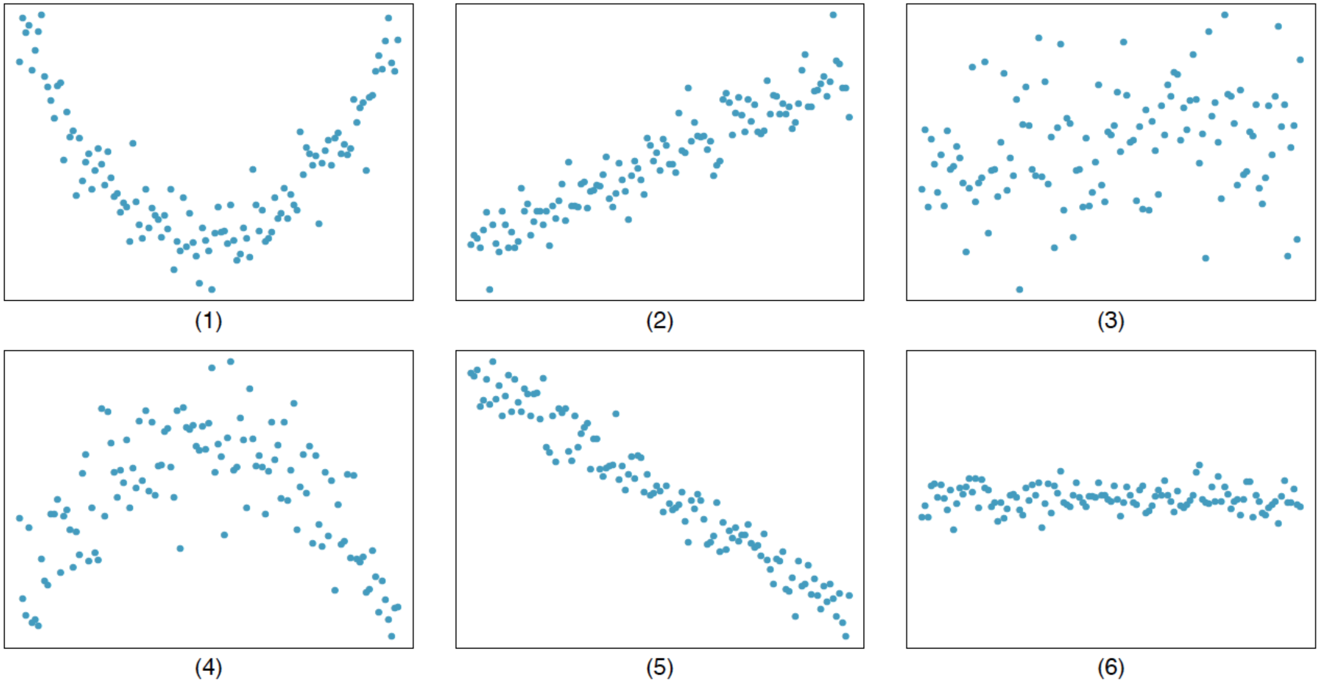
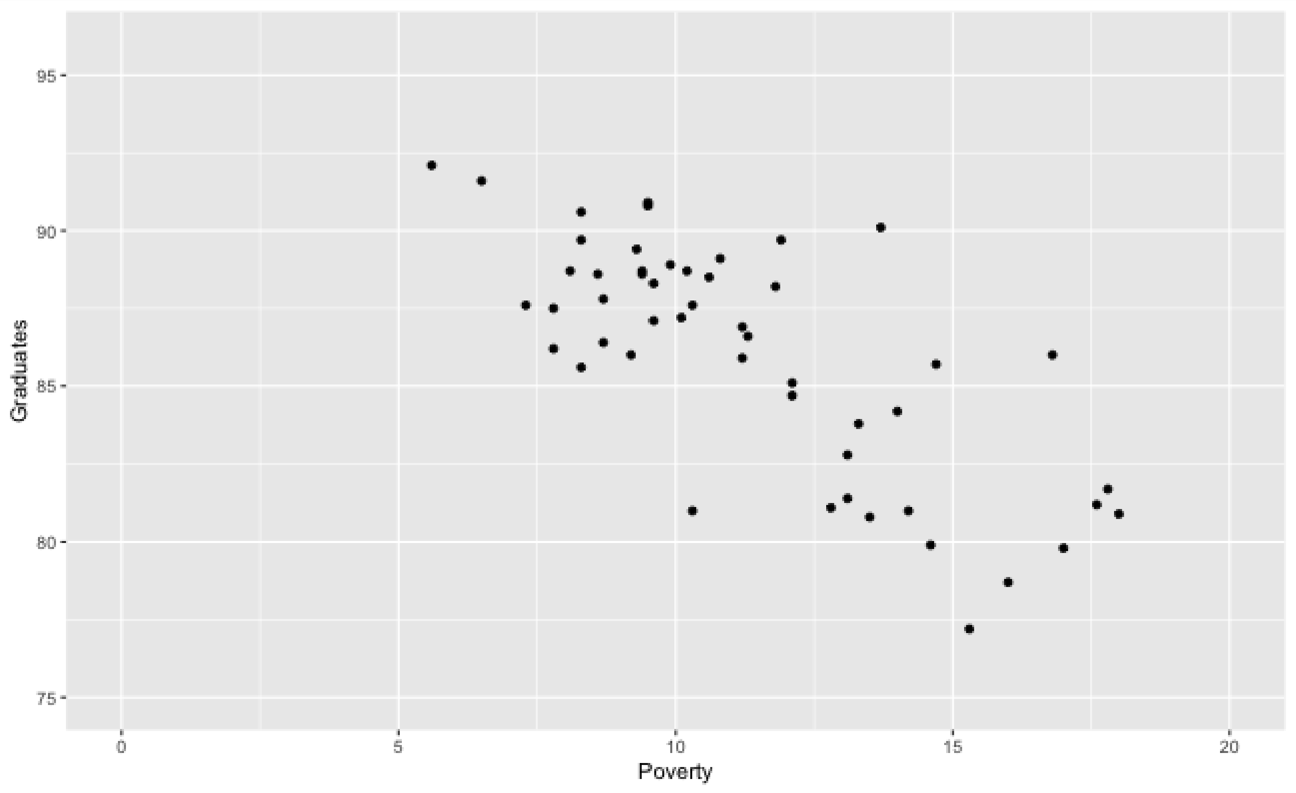
Simple Linear Regression II

Identify relationships For each of the six plots, identify the strength of the relationship (e.g. weak, moderate, or strong) in the data and whether fitting a linear model would be reasonable.

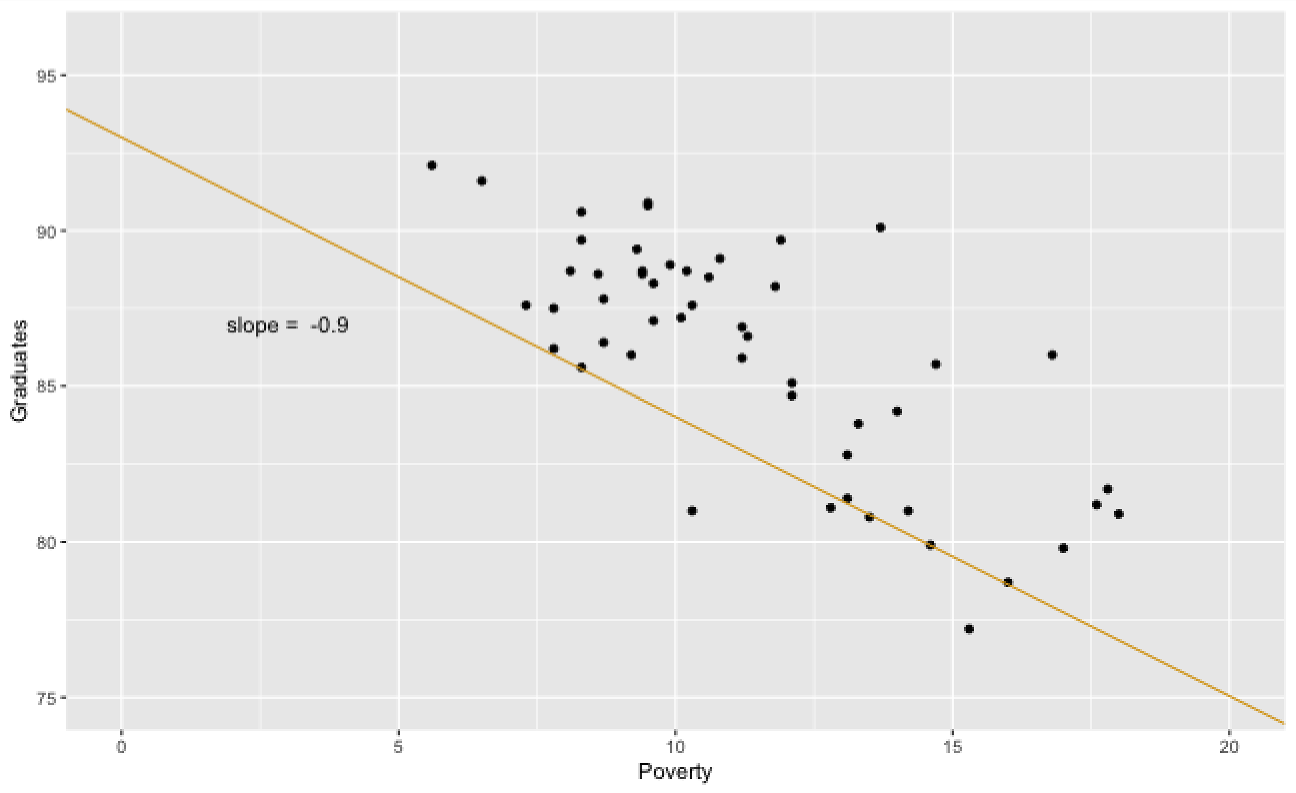


Estimating β_1



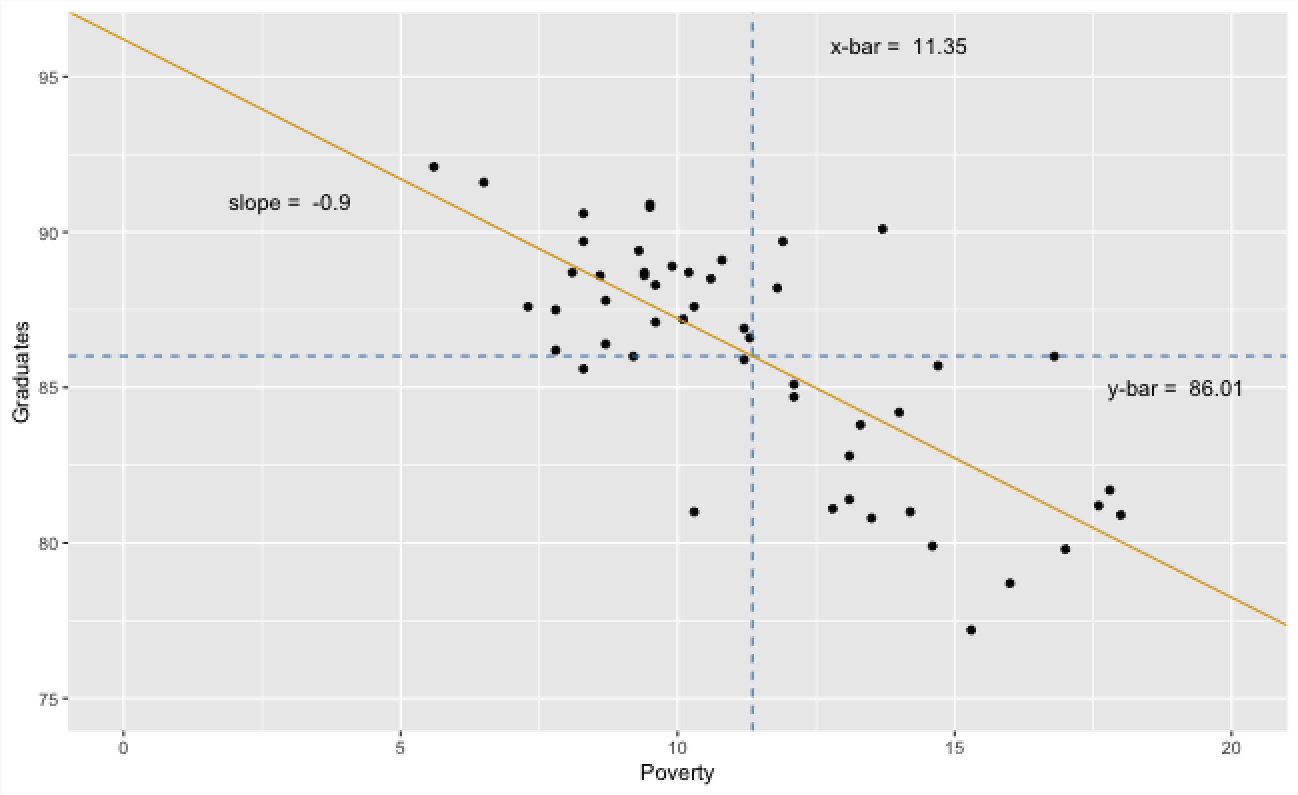
We use s_x, s_y, and R to calculate b_1.

Estimating β_1



We use s_x, s_y, and R to calculate b_1.

Estimating β_0



If the line of best fit *must* pass through $(\bar{x}, \bar{y})$, what is b_0?

Estimating β_0, cont.

Since $(11.35, 86.01)$ is on the line, the following relationship holds.

$$86.01 = b_0 - 0.9(11.35)$$

Then just solve for b_0.

$$b_0 = 86.01 + 0.9(11.35) = 96.22$$

More generally:

$$b_0 = \bar{y} - b_1\bar{x}$$

Estimation in R

```
m1 <- lm(Graduates ~ Poverty, data = poverty)
summary(m1)
```

```
##
## Call:
## lm(formula = Graduates ~ Poverty, data = poverty)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.954 -1.820  0.544  1.515  6.199
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   96.202      1.343   71.65  < 2e-16 ***
## Poverty       -0.898      0.114   -7.86  3.1e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0
##
## Residual standard error: 2.5 on 49 degrees of freedom
```

The `lm` object

```
attributes(m1)
```

```
## $names
##  [1] "coefficients"  "residuals"     "effects"       "
##  [5] "fitted.values" "assign"        "qr"            "
##  [9] "xlevels"       "call"          "terms"         "
##
## $class
## [1] "lm"
```

```
m1$coef
```

```
## (Intercept)     Poverty
##      96.202      -0.898
```

```
m1$fit
```

Interpretation of b_1

The **slope** describes the estimated difference in the y variable if the explanatory variable x for a case happened to be one unit larger.

```
m1$coef[2]
```

```
## Poverty
##  -0.898
```

For each additional percentage point of people living below the poverty level, we expect a state to have a proportion of high school graduates that is 0.898 lower.

Be Cautious: if it is observational data, you do not have evidence of a *causal link*, but of an association, which still can be used for prediction.

Interpretation of b_0

The **intercept** is the estimated y value that will be taken by a case with an x value of zero.

```
m1$coef[1]
```

```
## (Intercept)
##        96.2
```

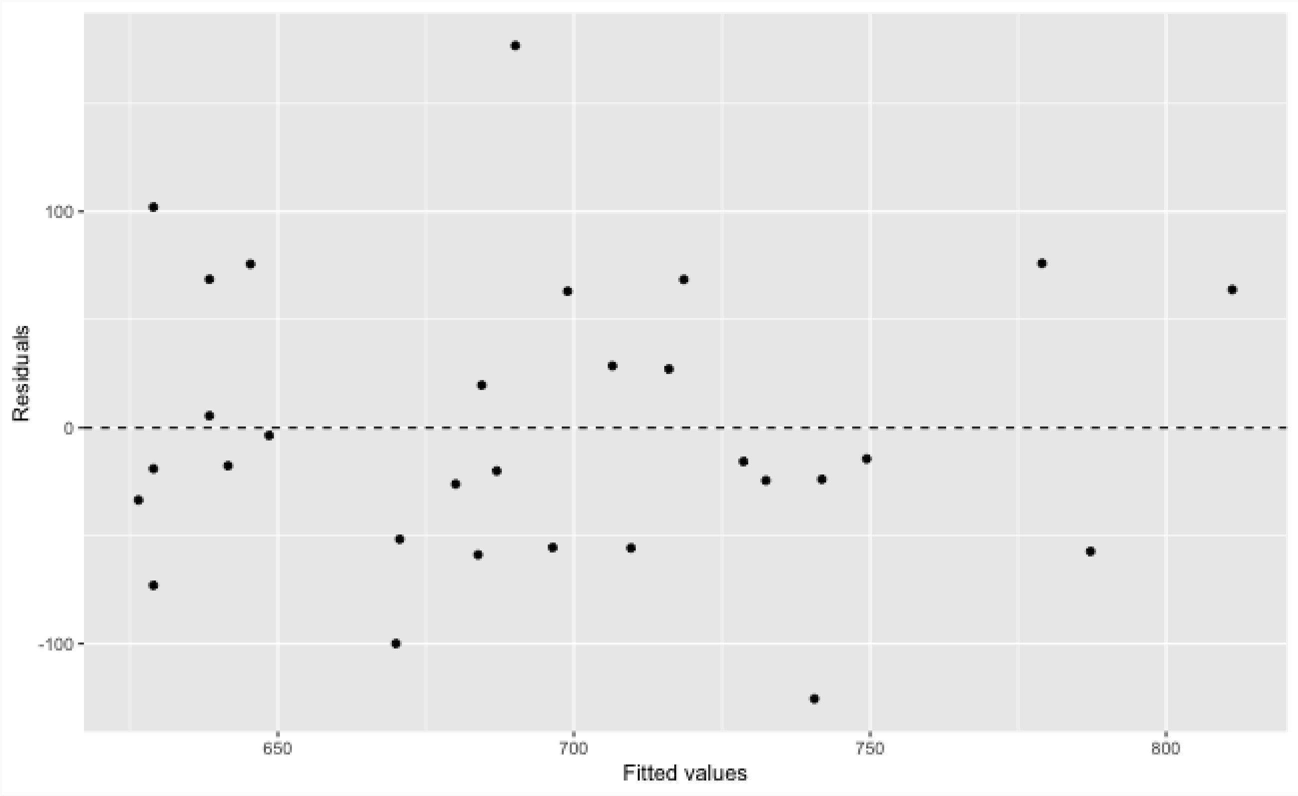
While necessary for prediction, the intercept often has no meaningful interpretation.

boardwork

Residual plot

```
m1 <- lm(runs ~ at_bats, data = mlb11)
ggplot(m1, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0,
             linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")
```

Residual plot

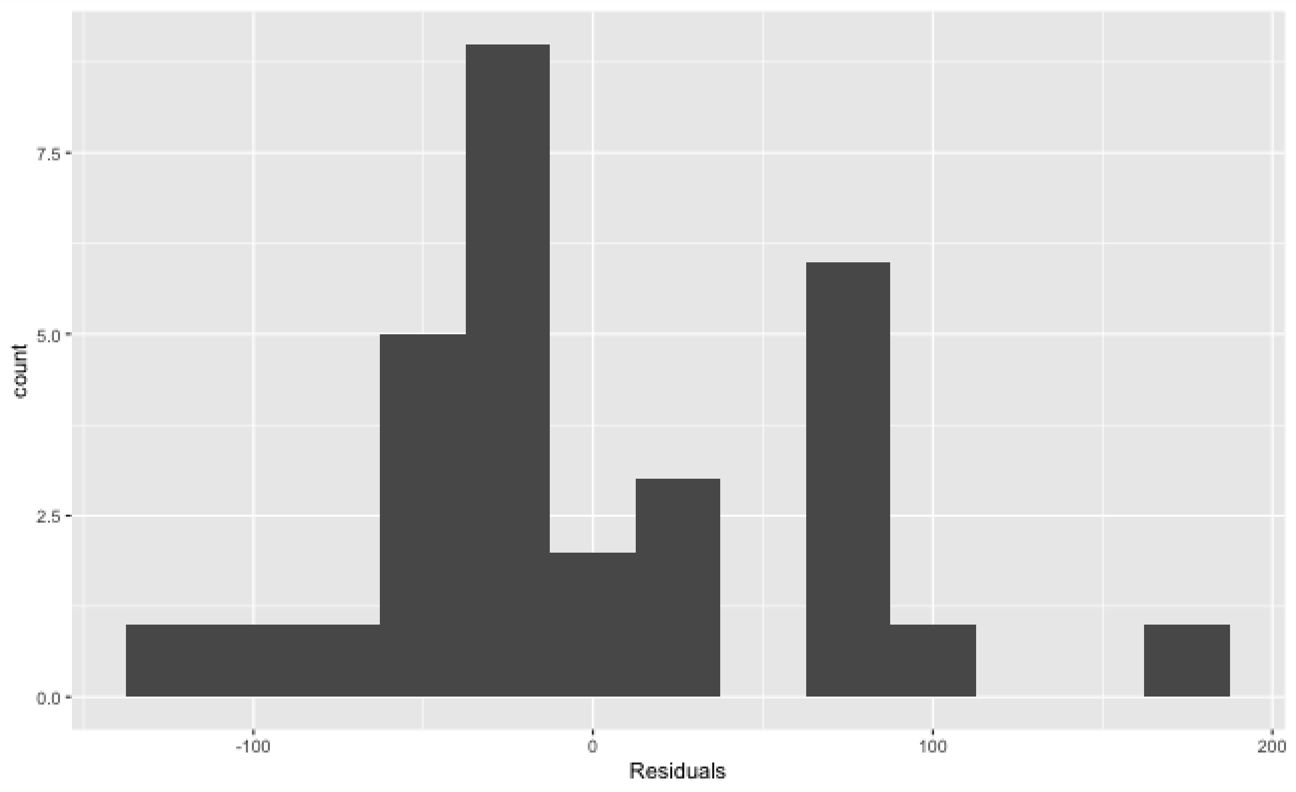


Distribution of the residuals

```
ggplot(m1, aes(x = .resid)) +
  geom_histogram(binwidth = 25) +
  xlab("Residuals")

ggplot(m1, aes(sample = .resid)) +
  geom_point(stat = "qq")
```

Distribution of the residuals



QQ plot