Simple Linear Regression: Misc. Topics

Sleuth3 Chapters 7, 8

Simple Linear Regression Model and Conditions

- Observations follow a normal distribution with mean that is a linear function of the explanatory variable
- $Y_i \sim \text{Normal}(\beta_0 + \beta_1 X_i, \sigma)$

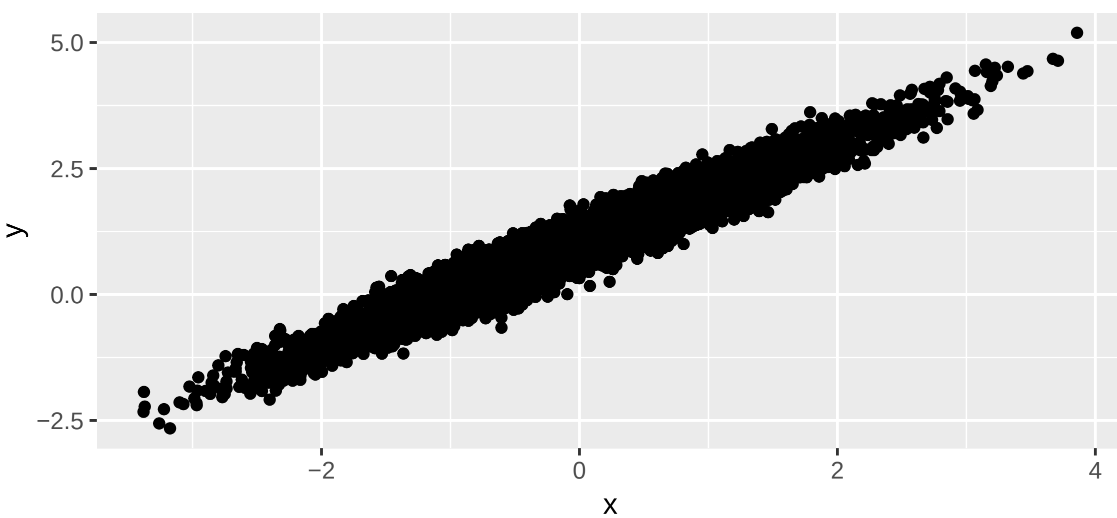
Conditions: spells "LINE-O"

- **Linear** relationship between explanatory and response variables: $\mu(Y|X) = \beta_0 + \beta_1 X$
- **Independent** observations (knowing that one observation is above its mean wouldn't give you any information about whether or not another observation is above its mean)
- **Normal** distribution of responses around the line
- **Equal standard deviation** of response for all values of X
- **no Outliers** (not a formal part of the model, but important to check in practice)

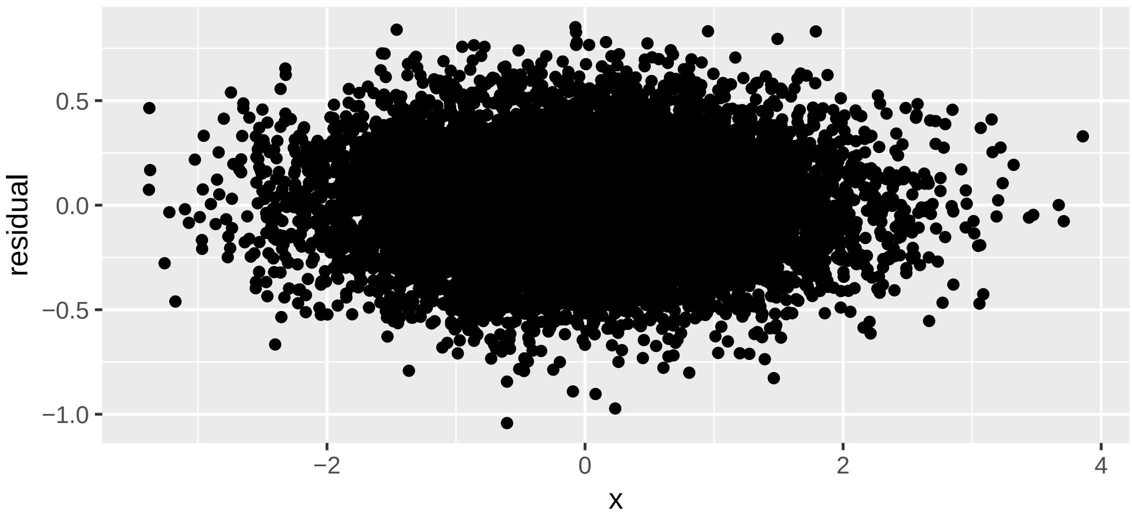
Some things that are not problems

Standard deviations may look narrower at the ends of the X axis due to fewer data points there

```r
ggplot(data = fake_data, mapping = aes(x = x, y = y)) +
  geom_point()
```




```r
lm_fit <- lm(y ~ x, data = fake_data)
fake_data <- fake_data %>%
  mutate(
    residual = residuals(lm_fit)
  )

ggplot(data = fake_data, mapping = aes(x = x, y = residual)) +
  geom_point()
```



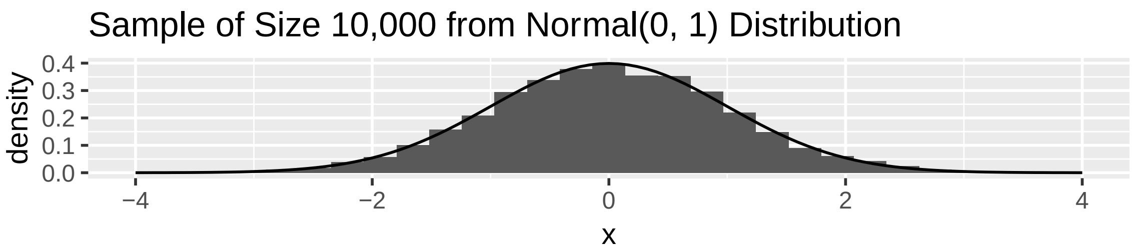

```
group_0 <- fake_data %>%
  filter(-0.5 <= x & x <= 0.5)

group_0 %>%
  summarize(
    sd(residual),
    residual_range = max(residual) - min(residual)
  )
```

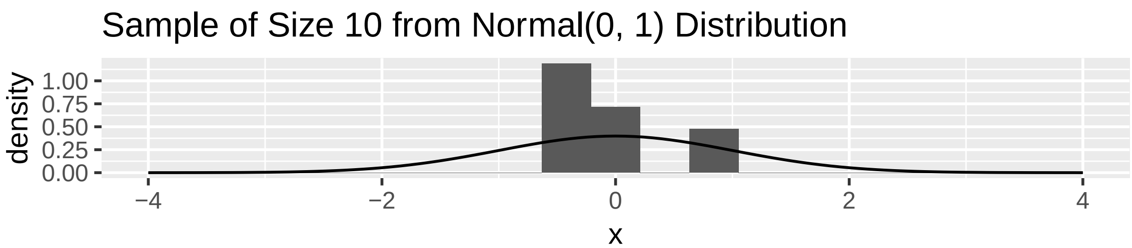
```
##   sd(residual) residual_range
## 1    0.2484015       1.824067
```

```
group_greater3 <- fake_data %>%
  filter(x > 3)

group_greater3 %>%
  summarize(
    sd(residual),
    residual_range = max(residual) - min(residual)
  )
```

```
##   sd(residual) residual_range
## 1    0.2489675      0.8810169
```

Why? A large sample will start to fill in the tails of the distribution, creating the appearance of more spread even though the distribution is the same.

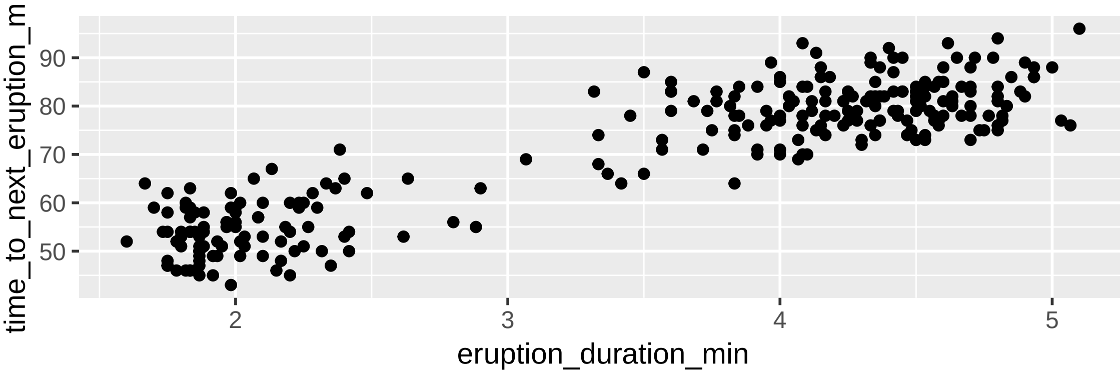




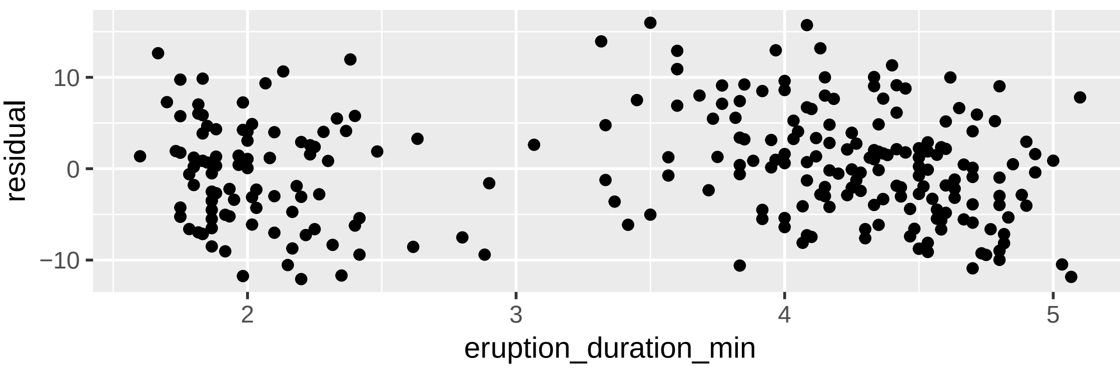
Areas with less data

Old Faithful is a geyser in Wyoming. X = duration in minutes of one eruption. Y = how long until the next eruption.

```
ggplot(data = old_faithful, mapping = aes(x = eruption_duration_min, y = time_to_next_eruption_min)) +
  geom_point()
```

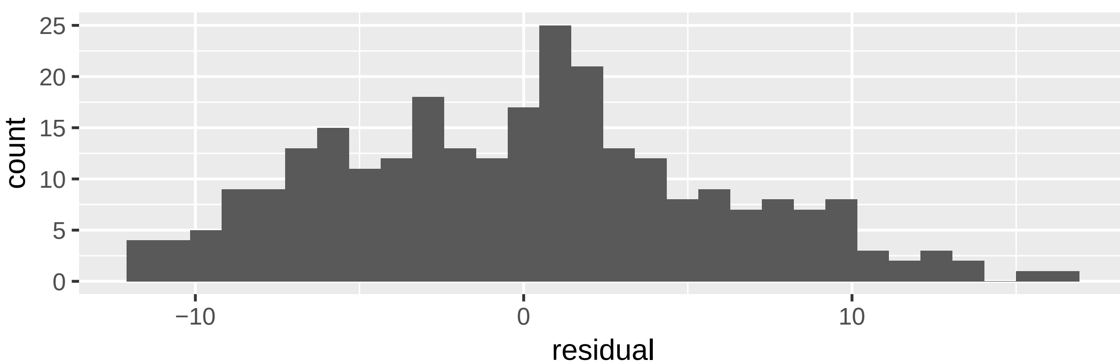



```
lm_fit <- lm(time_to_next_eruption_min ~ eruption_duration_min, data = old_faithful)

old_faithful <- old_faithful %>%
  mutate(
    residual = residuals(lm_fit)
  )

ggplot(data = old_faithful, mapping = aes(x = eruption_duration_min, y = residual)) +
  geom_point()
```




```
ggplot(data = old_faithful, mapping = aes(x = residual)) +
  geom_histogram()
```

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



Why? The model does not say anything about the distribution of the explanatory variable. It can have gaps. What matters is that at each value of X, Y follows a normal distribution.

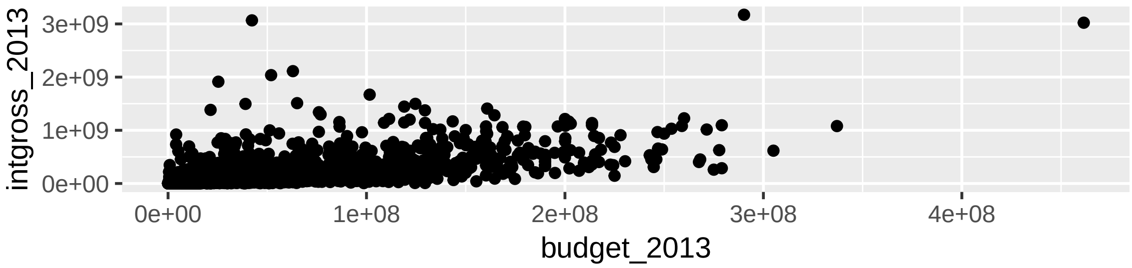
Checking Normality

- First Step: Fit the model, get the residuals, and make a histogram or density plot.
- Be cautious if outliers or long tails show up
- Possibly also: a Q-Q plot

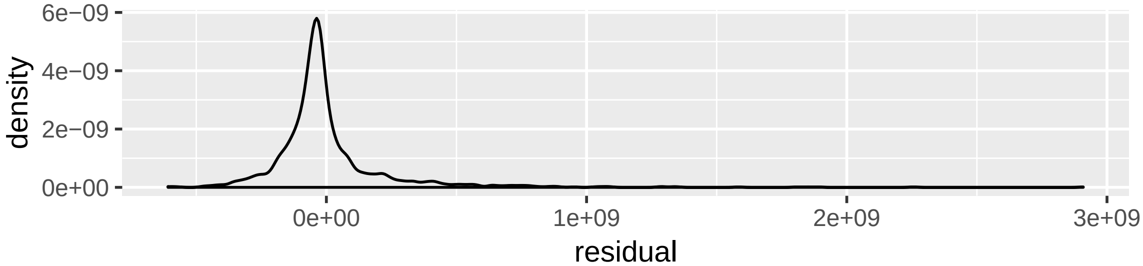
Example

Let's look at modeling a movie's international gross earnings in inflation-adjusted 2013 dollars (`intgross_2013`) as a function of its budget (`budget_2013`).

```
ggplot(data = movies, mapping = aes(x = budget_2013, y = intgross_2013)) +
  geom_point()
```




```
lm_fit <- lm(intgross_2013 ~ budget_2013, data = movies)
movies <- movies %>%
  mutate(
    residual = residuals(lm_fit)
  )

ggplot(data = movies, mapping = aes(x = residual)) +
  geom_density()
```



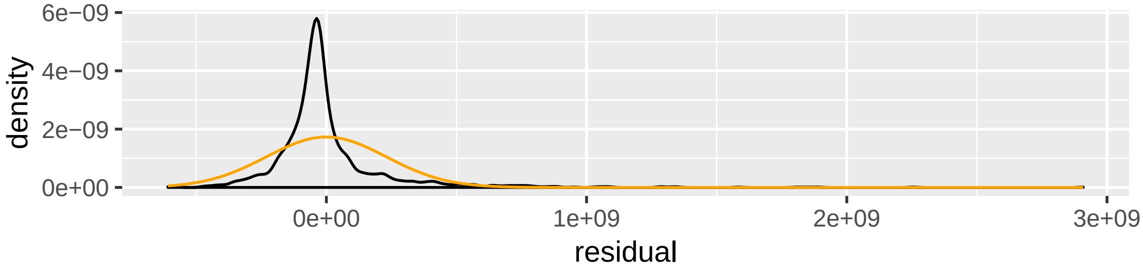
Is this close to a normal distribution?

No: In comparison to a normal distribution (orange), it is skewed right and **heavy tailed**:

- More movies have residuals close to 0 relative to the normal distribution
- More movies have residuals that are extremely large or extremely small relative to the normal distribution

Heavy tailed distributions are the one time when a lack of normality can cause problems.

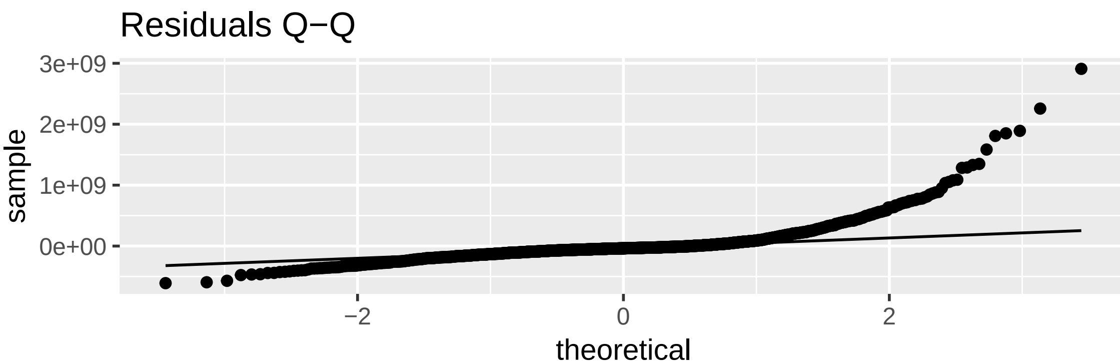
```
ggplot(data = movies, mapping = aes(x = residual)) +
  geom_density() +
  stat_function(fun = dnorm, args = list(mean = 0, sd = summary(lm_fit)$sigma), color = "orange")
```



To diagnose: a Q-Q plot.

- Q-Q stands for Quantile-Quantile
- Compare quantiles (percentiles) of the residuals to the corresponding quantiles (percentiles) from a normal distribution
- If the distribution of the residuals is approximately normal, points will fall along a line.
- If the distribution of the residuals is heavy tailed, the small residuals will be too small and the large residuals will be too large

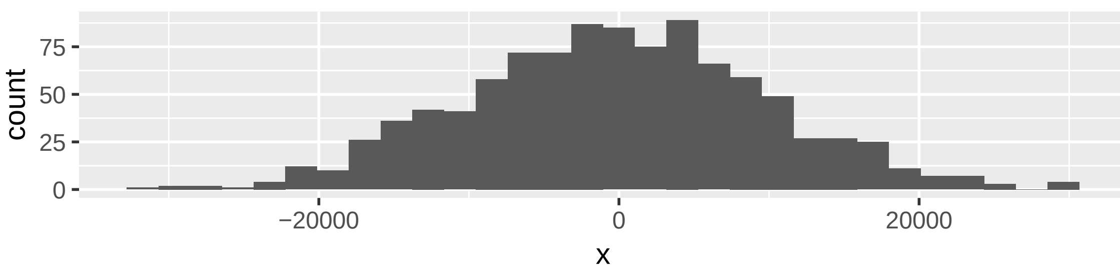
```
ggplot(data = movies, mapping = aes(sample = residual)) +
    stat_qq() +
    stat_qq_line() +
    ggtitle("Residuals Q-Q")
```



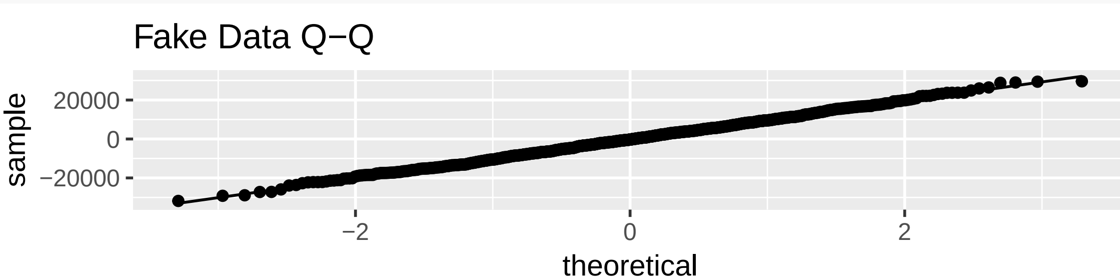
What we'd like to see:

```
fake_data <- data.frame(
  x = rnorm(1000, mean = 0, sd = 10000)
)

ggplot(data = fake_data, mapping = aes(x = x)) +
  geom_histogram()
```

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




```
ggplot(data = fake_data, mapping = aes(sample = x)) +
    stat_qq() +
    stat_qq_line() +
    ggtitle("Fake Data Q-Q")
```

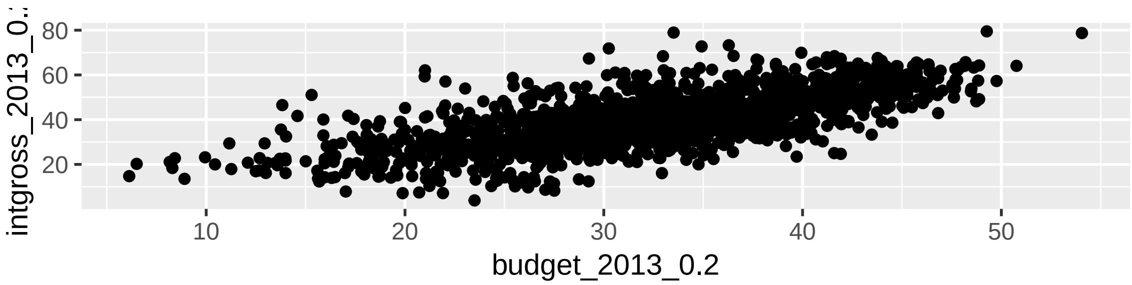


I use Q-Q plots as an indicator of whether I need to investigate more carefully; exact linearity in the Q-Q plot is not critical. (**An exactly normal distribution is not critical**)

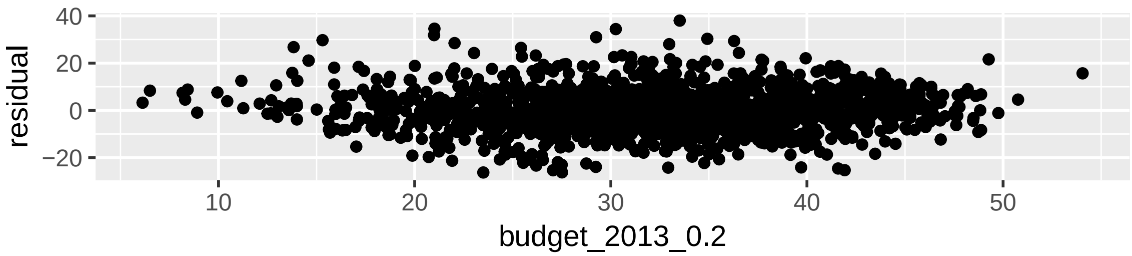
In this case, the problem can be fixed with a data transformation that also reduces the severity of the outliers.

```
movies <- movies %>% mutate(
  intgross_2013_0.2 = intgross_2013^{0.2},
  budget_2013_0.2 = budget_2013^{0.2}
)
```

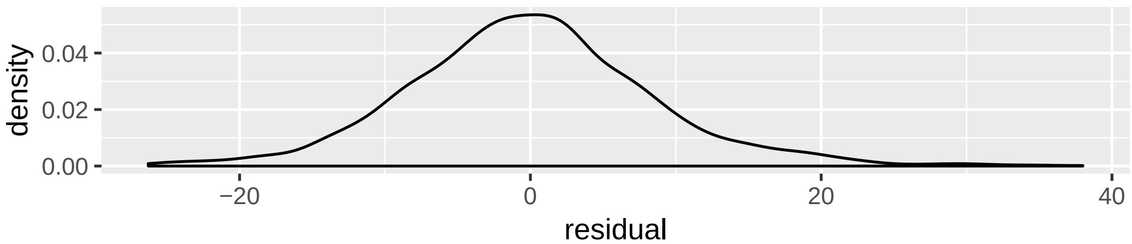
```
ggplot(data = movies, mapping = aes(x = budget_2013_0.2, y = intgross_2013_0.2)) +
  geom_point()
```



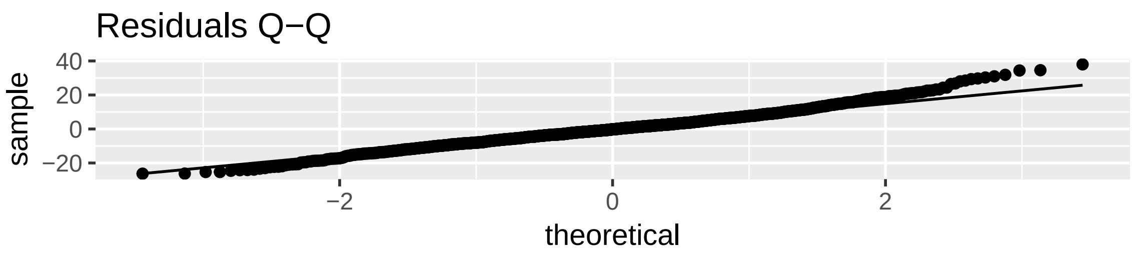

```
lm_fit <- lm(intgross_2013_0.2 ~ budget_2013_0.2, data = movies)
movies <- movies %>%
  mutate(
    residual = residuals(lm_fit)
  )

ggplot(data = movies, mapping = aes(x = budget_2013_0.2, y = residual)) +
  geom_point()
```




```
ggplot(data = movies, mapping = aes(x = residual)) +
  geom_density()
```




```
ggplot(data = movies, mapping = aes(sample = residual)) +
    stat_qq() +
    stat_qq_line() +
    ggtitle("Residuals Q-Q")
```



This is not perfect, but it is much better. Good enough.

Outliers

Suppose we were still worried about that movie with the largest budget (I'm not). We should:

- Figure out what movie it is and investigate whether there might have been a data entry error
- Fit the model both with and without that observation and **report both sets of results**.

Which movie is it? Use `filter` to find out:

```
movies %>%
  filter(budget_2013_0.2 > 50)
```

```
## # A tibble: 2 x 7
##    year title                    intgross_2013 budget_2013 residual intgross_2013_0~ budget_2013_0.2
##   <dbl> <chr>                            <dbl>       <dbl>    <dbl>            <dbl>           <dbl>
## 1  2009 Avatar                      3022588801   461435929    15.7              78.7            54.1
## 2  2007 Pirates of the Caribbea~    1079721346   337063045     4.57             64.1            50.8
```

- It was Avatar (larger budget). We confirm from our sources that the budget and gross earnings for Avatar were insane.
- Fit the model with Avatar:

```
lm_fit <- lm(intgross_2013_0.2 ~ budget_2013_0.2, data = movies)
summary(lm_fit)
```

```
##
## Call:
## lm(formula = intgross_2013_0.2 ~ budget_2013_0.2, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.278  -5.325  -0.237   4.852  37.997
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.89500    0.90335   5.419 6.84e-08 ***
## budget_2013_0.2  1.07580    0.02698  39.872  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.457 on 1744 degrees of freedom
## Multiple R-squared:  0.4769, Adjusted R-squared:  0.4766
## F-statistic:  1590 on 1 and 1744 DF,  p-value: < 2.2e-16
```

- Drop Avatar, fit it again without Avatar (`!=` means "not equal to")

```
movies_no_Avatar <- movies %>%
  filter(title != "Avatar")

lm_fit_no_Avatar <- lm(intgross_2013_0.2 ~ budget_2013_0.2, data = movies_no_Avatar)
summary(lm_fit_no_Avatar)
```

```
##
## Call:
## lm(formula = intgross_2013_0.2 ~ budget_2013_0.2, data = movies_no_Avatar)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.301  -5.319  -0.227   4.861  38.009
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.99810    0.90441   5.526 3.76e-08 ***
## budget_2013_0.2  1.07236    0.02703  39.679  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.451 on 1743 degrees of freedom
## Multiple R-squared:  0.4746, Adjusted R-squared:  0.4743
## F-statistic:  1574 on 1 and 1743 DF,  p-value: < 2.2e-16
```

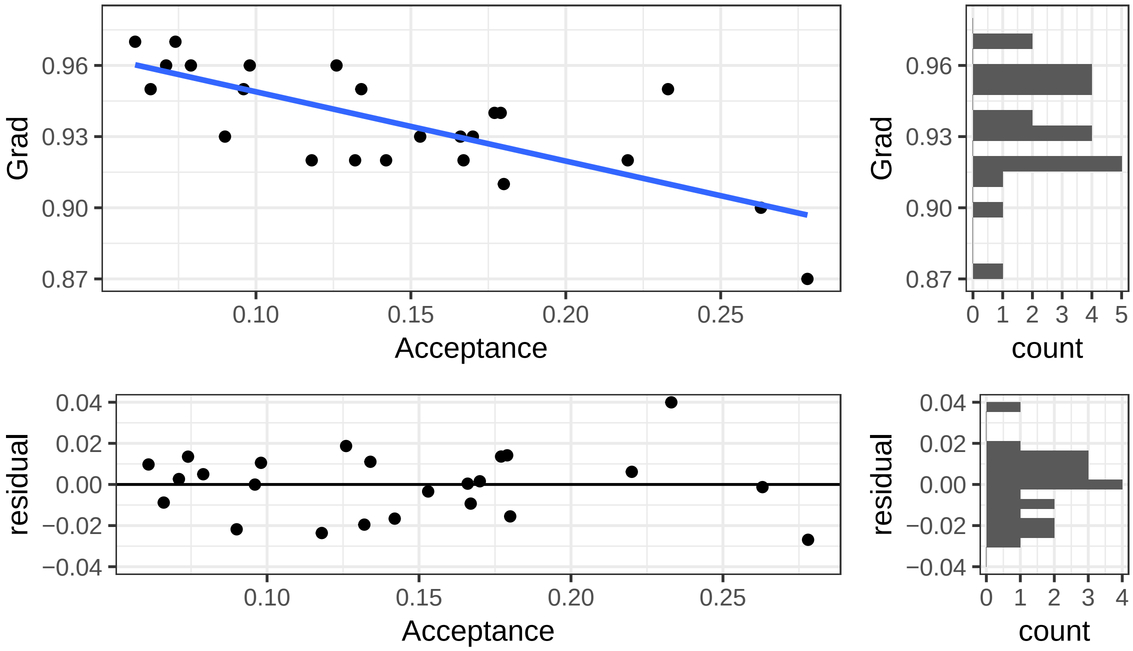
With Avatar included in the model, we estimate that a 1-unit increase in $\text{Budget}^{0.2}$ is associated with an increase of about 1.07580 in gross international earnings raised to the power of 0.2.

With Avatar not included in the model, we estimate that a 1-unit increase in $\text{Budget}^{0.2}$ is associated with an increase of about 1.07236 in gross international earnings raised to the power of 0.2.

Our conclusions about the association between a movie's budget and its gross international earnings are substantively the same whether or not Avatar is included.

R^2: The most useless statistic in statistics

Remember our example from last week with acceptance rate (explanatory variable) and graduation rate (response variable) for colleges in the US:



- Notice from the plots that the variance of the response variable is larger than the variance of the residuals.

```
colleges %>%
  summarize(
    var_Grad = var(Grad),
    var_resid = var(residual),
    var_resid_correct_df = sum((residual - mean(residual))^2)/(24 - 2)
  )
```

```
## # A tibble: 1 x 3
##   var_Grad var_resid var_resid_correct_df
##      <dbl>     <dbl>                <dbl>
## 1 0.000573  0.000250             0.000261
```

Our data set had $n = 24$ observations; the second variance of the residuals uses this correct degrees of freedom.

- $\frac{\text{Var(Residuals)}}{\text{Var(Response)}}$ can be interpreted as the proportion of the variance in the response variable that is still "left over" after fitting the model

```
0.000250 / 0.000573
```

```
## [1] 0.4363002
```

```
0.000261 / 0.000573
```

```
## [1] 0.4554974
```

44% or 46% of the variability in Graduation Rates is still there in the residuals.

- $R^2 = 1 - \frac{\text{Var(Residuals)}}{\text{Var(Response)}}$ can be interpreted as the proportion of the variance in the response variable that is accounted for by the linear regression on acceptance rate.

```
1 - 0.000250 / 0.000573
```

```
## [1] 0.5636998
```

```
1 - 0.000261 / 0.000573
```

```
## [1] 0.5445026
```

56% or 54% of the variability in Graduation Rates is accounted for by the linear regression on acceptance rate.

```r
summary(linear_fit)
```

```
##
## Call:
## lm(formula = Grad ~ Acceptance, data = colleges)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.026914 -0.010876  0.000968  0.010656  0.039947
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.978086   0.008582 113.966  < 2e-16 ***
## Acceptance  -0.291986   0.054748  -5.333 2.36e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01617 on 22 degrees of freedom
## Multiple R-squared:  0.5639, Adjusted R-squared:  0.544
## F-statistic: 28.44 on 1 and 22 DF,  p-value: 2.36e-05
```

- "Multiple R-squared" is the proportion of variability in the response accounted for by the model, but with the wrong degrees of freedom.
- "Adjusted R-squared" is the proportion of variability in the response accounted for by the model, but with the correct degrees of freedom.

Neither one of these is actually a useful indicator of anything. A model with low R^2 can still be useful. A model with high R^2 can still be wrong.

I never look at R^2.

Summary

- **Linear** relationship between explanatory and response variables: $\mu(Y|X) = \beta_0 + \beta_1 X$
  - **How to check:**
    - Look at scatter plots of the original data
    - Look at scatter plots of residuals vs. explanatory variable
  - **If not satisfied:**
    - Try a transformation
    - Fit a non-linear relationship
- **Independent** observations (knowing that one observation is above its mean wouldn't give you any information about whether or not another observation is above its mean)
  - **How to check:**
    - Be cautious of *time* effects or *cluster* effects
  - **If not satisfied:**
    - Use a different model that accounts for dependence
- **Normal** distribution of responses around the line
  - **How to check:**
    - Histogram or density plot of residuals; be cautious of outliers and/or long tails.
    - If any doubts, look at a Q-Q plot
  - **If not satisfied:**
    - Don't worry too much, unless the distribution is heavy tailed
    - If the distribution is heavy tailed (fairly rare), try a transformation or use a different method that is less affected by outliers
- **Equal standard deviation** of response for all values of X
  - **How to check:**
    - Look at scatter plots of the original data
    - Look at scatter plots of residuals vs. explanatory variable
  - **If not satisfied:**
    - Try a transformation (usually works)
    - Use weighted least squares
- **no Outliers** (not a formal part of the model, but important to check in practice)
  - **How to check:**
    - Look at scatter plots of the original data
    - Look at scatter plots of residuals vs. explanatory variable
  - **If not satisfied:**
    - Try to figure out what caused the outlier, and correct if a data entry error
    - Try a transformation
    - Conduct the analysis both with and without the outlier, **report both sets of results**