STA302 notes

June 1, 2022

STA302F in Summer 2022 with Mohammad Khan. Feel free to email **anton.sugolov@mail.utoronto.ca** if there are any mistakes, or edit the tex **here**.

Contents

May 9: Lecture 1

Syllabus

This is a course on linear regression. The focus is using R to do data analysis, and build the mathematical foundation for regression. We will understand how prediction works later, which is the foundation for data science.

Marking

- 2 HW - 15% each, due June 1, June 15
- Test - 25% on May 25
- Exam - 45% during June 22-27

Books J. Sheather, A Modern Approach to Regression w/ R and D. Montgomery, Linear Regression Analysis.

Review

Definition 1. A **sample space** S is the set of possible events. A **random variable** is a function $X : S \to \mathbb{R}$ assigning a number to elements of the sample space.

Constants can also be pseudo random variables. These are called **degenerate random variables** that have a **degenerate distribution** since they have infinite cdf.

Definition 2. For an event $A \subset S$, we define the **indicator function** I_A as

$$I_A(s) = \begin{cases} 1, & s \in A \\ 0, & s \notin A \end{cases}$$

These are important since we later use them to create dummy variables in linear regression. When we write an inequality involving random variables, we mean that it holds for all elements of the sample space. I.e. $X \geq Y \implies X(s) = Y(s), \forall s \in S$.

Example 1. Consider $S = \{1,2,3,4,5,6\}$. For $s \in S$, $X(s) = s$, let $Y(s) = X(s) + I_6(s)$. Then $Y = X$ for all $s \in S$ except 6, where $Y = 7$, $X = 6$.

Definition 3. **Discrete r.v.** are functions from a countable sample space, and **continuous r.v.** are functions from an uncountable sample space. There are also **mixture** random variables, which are continuous/discrete for different parts of the sample space. Random variables can be univariate and multivariate as well.

Example 2. The multinomial distribution is an example of a discrete multivariate random variable.

Definition 4. If X is a random variable, the p.d.f. is the derivative of the c.m.f. As well, $\mathscr{P}(a \leq X \leq b) = \int_a^b f(x)dx$ where $f(x)$ is pdf. Similar thing holds for discrete r.v.

Proposition 1. The expectation of two random variables is linear. For $Z = aX + bY$, X, Y r.v., then $E(Z) = aE(X) + bE(Y)$.

Definition 5. The **variance** of X is $V(X) = E(X - \mu_x)^2$. The **sample variance** $s^2 = \frac{\sum(x_i - \overline{x})^2}{n-1}$. Note we divide by $n-1$ so that it is an unbiased estimator (STA261).

Some properties:

- $V(X) \geq 0$
- $V(aX+b) = a^2 V(x)$
- $V(X) = E(X^2) - E(X)^2$
- $V(X) \leq E(X^2)$
- $\sigma_X = \sqrt{V(X)}$

Note: In linear regression, the variance of the predicted variable depends on the slope of regression line but not on the intercept (second property).

Let X_1, X_2, Y be r.v. and A be an event. Let $Z = aX_1 + bX_2$. Then

- $E(Z \mid A) = aE(X_1 \mid A) + bE(X_2 \mid A)$
- $E(Z \mid Y = y) = aE(X_1 \mid Y = y) + bE(X_2 \mid Y = y)$
- $E(Z \mid Y) = aE(X_1 \mid Y) + bE(X_2 \mid Y)$

Proposition 2. (Laws of Total Expecation and Variance) $E(E(Y \mid X)) = E(Y)$ and $V(X) = V(E(X \mid Y)) + E(V(X \mid Y))$.

We will see that linear regression is a conditional r.v., and the above will be very useful. For $X_1, \ldots, X_n$ i.i.d. random variables, $x_1 \ldots x_n$ realizations, then $\overline{x} = \frac{\sum x_i}{n}$. The **sample average** $\overline{X} = \frac{\sum X_i}{n}$ is a random variable. In general, any function of n i.i.d. random variables is a random variable, and called a **sampling statistic** that follows a **sampling distribution**.

Theorem 1. (Central Limit Theorem) For $X_1, \ldots, X_n$ i.i.d. $f(x, \theta)$, $E(X)$, $V(x) < \infty$, then $\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \to N(0,1)$ converges in distribution for sufficiently large n.

Proof. Proof with moment generating functions. □

Example 3. In the Cauchy distribution, this does not hold since it has infinite mean and variance.

Definition 6. The **covariance** $\text{Cov}(X,Y) = E[(X - \mu_x)(Y - \mu_Y)] = E(XY) - E(X)E(Y)$. Covariance quantifies the relationship between two variables, i.e. how much one varies with the other. The **correlation** $\text{Corr}(X,Y) = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}$.

- Covariance is an inner product, variance is norm.
- $V(X+Y) = V(X) + V(Y) + 2\text{Cov}(X,Y)$.
- If $X \perp Y$, $V(X+Y) = V(X) + V(Y)$.
- In general, $V(\sum_i X_i) = \sum_i V(X_i) + 2\sum_{i<j} \text{Cov}(X_i, X_j)$.

These will be useful in regression, where we try to identify relationships between r.v.s.

Definitions in statistics

In probability, we are given a mathematical model to work with. In statistics, we infer properties of a mathematical model. The steps of data analysis are: state the problem, identify what data is needed, decide on a model and collect data, clean data, estimate parameters of the model, and carry out appropriate tests, draw conclusions.

Introduction to Regression

Definition 7. The **corelation coefficient**

$$\rho_{X,Y} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2}\sqrt{\sum_i (y_i - \overline{y})^2}} = \frac{\text{Cov}(X, Y)}{s_x s_y}$$

The above value is somewhat like the $\cos(\theta)$ between the vectors X, Y; recall dot product. When we discuss corelation, we talk about linear relations only; the linear association between X, Y. We can see this by considering X and $Y = X^2$. Corelation is symmetric, it does not indicate the direction of the symmetry (which causes which/causation). Corelation only says the influence on the change of one variable when the other changes; think about moving along non-orthogonal vectors and projecting.

Galton investigated the effect of fathers heights on their sons height. Galton termed **regression** as a 'regression' of heights towards the mean; on average, heights of sons move towards the mean, so the average height across generations is the same.

In a linear regression, we assume there is a linear relation $Y = \beta_0 + \beta_1 X + \epsilon$ between the random variables X, Y where ϵ is an error random variable. The deviation not captured by linearity is incorporated to ϵ. Given two values of X, it is not guaranteed that the value of Y is the same. But for a unique X we get **unique average** Y. We want $E(Y \mid X = x) = \beta_0 + \beta_1 X$; the relationship between the mean of Y and a specific value of X is linear. Note $E(\epsilon) = 0$. We call X the **explanatory, predictor, independent** variable and Y as the **response, outcome, dependent** variable. Suppose we are given paired data $(x_1, y_1), \ldots, (x_n, y_n)$. We try to fit a linear regression to model the relationship between X and Y:

$$Y = \beta_0 + \beta_1 X + \epsilon \text{ and want } E(Y \mid X = x) = \beta_0 + \beta_1 X$$

The values of β_0, β_1 are not yet known and need to be estimated. In the sample, the error e_i replaces ϵ_i. The line best predicting Y as X changes should minimize the squares of the errors $e_i = y_i - \hat{y}_i$ where $\hat{y}_i = b_0 + b_1 x_i$ where b_0, b_1 are the intercept and slope of the regression line. We minimize the squares $\sum_i e_i^2$. The e_i are referred to as **residuals**; minimize residual sums squared. Note

$$RSS(b_0, b_1) = \sum_i e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Aside: What value of a minimizes (1) $\sum |x_i - a|$, and which minimizes (2) $\sum (x_i - a)^2$? Answer: (1) $a = \text{Med}(X)$, (2) $a = \overline{x}$. We do not minimize the sum of the residuals, since this must always be 0. We minimize the RSS with respect to b_0, b_1.

$$\frac{\partial RSS}{\partial b_0} = -2\sum_i (y_i - b_0 - b_1 x_i), \; \frac{\partial RSS}{\partial b_1} = -2\sum_i x_i(y_i - b_0 - b_1 x_i)$$

so setting these to 0, we get the **normal equations**

$$\sum_i y_i = b_0 n + b_1 \sum_i x_i, \quad \sum_i x_i y_i = b_0 \sum_i x_i + b_1 \sum_i x_i^2$$

Solving these, we get

$$\hat{\beta}_0 = b_0 = \overline{y} - \hat{\beta}_1 \overline{x}, \qquad \hat{\beta}_1 = b_1 = \frac{\sum_i x_i y_i - n\overline{x}\,\overline{y}}{\sum x_i^2 - n\overline{x}^2} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \frac{S_{X,Y}}{S_X}$$

The intercept is the average value of the response when $X = 0$.

Class Afterthoughts/Questions

When the errors have $E(\epsilon = 0)$, then $V(\epsilon) = E(\epsilon^2) - E(\epsilon)^2 = E(\epsilon^2)$. By minimizing this in the sample, we minimize the variance of the errors (?)

May 11: Lecture 2

Clarifying last class: $\hat{y}_i$ is the conditional mean of y_i. When this is true, then $\sum_i e_i = 0$. That is, we estimate $\hat{y}_i$ so that $\sum_i e_i = 0$.

Regression continued

We continue discussing linear regression; fitting a linear relation assuming it exists. The aim is to infer the true values of β_0, β_1 by inspecting their sampling distributions. We also make some assumptions regarding the error terms; the properties of their distributions (ϵ is r.v.).

Assumption: Linearity

The conditional mean of $Y \mid X = x$ is linear with respect to X. However, the relationship $E(Y \mid X)$ and X does not have to be linear, but the linearity assumption is linearity in the parameters.Our relationship must be realistic given the context; introducing linearity may produce unrealistic relationships.

R simulation: When generating random dataset, we set a seed so our results are reproducible. Always start with a seed in assignments. Note the Y variable is the transformation $\beta_0 + \beta_1 \log X + \epsilon$. Introducing linear relationship between X and Y is inaccurate. It is linear in the parameters β_0, β_1 however.

Qs: Chaos in random number generation? Look up random number generation algorithms. How do we quantify linearity in a data set? Mostly with plots but is there better way?

Assumption: Independence

The errors ϵ_i are independent. That is, the deviations from the mean are not related; they are i.i.d. r.v. This reduces predictive capibilities in some areas, but we can relax this assumption later (generalized least squares).

Assumption: Homoscedasticity (equal variance)

The error variance does not change depending on X. That is $V(\epsilon \mid X = x) = \sigma^2$ and is independent of x. In the R codes, we see that variance of errors increases with X, which decreases predictive power as X increases. Moreover, this implies some of the variation in the errors is explained by X, which violates our assumption. Variance **cannot** depend of X. $\epsilon \perp X$. This is relaxed in GLS.

In multiple linear regression, we talk about the Gauss-Markov assumption, but we need to make some assumptions about how ϵ_i is distributed in order to make inferences.

Assumption: Normality

$\epsilon \sim N(0, \sigma^2)$. The previous assumptions are required to obtain the least squares estimates, but normality is not required. Under this assumption, we can make confidence intervals and tests, and have nice properties following from normal distribution.

There are more assumpitons in general, but these are most important.

More about variance of ϵ

We have estimated β_0, β_1 using least squares. However, we have another parameter to estimate; $V(\epsilon) = \sigma^2$. From afterthoughts, $V(\epsilon) = E(\epsilon^2) = \sigma^2$. We take the average of e_i^2 using this, since we want summary measure. The mean residual squared (MRS) can be calculated as $s^2 = \dfrac{\sum_i e_i^2}{n-2}$. We show this estimator of $E(\epsilon^2)$ is unbiased as homework; prove this!.

Inferences about the regression model

Conditional expectation and variance of $\hat{\beta}_1$

Recall $\beta_1 = \dfrac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}$

Proposition 3. $\sum(x_i - \overline{x})(y_i - \overline{y}) = \sum_i (x_i - \overline{x}) y_i$

Proof.

$$\begin{aligned}
\sum(x_i - \overline{x})(y_i - \overline{y}) &= \sum(x_i y_i - \overline{x} y_i - \overline{y} x_i + \overline{x}\,\overline{y}) \\
&= \sum(x_i y_i - \overline{x} y_i) - n\overline{y}\,\overline{x} + n\overline{x}\,\overline{y} \\
&= \sum(x_i - \overline{x}) y_i
\end{aligned}$$

$\square$

A symmetric sum can be established for $\sum_i (y_i - \overline{y}) x_i$. However, the above is needed to simplify conditional expectation calculations. We may also show $\sum(x_i - \overline{x}) x_i = \sum(x_i - \overline{x})^2$. The idea of both of these proof is making the substitution $n\overline{x} = \sum x_i$.

Proposition 4. $\sum(x_i - \overline{x}) x_i = \sum(x_i - \overline{x})^2$

Proof.

$$\begin{aligned}\sum(x_i-\overline{x})x_i &= \sum(x_i^2-\overline{x}\,x_i)\\ &= \sum(x_i^2-2\overline{x}\,x_i)+n\overline{x}^2\\ &= \sum(x_i-\overline{x})^2\end{aligned}$$

□

Other way of writing: $\sum(x_i\overline{x})^2=\sum x_i^2-n\overline{x}^2$. Now, we calculate **conditional expectation of $\hat{\beta}_1$**

$$E(\hat{\beta}_1\mid X=x_i)=E\left(\frac{\sum(x_i-\overline{x})y_i}{\sum(x_i-\overline{x}^2)}\mid X=x_i\right)=\frac{\sum(x_i-\overline{x})E(Y_i\mid X=x_i)}{\sum(x_i-\overline{x})^2}$$

Substituting $E(Y_i\mid X_i=x)=\beta_0+\beta_1x$, then

$$E(\hat{\beta}_1\mid X=x_i)=\frac{\sum_i(x_i-\overline{x})\beta_0}{\sum(x_i-\overline{x})^2}+\frac{\sum_i(x_i-\overline{x})\beta_1x_i}{\sum(x_i-\overline{x})^2}=\frac{\beta_1\sum_i(x_i-\overline{x})^2}{\sum(x_i-\overline{x})^2}=\beta_1$$

Since $\sum(x_i-\overline{x})=\sum x_i-n\overline{x}=0$ and by above prop., $\sum_i(x_i-\overline{x})x_i=\sum(x_i-\overline{x})^2$. Therefore $\hat{\beta}_1$ does not depend on X, and has expected value of β_1; it is an unbiased estimator of β_1. That is, $E(\hat{\beta}_1\mid X=x_i)=E(\hat{\beta}_1)=\beta_1$. Next, we may calculate $V(\hat{\beta}_1)$. First, $V(Y_i\mid X=x_i)=\sigma^2$, that is, the variance of the error.

$$V(\hat{\beta}_1\mid X=x_i)=\left(\frac{\sum(x_i-\overline{x})y_i}{\sum(x_i-\overline{x}^2)}\mid X=x_i\right)=\frac{\sum_i(x_i-\overline{x})^2V(Y_i\mid X=x_i)}{(\sum_i(x_i-\overline{x})^2)^2}=\frac{\sigma^2}{\sum(x_i-\overline{x})^2}=\frac{\sigma^2}{S_{X,X}}$$

Inferences for variance of $\hat{\beta}_1$

Since $\epsilon_i\sim N(0,\sigma^2)$, then $Y_i\mid X\sim N(\beta_0+\beta_1X,\sigma^2)$. Letting $c_i=\frac{\sum(x_i-\overline{x})}{\sum(x_i-\overline{x})^2}$ then $\hat{\beta}_1=\sum c_iy_i$. Observe that this is a **linear combination** of normally distributed random variables, so $\hat{\beta}_1$ is normally distributed! Thus

$$\hat{\beta}_1\mid X=x_i\sim N\left(\beta_1,\frac{\sigma^2}{S_{X,X}}\right)$$

We can construct a $1-\alpha$ confidence interval for β_1 which has extremes $\hat{\beta}_1\pm Z_{1-\alpha/2}\dfrac{\sigma}{\sqrt{S_{X,X}}}$. When σ^2 is unknown, we construct a t-confidence using $S^2=\dfrac{\sum e_i^2}{n-2}$. We therefore make a confidence interval with critical values

$$\hat{\beta}_1\pm t_{1-\alpha/2,n-2}\frac{s^2}{\sqrt{S_{X,X}}}$$

Note our assumption of normality of errors.

Clarification $S_{X,X}=\sum(x_i-\overline{x})^2$ and $S_{X,Y}=\sum(x_i-\overline{x})(y_i-\overline{y})$.

Recall, the **p-value** can be calculated as $p=\mathscr{P}(Z\geq|z|)$ or $p=\mathscr{P}(T\geq|t|)$ where z,t are the calculated test statistics. The p-value is the probability of obtaining a sample that provides strong evidence against the hypothesized value of $H_0:\beta_1$, set by threshold α. α is the probability of making a type one error with repeated sampling.

Example 4. $\sum x_i = 4035$, $\sum y_i = 4041$, $\sum e_i^2 = 4753.125$, $\sum x_i^2 = 1005535$, $\sum x_i y_i = 864910$, $t_{0.975,18} = 2.10$.
We need to calculate $\hat{\beta}_1, s, S_{X,X}$ from this information; recall $\hat{\beta}_1 \pm t_{1-\alpha/2,n-2}\frac{s}{\sqrt{S_{X,X}}}$. The interval becomes $(0.18121, 0.33728)$. **Verify as homework.**

Do exercises from Montgomery (unassigned, do by chapter) and Sheather. Problems are similar to this, and this will appear on the midterm.

Properties of β_0

The conditional expectation of $\beta_0 \mid X$. Since $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$. Using this,

$$E(\hat{\beta}_0 \mid X = x_i) = \frac{\sum E(y_i \mid X = x_i)}{n} - \beta_1 \overline{x} = \left(\frac{n\beta_0 + n\beta_1 \overline{x}}{n}\right) - \beta_1 \overline{x} = \beta_0$$

Therefore $\hat{\beta}_0$ is an unbiased estimator of β_0. Now for the variance, (minor abuse of notation)

$$V(\hat{\beta}_0 \mid X = x_i) = V(oly - \hat{\beta}_1 \overline{x} \mid X = x_i) = V(\overline{y} \mid x_i) + \overline{x}^2 V(\hat{\beta}_1 \mid x_i) - 2\overline{x}\text{Cov}(\overline{y}, \hat{\beta}_1 \mid x_i)$$

Calculating each term separately,

$$V(\overline{y} \mid X = x_i) = V\left(\frac{\sum y_i}{n} \mid X = x_i\right) = \frac{\sum \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

To calculate covariance term, we use substitutions involving $\hat{\beta}_1 = \sum c_i y_i$ with c_i defined before

$$\text{Cov}(\overline{y}, \hat{\beta}_1 \mid X = x_i) = \text{Cov}\left(\frac{\sum_i y_i}{n}, \sum c_i y_i \mid X = x_i\right) = \frac{1}{n}\sum_i \text{Cov}(y_i, c_i y_i \mid X = x_i)$$

Recall $\text{Cov}(X, aY) = a\text{Cov}(X, Y)$. Also, given a particular x_i, c_i is a constant.

$$= \frac{1}{n}\sum_i c_i \text{Cov}(y_i, y_i \mid X = x_i) = \frac{1}{n}\sum_i c_i V(y_i \mid X = x_i) = \frac{1}{n}\sum_i c_i \sigma^2 = 0$$

From last section, $V(\hat{\beta}_1 \mid x_i) = \overline{x}^2 \frac{\sigma^2}{S_{X,X}}$. Therefore

$$V(\hat{\beta}_0 \mid X = x_i) = \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{X,X}}\right), \text{ and } \hat{\beta}_0 \mid X = x_i \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{X,X}}\right)\right)$$

Therefore the $(1-\alpha)$ confidence for β_0 is

$$\hat{\beta}_0 \pm Z_{1-\alpha/2}\sigma\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{X,X}}}$$

(fill in when σ^2 is unknown)

Confidence interval for the regression line

Denote x^*, y^* as an observation not currently in the sample. We use the model built with the current observations to see how far y^* observation can vary. It can easily be shown that

$$E(\hat{y}^* \mid X = x^*) = \beta_0 + \beta_1 x^*$$

Where $X = x^*$ new observation, y^* unknown. As well, $\hat{y}^*$ is the predicted value of y^* paired with x^*. Often, we are interested in calculating the variance of $E(Y \mid X = x^*) = \hat{y}^* \mid X = x^*$ and confidence interval for $E(Y \mid X = x^*)$. That is, calculating the variance and confidence of the regression line at each point. Note $E(\hat{y}^* \mid X = x^*) = \beta_0 + \beta_1 x^* = E(Y \mid X = x^*)$ implies the sample regression is an unbiased estimator of the true Linear relationship between X, Y. The variance can be calculated as

$$\begin{aligned} V(\hat{y}^* \mid X = x^*) &= V(\hat{\beta}_0 + \hat{\beta}_1 x^*) = V(-\overline{y} + \hat{\beta}_1(x^* - \overline{x})) \\ &= V(\overline{y}) + (x^* - \overline{x})^2 V(\hat{\beta}_1) = \frac{\sigma^2}{n} + \frac{\sigma^2 (x^* - \overline{x})^2}{S_{X,X}} \\ &= \sigma^2 \left(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{X,X}} \right) \end{aligned}$$

This is interpreted as the variance of the true location of the regression line at $X = x^*$. Note variance increases quadratically as x^* moves further from $\overline{x}$.

Prediction error and interval

Assuming we fit a regression line between X, Y with some sample. If a new data point $X = x^*$ is given, our predicted $\hat{y}^*$ lies exactly on the line in the model we have fitted, but y^* associated with x^* may deviate from the line. How much does this y vary? $y^* - \hat{y}^*$ is called the **prediction error** for $X = x^*$. We calculate its expectation and variance.

For expectation, the $*$ is redundant, so we write $E(y - \hat{y} \mid X = x^*)$. We can easily show this is 0 since $y - \hat{y} = 0$. Therefore

$$V(y^* - \hat{y}^* \mid X = x^*) = V(y - \hat{y} \mid X = x^*) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{X,X}} \right)$$

We just add the variance of y and variance of $\hat{y}$ by expansion of variance and since $\text{Cov}(\hat{y}, y) = 0$. The observation y is independent of the previous sample by assumption. The prediction interval is built in the same way as before using t distribution. The prediction interval is how much we expect the true value to deviate from the regression line.

R simulation:

The confidence interval is for the regression line. The prediction interval is for a new predicted value given x^*; how far y^* can deviate from the predicted $\hat{y}^*$.

Example 5. Calculate summary measures for the production data (in slides hw)

May 16: Lecture 3

Clarification In the derivations from last class, we used

$$\text{Cov}\left(\frac{\sum Y_i}{n}, \sum c_i y_i \mid X = x_i \right) = \frac{1}{n} \sum \text{Cov}(y_i, c_i y_i \mid X = x_i)$$

since $\text{Cov}(Y_i, Y_j) = 0$ by independence of Y_i, Y_j.

Understand theory and problem solving procedure for midterms. Data analysis will mostly be with R.

Assignment Task 1

The purpose of the assignment is using R for inference of parameters given simulated data. Use your student id as a seed. After data is generated, run the LM model. Repeating this procedure, get sampling distribution for $\hat{\beta}_i, \sigma^2$, and compare these to true variances.

Analysis of variance (ANOVA)

So far we have discussed inference about specific parameters, and hypothesis testing for their true values. For example, if we fail to reject $H_0 : \beta_1 = 0$, then there is no linear relationship between X, Y. In this case, $Y = \beta_0 + \epsilon$, $V(Y) = V(\epsilon) = \sigma^2$, so ϵ explains all the variance of Y. Usually, $V(Y) = \beta_1^2 V(X) + \sigma^2$, since $X \perp \epsilon$. Therefore when the above holds, part of the variance is given by $V(X)$. If most of the variation in Y is explained by X, then predictions are very accurate. We discuss this in ANOVA.

In the slides, points that are less scattered about the regression line have more of their variance explained by X.

As the residual variance σ^2 increases, the variation of Y is less explained by X. This increases prediction error. We want to answer how well the regression line might explain the variation we observe in the responses. ANOVA is another way of testing the significance of the regression line. The total varation of Y is explained by the **total sum of squares**, the numerator of s_Y

$$SST = \sum (y_i - \overline{y})^2$$

This can be decomposed by

$$\sum (y_i - \overline{y})^2 = \sum (y_i - \hat{y}_i + \hat{y}_i - \overline{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \overline{y})^2 + 2\sum (y_i - \hat{y}_i)(\hat{y}_i - \overline{y})$$

Where the third term becomes

$$\sum (y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = \sum (\hat{y}_i(y_i - \hat{y}_i) - \overline{y}(y_i - \hat{y}_i)) = \sum \hat{y}_i e_i - \overline{y} \sum e_i = 0$$

Since $\sum e_i = 0$ and $\sum x_i e_i = 0$ by the second normal equation, which gives $\sum \hat{y}_i e_i = 0$. Hint: $\sum (\beta_0 + \beta_1 x_i) e_i = \beta_0 \sum e_i + \beta_1 \sum x_i e_i$. Therefore the total variation of Y can be divided into

$$\sum (y_i - \overline{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \overline{y})^2$$

The term on the left is the **residual sum square**, $(n-2)s^2$. The second term explains the variance in $\hat{y}_i$, or the variation in fitted values from the regression. We may easily show $\sum \frac{\hat{y}_i}{n} = \overline{y}$. The second term on the right is the **regression sum squared**. The total variation in Y has been decomposed to come from the regression line, and from random errors.

Degrees of Freedom. This is the number of summed square normals. The proof for $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$ shows where one of the 'standard normal squares' are lost. (s^2 is sample variance). For each parameter we fix, we lose a degree of freedom. When $\overline{y}$ is fixed, we are free to have $n-1$ values, and are forced to choose one to get the fixed $\overline{y}$. That is, y_n, the n-th observation is fixed for a fixed $\overline{y}$. This is why sample variance, $\sum(y_i - \overline{y})^2/n-1$, uses $n-1$ degrees of freedom.

In the above SST, the **RSS** $\sum(y_i - \hat{y}_i)^2$ has $n-2$ degrees of freedom since $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ uses two estimated parameters. Since $\sum(y_i - \overline{y})^2$ has $n-1$ degrees of freedom, then the SS_{reg} $\sum(\hat{y}_i - \overline{y})^2$ must have 1 degree of freedom. This follows since the sum depends only on β_1 given fixed x_i:

$$\sum(\hat{y}_i - \overline{y})^2 = \sum(\hat{\beta}_0 + \hat{\beta}_1 x_i - \overline{y}^2) = \sum(\overline{y} - \hat{\beta}_1 \overline{x} + \hat{\beta}_1 x_i - \overline{y}^2) = \sum \hat{\beta}_1^2 (x_i - \overline{x})^2$$

We need degrees of freedom in order to test hypothesis. We will later show

$$\frac{SS_{reg}}{\sigma^2} \sim \chi^2_1, \frac{RSS}{\sigma^2} \sim \chi^2_{n-2}$$

Under $H_0 : \beta_0 = 0$ then $F_0 \sim F_{1,n-2}$. We want SS_{reg} as close to the SST as possible. The F-test here detects how close SS_{reg} is to TSS. The closer it is the bigger the value of F_0. We can show $t^2_{n-2} = F_{1,n-2}$. We can also show

$$E(SS_{reg}) = \sigma^2 + S_{X,X}\beta_1^2$$

So when $\beta_1 = 0$, the regression sum squared have variance equal to σ^2. Below is an ANOVA table:

Sources of Variation	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regression	1	SS_{reg}	$MS_{reg} = \frac{SS_{reg}}{1}$	$F_0 = \frac{MS_{reg}}{MRSS}$	etc
Residuals	n-2	RSS	$MRSS_{reg} = \frac{RSS}{n-2}$		
Total	n-1	SST			

In general, the F-test measures whether the means of two groups measure significantly. The F statistic is the ratio of explained variance (regression model attributes to $V(X)$) to unexplained variance (variance of e_i). Under the null, our data reflects the intercept only model $Y = \beta_0 + \epsilon$, and we test the departure from this.

The Coefficient of Determination

Another measure to assess whether the regression line explains enough of the variability in the response is the **coefficient of determination,** R^2. This gives the proportion of the total sample variability in the response that has been explained by the regression model.

$$R^2 = \frac{SS_{reg}}{SST} \text{ or } 1 - R^2 = \frac{RSS}{SST}$$

Note $0 \le R^2 \le 1$. If R^2 is close to 1, it is an important predictor of Y. If it is close to 0, then it offers little predictive power for Y. In simple linear regression, $\rho^2 = R^2$ where ρ is Pearson corelation coefficient.

Categorical predictors

So far we have required X to be continuous. However, X could be categorical. (X smoking status vs. Y blood pressure). Here the predictor is binary and the output is continuous. How would we test if the mean blood pressure varies between these groups?

We did this in STA261 with a two-sample t-test, and by homoescadicity we do one with equal variance. We may also use regression, by using **dummy variables** which are indicator variables. Setting 0 for non-smokers, 1 for smokers,

$$E(Y \mid X=0)=\beta_0,\ E(Y \mid X=1)=\beta_0+\beta_1$$

Using ANOVA this is essentially a t-test. $F_{1,n-2} \sim t^2_{n-1}$ so by squaring the t statistic we get F statistic; a significant F statistic indicate the change in means given by β_1 is significant. Therefore using hypothesis test with ANOVA for $\beta_1=0$, we get a test for differing means.

The 'slope' becomes the change in average. We can say β_1 reflects the average difference between two groups. The slope provides the magnitude of the difference, while the hypothesis test tells us whether the difference is statistically significant.

With categorical variables, R^2 may be low but the test will give significance.

Multiple Linear Regression

So far we have only had one predictor X, but we generalize to $X_1,\ldots,X_n$. That is

$$Y=\beta_0+\beta_1X_1+\ldots+\beta_pX_p+\epsilon$$

This implies Y is related to $X_1,\ldots X_p$ linearly. However, the predictor produces a p-dimensional subspace instead of a line. See image in 'Elements of Statistical Learning 2e'; with Y regressed on X_1,X_2 we get a regression plane.

The conditional mean of Y is given by $E(Y \mid X_1,\ldots,X_p)=\beta_0+\beta_1X_1+\ldots+\beta_pX_p$. For the sample dataset,

$$y_i=\beta_0+\beta_1x_{i,1}+\ldots+\beta_px_{p,1}+e_i$$

So we minimize $RSS(\beta_0,\ldots,\beta_p)=\sum(y_i-\sum^p\beta_jx_{ij})^2$. Differentiating with respect to each β_j,

$$\frac{\partial RSS}{\partial \beta_0}=\sum -2(y_i-\sum^p\beta_jx_{ij}) \qquad \frac{\partial RSS}{\partial \beta_j}\sum -2(y_i-\sum^p\beta_jx_{ij})x_{ij}$$

Setting these to 0, we get $p+1$ normal equations in $p+1$ unknowns, giving us a unique solution and therefore minimum, since it is the minimum for each β_j.

Matrix Notation

In order to simplify notation we use matrices. For this we write

$$\mathbf{Y}=\mathbf{X}\beta+\epsilon$$

$\mathbf{Y}$ is an $n\times 1$ vector, $\mathbf{X}$ is an $n\times(p+1)$ matrix, with the first column being a vector of 1s. β is $(p+1)\times 1$ vector, ϵ is $n\times 1$ vector.

We denote the transpose of matrix $\mathbf{A}$ as $\mathbf{A}'$. If $\mathbf{A}$ is a square matrix with $\mathbf{A} = \mathbf{A}'$ then it is symmetric (corrseponds to self adjoint operator). If $\mathbf{A}$ is invertible, we denote its inverse with $\mathbf{A}^{-1}$. A matrix is **orthogonal** if $\mathbf{A}^{-1} = \mathbf{A}'$; column vectors are orthogonal. An **idempotent** matrix satisfies $\mathbf{A}^2 = \mathbf{A}$. Some important properties are that

$$(\mathbf{A}+\mathbf{B})' = \mathbf{A}' + \mathbf{B}' \text{ and } (\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$$

Example 6. The projection matrix $P : \mathbb{R}^n \to \mathbb{R}^n$ of rank $p \leq n$ onto a subspace is a square matrix that is symmetric and idempotent.

May 18: Lecture 4

More properties

Definition 8. If $Y = (Y_1, \ldots, Y_n)$ is a random vector, then $E(Y) = (E(Y_1), \ldots, E(Y_n))$. The **covariance matrix** of Y is denoted

$$V(Y) = \begin{pmatrix} V(Y_1) & \text{Cov}(Y_1, Y_2) & \ldots & \text{Cov}(Y_1, Y_n) \\ \text{Cov}(Y_2, Y_1) & V(Y_2) & \ldots & \text{Cov}(Y_2, Y_n) \\ \vdots & & \ddots & \vdots \\ \text{Cov}(Y_n, Y_1) & \text{Cov}(Y_n, Y_2) & \ldots & V(Y_n) \end{pmatrix}$$

That is each entry $a_{i,j} = \text{Cov}(Y_i, Y_j)$. It is created by $\text{Cov}\{(Y - E(Y))(Y - E(Y))'\}$, the outer product.

Proposition 5. If A is a constant matrix, X a random vector, then $E(AX) = AE(X)$

Proposition 6. If b is a constant vector, Y a random vector, then $V(b'Y) = b'V(Y)b$.

Multiple Linear Regression Continued

Above, we wrote $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, that is $y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p} + \epsilon_i$ in matrix form. Explicitly,

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \ldots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \ldots & x_{2,p} \\ \vdots & & \ddots & & \vdots \\ 1 & x_{n,1} & x_{n,2} & \ldots & x_{n,p} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$\mathbf{Y}, \epsilon \in \mathbb{R}^n, \beta \in \mathbb{R}^{p+1}$, and $\mathbf{X}$ is $n \times (p+1)$ dimensional.

As before, we would like to minimize $\sum_i^n e_i^2$ given values in X. This evaluates to the scalar

$$RSS(\beta) = \sum_i^n e_i^2 = e'e = (Y - X\beta)'(Y - X\beta) = Y'Y - 2Y'X\beta + \beta'X'X\beta$$

Where $Y'X\beta = \beta'X'Y$ since the transpose of a scalar is the same scalar. Note $RSS : \mathbb{R}^{p+1} \to \mathbb{R}$ Differentiating with respect to β,

$$\frac{\partial RSS}{\partial \beta} = \frac{\partial}{\partial \beta}(Y'Y - 2\beta'X'Y + \beta'X'X\beta) = -2X'Y + 2X'X\beta$$

Setting this to 0, we see $\hat{\beta} = (X'X)^{-1}X'Y$. In the case of simple LR,

$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix},\ Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \implies X'X = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} = n\begin{pmatrix} 1 & \overline{x} \\ \overline{x} & \frac{1}{n}\sum x_i^2 \end{pmatrix}$$

We can compute $\det X'X = n^2 \cdot \left(\frac{1}{n}\sum x_i^2 - \overline{x}^2\right) = n \cdot \sum (x_i - \overline{x})^2 = n \cdot S_{X,X}$. Therefore

$$(X'X)^{-1} = \begin{pmatrix} \dfrac{\sum x_i^2}{n \cdot S_{X,X}} & -\dfrac{\overline{x}}{S_{X,X}} \\ -\dfrac{\overline{x}}{S_{X,X}} & \dfrac{1}{S_{X,X}} \end{pmatrix}$$

Multiplying by σ^2, we see this is the **covariance matrix for $\hat{\beta}_0, \hat{\beta}_1$;** $\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \dfrac{-\sigma^2\overline{x}}{S_{X,X}}$. **Important for midterm!**

Definition 9. The **projection** of Y on X is given by $\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = HY$. We call H the **hat** or **projection** matrix. Note it is $n \times n$, idempotent, and symmetric!

We let $e = Y - \hat{Y} = Y - X(X'X)^{-1}X'Y = (I - H)Y$

Proposition 7. H and $I - H$ are both idempotent.

Note that $HX = X$; this is easily checked by tracing definition and cancelling inverses. We can partition the first k and last $p+1-k$ columns of X into matrix $[X_1, X_2]$. Then $HX = [HX_1, HX_2] = X = [X_1, X_2]$. As well, $\mathrm{tr}(H) = p+1$ and $\dim \mathrm{range} H = p+1$.

Assumptions in Multiple LR

$E(Y \mid X) = X \cdot \beta$. Linearity, independence, homoescadicity, normality hold as assumptions for our model (same as before). We assume $\epsilon \sim N(0, \sigma^2 I)$. Then $Y \mid X \sim N(X\beta, \sigma^2 I)$. Now we discuss the distribution of $\hat{\beta}$.

$$E(\hat{\beta} \mid X) = E((X'X)^{-1}X'Y \mid X) = (X'X)^{-1}X'X\beta = \beta$$

so the estimator is consistent. For the variance, we carry out adjoints as in previous property

$$V(\hat{\beta} \mid X) = V((X'X)^{-1}X'Y \mid X) = (X'X)^{-1}X'\sigma^2 I X(X'X)^{-1} = (X'X)^{-1}\sigma^2$$

This is just the covariance matrix of $\hat{\beta}$! Look back to our example above. That is

$$C = (X'X)^{-1} \implies c_{ij} = \sigma^2 \mathrm{Cov}(\beta_i, \beta_j)$$

Least squares estimates are the **best linear unbiased estimators** according to the Gauss-Markov Theorem (which is stated later). The following assumptions are required for the theorem: (1) the errors ϵ_i are independent, (2) $E(\epsilon) = 0$, (3) $V(\epsilon) = \sigma^2$. Note normality is **not** assumed.

As in simple LR, the $\hat{\beta}_j$ are normally distributed; $\hat{\beta}_j \sim N(\beta_j, \sigma^2 c_{j,j})$. We can test hypotheses for β_j in the usual way. Given $H_O : \beta_j^0$, then we can calculate $Z = \frac{\hat{\beta}_j - \beta_j^0}{\sqrt{c_{j,j}}\sigma}$ and use a z-test.

May 30: Lecture 5

Term Test

Higher than expected. Expect lots of multiple linear regression questions in the final, like Question 5 on TT. Practice from Chapter 3 in Montgomery.

ANOVA for Multiple Linear Regression

Expectation of RSS, sample variance

The RSS for MLR is $\sum(y_i - \hat{y}_i) = e'e$. Recall $e = (I-H)y$ since $Y - \hat{Y} = Y - HY = (I-H)Y$, where $H = X(X'X)^{-1}X'$. Therefore

$$RSS = y'[I - X(X'X)^{-1}X']y = y'[I-H]y$$

In MLR, we have $p+1$ parameters to estimate so reasoning with degrees of freedom, the **sample variance** $s^2 = \dfrac{RSS}{n-p-1} = \dfrac{\sum e_i^2}{n-p-1}$. We show this by first calculating expectation of RSS by proving a theorem, and substituting $A = I - H$. Please see last lecture for properties of expectation and variance.

Theorem 2. If y is $n \times 1$ random vector, with mean vector μ and covariance matrix V, and A is a matrix of constants, then

$$E(y'Ay) = \text{tr}(AV) + \mu' A \mu$$

Proof. We multiply and use linearity of expectation, expansion of covariance

$$E(RSS) = E[Y'AY] = E\left(\sum_i^n \sum_j^n a_{i,j} y_i y_j\right) = \sum_i^n \sum_j^n a_{i,j} E\left(y_i y_j\right)$$

Expanding with covariance, and with $(\sigma_{i,j}) = Cov(Y) = V$

$$\begin{aligned}
&= \sum_i^n \sum_j^n a_{i,j}\left(\text{Cov}(y_i, y_j) + E(y_i)E(y_j)\right) = \sum_i^n \sum_j^n a_{i,j}\sigma_{i,j} + \sum_i^n \sum_j^n a_{i,j}\mu_i\mu_j \\
&= \text{tr}(AV) + \mu' A\mu
\end{aligned}$$

□

Proposition 8. $E(RSS) = (n-p-1)\sigma^2 + \mu' A \mu$ where $A = I - H$

Proof. Using the above, Set $A = I - H$, $V = \sigma^2 I$, then

$$\text{tr}(AV) = \text{tr}[(I-H)\sigma^2 I] = \sigma^2 \text{tr}(I-H)$$

Expanding, $\text{tr}(I-H) = \text{tr}(I_n) - \text{tr}(H) = n - p - 1$; where $\text{tr}(H) = \text{tr}(X(X'X)^{-1}X') = \text{tr}(X'X(X'X)^{-1}) = \text{tr} I_{p+1} = p+1$ since $(X'X)^{-1}$ is $(p+1)\times(p+1)$. □

This will be on the final!

Proposition 9. $\mu' A \mu = 0$, where $A = I - H$, $\mu = X\beta$.

Proof.

$$\begin{aligned}\mu' A\mu &= (X\beta)'(I - X(X'X)^{-1}X')X\beta = \beta'X'X\beta - \beta'X'X(X'X)^{-1}X'X\beta \\ &= \beta'X'X\beta - \beta'X'X'\beta \\ &= 0\end{aligned}$$

□

Proposition 10. $E(RSS) = (n-p-1)\sigma^2$

This follows from substitution into the past 3 statements. The following proposition also easily follows.

Proposition 11. $E(MRSS) = E(\frac{RSS}{n-p-1}) = \sigma^2$

RSS **and** SS_{reg} **for Multiple LR**

By Gauss-Markov assumptions, $\epsilon_i \sim N(0,\sigma^2)$, and so $\frac{\epsilon_i}{\sigma} \sim N(0,1)$ by Z-score. Also this gives $\frac{1}{\sigma}\epsilon \sim N(0,I)$. Note [1]

$$e = Y - X\hat{\beta} = Y - HY = AY$$

Our underlying model is assumed to be $Y = X\beta + \epsilon$, so therefore $Ay = AX\beta + A\epsilon$. Expanding and since $HX = X$, $AX\beta = (I-H)X\beta = 0$ so $e = Ay = A\epsilon$. That is our observed errors are the difference $\epsilon - H\epsilon$; the error vector minus its projection. This proves the following fact

Fact 1. $e = (I-H)\epsilon$.

We also showed $A = I - H$ is **symmetric and idempotent**; this implies

$$A'A = A^2 = A$$

Then

$$RSS = (y-\hat{y})'(y-\hat{y}) = e'e = \epsilon'A'A\epsilon = \epsilon A\epsilon = \sigma^2 Z'AZ$$

This implies $\dfrac{RSS}{\sigma^2} = Z'AZ$.

Theorem 3. If A is a symmetric and idempotent $n \times n$ matrix and $Z \sim N(0,I)$, then $Z'AZ \sim \chi^2(\text{tr}(A))$

No proof, try it yourself for practice. However, notice $Z'Z \sim \chi^2(n)$ and use a nice basis for a projection operator. Recall $A = I - H$ so this gives

$$\frac{RSS}{\sigma^2} \sim \chi^2(\text{tr}(A)) = \chi^2(n-p-1)$$

Proposition 12. $\overline{y} = (1'1)^{-1}1'y$.

Therefore we may rewrite the regression sum of squares involving y and H.

Proposition 13. $SS_{reg} = y'[H - 1(1'1)^{-1}1']y$

[1]The slides use $Q = I - H$, but we use A as before.

Proof. First, write

$$SS_{reg} = [\hat{y} - 1\overline{y}]'[\hat{y} - 1\overline{y}] = y'[H - 1(1'1)^{-1}1]'[H - 1(1'1)^{-1}1']y$$

Now we show $[H - 1(1'1)^{-1}1]'[H - 1(1'1)^{-1}1'] = H - 1(1'1)^{-1}1'$. Expanding,

$$[H - 1(1'1)^{-1}1']'[H - 1(1'1)^{-1}1'] = H^2 - 1(1'1)^{-1}1'H - H1(1'1)^{-1}1' + 1(1'1)^{-1}1'1(1'1)^{-1}1'$$

Note since $HX = X$ and 1 is a column in X, then $H \cdot \mathbf{1} = \mathbf{1}$, and taking transposes a similar result holds.

$$\begin{aligned} &= H - 1(1'1)^{-1}1' - 1(1'1)^{-1}1' + 1(1'1)^{-1}1' \\ &= H - 1(1'1)^{-1}1' \end{aligned}$$

□

This gives us a way to express the regression sum squared using y, H and a constant matrix. We use this to show that the regression sum squared, and residual sum squared from above are independent.

Proposition 14. The regression sum of squares $SS_{reg} = y'[H - 1(1'1)^{-1}1']y$ and residual sum of squares $RSS = y'[I - H]y$ are **independent**.

We do this by computing $[H - 1(1'1)^{-1}1']\sigma^2 I[I - X(X'X)^{-1}X'] = \sigma^2(H - B)(I - H) = 0$ as an exercise.

ANOVA Table for MLR

Sources of Variation	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regression	p	SS_{reg}	$MS_{reg} = \frac{SS_{reg}}{p}$	$F_0 = \frac{MS_{reg}}{MRSS}$	etc
Residuals	$n-p-1$	RSS	$MRSS_{reg} = \frac{RSS}{n-p-1}$		
Total	$n-1$	SST			

By independence, and since $SS_{reg} \sim \chi^2(p)$, $RSS \sim \chi^2(n-p-1)$, then we have

$$\frac{SS_{reg}/p}{RSS/n-p-1} \sim F(p, n-p-1)$$

We may perform an F test with the null hypothesis

$$H_0 : \beta_0 = \beta_1 = \ldots = \beta_p = 0 \text{ and } H_1 : \beta_i \neq 0 \text{ for some } i$$

Significance in the statistic gives evidence for at least one predictor being valid; at least some X_i explains a significant proportion of the variance in Y.

Recall that the coefficient of determination $R^2 = \frac{SS_{reg}}{SST}$. As the number of variables increases, so does R^2, since more predictors decrease RSS.

$$RSS = \sum (y_i - \beta_0 - \beta_1 x_i - \ldots - \beta_p x_p)^2$$

where an additional predictor will decrease each term in the sum. Note when we have n predictors for the sample size, we have a perfect fit and $R^2 = 1$. Geometrically, projection plane induced by $H = X(X'X)^{-1}X'$ is the whole space. In short, we get many predictors but none of them good and we overfit. We account for the number of predictors using an adjusted R^2

$$R^2_{adj} = 1 - \frac{RSS/n-p-1}{SST/n-1}$$

The interpretation is exactly the same, but is a more robust statistic in multiple linear regression due to the previous issues.

Partial F-test

One of the most important tests in an MLR is the partial F-test. In ANOVA we do a test for the full model; we identify whether there is any significant predictor. The **partial F-test** identifies whether a subset of predictors still significantly predicts the response. However, the **null hypothesis is that the reduced model is better** than the full model. Remember that we consider the ratio of the error sum squared; a significant increase in errors after removing predictors indicates a worse model, and larger F-statistic.

Suppose we have two models. The **full** $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon$. We test whether the model still explains the response when we remove the first k predictors; we consider the **reduced** $Y = \beta_{k+1} X_{k+1} + \ldots + \beta_p X_p + \epsilon$. First, write

$$RSS(\text{reduced}) - RSS(\text{full}) = y'[H - H_1]y$$

Without proof, but similar to for RSS before,

$$RSS(\beta_2 \mid \beta_1)/\sigma^2 = (RSS(\text{reduced}) - RSS(\text{full}))/\sigma^2 \sim \chi^2(k)$$

Thus

$$\frac{RSS(\beta_2 \mid \beta_1)/\sigma^2}{RSS(\text{full})/n-p-1} \sim F(k, n-p-1)$$

We test

$$H_0: \text{ reduced model is better fit}, \quad H_1: \text{ full model is better fit}$$

A large F value suggests that the reduced model explains much less variability than the full model, and fits the data worse. This implies we should be rejecting the null, so predictors cannot be removed from the model. Small values imply that both reduced and full models explain a similar amount of variability, so the additional predictors may not be necessary.

Opposite test hypotheses occur, since we test ratios of **residuals**; high ratio means large residuals in reduced model.

Diagnostic checking

The three assumptions of linear regression are (1) linearity, (2) homoscedasticity, (3) independence of the errors, with normality also being one. One of the most important tasks is **checking the assumptions** in our data. This is called diagnostic checking. Anscombe's datasets give an example of why checking these assumptions is important; the models give the same predictors but differ greatly in their structure.

Suppose we fit $Y = \beta_0 + \beta_1 X + \epsilon$. The fitted regression $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$ produces the estimate for $E(Y \mid X)$. e is an unbiased estimate for ϵ. A good way to check is to plot the residuals, there should be no pattern and should be a random scatter plot. We can also plot residuals against $\hat{y}$ as in multiple LR. Assumptions hold if there is **no pattern**. Other relationships, like a quadratic one, will become apparent in the residuals. The following steps are best practice:

1. Assess model assumptions using residual plot. There should be no pattern.
2. Determine which data points have x-values with large effect on Y. (**Leverage points.**)

3. Determine which points are outliers in their responses.
4. Assess the influence of bad leverage points on the fitted model.
5. Examine whether constant error variance assumption is reasonable. (Do residuals vary with X?)
6. If data is collected over prolonged period of time, see if it is corelated with time.
7. For small sample size or prediction intervals, assess whether normality of errors is reasonable. (Normality tests?)

If this is successful, then our assumptions are valid and our predictors can be trusted. If the assumptions fail, our analysis is invalid.

June 1: Lecture 6

Leverage Points

Leverage points are observations that are highly influential on the fitted regression line. Leverage points occur due to an a value of $X = x$ far from $\overline{x}$. The corresponding $Y = y$ greatly influences the line for a given $X = x$. Such a pair x, y that greatly changes the least square estimates is a **bad** leverage point. For extreme x, if y is close to the fitted line it is a good leverage point, but if it is far it is a bad one. An **outlier** is an observation that takes an extreme y value for an x that is not far from $\overline{x}$.

Numerical Summary

Recall

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \qquad \hat{\beta}_1 = \sum_j^n \frac{x_j - \overline{x}}{S_{X,X}} y_j = \sum_j^n c_j y_j$$

Then

$$\begin{aligned} \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \overline{y} + \hat{\beta}_1 (x_i - \overline{x}) &= \sum_j^n \left(\frac{y_j}{n} + \frac{(x_j - \overline{x})(x_i - \overline{x})}{S_{X,X}} y_j \right) \\ &= \sum_j^n \left(\frac{1}{n} + \frac{(x_j - \overline{x})(x_i - \overline{x})}{S_{X,X}} \right) y_j = \sum_j^n h_{i,j} y_j \end{aligned}$$

This $h_{i,j}$ is the entry in the hat matrix H. When $i = j$, then $h_{i,i} = \dfrac{1}{n} + \dfrac{(x_i - \overline{x})^2}{S_{X,X}}$. We show $\sum_j^n h_{i,j} = 1$. Further, we can write $\hat{y}_i = h_i i\, y_i + \sum_{j \neq i} h_{i,j} y_j$. If we have $h_{i,i} \approx 1$, then $\hat{y}_i$ is close to y_i, and it is a leverage point. It can also be shown mean$(h_{i,i}) = \frac{2}{n}$ (by definition). Using this, a popular way to identify a leverage point is to check if $h_{i,i} > \frac{4}{n}$, or twice the mean. This is a useful rule of thumb.

Leverage is concerned with a single observation far from the rest of the data in the x-space. We have two ways of dealing with bad leverage points. We can (1) remove the data point or (2) fit a different regression model. A quadratic or logarithmic transformation of X may be needed.

Standardized Residuals and Influential Points

In the real world, often people work in **sensitivity analysis**. This is essentially identifying influential points, which we discuss.

Residuals reflect the difference between observed and predicted response. We might want to use them to measure the influence a leverage point will have on the estimated line. It turns out that the estimated residuals do not always have the same variance; $V(e_i)$ is not the same for all i. Actually, we find $V(e_i) = \sigma^2(1-h_{i,i})$. We prove this.

Proposition 15. $\sum_{j=1}^n h_{i,j}^2 = h_{i,i}$

Proof.

$$\begin{aligned}
\sum_{j=1}^n h_{i,j}^2 &= \sum_{j=1}^n \left(\frac{1}{n} + \frac{(x_j-\overline{x})(x_i-\overline{x})}{S_{X,X}}\right)^2 = \sum_{j=1}^n \left(\frac{1}{n^2} + \frac{2}{n}\frac{(x_j-\overline{x})(x_i-\overline{x})}{S_{X,X}} + \frac{(x_j-\overline{x})^2(x_i-\overline{x})^2}{S_{X,X}^2}\right) \\
&= \frac{1}{n} + \frac{2}{n}\frac{\sum_{j=1}^n (x_j-\overline{x})(x_i-\overline{x})}{S_{X,X}} + \frac{\sum_{j=1}^n (x_j-\overline{x})^2(x_i-\overline{x})^2}{S_{X,X}^2} \\
&= \frac{1}{n} + \frac{(x_i-\overline{x})^2}{S_{X,X}} = h_{i,i}
\end{aligned}$$

□

We use this in the last steps to prove the following.

Proposition 16. $V(e_i) = \sigma^2(1-h_{i,i})$

Proof.

$$\begin{aligned}
V(e_i) &= V\left(y_i(1-h_{i,i}) + \sum h_{i.j}y_j\right) = (1-h_{i,i})^2 V(y_i) + \sum_{i\neq j} h_{i,j}^2 V(y_j) \\
&= \sigma^2\left((1-h_{i,i})^2 + \sum_{i\neq j} h_{i,j}^2\right) = \sigma^2\left((1-h_{i,i})^2 + \sum_{i}^n h_{i,j}^2 - h_{i,i}^2\right) \\
&= \sigma^2\left(1-2h_{i,i} + h_{i,i}^2 + h_{i,i} - h_{i,i}^2\right) = \sigma^2(1-h_{i,i})
\end{aligned}$$

□

We can now discuss the variation in each residual using our hat matrix. We see that the estimated residuals are not actually independent, event though we assume that the errors are. If e_i were independent, we would expect $Var(e_i) = \sigma^2$. However, we have an extra term of $-\sigma^2 h_{i,i}$, which indicates the variance of a residual depends on its distance from $\overline{x}$. Residuals are correlated, but the correlation is small.

This makes it difficult to know whether the patterns we see are due to model violations or variance of the residuals. To overcome this issue, we **standardize** the residuals by dividing by their **standard error**. By prop. 16,

$$\text{se}(e_i) = s\sqrt{1-h_{i,i}} \implies r_i = \frac{e_i}{s\sqrt{1-h_{i,i}}}$$

Where $s^2 = \frac{\sum e_i^2}{n-2}$. Note $r_i \sim t(n-2)$, so these are also called 'studentized' residuals. If high leverage points exist, it is more important to look at plots of standardized residuals; we can just check if $r_i \in [-2,2]$ or $[-4,4]$. It is expected that the variance of r_i will be larger for center values of X, and smaller for remote values. Then looking at the plot, we can identify whether a residual corresponds to an outlier; we plot standardizes residual against dependent variable.

Example 7. In our Treasury Bond example, we identify three bad leverage points by plotting studentized residual against dependent variable. Viewing these in detail, we find that they are 'flower bonds', so we remove them from the analysis. The remaining points are more or less linear, but a slight bend may give evidence that it is a logarithmic relationship.

Cook's Distance

How can we quantify the influence a small number of observations on the regression line with a single statistic? In 1977, Cook provided the following expression to calculate the influence of a single point on the regression line.

Definition 10. The **Cook's distance** for (x_i, y_i) is given by

$$D_i = \frac{(\hat{y}_{j(i)} - \hat{y}_j)^2}{2s^2} = \frac{r_i^2}{2} \cdot \frac{h_i}{1-h_i}$$

where the subscript i references the predicted value from a model fit without (x_i, y_i). Thus $\hat{y}_{j(i)}$ denotes the jth fitted value based on the fit when the ith observation is deleted from the fit.

There are similar metrics in MLR which we discuss later. The second expression is easier to work with since it does not require refitting of any models. Large Cook's distance means large r_i or large $h_{i,i}$. We use the cutoff $D_i > \dfrac{4}{n-2}$ as a rough cutoff guideline, but identifying unusual D_i is most important.

Example 8. In the previous Treasury Bond example, the 3 unusual observations have a very high Cook's distance when plotted, and are valid to be removed.

Normality of the errors

We need to assume ϵ_i is normally distributed to perform F, t, and Z tests, as well as construct confidence intervals. We will verify the normality assumption using residual plots. First, we can show $\sum h_{i,j} = 1$ and $\sum x_j h_{i,j} = x_i$,

Proposition 17. $e_i = \epsilon_i - \sum_j^n h_{i,j}\epsilon_j$

Proof.

$$e_i = y_i - \hat{y}_i = y_i - \sum_j^n h_{i,j} y_j = \beta_0 + \beta_1 x_i + \epsilon_i - \sum_j^n h_{i,j}(\beta_0 + \beta_1 x_j + \epsilon_j) = \epsilon_i - \sum_j^n h_{i,j}\epsilon_j$$

□

In small sample sizes, the second term may dominate, and the residuals may look normal even if the ϵ_i are not. As n increases, the second term in the last equation has a smaller variance than the first term, so the first term dominates the last equation. For large samples, the residuals can be used to assess normality of the samples.

A common way to assess normality is via a **QQ-plot**; the studentized errors are plotted against their quantiles. If the quantiles match that of a normal distribution, the plot is close to the $y = x$ line. We must also check that the constant variance assumption is met; we cannot use inferential tools if it is not true.

Variance stabilizing transformations

In the slides example, constant variances is violated. For inference, our prediction intervals depend on X. A transformation of Y can stabilize the variance: make it not depend on X.

When we are counting events, as in the slides (28) example, we typically fit a Poisson distribution. In a Poisson distribution, the mean and variance are both λ. Since in regression we model the conditional mean $E(Y \mid X) = \lambda_X$, we have also a conditional variance: λ_X changes by X, so should the variance. The **square root transformation** can help in this situation.

Taking the function of a random variable, $f(Y) \approx f(E(Y)) + f'(E(Y))(Y - E(Y))$. Taking the variance, we get

$$V(f(Y)) = \left(f'(E(Y)\right)^2 V(Y)$$

since $E(Y)$ is a constant, and using variance properties. This way of approximating variance is called the **delta method**. In the Poisson example, $E(Y \mid X) = V(Y \mid X) = \lambda(x)$. Letting $f(Y) = \sqrt{(}Y)$, then

$$V(Y^{0.5} \mid X) = \left(0.5E(Y \mid X)^{-0.5}\right)^2 V(Y \mid X) = \left(\frac{1}{2}\right)^2 \lambda(x)^{-1}\lambda(x) = \frac{1}{4}$$

which makes $V(f(Y) \mid X)$ constant. In the example, X, Y are both counts, so we perform the square root transformation on both and keep the same units. The variance stabilizing transformation stabilizes prediction error across the predictor variable. Our predictions may vary, but we keep the transformed model. We may not always get count data, so depending on the relationship between variance and mean we might use (table)

We need to make sure that interpretability is not lost: in practical data a transformation is chosen empirically. There is no specific rule.