Inference for Linear Regression

Grinnell College

December 8, 2025

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

A Note on Notation

There are a couples points of notation we will be using for the rest of the semester:

- Outcomes will be denoted with y, while explanatory variables will be denoted X
- β (beta) will represent our coefficients in our linear models
- The hat symbol ($\hat{y}$, $\hat{\beta}$) represents an estimated value, based on the data

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:

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

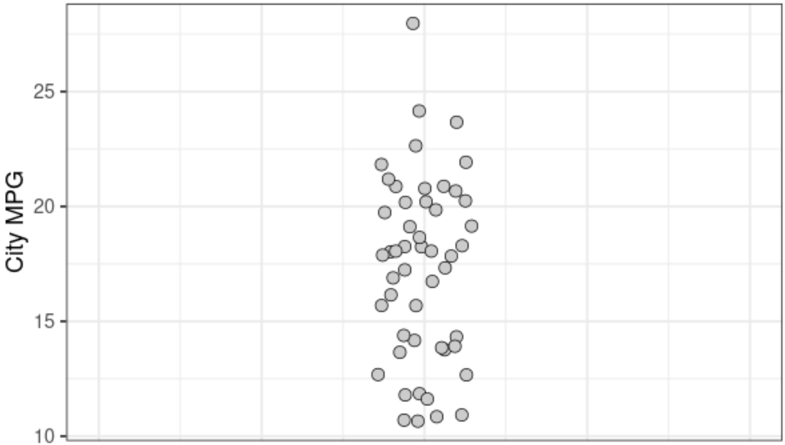
In this case by *better* we mean that we minimize the residual sum of squares, or the squared difference between our prediction and the true outcome

$$\text{Total Residuals} = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

$$= \sum_{i=1}^{n} r_i^2$$

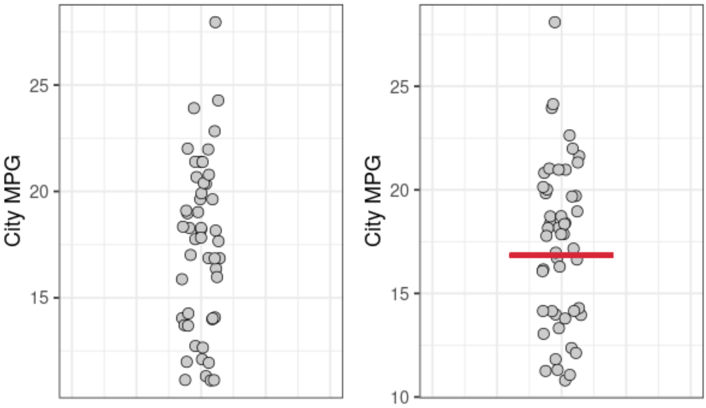
mpg Example

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



mpg Example

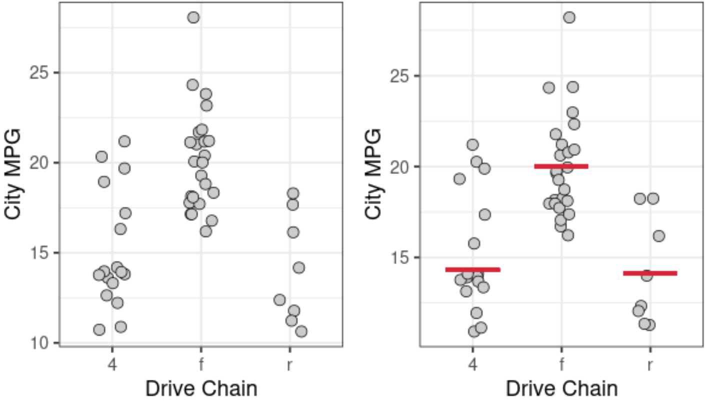
Using simply the overall mean, we would have total squared error of 4220



	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

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

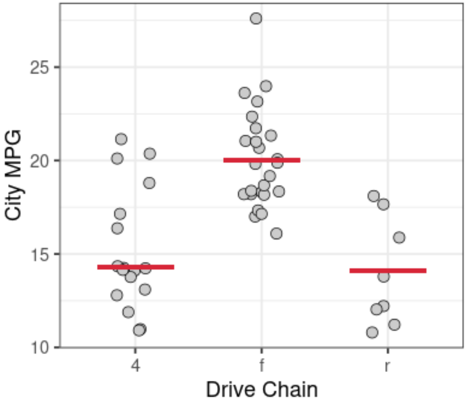
where $\mathbb{1}$ represents our *indicator variable* and, in the case of categorical variable regression, $\hat{\beta}$ represents the mean value for each group. This is precisely what we saw when we did this back in week 3

```
> 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
```



Inference and Regression

Similar to ANOVA, regression with a single categorical variable is concerned with minimizing residual error

However, instead of simply assessing whether or not there is *any* difference between groups, we are now interested specifically in estimating values of β in the expression

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

Inference and Regression

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

When considering a regression line, the null hypothesis represents the assumption that there is no linear relationship, and that the true value of β is equal to zero:

$$H_0 : \beta = 0$$

Given our estimate of $\hat{\beta}$, we are presented with a natural test statistic,

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

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

mpg Example

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

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

```
> 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

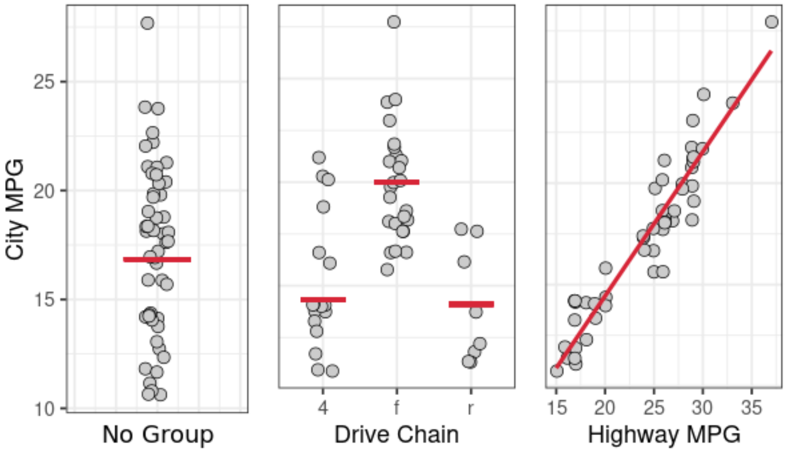
Just as ANOVA is a generalization of the t-test for multiple groups, regression is a generalization of ANOVA for any combination of variables

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?

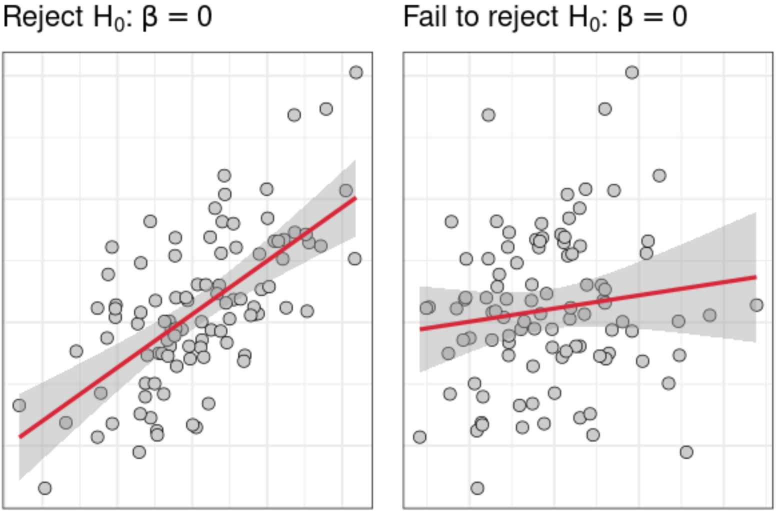


mpg Example

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

```
> 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



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
- R^2 gives an estimate of explained variance that, in the case of regression with a categorical variable, is identical to the sum of between-group variability
- Likewise, the residuals correspond to the total within-group variability