PSY 503: Foundations of Statistical Methods in Psychological Science

Modeling Continuous Relationships: Correlation and Linear Modeling

Jason Geller, Ph.D. (he/him/his)

Princeton University

Updated:2022-10-26

Housekeeping

- Data
  - Please let me know what data you intend to re-analyze (10/31)
- No more knowledge checks
  - Instead, you will submit 2-3 questions each week (Friday by 11:59 P.M.) over the material we have covered so far
  - What were 2-3 muddiest or unclear things from sessions this week? What are you still wondering about?

Power and Effect Size

- Branson's Question
    - 95% CI for Cohen's d: $[d - 1.96 \times \sigma(d), d + 1.96 \times \sigma(d)]$

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5133225/

Today

- Correlation
- Regression (linear modeling)

Dataset

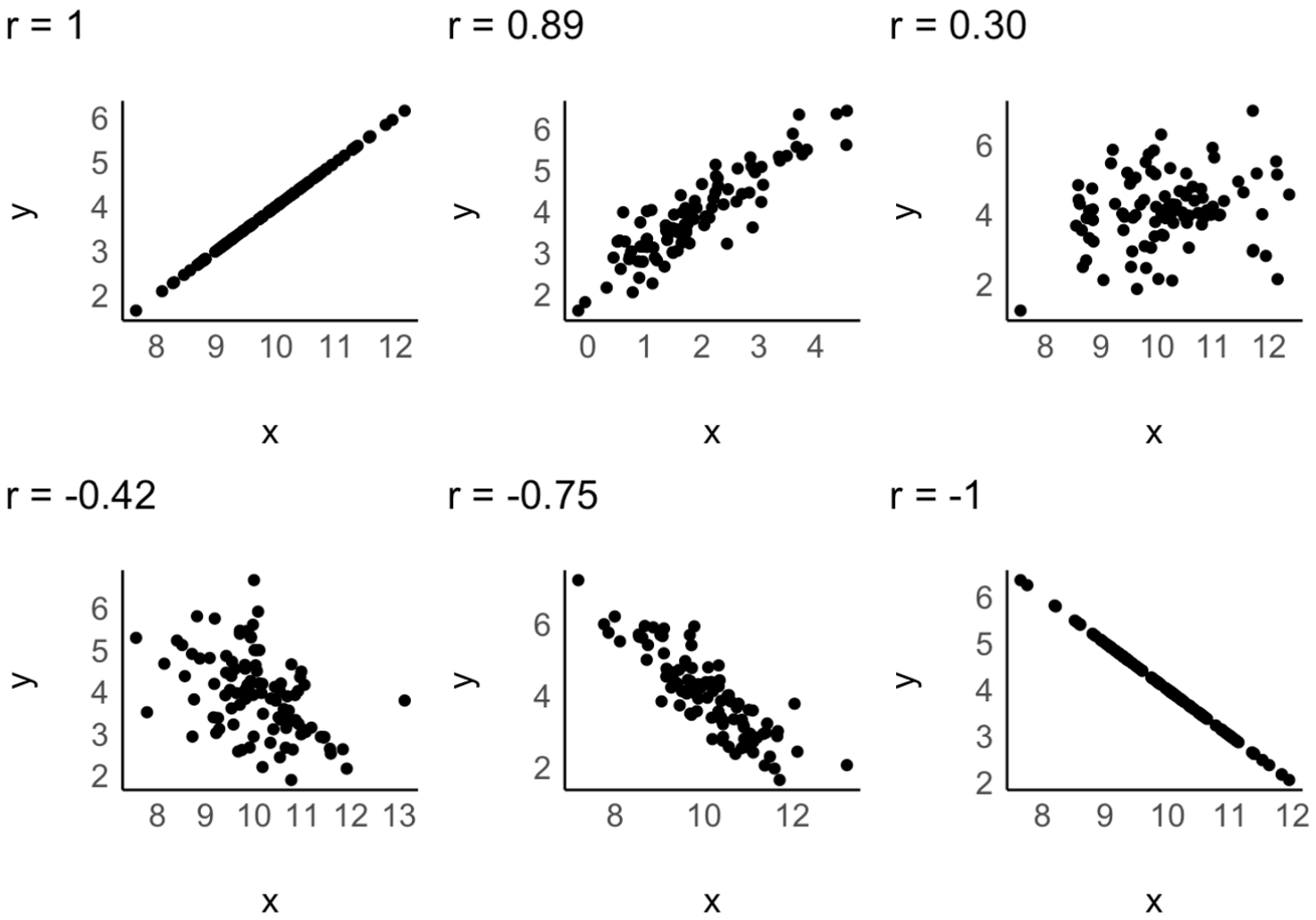
- Mental Health and Drug Use:
    - CESD = depression measure
    - PIL total = measure of meaning in life
    - AUDIT total = measure of alcohol use
    - DAST total = measure of drug usage

```
master <- read_csv("https://raw.githubusercontent.com/jgeller112/psy503-psych_stats/master/static/slides/10-lir
```

Correlation (*r*)

- Quantifies relationship between two variables
  - Direction (positive or negative)
  - Strength
- +1 is a perfect positive correlation
- 0 is no correlation (independence)
- -1 is a perfect negative correlation

Correlations



Effect Size Heuristics

- $r < 0.1$ very small
- $0.1 \leq r < 0.3$ small
- $0.3 \leq r < 0.5$ moderate
- $r \geq 0.5$ large

Covariance and Correlation

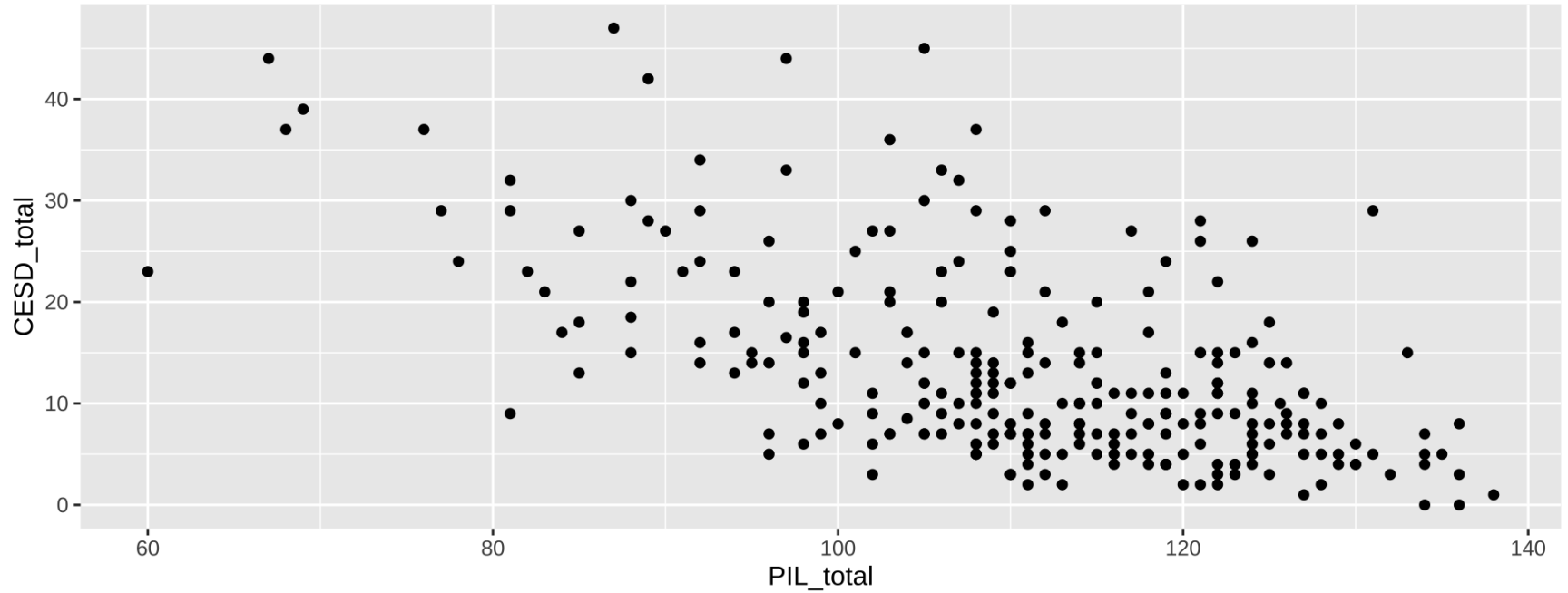
- Pearson's *r*

$$covariance = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

$$r = \frac{covariance}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y}$$

Dataset

- CESD = depression measure
- PIL total = measure of meaning in life
  - What do you think relationship looks like?



Statistical Test: Pearson's *r*

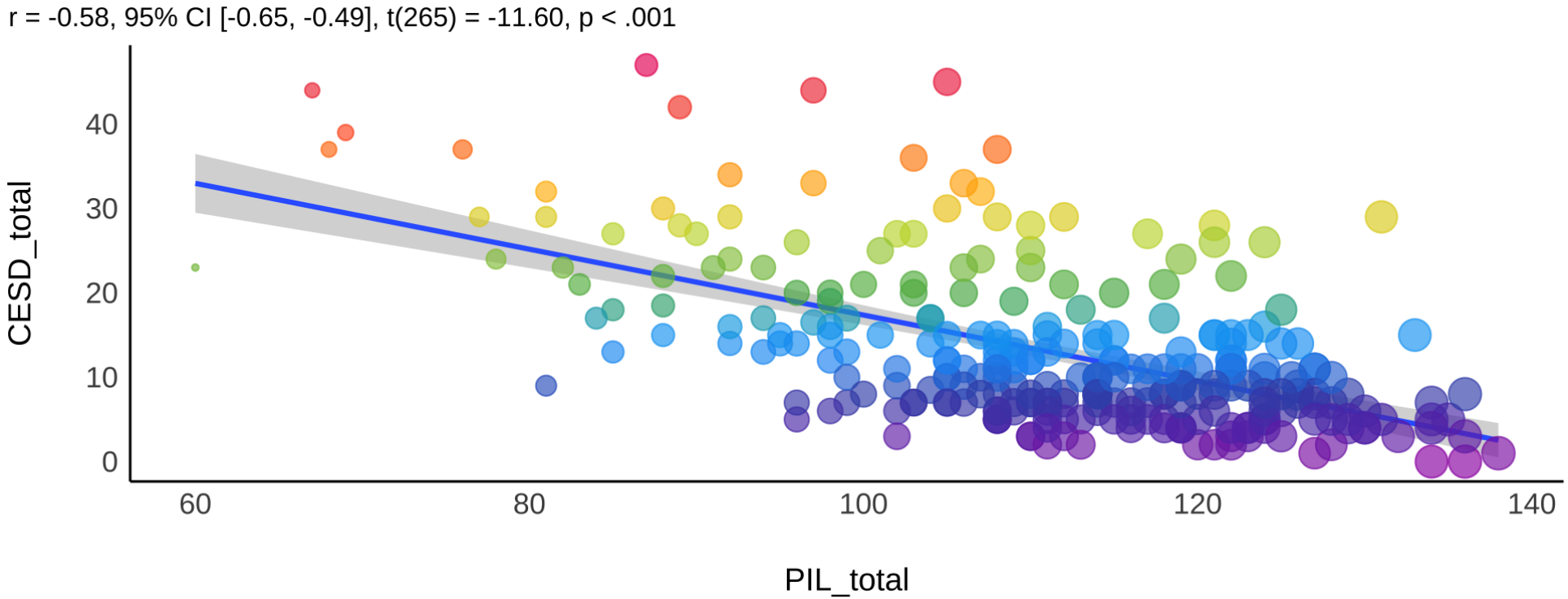
- H_0 $r = 0$
- H_1 $r \neq 0$

$$t_r = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

```
library(correlation) # easystats\

cor_result <- cor_test(master,"PIL_total", "CESD_total")
```

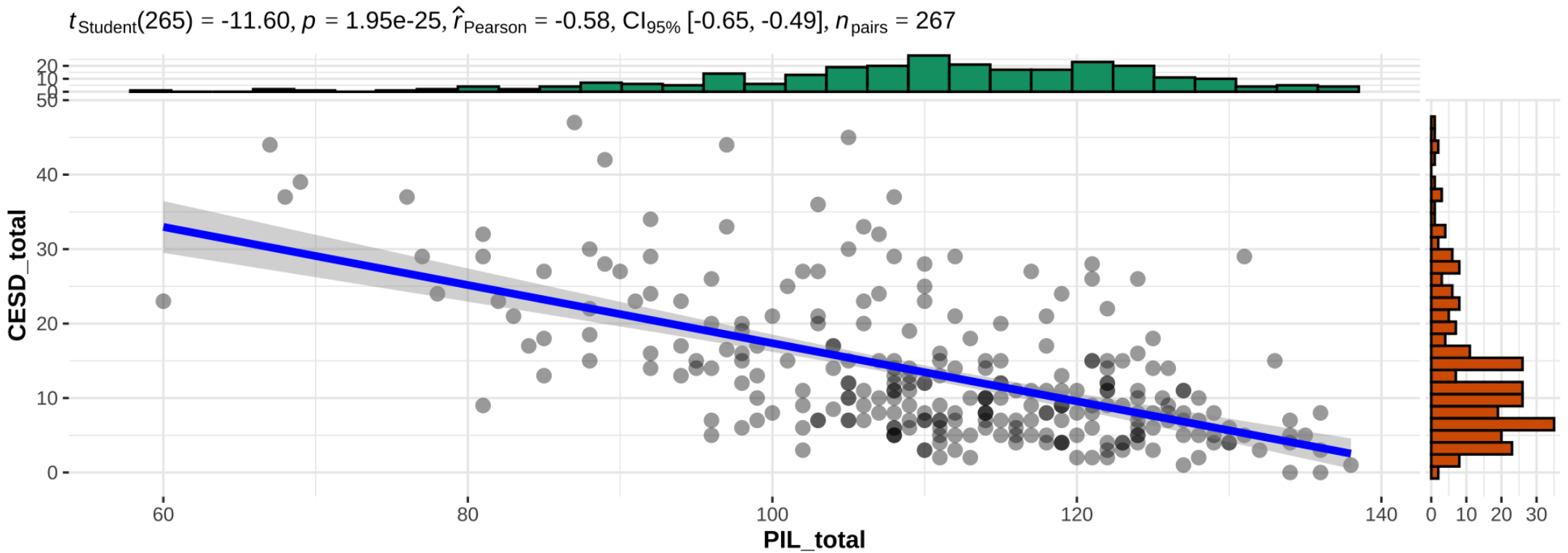
- Let's open R

Scatterplot



Scatterplot

```
library(ggstatsplot)

ggstatsplot::ggscatterstats(master,
                            x= "PIL_total",
                            y="CESD_total")
```



Non-paramteric Correlation

- Spearman’s rank correlation coefficient (ρ):

$$r_s = \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

- It assesses how well the relationship between two variables can be described using a monotonic (increasing or decreasing) function
- Rank order method
- Range [-1,+1]

Statistical Test: Spearman's *r*

```
cor_result_s <- cor_test(master, "CESD_total", "PIL_total", method = "spearman")
```

- Let's go to R!

What is Linear Modeling?

- It is a model of the relationship between two or more variables
  - The model commonly used is a linear one
  - BUT! This does not preclude testing non-linear or non-additive effects
- A way of describing/explaining a phenomenon/relationship between variables
- A way of predicting the value of one variable from other variables

Describing a Straight Line

- We describe the relationship between variables using the equation of a straight line

$$Y = mX + b$$

$$Y = \beta_0 + \beta_1 X + \varepsilon$$
$$\varepsilon \sim N(0, \sigma^2)$$

Describing a Straight Line

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

with

- reponse variable
- predictor variable
- error term

Regression Paramters

- What is b_1?
  - Regression coefficient for the predictor
  - Gradient (slope) of the regression line
  - Tells us how much we would expect y to change given a one-unit change in x
  - Direction/Strength of Relationship
- What is b_0?
  - Intercept (value of Y when X(s) = 0)
  - Point at which the regression line crosses the Y-axis

The Best Fit Line and Least Squares

- Many lines could fit the data, but which is best?
    - The best fitting line is one that produces the "least squares", or minimizes the squared difference between X and Y
- We use a method known as least squares to obtain estimates of b_0 and b_1

The Relation Between Correlation and Regression

$$\hat{r} = \frac{covariance_{xy}}{s_x * s_y}$$

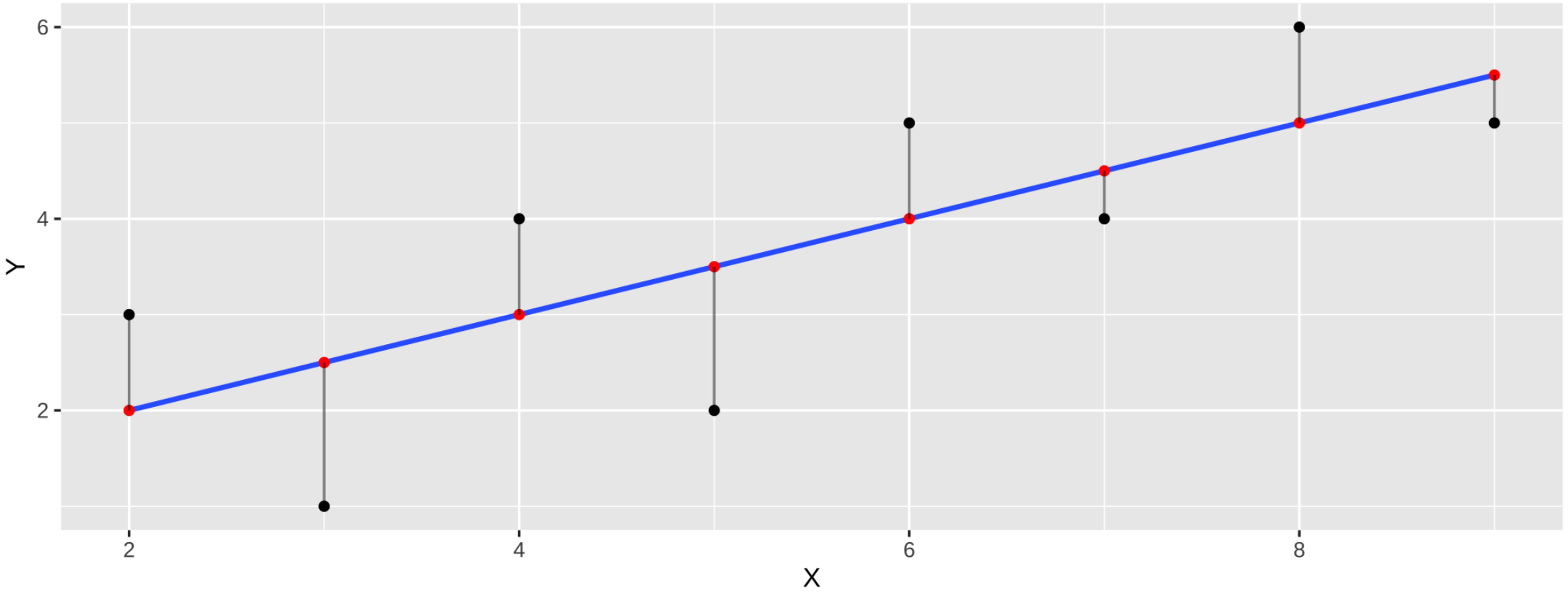
$$\hat{\beta}_x = \frac{covariance_{xy}}{s_x * s_y}$$

$$\hat{\beta}_x = \frac{\hat{r} * s_x * s_y}{s_x} = r * \frac{s_y}{s_x}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}_x$$

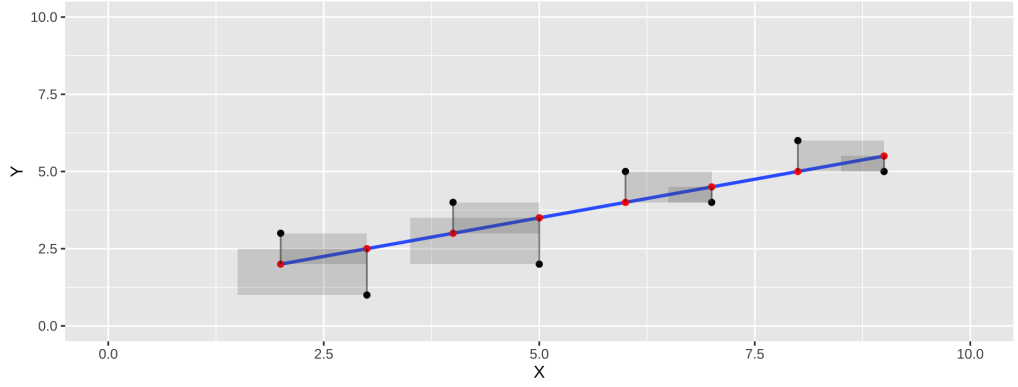
- What do squares have to do with it and why are they least squares?

Visualizing Error

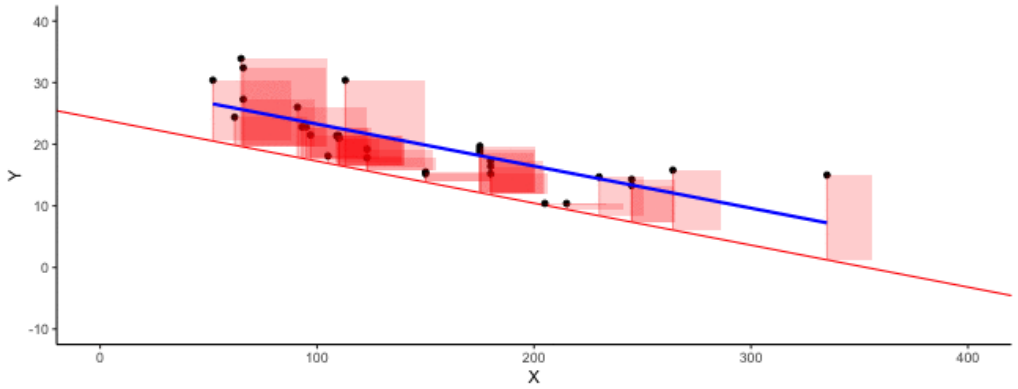


Visualize Errors as Squares

```
some_data <- data.frame(Y= c(1,2,4,3,5,4,6,5),
                        X= c(3,5,4,2,6,7,8,9)) %>%
  mutate(Y_pred = predict.lm(lm(Y~X))) %>%
  mutate(Y_error = Y - Y_pred)

g=ggplot(some_data, aes(x=X, y=Y))+
  geom_point()+
  geom_smooth(method='lm', se=FALSE)+
  geom_point(aes(y=Y_pred), color='red') +
  geom_segment(aes(xend = X, yend = Y-Y_error), alpl
  geom_rect(aes(ymin=Y,
                ymax=Y_pred,
                xmin=X,
                xmax=X+Y_error),
            alpha = .2)
```



Example

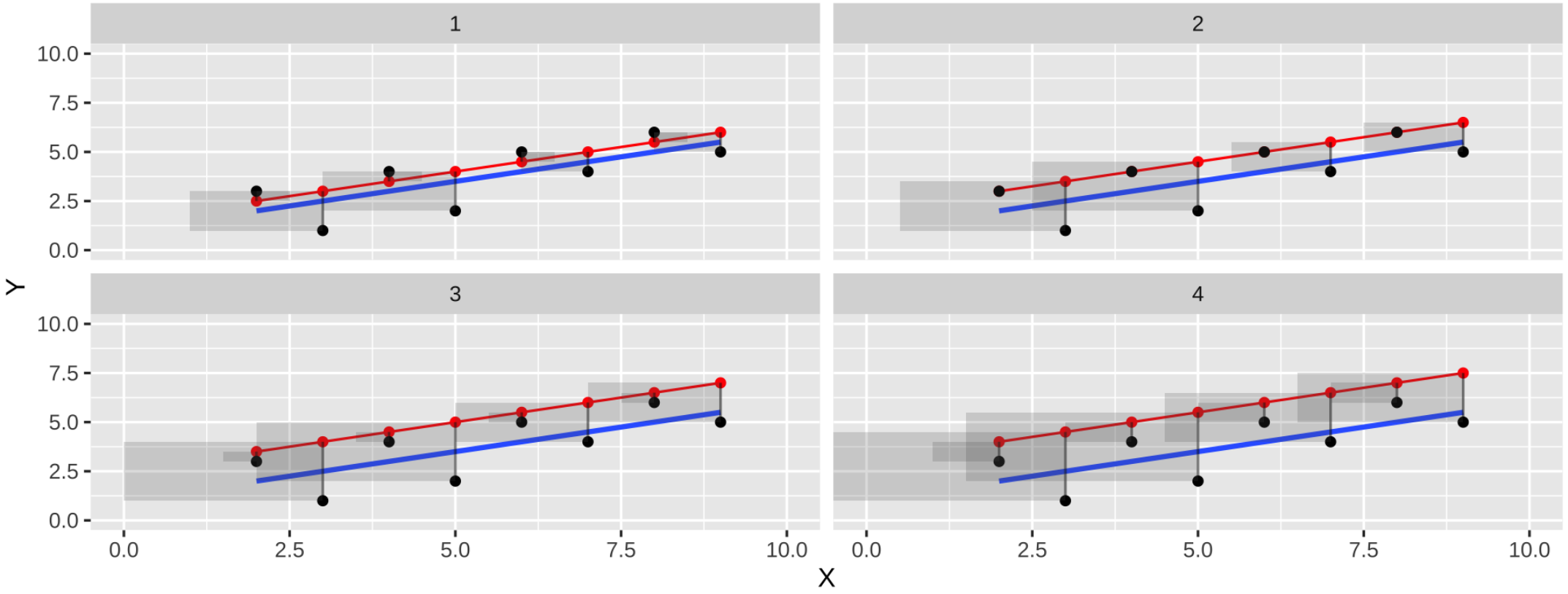


- Shows two concepts:

  1. Regression line is "best fit line"

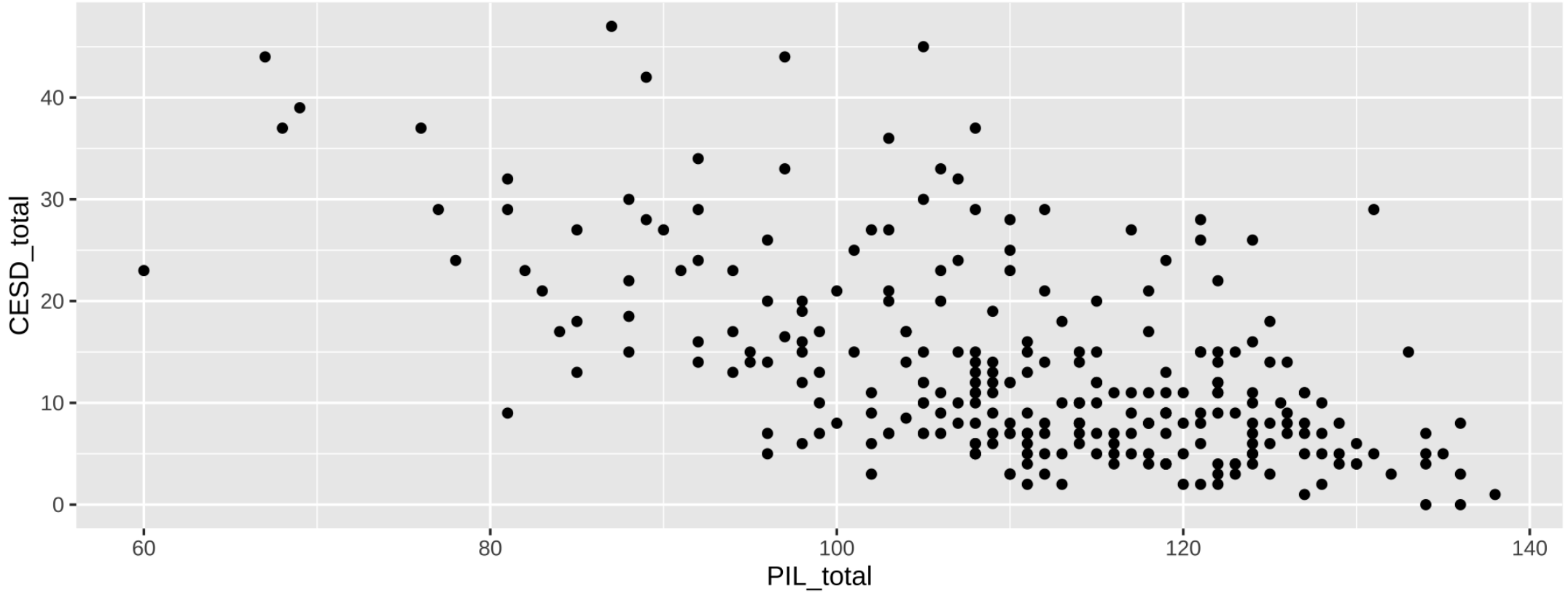
  2. The “best fit line” is the one that minimizes the sum of the squared deviations between each point and the line

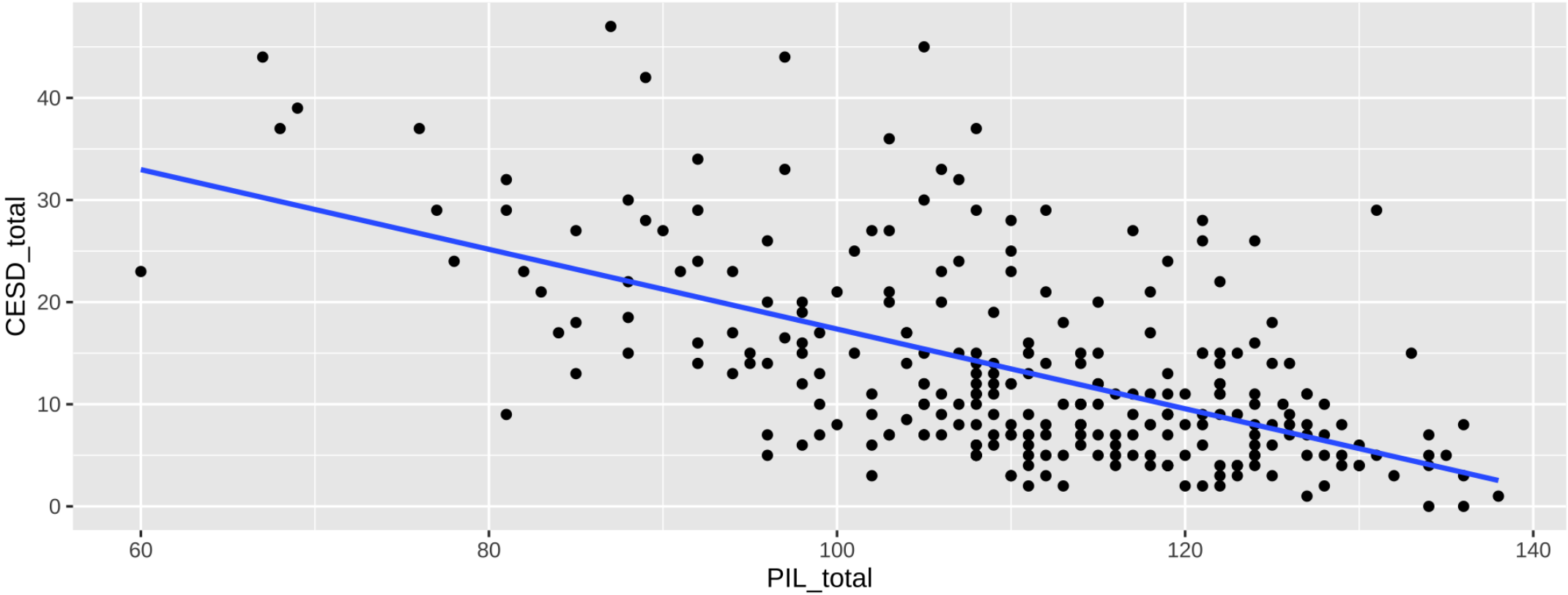
Worse Fit Lines



Simple Regression Example

- A linear model fit to data with a numeric X is classical regression
  - Depression scores and meaningfulness (in one's life)

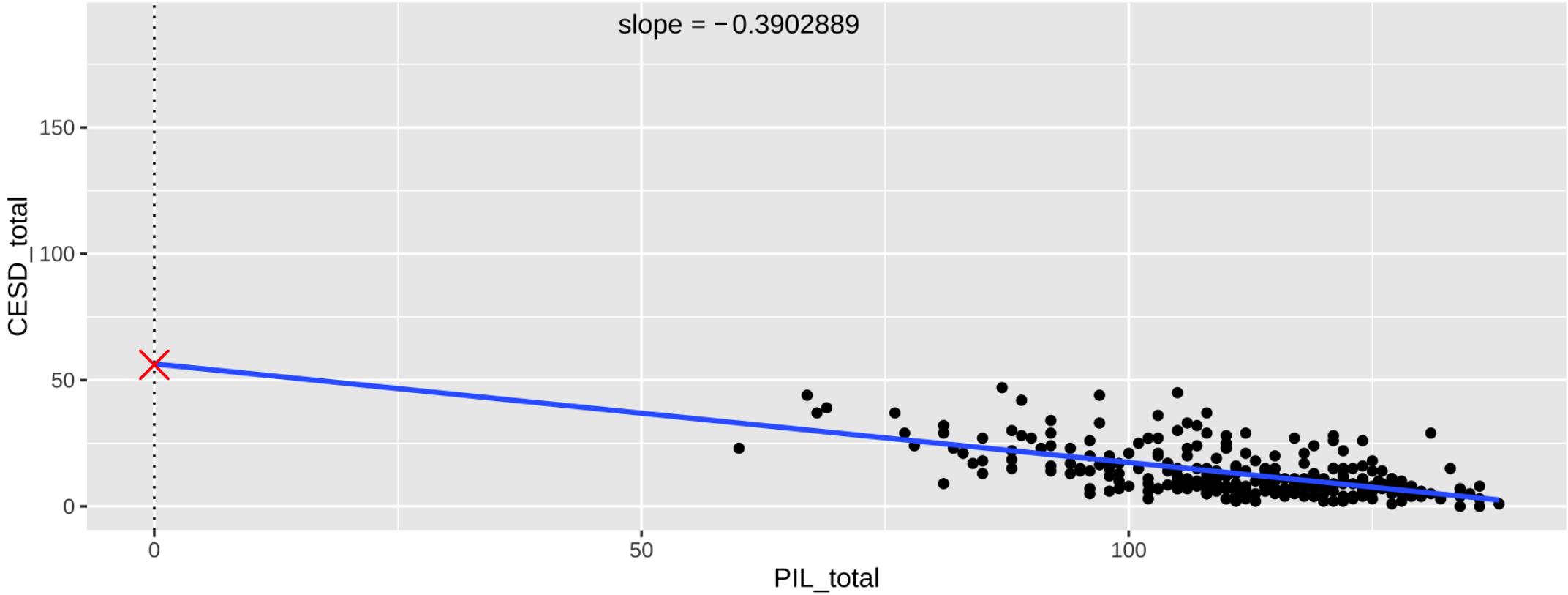


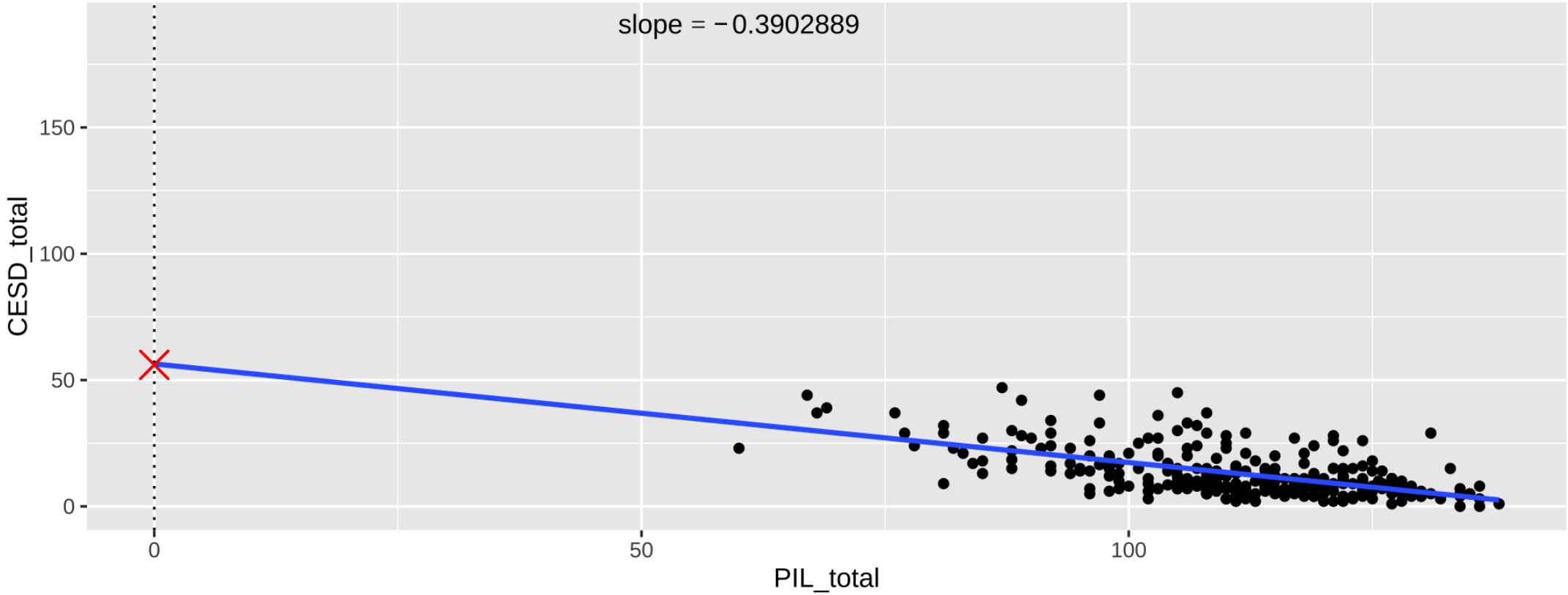

CESD_total
40
30
20
10
0
60
80
100
120
140
PIL_total

`lm()` in R

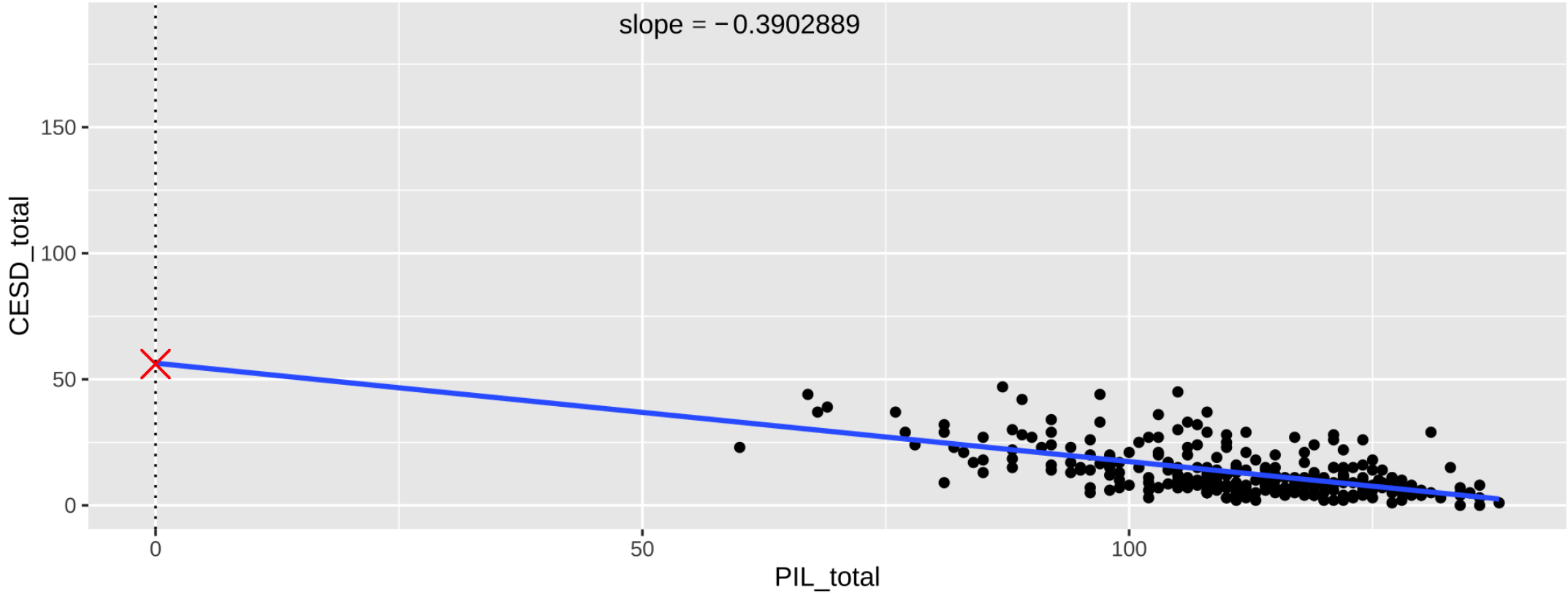
```
lm_reg <-lm(master$CESD_total~master$PIL_total)
```

- How would we interpret this?

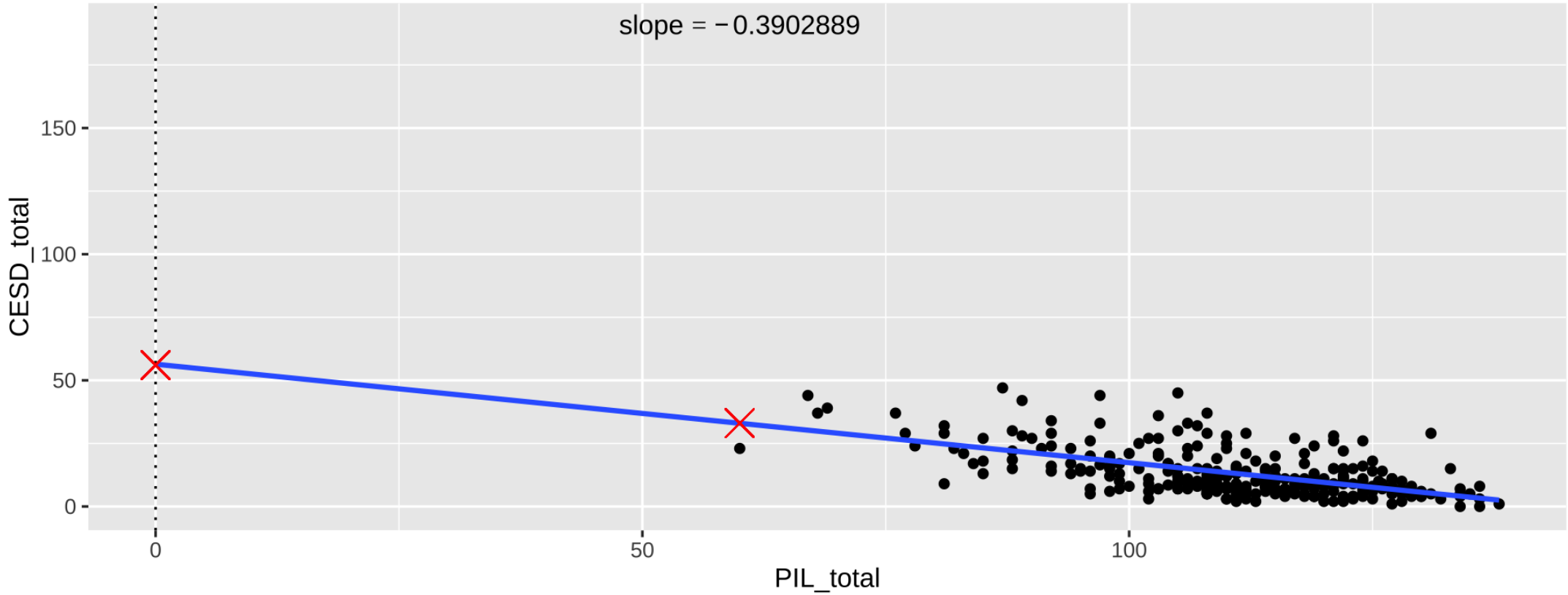

slope = −0.3902889
150
100
50
0
CESD_total
0
50
100
PIL_total


slope = −0.3902889
150
100
50
0
CESD_total
0
50
100
PIL_total

$$CESD_{total} = 56 + (-.39) * PIL_{total}$$



$$\hat{CESD}_{total} = 56 + (-.39) * 60$$



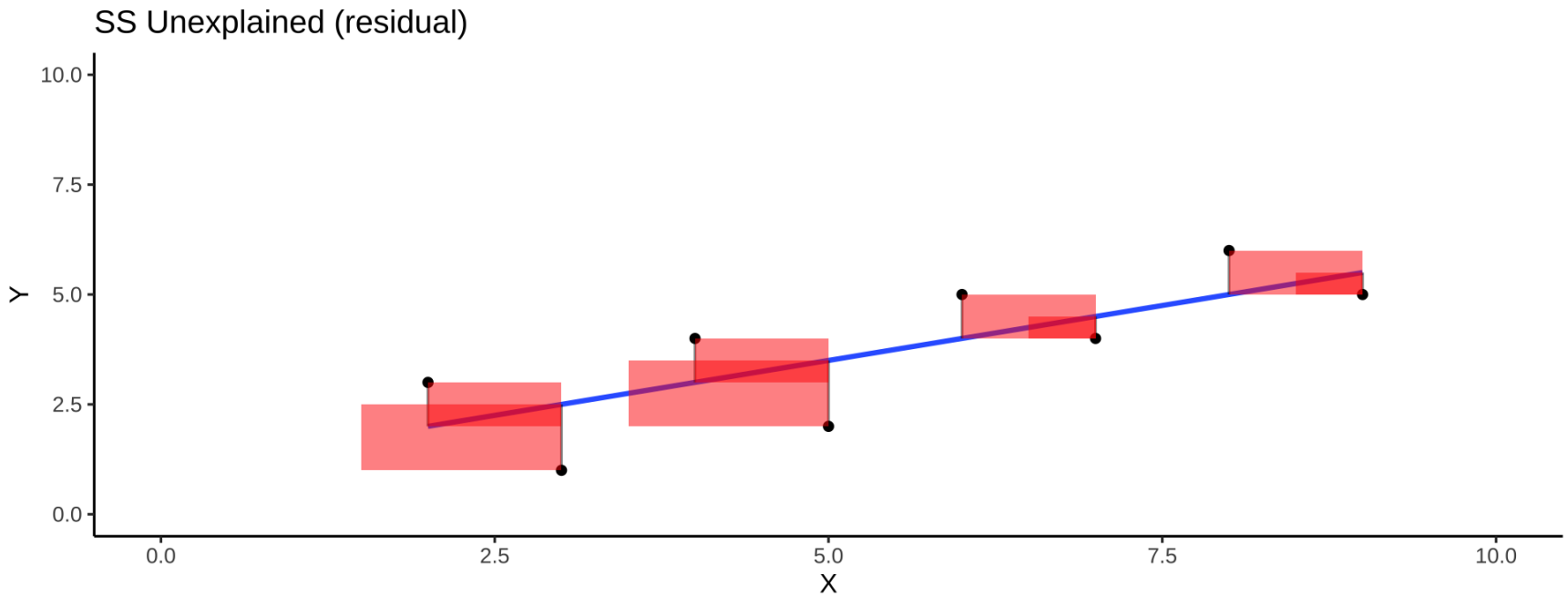
Residuals, Fitted Values, and Model Fit

- If we want to make inferences about the regression parameter estimates, then we also need an estimate of their variability
- We also need to know how well are data fits the linear model

SS Unexplained (Sums of Sqaures Error)

$$residual = y - \hat{y} = y - (x * \hat{\beta}_x + \hat{\beta}_0)$$

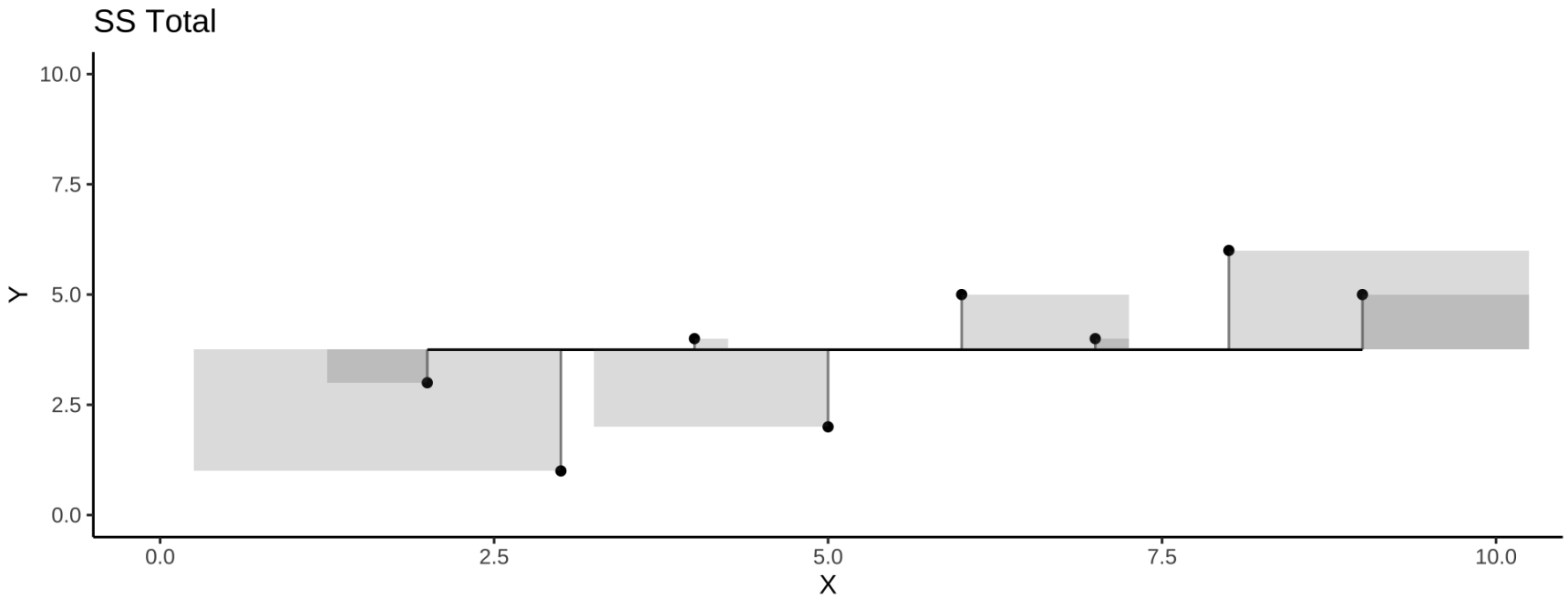
$$SS_{error} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} residuals^2$$



SS Total (Sums of Sqaures Total)

> Squared differences between the observed dependent variable and its mean.

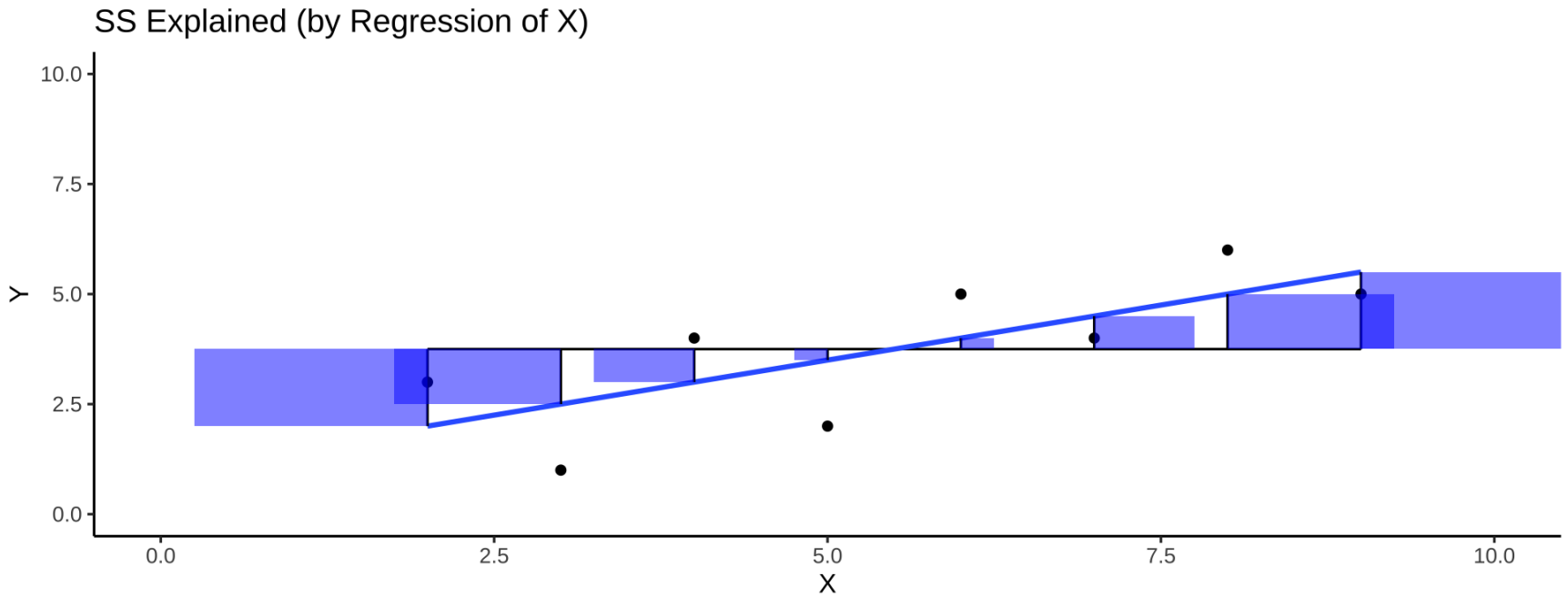
$$SS_{total} = \sum (Y_i - \bar{Y})^2$$



SS Explained (Sums of Squares Regression)

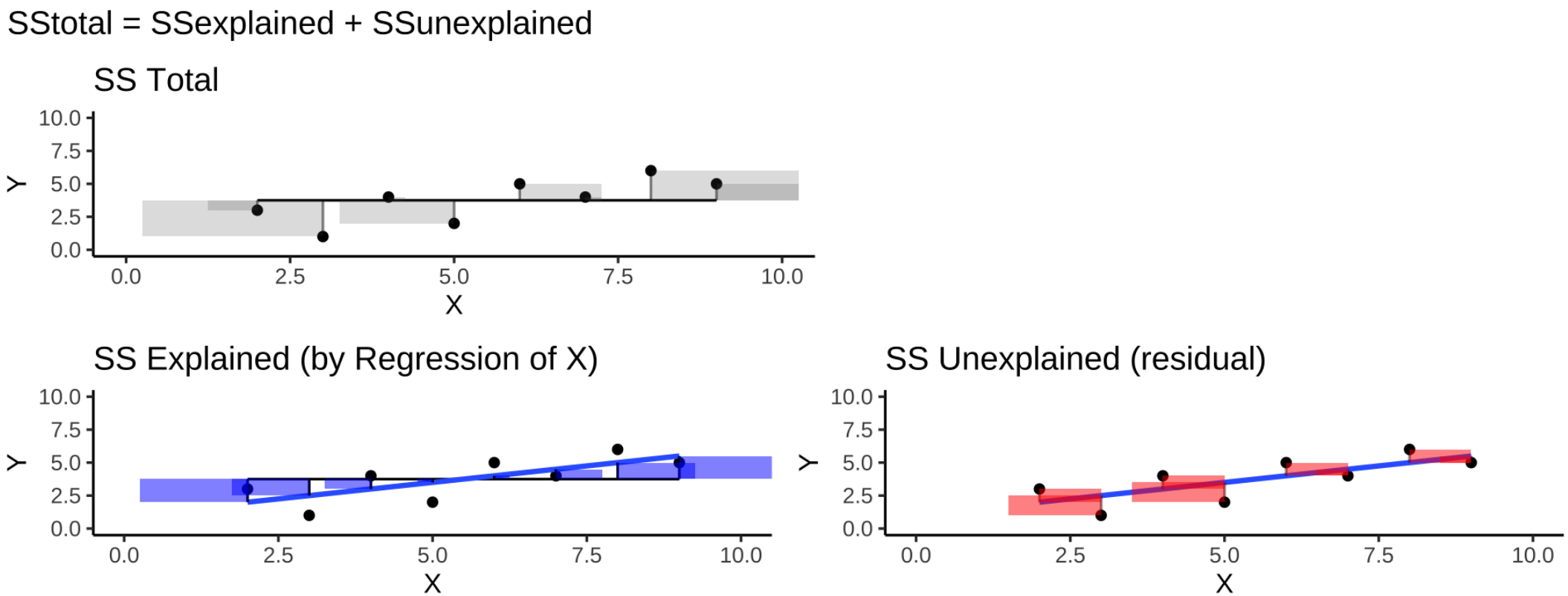
> The sum of the differences between the predicted value and the mean of the dependent variable

$$SS_{Explained} = \sum (Y_i' - \bar{Y})^2$$



All Together

```
library(patchwork)

(total_plot +plot_spacer())/(exp_plot+res_plot)+
  plot_annotation(title = 'SStotal = SSexplained + SSunexplained')
```



broom Regression

- tidy(): coefficent table
- glance(): model summary
- augment(): adds information about each observation

```
library(broom) # install
```

Regression: NHST

$$H_0\text{: } \beta_1 = 0$$

$$H_1\text{: } \beta_1 \neq 0$$

$$t_{N-p} = \frac{\hat{\beta} - \beta_{expected}}{SE_{\hat{\beta}}}$$

$$t_{N-p} = \frac{\hat{\beta} - 0}{SE_{\hat{\beta}}}$$

$$t_{N-p} = \frac{\hat{\beta}}{SE_{\hat{\beta}}}$$

```
model1 <- lm(master$CESD_total~master$PIL_total)
tidy(model1)
```

term	estimate	std.error	statistic	p.value
(Intercept)	56.4	3.75	15	2.43e-37
master$PIL_total	-0.39	0.0336	-11.6	1.95e-25

Calculate Standard Error

$$MS_{error} = \frac{SS_{error}}{df} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{N - p}$$

$$SE_{model} = \sqrt{MS_{error}} SE_{model} = \sqrt{MS_{error}}$$

$$SE_{\hat{\beta}_x} = \frac{SE_{model}}{\sqrt{\sum (x_i - \bar{x})^2}}$$

```
# get MSE
# 1
mse=performance_mse(model1)
# 2
SE<-sqrt(mse)
#3
x_de<- sum((master$CESD_total - mean(master$CESD_total))^2)

x_sqrt <- sqrt(x_de)

SE <- SE/x_sqrt

SE
```

95% CIs

$$b_1 \pm t^*(SE_{b_1})$$

```
tidy(model1, conf.int = TRUE)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	56.4	3.75	15	2.43e-37	49	63.8
master$PIL_total	-0.39	0.0336	-11.6	1.95e-25	-0.457	-0.324

Getting Residuals and Predicted Values

```
assump=augment(model1)# residuals and fitted values
```

Model Fit

```
assump=glance(model1)#model fit indices
```

Effect Size: R^2

- Coefficient of determination

$$R^2 = 1 - \frac{SS_{\text{error}}}{SS_{\text{tot}}}$$

$$R^2 = 1 - \frac{SS_{unexplained}}{SS_{Total}} = \frac{SS_{explained}}{SS_{Total}}$$

- Standardized effect size
  - Amount of variance explained
  - R^2 of .4 means 40% of variance in the outcome variable (Y) can be explained by the predictor (X)
  - Range: 0-1

R^2

```
library(performance)

rq2=r2(model1)

rq2=rq2$R2

SS_explained <- sum((assump$.fitted - mean(assump$`master$CESD_total`))^2)

SS_total <- sum((assump$`master$CESD_total` - mean(assump$`master$CESD_total`))^2)

r2<- SS_explained/SS_total
```

- R^2 of 0.337 means 33% of variance in depressions scores is explained by meaning in life

R^2_{adj}

$$R^2_{adj} = 1 - \frac{SS_{unexplained}}{SS_{Total}} = \frac{SS_{explained}(n-K)}{SS_{Total}(n-1)}$$

where:

- n = Sample size
- K = # of predictors

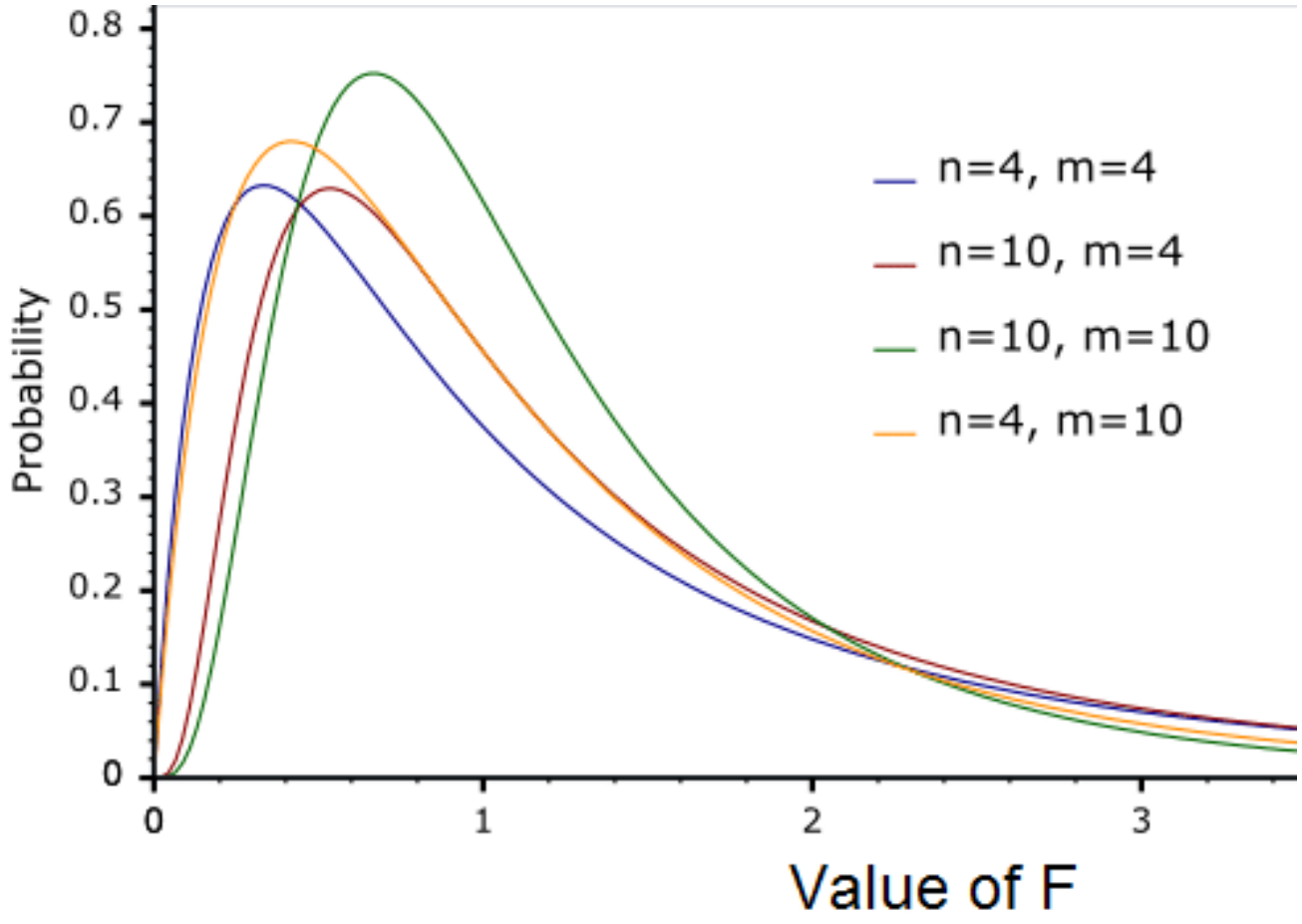
Regression Model

- How useful are each of the individual predictors for my model?
  - Use the coefficients and t-tests of the slopes
- Is my overall model *(i.e., the regression equation)* useful at predicting the outcome variable?
  - Use the model summary, F-test, and R^2

Overall Model Significance

- Our overall model uses an *F*-test
- However, we can think about the hypotheses for the overall test being:
  - H_0: We cannot predict the dependent variable (over and above a model with only an intercept)
  - H_1: We can predict the dependent variable (over and above a model with only an intercept)
- Generally, this form does not include two tailed tests because the math is squared, so it is impossible to get negative values in the statistical test

F-distribution



F-Statistic, Explained Over Unexplained

- F-statistics use measures of variance, which are sums of squares divided by relevant degrees of freedom

$$F = \frac{SS_{Explained}/df1(p-1)}{SS_{Unexplained}/df2(n-p)} = \frac{MS_{Explained}}{MS_{Unexplained}}$$

- If explained = unexplained, then F=1
- If explained > then, F >1
- If explained < unexplained, F < 1

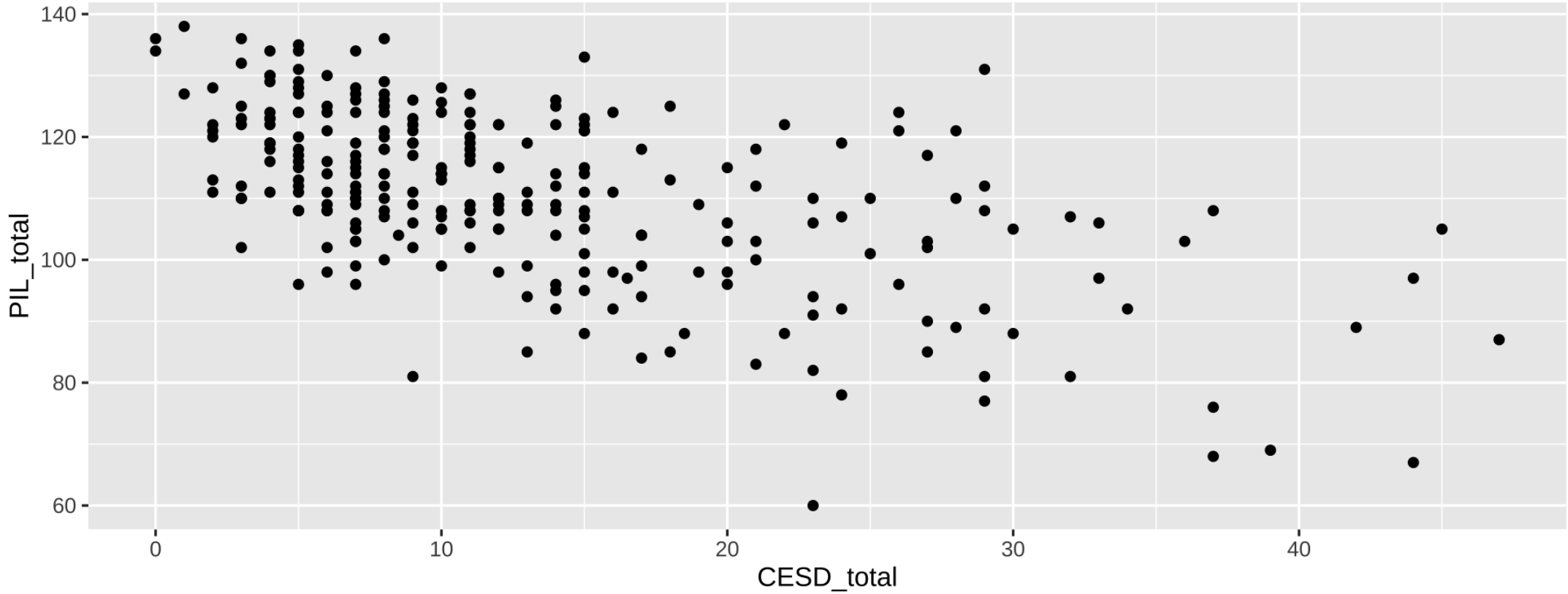
Calculating Mean Squares in R

```
#use augment to get fitted and resid information
SS_explained <- sum((assump$.fitted - mean(assump$`master$CESD_total`))^2)

SS_unexplained <- sum((assump$`master$CESD_total` - assump$.fitted)^2)

# easier solution
ms_error <- performance_mse(model1)

#TSS?
```

Linear Modeling Assumptions

1. Linearity
2. Independence of errors
    - There is not a relationship between the residuals and the variable
3. Normality of errors (on the residuals)
4. Equal variances (Homoscedacity)
    - The variance of the residuals is the same for all values of X

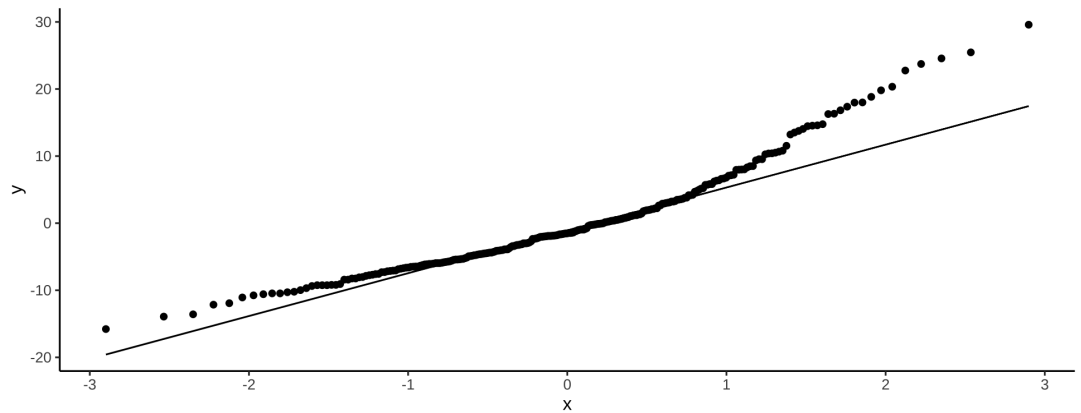
Assumptions: Linearity



Assumptions: Normality of Errors

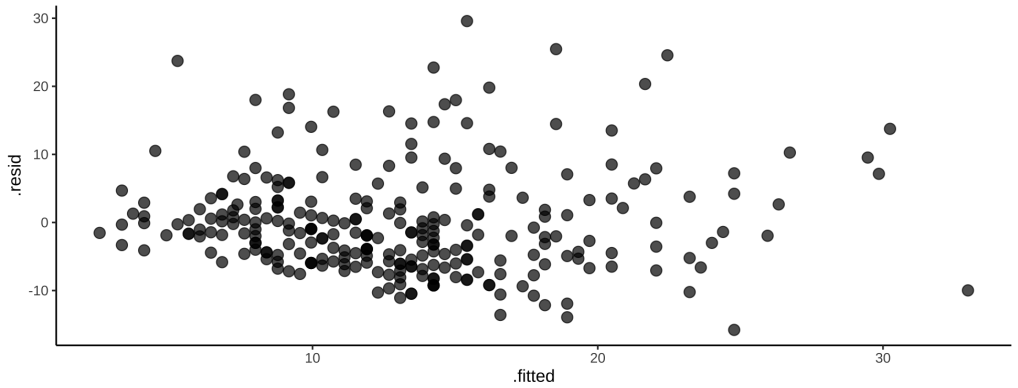
- Shaprio test
- QQ-Plot

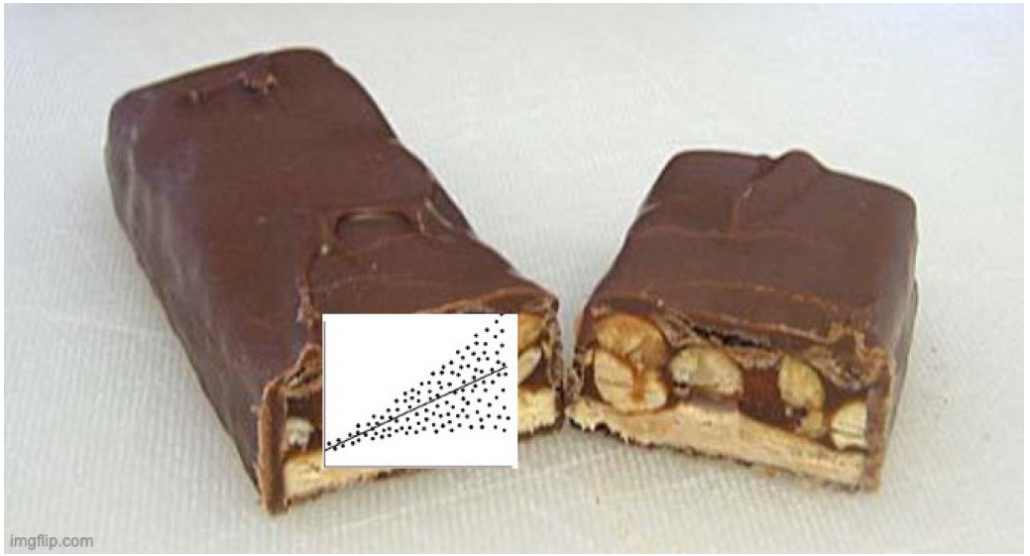
```
assump=augment(model1)# residuals and fitted values

  ggplot(assump, aes(sample = .resid)) +
  stat_qq() +
  stat_qq_line() +
  theme_classic()
```



Assumptions: Equal Variances and Independence

```
ggplot(assump, aes(x = .fitted, y = .resid)) +
geom_point(size = 3, alpha = 0.7) +
theme_classic()
```





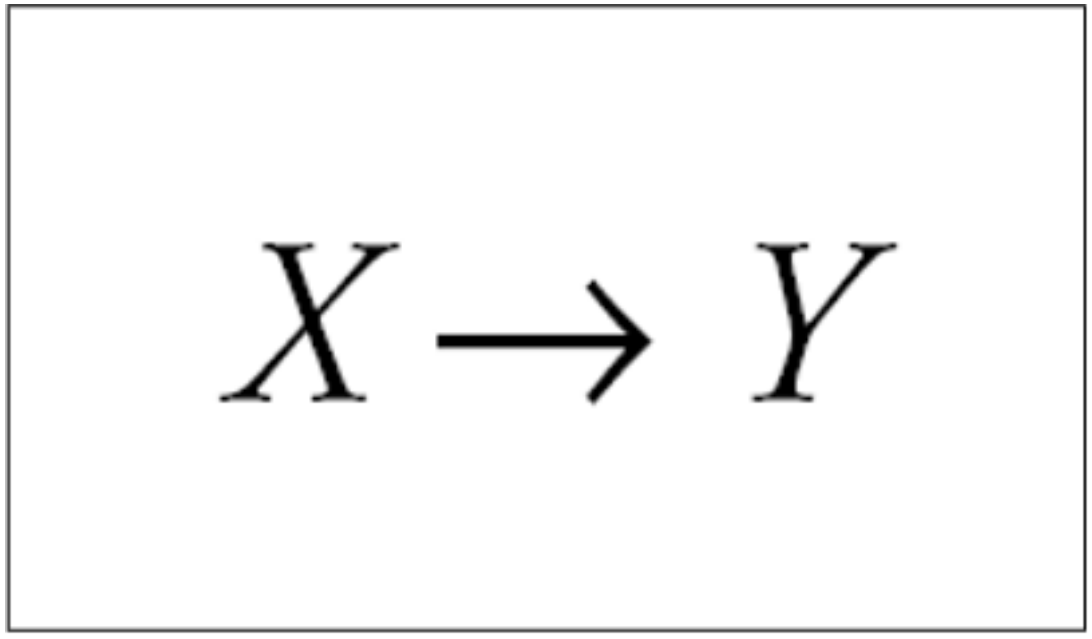
Reporting

- We fitted a linear model (estimated using OLS) to predict master$CESD_total with master (formula: master$CESD_total ~ master$PIL_total). The model explains a statistically significant and substantial proportion of variance (R2 = 0.34, F(1, 265) = 134.58, p < .001, adj. R2 = 0.33). The model's intercept, corresponding to PIL_total = 0, is at 56.40 (95% CI [49.01, 63.78], t(265) = 15.03, p < .001). Within this model:
    - The effect of master$PIL total is statistically significant and negative (beta = -0.39, 95% CI [-0.46, -0.32], t(265) = -11.60, p < .001; Std. beta = -0.39, 95% CI [-0.46, -0.32])

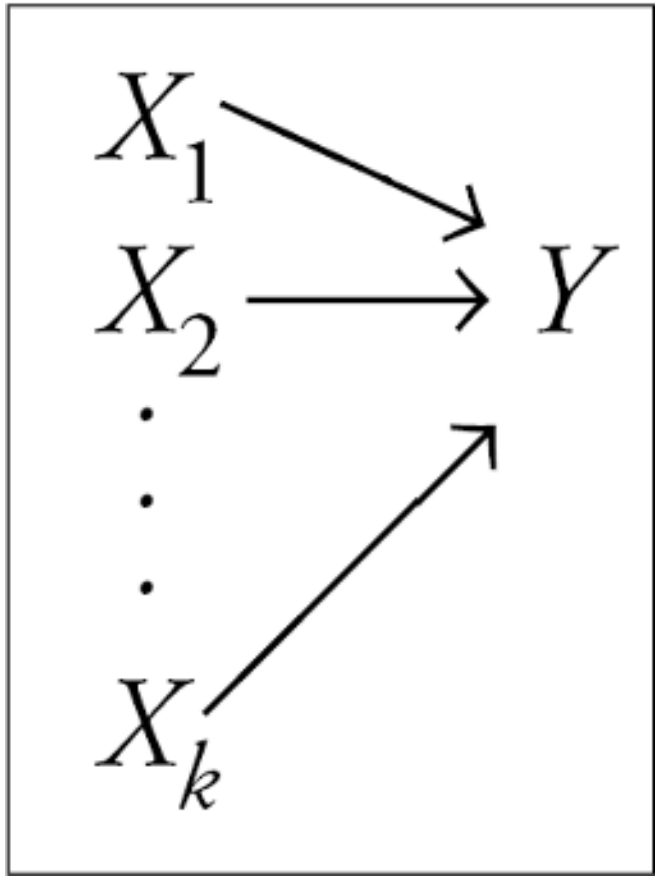
Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence Intervals (CIs) and p-values were computed using a Wald t-distribution approximation., We fitted a linear model (estimated using OLS) to predict master$CESD_total with PIL_total (formula: master$CESD_total ~ master$PIL_total). The model explains a statistically significant and substantial proportion of variance (R2 =

Linear Modeling with Multiple Continuous Predictors

Simple Linear Model



Multiple Predictors in Linear Regression



Multiple Regression Example

$$Y = \beta_0 + \beta_1 X + \varepsilon$$
$$\varepsilon \sim N(0, \sigma^2)$$

$$\hat{Y}_i = b_0 + b_1 X_1 i + b_2 X_2 i + b_3 X_3 i + \ldots + B_k X_k \varepsilon_i$$

- Simple as adding predictors to our linear equation

Multiple regression equation

$$\hat{Y}_i = b_0 + b_1 X_1 i + b_2 X_2 i + b_3 X_3 i + \ldots + B_k X_k \varepsilon_i$$

- $\hat{Y}$ = predicted value on the outcome variable Y
- B_0 = predicted value on Y when all Xs = 0
- X_k = predictor variables
- b_k = unstandardized regression coefficients
- k = the number of predictor variables

Straight Line to Hyperplane

- 2 predictors - 2-D Plane in 3-D
- Intercept α ($b_0) predicts where the regression plane crosses the Y axis
- Slope for variable X1 (b_1) predicts the change in Y per unit X1 holding X2 constant
- The slope for variable X2 (b_2) predicts the change in Y per unit X2 holding X1 constant

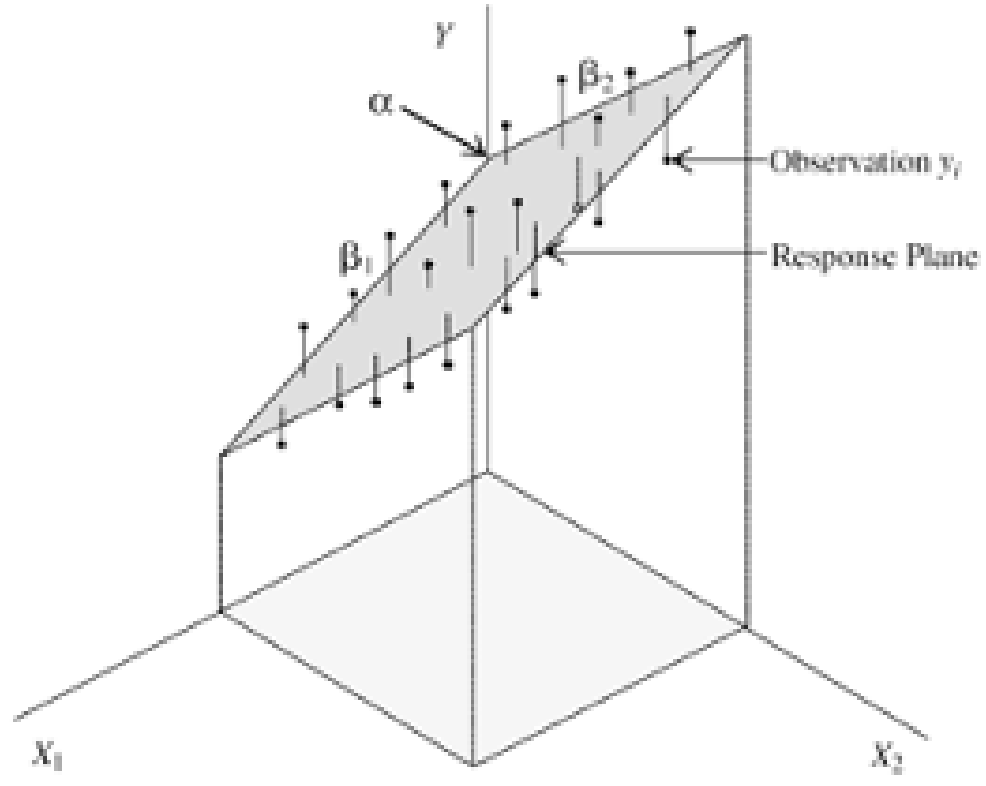


FIGURE 15.1 Three-dimensional response plane.

Multiple Regresson: Example

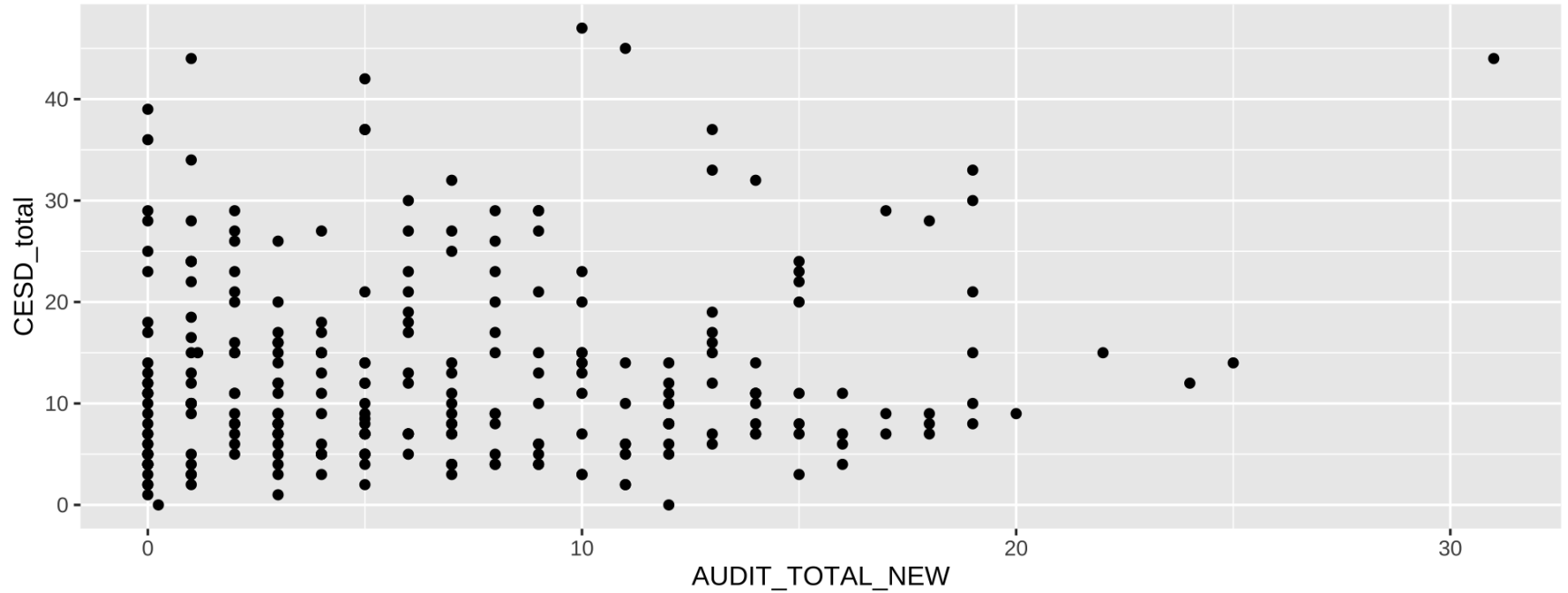
```
model2 <- lm(CESD_total~PIL_total + AUDIT_TOTAL_NEW + DAST_TOTAL_NEW, data=master)
```

$$\text{CESD\_total} = \alpha + \beta_1(\text{PIL\_total}) + \beta_2(\text{AUDIT\_TOTAL\_NEW}) + \beta_3(\text{DAST\_TOTA}$$

- Mental Health and Drug Use:
  - CESD = depression measure
  - PIL total = measure of meaning in life
  - AUDIT total = measure of alcohol use
  - DAST total = measure of drug usage

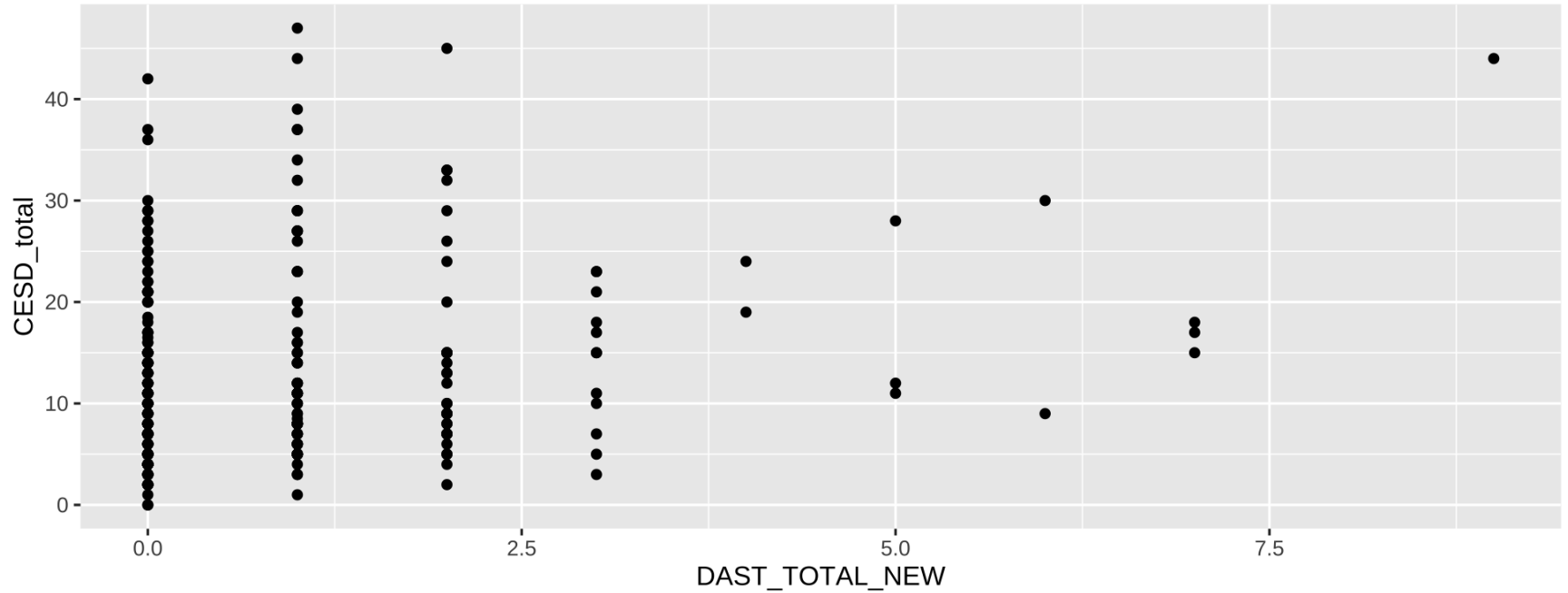
Scatterplot

```
# Change the point sizes manually

anim.1<- ggplot(master, aes(x=AUDIT_TOTAL_NEW, y=CESD_total))+
   geom_point()

anim.1
```



Scatterplot

```
# Change the point sizes manually

anim.1<- ggplot(master, aes(x=DAST_TOTAL_NEW, y=CESD_total))+
   geom_point()

anim.1
```



Linear Model Assumptions (multiple predictors)

- Linearity
- Independence of residuals
- Normality of residuals
- Equal error (“homoskedasticity”)
- Additive (more than one variable)

Problems:

- Missingness
- Factors are not correlated with one another(multicollinearity) (more than one variable)
- No outliers

Problems

- Missingness

```
summary(master)
```

```
##       ...1         PIL_total       CESD_total   AUDIT_TOTAL_NEW
##  Min.   :  1.0   Min.   : 60.0   Min.   : 0.0   Min.   : 0.000
##  1st Qu.: 67.5   1st Qu.:103.0   1st Qu.: 7.0   1st Qu.: 2.000
##  Median :134.0   Median :111.0   Median :11.0   Median : 5.000
##  Mean   :134.0   Mean   :110.7   Mean   :13.2   Mean   : 6.807
##  3rd Qu.:200.5   3rd Qu.:121.0   3rd Qu.:17.0   3rd Qu.:11.000
##  Max.   :267.0   Max.   :138.0   Max.   :47.0   Max.   :31.000
##
##  DAST_TOTAL_NEW
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.906
##  3rd Qu.:1.000
##  Max.   :9.000
##  NA's   :1
```

```
nomiss <- na.omit(master)
nrow(master)
```

Problems

- Multicolinearity
  - You want X and Y to be correlated
  - You do not want the Xs to be highly correlated

Multicolinearity

- Problems
  - Extreme cases (complete collinearity) = Nonidentifiable model
  - “Unstable” regression coefficients ("Bouncing betas")
  - Large standard errors

Multicolinearity

- Tolerance

$$tolerance = 1 - R^2$$

- VIF (variance inflation factor)

$$VIF = \frac{1}{1 - R^2}$$

- Rule of thumb:
  - 10 indicates issues

Multicolinearity

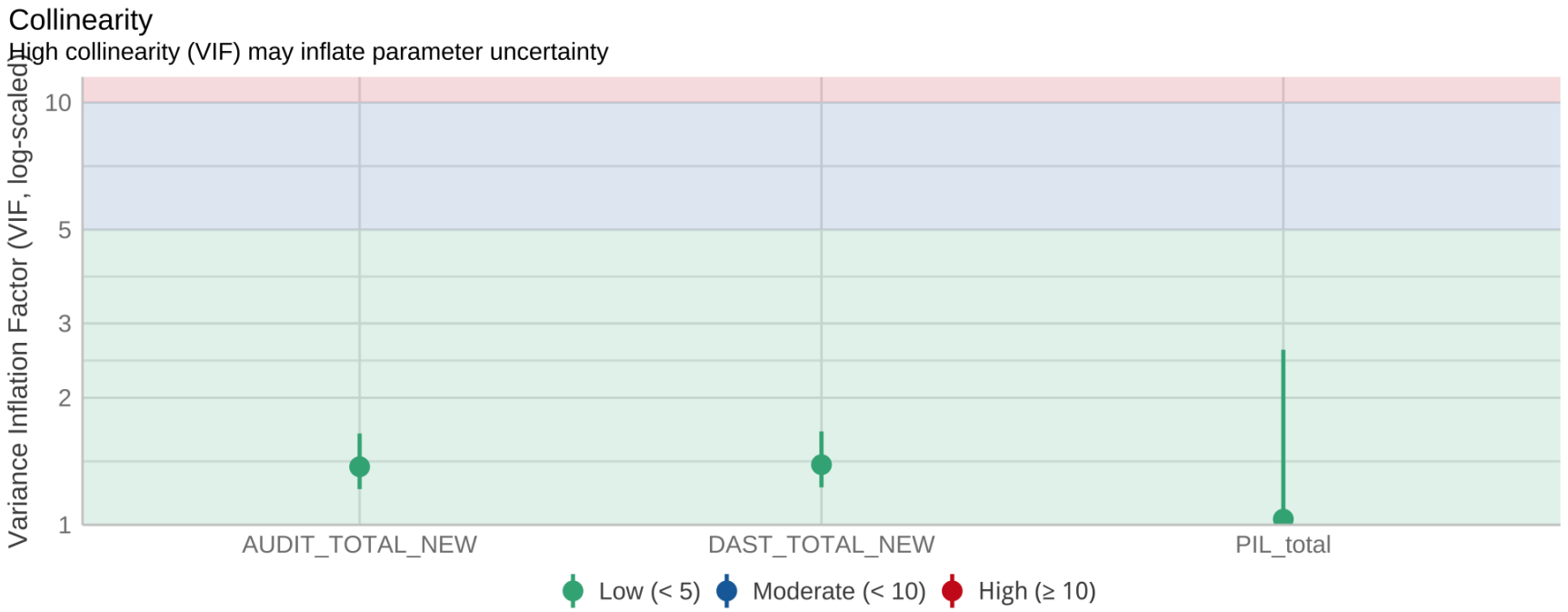
- Strategies
  - Anticipate collinearity issues at the design stage
  - Depends: Drop variable if there's no theoretical pay-off anyway
  - Depends: Fit separate models and compare fit
  - Depends: Increase sample size
  - Depends: Orthogonalize predictors experimentally
  - Depends: Use alternative approaches, such as ridge regression or LASSO

Do not residualize

Problems

- Check multicolinearity in your data

```
#easystats performance package
plot(check_collinearity(model2))
```



Problems: Influential points

- A few outlier checks:
    - Mahalanobis
    - Leverage scores
    - Cook's distance

Mahalanobis Distance

- The `mahalanobis()` function
- Since we are going to use multiple criteria, we are going to save if they are an outlier or not
- The table tells us: 0 (not outliers) and 1 (considered an outlier) for just Mahalanobis values

```
mahal <- mahalanobis(nomiss,
                     colMeans(nomiss),
                     cov(nomiss))
cutmahal <- qchisq(1-.001, ncol(nomiss))
badmahal <- as.numeric(mahal > cutmahal) ##note the direction of the >
table(badmahal)
```

```
## badmahal
##   0   1
## 261   5
```

Other Outliers

- To get the other outlier statistics, we have to use the regression model we wish to test.
- We will use the `lm()` function with our regression formula
- So we will predict depression scores (CESD) with meaning, drugs, and alcohol

```
model3 <- lm(CESD_total ~ PIL_total + AUDIT_TOTAL_NEW + DAST_TOTAL_NEW,
             data = nomiss)
```

Leverage

- Influence of that data point on the slope
- Each score is the change in slope if you exclude that data point
- How do we calculate how much change is bad?
  - $\frac{2K+2}{N}$
  - K is the number of predictors
  - N is the sample size

Leverage

```
k <- 3 ##number of IVs
leverage <- hatvalues(model3)
cutleverage <- (2*k+2) / nrow(nomiss)
badleverage <- as.numeric(leverage > cutleverage)
table(badleverage)
```

```
## badleverage
##   0   1
## 247  19
```

Cook's Distance

- Influence (Cook's Distance)
  - A measure of how much of an effect that single case has on the whole model
  - Often described as leverage + discrepancy
- How do we calculate how much change is bad?
  - $\frac{4}{N-K-1}$

```
cooks <- cooks.distance(model3)
cutcooks <- 4 / (nrow(nomiss) - k - 1)
badcooks <- as.numeric(cooks > cutcooks)
table(badcooks)
```

```
## badcooks
##   0   1
## 251  15
```

Combine Outlier Metrics

- What do I do with all these numbers?
- Create a total score for the number of indicators a data point has
- You can decide what rule to use, but a suggestion is 2 or more indicators is an outliers

```
##add them up!

totalout <- badmahal + badleverage + badcooks

table(totalout)
```

```
## totalout
##   0   1   2   3
## 239  17   8   2
```

```
noout <- filter(nomiss, totalout <= 2)
```

Run Model Again

- Now that we got rid of outliers, we need to run that model again, without the outliers

```
model4 <- lm(CESD_total ~ PIL_total + AUDIT_TOTAL_NEW + DAST_TOTAL_NEW,
             data = noout)
```

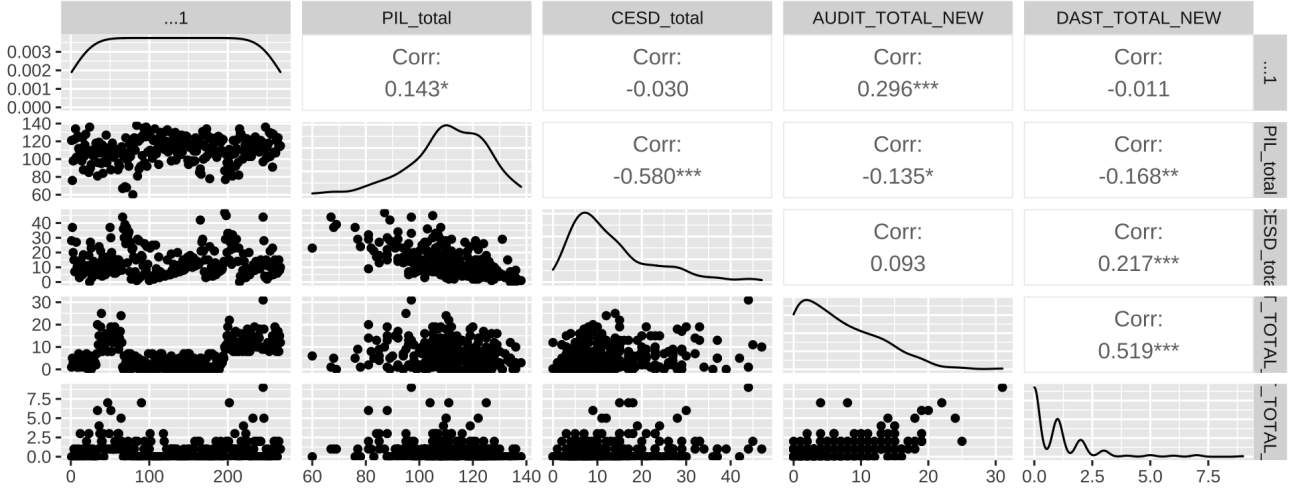
Assumptions

- Additivity
    - Implies that for an existing model, the effect of a predictor on a response variable (whether it be linear or non-linear) is not affected by changes in other existing predictors

Assumptions

- Linearity
  - Check (use a plot) that your variables are linearly related

```
library(GGally)
ggpairs(master)
```

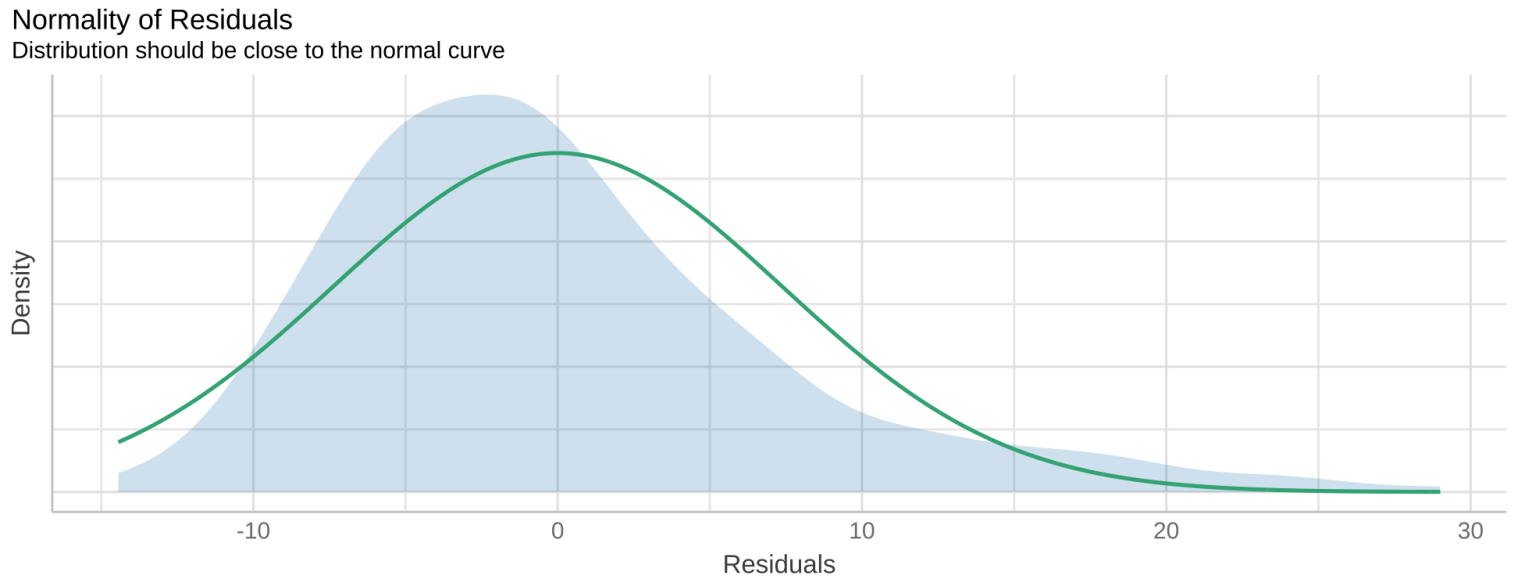


Assumptions

- Normality

Applies to residuals and not the distribution of the data

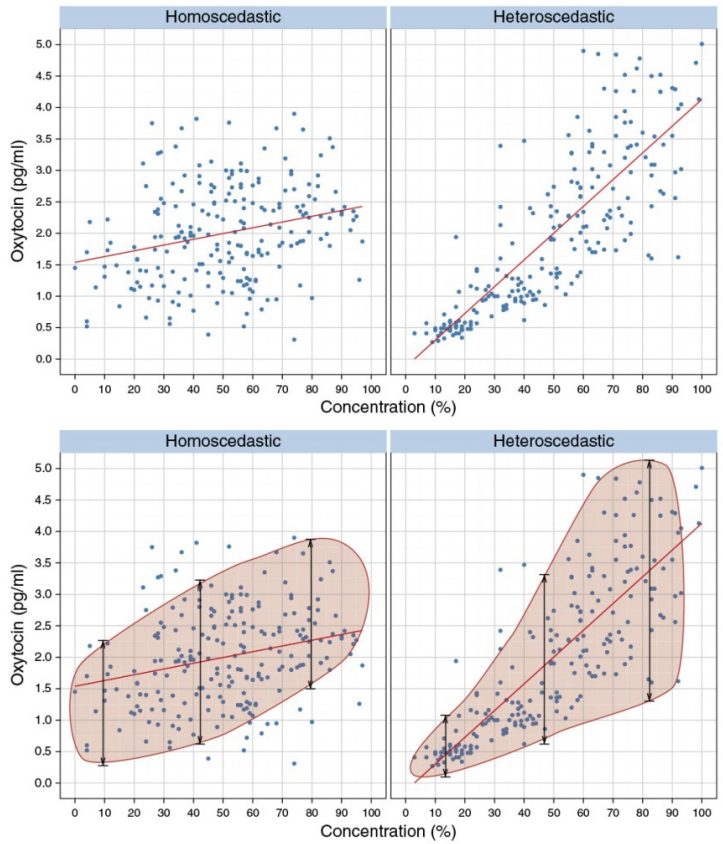
```
model4 <- lm(CESD_total ~ PIL_total + AUDIT_TOTAL_NEW + DAST_TOTAL_NEW,
             data = noout)

plot(check_normality(model4))
```



Assumptions

- Homogeneity & Homoscedasticity
  - Constant error variance

```
knitr::include_graphics("images/homohetero.svg")
```



Assumptions

- Homogeneity & Homoscedasticity
    - Uses Levene's test (we know that one) or Bartlett's test

```
#easystats
performance::check_homogeneity(model4)
```

Check Assumptions

```
check_normality(model4)
```

```
## Warning: Non-normality of residuals detected (p < .001).
```

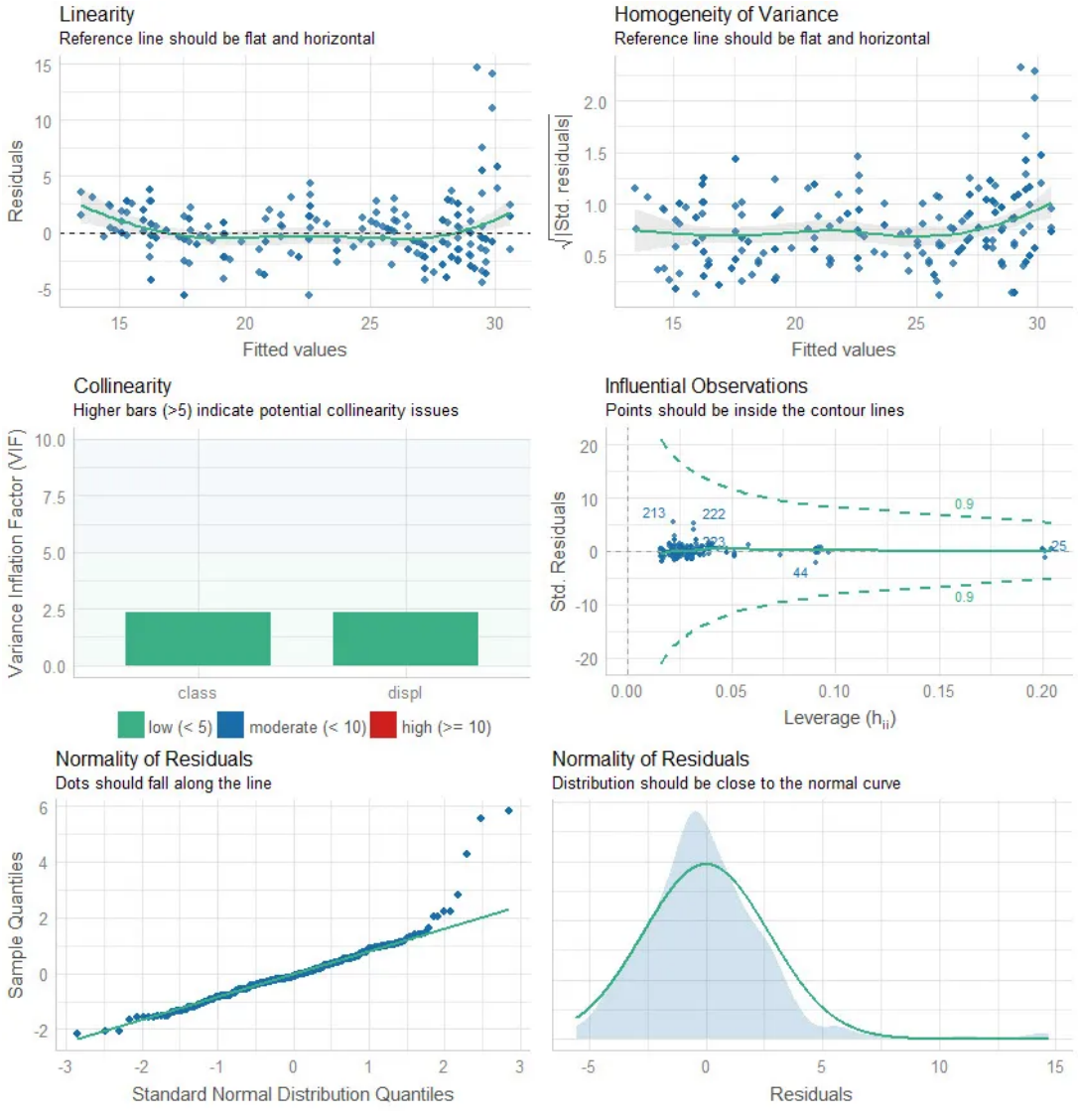
```
check_heteroscedasticity(model4)
```

```
## Warning: Heteroscedasticity (non-constant error variance) detected (p < .001).
```

```
check_collinearity(model4)
```

Term	VIF	VIF_CI_low	VIF_CI_high	SE_factor	Tolerance	Tolerance_C
PIL_total	1.02	1	10.9	1.01	0.981	0
AUDIT_TOTAL_NEW	1.27	1.14	1.52	1.13	0.79	0
DAST_TOTAL_NEW	1.27	1.14	1.53	1.13	0.785	0

Check Assumptions



Assumption Alternatives

- If your assumptions go wrong:
  - Linearity - try nonlinear regression or nonparametric regression
  - Normality - more subjects, still fairly robust
  - Homogeneity/Homoscedasticity - bootstrapping
    - Use robust methods for SE
    - model_parameters(model, vcov = "HC3")
  - Multicolineary - drop one of the predictors; ridge regression; lasso regression

Individual Predictors

- We test the individual predictors with a t-test:
  - $t = \frac{b}{SE}$
  - Therefore, the model for each individual predictor is our coefficient b
  - Single sample t-test to determine if the b value is different from zero

Linear Model with Multiple Predictors: NHST

- All slopes

$$b_1 = b_2 = b_3 = 0$$

$$b_1 \neq b_2 \neq b_3$$

- Or at least one of the slopes is not 0

Overall Model Fit

- Is the overall model significant?

```
summary(model4)
```

```
##
## Call:
## lm(formula = CESD_total ~ PIL_total + AUDIT_TOTAL_NEW + DAST_TOTAL_NEW,
##     data = noout)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.438  -5.139  -1.219   3.327  29.002
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     56.28852    3.81631  14.749   <2e-16 ***
## PIL_total       -0.38986    0.03324 -11.729   <2e-16 ***
## AUDIT_TOTAL_NEW -0.10774    0.09220  -1.169   0.2436
## DAST_TOTAL_NEW   0.91509    0.41320   2.215   0.0277 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.416 on 260 degrees of freedom
## Multiple R-squared:  0.3655,    Adjusted R-squared:  0.3582
## F-statistic: 49.92 on 3 and 260 DF,  p-value: < 2.2e-16
```

```r
library(papaja)
apa_style <- apa_print(model4)
apa_style$full_result$modelfit
```

```
## $r2
## [1] "$R^2 = .37$, 90\\% CI $[0.28, 0.44]$, $F(3, 260) = 49.92$, $p < .001$"
```

- $R^2 = .37$, 90\% CI $[0.28, 0.44]$, $F(3, 260) = 49.92$, $p < .001$

Predictors

- What about the predictors?

```
summary(model4) %>%
tidy()
```

term	estimate	std.error	statistic	p.value
(Intercept)	56.3	3.82	14.7	3.45e-36
PIL_total	-0.39	0.0332	-11.7	8.78e-26
AUDIT_TOTAL_NEW	-0.108	0.0922	-1.17	0.244
DAST_TOTAL_NEW	0.915	0.413	2.21	0.0277

Predictors

```
apa_style$full_result$PIL_total
```

```
## [1] "$b = -0.39$, 95\\% CI $[-0.46, -0.32]$, $t(260) = -11.73$, $p < .001$"
```

```
apa_style$full_result$AUDIT_TOTAL_NEW
```

```
## [1] "$b = -0.11$, 95\\% CI $[-0.29, 0.07]$, $t(260) = -1.17$, $p = .244$"
```

```
apa_style$full_result$DAST_TOTAL_NEW
```

```
## [1] "$b = 0.92$, 95\\% CI $[0.10, 1.73]$, $t(260) = 2.21$, $p = .028$"
```

- Meaning: $b = -0.39$, 95\% CI $[-0.46, -0.32]$, $t(260) = -11.73$, $p < .001$
- Alcohol: $b = -0.11$, 95\% CI $[-0.29, 0.07]$, $t(260) = -1.17$, $p = .244$
- Drugs: $b = 0.92$, 95\% CI $[0.10, 1.73]$, $t(260) = 2.21$, $p = .028$

Predictors

- Two concerns:
  - What if I wanted to use beta because these are very different scales?
  - What about an effect size for each individual predictor?

Individual Predictors: Standardization

- b = unstandardized regression coefficient
  - For every one unit increase in X, there will be b units increase in Y.
- β = standardized regression coefficient
  - b in standard deviation units
  - For every one SD increase in X, there will be β SDs increase in Y
- b or β?:
  - b is more interpretable given your specific problem
  - β is more interpretable given differences in scales for different variables

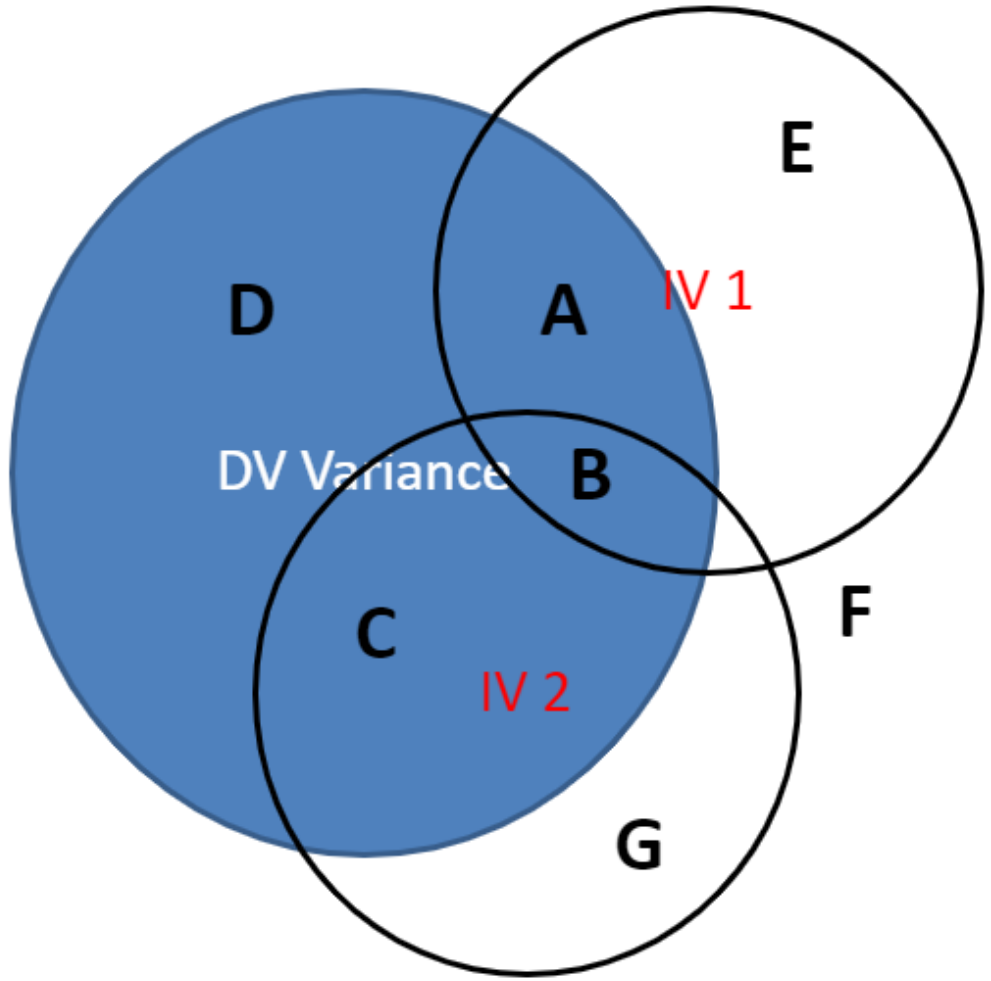
Beta

```
#easystats
library(parameters)
standardize_parameters(model2)
```

Parameter	Std_Coefficient	CI	CI_low	CI_high
(Intercept)	1.73e-16	0.95	-0.0972	0.0972
PIL_total	-0.567	0.95	-0.666	-0.468
AUDIT_TOTAL_NEW	-0.0569	0.95	-0.171	0.0573
DAST_TOTAL_NEW	0.151	0.95	0.0364	0.266

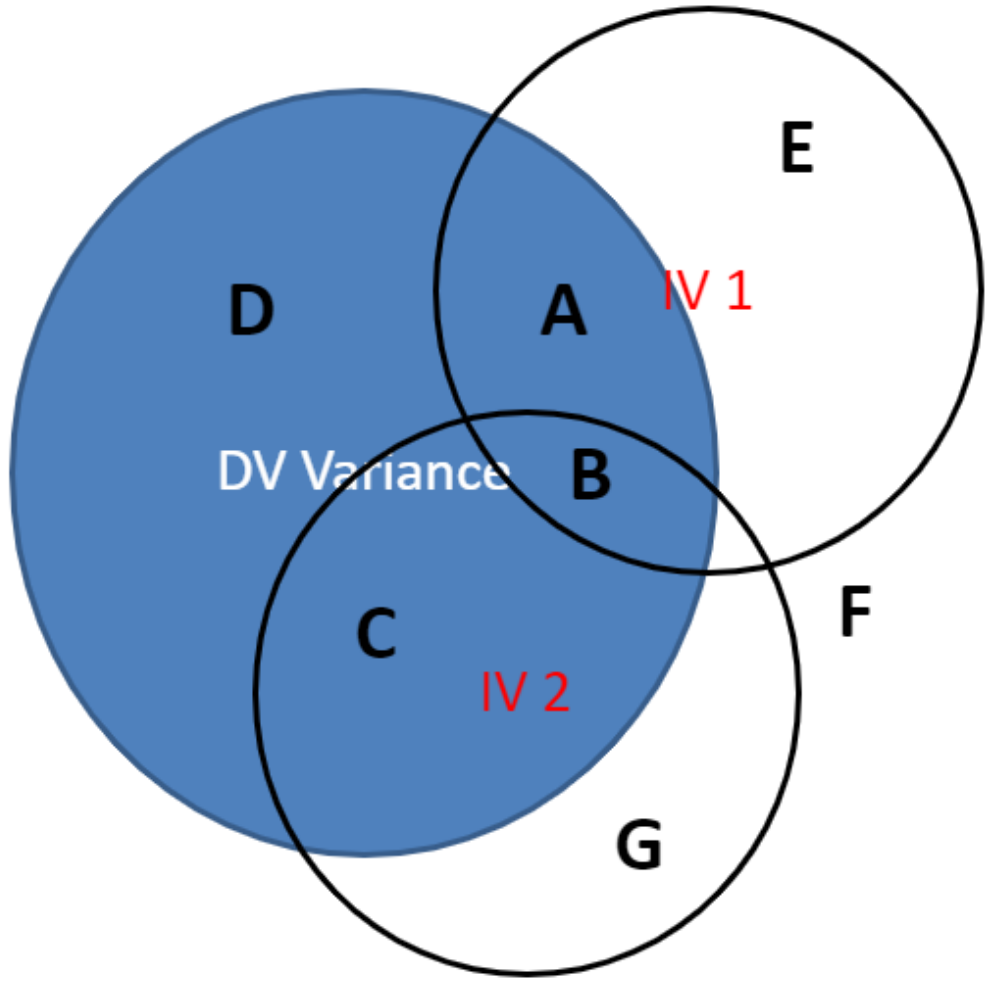
The Impact of Individual Predictors on the Model: Effect Size

- R is the multiple correlation
- R^2 is the multiple correlation squared
- All overlap in Y, used for overall model
- $A + B + C/(A + B + C + D)$



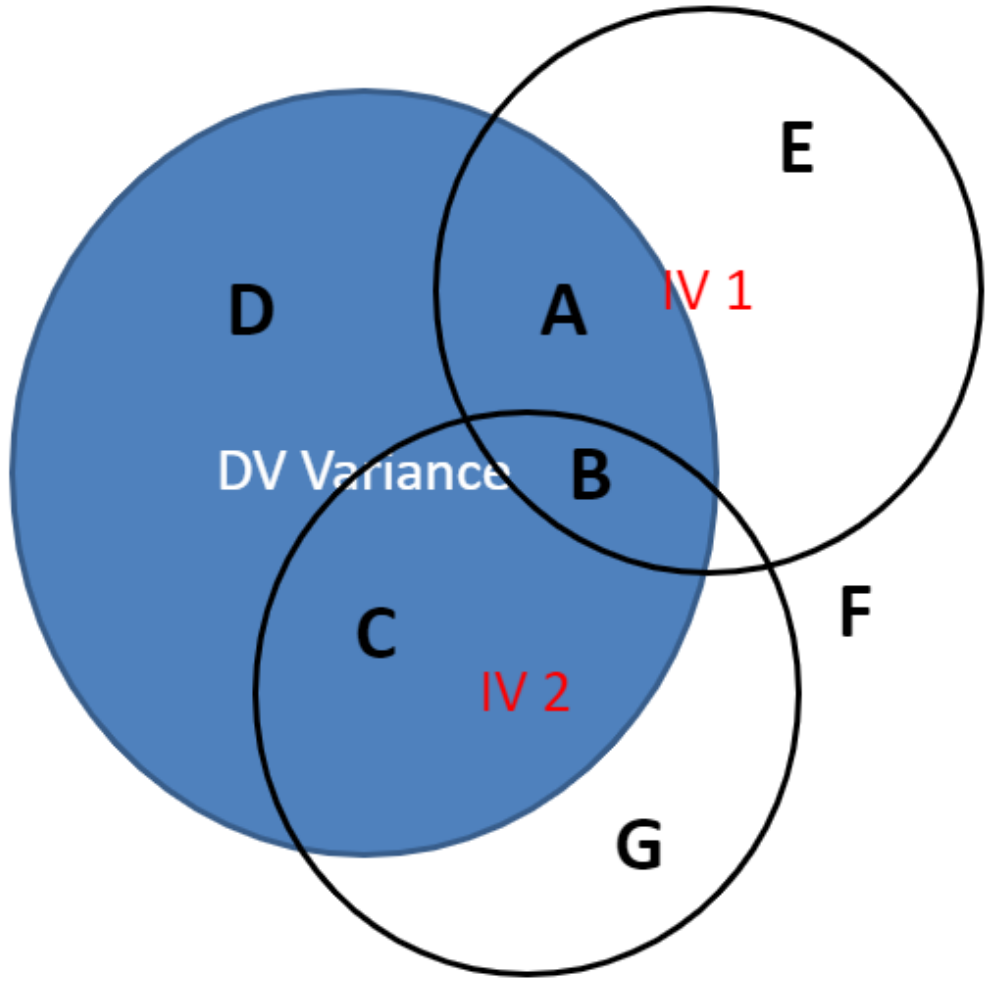
Effect Size

- sr is the semi-partial correlations
- Unique contribution of IV to R^2 for those IVs
- Increase in proportion of explained Y variance when X is added to the equation
- $A/(A+B+C+D)$



Effect Size

- pr is the partial correlation
- Partial correlation asks how much of the Y variance, which is not estimated by the other IVs, is estimated by this variable.
- $A/(A+D)$
- Removes the shared variance of the control variable (Say X2) from both Y and X1
- Pr > sr



Partial Correlations

- We would add these to our other reports:
  - Meaning: $b = -0.39$, 95\% CI $[-0.46, -0.32]$, $t(260) = -11.73$, $p < .001$, $\left(pr^2 = .30\right)$`
  - Alcohol: $b = -0.11$, 95\% CI $[-0.29, 0.07]$, $t(260) = -1.17$, $p = .244$, $\left(pr^2 < .01\right)$`
  - Drugs: $b = 0.92$, 95\% CI $[0.10, 1.73]$, $t(260) = 2.21$, $p = .028$, $\left(pr^2 < .01\right)$`

```
library(ppcor)
partials <- pcor(noout)
partials$estimate^2
```

```
##                       ...1   PIL_total  CESD_total AUDIT_TOTAL_NEW
## ...1            1.000000000 0.028851596 0.004423368     0.138166386
## PIL_total       0.028851596 1.000000000 0.347759029     0.019113396
## CESD_total      0.004423368 0.347759029 1.000000000     0.008404461
## AUDIT_TOTAL_NEW 0.138166386 0.019113396 0.008404461     1.000000000
## DAST_TOTAL_NEW  0.036728054 0.001382052 0.021314715     0.234879429
```

Multiple Regression: Power

- We can use the `pwr` library to calculate the required sample size for any particular effect size
- First, we need to convert the R^2 value to f^2, which is a different effect size, not the ANOVA *F*

```
library(pwr)
R2 <- model2$r.squared
f2 <- R2 / (1-R2)
R2
```

```
## NULL
```

```
f2
```

```
## numeric(0)
```

Multiple Regression: Power

- `u` is degrees of freedom for the model, first value in the F-statistic
- `v` is degrees of freedom for error, but we are trying to figure out sample size for each condition, so we leave this one blank.
- `f2` is the converted effect size.
- `sig.level` is our α value
- `power` is our power level
- The final sample size is *v + k + 1 where k is the predictors*

```
#f2 is cohen f squared
pwr.f2.test(u = model2$df[1],
            v = NULL, f2 = f2,
            sig.level = .05, power = .80)
```