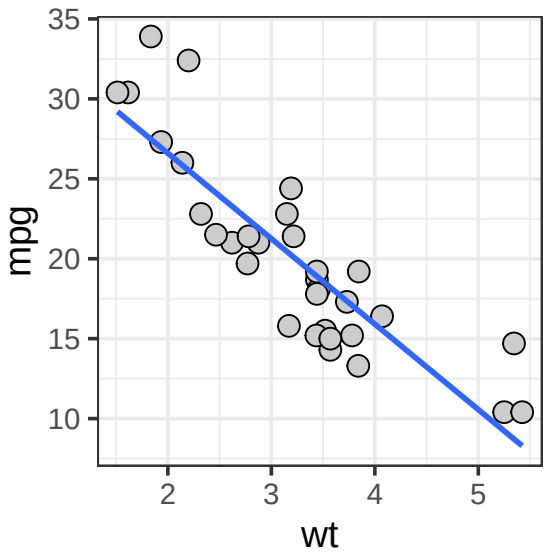
MLR Worksheet

Question 1

Below is a model illustrating the relationship between vehicle weight and miles per gallon:

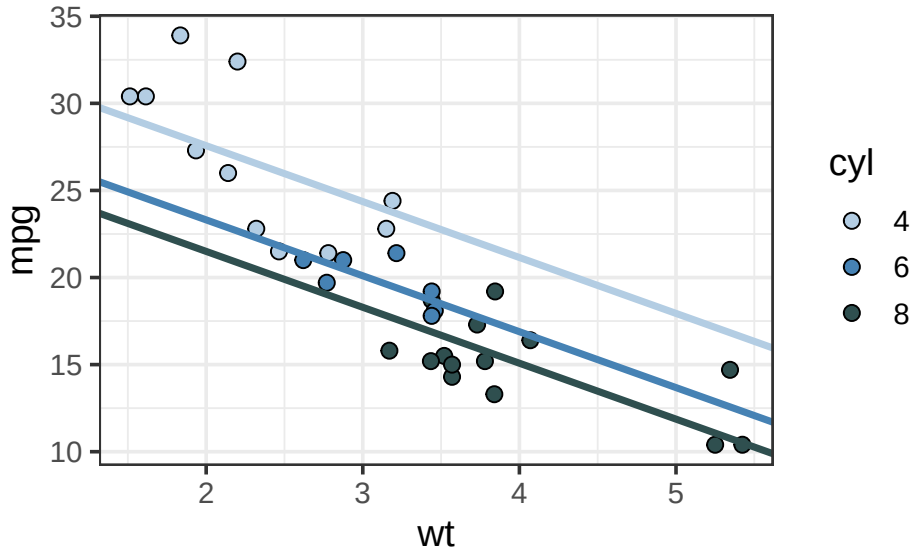



```
lm(formula = mpg ~ wt, data = mtcars)

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept)   37.285      1.878   19.86 < 0.0000000000000002 ***
wt            -5.344      0.559   -9.56        0.00000000013 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.05 on 30 degrees of freedom
Multiple R-squared:  0.753, Adjusted R-squared:  0.745
F-statistic: 91.4 on 1 and 30 DF,  p-value: 0.000000000129
```

1. What is the null hypothesis associated with the coefficient for weight?
2. Provide an interpretation of this coefficient. Do we have evidence to reject the null?
3. How much variance is explained by this model?
4. What is the null hypothesis associated with the F statistic?

Question 2

Below is a model predicting miles per gallon with weight and number of cylinders as a categorical variable

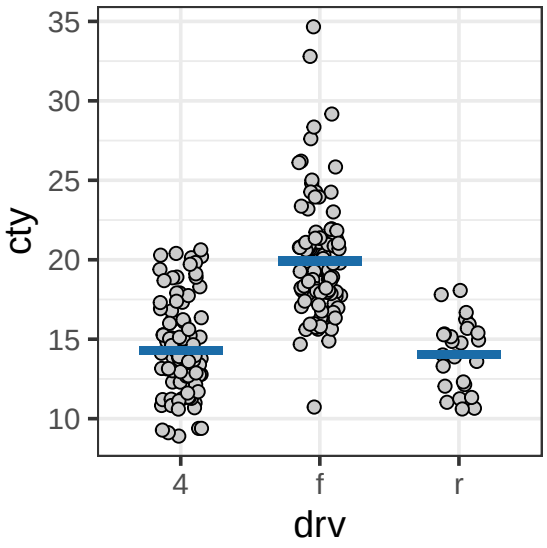



```
> lm(mpg ~ wt + cyl, mtcars) %>% summary()

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept)   33.991      1.888   18.01 < 0.0000000000000002 ***
wt            -3.206      0.754   -4.25              0.00021 ***
cyl6          -4.256      1.386   -3.07              0.00472 **
cyl8          -6.071      1.652   -3.67              0.00100 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.56 on 28 degrees of freedom
Multiple R-squared:  0.837, Adjusted R-squared:  0.82
F-statistic: 48.1 on 3 and 28 DF,  p-value: 0.0000000000359
```

1. Write out the linear equation represented by this model
2. For the coefficient for cylinder 6, what is the null hypothesis? Interpret the coefficient returned by the model. Do we have evidence to reject?
3. Using the model output, provide an assessment of how the model including weight and cylinders compares to the model with just weight. Which would you prefer?

Question 3

Below is a model predicting city miles per gallon using drvie train (with 4wd, fwd, and read rwd (`4`, `f`, and `r`, respectively)).

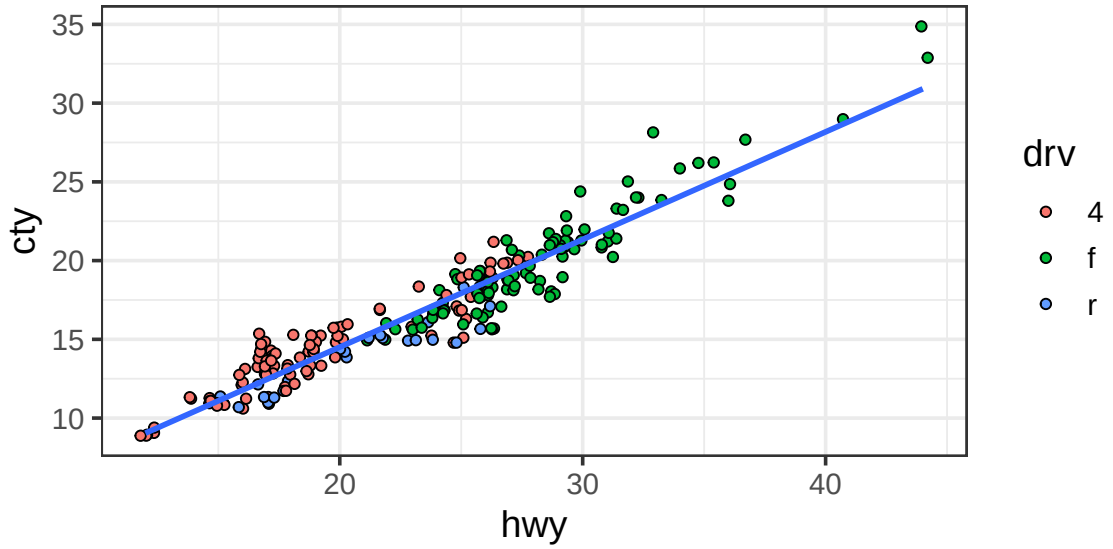



```
> lm(formula = cty ~ drv, data = mpg)

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)
(Intercept)   14.330      0.314   45.68 <0.0000000000000002 ***
drvf           5.642      0.440   12.81 <0.0000000000000002 ***
drvr          -0.250      0.710   -0.35                0.72
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.18 on 231 degrees of freedom
Multiple R-squared:  0.445, Adjusted R-squared:  0.44
F-statistic: 92.7 on 2 and 231 DF,  p-value: <0.0000000000000002
```

1. What is the reference variable for this model?
2. Interpret the coefficient for rear-wheel drive. State the hypothesis associated with this coefficient and draw a conclusion testing at $\alpha = 0.05$. Based on this, can we say that our variables are associated?
3. Consider the F statistic. Does it appear there is an association between city mpg and drive train? Reconcile this with your answer to Part 2

Question 4

Consider the same model as the previous question, now including highway mpg as a predictor

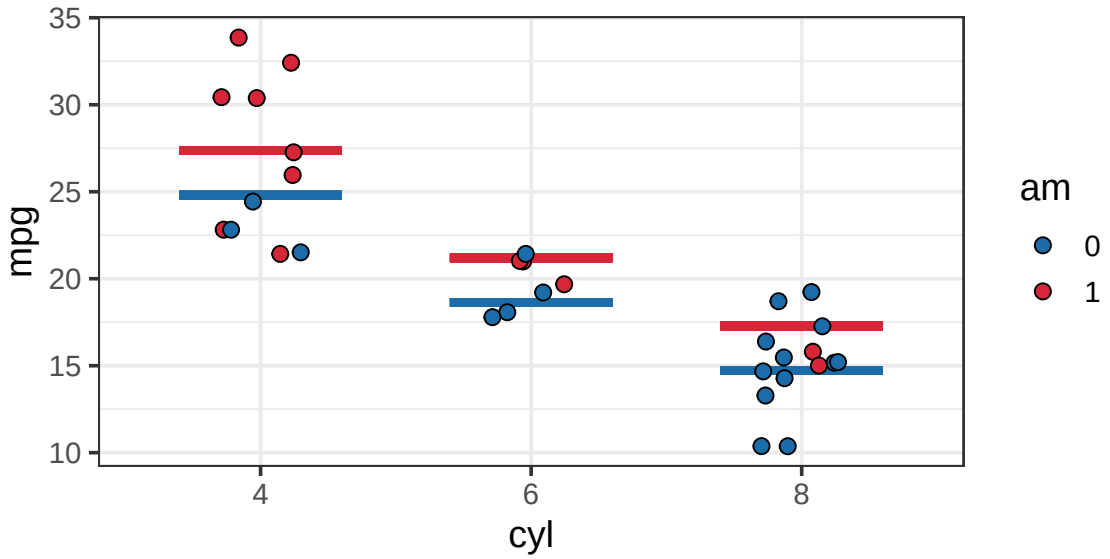



```
> lm(cty ~ drv + hwy, mpg) %>% summary()

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept)   0.4501     0.3753    1.20              0.23175
drvf         -0.8628     0.2321   -3.72              0.00025 ***
drvr         -1.5713     0.2611   -6.02         0.0000000069 ***
hwy           0.7239     0.0186   38.83 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.16 on 230 degrees of freedom
Multiple R-squared:  0.927, Adjusted R-squared:  0.926
F-statistic:  967 on 3 and 230 DF,  p-value: <0.0000000000000002
```

1. Compare the coefficients for drive train in this model to the previous one. Have the differences from the reference variable gotten bigger or smaller? Does this suggest more or less evidence for a difference?
2. How have the t statistics associated with drive train covariate changed? Why is this better evidence for a difference than the size of the coefficients?
3. How has including `hwy` in the model changed our interpretation of the drive train coefficients? What else do we have to take into consideration?
4. In your own words, explain why including `hwy` had such a large effect on the differences and statistical significance of the differences for the drive train coefficients.

Question 5

For this, `am` represents the variable automatic/manual, with 0 corresponding to an automatic transmission




```
> lm(mpg ~ cyl + am, mtcars) %>% summary()

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept)    24.80       1.32   18.75 < 0.0000000000000002 ***
cyl6           -6.16       1.54   -4.01              0.00041 ***
cyl8          -10.07       1.45   -6.93           0.00000015 ***
am1             2.56       1.30    1.97              0.05846 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.07 on 28 degrees of freedom
Multiple R-squared:  0.765, Adjusted R-squared:  0.74
F-statistic: 30.4 on 3 and 28 DF,  p-value: 0.00000000596
```

1. Write out the linear equation for this model. Explain what variables are included and how they are being used.
2. Does it appear the manual or automatic transmissions get better gas mileage? Testing at $\alpha = 0.05$, do we have enough evidence to conclude that there is a difference?
3. How does this model compare to the one with cylinders and weight? Justify your answer