
plot(dlogWage, lty = "dashed", main = "Estimated Effect of Experience on logWage",
     col = "blue", ylab = "logWage", xlab = "Experience")
```



```
# Calculate the value of the effect of experience on wage when
# experience is 10 years.
dlogWage10 = 0.0924 - 2 * (0.00288) * 10
dlogWage10
```

```
## [1] 0.0348
```

## Question 4.5

### Part 1

The number of observations used in this regression 723 (out of 1,000). The participants with missing mom\_education or dad\_education (mdMiss) values compare to participants that have both a mom\_education and a dad\_education (mdHave) value as follows.

- **wage** - The mdMiss participants have lower wages than the mdHave participants. The minimum value for wage for mdMiss is \$127 vs. \$136 for mdHave. The median and mean for values for mdMiss are lower at \$481 vs. \$570 and \$531 vs. \$597, respectively. The maximum values are also lower at \$2,083 vs. \$2,404.
- **education** - The mdMiss participants have less education than the mdHave participants. This could indicate that they stopped education sooner and went to work earlier than the mdHave participants. The mean value for mdMiss is 12.09 vs. 13.65 for mdHave, a difference of 1.56 years or a 11.4% decrease.
- **experience** - The mdMiss participants have more experience than the mdHave participants. This is further evidence that they stopped education sooner and went to work earlier than the mdHave participants. The mean value for mdMiss is 10.47 vs. 8.145 for mdHave, a difference of 2.32 years or a 28.5% increase.
- **raceColor** - The mdMiss participants have a much higher percentage of people with the raceColor variable set to 1 than mdHave. 44.77% (124 people) with 1's for mdMiss vs. 15.77% (114 people) with 1's for mdHave.

### Part 2

We do not think we can just throw away the participants with the missing values. They are important to the analysis since they represent a disproportional amount of people with lower wages, less education, more experience and more raceColor variables equal to 1 than participants without missing values.

### Part 3

Blindly replace all of the missing values with the average of the observed values of the corresponding variable. See the re-run of the original regression using all of the observations below.

### Part 4

Regress the variable(s) with missing values on education, experience, and raceColor, and use this regression(s) to predict (i.e. "impute") the missing values. See the re-run of the original regression using all of the observations below.

### Part 5

We would not use any of the previous 3 models that included the mom\_education and dad\_education variables. The mom\_education and dad\_education are not significant to even the 10% significance level in the models. We would take them out and use a 6 variable model without them. The Adjusted R-squared goes up slightly when we do this from 0.2925 to 0.2935 even though we are using fewer variables.

```
# Part 1 Create the model
OLS.logWage.8var = lm(logWage ~ education + experience + experienceSquare +
  raceColor + dad_education + mom_education + rural + city, data = data)
# Print the model
summary(OLS.logWage.8var)
```

```
##
## Call:
## lm(formula = logWage ~ education + experience + experienceSquare +
##     raceColor + dad_education + mom_education + rural + city,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.2961 -0.2240  0.0160  0.2454  1.0404
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.6422296   0.1408825   32.951 < 2e-16 ***
## education      0.0681701   0.0077409    8.806 < 2e-16 ***
## experience     0.0973419   0.0133133    7.312 7.1e-13 ***
## experienceSquare -0.0029568  0.0006678   -4.428 1.1e-05 ***
## raceColor     -0.2130226   0.0425014   -5.012 6.8e-07 ***
## dad_education  -0.0011474   0.0050988   -0.225 0.82202
## mom_education   0.0113176   0.0061886    1.829 0.06785 .
## rural          -0.0919377   0.0314151   -2.927 0.00354 **
## city           0.1782137   0.0323826    5.503 5.2e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3786 on 714 degrees of freedom
## (277 observations deleted due to missingness)
## Multiple R-squared:  0.2746, Adjusted R-squared:  0.2665
## F-statistic: 33.79 on 8 and 714 DF, p-value: < 2.2e-16
```

```
# Create 2 temporary datasets. mdMiss contains all of the rows with
# either mom_education or dad_education equal to NA. mdHave contains
# all of the rows with both mom_education and dad_education equal to a
# non-NA value.
mdMiss = data[(is.na(data$mom_education) | is.na(data$dad_education)),
]
mdHave = data[(!is.na(data$mom_education) & !is.na(data$dad_education)),
]
# Use summary to confirm that mdMiss has the correct number of NA's for
# mom_education and dad_education
summary(mdMiss)
```

```
##           X           wage           education           experience
## Min.      : 15      Min.    : 127      Min.      : 2.00      Min.      : 0.00
## 1st Qu.: 842      1st Qu.: 358      1st Qu.:11.00      1st Qu.: 7.00
## Median :1688      Median : 481      Median :12.00      Median :10.00
## Mean     :1643      Mean     : 531      Mean      :12.09      Mean      :10.47
## 3rd Qu.:2495      3rd Qu.: 640      3rd Qu.:13.00      3rd Qu.:14.00
## Max.     :3009      Max.      :2083      Max.      :18.00      Max.      :23.00
##
##           age           raceColor           dad_education           mom_education
## Min.      :24.00      Min.      :0.0000      Min.      : 2.000      Min.      : 0.000
## 1st Qu.:26.00      1st Qu.:0.0000      1st Qu.: 6.000      1st Qu.: 7.000
## Median :28.00      Median :0.0000      Median : 9.500      Median : 9.000
## Mean      :28.56      Mean      :0.4477      Mean      : 9.184      Mean      : 8.987
```



```
## 3rd Qu.:31.00 3rd Qu.:1.0000 3rd Qu.:12.000 3rd Qu.:12.000
## Max. :34.00 Max. :1.0000 Max. :16.000 Max. :18.000
## NA's :239 NA's :128
## rural city z1 z2
## Min. :0.000 Min. :0.0000 Min. :0.0000 Min. :0.0000
## 1st Qu.:0.000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000
## Median :1.000 Median :1.0000 Median :0.0000 Median :1.0000
## Mean :0.509 Mean :0.6643 Mean :0.4188 Mean :0.6895
## 3rd Qu.:1.000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000
## Max. :1.000 Max. :1.0000 Max. :1.0000 Max. :1.0000
##
## IQscore logWage logWageOLD experienceSquare
## Min. : 50 Min. :4.844 Min. :4.844 Min. : 0.0
## 1st Qu.: 85 1st Qu.:5.881 1st Qu.:5.881 1st Qu.: 49.0
## Median : 98 Median :6.176 Median :6.176 Median :100.0
## Mean : 96 Mean :6.174 Mean :6.174 Mean :128.1
## 3rd Qu.:107 3rd Qu.:6.461 3rd Qu.:6.461 3rd Qu.:196.0
## Max. :135 Max. :7.642 Max. :7.642 Max. :529.0
## NA's :124
## age.formula age.difference
## Min. :24.00 Min. :0
## 1st Qu.:26.00 1st Qu.:0
## Median :28.00 Median :0
## Mean :28.56 Mean :0
## 3rd Qu.:31.00 3rd Qu.:0
## Max. :34.00 Max. :0
##
```

`summary(mdHave)`

```
## X wage education experience
## Min. : 5.0 Min. :136.0 Min. : 3.00 Min. : 0.000
## 1st Qu.: 680.5 1st Qu.: 409.0 1st Qu.:12.00 1st Qu.: 5.000
## Median :1314.0 Median : 570.0 Median :13.00 Median : 8.000
## Mean :1399.0 Mean : 597.1 Mean :13.65 Mean : 8.145
## 3rd Qu.:2125.0 3rd Qu.: 721.0 3rd Qu.:16.00 3rd Qu.:10.000
## Max. :2998.0 Max. :2404.0 Max. :18.00 Max. :21.000
##
## age raceColor dad_education mom_education
## Min. :24.0 Min. :0.0000 Min. : 0.00 Min. : 0.00
## 1st Qu.:25.0 1st Qu.:0.0000 1st Qu.: 8.00 1st Qu.: 9.00
## Median :27.0 Median :0.0000 Median :11.00 Median :12.00
## Mean :27.8 Mean :0.1577 Mean :10.23 Mean :10.75
## 3rd Qu.:30.0 3rd Qu.:0.0000 3rd Qu.:12.00 3rd Qu.:12.00
## Max. :34.0 Max. :1.0000 Max. :18.00 Max. :18.00
##
## rural city z1 z2
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.0000
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000
## Median :0.0000 Median :1.0000 Median :0.0000 Median :1.0000
## Mean :0.3458 Mean :0.7303 Mean :0.4481 Mean :0.6846
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.0000
##
```

```
##      IQscore      logWage      logWageOLD      experienceSquare
## Min.   : 60.0    Min.   :4.913    Min.   :4.913    Min.   : 0.00
## 1st Qu.: 95.0    1st Qu.:6.014    1st Qu.:6.014    1st Qu.: 25.00
## Median :105.0    Median :6.346    Median :6.346    Median : 64.00
## Mean   :104.1    Mean   :6.297    Mean   :6.297    Mean   : 82.37
## 3rd Qu.:114.0    3rd Qu.:6.581    3rd Qu.:6.581    3rd Qu.:100.00
## Max.   :144.0    Max.   :7.785    Max.   :7.785    Max.   :441.00
## NA's   :192
## age.formula age.difference
## Min.   :24.0    Min.   :0
## 1st Qu.:25.0    1st Qu.:0
## Median :27.0    Median :0
## Mean   :27.8    Mean   :0
## 3rd Qu.:30.0    3rd Qu.:0
## Max.   :34.0    Max.   :0
##
```

```
# Use str to confirm that the datasets have the appropriate number of
# observations
str(mdMiss)
```

```
## 'data.frame': 277 obs. of 18 variables:
## $ X : int 191 2059 1927 1481 1484 2548 574 2061 2700 2689 ...
## $ wage : int 951 288 454 565 670 624 400 673 575 340 ...
## $ education : int 12 8 10 12 13 9 12 12 12 9 ...
## $ experience : int 10 11 11 10 8 9 8 14 8 19 ...
## $ age : int 28 25 27 28 27 24 26 32 26 34 ...
## $ raceColor : int 0 1 1 1 1 1 0 0 1 1 ...
## $ dad_education : int NA NA NA NA NA NA NA NA NA NA ...
## $ mom_education : int 12 7 1 NA NA 7 12 6 NA NA ...
## $ rural : int 0 1 1 1 1 1 0 1 0 1 ...
## $ city : int 1 0 0 1 1 0 0 0 1 0 ...
## $ z1 : int 1 0 0 0 0 1 0 0 1 1 ...
## $ z2 : int 1 1 1 1 1 0 1 1 0 0 ...
## $ IQscore : int 122 NA NA NA 99 NA 117 93 NA NA ...
## $ logWage : num 6.86 5.66 6.12 6.34 6.51 ...
## $ logWageOLD : num 6.86 5.66 6.12 6.34 6.51 ...
## $ experienceSquare: int 100 121 121 100 64 81 64 196 64 361 ...
## $ age.formula : num 28 25 27 28 27 24 26 32 26 34 ...
## $ age.difference : num 0 0 0 0 0 0 0 0 0 0 ...
```

```
str(mdHave)
```

```
## 'data.frame': 723 obs. of 18 variables:
## $ X : int 2072 945 1920 2571 437 1265 603 2936 1123 2080 ...
## $ wage : int 509 647 225 479 615 641 740 619 583 813 ...
## $ education : int 12 18 10 13 16 12 13 17 12 12 ...
## $ experience : int 6 5 11 15 7 16 10 6 10 6 ...
## $ age : int 24 29 27 34 29 34 29 29 28 24 ...
## $ raceColor : int 0 0 1 0 0 0 0 0 0 0 ...
## $ dad_education : int 12 12 5 7 12 4 16 8 14 9 ...
## $ mom_education : int 9 12 5 12 12 8 16 13 8 9 ...
## $ rural : int 1 0 1 1 0 0 0 0 0 1 ...
```

```
## $ city      : int  1 1 0 1 1 0 0 1 1 1 ...
## $ z1        : int  0 0 0 0 1 0 0 0 1 1 ...
## $ z2        : int  0 1 1 1 1 1 0 0 1 1 ...
## $ IQscore    : int  127 110 NA NA 113 92 108 138 94 NA ...
## $ logWage    : num  6.23 6.47 5.42 6.17 6.42 ...
## $ logWageOLD : num  6.23 6.47 5.42 6.17 6.42 ...
## $ experienceSquare: int  36 25 121 225 49 256 100 36 100 36 ...
## $ age.formula : num  24 29 27 34 29 34 29 29 28 24 ...
## $ age.difference : num  0 0 0 0 0 0 0 0 0 0 ...
```

```
# Compare the summaries of the mdMiss and mdHave datasets to see the
# differences in wage, eduction, experience and raceColor
summary(mdMiss$wage)
```

```
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##      127      358      481      531      640     2083
```

```
summary(mdHave$wage)
```

```
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##     136.0   409.0   570.0   597.1   721.0  2404.0
```

```
summary(mdMiss$education)
```

```
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##       2.00   11.00   12.00   12.09   13.00   18.00
```

```
summary(mdHave$education)
```

```
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##       3.00   12.00   13.00   13.65   16.00   18.00
```

```
summary(mdMiss$experience)
```

```
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##       0.00    7.00   10.00   10.47   14.00   23.00
```

```
summary(mdHave$experience)
```

```
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##      0.000    5.000    8.000    8.145   10.000   21.000
```

```
summary(mdMiss$raceColor)
```

```
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##      0.0000  0.0000  0.0000  0.4477  1.0000  1.0000
```

```
summary(mdHave$raceColor)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.0000  0.0000  0.0000  0.1577  0.0000  1.0000
```

```
# Part 3 Copy the dataset to a new variable
data.avgForNA = data
# Set all of the values with dad_education = NA to the mean of
# dad_education
data.avgForNA$dad_education[is.na(data.avgForNA$dad_education)] = mean(data.avgForNA$dad_education,
  na.rm = TRUE)
# Set all of the values with mom_education = NA to the mean of
# mom_education
data.avgForNA$mom_education[is.na(data.avgForNA$mom_education)] = mean(data.avgForNA$mom_education,
  na.rm = TRUE)
# Rerun the regression
OLS.logWage.8var.avgNA = lm(logWage ~ education + experience + experienceSquare +
  raceColor + dad_education + mom_education + rural + city, data = data.avgForNA)
```

```
# Part 4 Copy the dataset to a new variable
data.regressForNA = data
# Regress dad_education on the education, experience and raceColor
# variables
m1 = lm(dad_education ~ education + experience + raceColor, data = data)
# Regress mom_education on the education, experience and raceColor
# variables
m2 = lm(mom_education ~ education + experience + raceColor, data = data)

# Set all of the values with dad_education = NA to the value output
# from using the regression coefficients from m1 above.
data.regressForNA$dad_education[is.na(data.regressForNA$dad_education)] = m1$coefficients[1] +
  m1$coefficients[2] * data.regressForNA$education + m1$coefficients[3] *
  data.regressForNA$experience + m1$coefficients[4] * data.regressForNA$raceColor
```

```
## Warning in data.regressForNA$dad_education[is.na(data.regressForNA
## $dad_education)] = m1$coefficients[1] + : number of items to replace is not
## a multiple of replacement length
```

```
# Set all of the values with mom_education = NA to the value output
# from using the regression coefficients from m2 above.
data.regressForNA$mom_education[is.na(data.regressForNA$mom_education)] = m2$coefficients[1] +
  m2$coefficients[2] * data.regressForNA$education + m2$coefficients[3] *
  data.regressForNA$experience + m2$coefficients[4] * data.regressForNA$raceColor
```

```
## Warning in data.regressForNA$mom_education[is.na(data.regressForNA
## $mom_education)] = m2$coefficients[1] + : number of items to replace is not
## a multiple of replacement length
```

```
# Rerun the regression
OLS.logWage.8var.regressNA = lm(logWage ~ education + experience + experienceSquare +
  raceColor + dad_education + mom_education + rural + city, data = data.regressForNA)
```

*# Part 5 Print the summaries of the 2 new models*

```
summary(OLS.logWage.8var.avgNA)
```

```
##
## Call:
## lm(formula = logWage ~ education + experience + experienceSquare +
##      raceColor + dad_education + mom_education + rural + city,
##      data = data.avgForNA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.30741 -0.23286  0.01943  0.24786  1.28807
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.729e+00  1.226e-01  38.584 < 2e-16 ***
## education      7.097e-02  6.499e-03  10.920 < 2e-16 ***
## experience     8.958e-02  1.124e-02   7.970 4.36e-15 ***
## experienceSquare -2.678e-03  5.318e-04  -5.036 5.65e-07 ***
## raceColor     -2.313e-01  3.099e-02  -7.464 1.84e-13 ***
## dad_education  -3.513e-05  4.416e-03  -0.008 0.993656
## mom_education   3.485e-03  5.009e-03   0.696 0.486742
## rural          -9.529e-02  2.638e-02  -3.612 0.000319 ***
## city           1.671e-01  2.703e-02   6.183 9.21e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3764 on 991 degrees of freedom
## Multiple R-squared:  0.2981, Adjusted R-squared:  0.2925
## F-statistic: 52.62 on 8 and 991 DF, p-value: < 2.2e-16
```

```
summary(OLS.logWage.8var.regressNA)
```

```
##
## Call:
## lm(formula = logWage ~ education + experience + experienceSquare +
##      raceColor + dad_education + mom_education + rural + city,
##      data = data.regressForNA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.30770 -0.23222  0.02095  0.24785  1.29770
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.7278751  0.1228090  38.498 < 2e-16 ***
## education      0.0710341  0.0064659  10.986 < 2e-16 ***
## experience     0.0896724  0.0112433   7.976 4.16e-15 ***
## experienceSquare -0.0026820  0.0005318  -5.043 5.45e-07 ***
## raceColor     -0.2313406  0.0311112  -7.436 2.24e-13 ***
## dad_education  -0.0003385  0.0041318  -0.082 0.934718
## mom_education   0.0037753  0.0047649   0.792 0.428365
```

```
## rural          -0.0952834  0.0263780  -3.612 0.000319 ***
## city           0.1673210  0.0270228   6.192 8.70e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3764 on 991 degrees of freedom
## Multiple R-squared:  0.2982, Adjusted R-squared:  0.2925
## F-statistic: 52.64 on 8 and 991 DF,  p-value: < 2.2e-16

# Run a 6-variable model without dad_education and mom_education
OLS.logWage.6var = lm(logWage ~ education + experience + experienceSquare +
  raceColor + rural + city, data = data)
# Print the summary of the new model
summary(OLS.logWage.6var)

##
## Call:
## lm(formula = logWage ~ education + experience + experienceSquare +
##     raceColor + rural + city, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.31258 -0.23242  0.02192  0.24694  1.28360
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.7510008  0.1177400  40.352 < 2e-16 ***
## education     0.0722416  0.0061959  11.660 < 2e-16 ***
## experience    0.0892966  0.0112131   7.964 4.55e-15 ***
## experienceSquare -0.0026714  0.0005312  -5.029 5.86e-07 ***
## raceColor    -0.2345897  0.0305852  -7.670 4.09e-14 ***
## rural        -0.0963238  0.0263220  -3.659 0.000266 ***
## city          0.1677263  0.0269991   6.212 7.67e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3761 on 993 degrees of freedom
## Multiple R-squared:  0.2977, Adjusted R-squared:  0.2935
## F-statistic: 70.16 on 6 and 993 DF,  p-value: < 2.2e-16
```

## Question 4.6

### Part 1

The assumptions needed are  $\text{Cov}(z1, \text{education}) \neq 0$  and  $\text{Cov}(z1, u) = 0$ .

### Part 2

Suppose  $z1$  is an indicator representing whether or not an individual lives in an area in which there was a recent policy change to promote the importance of education. Yes,  $z1$  could be correlated with other unobservables captured in the error term. Some examples are 1. Income. People with higher incomes might

be more educated and thus might place a higher importance on education and thus be more likely to live in an area that promotes education, 2. Political party. A particular political party might be more aligned with education and therefore people in that political party might be more inclined to live in an area that promotes education, and 3. Whether you voted or not. It's possible that people who vote might be more educated and more likely to live in an area that promotes education. These are just a few examples. There could be many more.

## Part 3

Using the same specification as that in question 4.5, estimate the equation by 2SLS, using both  $z_1$  and  $z_2$  as instrument variables.

The coefficient estimate on education goes from 0.0681701 in the original model to 0.0950302, however, in the new model, the education estimate is not significant at the 5% level, so the increase in the coefficient can no longer be used in our interpretation.

However, if we remove `mom_education` and `dad_education` from both the TSLS and original models, the education coefficient becomes significant again at the 5% level. The value of the education coefficient now goes from 0.0722416 in the original model to 0.1042749 in the TSLS model. This means that using  $z_1$  and  $z_2$  as instrumental variables the effect of education on `logWage` increases from about 7.2% to 10.4% (an increase of about 3 percentage points). This is a 44% increase which is a large practical effect.

```
# Run the IV TSLS regression with z1 and z2
TSLS.logWage.8var = ivreg(logWage ~ education + experience + experienceSquare +
  raceColor + dad_education + mom_education + rural + city | z1 * z2 +
  experience + experienceSquare + raceColor + dad_education + mom_education +
  rural + city, data = data)
# Print the summary of TSLS the model
summary(TSLS.logWage.8var)
```

```
##
## Call:
## ivreg(formula = logWage ~ education + experience + experienceSquare +
##       raceColor + dad_education + mom_education + rural + city |
##       z1 * z2 + experience + experienceSquare + raceColor + dad_education +
##       mom_education + rural + city, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.31628 -0.23169  0.03689  0.23949  1.03574
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.2815365  0.8256004   5.186 2.80e-07 ***
## education     0.0950302  0.0610647   1.556  0.12010
## experience    0.1069713  0.0255275   4.190 3.13e-05 ***
## experienceSquare -0.0030032  0.0006815  -4.407 1.21e-05 ***
## raceColor    -0.2001502  0.0517616  -3.867  0.00012 ***
## dad_education -0.0041758  0.0085477  -0.489  0.62533
## mom_education  0.0071767  0.0112304   0.639  0.52300
## rural        -0.0888567  0.0324316  -2.740  0.00630 **
## city          0.1670192  0.0412727   4.047 5.76e-05 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3818 on 714 degrees of freedom
## Multiple R-Squared:  0.2624, Adjusted R-squared:  0.2541
## Wald test:      24 on 8 and 714 DF, p-value: < 2.2e-16

# Print the summary of the original model
summary(OLS.logWage.8var)

##
## Call:
## lm(formula = logWage ~ education + experience + experienceSquare +
##     raceColor + dad_education + mom_education + rural + city,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.2961 -0.2240  0.0160  0.2454  1.0404
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.6422296   0.1408825   32.951 < 2e-16 ***
## education      0.0681701   0.0077409    8.806 < 2e-16 ***
## experience     0.0973419   0.0133133    7.312 7.1e-13 ***
## experienceSquare -0.0029568  0.0006678   -4.428 1.1e-05 ***
## raceColor     -0.2130226   0.0425014   -5.012 6.8e-07 ***
## dad_education -0.0011474   0.0050988   -0.225 0.82202
## mom_education  0.0113176   0.0061886    1.829 0.06785 .
## rural         -0.0919377   0.0314151   -2.927 0.00354 **
## city           0.1782137   0.0323826    5.503 5.2e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3786 on 714 degrees of freedom
## (277 observations deleted due to missingness)
## Multiple R-squared:  0.2746, Adjusted R-squared:  0.2665
## F-statistic: 33.79 on 8 and 714 DF, p-value: < 2.2e-16

# Run the IV TSLS regression with z1 and z2 with only 6 variables,
# removing mom_education and dad_education.
TSLS.logWage.6var = ivreg(logWage ~ education + experience + experienceSquare +
    raceColor + rural + city | z1 * z2 + experience + experienceSquare +
    raceColor + rural + city, data = data)
# Print the summary of the 6 variable TSLS model
summary(TSLS.logWage.6var)

##
## Call:
## ivreg(formula = logWage ~ education + experience + experienceSquare +
##     raceColor + rural + city | z1 * z2 + experience + experienceSquare +
##     raceColor + rural + city, data = data)
##
```



```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.33618 -0.23434  0.02741  0.23425  1.25226
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.2132904   0.7968306   5.288 1.52e-07 ***
## education     0.1042749   0.0473529   2.202  0.02789 *
## experience    0.1024823   0.0224135   4.572 5.43e-06 ***
## experienceSquare -0.0026811  0.0005385  -4.979 7.55e-07 ***
## raceColor    -0.1978820   0.0620751  -3.188  0.00148 **
## rural        -0.0873090   0.0297651  -2.933  0.00343 **
## city          0.1495403   0.0381914   3.916 9.64e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3811 on 993 degrees of freedom
## Multiple R-Squared:  0.2788, Adjusted R-squared:  0.2745
## Wald test: 47.07 on 6 and 993 DF, p-value: < 2.2e-16
```

```
# Print the summary of the original 6 variable model
summary(OLS.logWage.6var)
```

```
##
## Call:
## lm(formula = logWage ~ education + experience + experienceSquare +
##      raceColor + rural + city, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.31258 -0.23242  0.02192  0.24694  1.28360
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.7510008   0.1177400  40.352 < 2e-16 ***
## education     0.0722416   0.0061959  11.660 < 2e-16 ***
## experience    0.0892966   0.0112131   7.964 4.55e-15 ***
## experienceSquare -0.0026714  0.0005312  -5.029 5.86e-07 ***
## raceColor    -0.2345897   0.0305852  -7.670 4.09e-14 ***
## rural        -0.0963238   0.0263220  -3.659 0.000266 ***
## city          0.1677263   0.0269991   6.212 7.67e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3761 on 993 degrees of freedom
## Multiple R-squared:  0.2977, Adjusted R-squared:  0.2935
## F-statistic: 70.16 on 6 and 993 DF, p-value: < 2.2e-16
```