Regression Course Project

Gavin Leeper

December 1, 2016

Executive Summary

I’ve completed some simple linear regression analysis on the mtcars dataset. After fitting a few different models to the data, I ultimately found that manual transmission vehicles in the data would on average get 7.245 more miles to the gallon than their automatic counterparts. This relationship was significant at th .02% level, making it rather strong. I examined the residuals and concluded that there was no discernable pattern to the error terms except a bit of a change in variance.

Preliminary Analysis

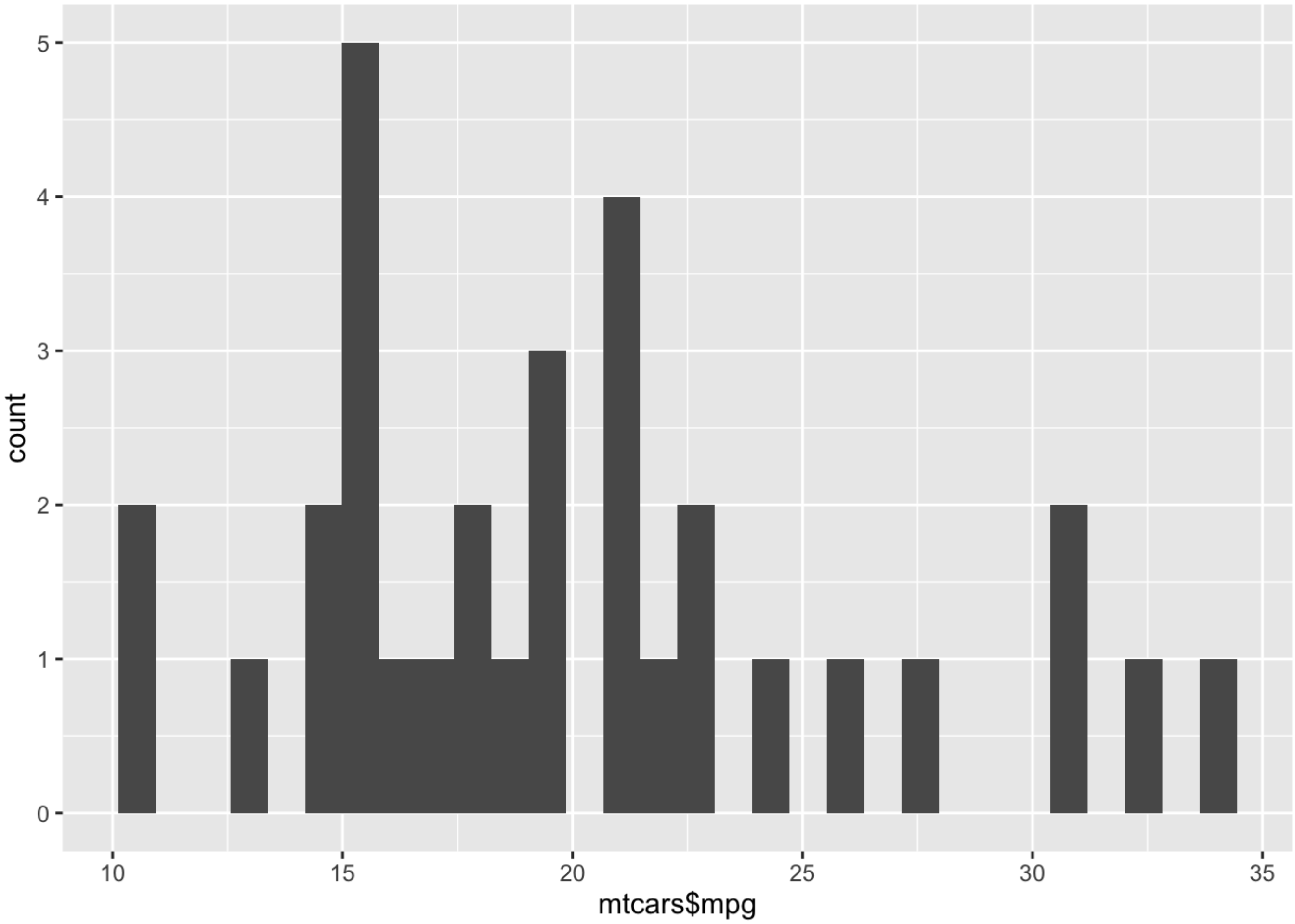
Let’s first load in the “mtcars” dataset and check look at the distributions of the variables we’re curious about.

```
library(ggplot2)
data(mtcars)
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

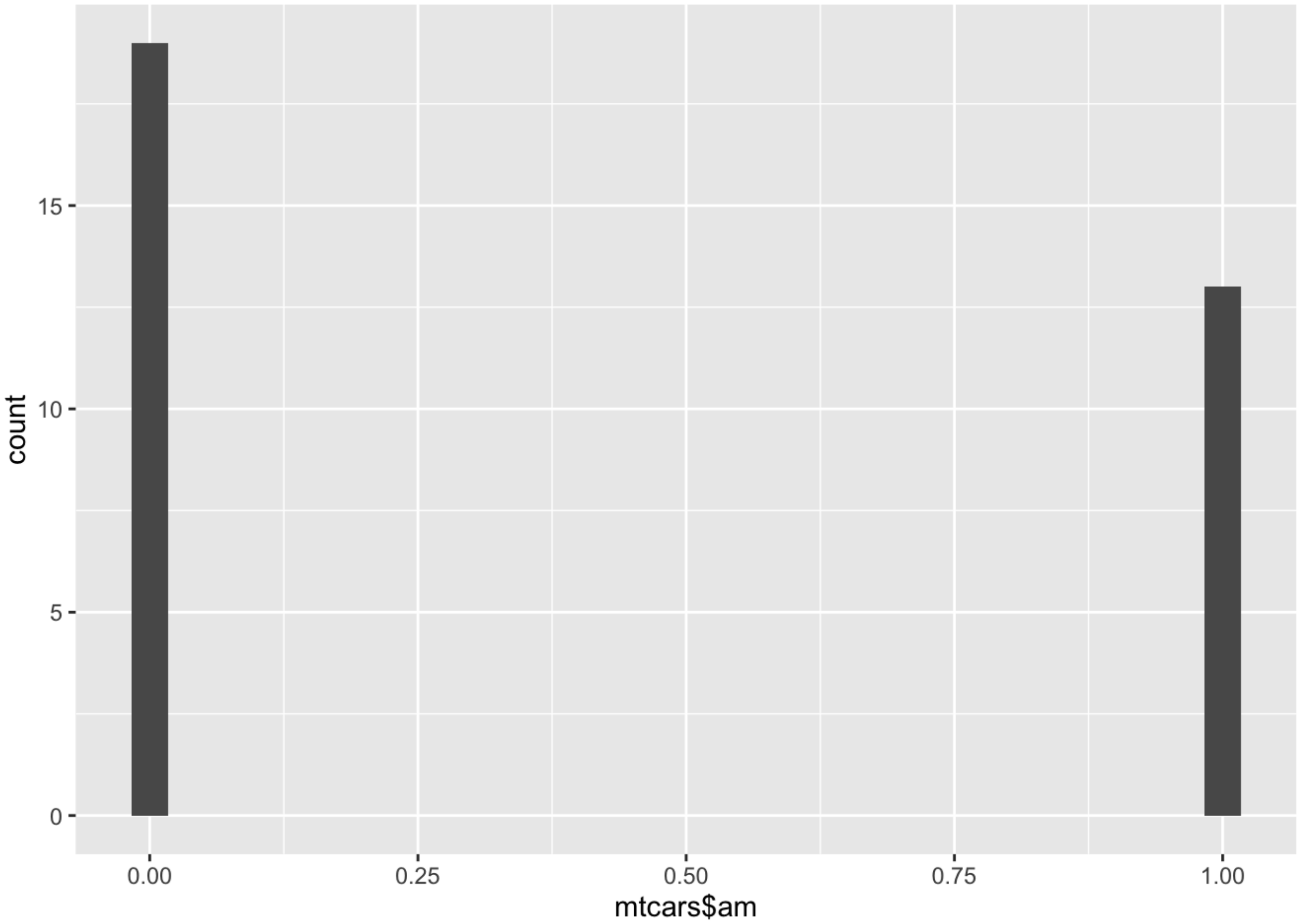
```
qplot(mtcars$mpg)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

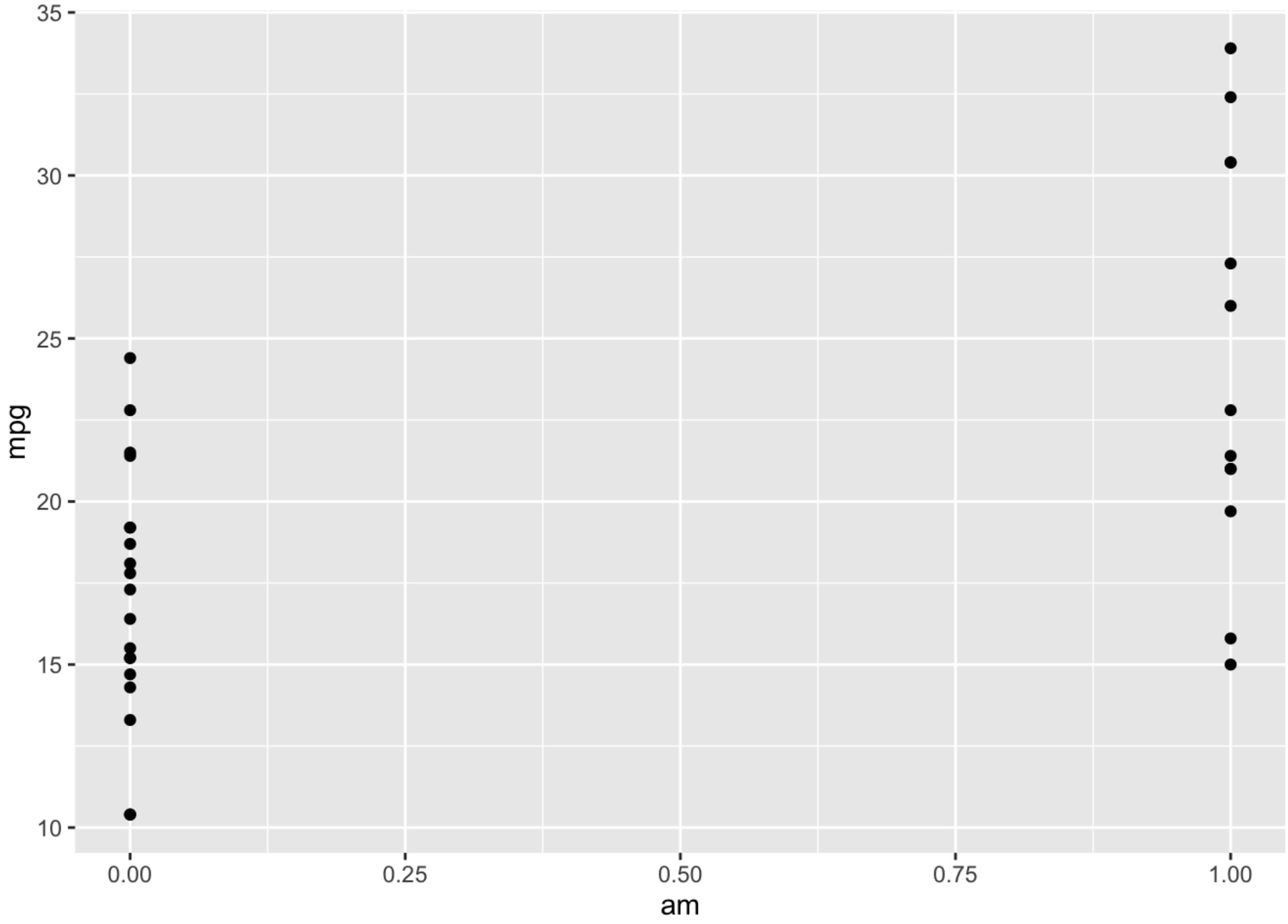


```
qplot(mtcars$am)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
qplot(am,mpg,data=mtcars)
```



Simple Linear Fit

Now let’s test some simple linear regression models to see if they reveal any relationship. We can start with predicting mpg using all the other available variables, and then pare down from there.

```
fit1<-lm(mpg~., mtcars)
summary(fit1)
```

```
## 
## Call:
## lm(formula = mpg ~ ., data = mtcars)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.4506 -1.6044 -0.1196  1.2193  4.6271 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 12.30337   18.71788   0.657   0.5181  
## cyl         -0.11144    1.04502  -0.107   0.9161  
## disp         0.01334    0.01786   0.747   0.4635  
## hp          -0.02148    0.02177  -0.987   0.3350  
## drat         0.78711    1.63537   0.481   0.6353  
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739  
## vs           0.31776    2.10451   0.151   0.8814  
## am           2.52023    2.05665   1.225   0.2340  
## gear         0.65541    1.49326   0.439   0.6652  
## carb        -0.19942    0.82875  -0.241   0.8122  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066 
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

Judging from the p value for the coefficient for the transmission variable (“am”), there seems to be no significant relationship between transmission type and mpg when all of these other variables are jointly considered. In fact, none of the variables have coefficients significant at the 5% level if we choose this model.

Interestingly, if we take the intercept term out, we do get find a significant relationship between mpg and the quarter mile time in seconds (qsec)

```
fit2<-lm(mpg~.-1, mtcars)
summary(fit2)
```

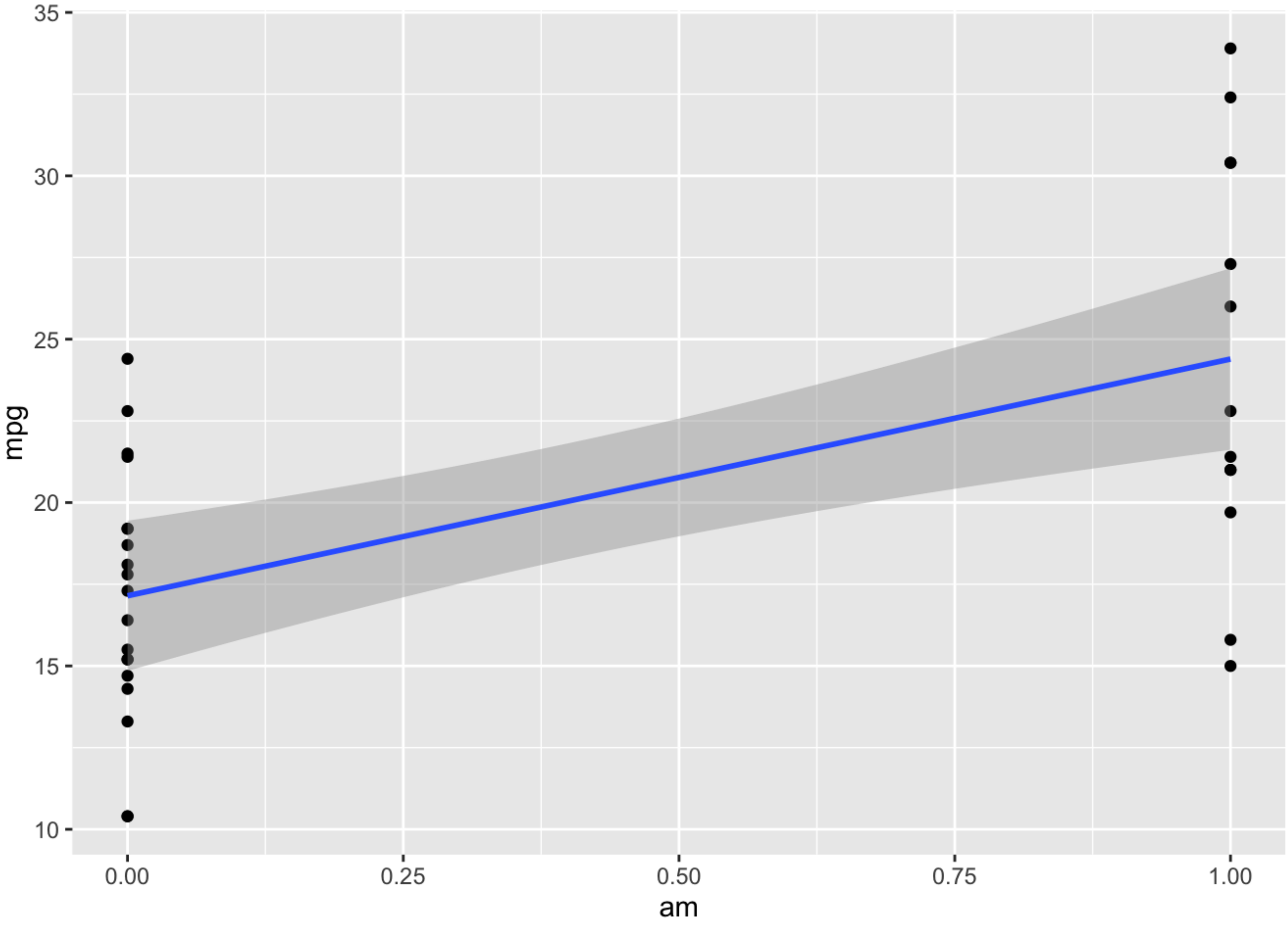
```
##
## Call:
## lm(formula = mpg ~ . - 1, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7721 -1.6249  0.1699  1.1068  4.4666
##
## Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## cyl   0.35083    0.76292   0.460   0.6501
## disp  0.01354    0.01762   0.768   0.4504
## hp   -0.02055    0.02144  -0.958   0.3483
## drat  1.24158    1.46277   0.849   0.4051
## wt   -3.82613    1.86238  -2.054   0.0520 .
## qsec  1.19140    0.45942   2.593   0.0166 *
## vs    0.18972    2.06825   0.092   0.9277
## am    2.83222    1.97513   1.434   0.1656
## gear  1.05426    1.34669   0.783   0.4421
## carb -0.26321    0.81236  -0.324   0.7490
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.616 on 22 degrees of freedom
## Multiple R-squared:  0.9893, Adjusted R-squared:  0.9844
## F-statistic:   203 on 10 and 22 DF,  p-value: < 2.2e-16
```

We can next try the other extreme, in which we only try to predict mpg based on the transmission variable and an intercept term.Our resulting p values show a high significance for both the intercept term and the am term.

```r
library(ggplot2)
fit3<-lm(mpg~am, mtcars)
summary(fit3)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am             7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

```
g<-qplot(am,mpg,data=mtcars)
g+geom_smooth(method="lm")
```

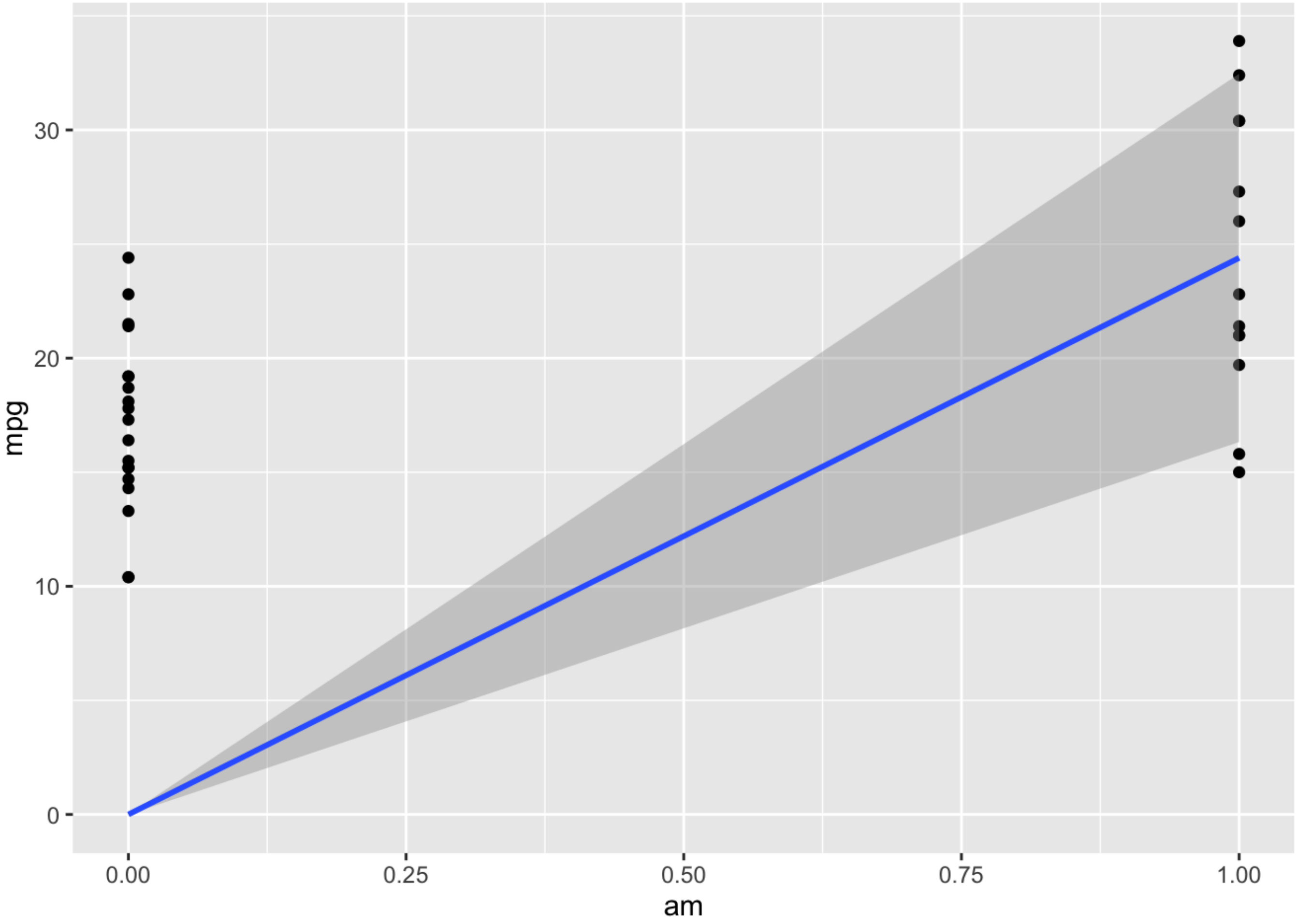


Given the significance of the intercept term, it seems like we're best off leaving it in our model, but let's check the model fitted with no intercept as well for completeness.

```
library(ggplot2)
fit4<-lm(mpg~am-1, mtcars)
summary(fit4)
```

```
##
## Call:
## lm(formula = mpg ~ am - 1, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.392  2.583 13.800 17.875 24.400
##
## Coefficients:
##    Estimate Std. Error t value Pr(>|t|)
## am   24.392      3.956   6.166 7.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.26 on 31 degrees of freedom
## Multiple R-squared:  0.5508, Adjusted R-squared:  0.5363
## F-statistic: 38.01 on 1 and 31 DF,  p-value: 7.666e-07
```

```
g<-qplot(am,mpg,data=mtcars)
g+geom_smooth(method="lm",formula=y~x-1)
```



It turns out that at least by measure of R squared, this model actually fits the data better. That said, it might not make practical sense because it forces our regression line through 0, which essentially predict that all automatic transmission cars have an mpg of 0. In fact, if we look at the predicted values of fit4, we see that all automatic tranmission cars get a mpg prediction of 0 and all manual transmission cars get 24.39.

```
predict(fit4)
```

```
##           Mazda RX4       Mazda RX4 Wag          Datsun 710
##            24.39231            24.39231            24.39231
##      Hornet 4 Drive   Hornet Sportabout             Valiant
##             0.00000             0.00000             0.00000
##          Duster 360           Merc 240D            Merc 230
##             0.00000             0.00000             0.00000
##            Merc 280           Merc 280C          Merc 450SE
##             0.00000             0.00000             0.00000
##          Merc 450SL         Merc 450SLC  Cadillac Fleetwood
##             0.00000             0.00000             0.00000
## Lincoln Continental   Chrysler Imperial            Fiat 128
##             0.00000             0.00000            24.39231
##         Honda Civic      Toyota Corolla       Toyota Corona
##            24.39231            24.39231             0.00000
##    Dodge Challenger         AMC Javelin          Camaro Z28
##             0.00000             0.00000             0.00000
##    Pontiac Firebird           Fiat X1-9       Porsche 914-2
##             0.00000            24.39231            24.39231
##        Lotus Europa      Ford Pantera L        Ferrari Dino
##            24.39231            24.39231            24.39231
##       Maserati Bora          Volvo 142E
##            24.39231            24.39231
```

Of these two, I would go with fit3 due to its predictions being more logical. From this model, we can conclude that manual transmission cars have a mileage that is 7.245 mile per gallon greater than automatic transmission vehicles. As calculated in the associated p value, we can be confident in this conclusion to a .02% level of significance, which makes it quite strong.

Residual analysis

We do see some heteroskedasticity here in that the variance of the residuals for manual transmission cars looks greater than the variance for automatic transmission cars, indeed, the low p value from the Breuch Pagan test suggest that there is significant heteroskedasticity here. However, since our x value here is non-continuous, it’s challenging to confidently discern any pattern in the errors.

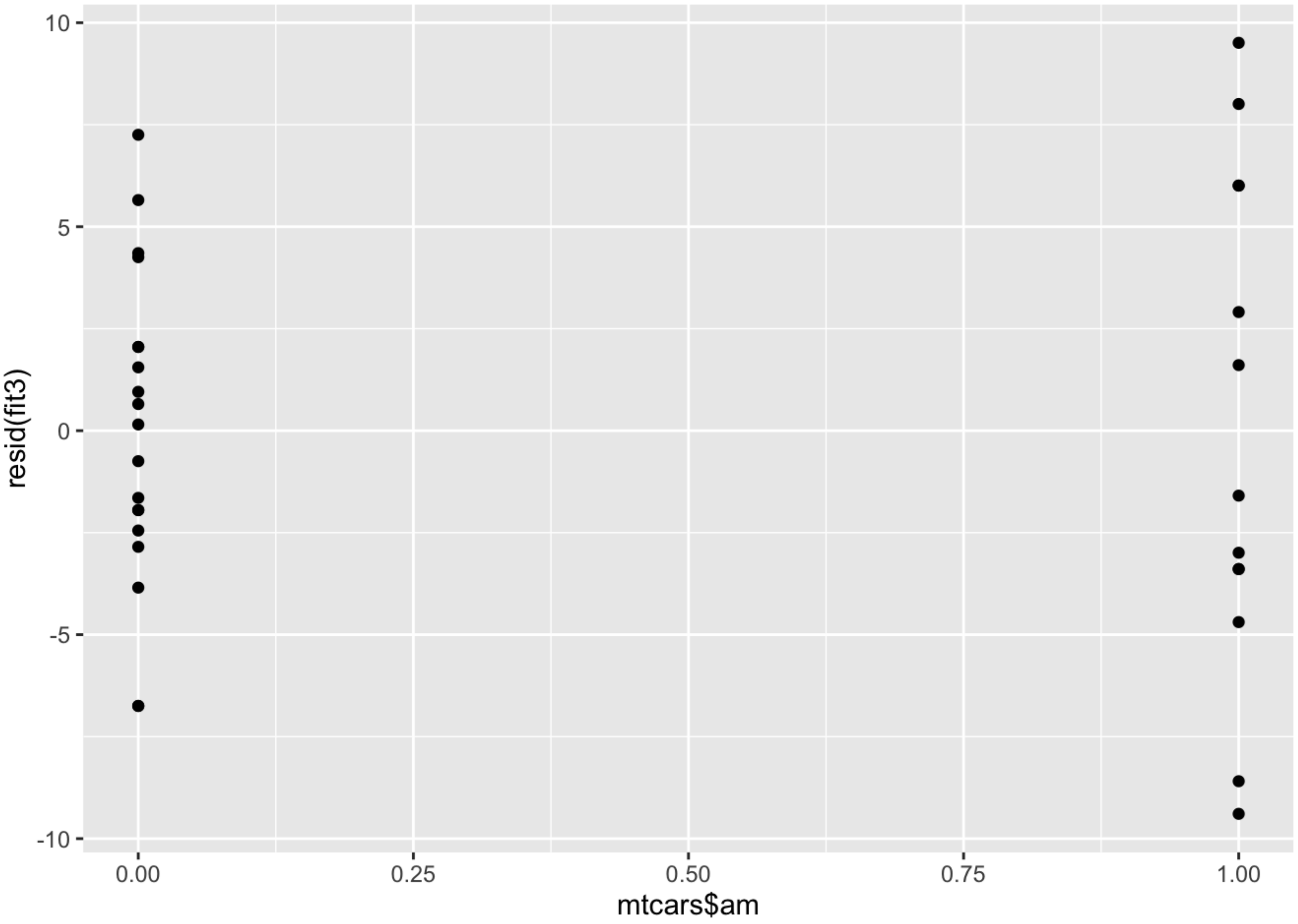
```r
library(ggplot2)
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
qplot(mtcars$am,resid(fit3))
```



```
var(resid(fit3)[mtcars$am==1])
```

```
## [1] 38.02577
```

```
var(resid(fit3)[mtcars$am==0])
```

```
## [1] 14.6993
```

```
bptest(fit3)
```

```
## 
##  studentized Breusch-Pagan test
## 
## data:  fit3
## BP = 5.0771, df = 1, p-value = 0.02424
```

A more thorough analysis could try out different transormations of mpg as the predicted variable and see if c regression.