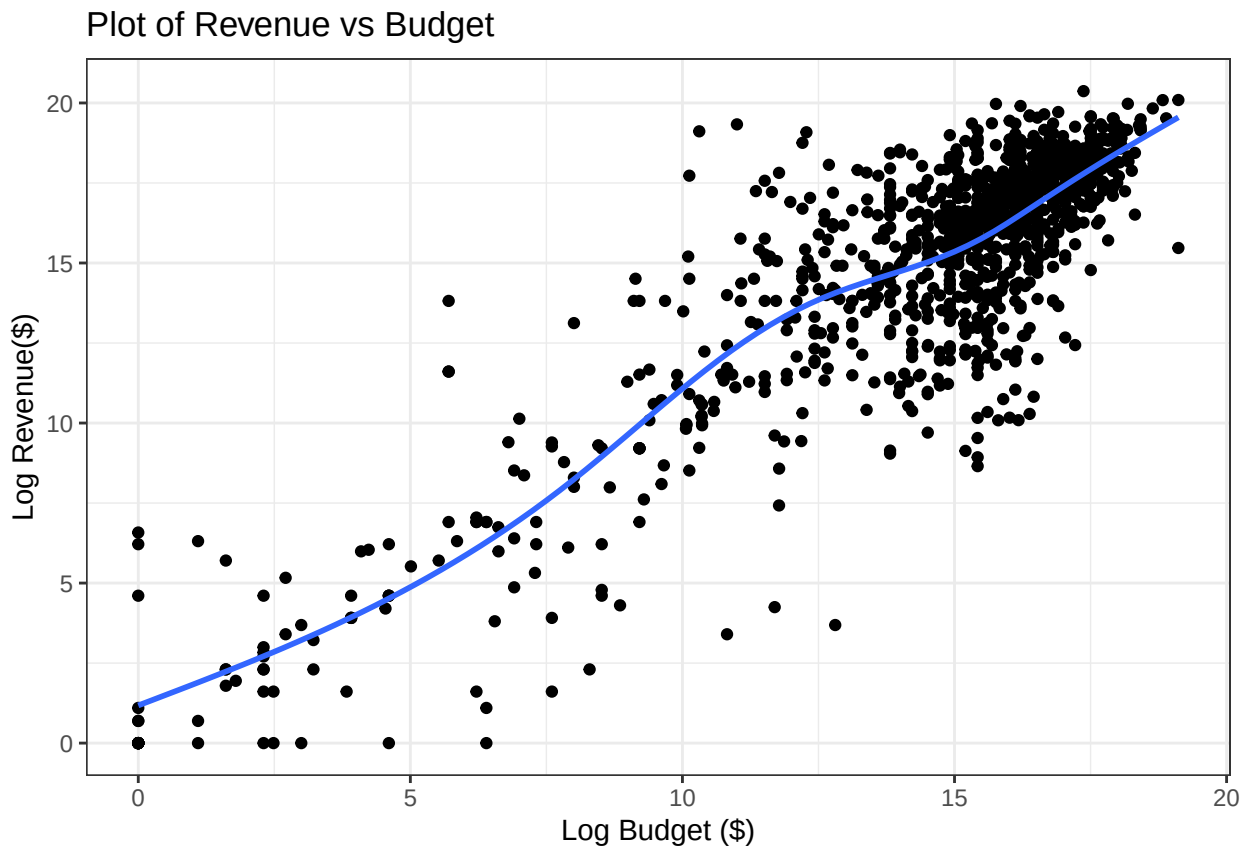
Estimating the Market Value of House Remodels

Lab 2 Sample Answer - Paul Laskowski

#look into assumptions here through EDA, check distributions. Money variables should be log vs no log.

```
horror_movies %>% ggplot() + aes(x = (budget), y = (revenue)) + geom_point() + geom_smooth(se = FALSE) +
  labs(x = 'Log Budget ($)', y = 'Log Revenue($)')
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

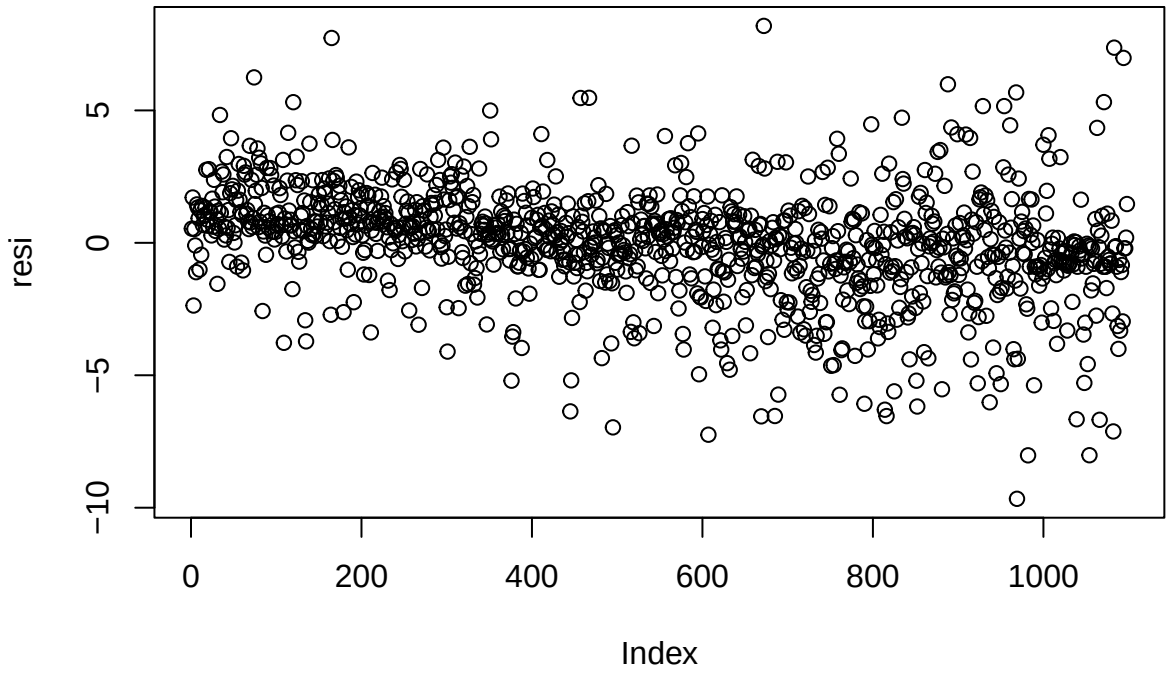



```
#looks less heteroskedastic, use this instead and into our model. Be wary of  how you interpret your va
```

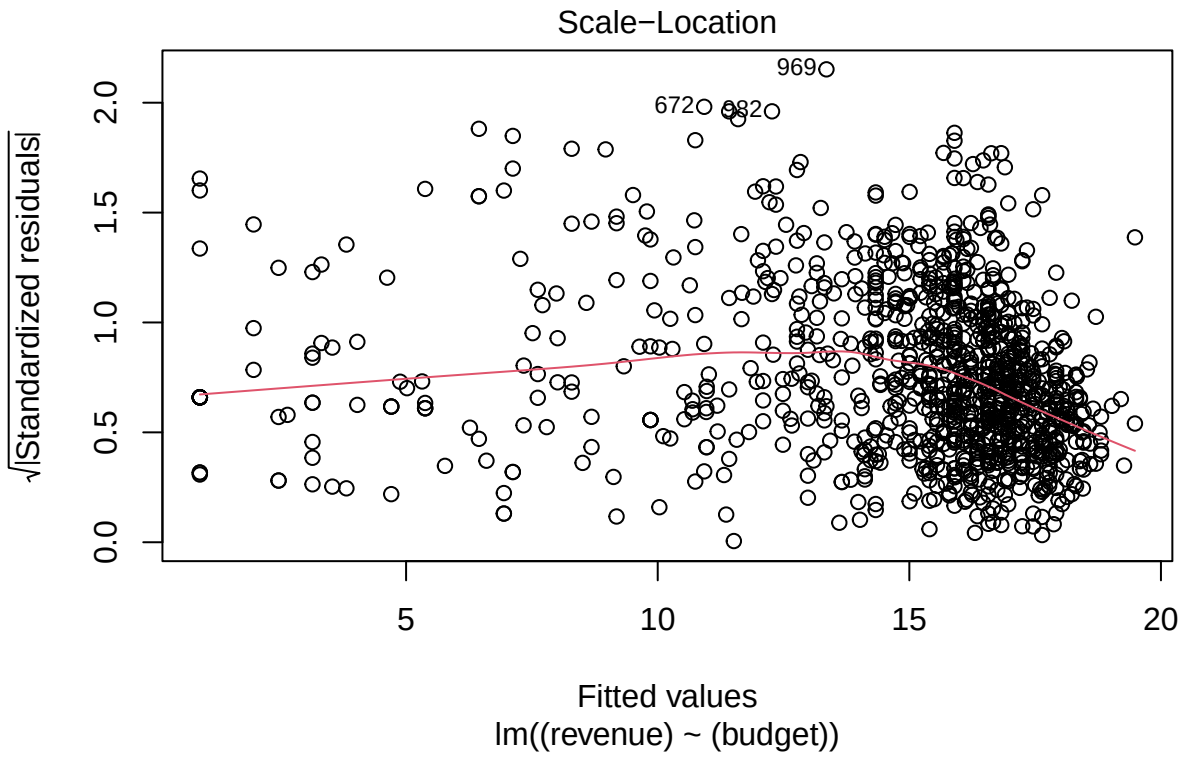
```
m_minimal <- lm((revenue) ~ (budget), horror_movies)
se_minimal <- m_minimal %>%
  vcovHC(type = "HC1") %>%
  diag() %>%
  sqrt()

m_central <- lm(revenue ~ budget  + runtime + vote_count, horror_movies)
se_central <- m_central %>%
```

```
  vcovHC(type = "HC1") %>%
  diag() %>%
  sqrt()

m_verbose<- lm(revenue ~ budget + runtime + vote_count + vote_average + year_diff + english + collectio
se_verbose <- m_verbose %>%
  vcovHC(type = "HC1") %>%
  diag() %>%
  sqrt()
```

```
resi <- m_minimal$residuals

plot(resi)
```




```
plot(m_minimal, which = 3) + stat_smooth()
```

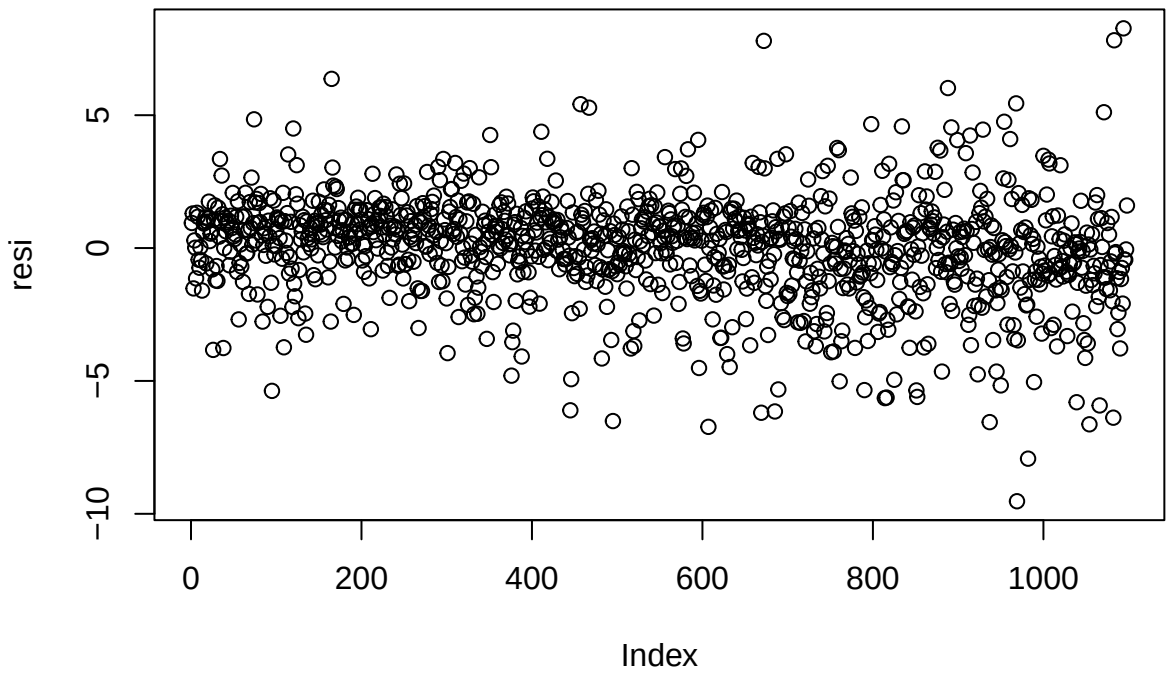



```
## NULL
```

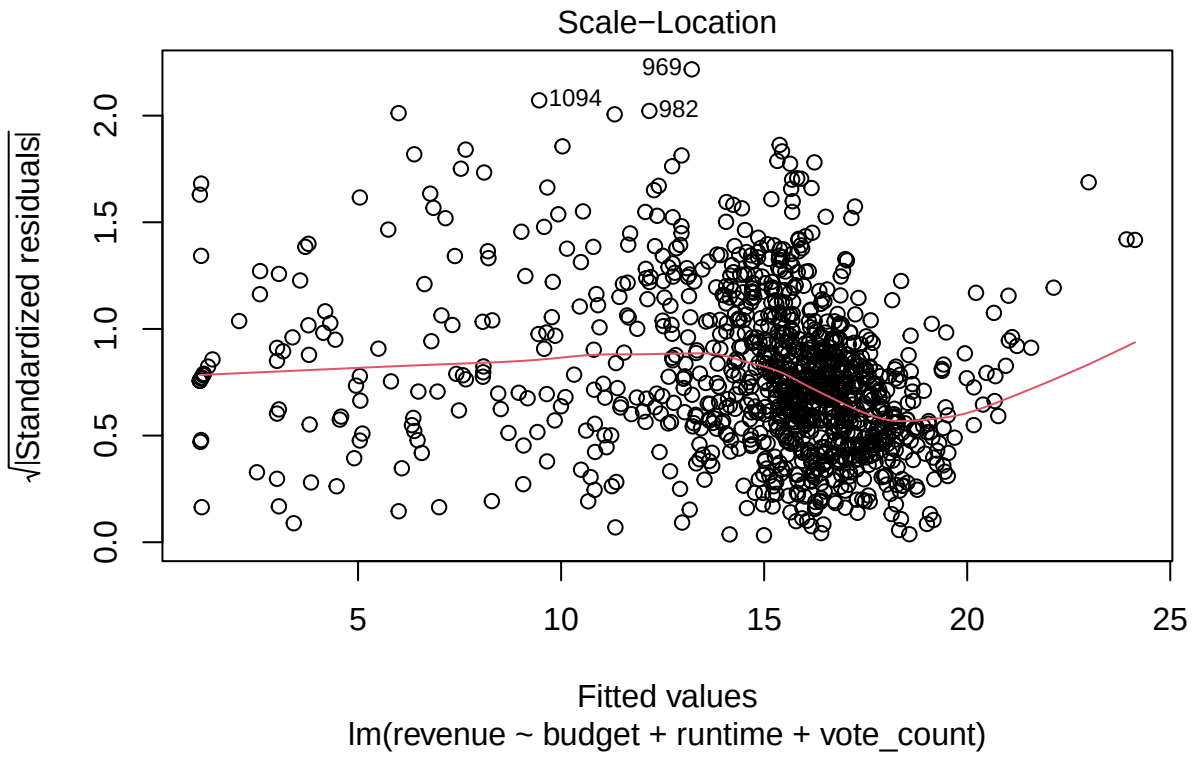
```
bptest(m_minimal)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  m_minimal
## BP = 32.518, df = 1, p-value = 1.181e-08
```

```
resi <- m_central$residuals

plot(resi)
```




```
plot(m_central, which = 3) + stat_smooth()
```

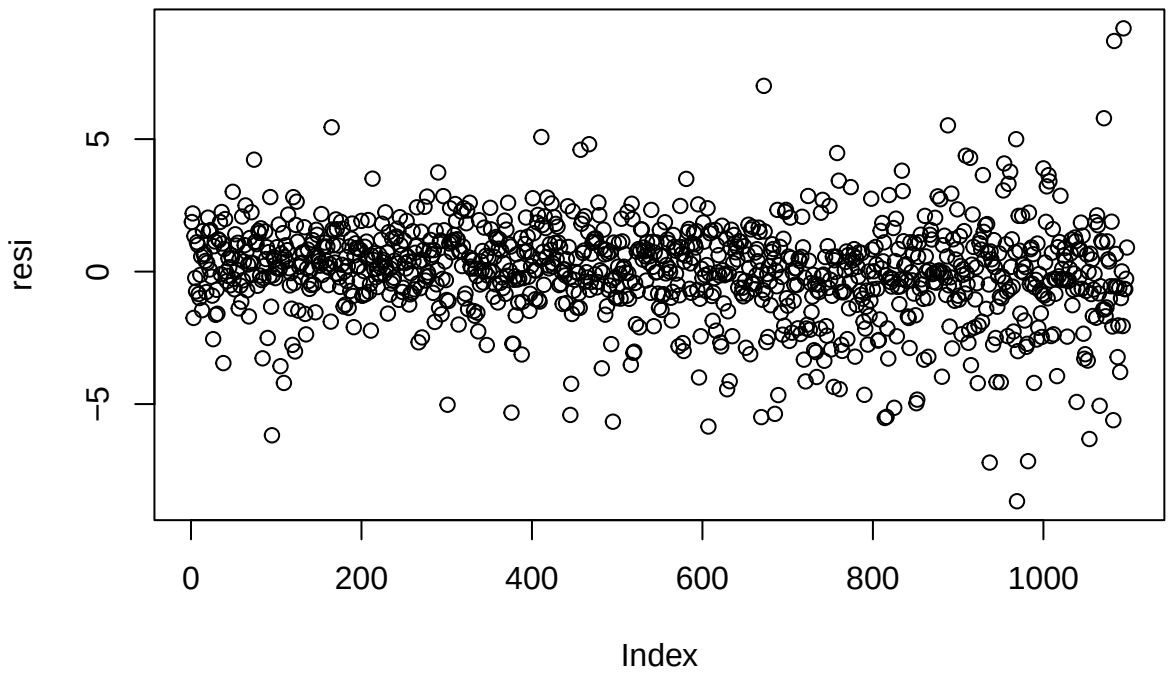



```
## NULL
```

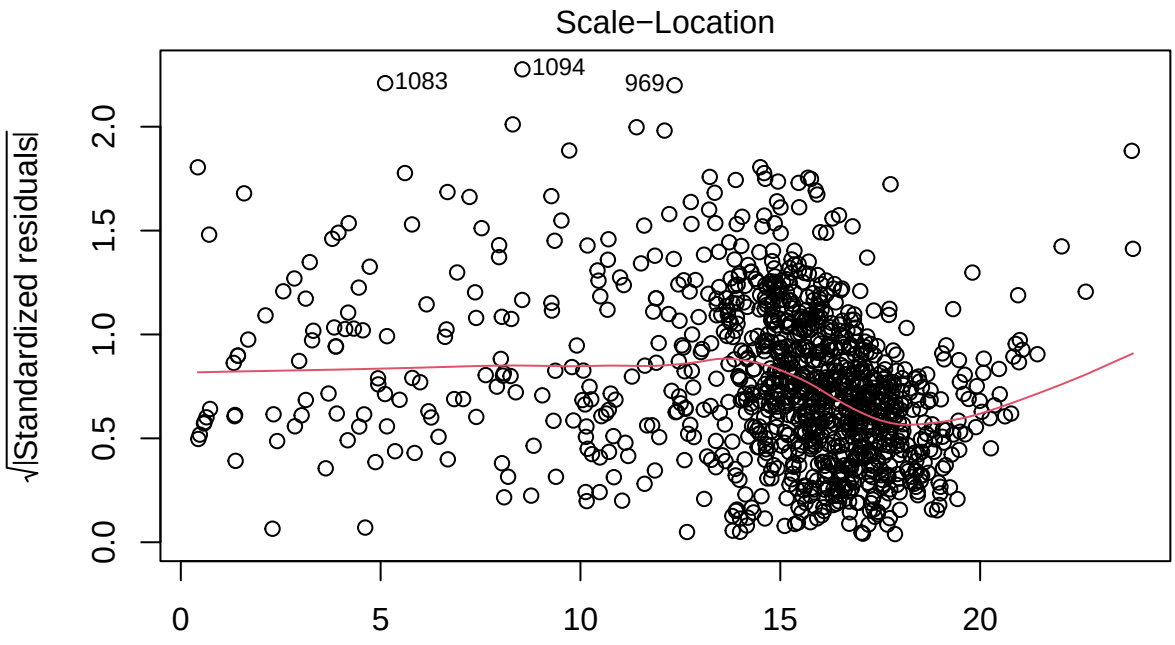
```
bptest(m_central)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  m_central
## BP = 45.13, df = 3, p-value = 8.683e-10
```

```
resi <- m_verbose$residuals

plot(resi)
```




```
plot(m_verbose, which = 3) + stat_smooth()
```




```
## NULL
```

```
bptest(m_verbose)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  m_verbose
## BP = 65.546, df = 7, p-value = 1.168e-11
```

```
  cor(horror_movies, method = c("pearson"))
```

```
##              vote_count vote_average     budget    revenue     runtime
## vote_count    1.0000000   0.27696018  0.3344668  0.4512944  0.30180694
## vote_average  0.2769602   1.00000000  0.2383073  0.2939443  0.22653674
## budget        0.3344668   0.23830729  1.0000000  0.8687186  0.72244035
## revenue       0.4512944   0.29394430  0.8687186  1.0000000  0.68642580
## runtime       0.3018069   0.22653674  0.7224404  0.6864258  1.00000000
## collection    0.2422896   0.11801859  0.2135870  0.3568736  0.16765222
## year_diff     0.1104494  -0.09048561 -0.1108616 -0.1886906 -0.14240845
## english       0.1874672   0.04721560  0.2066673  0.2012218  0.09465987
##               collection   year_diff     english
## vote_count    0.24228962  0.11044944  0.18746718
## vote_average  0.11801859 -0.09048561  0.04721560
## budget        0.21358701 -0.11086155  0.20666725
## revenue       0.35687359 -0.18869057  0.20122179
## runtime       0.16765222 -0.14240845  0.09465987
## collection    1.00000000 -0.16901755  0.08932781
```

Table 1: Estimated Regressions

	Output Variable: Log Revenue in Dollars		
		revenue	
	(1)	(2)	(3)
Log Budget	0.97***	0.83***	0.79***
	(0.02)	(0.03)	(0.03)
Runtime		0.02***	0.01***
		(0.004)	(0.004)
Vote Count		0.0004***	0.0004***
		(0.0000)	(0.0000)
Vote Average			0.09*
			(0.04)
(year released - oldest release year)			−0.03***
			(0.004)
English			−0.10
			(0.19)
Collection			1.18***
			(0.11)
Constant	0.90**	1.10***	2.27***
	(0.29)	(0.30)	(0.47)
Observations	1,098	1,098	1,098
R^2	0.75	0.79	0.82
Adjusted R^2	0.75	0.79	0.82
Residual Std. Error	2.09 (df = 1096)	1.94 (df = 1094)	1.80 (df = 1090)
F Statistic	3,371.49*** (df = 1; 1096)	1,360.43*** (df = 3; 1094)	706.00*** (df = 7; 1090)

Note: HC_1 robust standard errors in parentheses. Additional features are vote count, vote average, year diff, english, collection

```
## year_diff    -0.16901755  1.00000000 -0.12671740
## english       0.08932781 -0.12671740  1.00000000
```