Lecture 13 Notes

2024-02-29

Opening the data

```
require(tidyverse)
```

```
## Loading required package: tidyverse
```

```
## Warning: package 'tidyverse' was built under R version 4.3.2
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.2     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.4     ✔ tibble    3.2.1
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts t
o become errors
```

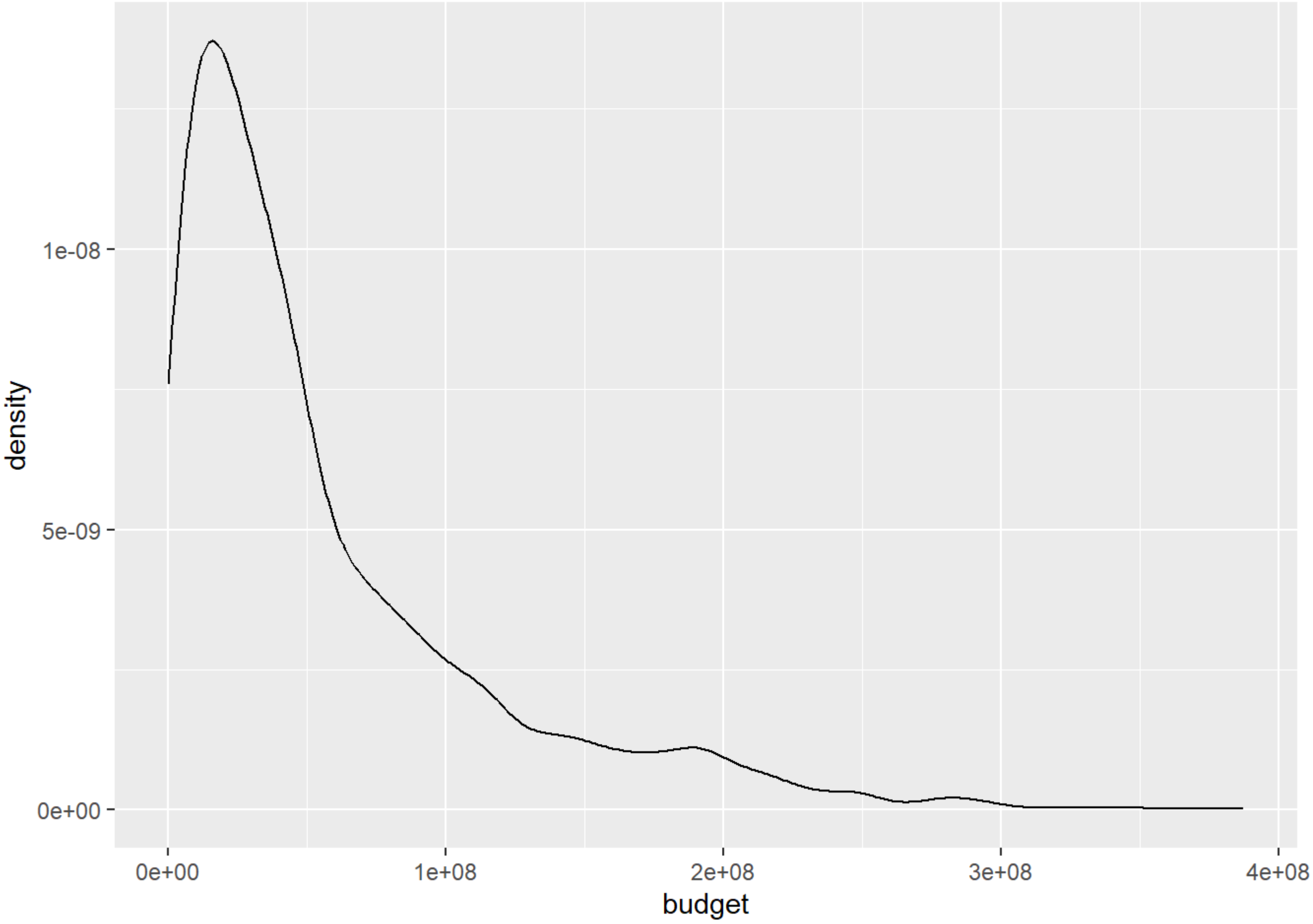
```
mv <- read_rds('https://github.com/jbisbee1/DS1000_S2024/raw/main/data/mv.Rds')
mv
```

```
## # A tibble: 7,673 × 20
##    title  rating genre  year released score  votes director writer star  country
##    <chr>  <chr>  <chr> <dbl> <chr>    <dbl>  <dbl> <chr>    <chr>  <chr> <chr>
##  1 The S… R      Drama  1980 June 13…   8.4 9.27e5 Stanley… Steph… Jack… United…
##  2 The B… R      Adve…  1980 July 2,…   5.8 6.5 e4 Randal … Henry… Broo… United…
##  3 Star … PG     Acti…  1980 June 20…   8.7 1.20e6 Irvin K… Leigh… Mark… United…
##  4 Airpl… PG     Come…  1980 July 2,…   7.7 2.21e5 Jim Abr… Jim A… Robe… United…
##  5 Caddy… R      Come…  1980 July 25…   7.3 1.08e5 Harold … Brian… Chev… United…
##  6 Frida… R      Horr…  1980 May 9, …   6.4 1.23e5 Sean S.… Victo… Bets… United…
##  7 The B… R      Acti…  1980 June 20…   7.9 1.88e5 John La… Dan A… John… United…
##  8 Ragin… R      Biog…  1980 Decembe…   8.2 3.30e5 Martin … Jake … Robe… United…
##  9 Super… PG     Acti…  1980 June 19…   6.8 1.01e5 Richard… Jerry… Gene… United…
## 10 The L… R      Biog…  1980 May 16,…   7   1    e4 Walter … Bill … Davi… United…
## # ℹ 7,663 more rows
## # ℹ 9 more variables: budget <dbl>, gross <dbl>, company <chr>, runtime <dbl>,
## #   id <dbl>, imdb_id <chr>, bechdel_score <dbl>, boxoffice_a <dbl>,
## #   language <chr>
```

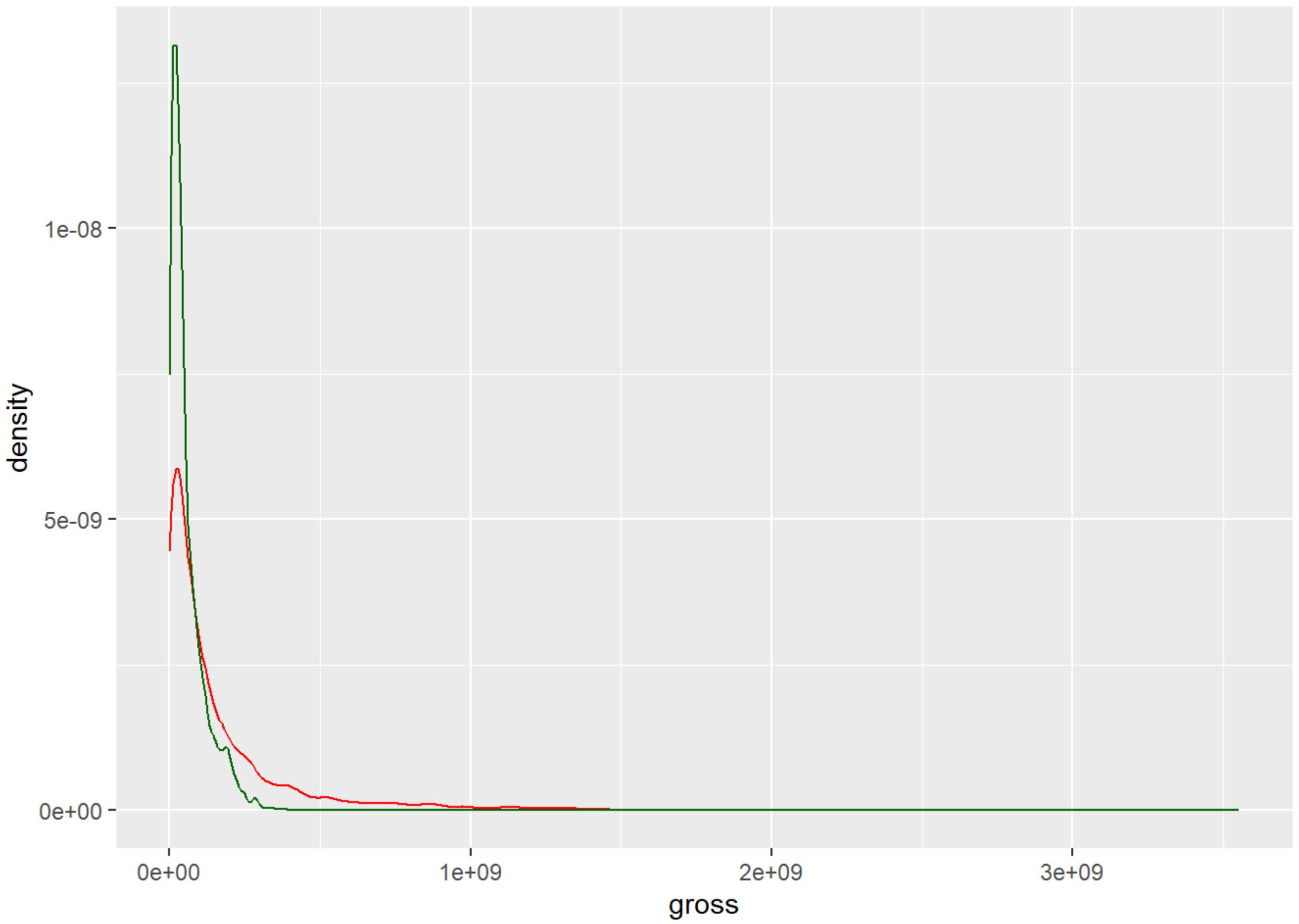
Univariate visualization

```
mv %>%
  ggplot(aes(x = budget)) +
  geom_density()
```

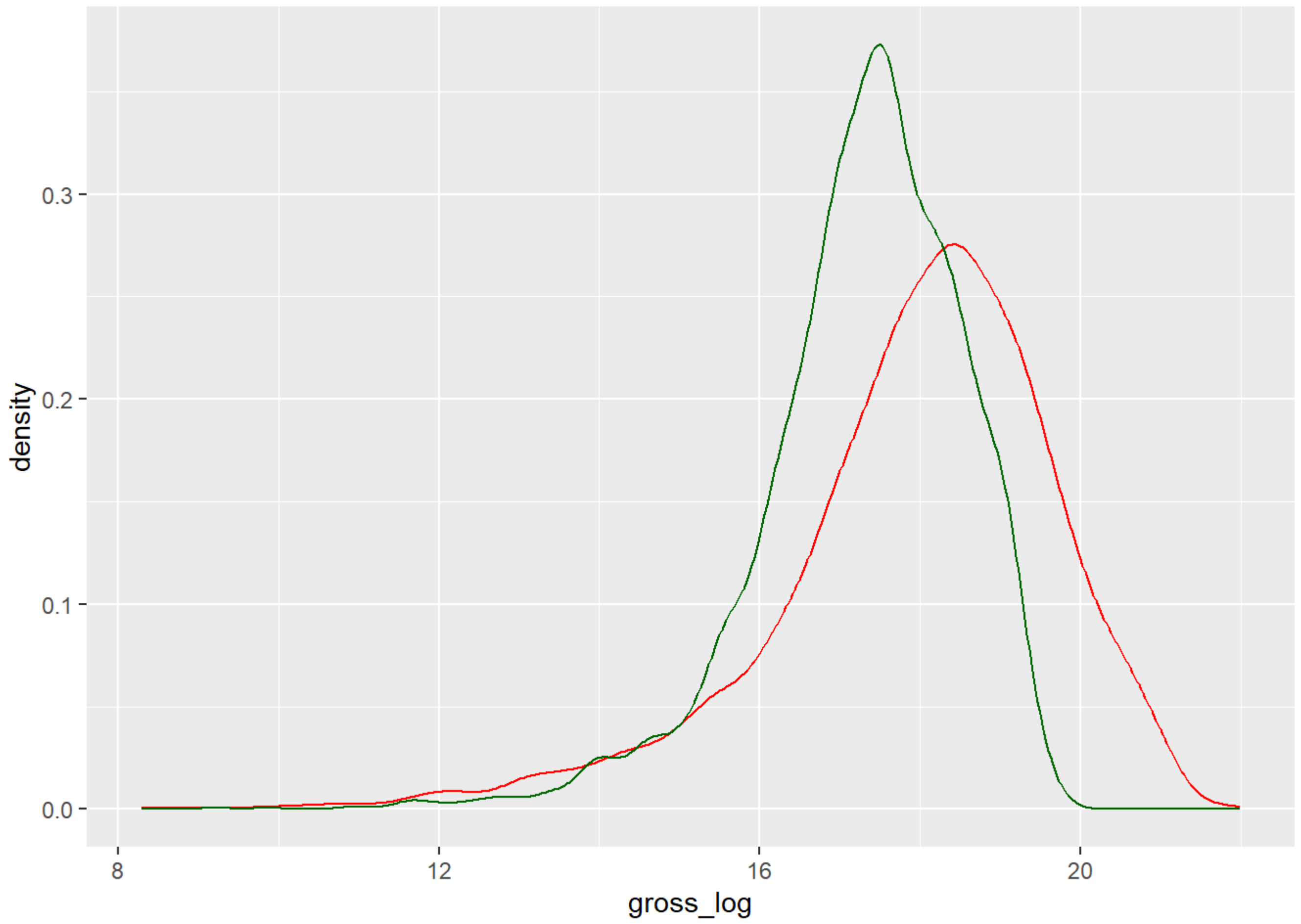
```
## Warning: Removed 4482 rows containing non-finite values (`stat_density()`).
```



```
# Combine both on a single plot
mv %>%
  drop_na(budget,gross) %>%
  ggplot() +
  geom_density(aes(x = gross),color = 'red') +
  geom_density(aes(x = budget),color = 'darkgreen')
```



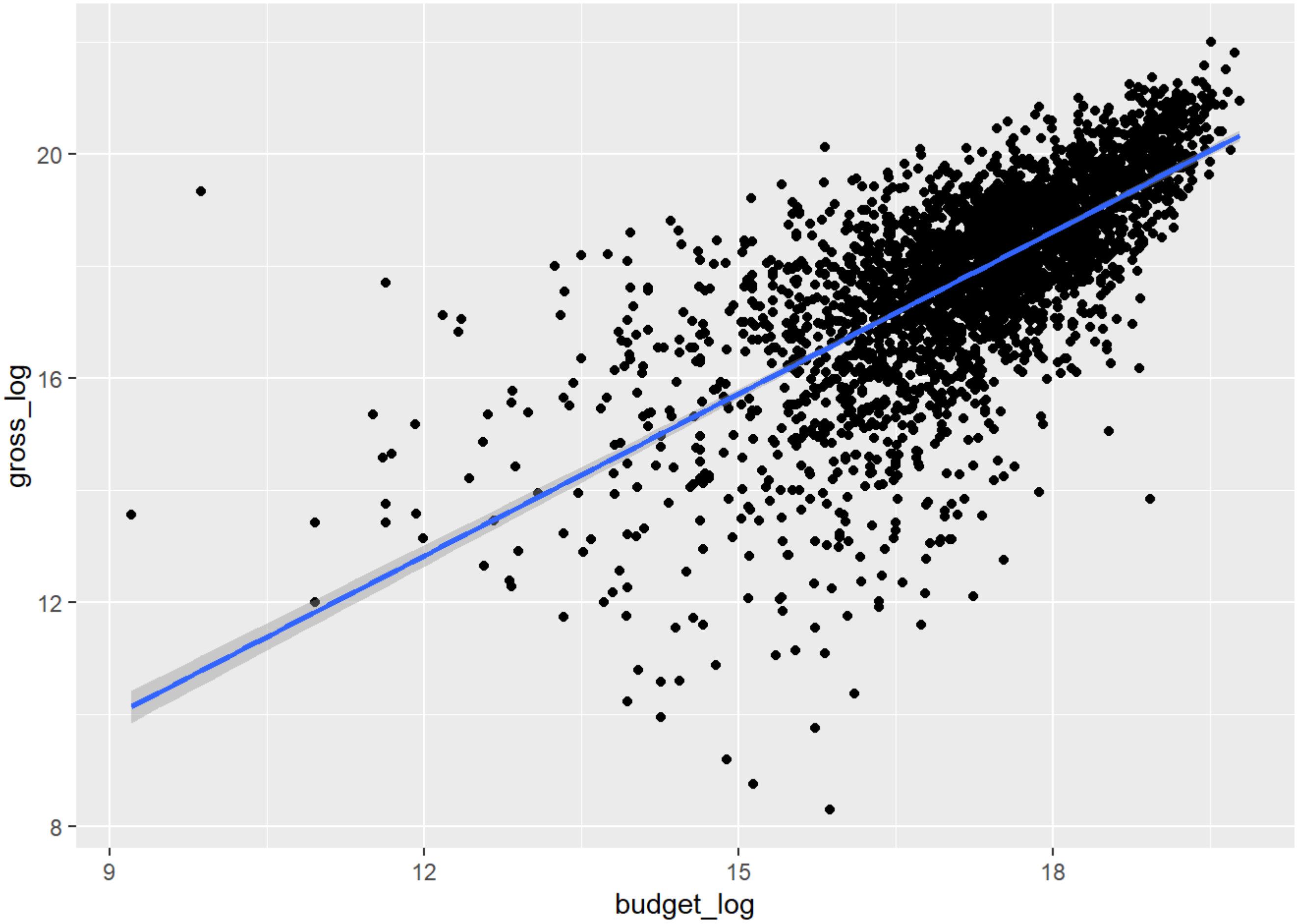
```
# Transforming with log()
mv %>%
  drop_na(budget,gross) %>%
  mutate(budget_log = log(budget),
         gross_log = log(gross)) %>%
  ggplot() +
  geom_density(aes(x = gross_log),color = 'red') +
  geom_density(aes(x = budget_log),color = 'darkgreen')
```



Multivariate visualization

```
mv_analysis <- mv %>%
  drop_na(budget,gross) %>%
  mutate(budget_log = log(budget),
         gross_log = log(gross))

mv_analysis %>%
  ggplot(aes(x = budget_log,
             y = gross_log)) +
  geom_point() +
  geom_smooth(method = 'lm')
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Regression

```
model_gross_budget <- lm(formula = gross_log ~ budget_log,
                         data = mv_analysis)


exp(1.26)
```

```
## [1] 3.525421
```

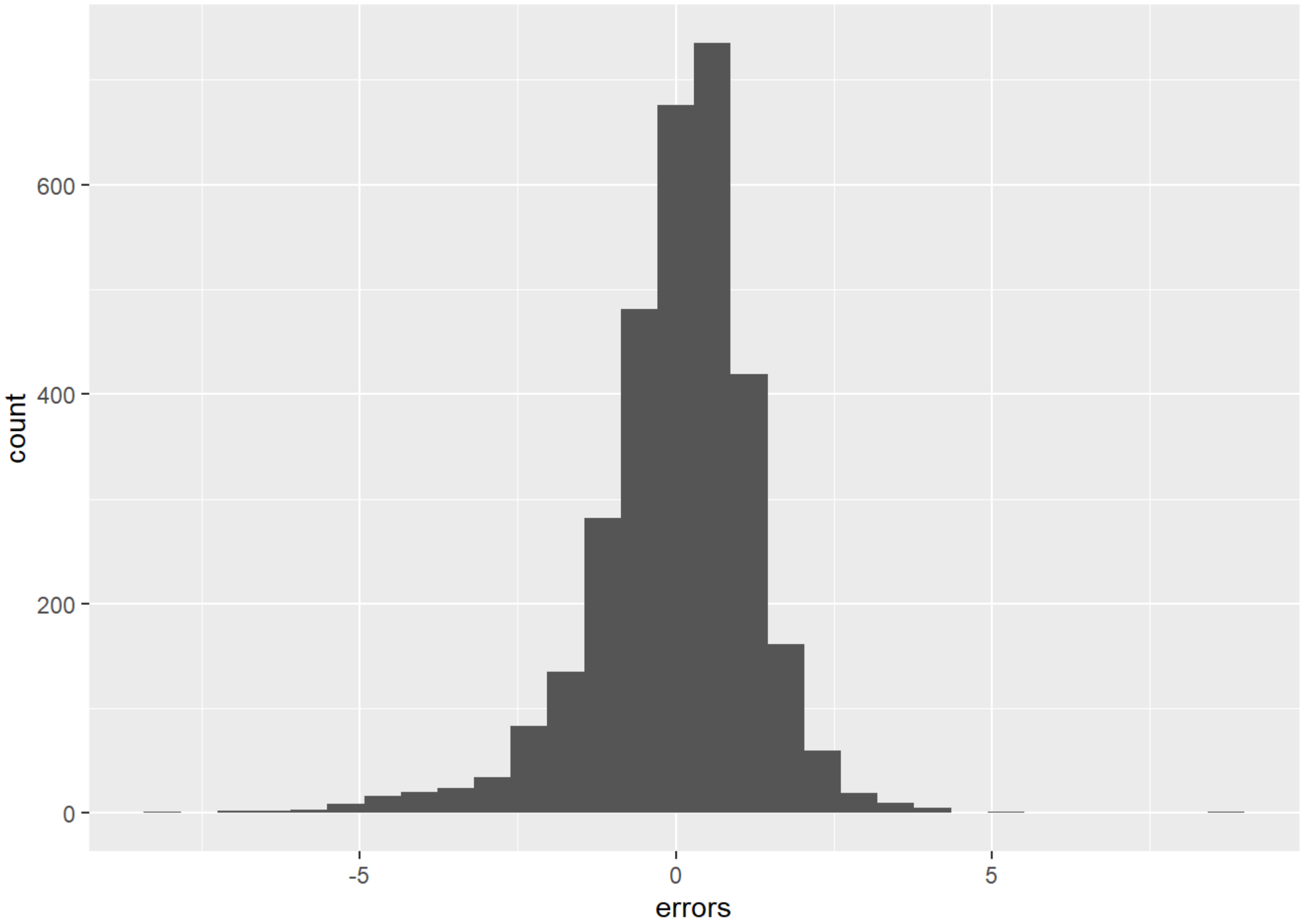
```
exp(0)
```

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

Calculating errors

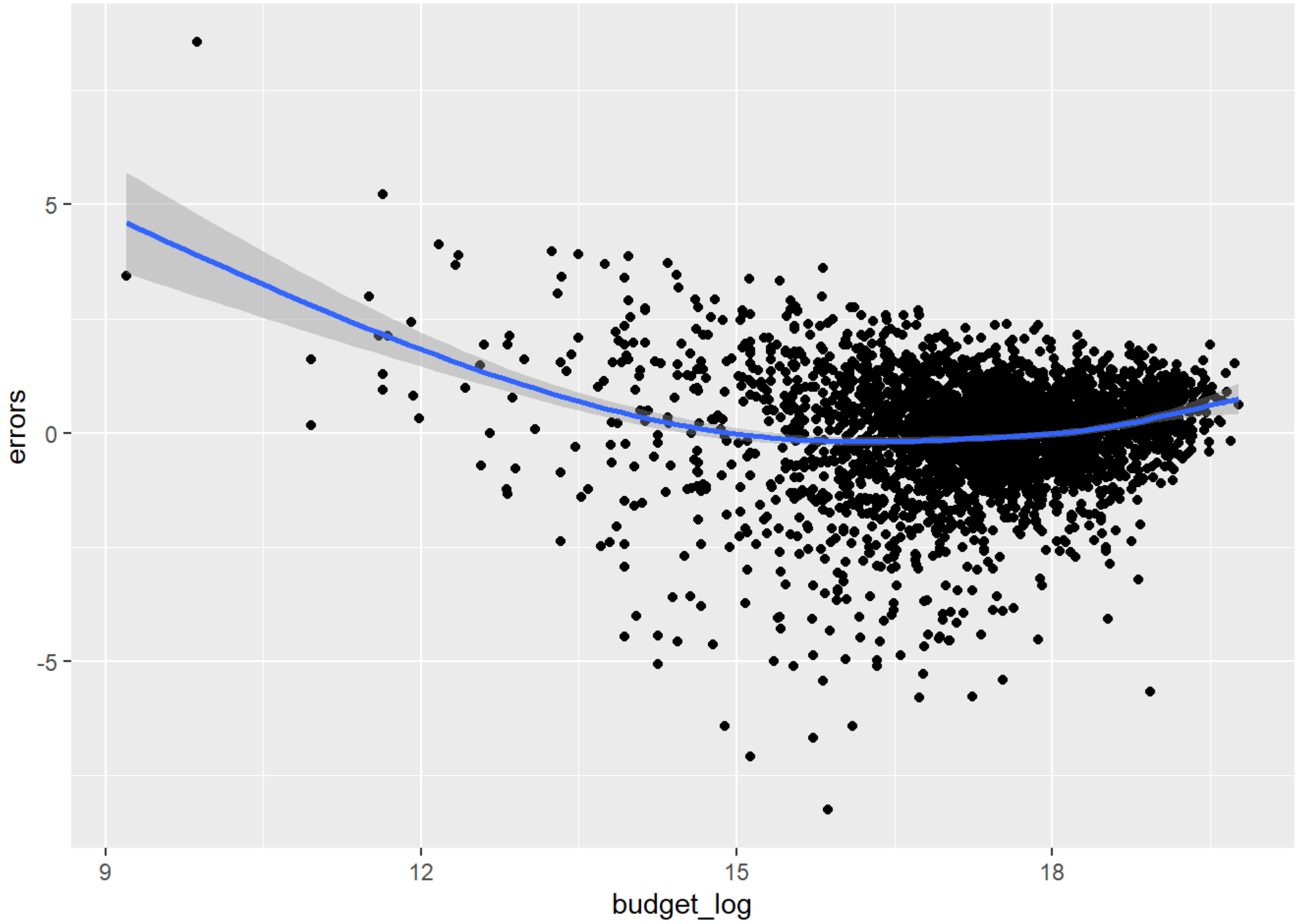
```
mv_analysis <- mv_analysis %>%
  mutate(preds = predict(model_gross_budget))

mv_analysis <- mv_analysis %>%
  mutate(errors = gross_log - preds)

mv_analysis %>%
  ggplot(aes(x = errors)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
mv_analysis %>%
  ggplot(aes(x = budget_log,y = errors)) +
  geom_point() +
  geom_smooth()
```

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



RMSE

```
rmse <- mv_analysis %>%
  mutate(se = errors^2) %>%
  summarise(mse = mean(se)) %>%
  mutate(rmse = sqrt(mse))
```

Evaluating RMSE

```
model_gross_budget
```

```
## 
## Call:
## lm(formula = gross_log ~ budget_log, data = mv_analysis)
## 
## Coefficients:
## (Intercept)   budget_log  
##      1.2611       0.9639
```

```
predLog_gross <- 1.26 + .96*log(10000000)
exp(predLog_gross)
```

```
## [1] 18501675
```

```
# Range: upper bound
predLog_gross_ub <- 1.26 + .96*log(10000000) + rmse$rmse
exp(predLog_gross_ub)
```

```
## [1] 66599457
```

```
# Range: lower bound
predLog_gross_lb <- 1.26 + .96*log(10000000) - rmse$rmse
exp(predLog_gross_lb)
```

```
## [1] 5139861
```