Lecture 3: Inference in SLR

Agenda:

- Measures of confidence in parameter estimates.
- Logic of hypothesis testing
- Tests of parameters in SLR
- Residual diagnostics and checking the validity of SLR.

Inference and SLR

- We have estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$. How good are they?
- Estimated parameters depend on random data – are themselves random.
- So how different might they be if we repeated the experiment?
- Variance of parameter *estimates* from repeated experiments is

$$\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{S_{XX}}, \ \sigma^2_{\hat{\beta}_0} = \sigma^2_{\hat{\beta}_1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right)$$

- Usually, we deal with $\sqrt{\sigma^2_{\hat{\beta}_1}}$: easier to understand.
- Estimated standard error is

$$\hat{\sigma}_{\hat{\beta}_1} = \frac{\hat{\sigma}}{\sqrt{S_{XX}}}$$

A Concrete Simulation

Meaning of variance of $\hat{\beta}_1$:

If we obtained new data 1000 times and recorded $\hat{\beta}_1$ each time, what would the variance be?

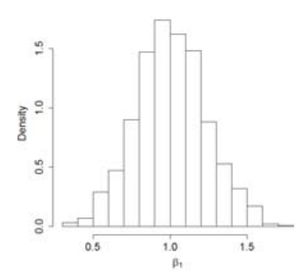
Let's do it!

```
coefmat = matrix(0,1000,2)              # Let's record the coefficients
predmat = matrix(0,1000,11)

for(sim in 1:1000){
  epsilon = rnorm(11,mean=0,sd=sigma)   # Only epsilon is random
  Y =  beta0 + beta1*X + epsilon        # Here is my response

  mod = lm(Y~X)                         # And I will re-fit the model
  coefmat[sim,] = mod$coefficients

  predmat[sim,] = mod$fit
}
```


Seeing The Result

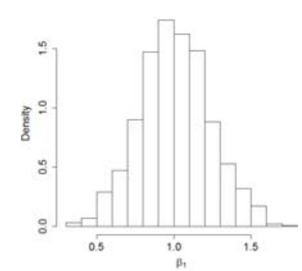
```
SXX = sum( (X-mean(X))^2 )
```

```
hist(coefmat[,1])
```




```
> var(coefmat[,1])
[1] 0.01916991
var.beta0 = var.beta1 * mean( X^2 )
[1] 0.01988636
```

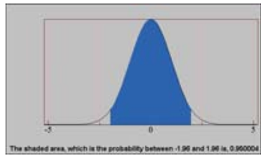
```
hist(coefmat[,2])
```




```
> var(coefmat[,2])
[1] 0.05255677
var.beta1 = sigma^2/SXX
[1] 0.05681818
```

Representing Uncertainty

Confidence intervals for β_1 are



$$\hat{\beta}_1 \pm z^{\alpha/2}\sqrt{\sigma^2_{\hat{\beta}_1}}$$

- $z^{\alpha/2}$ chosen so that a Normal random variable falls into $[-z^{\alpha/2},\ z^{\alpha/2}]$ $1-\alpha$% of the time.
- But we also estimate $\hat{\sigma}^2_{\hat{\beta}_1}$ and plug this in.

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}} \sim t_{n-2}$$

has heavier tails than Gaussian, because of uncertainty in $\hat{\sigma}^2_{\hat{\beta}_1}$.

- Use $t^{\alpha/2}_{n-2}$ instead of $z^{\alpha/2}$.

Computing Confidence Intervals in R

```
# First estimate the variance
> sig.hat = sum( mod$resid^2 )/9

# Then plug these into the variance family
> sd.beta1 = sqrt(sig.hat/SXX)
> sd.beta0 = sqrt(sd.beta1^2 * mean( X^2 ))

# Estimate plus and minus variance times critical value of t-distribution
# Intercept
> c( mod$coef[1] - qt(0.975,9)*sd.beta0,mod$coef[1] + qt(0.975,9)*sd.beta0 )
 -0.2797420   0.5021873

# Slope
> c( mod$coef[2] - qt(0.975,9)*sd.beta1,mod$coef[2] + qt(0.975,9)*sd.beta1 )
0.1371975 1.4588992

# Or much more easily use the following function
> confint(mod)
                 2.5 %    97.5 %
(Intercept) -0.2797420 0.5021873
X            0.1371975 1.4588992
```


Confidence Intervals for the Regression

So how certain are we about the average value of Y for a given X, $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$?

$$\mathrm{var}(\hat{Y}|X) = \sigma^2\left(\frac{1}{n} + \frac{(X-\bar{X})^2}{S_{XX}}\right)$$

(variance at $\bar{X}$ + correction as we get towards the edge of X.)

Interval for predicted value (expectation of Y) is

$$\hat{Y} \pm t^{\alpha/2}_{n-2}\hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{(X-\bar{X})^2}{S_{XX}}\right)}$$

Prediction for a Future Response

- Suppose we have a new X and want to know where Y will fall $(1-\alpha)$% of the time?
- We predict the mean of Y by

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

already a random quantity; has some uncertainty.

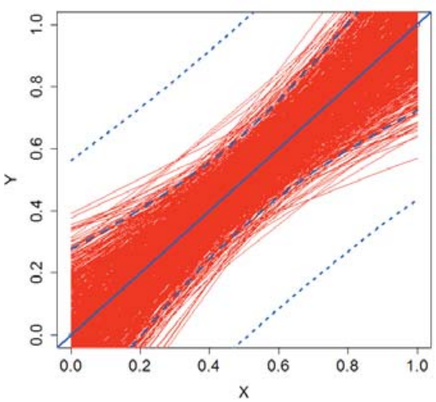
- But new Y also has error ϵ with variance σ^2.

$$\mathrm{var}(\hat{\beta}_0 + \hat{\beta}_1 X + \epsilon) = \sigma^2\left(1 + \frac{1}{n} + \frac{(X-\bar{X})^2}{S_{XX}}\right)$$

- Over-all standard error is

$$\mathrm{se}(Y - \hat{Y}) = \sigma\sqrt{\left(1 + \frac{1}{n} + \frac{(X-\bar{X})^2}{S_{XX}}\right)}$$

Graphically



- Confidence intervals (narrower) = where is the *average* response at each X.
- Prediction intervals (wider) = where might a new observation fall?
- You do not need to remember specific formulas for these.

Why do confidence intervals increase away from the mean?

- Suppose X is centered, $\bar{X} = 0$.
- Prediction formula is

$$\begin{aligned}\hat{Y} &= \hat{\beta}_0 + \hat{\beta}_1 X \\ &= (\beta_0 + e_{\beta_0}) + (\beta_1 + e_{\beta_1})X\end{aligned}$$

  (e_{β_0} = random error in estimate $\hat{\beta}_0$).
- Variance is

$$\mathrm{var}(\hat{Y}) = \sigma^2_{\hat{\beta}_0} + X^2\sigma^2_{\hat{\beta}_1}$$

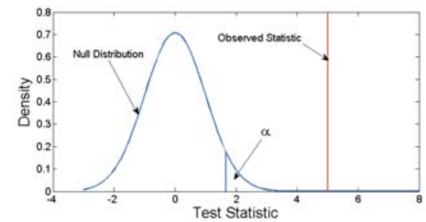
  because e_{β_0} independent of e_{β_1} (when $\bar{X} = 0$).
- So standard deviation is

$$\sigma_{\hat{Y}} = \sqrt{\sigma^2_{\hat{\beta}_0} + X^2\sigma^2_{\hat{\beta}_1}}$$

Testing Statistical Hypotheses

Reject $H_0 : \beta_1 = 0$ if

$$\frac{|\hat{\beta}_1 - 0|}{\sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}} > t^{\alpha/2}_{n-2}$$



- If the null hypothesis were true, the probability of the data producing a test statistic this extreme is less than 0.05.
- *p-value* The probability of seeing data in worse agreement with H_0 than those actually observed.

Some Thoughts About Hypothesis Tests

Most used, most misunderstood and least informative statistical procedure.

Can I tell that my data did not come from the null hypothesis?

This is a statement about the amount and accuracy of your data.

It is not:

- A statement about how useful/important your results are.
- An indication of the reliability of your estimates.

In science: minimum standard of evidence, but given much more weight than that.

Pop Question

Your test has a p-value of 0.089. You should:

1. Give up
2. Publish anyway (decide $p = 0.1$ is significant).
3. Find another statistician (or at least another test)
4. Add more data to your set.
5. None of the above

Aside: Mendel and Fisher

- Grygor Mendel's experimented with crossing strains of peas (long and short).
- Results laid foundation for genetic inheritance and the notion of dominant/recessive traits.
- R.A. Fisher (100 years later) showed that for his data $p > 0.95$:

  The results should only be this perfect 5% of the time.

- Most likely, Mendel kept collecting data until his results "looked" right.
- Lesson: you cannot include data that you used to design your experiment.

Hypothesis Tests in SLR

Why $H_0 : \beta_1 = 0$?

- Then $Y = \beta_0 + 0X + \epsilon = \beta_0 + \epsilon$
- X tells us nothing about Y; most common "null" case.

But there is no reason that we can't test $H_0 : \beta_1 = b$.

e.g. predicting a daughter's height from her mothers, consider $b = 1$ ($\Rightarrow$ height should be about the same as mothers.

In this case we reject if

$$\frac{|\hat{\beta}_1 - b|}{\sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}} > t_{n-2}^{\alpha/2}$$

i.e. $\hat{\beta}_1$ is too far away from the null value.

Hypothesis Tests and Confidence Intervals

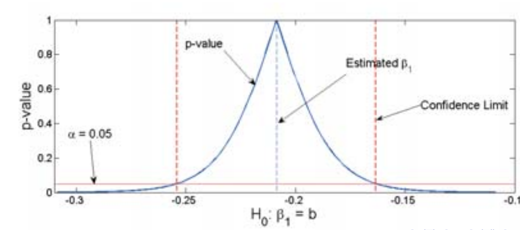
Hypothesis tests do help to define confidence intervals:

A confidence interval is all the values of a parameter that would not be rejected by a hypothesis test.

- For any value b test

$$H_0 : \beta_1 = b, \text{ versus } H_a : \beta_1 \neq b$$

- If the p-value is greater than α, b is in the confidence interval for β_1.



Hypothesis Tests and Confidence Intervals

Algebraically: reject $H_0 : \beta_1 = b$ if

$$\frac{|\hat{\beta}_1 - b|}{\sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}} > t_{n-2}^{\alpha/2}$$

otherwise we accept.

Re-arranging, we reject if

$$|\hat{\beta}_1 - b| > \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}\, t_{n-2}^{\alpha/2}$$

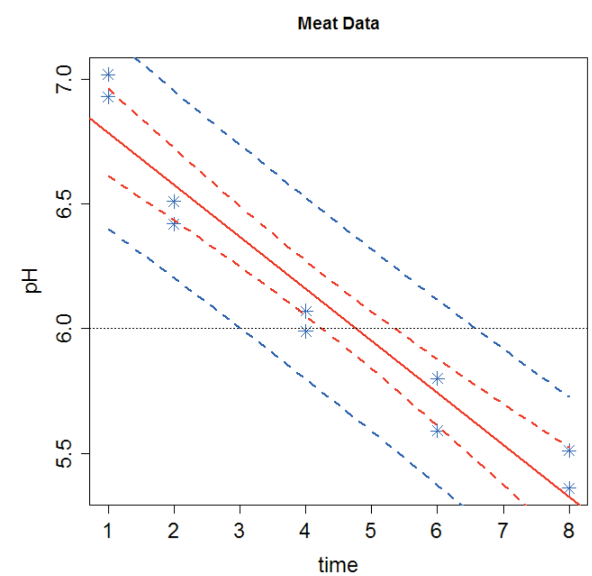
or accept if b is in the range

$$\left[\hat{\beta}_1 - \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}\, t_{n-2}^{\alpha/2},\ \hat{\beta}_1 + \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}\, t_{n-2}^{\alpha/2}\right]$$

Example Inference: pH Data

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.17174 -0.13870 -0.01805  0.12056  0.23220

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.99649    0.09691   72.20 1.51e-12 ***
time        -0.20869    0.01970  -10.59 5.51e-06 ***
---
Signif. codes:  '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1595 on 8 degrees of freedom
Multiple R-squared: 0.9335,     Adjusted R-squared: 0.9251
F-statistic: 112.2 on 1 and 8 DF,  p-value: 5.509e-06
```


Example Inference: pH Data



Linear Regression: Terminology Reminder

When we are interested in the value that Y might take at a particular X we refer to

Confidence Interval where are we 95% confident the mean value of Y is? (uncertainty in model parameters).

Prediction Interval where 95% of future Y's will fall, accounting for uncertainty in model parameters.

Calibration Interval what values of X could reasonably result in a particular value of Y?

Checking Assumptions

All of our inference works only if our model is correct

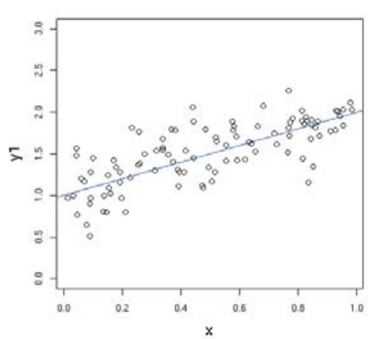
$$Y = \beta_0 + \beta_1 X + \epsilon$$

1. $E(\epsilon) = 0$.
   If we took many observations at a particular X, they should average to $E(Y) = \beta_0 + \beta_1 X$.
2. ϵ is normally distributed.
3. The variance of each ϵ is a constant σ^2.
4. The values ϵ associated with any two observations of Y are independent.

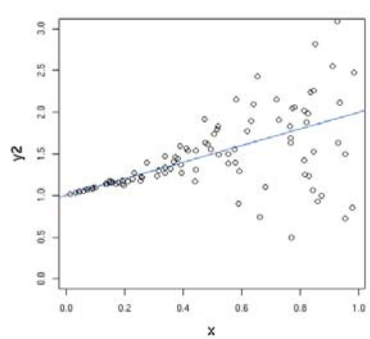
Mild violations are not important - often fine in practice.

Violations I

Assumptions Met

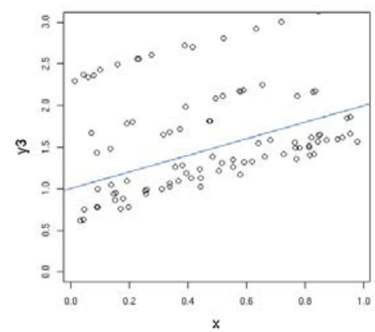


Non-Constant variance

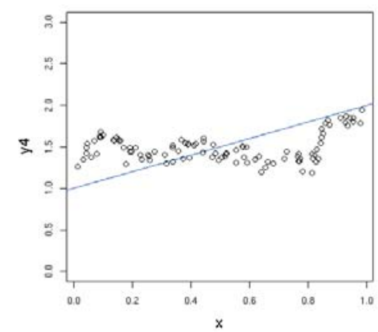


Violations II

Non-Normal



Not Independent



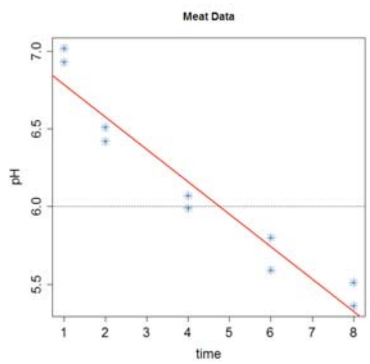
Doing Something About Model Violations

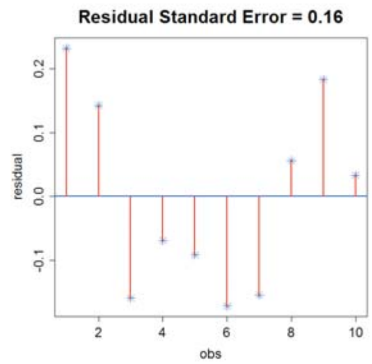
A few options:

1. Transform X
2. Transform Y
3. Transform both
4. Make the model more complicated

pH Data

Residuals show some noticeable patterns:

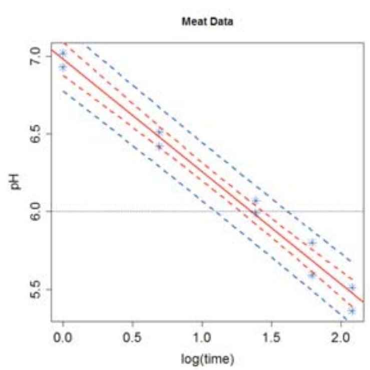




Try using $\log(X)$ instead.

Calibration Intervals

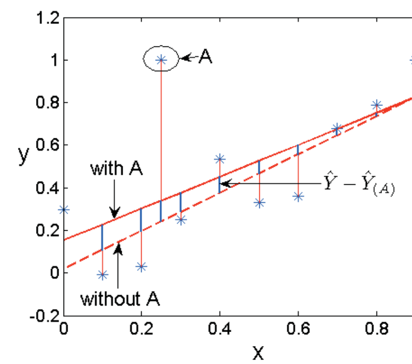
Using $\log(X)$ looks much better.



- *Calibration Interval*: what are the times in which a sample passes through pH 6.0?
- Find where prediction interval lines cross threshold.

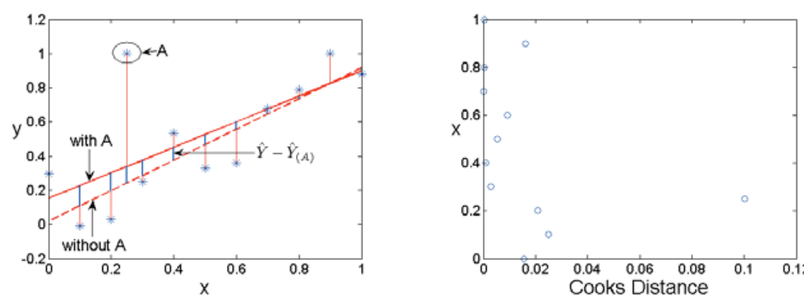
Violations: Outliers and Influence

May need to remove large outliers or influential points.



Notation: $\hat{Y}_{j(i)}$ prediction for Y_j when observation i is removed from data; similarly write $\hat{\beta}_{1(i)}$ or $\hat{\sigma}^2_{(i)}$.

Cooks Distance



- Cooks distance (use `cooks.distance()` in R): how much does the result change if I leave out one data point?

$$D_i = \frac{1}{\hat{\sigma}^2} \sum_j (\hat{Y}_j - \hat{Y}_{j(i)})^2$$

- Influence can be due to extreme values in Y or in X.

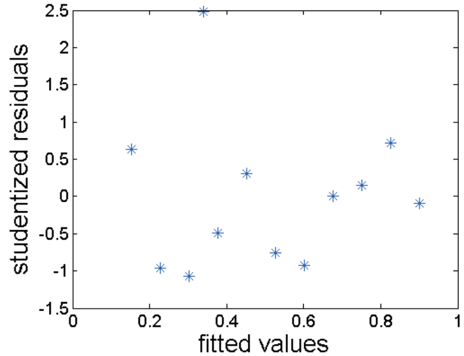
Violations: Distributional Assumptions

- Raw residuals: $Y_i - \hat{Y}_i$
- Standardized residuals: $\frac{(Y_i - \hat{Y}_i)}{\text{se}(Y_i - \hat{Y}_i)} =$

$$\frac{(Y_i - \hat{Y}_i)}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{S_{XX}}}}$$

- *Studentized* residuals:

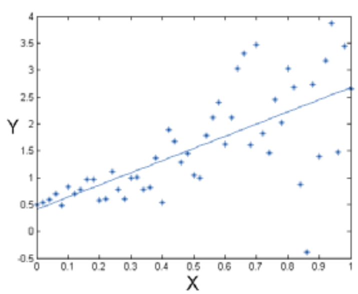
$$\frac{(Y_i - \hat{Y}_i)}{\hat{\sigma}_{(i)}\sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{S_{XX}}}}$$

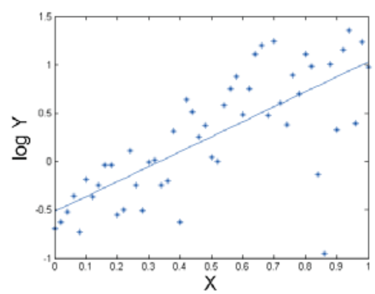


Studentized residuals (`studres()` in the package `MASS`) larger than 2 may be problematic.

Heteroskedasticity

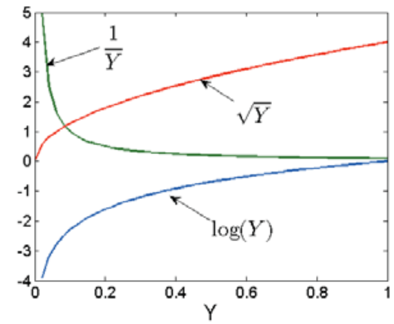
- Common problem is that measurement variance increases.
- Estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are still reasonable, but estimates of uncertainty are off.
- log transformation is common in this case





Common Transforms of Y

Frequently use $\log(Y)$, $\sqrt{Y}$ or $1/Y$



- Must have $Y > 0$ (add a constant so this is true)
- All tend to spread-out small values of Y, shrink large values.
- $\sqrt{Y}$ least severe, then $\log(Y)$ then $1/Y$.
- Try them and see which works best. You may also need to transform X.

How to do diagnostics

In assignments, what do you need to do for an analysis of residuals?

1. Plot studentized residuals and Cooks distances – look for outliers and influential points.
2. Plot residuals versus predicted values; look for heteroskedasticity and patterns of curvature.
3. Plot residuals versus covariate values (esp in multiple regression later) to look for curvature.

Indicate any action you take as a result of these plots.

What plots to report?

- Anything that indicates a violation of assumptions.
- If you believe all assumptions are met, provide a plot of influence and residuals versus predicted as an indication of fit.

On Transforms and Data Analysis

Why is it ok to take $\log(y_i)$ as a response?

$$\log(y_i) = \beta_0 + \beta_1 x_i + \epsilon_i$$

implies a nonlinear model for the y_i:

$$y_i = e^{\beta_0 + \beta_1 x_i + \epsilon} = e^{\beta_0} e^{\beta_1 x_i} e^{\epsilon_i}$$

- Each unit increase of x_i *multiplies* response by e^{β_1}.
- Errors also multiplicative – larger predicted value = wider spread.

Are you allowed to choose a model after you've seen the data?

Technically you shouldn't, but here it doesn't make much difference.

End of Simple Linear Regression

Review of

1. Linear models and estimation
2. Confidence intervals, standard errors and hypothesis tests
3. Assumptions and diagnostics

(more to come in Multiple Linear Regression)
Formulas:

- do not need to be memorized
- are helpful for understanding what you are doing

Next: some matrix algebra

Readings: Fox, 5, 6, 9.1