Lecture 15: Diagnostics and Inference for Multiple Linear Regression

36-401, Section B, Fall 2015

20 October 2015

Contents

1 Lighting Review of Multiple Linear Regression

In the multiple linear regression model, we assume that the response Y is a linear function of all the predictors, plus a constant, plus noise:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p + \epsilon \tag{1}$$

We make no assumptions about the (marginal or joint) distributions of the X_i, but we assume that $\mathbb{E}\left[\epsilon|X\right] = 0$, $\mathrm{Var}\left[\epsilon|X\right] = \sigma^2$, and that ϵ is uncorrelated across measurements. In matrix form, the model is

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \tag{2}$$

where $\mathbf{X}$ includes an initial column of all 1s.

When we add the Gaussian noise assumption, we are making all of the assumptions above, and further assuming that

$$\epsilon \sim MVN(\mathbf{0}, \sigma^2\mathbf{I}) \tag{3}$$

independently of $\mathbf{X}$.

The least squares estimate of the coefficients is

$$\widehat{\beta} = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y} \tag{4}$$

Under the Gaussian noise assumption, this is also the maximum likelihood estimate.

The fitted values (i.e., estimates of the conditional means at data points used to estimate the model) are given by the "hat" or "influence" matrix:

$$\widehat{\mathbf{m}} = \mathbf{x}\widehat{\beta} = \mathbf{H}\mathbf{y} \tag{5}$$

which is symmetric and idempotent. The residuals are given by

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y} \tag{6}$$

and $\mathbf{I} - \mathbf{H}$ is also symmetric and idempotent.

The expected mean squared error, which is the maximum likelihood estimate of σ^2, has a small negative bias:

$$\mathbb{E}\left[\hat{\sigma}^2\right] = \mathbb{E}\left[\frac{1}{n}\mathbf{e}^T\mathbf{e}\right] = \sigma^2\frac{n-p-1}{n} = \sigma^2\left(1 - \frac{p+1}{n}\right) \tag{7}$$

Since $\mathbf{H}\mathbf{x}\beta = \mathbf{x}\beta$, the residuals can also be written

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\epsilon \tag{8}$$

hence

$$\mathbb{E}\left[\mathbf{e}\right] = \mathbf{0} \tag{9}$$

and

$$\mathrm{Var}\left[\mathbf{e}\right] = \sigma^2(\mathbf{I} - \mathbf{H}) \tag{10}$$

Under the Gaussian noise assumption, $\widehat{\beta}$, $\widehat{\mathbf{m}}$ and $\mathbf{e}$ all have Gaussian distributions (about which more below, §3.1).

1.1 Point Predictions

Say that $\mathbf{x}'$ is the $m \times (p+1)$ dimensional matrix storing the values of the predictor variables at m points where we want to make predictions. (These may or may not include points we used to estimate the model, and m may be bigger, smaller or equal to n.) Similarly, let $\mathbf{Y}'$ be the $m \times 1$ matrix of random values of Y at those points. The point predictions we want to make are

$$\mathbb{E}\left[\mathbf{Y}'|\mathbf{X}' = \mathbf{x}'\right] = \mathbf{m}(\mathbf{x}') = \mathbf{x}'\beta \tag{11}$$

and we *estimate* this by

$$\widehat{\mathbf{m}}(\mathbf{x}') = \mathbf{x}'\widehat{\beta} \tag{12}$$

which is to say

$$\widehat{\mathbf{m}}(\mathbf{x}') = \mathbf{x}'(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y} \tag{13}$$

(It's easy to verify that when $\mathbf{x}' = \mathbf{x}$, this reduces to $\mathbf{H}\mathbf{y}$.)

Notice that the point predictions we make *anywhere* are always weighted sums (linear combinations) of the values of the response we happened to observe when we estimated the model. The weights just depend on the values of the predictors at the original data points, and at the points where we'll be making predictions — changing the responses doesn't change those weights.

2 Diagnostics for Multiple Linear Regression

Before proceeding to detailed statistical inference, we need to check our modeling assumptions, which means we need diagnostics.

2.1 Plot All the Things!

All of the plots we learned how to do for simple linear regression remain valuable:

1. *Plot the residuals against the predictors.* This now means p distinct plots, of course. Each of them should show a flat scatter of points around 0 (because $\mathbb{E}\left[\epsilon|X_i\right] = 0$), of roughly constant width (because $\mathrm{Var}\left[\epsilon|X_i\right] = \sigma^2$). Curvature or steps to this plot is a sign of potential nonlinearity, or of an omitted variable. Changing width is a potential sign of non-constant variance.

2. *Plot the squared residuals against the predictors.* Each of these p plots should show a flat scatter of points around $\hat{\sigma}^2$.

3. *Plot the residuals against the fitted values.* This is an extra plot, redundant when we only have one predictor (because the fitted values were linear in the predictor).

4. *Plot the squared residuals against the fitted values.*

5. *Plot the residuals against coordinates.* If observations are dated, time-stamped, or spatially located, plot the residuals as functions of time, or make a map. If there is a meaningful order to the observations, plot residuals from successive observations against each other. Because the ϵ_i are uncorrelated, all of these plots should show a lack of structure.

6. *Plot the residuals' distribution against a Gaussian.*

Out-of-sample predictions, with either random or deliberately selected testing sets, also remain valuable.

2.1.1 Collinearity

A linear dependence between two (or more) columns of the $\mathbf{x}$ matrix is called **collinearity**, and it keeps us from finding a solution by least squares. (In fact, collinearity at the population level makes the coefficients ill-defined, not just impossible to estimate.) Collinearity between a pair of variables will show up in a pairs plot as an exact straight line. Collinearity among more than two variables will not. For instance, if $X_3 = (X_1 + X_2)/2$, we can't include all three variables in a regression, but we'd not see that from any of the pairs.

Computationally, collinearity will show up in the form of the determinant of $\mathbf{x}^T\mathbf{x}$ being zero. Equivalently, the smallest eigenvalue of $\mathbf{x}^T\mathbf{x}$ will be zero. If `lm` is given a collinear set of predictor variables, it will sometimes give an error messages, but more often it will decide not to estimate one of the collinear variables, and return an `NA` for the offending coefficient.

We will return to the subject of collinearity in lecture 17.

```
# Minimal simulation of interactions
# Model: Y = sin(X1*X2) + noise
X <- matrix(runif(200),ncol=2)
Y <- sin(X[,1]*X[,2])+rnorm(200,0,0.1)
df <- data.frame(Y=Y, X1=X[,1], X2=X[,2])
missed.interact <- lm(Y ~ X1+X2, data=df)
```

FIGURE 1: *Simulating data from the model $Y = \sin X_1 X_2 + \epsilon$, to illustrate detecting interactions. Self-checks: what is the distribution of X_1 and X_2? what is σ^2?*

2.1.2 Interactions

Another possible complication for multiple regression which we didn't have with the simple regression model is that of *interactions* between variables. One of our assumptions is that each variable makes a distinct, additive contribution to the response, and the size of this contribution is completely insensitive to the contributions of other variables. If this is *not* true — if the relationship between Y and X_i changes depending on the value of another predictor, X_j — then there is an **interaction** between them.

There are several ways of looking for interactions. We will return to this subject in Lecture 19, but, for now, I'll stick with describing some diagnostic procedures.

Sub-divide and re-estimate The simplest thing to do, if you suspect an interaction between X_i and X_j, is to sub-divide the data based on the value of X_j, into two or more parts, and re-estimate the model. If there is no interaction, the coefficient on X_i should be the same, up to estimation error, in each part of the data. (That is, there should be no significant difference in the estimated coefficients.) While in principle straightforward, drawbacks to this include having to guess how to sub-divide the data (into two parts? three? more?), and at what values of X_j to make the cuts.

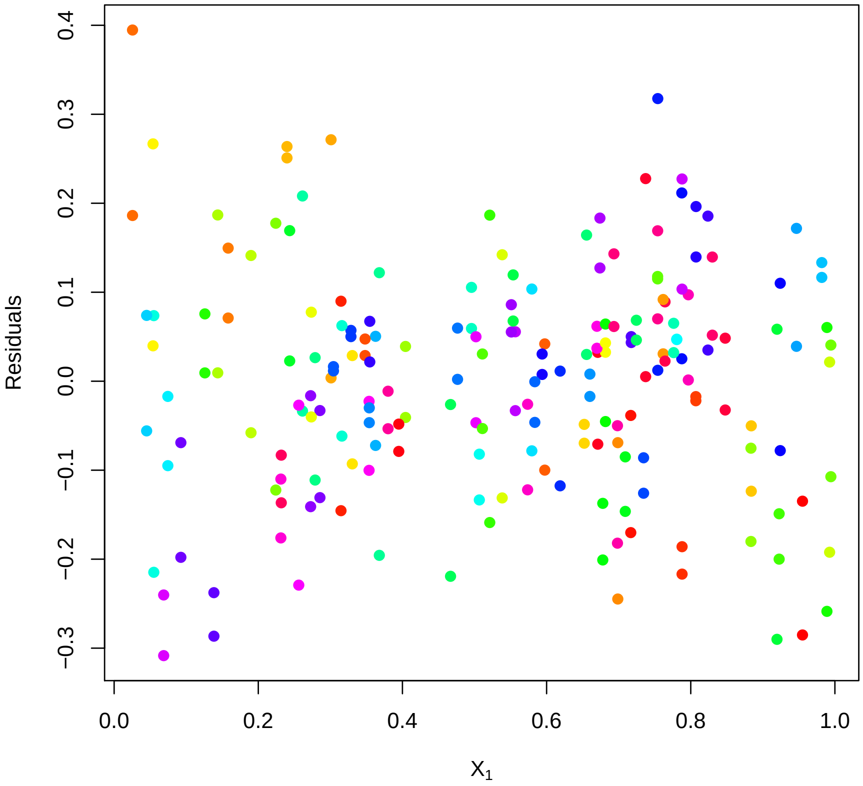
Scatterplot with color or symbols A more visual alternative is to plot the residuals against X_i, as usual, but to give each point a color which varies continuously with the value of X_j. In the absence of interactions, there should be no pattern to the colors. If there are interactions, however, we could predict what the residuals will be from knowing both variables, so we should tend to see similarly-colored regions in the plot.

If color is not available, a similar effect can be obtained by using different plotting symbols, corresponding to different values of X_j.

3D Plots

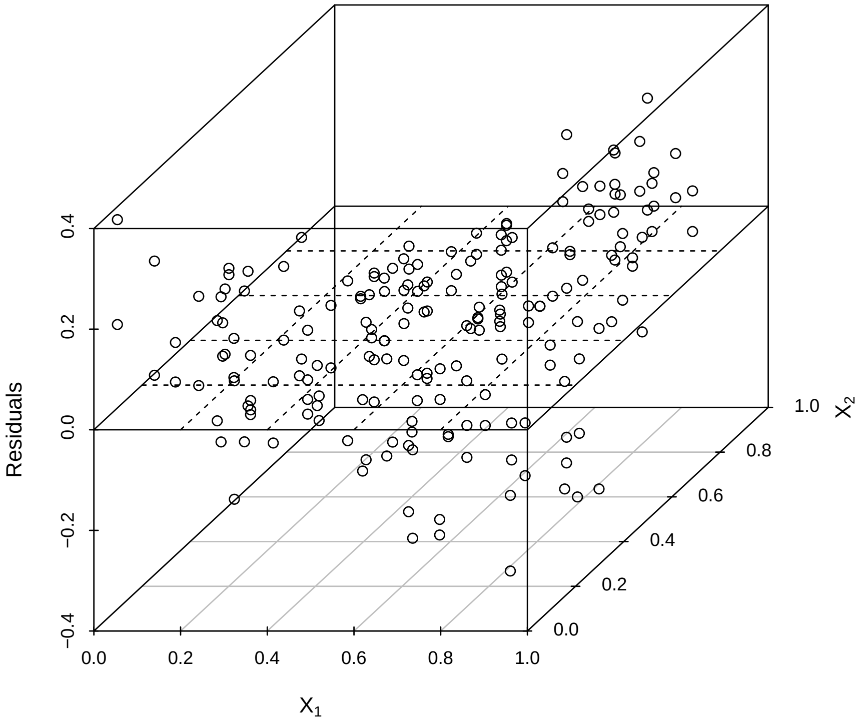
```
coefficients(summary(lm(Y~X1+X2, data=df, subset=which(df$X2 < median(df$X2)))))

##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -0.09257462 0.03356206 -2.758311 6.944694e-03
## X1           0.13546201 0.04375679  3.095794 2.566032e-03
## X2           0.50563560 0.08964297  5.640549 1.673144e-07

coefficients(summary(lm(Y~X1+X2, data=df, subset=which(df$X2 > median(df$X2)))))

##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -0.3024995 0.04246205 -7.123997 1.851872e-10
## X1           0.6441399 0.03553877 18.124990 6.582246e-33
## X2           0.4539010 0.05363622  8.462584 2.764062e-13
```

FIGURE 2: *Here we have sub-setted the data based on the value of the second predictor (dividing it, somewhat arbitrarily, at its median). Notice that the difference in the two coefficients for X_1 is much larger than their standard errors. Can you give a significance level for the difference in means?*



```
# Create a vector of gradually-changing colors, with one entry for
# each data point
the.colors <- rainbow(n=nrow(df))
# For each data point, see how it ranks according to X2, from smallest (1)
# to largest
the.ranks <- rank(df$X2)
# Plot residuals vs. X1, colored according to X2
  # Defining the color and rank vectors makes this next line a bit less
  # mysterious, but it's not necessary; this could all be a one-liner.
plot(df$X1, residuals(missed.interact), pch=19, col=the.colors[the.ranks],
     xlab=expression(X[1]), ylab="Residuals")
```

FIGURE 3: *Plotting residuals from the linear model against X_1, with the color of the point set by the value of X_2. Notice the clumping of points with similar colors: this means that knowing both X_1 and X_2 lets us predict the residual. Horizontal bands of the same color, on the other hand, would show that X_2 helped predict the residuals but X_1 did not, pointing to a mis-specification for the dependence of Y on X_2.*



```
library(scatterplot3d)
# Make a 3D scatterplot of residuals against the two predictor variables
s3d <-scatterplot3d(x=df$X1, y=df$X2, z=residuals(missed.interact),
                    tick.marks=TRUE, label.tick.marks=TRUE,
                    xlab=expression(X[1]), ylab=expression(X[2]),
                    zlab="Residuals")
# Add a plane with intercept 0 and both slopes also 0, for visual
# reference
s3d$plane3d(c(0,0,0), lty.box="solid")
```

FIGURE 4: *Residuals (vertical axis) vs. predictor variables. Notice that there are regions where the residuals are persistent positive or negative, but that these are defined by the value of* both *variables, not one or the other alone.*

2.2 Remedies

All of the remedies for model problems we discussed earlier, for the simple linear model, are still available to us.

Transform the response We can change the response variable from Y to $g(Y)$, in the hope that the assumptions of the linear-Gaussian model are more nearly satisfied for this new variable. That is, we hope that

$$g(Y) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon, \ \epsilon \sim N(0, \sigma 0^2 \tag{14}$$

The Box-Cox method, if you want to use it, will work as well as it did for simple linear models. Computationally, we'd just fill the $n \times 1$ response matrix with $[g(y_1) \ g(y_2) \ \ldots g(y_n)]^T$, and proceed as with any other multiple regression.

However, see the handout on transformations for cautions on interpretation after such transformations.

Transform the predictors We can also transform each of the predictors, making the model

$$Y = \beta_0 + \beta_1 f_1(X_1) + \ldots \beta_p f_p(X_p) + \epsilon, \ \epsilon \sim N(0, \sigma^2) \tag{15}$$

As the notation suggests, each X_i could be subject to a different transformation. Again, it's just a matter of what we put in the columns of the $\mathbf{x}$ matrix before solving for $\widehat{\beta}$. Again, see the handout on transformations for cautions on interpretations.

(A model of this form is called an **additive** model; in 402 we will look extensively at how they can be estimated, by automatically searching for near-optimal transformations.)

An alternative is to transform, not each predictor variable, but their linear combination:

$$Y = h\left(\beta_0 + \beta_1 X_1 + \ldots \beta_p X_p\right) + \epsilon, \ \epsilon \sim N(0, \sigma^2) \tag{16}$$

This is called a "single index" model, because there is only one combination of the predictors, the weighted sum $\beta_1 X_1 + \ldots \beta_p X_p$, which matters to the response. Notice that this is *not* the same model as the transform-Y model, even if $h = g^{-1}$, because of the different role of the noise.

Changing the variables used One option which is available to us with multiple regression is to add in new variables, or to remove ones we're already using. This should be done carefully, with an eye towards satisfying the model assumptions, rather than blindly increasing some score. We will discuss this extensively in lectures 20 and 26.

2.3 Plot *All* the Things?

There is one important caution about exuberant diagnostics plotting. This is that the more checks you run, the bigger the chance that you will find something which looks weird *just by chance*. If we were doing formal hypothesis tests, and insisted on a uniform false positive rate of α, then after running r tests, we'd expect to make $\approx r\alpha$ rejections, *even if all of our null hypotheses are true.* (Why?) If you are doing lots of diagnostic plots — say, 20 or 30 or more — it becomes a very good idea to do some randomization to see whether the *magnitude* of the bad-looking things you're seeing is about what you should be anticipating from one plot or another, even if everything was absolutely fine.

3 Inference for Multiple Linear Regression

Unless I say otherwise, all results in this section presume that all of the modeling assumptions, Gaussian noise very much included, are correct. Also, all distributions stated are conditional on $\mathbf{x}$.

3.1 Sampling Distributions

As in the simple linear model, the sampling distributions are the basis of all inference.

3.1.1 Gaussian Sampling Distributions

Gaussian distribution of coefficient estimators In the simple linear model, because the noise ϵ is Gaussian, and the coefficient estimators were linear in the noise, $\widehat{\beta}_0$ and $\widehat{\beta}_1$ were also Gaussian. This remains true in for Gaussian multiple linear regression models:

$$\begin{aligned} \widehat{\beta} &= (\mathbf{x}^T\mathbf{x})\mathbf{x}^T\mathbf{Y} && (17)\\ &= (\mathbf{x}^T\mathbf{x})\mathbf{x}^T(\mathbf{x}\beta+\epsilon) && (18)\\ &= \beta + (\mathbf{x}^T\mathbf{x})\mathbf{x}^T\epsilon && (19)\end{aligned}$$

Since $(\mathbf{x}^T\mathbf{x})\mathbf{x}^T\epsilon$ is a constant times a Gaussian, it is also a Gaussian; adding on another Gaussian still leaves us with a Gaussian. We saw the expectation and variance last time, so

$$\widehat{\beta} \sim MVN(\beta, \sigma^2(\mathbf{x}^T\mathbf{x})^{-1}) \tag{20}$$

It follows that

$$\widehat{\beta}_i \sim N\left(\beta_i, \sigma^2(\mathbf{x}^T\mathbf{x})^{-1}_{ii}\right) \tag{21}$$

Gaussian distribution of estimated conditional means The same logic applies to the estimates of conditional means. In §1.1, we saw that the estimated conditional means at new observations $\mathbf{x}'$ are given by

$$\widehat{\mathbf{m}}(\mathbf{x}') = \mathbf{x}'(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y} \tag{22}$$

so (Exercise)

$$\widehat{\mathbf{m}}(\mathbf{x}') \sim MVN(\mathbf{x}'\beta, \sigma^2\mathbf{x}'(\mathbf{x}^T\mathbf{x})^{-1}(\mathbf{x}')^T) \tag{23}$$

Gaussian distribution of fitted values Eq. 23 simplifies for the special case of the fitted values, i.e., the estimated conditional means on the original data.

$$\widehat{\mathbf{m}}(\mathbf{x}') \sim MVN(\mathbf{x}\beta, \sigma^2\mathbf{H}) \tag{24}$$

Gaussian distribution of residuals Similarly, the residuals have a Gaussian distribution:

$$\mathbf{e} \sim MVN(0, \sigma^2(\mathbf{I} - \mathbf{H}) \tag{25}$$

3.1.2 $\hat{\sigma}^2$ and Degrees of Freedom

The in-sample mean squared error $\hat{\sigma}^2 = n^{-1}\mathbf{e}^T\mathbf{e}$ has the distribution

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-(p+1)} \tag{26}$$

I won't prove this here, because it involves the same sort of tedious manipulations of Gaussians as I evaded in showing the special-case χ^2_{n-2} result for simple linear models. To give a hint of what's going on, though, I'll make two (related) observations.

Constraints on the residuals The residuals are not all independent of each other. In the case of the simple linear model, the fact that we estimated the model by least squares left us with two constraints, $\sum_i e_i = 0$ and $\sum_i e_i x_i = 0$. If we had only one constraint, that would let us fill in the last residual if we knew the other $n-1$ residuals. Having two constraints meant that knowing any $n-2$ residuals determined the remaining two.

We got those constraints from the normal or estimating equations, which in turn came from setting the derivative of the mean squared error (or of the log-likelihood) to zero. In the multiple regression model, when we set the derivative to zero, we get the matrix equation

$$\mathbf{x}^T(\mathbf{y} - \mathbf{x}\widehat{\beta}) = \mathbf{0} \tag{27}$$

But the term in parentheses is just $\mathbf{e}$, so the equation is

$$\mathbf{x}^T\mathbf{e} = \mathbf{0} \tag{28}$$

Expanding out the matrix multiplication,

$$\left[\begin{array}{c} \sum_i e_i \\ \sum_i x_{i1}e_i \\ \vdots \\ \sum_i x_{ip}e_i \end{array}\right] = \left[\begin{array}{c} 0 \\ 0 \\ \vdots \\ 0 \end{array}\right] \tag{29}$$

Thus the residuals are subject to $p+1$ linear constraints, and knowing any $n-(p+1)$ of them will fix the rest.

Geometric interpretation of constraints The vector of residuals $\mathbf{e}$ is a point in an n-dimensional space. As a random vector, without any constraints it could lie anywhere in that space, as, for instance, ϵ can. The constraints, however, for it to live in a lower-dimensional subspace, specifically, a space of dimension $n-(p+1)$.

Bias of $\hat{\sigma}^2$ As more of a formal manipulation, when we look at the expectation of $\hat{\sigma}^2$, we get

$$\begin{aligned}
\mathbb{E}\left[\hat{\sigma}^2\right] &= \frac{1}{n}\mathbb{E}\left[\mathbf{e}^T\mathbf{e}\right] && (30)\\
&= \frac{1}{n}\mathbb{E}\left[((\mathbf{I}-\mathbf{H})\mathbf{e})^T((\mathbf{I}-\mathbf{H})\mathbf{e})\right] && (31)\\
&= \frac{1}{n}\mathbb{E}\left[\mathbf{e}^T(\mathbf{I}^T-\mathbf{H}^T)(\mathbf{I}-\mathbf{H})\mathbf{e}\right] && (32)\\
&= \frac{1}{n}\mathbb{E}\left[\mathbf{e}^T(\mathbf{I}-\mathbf{H}-\mathbf{H}^T+\mathbf{H}^T\mathbf{H})\mathbf{e}\right] && (33)\\
&= \frac{1}{n}\mathbb{E}\left[\mathbf{e}^T(\mathbf{I}-\mathbf{H})\mathbf{e}\right] && (34)
\end{aligned}$$

using the easily-checked facts that $\mathbf{H}=\mathbf{H}^T$, and that $\mathbf{H}^2=\mathbf{H}$. We've therefore reduced the expectation to a quadratic form, and so (Lecture 13)

$$\begin{aligned}
\mathbb{E}\left[\hat{\sigma}^2\right] &= \frac{1}{n}\operatorname{tr}\left((\mathbf{I}-\mathbf{H})\operatorname{Var}\left[\mathbf{e}\right]\right) && (35)\\
&= \frac{1}{n}\operatorname{tr}\left((\mathbf{I}-\mathbf{H})\sigma^2(\mathbf{I}-\mathbf{H})\right) && (36)\\
&= \frac{\sigma^2}{n}\operatorname{tr}\left(\mathbf{I}-\mathbf{H}\right)^2 && (37)\\
&= \frac{\sigma^2}{n}\operatorname{tr}\left(\mathbf{I}-\mathbf{H}\right) && (38)
\end{aligned}$$

since we've just seen that $(\mathbf{I}-\mathbf{H})^2=(\mathbf{I}-\mathbf{H})$, and (Eq. 10) $\operatorname{Var}\left[\mathbf{e}\right]=\sigma^2(\mathbf{I}-\mathbf{H})$. Making one last push,

$$\mathbb{E}\left[\hat{\sigma}^2\right] = \frac{\sigma^2}{n}(n-p-1) \qquad (39)$$

since $\operatorname{tr}\mathbf{I}=n$ while (as you proved in the homework) $\operatorname{tr}H=p+1$.

3.2 t Distributions for Coefficient and Conditional Mean Estimators

From Eq. 21, it follows that

$$\frac{\hat{\beta}_i-\beta_i}{\sigma^2(\mathbf{x}^T\mathbf{x})^{-1}_{ii}} \sim N(0,1) \qquad (40)$$

This would be enough to let us do hypothesis tests and form confidence intervals, if only we knew σ^2, Since that's estimated itself, and $\hat{\sigma}^2$ has a distribution derived from a χ^2_{n-p-1}, we can go through the same arguments we did in the simple linear model case to get t distributions. Specifically,

$$\frac{\hat{\beta}_i - \beta_i}{\widehat{\mathrm{se}}\left[\hat{\beta}_i\right]} \sim t_{n-p-1} \tag{41}$$

The same applies to the estimated conditional means, and to the distribution of a new Y' around the estimated conditional mean (in a prediction interval). Thus, all the theory we did for parametric and predictive inference in the simple model carries over, just with a different number of degrees of freedom.

As with the simple model, $t_{n-p-1} \rightarrow N(0,1)$, so t statistics approach z statistics as the sample size grows.

3.3 What, Exactly, Is R Testing?

The `summary` function lists a p-value for each coefficient in a linear model. For each coefficient, say β_i, this is the p-value in testing the hypothesis that $\beta_i = 0$. It is important to be very clear about exactly what this means.

The hypothesis being tested is "Y is a linear function of all of the X_i, $i \in 1 : p$, with constant-variance, independent Gaussian noise, and it just so happens that $\beta_i = 0$". Since, as we saw in Lecture 14, the optimal coefficients for each predictor variable depend on which other variables are included in the model (through the off-diagonal terms in $(\mathbf{x}^T\mathbf{x})^{-1}$), this is a *very* specific hypothesis. In particular, whether the null hypothesis that $\beta_i = 0$ is true or not can easily depend on what other variables are included in the regression. What is really being checked here is, in ordinary language, something like "If you included all these other variables, would the model really fit *that* much better if you gave X_i a non-zero slope?"

3.3.1 Why, on Earth, Would You Want to Test That?

I am afraid that usually the answer is "you do not actually want to test that". You should ask yourself, carefully, whether it would really make any difference to you to know that the coefficient was precisely zero. (See Lecture 8, for some ideas about when that's worth testing and when it isn't.)

3.3.2 What Will Tend to Make a $\hat{\beta}$ Significant?

The t statistic for testing $\beta_i = 0$ is

$$\frac{\hat{\beta}_i}{\widehat{\mathrm{se}}\left[\hat{\beta}_i\right]} \tag{42}$$

We know that $\hat{\beta}_i$, being unbiased, will have a distribution centered on β_i, and the typical deviation away from that will in fact be about $\widehat{\mathrm{se}}\left[\hat{\beta}_i\right]$ in size, so we need to get a grip on that standard error.

From the theory above,

$$\widehat{\mathrm{se}}\left[\hat{\beta}_i\right] = \sqrt{\frac{\hat{\sigma}^2}{n}\left(\frac{1}{n}\mathbf{x}^T\mathbf{x}\right)^{-1}} \tag{43}$$

You showed in the homework (problem 3) that

$$\left(\frac{1}{n}\mathbf{x}^T\mathbf{x}\right) = \left[\begin{array}{ccccc} 1 & \overline{x_1} & \overline{x_2} & \ldots & \overline{x_p} \\ \overline{x_1} & \overline{x_1^2} & \overline{x_1 x_2} & \ldots & \overline{x_1 x_p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \overline{x_p} & \overline{x_1 x_p} & \ldots & \overline{x_p^2} & \end{array}\right] \tag{44}$$

What will happen when we invert this? You can check (Exercise **??**) that if $\overline{x_i x_j} = \bar{x}_i \bar{x}_j$ for all i, j, we'll get a diagonal matrix. Except for the very first entry on the diagonal (corresponding to the intercept), the entries will be inversely proportional to the variances of the predictor variables. If $\overline{x_i x_j} \neq \bar{x}_i \bar{x}_j$, the predictors are correlated, and this is going to increase the variance of their coefficients.

So, to sum up, four things control the standard error in $\hat{\beta}_i$: σ^2, the variance around the true regression function, since all standard errors are proportional to σ; n, since (all else being equal) all the standard errors are proportional to $1/\sqrt{n}$; the sample variance of X_i (since having data more widely spread on that axis makes it easier to find the slope); and the sample correlation between X_i and the other X_j (since strong correlations, positive or negative, make it harder to find their *specific* slopes).

What are the consequences?

1. Since, on any one data set, σ^2 and n are the same for all coefficients, the ones which are going to have the biggest test statistics, and so be "most significant", are the ones where (i) $|\beta_i|$ is large, (ii) the sample variance of X_i is large, and (iii) the sample correlation of X_i with other predictors is small.

2. The coefficients with the smallest p-values aren't necessarily the largest, let alone the most important; they may just be the most precisely measured.

3. Two people dealing with the same system, with precisely the same parameters and even the same n, can find different sets of coefficients to be significant, if their design matrices $\mathbf{x}$ differ. In fact, there need be no overlap in which coefficients are significant at all[1].

4. Adding or removing predictors will change which coefficients are significant, not just by changing the β_i, but also changing the standard error.

5. Holding all the parameters fixed and letting n grow, the t statistic will go off to $\pm\infty$, unless $\beta_i = 0$ exactly. Every non-zero coefficient eventually becomes significant at arbitrarily small levels.

[1] In this case, the natural thing to do would be to combine the data sets.

The same reasoning as in lecture 8 shows that p-values will tend to go to zero exponentially fast as n grows, unless of course $\beta_i = 0$.

3.3.3 Things It Would Be Very Stupid to Do, So Of Course You Would Never Even *Think* of Doing Them

- Saying "β_i wasn't significantly different from zero, so X_i doesn't matter for Y". After all, X_i could still be an important cause of Y, but we don't have enough data, or enough variance on X_i, or enough variance in X_i uncorrelated with other X's, to accurately estimate its slope. All of these would prevent us from saying that β_i was *significantly* different from 0, i.e., distinguishable from 0 with high reliability.

- Saying "β_i was significantly different from zero, so X_i really matters to Y". After all, any β_i which is not *exactly* zero can be made arbitrarily significant by increasing n and/or the sample variance of X_i. That is, its t statistic will go to $\pm\infty$, and the p-value as small as you have patience to make it.

- Deleting all the variables whose coefficients didn't have stars by them, and re-running the regression. After all, since it makes no sense to pretend that the statistically significant variables are the only ones which matter, limiting the regression to the statistically significant variables is even less sensible.

- Saying "all my coefficients are really significant, so the linear-Gaussian model must be right". After all, all the hypothesis tests on individual coefficients *presume* the linear Gaussian model, both in the null and in the alternative. The tests have no power to notice nonlinearities, non-constant noise variance, or non-Gaussian noise.

4 Further Reading

The marginal figures are taken from Allie Brosh, "This Is Why I'll Never Be an Adult", *Hyperbole and a Half*, 17 June 2010, without permission but with the deepest possible respect. If these notes do nothing beyond inspiring you to read one of the greatest moral psychologists of our age, they will have done more than many classes.

On a profoundly lower plane, **?** has one of the most sensible discussions of the uses and abuses of statistical inference in multiple regression I know of.

We will discuss additive models (where we automatically search for transformations of the predictors) extensively in 402 (**?**, ch. 9). Single-index models are used widely in econometrics; see, for instance, **?**.

5 Exercises

To think through or practice on, not to hand in.

1. Prove Eq. 10.
2. Prove Eq. 23
3. *What if all null hypotheses were true?* Draw a $\mathbf{Y}$ from a standard Gaussian distribution with 1000 observations. Draw $\mathbf{X}$ by setting $p = 100$, and giving each X_i a standard Gaussian distribution.

   (a) Regress Y on all 100 X's (plus an intercept). How many of the β_is are significant at the 10% level? At the 5% level? At the 1% level? What is the R^2? The adjusted R^2?

   (b) Re-run the regression using just the variables which are significant at the 5% level. Plot a histogram of the change in coefficient for each variable from the old regression to the new regression. How many variables are now significant at the 1% level? What is the R^2? The adjusted R^2?

   (This problem is inspired by an old example of David Freedman's.)

4. *Standard errors and correlations among the predictors* Assume that $p = 2$, so $n^{-1}\mathbf{x}^T\mathbf{x}$ is a 3×3 matrix.

   (a) Suppose that $\overline{x_1 x_2} = \bar{x_1}\bar{x_2}$, so there is no sample covariance between the two predictors. Find $(\frac{1}{n}\mathbf{x}^T\mathbf{x})^{-1}$ in terms of $\bar{x_1}$, $\bar{x_2}$, $\overline{x_1^2}$ and $\overline{x_2^2}$. Simplify, where possible, to eliminate second moments in favor of variances.

   (b) Give the general form of the inverse, $(\frac{1}{n}\mathbf{x}^T\mathbf{x})^{-1}$, without assuming $\overline{x_1 x_2} = \bar{x_1}\bar{x_2}$. How, qualitatively, do the variances of the slope estimates depend on the variances and covariances of the predictors?