BTRY 6020 Lab III Solutions

February 13, 2017

The goals of the following lab are:

1) To illustrate the source of multicollinearity in multiple regression;
2) To illustrate the effects of multicollinearity in multiple regression;
3) To illustrate the usefulness of polynomial regression in approximating curves.

Question 1: Introductory Multiple Linear Regression

With health care costs spiraling out of control, administrators look for inefficiencies everywhere. One particular study examined the number of nurses used to staff a hospital (NumNurses) based on the number of beds (NumBeds) and the average number of patients in the hospital (NumPats).

A) Regress the number of nurses (NumNurses) on the number of patients (NumPats). What is the coefficient of NumPats? Its estimated standard error?

We load the data and fit the linear regression $NumNurses = \beta_0 + \beta_1 NumPats + \epsilon$.

```
#load the data
library(readxl)
NurDat <- read_excel("~/BTRY6020Sp17/Labs/Lab3/Lab3NurseData.xlsx")
head(NurDat)
```

```
##   NumBeds NumPats NumNurses
## 1     279     207       241
## 2      80      51        52
## 3     107      82        54
## 4     147      53       148
## 5     180     134       151
## 6     150     147       106
```

```
#fit the model
NurDat.lm = lm(NumNurses ~ NumPats, data = NurDat)
summary(NurDat.lm)
```

```
##
## Call:
## lm(formula = NumNurses ~ NumPats, data = NurDat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -321.88  -22.89    1.50   28.50  181.61
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.71420   11.30846   0.859    0.392
## NumPats      0.79971    0.04815  16.609   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 73.36 on 111 degrees of freedom
```

```
## Multiple R-squared:  0.7131, Adjusted R-squared:  0.7105
## F-statistic: 275.9 on 1 and 111 DF,  p-value: < 2.2e-16
```

From this output, we see that the estimated coefficient of `NumPats` is 0.79971. The etimated standard error is 0.04815.

B) Regress the number of nurses NumNurses on both the number of patients (NumPats) and the number of beds (NumBeds). What is the estimated regression equation? Give a point estimate for the number of nurses that would staff a 280 bed hospital that averaged 228 patients per day.

Here we fit the multiple linear regression formula $NumNurses = \beta_0 + \beta_1 NumPats + \beta_2 NumBeds + \epsilon$.

```
#fit the model
NurDat.mlm = lm(NumNurses ~ NumPats + NumBeds, data = NurDat)
summary(NurDat.mlm)
```

```
##
## Call:
## lm(formula = NumNurses ~ NumPats + NumBeds, data = NurDat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -194.471  -20.784   -1.816   19.605  171.746
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.30807    7.94758   0.668  0.50561
## NumPats      0.20916    0.06454   3.241  0.00158 **
## NumBeds      0.51750    0.04819  10.739  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 51.49 on 110 degrees of freedom
## Multiple R-squared:  0.8599, Adjusted R-squared:  0.8574
## F-statistic: 337.7 on 2 and 110 DF,  p-value: < 2.2e-16
```

The estimated regression equation is $NumNurses = 5.30807 + 0.20916 NumPats + 0.51750 NumBeds + \epsilon$.

With this equation, we estimate that a 280 bed hostpital with 228 patients will have $5.30807 + 0.20916 * 280 + 0.51750 * 228 = 181.8629$ nurses.

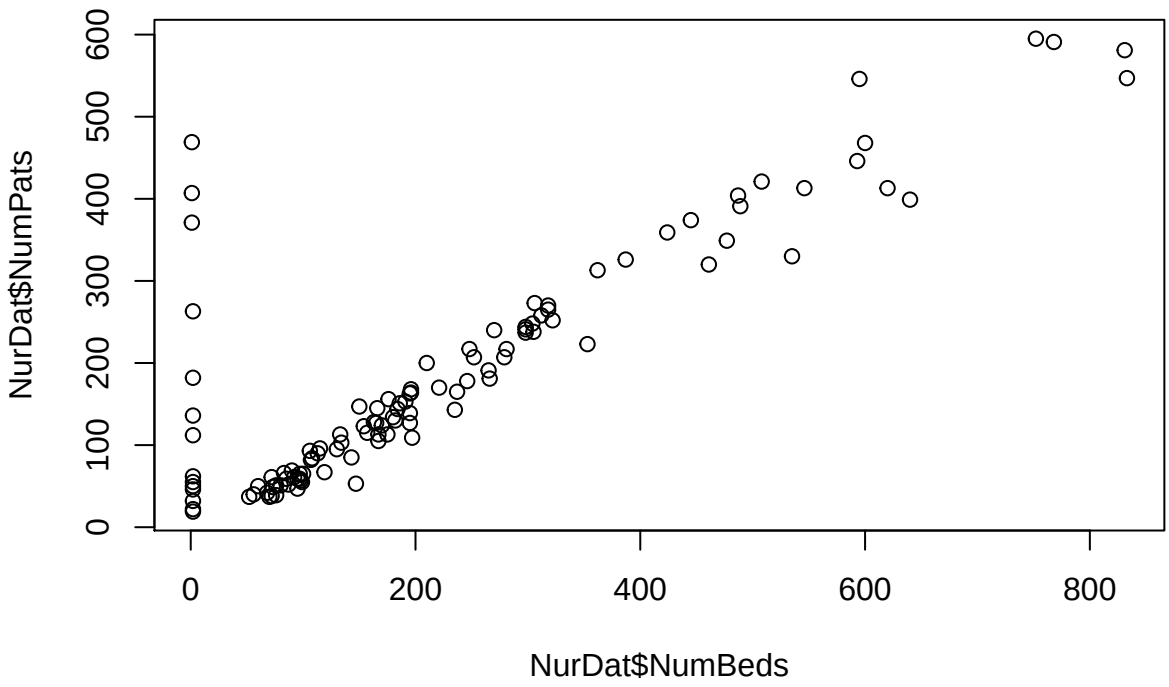
C) Now with Numbeds in the regression, what is the coefficient of NumPats? Its standard error? How has this changed from the simple linear regression in part A above?

The coefficient of `NumPats` is now 0.20916 with a standard error of 0.06464. The slope term is almost 4 times smaller than than in the simple linear regression, and the standard error is about 33% larger.

D) Plot the Numbeds vs NumPats; does your answer in part C appear to be correct?

Plotting the data

```
plot(NurDat$NumBeds, NurDat$NumPats)
```



This plot appears to be consistent with our answer in Part C. The correlation between the variables is increasing the standard error for any individual coefficient. Also, in the presence of multicollinearity we tend to see a trade-off in the slope from multiple linear regression (i.e. the once large slope of 0.79971 is now 0.20916 since part of this effect is captured by `NumBeds`)

E) Get the correlation coefficient between NumBeds and NumPats; does this confirm what you answered for part C.

Obtaining correlation

```
cor(NurDat$NumBeds, NurDat$NumPats)
```

```
## [1] 0.8519789
```

The correlation is calculated as 0.8519789 confirming multicollinearity in the data.

F) From part F above, compute the VIF between the two predictors. Confirm this by getting the VIFs from R.

The VIF can be calculated as $\frac{1}{1-corr(NumBeds,NumPats)^2}$ which is calculated below.

```
1/(1-cor(NurDat$NumBeds, NurDat$NumPats)^2)
```

```
## [1] 3.647878
```

This can be calculated in `R` using the `vif` function.

```
library(car)
vif(NurDat.mlm)
```

```
##  NumPats  NumBeds
## 3.647878 3.647878
```

Which is consistent with our calculation.

G) Interpret the coefficient of NumPats you got in the simple linear regression in Part A; then interpret the coefficient of NumPats you got in Part B. Is your interpretation of the coefficient of NumPats in the multiple regression setting valid? Why?

In the simple linear regression, we estimate that for every increase in `NumPats` by one patient, the number of nurses increased by on average 0.79971.

In the multiple linear regression, we estimate that when holding all other variables cconstant, every increase in `NumPats` by one patient will correspond to an avereage increase in number of nurses by 0.20916.

The interpretation of the coefficient in the multiple regression setting is dubious as there is a large amount of correlation between `NumPats` and `NumBeds` so, "holding all other variables constant" does not lend to a necessarily realistic interpretation of this coefficient.

H) What is the inferential model for this multiple linear regression? What assumptions does this include?

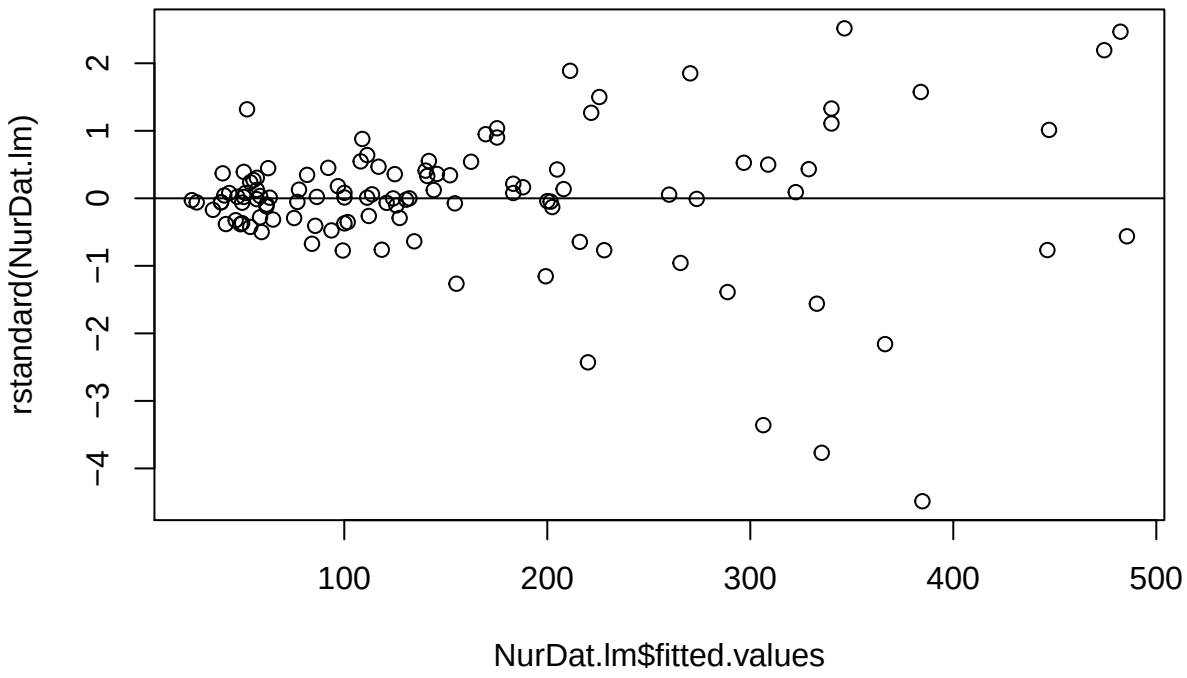
Our linear regression model used for inference is '$NumNurses$' $\sim N(\beta_0 + \beta_1 NumPats + \beta_2 NumBeds, \sigma^2)$ independently over all observations.

Therefore we assume that the data is independent and normallty distributed, with means linearly related to the predictor(s), and with equal variance. When doing inference we should also chec k to see if outliers are driving our conclusions.

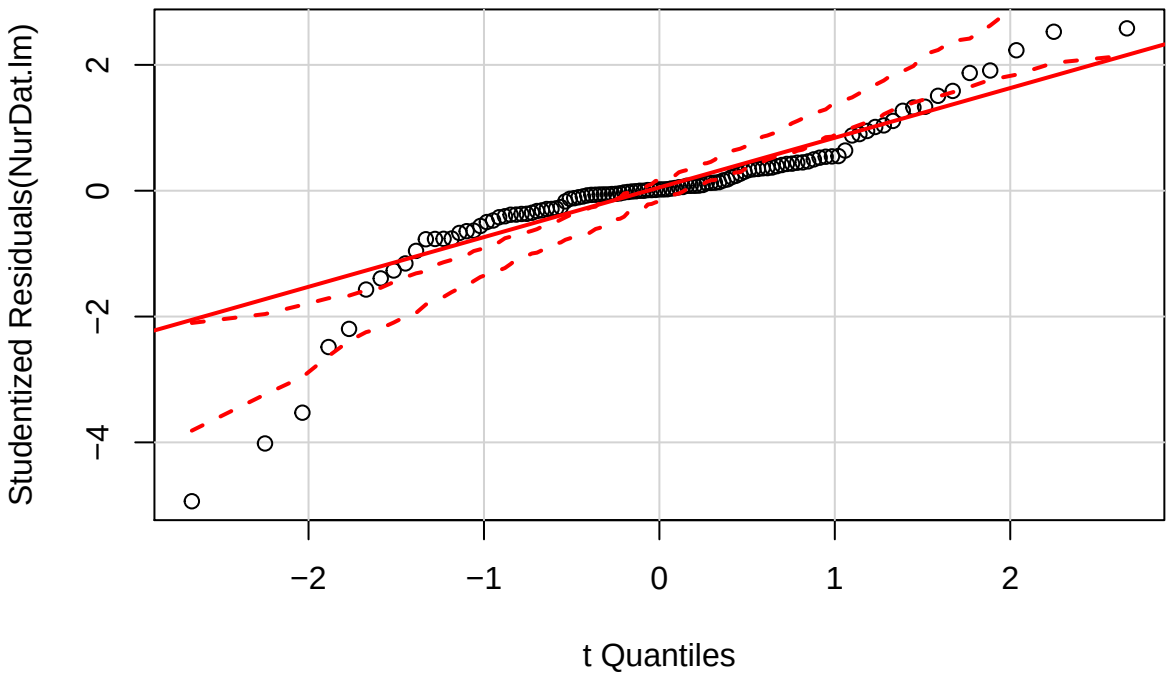
I) Get the standardized (or studentized) residual plot and qqPlot for the residuals in this multiple regression. Comment on the usefulness of this regression for inference.

Below we obtain the residual plots

```
plot(NurDat.lm$fitted.values, rstandard(NurDat.lm))
abline(h = 0)
```




```
qqPlot(NurDat.lm)
```

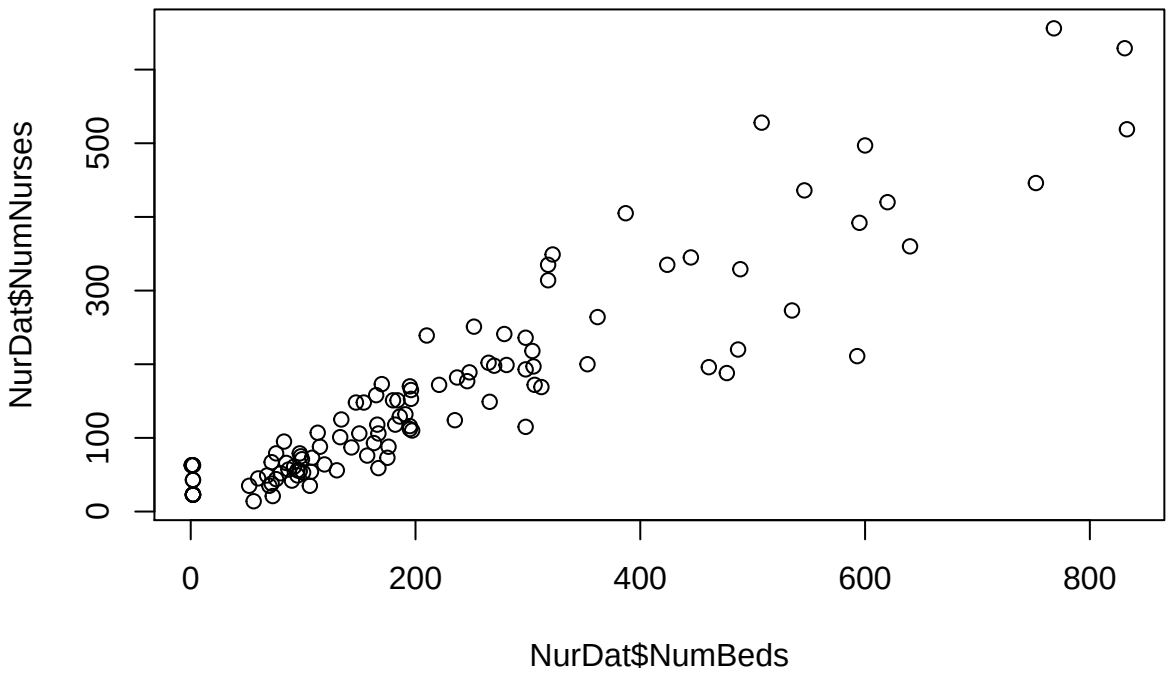


In the residual plot we see an increase in variance of the data associated with the fitted values. In the qqPlot we see a clear deviation from normality.

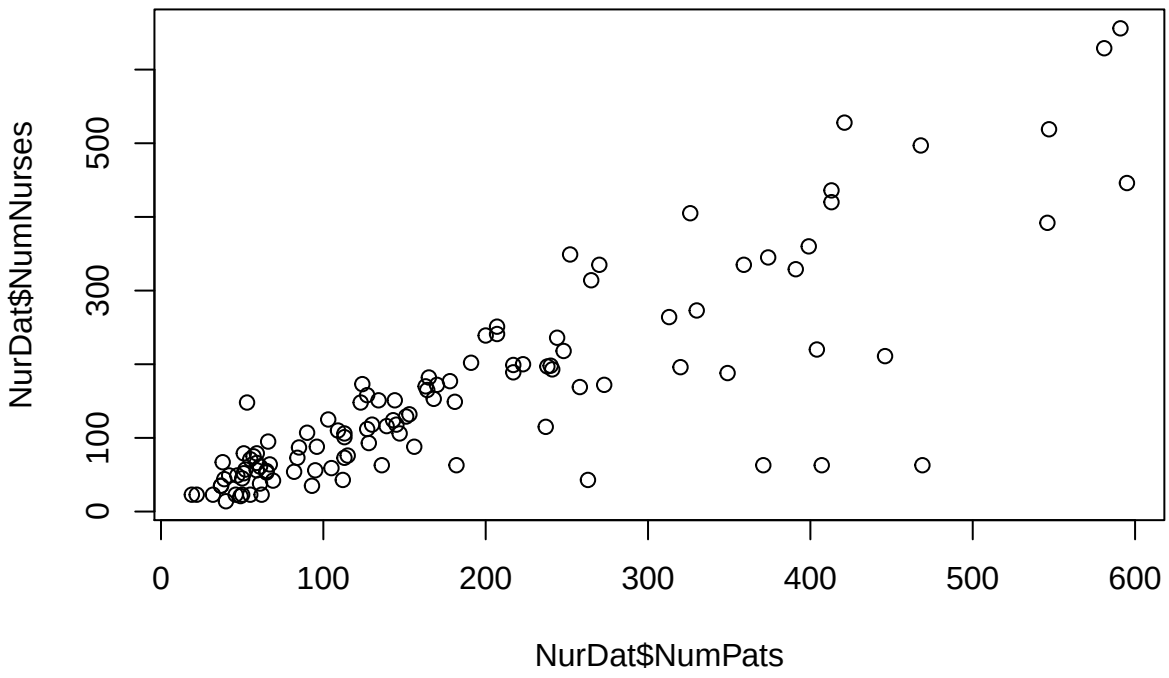
Since the regression assumptions do not appear to hold, this model may not be particularly useful for inference.

J) Plot NumNurses against Numbeds and NumNurses against NumPats. Based on these plots and the plots obtained in part J above, how would you suggest proceeding before doing inference?

```
plot(NurDat$NumBeds, NurDat$NumNurses)
```




```
plot(NurDat$NumPats, NurDat$NumNurses)
```



From the above two plots we see that the relationships between NumNurses and the two predictors both appear linear but with increasing variance. This suggests we should try transforming down Y and both Xs; Y, to stabilize variance, and the Xs to maintain linearity.

Question 2: Polynomial Regression

The effectiveness of a new experimental overdrive gear in reducing gasoline consumption was evaluated in 12 trials with a light truck equipped with this gear. A study was undertaken to determine the effectiveness of this in increasing gas mileage at different speeds.

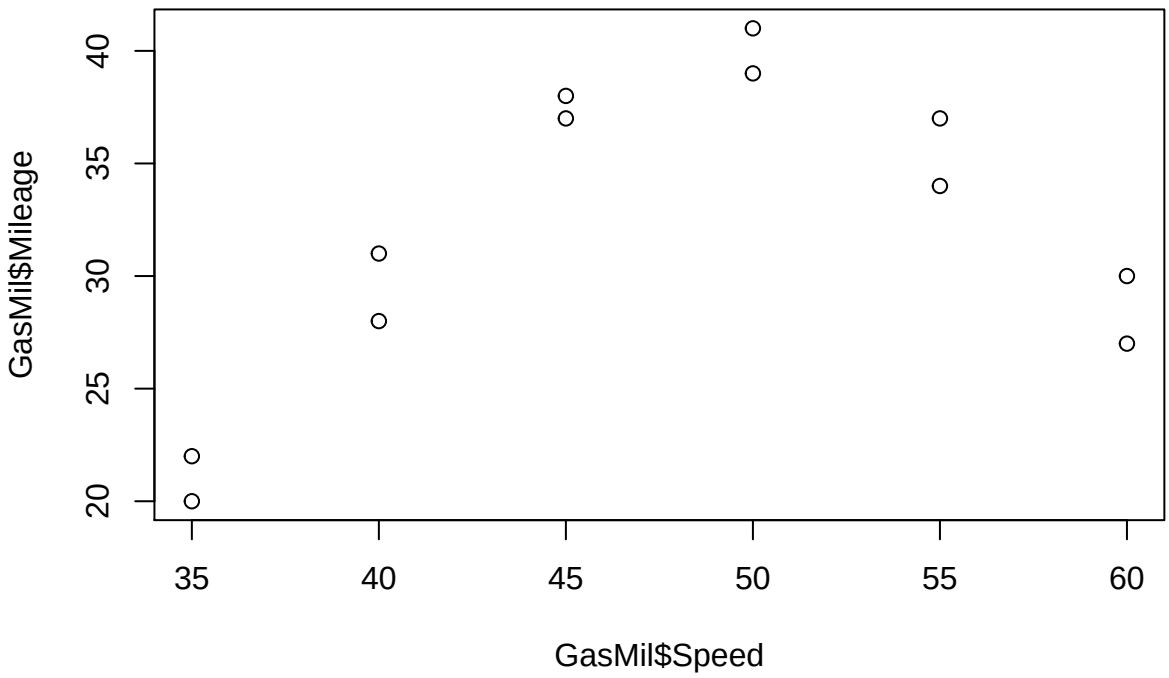
A) Plot the data (of course!).

Loading and plotting the data

```
GasMil = read_excel("Lab3GasMileage.xlsx")
head(GasMil)
```

```
##   Speed Mileage
## 1    35      22
## 2    35      20
## 3    40      28
## 4    40      31
## 5    45      37
## 6    45      38
```

```
plot(GasMil$Speed, GasMil$Mileage)
```



B) Based on what you observe in the plot, what order polynomial appears appropriate here?

Since we see one "hump" in the relationship between these variables, a 2nd order polynomial (a parabola) appears appropriate.

C) Run a polynomial regression with the order (highest power of a polynomial) one greater than what you answered in part b above. Is the highest power statistically significant?

Below we run a polynomial regression of order 3.

```
GasMil.plm3 = lm(Mileage~Speed + I(Speed^2) + I(Speed^3), data = GasMil)
summary(GasMil.plm3)
```

```
##
## Call:
## lm(formula = Mileage ~ Speed + I(Speed^2) + I(Speed^3), data = GasMil)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5675 -0.7907  0.2560  1.1419  1.8492
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -73.912698 125.695542  -0.588    0.573
## Speed         1.848677   8.204499   0.225    0.827
## I(Speed^2)    0.061984   0.175416   0.353    0.733
## I(Speed^3)   -0.001074   0.001230  -0.873    0.408
##
```

```
## Residual standard error: 1.75 on 8 degrees of freedom
## Multiple R-squared:  0.952,  Adjusted R-squared:  0.934
## F-statistic: 52.85 on 3 and 8 DF,  p-value: 1.284e-05
```

The highest power is not statistically significant in this fit.

D) Based on your answer above, make any changes necessary to your regression and rerun this. Remember, the objective of polynomial regression is to run an order one higher than necessary (that is non-significant), then drop back to an appropriate level (so the highest power is statistically significant).

Fitting the order two polynomial gives the following fit

```
GasMil.plm2 = lm(Mileage~Speed + I(Speed^2), data = GasMil)
summary(GasMil.plm2)
```

```
##
## Call:
## lm(formula = Mileage ~ Speed + I(Speed^2), data = GasMil)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.03214 -0.58929  0.03393  1.10893  2.10000
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.826e+02  1.768e+01  -10.33 2.73e-06 ***
## Speed        8.983e+00  7.616e-01   11.80 8.91e-07 ***
## I(Speed^2)  -9.107e-02  7.993e-03  -11.39 1.20e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.727 on 9 degrees of freedom
## Multiple R-squared:  0.9474, Adjusted R-squared:  0.9357
## F-statistic: 81.03 on 2 and 9 DF,  p-value: 1.757e-06
```

E) Without giving thought to being correct, use the definition of slope in multiple regression to interpret the coefficient of `Speed` . Can you see why this makes no sense whatsoever, and so interpreting these coefficients is incorrect?

We estimate that holding all other variables constant, if we increase Speed by one mph, milage will increase by 8.983 mpg.

This interpreting makes no sense, because it is impossble to increase Speed while holding Speed^2 constant. Therefore we cannot apply the usual interpretation of slope parameters in the polynomial regression model.

F) Currently used transmissions get 31 mpg at 52 miles per hour. Can we get a greater average than this with the new overdrive gear? State hypotheses, test statistic, p-value, and conclusions, assuming all assumptions for statistical inference have been met.

If $\mu_{Y|X=52}$ is the mean mileage at 52 mph, then our hypotheses are $H_0 : \mu_{Y|X=52} \leq 31$ vs $H_a : \mu_{Y|X=52} > 31$.

The estimated value for the mean mileage at this speed and its standard error can be calculated using the `predict` function and the `se.fit` option.

```
newData = data.frame(Speed = 52)
pred = predict(GasMil.plm2, newData, se.fit = TRUE)
pred
```

```
## $fit
##        1
```

```
## 38.28786
##
## $se.fit
## [1] 0.7033125
##
## $df
## [1] 9
##
## $residual.scale
## [1] 1.726658
```

From this we can calculate the test statistic and the (one-sided) pvalue.

```
#caculating the test statistic
tstat = (pred$fit-31)/pred$se.fit
tstat
```

```
##        1
## 10.36219
```

```
#calculating the pvalue
1-pt(tstat, df = 9)
```

```
##           1
## 1.32951e-06
```

Therefore, with $p < .00001$, at the 0.05 significance level we have sufficient evidence to conclude that we get a greater than 31 mpg at 52 mph with this new overdrive gear.