Statistical Inference for the Linear Regression Model

Brad Jones$^1$

$^1$Department of Political Science
University of California, Davis

January 22, 2010
Today: Variance Components and Inference
Useful to think about the “standard error” of the regression.

Quantity minimized: $\sum r_i^2$

Suppose we compute the variance of the residuals:

$$\text{var}(r) = \frac{\sum r_i^2}{n - k - 1}$$

Why $n - k - 1$? [These are the consumed degrees of freedom. Note again what must happen as $k \to n$.]
Variance Components from a Regression Model

Since the variance gives us the average squared deviation between the observed $Y$ and $\hat{Y}$, we take the square root:

$$s.e(r) = \sqrt{\frac{\sum r_i^2}{n - k - 1}}$$

This gives us the *standard error of the regression*.

...or the “average prediction error.” The smaller the residual component, the smaller the s.e. of the regression.

For the pedagogical regression using the calcount data, the s.e. is about 6.15.

Average prediction error in the model is about 6 percent.

The s.e. is scaled by $Y$ so it is easy to interpret.
It is clear the residual sums of squares are just half of the overall variance in the regression model.

If the RSS gives us “error variance” then what informs us about predictive improvement over and above the mean?

Recall that if $\beta_j = 0$ then $\beta_0 = \bar{Y}$.

Deviations in predictions, $\hat{Y}$ from the mean, $\bar{Y}$ tells the improvement gain in using $X$ to predict $Y$ over simply guess the mean every time.

The calcounty data:
Regression Model

[Image of a scatter plot with a regression line]
Variance Components from a Regression Model

- So the deviation $\hat{Y} - \bar{Y}$ gives the signed difference between predicted and the mean.
- Intuition: if the fitted values do not depart from the mean, $X$ is not doing a “good job” of predicting $Y$.
- Square and sum:
  \[ \sum_{i=1}^{n} (\hat{Y} - \bar{Y})^2 \]
- This is regression sum of squares (or sum of squares due to the regression). Fox refers to it as RegSS.
- It should be clear that the sum of RSS and RegSS accounts for the total variance in the model.
Variance Components from a Regression Model

- Total Sums of Squares (TSS):

\[ TSS = \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y} - \bar{Y})^2 \]

\[ = \sum (Y_i - \bar{Y})^2 \]

\[ = \text{RegSS} + \text{RSS} \]

- This shows us again that the regression function must pass through the point of averages.

- From these variance components, an intuitive fit measure emerges:

\[ R^2 = \frac{\sum (\hat{Y} - \bar{Y})^2}{\sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y} - \bar{Y})^2} \]

\[ = \frac{\text{RegSS}}{\text{TSS}} \]
Variance Components from a Regression Model

- In multiple regression, this is the squared multiple correlation or equivalently the square of $r_{\hat{Y}Y}$.
- Obama model: $RSS = 2119$; $RegSS = 7667$.
- $R^2 = 7667/(2119 + 7667) \approx .78$.
- In terms of the total variance in the model, about 78 percent is accounted for by the linear regression of votes on Prop. 8 support (n.b.).
- Issues with the $R^2$
Variance Components from a Regression Model

- The $R^2$ is nondecreasing in $X$ (why must this be the case?; be able to show this mathematically)
- Usually better to use “adjusted $R^2$:

$$\tilde{R}^2 = 1 - (1 - R^2) \frac{n - 1}{n - k - 1}$$

$$= 1 - \frac{RSS}{TSS} \times \frac{df_{TSS}}{df_{RSS}}$$

$$= 1 - \frac{RSS/n - k - 1}{TSS/n - 1}$$  \hfill (1)$$

- The degrees-of-freedom are used as a correction factor.
- In passing, note that $\tilde{R}^2$ can be negative (see next slide).
- If $R^2$ is nondecreasing, it is not very useful for model comparisons.
Adj. $R^2$ from Regression Model
Properties

- What are some properties of the model?
- Start with assumption that the error is not systematic implying:
  \[ E(\epsilon_i) = E(\epsilon_i \mid X) = 0 \]  \hfill (2)
- Linearity: \( E(Y) \) is a linear function of \( X_k \):
  \[
  \mu_i = E(Y \mid x_i) = E(\beta_0 + \beta_1x_i + \epsilon_i) \\
  = \beta_0 + \beta_1x_i + E(\epsilon_i) \\
  = \beta_0 + \beta_1x_i + 0 \\
  = \beta_0 + \beta_1x_i 
  \]  \hfill (3)
Homoskedasticity (constant variance):

\[
\text{var}(\epsilon_i \mid X_i) = E[\epsilon_i - E(\epsilon_i)]^2 \mid X_i] = E(\epsilon_i^2 \mid X_i) = \sigma^2
\]

This implies the variance of \( \epsilon_i \) for each \( X_i \) is equal to some positive constant (which is equal to \( \sigma^2 \)).

(Q: Since we usually do not observe \( \sigma^2 \) directly, what do you think is used as its estimator?)

When this assumption does not hold, we have a condition known as heteroskedasticity, and the variance is equal to \( \sigma_i^2 \).

Why might you care about this assumption?
Assumptions and Properties

- Independence assumptions:

\[
\text{cov}(\epsilon_i, \epsilon_j \mid X_i, X_j) = E[\epsilon_i - E(\epsilon_i) \mid X_i][\epsilon_j - E(\epsilon_j) \mid X_j] = E(\epsilon_i \mid X_i)(\epsilon_j \mid X_j)(\text{Why?}) = 0
\]  

- This implies that there is no correlation of the disturbances across the observations.
- With respect to sampling, the observations are sampled independently.
- Problem with time-series data: If \(\epsilon_{ti}\) and \(\epsilon_{t-1,i}\) are positively correlated, then \(Y\) is a function of not only \(X_i\) and \(\epsilon_{ti}\), but also \(\epsilon_{t-1,i}\).
Assumptions and Properties

- $X_k$ are fixed in repeated sampling.
- *Very strong assumption!*
- Experimental designs (in principle) will satisfy this condition...
- Unfortunately, we often work with observational data.
- This is why causal inference is difficult (or at least one reason why).
Covariance result: the covariance between $\epsilon_i$ and $X_i$ is 0:

$$\text{cov}(\epsilon_i, X_i) = E[\epsilon_i - E(\epsilon_i)][X_i - E(X_i)]$$

$$= E[\epsilon_i(X_i - E(X_i))](\text{Why?})$$

$$= E(\epsilon_i X_i) - E(X_i)E(\epsilon_i)$$

$$= E(\epsilon_i X_i)$$

$$= 0.$$  \hfill (6)

The import of it is to say that the unsystematic component (given by $\epsilon_i$) is not related to the systematic component (given by the $X_i$).
Assumptions and Properties

- $X$ is not a constant
- $n > k + 1$
- There is no *perfect* collinearity.

$$-1 < r_{X_i, X_j} < 1$$

i.e. one variable is not a linear combination of another variable such that the correlation between the variables is 1 (or -1).

- The model is correctly specified.
- Note we have said nothing about distributions at this point.
Inference for the Regression Model

- The regression assumptions give us a baseline to evaluate the adequacy of the model.
- But we need more precision in connecting our estimates back to the population parameters.
- $\hat{\beta}_k$ are derived from the sample data so there will be variability.
- We want to estimate the parameter’s precision or its reliability.
Inference for the Regression Model

- The usual measure of precision in statistics is the *standard error*.
- It is taken as the standard deviation of the sampling distribution of the estimator.
- Given that our estimator has a probability distribution (for a given sample size from a given population), it is natural to ask what the variance is *of* that distribution.
- This leads directly to the consideration of the variance of the estimators.
Inference for the Regression Model

- Bivariate model first (extension to the $n$-variable case is straightforward).
- The variance of the regression slope, $\hat{\beta}$ is given by

$$\text{var}(\hat{\beta}) = \frac{\sigma^2}{\sum(X_i - \bar{X})^2},$$

and the standard error is the square root of the variance, giving us

$$\text{se}(\hat{\beta}) = \frac{\sigma}{\sqrt{\sum(X_i - \bar{X})^2}}.$$
The variance of the regression intercept $\hat{\beta}_0$ is given by

$$\text{var}(\hat{\beta}_0) = \left( \frac{\sum X_i^2}{n \sum (X_i - \bar{X})^2} \right) \sigma^2,$$

and the standard error is given by

$$\text{se}(\hat{\beta}_0) = \sqrt{\left( \frac{\sum X_i^2}{n \sum (X_i - \bar{X})^2} \right) \sigma}.$$

In general, we will be more interested in the precision around the slope coefficient than the intercept.
We have seen $\sigma^2$ before: variance of the error component. Which is assumed to be constant. We usually will not directly observe this term (why?) but instead must estimate it directly from the data. What is the estimator we use?

Recall the “standard error of the estimate”:

$$s.e(r) = \sqrt{\frac{\sum r_i^2}{n - k - 1}}$$
Inference for the Regression Model

For the bivariate setting:

\[ \text{var}(r_i) = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum(r_i)^2}{n-2} = \frac{SSE}{n-2}, \]

which, after taking the square root, gives us

\[ \sqrt{\frac{SSE}{n-2}} \]

The square root is the s.e. of the estimate aka the “root mean square error.”

... and the MSE is?

\[ \sum(Y_i - \hat{Y}_i)^2 / n - 2 \]
Call: lm(formula = obamapercent ~ proportionforprop8)

Residuals:
   Min     1Q Median     3Q    Max
-8.795  -5.392  -0.669   4.117  19.317

Coefficients:            Estimate  Std. Error   t value  Pr(>|t|)
(Intercept)             102.3339     3.5434     28.9     <2e-16 ***
proportionforprop8     -0.8658      0.0608     -14.2     <2e-16 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 6.15 on 56 degrees of freedom
Multiple R-squared: 0.783,    Adjusted R-squared: 0.78
F-statistic: 203 on 1 and 56 DF,  p-value: <2e-16

> anova(regmod)
Analysis of Variance Table

Response: obamapercent

Df  Sum Sq  Mean Sq  F value Pr(>F)
proportionforprop8  1    7667    7667    203 <2e-16 ***
Residuals          56   2119      38
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
Inference for the Regression Model

- Model output:
  - RSS: 2119
  - RSS df: 56 (n-2)
  - MSE: 38
  - RegSS: 7667
  - RegSS df: 1 (k)
  - MSR: 7667 (RegSS/df)
  - TSS: RSS + RegSS (not shown)
  - TSS df: 57 (n-1)
  - RMSE: 6.15
Inference for the Regression Model

- More model output . . .
- Standard error of the regression coefficient:
  \[ \text{se}(\hat{\beta}) = \frac{\sigma}{\sqrt{\sum(X_i - \bar{X})^2}}. \]
- RMSE = 6.15 and \[ \sqrt{\sum(X_i - \bar{X})^2} \approx 10620 \]
- s.e.(\hat{\beta}) = 6.15/\sqrt{(10620)} \approx 0.06
- You can verify the s.e. of the constant on your own.
Inference for the Regression Model

- The extension to multiple regressors is straightforward (although like the least squares estimators, the presentation in scalar form gets ugly).
- Model with $\beta_0$, $\beta_1 \beta_2$ gives variances for the intercepts of:

\[
\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum (X_1 - \bar{X}_1)^2 (1 - r_{1,2}^2)},
\]

for $\hat{\beta}_1$,

\[
\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum (X_2 - \bar{X}_2)^2 (1 - r_{1,2}^2)},
\]

for $\hat{\beta}_2$

- The variance function for the constant is a bit ugly; consult Fox to see it.
Inference for the Regression Model

- Standard errors:

\[
\text{se}(\hat{\beta}_1) = \frac{\sigma}{\sqrt{\sum(X_1 - \bar{X}_1)^2(1 - r_{1,2}^2)}},
\]

for \( \hat{\beta}_1 \),

\[
\text{se}(\hat{\beta}_2) = \frac{\sigma}{\sqrt{\sum(X_2 - \bar{X}_2)^2(1 - r_{1,2}^2)}},
\]

for \( \hat{\beta}_2 \).

- The term, \( 1 - r_{1,2}^2 \), is known as the “auxiliary regression” where the \( r^2 \) is obtained by the regression of \( X_1 \) on \( X_2 \).

- Equivalently, the square root of the \( r^2 \) term gives you the correlation coefficient between \( X_2 \) and \( X_1 \).

- It is a measure of how collinear the covariates are.
Inference for the Regression Model

Fox uses this notation for the variance:

\[
\text{var}(\beta_k) = \frac{1}{1 - R^2_k} \times \frac{\sigma^2}{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}
\]

Obviously the same result will be gotten if you take the square root.

When \( k > 2 \), \( R^2_k \) is the squared multiple correlation from the regression of some \( X \) on all the other \( X_k \).

Note that the first factor is sometimes called the “variance inflation factor.”
Inference for the Regression Model

- We now have $\beta$ and we have the s.e.$(\beta)$
- What can we do in the way of inference?
- It’s time to overlay some distributional assumptions here.
- Conventional to assume normality.
Now understand, we’ve gotten pretty far without the normality assumption.

The only assumptions regarding $\epsilon_i$ have been:
- conditional mean is 0.
- variance is homoskedastic.
- 0 covariance with $x_i$.

But now we need to go beyond point estimation and enter the world of hypothesis testing. This requires us to say something about the distribution of the error term.

The regression coefficients are a linear function of $\epsilon_i$ (recall the least squares estimator).

Therefore, the sampling distribution of our least squares estimator will depend on the sampling distribution of $\epsilon$. 
Inference for the Regression Model

- The assumptions:

\[
E(\epsilon_i) = 0 \\
E(\epsilon_i^2) = \sigma^2 \\
E(\epsilon_i, \epsilon_j) = 0, \ i \neq j,
\]

which are the assumptions discussed earlier.

- But in addition to this, we’re going to assume the \( \epsilon \) is normally distributed.

- This leads to the following assumption:

\[
\epsilon_i \sim N(0, \sigma^2),
\]

which says that \( \epsilon \) is normally distributed with mean 0 and variance \( \sigma^2 \).
Inference for the Regression Model

- We can state this more explicitly by recognizing that for any two normally distributed random variables, a zero covariance between them implies independence.
- This means that if \( \epsilon_i \) and \( \epsilon_j \) have a 0 covariance (which they do by assumption), then they can be said to be independently distributed, leading to:

\[
\epsilon_i \sim \text{NID}(0, \sigma^2),
\]

where NID means \textit{normally and independently distributed}.
- Why assume the normal?
Inference for the Regression Model

▶ The principle reason is given by the central limit theorem.
▶ Under the CLT, if there are large number of iid random variables, the distribution of their sum will tend to a normal distribution as $n$ increases.
▶ So it is the central limit theorem that provides us with a strong justification to assume normality.
▶ An important result of the normal distribution is that any linear function of normally distributed random variables is itself, normally distributed.
▶ The regression coefficients are linear functions of $\epsilon_i$, so it must be the case that the sampling distributions for the regression estimates are also normally distributed.
Inference for the Regression Model

- For the multiple regression setting we now can say
  \[ \hat{\beta}_k \sim N(\beta_k, \sigma^2_{\hat{\beta}_k}) \]

- Additional results: under the normal distribution, we can define a distribution for our estimator \( \hat{\sigma}^2 \) as
  \[ \frac{(n - k - 1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-k-1}, \]
  where \( \chi^2 \) denotes the chi-square distribution with \( n - k - 1 \) degrees of freedom. Use of the \( \chi^2 \) statistic will allow us to compute confidence intervals around the estimator \( \sigma^2 \).

- Under the normal distribution, the regression estimates have minimum variance in the entire class of unbiased estimators.

- Finally if \( \epsilon_i \) is distributed normally, then \( Y_i \) itself must be normally distributed:
  \[ Y_i \sim N(\beta_0 + \beta_k X_i, \sigma^2). \]
Inference for the Regression Model

- Under the normality condition, we can specify \( Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \).
- A fundamental problem exists because \( \sigma \) is usually unknown. In its place, we estimate \( \sigma \) by using the standard error of \( \hat{\beta}_1 \).
- This gives rise to a \( t \)-statistic:

\[
t = \frac{\hat{\beta}_1 - \beta_1}{\text{s.e.}(\hat{\beta}_1)},
\]

which follows the \( t \) distribution with \( n - k - 1 \) degrees of freedom.

- Now since \( \frac{\hat{\beta}_1 - \beta_1}{\text{s.e.}(\hat{\beta}_1)} \sim t(n - k - 1) \), we can use the \( t \) distribution to establish a confidence interval:

\[
\Pr(-t_{\alpha/2} \leq t \leq t_{\alpha/2}) = 1 - \alpha.
\]

The term \( t_{\alpha/2} \) denotes our critical value and \( \alpha \) denotes the significance level. The level \( \alpha = .05 \) is common, but .01 or .10 levels are also commonly used as well.
Inference for the Regression Model

Substituting terms for the interval, we can rewrite the previous statement as

$$\Pr(-t_{\alpha/2} \leq \frac{\hat{\beta}_1 - \beta_1}{\text{s.e.}(\hat{\beta}_1)} \leq t_{\alpha/2}) = 1 - \alpha,$$

Rearranging, gives

$$\Pr[\hat{\beta}_1 - t_{\alpha/2}\text{s.e.}(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2}\text{s.e.}(\hat{\beta}_1)]$$

which is the 100(1 – \alpha) percent confidence interval.

Hence, \(\alpha = .05\) yields a 95 percent confidence interval:

$$\hat{\beta}_1 \pm t_{\alpha/2}\text{s.e.}(\hat{\beta}_1).$$
Inference for the Regression Model

- One important thing to note is the fact that we’re dividing the significance level by two.
- Note also that the width of the c.i. is proportional to the standard error of the coefficient.
- We can now see why the standard error is a measure of precision: it directly effects the interval in which the population parameters will probabilistically reside (over repeated samples).
- Simple tests-of-significance can now be done.
Inference for the Regression Model

- We can test the condition of the null hypothesis using a \( t \) statistic.
- The \( t \):

\[
t = \frac{\hat{\beta}_1 - \beta_1}{\text{s.e.}(\hat{\beta}_1)}.
\]

- We could state the null as

\[
H_0 : \beta_1 = 0,
\]

but we could easily specify \( \beta_1 \) under the null as being equal to any hypothetical value (i.e. 1, .5, 3.14, etc.).

- Define \( \beta_1^* \) as the value of \( \beta_1 \) under the null and rewrite \( t \) as

\[
t = \frac{\hat{\beta}_1 - \beta_1^*}{\text{s.e.}(\hat{\beta}_1)}.
\]

where \( \beta_1^* \) now reflects the condition of the null (and \( t_{\alpha/2} \) denote the critical \( t \) values).
Inference for the Regression Model

- We can utilize $p$ values to determine the probability of a $t$ value.
- In consulting a $t$ table, we can look up the appropriate degrees of freedom and derive the probability for a given $t$ value.
- Suppose we have 8 degrees of freedom and obtain a $t$ value of 2.306.
- In looking at the $t$ table, we see that the probability of obtaining a $t$ value of 2.306 or greater is 5 percent. This means that this result could have occurred by chance alone only about 5 percent of the time.
- This is all based on classical statistics.
Inference for the Regression Model

- Joint tests-of-significance are possible.
- Omnibus $F$-test: $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$
- Easy to compute using variance components:

$$F_0 = \frac{\text{RegSS}/k}{RSS/(n-k-1)} = \frac{\text{MSR}/\text{MSE}}{}$$

Consult an F-table for $k$ and $n-k-1$ degrees-of-freedom and you can obtain a $p$-value for the test.
Inference for the Regression Model

- Next week: matrix form of the model as well as further proofs of the assumptions.
- Model matrices and diagnostic methods.
- Review matrix algebra.