Stat 21 Test 2

Due: Dec 7, 2020 by noon ET

This test is due on to be submitted on Gradescope on **December 7** by **12:00pm ET**. Please use the `#test_questions` channel on Slack to post any clairfication questions. Do not ask questions like "Is [this] the right answer?"

You must submit your solutions as a single **PDF** document uploaded to **Gradescope**. You may use R markdown to write up your solutions alone or you may use R markdown and hand-written solutions. You must show all of your work, including code input and output. Please make sure each problem is **clearly labeled** and that any handwritten components (such as pictures or equations) are easily readable in the PDF document. You may want to use a service like CamScanner (https://www.camscanner.com/) to help you upload handwritten pages and Small PDF (https://smallpdf.com/merge-pdf) to merge multiple PDFs into a single document.

You are permitted to reference all class material and use the internet (though I am not sure it will be very helpful). You are not permitted however, to get assistance from any person online or otherwise.

- Your file should contain the code to answer each question in its own code block. Your code should produce plots/output that will be automatically embedded in the output pdf file.
- Each answer must be supported by written statements and relevant plots.
- Each problem is worth 50 points for a total of 100 points possible.
- In order to knit this document, make sure you have installed the following packages in your version of RStudio: `ggplot2`, `tidyverse`, `gridExtra`, `gcookbook`, `knitr`

Suppose we have randomly surveyed 91 roller coasters across the US. We are interested in investigating the relationship among a set of predictor variables with the quantitative response variable, the speed of the coaster (mph). The quantitative predictor variables we are going to consider include

- the length of the track (in feet),
- the duration of the ride (in seconds),
- the highest climb the ride reaches (in feet), and
- the lowest drop the ride reaches (in feet).

The categorical predictor variables we are going to consider are

- the type of track (wooden or steel) and
- whether or not the ride has a loop/inversion (1 for yes, 0 for no).

Use the R code below to import this data set into RStudio. (This code makes sure that there are no missing data entries in the sample.)

```r
library('tidyverse')
coasters <- read_table2(url("http://www.swarthmore.edu/NatSci/sthornt1/DataFiles/roller_coasters.txt"))
```

Note: For this data set, we are given that the data is a random sample so we can assume that it is representative of all roller coasters in the US. You can also assume that the data is independent for this example.

Problem 1 (5 points)

Before fitting a model, the first step is to process your data. Preform any necessary processing steps here and briefly justify each step. .~~We are eventually going to try to determine which predictor variables have the largest effect on coaster speed.~~

Solution

```
coasters2 <- coasters %>% mutate(Inversions_cat = factor(Inversions), ## Step 1
                                 Track_cat = factor(Track))           ## Step 2
```

Full credit for students who note that the purpose of Steps 1 and 2 is to make sure that both categorical variables are recognized as factors in R.

Problem 2 (5 points)

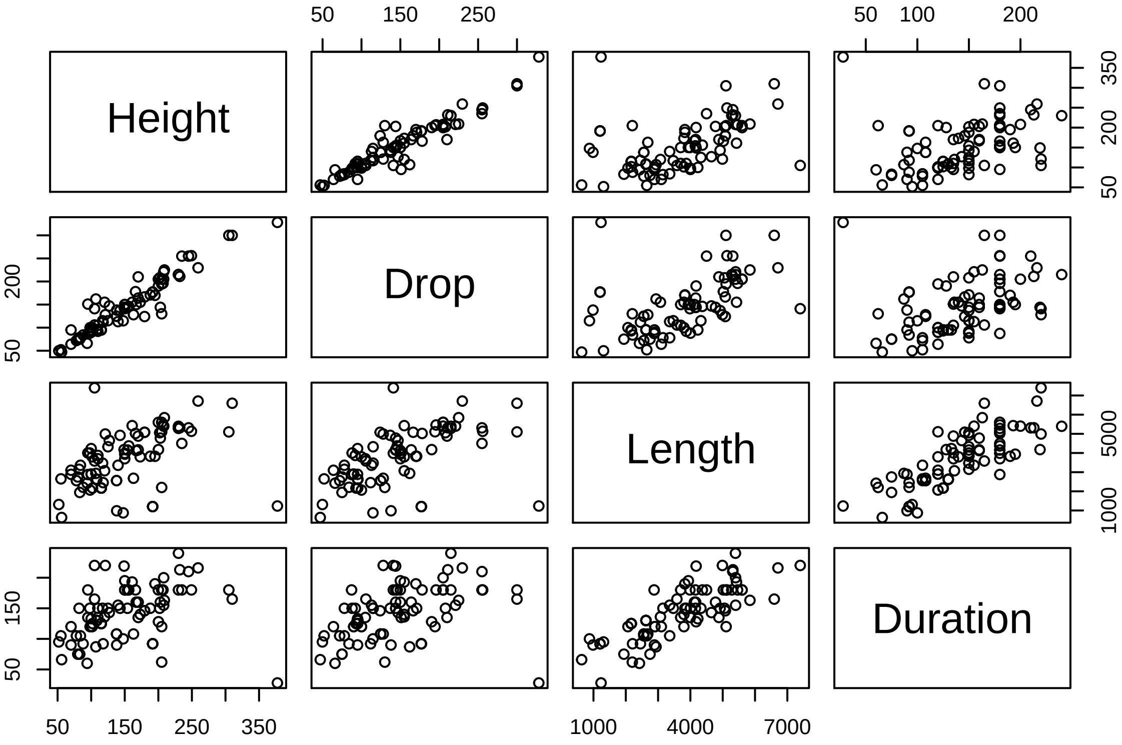
Create a scatter plot matrix for all of the quantitative variables and create box plots for each of the categorical variables (with coaster speed on the vertical axis). Does there appear to be any evidence of multicollinearity among the quantitative predictors?

Solution

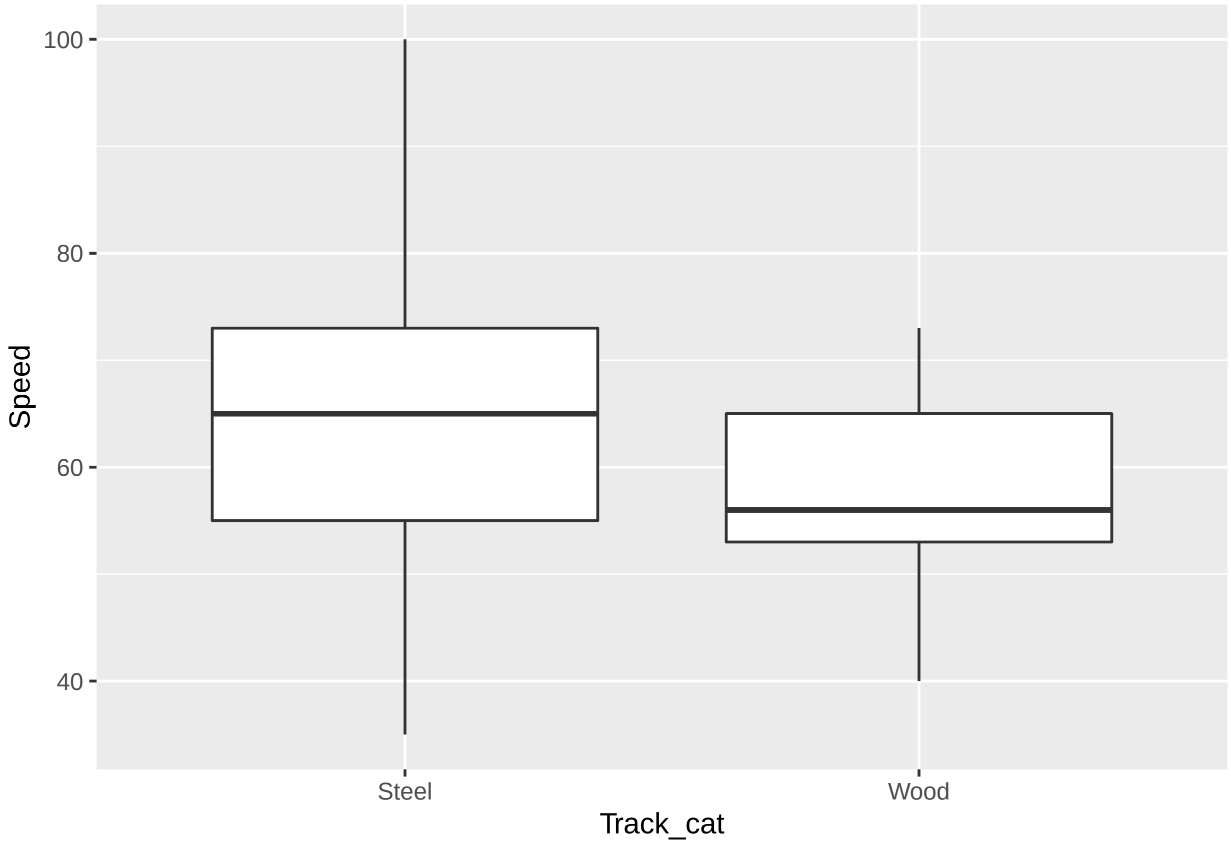
For full credit students must have all three plots below. In the matrix of scatterplots they must only display these four quantitative variables (do not give full credit if they included the variable Speed or the categorical predictors in the matrix scatter plot).

The matrix of scatter plots for the quantitative predictors shows strong evidence of multicolllinearity, especially between the variables Height and Drop. There is also a somewhat strong linear relationship between the variables Length and Duration. The other plots indicate a somewhat positive linear relationship but nowhere near as strong.

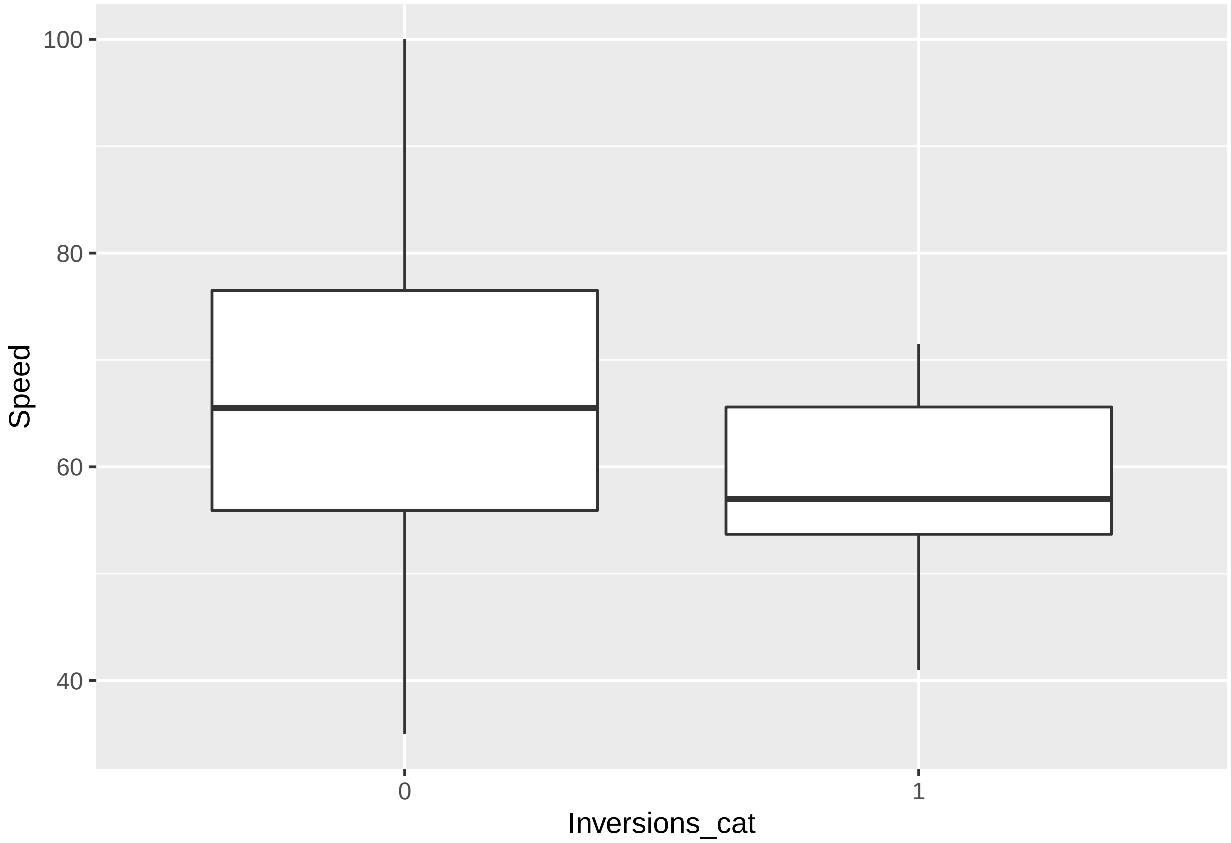
```
coasters2 %>% select(c(Height, Drop, Length, Duration)) %>% pairs
```



```
ggplot(coasters2, aes(x=Track_cat, y=Speed)) +
  geom_boxplot()
```



```
ggplot(coasters2, aes(x=Inversions_cat, y=Speed)) +
  geom_boxplot()
```



Problem 3 (5 points)

Write the equation for the estimated MLR model that includes all of the predictor variables. Make sure you clearly define all of your variables including any indicator variables.

Solution:

```
mod <- lm(Speed ~ Height + Drop + Length + Duration + Inversions_cat + Track_cat, coasters2)
```

$$\text{Average Speed} = 32.23 + 0.07Height + 0.12Drop + 0.002Length - 0.02Duration + 0.87Inversions + 1.99Track,$$

$$\text{where } Inversions = \begin{cases} 1, & \text{if ride has a loop/inversion} \\ 0, & \text{otherwise} \end{cases} \quad \text{and } Track = \begin{cases} 1, & \text{if ride is made of wood} \\ 0, & \text{if ride is made of steel} \end{cases}.$$

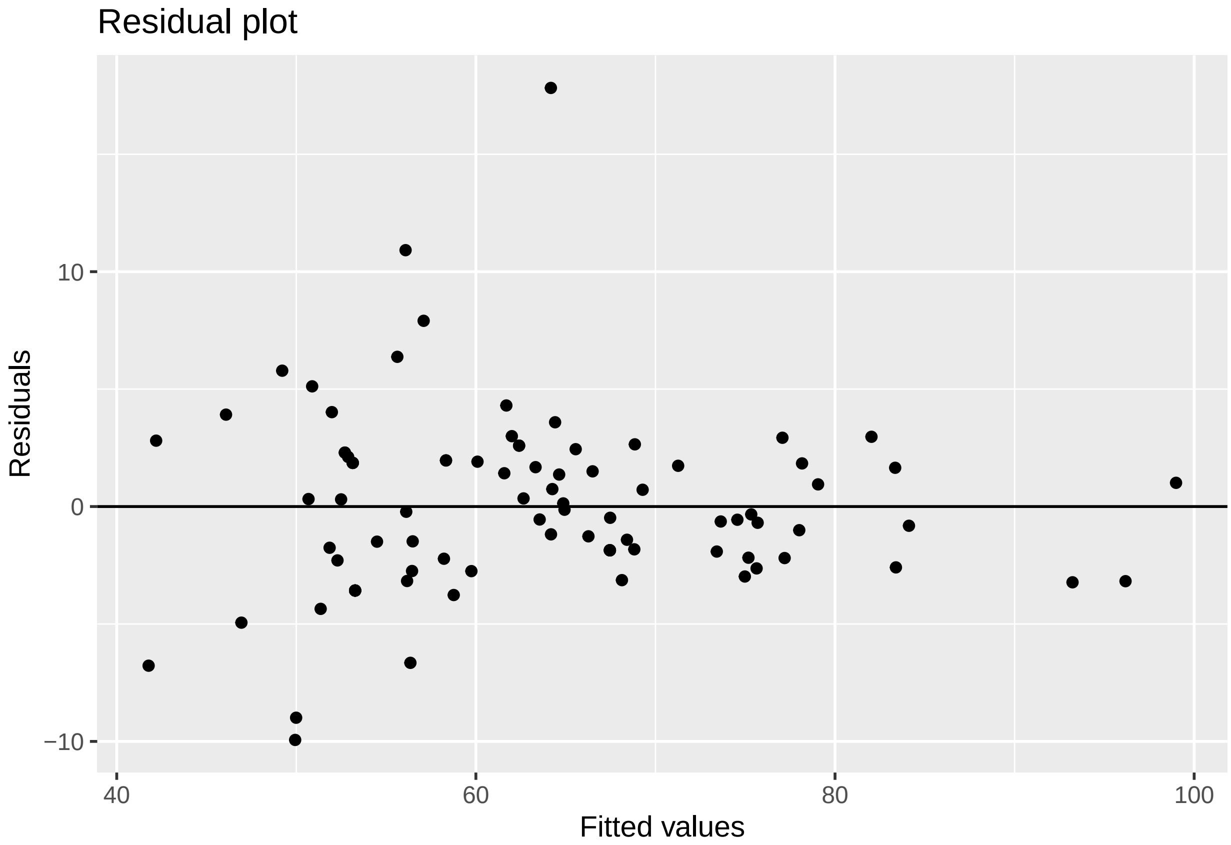
For full credit they must have the same coefficients (unless they standardized any of the quantitative variables), they must clearly define both indicator variables (what does 1 represent, what does 0 represent) and they must denote that the response variable is the (maximum) speed of the coaster but the regression equation predicts the average (maximum) speed.

Problem 4 (10 points)

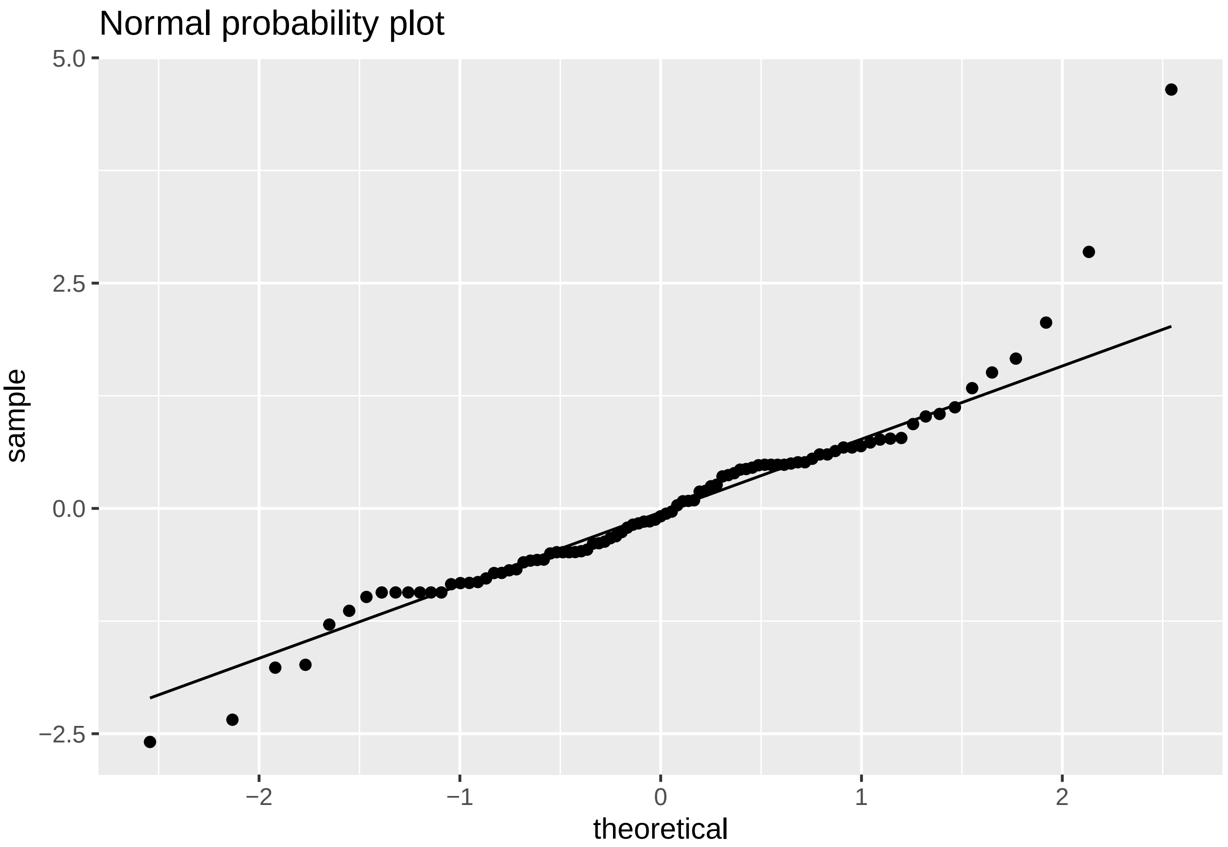
Before we can use the model from Problem 3, we need to check some assumptions by investigating the residual plot. Plot the standardized residuals against the fitted values and create a normal probability plot for the standardized residuals. What can you conclude about your estimated MLR model for this data set based on these two plots?

Solution:

```
coasters3 <- coasters2 %>% mutate(resids = mod$residuals,
                                  fits = mod$fitted.values)
ggplot(coasters3, aes(x=fits, y=resids)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  labs(title = "Residual plot", x="Fitted values", y="Residuals")
```



```
ggplot(coasters3, aes(sample=scale(resids))) +
  stat_qq() +
  stat_qq_line() +
  labs(title="Normal probability plot")
```



We are given that the data is a random, independent sample. We also don't really need to check the zero mean assumption (especially if the response variable has been standardized).

In the residual plot, we look for evidence of heteroskedasticity and any non-random trends. The residuals are pretty evenly and randomly scattered above and below the horizontal line at zero so the linear model is appropriate. The residuals do seem to change in spread about the line at zero however as the predicted response increases. In particular, the spread is larger for smaller values of speed and becomes more narrow as speed increases. So there is some indication of heteroskedasticity.

In the normal probability plot, we look for evidence that our residuals are not normally distributed. In the QQ-plot of the standardized residuals above, there is a strong deviation from normality especially in the upper right hand side of the plot. The upper quantiles of the residuals are much larger than we'd expect if they came from a normal distribution. This indicates that the distribution of the random error in the regression model is skewed left. The most important thing to note here is that it does not appear like the normality assumption is reasonable.

Problem 5 (20 points)

(a) In the summary output for this regression model there are two R-squared values provided. Which R-squared value should we use and why? What does this value represent (in the context of these data)? (10 points)

(b) What is $\hat{\sigma}$ based on this model? What does this number represent in the context of these data? (5 points)

(c) What is the average difference in the speed of a steel roller coaster versus a wooden roller coaster (given all other input variables are the same)? (5 points)

Solution:

(a) We should use the adjusted R squared value (not the multiple R squared value) because this mitigates the artificial increase in R-squared due to adding predictor variables. The R-squared value here represents the proportion of the variability in the maximum speeds of the roller coasers that can be explained by the predictor variables (Height, Drop, Length, Duration, Track materia, and Inversions). The adjusted R-squared value represents the same thing but with a small penalty term to adjust for the fact that there are multiple predictor variables.

(b)

```
summary(mod)
```

```
##
## Call:
## lm(formula = Speed ~ Height + Drop + Length + Duration + Inversions_cat +
##     Track_cat, data = coasters2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9368 -2.2575 -0.3345  1.9374 17.8231
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     32.2281890  2.3541045  13.690  < 2e-16 ***
## Height           0.0712878  0.0224073   3.181  0.00205 **
## Drop             0.1175544  0.0237273   4.954 3.71e-06 ***
## Length           0.0015250  0.0006109   2.496  0.01450 *
## Duration        -0.0210553  0.0176606  -1.192  0.23653
## Inversions_cat1  0.8710291  1.2144355   0.717  0.47522
## Track_catWood    1.9940675  1.4268725   1.398  0.16594
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.969 on 84 degrees of freedom
## Multiple R-squared:  0.9023, Adjusted R-squared:  0.8954
## F-statistic: 129.4 on 6 and 84 DF,  p-value: < 2.2e-16
```

```
summary(mod)$sigma
```

```
## [1] 3.969354
```

The spread in the maximum speed of the roller coasters is about 3.97 mph. (Must include units and have the correct value). They should also mention that this is not a very good estimate for the variability in coaster speed because of the heteroskedasticity (but this last part is only worth 1 point).

(c) Given that all other predictor variables are the same, the average difference in the maximum speed of a steel roller coaster and a wooden roller coaster is about 2 mph. (That is, the speed of a wooden roller coaster is about 2 mph faster than a steel roller coaster.)

Problem 6 (10 points)

Use the `filter` function to create a new data set that contains only the rows of data corresponding to roller coasters that do not have any inversions (i.e. loops).

(a) Write out the estimated regression equation. (4 points)

(b) Create a residual plot. (4 points)

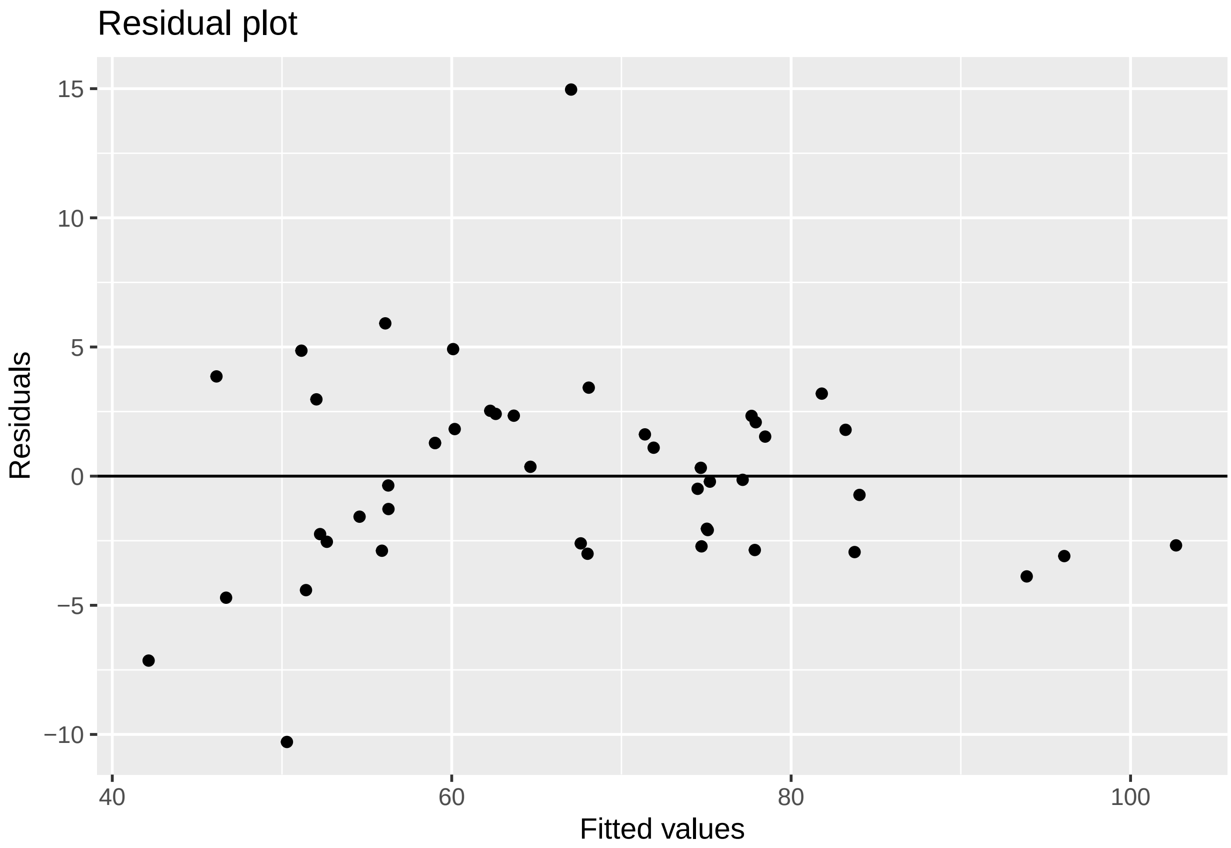
(c) Create a normal probability plot for the standardized residuals.(2 points)

Solution:

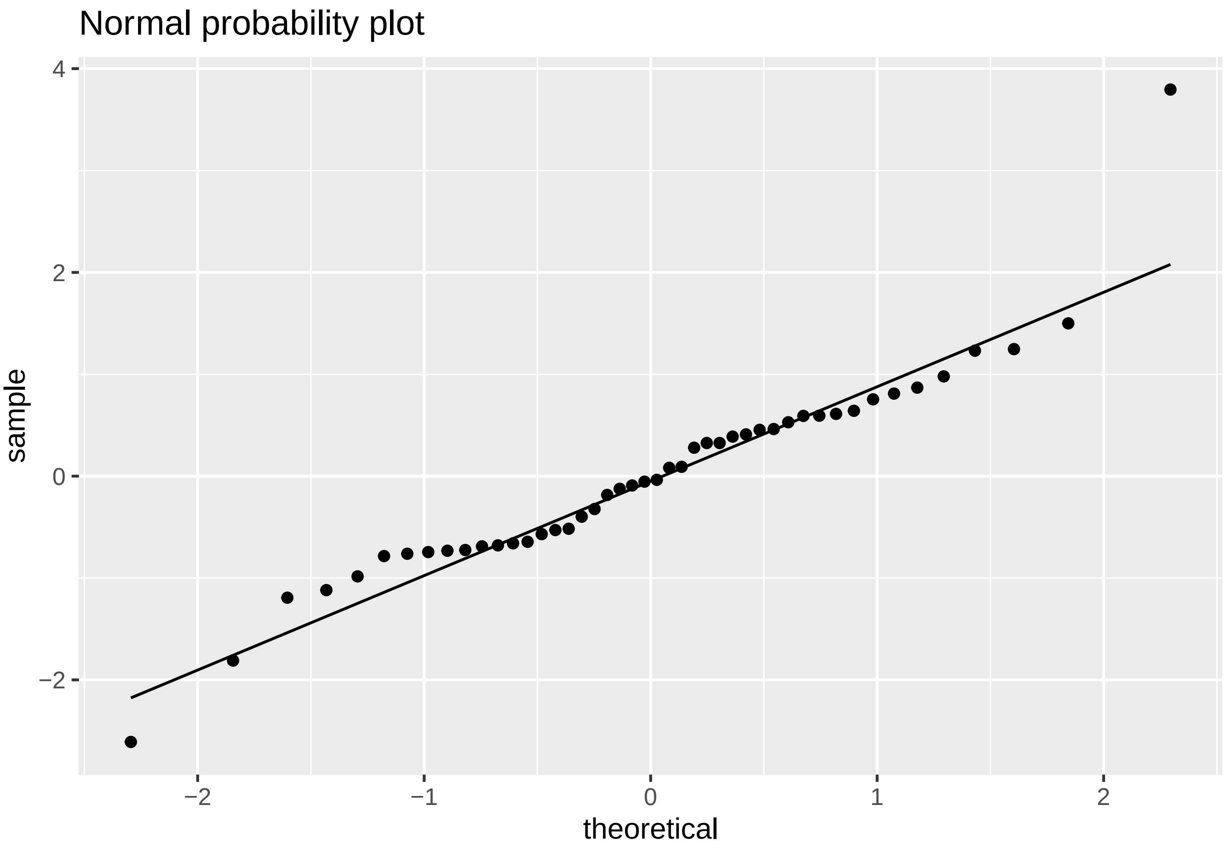
```
coasters_no_inversions <- coasters2 %>% filter(Inversions==0)
mod6 <- lm(Speed ~ Track_cat + Height + Drop + Length + Duration, coasters_no_inversions)
summary(mod6)
```

```
##
## Call:
## lm(formula = Speed ~ Track_cat + Height + Drop + Length + Duration,
##     data = coasters_no_inversions)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2906  -2.6613  -0.1781   2.2706  14.9653
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   32.3353164  3.0997286  10.432 5.63e-13 ***
## Track_catWood  2.3143358  1.7519074   1.321 0.193996
## Height         0.1070115  0.0296404   3.610 0.000842 ***
## Drop           0.0889787  0.0335266   2.654 0.011359 *
## Length         0.0009699  0.0008665   1.119 0.269683
## Duration      -0.0151751  0.0240875  -0.630 0.532277
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.183 on 40 degrees of freedom
## Multiple R-squared:  0.9255, Adjusted R-squared:  0.9161
## F-statistic: 99.33 on 5 and 40 DF,  p-value: < 2.2e-16
```

```
coasters_no_inversions2 <- coasters_no_inversions %>% mutate(resids = mod6$residuals,
                                                           fits = mod6$fitted.values)
ggplot(coasters_no_inversions2, aes(x=fits, y=resids)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  labs(title = "Residual plot", x="Fitted values", y="Residuals")
```



```r
ggplot(coasters_no_inversions2, aes(sample=scale(resids))) +
  stat_qq() +
  stat_qq_line() +
  labs(title="Normal probability plot")
```



$$AverageMaximumSpeed = 32.34 + 2.31Track + 0.11Height + 0.09Drop + 0.001Length - 0.02Duration,$$

where $Track = \begin{cases} 1, & \text{if track is made of wood} \\ 0, & \text{if track is made of steel} \end{cases}$ (same requirements as in Problem 3 for full credit for part (a))

Problem 7 (10 points)

Use the `filter` function to create a new data set that contains only the rows of data corresponding to roller coasters that do have inversions.

(a) Write out the estimated regression equation. (4 points)

(b) Create a residual plot. (4 points)

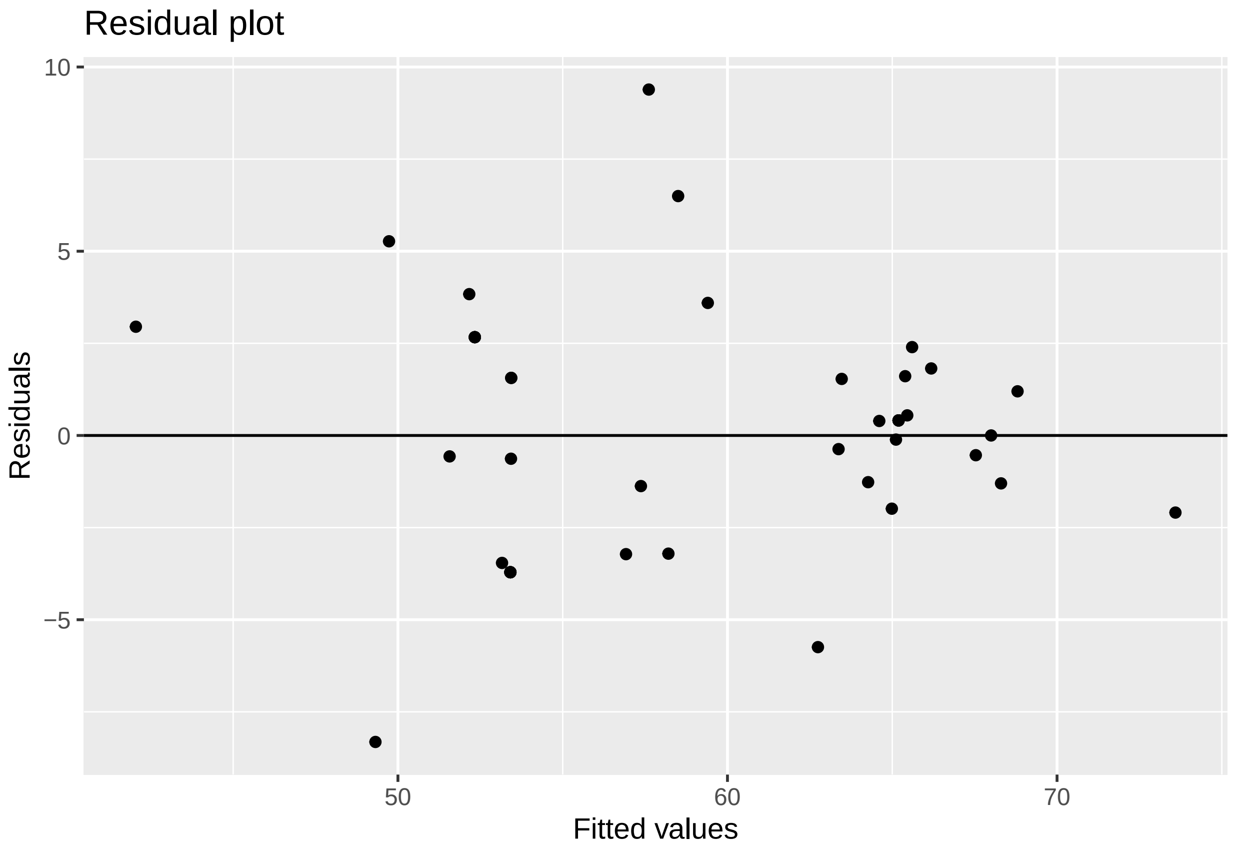
(c) Create a normal probability plot for the standardized residuals. (2 points)

Solution:

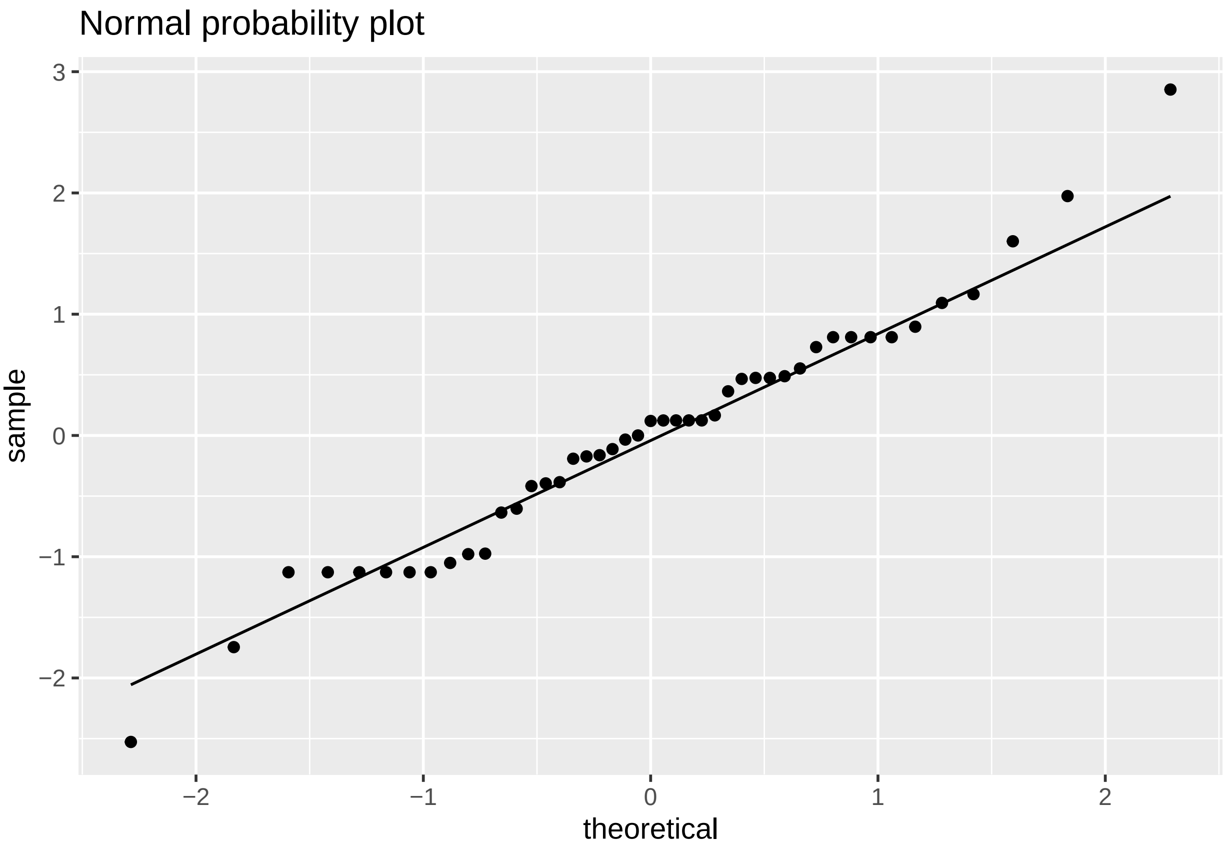
```
coasters_inversions <- coasters2 %>% filter(Inversions==1)
mod7 <- lm(Speed ~ Track_cat + Height + Drop + Length + Duration, coasters_inversions)
summary(mod7)
```

```
##
## Call:
## lm(formula = Speed ~ Track_cat + Height + Drop + Length + Duration,
##     data = coasters_inversions)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3180 -2.0926  0.3936  1.8182  9.3876
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   34.1578532  2.5129763  13.593 2.23e-16 ***
## Track_catWood  0.1475874  4.8479319   0.030    0.976
## Height        -0.0162332  0.0406539  -0.399    0.692
## Drop           0.1896081  0.0437540   4.334 9.98e-05 ***
## Length         0.0023152  0.0008669   2.671    0.011 *
## Duration      -0.0239620  0.0269478  -0.889    0.379
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.496 on 39 degrees of freedom
## Multiple R-squared:  0.8177, Adjusted R-squared:  0.7943
## F-statistic: 34.98 on 5 and 39 DF,  p-value: 2.065e-13
```

```
coasters_inversions2 <- coasters_inversions %>% mutate(resids = mod7$residuals,
                                                       fits = mod7$fitted.values)
ggplot(coasters_inversions2, aes(x=fits, y=resids)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  labs(title = "Residual plot", x="Fitted values", y="Residuals")
```



```
ggplot(coasters_inversions2, aes(sample=scale(resids))) +
  stat_qq() +
  stat_qq_line() +
  labs(title="Normal probability plot")
```



$$AverageMaximumSpeed = 34.16 + 0.15Track - 0.02Height + 0.19Drop + 0.002Length - 0.02Duration,$$

where $Track = \begin{cases} 1, & \text{if track is made of wood} \\ 0, & \text{if track is made of steel} \end{cases}$ (same requirements as in Problem 3 for full credit for part (a))

Problem 8 (Table 1 - 5 points, Table 2 - 10 points)

Based on the previous problems, fill out the following tables (by replacing the XX's with the correct numbers). The first table compares the coefficient of determination of the three different models from Problems 3, 6, and 7. The second table compares the confidence intervals and prediction intervals for each of the three models. For the second table, use the input values of:

- Track = Steel
- Height = 111 ft
- Drop = 95 ft
- Length = 2555 ft
- Duration = 75 s

Model	R squared	$\hat{\sigma}$
MLR - Problem 3	0.895	3.969
MLR - Problem 6	0.916	4.183
MLR - Problem 7	0.794	3.496

Model	CI for mean response	PI for new response
MLR - Problem 3	[52.31, 56.69] (Inversions)	[46.31, 62.69] (Inversions)
	[50.86, 56.39] (No Inversions)	[45.26, 61.99] (No Inversions)
MLR - Problem 6	[50.38, 57.64]	[44.81, 63.21]
MLR - Problem 7	[51.66, 57.32]	[46.87, 62.10]

Solution:

Note: if the student used a different confidence level (besides 0.95) that is ok, just make sure that they used the same level for ALL intervals and that the prediction intervals are all wider than the confidence intervals.

```
## To find the R-squared values
summary(mod)$adj.r.squared
```

```
## [1] 0.8953694
```

```
summary(mod6)$adj.r.squared
```

```
## [1] 0.9161456
```

```
summary(mod7)$adj.r.squared
```

```
## [1] 0.7942843
```

```
## To find sigma_hat
summary(mod)$sigma
```

```
## [1] 3.969354
```

```
summary(mod6)$sigma
```

```
## [1] 4.182835
```

```
summary(mod7)$sigma
```

```
## [1] 3.495534
```

```r
## To find intervals
## Problem 3
new_pt1 <- data.frame(Height = 111, Drop = 95, Length = 2555, Duration = 75,
                      Inversions_cat = "0", Track_cat = "Steel")
predict(mod, new_pt1, interval="confidence", level=0.95)
```

```
##        fit     lwr      upr
## 1 53.62607 50.8644 56.38774
```

```r
predict(mod, new_pt1, interval="predict", level=0.95)
```

```
##        fit      lwr      upr
## 1 53.62607 45.26341 61.98873
```

```r
new_pt2 <- data.frame(Height = 111, Drop = 95, Length = 2555, Duration = 75,
                      Inversions_cat = "1", Track_cat = "Steel")
predict(mod, new_pt2, interval="confidence", level=0.95)
```

```
##       fit      lwr     upr
## 1 54.4971 52.30869 56.6855
```

```r
predict(mod, new_pt2, interval="predict", level=0.95)
```

```
##       fit      lwr      upr
## 1 54.4971 46.30586 62.68833
```

```r
new_pt <- data.frame(Height = 111, Drop = 95, Length = 2555, Duration = 75,
                      Track_cat = "Steel")
## Problem 6
predict(mod6, new_pt, interval="confidence", level=0.95)
```

```
##        fit      lwr      upr
## 1 54.00645 50.37753 57.63537
```

```r
predict(mod6, new_pt, interval="predict", level=0.95)
```

```
##        fit      lwr      upr
## 1 54.00645 44.80665 63.20624
```

```r
## Problem 7
predict(mod7, new_pt, interval="confidence", level=0.95)
```

```
##        fit      lwr      upr
## 1 54.48688 51.65568 57.31808
```

```r
predict(mod7, new_pt, interval="predict", level=0.95)
```

```
##        fit      lwr      upr
## 1 54.48688 46.87071 62.10305
```

Problem 9 (20 points)

You've now analyzed this data with two different methods (and three different models). The question I'm sure you've been waiting for is… which analysis should we use and why? Using the information from the previous problems, should we use `Inversions` as a predictor variable in a single MLR model for this data OR should we separate the data and analyze the speed of roller coasters with and without loops separately OR does it even matter? Provide a definite, but succinct answer. (No more than 5 sentences.)

[Write your answer here.]

Extra Credit

If the response rate for the course evaluation is higher than 85%, everyone will get 2 points added to their grade for this test ;)