Basic Regression

Jameson Watts, Ph.D.

Agenda

1. Mid-term review
2. Basic Regression
    1. Numerical predictors
    2. Categorical predictors
    3. Residual analysis

Environment setup

```
library(tidyverse)
wine <- read_csv("../resources/winemag-data.csv") %>%
  filter(!is.na(price)) %>%
  mutate(year = as.numeric(str_extract(title,"(\\d{4})"))) %>%
  mutate(lprice=log(price))
```

Mid-term Review

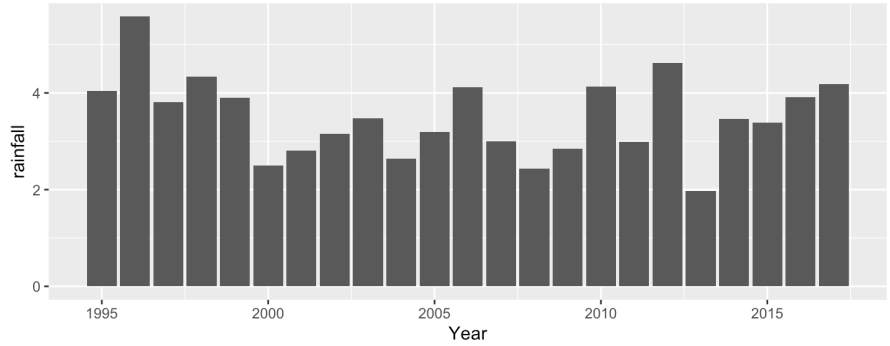
Problem 10

Create bar plot that shows average rainfall from 1995 onwards.

Hint: you may want to mutate using if_else to change NA to 0 before summarize

Solution

```
read_csv("../resources/rainfall.csv") %>%
  pivot_longer(-Year,names_to = "month", values_to = "rainfall") %>%
  mutate(rainfall=if_else(is.na(rainfall),0,rainfall)) %>%
  group_by(Year) %>%
  summarise(rainfall=mean(rainfall)) %>%
  filter(Year >= 1995) %>%
  ggplot(aes(Year,rainfall))+
    geom_col()
```

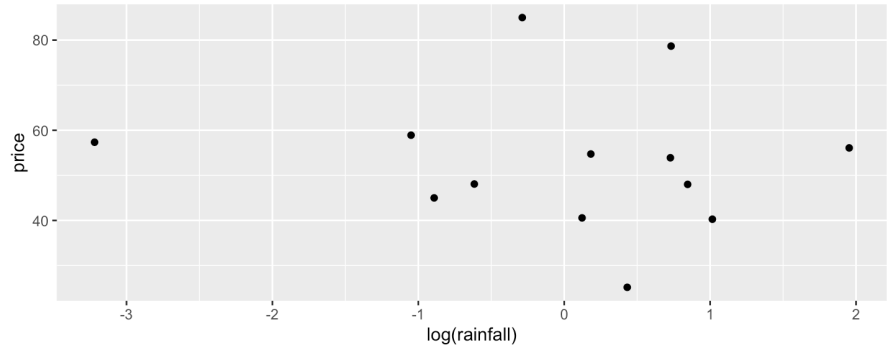


Problem 11

Create a scatter (i.e., point) plot that shows the relationship between the log of september rainfall and average price for wines in the Dundee Hills region. Does it look like there is a relationship?

Solution

```
read_csv("../resources/rainfall.csv") %>%
  pivot_longer(-Year,names_to = "month", values_to = "rainfall") %>%
  rename("year"="Year") %>%
  filter(month=="Sep") %>%
  right_join(wine) %>%
  filter(region_1=="Dundee Hills") %>%
  group_by(rainfall) %>%
  summarise(price=mean(price)) %>%
  ggplot(aes(log(rainfall),price))+
    geom_point()
```



Bonus

What are the top scoring wines from the ten french wineries with the highest average prices? Only include wineries that produce 5 or more wines.

Note: For full credit, use 10 operations (piped lines of code) or less.

Solution

```
wine %>%
  filter(country=="France") %>%
  group_by(winery) %>%
  summarize(
    count=n(),
    avg_price=mean(price)) %>%
  filter(count>=5) %>%
  top_n(10,avg_price) %>%
  left_join(wine) %>%
  arrange(winery,desc(points)) %>%
  group_by(winery) %>%
  summarize(title=first(title), points=first(points))
```

```
## # A tibble: 10 x 3
##    winery                title                                        points
##    <chr>                 <chr>                                         <dbl>
##  1 Château Haut-Brion    Château Haut-Brion 2014  Pessac-Léognan         100
##  2 Château La Mission H… Château La Mission Haut-Brion 2009  Pessac…      97
##  3 Château Lafite Roths… Château Lafite Rothschild 2010  Pauillac        100
##  4 Château Margaux       Château Margaux 2009  Margaux                   98
##  5 Château Mouton Roths… Château Mouton Rothschild 2009  Pauillac         96
##  6 Château Trotanoy      Château Trotanoy 2009  Pomerol                   96
##  7 Domaine Jacques Prie… Domaine Jacques Prieur 2009 Musigny  (Musi…      96
##  8 Domaine Leflaive      Domaine Leflaive 2010  Bâtard-Montrachet         99
##  9 Krug                  Krug 2002 Brut  (Champagne)                     100
## 10 Salon                 Salon 2006 Le Mesnil Blanc de Blancs Brut …    100
```

Alternative solution (and upgrade)

```
wine %>%
  filter(country=="France") %>%
  group_by(winery) %>%
  mutate(
    count=n(),
    avg_price=mean(price)) %>%
  filter(count>=5) %>%
  arrange(winery,desc(points)) %>%
  summarize(title=first(title), points=first(points), avg_price=first(avg_price)) %>%
  top_n(10,avg_price)
```

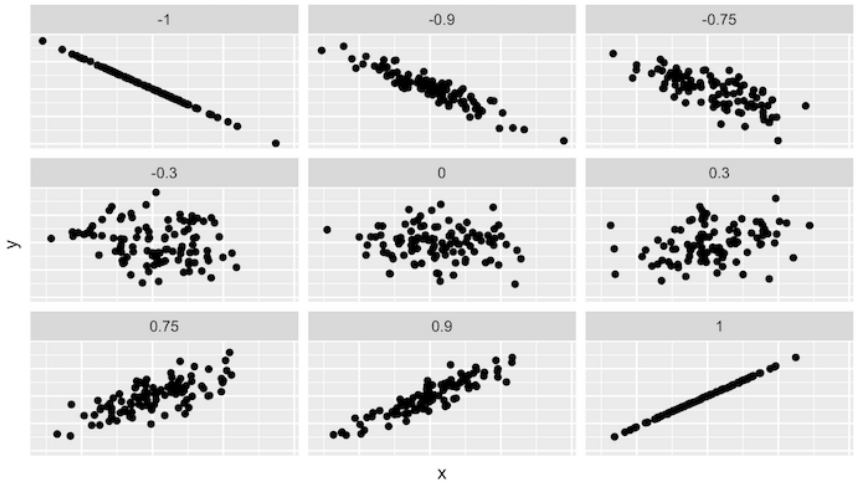
```
## # A tibble: 10 x 4
##    winery              title                                points avg_price
##    <chr>               <chr>                                 <dbl>     <dbl>
##  1 Château Haut-Brion  Château Haut-Brion 2014  Pessac-Lé…    100      572.
##  2 Château La Mission… Château La Mission Haut-Brion 2009…     97      546.
##  3 Château Lafite Rot… Château Lafite Rothschild 2010  Pa…    100      472.
##  4 Château Margaux     Château Margaux 2009  Margaux           98      448.
##  5 Château Mouton Rot… Château Mouton Rothschild 2009  Pa…     96      479.
##  6 Château Trotanoy    Château Trotanoy 2009  Pomerol          96      325
##  7 Domaine Jacques Pr… Domaine Jacques Prieur 2009 Musign…     96      261.
##  8 Domaine Leflaive    Domaine Leflaive 2010  Bâtard-Mont…     99      259.
##  9 Krug                Krug 2002 Brut  (Champagne)            100      331.
## 10 Salon               Salon 2006 Le Mesnil Blanc de Blan…    100      381.
```

Numerical predictors

Correlation

Credit: Modern Dive

http://guessthecorrelation.com/ …my high score is 122



Calculating correlation

```
library(moderndive)
wine %>% get_correlation(formula = price ~ points)
```

```
## # A tibble: 1 x 1
##   correlation
##         <dbl>
## 1       0.416
```

```
wine %>% summarise(correlation=cor(price,points))
```

```
## # A tibble: 1 x 1
##   correlation
##         <dbl>
## 1       0.416
```

```
wine %>% summarise(correlation=cor(lprice,points))
```

```
## # A tibble: 1 x 1
##   correlation
##         <dbl>
## 1       0.612
```

Exercise

1. Calculate the correlation between log(price) and points
2. by variety
3. for Oregon Chardonnay, Pinot Noir and Pinot Gris
4. in the same tibble

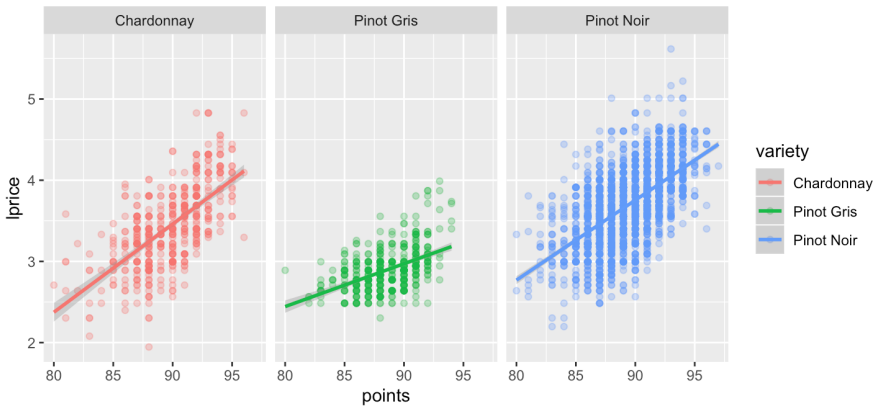
Solution

```
wine %>%
  filter(province=="Oregon") %>%
  filter(variety %in% c("Chardonnay","Pinot Noir","Pinot Gris")) %>%
  group_by(variety) %>%
  summarise(correlation=cor(lprice,points))
```

```
## # A tibble: 3 x 2
##   variety    correlation
##   <chr>            <dbl>
## 1 Chardonnay       0.646
## 2 Pinot Gris       0.486
## 3 Pinot Noir       0.592
```

Visualizing these different correlations

```
wine %>%
  filter(province=="Oregon") %>%
  filter(variety %in% c("Chardonnay","Pinot Noir","Pinot Gris")) %>%
  ggplot(aes(points,lprice, color=variety)) +
    geom_point(alpha=0.3)+
    facet_wrap(~variety)+
    geom_smooth(method = lm)
```



Simple linear regression

```
model <- lm(lprice~points, filter(wine,province=="Oregon"))
get_regression_table(model)
```

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept   -4.87      0.19      -25.7       0   -5.24    -4.50
## 2 points       0.094     0.002      43.9       0    0.089    0.098
```

Interpreting the coefficients

```
pct = (exp(coef(model)["points"]) - 1) * 100
```

Since we logged the DV, a 1 point ratings increase = `9.81`% increase in price on average.

Note:

$$(e^x - 1) * 100$$

Exercise

1. Pair up and calculate the percent increase in price due a 1 point increase in quality
2. for Oregon Chardonnay, Pinot Gris and Pinot Noir

Solution (and upgrade)

```
model = list()
for(v in c("Chardonnay", "Pinot Gris","Pinot Noir")){
  model[[v]] <- lm(lprice~points, filter(wine,province=="Oregon", variety==v))
  print(get_regression_table(model[[v]]))
}
```

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept   -6.32      0.518     -12.2       0   -7.34     -5.30
## 2 points       0.109     0.006      18.8       0    0.097     0.12
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept   -1.78      0.388     -4.59       0   -2.54    -1.02
## 2 points       0.053     0.004     12.1        0    0.044    0.061
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept   -5.11      0.228     -22.4       0   -5.56    -4.66
## 2 points       0.099     0.003      38.7       0    0.094    0.103
```

Human-readable solution

```
for(v in names(model)){
  pct <- round((exp(coef(model[[v]])["points"]) - 1) * 100,2)
  print(str_c("For ",v,", a 1 point ratings increase leads to a ",pct,"% increase in price."))
}
```

```
## [1] "For Chardonnay, a 1 point ratings increase leads to a 11.48% increase in price."
## [1] "For Pinot Gris, a 1 point ratings increase leads to a 5.42% increase in price."
## [1] "For Pinot Noir, a 1 point ratings increase leads to a 10.35% increase in price."
```

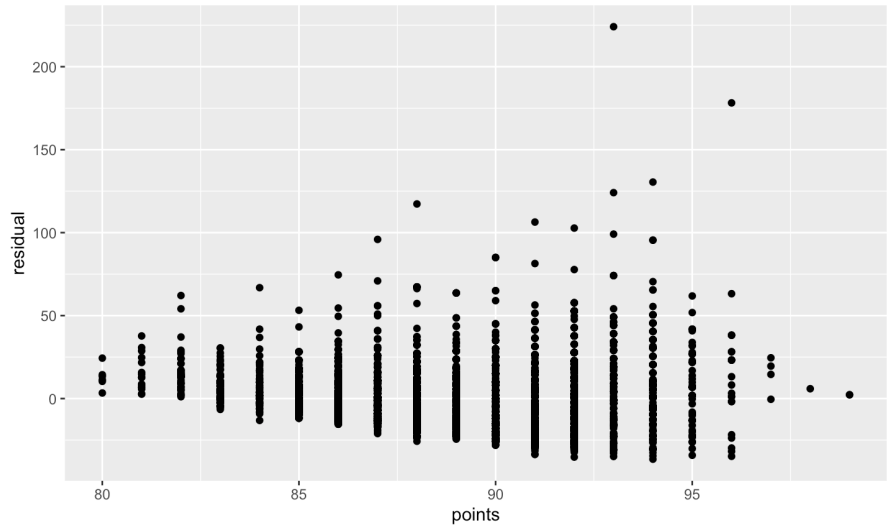
Looking at the residuals

```
model <- lm(lprice~points, filter(wine,province=="Oregon"))
get_regression_points(model)
```

```
## # A tibble: 5,359 x 5
##       ID lprice points lprice_hat residual
##    <int>  <dbl>  <dbl>      <dbl>    <dbl>
##  1     1   2.64     87       3.27   -0.635
##  2     2   4.17     87       3.27    0.9
##  3     3   3.00     87       3.27   -0.278
##  4     4   3.91     86       3.18    0.732
##  5     5   3.09     86       3.18   -0.089
##  6     6   3.22     86       3.18    0.039
##  7     7   3.64     91       3.65   -0.011
##  8     8   3.33     85       3.09    0.246
##  9     9   3.81     85       3.09    0.72
## 10    10   3.09     85       3.09    0.004
## # … with 5,349 more rows
```

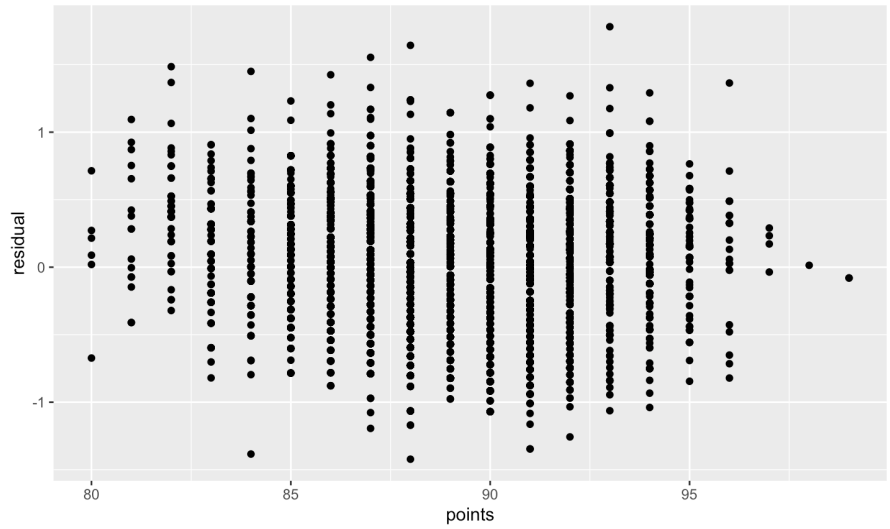
Graphing residuals (bad)

```
model <- lm(price~points, filter(wine,province=="Oregon"))
get_regression_points(model) %>%
  ggplot(aes(points, residual))+
    geom_point()
```



Graphing residuals (good)

```
model <- lm(lprice~points, filter(wine,province=="Oregon"))
get_regression_points(model) %>%
  ggplot(aes(points, residual))+
    geom_point()
```

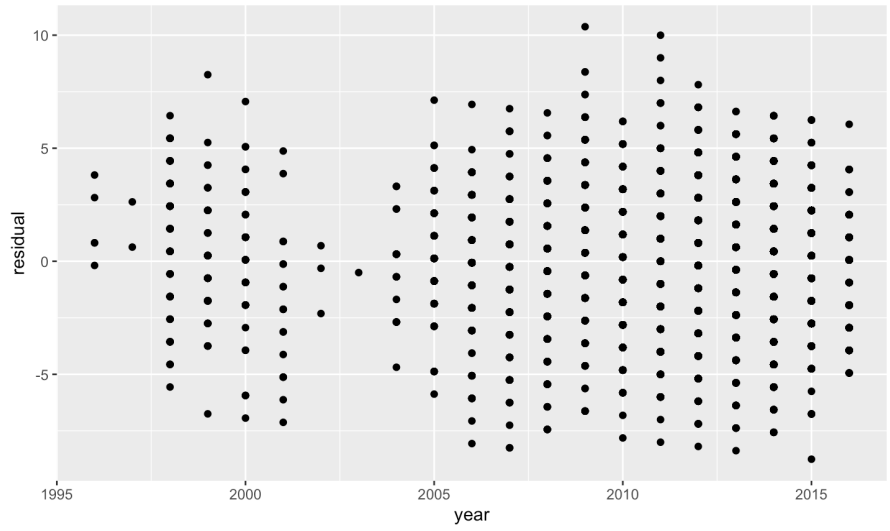


Exercise

1. model the relationship between year and points
2. after 1995
3. for oregon wine
4. then graph the residuals

Solution

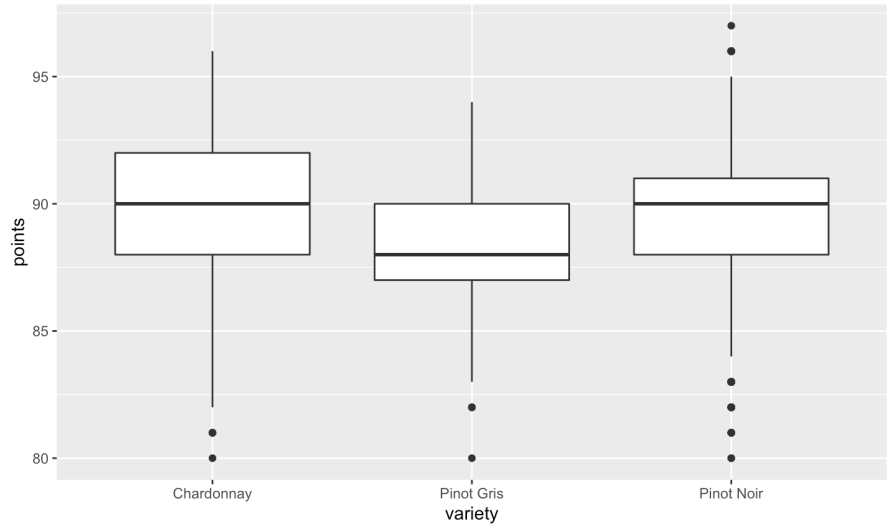
```
model <- lm(points~year, filter(wine,province=="Oregon",year>1995))
get_regression_points(model) %>%
  ggplot(aes(year, residual))+
    geom_point()
```



Categorical predictors

Graphing points by variety

```
wine %>%
  filter(province=="Oregon") %>%
  filter(variety %in% c("Chardonnay","Pinot Noir","Pinot Gris")) %>%
  ggplot(aes(variety,points))+
    geom_boxplot()
```



Summary

```
(tmp <- wine %>%
  filter(province=="Oregon") %>%
  filter(variety %in% c("Chardonnay","Pinot Noir","Pinot Gris")) %>%
  group_by(variety) %>%
  summarise(mean=mean(points)))
```

```
## # A tibble: 3 x 2
##   variety     mean
##   <chr>      <dbl>
## 1 Chardonnay  89.7
## 2 Pinot Gris  88.5
## 3 Pinot Noir  89.5
```

Note:

1. The difference between Pinot Gris and Chardonnay is -1.1829692
2. The difference between Pinot Noir and Chardonnay is -0.2433158

Regression

```
model <- lm(points~variety,
            filter(wine,province=="Oregon",variety %in% c("Chardonnay","Pinot Noir","Pinot Gris")))
get_regression_table(model)
```

```
## # A tibble: 3 x 7
##   term              estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept            89.7      0.12     751.     0       89.5     90.0
## 2 varietyPinot Gris    -1.18     0.171     -6.90   0       -1.52    -0.847
## 3 varietyPinot Noir    -0.243    0.13      -1.88   0.061   -0.498    0.011
```

What is the equation for this regression?

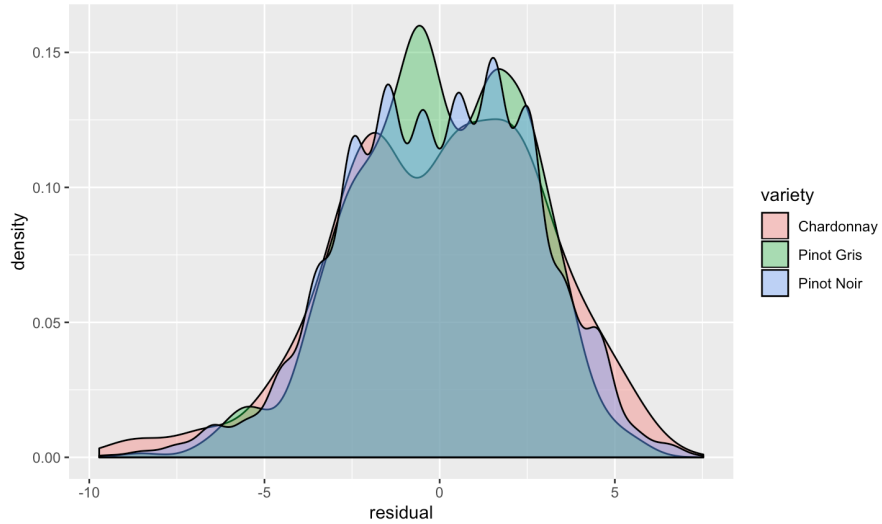
Residuals

```
get_regression_points(model)
```

```
## # A tibble: 3,749 x 5
##       ID points variety    points_hat residual
##    <int>  <dbl> <chr>           <dbl>    <dbl>
##  1     1     87 Pinot Gris       88.5    -1.53
##  2     2     87 Pinot Noir       89.5    -2.47
##  3     3     87 Pinot Noir       89.5    -2.47
##  4     4     86 Pinot Noir       89.5    -3.47
##  5     5     86 Pinot Noir       89.5    -3.47
##  6     6     86 Pinot Noir       89.5    -3.47
##  7     7     91 Pinot Noir       89.5     1.53
##  8     8     85 Pinot Noir       89.5    -4.47
##  9     9     85 Pinot Noir       89.5    -4.47
## 10    10     85 Pinot Noir       89.5    -4.47
## # … with 3,739 more rows
```

Plot residuals

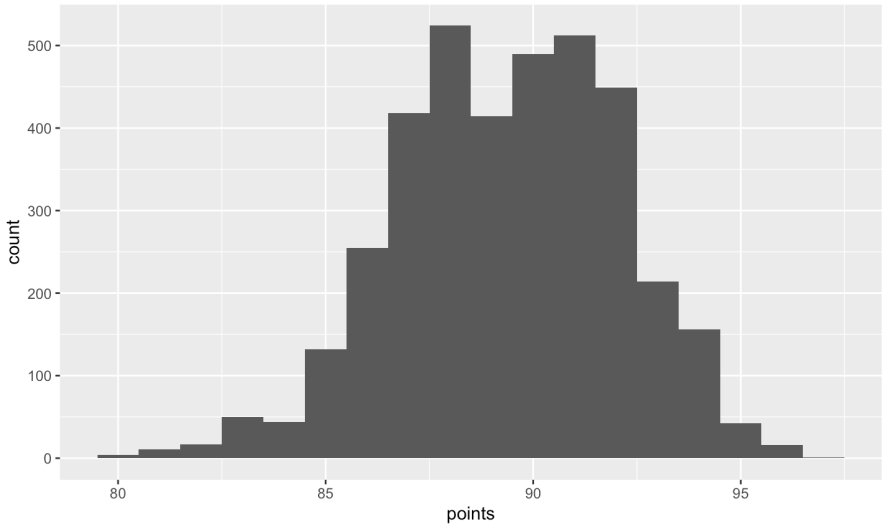
```
get_regression_points(model) %>%
  ggplot(aes(residual, fill=variety))+
    geom_density(alpha=0.4)
```



What is causing that left skew?

Points histogram

```
wine %>%
filter(province=="Oregon") %>%
filter(variety %in% c("Chardonnay","Pinot Noir","Pinot Gris")) %>%
ggplot(aes(points))+
  geom_histogram(binwidth = 1)
```

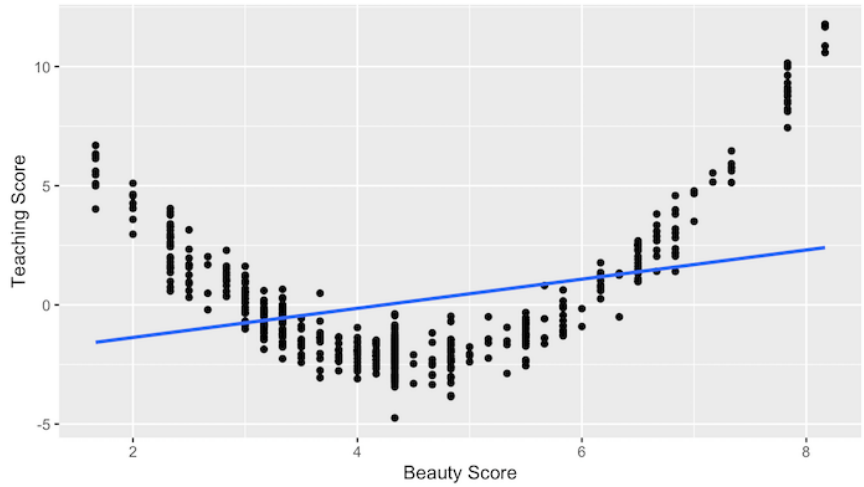


Assumptions of linear regression

1. **L**inearity of relationship between variables
2. **I**ndependence of the residuals
3. **N**ormality of the residuals
4. **E**quality of variance of the residuals

Linearity of relationship

Credit: Modern Dive



What would the residuals look like?

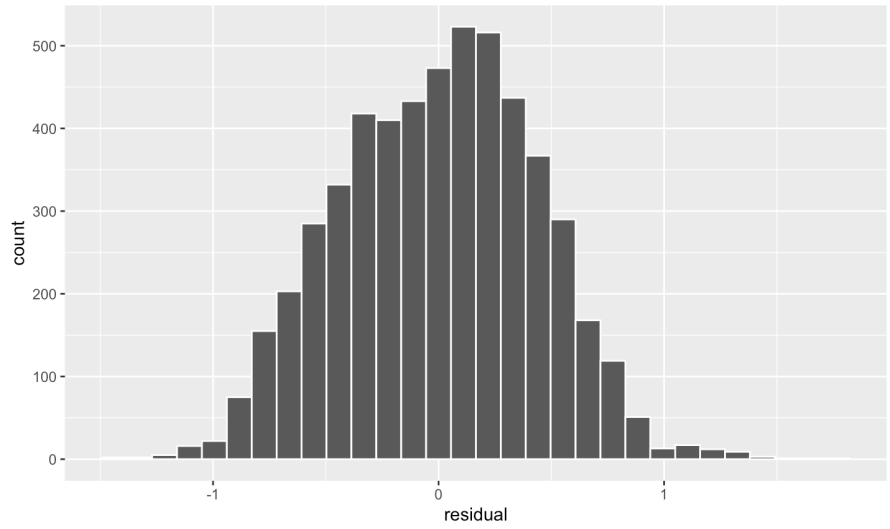
Independence

Given our original model of $log(price) = m * Points + b$…

…are there any problems with independence?

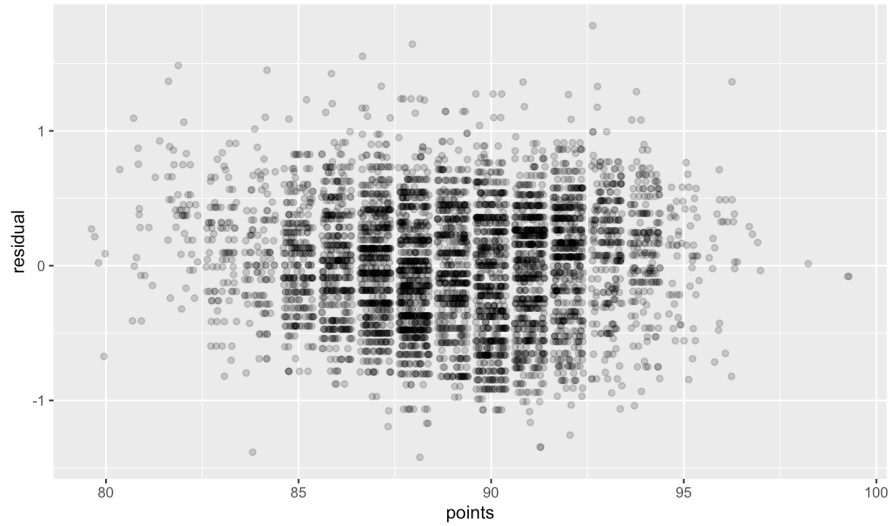
Normality

```
model <- lm(lprice~points, filter(wine,province=="Oregon"))
get_regression_points(model) %>%
  ggplot(aes(residual))+
    geom_histogram(color="white")
```



Equality of variance

```
get_regression_points(model) %>%
  ggplot(aes(points, residual))+
    geom_jitter(alpha=0.2)
```



No equality in the variance

Credit: Modern Dive