STAT-S631

Assignment 10

John Koo

```r
dp <- loadNamespace('dplyr')
import::from(magrittr, `%>%`, `%<>%`)
import::from(car, Anova, ncvTest)
library(ggplot2)
theme_set(theme_bw())
```

Problem 1

```r
# get the data
robey.df <- read.table('~/dev/stats-hw/stat-s631/Robey.txt') %>%
  dp$mutate(country = rownames(.))

# full model with region first
model.1 <- lm(tfr ~ region * contraceptors, data = robey.df)
summary(model.1)
```

```
Call:
lm(formula = tfr ~ region * contraceptors, data = robey.df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.54546 -0.26527 -0.04661  0.34689  1.30579

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                    6.832351   0.194090  35.202  < 2e-16 ***
regionAsia                    -0.322375   0.563627  -0.572    0.570
regionLatin.Amer              -0.237356   0.520948  -0.456    0.651
regionNear.East                0.631733   0.632999   0.998    0.324
contraceptors                 -0.054099   0.007718  -7.009 1.41e-08 ***
regionAsia:contraceptors      -0.003795   0.012389  -0.306    0.761
regionLatin.Amer:contraceptors  0.003136   0.012044   0.260    0.796
regionNear.East:contraceptors -0.013920   0.016141  -0.862    0.393
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5732 on 42 degrees of freedom
Multiple R-squared:  0.8667,    Adjusted R-squared:  0.8445
F-statistic: 39.01 on 7 and 42 DF,  p-value: < 2.2e-16
```

```r
anova(model.1)
```

```
Analysis of Variance Table

Response: tfr
```

```
                     Df Sum Sq Mean Sq  F value    Pr(>F)
region                3 44.304  14.768  44.9534 3.576e-13 ***
contraceptors         1 45.045  45.045 137.1158 8.226e-15 ***
region:contraceptors  3  0.365   0.122   0.3706    0.7746
Residuals            42 13.798   0.329
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
Anova(model.1)
```

```
Anova Table (Type II tests)

Response: tfr
                     Sum Sq Df  F value    Pr(>F)
region                1.677  3   1.7018    0.1812
contraceptors        45.045  1 137.1158 8.226e-15 ***
region:contraceptors  0.365  3   0.3706    0.7746
Residuals            13.798 42
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# full model with contraception rate first
model.2 <- lm(tfr ~ contraceptors * region, data = robey.df)
summary(model.2)
```

```

Call:
lm(formula = tfr ~ contraceptors * region, data = robey.df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.54546 -0.26527 -0.04661  0.34689  1.30579

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     6.832351   0.194090  35.202  < 2e-16 ***
contraceptors                  -0.054099   0.007718  -7.009 1.41e-08 ***
regionAsia                     -0.322375   0.563627  -0.572    0.570
regionLatin.Amer               -0.237356   0.520948  -0.456    0.651
regionNear.East                 0.631733   0.632999   0.998    0.324
contraceptors:regionAsia       -0.003795   0.012389  -0.306    0.761
contraceptors:regionLatin.Amer  0.003136   0.012044   0.260    0.796
contraceptors:regionNear.East  -0.013920   0.016141  -0.862    0.393
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5732 on 42 degrees of freedom
Multiple R-squared:  0.8667,    Adjusted R-squared:  0.8445
F-statistic: 39.01 on 7 and 42 DF,  p-value: < 2.2e-16
```

```r
anova(model.2)
```

```
Analysis of Variance Table

Response: tfr
                     Df Sum Sq Mean Sq  F value Pr(>F)
```

```
contraceptors         1 87.672  87.672 266.8706 <2e-16 ***
region                3  1.677   0.559   1.7018 0.1812
contraceptors:region  3  0.365   0.122   0.3706 0.7746
Residuals            42 13.798   0.329
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
Anova(model.2)
```

```
Anova Table (Type II tests)

Response: tfr
                     Sum Sq Df  F value    Pr(>F)
contraceptors        45.045  1 137.1158 8.226e-15 ***
region                1.677  3   1.7018    0.1812
contraceptors:region  0.365  3   0.3706    0.7746
Residuals            13.798 42
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Part a

Here, for simplicity, we will just subscript in the order of the model. So for example, β_1 corresponds to `regionAsia`, β_4 corresponds to `contraceptors`, β_7 corresponds to `regionNear.East:contraceptors`, etc.

Type I anova (`anova(model.1)`)

This test is sequential. So the first line tests $H_0 : \beta_i = 0$ for $0 < i \leq 7$ and $H_A : \beta_1 \neq 0$ and $\beta_2 \neq 0$ and $\beta_3 \neq 0$ since these all fall under `region`. In other words, we want to know if `region` adds anything to an intercept-only model. Since p is small, we can conclude that it does.

The second line tests $H_0 : \beta_i = 0$ for $3 < i \leq 7$ and $H_A : \beta_i \neq 0$ for all of $3 < i \leq 4$. In other words, given a model that just uses `region` and `contraceptors` without the interaction, does `contraceptors` add significant predictive power to the model? Since p is small, we can conclude that a model with both (but no interaction term) is significantly different from a model that just uses `region`.

The third line tests $H_0 : \beta_i = 0$ for $4 < i \leq 7$ and $H_A : \beta_i \neq 0$ for all $4 < i \leq 7$. In other words, we want to know if the interaction terms add anything to a model without interaction terms. Since p is close to 1, we can say that it does not.

Type II anova (`Anova(model.1)`)

In the first line, we test if a model containing `region` and the interaction term is significantly different from the full model. Here we remove the interaction term per the marginality principle. So H_0 : `tfr` $\sim$ `contraceptors` and H_A : `tfr` $\sim$ `contraceptors` + `region` + `contraceptors:region`. Since p is large, we can say that the alternative model does not significantly add to the null model.

In the second line, we switch `contraceptors` and `region`. Since p is small, we can say that `contraceptors` + `region:contraceptors` does add to a model with just `region`.

In the third line, we test the non-interaction "parallel" model vs. the full model with interactions. H_0 : `tfr` $\sim$ `region` + `contraceptors`, and H_A : `tfr` $\sim$ `region` + `contraceptors` + `region:contraceptors`. Since p is large, we can fail to reject the null hypothesis and say that there is no significant addition to the model by adding the interaction term.

Part b

Type II anova is not sequential. It looks at each term separately (other than the marginality principle). So the order in which the covariates are written in the model call does not matter. On the other hand, type I anova is sequential so the order makes an effect. In both models, the same set of covariates are used, so the type II anova does not change, but since the order in which they are listed changes, the type I anova results differ.

Part c

From the type II anova tests (doesn't matter which one), we might say that `region` does not make a difference in the model (i.e., `contraceptors` already explains all of the variance in `tfr` that `region` can). So a model we might consider is:

```
model.3 <- lm(tfr ~ contraceptors, data = robey.df)
summary(model.3)
```

```
Call:
lm(formula = tfr ~ contraceptors, data = robey.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.5493 -0.3013  0.0254  0.3957  1.2021 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    6.875085   0.156860   43.83   <2e-16 ***
contraceptors -0.058416   0.003584  -16.30   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5745 on 48 degrees of freedom
Multiple R-squared:  0.847, Adjusted R-squared:  0.8438 
F-statistic: 265.7 on 1 and 48 DF,  p-value: < 2.2e-16
```

```
anova(model.3, model.1)
```

```
Analysis of Variance Table

Model 1: tfr ~ contraceptors
Model 2: tfr ~ region * contraceptors
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     48 15.840                           
2     42 13.798  6    2.0425 1.0362 0.4158
```

So we can conclude that `region` does not add to a model with `contraceptors`. This is the same result we obtained in Homework Assignment 8, where we made this decision largely based on visualizations of the data.

Problem 2

Note that $[W]_{ij} = 0 \; \forall i \neq j$, and $[W]_{ii} > 0 \; \forall i$ such that $0 < i \leq n$. Furthermore, $[W^{1/2}]_{ij} = \sqrt{[W]_{ij}} \; \forall i, j \leq n$. Then since W and $W^{1/2}$ are diagonal matricies, they are symmetric, i.e., $W^T = W$ and $(W^{1/2})^T = W^{1/2}$.

Since $W^{1/2}$ is a diagonal matrix, $W^{1/2}W^{1/2}$ is also diagonal. Note that $[W^{1/2}W^{1/2}]_{ij} = \sum_k \sqrt{w_{ik}w_{kj}} = \sqrt{w_{ii}w_{ij}}$ where $w_{ij} = [W]_{ij}$ and the other terms in the sum are zero since $w_{ij} = 0$ when $i \neq j$. Then $[W^{1/2}W^{1/2}]_{ij} = 0$ when $i \neq j$ and $[W^{1/2}W^{1/2}]_{ii} = w_{ii}$. Then $W^{1/2}W^{1/2} = W$.

Since $Y|X \sim \mathcal{N}(X\beta, \sigma^2 W^{-1})$, $\hat{\beta} = (X^T W X)^{-1} X^T W Y$. But if we start with the claim $\hat{\beta} = (X^{*T}X^*)^{-1}X^{*T}Y^*$ where $X^* = W^{1/2}X$ and $Y^* = W^{1/2}Y$:

$$
\begin{aligned}
&(X^{*T}X^*)^{-1}X^{*T}Y^* \\
&= ((W^{1/2}X)^T W^{1/2}X)^{-1}(W^{1/2}X)^T W^{1/2}Y \\
&= (X^T W^{1/2}W^{1/2}X)^{-1}X^T W^{1/2}W^{1/2}Y \\
&= (X^T W X)^{-1}X^T W Y
\end{aligned}
$$

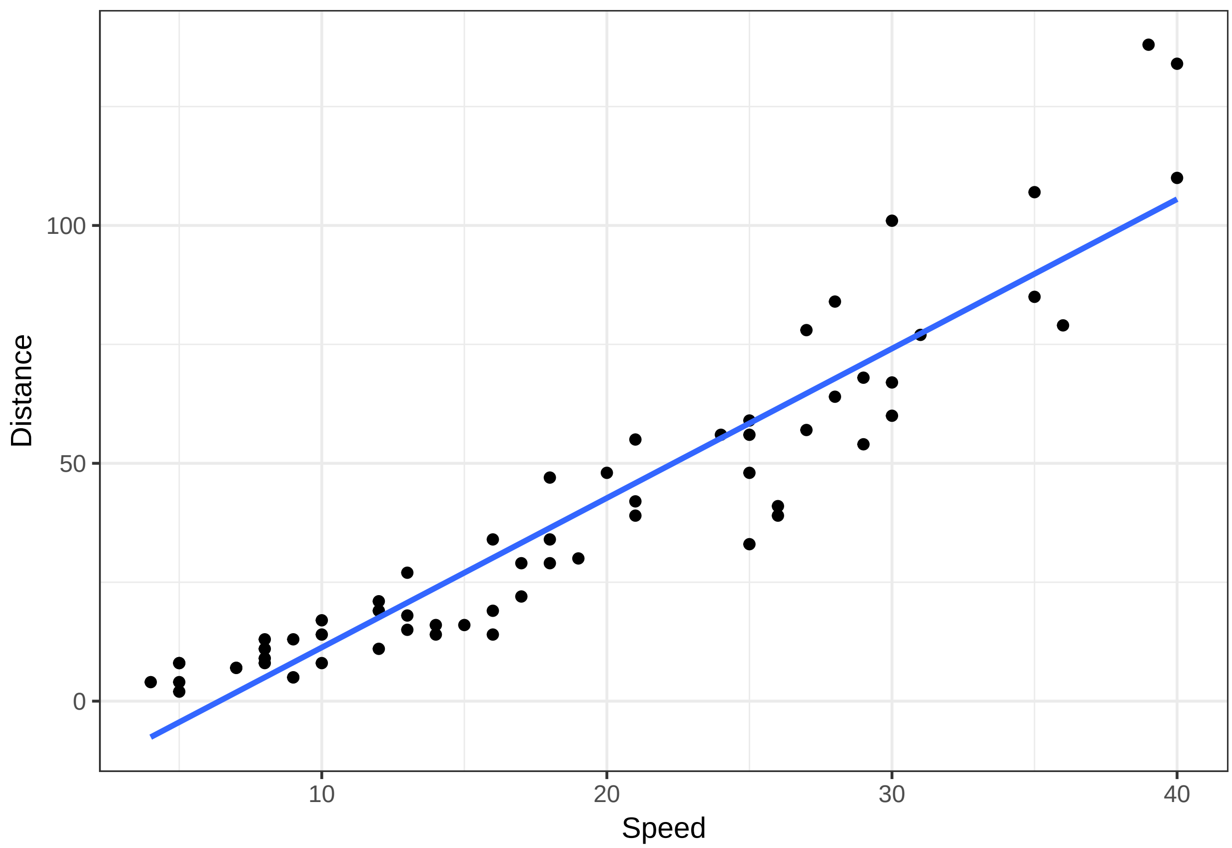
Which is just our WLS estimator for $\hat{\beta}$.

Problem 3

[From ALR 7.6]

Part 1

```r
stopping.df <- alr4::stopping

ggplot(stopping.df) +
  geom_point(aes(x = Speed, y = Distance)) +
  stat_smooth(aes(x = Speed, y = Distance),
              se = FALSE, method = 'lm')
```



E[`Distance` | `Speed`] appears to change as we move along x. From the plot, we can see that it curves along the OLS line, starting above it, dipping below, and then increasing above it again.

Based on this scatterplot and our intuition of the problem, I believe it makes sense to exclude the intercept term. If a car is not moving (i.e., `Speed` $= 0$), then it cannot have a stopping distance, so `Distance` $= 0$, which eliminates the intercept term.

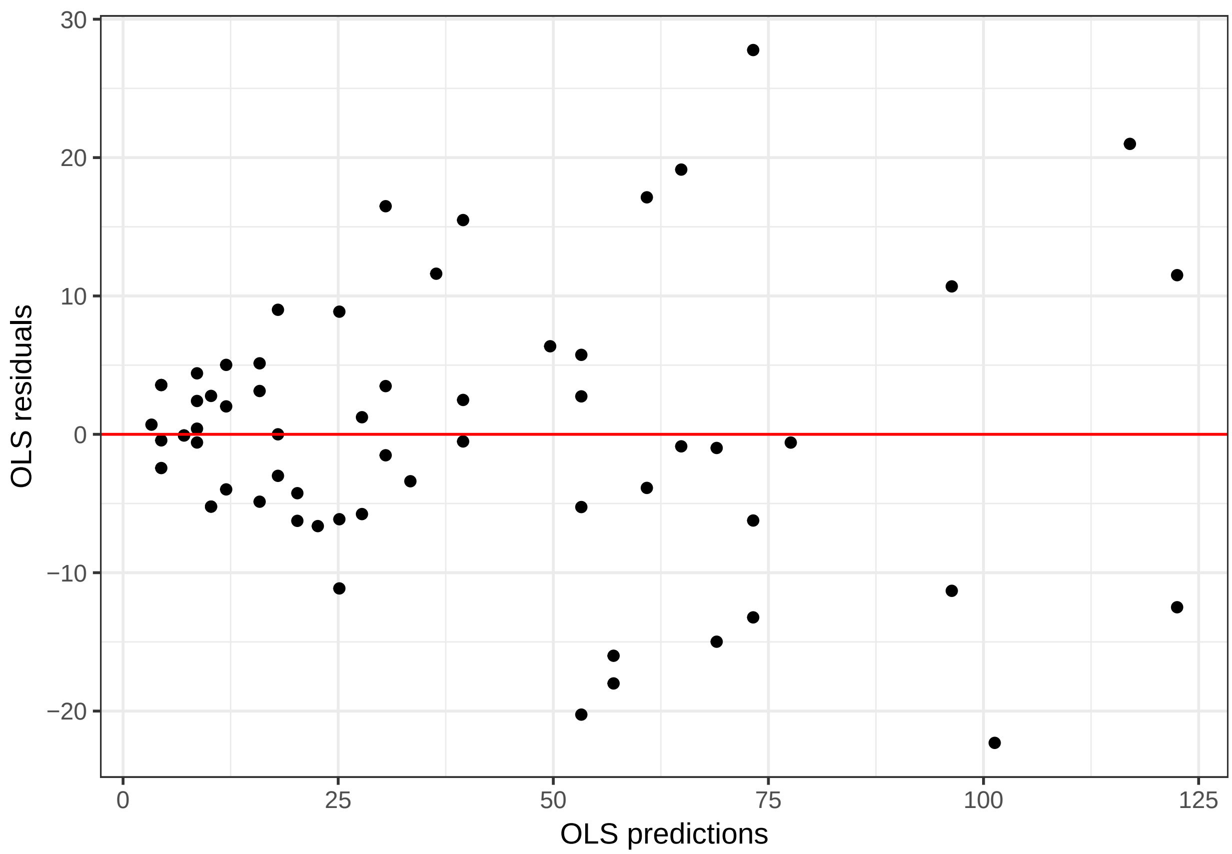
Part 2

```r
# const.var.mod <- lm(Distance ~ Speed + I(Speed ** 2), data = stopping.df)
# const.var.mod <- lm(Distance ~ I(Speed ** 2) - 1, data = stopping.df)
const.var.mod <- lm(Distance ~ Speed + I(Speed ** 2) - 1, data = stopping.df)
summary(const.var.mod)
```

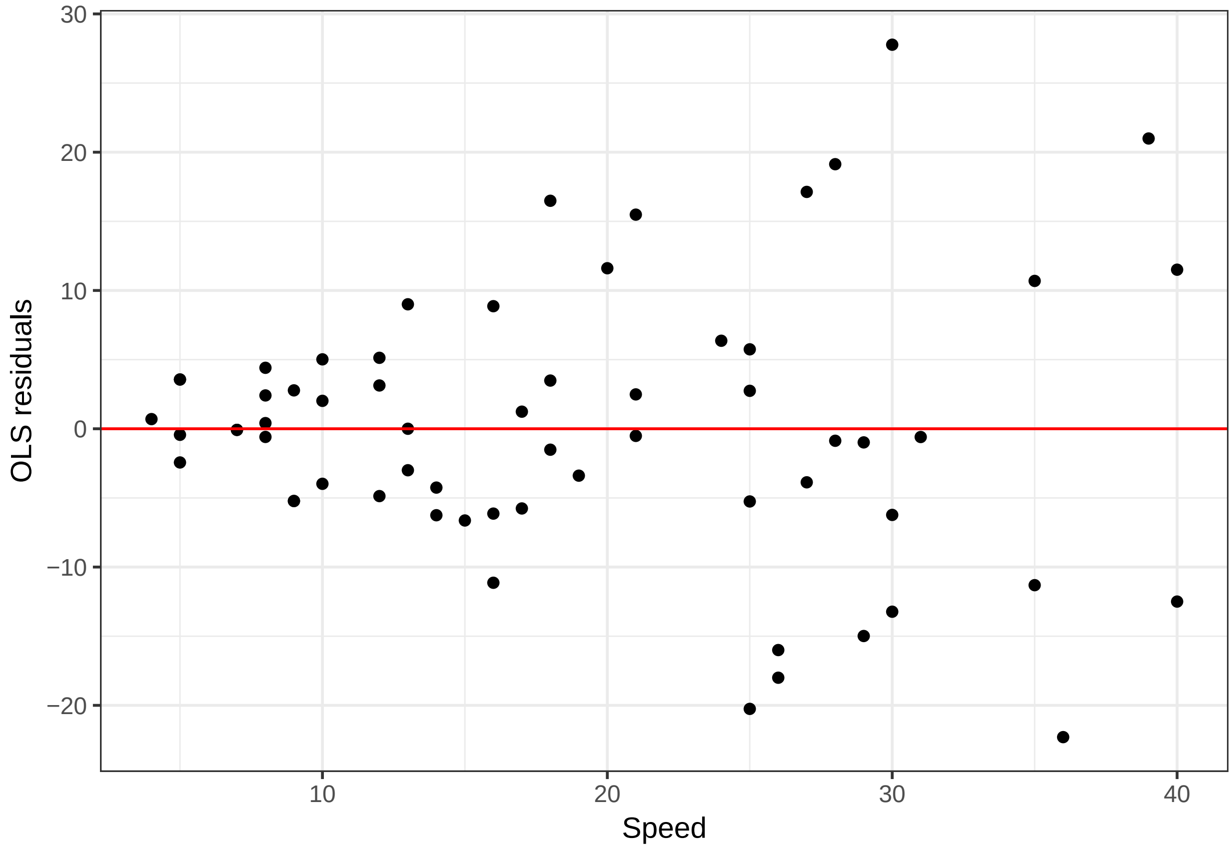
```
Call:
lm(formula = Distance ~ Speed + I(Speed^2) - 1, data = stopping.df)

Residuals:
    Min      1Q  Median      3Q     Max
-22.298  -5.223  -0.259   4.198  27.771

Coefficients:
           Estimate Std. Error t value Pr(>|t|)
Speed      0.576599   0.200804   2.871  0.00564 **
I(Speed^2) 0.062145   0.006904   9.001 9.83e-13 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.852 on 60 degrees of freedom
Multiple R-squared:  0.9644,    Adjusted R-squared:  0.9632
F-statistic: 813.5 on 2 and 60 DF,  p-value: < 2.2e-16
```

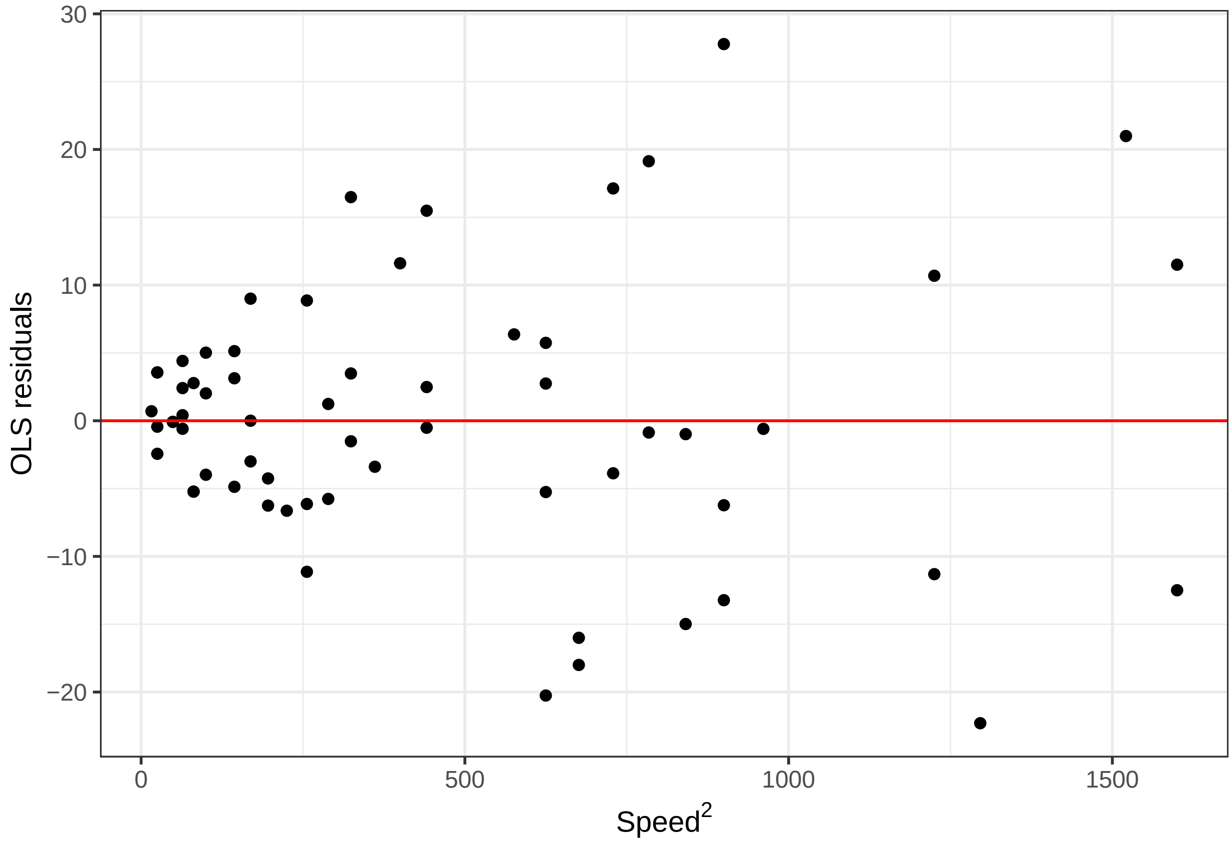
```
# add the predictions to the data
stopping.df %<>%
  dp$mutate(distance.hat = predict(const.var.mod, stopping.df)) %>%
  dp$mutate(resid.ols = Distance - distance.hat)

ggplot(stopping.df) +
  geom_point(aes(x = distance.hat, y = resid.ols)) +
  geom_abline(slope = 0, colour = 'red') +
  labs(x = 'OLS predictions', y = 'OLS residuals')
```




```
ggplot(stopping.df) +
  geom_point(aes(x = Speed, y = resid.ols)) +
  geom_abline(slope = 0, colour = 'red') +
  labs(y = 'OLS residuals')
```




```
ggplot(stopping.df) +
  geom_point(aes(x = Speed ** 2, y = resid.ols)) +
  geom_abline(slope = 0, colour = 'red') +
  labs(y = 'OLS residuals', x = expression(Speed^2))
```




```r
# part a
fv.test <- ncvTest(const.var.mod)
fv.test
```

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 23.20677    Df = 1     p = 1.454844e-06
```

```r
# part b
speed.test <- ncvTest(const.var.mod, ~ Speed)
speed.test
```

```
Non-constant Variance Score Test
Variance formula: ~ Speed
Chisquare = 23.47833    Df = 1     p = 1.263289e-06
```

```r
# part c
speed.speed2.test <- ncvTest(const.var.mod, ~ Speed + I(Speed ** 2))
speed.speed2.test
```

```
Non-constant Variance Score Test
Variance formula: ~ Speed + I(Speed^2)
Chisquare = 23.57714    Df = 2     p = 7.590831e-06
```

```r
# difference between b and c?
1 - pchisq(speed.speed2.test$ChiSquare - speed.test$ChiSquare, 1)
```

```
[1] 0.7532587
```

We do not gain any significant information by including both `Speed` and `Speed`^{2} compared to just `Speed`.

Part 3

```
# build the WLS model
speed.var.mod <- lm(Distance ~ Speed + I(Speed ** 2) - 1,
                    weights = Speed ** -1,
                    data = stopping.df)
summary(speed.var.mod)
```

```

Call:
lm(formula = Distance ~ Speed + I(Speed^2) - 1, data = stopping.df,
    weights = Speed^-1)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-4.0703 -1.4275 -0.1057  1.3737  5.0866

Coefficients:
           Estimate Std. Error t value Pr(>|t|)
Speed      0.614461   0.158903   3.867 0.000274 ***
I(Speed^2) 0.060784   0.006141   9.898 3.14e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.997 on 60 degrees of freedom
Multiple R-squared:  0.958, Adjusted R-squared:  0.9566
F-statistic: 683.6 on 2 and 60 DF,  p-value: < 2.2e-16
```

```
summary(const.var.mod)$coefficients
```

```
             Estimate  Std. Error  t value     Pr(>|t|)
Speed      0.57659878 0.200803820 2.871453 5.638438e-03
I(Speed^2) 0.06214515 0.006904165 9.001110 9.828141e-13
```

```
summary(speed.var.mod)$coefficients
```

```
             Estimate  Std. Error  t value     Pr(>|t|)
Speed      0.61446093 0.158903019 3.866893 2.737223e-04
I(Speed^2) 0.06078404 0.006140893 9.898240 3.143928e-14
```

Note that now our estimate for $\hat{\beta} = (X^T W X)^{-1} X^T W Y$ and $var(\hat{\beta}|X) = \sigma^2 (X^T X)^{-1} X^T W^{-1} X (X^T X)^{-1}$. w_{ii} is small when both x_{i1} and x_{i2} are large.

Part 4

```
# weight matrix
# it's diagonal since we only have one covariate
W <- diag(stopping.df$Speed ** -1)

# model matrix
X <- model.matrix(~ Speed + I(Speed**2) - 1, data = stopping.df)

# response
Y <- stopping.df$Distance
```

```r
# projection matrix
H <- X %*% solve(t(X) %*% X) %*% t(X)

# sandwich term
S <- t(X) %*% diag(residuals(const.var.mod) ** 2 / (1 - diag(H)) ** 2) %*% X

# copute standard errors
var.beta.hat <- solve(t(X) %*% X) %*% S %*% solve(t(X) %*% X)
diag(sqrt(var.beta.hat))
```

```
      Speed  I(Speed^2)
0.216683874 0.009022654
```

```r
# equivalent
diag(sqrt(sandwich::vcovHC(const.var.mod)))
```

```
      Speed  I(Speed^2)
0.216683874 0.009022654
```

We end up with much larger standard errors for $var(\hat{\beta}|X)$.