Wages Dataset

The Dataset

The *Wages* dataset—stored in the *Ecdat* library—looks at variables pertaining to the wages of individuals during 1976-1982.

- *exp*: The subject's years of experience.
- *wks*: The subject's weeks worked per year.
- *bluecol*: A factor indicating whether a subject's job was blue collar.
- *ind*: A factor indicating whether the subject works in the manufacturing industry.
- *south*: A factor indicating whether the subject lived in the south.
- *smsa*: A factor indicating if the subject lives in a standard metropolitan statistical area.
- *married*: A factor indicating if the subject is married.
- *sex*: A factor indicating the sex of the subject.
- *union*: A factor indicating if the subject's wage is set by a union contract.
- *ed*: The subject's years of education.
- *black*: A factor indicating if the subject is black.
- *lwage*: The logarithm of the individuals wage.

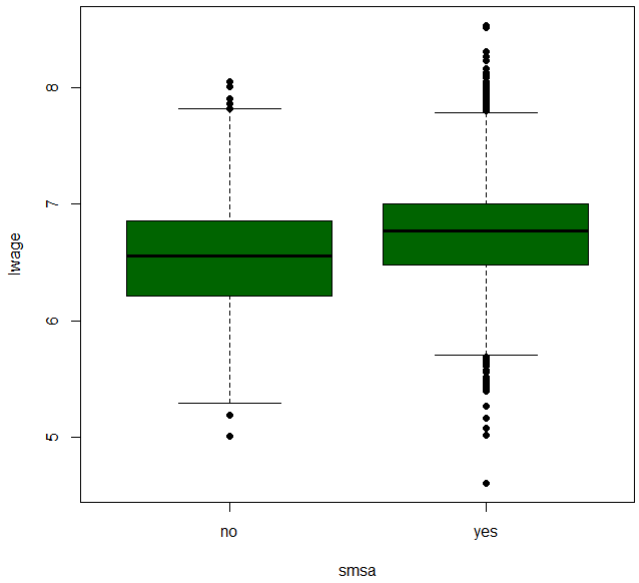
Potential Questions

There are many great opportunities in this dataset, looking at the effects of sex, race, education, and geography on wages. Beyond that, we can look into the ways that these variables interplay with each other. This gives students plenty of chances to explore, along with—if they care to—trying to find updated versions of the dataset. It should be noted that this is time series data, as each subject appears seven times across the seven years in the dataset. Advanced questions—well beyond the scope of a first class—would include looking at mixed model effects for each individual or other time series models. However, even with just this data, there are a few routes to go.

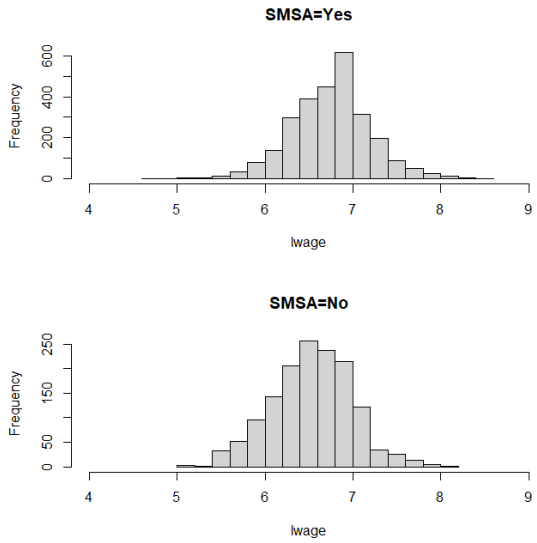
1. **Do individuals in a metropolitan area earn higher wages? (2-sample mean)**
2. Do married workers earn more money than non-married workers? (2-sample mean)
3. Do men earn more money than women? (2-sample mean)
4. Do individuals who have their contract set by unions earn differently from those who do not? (2-sample mean)
5. **Does the south have a higher rate of blue collar workers? (2-sample proportion)**
6. Do men work in the manufacturing industry at higher rates than women (2-sample proportion)
7. Are men more likely to belong to a union than women? (2-sample proportion)
8. **Is education associated with log-wage? (Correlation)**
9. **How does years of experience affect log-wage? (Regression)**
10. What are the most important factors in predicting wage?*

Question 1

The urban/rural divide has always been an interesting distinction for statistical analyses, particularly in socioeconomic data. There are such distinct differences in types of employment, social supports, and demographics—among other things—that comparing the two locations almost always yields interesting results and commentary. Here, we look at one of the standard comparisons: where people make more money. Now, just like in many of these questions, the answer may come down to the jobs and opportunities available in the areas, but here we'll just work with the data. A brief look at the exploratory analyses suggests a specific answer.



It seems like individuals living in an MSA seem to probably make a higher wage, but there's definitely higher variability and the possibility for more outliers—in both directions. These outliers could actually drive the results of the hypothesis tests, meaning that we need to go the actual inferential techniques—after checking the equality of variances of course. In this case, we'll still just test $H_A : \mu_{YesSMSA} - \mu_{NoSMSA} \neq 0$ at the $\alpha = 0.05$ level, where μ is the average log-wage. The test returns a test statistic of $t = -14.828$ with a p-value of $2 \times P\big(t_{4163} > |-14.828|\big) = 1.638 \times 10^{-48}$. Thus, we'd reject the null hypothesis and conclude that there does appear to be a difference in the population mean wages for individuals living in a standard metropolitan statistical area and those not.



I will note that the distributions of log-wage for these two groups are rather interesting. They have nearly identical variances with different distributional shapes and concentration of data (Evidenced by the boxplots as well as looking at their respective IQRs). There's something that can be gleaned from this that would be an interesting analysis, again pertaining to the opportunities available in different locations as well as the distribution of the wages from those opportunities.

```
library(Ecdat)
data(Wages)
boxplot(lwage~smsa,data=Wages,
col=''darkgreen'',pch=16)
var(lwage[smsa==''no''])
var(lwage[smsa==''yes''])
t.test(lwage~smsa,data=Wages,
var.equal=T)
par(mrow=c(1,2))
hist(lwage[smsa=="yes"],breaks=20,
xlim=c(4,9),main="SMSA=Yes",
xlab="lwage")
hist(lwage[smsa=="no"],breaks=20,
xlim=c(4,9),main="SMSA=No",
xlab="lwage")
```

Question 5

Over time, the southern United States is a region that has become more synonymous with blue collar jobs. Some of this thought may be in part confounded with more rural farming jobs—which are not technically blue collar by some definitions—but is still worth testing. In short, we'll be testing $H_A : p_{South} - p_{NonSouth} > 0$, where p is the proportion of blue collar jobs, at the $\alpha = 0.05$ level. The exploratory analyses of contingency tables suggest it's possible for these to be equal, though as we know it doesn't take a large difference to find statistical significance in large enough sample sizes.

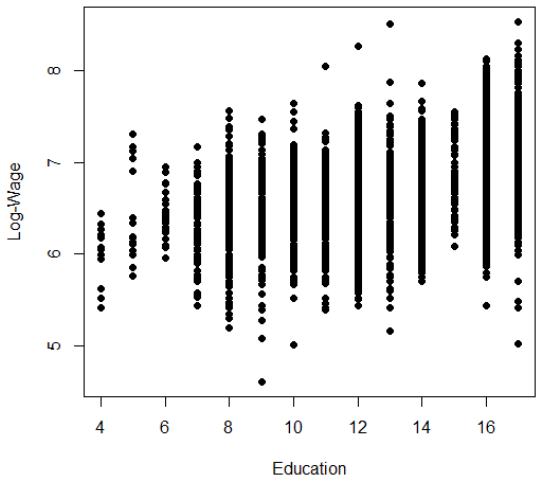
	Blue Collar	Non-Blue Collar
Non-South	1484	1472
South	552	657

The test returns a test statistic of $t = \sqrt{7.0948} = 2.664$ with a p-value of $P\big(N(0,1) > 2.664\big) = 0.0039$, thus rejecting the null hypothesis and concluding that the rate of blue collar jobs is indeed higher in the south. We should note that all our assumptions regarding sample size check out.

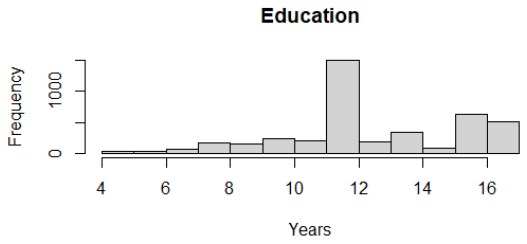
```
library(Ecdat)
data(Wages)
table(south=Wages$south,
blue=Wages$bluecollar)
prop.test(x=c(657,1472),n=c(552+657,1484+1472),
alternative="greater",correct=F)
sqrt(7.0948)
```

Question 8

Education is generally thought to be associated with salary. This effect has likely crystallized over the past few decades as more students complete high school and college becomes required of more jobs. However, in this dataset in the 1970s and early 80s, it's possible that the effect is less pronounced. To test this, we can look at if education and log-wage are correlated or not.



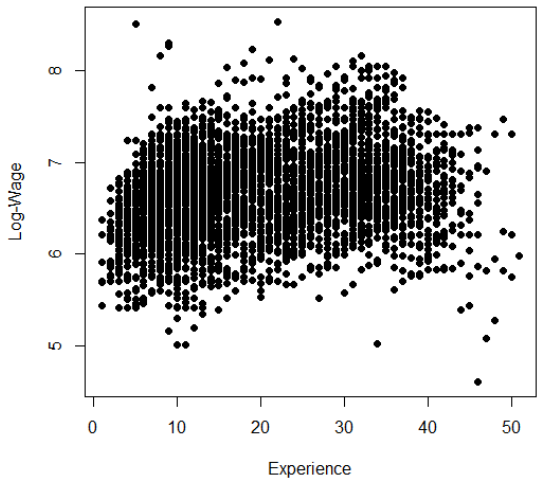
The linearity assumption for the test seems very reasonable, and with a sample correlation of $r = 0.3939$, it would seem likely that we'll find the two variables are indeed correlated. We'll test $H_A : \rho \neq 0$ at the $\alpha = 0.05$. The test returns a test statistic of $t = 27.65$, with a p-value of $2 \times P\big(t_{4163} > |27.65|\big) = 1.21 \times 10^{-154}$. So we'd definitely reject the null hypothesis and conclude that education and wage are correlated. However, we should note that while log-wage is roughly Normally distributed, education is decidedly not Normal. Additionally we should not that for each subject, education is generally a static variable through the dataset while their wage increases. Again, this is a vestige of the time series nature of the dataset, but something that maybe should be considered or thought about when analyzing the data.




```
library(Ecdat)
data(Wages)
plot(Wages$edu, Wages$lwage, pch=16,
xlab=''Education'', ylab=''Log-Wage'')
cor(Wages$edu, Wages$lwage)
cor.test(Wages$edu, Wages$lwage)
par(mfrow=c(2,1))
hist(Wages$edu, main=''Education'', xlab=''Years'')
hist(Wages$lwage, main=''Log-Wage'', xlab=''Log-Wage'')
```

Question 9

Similar to education, years of experience is expected to affect an individual's salary. However, the question is by how much? We can look at this through the lens of simple linear regression. First however, we'll look at the scatterplot of the data.



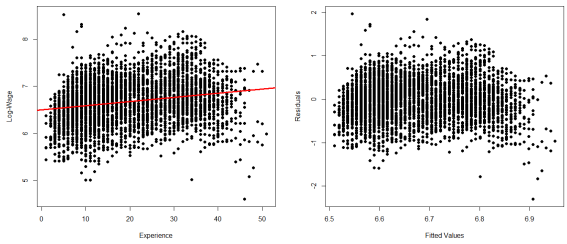
Unsurprisingly, as experience increase, so does salary. Quantifying this using the *lm* function, we see that for every extra year worked, the individual's expected log-wage increase by 0.0088, which is statistically significantly different from 0. However, we should note that this model isn't the strongest fit, with an $R^2 = 0.04383$, an acknowledgment of the many factors that likely go into this prediction.

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.30153 -0.29144  0.02307  0.27927  1.97171

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.5014318  0.0144657  449.44   <2e-16 ***
exp         0.0088101  0.0006378   13.81   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4513 on 4163 degrees of freedom
Multiple R-squared:  0.04383,   Adjusted R-squared:  0.0436
F-statistic: 190.8 on 1 and 4163 DF,  p-value: < 2.2e-16
```

Additionally, when we look at our residuals, though I would say that they are met there is a little odd "fan-like" pattern. Some of this likely is due to the through time nature of our data, and could actually bring into question whether our residuals are independent. To check this, we'd need to look at the correlation of our residuals—the proxies of our errors—through time for each individual. This is generally referred to as checking for autocorrelation, and is an essential part of times series analysis.




```
library(Ecdat)
data(Wages)
plot(Wages$exp, Wages$lwage, pch=20,
xlab=''Experience'', ylab=''Log-Wage'')
lm=lm(lwage~exp, data=Wages)
summary(lm)
par(mfrow=c(1,2))
plot(Wages$exp, Wages$lwage, pch=20,
xlab=''Experience'', ylab=''Log-Wage'')
abline(lm, col=''red'', lwd=3)
plot(lm$fitted, lm$resid, pch=16,
xlab=''Fitted Values'', ylab=''Residuals'')
```