2.1: Introduction to Simple Linear Regression

Dr. Bean - Stat 5100

See **Handout 2.1.1** for information regarding the Toluca power company example.

1 Why Linear Regression?

Linear regression is good for:

- **Inference:** determine if there is a statistically significant linear relationship between two variables, while possibly accounting for the effect of additional variables.
  - Example: after accounting for the effects of square footage and age, are lot size and home sale price significantly linearly related?
- **Prediction:** use variables that are "easy" to measure to predict variables that are harder to measure.
  - Example: Use elevation (easy to measure) to predict annual snow accumulation (hard to measure).

Linear regression only works for variables that share a statistical relationship.

Terminology:

- Y - response variable
- X_i - predictor variables
- ϵ - error (or difference) term
- β_i - model parameters (true values are unknown and are estimated)

Linear Regression focuses on finding appropriate estimates of the model parameters (b_i):

The idea is that we want to select paramater estimates that make the predicted values of Y ($\hat{Y}$) close to the actual values of Y.

2 Ordinary Least Squares (OLS) Regression

If assumptions regarding residuals are satisfied (more in Handout 2.2), then the OLS estimates of the model parameters are "best."

What does it mean to be "best"?

- **unbiased** - given an infinite number of different samples of data, the average of my estimates will be equal to true (and unknown) value of the parameter.
  - In other words, my estimates are "centered" on the truth.
- **minimum variance** - the variation in the estimate from sample to sample is the smallest of all possible estimation methods.

Applications - Toluca Example:

Let X represent the lot size and let Y represent the total work hours. Based on the initial scatterplot, we assume that the relationship between X and Y can be modeled as

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

OLS seeks to minimize:

$$Q = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2,$$

which requires us to select estimates b_0 and b_1 that minimize

$$Q = \sum_{i=1}^{n} \left(Y_i - (b_0 + b_1 X_i)\right)^2 = f(\mathbf{X}).$$

We can use multivariable calculus to find the minimum of Q by finding the critical points, i.e.

$$\nabla Q = \nabla f(\mathbf{X}) = 0.$$

The single critical point that minimizes Q is

$$b_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i \left(X_i - \bar{X}\right)^2}$$
$$b_0 = \bar{Y} - b_1 \bar{X}.$$

Obtain OLS estimates automatically in SAS with:

```
proc reg data=toluca;
  model workhours = lotsize;
  title1 'Simple linear model';
run;
```

Or in R with:

```
data(toluca)
toluca_lm <- lm(workhours ~ lotsize, data = toluca)
toluca_lm
```

Equation Estimates:

$b_0 = 62.37,\ b_1 = 3.57$

Model Equation:

$\hat{Y} = 62.37 + 3.57(lotSize)$

The Critical Assumption

OLS least squares hinges on the assumption that

$$\epsilon \overset{iid}{\sim} N(0, \sigma^2)$$

- **independent:** Knowing the value of one of the model residuals tells you nothing about any of the others.
- **identically distributed:** All of the residuals come from the same distribution.
- **Normal Distribution:** The model residuals follow a normal (bell shaped) distribution.
- **zero mean:** The average of the residuals is zero (unbiased estimates).
- **constant variance:** The spread of the residuals about the line is the same across the range of X and the range of predicted values.

If the assumptions hold, then the simple linear regression can be visualized as in Figure 1.

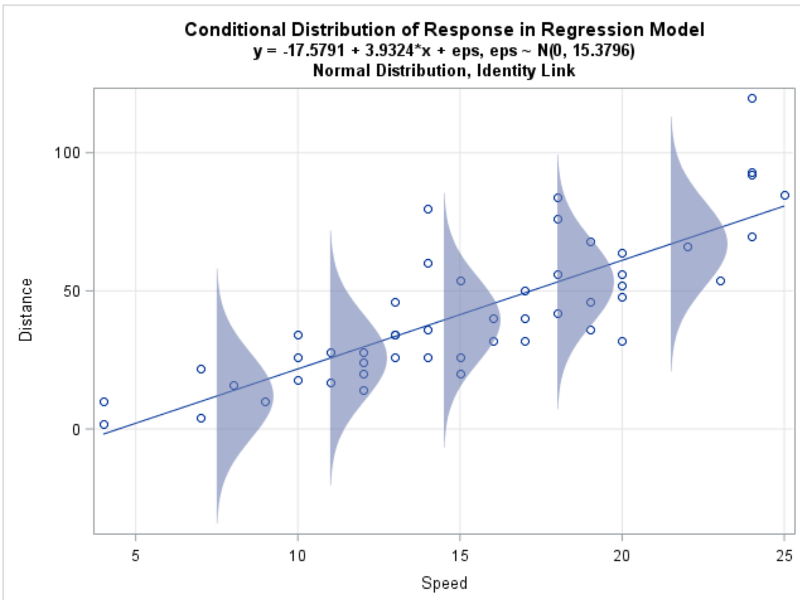


Figure 1: Sample visualization taken from Rick Wicklin on The DO Loop.

In other words, Y follows a normal distribution with a center that is conditional on X.

Estimating σ

Estimating the variance about the regression line:

- Allows us to get a measure of the model fit: lower relative MSE $\rightarrow$ better model.
- All significance tests of model coefficients are based on our estimate of σ.

Estimation of ϵ in Theory

Suppose that $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ were an observed sample from some population. (In practice, ϵ is estimated as the residuals of our OLS model, represented as e_i.)

We could then estimate $\text{Var}(\epsilon)$ as

$$\frac{1}{n-1}\sum_{i=1}^{n}(\epsilon_i - \bar{\epsilon})^2$$

.

Note that the variance calculation requires the estimation of $\mu_\epsilon = \bar{\epsilon} = \frac{1}{n}\sum_i \epsilon_i$.

This calculation "constrains" one of the ϵ_i. This means that if we know $e\bar{psilon}$ and $\epsilon_1, \ldots \epsilon_{n-1}$, then we can know ϵ_n.

We call the number of unconstrained observations the "degrees of freedom" (DF).

Every time you estimate a parameter, **you lose one degree of freedom.**

Think of observations as currency. We spend money to estimate things and our degrees of freedom are the leftover cash.

Estimation of ϵ in Practice

Why is it that we can't directly obtain the values of ϵ?

We don't know the true regression line, so we cannot know the true values of epsilon.

We can obtain estimates of the residuals e_i through the regression line:

$$e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i).$$

OLS, by design, makes $\sum_i e_i = 0 \to \bar{e} = 0$, meaning I don't have to spend any DF to obtain $\bar{e}$.

Variance of the Residuals

$$\hat{\sigma}^2 = s^2 = \frac{1}{df_E} \sum_{i=1}^{n} e_i^2 = \frac{1}{n-2} \sum_{i=1}^{n} (e_i - \bar{e})^2$$

We call this estimate the **"mean square error"** or MSE.

2.2: Diagnostics and Remedial Measures

Dr. Bean - Stat 5100

1 Why Diagnostics

Recall that the nice properties of the OLS coefficient estimates relied on the assumption that

$$\epsilon \overset{iid}{\sim} N(0, \sigma^2).$$

Model Assumptions in Linear Regression

1. X and Y share a linear relationship
   - X and Y can be related in a non-linear way, but OLS regression cannot be used in this case
2. model describes all observations
   - no outliers or influential points
3. additional predictor variables are unnecessary
   - there is no additional information to "extract" from ϵ
4. ϵ's follow a normal distribution
   - Crucial for small sample sizes, not so critical for large (> 500) sample sizes due to central limit theorem.
5. ϵ's have constant variance
6. ϵ's are independent (possibly related to item #3)

We check assumptions using **diagnostics** and fix violated assumptions using **remedial measures**. Violations are most apparent in the **error terms** ($\epsilon_1, \ldots, \epsilon_n$) so we focus on **residuals** ($e_1 \ldots e_n$).

There are both **graphical** and **numerical** checks of the assumptions regarding residuals, but the graphical assumptions are more informative.

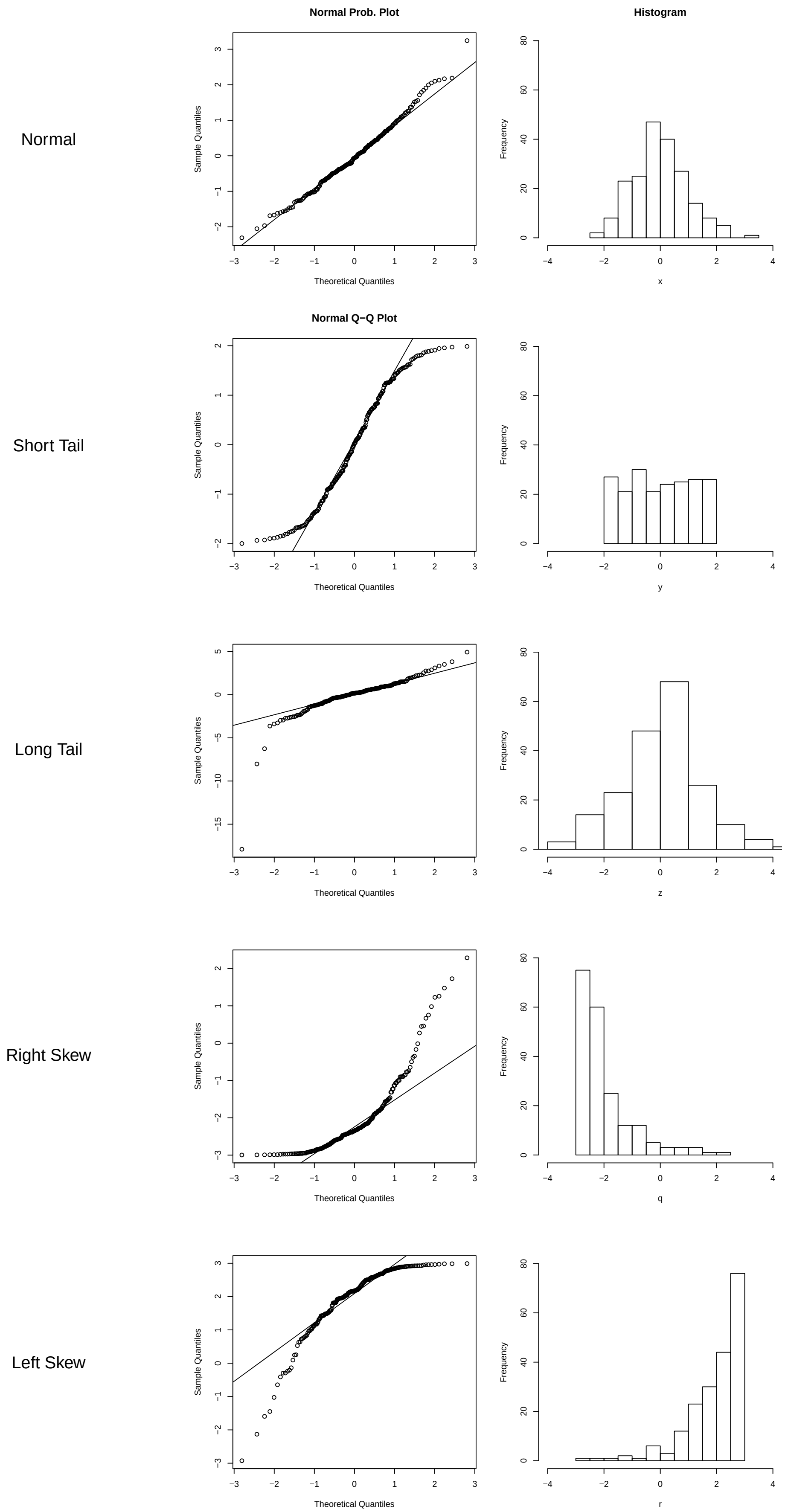


Figure 1: Example scatterplot and histograms for different data distributions.

2 Graphical Diagnostics

- **Boxplot:** quick way to check for the symmetry of residuals.
- **Histogram:** Way to check the shape of a distribution.
  - SAS will overlay a normal curve on the histogram of residuals to help check for normality.
- **Normal Probability Plot:** a qq-plot where the quantiles of the data are compared to the expected quantiles under a normal distribution
  - Expected values under normality have a mean of 0 and a SD = $\sqrt{\text{MSE}}$.
  - See page 111 in textbook for method for details about how to approximate expected observations under normality.
  - If data are approximately normal, the residuals in the Normal Probability Plot should closely follow a straight line.
- **Sequence Plot:** Line plot with residual values on the Y axis and observation number on the X-axis.
  - Can "connect" the dots because there is only one Y value for every X value.
  - Look for patterns in the residuals across time/observation number.
    - Patterns would suggest that the residuals are **not independent.**
- **Residual Plot**
  - Plot e vs X or e vs $\hat{Y}$
  - Look for non-linearity and or non-constant variance in these scatterplots.

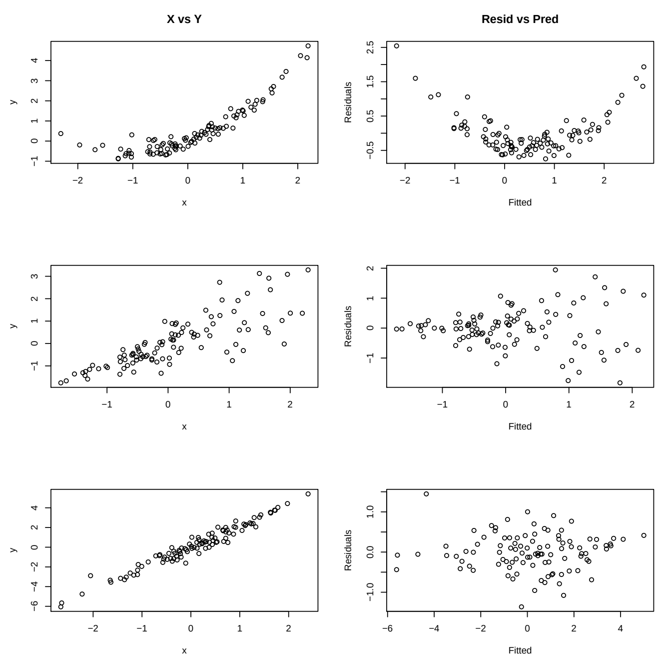


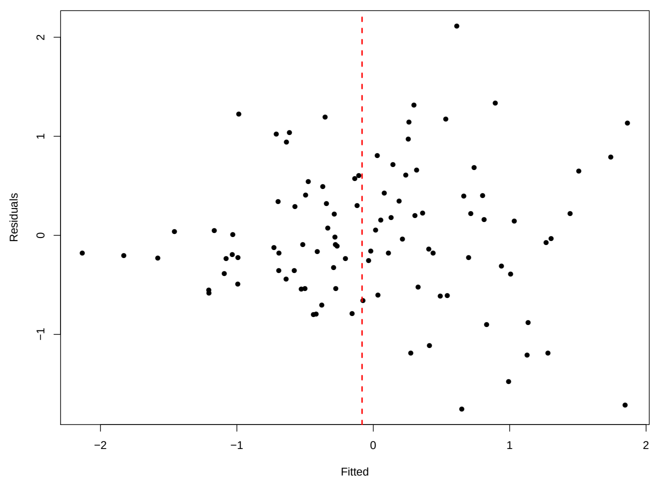
Figure 2: Plots showing 1) non-linearity, 2) non-constant variance, and 3) satisfied assumptions.

3 Numerical Diagnostics

Numerical diagnostics seek to determine if violations of model assumptions are statistically significant.

3.1 Some Numerical Tests

- Brown-Forsythe (BF) test of constant variance:



  - Split the data into two groups based on the median predicted value.
  - Calculate the median absolute deviation (MAD) $d_i = |e_i - \tilde{e}|$, where $\tilde{e}$ is the median within each group (lower and upper).
  - Conduct a two-sample t-test of d's to determine if the average of d_i's within each group are significantly different.
  - **Null Hypothesis: The variance of ϵ is constant.** (Estimated with residuals e).

  Toluca Example: BF p-value is .2, which suggests there is not significant evidence of non-constant variance.

- Correlation Test of Normality
  - Calculate correlation between observed e's and the "normal-expected" e's ($e*$). Similar to expected residuals in qqplot.
  - **Null hypothesis: ϵ follows a normal distribution.**
  - If the correlation isn't at least as big as the critical value for $\alpha = 0.05$ in Table B.6 for a given sample size n, then reject H_0.
  - **NOTE: The p-values provided in the SAS macro output mean *nothing* for the correlation test of normality.**

  Toluca Example: Correlation of $0.992 > 0.96$, so there is not significant evidence of non-normality.

- F-test for lack of fit (test of linearity between X and Y)
  - See textbook 3.7 for details.
  - Requires multiple observations at one or more X-levels. (Hard to do in observational studies or studies with multiple X-variables).
  - Basically, the test compares the regression predictions to the empirical average of Y at X-levels with multiple observations.
  - **Null hypothesis: The regression function is linear.**

Toluca Example: p-value 0.69 $>$ .05 suggests there is no significant evidence of non-linearity.

TABLE B.6 Critical Values for Coefficient of Correlation between Ordered Residuals and Expected Values under Normality when Distribution of Error Terms Is Normal.

	Level of Significance α				
n	.10	.05	.025	.01	.005
5	.903	.880	.865	.826	.807
6	.910	.888	.866	.838	.820
7	.918	.898	.877	.850	.828
8	.924	.906	.887	.861	.840
9	.930	.912	.894	.871	.854
10	.934	.918	.901	.879	.862
12	.942	.928	.912	.892	.876
14	.948	.935	.923	.905	.890
16	.953	.941	.929	.913	.899
18	.957	.946	.935	.920	.908
20	.960	.951	.940	.926	.916
22	.963	.954	.945	.933	.923
24	.965	.957	.949	.937	.927
26	.967	.960	.952	.941	.932
28	.969	.962	.955	.944	.936
30	.971	.964	.957	.947	.939
40	.977	.972	.966	.959	.953
50	.981	.977	.972	.966	.961
60	.984	.980	.976	.971	.967
70	.986	.983	.979	.975	.971
80	.987	.985	.982	.978	.975
90	.988	.986	.984	.980	.977
100	.989	.987	.985	.982	.979

Source: Reprinted, with permission, from S. W. Looney and T. R. Gulledge, Jr., "Use of the Correlation Coefficient with Normal Probability Plots," *The American Statistician* 39 (1985), pp. 75–79.

4 Remedial Measures

If assumptions are violated, your options are:

- Give up (at least on linear regression).

- Alternatives to OLS like Regression Trees, Quantile Regression etc. (more later in the semester).
- Depending on the X vs Y relationship, try non-linear regression (more later in the semester).
- *For non-normality and heteroskedasticity:* Variable Transformations on X or Y (not e).
  - Sometimes, a combination of transformations on both X and Y variables is needed.
  - NOTE: Removing "outlier" points from a model should be seen as a measure of last resort.

Box-Cox Approach

Great starting point to determine candidate transformations, but no "magical" solution.

- Define new response variable
$$Y' = \begin{cases} \text{sign}(\lambda)Y^{\lambda} & \lambda \neq 0 \\ \log(Y) & \lambda = 0 \end{cases}$$
  (Note that sign(λ) preserves the original ordering of the response variable).
- Consider the theoretical model
$$Y_i^{\lambda} = \beta_0 + \beta_1 X_i + \epsilon_i$$
- Consider a set of candidate lambda values, use maximum likelihood estimation to determine the "best value."
  - Maximum likelihood estimation: "which value of λ is the most likely, given the data that I have observed?
- **When possible: pick an *interpretable* transformation that is close to the transformation recommended by SAS.**

  Ex: if $\lambda = .009$ is recommended, probably go with $\lambda = 0 \rightarrow \log(\lambda)$

In SAS:

```
proc transreg data=plasma;
 model boxcox(<response variable> / lambda=<lower> to <upper> by <step size>)
       = identity(<explanatory variable>) ...;
run;
```

In R:

```
library(MASS)
boxcox(<response variable> ~ <explanatory variable>, data = <dataframe>)
```

Omitted Predictors

Think of regression as a form of data mining: we want to extract as much *information* as we can from our *data*.

If we failed to extract all the information from the data, this may show up as a trend in the plot the residuals (which SHOULD have a constant mean of 0).

We will discuss more about how to extract *time* related information in data at the end of the semester.

If you apply multiple methods to fix violations of assumptions, make sure to check that the final model actually fixed the violations of assumptions.

2.3: Simple Model Inference

Dr. Bean - Stat 5100

Recall the simple linear model

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \epsilon_i.$$

Inference is the process by which we make a decision about whether an observed difference from an expectation was simply due to chance or not.

In other words, inference is the process of making conclusions given incomplete information.

1 Why Inference?

Hypothetical questions:

- Suppose you found out that there is not significant relationship between study time and final grades in Stat 5100, how would this effect your approach to this course?
- Suppose you have the flu and you find out from a clinical trial of a new flu drug that those who took the drug had slightly shorter flu durations than those who took the placebo, but that the difference was likely due to chance. How likely would you be to purchase this drug?

In the absence of complete information, inference is an efficient way to decide what associations are "real" and which are not.

2 Hypothesis Testing

Recall that hypothesis testing is the formal way by which we determine if an observed difference from an expectation was due to chance.

Process

- Define a null and alternative hypothesis.
  - H_0 : "no effect"
  - H_a : "some effect"
- Define a test statistic:
  - Compares what we observed to what we expected if the null hypothesis was true.
- Determine the "sampling distribution"
  - Determines the natural variation in the test statistic that we would expect if we took many different samples from the same population.
  - In practice, we only ever take one sample. Statistical theory is what allows us to determine what the distribution would look like if we could take many samples.
  - The distribution often relies on **model assumptions**.
- Get p-value

- This is the probability of obtaining an observation as far, or farther, away from what we expected if the null was true.

- Make conclusion in context.
  - If the p-value is small ($< \alpha$), then it is unlikely that we would have obtained our observation if the null hypothesis is true. This provides evidence that the observed difference between our observation and expectation is REAL, and not simply due to chance.

2.1 Toluca Example:

If model assumptions are satisfied, then $b_1 \sim N(\beta_1, \sigma^2\{b_1\})$.

$\sim$ means "follows" while $\sigma^2\{b_1\}$ represents the true variance of b_1, as estimated by $s\{b_1\}$.

Recall that, if the null hypothesis is true, then $\beta_1 = 0$. Thus, our test statistic becomes

$$t = \frac{b_1 - 0}{s\{b_1\}} \sim t_{df_E} = 10.29$$

with $25 - 2 = 23$ degrees of freedom with a **p-value** $<$ **0.0001**.

where df_E is the degrees of freedom for the residuals, which is $n-2$ in the simple linear model case

draw t-distribution and shade the area that represents the p-value

Since our p-value is lower than our level of significance (which is typically 0.05 and something we set beforehand), we would **reject** the null hypothesis **and conclude** that there is significant evidence that lot size and work hours are linearly related.

Where did $\alpha = 0.05$ come from?

Short answer: Sir Ronald Fisher, a prominent statistician, made it up:

> It is a common practice to judge a result significant, if it is such a magnitude that it would have been produced by chance not more frequently than once in twenty trials. This is an arbitrary, but convenient, level of significance for the practical investigator...[1]

However, $\alpha = 0.05$ has proven to be a good level of significance that balances the probability of Type I (claiming a difference when there isn't one) and Type II (claiming *no* difference when there *is* one).

Consider the following

You wish to determine if Aggie ice cream is more fattening than other ice cream shops in Logan. Suppose your null hypothesis is: "Aggie ice cream has the same number of calories per cup as Charlie's ice cream." You then conduct a test and obtain a p-value of 0.048, indicating that there is evidence that the average caloric counts are significantly different. You then realize that you forgot

[1] As on p99 of "The Lady Tasting Tea" (2001) by David Salsburg. See `http://jse.amstat.org/v16n2/velleman.pdf` for more discussion about statistical theory.

to include five recorded observations in your study. When you include these additional observations, you obtain a p-value of 0.052, indicating no significant difference.

P-values should inform an analysis, rather than become the analysis.

Confidence Intervals

- General Form:
$$\text{estimate} \pm (\text{critical value}) \times (\text{SE of estimate})$$
- For β_1:
$$b_1 \pm t^* \times s\{b_1\}$$
- Interpretation:
  - We are 95% confident that the true value of β_1 is contained in this interval.
  - If we were to create 100 confidence intervals from 100 different samples, we would expect about 95 of them to contain the true β_1.

Testing $H_0 : \beta_1$ at level α is the same as checking whether 0 is inside the $(1-\alpha)100\%$ CI for β_1.

Model Inference

All previous examples test whether an individual X variable has a significant linear relationship with Y. We will now look at some measures of model usefulness that apply when there is more than one X variable.

Ingredients of Model Inference

- Sum of Squares
  - $SS_{total} = \sum_i \left(Y_i - \bar{Y}\right)^2 \propto$ variance of Y
  - $SS_{error} = \sum_i \left(Y_i - \hat{Y}_i\right)^2 = \sum_i e_i^2 \propto$ variance not explained by model
  - $SS_{model} = SS_{total} - SS_{error}$

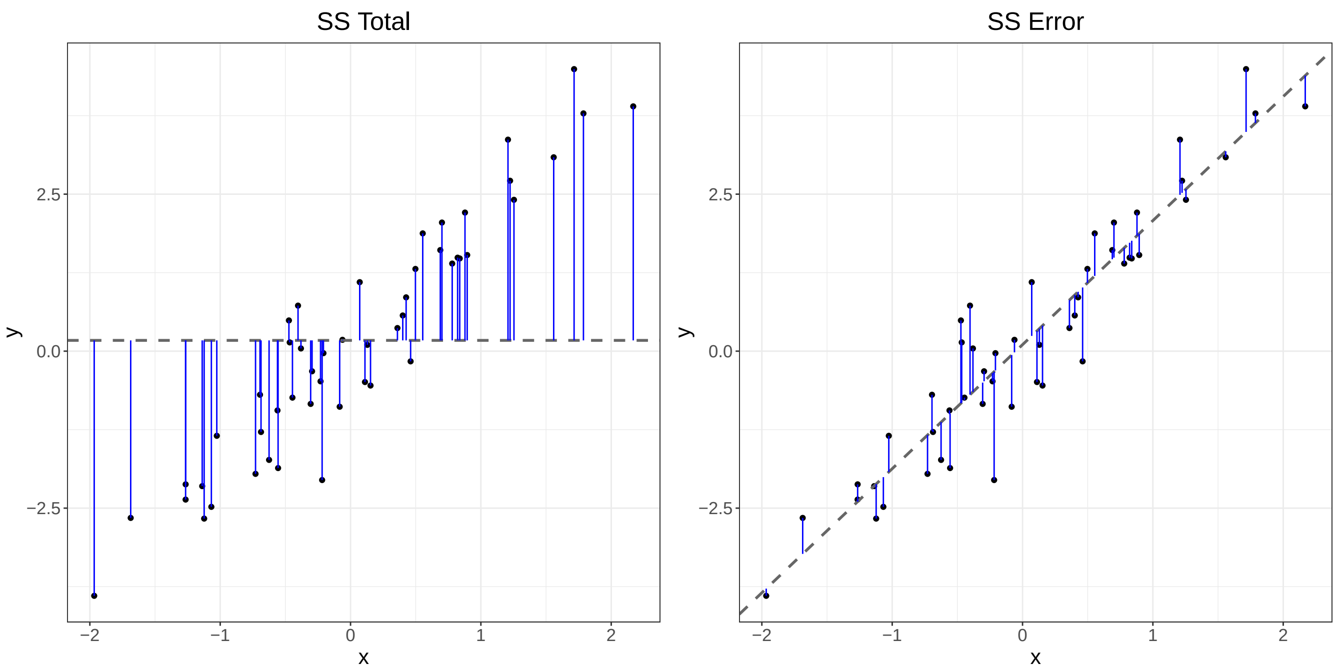


Figure 1: Illustration of SS_{total} and SS_{error}.

- Mean Square: $MS = \frac{SS}{df}$
- $F = \frac{MS_{model}}{MS_{error}}$
- $R^2 = \frac{SS_{model}}{SS_{total}} = 1 - \frac{SS_{error}}{SS_{total}}$
  - Interpretation: the percent of the variation in Y that is explained by the model.
- MSE = Mean Square Error = $\hat{\sigma}^2$ = our best estimate of the error variance ($\epsilon \sim N(0, \sigma^2)$).

Toluca Example:

$R^2 = 0.82$ (from Handout 2.1.1) which means that about 82% of the variation in work hours is explained by lot size.

Two other ways to look at $H_0 : \beta_1 = 0$:

1. How much worse would the model fit be if we dropped the β_1 term?

   Reduced Model (null hypothesis): $Y_i = \beta_0 + \epsilon_i$.

   Full Model: $Y_i = \beta_0 + \beta_1 X_{i,1} + \epsilon_i$.

   F-statistic looks at chance in SS_{error} between these two models.

   Can be extended to consider removal of *subsets* of X variables (more later in the semester).

2. Let $\rho = Corr(X, Y)$ = true, unknown correlation coefficient, which we estimate with the sample correlation (r).

   - When there is only one x-variable in the model it follows that

$$H_0 : \beta_1 = 0 \equiv H_0 : \rho = 0.$$

Inference on the Response Variable Y

We can create interval estimates for the response variable.

$$\hat{Y} \pm t_{df_E}(1 - \frac{\alpha}{2}) * SE\{\hat{Y}\}$$

Two Intervals:

- **Confidence Interval**: Interval estimate of **mean** (or expected) Y for **population** of all $X = X_h$.
$$SE\{\hat{Y}\} = s\{\hat{Y}_h\} = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_i (X_i - \bar{X})^2} \right]$$
- **Prediction Interval**: Interval estimate of **predicted** Y for a single [new] observation at $X = X_h$
$$SE\{\hat{Y}\} = s\{\hat{Y}_{h(new)}\} = \hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_i (X_i - \bar{X})^2} \right]$$

Toluca Example (with $X_h = 10$)

- If we were to go to a new (single) lot of 10 acres, we are 95% confident that the work hours would be between

  -13.6 (truncate at 0) and 209.7.

- If we were to consider all possible 10 acre sized lots, we are 95% confident that the mean work hours across all these lots would be between

  50.5 and 145.6.

Note: Most models have more than one predictor variable, we will use the following common notation throughout the remainder of this course:

- n = sample size
- p = number of β_j's in the model (including the intercept)
- $df_E = n - p$

2.4: Simultaneous Inference and Important Considerations

Dr. Bean - Stat 5100

Simultaneous inference is when we want to conduct multiple tests of significance at the same time.

1 Why Simultaneous Inference?

In handout 2.3, we conducted inference for parameters one at a time. We need to change our approach when looking at multiple parameters simultaneously.

How and why do we need to change our approach when conducting simultaneous inference?
(check out **this comic** for help).

If we conduct several tests at the same level of significance, the probability of getting one false positive result (a type I) error becomes much higher than α.

As a result, we need to adjust the level of significance to account for a "multiplicity" of testing.

2 Bonferroni Adjustment

Multiplicity:

- Let A_j = event that an individual $(1-\alpha)100\%$ CI does not contain the true value of β_j.
- $P(A_0) = P(A_1) = \alpha \rightarrow$ Type I Error
  - $P(\text{NOT}A_j)$ = probability that an interval contains the true value of β_j.
- **Bonferroni Inequality:** $P(\text{NOT}A_0 \text{ AND NOT } A_1) \geq 1 - P(A_0) - P(A_1)$

This means that if we conduct g tests at a confidence level of $(1 - \frac{\alpha}{g})$, then we are guaranteed that overall level of confidence for all intervals *considered jointly* will be at least $(1-\alpha)$, we call this the **Bonferroni adjustment**.

- **Bonferroni Advantage:** Can be applied in *any* situation that requires a multiplicity adjustment, including simultaneous intervals for $\hat{Y}$ at multiple X_h levels.
- **Bonferroni Disadvantage:** Can be overly conservative, producing inefficient (unnecessarily wide) intervals.

Comparison of Simultaneous Intervals for $\hat{Y}$

- Confidence intervals (mean response)
  - Bonferroni

$$\hat{Y} \pm t_{n-p}(1 - \frac{\alpha}{2g}) * s\{\hat{Y}_h\}$$

- Working-Hotelling (WH)

$$\hat{Y} \pm W * s\{\hat{Y}_h\} \qquad \left(W = \sqrt{pF_{p,n-p}(1-\alpha)}\right)$$

Notice that the W-statistic does not consider g

  * WH provides a "confidence band" for the entire regression line (all possible X_h levels).
  * This means the WH interval at any individual X_h will be wider than the t-based confidence interval, but the WH intervals will eventually be narrower than Bonferroni confidence intervals if enough X_h are considered.

- Prediction intervals (new response)
  - Bonferroni

$$\hat{Y} \pm t_{n-p}(1 - \frac{\alpha}{2g}) * s\{\hat{Y}_{h(new)}\}$$

  - Scheffe (chef-eh)

$$\hat{Y} \pm S * s\{\hat{Y}_{h(new)}\} \qquad \left(S = \sqrt{gF_{g,n-p}(1-\alpha)}\right)$$

Rule of Thumb: Always pick the most efficient interval that guarantees your intended type I error (α).

Table 1: Summary of Methods for Simultaneous Intervals

Simultaneous Interval on:	**Methods**
β's	Bonferroni
Population means of Y at multiple X_h	Bonferroni or Working-Hotelling
Predictions for Y at multiple X_h	Bonferroni of Scheffe

3 Inverse Prediction

Problem: What is the value of X_h necessary to achieve a specific value of $\hat{Y}$.

Solution: solve for X.

$$\begin{aligned} \hat{Y} &= b_0 + b_1 X_h \\ b_1 X_h &= \hat{Y} - b_0 \\ X_h &= \frac{\hat{Y} - b_0}{b_1} \end{aligned}$$

Problem: Use Y to predict values of X.

Solution: DO NOT solve for X.

Why?

- The least squares slope estimate of regression model that predicts Y using X: $\rho\frac{SD\{Y\}}{SD\{X\}}$.
- The least squares slope estimate of regression model that predicts X using Y: $\rho\frac{SD\{X\}}{SD\{Y\}}$.
- Notice that the slopes are NOT inverses of each other.

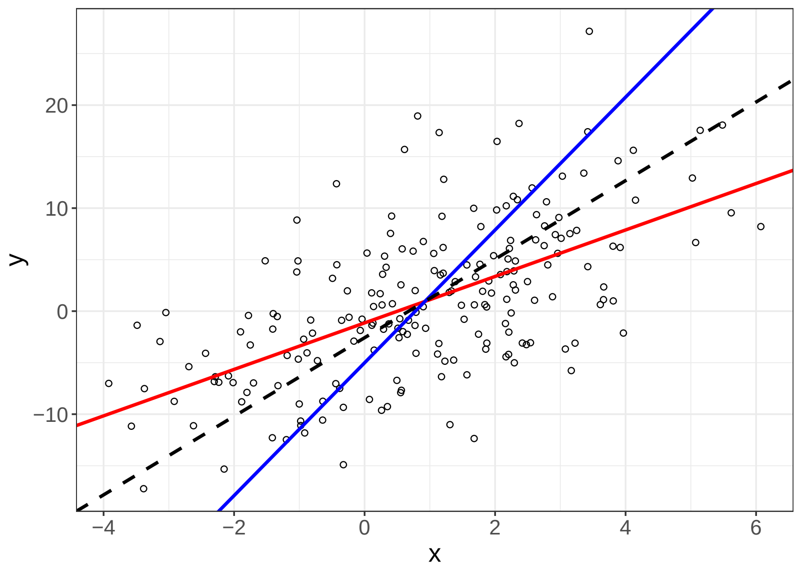


Figure 1: Scatterplot of points along with the regression line that uses X to predict Y (red), the regression line that uses Y to predict X (blue), the SD line (black).

4 Cautions for Linear Regression

- Remedial measures may not fix violations of assumptions
  - May need to abandon OLS regression altogether
- Interpretation: Sometimes the X vs Y relationship may look counterintuitive
  - May be the result of omitted predictors
- R^2 can be abused
  - Higher R^2 $\rightarrow$ not always better model
  - Lower R^2 $\rightarrow$ does not mean there is no linear relationship

2.5: Multiple Linear Regression

Dr. Bean - Stat 5100

1 Why Multiple Linear Regression?

- Models that use a single explanatory variable to predict a response are very limited in terms of its capability.
- We are often interested in determining the effect of an explanatory variable on the response variable *after* accounting for the effects due to other explanatory variables.
  - Example: Is there a difference in the pay based on gender after accounting for job type and hours worked?

2 What Changes from Simple Linear Regression?

1. Interpretation of coefficients

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \\ \beta_0 &= E[Y|X_1 = X_2 = \cdots = X_{p-1} = 0] \\ \beta_k &= \text{expected (or average) change in Y} \\ &\quad \text{for every unit increase in predictor } X_k, \\ &\quad \underline{\text{while holding all other predictors constant}} \end{aligned}$$

Need all three elements for a correct interpretation.

β_k sometimes called "partial regression coefficient" because it reflects partial effect of X_k on Y after accounting for effects of other predictors

2. ANOVA table

- model $df = p - 1 = \#$ of predictor variables
- error $df = n - p$
  – we have to "spend" more degrees of freedom to calculate the additional coefficients
- model F-test more meaningful:

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0 \\ H_a &: \beta_k \neq 0 \text{ for at least one } k = 1, \ldots, p-1 \end{aligned}$$

- R^2 called coefficient of multiple determination (still interpret as % variance in Y explained by model); $\sqrt{R^2}$ called coeff. of multiple correlation

3. Refer to regression "surface" instead of "line"

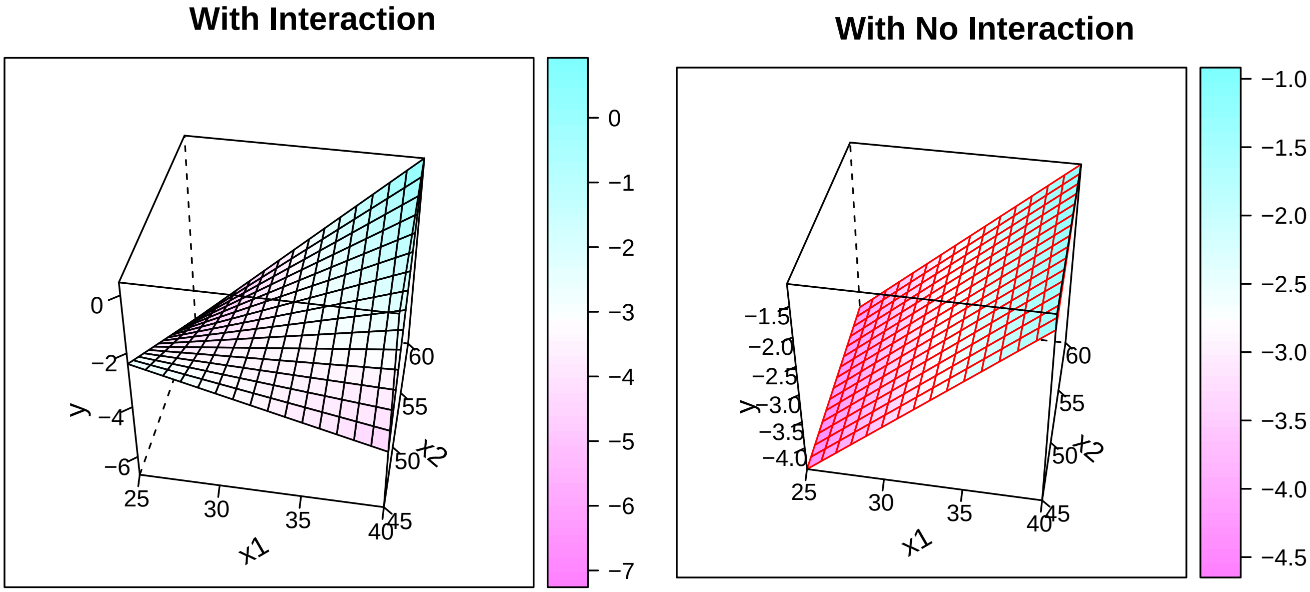


Figure 1: Regression surface using two X variables to predict Y. Harder to visualize when there are more than two predictor variables.

4. F-test for lack of fit less practical
   - requires multiple observations at one or more X profiles, which is hard to achieve when the number of X's is large.
   - "X-profile" or "covariate profile" refers to specific values for all predictors
5. More assumptions to check later – regarding inter-related predictors
   - basically, if predictors are related to each other, the model becomes very hard to interpret
6. Other variable types can be included
   (interactions, qualitative, higher-order) – (more in Module 3)

3 Matrix Approach to Multiple Linear Regression

When the number of X variables gets large, the matrix representation of linear regression models is easier to write and understand.

$$
\begin{aligned}
Y &= (Y_1, \ldots, Y_n)' = \text{vector of response variable} \\
\varepsilon &= (\varepsilon_1, \ldots, \varepsilon_n)' = \text{vector of error terms} \\
X_k &= (X_{1k}, \ldots, X_{nk})' = \text{vector of predictor variable \#k} \quad (k = 1, \ldots, p-1) \\
X &= [\ 1 \quad X_1 \quad \ldots \quad X_{p-1}\ ] = \text{matrix with } p \text{ columns and } n \text{ rows} \\
\beta &= (\beta_0, \beta_1, \ldots, \beta_{p-1})' = \text{vector of coefficients} \\
b &= (b_0, b_1, \ldots, b_{p-1})' = \text{vector of coefficient } \underline{\text{estimates}}
\end{aligned}
$$

Then regression model is

$$
\begin{aligned}
Y &= X\beta + \varepsilon \\
\varepsilon &\sim N(0, \sigma^2 I) \qquad I = \text{"identity" matrix}
\end{aligned}
$$

Estimates:

$$
b = (X'X)^{-1}X'Y
$$

Matrices with variance on diag., covariance on off-diag.

$$
\begin{aligned}
\text{truth:} \quad Cov(b) &= (X'X)^{-1}\sigma^2 \\
\text{estimated:} \quad s^2\{b\} &= (X'X)^{-1} \cdot \text{MSE}
\end{aligned}
$$

$\sqrt{\text{diag. elements}}$ gives SE's of b_k's

We'll come back to this, but for now, note that

$$
\begin{aligned}
\hat{Y} &= Xb \\
&= X(X'X)^{-1}X'Y \\
&= HY
\end{aligned}
$$

H projects Y down to column space of X:

- Y = observed response values vector; <u>is not</u> a [perfect] linear combination of predictor variables
- $\hat{Y}$ = predicted response values vector; <u>is</u> a [perfect] linear combination of predictor variables

2.6: Multiple Inference and Multicollinearity

Dr. Bean - Stat 5100

1 Why Multiple Inference?

We already have tools to test the significance of model coefficients:

- Individual coefficients: t-tests ($H_0 : \beta_k = 0$)
- *All* coefficients: model F-test ($H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$)

What if we want to consider the significance of a subset of the X predictor variables? (More than one, but not all of them).

2 Subset Testing

Example: Bodyfat Dataset (Handout 2.6.1)

Y = body, X_1 = triceps, X_2 = thigh, X_3 = midarm

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

Consider $H_0 : \beta_2 = \beta_3 = 0$.

How to test: See how much better the full model is (using tricep, thigh, and midarm) compared to the reduced one (using only triceps).

- Notation: $SSE(X_1, X_2, X_3) = SS_{error}$ when model has predictors X_1, X_2, and X_3
  – represents amount variation in Y left unexplained by the full model
- Assuming $H_0 : \beta_2 = \beta_3 = 0$ is true, fit "reduced" model (only predictor X_1) and calculate $SSE(X_1)$
- Note that $SSE(X_1) > SSE(X_1, X_2, X_3)$
  - ALWAYS true, as a "worthless" X variable won't ever increase the SSE, but may reduce it slightly by chance.
  - NOT true of validation error (more discussion in Module 4).

  – then define "extra sum of squares"

  $$SSR(X_2, X_3 | X_1) = SSE(X_1) - SSE(X_1, X_2, X_3)$$

  Note: this represents amount variation in Y accounted for by X_2 & X_3 when X_1 already in model
- Define

  $$MSR(X_2, X_3 | X_1) = \frac{SSR(X_2, X_3 | X_1)}{2}$$

  – think of this as the mean square reduction

- Build test statistic for $H_0: \beta_2 = \beta_3 = 0$

$$\begin{aligned} F^* &= \frac{MSR(X_2, X_3|X_1)}{MSE(X_1, X_2, X_3)} \\ &= \frac{SSR(X_2, X_3|X_1)/(2)}{SSE(X_1, X_2, X_3)/(16)} \end{aligned}$$

- When $H_0: \beta_2 = \beta_3 = 0$ is true, $F^* \sim F_{2,16}$

General test of any # of β_k's:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{p-1} X_{p-1} + \epsilon \\ H_0 &: \beta_q = \beta_{q+1} = \ldots = \beta_{p-1} = 0 \\ p &= \#\text{ of } \beta\text{'s in full model (incl. intercept)} \\ q &= \#\text{ of } \beta\text{'s in reduced model (incl. intercept)} \\ p - q &= \#\text{ of } \beta\text{'s being tested in } H_0 \\ F^* &= \frac{[(\text{SSE in reduced model}) - (\text{SSE in full model})]/(p-q)}{[\text{SSE in full model}]/(n-p)} \end{aligned}$$

Under H_0, $F^* \sim F_{p-q,n-p}$

Recall the t-statistic from test of individual predictor ($H_0: \beta_k = 0$)?

$$t^* = \frac{b_k}{s\{b_k\}}$$

– if only have one predictor in model then $(t^*)^2 \sim F_{1,n-p}$

SSR also called sequential sums of squares or Type I SS; example in SAS:

- $SSR(X_1) \approx 352.27$
- $SSR(X_2|X_1) \approx 33.17$
- $SSR(X_3|X_1, X_2) \approx 11.55$

Related concept: "Coefficients of Partial Determination"

- what proportion of [previously unexplained] variation in Y can be explained by addition of predictor X_k to model

$$R^2_{Y3|12} = \frac{SSR(X_3|X_1, X_2)}{SSE(X_1, X_2)}$$

  – $SSR(X_3|X_1, X_2)$ - reduction in SSE that occurs when X_3 is added to the model when X_1 and X_2 are already in the model.

- $SSE(X_1, X_2)$ - amount of unexplained variation in Y when X_1 and X_2 are in the model.

- example in SAS:
  - $R^2_{Y1} \approx 0.711$
  - $R^2_{Y2|1} \approx 0.232$
  - $R^2_{Y3|12} \approx 0.105$

3 Multicollinearity

Textbook sections 7.6 and 10.5

The model F test says that the coefficients *collectively* are highly significant, but *none* of the individual variables are significant.

This is a symptom of **multicollinearity** (i.e. collinearity):

- Two X variables share a strong linear relationship *with each other* (independent of Y)
- One X variable is a near linear combination of two or more X variables

Problems with Multicollinearity:

- β_k hard to interpret as it no longer makes sense to "hold all other predictor variables constant."
- The variance of b_k will be very large (inflated) as our estimates are starting to become non-unique $\rightarrow$ makes inference of β_k difficult if not impossible.
  - Could make estimate of b_k counter-intuitive (example: getting a negative estimate of b_k despite knowing that X and Y are positively correlated)).
- Contradictory results between individual t-tests and model F tests (or subset F tests).

NOT Problems with Multicollinearity:

- Multicollinearity does NOT affect a model's predictive ability.

3.1 Standardizing Variables

One way to better understand multicollinearity is by standardizing variables.

$$Y_i^* = \frac{1}{\sqrt{n-1}} \left(\frac{Y_i - \bar{Y}}{\text{SD of } Y} \right) \qquad , \qquad X_{ik}^* = \frac{1}{\sqrt{n-1}} \left(\frac{X_{ik} - \bar{X}_k}{\text{SD of } X_k} \right)$$

– sometimes called "correlation transformation" because

$$Corr(X_k, Y) \;=\; \sum_i X_{ik}^* Y_i^*$$

If all variables have been standardized, then consider matrix approach
(with no Intercept column in matrix X^*):

$$\begin{aligned} Y^* &= X^*\beta^* + \varepsilon \\ b^* &= (X^{*\prime}X^*)^{-1}X^{*\prime}Y^* \\ Cov(b^*) &= (X^{*\prime}X^*)^{-1}\sigma^2 \end{aligned}$$

There is no intercept column because, by construction, the intercept will be Y=0 as all points must past through $(\bar{X}, \bar{Y}) = (0, 0)$

To un-standardize regression coefficient estimates:

$$\begin{aligned} b_k &= \left(\frac{\text{SD of } Y}{\text{SD of } X_k}\right) \cdot b_k^* \\ b_0 &= \bar{Y} - \sum_{k=1}^{p-1} b_k \bar{X}_k \end{aligned}$$

Relevance to multicollinearity:

- the correlation matrix among the [original] predictor variables is $X^{*\prime}X^*$
- the "closer" X_j and X_h are, the larger will be the j^{th} and h^{th} diagonal elements of $Cov(b^*)$, so the estimated variance is higher for b_j and b_h
- We can use the correlation matrix to obtain a set of **condition indices** as obtained from the **eigenvalues** of the matrix.

While standardizing helps to better mathematically understand the effect of multicollinearity, it is not necessary to standardize to detect multicollinearity.

3.2 Ways to Diagnose Multicollinearity

3.2.1 Condition Index/Principal Components

- Recall from linear algebra: λ is an **eigenvalue** of a symmetric, square matrix A iff there exists a vector x (the **eigenvector** for λ) such that $Ax = \lambda x$.
- Let $\lambda_1, \ldots, \lambda_k$ be the eigenvalues of $X^{*\prime}X^*$, and let

$$\text{Condition Index}_i = \left(\frac{\lambda_{max}}{\lambda_i}\right)^{1/2}$$

- Each condition index is associated with a **principal component**
  - Each principal component is a linear combination of the original predictor variables. Each principal component shares no correlation with any other principal component (i.e. $cor(PC_1, PC_2) = 0$).

$$
\begin{aligned}
PC_1 &= a_1 X_1^* + \ldots + a_{p-1} X_{p-1}^* \\
PC_2 &= c_1 X_1^* + \ldots + c_{p-1} X_{p-1}^* \\
&\vdots
\end{aligned}
$$

- Each principal component explains some percentage of the variation in the original predictors.

IF the condition index is high (more than 10 or so) **AND** the associated principal component explains a high proportion of the variance (usually more than 50% variability) *in the beta coefficients associated with* two or more predictor variables, then we have potentially problematic multicollinearity.

3.2.2 Variance Inflation Factor (VIF)

- Let R_k^2 be the coefficient of multiple determination (the R^2 value) when predictor X_k^* is regressed on the other predictors
  - This is a measure of how much of the variance of X_k^* is explained by the other X variables.
- Define $VIF_k = (1 - R_k^2)^{-1}$, for $k = 1, \ldots, p-1$ as the "Variance Inflation Factor" for b_k (the estimate of β_k)

IF the largest VIF is much more than 10 **OR** the average VIF is much more than 1, then we have evidence of potentially problematic multicollinearity.
We usually use a combination of the VIF and condition index to asses multicollinearity.

3.2.3 Important things to remember about standardization

- Relative magnitude of b_k^* estimates not meaningful if predictors are on different scales
- Standardization most common when predictors $X_1, \ldots, X_{p-1}$ have <u>very</u> different scales
- β_k^* is expected change in Y for every <u>SD</u> (not unit) increase in predictor X_k, while all other predictors are held constant
- Standardizing has:
  - no effect on VIF
  - marginal effect on proportions of variance in Condition Index output
  - possibly substantial effect on magnitude of Condition Indexes
- Recommendations:
  - Standardize if either:
    * desire common scale of b_k^* estimates
    * <u>need</u> uncorrelated, higher-order predictors

3.3 Multicollinearity Summary

Three ways to diagnose multicollinearity:

1.
   combination of condition index and proportion of variation

2.
   variance inflation factors

3.
   model F-test vs. individual t-tests

Possible remedial measures for multicollinearity:

- Collect more data
- Choose a subset of predictor variables
- Ridge regression
- Latent root regression – use Principal Components as predictors (may lack interpretability)

$$PC = a_1X_1 + a_2X_2 + \ldots + a_{p-1}X_{p-1}$$