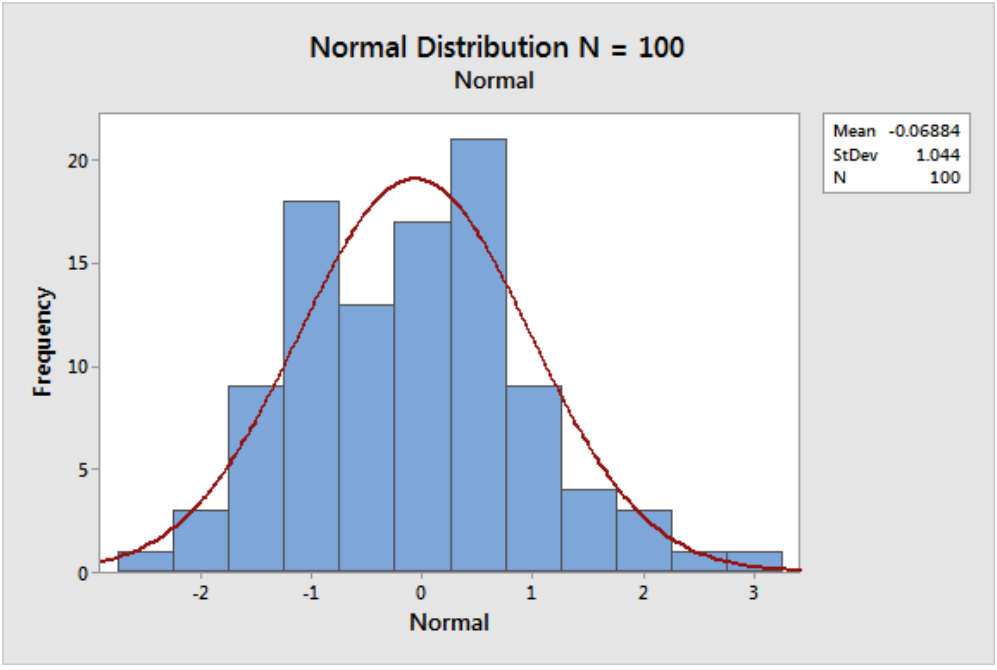
Samples and Parameters

One of the key ideas in introductory statistics is that we can use samples (e.g., $\bar{x}$) to estimate parameters (μ): (Meaning?)

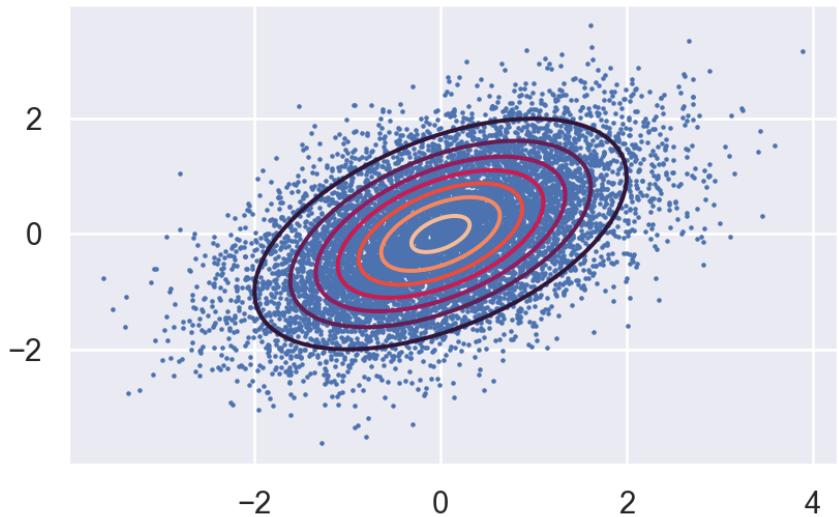
$$\mu \text{ is in } \bar{x} \pm Z_\alpha \left(\frac{\sigma}{\sqrt{n}} \right)$$



The point estimate $\bar{x}$ for μ depends on the sample . . .

Samples of the Bivariate Normal Distribution

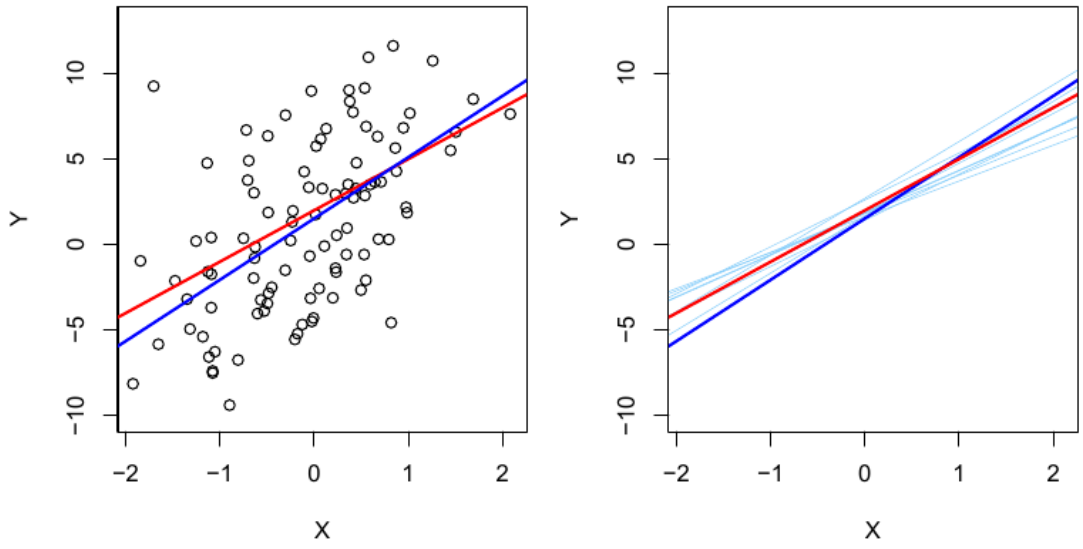
The same things is happening when we perform linear regression: We assume there is an underlying population linear regression line $y = \beta_o + \beta_1 x$ depending on the population distribution. But we *sample* from this distribution to get points $(x_1, y_1), \ldots, (x_n, y_n)$ and derive an *estimate*: $\hat{y} = \hat{\beta}_o + \hat{\beta}_1 x$



Question: How far from the true coefficients are our coefficients?

Confidence Intervals for Regression Coefficients

Example:



Confidence intervals for β_0 and β_1 are given by

$$\beta_i \text{ is in } \hat{\beta}_i \pm Z_\alpha SE(\hat{\beta}_i)$$

where:

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad (3.8)$$

Hypothesis Tests for Regression Coefficients

You also did hypothesis tests in introductory statistics . . .

We can also do them for regression coefficients:

$$H_0 : \beta_1 = 0 \qquad\qquad H_a : \beta_1 \neq 0$$

where

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

What does this mean?

These values are usually determined by linear regression model software:

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	7.0325	0.4578	15.36	< 0.0001
`TV`	0.0475	0.0027	17.67	< 0.0001

Measures of Model Accuracy

If we *believe* that a linear model is appropriate, we can ask *Just how good is the model*? There are two statistics for that:

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \qquad R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

Meaning?

These are also typically given by software:

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

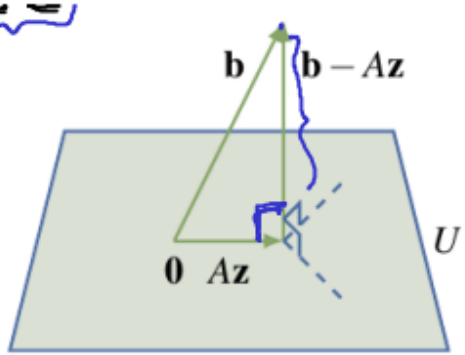
Multiple regression

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

Linear Regression With Several Variables

This can be converted to a system of linear equations $A\mathbf{x} = \mathbf{b}$, which doesn't have a solution

$\Rightarrow \|A\vec{x} - \vec{b}\|$ minimize at z



$U = \{A\vec{x} \mid \vec{x} \in \mathbb{R}^n\}$

$$\Rightarrow 0 = Az \cdot (Az - b) = (Az)^T (Az - b)$$
$$= z^T A^T (Az - b)$$
$$= z^T A^T A z - z^T A^T b$$

The end result: $A^T Az = A^T b$. "normal equations"

Solve for z.

Issues with multiple regression

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

Last time:

8	57.5	32.8	23.5	11.8

1. Is at least one of the predictors $X_1, \ldots, X_p$ even useful for predicting Y?
2. Do all of the predictors help explain Y, or can we get away with a subset
3. How well does the model fit the data?
4. How accurate is our prediction?

Issues with multiple regression

Hypothesis testing: Is there really a linear relationship?

We can use a hypothesis test . . .

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

versus the alternative

$$H_a : \text{ at least one } \beta_j \text{ is non-zero.}$$

This hypothesis test is performed by computing the *F-statistic*,

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)},$$

Issues with multiple regression

Model selection: Which variables to use?

- We can use different statistics to compare different models: Akaike information criterion (AIC), Bayesian information criterion (BIC), and *adjusted* R^2.
- We can create different models by choosing different variables . . .
- There are various *algorithms* for selecting among models such as . . . (See [I] p. 87.)

  Forward selection. We begin with the null model—a model that contains an intercept but no predictors. We then fit p simple linear regressions and add to the null model the variable that results in the null model lowest RSS. We then add to that model the variable that results in the lowest RSS for the new two-variable model . . .

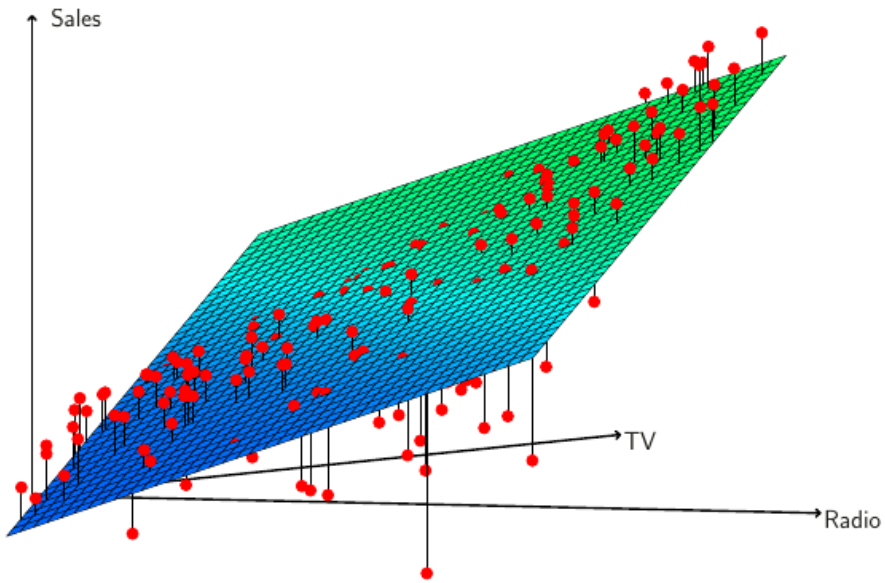
Issues with multiple regression

Model Fit: Which models gives the best "fit"?

- Use a modified RSE:

$$\mathrm{RSE} = \sqrt{\frac{1}{n-p-1}\mathrm{RSS}},$$

- Plot the data:

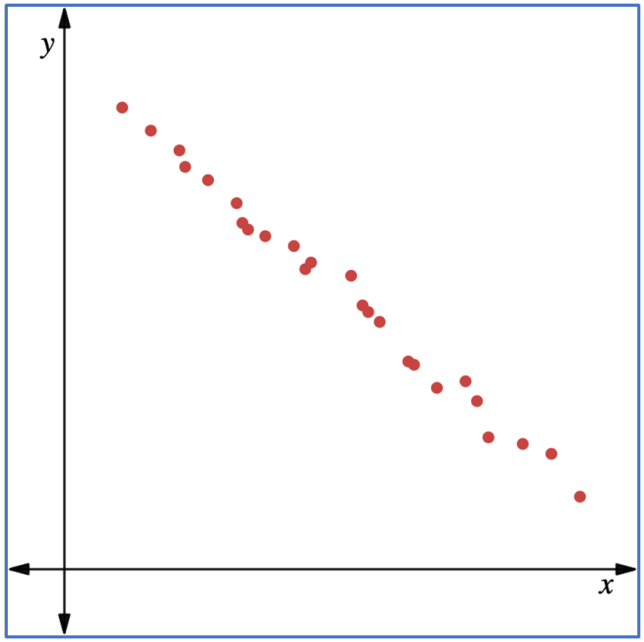


Still More Issues with multiple regression

The Homoscedasticity Requirement: Under the hood, all of this requires that the model be of the form
$f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \epsilon$, where $\epsilon \sim N(\mu, \sigma^2)$.

This is usually checked graphically . . .

Multicollinearity in Multivariable Regression: Sometimes *input* variables x_i, x_j can be correlated . . . and this is really bad.



A Linear Algebra Problem . . .

Theorem 5.6.1: Best Approximation Theorem

Let A be an $m \times n$ matrix, let $\mathbf{b}$ be any column in $\mathbb{R}^m$, and consider the system

$$A\mathbf{x} = \mathbf{b}$$

of m equations in n variables.

1. *Any solution $\mathbf{z}$ to the normal equations*

$$(A^T A)\mathbf{z} = A^T \mathbf{b}$$

is a best approximation to a solution to $A\mathbf{x} = \mathbf{b}$ in the sense that $\|\mathbf{b} - A\mathbf{z}\|$ is the minimum value of $\|\mathbf{b} - A\mathbf{x}\|$ as $\mathbf{x}$ ranges over all columns in $\mathbb{R}^n$.

2. *If the columns of A are linearly independent, then $A^T A$ is invertible and $\mathbf{z}$ is given uniquely by $\mathbf{z} = (A^T A)^{-1} A^T \mathbf{b}$.*

"If the columns of A are linearly independent . . ." **Why?**

4. *If two distinct rows (or columns) of A are identical, $\det A = 0$.*

5. *If a multiple of one row of A is added to a different row (or if a multiple of a column is added to a different column), the determinant of the resulting matrix is $\det A$.*

Idea: If two input variables are *correlated*, $det(A) \approx 0$.

Matrix Inverses and Regularization

If $det(A) \approx 0$, this means the coefficients of A^{-1} are going to be really big . . .

Theorem 3.2.4: Adjugate Formula

If A is any square matrix, then

$$A(\operatorname{adj} A) = (\det A)I = (\operatorname{adj} A)A$$

In particular, if $\det A \neq 0$*, the inverse of A is given by*

$$A^{-1} = \frac{1}{\det A} \operatorname{adj} A$$

In our case, this means $\beta = (A^T A)^{-1} A^T \mathbf{b} \approx \infty$. Why is this a problem?

Key point: Small changes in the data set (A) can lead to *big* changes in β.

Idea: We need a way to force the coefficients β to be smaller . . .

Ridge Regression

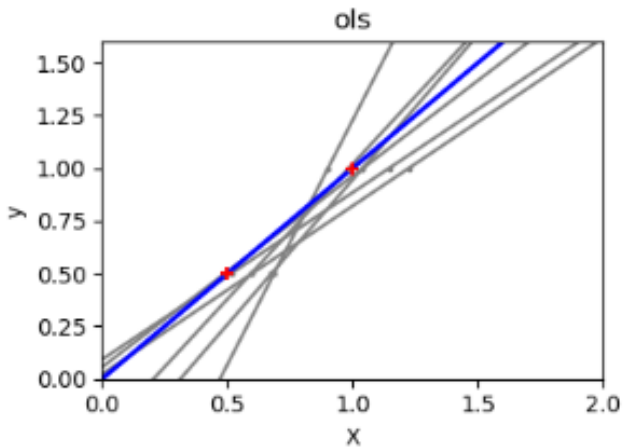
Recall: The coefficients in OLS(?) are found by minimizing

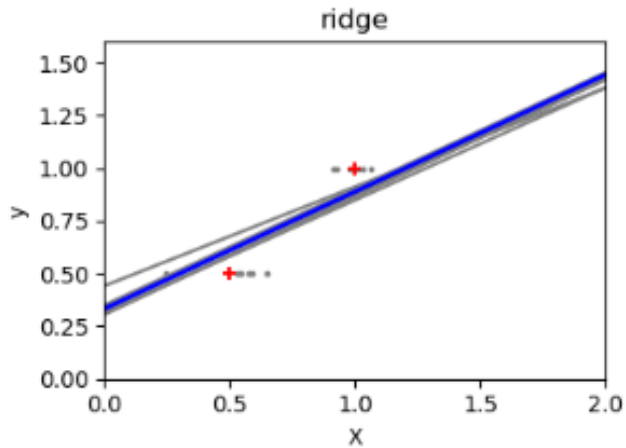
$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

In **ridge regression**, we find β by minimizing

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

where λ is a **hyperparameter** determined by the model builder.





Lasso Regression

Ridge regression is also called **L2 regularization**.

There is also **L1 regularization**:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

L^1 regularization is also called the **Least Absolute Shrinkage and Selection Operator (LASSO)**.

Key point: It can shrink some coefficients to zero, thereby performing qutomatic feature selection.

The Linear Regression Family

All of these regression types vary only in what function is being optimized:

$$\text{Linear Regression} : \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

$$\text{Lasso Regression} : \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

$$\text{Ridge Regression} : \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j{}^2$$