Inference for Linear Regression

Grinnell College

April 29, 2026

ANOVA and Regression

We stated last week that the null hypothesis for ANOVA was of the form

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

where we are comparing the mean value of a continuous variable across $j = 1, \ldots, k$ different groups.

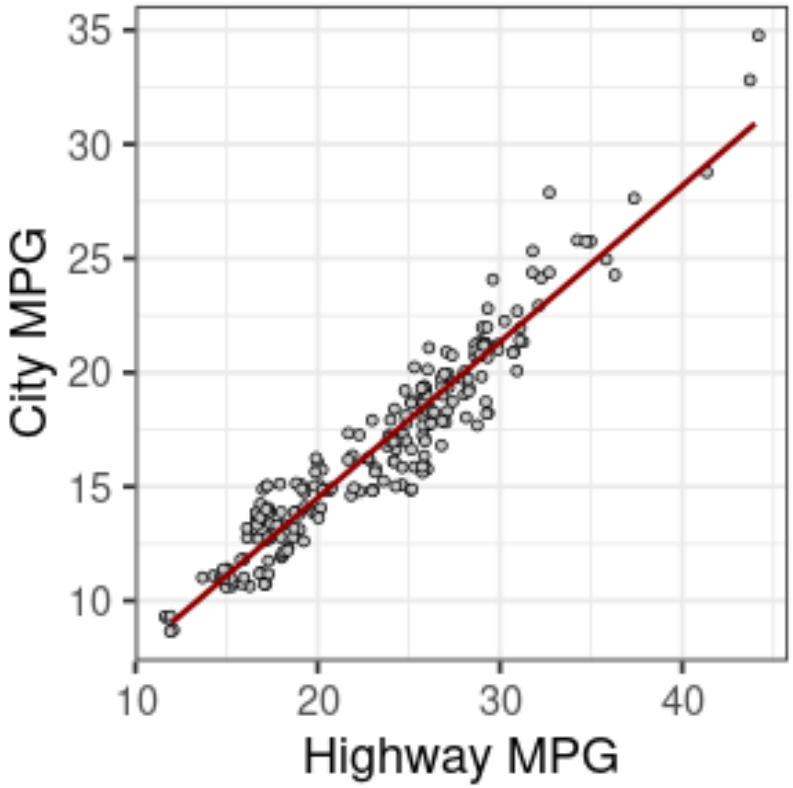
If the null hypothesis were true, then each of the groups would share the same *overall* mean μ

We will now consider reframing this question in terms of linear regression

What is Regression?

$$\widehat{\text{City mpg}} = 0.84 + 0.68 \times \text{Highway mpg}$$

- Can I represent the relationship between X and y as a line?
- Association
- Is this better than just guessing the mean city mpg?
- The slope tells us how changes in X create change in y



Characters Involved

The general assumption about our line is that it is of the form

$$y = \beta_0 + \beta_1 X + \epsilon$$

Where

- y represents our outcome of interest; is it better to use $\overline{y}$ to predict it, or should we use X?
- X represents a *covariate* or predictor variable. This is what we are using to estimate y
- β (beta) is our *coefficient*, telling us how much change in X leads to change in y
  - $\beta_1 = 0$ corresponds to using $\overline{y}$ to estimate y. Why?
- ϵ (epsilon) represents our *error* term. We assume any value of y is equal to the line $\beta_0 + \beta_1 X$ plus or minus some error (more on this term later)

Recall also

Recall also that in linear regression, we can use both quantitative and categorical variables to predict y

$$y = \beta_0 + \mathbb{1}_A \beta_1 + X\beta_2$$

Here we remember that

- β_0 represents the *intercept*; the value of y when all of the covariates (are equal to zero
- $\mathbb{1}_A$ represents an *indicator variable*. β_1 represents a change in intercept when an observation is in group A
- X represents a quantitative variable, with β_2 representing the slope: how much does y change when X changes by 1?

Almost Finally

Just as before, we need to consider the relationship between *parameters* and *statistics*

If there is a linear relationship between X and y, the coefficients, β represent the parameters we are trying to estimate

In that light, we will use $\hat{\beta}$ to represent our *statistic*, or estimate of the parameter based on our data. Similarly, $\hat{y}$ will be our predicted value of y, based on the data

Finally

In light of this, we can finally consider the nature of hypothesis testing in regression. For variables y and X related by the line

$$y = \beta_0 + \beta_1 X + \epsilon$$

we have that *no association between X and y* corresponds to the assumption

$$H_0 : \beta_1 = 0$$

In other words, if changes in X have zero impact on y, then our best guess for y remains $\hat{y}$

ANOVA and Regression

Relating to the case of ANOVA, we might ask if it is best to predict an outcome using an overall mean or if we are better off predicting with a group mean:

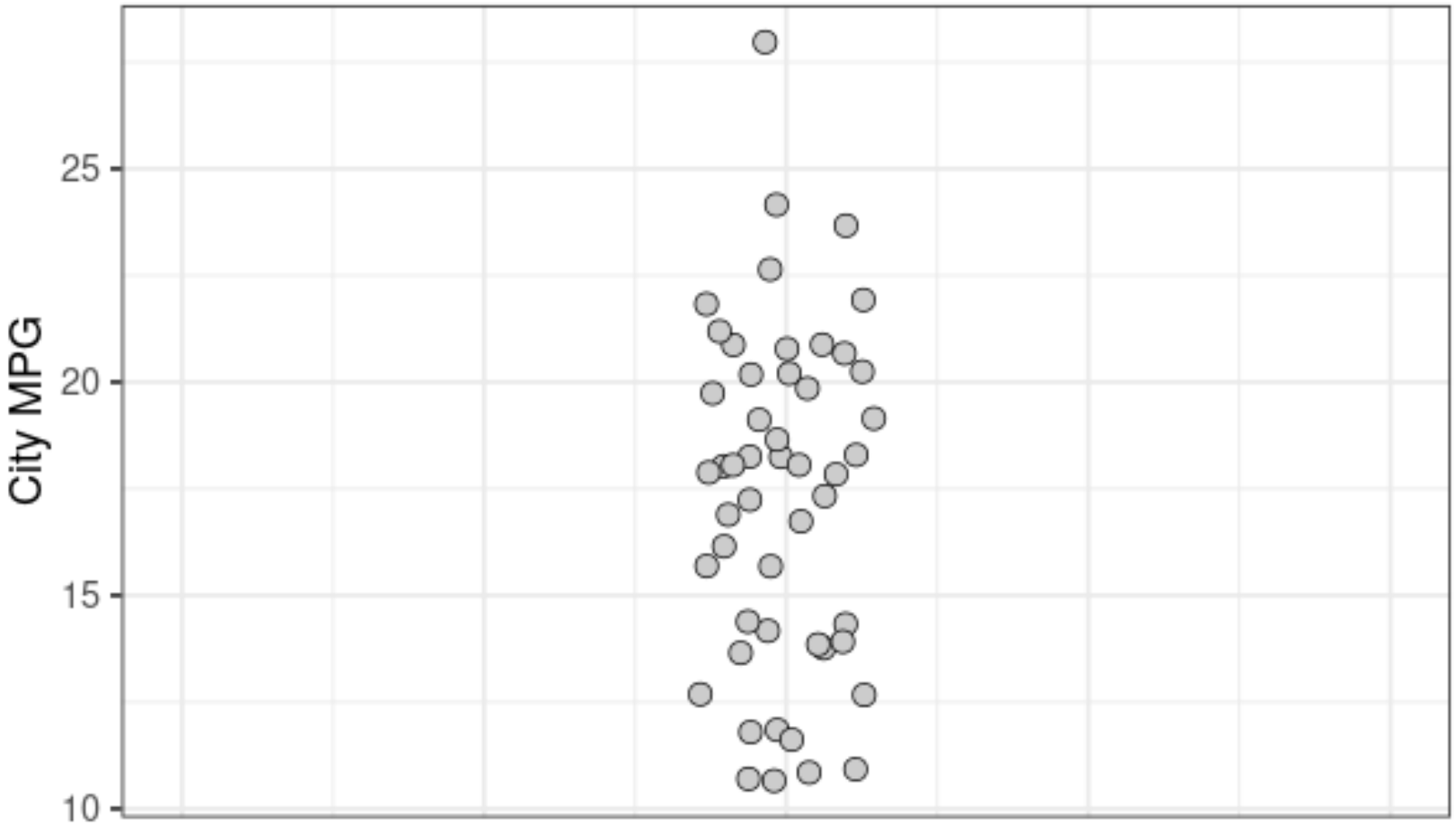
$$H_0 : y_j = \mu, \qquad H_A : y_j = \mu_j$$

In this sense, we can thing of ANOVA as a *statistical model*, or a set of assumptions relating our sample data to an outcome.

Let's begin by seeing how ANOVA is but a special case of linear regression

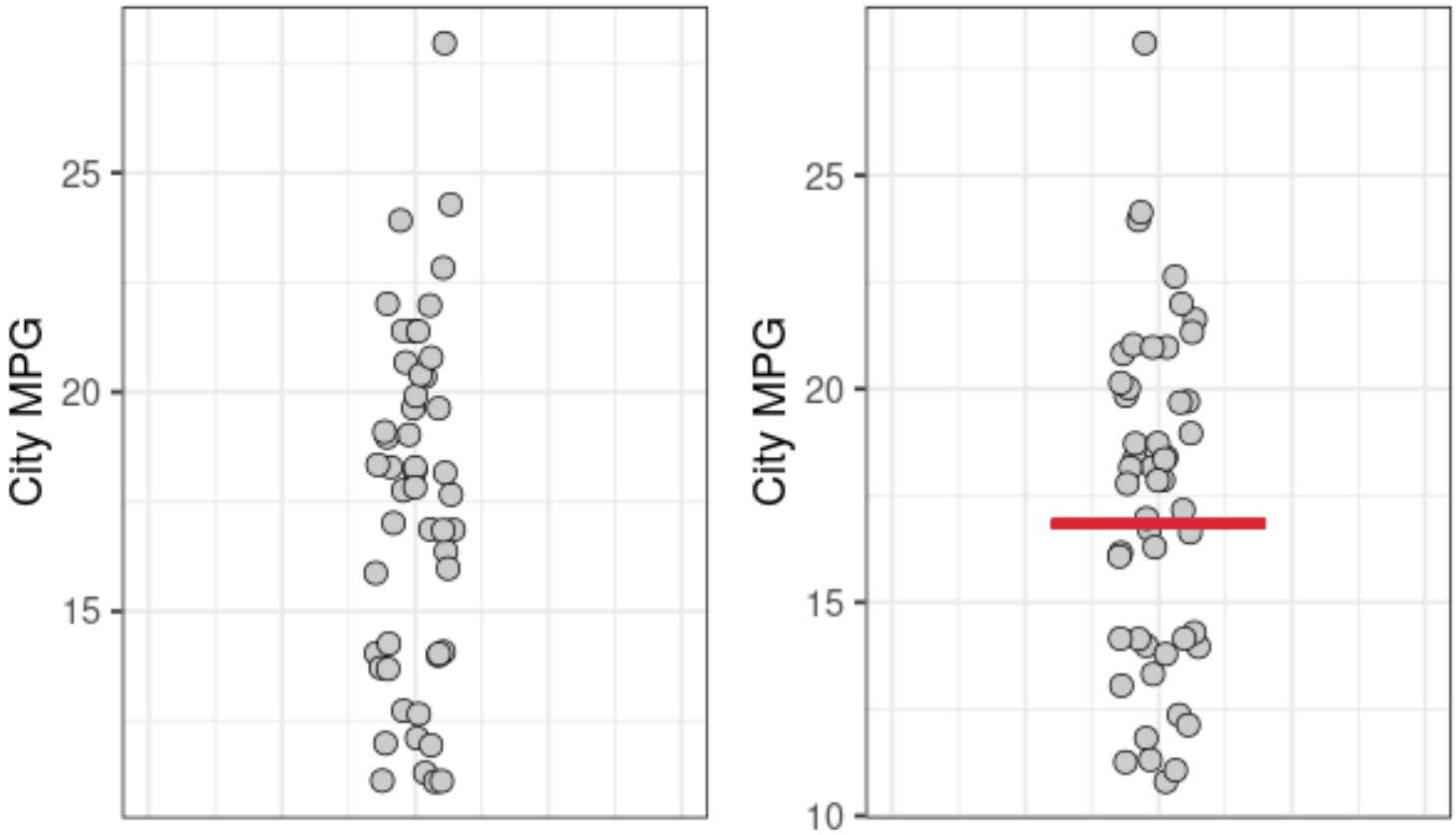
mpg Example

Consider again our mpg dataset, where we might be interested in estimating the city miles per gallon of various vehicles



mpg Example

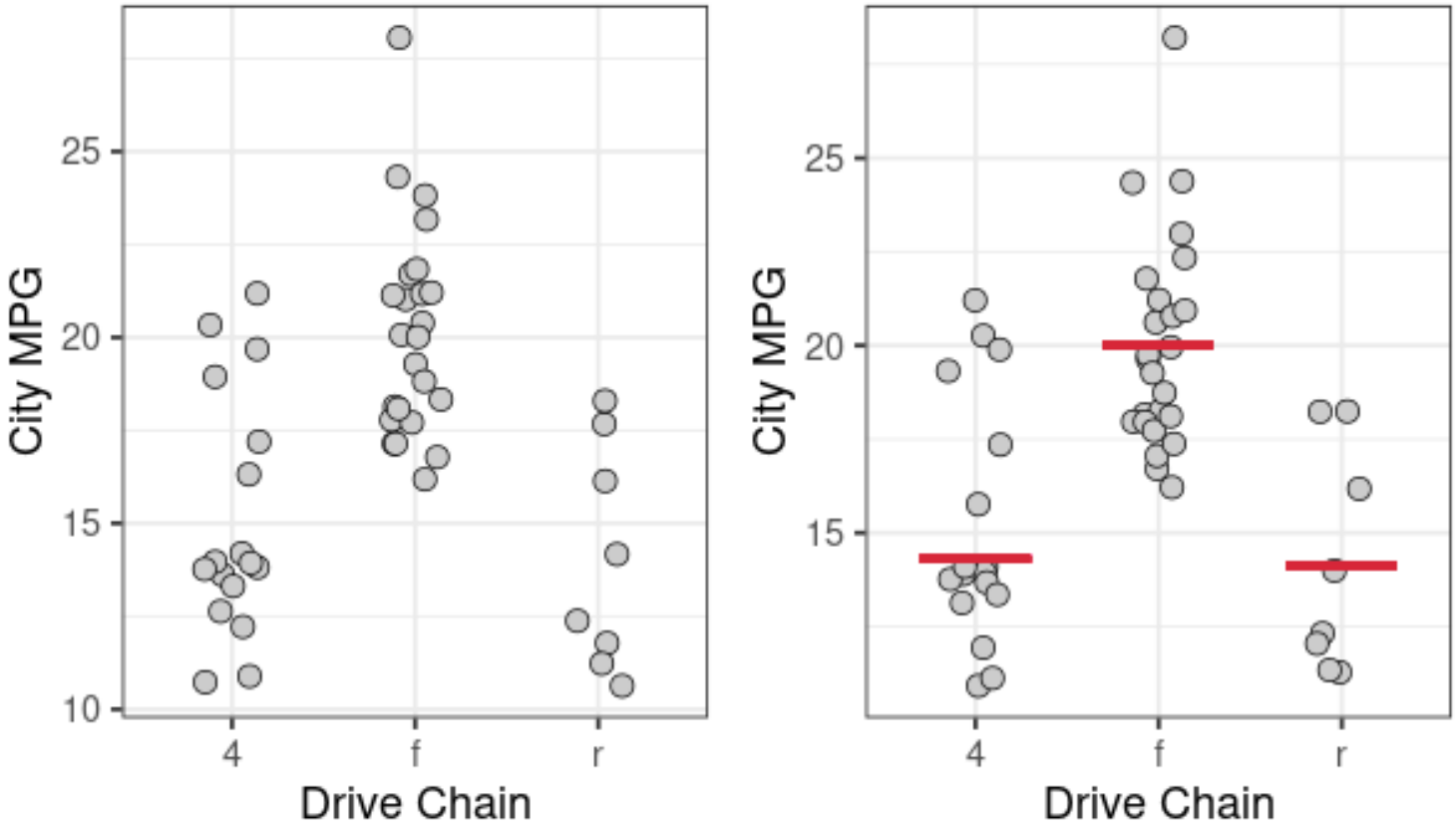
Using simply the overall mean, we would have total squared error of 4220



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Residuals	233	4220.35	18.11		

mpg Example

Consider the alternative, where we predict city mileage based on drive train



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
drv	2	1878.81	939.41	92.68	<0.0001
Residuals	231	2341.53	10.14		

mpg Example

In terms of a regression model, we could frame this as

$$\hat{y} = \mathbb{1}_{\text{4wd}}\hat{\beta}_1 + \mathbb{1}_{\text{Fwd}}\hat{\beta}_2 + \mathbb{1}_{\text{Rwd}}\hat{\beta}_3$$

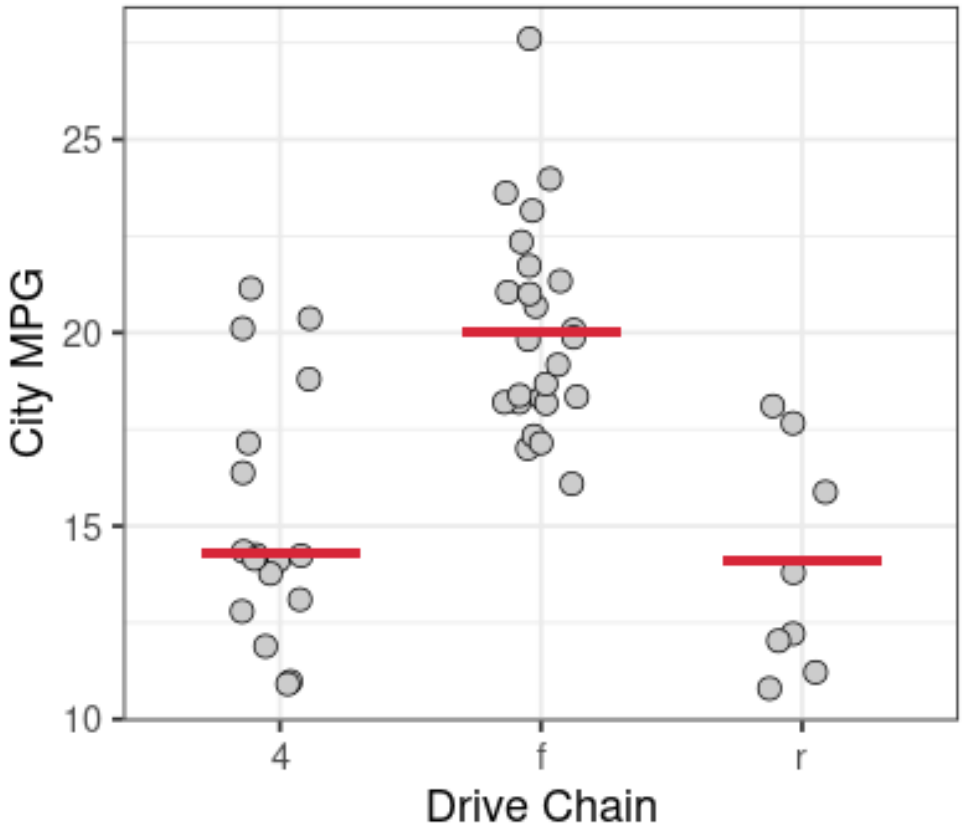
where $\mathbb{1}$ represents our *indicator variable* and, in the case of categorical variable regression, $\hat{\beta}$ represents the mean value for each group. This is precisely what we saw when we did this back in week 3

```
> lm(cty ~ -1 + drv, mpg)

Coefficients:
 drv4   drvf   drvr
14.33  19.97  14.08
```

By default, R will choose one category as the "reference" variable

$$\hat{y} = 14.33 + 5.64 \times \mathbb{1}_{\text{Fwd}} - 0.25 \times \mathbb{1}_{\text{Rwd}}$$

```
> lm(cty ~ drv, mpg)

(Intercept)         drvf         drvr
    14.3301       5.6416      -0.2501
```



By default, R will choose one category as the "reference" variable

$$\hat{y} = 14.33 + 5.64 \times \mathbb{1}_{\text{Fwd}} - 0.25 \times \mathbb{1}_{\text{Rwd}}$$

The motivation for this becomes clear when we consider the null hypothesis with regards to β:

$$H_0 : \beta = 0$$

What, specifically, does β correspond to in this instance?

What would we need of $\hat{\beta}$ in order to test this hypothesis?

β as a test statistic

In assessing the hypothesis

$$H_0 : \beta = 0$$

we estimate the value of $\hat{\beta}$ using our sample to construct a line of best fit. Testing this against the null, we have a natural *test statistic*,

$$t = \frac{\hat{\beta} - 0}{SE_\beta} = \frac{\hat{\beta}}{SE_\beta}$$

where $t \sim t(n-k)$, n being the number of observations and k being the number of predictors used.

mpg Example

$$\hat{y} = 14.33 + 5.64 \times \mathbb{1}_{\text{Fwd}} - 0.25 \times \mathbb{1}_{\text{Rwd}} \hat{\beta}_2$$

```
> lm(cty ~ drv, mpg) %>% summary()

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.3301     0.3137  45.680   <2e-16 ***
drvf          5.6416     0.4405  12.807   <2e-16 ***
drvr         -0.2501     0.7098  -0.352    0.725


Residual standard error: 3.184 on 231 degrees of freedom
Multiple R-squared:  0.4452,  Adjusted R-squared:  0.4404
F-statistic: 92.68 on 2 and 231 DF,  p-value: < 2.2e-16
```

mpg Example

Comparing residuals and F statistic for ANOVA and regression

```
> aov(cty ~ drv, mpg) %>% summary()
             Df Sum Sq Mean Sq F value Pr(>F)
drv           2   1879   939.4   92.68 <2e-16 ***
Residuals   231   2342    10.1
```

```
> lm(cty ~ drv, mpg) %>% summary()

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.3301     0.3137  45.680   <2e-16 ***
drvf          5.6416     0.4405  12.807   <2e-16 ***
drvr         -0.2501     0.7098  -0.352    0.725

Residual standard error: 3.184 on 231 degrees of freedom
Multiple R-squared:  0.4452,  Adjusted R-squared:  0.4404
F-statistic: 92.68 on 2 and 231 DF,  p-value: < 2.2e-16
```

mpg Example

Comparing pairwise differences for TukeyHSD and regression (reference/intercept var is 4WD)

```
> aov(cty ~ drv, mpg) %>% TukeyHSD()
  Tukey multiple comparisons of means
    95% family-wise confidence level

           diff       lwr       upr     p adj
f-4  5.6416010  4.602497  6.680705 0.0000001
r-4 -0.2500971 -1.924554  1.424359 0.9338857
r-f -5.8916981 -7.561520 -4.221876 0.0000001
```

```
> lm(cty ~ drv, mpg) %>% summary()

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.3301     0.3137  45.680   <2e-16 ***
drvf          5.6416     0.4405  12.807   <2e-16 ***
drvr         -0.2501     0.7098  -0.352    0.725
```

ANOVA and Regression

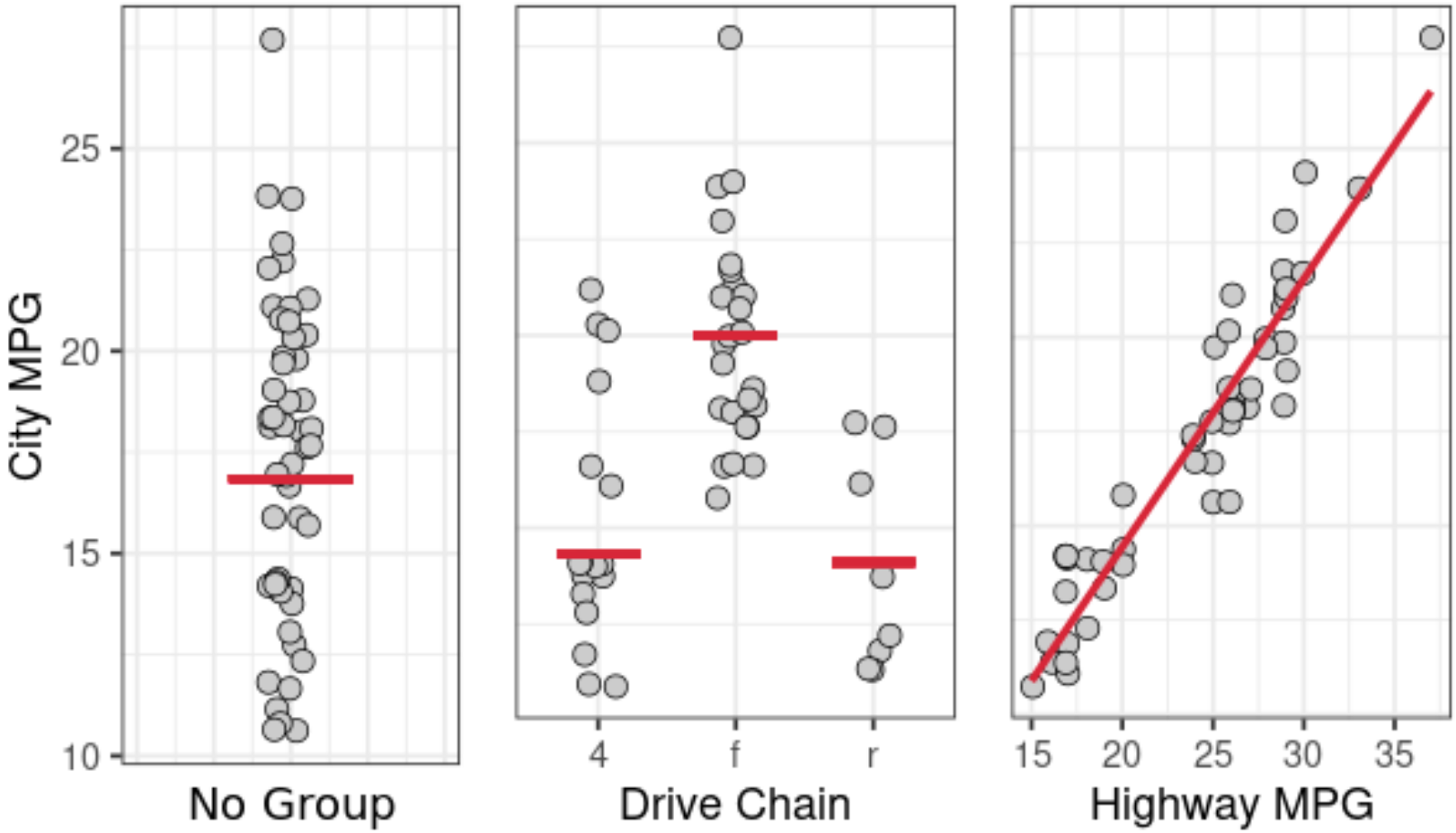
Just as ANOVA is a generalization of the t-test for multiple groups, regression is a generalization of ANOVA for any combination of variables

In most cases, regression is more robust, requiring fewer assumptions about the data while also providing statistical tests for each of the group categories

Most importantly, regression also allows us to predict a continuous outcome using continuous variables

Regression Example

Which of these do you suspect will have the smallest residual error?



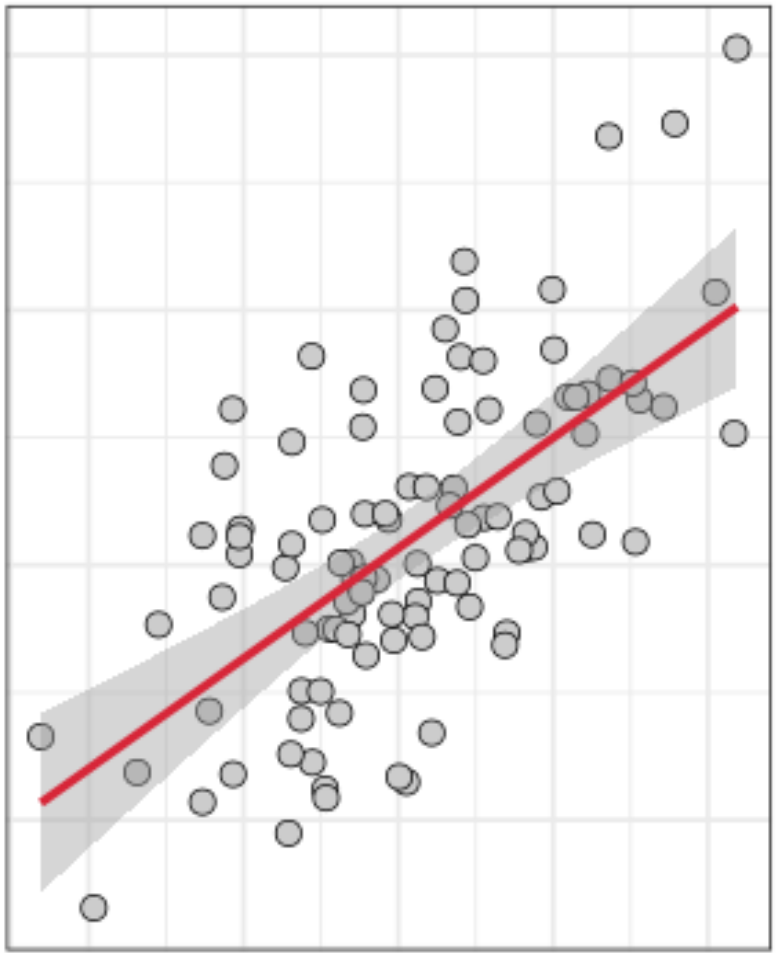
mpg Example

$$\hat{y} = \ldots$$

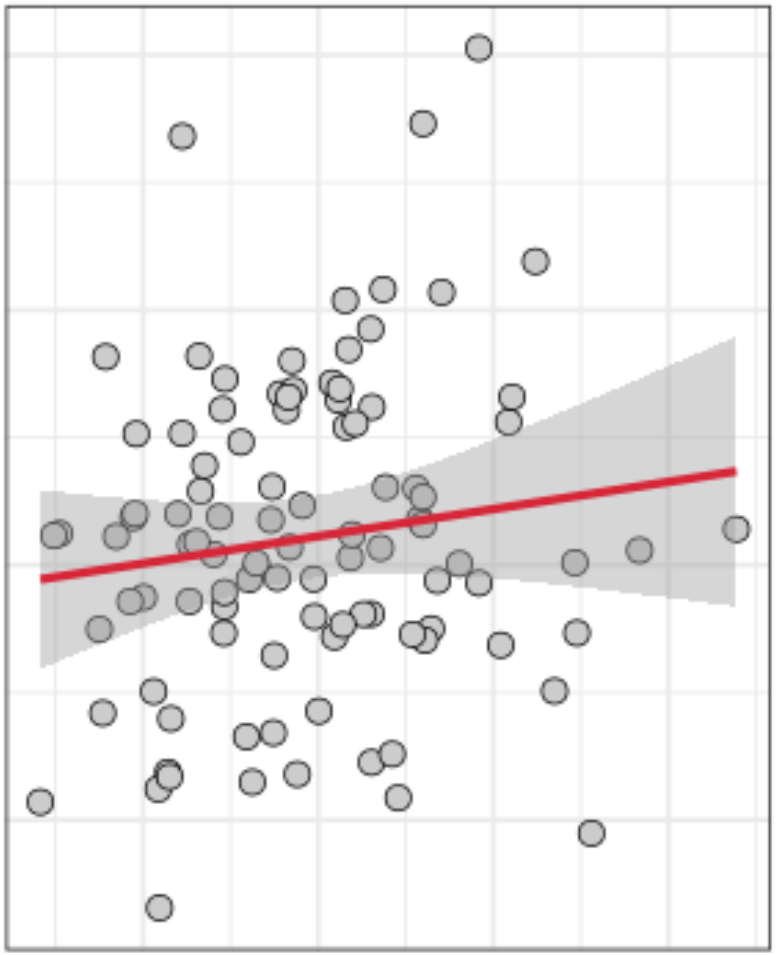
```
> lm(cty ~ hwy, mpg) %>% summary()


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.84420    0.33319   2.534   0.0119 *
hwy          0.68322    0.01378  49.585   <2e-16 ***


Residual standard error: 1.252 on 232 degrees of freedom
Multiple R-squared:  0.9138,  Adjusted R-squared:  0.9134
F-statistic:  2459 on 1 and 232 DF,  p-value: < 2.2e-16
```

Visualizing Hypothesis Testing

Reject H_0: $\beta = 0$



Fail to reject H_0: $\beta = 0$



Key Takeaways

- Regression is a generalization of ANOVA
- The β coefficients indicate how much a change in X impacts a change in Y
- Under the null, $H_0 : \beta = 0$, i.e., there is no relationship between predictor and outcome
- Likewise, the residuals correspond to the total within-group variability