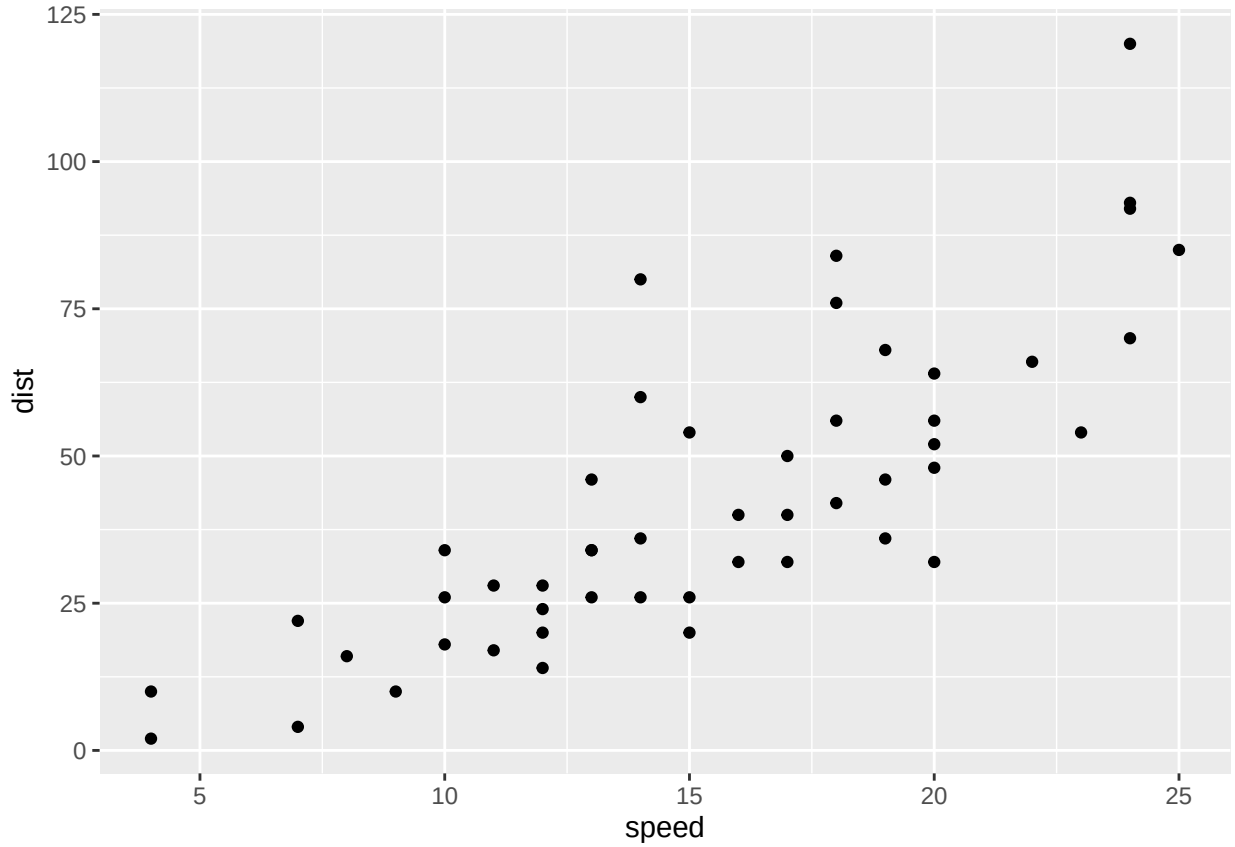
Module 3 Lab Submission

Ben Tankus

Consider the `cars` data, which contains cars speed in MPH and stopping distance in feet. Load the data with `data("cars")`.

```
data("cars")
qplot(speed, dist, data = cars)
```



- Fit a simple linear model with `dist` as the response and `speed` as the explanatory variable.

```
fit <- lm(dist ~ speed, data = cars)
summary(fit)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
```

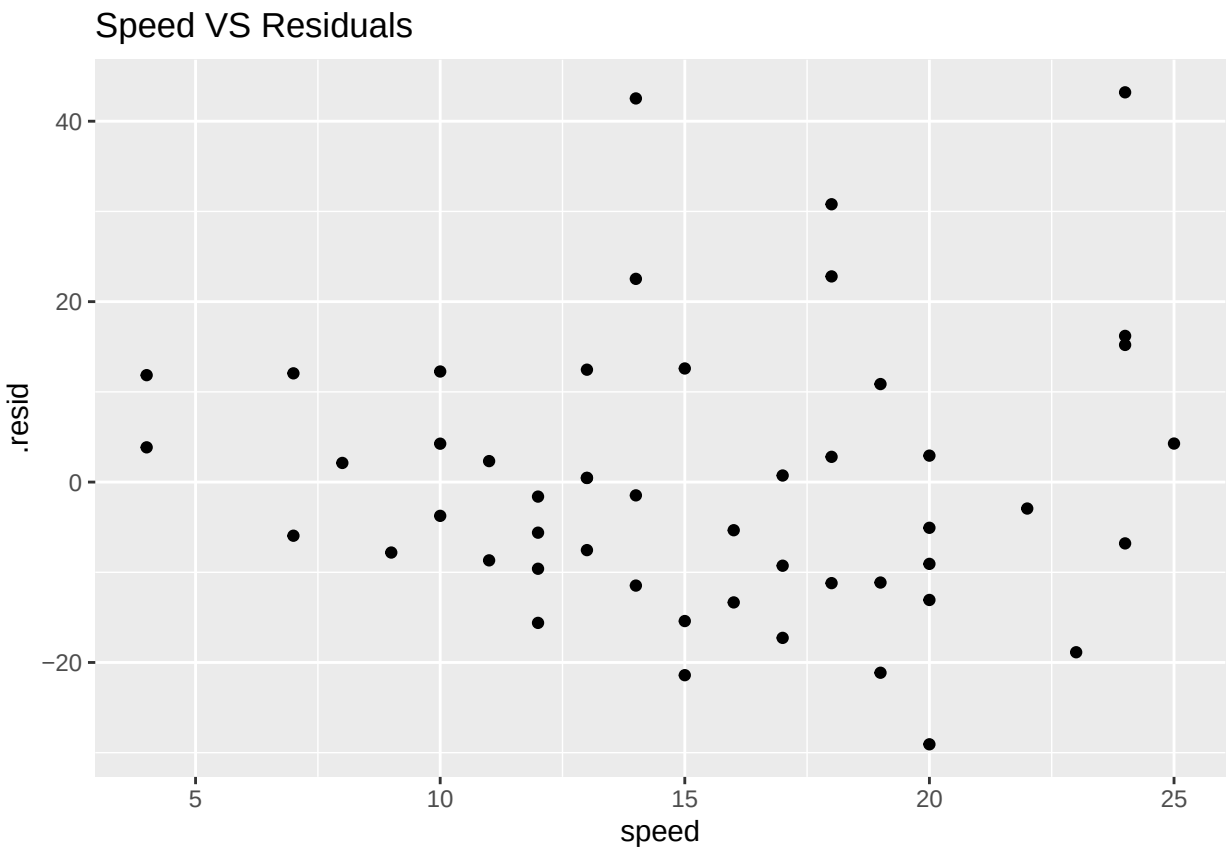
```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

- Create two diagnostic plots using the residuals, one with **speed** on the x-axis, and the other with the fitted values from the model. Do the plots look good: do these data seem to satisfy the assumptions for a linear regression model?

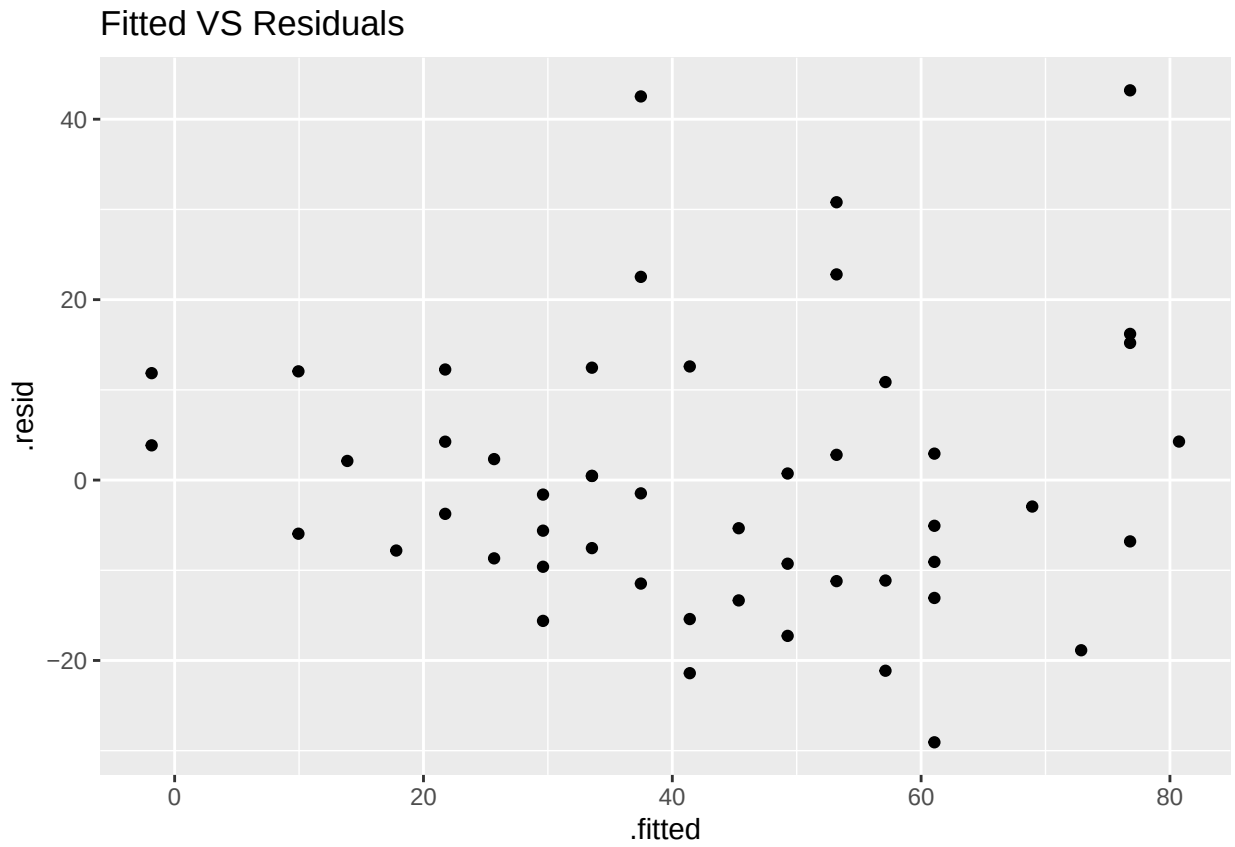
```
augFit <- augment(fit)

mean(fit$residuals)
```

```
## [1] 8.65974e-17
```

```
qplot(speed, .resid, data = augFit, main = 'Speed VS Residuals')
```




```
qplot(.fitted, .resid, data = augFit, main = 'Fitted VS Residuals')
```




```
augFit
```

```
## # A tibble: 50 x 8
##     dist speed .fitted .resid .std.resid   .hat .sigma  .cooksd
##    <dbl> <dbl>   <dbl>  <dbl>      <dbl>  <dbl>  <dbl>    <dbl>
##  1     2     4   -1.85   3.85      0.266 0.115    15.5 0.00459
##  2    10     4   -1.85  11.8       0.819 0.115    15.4 0.0435
##  3     4     7    9.95  -5.95     -0.401 0.0715   15.5 0.00620
##  4    22     7    9.95  12.1       0.813 0.0715   15.4 0.0255
##  5    16     8   13.9    2.12      0.142 0.0600   15.5 0.000645
##  6    10     9   17.8   -7.81     -0.521 0.0499   15.5 0.00713
##  7    18    10   21.7   -3.74     -0.249 0.0413   15.5 0.00133
##  8    26    10   21.7    4.26      0.283 0.0413   15.5 0.00172
##  9    34    10   21.7   12.3       0.814 0.0413   15.4 0.0143
## 10    17    11   25.7   -8.68     -0.574 0.0341   15.5 0.00582
## # ... with 40 more rows
```

I do not belive the data is good to analyse. The variance on the right is much higher than on the left (non-constant varience), and the r-squared value is only 0.65 which is quite low.This r-squared value means the model does not fit the data well.

- Use `predict()` to get the confidence and prediction intervals using the following new data.

```r
new <- data.frame(speed = c(6, 10.5, 14.7, 18.3, 21))
new
```

```
##   speed
## 1   6.0
## 2  10.5
## 3  14.7
## 4  18.3
## 5  21.0
```

```r
print('Prediction')
```

```
## [1] "Prediction"
```

```r
predict(fit, newdata = new, interval = 'prediction')
```

```
##         fit        lwr      upr
## 1  6.015358 -26.187314 38.21803
## 2 23.711197  -7.786388 55.20878
## 3 40.227314   8.991411 71.46322
## 4 54.383985  23.059721 85.70825
## 5 65.001489  33.422574 96.58040
```

```r
print('Confidence')
```

```
## [1] "Confidence"
```

```r
predict(fit, newdata = new, interval = 'confidence')
```

```
##         fit       lwr      upr
## 1  6.015358 -2.973341 15.00406
## 2 23.711197 17.720996 29.70140
## 3 40.227314 35.815250 44.63938
## 4 54.383985 49.384564 59.38341
## 5 65.001489 58.597384 71.40559
```

Now note that there are many speeds for which there were multiple observations at that speed. This means we can perform a lack-of-fit test on this data.

- Fit a separate means model using `lm()` and `factor()` to treat `speed` as a categorical variable.

```r
fitSSM <- lm(dist ~ factor(speed), data = cars )
summary(fitSSM)
```

```
##
## Call:
## lm(formula = dist ~ factor(speed), data = cars)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -24.5000  -7.4583  -0.3333   6.2750  29.5000
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)         6.00      10.45   0.574 0.569837
## factor(speed)7      7.00      14.77   0.474 0.638919
## factor(speed)8     10.00      18.09   0.553 0.584416
## factor(speed)9      4.00      18.09   0.221 0.826473
## factor(speed)10    20.00      13.48   1.483 0.148140
## factor(speed)11    16.50      14.77   1.117 0.272594
## factor(speed)12    15.50      12.79   1.212 0.234825
## factor(speed)13    29.00      12.79   2.267 0.030525 *
## factor(speed)14    44.50      12.79   3.478 0.001518 **
## factor(speed)15    27.33      13.48   2.027 0.051345 .
## factor(speed)16    30.00      14.77   2.031 0.050923 .
## factor(speed)17    34.67      13.48   2.571 0.015172 *
## factor(speed)18    58.50      12.79   4.573 7.28e-05 ***
## factor(speed)19    44.00      13.48   3.263 0.002686 **
## factor(speed)20    44.40      12.36   3.592 0.001117 **
## factor(speed)22    60.00      18.09   3.316 0.002334 **
## factor(speed)23    48.00      18.09   2.653 0.012465 *
## factor(speed)24    87.75      12.79   6.859 1.09e-07 ***
## factor(speed)25    79.00      18.09   4.367 0.000131 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.77 on 31 degrees of freedom
## Multiple R-squared:  0.7921, Adjusted R-squared:  0.6714
## F-statistic: 6.562 on 18 and 31 DF,  p-value: 2.846e-06
```

- Compare the separate means model to the simple linear regression model using the `anova()` function.

```
anova(fit, fitSSM)
```

```
## Analysis of Variance Table
##
## Model 1: dist ~ speed
## Model 2: dist ~ factor(speed)
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1     48 11353.5
## 2     31  6764.8 17    4588.7 1.2369 0.2948
```

Based on the results of this lack-of-fit F-test, we would fail to reject the null hypothesis that the linear model is adequate (p-value = 0.29, F-statistic = 1.24 on 17 and 31 degrees of freedom). We would therefore conclude that there is no departure from linearity in the relationship between Distance and speed.