Statistics 512: Applied Linear Models

Topic 1

Topic Overview

This topic will cover

- Course Overview & Policies
- SAS
- KNNL Chapter 1 (emphasis on Sections 1.3, 1.6, and 1.7; much should be review) – Simple linear regression; method of least squares (LS)
- KNNL Chapter 2 (emphasis on Sections 2.1-2.9) – Inference in simple linear regression, Prediction intervals and Confidence bands, ANOVA tables, General Linear Test

Class Policies

Refer to handout

Overview

We will cover

- simple linear regression – KNNL Chapters 1-5
- multiple regression – KNNL Chapters 6-11
- analysis of variance (ANOVA) – KNNL Chapters 15-25

and possibly throw in some other stuff for fun...

Emphasis will be placed on using selected practical tools (such as SAS) rather than on mathematical manipulations. We want to *understand* the theory so that we can *apply* it appropriately. Some of the material on SLR will be review, but our goal with SLR is to be able to generalize the methods to MLR.

SAS

SAS is the program we will use to perform data analysis for this class. **Learning to use SAS will be a large part of the course.**

Getting Help with SAS

Several sources for help:

- SAS Help Files (not always best)
- World Wide Web (look up the syntax in your favorite search engine)
- SAS Getting Started and Tutorials
- Statistical Consulting Service
- Evening Help Sessions
- *Applied Statistics and the SAS Programming Language, 5th edition* by Cody and Smith; most relevant material in Chapters 1, 2, 5, 7, and 9.

Statistical Consulting Service

Math G175 Hours 10-4 M through F
`http://www.stat.purdue.edu/scs/` I will often give examples from SAS in class. The programs used in lecture (and any other programs you should need) will be available for you to download from the website.

I will usually have to edit the output somewhat to get it to fit on the page of notes. You should run the SAS programs yourself to see the real output and experiment with changing the commands to learn how they work. Let me know if you get confused about what is input, output, or my comments. I will tell you the names of all SAS files I use in these notes. If the notes differ from the SAS file, take the SAS file to be correct, since there may be cut-and-paste errors.

There is a tutorial in SAS to help you get started.
`Help` → `Getting Started with SAS Software`
You should spend some time before next week getting comfortable with SAS (see HW #0).

For today, don't worry about the detailed syntax of the commands. Just try to get a sense of what is going on.

Example (Price Analysis for Diamond Rings in Singapore)

Variables

- *response variable* – price in Singapore dollars (Y)
- *explanatory variable* – weight of diamond in carats (X)

Goals

- Create a scatterplot
- Fit a regression line
- Predict the price of a sale for a 0.43 carat diamond ring

SAS Data Step

File `diamond.sas` on website.

> One way to input data in SAS is to type or paste it in. In this case, we have a sequence of ordered pairs (weight, price).

```
data diamonds;
    input weight price @@;
    cards;
.17 355 .16 328 .17 350 .18 325 .25 642 .16 342 .15 322 .19 485
.21 483 .15 323 .18 462 .28 823 .16 336 .20 498 .23 595 .29 860
.12 223 .26 663 .25 750 .27 720 .18 468 .16 345 .17 352 .16 332
.17 353 .18 438 .17 318 .18 419 .17 346 .15 315 .17 350 .32 918
.32 919 .15 298 .16 339 .16 338 .23 595 .23 553 .17 345 .33 945
.25 655 .35 1086 .18 443 .25 678 .25 675 .15 287 .26 693 .15 316
.43 .
;

data diamonds1;
    set diamonds;
    if price ne .;
```

Syntax Notes

- Each line must end with a semi-colon.
- There is no output from this statement, but information does appear in the `log` window.
- Often you will obtain data from an existing SAS file or import it from another file, such as a spreadsheet. Examples showing how to do this will come later.

SAS `proc print`

Now we want to see what the data look like.

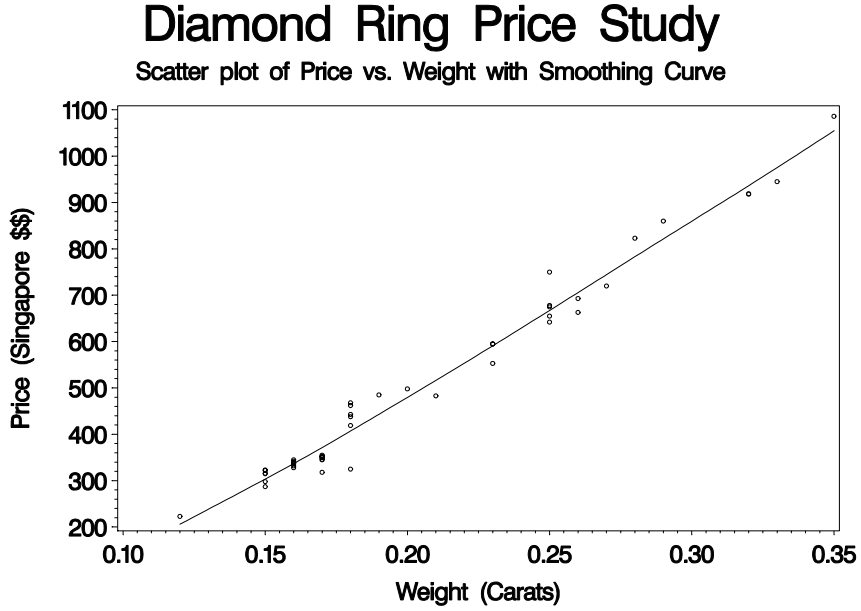
```
proc print data=diamonds;
run;
```

```
Obs    weight    price
  1     0.17       355
  2     0.16       328
  3     0.17       350

          ...

 47     0.26       693
 48     0.15       316
 49     0.43         .
```

SAS `proc gplot`

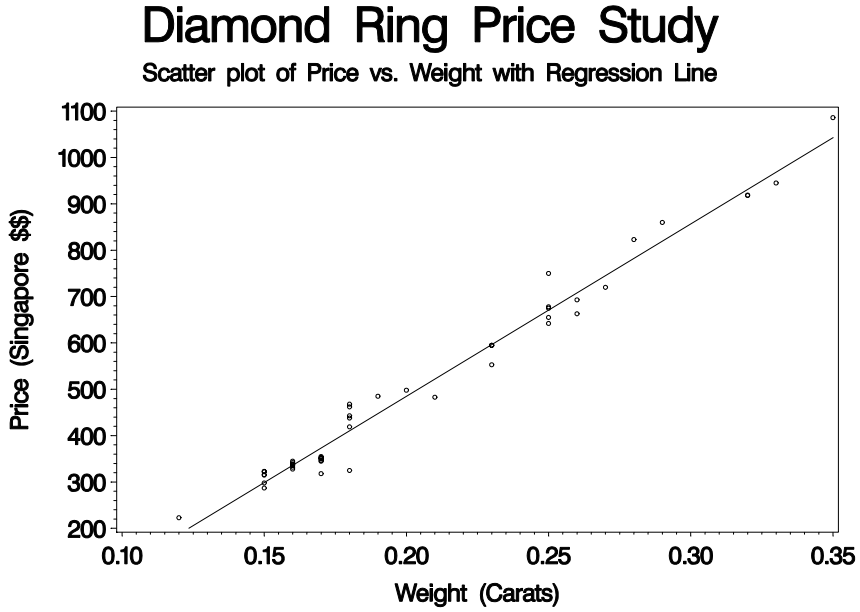
We want to plot the data as a scatterplot, using circles to represent data points and adding a curve to see if it looks linear. The `symbol` statement "`v = circle`" (`v` stands for "value") lets us do this. The symbol statement "`i = sm70`" will add a smooth line using splines (interpolation = smooth). These are options which stay on until you turn them off. In order for the smoothing to work properly, we need to sort the data by the X variable.

```
proc sort data=diamonds1; by weight;
symbol1 v=circle i=sm70;
title1 'Diamond Ring Price Study';
title2 'Scatter plot of Price vs. Weight with Smoothing Curve';
axis1 label=('Weight (Carats)');
axis2 label=(angle=90 'Price (Singapore $$)');
proc gplot data=diamonds1;
     plot price*weight / haxis=axis1 vaxis=axis2;
run;
```



Now we want to use the simple linear regression to fit a line through the data. We use the `symbol` option "`i = rl`", meaning "interpolation = regression line" (that's an "L", not a one).

```
symbol1 v=circle i=rl;
title2 'Scatter plot of Price vs. Weight with Regression Line';
proc gplot data=diamonds1;
    plot price*weight / haxis=axis1 vaxis=axis2;
run;
```



SAS `proc reg`

We use `proc reg` (regression) to estimate a regression line and calculate predictors and residuals from the straight line. We tell it what the data are, what the model is, and what options we want.

```
proc reg data=diamonds; model price=weight/clb p r;
   output out=diag p=pred r=resid;
   id weight; run;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2098596	2098596	2069.99	<.0001
Error	46	46636	1013.81886		

```
Corrected Total          47        2145232

Root MSE             31.84052    R-Square     0.9783
Dependent Mean      500.08333    Adj R-Sq     0.9778
Coeff Var             6.36704
```

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-259.62591	17.31886	-14.99	<.0001
weight	1	3721.02485	81.78588	45.50	<.0001

```
proc print data=diag;
run;
```

Output Statistics

Obs	weight	Dep Var price	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual
1	0.17	355.0000	372.9483	5.3786	-17.9483	31.383
2	0.16	328.0000	335.7381	5.8454	-7.7381	31.299
3	0.17	350.0000	372.9483	5.3786	-22.9483	31.383
4	0.18	325.0000	410.1586	5.0028	-85.1586	31.445
5	0.25	642.0000	670.6303	5.9307	-28.6303	31.283
...						
46	0.15	287.0000	298.5278	6.3833	-11.5278	31.194
47	0.26	693.0000	707.8406	6.4787	-14.8406	31.174
48	0.15	316.0000	298.5278	6.3833	17.4722	31.194

Simple Linear Regression

Why Use It?

- Descriptive purposes (cause/effect relationships)
- Control (often of cost)
- Prediction of outcomes

Data for Simple Linear Regression

- Observe $i = 1, 2, \ldots, n$ pairs of variables (explanatory, response)
- Each pair often called a *case* or a data point
- $Y_i = i$th response
- $X_i = i$th explanatory variable

Simple Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \text{ for } i = 1, 2, \ldots, n$$

Simple Linear Regression Model Parameters

- β_0 is the intercept.
- β_1 is the slope.
- ϵ_i are independent, normally distributed random errors with mean 0 and variance σ^2, i.e.,

$$\epsilon_i \sim N(0, \sigma^2)$$

Features of Simple Linear Regression Model

- Individual Observations: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
- Since ϵ_i are random, Y_i are also random and

$$\begin{aligned} \mathrm{E}(Y_i) &= \beta_0 + \beta_1 X_i + \mathrm{E}(\epsilon_i) = \beta_0 + \beta_1 X_i \\ \mathrm{Var}(Y_i) &= 0 + \mathrm{Var}(\epsilon_i) = \sigma^2. \end{aligned}$$

  Since ϵ_i is Normally distributed, $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$ (See A.4, page 1302)

Fitted Regression Equation and Residuals

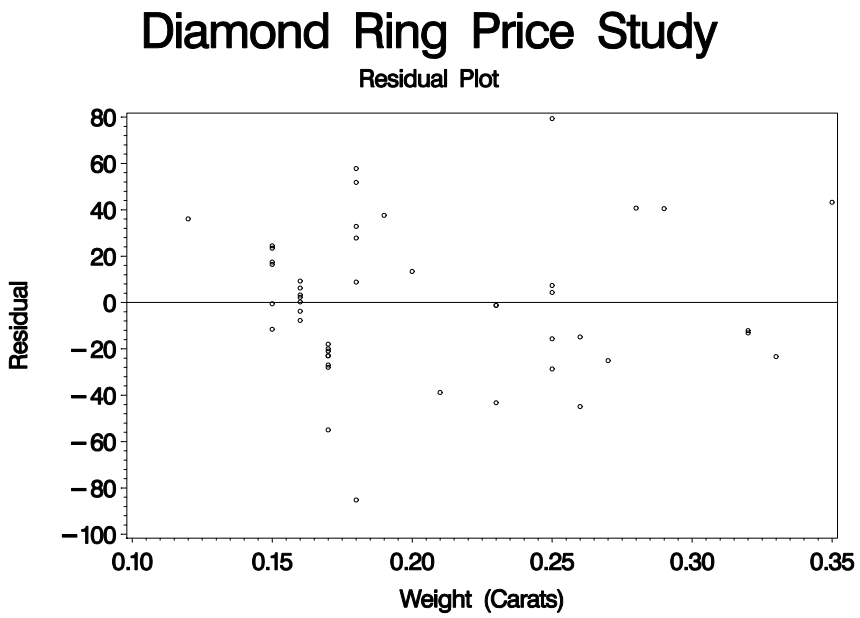
We must estimate the parameters β_0, β_1, σ^2 from the data. The estimates are denoted b_0, b_1, s^2. These give us the *fitted* or *estimated regression line* $\hat{Y}_i = b_0 + b_1 X_i$, where

- b_0 is the estimated intercept.
- b_1 is the estimated slope.
- $\hat{Y}_i$ is the estimated mean for Y, when the predictor is X_i (i.e., the point on the fitted line).
- e_i is the residual for the ith case (the vertical distance from the data point to the fitted regression line). Note that $e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i)$.

Using SAS to Plot the Residuals (Diamond Example)

When we called `proc reg` earlier, we assigned the residuals to the name "`resid`" and placed them in a new data set called "`diag`". We now plot them vs. X.

```
proc gplot data=diag;
  plot resid*weight / haxis=axis1 vaxis=axis2 vref=0;
  where price ne .;
run;
```



Notice there does not appear to be any obvious pattern in the residuals. We'll talk a lot more about diagnostics later, but for now, you should know that looking at residuals plots is an important way to check assumptions.

Least Squares

- Want to find "best" estimators b_0 and b_1.
- Will minimize the sum of the squared residuals $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(Y_i - (b_0 + b_1 X_i))^2$.
- Use calculus: take derivative with respect to b_0 and with respect to b_1 and then set the two result equations equal to zero and solve for b_0 and b_1 (see KNNL, pages 17-18).

Least Squares Solution

- These are the best estimates for β_1 and β_0.

$$\begin{aligned} b_1 &= \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{SS_{XY}}{SS_X} \\ b_0 &= \bar{Y} - b_1\bar{X} \end{aligned}$$

- These are also maximum likelihood estimators (MLE), see KNNL, pages 26-32.
- This estimate is the "best" because because it is *unbiased* (its expected value is equal to the true value) with *minimum variance*.

Maximum Likelihood

$$\begin{aligned} Y_i &\sim N(\beta_0 + \beta_1 X_i, \sigma^2) \\ f_i &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma}\right)^2} \\ L &= f_1 \times f_2 \times \ldots \times f_n \text{ – likelihood function} \end{aligned}$$

Find values for β_0 and β_1 which maximize L. These are the SAME as the least squares estimators b_0 and b_1!!!

Estimation of σ^2

We also need to estimate σ^2 with s^2. We use the sum of squared residuals, SSE, divided by the degrees of freedom $n - 2$.

$$\begin{aligned} s^2 &= \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum e_i^2}{n-2} \\ &= \frac{SSE}{df_E} = MSE \\ s &= \sqrt{s^2} = \text{Root } MSE, \end{aligned}$$

where $SSE = \sum e_i^2$ is the sum of squared residuals or "errors", and *MSE* stands for "mean squared error".

There will be other estimated variance for other quantities, and these will also be denoted s^2, e.g. $s^2\{b_1\}$. Without any $\{\}$, s^2 refers to the value above – that is, the estimated variance of the residuals.

Identifying these things in the SAS output

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2098596	2098596	2069.99	<.0001
Error	46	46636	1013.81886		
Corrected Total	47	2145232			

Root MSE	31.84052	R-Square	0.9783
Dependent Mean	500.08333	Adj R-Sq	0.9778
Coeff Var	6.36704		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-259.62591	17.31886	-14.99	<.0001	-294.48696	-224.76486
weight	1	3721.02485	81.78588	45.50	<.0001	3556.39841	3885.65129

Review of Statistical Inference for Normal Samples

This should be review!
In Statistics 503/511 you learned how to construct confidence intervals and do hypothesis tests for the mean of a normal distribution, based on a random sample. Suppose we have an iid (random) sample $W_1, \ldots, W_n$ from a normal distribution. (Usually, I would use the symbol X or Y, but I want to keep the context general and not use the symbols we use for regression.)

We have

$$
\begin{aligned}
W_i &\sim^{iid} N(\mu, \sigma^2) \text{ where } \mu \text{ and } \sigma^2 \text{ are unknown} \\
\bar{W} &= \frac{\sum W_i}{n} = \text{ sample mean} \\
SS_W &= \sum (W_i - \bar{W})^2 = \text{ sum of squares for } W \\
s^2\{W\} &= \frac{\sum (W_i - \bar{W})^2}{n-1} = \frac{SS_W}{n-1} = \text{ sample variance (estimating variance of entire population)} \\
s\{W\} &= \sqrt{s^2\{W\}} = \text{ sample standard deviation} \\
s\{\bar{W}\} &= \frac{s\{W\}}{\sqrt{n}} = \text{ standard error of the mean (standard deviation of mean)}
\end{aligned}
$$

and from these definitions, we obtain

$$\begin{aligned} \bar{W} &\sim N\left(\mu, \frac{\sigma^2}{n}\right), \\ T &= \frac{\bar{W}-\mu}{s\{\bar{W}\}} \text{ has a } t\text{-distribution with } n-1 \text{ df (in short, } T \sim t_{n-1}) \end{aligned}$$

This leads to *inference*:

- *confidence intervals* for μ
- *significance tests* for μ.

Confidence Intervals

We are $100(1-\alpha)\%$ confident that the following interval contains μ:

$$\bar{W} \pm t_c s\{\bar{W}\} = \left[\bar{W} - t_c s\{\bar{W}\}, \bar{W} + t_c s\{\bar{W}\}\right]$$

where $t_c = t_{n-1}(1-\frac{\alpha}{2})$, the upper $(1-\frac{\alpha}{2})$ percentile of the t distribution with $n-1$ degrees of freedom, and $1-\alpha$ is the confidence level (e.g. $0.95 = 95\%$, so $\alpha = 0.05$).

Significance Tests

To test whether μ has a specific value, we use a t-test (one sample, non-directional).

$$\mathrm{H}_0 : \mu = \mu_0 \quad \text{vs} \quad \mathrm{H}_a : \mu \neq \mu_0$$

- $t = \frac{\bar{W}-\mu_0}{s\{\bar{W}\}}$ has a t_{n-1} distribution under H_0.
- Reject H_0 if $|t| \geq t^c$, where $t^c = t_{n-1}(1-\frac{\alpha}{2})$.
- p-value $= \mathrm{Prob}_{\mathrm{H}_0}(|T| > |t|)$, where $T \sim t_{n-1}$.

The p-value is twice the area in the upper tail of the t_{n-1} distribution above the observed $|t|$. It is the probability of observing a test statistic at least as extreme as what was actually observed, when the null hypothesis is really true. We reject H_0 if $p \leq \alpha$. (Note that this is basically the same – more general, actually – as having $|t| \geq t^c$.)

Important Notational Comment

The text says "conclude H_A" if t is in the rejection region ($|t| \geq t_c$), otherwise "conclude H_0". This is shorthand for

- "conclude H_A" means "there is sufficient evidence in the data to conclude that H_0 is false, and so we assume that H_A is true."
- "conclude H_0" means "there is insufficient evidence in the data to conclude that either H_0 or H_A is true or false, so we default to assuming that H_0 is true."

Notice that a failure to reject H_0 does *not* mean that there was any evidence in favor of H_0

NOTE: In this course, $\alpha = 0.05$ unless otherwise specified.

Section 2.1: Inference about β_1

$$\begin{aligned} b_1 &\sim N(\beta_1, \sigma^2\{b_1\}) \\ \text{where } \sigma^2\{b_1\} &= \frac{\sigma^2}{SS_X} \\ t &= \frac{(b_1 - \beta_1)}{s\{b_1\}} \\ \text{where } s\{b_1\} &= \sqrt{\frac{s^2}{SS_X}} \\ t &\sim t_{n-2} \text{ if } \beta_1 = 0 \end{aligned}$$

According to our discussion above for "W", you therefore know how to obtain CI's and t-tests for β_1. (I'll go through it now but not in the future.) There is one important difference: the degrees of freedom (df) here are $n-2$, not $n-1$, because we are also estimating β_0.

Confidence Interval for β_1

- $b_1 \pm t_c s\{b_1\}$,
- where $t_c = t_{n-2}(1 - \frac{\alpha}{2})$, the upper $100(1 - \frac{\alpha}{2})$ percentile of the t distribution with $n-2$ degrees of freedom
- $1 - \alpha$ is the confidence level.

Significance Tests for β_1

- $\mathrm{H}_0 : \beta_1 = 0$ vs $\mathrm{H}_a : \beta_1 \neq 0$
- $t = \frac{b_1 - 0}{s\{b_1\}}$
- Reject H_0 if $|t| \geq t^c$, $t^c = t_{n-2}(1 - \alpha/2)$
- p-value = Prob$(|T| > |t|)$, where $T \sim t_{n-2}$

Inference for β_0

- $b_0 \sim N(\beta_0, \sigma^2\{b_0\})$ where $\sigma^2\{b_0\} = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{SS_X}\right]$.
- $t = \frac{b_0 - \beta_0}{s\{b_0\}}$ for $s\{b_0\}$ replacing σ^2 by s^2 and take $\sqrt{}$
- $s\{b_0\} = s\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{SS_X}}$
- $t \sim t_{n-2}$

Confidence Interval for β_0

- $b_0 \pm t_c s\{b_0\}$
- where $t_c = t_{n-2}(1 - \frac{\alpha}{2})$, $1 - \alpha$ is the confidence level.

Significance Tests for β_0

- $\mathrm{H}_0 : \beta_0 = 0$ vs $\mathrm{H}_A : \beta_0 \neq 0$
- $t = \frac{b_0 - 0}{s\{b_0\}}$
- Reject H_0 if $|t| \geq t_c$, $t_c = t_{n-2}(1 - \frac{\alpha}{2})$
- p-value = Prob$(|T| > |t|)$, where $T \sim t_{n-2}$

Notes

- The normality of b_0 and b_1 follows from the fact that each is a linear combination of the Y_i, themselves each independent and normally distributed.
- For b_1, see KNNL, page 42.
- For b_0, try this as an exercise.

- Often the CI and significance test for β_0 is not of interest.
- If the ϵ_i are not normal but are approximately normal, then the CI's and significance tests are generally reasonable approximations.
- These procedures can easily be modified to produce one-sided confidence intervals and significance tests
- Because $\sigma^2(b_1) = \frac{\sigma^2}{\sum(X_i - \bar{X})^2}$, we can make this quantity small by making $\sum(X_i - \bar{X})^2$ large, i.e. by spreading out the X_i's.

SAS `proc reg`

Here is how to get the parameter estimates in SAS. (Still using `diamond.sas`). The option "`clb`" asks SAS to give you confidence limits for the parameter estimates b_0 and b_1.

```
proc reg data=diamonds;
    model price=weight/clb;
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-259.62591	17.31886	-14.99	<.0001	-294.48696	-224.76486
weight	1	3721.02485	81.78588	45.50	<.0001	3556.39841	3885.65129

Points to Remember

- What is the default value of α that we use in this class?
- What is the default confidence level that we use in this class?
- Suppose you could choose the X's. How would you choose them if you wanted a precise estimate of the slope? intercept? both?

Summary of Inference

- $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
- $\epsilon_i \sim N(0, \sigma^2)$ are independent, random errors

Parameter Estimators

$$\begin{aligned} \text{For } \beta_1 \quad &: \quad b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \\ \beta_0 \quad &: \quad b_0 = \bar{Y} - b_1\bar{X} \\ \sigma^2 \quad &: \quad s^2 = \frac{\sum(Y_i - b_0 - b_1X_i)^2}{n-2} \end{aligned}$$

95% Confidence Intervals for β_0 and β_1

- $b_1 \pm t_c s\{b_1\}$
- $b_0 \pm t_c s\{b_0\}$
- where $t_c = t_{n-2}(1 - \frac{\alpha}{2})$, the $100(1 - \frac{\alpha}{2})$ upper percentile of the t distribution with $n-2$ degrees of freedom

Significance Tests for β_0 and β_1

$$\begin{aligned} \text{H}_0 : \beta_0 = 0, \qquad & \text{H}_a : \beta_0 \neq 0 \\ & t = \tfrac{b_0}{s\{b_0\}} \sim t_{(n-2)} \text{ under H}_0 \\ \text{H}_0 : \beta_1 = 0, \qquad & \text{H}_a : \beta_1 \neq 0 \\ & t = \tfrac{b_1}{s\{b_1\}} \sim t_{(n-2)} \text{ under H}_0 \end{aligned}$$

Reject H_0 if the p-value is small (< 0.05).

KNNL Section 2.3 Power

The *power* of a significance test is the probability that the null hypothesis will be rejected when, in fact, it is false. This probability depends on the particular value of the parameter in the alternative space. When we do power calculations, we are trying to answer questions like the following:

> "Suppose that the parameter β_1 truly has the value 1.5, and we are going to collect a sample of a particular size n and with a particular SS_X. What is the probability that, based on our (not yet collected) data, we will reject H_0?"

Power for β_1

- $H_0 : \beta_1 = 0$, $H_a : \beta_1 \neq 0$
- $t = \frac{b_1}{s\{b_1\}}$
- $t_c = t_{n-2}(1 - \frac{\alpha}{2})$
- for $\alpha = 0.05$, we reject H_0, when $|t| \geq t_c$
- so we need to find $P(|t| \geq t_c)$ for arbitrary values of $\beta_1 \neq 0$
- when $\beta_1 = 0$, the calculation gives α (H_0 is true)
- $t \sim t_{n-2}(\delta)$ – noncentral t distribution: t-distribution not centered at 0
- $\delta = \frac{\beta_1}{\sigma\{b_1\}}$ is the noncentrality parameter: it represents on a "standardized" scale how far from true H_0 is (kind of like "effect size")
- We need to assume values for $\sigma^2(b_1) = \frac{\sigma^2}{\sum(X_i - \bar{X})^2}$ and n
- KNNL uses tables, see pages 50-51
- we will use SAS

Example of Power for β_1

- Response Variable: Work Hours
- Explanatory Variable: Lot Size
- See page 19 for details of this study, page 50-51 for details regarding power
- We assume $\sigma^2 = 2500$, $n = 25$, and $SS_X = 19800$, so we have $\sigma^2(b_1) = \frac{\sigma^2}{\sum(X_i - \bar{X})^2} = 0.1263$.
- Consider $\beta_1 = 1.5$.
- We now can calculate $\delta = \frac{\beta_1}{\sigma\{b_1\}}$
- with $t \sim t_{n-2}(\delta)$; we want to find $P(|t| \geq t_c)$
- We use a function that calculates the cumulative distribution function (cdf) for the noncentral t distribution.

See program `knnl050.sas` for the power calculations.

```
data a1;
   n=25; sig2=2500; ssx=19800; alpha=.05;
   sig2b1=sig2/ssx; df=n-2;
   beta1=1.5;
   delta=abs(beta1)/sqrt(sig2b1);
   tstar=tinv(1-alpha/2,df);
   power=1-probt(tstar,df,delta)+probt(-tstar,df,delta);
   output;

proc print data=a1;run;
```

```
 Obs     n    sig2     ssx     alpha     sig2b1    df    beta1     delta      tstar
  1     25    2500    19800     0.05    0.12626    23      1.5   4.22137    2.06866
   power
 0.98121
```

```
data a2;
   n=25; sig2=2500; ssx=19800; alpha=.05;
   sig2b1=sig2/ssx; df=n-2;
   do beta1=-2.0 to 2.0 by .05;
      delta=abs(beta1)/sqrt(sig2b1);
      tstar=tinv(1-alpha/2,df);
      power=1-probt(tstar,df,delta)+probt(-tstar,df,delta);
      output;
   end;

proc print data=a2;
run;

title1 'Power for the slope in simple linear regression';
symbol1 v=none i=join;
proc gplot data=a2; plot power*beta1; run;
```

Section 2.4: Estimation of $\mathbf{E}(Y_h)$

- $\mathrm{E}(Y_h) = \mu_h = \beta_0 + \beta_1 X_h$, the mean value of Y for the subpopulation with $X = X_h$.
- We will estimate $E(Y_h)$ with $\hat{Y}_h = \hat{\mu}_h = b_0 + b_1 X_h$.
- KNNL uses $\hat{Y}_h$ to denote this estimate; we will use the symbols $\hat{Y}_h = \hat{\mu}_h$ interchangeably.
- See equation (2.28) on page 52.

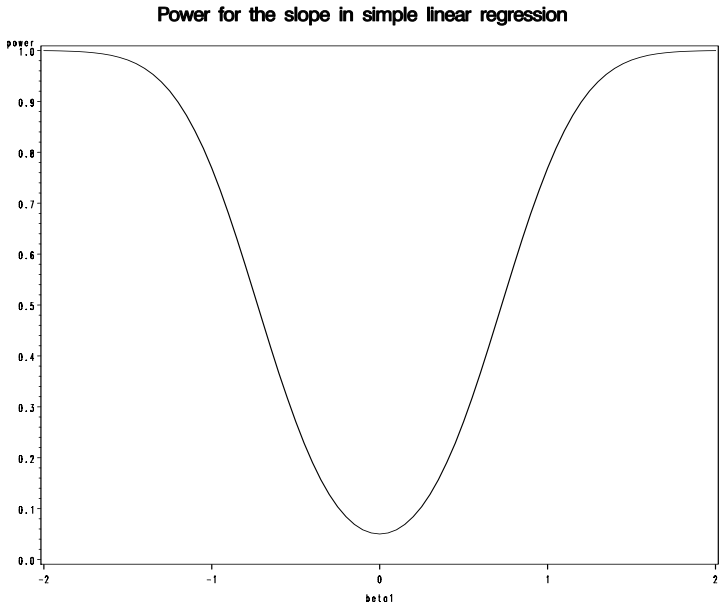


Figure 1: Power of the t-test for detecting different values of β_1

Theory for Estimation of $\mathbf{E}(Y_h)$

$\hat{Y}_h$ is normal with mean μ_h and variance $\sigma^2\{\hat{Y}_h\} = \sigma^2\left[\frac{1}{n} + \frac{(X_h-\bar{X})^2}{\sum(X_i-\bar{X})^2}\right]$.

- The normality is a consequence of the fact that $b_0 + b_1X_h$ is a linear combination of Y_i's.
- The variance has two components: one for the intercept and one for the slope. The variance associated with the slope depends on the distance $X_h - \bar{X}$. The estimation is more accurate near $\bar{X}$.
- See KNNL pages 52-55 for details.

Application of the Theory

We estimate $\sigma^2\{\bar{Y}_h\}$ with $s^2\{\hat{Y}_h\} = s^2\left[\frac{1}{n} + \frac{(X_h-\bar{X})^2}{\sum(X_i-\bar{X})^2}\right]$
It follows that $t = \frac{\hat{Y}_h - E(Y_h)}{s(\hat{\mu}_h)} \sim t_{n-2}$; proceed as usual.

95% Confidence Interval for $\mathbf{E}(Y_h)$

$\hat{Y}_h \pm t^c s\{\hat{Y}_h\}$, where $t^c = t_{n-2}(0.975)$.
NOTE: Significance tests can be performed for $\hat{Y}_h$, but they are rarely used in practice.

Example

See program `knnl054.sas` for the estimation of subpopulation means. The option "`clm`" to the `model` statement asks for confidence limits for the mean $\hat{Y}_h$.

```
data a1;
   infile 'H:\Stat512\Datasets\Ch01ta01.dat';
   input size hours;
data a2; size=65;  output;
         size=100; output;
data a3; set a1 a2;
proc print data=a3; run;
proc reg data=a3;
   model hours=size/clm;
   id size;
run;
```

Obs	size	Dep Var hours	Predicted Value	Std Error Mean Predict	95% CL Mean	
25	70	323.0000	312.2800	9.7647	292.0803	332.4797
26	65	.	294.4290	9.9176	273.9129	314.9451
27	100	.	419.3861	14.2723	389.8615	448.9106

Section 2.5: Prediction of $Y_{h(new)}$

We wish to construct an interval into which we predict the next observation (for a given X_h) will fall.

- The only difference (operationally) between this and $\mathrm{E}(Y_h)$ is that the variance is different.
- In prediction, we have two variance components: (1) variance associated with the estimation of the mean response $\hat{Y}_h$ and (2) variability in a single observation taken from the distribution with that mean.
- $Y_{h(new)} = \beta_0 + \beta_1 X_h + \epsilon$ is the value for a new observation with $X = X_h$.

We estimate $Y_{h(new)}$ starting with the predicted value $\hat{Y}_h$. This is the center of the confidence interval, just as it was for $\mathrm{E}(Y_h)$. However, the width of the CI is different because they have different variances.

$$\begin{aligned}
\mathrm{Var}(Y_{h(new)}) &= \mathrm{Var}(\hat{Y}_h) + \mathrm{Var}(\epsilon) \\
s^2\{pred\} &= s^2\{\hat{Y}_h\} + s^2 \\
s^2\{pred\} &= s^2\left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right] \\
\frac{Y_{h(new)} - \hat{Y}_h}{s\{pred\}} &\sim t_{n-2}
\end{aligned}$$

$s\{pred\}$ denotes the estimated standard deviation of a new observation with $X = X_h$. It takes into account variability in estimating the mean $\hat{Y}_h$ as well as variability in a single observation from a distribution with that mean.

Notes

The procedure can be modified for the mean of m observations at $X = X_h$ (see 2.39a on page 60). Standard error is affected by how far X_h is from $\bar{X}$ (see Figure 2.3). As was the case for the mean response, prediction is more accurate near $\bar{X}$.

See program `knnl069.sas` for the prediction interval example. The "`cli`" option to the `model` statements asks SAS to give confidence limits for an individual observation. (c.f. `clb` and `clm`)

```
data a1;
   infile 'H:\Stat512\Datasets\Ch01ta01.dat';
   input size hours;
data a2;
   size=65;  output;
   size=100; output;
data a3;
   set a1 a2;
proc reg data=a3;
   model hours=size/cli;
run;
```

Obs	size	Dep Var hours	Predicted Value	Std Error Mean Predict	95% CL Predict		Residual
25	70	323.0000	312.2800	9.7647	209.2811	415.2789	10.7200
26	65	.	294.4290	9.9176	191.3676	397.4904	.
27	100	.	419.3861	14.2723	314.1604	524.6117	.

Notes

- The standard error (`Std Error Mean Predict`) given in this output is the standard error of $\hat{Y}_h$, not $s\{pred\}$. (That's why the word *mean* is in there.) The `CL Predict` label tells you that the confidence interval is for the prediction of a new observation.
- The prediction interval for $Y_{h(new)}$ is wider than the confidence interval for $\hat{Y}_h$ because it has a larger variance.

Section 2.6: Working-Hotelling Confidence Bands for Entire Regression Line

- This is a confidence limit for the whole line at once, in contrast to the confidence interval for just one $\hat{Y}_h$ at a time.

- Regression line $b_0 + b_1 X_h$ describes $\mathrm{E}(Y_h)$ for a given X_h.
- We have 95% CI for $\mathrm{E}(Y_h) = \hat{Y}_h$ pertaining to specific X_h.
- We want a confidence band for all X_h – this is a confidence limit for the whole line at once, in contrast to the confidence interval for just one $\hat{Y}_h$ at a time.
- The confidence limit is given by $\hat{Y}_h \pm W s\{\hat{Y}_h\}$, where $W^2 = 2F_{2,n-2}(1-\alpha)$. Since we are doing all values of X_h at once, it will be wider at each X_h than CI's for individual X_h.
- The boundary values define a hyperbola.
- The theory for this comes from the joint confidence region for (β_0, β_1), which is an ellipse (see Stat 524).
- We are used to constructing CI's with t's, not W's. Can we fake it?
- We can find a new smaller α for t^c that would give the same result – kind of an "effective alpha" that takes into account that you are estimating the entire line.
- We find W^2 for our desired α, and then find the effective α^t to use with t^c that gives $W(\alpha) = t^c(\alpha^t)$.

Confidence Band for Regression Line

See program `knnl061.sas` for the regression line confidence band.

```
data a1;
   n=25; alpha=.10; dfn=2; dfd=n-2;
   w2=2*finv(1-alpha,dfn,dfd);
   w=sqrt(w2);  alphat=2*(1-probt(w,dfd));
   tstar=tinv(1-alphat/2, dfd); output;
proc print data=a1;run;
```

Note `1-probt(w, dfd)` gives the area under the t-distribution to the right of w. We have to double that to get the total area in both tails.

Obs	n	alpha	dfn	dfd	w2	w	alphat	tstar
1	25	0.1	2	23	5.09858	2.25800	0.033740	2.25800

```
data a2;
   infile 'H:\System\Desktop\CH01TA01.DAT';
   input size hours;
symbol1 v=circle i=rlclm97;
proc gplot data=a2;
   plot hours*size;
```

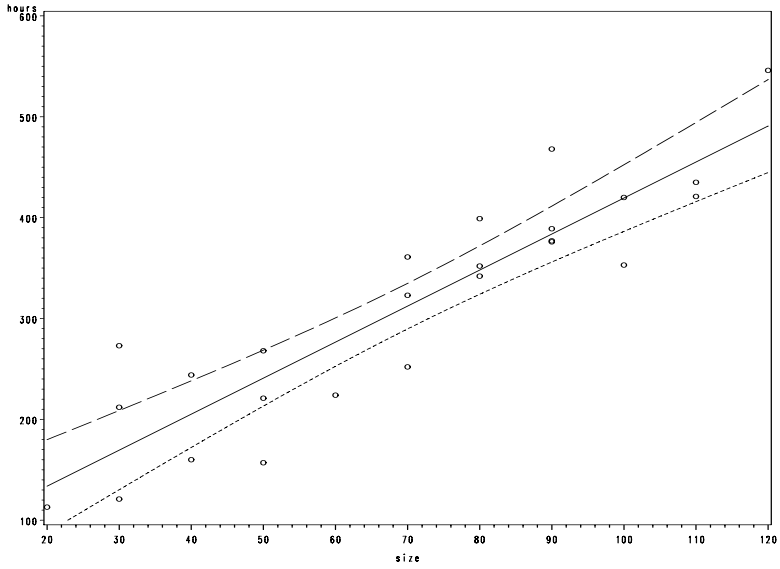


Figure 2: Working-Hotelling 95% Confidence Region for the Line

Estimation of $\mathbf{E}(Y_h)$ Compared to Prediction of Y_h

$$\begin{aligned} \hat{Y}_h &= b_0 + b_1 X_h \\ s^2\{\hat{Y}_h\} &= s^2\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right] \\ s^2\{pred\} &= s^2\left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right] \end{aligned}$$

See the program `knnl061x.sas` for the `clm` (mean) and `cli` (individual) plots.

```
data a1;
   infile 'H:\System\Desktop\CH01TA01.DAT';
   input size hours;
```

Confidence intervals:

```
symbol1 v=circle i=rlclm95;
proc gplot data=a1;
   plot hours*size;
```

Prediction Intervals:

```
symbol1 v=circle i=rlcli95;
proc gplot data=a1;
   plot hours*size;
run;
```

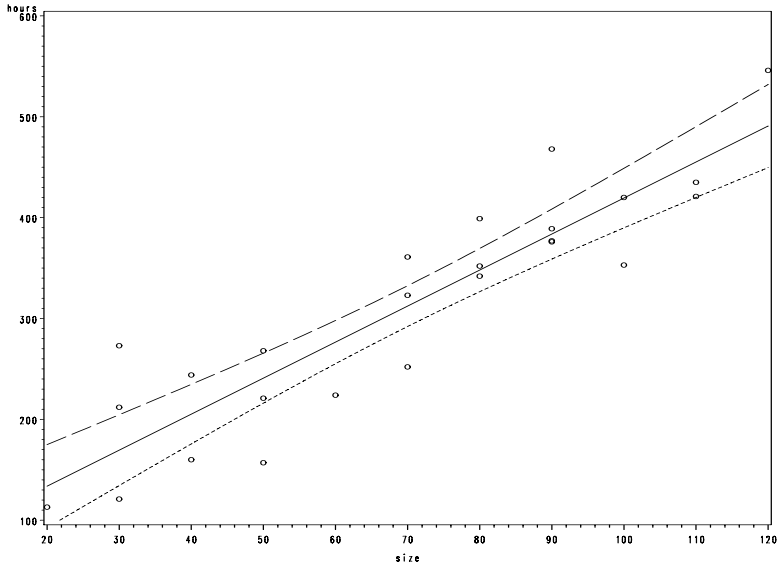


Figure 3: 95% Confidence Intervals for the Mean

Section 2.7: Analysis of Variance (ANOVA) Table

- Organizes results arithmetically
- Total sum of squares in Y is $SS_Y = \sum(Y_i - \bar{Y})^2$
- Partition this into two *sources*
  - Model (explained by regression)
  - Error (unexplained / residual)

$$\begin{aligned} Y_i - \bar{Y} &= (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}) \\ \sum(Y_i - \bar{Y})^2 &= \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \bar{Y})^2 \end{aligned}$$

(cross terms cancel: see page 65)

Total Sum of Squares

- Consider ignoring X_h to predict $E(Y_h)$. Then the best predictor would be the sample mean $\bar{Y}$.
- SST is the sum of squared deviations from this predictor $SST = SS_Y = \sum(Y_i - \bar{Y})^2$.
- The total degrees of freedom is $df_T = n - 1$.

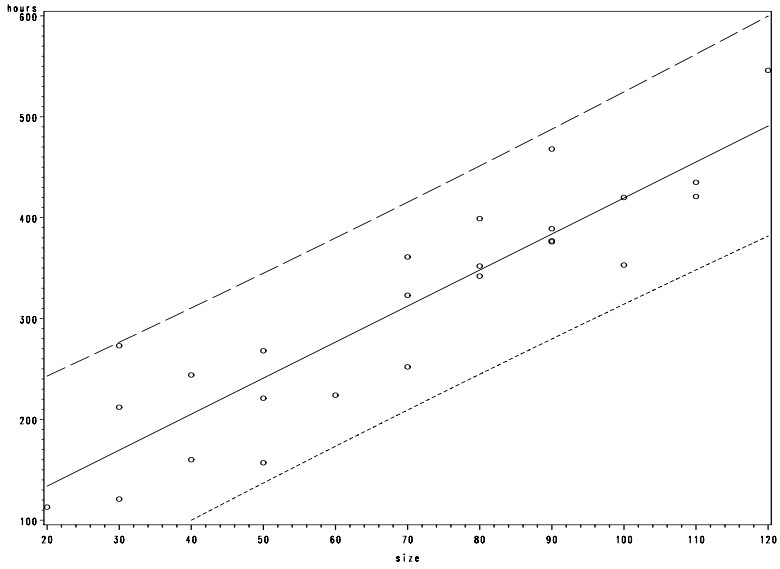


Figure 4: Prediction Intervals

- $MST = SST/df_T$
- MST is the usual estimate of the variance of Y if there are no explanatory variables, also known as $s^2\{Y\}$.
- SAS uses the term `Corrected Total` for this source. "Uncorrected" is $\sum Y_i^2$. The term "corrected" means that we subtract off the mean $\bar{Y}$ before squaring.

Model Sum of Squares

- $SSM = \sum(\hat{Y}_i - \bar{Y})^2$
- The model degrees of freedom is $df_M = 1$, since one parameter (slope) is estimated.
- $MSM = SSM/df_M$
- KNNL uses the word *regression* for what SAS calls *model*
- So SSR (KNNL) is the same as $SS\ Model$ (SAS). I prefer to use the terms SSM and df_M because R stands for regression, residual, and reduced (later), which I find confusing.

Error Sum of Squares

- $SSE = \sum(Y_i - \hat{Y}_i)^2$
- The error degrees of freedom is $df_E = n - 2$, since estimates have been made for both slope and intercept.

- $MSE = SSE/df_E$
- $MSE = s^2$ is an estimate of the variance of Y taking into account (or *conditioning on*) the explanatory variable(s)

ANOVA Table for SLR

Source	df	SS	MS
Model (Regression)	1	$\sum(\hat{Y}_i - \bar{Y})^2$	$\frac{SSM}{df_M}$
Error	$n-2$	$\sum(Y_i - \hat{Y}_i)^2$	$\frac{SSE}{df_E}$
Total	$n-1$	$\sum(Y_i - \bar{Y}_i)^2$	$\frac{SST}{df_T}$

Note about degrees of freedom
Occasionally, you will run across a reference to "degrees of freedom", without specifying whether this is model, error, or total. Sometimes is will be clear from context, and although that is sloppy usage, you can generally assume that if it is not specified, it means error degrees of freedom.

Expected Mean Squares

- MSM, MSE are random variables
- $\mathrm{E}(MSM) = \sigma^2 + \beta_1^2 SS_X$
- $\mathrm{E}(MSE) = \sigma^2$
- When $\mathrm{H}_0 : \beta_1 = 0$ is true, then $\mathrm{E}(MSM) = \mathrm{E}(MSE)$.
- This makes sense, since in that case, $\hat{Y}_i = \bar{Y}$.

F-test

- $F = MSM/MSE \sim F_{df_M, df_E} = F_{1,n-2}$
- See KNNL, pages 69-70
- When $\mathrm{H}_0 : \beta_1 = 0$ is false, MSM tends to be larger than MSE, so we would want to reject H_0 when F is large.
- Generally our decision rule is to reject the null hypothesis if

$$F \geq F_c = F_{df_M, df_E}(1-\alpha) = F_{1,n-2}(0.95)$$

- In practice, we use p-values (and reject H_0 if the p-value is less than α).

- Recall that $t = b_1/s(b_1)$ tests $H_0 : \beta_1 = 0$. It can be shown that $t^2_{df} = F_{1,df}$. The two approaches give the same p-value; they are really the same test.
- Aside: When $H_0 : \beta_1 = 0$ is false, F has a *noncentral* F distribution; this can be used to calculate power.

ANOVA Table

Source	df	SS	MS	F	p
Model	1	SSM	MSM	$\frac{MSM}{MSE}$	p
Error	$n-2$	SSE	MSE		
Total	$n-1$				

See the program `knnl067.sas` for the program used to generate the other output used in this lecture.

```
data a1;
   infile 'H:\System\Desktop\CH01TA01.DAT';
   input size hours;
proc reg data=a1;
   model hours=size;
run;
```

```
                                 Analysis of Variance
                              Sum of           Mean
Source                    DF       Squares         Square    F Value    Pr > F
Model                      1        252378         252378     105.88    <.0001
Error                     23         54825     2383.71562
Corrected Total           24        307203

                                 Parameter Estimates
                     Parameter       Standard
Variable     DF       Estimate          Error    t Value    Pr > |t|
Intercept     1       62.36586       26.17743       2.38      0.0259
size          1        3.57020        0.34697      10.29      <.0001
```

Note that $t^2 = 10.29^2 = 105.88 = F$.

Section 2.8 General Linear Test

- A different view of the same problem (testing $\beta_1 = 0$). It may seem redundant now, but the concept is extremely useful in MLR.
- We want to compare two models:

$$\begin{aligned} Y_i = &\quad \beta_0 + \beta_1 X_i + \epsilon_i \quad (\textit{full} \text{ model}) \\ Y_i = &\quad \beta_0 + \epsilon_i \quad (\textit{reduced} \text{ model}) \end{aligned}$$

Compare using the error sum of squares.

Let $SSE(F)$ be the SSE for the Full model, and let $SSE(R)$ be the SSE for the Reduced Model.

$$F = \frac{(SSE(R) - SSE(F))/(df_{E(R)} - df_{E(F)})}{SSE(F)/df_{E(F)}}$$

Compare to the critical value $F_c = F_{df_{E(R)}-df_{E(F)},df_{E(F)}}(1-\alpha)$ to test $\mathrm{H}_0 : \beta_1 = 0$ vs. $\mathrm{H}_a : \beta_1 \neq 0$.

Test in Simple Linear Regression

$$\begin{aligned}
SSE(R) &= \sum (Y_i - \bar{Y})^2 = SST \\
SSE(F) &= SST - SSM (\text{the usual } SSE) \\
df_{E(R)} = n-1, &\quad df_{E(F)} = n-2, \\
df_{E(R)} - df_{E(F)} &= 1 \\
F &= \frac{(SST - SSE)/1}{SSE/(n-2)} = \frac{MSM}{MSE} \text{ (Same test as before)}
\end{aligned}$$

This approach ("full" vs "reduced") is more general, and we will see it again in MLR.

Pearson Correlation

ρ is the usual correlation coefficient (estimated by r)

- It is a number between -1 and +1 that measures the strength of the *linear* relationship between two variables

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

- Notice that

$$\begin{aligned}
r &= b_1 \sqrt{\frac{\sum (X_i - \bar{X})^2}{\sum (Y_i - \bar{Y})^2}} \\
&= b_1 \frac{s_X}{s_Y}
\end{aligned}$$

  Test $\mathrm{H}_0 : \beta_1 = 0$ similar to $\mathrm{H}_0 : \rho = 0$.

R^2 and r^2

- R^2 is the ratio of explained and total variation: $R^2 = SSM/SST$
- r^2 is the square of the correlation between X and Y:

$$\begin{aligned} r^2 &= b_1^2 \left(\frac{\sum (X_i - \bar{X})^2}{\sum (Y_i - \bar{Y})^2} \right) \\ &= \frac{SSM}{SST} \end{aligned}$$

In SLR, $r^2 = R^2$ are the same thing.
However, in MLR they are different (there will be a different r for each X variable, but only one R^2).
R^2 is often multiplied by 100 and thereby expressed as a percent.
In MLR, we often use the adjusted R^2 which has been adjusted to account for the number of variables in the model (more in Chapter 6).

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	252378	252378	105.88	<.0001
Error	23	54825	2383		
C Total	24	307203			

R-Square 0.8215

$$\begin{aligned} &= SSM/SST = 1 - SSE/SST \\ &= 252378/307203 \end{aligned}$$

Adj R-sq 0.8138

$$\begin{aligned} &= 1 - MSE/MST \\ &= 1 - 2383/(307203/24) \end{aligned}$$